SEDE ADMINISTRATIVA

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
FACULDADE DE CIÊNCIAS

The role of pancreatic enhancers disruption in the development of pancreatic cancer
Joana Maria Santiago Teixeira

U.PORTO

Joana Maria Santiago Teixeira . The role of pancreatic enhancers disruption
in the development of pancreatic cancer

D.ICBAS **2022**

U.PORTO

DOUTORAMENTO

BIOLOGIA MOLECULAR E CELULAR

The role of pancreatic enhancers disruption in the development of pancreatic cancer

Joana Maria Santiago Teixeira

D

2022

JOANA MARIA SANTIAGO TEIXEIRA

THE ROLE OF PANCREATIC ENHANCERS DISRUPTION IN THE DEVELOPMENT OF PANCREATIC CANCER

Tese de Candidatura ao grau de Doutor em Biologia Molecular e Celular;
Programa Doutoral da Universidade do Porto (Instituto de Ciências Biomédicas Abel Salazar e Faculdade de Ciências).

Orientador – Doutor José Carlos Ribeiro Bessa

Categoria – Investigadora Principal

Afiliação – Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S) da Universidade do Porto.

Life is not easy for any of us. But what of that?

We must have perseverance and above all confidence in ourselves.

We must believe that we are gifted for something, and that this thing must be attained.

Marie Curie

Financial support

This work was financially supported by:

• PhD fellowship SFRH/BD/126467/2016, Fundação para a Ciência e Tecnologia (FCT).

• Programa Graduado em Biologia Molecular e Celular, Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto (ICBAS-UP), Faculdade de Ciências da Universidade do Porto (FCUP).

• European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG-680156-ZPR).

• National Funds through FCT—Fundação para a Ciência e a Tecnologia, I.P., under the project UIDB/04293/2020.

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

NORTE 2020
PROGRAMA OPERACIONAL REGIONAL DO NORTE

PORTUGAL 2020

União Europeia
Fundo Europeu de Desenvolvimento Regional

erc
European Research Council
Established by the European Commission

U.PORTO
ICBAS | INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
SCHOOL OF MEDICINE AND BIOMEDICAL SCIENCES

U.PORTO
FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

DOCTORAL PROGRAM:
Molecular and Cell Biology
Programa Doutoral em Biologia Molecular e Celular

i3S INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

IBMC
INSTITUTO DE BIOLOGIA MOLECULAR E CELULAR
INSTITUTE FOR MOLECULAR AND CELL BIOLOGY

Acknowledgements/Agradecimentos

Chega então o fim a maior aventura da minha vida. Ri, chorei, desesperei, entusiasmei e sonhei. Acho que só quem passa por esta jornada consegue perceber o que se sente, e mesmo assim acho que cada um tem uma experiência diferente. Todos os caminhos longos só se fazem bem quando temos companhia. Além disso, Ciência nunca se faz sozinha, Ciência faz-se em grupo, Ciência faz-se partilhada. Sendo assim, não podia deixar passar em branco esta oportunidade, e agradecer a todos os meus companheiros de viagem.

Ao meu orientador, José Bessa, pela oportunidade de realizar este trabalho da melhor forma possível. Obrigada por todos os ensinamentos, obrigada por toda a paciência com a minha impaciência, mas sobretudo obrigada pela confiança no meu trabalho e nas minhas capacidades. Nos dias mais escuros lembrei-me sempre de um email que me mandou "Tenho high expectations de ti!!!". E isso sempre me guiou ao caminho certo.

Aos meus queridos colegas e amigos do VDR lab (passados e presentes), por de uma forma ou de outra, acrescentaram algo em mim e me fizeram ser melhor pessoa e cientista. Não posso nunca esquecer todos os momentos divertidos, jantares, lanches e festas. Obrigada, Joana Marques, Ana Eufrásio, Fábio Ferreira, Mafalda Galhardo, Pablo Cabezas-Sainz, Chiara Perrod, Marta Duque, João Amorim, Hugo Marcelino, Ana Gali, Renata Carriço, Isabel, Leonor Carvalho. Existem também outros membros do i3S que não posso deixar de mencionar e de agradecer por todo o apoio científico, técnico e pessoal. Obrigada, Mafalda Pereira, Nuno Padrão, Nádia Eusébio, Sofia Moreira, Elsa Logarinho, Monika Vilares, Sara Vaz, Ana Rita Araújo, Joana Macedo, Ana Rita Castro, Renata Freitas, Isabel Silveira, Joana Loureiro, Filipa Castro, Tânia Meireles, Paula Magalhães, Catarina Meireles, Emília Cardoso, Mafalda Rocha, Rob Mensink, Paula Sampaio, Maria Azevedo, André Torres, Mafalda Sousa, Nuno Mendes, Rossana Correia. Aos colaboradores, pertencentes a laboratórios externos ao i3s, deixo também o meu agradecimento. Obrigada, Panos Firbas, Juan Tena, Rafael Acemel e José Luís Goméz-Skarmeta.

Às mulheres da minha vida, Mãe e Avós, por me terem educado da melhor forma possível e me terem ensinado a lutar por tudo aquilo em que acredito. Aos homens da minha vida, Pai, Avôs, Tio, Irmão e Primo, por acreditarem sempre no meu valor.

Às duas irmãs que a vida me deu, Ana Morais (por estares sempre aí, por aceitares tudo que sou, por compreenderes tudo que sou, por todos os beliscões de realidade, mas sobretudo

por me mostrares o poder da amizade-família) e Mariana Barros (por me ensinares o poder da liberdade e da frontalidade, o poder do orgulho no que somos e no que queremos ser. Obrigada, também pelo “colo” bom quando acho que este mundo é insano).

Aos meus amigos-família, por todo carinho, força e ajuda. João Mesquita, Nuno Magalhães, Daniela Carvalho, Sara Azevedo, Pedro Matos, Raul Reis, Elsa Rocha.

Aos juniores, por serem uns amigos do caraças, por me terem acolhido no nosso círculo “folclórico” e por terem salvo a minha sanidade durante esta jornada: Helena Manzoni, Lisete Moreira, Daniela Monteiro, Bruno Silva, Fábio Oliveira, Joel Pinto, Patrícia Raquel, Daniel Amorim, Marina Rosa, Luís França, Pedro Silva, Raquel Baleixo, Sérgio Fernandes e Alexandra Barros.

E por fim, mas não menos importante, ao Zé Manel e ao RPG, pela ajuda no meu crescimento pessoal.

“Não sei quantas almas tenho.
Cada momento mudei.
Continuamente me estranho.
Nunca me vi nem achei.
De tanto ser, só tenho alma.
Quem tem alma não tem calma.
Quem vê é só o que vê,
Quem sente não é quem é,
Atento ao que sou e vejo,
Torno-me eles e não eu.
Cada meu sonho ou desejo
É do que nasce e não meu.

Sou minha própria paisagem
Assisto à minha passagem,
Diverso, móbil e só,
Não sei sentir-me onde estou.
Por isso, alheio, vou lendo
Como páginas, meu ser
O que segue não prevendo,
O que passou a esquecer.
Noto à margem do que li
O que julguei que senti.
Releio e digo: «Fui eu?»
Deus sabe, porque o escreveu.”

Fernando Pessoa

Dedico esta tese ao meu maior fã, á pessoa que me ensinou a forma mais pura e desinteressada de amor. A ti, Avô Firmino.

Publication list

Original research article, published

Bordeira-Carriço, R.*, **Teixeira, J.***, Duque, M.*, Galhardo, M., Ribeiro, D., Dominguez-Acemel, R., Firbas, P. N., Tena, J. J., Eufrasio, A., Marques, J., Ferreira, F.J., Freitas, T., Carneiro, F., Goméz-Skarmeta, J-L., Bessa, J. (2022) Multidimensional chromatin profiling of zebrafish pancreas to uncover and investigate disease-relevant enhancers. Nat Commun. 13:1945.
https://doi.org/10.1038/s41467-022-29551-7
*Equal contribution as first author

Original research article, published

Amorim, J.P., Macedo-Gali, A., Marcelino, H., Bordeira-Carriço, R., Naranjo, S., Rivero-Gil, S., **Teixeira, J.,** el al. (2020) A Conserved Notochord Enhancer Controls Pancreas Development in Vertebrates. Cell Rep. 32:107862.
doi: 10.1016/j.celrep.2020.107862

Original research article, in preparation

Teixeira, J., Galhardo, M., Bessa, J*. Nucleotide polymorphisms associated to increased risk of pancreatic cancer modulate pancreatic enhancer's function.

Original research article, in preparation

Ferreira, F.J., **Teixeira, J.**, Galhardo, M., Logarinho, E., Bessa, J. A master cis-regulatory element in the FOXM1/RHNO1/TEAD4 landscape drives cellular aging.

Original research article, in preparation

Ferreira, F.J., **Teixeira, J.**, Galhardo, M., Logarinho, E., Bessa, J. FOXM1 regulates age dependent chromatin accessibility.

Authorization of the Journal “Nature Communications” for inclusion in the Doctoral Thesis:

“Nature Communications” is a peer reviewed academic journal, part of the Springer Nature Group. Springer Nature’s Copyright Policies state that: “The author of articles published by Springer Nature do not usually need to seek permission for re-use of their material as long as the journal is credited with initial publication. Ownership of copyright in in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights: To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s). (...) Authors have the right to reuse their article’s Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.”

See full policy information in:
https://www.nature.com/nature-research/reprints-and-permissions/permissions-requests
https://www.springer.com/gp/rights-permissions/obtaining-permissions/882

Resumo

O cancro pancreático (CP) é uma doença maligna agressiva do pâncreas, que representa a sétima causa de morte relacionada com cancro no mundo. Estudos de associação genómica tem descoberto polimorfismos de nucleotídeo simples (PNS) associados a CP, muitos deles localizados em regiões não-codificantes do genoma. Os elementos cis-regulatórios (ECR) são elementos maioritariamente localizados em regiões não-codificantes do genoma, sendo uma das entidades funcionais mais sujeitas a serem afetadas pelos PNS associados a CP. Estas observações sugerem que os PNSs associados a CP podem interromper os ECRs pancreáticos, levando a desregulação transcricional dos genes-alvo, contribuindo para um aumento do risco de desenvolver CP. Até que ponto os PNSs associados ao CP podem interrompem a função dos ECRs ainda não foi totalmente esclarecida. Adicionalmente, o CP é uma doença complexa caracterizada por potenciais origens celulares múltiplas, progressão de diferentes estados celulares e interação com complexos ambientes celulares. Assim, para abordar adequadamente a interrupção dos ECRs como contribuintes para o desenvolvimento de CP, é imperativo o uso de modelos animais como o peixe-zebra, que contêm um pâncreas homologo aos dos humanos. Contudo, muitos dos ECRs pancreáticos e seus genes-alvo em peixe-zebra são desconhecidos.

No capítulo II desta tese, nós analisamos as modificações de histonas, de transcrição, e de acessibilidade e interação da cromatina, para identificar os ECRs pancreáticos em peixe-zebra e os seus equivalentes funcionais em humano, descobrindo sequências associadas a doenças entre espécies, incluindo ECRs potencialmente associadas a PC. Adicionalmente, através da realização de deleções genómicas em um desses ECRs identificados, usando uma linha celular ductal humana, nós conseguimos demostrar a sua habilidade para interromper a expressão do gene *ARID1A*, um gene supressor de tumores. No capítulo III, nós realizamos deleções no ECR equivalente funcional em peixe-zebra do gene *arid1ab* e começamos a avaliar o seu impacto fenotípico. Finalmente, no capítulo IV, nós combinamos diversos recursos genómicos e demostramos que a localização genómica dos PNSs associados a CP está enriquecida em ECRs pancreáticos, demostrando que alguns destes alelos podem ter impacto na regulação dos ECRs em que estão presentes.

No geral, as descobertas reportadas nesta tese doutoral suportam a hipótese que a interrupção dos ECRs pancreáticos pode contribuir para o desenvolvimento de CP, expandindo o nosso conhecimento sobre as os jogadores genéticos implicados e genes alvo-associados.

Abstract

Pancreatic cancer (PC) is an aggressive malignant disease of the pancreas, representing the seventh leading cause of cancer-related deaths worldwide. Genome-wide association studies have uncovered single nucleotide polymorphisms (SNPs) associated with PC, many localized in the non-coding genome. Transcriptional cis-regulatory elements (CREs) are mostly localized in the non-coding genome, being one of the functional entities of the DNA to be potentially affected by PC associated SNPs. These observations suggest that PC associated SNPs might disrupt pancreatic CREs, leading to transcriptional dysregulation of target genes, contributing to an increased risk of PC development. To what extent PC associated SNPs disrupt CREs function is yet to be fully explored. Also, PC is a complex disease characterized by potential multiple cellular origins, a progression of different cellular states and interaction with complex environmental cellular contexts. Therefore, to properly address CREs disruption as contributors for PC development, it is imperative the use of animal models such as the zebrafish, that contains a pancreas homologue to its human counterpart. However, many of the zebrafish pancreatic CREs and their target genes remain unknown.

In the Chapter II of this thesis, we have analysed histone modifications, transcription, chromatin accessibility and interactions, to identify zebrafish pancreas CREs and their human functional equivalents, uncovering disease-associated sequences across species, including PC potentially associated CREs. Importantly, by performing genomic deletions in one of these CREs in pancreatic human cell lines, we have demonstrated their ability to disrupt the expression of ARID1A, a tumour suppressor gene. In the Chapter III, we have performed deletions in the zebrafish functional equivalent CRE of the gene *arid1ab*, and we started to evaluate its phenotypic impact. Finally, in the Chapter IV, we have combined available genome-wide resources and demonstrated that the genomic location of PC risk SNPs is enriched in pancreatic CREs, also demonstrating that some PC associated alleles impact in the regulatory output of the overlapping CREs.

Overall, the discoveries reported in this doctoral thesis support the hypothesis that the disruption of pancreatic CREs might contribute to the development of PC, expanding our understanding about the implicated genetic players and associated targeted genes.

Table of contents

Abbreviations

ADM – Acinar–ductal metaplasia

AllG – The average expression of all genes

ARID1A – AT-rich interactive domain-containing protein 1A

ATAC-seq – Assay for transposase-accessible chromatin using sequencing

BP – Base pair

CDKN2A – Cyclin-dependent kinase inhibitor 2A

ChIP-seq – Chromatin immunoprecipitation followed by sequencing

CRE – Cis-regulatory element

CTCF – CCCTC-binding factor

ETAA1 – Ewing tumor-associated antigen 1

D80 – dome and 80%epiboly

DevE – Developmental enhancers

FDR – False discovery rate

FOXA2 – Forkhead box A2

GATA6 – GATA-binding factor 6

GNAS – Guanine nucleotide binding protein alpha stimulating

GO – Gene ontology

GWAS – Genome-wide association study

H.A – Human ari1ab enhancer

H.P – Human ptf1a enhancer or human pancreas

H3K27ac – Acetylation of lysine 27 on histone H3

H3K27me3 – Trimethylation of lysine 27 on histone H3

H3K4me1 – Monomethylation of lysine 4 on histone H3

H3K4me3 – Trimethylation of lysine 4 on histone H3

HC – Average expression of genes interacting with putative pancreas-specific enhancer sequences

HG – High grade

HNF1B – Hepatocyte nuclear factor 1B

HNF6 – Hepatocyte nuclear factor 6

HPF – Hours post-fertilization

INSR – Insulin receptor

IPMN – Intraductal papillary mucinous neoplasm

KLF5 – Kruppel like factor 5

KRAS – Kirsten ras homolog

LG – Low grade

LINC00673 – Long intergenic non-protein coding RNA 673

LINC01829 – Long intergenic non-protein coding RNA 1829

MCN – Mucinous cystic neoplasm

MEIS1 – Meis Homeobox 1

miR-1231 – MicroRNA 1231

MPC – Multipotent cell

ncRNA – Non-coding RNA

NGN3 – Neurogenin-3

NKX6.1 – Homeobox protein 6.1

NR5A2 – Nuclear Receptor Subfamily 5 Group A Member 2

PANIN – Pancreatic intraepithelial neoplasia

PC – Pancreatic cancer

PDAC – Pancreatic ductal adenocarcinoma

PDX1 – Pancreatic and duodenal homeobox 1

Pol II – RNA-polymerase II

PsE – Pancreatic specific enhancers

PTF1A – Pancreas-specific transcription factor 1A

PTPN11 – Protein tyrosine phosphatase non-receptor type 11

RNF43 – Ring finger protein 43

RT – Room temperature

SMAD4 – Mothers against decapentaplegic homolog 4

SOX9 – Sry-box9

O.N – Overnight

TAD – Topologically associated domain

TF – Transcription factors

TFBS – Transcription factor binding site

TP53 – Tumour protein p53

TSS – Transcription starting site

V – Ventricle

WT – Wild-type

Z.A – Zebrafish arid1ab enhancer

zE – Zebrafish enhancer

hE – Human enhancer

Z.P – Zebrafish ptf1a enhancer or Zebrafish pancreas

3C – Chromosome conformation capture

3D – Three-dimensional

Chapter I – General Introduction

The genetics of human pancreatic dysfunction in pancreatic cancer

1.1 The pancreas

1.1.1 An overview of the anatomy, morphology, and physiology of the pancreas

The pancreas is a flattened and lobulated organ that is an integral part of the digestive system. The pancreas is located, adjacent to other organs, including the small intestine, liver, and spleen, on the posterior wall of the abdominal cavity (Fig.1.1a). Macroscopically, four main parts can be distinguished in this organ: head, neck, body, and tail (Longnecker, 2014; Tsuchitani et al., 2016; Fig1.1b). The head and tail portions mark the right and left extremities of the organ, while the neck lies slightly to the right of the midline. Additionally, the body of the pancreas passes to the left, inclining slightly upwards to become continuous with the tail (Longnecker, 2021). Regarding its function, the pancreas is often described as a two-in-one organ, because it comprises two main cellular compartments with distinct functions: endocrine and exocrine (Locci et al., 2016; Fig1.1b). Together, these two cellular compartments can intervene in a variety of physiological functions in the organism (Gittes, 2009). The endocrine tissue, which makes ~5% of the total pancreatic mass, comprehends the hormone-secreting-cells (islets of Langerhans), important for the maintenance of glucose homeostasis. The islets of Langerhans, include numerous different cell types that secrets different hormones into the bloodstream regulating the glucose homeostasis and nutrient metabolism in the whole body (α-cells secreting glucagon; β-cells secreting insulin; δ-cells secreting somatostatin; ε-cells secreting ghrelin; and γ [or PP]-cells secreting pancreatic polypeptide; Fig1.1c). In contrast, the exocrine tissue, which constitutes 95% of the total pancreatic mass, comprises enzyme-secreting-cells (acinus) with essential gastrointestinal functions. Essentially, the exocrine tissue is composed by acinar cells that are responsible for the secretion of an ample digestive enzymes including trypsin, lipase, protease, and amylase, that are guided into the gastrointestinal tract through a complex ductal network system, aiding in the digestion process (Habener et al., 2005; Jennings et al., 2020; Fig1.1c).

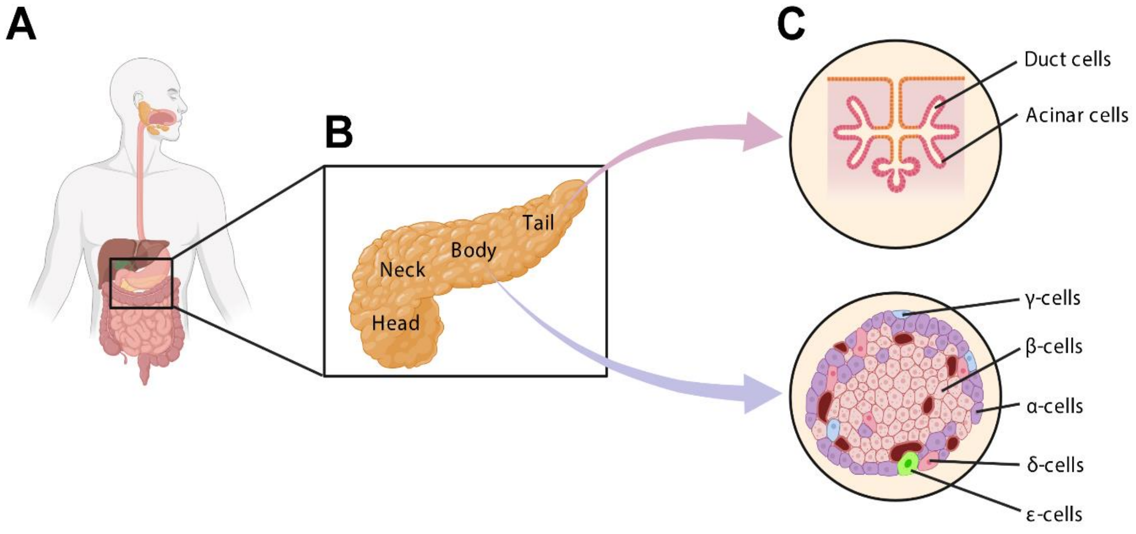


Figure 1.1 Overview of adult pancreas. a) Localization of pancreas in the human body. The mature organ is adjacent to the duodenum, the most anterior part of the small intestine. **b)** The macroscopic anatomy of pancreas. This organ can be classified in four major sections: head, neck, body, and tail. **c)** The composition of endocrine and exocrine compartments in pancreas. The exocrine part is composed by acinar (in rose) and duct cells that secrete and transport digestive enzymes, assisting the digestion. The endocrine part, comprising islets of Langerhans [composed by α-cells (in purple) secreting glucagon; β-cells secreting insulin; δ-cells (in pink) secreting somatostatin; ε-cells (in green) secreting ghrelin; and γ [or PP]-cells (in blue) secreting pancreatic polypeptide], secretes hormones responsible for the maintenance of glucose homeostasis and nutrient metabolism. Adapted from Atkinson et al., 2020; Longnecker, 2014, by Biorender.com (2022).

1.1.2 Introduction to the gene networks involved in vertebrate pancreatic development

Pancreas organogenesis is characterized by a highly and organized process comprising multiple gene regulatory networks and signalling events that controls a stepwise process of organ formation since early bud specification to a final mature and differentiated organ state (Bastidas-Ponce et al., 2017; Pan and Wright, 2011).

The pancreas development and formation begin with the thickening of the distal foregut endoderm and two evaginations in the epithelial buds in opposing sides of the endoderm of the gut: dorsal and ventral epithelial pancreatic buds (Bastidas-Ponce et al., 2017; Pan and Wright, 2011). These two pancreatic buds comprised a pool of proto-differentiated multipotent cells (MPCs) co-expressing *pancreatic and duodenal homeobox 1* (*Pdx1*) and *pancreas-specific transcription factor 1a* (*Ptf1a*) encoding genes, committing the gut endoderm to a pancreatic fate (Bastidas-Ponce et al., 2017; Burlison et al., 2008; Fig.1.2). Several other genes have been described as having a role upstream of *Pdx1* and *Ptf1a*. The existence of transcriptional binding sites for sry-box9 (Sox9) on the promoter of *Pdx1*, along with its regulatory role in *hepatocyte nuclear factor 1b* (*Hnf1b*), *hepatocyte nuclear factor 6* (*Hnf6*) and *forkhead box a2* (*Foxa2*) expression, other genes recognized to have significant roles in pancreas development, indicates that *Sox9* can be an important player in the pancreatic regulatory networks (McCracken and Wells, 2012; Fig.1.2). During the vertebrate pancreatic development, another crucial player is *Ptf1a*, since it is required for the pancreas speciation and fate, along with the maintenance of acinar cells identify (Duque et al., 2021; Kawaguchi et al., 2002). Several studies have been showed that genetic alterations in this gene and in the elements that regulates their expression are associated to pancreatic agenesis, permanent neonatal diabetes and exocrine insufficiency (Sellick et al., 2004; Weedon et al., 2014).

Following the formation of the pancreatic buds, the epithelial cells start to organize themselves, in a firmly synchronized process, that includes epithelial stratification, cellular polarization and arrangement into microlumen structures – pancreatic morphogenesis (Marty-Santos and Cleaver, 2016). Pancreatic organogenesis, in rodents, is described as having two distinct temporal transitions. In the first transition, occurs induction, budding and fusion of the pancreas, accompanied by the development of the microlumen and growth of a pool of MPCs (Marty-Santos and Cleaver, 2016). After the pancreatic budding, cells more distal form a “tip” domain containing multipotent pancreatic progenitor cells, marked by the expression of several important factors, such as *Ptf1a* (Fig.1.2). In contrast, proximal cells form the “trunk” domain that express other important factors, such as *homeobox protein 6.1* (*Nkx6.1*), *Sox9*, *Hnf1b* and *Pdx1*. Trunk cells are bipotent, being committed to differentiate into endocrine islets cells or exocrine ducts (Arda et al., 2013; Davidson, 2010; Fig.1.2). In the second transition, the microlumen go through a morphogenic alteration process to establish the luminal network and the second wave of endocrine cell formation. The expression of *neurogenin-3* (*Ngn3*), in this second transition, is relevant for the endocrine differentiation. The emerging expression of *Ngn3* in pancreatic progenitors determines the transition to endocrine precursors that consequently will give rise to endocrine cells (Arda et al., 2013; Wang et al., 2010; Fig.1.2).

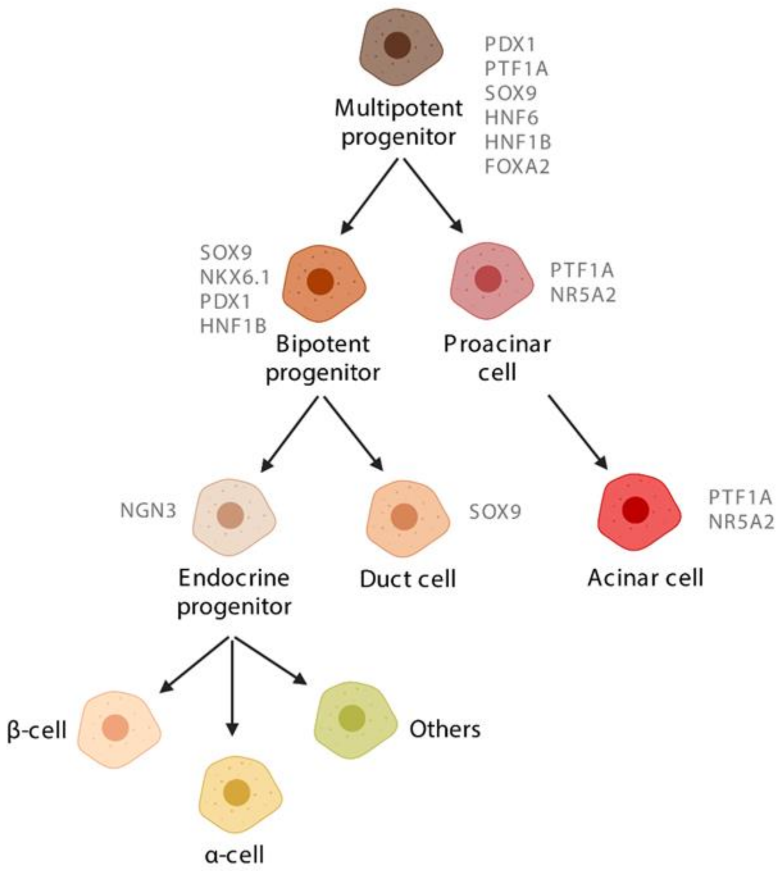


Figure 1.2 An overview of pancreatic cell lineage and its principal gene regulatory motifs during pancreatic development. Key genes (in grey) that mark each lineage includes: *Pdx1*, *Ptf1a*, *Sox9*, *Hnf6*, *Hnf1b* and *Foxa2* (multipotent progenitor); *Sox9*, *Nkx6.1*, *Pdx1* and *Hnf1b* (bipotent progenitor); *Ptf1a* and *Nr5a2* (proacinar and acinar cells); *Ngn3* (endocrine progenitor); *Sox9* (duct cell). Adapter from Arda et al., 2013; Pan and Wright, 2011, by Biorender.com (2022).

1.2 Pancreatic cancer

1.2.1 A general perspective of pancreatic cancer epidemiology

Pancreatic cancer (PC) is highly fatal malignancy, classified as the seventh leading cause of cancer death in both genders worldwide, accounting for approximately 459 000 new cases and 466 000 deaths in 2020 (Sung et al., 2021). Due to the increasing rates of incidence and/or mortality of this disease, it has been predicted that PC will soon exceed breast cancer becoming the third leading cause of cancer death in European populations (Luo et al., 2020).

The most common subtype of pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC), contributing for 90% of the cases (Gao et al., 2020; Haeberle and Esposito, 2019). Thus, the terms "pancreatic cancer" and "pancreatic ductal adenocarcinoma" are frequently used interchangeably. Hence, pancreatic ductal adenocarcinoma will be termed as PC throughout this dissertation. PC originates in the exocrine pancreas (Backx et al., 2021; Wood and Maitra, 2021), a tissue composed of acinar and duct cells, as previously described. During many years, PC was thought to arise from duct cells, due to its typical tumour ductal morphology and the expression of ductal markers (Backx et al., 2021). However, several studies have shown transitionary phenotypic stages which acinar cells adopt duct cell features, leading to the discussion, until now, about the true cell of origin of exocrine PC tumours (Backx et al., 2021; Wood and Maitra, 2021).

Other types of PC, however less frequent, are also known: neuroendocrine tumours, acinar carcinomas, solid-pseudopapillary neoplasms, pancreatoblastomas, and colloid carcinomas (Naqvi et al., 2018).

PC has an extremely poor prognosis and, typically after diagnosis, only 24% of the patients survive 1 year, and 9% lives for 5 years (Rawla et al., 2019). Several factors can be pointing out as the responsible for the poor survival rate associated to this malignancy: high aggressiveness, chemotherapeutics resistance and absence of successfully targetable oncogenic drivers (Lai et al., 2019). Over the last decades, improvements in the diagnostic approaches along with the progress of novel therapies for this disease have been made but resulting in only a limited improvement in patient outcomes (Mizrahi et al., 2020).

Many risk factors have been established as hight contributors to the emergence of PC: lifestyle (cigarette smoking, alcohol intake), specific diseases (obesity, diabetes, pancreatitis, allergies), inherited genetic factors (hereditary and familial predisposition syndromes) as well as the shifting age structure of the global population, especially in developing countries, since the risk for PC increases with the age and the world's older population continues to grow (Klein, 2021; Mizrahi et al., 2020; Rawla et al., 2019). Nevertheless, PC is a complex and multifactorial disease, being all these factors insufficient to explain its etiology (Klein, 2021; Rawla et al., 2019).

1.2.2 The genetics and molecular mechanisms of pancreatic cancer

When exposed to cellular stress and inflammatory conditions, acinar cells can go through a dedifferentiation process known as acinar–ductal metaplasia (ADM). During this process, the pancreatic acinar cells differentiate into ductal-like cells (van Roey et al., 2021; Wang et al., 2019a). However, due to oncogenic alterations and/or continuous exposure to stress, ADM may lead to a precancerous lesions commonly labelled as pancreatic intraepithelial neoplasias

(PanIN), that progress in a stepwise process culminating in development of PC (Morani et al., 2020; Orth et al., 2019). Histologically, PanINs are classified into three stages of increasingly dysplastic growth: PanIN-1, PanIN-2, and PanIN-3. The first two are classified as low-grade tumours and can be found in the normal pancreas after the age of 40 years, while the last stage (PanIN-3) is a high-grade tumour that almost always (~95% of the cases) occurs with concomitant cancer (Kim and Hong, 2018; Morani et al., 2020). Moreover, a significant proportion of PC also arise from mucinous neoplasms such as intraductal papillary mucinous neoplasmas (IPMN) and mucinous cystic neoplasmas (MCN), however these types of lesions are less frequent and studied (Tanaka et al., 2006).

It is well described that the progression of invasive PC from normal pancreatic cells involves a continuous accumulation of genetic and epigenetic alterations in fundamental signalling pathways. So, each stage of PC is associated with specific mutational profiles, that are acquired during time (Bardeesy and DePinho, 2002; Orth et al., 2019). Apart from some exceptions, four driver mutations are frequently identified in PC: an activating mutation of *kirsten ras homolog* (*KRAS*) oncogene and subsequent inactivation of *cyclin-dependent kinase inhibitor 2A (CDKN2A/p16*), *tumour protein p53* (*TP53*), and *mothers against decapentaplegic homolog 4* (*SMAD4*) tumour suppressor genes. Each of these driver mutations causes deregulation of specific signalling pathways, inducing the development and progression of cancer clones (Gu et al., 2020; Morani et al., 2020). The earliest genetic alterations identified during PC progression are the *KRAS* mutations and telomere shortening, seen approximately in 90% of low-grade PanINs (PanIN-1) and 80% of IPMNs (Morani et al., 2020; Rishi et al., 2015; Fig.1.3). These mutations trigger the initiation of the disease, and they are crucial for rapid stromal remodelling and tumour progression. The mutations that occur in the three tumour suppressor genes, *CDKN2A/p16*, TP53 and *SMAD4,* are commonly found in high-grade tumours (Bardeesy and DePinho, 2002; Brosens et al., 2015; Morani et al., 2020; Orth et al., 2019). *CDKN2A/p16* is an essential tumour suppressor gene with an important function in cell cycle regulation. It is the most common inactivated gene during the PC progression (95% of cases), and it is induced hypermethylation, mutations or deletions in the promoter region (Bardeesy and DePinho, 2002; Brosens et al., 2015; Morani et al., 2020; Orth et al., 2019; Fig.1.3). *TP53*, the “guardian” of genome, is a tumour suppressor gene that shows a vital function in cell cycle arrest, DNA repair and apoptosis. The inactivation of this gene is a later event in PC progression (Morani et al., 2020). Mutations in *TP53* occur in 75% of PC cases by missense mutations of the DNA-binding domain, leading to genetic instability (Bardeesy and DePinho, 2002; Brosens et al., 2015; Morani et al., 2020; Orth et al., 2019; Fig.1.3). In its turn, *SMAD4* inactivation arises in around 50% of PC cases also as a late event during the PC progression (Bardeesy and DePinho, 2002; Morani et al., 2020). The genetic alterations in this gene leads to aberrant cell cycle regulation, usually connected to aggressive phenotypes (Morani et al., 2020; Yamada et al., 2015; Fig.1.3).

PanINs and IPMNs share the most frequent genetic alterations, however, it has also been described some specific and relevant IPMN mutations (Morani et al., 2020; Wood and Hruban, 2012). Around 75% of IPMNs have inactivating mutations in *ring finger protein 43* (*RNF43*), a tumour suppressor gene, which encodes a ubiquitin ligase that negatively regulates the Wnt/β-catenin signalling pathway by ubiquitinating frizzled receptors. Thus, inactivating mutations in *RNF43*, mostly frameshift or nonsense mutations, promote Wnt signalling activity, leading to a neoplastic transformation (Chang et al., 2020; Sakihama et al., 2022; Fig.1.3). Moreover, in 60% of IPMNs activating mutations were found in a hotspot codon of guanine nucleotide binding protein alpha stimulating (*GNAS*), a relevant oncogene, that encodes the Gsα protein that works as a mediator in the G-protein-coupled receptor signalling pathway (Taki et al., 2016). Mutations in *GNAS* lead to the activation of G-protein signalling contributing to PC development and progression (Furukawa et al., 2011; Fig.1.3).

Apart from the most known driver mutations of PC, previously described, an enormous number of genes with a small frequency of mutations have been associated to this disease (Bailey et al., 2016; Liu et al., 2021; The Cancer Genome Atlas Research Network, 2017). The statistical significance associated to the low frequency of these mutations points out that they potentially have functional roles in the tumour development and progression. Additionally, the chance of a low-frequency-mutation gene acquire a genetic alteration is significantly higher than a high-frequency-mutation gene (Liu et al., 2021). Hence, it is plausible to consider that few of these low-frequency mutations could be developed at primary stages of the tumours and have crucial functions in tumorigenesis. The highest hit among the genes with low-frequency-mutations in PC is *AT-rich interactive domain-containing protein 1A* (*ARID1A*), comprising an 8% of mutation rate (Liu et al., 2021). This subunit of the SWI/SNF chromatin remodelling complex plays a relevant role controlling numerous biological cell process such as differentiation, proliferation, and apoptosis (Castellanos and Grippo, 2019; Wu and Roberts, 2013). Recent studies in mice have been showing that *Arid1a* has a relevant function in the ADM initiating stage. Acinar-specific *Arid1a* deletions alone are sufficient to initiate pancreatic inflammation, however, they are not able to support further progression. Nevertheless, the suppression of *Arid1a* in adult acinar cells harbouring oncogenic *Kras* mutations result in the accelerated formation of ADM and PanIN lesions (Livshits et al., 2018; Wang et al., 2019b). Additionally, the *nuclear receptor subfamily 5 group A member 2* (*NR5A2*), a member of the orphan nuclear hormone receptors family and a tumour suppressor gene, has been associated to the development of PC. Several studies have been described that *NR5A2* overexpression in PC cell lines promotes the cell migration and invasion, leading to the formation of epithelial-to-mesenchymal transition (Lin et al., 2014). In contract, other studies have been showed that in mice, *Nr5a2* gene in heterozygosity makes the pancreas more susceptible to damage, and in cooperation with other mutations can drive pancreatic tumorigenesis (Flandez et al., 2014).

PC starts developing from various precursor or preneoplastic lesions, due to activating and inactivating mutations in oncogenes and tumour suppressor genes, as described before. However, recent studies have directed their efforts to investigate the role of important pancreatic developmental genes in the development and progression of PC, such as *pancreas associated transcription factor 1a* (*PTF1A*; Mansoori et al., 2017; Naqvi et al., 2018; Reichert et al., 2016). As described previously, *PTF1A* has an important role during the pancreatic development, especially during the formation of acinar cells (Arda et al., 2013; Davidson, 2010). In addition, its continuous expression, in the adult pancreas, is also relevant to maintain the mature state of acinar cells (van Roey et al., 2021). However, during the early event of tumorigenesis, this gene can be deregulated, contributing strongly for the PC progression. Several studies have reported that *PTF1A* is downregulated during the inflammation-induced ADM (De La O et al., 2008). Additionally, it has also been described that in mice *PTF1A* is epigenetically silenced in ADM and PC cells harbouring an oncogenic *KRAS* allele (Benitz et al., 2016). In contrast, some studies described that the constant expression of *PTF1A* is able to prevent and revert *KRAS*-driven pancreas tumorigenesis, rescuing the gene program of acinar cells and restraining the tumour progression (Jakubison et al., 2018; Krah et al., 2019).

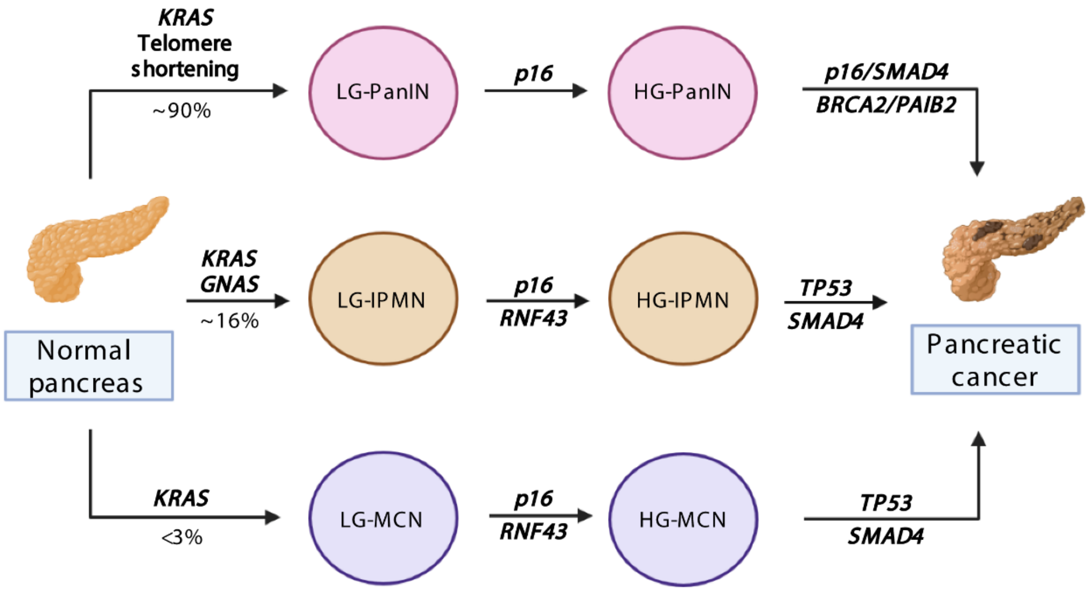


Figure 1.3 Schematic representation of precursor lesions of pancreatic cancer and its respective genetic mutations. The percentages represent the frequency of occurrence of genetic alterations in each gene. LG-PanIN = Low-grade pancreatic intraepithelial neoplasm, HG=PanIN = High-grade pancreatic intraepithelial neoplasm, LG-IPMN = Low-grade intraductal papillary mucinous neoplasm, HG-IPMN = High-grade Low-grade intraductal papillary mucinous neoplasm, LG-MCN = Low-grade mucinous cystic neoplasm, HG-MCN = High-grade mucinous cystic neoplasm. Adapted from Morani et al., 2020; Naqvi et al., 2018, by biorender.com (2022).

1.2.3 The relevance of non-coding regulatory regions in pancreatic cancer initiation, progression, and maintenance

Most of PC analysis has been largely focused on the identification of driver mutations within the protein-coding regions, where the most-well characterized pathogenic alterations are known to occur, being the contributions of non-coding regions disregard. However, it is well known that in mammal's genome, these regions are significantly larger than its protein-coding counterpart and some of these regions contain cis-regulatory elements (CREs) that are important in the regulation of gene expression (Venkat et al., 2021). During many years, due to the limited knowledge of CRE functionality in cancer genomes, the impact of non-coding regulatory mutations has been poorly studied and explored.

Recently, some studies have shown that CREs harbouring genetic mutations can have a significant impact in the regulation of pancreatic pathways, triggering the development of this disease (Scarpa and Mafficini, 2018). The impact of RNA splicing and non-coding variants as a relevant contributor to PC initiation, has gained an increasing attention in studies focus on the genetic networks of PC (Venkat et al., 2021). Zheng and colleagues described the tumour-suppressor role of the ncRNA *LINC00673* in PC context. Essentially, *LINC00673* promotes the ubiquitination and degradation of the tyrosine phosphatase *PTPN11*, leading to the inhibition of cell proliferation. However, a genome-wide association study (GWAS) found a single nucleotide polymorphism within *LINC00673* that are associated to PC risk and this genetic alteration generates a binding site for miR-1231 on *LINC00673*, causing its suppression, and this correlates with an increased in PC susceptibility (Wu et al., 2011; Zheng et al., 2016). Like the other cancer cells, PC cells frequently develop epigenetic profiles that drive the dysregulation of expression programs and the maintenance of PC phenotypic state. The kruppel like factor 5 (KLF5) is a relevant transcription factor, that is responsible for the maintenance of the chromatin acetylation of a group of enhancers, that regulates the pancreatic epithelial gene expression program (Diaferia et al., 2016). Recently, Natoli and colleagues found that the knockout of KLF5, leads to drastic reduction in the level of epigenetic marks for enhancer activity, that consequently generate a dramatic epigenomic phenotype, with a partial loss of epithelial identity in PC tumours (Diaferia et al., 2016). Additionally, a recent study of Feigin and colleagues identified regulatory non-coding mutations in the promoter region of several genes, that promotes a significant decrease in its gene expression and consequently promotes PC growth pathways, such as the Wnt/β-catenin signalling pathway, cell adhesion and axon guidance (Feigin et al., 2017).

1.2.4 The zebrafish as a pancreatic cancer model

The zebrafish (*Danio rerio*) is a small tropical freshwater fish, belonging the teleostei infraclass and that lives in rivers and rice paddles in India, Nepal, and Bangladesh (Raby et al., 2020; Howe et al., 2013). In the last few decades, this small vertebrate has emerged as a popular and powerful animal model in several scientific fields, such as toxicology, developmental biology, and human diseases (Adamson et al., 2018). The first mention in the literature of the usage of zebrafish as a model organism for developmental genetics was in 1960s, with the work of George Streisinger (reviewed in Grunwald and Eisen, 2002). Since then, this vertebrate disease model has been studied in thousands of scientific articles.

The zebrafish is attractive and useful for PC studies. As in all vertebrates, zebrafish share nearly all organs with mammals, including the liver and pancreas. Additionally, pancreas anatomy, histology and physiology are similar between teleost and several mammals (Pack et al., 1996; Youson and Al-Mahrouki, 1999). The zebrafish pancreas is composed by a principal islet that is located adjacent to the gallbladder, and numerous secondary islets that are embedded within exocrine tissue located in the intestinal mesentery. Additionally, the exocrine cells are organized in acinus, surrounding the islets, which are linked with the intestine through a numerous and complex ductular network. Besides the identical fashion of pancreas, the zebrafish exocrine and endocrine compartments also produce the same type of enzymes and hormones that can be easily localized immunohistochemically using antibodies raised against mammals (Pack et al., 1996; Youson and Al-Mahrouki, 1999; Farber et al., 2001). Additionally, there is some orthologous signalling pathways and transcription factors that regulate pancreas development in both organisms (Yee and Pack, 2005). Based on all these attributes, several transient and stable transgenic zebrafish lines have been developed in pancreatic cancer field, in order to understand if this in vivo model is able to develop tumours in endocrine and exocrine pancreas (Hwang and Goessling, 2016). Look and colleagues established a transient transgenic zebrafish line, where a member of *MYC* proto-oncogene family, with pathogenic functions in various neoplastic diseases, is expressed under control of *myod* promoter, which targets gene expression in pancreatic neuroendocrine β cells along with muscle and neuron cells (Yang et al., 2004). This line was able to develop neuroendocrine tumours in 3-6 months of age, with close similarities with human pancreatic neuroendocrine tumours at histological level (Yang et al., 2004). Additionally, Leach and colleagues established a stable transgenic zebrafish line, where an oncogenic *Kras* fused with an eGFP marker is expressed under the control of *ptf1a* regulatory elements, which targets gene expression in exocrine portion of pancreas. This line was able to develop pancreatic intraepithelial neoplasia, similar to the ones that appears in human pancreatic ductal adenocarcinoma (Park et al., 2008).

All these characteristics makes zebrafish an appealing tool to applied in PC studies. However, the successful animal model has many other attributes, which make it attractive not only to PC field, but to study human diseases in general. The large number of progenies produced (~100-200 embryos can be obtained by a single adult mating pair per week) can contribute for a high confidence in statistical analysis (Raby et al., 2020; White et al., 2013). Additionally, the production of optical clear embryos that undergoes rapid development *ex utero* as well as the existence of transparent adults, allows an in vivo imaging of the cancer growth and progression, including cell invasion, metastasis, and angiogenesis at a single-cell resolution (Stoletov et al., 2007; White et al., 2008). The efficient ability of zebrafish to absorb small molecular weight compounds that are directly dissolved in water makes this small vertebrate also attractive for anticancer drug screenings (Rennekamp and Peterson, 2015; Dang et al., 2016). Although the large phylogenetic distance between humans and zebrafish, the strong genetic conservation, development and physiology between fish and humans make zebrafish a measureless genetic tool (Hwang and Goessling, 2016; Matsuda, 2018). Thanks to the zebrafish genome-sequencing project, novel insights about orthology between human and zebrafish genomes were discovered (Jekosch, 2004). Comparative studies have estimated that 71.4% of human genes contain leastwise a single orthologue in zebrafish genome (Howe et al., 2013) and 69% of zebrafish genes contain leastwise a single orthologue in the human genome (Howe et al., 2013). Among these human orthologous, 47% of these genes contains a one-to-one link with a zebrafish orthologue (Howe et al., 2013). Additionally, comparative analysis also described that 82% of human genes linked to a human disease have an equivalent orthologue in zebrafish (Howe et al., 2013). Finally, the reduced expenses and the minimal care to maintain a zebrafish husbandry is also an attribute that makes zebrafish an attractive tool to applied in cancer studies (Hason and Bartůněk, 2019; Raby et al., 2020).

1.3 Regulation of gene expression in Eukaryotes

During the development of eukaryotic multicellular organisms, a single fertilized cell gives rise to a high diversity of cell types and tissues. This huge diversity of cells is achieved by a combinatorial and dynamic spatiotemporal expression of genes and activation of gene networks. Therefore, precise regulation of gene expression is mandatory for the development, growth, differentiation, and survival of cells (Gahan, 2005; Schvartzman et al., 2018).

The expression of genes is controlled at several levels: transcription, messenger RNA processing, transport, translation, and protein stability. Each step of control of gene expression is precisely determined and mediated by specific factors. Transcription control is the first step that occurs and plays an important role, determining RNA availability for a latter protein translation, contributing to define protein levels (Maston et al., 2006; Buccitelli and Selbach, 2020).

The eukaryotic transcription machinery, driven mostly by RNA-polymerase II (Pol II), involves two complementary regulatory components based on their structure: the cis and trans-regulatory elements. Cis-regulatory elements (CREs) are DNA sequences in the coding or non-coding regions of the genome, and the trans-regulatory elements are mostly composed by transcription factors (TFs), that are DNA-binding proteins (Maston et al., 2006; Shibata et al., 2015). TFs recognize and bind to specific sequences in the CREs to initiate, enhance or suppress transcription (Maston et al., 2006; Reinke et al., 2013; Mitsis et al., 2020). This interaction, modulated by many epigenetic processes, have been described as complex and dynamic (Müller and Stelling, 2009). CREs are controlled by the cooperative or competitive binding of different TFs. Furthermore, TFs, operating via CREs, can regulate transcription synergistically with the help of transcriptional cofactors, RNA-binding proteins, non-coding RNAs epigenetic and chromatin modifications (Shibata et al., 2015; Levine and Davidson, 2005; Son and Crabtree, 2014; Maurano et al., 2012). This cooperative and multi-level transcriptional regulation greatly contribute to the complexity of the transcriptional regulation of gene expression, leading to unique spatiotemporal patterns essential for development of multicellular organisms and proper cell function (Spitz and Duboule, 2008).

1.3.1 The function of cis-regulatory elements in regulation of gene expression

CREs are typically categorized based on their distance to the gene transcription starting site (TSS) and their detected effect on the transcription levels of their target gene. Essential for the proper transcription of genes, promoters, a proximal element that reside within 1 kb of the TSS of a gene, have a relevant role in the assembly of the transcriptional machinery that recruits Pol II to the TSS (Creyghton et al., 2010; Rada-Iglesias et al., 2011; Schier and Taatjes, 2020; Fig.1.4). For many genes, the transcriptional information contained in the promoter regions is enough to control their transcription (Danino et al., 2015; Bessa et al., 2014). However, in more complex and dynamic gene networks, other distal CREs, as enhancers, silencers and insulators are described to have an indispensable function in the regulation of gene expression (Chen and Lei, 2019; Panigrahi and O'Malley, 2021; Segert et al., 2021). In brief, enhancers contain clusters of binding sites for multiple TFs and structural proteins that positively control the Pol II activity (Ong and Corces, 2011; Fig.1.4). In contract, silencers recruit TFs and structural proteins that repress or impair the gene transcription (Fig.1.4). On the other hand, insulators are specific regulatory sequences, often enriched in binding proteins, such as CCCTC-binding factor (CTCF), responsible for the chromatin structure, which work in a position and orientation independent manner to prevent the communication between genes and nearby CREs (Maston et al., 2006; Van Bortle et al., 2014; Fig.1.4).

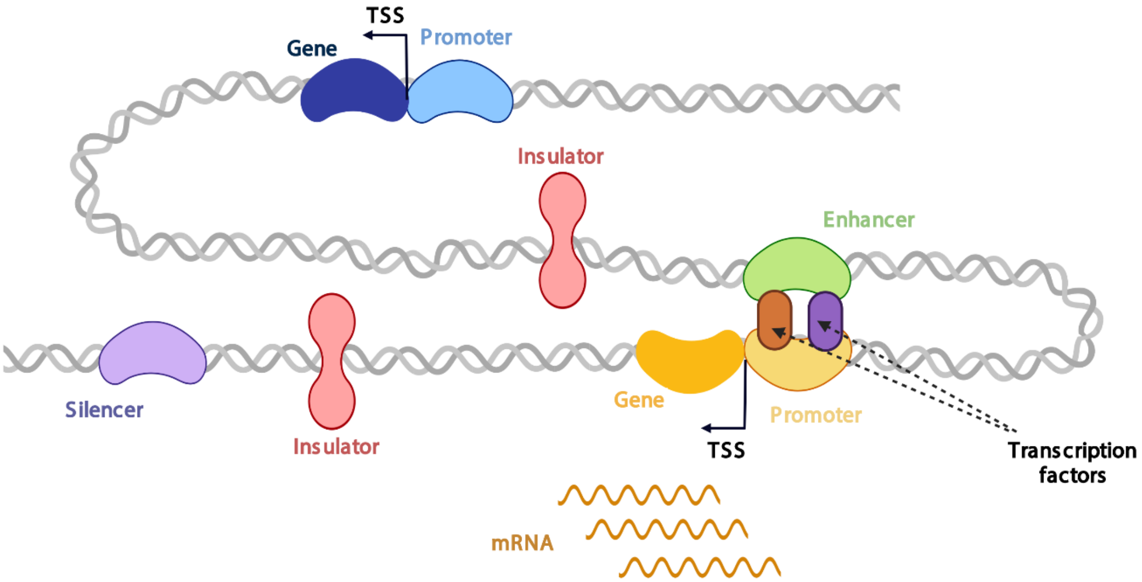


Figure 1.4 Overview of gene regulatory regions. Schematic representation of proximal and distal CREs and the respective interactions established between them. Promoters (light blue and light yellow) are next to transcription starting site (TSS) and have transcription factor binding sites (TFBSs) that serve as anchoring points for enhancers. Insulators have an opposite effect compared to enhancers, they repress the gene expression in specific tissues and specific timepoints acting as "barriers" for enhancers and silencers (Maston et al., 2006; Luizon and Ahituv, 2015). Adapter from Luizon and Ahituv, 2015, by Biorender.com (2022).

1.3.1.1 Enhancers

Enhancers, the second major category of CREs, are described as segments of DNA with hundreds of bp, located in intergenic regions, introns, or exons, and frequently present in "gene deserts" (Kleinjan and van Heyningen, 2005). While the promoter regions lay upstream of a TSS, enhancers can be found both upstream and downstream of genes. The first enhancer was found in 1981 by Schanffer and collaborators (Banerji et al., 1981). They described it as a 72 bp region of the SV40 tumour virus genome that could boost the transcription of human genes. Since then, many other enhancers have been found in many eukaryotic genomes and their biochemical and functional characteristics have been exhaustively studied (Claringbould and Zaugg, 2021; Wang et al., 2020; Ding et al., 2019).

Enhancers are small DNA regulatory elements that controls transcription of specific gene or genes. They contain specific grouped cluster of transcription factor binding sites (TFBSs) that labour cooperatively, recruiting co-activators and co-repressors, to activate the promoter and enhance the transcription of genes (Panigrahi and O'Malley, 2021; Mora et al., 2016). An interaction, between the enhancer-bound TFs and the core promoter, is thought to regulate the transcription. Some genome profiling results have uncovered that common TFs and Pol II are recruited to enhancers, indicating that enhancers could be the centres for the assembly of the core promoter (Haberle and Stark, 2018; Pennacchio et al., 2013).

Enhancers do not necessarily play a role in the nearest promoter region but can circumvent neighbouring genes to control genes placed more distantly, being one of the most distant enhancers reported located at 1 Mb from its target gene (Arnold et al., 2019; Laverré et al., 2022). Additionally, enhancers can also act independently of their orientation to target genes and can control transcription in a specific spatiotemporal manner. Different TFs factors can bind in enhancer regions and the multiple binding of specific TFs is crucial to have a tissue-specific enhancer activity (Rao et al., 2020; Lagha et al., 2012). Different tissue specific enhancers can interact with the promoter of the same target gene, combining their activities to compose complex and dynamic expression patterns (Kyrchanova and Georgiev, 2021; Snetkova and Skok, 2018).

Regarding its activity, enhancers are usually classified as either active or inactive. However, enhancers exist in multiple regulatory states during development and in reaction to multiple external stimuli (Calo and Wysocka, 2013; Heinz et al., 2015). While an active enhancer can be clearly characterized as the one that promotes transcription from the target promoter, an inactive enhancer can in fact correspond to several states, which differ not only in terms of their regulatory potential but also chromatin organization (Bozek and Gompel, 2020). Thus, in a simplistic mode, inactive enhancers can be classified as: silenced, that are sheltered in compact chromatin, depleted of active histone modifications and devoid of TF binding; repressed, that are occupied by inactivating TFs blocking the communication with the target promoters; and primed or poised enhancers, occupied by TFs and co-regulators that they do not receive sufficient regulatory input to promote transcription from the target promoters. Additionally, poised enhancers also associated with Polycomb Repressive Complex (Bozek and Gompel, 2020; Kulkarni and Arnosti, 2005; Ostuni et al., 2013; Koenecke et al., 2017).

Over the last decade, genome-wide sequencing assays taking advantage of chromatin features such histone modifications or chromatin accessibility, have revolutionized the ability to look for enhancers throughout whole genomes. However, functionally validating these sequences remains a fundamental challenge (Ryan and Farley, 2020).

1.3.2 Histone modifications as a marker for finding cis-regulatory elements

As described previously, several factors contribute to the proper transcriptional regulation of genes, including CREs. However, post-translational changes on histones have also been described as having a relevant role in the regulation of gene expression during development, and in response to several stimuli (Starks et al., 2021; Taatjes et al., 2004). Histone modifications are usually mediated by enzymes, including histone acetyltransferases, that used an acetyl CoA as cofactor and catalyse the acetylation of lysine residues, and histone methyltransferases, that mediate the methylation of lysine or arginine residues of histones (Bannister and Kouzarides, 2011; Yu and Zhuang, 2019; Taatjes et al., 2004). Some studies have described that perturbations in histone modifications can change the chromatin structure, blocking interactions among specific regulatory chromatin factors, and that consequently contribute to the development and/or progression of several diseases (Li et al., 2018a; Kurdistani, 2007; Li et al., 2018a).

Most of the histone acetylation is associated with the activation of gene transcription. Acetylation of lysine 27 on histone H3 (H3K27ac) is enriched in active promoter and enhancer regions (Wang et al., 2008; Kimura, 2013). In contrast, trimethylation of lysine 27 on histone H3 (H3K27me3) has commonly been correlated with repression of genes (Saksouk et al., 2015). In addition, monomethylation and trimethylation of lysine 4 on histone H3 (H3K4me1/H3K4me3) have been correlated with particular regulatory functions. H3K4me3 is usually enriched in TSS region of active genes, overlapping with genes' active promoters, while H3K4me1 has been described to be enriched in poised enhancer regions (Kimura, 2013; Hon et al., 2009; Creyghton et al., 2010). Each histone modification per se greatly contributes to a better knowledge of gene regulation, however, when combined, these histone changes can meticulously annotate the genome in functional domains. Thus, for example, the presence of H3K27ac and H3K4me1 marks active enhancers, while poised enhancers are characterized by an absence of H3K27ac and are enriched for H3K27me3 and H3K4me1 (Creyghton et al., 2010; Hawkins et al., 2011). On the other hand, when H3K4me3 and H3K27me3 are find together, these regions are commonly charactered as poised genes (Starks et al., 2021; Voigt et al., 2013).

Hence, many genome-wide assays have been developed to explore and investigate the gene regulation through histone modifications, being chromatin immunoprecipitation followed by sequencing (ChIP-seq) the method of choice for the genome-wide identification of histone modifications (Barski et al., 2007). This method also allows a genome-wide profiling of DNA-binding proteins, TFBSs or nucleosomes, that are also crucial for a better understanding of the gene regulatory networks involved in several biological processes (Barski et al., 2007; Park, 2009). ChIP-seq uses a specific antibody, which binds to the protein of interest, to immunoprecipitate the DNA-protein complex. Then, the DNA released from the proteins is purified and assayed directly by sequencing. The sequencing results allows the identification of the genomic regions where the protein of interest is bound (Park, 2009; Schmidt et al., 2009). Thus, the application of ChIP-seq technique allows a genome-wide analysis of histone modifications, enabling a systematic analysis of the epigenomic landscapes, that consequently contributes to a better understanding of gene regulatory networks.

1.3.3 Chromatin accessibility as a marker for finding cis-regulatory elements

Most of eukaryotic chromatin is usually found in a tightly packed chromatin state occupied by nucleosomes, making binding sites unavailable for most of TFs. Thus, chromatin reconfigurations need to occur, allowing the binding of TFs, facilitating consequently the transcription of genes. The regulation of gene expression is a dynamic competition between nucleosomes and TFs for relevant cis-regulatory sequences across the genome. This competition is mediated by chromatin modifiers, enzymes that covalently alter nucleosomes, and chromatin remodelers, enzymes that reposition, reconfigure, and eject nucleosomes. Thus, the identification of the open or accessible chromatin regions is essential to better understand the regulation of gene expression. Different techniques have been developed in order to pinpoint the accessible chromatin (Minnoye et al., 2021), being the assay for transposase-accessible chromatin using sequencing (ATAC-seq; Buenrostro et al., 2015) one of them. ATAC-seq probes DNA accessibility with a hyperactive Tn5 transposase, that simultaneously cut and inserts sequencing adapters into accessible chromatin regions for high-throughput sequencing. The resulting sequencing reads allows a multidimensional analysis of the regulatory landscapes. Besides allowing the determination of accessible chromatin, the nature of the sequence reads also allows the inference of nucleosomal positions and the detection of TF binding sites (Buenrostro et al., 2015). Overall, the application of ATAC-seq allows genome-wide mapping of chromatin accessibility, nucleosome positioning and prediction of TF binding sites that contributes to decipher the regulation of gene expression (Chen et al., 2020).

1.3.4 Importance of chromatin architecture in the regulation of gene expression

Understanding the elegantly complex nature of the three-dimensional architecture of chromatin and how it affects the gene regulation remains a major challenge in molecular biology (Pratt and Won, 2022). In most of the cases, enhancers control gene expression through binding of TFs and contacting promoters through long-range chromatin loops (Fig.1.4). In addition, these enhancer-promoter interactions can be divided into different types of architectural units, A/B compartments, which are mega-sized cell-type specific chromatin structures (Pratt and Won, 2022; Feng and Pauklin, 2020). These compartments can be further divided into finer structural domains named topologically associated domains (TADs; Pratt and Won, 2022; Feng and Pauklin, 2020). Additionally, TADs are demarcated by particular boundary elements, namely CTCF and cohesins, and represents genomic regions in which chromosomal interactions occurs more frequently with each other compared to nearby regions in the genome (Pratt and Won, 2022; Feng and Pauklin, 2020). The disruption of these topological domains can lead to aberrant enhancer-promoter interactions, contributing for the development of diseases, and promote the formation of cancer, demonstrating that TADs domains are fundamental for a proper gene transcription (Boltsis et al., 2021; Akdemir et al., 2020).

Several methods have been developed to search and study the promoter-enhancer interactions. One of the most popular genomic approaches to identify chromatin conformation is to the use chromosome conformation capture (3C) and its derivative methods, such as 4C (circularized 3C), 5C (carbon-copy 3C) and Hi-C (Dekker et al., 2013; Belton et al., 2012). All these methodologies, based on the principles of 3C, cross-link DNA using formaldehyde, maintaining regions within a three-dimensional (3D) spatial proximity linked together with protein complexes. DNA is then fragmented, and ligated, favouring ligation of DNA fragments that remain in their 3D physical proximity due to cross-link (Dekker et al., 2002; Carty et al., 2017). Additionally, the readout of these methodologies varies with the C technique involved. In general, 3C method is able to detect individual chromatin interactions between a given set of genomic loci of interest ("one-to-one"; Han et al., 2018; Pratt and Won, 2022); 4C assay is capable to detect all the interactions associated with the genomic locus of interest at the genome-wide level ("one-to-all") and 5C is a complex variation of 3C method and can uncover interactions between numerous loci in a high-throughput manner ("many-to-many"; Han et al., 2018; Pratt and Won, 2022). In contrast, Hi-C assay allows an "all-to-all" interaction profile, used for mapping all the chromatin interactions occurred in a nucleus, being a powerful tool to study 3D genomic architecture in a genome-wide manner (Han et al., 2018; Pratt and Won, 2022).

Recently, a novel technique to analysed chromatin configuration was developed, the highly integrative chromatin immunoprecipitation (HiChIP) technique (Mumbach et al., 2016). This assay is a combination between HI-C assay and a chromatin immunoprecipitation step, using a specific antibody against a protein of interest. The readout of this technique is map of all the chromatin interactions that occurs with a specific CREs (Ando-Kuri et al., 2018; Shi et al., 2020). The development of novel methodologies that describe the 3D conformation and organization of the genome, as well as the integration of multi-omic data, will provide comprehensive insights and in-depth understanding of gene regulatory networks and how genomic alterations can impact in transcriptional gene regulation, and their relevance in the development of diseases (Pratt and Won, 2022; Zarayeneh et al., 2017; Li et al., 2018b).

1.4 Research aims

The main goal of this doctoral thesis is to better understand the role of pancreatic transcriptional cis-regulatory elements in the development of pancreatic cancer. With this work, we aim to identify novel genetic players in pancreatic cancer development, as cis-regulatory sequences, and their target genes.

The specific aims are:

i) Identification of pancreatic enhancers in the zebrafish and their functional equivalents in human (Chapter II);

ii) Explore the impact of enhancer mutations in the development of pancreatic cancer, using the zebrafish as a model organism (Chapter III);

iii) Identification and functional assessment of human pancreatic enhancers in pancreatic cancer development (Chapter IV).

1.5 References

Adamson, K.I., Sheridan, E., and Grierson, A.J. (2018). Use of zebrafish models to investigate rare human disease. J Med Genet *55*, 641–649.

Akdemir, K.C., Le, V.T., Chandran, S., Li, Y., Verhaak, R.G., Beroukhim, R., Campbell, P.J., Chin, L., Dixon, J.R., Futreal, P.A., et al. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat Genet *52*, 294–305.

Ando-Kuri, M., Rivera, I.S.M., Rowley, M.J., and Corces, V.G. (2018). Analysis of Chromatin Interactions Mediated by Specific Architectural Proteins in Drosophila Cells. Methods Mol Biol *1766*, 239–256.

Arda, H.E., Benitez, C.M., and Kim, S.K. (2013). Gene regulatory networks governing pancreas development. Dev Cell *25*, 5–13.

Arnold, P.R., Wells, A.D., and Li, X.C. (2019). Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. Front Cell Dev Biol *7*, 377.

Atkinson, M.A., Campbell-Thompson, M., Kusmartseva, I., and Kaestner, K.H. (2020). Organisation of the human pancreas in health and in diabetes. Diabetologia *63*, 1966–1973.

Backx, E., Coolens, K., Van den Bossche, J.-L., Houbracken, I., Espinet, E., and Rooman, I. (2021). On the Origin of Pancreatic Cancer. Cell Mol Gastroenterol Hepatol S2352-345X(21)00248-4.

Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.-M., Gingras, M.-C., Miller, D.K., Christ, A.N., Bruxner, T.J.C., Quinn, M.C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature *531*, 47–52.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res *21*, 381–395.

Bardeesy, N., and DePinho, R.A. (2002). Pancreatic cancer biology and genetics. Nat Rev Cancer *2*, 897–909.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bastidas-Ponce, A., Scheibner, K., Lickert, H., and Bakhti, M. (2017). Cellular and molecular mechanisms coordinating pancreas development. Development *144*, 2873–2888.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods *58*, 268–276.

Benitz, S., Regel, I., Reinhard, T., Popp, A., Schäffer, I., Raulefs, S., Kong, B., Esposito, I., Michalski, C.W., and Kleeff, J. (2016). Polycomb repressor complex 1 promotes gene silencing through H2AK119 mono-ubiquitination in acinar-to-ductal metaplasia and pancreatic cancer cells. Oncotarget *7*, 11424–11433.

Bessa, J., Luengo, M., Rivero-Gil, S., Ariza-Cosano, A., Maia, A.H.F., Ruiz-Ruano, F.J., Caballero, P., Naranjo, S., Carvajal, J.J., and Gómez-Skarmeta, J.L. (2014). A mobile insulator system to detect and disrupt cis-regulatory landscapes in vertebrates. Genome Res *24*, 487–495.

Boltsis, I., Grosveld, F., Giraud, G., and Kolovos, P. (2021). Chromatin Conformation in Development and Disease. Front Cell Dev Biol *9*, 723859.

Bozek, M., and Gompel, N. (2020). Developmental Transcriptional Enhancers: A Subtle Interplay between Accessibility and Activity: Considering Quantitative Accessibility Changes between Different Regulatory States of an Enhancer Deconvolutes the Complex Relationship between Accessibility and Activity. Bioessays *42*, e1900188.

Brosens, L.A.A., Hackeng, W.M., Offerhaus, G.J., Hruban, R.H., and Wood, L.D. (2015). Pancreatic adenocarcinoma pathology: changing “landscape.” J Gastrointest Oncol *6*, 358–374.

Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet *21*, 630–644.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol *109*, 21.29.1-21.29.9.

Burlison, J.S., Long, Q., Fujitani, Y., Wright, C.V.E., and Magnuson, M.A. (2008). Pdx-1 and Ptf1a concurrently determine fate specification of pancreatic multipotent progenitor cells. Dev Biol *316*, 74–86.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol Cell *49*, 825–837.

Carty, M., Zamparo, L., Sahin, M., González, A., Pelossof, R., Elemento, O., and Leslie, C.S. (2017). An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun *8*, 15454.

Castellanos, K.J., and Grippo, P.J. (2019). ARID1A: guardian of normal pancreatic ducts. Transl Cancer Res *8*, S133–S134.

Chang, X.Y., Wu, Y., Jiang, Y., Wang, P.Y., and Chen, J. (2020). RNF43 Mutations in IPMN Cases: A Potential Prognostic Factor. Gastroenterol Res Pract *2020*, 1457452.

Chen, D., and Lei, E.P. (2019). Function and regulation of chromatin insulators in dynamic genome organization. Curr Opin Cell Biol *58*, 61–68.

Chen, M., Zhang, Z., Meng, Z.Y., and Zhang, X.J. (2020). [ATAC-seq and its applications in complex disease]. Yi Chuan *42*, 347–353.

Claringbould, A., and Zaugg, J.B. (2021). Enhancers in disease: molecular basis and emerging treatment strategies. Trends Mol Med *27*, 1060–1073.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931–21936.

Dang, M., Fogley, R., and Zon, L.I. (2016). Identifying Novel Cancer Therapies Using Chemical Genetics and Zebrafish. Adv Exp Med Biol *916*, 103–124.

Danino, Y.M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The core promoter: At the heart of gene expression. Biochim Biophys Acta *1849*, 1116–1131.

Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. Nature *468*, 911–920.

De La O, J.-P., Emerson, L.L., Goodman, J.L., Froebe, S.C., Illum, B.E., Curtis, A.B., and Murtaugh, L.C. (2008). Notch and Kras reprogram pancreatic acinar cells to ductal intraepithelial neoplasia. Proc Natl Acad Sci U S A *105*, 18907–18912.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet *14*, 390–403.

Diaferia, G.R., Balestrieri, C., Prosperini, E., Nicoli, P., Spaggiari, P., Zerbi, A., and Natoli, G. (2016). Dissection of transcriptional and cis-regulatory control of differentiation in human pancreatic cancer. EMBO J *35*, 595–617.

Ding, X., Jiang, X., Tian, R., Zhao, P., Li, L., Wang, X., Chen, S., Zhu, Y., Mei, M., Bao, S., et al. (2019). RAB2 regulates the formation of autophagosome and autolysosome in mammalian cells. Autophagy *15*, 1774–1786.

Duque, M., Amorim, J.P., and Bessa, J. (2021). Ptf1a function and transcriptional cis-regulation, a cornerstone in vertebrate pancreas development. FEBS J.

Farber, S.A., Pack, M., Ho, S.Y., Johnson, I.D., Wagner, D.S., Dosch, R., Mullins, M.C., Hendrickson, H.S., Hendrickson, E.K., and Halpern, M.E. (2001). Genetic analysis of digestive physiology using fluorescent phospholipid reporters. Science *292*, 1385–1388.

Feigin, M.E., Garvin, T., Bailey, P., Waddell, N., Chang, D.K., Kelley, D.R., Shuai, S., Gallinger, S., McPherson, J.D., Grimmond, S.M., et al. (2017). Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma. Nat Genet *49*, 825–833.

Feng, Y., and Pauklin, S. (2020). Revisiting 3D chromatin architecture in cancer development and progression. Nucleic Acids Res *48*, 10632–10647.

Flandez, M., Cendrowski, J., Cañamero, M., Salas, A., del Pozo, N., Schoonjans, K., and Real, F.X (2014) Nr5a2 heterozygosity sensitises to, and cooperates with, inflammation in KRas(G12V)-driven pancreatic tumourigenesis. Gut *63*, 645-655.

Furukawa, T., Kuboki, Y., Tanji, E., Yoshida, S., Hatori, T., Yamamoto, M., Shibata, N., Shimizu, K., Kamatani, N., and Shiratori, K. (2011). Whole-exome sequencing uncovers frequent GNAS mutations in intraductal papillary mucinous neoplasms of the pancreas. Sci Rep *1*, 161.

Gahan, P.B. (2005). Molecular biology of the cell (4th edn) B. Alberts, A. Johnson, J. Lewis, K. Roberts and P. Walter (eds), Garland Science, 1463 pp., ISBN 0-8153-4072-9 (paperback) (2002). Cell Biochemistry and Function *23*, 150–150.

Gao, H.-L., Wang, W.-Q., Yu, X.-J., and Liu, L. (2020). Molecular drivers and cells of origin in pancreatic ductal adenocarcinoma and pancreatic neuroendocrine carcinoma. Exp Hematol Oncol *9*, 28.

Gittes, G.K. (2009). Developmental biology of the pancreas: a comprehensive review. Dev Biol *326*, 4–35.

Grunwald, D.J., and Eisen, J.S. (2002). Headwaters of the zebrafish -- emergence of a new model vertebrate. Nat Rev Genet *3*, 717–724.

Gu, Y., Ji, Y., Jiang, H., and Qiu, G. (2020). Clinical Effect of Driver Mutations of KRAS, CDKN2A/P16, TP53, and SMAD4 in Pancreatic Cancer: A Meta-Analysis. Genet Test Mol Biomarkers *24*, 777–788.

Habener, J.F., Kemp, D.M., and Thomas, M.K. (2005). Minireview: transcriptional regulation in pancreatic development. Endocrinology *146*, 1025–1034.

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol *19*, 621–637.

Haeberle, L., and Esposito, I. (2019). Pathology of pancreatic cancer. Transl Gastroenterol Hepatol *4*, 50.

Han, J., Zhang, Z., and Wang, K. (2018). 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. Mol Cytogenet *11*, 21.

Hason, M., and Bartůněk, P. (2019). Zebrafish Models of Cancer-New Insights on Modeling Human Cancer in a Non-Mammalian Vertebrate. Genes (Basel) *10*, E935.

Hawkins, R.D., Hon, G.C., Yang, C., Antosiewicz-Bourget, J.E., Lee, L.K., Ngo, Q.-M., Klugman, S., Ching, K.A., Edsall, L.E., Ye, Z., et al. (2011). Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. Cell Res *21*, 1393–1409.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol *16*, 144–154.

Hon, G.C., Hawkins, R.D., and Ren, B. (2009). Predictive chromatin signatures in the mammalian genome. Hum Mol Genet *18*, R195-201.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498–503.

Hwang, K.L., and Goessling, W. (2016). Baiting for Cancer: Using the Zebrafish as a Model in Liver and Pancreatic Cancer. Adv Exp Med Biol *916*, 391–410.

Jakubison, B.L., Schweickert, P.G., Moser, S.E., Yang, Y., Gao, H., Scully, K., Itkin-Ansari, P., Liu, Y., and Konieczny, S.F. (2018). Induced PTF1a expression in pancreatic ductal adenocarcinoma cells activates acinar gene networks, reduces tumorigenic properties, and sensitizes cells to gemcitabine treatment. Mol Oncol *12*, 1104–1124.

Jekosch, K. (2004). The zebrafish genome project: sequence analysis and annotation. Methods Cell Biol *77*, 225–239.

Jennings, R.E., Scharfmann, R., and Staels, W. (2020). Transcription factors that shape the mammalian pancreas. Diabetologia *63*, 1974–1980.

Kawaguchi, Y., Cooper, B., Gannon, M., Ray, M., MacDonald, R.J., and Wright, C.V.E. (2002). The role of the transcriptional regulator Ptf1a in converting intestinal to pancreatic progenitors. Nat Genet *32*, 128–134.

Kim, J.Y., and Hong, S.-M. (2018). Precursor Lesions of Pancreatic Cancer. Oncol Res Treat *41*, 603–610.

Kimura, H. (2013). Histone modifications for human epigenome analysis. J Hum Genet *58*, 439–445.

Klein, A.P. (2021). Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. Nat Rev Gastroenterol Hepatol *18*, 493–502.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet *76*, 8–32.

Koenecke, N., Johnston, J., He, Q., Meier, S., and Zeitlinger, J. (2017). Drosophila poised enhancers are generated during tissue patterning with the help of repression. Genome Res *27*, 64–74.

Krah, N.M., Narayanan, S.M., Yugawa, D.E., Straley, J.A., Wright, C.V.E., MacDonald, R.J., and Murtaugh, L.C. (2019). Prevention and Reversion of Pancreatic Tumorigenesis through a Differentiation-Based Mechanism. Dev Cell *50*, 744-754.e4.

Kulkarni, M.M., and Arnosti, D.N. (2005). cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. Mol Cell Biol *25*, 3411–3420.

Kurdistani, S.K. (2007). Histone modifications as markers of cancer prognosis: a cellular view. Br J Cancer *97*, 1–5.

Kyrchanova, O., and Georgiev, P. (2021). Mechanisms of Enhancer-Promoter Interactions in Higher Eukaryotes. Int J Mol Sci *22*, E671.

Lagha, M., Bothma, J.P., and Levine, M. (2012). Mechanisms of transcriptional precision in animal development. Trends Genet *28*, 409–416.

Lai, E., Puzzoni, M., Ziranu, P., Pretta, A., Impera, V., Mariani, S., Liscia, N., Soro, P., Musio, F., Persano, M., et al. (2019). New therapeutic targets in pancreatic cancer. Cancer Treat Rev *81*, 101926.

Laverré, A., Tannier, E., and Necsulea, A. (2022). Long-range promoter-enhancer contacts are conserved during evolution and contribute to gene expression robustness. Genome Res *32*, 280–296.

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc Natl Acad Sci U S A *102*, 4936–4942.

Li, F., Wan, M., Zhang, B., Peng, Y., Zhou, Y., Pi, C., Xu, X., Ye, L., Zhou, X., and Zheng, L. (2018a). Bivalent Histone Modifications and Development. Curr Stem Cell Res Ther *13*, 83–90.

Li, Y., Hu, M., and Shen, Y. (2018b). Gene regulation in the 3D genome. Hum Mol Genet *27*, R228–R233.

Lin, Q., Aihara, A., Chung, W., Chen, X., Huang, Z., Weng, S., Carlson, R., Nadolny, C., Wands, J., and Dong, X. (2014). LRH1 promotes pancreatic cancer metastasis. Cancer Lett *350*: 15-24.

Liu, S., Cao, W., Niu, Y., Luo, J., Zhao, Y., Hu, Z., and Zong, C. (2021). Single-PanIN-seq unveils that ARID1A deficiency promotes pancreatic tumorigenesis by attenuating KRAS-induced senescence. Elife *10*, e64204.

Livshits, G., Alonso-Curbelo, D., Morris, J.P., Koche, R., Saborowski, M., Wilkinson, J.E., and Lowe, S.W. (2018). Arid1a restrains Kras-dependent changes in acinar cell identity. Elife *7*, e35216.

Locci, G., Pinna, A.P., Dessì, A., Obinu, E., Gerosa, C., Marcialis, M.A., Pintus, M.C., Angiolucci, M., Fanos, V., Ambu, R., et al. (2016). Stem progenitor cells in the human pancreas. Journal of Pediatric and Neonatal Individualized Medicine (JPNIM) *5*, e050223–e050223.

Longnecker, D.S. (2014). Anatomy and Histology of the Pancreas (Version 1.0). Pancreapedia: The Exocrine Pancreas Knowledge Base.

Longnecker, D.S. (2021). Anatomy and Histology of the Pancreas. Pancreapedia: The Exocrine Pancreas Knowledge Base.

Luizon, M.R., and Ahituv, N. (2015). Uncovering drug-responsive regulatory elements. Pharmacogenomics *16*, 1829–1841.

Luo, W., Tao, J., Zheng, L., and Zhang, T. (2020). Current epidemiology of pancreatic cancer: Challenges and opportunities. Chin J Cancer Res *32*, 705–719.

Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S., and Baradaran, B. (2017). The Different Mechanisms of Cancer Drug Resistance: A Brief Review. Adv Pharm Bull *7*, 339–348.

Marty-Santos, L., and Cleaver, O. (2016). Pdx1 regulates pancreas tubulogenesis and E-cadherin expression. Development *143*, 101–112.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29–59.

Matsuda, H. (2018). Zebrafish as a model for studying functional pancreatic β cells development and regeneration. Dev Growth Differ *60*, 393–399.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McCracken, K.W., and Wells, J.M. (2012). Molecular pathways controlling pancreas induction. Semin Cell Dev Biol *23*, 656–662.

Minnoye, L., Marinov, G.K., Krausgruber, T., Pan, L., Marand, A.P., Secchia, S., Greenleaf, W.J., Furlong, E.E.M., Zhao, K., Schmitz, R.J., et al. (2021). Chromatin accessibility profiling methods. Nat Rev Methods Primers *1*, 1–24.

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G.P., and Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression (Review). World Academy of Sciences Journal *2*, 3–8.

Mizrahi, J.D., Surana, R., Valle, J.W., and Shroff, R.T. (2020). Pancreatic cancer. Lancet *395*, 2008–2020.

Mora, A., Sandve, G.K., Gabrielsen, O.S., and Eskeland, R. (2016). In the loop: promoter-enhancer interactions and bioinformatics. Brief Bioinform *17*, 980–995.

Morani, A.C., Hanafy, A.K., Ramani, N.S., Katabathina, V.S., Yedururi, S., Dasyam, A.K., and Prasad, S.R. (2020). Hereditary and Sporadic Pancreatic Ductal Adenocarcinoma: Current Update on Genetics and Imaging. Radiol Imaging Cancer *2*, e190020.

Müller, D., and Stelling, J. (2009). Precise regulation of gene expression dynamics favors complex promoter architectures. PLoS Comput Biol *5*, e1000279.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods *13*, 919–922.

Naqvi, A.A.T., Hasan, G.M., and Hassan, M.I. (2018). Investigating the role of transcription factors of pancreas development in pancreatic cancer. Pancreatology *18*, 184–190.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet *12*, 283–293.

Orth, M., Metzger, P., Gerum, S., Mayerle, J., Schneider, G., Belka, C., Schnurr, M., and Lauber, K. (2019). Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. Radiat Oncol *14*, 141.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. Cell *152*, 157–171.

Pack, M., Solnica-Krezel, L., Malicki, J., Neuhauss, S.C., Schier, A.F., Stemple, D.L., Driever, W., and Fishman, M.C. (1996). Mutations affecting development of zebrafish digestive organs. Development *123*, 321–328.

Pan, F.C., and Wright, C. (2011). Pancreas organogenesis: from bud to plexus to gland. Dev Dyn *240*, 530–565.

Panigrahi, A., and O'Malley, B.W. (2021). Mechanisms of enhancer action: the known and the unknown. Genome Biol *22*, 108.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669–680.

Park, S.W., Davison, J.M., Rhee, J., Hruban, R.H., Maitra, A., Leach, S.D. (2008). Oncogenic KRAS induces progenitor cell expansion and malignant transformation in zebrafish exocrine pancreas. Gastroenterology *134*, 2080-2090.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. Nat Rev Genet *14*, 288–295.

Pratt, B.M., and Won, H. (2022). Advances in profiling chromatin architecture shed light on the regulatory dynamics underlying brain disorders. Semin Cell Dev Biol *121*, 153–160.

Raby, L., Völkel, P., Le Bourhis, X., and Angrand, P.-O. (2020). Genetic Engineering of Zebrafish in Cancer Research. Cancers (Basel) *12*, E2168.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rao, S., Ahmad, K., and Ramachandran, S. (2020). Cooperative Binding of Transcription Factors is a Hallmark of Active Enhancers. 2020.08.17.253146.

Rawla, P., Sunkara, T., and Gaduputi, V. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. World J Oncol *10*, 10–27.

Reichert, M., Blume, K., Kleger, A., Hartmann, D., and von Figura, G. (2016). Developmental Pathways Direct Pancreatic Cancer Initiation from Its Cellular Origin. Stem Cells Int *2016*, 9298535.

Reinke, V., Krause, M., and Okkema, P. (2013). Transcriptional regulation of gene expression in C. elegans. WormBook 1–34.

Rennekamp, A.J., and Peterson, R.T. (2015). 15 years of zebrafish chemical screening. Curr Opin Chem Biol *24*, 58–70.

Rishi, A., Goggins, M., Wood, L.D., and Hruban, R.H. (2015). Pathological and molecular evaluation of pancreatic neoplasms. Semin Oncol *42*, 28–39.

van Roey, R., Brabletz, T., Stemmler, M.P., and Armstark, I. (2021). Deregulation of Transcription Factor Networks Driving Cell Plasticity and Metastasis in Pancreatic Cancer. Front Cell Dev Biol *9*, 753456.

Ryan, G.E., and Farley, E.K. (2020). Functional genomic approaches to elucidate the role of enhancers during development. Wiley Interdiscip Rev Syst Biol Med *12*, e1467.

Sakihama, K., Koga, Y., Yamamoto, T., Shimada, Y., Yamada, Y., Kawata, J., Shindo, K., Nakamura, M., and Oda, Y. (2022). RNF43 as a predictor of malignant transformation of pancreatic mucinous cystic neoplasm. Virchows Arch.

Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin *8*, 3.

Scarpa, A., and Mafficini, A. (2018). Non-coding regulatory variations: the dark matter of pancreatic cancer genomics. Gut *67*, 399–400.

Schier, A.C., and Taatjes, D.J. (2020). Structure and mechanism of the RNA polymerase II transcription machinery. Genes Dev *34*, 465–488.

Schmidt, D., Wilson, M.D., Spyrou, C., Brown, G.D., Hadfield, J., and Odom, D.T. (2009). ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods *48*, 240–248.

Schvartzman, J.M., Thompson, C.B., and Finley, L.W.S. (2018). Metabolic regulation of chromatin modifications and gene expression. J Cell Biol *217*, 2247–2259.

Segert, J.A., Gisselbrecht, S.S., and Bulyk, M.L. (2021). Transcriptional Silencers: Driving Gene Expression with the Brakes On. Trends Genet *37*, 514–527.

Sellick, G.S., Barker, K.T., Stolte-Dijkstra, I., Fleischmann, C., Coleman, R.J., Garrett, C., Gloyn, A.L., Edghill, E.L., Hattersley, A.T., Wellauer, P.K., et al. (2004). Mutations in PTF1A cause pancreatic and cerebellar agenesis. Nat Genet *36*, 1301–1305.

Shi, C., Rattray, M., and Orozco, G. (2020). HiChIP-Peaks: a HiChIP peak calling algorithm. Bioinformatics *36*, 3625–3631.

Shibata, M., Gulden, F.O., and Sestan, N. (2015). From trans to cis: transcriptional regulatory networks in neocortical development. Trends Genet *31*, 77–87.

Snetkova, V., and Skok, J.A. (2018). Enhancer talk. Epigenomics *10*, 483–498.

Son, E.Y., and Crabtree, G.R. (2014). The role of BAF (mSWI/SNF) complexes in mammalian neural development. Am J Med Genet C Semin Med Genet *166C*, 333–349.

Spitz, F., and Duboule, D. (2008). Global control regions and regulatory landscapes in vertebrate development and evolution. Adv Genet *61*, 175–205.

Starks, R.R., Kaur, H., and Tuteja, G. (2021). Mapping cis-regulatory elements in the midgestation mouse placenta. Sci Rep *11*, 22331.

Stoletov, K., Montel, V., Lester, R.D., Gonias, S.L., and Klemke, R. (2007). High-resolution imaging of the dynamic tumor cell vascular interface in transparent zebrafish. Proc Natl Acad Sci U S A *104*, 17406–17411.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin *71*, 209–249.

Taatjes, D.J., Marr, M.T., and Tjian, R. (2004). Regulatory diversity among metazoan co-activator complexes. Nat Rev Mol Cell Biol *5*, 403–410.

Taki, K., Ohmuraya, M., Tanji, E., Komatsu, H., Hashimoto, D., Semba, K., Araki, K., Kawaguchi, Y., Baba, H., and Furukawa, T. (2016). GNAS(R201H) and Kras(G12D) cooperate to promote murine pancreatic tumorigenesis recapitulating human intraductal papillary mucinous neoplasm. Oncogene *35*, 2407–2412.

Tanaka, M., Chari, S., Adsay, V., Fernandez-del Castillo, C., Falconi, M., Shimizu, M., Yamaguchi, K., Yamao, K., Matsuno, S., and International Association of Pancreatology (2006). International consensus guidelines for management of intraductal papillary mucinous neoplasms and mucinous cystic neoplasms of the pancreas. Pancreatology *6*, 17–32.

The Cancer Genome Atlas Research Network (2017). Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell *32*, 185-203.e13.

Tsuchitani, M., Sato, J., and Kokoshima, H. (2016). A comparison of the anatomical structure of the pancreas in experimental animals. J Toxicol Pathol *29*, 147–154.

Van Bortle, K., Nichols, M.H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z.S., and Corces, V.G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol *15*, R82.

Venkat, S., Alahmari, A.A., and Feigin, M.E. (2021). Drivers of Gene Expression Dysregulation in Pancreatic Cancer. Trends Cancer *7*, 594–605.

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes Dev *27*, 1318–1338.

Wang, L., Xie, D., and Wei, D. (2019a). Pancreatic Acinar-to-Ductal Metaplasia and Pancreatic Cancer. Methods Mol Biol *1882*, 299–308.

Wang, S., Yan, J., Anderson, D.A., Xu, Y., Kanal, M.C., Cao, Z., Wright, C.V.E., and Gu, G. (2010). Neurog3 gene dosage regulates allocation of endocrine and exocrine cell fates in the developing mouse pancreas. Dev Biol *339*, 26–37.

Wang, S.C., Nassour, I., Xiao, S., Zhang, S., Luo, X., Lee, J., Li, L., Sun, X., Nguyen, L.H., Chuang, J.-C., et al. (2019b). SWI/SNF component ARID1A restrains pancreatic neoplasia formation. Gut *68*, 1259–1270.

Wang, W., Hu, C.-K., Zeng, A., Alegre, D., Hu, D., Gotting, K., Ortega Granillo, A., Wang, Y., Robb, S., Schnittker, R., et al. (2020). Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates. Science *369*, eaaz3090.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet *40*, 897–903.

Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., De Franco, E., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., Allen, H.L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet *46*, 61–64.

White, R., Rose, K., and Zon, L. (2013). Zebrafish cancer: the state of the art and the path forward. Nat Rev Cancer *13*, 624–636.

White, R.M., Sessa, A., Burke, C., Bowman, T., LeBlanc, J., Ceol, C., Bourque, C., Dovey, M., Goessling, W., Burns, C.E., et al. (2008). Transparent adult zebrafish as a tool for in vivo transplantation analysis. Cell Stem Cell *2*, 183–189.

Wood, L.D., and Hruban, R.H. (2012). Pathology and molecular genetics of pancreatic neoplasms. Cancer J *18*, 492–501.

Wood, L.D., and Maitra, A. (2021). Insights into the origins of pancreatic cancer. Nature *597*, 641–642.

Wu, J.N., and Roberts, C.W.M. (2013). ARID1A mutations in cancer: another epigenetic tumor suppressor? Cancer Discov *3*, 35–43.

Wu, C., Miao, X., Huang, L., Che, X., Jiang, G., Yu, D., Yang, X., Cao, G., Hu, Z., Zhou, Y., et al. (2011). Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. Nat Genet *44*, 62–66.

Yamada, S., Fujii, T., Shimoyama, Y., Kanda, M., Nakayama, G., Sugimoto, H., Koike, M., Nomoto, S., Fujiwara, M., Nakao, A., et al. (2015). SMAD4 expression predicts local spread and treatment failure in resected pancreatic cancer. Pancreas *44*, 660–664.

Yang, H.W., Kutok, J.L., Lee, N.H., Piao, H.Y., Fletcher, C.D.M., Kanki, J.P., Look, A.T. (2004). Targeted expression of human MYCN selectively causes pancreatic neuroendocrine tumors in transgenic zebrafish. Cancer Res *64*, 7256-7262.

Yee, N.S., and Pack, M. (2005). Zebrafish as a model for pancreatic cancer research. Methods Mol Med *103*, 273–298.

Youson, J.H., and Al-Mahrouki, A.A. (1999). Ontogenetic and phylogenetic development of the endocrine pancreas (islet organ) in fish. Gen Comp Endocrinol *116*, 303–335.

Yu, C., and Zhuang, S. (2019). Histone Methyltransferases as Therapeutic Targets for Kidney Diseases. Front Pharmacol *10*, 1393.

Zarayeneh, N., Ko, E., Oh, J.H., Suh, S., Liu, C., Gao, J., Kim, D., and Kang, M. (2017). Integration of multi-omics data for integrative gene regulatory network inference. Int J Data Min Bioinform *18*, 223–239.

Zheng, J., Huang, X., Tan, W., Yu, D., Du, Z., Chang, J., Wei, L., Han, Y., Wang, C., Che, X., et al. (2016). Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. Nat Genet *48*, 747–757.

Chapter II

Identification of pancreatic enhancers in the zebrafish and their functional equivalents in human

This chapter was accepted for publication in *Nature Communications* journal:

Renata Bordeira-Carriço†, **Joana Teixeira**†, Marta Duque†, Mafalda Galhardo, Diogo Ribeiro, Rafael D. Acemel, Panos. N. Firbas, Juan J. Tena, Ana Eufrásio, Joana Marques, Fábio J. Ferreira, Telmo Freitas, Fátima Carneiro, José Luís Goméz-Skarmeta and José Bessa*. (2022) Multidimensional chromatin profiling of zebrafish pancreas to uncover and investigate disease-relevant enhancers. Nat Commun.13:1945. https://doi.org/10.1038/s41467-022-29551-7

†These authors contributed equally to this work

*Corresponding author.

Author contributions:

JB designed and supervised the study. JLG-S supervised the work and gave important inputs to the study. RBC obtained biological material and generated ATAC-seq, ChIP-seq, 4C-seq and HiChIP data from zebrafish pancreas. RBC and FJF collected biological material and generated RNA-seq data from zebrafish pancreas and muscle. JT, MG, RDA, PNF and JTx performed computational analyses and data interpretation. RBC, MD, AE and DR performed enhancer-assays in zebrafish and CRISPR-Cas9 in zebrafish as well as immunohistochemistry, microscopy acquisition and analysis. JTx performed transfection, CRISPR-Cas9 and image acquisition in human cell lines with support of FF. FC contributed with histology of human pancreas. TF and JM contributed for plasmid and zebrafish lines generation. JB wrote the manuscript with input from all authors, and all contributed for the development and discussion of the work.

2.1 Introduction

The mechanisms that tightly control transcription are essential for organ function. The transcriptional regulation of genes is controlled by non-coding cis-regulatory elements (CREs) spread over large genomic distances (Furlong and Levine, 2018a). Genome-Wide Association Studies (GWAS) have identified many non-coding disease-associated alleles that have a hereditary component and overlap with CREs epigenetic signatures, suggesting that the disruption of CREs may be one of the genetic bases of human disease. This is the case of some pancreatic diseases, such as pancreatic cancer and diabetes (Klein et al., 2018; Mahajan et al., 2018; Morris et al., 2012; Pasquali et al., 2014; Wolpin et al., 2014), that have a heavy societal burden, with incidence and death rates increasing worldwide (GBD 2017 Pancreatic Cancer Collaborators, 2019; Huang et al., 2021; Lascar et al., 2018; Lippi and Mattiuzzi, 2020; Saeedi et al., 2019; Sinclair et al., 2020). Many previous studies demonstrated an enrichment of diabetes-associated variants in adult human islet enhancers (Greenwald et al., 2019; Khetan et al., 2018; Mahajan et al., 2018; Miguel-Escalada et al., 2019; Parker et al., 2013; Pasquali et al., 2014), corroborating the hypothesis of pancreatic diseases being caused by alterations in CREs. Likewise, experimental in vivo and in vitro enhancer reporter assays also showed that specific islet enhancer variants correlate with altered regulatory functions (Eufrásio et al., 2020; Gaulton et al., 2010; Khetan et al., 2018; Kycia et al., 2018; Roman et al., 2017). Studies of the role of CREs' mutations in the development of pancreatic diseases using in vivo models would provide invaluable insight given the complex regulatory networks involved; however, such studies are very still scarce (Akerman et al., 2021; van Arensbergen et al., 2017; Fujitani et al., 2006).

The zebrafish is a vertebrate model suitable for genetic manipulation (Hwang et al., 2013), with a pancreas that shares many similarities with the human pancreas, including similar transcription factors (TFs) and genetic networks of pancreatic development and function (Kinkel and Prince, 2009; Prince et al., 2017). Thus, the zebrafish is a suitable in vivo model to validate causal regulatory variants. Yet, the identification of interspecies functionally equivalent CREs faces unsolved fundamental challenges, such as low conservation of interspecies non-coding sequences (Elgar and Vavouri, 2008) and, for the minority of CREs whose sequence is conserved, their fast-evolving functionality (Prescott et al., 2015). Indeed, although sequence conservation of non-coding sequences has successfully been used to find enhancers, many with interspecies orthologous identities (modENCODE Consortium et al., 2010; Visel et al., 2009), it has also been demonstrated to be insufficient for identifying all enhancers within a genome and between species (Fisher et al., 2006; Wittkopp and Kalay, 2011). To bypass these limitations, in this work we profiled the chromatin state of zebrafish pancreas cells and chromatin interaction points. We were able to accurately identify zebrafish pancreatic enhancers and, by comparisons with similar human datasets, we predicted functionally equivalent pancreatic enhancers. These findings revealed a previously unidentified human enhancer in the landscape of the tumour suppressor *ARID1A* (Jones et al., 2012; Wu and Roberts, 2013), with a potential role in the susceptibility to pancreatic cancer. Additionally, we explored the regulatory landscape of *PTF1A*, known to contain a human distal enhancer whose deletion leads to pancreatic agenesis/hypoplasia (Demirbilek et al., 2020; Evliyaoğlu et al., 2018; Gabbay et al., 2017; Weedon et al., 2014), and found a zebrafish distal *ptf1a* enhancer that contains similar regulatory information to its human counterpart. We further demonstrated its functional equivalency by showing that its ablation induces pancreatic agenesis, explained by a reduction in the pancreatic progenitor domain early in development. Taken together, the multidimensional chromatin profiling used here allowed the establishment of previously unknown functional connections between human and zebrafish enhancers. These bridges between different species are invaluable for the prediction of new disease-relevant enhancers and the study of their role in human disease.

2.2 Results

2.2.1 Zebrafish putative pancreatic enhancers share developmental roles

When comparing the basic structure of the human and zebrafish adult pancreas we observed that the organ structure is analogous between the two species (Fig.2.1a). We further extended this comparison to the cellular composition of the main cell types of the pancreas between zebrafish, mouse (Alvarsson et al., 2020) and human (Alvarsson et al., 2020; Rahier et al., 1981; Saito et al., 1978), and found that the predominance of the major cellular types is maintained in these three vertebrates (Supplementary Fig.2.1). Because of these extended similarities between the zebrafish and mammal pancreas, the zebrafish has been used as a model to study pancreatic diseases (Kinkel and Prince, 2009; Park and Leach, 2018). Furthermore, these similarities hint at the existence of shared genetic networks that operate, likely through equivalent sets of CREs, in these three species. Thus, we explored the chromatin state and chromatin interaction points of zebrafish whole pancreas, to gather information about endocrine and exocrine cells, and compared it to human data sets. To identify CREs active in the zebrafish adult pancreas, we performed ChIP-seq for H3K27ac (Rada-Iglesias et al., 2011), a key histone modification associated with active enhancers, and ATAC-seq (Buenrostro et al., 2013), an assay that identifies regions of open chromatin (Fig.2.1b). We also performed HiChIP (Gaulton et al., 2010) against H3K4me3 (Guenther et al., 2007) to detect active promoters interacting with the uncovered enhancers (Fig.2.1b). We found 14753 putative active enhancers, mostly in intergenic regions (57.8%), and 23298 putative active promoters corresponding to 9848 genes (Fig.2.1c; Supplementary Dataset 1a-c). To identify a subset of pancreatic enhancers with higher tissue-specificity, we compared the H3K27ac data from adult zebrafish pancreas to whole zebrafish embryos at four developmental stages, Dome, 80% epiboly, 24 hours post-fertilization (hpf) and 48hpf (Bogdanovic et al., 2012), since these comprise differentiated and non-differentiated cells from many different tissues. We found that 7115 putative enhancers (48.2%) are active only in the differentiated adult pancreas (PsE; Fig.1c; Supplementary Dataset 1a-c) while the remaining 7638 (51.8%) are also broadly active in developing embryos (DevE), suggesting that their activity is not restricted to the pancreas. DevE presented 4 clusters (C1-4) with different H3K27ac abundance profiles during the different developmental stages (Fig.2.1d; Supplementary Fig.2.2a; Supplementary Dataset 1e-l), suggesting that, apart from their activity in the adult pancreas, these enhancers might function in other cell types. C1 and C4 show similar levels of H3K27ac in all developmental stages, compatible with a putative ubiquitous enhancer activity, while C2 and C3 show different levels of H3K27ac during development, which may reflect a dynamic state of repression (C2) and activation (C3) of enhancers, or, alternatively, differences in the abundance of cells where these enhancers are active during development.

2.2.2 Functional similarities between human and zebrafish pancreatic enhancers

Pancreatic enhancers are expected to activate the expression of genes in the pancreas. To test if the predicted enhancers correlate with the expression of target genes in the pancreas, we identified the nearest genes to each putative pancreatic enhancer (Hiller et al., 2013; McLean et al., 2010) and observed that genes nearby PsE are enriched for exocrine pancreas expression (*p*<4.27E-9; Supplementary Fig.2.2b; Supplementary Dataset 2a-c), detected by *in situ* hybridization (Hiller et al., 2013; McLean et al., 2010). These results contrast with the ones obtained for DevE, for which nearby genes are enriched for expression in several other tissues, including epidermis and endothelial cells (Supplementary Fig.2.2; Supplementary Dataset 2d-f), suggesting a higher tissue-specificity of PsE. Additionally, the presence of endothelial expression also in genes associated to the PsE group suggests the detection of endothelial enhancers, likely derived from the vasculature present in the zebrafish adult pancreas (Supplementary Dataset 2d-f).

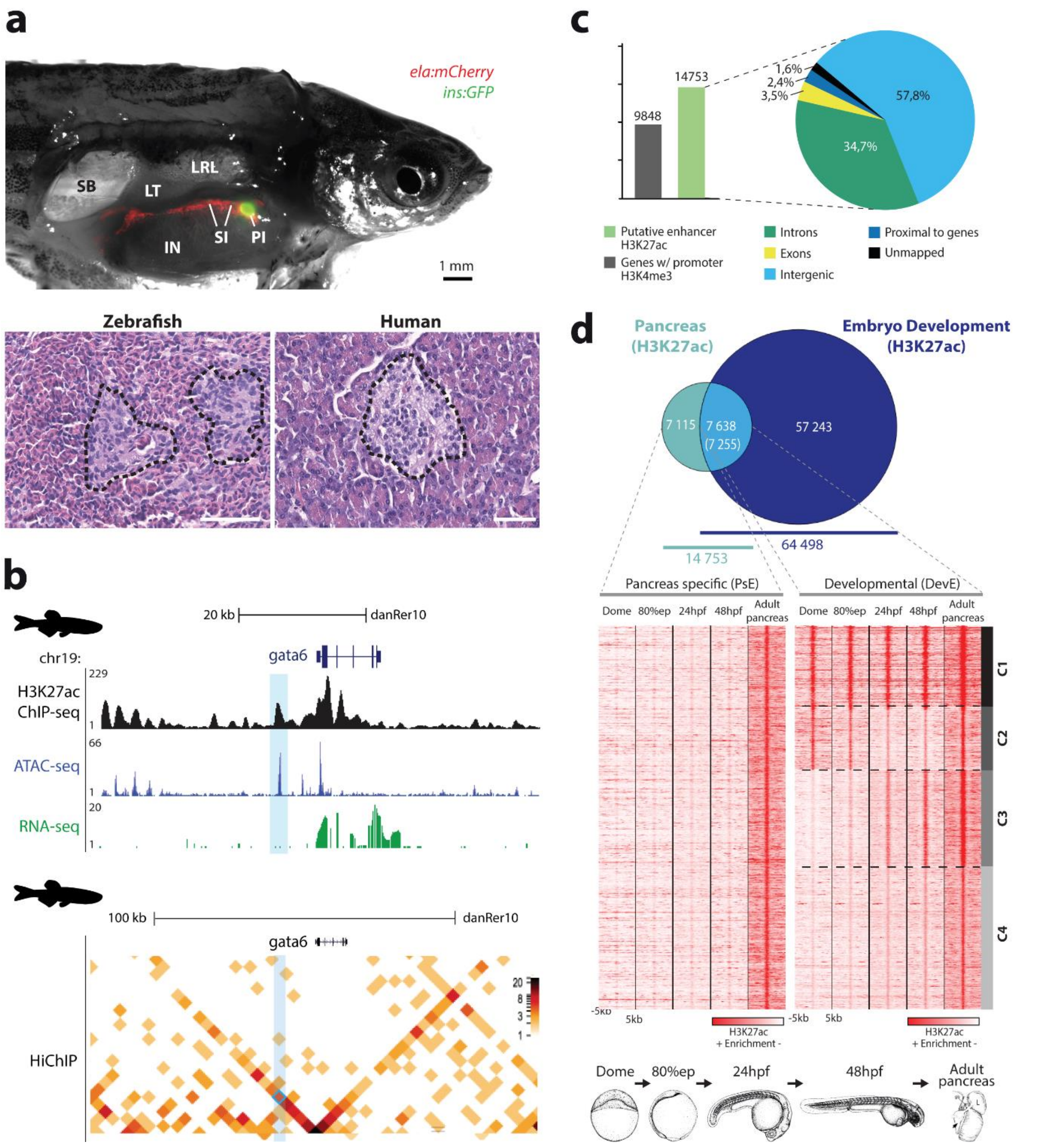


Figure 2.1 The zebrafish pancreas, from histology to chromatin state. **a)** Comparison of the basic structure of the human and zebrafish adult pancreas. Above: Dissected adult male Tg(insulin:GFP, elastase:mCherry) zebrafish; insulin and elastase promoters drive GFP expression in beta-cells (green) and mCherry in acinar cells (red), respectively. IN, intestine; LRL, Liver right lobe; LT, left testis; PI, principal islet; SI, secondary islets; SB, swim bladder. Below: Histology of the pancreas; transverse sections with hematoxylin/eosin staining showing islets of Langerhans (black dashed lines) surrounded by exocrine tissue in zebrafish and human pancreas. Magnification:

40x and scale bar: 1mm **b)** Genomic landscape of *gata6* in the zebrafish adult pancreas showing the H3K27ac ChIP-seq profile (black) and ATAC-seq peaks (blue) from whole pancreas, RNA-seq from exocrine pancreas (green) and a heat map for chromatin interactions with *gata6* promoter detected by HiChIP for H3K4me3 from whole pancreas (below). A putative enhancer sequence that interacts with the *gata6* promoter is highlighted by the light blue box. **c)** Bar plot (left panel) showing the number of genes with active promoters (defined by H3K4me3 signal, gray bar) and putative active enhancers in adult zebrafish pancreas (defined by H3K27ac mark, green bar), and their distribution throughout the regions of the genome (right panel). **d)** Above: Venn diagram showing the overlap of putative active enhancers in adult zebrafish pancreas and stages of zebrafish embryonic development. Putative active enhancers exclusive to the adult pancreas form the pancreas-specific enhancers (PsE) group, while the shared enhancers belong to the developmental shared enhancers (DevE) group (Supplementary Dataset 1e-f). Below: Heat maps showing clusters of H3K27ac mark for PsE and DevE enhancers during embryonic development [dome, 80% epiboly (80%epi), 24hpf, 48hpf] and in adult pancreas. A window of 10 kb around the reference coordinates for each sequence was used and the density files were subjected to k-means clustering, obtaining four different clusters in DevE: C1, Cluster 1; C2, Cluster 2; C3, Cluster 3; and C4, Cluster 4. For © and (**d**), source data are provided as a Source Data file.

To improve the enhancer to gene association, we used H3K4me3 HiChIP to detect chromatin interactions between active promoters and putative enhancers in the zebrafish adult pancreas (Fig.2.1b; Supplementary Dataset 3a) and used RNA-seq to evaluate transcription (Fig.2.1b). We found that, compared to all genes, PsE-associated genes have a higher average expression in multiple pancreatic cell types (Fig.2.2a, Supplementary Dataset 3b). As expected, these expression results contrast with the lower average expression levels of the PsE-associated genes compared to all genes in a distantly related control tissue such as the muscle (Fig.2.2a, Supplementary Dataset 3b). Similar results were obtained when analysing genes associated to the other identified clusters of pancreatic enhancers, specifically, DevE, C1-C4 and the total dataset of pancreatic enhancers altogether (PsEs+DevE; Supplementary Fig.2.2c-d, Supplementary Dataset 3c-g), which had higher expression levels for at least one pancreatic adult tissue and lower expression levels in the muscle (control tissue), when compared to all transcribed genes. Next, we performed a similar analysis by calculating the ratio of the average expression level of genes associated to C1-4 and PsE putative enhancers (HC) divided by the average expression of all genes (AllG), using the previously published transcriptome of whole zebrafish embryos from 18 developmental stages (White et al., 2017). We found that the genes associated to C1-4 and PsE have a HC/AllG ratio ≥ 1 (Fig.2.2b; Supplementary Fig.2.2e) and that the HC/AllG ratio of the DevE associated genes is higher than the one of PsE associated genes, for most of the analysed developmental time points (Fig.2.2b). These results suggest that DevE enhancers likely control gene expression during development in embryonic stages of the zebrafish. This hypothesis is further supported by the observed variation of the HC/AllG ratio during development that partially reflects the variation of H3K27ac signal observed in the enhancers of the C1-4 clusters (Fig.2.1d, Fig.2.2b and Supplementary Fig.2.2e). For instance, the C2 group that shows an increased presence of H3K27ac signal at Dome and 80% epiboly developmental time-points (Fig.2.1d), also shows an increased HC/AllG ratio in the earliest developmental time points (BDO:blastula to G75: 75%epiboly; Fig.2.2b and Supplementary Fig.2.2e). These results suggest that C1-4 enhancers control gene expression in the adult differentiated pancreas, in addition to other cell types during development. Overall, these results increase the robustness of the pancreatic enhancers predictions, since it is possible to correlate with the transcription of the respective putative target genes.

To determine if the detected H3K27ac signal is a good predictor of active pancreatic enhancers, we performed in vivo enhancer reporter assays for 17 regions within the regulatory landscapes of known pancreatic genes. We selected sequences with detectable, but variable, H3K27ac signal overlapping with open chromatin, detected by ATAC-seq (Buenrostro et al., 2013). Of the 10 sequences with the highest H3K27ac values (-log10(p-value) from 36.5 to 92.1), 6 were validated in vivo as pancreatic enhancers (60%; Fig.2.2c and d, Supplementary Fig.2.3a and Supplementary Dataset 4a). Conversely, of the remaining 7 sequences with the lowest H3K27ac values (-log10(p-value) from 18.5 to 28.4), only 1 showed strong and reproducible evidence of pancreatic enhancer activity (14%, Supplementary Fig.2.3a-c and Supplementary Dataset 4a). Previous studies described similar percentages of validated enhancers from H3K27ac positive sequences (Gorkin et al., 2020; Nord et al., 2013; Shen et al., 2012). These results validate the robustness of pancreatic enhancers prediction based on chromatin state and further suggest that the abundance of H3K27ac mark in genomic locations might improve such predictions.

We observed that out of 14753 putative zebrafish pancreatic enhancers, only 12.49% (n=1842) could be directly aligned to the human genome (Hinrichs et al., 2006; Fig.2.3a and Supplementary Dataset 3i-l). A similar proportion was found in the group of developmental enhancers (11.36%; 7326 out of 64498; Fig.2.3a). Using the corresponding human sequences from the pancreas and developmental enhancers groups, we found that they share similar PhastCons conservation scores (Fig.2.3b; Supplementary Fig.2.3d and Supplementary Dataset 3m-p). Next, we wanted to determine if the zebrafish putative pancreatic enhancers that align to the human genome also overlap with H3K27ac signal from human pancreas. Only a minority of interspecies aligned sequences shared H3K27ac signal (total pancreas data set:

227 out of 1842; PsE: 115 out of 1052; DevE: 112 out of 790). The human sequences, that shared H3K27ac signal with zebrafish, did not show a higher average conservation score than the aligned sequences that showed H3K27ac signal in zebrafish alone (Fig.2.3b and Supplementary Fig.2.3e; Average sequence conservation score for H3K27ac non-shared vs shared signal, Pancreas: 0.40vs0.36, PsE:0.42vs0.41, DevE:0.36vs0.34). Notwithstanding the low absolute numbers of aligned sequences that share H3K27ac signal in human and zebrafish pancreas, these sequences represent a clear enrichment compared to the overlap obtained by randomized set of sequences in the human genome (3.21 times higher for pancreas, 2.79 times higher for PsE, 3.76 times higher for DevE and 1.76 times higher for embryo, Fig.2.3c; Supplementary Dataset 3q). Overall, these results suggest that pancreatic enhancer function is not a strong condition to impose sequence conservation.

Following these data, we assessed whether functionally equivalent pancreatic CREs exist between human and zebrafish, despite an overall lack of sequence conservation. To explore this possibility, we investigated if the genes interacting with each cluster of zebrafish enhancers were enriched for homologs of human genes associated with pancreatic diseases, which would suggest the existence of functionally equivalent pancreatic CREs with potential biomedical relevance. Such enrichment was observed for the clusters of late development and adult pancreas (PsE, C3 and C4; Fig.2.3d; Supplementary Dataset 3r-s). Human gene-disease associations were retrieved from DisGeNET (Piñero et al., 2015) and we observed that 306 out of 836 zebrafish genes (36.6%) homologous to human pancreas disease-associated genes also interact with zebrafish pancreatic enhancers.

Enhancers can exist in their typical form, as short and restricted regions of DNA, or they can be present as large regions of hyperactive chromatin referred to as super enhancers (Lovén et al., 2013; Parker et al., 2013; Whyte et al., 2013). Several computational approaches have been applied to identify super enhancers in vertebrate genomes, including in human and zebrafish (Pérez-Rico et al., 2017). We searched for super enhancers active in the pancreas of human and zebrafish (Supplementary Dataset 1m-n; 275 in zebrafish and 875 in human), to understand if pancreatic super enhancers control the same genes in both species, further suggesting an equivalency in function. Gene ontology for putative target genes showed a similar enrichment for transcriptional regulation in both species and several of these genes corresponded to the same orthologues (32 out of the 271 zebrafish genes; Supplementary Fig.2.3f-g), some with important pancreatic functions, such as *INSR,* a critical regulator of glucose homeostasis (Shirakawa et al., 2017) and *GATA6,* which plays a crucial role in pancreas development and β-cell function (Tiyaboonchai et al., 2017; Supplementary Fig.2.3h). We further inquired if human and zebrafish enhancers might operate similarly, using equivalent TFs. To test this, we performed a motif enrichment search for TF binding sites (TFBS) in regions of open chromatin identified by ATAC-seq (Buenrostro et al., 2013), within the 14753 pancreatic enhancers, and found several TFBS for known pancreatic TFs (ZP; Fig.2.3f, Supplementary Fig.2.4a, and Supplementary Dataset 3t-u). We also performed a similar analysis using available human whole pancreas datasets (HP; ENCODE Project Consortium et al., 2020; Datasets summarized in Supplementary Dataset 4g). To compare the extent of overlap of enriched motifs in human and zebrafish pancreatic enhancers with motifs enriched in other pancreas unrelated enhancers, we have performed a similar motif enrichment search for datasets of zebrafish embryos (D80, dome and 80%epiboly; 24HPF, 24 hpf) and human heart ventricle (V; ENCODE Project Consortium et al., 2020; Datasets summarized in Supplementary Dataset 4g). We selected the top 140 enriched motifs from each dataset and observed that the majority of the common motifs were found in zebrafish (ZP) and human (HP) pancreas datasets (ZP,HP:98; ZP,D80:63; HP,D80:61; Fig.2.3g, Supplementary Fig.2.4b), while comparisons with the human ventricle (V) showed that ZP,HP was the second largest group following HP,V (Supplementary Fig.2.4c).

Several TFs, such as Ptf1a, Pdx1, Pax6 and Sox9, are known to be important for pancreas function or development in several vertebrate species, including human and zebrafish (Cebola et al., 2015; Duque et al., 2021; Jennings et al., 2020; Pasquali et al., 2014). As shown above, human and zebrafish pancreatic enhancers are enriched for many shared TFBS, therefore it is reasonable to expect that many of these TFBS are from TFs known to have an important pancreatic function. To test this hypothesis, we have selected 25 TFs known to be required for pancreas function and development and calculated the distribution of the respective TFBS motifs within the previously identified enriched motifs described in Supplementary Dataset 3t. We found that the majority of the TFBS motifs from the pancreatic TFs were within the ZP,HP overlapping datasets, regardless of the compared groups (Supplementary Fig.4d-f). These results suggest that the same set of TFs operates in zebrafish and human pancreatic enhancers. Overall, these results argue in favour of interspecies functional equivalency of enhancers.

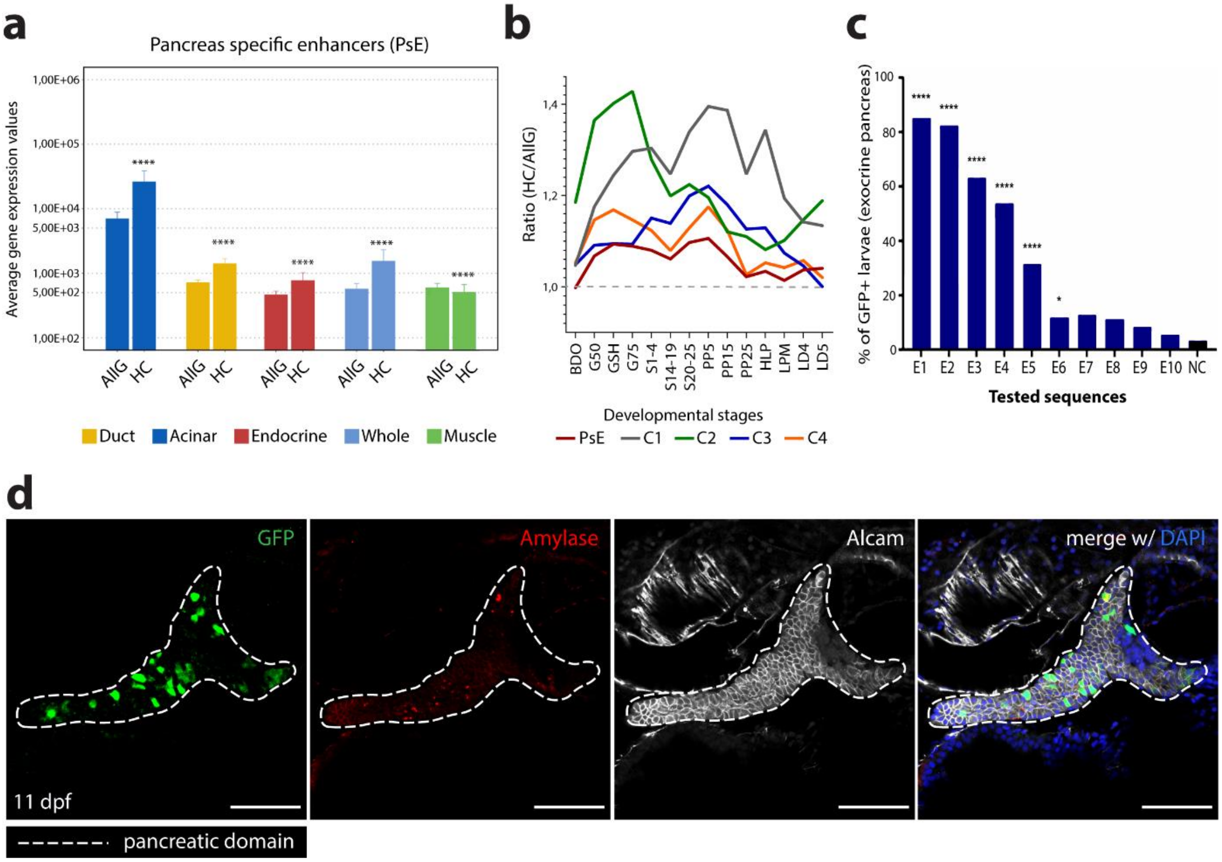


Figure 2.2 ChIP-seq and ATAC-seq data accurately predict functional pancreatic enhancers. a) Average expression of genes interacting with putative pancreas-specific enhancer sequences (PsE), detected by HiChIP for H3K4me3 (HC, n=6174 genes), compared to the average expression of all genes (AllG, n=33737 genes). Gene expression was determined from RNA-seq data from different pancreatic cells (acinar n=4, duct n=3, endocrine pancreas n=4), whole pancreas(n=2), and muscle (control; n=2). One-sided Wilcoxon test (≥), *p*-values<0.05 were considered statistically significant (*****p*<2E-16). Error bars represent the 95% confidence interval. **B)** Ratio between average expression of genes interacting with putative pancreatic enhancers (PsE, C1, C2, C3 and C4 clusters) and the average expression of all genes throughout zebrafish development. C1, C2, C3 and C4 are different clusters that compose the DevE category. BDO: blastula, dome; G50: gastrula, 50% epiboly; GSH: gastrula, shield; G75: gastrula, 75% epiboly; S1-4: segmentation, 1-4 somites; S14-19: segmentation, 14-19 somites; S20-25: segmentation, 20-25 somites; PP5: pharyngula, Prim-5; PP15: pharyngula, Prim-15; PP25: pharyngula, Prim-25; HLP: hatching, long-pec; LPM: larval, protruding-mouth; LD4: larval, day 4; LD5: larval, day 5. **c)** Percentage of F0 zebrafish larvae with GFP expression in the exocrine pancreas following in vivo transient transgenesis reporter assays. The empty enhancer reporter vector was used as the negative control (NC). The depicted sequences (E1 to 10) represent the top 10 putative enhancer sequences with higher H3K27ac signal ("high H3K27ac" group). Values are represented as percentages and compared by two-sided Chi-square with Yates' correction test. *P*-values<0.05 were considered significant (*****p*<0.0001, **p*<0.05). The exact *p*-value and n are discriminated in Supplementary Dataset 4. **d)** Representative confocal image of the in vivo transient transgenesis reporter assays for the E3 sequence (n=30). Depicted in c) showing expression of GFP (green) in 11dpf zebrafish pancreas (white dashed line), labelled by anti-Alcam staining (white) and anti-Amylase (red) antibodies (n=30, from 2 independent

injections, with 63.33% of larvae showing GFP expression in the exocrine pancreas). Nuclei were stained with DAPI (blue). Images were captured with a Leica SP5II confocal microscope. Scale bar: 60 µm. For (**a**), (**b**) and (**c**), source data are provided as a Source Data file.

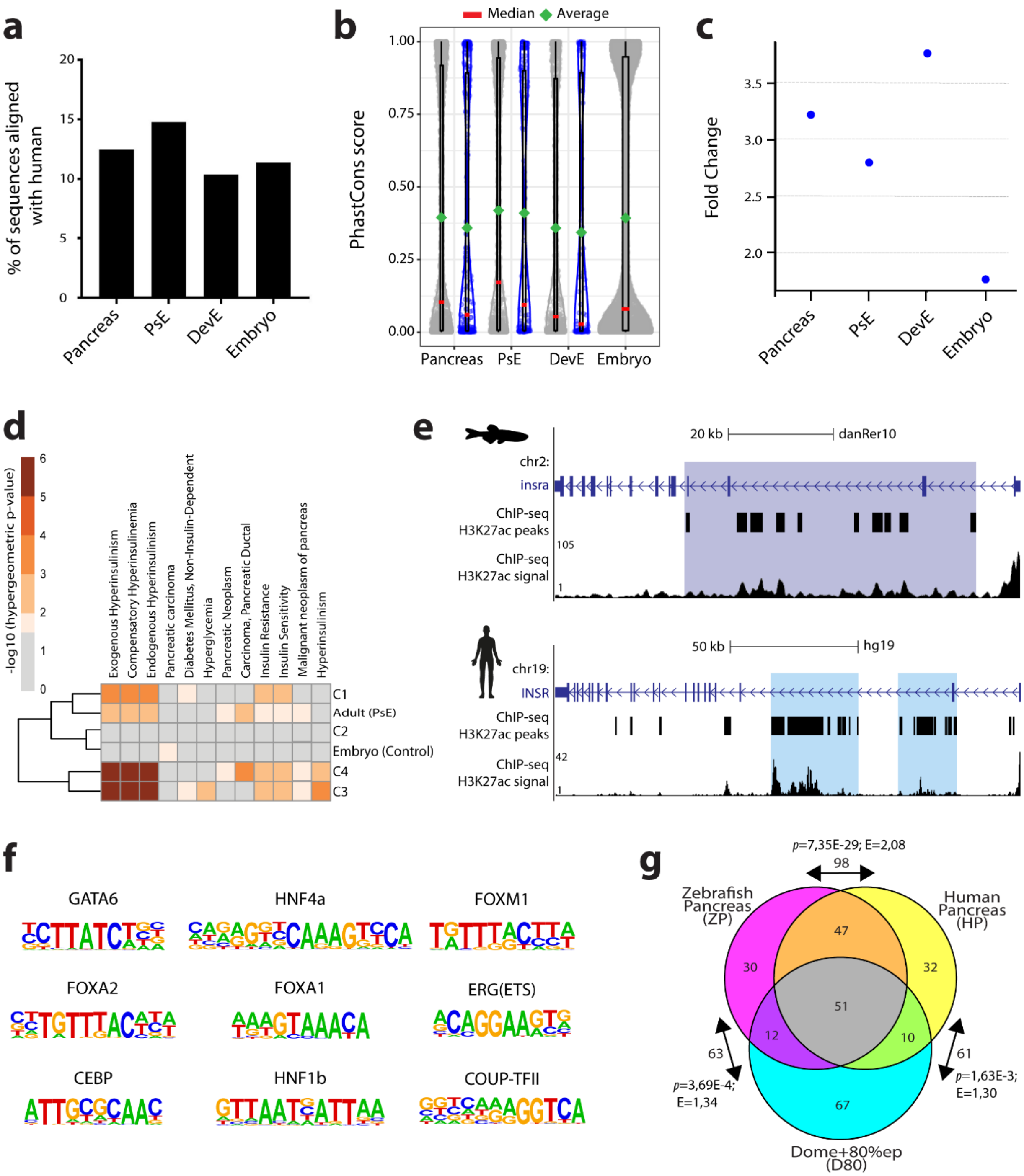


Figure 2.3 The zebrafish and human pancreas share cis-regulatory similarities. a) Percentage of predicted zebrafish pancreatic enhancer sequences aligned to the human genome. Sequences are grouped in different clusters: "Pancreas" that includes PsE and DevE; "PsE"; "DevE"; "Embryo" that include putative enhancers active only during embryonic development. **b)** PhastCons scores (99 vertebrate genomes against hg38) for human sequences converted from zebrafish putative enhancers. Grey dots label conserved sequences that do not overlap with H3K27ac mark in human pancreas (Pancreas-1801, PsE-1017, DevE-784 and Embryo-6792). Blue dots label conserved sequences that also show H3K27ac signal in human pancreas (ENCODE data; Pancreas-227, PsE-112, DevE-115). Green diamonds: average (grey dots: 0.40, 0.42, 0.36, 0.39; blue dots: 0.36, 0.41, 0.34, respectively for Pancreas, PsE, DevE and Embryo). Red line: median (grey dots: 0.10, 0.17, 0.05, 0.08; blue dots: 0.06, 0.09, and 0.03, respectively for pancreas, PsE, DevE and Embryo). The Embryo dataset is composed by different developmental stages (Dome, 80% Epiboly, 24hpf and 48hpf). **c)** Ratio between the number of human sequences conserved with the zebrafish putative active enhancers (Pancreas-3.21, PsE-2.79, DevE-3.76 or Embryo-1.76) overlapping H3K27ac signal in human pancreas (ENCODE data) over the average of a 10^5 random shuffling of human sequences overlapping with H3K27ac signal in human pancreas (Supplementary Dataset 3q; empirical *p*-value < 1E-5). **d)** Heatmap showing -log$_{10}$(p-values) from hypergeometric enrichment test for pancreatic disease association on the genes linked by HiChIP to each enhancer cluster. Represented values meet the criteria: q-value≤0.05 and fold enrichment≥1.5. **e)** Genomic landscape of the human *INSR* gene (top) and zebrafish *arid1ab* ortholog (bottom), showing H3K27ac signal and predicted super-enhancers (blue). **f)** Relevant pancreas transcription factors whose binding motifs are enriched in zebrafish pancreas H3K27ac ChIP-seq data. **g)** Venn diagram of the top 140 enriched TFBS motifs in H3K27ac positive sequences in three different datasets: zebrafish pancreas (ZP), human pancreas (HP) and dome+80%epiboly embryos (D80). Number of motifs shared between pairs of groups (arrows). *p*-values are described (*p;* hypergeometric enrichment test). The enrichment of the observed *vs* expected is represented (E). *p*-values≤0.05 were considered significant. For (**a**), (**b**), (**c**), (**d**) and (**g**), source data provided in Source Data file.

2.2.3 Landscape of *arid1a* reveals potential pancreatic cancer associated enhancer

To better address the hypothesis of interspecies functional equivalency of enhancers, we focused on the regulatory landscape of a gene that is potentially linked to human pancreatic diseases. We selected *arid1ab*, the orthologue of human *ARID1A*, a tumour-suppressor gene associated with cancer in several different cell types (Jones et al., 2012; Wu and Roberts, 2013), including pancreatic ductal adenocarcinoma (Kimura et al., 2018). ARID1A plays a key role in the regulation of DNA damage repair, by promoting an efficient processing of double-strand breaks into single-strand ends, being required to sustain DNA damage signaling and repair, hence suppressing tumorigenesis (Shen et al., 2015).

We identified several putative enhancers (zA.E1-4, Fig.2.4a), that we tested in vivo using enhancer reporter assays (Supplementary Dataset 4a). Of these, zA.E2 and zA.E4 were validated as pancreatic enhancers. zA.E4 was the most robust pancreatic enhancer of this set

(Fig.4a and Supplementary Dataset 4a), driving expression in endocrine, acinar and duct cells of the zebrafish pancreas (Fig.2.4b and Supplementary Fig.2.5a) and interacting with the promoter of *arid1ab* (Fig.2.4a and Supplementary Fig.2.5b). Additionally, we detected a human/zebrafish syntenic block containing the zebrafish zA.E4 enhancer and a human pancreatic CRE (hA.E4) (Fig.4a). In vivo enhancer assays for hA.E4 demonstrated its ability to drive expression in endocrine cells of the zebrafish pancreas, and in vitro in a human pancreatic duct cell line (hTERT-HPNE), suggesting a functional equivalency to the zebrafish zA.E4 enhancer (Fig.2.4b-c and Supplementary Fig.2.5a). To study the influence of this human enhancer on *ARID1A* expression, we deleted the hA.E4 enhancer in the hTERT-HPNE cell line, relevant for the pancreatic tumor suppressor role of *ARID1A*, through CRISPR-Cas9 system (Fig.2.4d and Supplementary Fig.2.5c-e), using a deletion in an unrelated genomic region (Miguel-Escalada et al., 2019) as a control. We observed lower levels of ARID1A upon deletion of hA.E4 compared to the control (Fig.2.4e-f and Supplementary Fig.2.5e), suggesting that the loss of this enhancer may interfere with the DNA-damage response, with possible implications in the increased risk for pancreatic cancer (Wang et al., 2019a, 2019b).

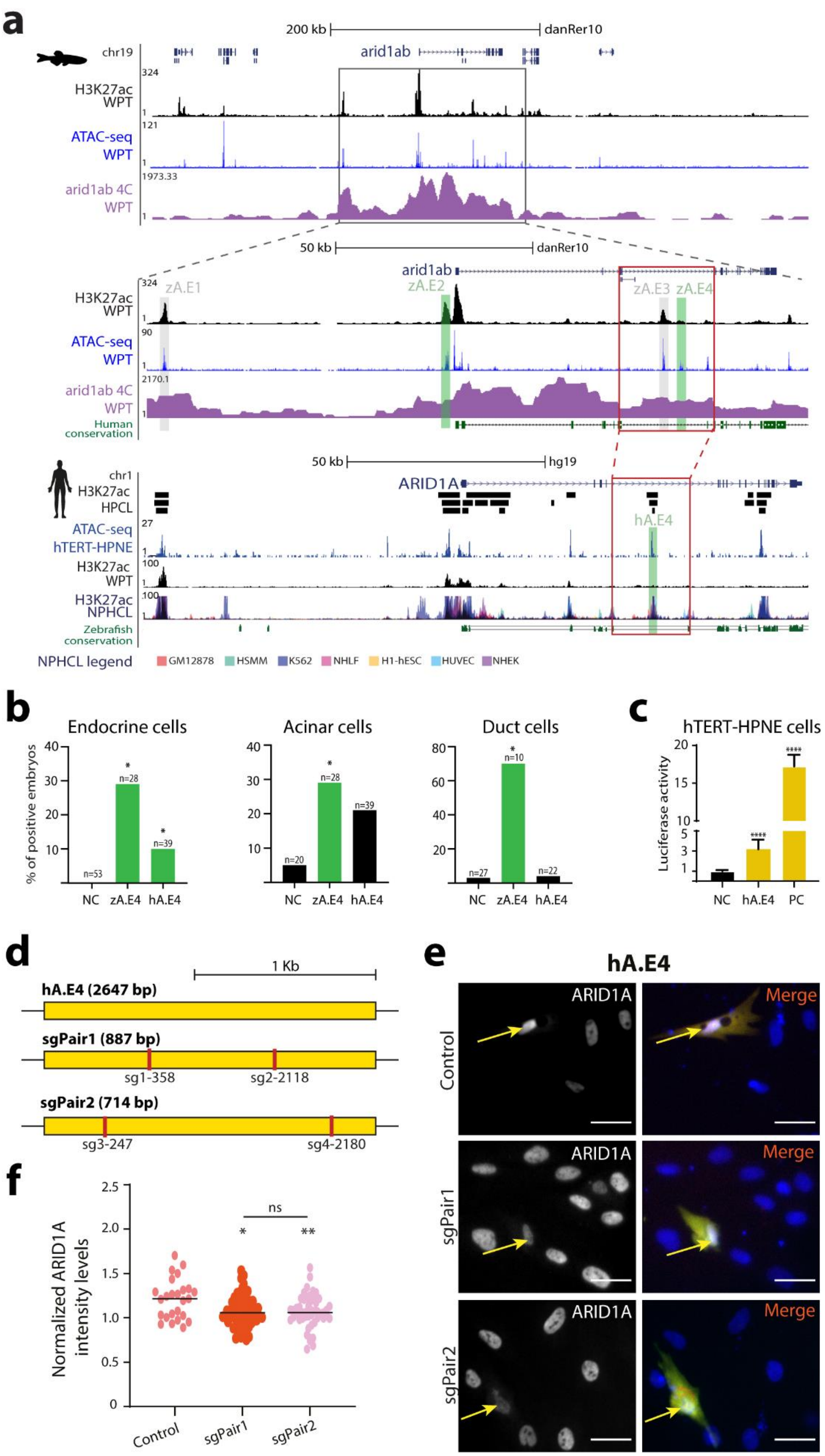

a
200 kb
danRer10
chr19
arid1ab
H3K27ac
WPT
324
1
ATAC-seq
WPT
121
1
arid1ab 4C
WPT
1973.33
1
50 kb
danRer10
arid1ab
zA.E1
zA.E2
zA.E3
zA.E4
H3K27ac
WPT
324
1
ATAC-seq
WPT
90
1
arid1ab 4C
WPT
2170.1
1
Human conservation
50 kb
hg19
chr1
ARID1A
H3K27ac
HPCL
hA.E4
ATAC-seq
hTERT-HPNE
27
1
H3K27ac
WPT
100
1
H3K27ac
NPHCL
100
1
Zebrafish conservation
NPHCL legend
GM12878
HSMM
K562
NHLF
H1-hESC
HUVEC
NHEK
b
Endocrine cells
% of positive embryos
40
30
20
10
n=53
n=28
n=39
*
*
NC
zA.E4
hA.E4
Acinar cells
40
30
20
10
n=20
n=28
n=39
*
NC
zA.E4
hA.E4
Duct cells
80
60
40
20
n=27
n=10
n=22
*
NC
zA.E4
hA.E4
c
hTERT-HPNE cells
Luciferase activity
20
15
10
5
3
1
****
****
NC
hA.E4
PC
d
1 Kb
hA.E4 (2647 bp)
sgPair1 (887 bp)
sg1-358
sg2-2118
sgPair2 (714 bp)
sg3-247
sg4-2180
e
hA.E4
ARID1A
Merge
Control
sgPair1
sgPair2
f
Normalized ARID1A intensity levels
2.5
2.0
1.5
1.0
0.5
0
ns
*
**
Control
sgPair1
sgPair2

Figure 2.4 The zebrafish and human *arid1ab/ARID1A* regulatory landscapes contain an equivalent pancreatic enhancer. **a)** Genomic landscape of the zebrafish *arid1ab* gene, showing profiles for H3K27ac ChIP-seq (black), ATAC-seq (blue) and 4C with viewpoint in the *arid1ab* promoter (magenta) in adult zebrafish pancreas (top); zoom-in in *arid1ab* regulatory landscape (middle). Human *ARID1A* genomic landscape (bottom) with H3K27ac enriched intervals from human pancreatic cell lines (HPCL, black bars, top-to-bottom: PT-45-P1, CFPAC-1 and HPAF-II), H3K27ac profile from human pancreas (WPT, black) and from non-pancreatic human cell lines (NPHCL; GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK and NHLF; Data from ENCODE). Human/zebrafish sequence conservation (dark green). Tested putative enhancers are highlighted in grey (zA.E1 and zA.E3; no enhancer activity) and green (zA.E2, zA.E4 and hA.E4; enhancer activity). Zebrafish/human syntenic box (red box). **b)** Transient in vivo enhancer reporter assays of zA.E4 and hA.E4 showing the percentage of zebrafish embryos with GFP expression in endocrine, acinar and duct cells (two-sided chi-square test with Yates correction; **p*<0.05; Endocrine cells: zA.E4, *p*=0.0001; hA.E4, *p*=0.0294; Acinar cells: zA.E4, *p*=0.0391; hA.E4, *p*=0.1167; Duct cell: zA.E4, *p*=0.00001; hA.E4, *p*=0.9731). Number of analysed embryos (n). Negative control (NC). **c)** Luciferase enhancer reporter assays performed in human hTERT-HPNE cells for hA.E4, showing luc2/Nluc ratios, relative to the negative control (two-sided t-test; *****p*<0.0001; hA.E4 *p*-value=0.0001; PC *p*-value<0.0001). Data from three biological replicates (grey dots, n=3) and Mean±SD (error bar). Negative control (NC). Positive control (PC). **d)** Strategy for CRISPR-Cas9 deletions in the hA.E4 locus, indicating sgRNA target sites. **e)** Representative images of transfected hTERT-HPNE human cells expressing pairs of sgRNAs and Cas9 (arrows). In control, sgRNAs target a H3K27ac depleted region, while sgRNAs in sgPair1 and sgPair2 target the hA.E4 locus. Left column show anti-ARID1A (grey) and right column GFP (green), mCherry (red) and DAPI (blue; nuclei). Representative images from three biological replicates. Scale bar: 40µm. **f)** Normalized ARID1A levels from immunocytochemistry images. Two-sided t-test depicted for p≤0.05(*), p≤0.01(**) and not significant (ns; *p*-values of: Control vs sgPair1=0.0208, Control vs sgPair2=0.0044, sgPair1 vs sgPair2=0.6227). A black line represents the mean of values. Data from three biological replicates. Data included in Source Data file for (**b**), (**c**) and (**f**).

2.2.4 A *ptf1a* enhancer explains pancreatic agenesis causal variant in vivo

To further evaluate the interspecies functional equivalency of enhancers and their role in human pancreatic diseases, we focused on the human *PTF1A* locus, known to be controlled by a distal downstream enhancer whose deletion causes pancreatic agenesis (Weedon et al., 2014; Fig.5a; hP.E3). Concomitantly, we detected a zebrafish distal *ptf1a* enhancer, downstream of *ptf1a* (zP.E3), as well as two previously identified proximal enhancers (zP.E1 and zP.E2; Pashos et al., 2013). zP.E3 interacts with the promoter of *ptf1a*, observed by Hi-ChIP and 4C-seq (Fig.2.5a and Supplementary Fig.2.5b), and could correspond to the functional equivalent enhancer whose deletion causes pancreatic agenesis in humans (hP.E3), although its sequence partially aligns with a more distal human sequence likely inactive in human pancreatic cells (Supplementary Fig.2.6). In vivo enhancer assays for zP.E3 and hP.E3 showed strong and robust expression in progenitor cells (Fig.2.5b), a result that is in agreement with the described activity of hP.E3 in vitro as a human developmental enhancer (Weedon et al., 2014). These results suggest that the human and zebrafish enhancers share some regulatory information. This is further supported by binding sites for FOXA2 and PDX1 in the human hP.E3, also predicted to bind to the zebrafish zP.E3 (Supplementary Fig.2.7a-b). To further evaluate the role of zP.E3, we generated genomic deletions in the zP.E3 sequence (Fig.2.5c-g, Supplementary Fig.2.8 and Fig.2.9). Deletion1, a 632bp deletion that includes the predicted Foxa2 and Pdx1 binding sites and the majority of transposase-accessible chromatin within zP.E3 (Supplementary Fig.2.9a), results in a decrease of the pancreatic progenitor domain area in homozygous mutants (Fig.2.5 c, d and f), as well as a reduction in the expression levels of *ptf1a* (Supplementary Fig.2.9b). Furthermore, after pancreatic differentiation, the Deletion1 mutants displayed pancreatic hypoplasia (Fig.2.5e and g; Supplementary Fig.2.9c-e), and we observed the same phenotype for multiple independent deletions of zP.E3 generated in somatic cells (Supplementary Fig.2.8). In contrast, no phenotypes were observed for a 517bp deletion within the zP.E3 enhancer, adjacent to Deletion1, which excludes the majority of accessible chromatin and predicted TF binding sites (Deletion2; Supplementary Fig.2.9a, d and e), suggesting that the functional core of zP.E3 coincides with the regions of available chromatin that overlap with the predicted binding of Foxa2 and Pdx1. In agreement with the observed phenotypes, pancreatic hypoplasia is compatible with the described loss-of-function of *ptf1a* in zebrafish (Pashos et al., 2013) and the loss of hP.E3 function in humans (Weedon et al., 2014). In light of these results, we suggest that pancreatic hypoplasia is the consequence of the reduction in the pancreatic progenitor domain caused by decreased levels of *ptf1a* due to the loss of an important pancreatic progenitor enhancer.

Later on, after pancreatic differentiation, zP.E3 and hP.E3 enhancers acquire distinct activity patterns. The zebrafish zP.E3 enhancer is able to drive a consistent expression in differentiated pancreatic cells from late embryos up to adults (Supplementary Fig.2.10), including acinar and duct cells, while the human hP.E3 enhancer shows almost a total lack of activity in differentiated acinar and duct cells, as previously observed in vitro (Weedon et al., 2014) driving expression only in very few cells (Supplementary Fig.2.10). Overall, these results suggest that zebrafish and humans share a functionally equivalent distal enhancer of *PTF1A* during development, whose loss-of-function results in a reduction of the pancreatic progenitor domain, elucidating, in vivo, the causal link between the disruption of this enhancer in humans and pancreatic agenesis.

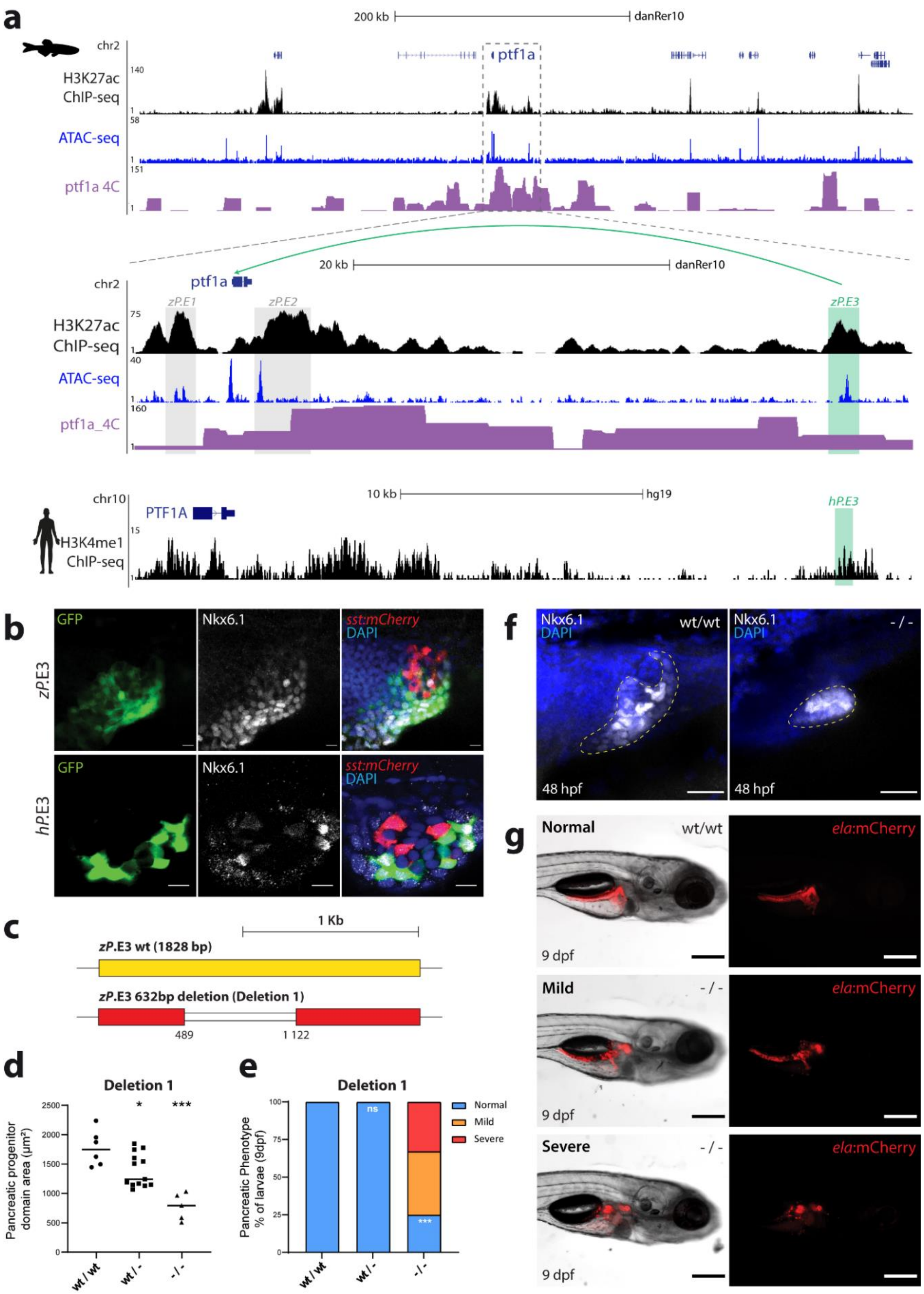

a
200 kb
danRer10
chr2
ptf1a
H3K27ac ChIP-seq
140
1
ATAC-seq
58
1
ptf1a 4C
151
1
20 kb
danRer10
chr2
ptf1a
zP.E1
zP.E2
zP.E3
H3K27ac ChIP-seq
75
1
ATAC-seq
40
1
ptf1a_4C
160
1
chr10
10 kb
hg19
PTF1A
hP.E3
H3K4me1 ChIP-seq
15
1
b
zP.E3
hP.E3
GFP
Nkx6.1
sst:mCherry
DAPI
c
1 Kb
zP.E3 wt (1828 bp)
zP.E3 632bp deletion (Deletion 1)
489
1 122
d
Deletion 1
Pancreatic progenitor domain area (µm²)
2500
2000
1500
1000
500
0
*
***
wt / wt
wt / -
- / -
e
Deletion 1
Pancreatic Phenotype % of larvae (9dpf)
100
75
50
25
0
ns
***
Normal
Mild
Severe
wt / wt
wt / -
- / -
f
Nkx6.1
DAPI
wt/wt
- / -
48 hpf
48 hpf
g
Normal
wt/wt
ela:mCherry
9 dpf
Mild
- / -
ela:mCherry
9 dpf
Severe
- / -
ela:mCherry
9 dpf

Figure 2.5 The zebrafish and human *ptf1a/PTF1A* regulatory landscapes contain a functional equivalent enhancer essential for pancreas development. **a)** UCSC Genome Browser view of the zebrafish *ptf1a* and human *PTF1A* genomic landscapes showing H3K27ac ChIP-seq (black), ATAC-seq (blue) and *ptf1a* 4C interactions (purple) from whole zebrafish pancreas samples (upper panel), with a zoom-in (middle panel), and H3K4me1 ChIP-seq data[2] (black) from human embryonic pancreatic progenitors (lower panel). Grey boxes highlight two previously validated zebrafish enhancers, zP.E1 and zP.E2 in the vicinity of the *ptf1a* gene. Green boxes highlight a distal enhancer in zebrafish, zP.E3, and the location of its putative human functional ortholog hP.E3. **b)** Confocal images of zebrafish reporter stable transgenic lines Tg(zP.E3:GFP) (n=10) and Tg(hP.E3:GFP) (n=3), showing co-localization of GFP expression (green) with Nkx6.1 (white), a marker of pancreatic progenitors, at 48hpf. Delta-cells of the endocrine pancreas express mCherry (red) and nuclei are labelled with DAPI (blue). Scale bar: 25 μm. **c)** Schematic depiction of the CRISPR-Cas9 mediated 632 bp deletion (Deletion 1) of the zP.E3 enhancer. **d)** Pancreatic progenitor domain area, defined by Nkx6.1 (white), of homozygous (-/-; n=5), heterozygous (wt/-; n=13) and wild type (wt/wt; n=6) embryos for Deletion1 of zP.E3, at 48hpf. Unpaired student's t-test (two-tailed), *p*-values<0.05 were considered significant (**p*=0.017, ****p*=0.0002). **e)** Percentage of larvae (-/-, n=12; wt/-, n=14 and wt/wt, n=12) with different pancreatic phenotypic defects (normal, mild and severe) at 9dpf. Fisher's exact test (two-sided), p-values<0.05 were considered significant (****p*=0.0003). **f)** Representative confocal images (maximum intensity projections) of the pancreatic progenitor domain (yellow dashed line) of zP.E3wt/wt (n=6) and zP.E3-/- sibling embryos (n=5) at 48hpf. Nuclei are stained with DAPI. Scale bar: 25 μm. **g)** Epifluorescence live images of representative phenotypes quantified in e). Scale bar: 250 μm. Abbreviations: ela, elastase; sst, somatostatin. For (**d**), and (**e**), source data are provided as a Source Data file**.**

2.3 Discussion

Cis-regulatory mutations and sequence variations are associated with pancreatic cancer and diabetes (Furlong and Levine, 2018b; Pasquali et al., 2014; Klein et al., 2018; Wolpin et al., 2014; Mahajan et al., 2018; Morris et al., 2012). However, the in vivo implications of these genetic changes are still unknown. Here, we explore the chromatin state of the zebrafish pancreas to uncover pancreatic enhancers and establish comparisons with humans, so that we can predict and model human pancreas disease-associated enhancers. We found that, although most of the zebrafish pancreatic enhancers do not share significant sequence identity with human pancreatic enhancers, they share many TFBS, and their target genes are enriched for human pancreas diseases. These results suggest the existence of functionally equivalent enhancers in zebrafish and humans, as proposed for other tissues and species (Khoueiry et al., 2017; Yang et al., 2015). Indeed, recent studies looking into highly divergent species as human and sponges have located similarly functional enhancers within microsyntenic regions that, although do not share significant sequence identity, clearly recapitulate similar expression patterns in enhancer reporter assays, arguing in favour of functional equivalency (Wong et al., 2020). This is likely the consequence of enhancers being fast evolving sequences operating with a high degree of sequence flexibility (Snetkova et al., 2021). Several mechanisms that may operate together during evolution can illustrate the potential for sequence flexibility of enhancers while retaining a consistent TFBS code. Among them, nucleotide alterations within similar TFBS (Deplancke et al., 2016), reshuffle of TFBSs within enhancers, compatible with a billboard model (Arnosti and Kulkarni, 2005; Buffry et al., 2016), and substitution of enhancer's sequence by acquisition of redundant enhancers within the same regulatory landscape (Eichenlaub and Ettwiller, 2011). In the current work we show several examples compatible with the potential for enhancers' sequence flexibility. Focusing on the regulatory landscape of *Arid1a*, a tumour-suppressor gene active in the pancreas (Kimura et al., 2018; Wang et al., 2019a) and other tissues (Jones et al., 2012), we show that within a microsyntenic region within the *arid1a* locus in humans and zebrafish, there are pancreatic enhancers that share regulatory information, although not sharing significant sequence identity. We further show that the deletion of the human *ARID1A* pancreatic enhancer impairs ARID1A expression, defining a locus for non-coding mutations that may increase the risk for pancreatic cancer. We further explored the potential of functional equivalency for an enhancer of *ptf1a* (Jin and Xiang, 2019), in which both zebrafish and human enhancers share regulatory information and biological requirements during pancreas development. The loss-of-function of the zebrafish enhancer results in a decrease of the pancreatic progenitor domain and ultimately in pancreatic hypoplasia, a phenotype consistent with the impact of mutations described in the human regulatory landscape, which are associated with pancreatic agenesis (Weedon et al., 2014). The reduction of the pancreatic progenitor domain in zebrafish may explain the phenotype observed in humans, contributing to the clarification of its molecular and cellular origin. Interestingly, the deletion of the zebrafish *ptf1a* enhancer does not show a complete phenotypic penetrance, with approximately 25% of the embryos having a pancreas morphologically similar to the controls, suggesting that other redundant enhancers may exist in the zebrafish regulatory landscape of *ptf1a*, compatible with a shadow enhancer identity (Kvon et al., 2021). Additionally, human and zebrafish *ptf1a* enhancers show divergent functions after differentiation. While the human enhancer shows very little activity in differentiated pancreatic cells, the zebrafish enhancer drives persistent reporter expression, suggesting that the phenotype in zebrafish after pancreatic differentiation could have the extra contribution of this late zebrafish specific function of the *ptf1a* enhancer.

Sequence conservation of CREs can be a good predictor of sequence functionality, however it holds important limitations in the prediction of equivalent functions. This is observed in the current work, where the vast majority of the zebrafish pancreatic enhancers that could be aligned to the human genome did not share marks of enhancer activity in pancreatic cells. This is further illustrated by zP.E3, which shows some partial alignment with a human sequence that has no active marks of enhancer in pancreatic cells. Many examples have been described showing how conserved sequences among divergent species might harbour divergent functions. These include differences in conserved enhancer sequences resulting in functional divergence (Ariza-Cosano et al., 2012; Vierstra et al., 2014), to more striking examples of coding exons sequences repurposed to cis-regulatory functions (Eichenlaub and Ettwiller, 2011). Additionally, recent studies have shown that the ultra-conservation at sequence level observed in some enhancers is not necessary for the maintenance of tissue specific regulatory functions, suggesting that sequence constraint may partially result from other regulatory or unknown functions (Snetkova et al., 2021).

The use of animal models to understand the role of CREs in the development of human diseases requires the identification of functionally equivalent sequences. As discussed above, sequence conservation is not a reliable predictor of functional conservation (Cooper and Brown, 2008) and functional equivalent sequences might not present high sequence conservation (Pennacchio and Visel, 2010). This problem can be partially bypassed by combining the use of biochemical marks associated to CREs activity with enhancer reporter assays to identify similar regulatory information harboured by such sequences. In the current work we used this strategy, allowing us to identify and test in vivo enhancers that, when altered, can affect the expression of disease-associated genes. This strategy can help to identify where in the genome disease-causing non-coding mutations may occur by predicting disease relevant CREs based on phenotypic description of CRE's loss-of-function. Furthermore, in the near future this strategy may be further improved by computational methods as well as the detection of TFBS in both species. These improvements could help to establish a correspondence of enhancers' identity genome wide.

The pancreas is a complex structure composed by multiple cell types. In this work we assessed the chromatin state of the whole pancreas of adult zebrafish in order to identify pancreatic CREs and their target genes. By associating CREs to the expression of target genes, we have shown that our dataset includes exocrine and endocrine CREs. This broad pancreatic enhancer map is very advantageous since it allows us to approach different biological and biomedical questions related with different pancreatic cell types. The pancreas also contains other cell types that are heavily intertwined, as is the case of endothelial cells. Indeed, several of our observations indicate the presence of endothelial enhancers in the described CREs datasets, namely the enrichment of endothelial expressing genes located nearby DevEs (Supplementary Dataset 2d-f) and the extended overlap of common motifs between pancreatic enhancers and heart ventricle enhancers.

Enhancers can be highly tissue specific, while others can be active in multiple tissues, as observed by the identification of PsE and DevE. The former showed H3K27ac profiles more restricted to the zebrafish adult pancreas, while the latter had broad profiles throughout development, suggesting their activity to be present in multiple tissues. The zP.E3 enhancer is not detected in the embryonic H3K27ac dataset, likely because its activity is highly restricted to pancreatic progenitor cells during development, resulting in its inclusion in the PsE group. A detailed analysis of the activity of this enhancer, from the larval stage to adulthood, shows it to be almost exclusively active in exocrine pancreatic cells (Supplementary Fig.10e), illustrating the expected tissue specificity of PsE enhancers.

In this work, we identified pancreatic CREs in zebrafish, a model organism that is amenable to genetic manipulation and phenotyping. By establishing a correlation between human and zebrafish pancreatic CREs, functional testing of CREs can be performed in vivo, helping to clarify the role of CREs in pancreatic function and disease. In summary, the combination of techniques used in this work, allowed the identification of human cis-regulatory elements involved in disease. We show that transcriptional cis-regulation of the human and zebrafish adult pancreas have a high degree of similarity, allowing the functional exploration of cis-regulatory sequences in zebrafish, with the potential of translation to human pancreatic diseases.

2.4 Materials and Methods

2.4.1 Experimental procedures

2.4.1.1 Zebrafish stocks, husbandry, breeding and embryo rearing

Adult zebrafish AB/TU WT strains were obtained from the Gomez-Skarmeta's laboratory in Seville (CABD). WT, transgenic and mutant lines were maintained at 26-28ºC under a 10h dark/14h light cycle in a recirculating housing system according to standard protocols (Westerfield, M, 2000). Embryos were grown at 28ºC in E3 medium [5mM NaCl (#S/3161/60, Fisher Chemical), 0.17mM KCl (#2676.298, VWR), 0.33mM $CaCl_2•2H_2O$ (#C3881, Sigma-Aldrich), 0.33mM $MgSO_4•7H_2O$ (#63140, Sigma-Aldrich) and 0.01% methylene blue (#66120, Sigma-Aldrich), pH 7.2] or E3 supplemented with 0.01% PTU (1-phenyl-2-thiourea, #P7629, Sigma-Aldrich; Ishibashi et al., 2013). For the in vivo enhancer assays, embryos were anesthetized by adding tricaine (MS222; ethyl-3-aminobenzoate methanesulfonate, #E10521-10G, Sigma-Aldrich) to the medium and selected by the internal positive control of transgenesis. For the establishment of transgenic and mutant zebrafish lines, embryos were microinjected, selected, bleached, and grown until adulthood. Adult F0s were outcrossed with WT adults and the offspring screened for the internal control of transgenesis and the pattern of expression of the regulatory element, or for the respective mutations, by genotyping. In vivo reporter lines, Tg(ela:mCherry) and Tg(sst:mCherry), were used to label the exocrine and endocrine domain, respectively. The i3S animal facility and this project were licensed by *Direcção Geral de Alimentação e Veterinária (DGAV)* and all the protocols used for the experiments were approved by the i3S Animal Welfare and Ethics Review Body.

2.4.1.1 Cell culture

hTERT-HPNE (ATCC CRL-4023) cells were cultured in a 5% CO_2-humidified chamber at 37ºC in DMEM (1x, 4.5 g/L D-glucose with pyruvate; #D6429, Gibco, ThermoFisher Scientific), supplemented with 10% fetal bovine serum (#BCS0615, biotecnomica), 10ng/mL human recombinant EGF (#11343406, Immunotools) and 750ng/mL puromycin (#P8833-25MG, Sigma-Aldrich) in TC Dish 100 (SARSTEDT). When cells reached 90% of confluence, they were split using TrypLE Express (#12604-021, Gibco, ThermoFisher Scientific; approximately 0.5 mL per 10 cm2).

2.4.1.2 ChIP-seq

Whole pancreas was dissected from 25 adult zebrafish (~50x10^6 cells; both genders and with 12-24 months), kept on ice in PBS [137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)] with 1x Complete Proteinase Inhibitor (#11697498001, Roche), fixed in 2% formaldehyde (#F1635-500ML, Sigma-Aldrich) for 10 min, and stored at -80ºC. ChIP was performed as previously described for zebrafish embryos (Wittkopp and Kalay, 2011) with minor alterations. Cell lysis was performed on ice, using a 15 mL Tenbroeck Homogenizer, in cell lysis buffer [10mM Tris-HCl pH7.5 (Tris Base #BP152-1, Fisher bioreagents, HCL #20255.290, VWR), 10mM NaCl (#S/3161/60, Fisher Chemical), 0.5% NP-40 (#85124, ThermoFisher Scientific), 1x Complete Proteinase Inhibitor (#11697498001, Roche)] for 15 min. Nuclei were washed and re-suspended in nuclei lysis buffer (50mM Tris-HCl pH7.5 (Tris Base #BP152-1, Fisher bioreagents, HCL #20255.290, VWR), 10mM EDTA (#20301.290, VWR), 1% SDS (#MB11601, NZYTech), 1x Complete Proteinase Inhibitor (#11697498001, Roche)). Chromatin was sheared using a BioruptorPlus (Diagenode) device with the following cycling conditions: 10 min high–30 sec on, 30 sec off; 15 min on ice; 10 min high–30 sec on, 30 sec off. The sonicated chromatin had a size in the range of 100–500 bp and was incubated overnight at 4ºC with the anti-H3K27ac antibody (1:2, #ab4729, Abcam). Samples were incubated for 1h at 4ºC with Dynabeads Protein G for Immunoprecipitation (#10003D, Invitrogen, ThermoFisher Scientific). Final DNA was purified with MinElute (#28004, Qiagen) and sequenced on Illumina HiSeq 2000 platform.

2.4.1.3 ATAC-seq

ATAC-seq was performed as previously described (Fernández-Miñán et al., 2016), with minor changes. Whole pancreas was dissected from 2-3 adult zebrafish (both genders and with 12-24 months). Following cell lysis, 50000-100000 nuclei were submitted to tagmentation with Nextera DNA Library Preparation Kit (#FC-121-1030, Illumina). ATAC-seq libraries were amplified using KAPA HiFi HotStart PCR Kit (#KK2500, Roche) with the primers Ad1, Ad2.2 and Ad2.3 (Buenrostro et al., 2013), and further purified with PCR Cleanup Kit (#28104, Qiagen).

2.4.1.4 4C-seq

4C-seq was performed as previously described (Fernández-Miñán et al., 2016), with minor alterations. Whole pancreas was dissected from 6-12 adult zebrafish (7-15x10^6 cells; both genders and with 12-24 months), kept on ice in PBS [137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)] with 1x Complete Proteinase Inhibitor (#11697498001, Roche), fixed in 2% formaldehyde (#F1635-500ML, Sigma-Aldrich) for 10 min, and stored at -80ºC. Cell lysis was performed on ice, with a 15 mL Tenbroeck Homogenizer, not exceeding 10 min. Ligation was performed with 60U T4 DNA Ligase (#EL0012, ThermoFisher Scientific). The restriction enzymes used were DpnII (#R0543M, NEB) and Csp6I (#ER0211, ThermoFisher Scientific) for the first and second cuts, respectively. Chromatin was purified by Amicon Ultra-15 Centrifugal Filter Device (#UFC901024, Milipore). 4C libraries were prepared for Illumina sequencing by the Expand Long Template Polymerase (#11759060001, Roche) with primers targeting the TSSs of each gene and including Illumina adapters (Supplementary Dataset 4c). Final PCR products were purified with the High Pure PCR Product Purification Kit (#11796828001, Roche) and AMPure XP PCR purification kit (#B37419AB, Agencourt AMPure XP).

2.4.1.5 HiChIP-seq

HiChIP-seq was performed as previously described (Mumbach et al., 2016), with minor alterations. Whole pancreas, from both genders and with 12-24 months, was dissected, fixed in 1% formaldehyde (#F1635-500ML, Sigma-Aldrich) and cells lysed as described for 4C-seq. Immediately after lysis, samples were washed with HiChIP Wash Buffer [Tris-HCl pH 8 50mM (Tris Base #BP152-1, Fisher bioreagents, HCL #20255.290, VWR), NaCl 50 mM (#S/3161/60, Fisher Chemical), EDTA 1 mM (#20301.290, VWR)]. Chromatin was sonicated using the BioruptorPlus (Diagenode) with the following cycling conditions: 10 min high–30 sec on, 30 sec off; 15 min on ice, to obtain a size in the range of 100–500 bp. Samples were incubated with anti-H3K4me3 antibody (1:5, #AB8580, Abcam) and Dynabeads Protein G for Immunoprecipitation (#10003D, Invitrogen, ThermoFisher Scientific) and purified with DNA Clean and Concentrator columns (#D4004, Zymo Research). Up to 150 ng of the DNA was then biotinylated with Streptavidin C-1 beads (#65001, ThermoFisher Scientific). Tagmentation was performed using Nextera DNA Library Preparation Kit (#FC-121-1030, Illumina). Libraries were amplified using NEBNext® High-Fidelity 2X PCR Master Mix (#M0541S, NEB) with primers Ad1, Ad2.23 and Ad2.24(Buenrostro et al., 2013). The final product was purified with DNA Clean and Concentrator kit (#D4004, Zymo Research).

2.4.1.6 Generation of plasmids for enhancer assays

Putative enhancer sequences were selected based on the overlap between H3K27Ac ChIP-seq and ATAC-seq signal in non-coding regions within the landscape of each pancreas-relevant gene. Sequences were PCR amplified from zebrafish genomic DNA using the primers in Supplementary Dataset 4b (designed to span the ChIP-seq and ATAC-seq signals) (Sigma-Aldrich), with the proof-reading iMax ™ II DNA polymerase (#25261, INtRON Biotechnology) following the manufacturer's instructions for a standard 20µl PCR reaction. PCR products were visualized by electrophoresis on an 1% agarose gel, the bands excised, purified with NZYGelpure kit (#MB011, NZYTech) and cloned into the entry vector pCR®8/GW/TOPO (#250020 Invitrogen, ThermoFisher Scientific) according to manufacturer's instructions. The vectors were then recombined into the destination vectors Z48(de la Calle-Mustienes et al., 2005), for transient enhancer assays, and ZED (Bessa et al., 2009, 2014), for stable transgenic lines, using Gateway® LR Clonase® II Enzyme mix (#11791020, Invitrogen, ThermoFisher Scientific), following manufacturer's instructions.

Standard chemical transformation was performed with MultiShotTM FlexPLate Mach1™ T1R (#C8681201, Invitrogen, ThermoFisher Scientific), grown O.N. at 37ºC. Vector selection was performed with 100 µg/ml Spectinomycin (#S4014, Sigma-Aldrich) in the growth medium for the pCR®8/GW/TOPO vectors, or 100 µg/ml Ampicillin (#624619.1, Normon) for the Z48 and ZED vectors. Plasmids were purified with NZYMiniprep kit (#MB010, NZYTech) and confirmed by Sanger sequencing using the primers in Supplementary Dataset 4b. Final plasmids were purified with phenol/chloroform (#A931I500 and #C/4920/15, Fisher Chemical) and concentration was determined by NanoDrop 1000 Spectrophotometer (ThermoFisher Scientific).

2.4.1.7 In vitro mRNA synthesis, Microinjection and Transgenesis

Z48 and ZED zebrafish lines were generated through TOL2-mediated transgenesis (Kawakami et al., 2004). TOL2 cDNA was transcribed by Sp6 RNA polymerase (#EP0131, ThermoFisher Scientific) after Tol2-pCS2FA vector linearization with NotI restriction enzyme (#IVGN0016, Anza, Invitrogen, ThermoFisher Scientific). TOL2 mRNA was purified as previously described (Bessa et al., 2009). One-cell stage embryos were injected with 1nL solution containing

25ng/µL of transposase mRNA, 25ng/µL of phenol/chloroform (#A931I500 and #C/4920/15, Fisher Chemical) purified plasmid (Z48 or ZED), and 0.05% phenol red (#P0290, Sigma-Aldrich).

2.4.1.8 Luciferase reporter assays

The h.A.E4 enhancer were cloned in the pGL4.23GW[luc2/minP] vector (Addgene #60323) and co-transfected along with pNL1.1PGK[Nluc/PGK] (Promega #N1441) in hTERT-HPNE cells using Lipofectamine 3000 (#L3000008, ThermoFisher), following manufacturer's instructions. The promoter of tyrosine kinase was cloned into the pGL4.23GW[luc2/minP] vector and used as positive control (pGL4.23GW[luc2/Tkp]; Vaz et al., 2021). As negative control, a region without marks of enhancer activity (H3K27ac) was cloned into the pGL4.23GW [luc2/minP] vector. The luciferase activity was measured 48 hours post transfection with the Nano-Glo Luciferase Assay System (#N1610, Promega) on a Synergy 2 microplate reader (BioTek). Results were presented as luc2/Nluc ratios, relative to the negative control. Two-sided t-test was used to calculate statistical significance. Three independent replicates of the transfection were performed.

2.4.1.9 Cas9 target design, sgRNA synthesis and mutant generation

Small guide RNAS (sgRNAs) targeting regions flanking zP.E3 were designed using the CRISPRscan algorithm (Moreno-Mateos et al., 2015) to include H3K27ac ChIP-seq and ATAC-seq signal (Supplementary Dataset 4f). Oligonucleotides (1.5µL at 100 µM each, from Sigma-Aldrich) were annealed in vitro by incubation at 95ºC for 5 min in 2x Annealing Buffer [10mM Tris, pH7.5-8.0 ((Tris Base #BP152-1, Fisher bioreagents, HCL #20255.290, VWR), 50mM NaCL (#S/3161/60, Fisher Chemical), 1mM EDTA (#20301.290, VWR)] followed by slow cooling at RT, and inserted into 100ng of pDR274 vector (#42250, Addgene) previously cut with BsaI (#IVGN0366, Anza, Invitrogen, ThermoFisher Scientific; 1:10). The pDR274 vectors carrying sgRNA sequences were linearized with HindIII (#IVGN0168, Anza, Invitrogen, ThermoFisher Scientific; 1:10), purified with phenol/chloroform (#A931I500 and #C/4920/15, Fisher Chemical) and transcribed with T7 RNA polymerase (#EP0111, ThermoFisher Scientific). Final sgRNAs were purified as described previously (Bessa et al., 2009). One cell-stage zebrafish embryos were co-injected with two sgRNAs (40 ng/µl each) and Cas9 protein (300 ng/µl; #CP01-50 PNA Bio, Inc). Zebrafish mutant lines for zP.E3 deletion were generated using the combinations sgRNA1+sgRNA2 (sgPair1) and sgRNA3+sgRNA2 (sgPair2;

Supplementary Dataset 4f). Enhancer deletions in zebrafish were detected with PCR using HOT FIREPol DNA Polymerase (#01-02-00500, Solis BioDyne) with the flanking primers used to amplify the enhancers (Supplementary Dataset 4b). PCR products were visualized by electrophoresis in 2% agarose gel and confirmed by Sanger sequencing. The mutations were further verified in the F1 mutants by sequencing.

2.4.1.10 CRISPR-Cas9 in human cell lines

Four single-guide sequences named sg1, sg2, sg3, sg4, targeting hA.E4 enhancer were designed (Supplementary Dataset 4f). sg1 and sg3 were designed upstream of the enhancer, while sg2 and sg4 were designed downstream, based on H3K27ac ChIP-seq and ATAC-seq signal. Two complementary oligonucleotides containing the single-guide sequences and BbsI ligation adapters were synthesized by Sigma. Two single-guide sequences designed to delete a genomic region lacking enhancer activity marks (based on H3K27ac), named ng1 and ng2, were used as negative control of the experiment (Miguel-Escalada et al., 2019). Oligonucleotides were annealed in T4 Ligation Buffer (ThermoFisher Scientific). sgRNA was cloned into the BbsI-linearized pSpCas9-T2A-GFP (#R3539S, NEB; #48138, Addgene; sg1, sg3, ng1) and pU6-(BbsI)CBh-Cas9-T2A-mCherry (#64324, Addgene; sg2, sg4, ng2) vectors using T4 Ligase (ThermoFisher Scientific). The plasmid DNA was purified with Plasmid Midi Kit (#12143, Qiagen).

hTERT-HPNE cells were seeded in 6-well plates (1.1x10^5 cells/well, at early passage number) and transfected (~70-90% of confluency) using the following combinations: ng1+ng2 (control); sg1+sg2 (sgPair1); sg3+sg4 (sgPair2). The transfection (1.5 µg of each sgRNA plasmid) was performed using Lipofectamine 3000 (#L3000008, ThermoFisher Scientific), according to the manufacture instructions. Then, cells were changed to fresh culture medium after 24 h. Three independent replicates of the transfection were performed. After 48h of recovery, cells were used in subsequent experiments.

2.4.1.11 Nucleic acid extraction from zebrafish and human cell lines

Genomic DNA was extracted from whole zebrafish embryos at 24 hpf, after removal of the chorion, with a standard phenol-chloroform DNA extraction (#A931I500 and #C/4920/15, Fisher Chemical), and used as template for PCR amplification in order to genotype the tested conditions (Supplementary Dataset 4b). The DNA samples were resuspended in 20 µl of TE buffer with RNase [10mM Tris, pH 8.0 (Tris Base #BP152-1, Fisher bioreagents, HCl

#20255.290, VWR); 1mM EDTA pH 8.0 (#20301.290, VWR) and 100 µg/ml RNAse (#10109142001, Sigma-Aldrich)], incubated for 1 hour at 37ºC, and stored at -20ºC.

Genomic DNA from hTERT-HPNE cells was extracted 48h after transfection and used as template for PCR amplification in order to genotype the tested conditions (Supplementary Dataset 4b).

RNA was extracted from zebrafish embryos, pancreas and muscle, with 500µl TRIzol (#15596026, Invitrogen, ThermoScientific), following the manufacturer's instructions. Samples were incubated 30min at 37ºC with 1 µl DNAse I (#EN0521, ThermoScientific), 1µl 10x reaction buffer and 0.5µl NZY Ribonuclease Inhibitor (40U/µl; # MB084, NZYTech) at 0.05µl/µl final concentration. After adding 1µl EDTA (#20301.290, VWR) 50mM per 1µg of estimated RNA, final volume was completed to 60µl with H2O, phenol-chloroform (#A931I500 and #C/4920/15, Fisher Chemical) standard purification was performed and the RNA stored at -80ºC.

Zebrafish pancreatic progenitor cells were extracted from 48hpf embryos, immediately following euthanasia by rapid chilling, by repeated pipetting up and down in a gentle motion with 300 µL of Ginzburg fish Ringer's solution [55mM NaCl (#S/3161/60, Fisher Chemical), 1.8mM KCl (#2676.298, VWR), 1.25mM NaHCO3 (# S5761, Sigma-Aldrich)]. Embryos were allowed to settle to the bottom and the suspension containing the detached pancreatic progenitor cells and yolk was collected, washed with PBS [137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)], and RNA was extracted using Quick-RNA Microprep Kit (#R10150, Zymo Research), according to the manufacturer's instructions. For Real-time qPCR, RNA samples were treated with DNaseI (#EN0521, ThermoScientific) and reverse transcribed using the iScript cDNA Synthesis Kit (#1708890, Bio-Rad) according to the manufacturer's instructions.

2.4.1.12 Immunohistochemistry in zebrafish embryos and human cell lines

Zebrafish embryos/larvae were euthanized by prolonged immersion in 200-300 mg/L tricaine (MS222; ethyl-3-aminobenzoate methanesulfonate, #E10521-10G, Sigma-Aldrich). Whenever necessary the chorion was removed, and the zebrafish were fixed in formaldehyde 4% (#F1635-500ML, Sigma-Aldrich) for 1h at RT (8-12dpf larvae) or O.N. at 4ºC (48hpf embryos). Permeabilization was carried out by incubation with 1% Triton X-100 (#X100, Sigma-Aldrich) in PBS [137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)] for 1h at RT, followed by blocking with 5% bovine serum albumin (BSA; #MB04602, NZYTech) in 0.1% Triton X-100 (#X100, Sigma-Aldrich) for 1h at RT. Zebrafish were incubated with the primary antibody diluted in blocking solution at 4ºC O.N., and then incubated with the secondary antibody plus DAPI (1:1000, D1306 Invitrogen, ThermoFisher Scientific) diluted in blocking solution for 4 hours at RT. After each antibody incubation, embryos were washed 6 times in PBS-T (0.5 % Triton X-100 (#X100, Sigma-Aldrich) in PBS-1x[137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)] 5 minutes at RT. Embryos were stored in 50% Glycerol/PBS (#BP229-1, Fisher bioreagents) at 4ºC before microscopy slides preparation in the mounting medium 50% Glycerol/PBS; (#BP229-1, Fisher bioreagents). Images were acquired with a Leica TCS SP5 II confocal microscope (Leica Microsystems, Germany; LAS AF software (v.2.6.3.8173) and processed by ImageJ software (v.1.8.0). Primary antibodies: rabbit anti-Amylase (1:50, #A8273-1VL, Sigma-Aldrich), mouse anti-Alcam (1:50, #ZN-8, DSHB) and mouse anti-Nkx6.1 (1:50, #F55A10, DSHB). Secondary antibodies: goat anti-mouse AlexaFluor647 (1:800, #A-21236 Invitrogen, ThermoFisher Scientific), goat anti-rabbit AlexaFluor568 (1:800, #A-11036 Invitrogen, ThermoFisher Scientific).

The hTERT-HPNE cells were fixed at 48h after transfection in formaldehyde 4% (#F1635-500ML, Sigma-Aldrich) in PBS [137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)] for 15 min at RT, permeabilized with 1% Triton X-100 (#X100, Sigma-Aldrich) in PBS and blocked with 2% BSA (#MB04602, NZYTech) in PBS for 20 min at RT. Incubation with primary antibody in 2% BSA/PBS (#MB04602, NZYTech) was O.N. at 4ºC and in secondary antibody plus DAPI (1:1000, D1306 Invitrogen, ThermoFisher Scientific) was 3h at 4ºC in 2% BSA/PBS (#MB04602, NZYTech) for 3h. Human cells were washed once after fixation and permeabilization, and three times after each incubation with primary and secondary antibodies with PBS for 10 minutes at RT. Fluorescence images were obtained at 40x magnification on a Leica DMI6000 FFW microscope (v.3.7.4.23463). Primary antibody used: anti-ARID1A (1:1000; #HPA005456 Sigma-Aldrich). Secondary antibody used: anti-rabbit Alexa Fluor 647 (1:1000, #A31573, ThermoFisher Scientific). In hTERT-HPNE immunohistochemistry images, the ARID1A nuclear staining was measured for each cell GFP+/mCherry+ and normalized for the average staining of the nucleus of all other cells in the same field (ratio=ARID1A expression/mean of ARID1A expression in the field). Then, we normalized the ratios using the control values.

2.4.1.13 Flow Cytometry

The whole pancreases were dissected from double transgenic adult zebrafish [Tg(ins:GFP, ela:mCherry), Tg(ins:GFP, gcga:mCherry), and Tg(ins:GFP, sst:mCherry)] and fixed using 4% formaldehyde (#F1635, Sigma-Aldrich) in 1xPBS [137mM NaCl (#S/3161/60, Fisher Chemical), 2.7mM KCl (#2676.298, VWR), 10mM NaHPO4 (#1.06342.0250, Merk), and 1.8 mM KH2PO4 (#1.06585.1000, Merk)]. Cells were dissociated, on ice, using a 15 mL Dounce homogenizer in 1 mL of ice-cold sort buffer [1% EDTA (#20301.290, VWR), 2mM HEPES (#83264, Sigma-Aldrich) pH 7.0 in 1xPBS, and then passed through a 40-µm cell strainer. Immediately following dissociation, the mCherry and GFP fluorescence were analysed on a BD FACS-ARIATM II cell sorter (BD Biosciences).

2.4.1.14 Statistical Analysis

Two-tailed paired Student's t-test was applied to area quantifications, and in expression analysis. Chi-square test was applied to the in vivo validation of selected putative pancreatic enhancers and TFs motif comparisons. Wilcoxon test was applied to gene-to-enhancer association by chromatin interaction points comparisons. Fisher's exact test was applied to analyse the percentage of larvae in each phenotypic class. In all analyses, *p*<0.05 was required for statistical significance and calculated in GraphPad Prism 5 (San Diego, CA, USA).

2.4.2 Processing and Bioinformatic analysis

2.4.2.1 ChIP-seq analysis

High quality raw reads for the two replicates of H3K27ac ChIP-seq (FASTQC v.0.11.5(Andrews, S, 2010), Supplementary Data 1 and 2) were aligned to the zebrafish genome (GRCz10/danRer10) using Bowtie2 (v.2.2.6) with default settings (Langmead and Salzberg, 2012). Before the alignment, the sequencing adapters were removed from the raw reads applying Skewer (v.0.2.1; Jiang et al., 2014). The alignment file was converted into a bed file (Bedtools v.2.27; Quinlan and Hall, 2010) and the data extended 300 bp, bigwig tracks generated and uploaded to UCSC Genome Browser (Fig.2.1b). Highly enriched regions (peaks) were obtained by MACS14 (v.1.4.2) with the parameters "--nomodel, --nolambda and --space=30" (Zhang et al., 2008). During the ChIP-seq analysis the two replicates were processed independently. Reproducibility of the two biological replicates was measured by Pearson's correlation coefficient (Bailey et al., 2013) in R. The same pipeline was applied to analyse human dataset from the ENCODE project (https://www.encodeproject.org/):

ENCSR340GAZ; ENCSR748TFF. Regarding the embryo ChIP-seq datasets from the work by Bogdanovic and colleagues (Bogdanovic et al., 2012), the data processed by the authors was used.

2.4.2.2 Identification of putative enhancers

To identify the best putative active enhancers in the zebrafish adult pancreas, we intersected the peaks from the two H3K27ac ChIP-seq replicates, generated by peak calling, selecting only the enriched regions present in both replicates (Bedtools intersect v.2.27 with the default parameters (Quinlan and Hall, 2010). Since H3K27ac is also present in promoter regions, we excluded peaks overlapping with TSS by intercepting our set of putative active enhancers with the TSS coordinates (Bedtools intersect with the parameter "-v"). To determine the presence of unreliable peaks, a "blacklist" was generated using H3K27ac ChIP-seq of different zebrafish tissues to identify putative false positive peaks. The used datasets from the DANIO-CODE consortium were the following(https://danio-code.zfin.org).: DCD002894SQ, DCD002921SQ, DCD003653SQ, DCD003654SQ, DCD003671SQ and DCD002742SQ. Then, MACS software was performed in these datasets using the same parameters described in the last section and the peaks that were present in at least 5 out 6 datasets were selected. This analysis generated 156 peaks, from which 102 overlapped with 69 peaks from the list of 14753 putative pancreatic enhancers, representing less than 0,5% of the total dataset. We have used a published human "blacklist" of unreliable peaks (Amemiya et al., 2019) and observed that these represent 192 out of 102548 of the human pancreas H3K27ac ChIP-seq called peaks (0.2% of the identified peaks). The zebrafish and human "backlist" of peaks is included in Supplementary Dataset 1o and annotated in Supplementary Dataset 1a.

The genomic distribution of putative enhancers was performed using the annotatePeaks.pl module of HOMER (v.4.11.1; Heinz et al., 2010; Fig.2.1c). The adult pancreas putative active enhancer dataset (PsE+DevE) was crossed with the H3K27ac zebrafish embryonic dataset (dome, 80% epiboly, 24 hpf and 48 hpf) (Supplementary Dataset 4g; Bogdanovic et al., 2012) to identify enriched regions present only in adult pancreas (PsE; Fig.2.1d). All genomic intersections were performed using Bedtools "intersect" (Quinlan and Hall, 2010). We superimposed the H3K27ac mapped reads from adult pancreas and the embryonic dataset with the adult pancreas H3K27ac peaks using seqMINER (v1.3.4) with default settings (Fig.2.1d), showing read densities ±5 kb from the acetylation peak center (Zhan and Liu, 2015). Gene enrichment and functional annotation of our dataset were obtained with GREAT (v.3.0.0;

Hiller et al., 2013; McLean et al., 2010), using the basal plus extension association rule (proximal: 5kb upstream, 1kb downstream, plus distal: up to 1000 kb (Supplementary Fig.2.2b).

2.4.2.3 ATAC-seq analysis

High quality raw reads for the two replicates of pancreas ATAC-seq (FASTQC v.0.11.5; Andrews, S, 2010) were trimed for adapter sequences using Skewer (v.0.2.1; Jiang et al., 2014). All libraries were sequenced on Illumina HiSeq 2500 platform and raw reads were mapped to the reference zebrafish genome (GRCz10/danRer10) using Bowtie2 (v.2.2.6) with parameters "-X 2000 and --very-sensitive" (Langmead and Salzberg, 2012). To avoid clonal artefacts, the duplicated mapped reads were removed using Samtools (v.1.9; Li et al., 2009). Mapped reads were filtered by the fragment size (≤120 bp) and mapping quality (≥10). For a better visualization, data were extended 10 bp, generated bigwig tracks and uploaded to the UCSC browser (Fig.2.1b). To call for enriched regions, MACS2 (v.2.1.0; Zhang et al., 2008) was used with the parameters "--nomodel, --keep-dup 1, --llocal 10000, --extsize 74, --shift –37 and -p 0.07". For the ATAC-seq analysis, the two replicates were processed independently. Reproducibility of the biological replicates was measured using the Pearson's correlation coefficient (Bailey et al., 2013) in R. Then, we applied the Irreproducible Discovery Rate (IDR, v.2.0.4) in order to obtain a confident and reproducible set of peaks (Li et al., 2011). The same pipeline was applied to analyse human dataset from ENCODE project (https://www.encodeproject.org/; ENCSR340GAZ; ENCSR515CDW) and ATAC-seq dataset from the work by Bogdanovic and colleagues (Bogdanovic et al., 2012).

2.4.2.4 4C-seq analysis

4C-seq libraries were first inspected for quality control using FASTQC (Andrews, S, 2010; v.0.11.5, Supplementary Data 3-5) and demultiplexed using the script "demultiplex.py" from the FourCSeq package (Klein et al., 2015), allowing for 1 mismatch in the primer sequence. 4C-seq data were analysed as previously described (Noordermeer et al., 2011; Splinter et al., 2012). Briefly, reads were aligned to the zebrafish genome (GRCz10/danRer10) using Bowtie (v.1.1.2; Emera et al., 2016), keeping only uniquely mapping reads (-m 1). Reads within fragments flanked by restriction sites of the same enzyme or if fragments smaller than 40 bp were filtered out. In addition, reads falling ±5kb from the viewpoint were filtered out. Mapped reads were then converted to reads-per-first-enzyme-fragment-end units, and smoothed using a 30 fragment mean running window algorithm (Fig2.4a and 2.5a).

2.4.2.5 HiChIP-seq analysis

H3K4me3 HiChIP-seq analysis from paired-end fastq files to pairs of interacting chromatin fragments were performed using a custom python script based on the default function of the pytadbit python library (Serra et al., 2017). This library first uses GEM mapper (v.3.6; Marco-Sola et al., 2012) to map paired reads independently to the zebrafish reference genome (GRCz10/danRer10, flags used by GEM mapper --max-decoded-strata 1; --min-decoded-strata 0; -e 0.04). Then, reads are associated to a particular restriction fragment and paired together according to their read names. Once the reads are paired, the pairs of reads are filtered so that only those belonging to different restriction fragments are kept. Compressed sparse matrix files in cooler and hic formats were generated from those filtered reads using Cooler ("cload pairix" utility) and Juicer tools ("pre" utility) respectively for both visualization and further analysis. From the hic file we obtained contact matrices detailing the coordinates of 2 interacting 5kb chunks and the respective number of interactions, using Juicer tools ("dump" utility) and filtering for ≥2 interactions between chunks ≤100kb apart. To predict the target promoters of putative active enhancers, only contacts connecting zebrafish pancreas active TSSs and putative active enhancers given by H3K27ac ChIP-seq peaks from whole pancreas, adult pancreas (PsE), developing pancreas (DevE) and the different enhancer clusters (C1-C4) were selected. An output table was produced with genes targeted by enhancers, per enhancer cluster (Supplementary Dataset 3a-g). Custom scripts are provided in a GitLab repository (https://gitlab.com/rdacemel/pancreasregulome).

2.4.2.6 Identification of active promoters

H3K4me3 sequencing datasets (2 replicates performed in the HiChIP assay; Supplementary Data 6-9) were aligned to the zebrafish genome (GRCz10/danRer10) using Bowtie2 (v.2.2.6) with default settings. Highly enriched regions (peaks) were obtained by MACS14 (v.1.4.2) algorithm with the parameters "--nomodel, --nolambda and --space=30" (Zhang et al., 2008). Then, the peaks present in both replicates were filtered with the transcription start site (TSS) position to identify the active promoters using Bedtools "intersect" (v.2.27; Quinlan and Hall, 2010).

2.4.2.7 RNA-seq analysis

Total RNA extracted from adult zebrafish (exocrine, endocrine and muscle) and sequenced on Illumina HiSeq 2000 platform was inspected for quality control using FASTQC (Andrews, S, 2010) (v.0.11.5, Supplementary Data 10-17). Then, sequences were trimmed to remove adaptors, sequencing artefacts and low-quality reads (Q<20; Gordon A, Hannon G., 2003). The BWA-MEM software (v.0.7.17) was used to map the clean reads to the reference genome (ZV9/danRer7) with the parameters "-w 2 and -c 3" (Li and Durbin, 2009). Gene expression was measured from the mapped reads using HT-seq-count (v0.9.0; Anders et al., 2015). In addition, two public RNA-seq datasets were used (Supplementary Dataset 4g).

2.4.2.8 Gene expression barplots

The average expression of genes associated with each enhancer cluster (PsE, DevE, C1-C4), as defined by HiChIP, was compared to the average expression of all genes present in the RNA-seq datasets using R and ggplot for drawing barplots (Fig.2.2a, Supplementary Fig.2.2c, Supplementary Dataset 3h, Fig.2.2a R in https://gitlab.com/rdacemel/pancreasregulome).

2.4.2.9 Identification of Human/zebrafish syntenic blocks

Human/zebrafish syntenic blocks were defined by two aligned regions between both species that kept their relative position among each other. Pre-existing alignments available in the UCSC genome browser were used. Then, enhancers were searched within these blocks in both species.

2.4.2.10 Conservation between zebrafish and human and PhastCons scores

To obtain the percentage of zebrafish putative active enhancers conserved with human, the coordinates of putative active enhancers from adult zebrafish pancreas and embryos at different development stages (GRCz10/danRer10) were used as input to the UCSC genome coordinate conversion tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver, liftover (v.1.04.00) to hg19, October 2019; Fig.2.3a). To visualize the conservation of the respective sequences, liftOver (v.1.04.00) to hg38 was done and their average PhastCons conservation score plotted (Fig.2.3b). For this, we downloaded PhastCons scores in bigWig format from a 100-way multiple species alignment of 99 vertebrates against human (hg38; hg38.phastCons100way.bw, October 2019; Siepel et al., 2005) and converted to BedGraph text format using the UCSC's utility *bigWigToBedGraph* (v.1.04.00). Then, the Bedtools (Quinlan and Hall, 2010) suite (v.2.27) was used to intersect and map different putative enhancer clusters in bed format with the conservation scores, storing for each putative enhancer the median and average PhastCons score. To know which of them overlap putative active enhancers in human pancreas, we used the Bedtools “intersect” tool with default ≥1 bp of overlap (Fig.2.3b, blue). To calculate the Fold Change (FC) of the graph displayed in Fig.2.3c, we have quantified the number of zebrafish H3K27ac positive sequences aligned with the human genome that also showed H3K27ac signal in human pancreas (ZebraHumanK27). As a control, we have performed a similar analysis, randomizing the aligned human sequences, quantifying the number of those that also showed H3K27ac signal in human pancreas, repeating this operation 10^5 times (randomZebraHumanK27). FC was calculated by the ratio: ZebraHumanK27/[average(randomZebraHumanK27)] (Supplementary Dataset 3q). This was performed for the different populations of zebrafish enhancers (Pancreas, PsE, DevE, and embryo).

2.4.2.11 Transcription factor binding motifs enrichment

To refine our data, H3K27ac peaks were filtered with the ATAC-seq peaks. Then, the transcription factor binding site (TFBS) predictor program Hypergeometric Optimization of Motif EnRichment (HOMER v.4.11.1) was used to identify conserved sequence motifs enriched (Heinz et al., 2010). To evaluate our results, we also analysed, using HOMER, different acetylation data from: human pancreas, human ventricle, zebrafish embryos at 24hpf and at dome+80%epiboly (Supplementary Dataset 3t-u and 4g). From the resulting analysis, we selected the top 140 enriched motifs for each dataset. These motifs were selected based on ranking and the groups were compared by performing hypergeometric enrichment tests. Fisher exact test from GraphPad Prism 7 (v.7.04) was performed to evaluate the enrichment in 25 known pancreas-related TFs (with Bonferroni correction). The HOMER software was also similarly applied in PsE, C1, C2, C3 and C4 in order to identify TFBS (Supplementary Fig.2.3f-g, Fig.2.4 and Supplementary Dataset 3t-u).

2.4.2.12 Identification of super-enhancers

We applied ROSE (Ranking Ordering of Super-Enhancers, v.1) algorithm with default parameters to define super-enhancers in our whole pancreas acetylation data and in human pancreas acetylation data (Whyte et al., 2013). Then, we performed gene ontology analysis in both data using PANTHER software (v.14.0, on April 2019) and compared the molecular functions obtained (http://pantherdb.org). To identify the genes shared between the two groups, we identified the human orthologous genes in our zebrafish list using Biomart (https://www.ensembl.org/biomart; on April 2019) and compared the groups (Fig.2.3e, Supplementary Fig.2.3h).

2.4.2.13 Disease association enrichment of genes from different enhancer clusters

To know whether the genes interacting with the pancreatic enhancer sets (PsE, C1-C4) include homologs of human genes associated with pancreatic diseases in a higher proportion than expected by chance, we took human gene-disease associations from DisGeNET (v.6.0; Piñero et al., 2015), for the available pancreatic diseases. Then, we derived for each disease, the set of zebrafish genes homologous to the human disease-associated genes. In detail, pancreatic diseases and their associated genes were selected from the file containing all gene-disease links from DisGeNET (all_gene_disease_associations.tsv, downloaded from the DisGeNET website on April 2019, v6.0, http://www.disgenet.org/, Integrative Biomedical Informatics Group GRIB/IMIM/UPF), filtering for associations with a score > 0.1 to exclude those based only on text-mining. The disease search term used was "pancrea*", followed by manually filtering for pancreas-related diseases and their human associated genes.

Gene annotations were obtained from Ensembl via BioMart on April 2019 selecting protein coding genes in zebrafish and gene homologs between human and zebrafish. We required a minimum of 15 zebrafish genes relating to a disease to avoid significant gene set enrichments only due to small group ratios without real over/under representations, yielding 16 pancreatic diseases totalling 836 zebrafish homologs of human genes associated to pancreatic diseases (Supplementary Dataset 3r). To check whether the genes interacting with various enhancer clusters (Embryo only, C1, C2, C3, C4, PsE) are enriched for pancreas disease-association, we performed hypergeometric tests for gene set enrichment with the 16 pancreatic diseases left (R phyper function, X: number of genes in disease Ai and in enhancer set Bi; M: number of genes in disease Ai, N: non-disease genes – number of zebrafish protein coding genes minus M; K: number of genes in enhancer set Bi) The R package "qvalue" was used to correct for multiple testing using FDR and convert unadjusted p-values into q-values(MacDonald et al., 2019). Hypergeometric enrichment was obtained as the ratio "(number disease genes in clusterX / number of genes in clusterX) / (number disease genes / number of protein coding genes)”. Finally, diseases with an absolute enrichment ≥ 1.5 and a q-value ≤ 0.05 were considered significantly enriched in the respective cluster (Fig.2.3d).

2.5 Data availability

All the sequencing data (raw data) generated within this study has been submitted to ENA under accession number “PRJEB40292”. The analysed data are available on USCS browser (http://genome-euro.ucsc.edu/s/VDR_group_public_data/Carrico_et_al_2020_ZebrafishPancreasRegulome) and in supplementary material.

Other datasets used in this study can be downloaded from ENCODE project (https://www.encodeproject.org/): ChIP-seq and ATAC-seq of Human pancreas “ENCSR340GAZ” ChIP-seq and ATAC-seq of left ventricle “ENCSR464TTP”; from Expression Atlas data (http://www.ebi.ac.uk/gxa/experiments/): RNA-seq of zebrafish development stages “E-ERAD-475”; NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/): ChIP-seq of developmental stages of zebrafish “GSE32483”; European Nucleotide Archive (ENA) browser (https://www.ebi.ac.uk/ena): RNAseq of the pancreatic acinar, alpha, beta and delta cells from zebrafish “PRJEB10140”, RNA-seq of developmental stages of zebrafish “PRJEB12296”; “PRJEB7244”; “PRJEB12982”. ChIP-seq from the DANIO-CODE consortium to create the blacklist were the following(https://danio-code.zfin.org): “DCD002894SQ”, “DCD002921SQ”, “DCD003653SQ”, “DCD003654SQ”, “DCD003671SQ” and “DCD002742SQ”. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request.

2.6 Code availability

The custom code for analysis of optical action potential traces is available in gitbub (https://gitlab.com/rdacemel/pancreasregulome) and in Zenodo (https://doi.org/10.5281/zenodo.6340878).

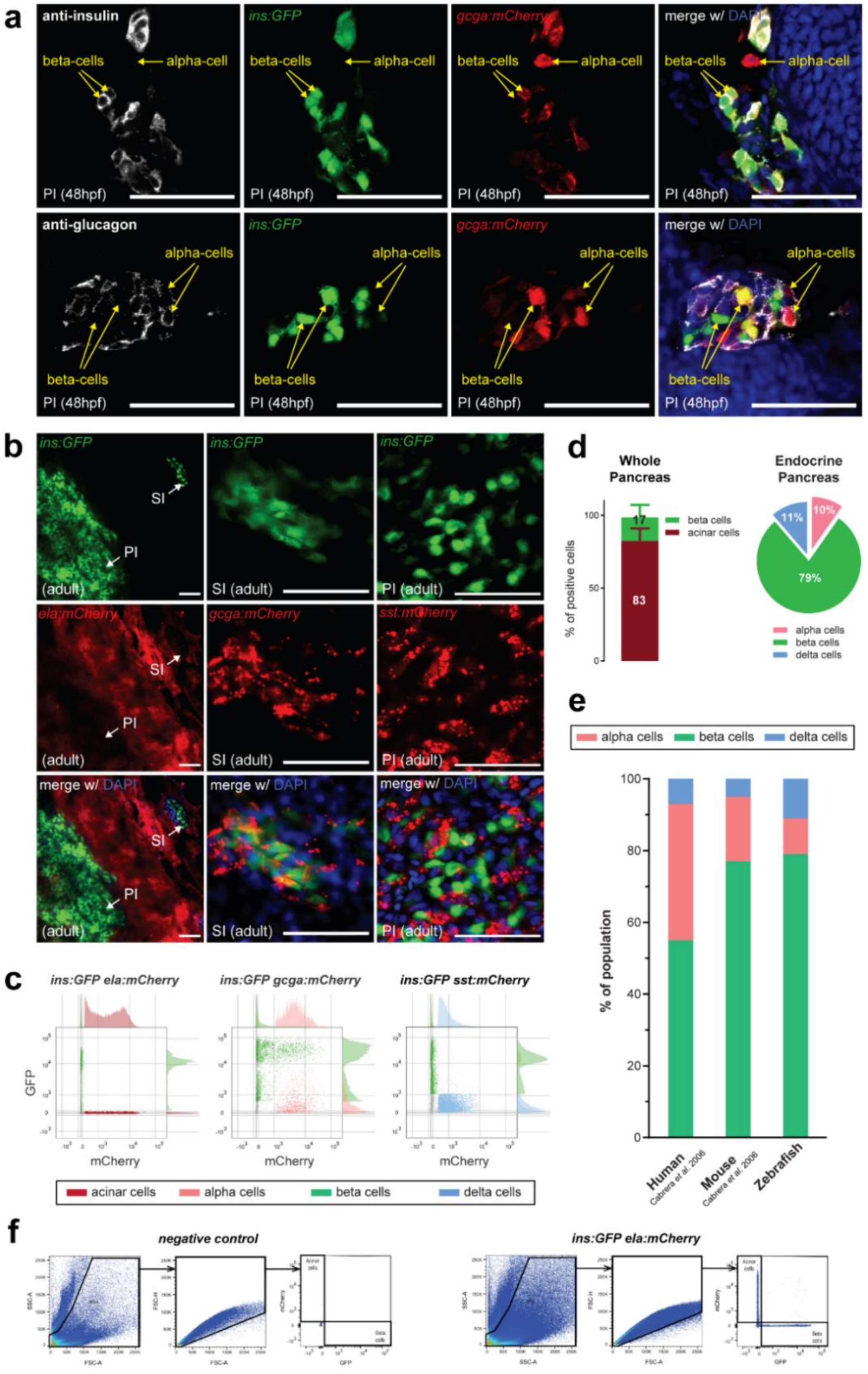
2.7 Acknowledgements

We thank to the i3S Scientific Platform: Advanced Light Microscopy, member of PPBI (POCI-01-0145-FEDER-022122); Translational Cytometry; Cell Culture and Genotyping and i3s HPC facility. We also thank to Carla Oliveira for statistical support; Mafalda Sousa for the automated cell analysis; Guilherme Cardoso from the Histology Service and Isabel Guedes for maintenance of the zebrafish lines. From CABD, we thank to Elisa de la Calle-Mustienes (ChIP-seq), Sandra Jimenez Gancedo (ChIP-seq), Ana Fernandez-Minñán (4C-seq) and Ensieh Farahani (ATAC-seq) for protocol support.

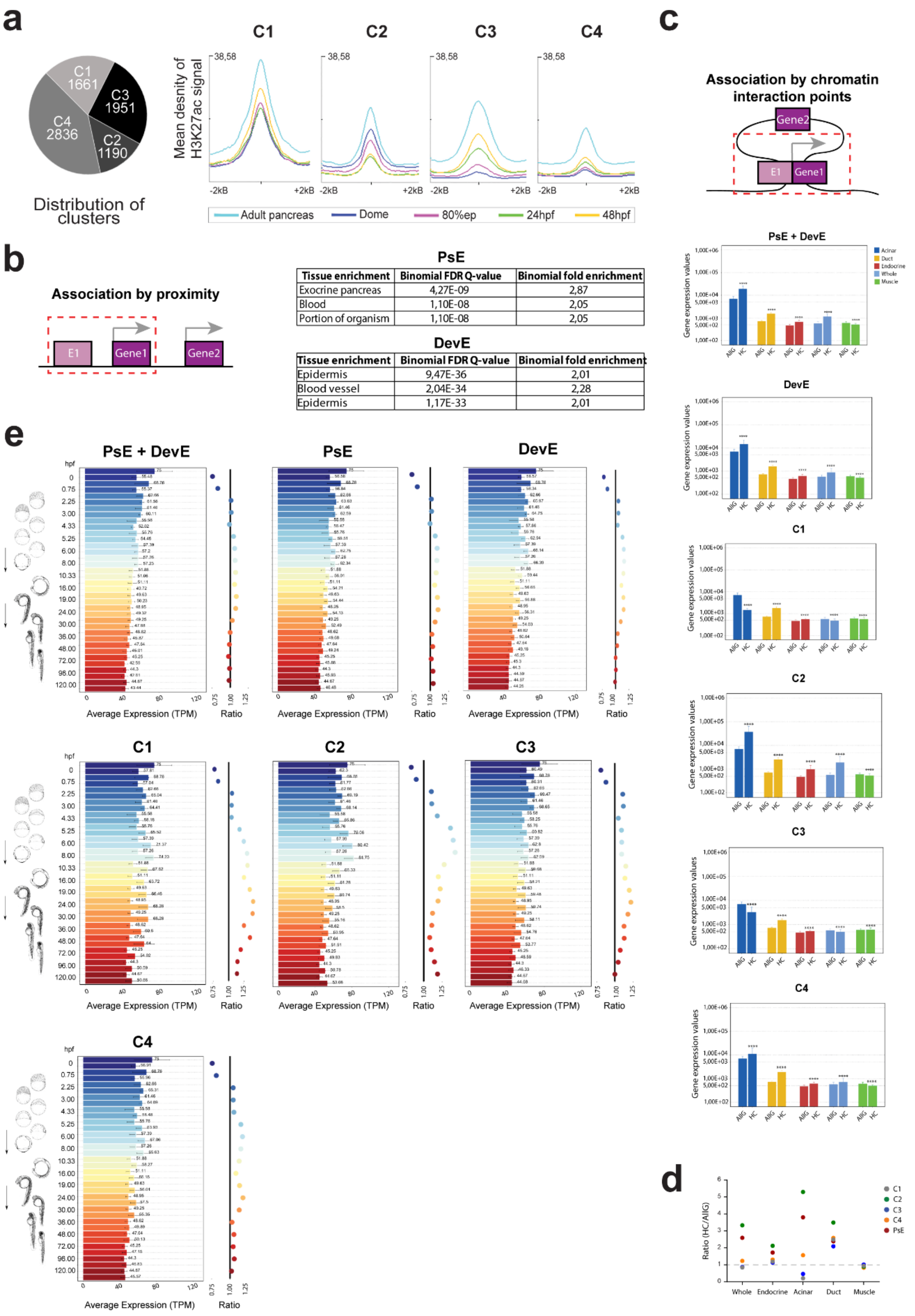
This study was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-2015-StG-680156-ZPR and ERC-2016-AdG-740041-EvoLand to JLG-S). JB is supported by an FCT CEEC grant (CEECIND/03482/2018). JLG-S is supported by the Spanish Ministerio de Economía y Competitividad (BFU2016-74961-P), the Marató TV3 Fundacion (Grant 201611) and the institutional grant Unidad de Excelencia María de Maeztu (MDM-2016-0687). RBC was funded by FCT (ON2201403-CTO-BPD), IBMC (BIM/04293-UID991520-BPD) and EMBO (Short-Term Fellowship). JTx (SFRH/BD/126467/2016), MD (SFRH/BD/135957/2018), AE (SFRH/BD/147762/2019), and FJF (PD/BD/105745/2014) are PhD fellows from FCT. MG was supported by the EnvMetaGen project via the European Union's Horizon 2020 research and innovation programme (grant 668981). This work was funded by National Funds through FCT—Fundação para a Ciência e a Tecnologia, I.P., under the project UIDB/04293/2020".

2.8 Supplementary information

2.8.1 Supplementary figures



Supplementary Figure 2.1 Basic constitution of the zebrafish adult pancreas. a) Representative confocal images of the principal islet (PI) from 48hpf Tg(*ins:GFP, gcga:mCherry*) zebrafish embryos stained with antibodies directed against insulin (white, top panels, n=3) or glucagon (white, bottom panels, n=8). The GFP reporter (green) is expressed exclusively in beta cells (defined by insulin immunostaining), while the mCherry reporter (red) is expressed in alpha cells (defined by glucagon immunostaining) as well as a subset of the beta-cell population. Arrows show the location of some insulin-producing beta cells and glucagon-producing alpha cells. Nuclei are stained with DAPI. Scale bar: 50 µm. **b)** Representative confocal images of whole-mounted pancreatic tissue from adult Tg(*ins:GFP*, *ela:mCherry*) (left panels, n=4), Tg(*ins:GFP, gcga:mCherry*) (middle panels, n=2), and Tg(*ins:GFP*, *sst:mCherry*) zebrafish (right panels, n=1). The adult endocrine pancreas consists of a larger principal islet (PI) and smaller secondary islets (SI) embedded within the pancreatic exocrine tissue composed of acinar cells (red) and a network of duct cells. The PI and SIs are composed of the three major cell populations of beta cells, alpha cells and the somatostatin-producing delta cells, among others. Nuclei are stained with DAPI. Scale bar: 50 µm. **c)** Representative scatter plots of flow cytometry analysis of adult Tg(*ins:GFP*, *ela:mCherry*), Tg(*ins:GFP, gcga:mCherry*), and Tg(*ins:GFP*, *sst:mCherry*) zebrafish pancreas. **d)** Left panel: Relative percentage of pancreatic endocrine beta cells and exocrine acinar cells (mean ± SD) quantified by flow cytometry from adult Tg(*ins:GFP*, *ela:mCherry*) zebrafish pancreas (n=5). From this quantification we propose that the zebrafish exocrine pancreas is around 5-fold more abundant than the endocrine pancreas. The overrepresentation of the exocrine compartment of the pancreas compared to the endocrine compartment is also observed in the mammalian pancreas with 1-2% of the mouse adult pancreas being made up of beta cells[39] and human islets occupying between 1-3% of the total adult pancreatic mass[39-41]. Right panel: Relative percentage of pancreatic endocrine beta cells, alpha cells and delta cells, quantified by flow cytometry from adult Tg(*ins:GFP*, *gcga:mCherry*) (n=65) and Tg(*ins:GFP*, *sst:mCherry*) pancreas (n=30). **e)** Comparison of the relative cell composition of the adult zebrafish endocrine pancreas with that of human and mouse islets[117]. In the three species, the endocrine pancreas is composed mainly of beta-cells, followed by alpha cells and delta cells. Abbreviations: ela, elastase; gcga, glucagon; ins, insulin; sst, somatostatin. **f)** Representative plots for adult wild-type zebrafish (negative control, left panels) and adult Tg(ins:GFP, ela:mCherry) whole pancreas (right panels) illustrating the gating strategy for flow cytomety analysis: FSC/SSC gate to identify living cells, FSC-H/FSC-A to identify single cells types, and single cells types are gated according with positivity/negativity for reporter expression (representative plots can be found in c): in Tg(ins:GFP, ela:mCherry) animals, the beta-cell population is defined by gating ins:GFP positive ela:mCherry negative single cells, and the acinar cell population is defined by gating ins:GFP negative ela:mCherry positive single cells; in Tg(ins:GFP, gcga:mCherry) animals, the beta-cell population is defined by gating ins:GFP positive single cells, and the alpha-cell population is defined by gating ins:GFP negative gcga:mCherry positive single cells; in Tg(ins:GFP, sst:mCherry) animals, the beta-cell population is defined by gating ins:GFP positive sst:mCherry negative single cells, and the delta-cell population was defined by gating ins:GFP negative sst:mCherry positive single cells. Data included in Source Data file for (**d**), and (**e**).



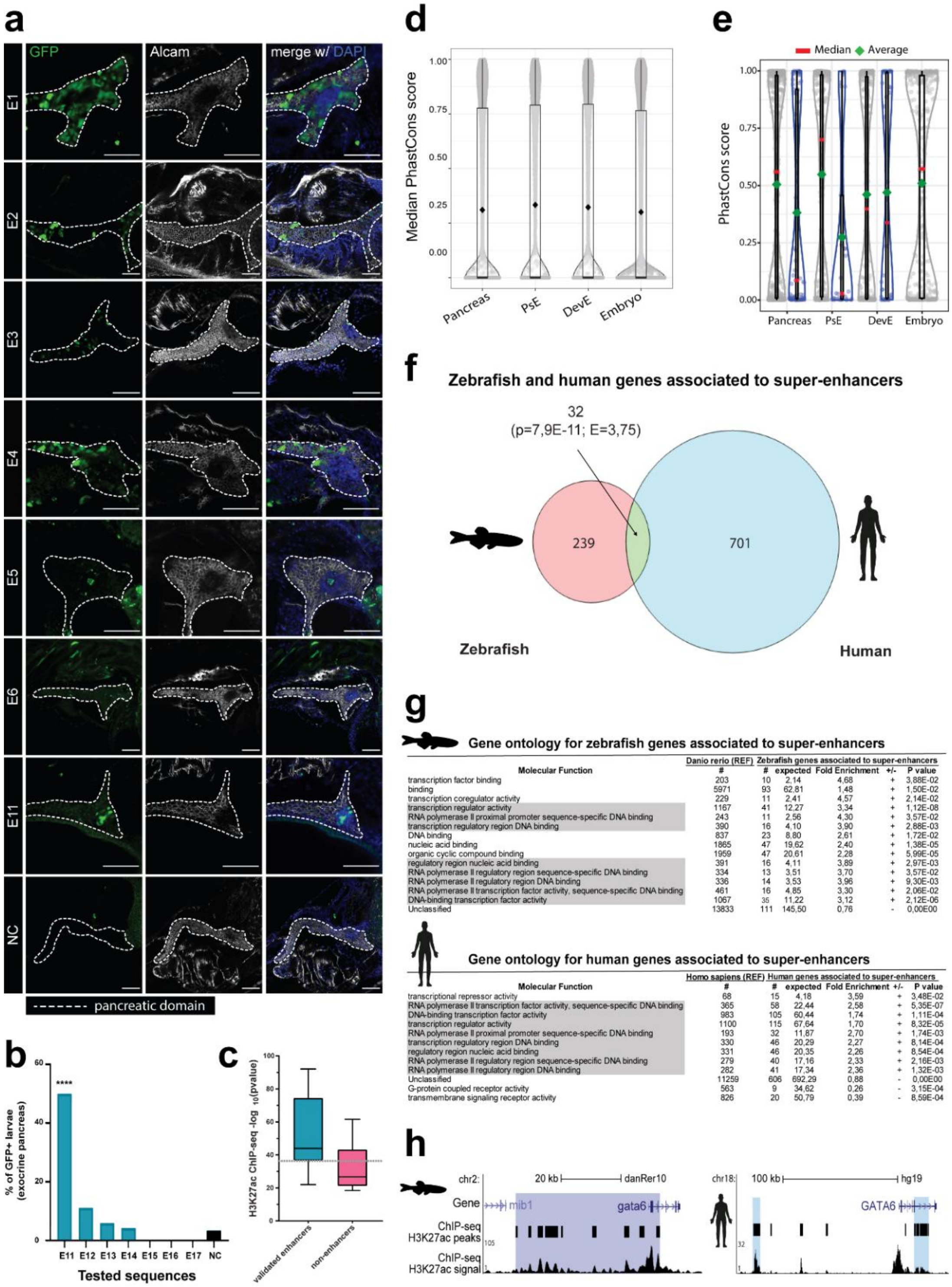
PsE

Tissue enrichment	Binomial FDR Q-value	Binomial fold enrichment
Exocrine pancreas	4,27E-09	2,87
Blood	1,10E-08	2,05
Portion of organism	1,10E-08	2,05

DevE

Tissue enrichment	Binomial FDR Q-value	Binomial fold enrichment
Epidermis	9,47E-36	2,01
Blood vessel	2,04E-34	2,28
Epidermis	1,17E-33	2,01

Supplementary Figure 2.2 Average gene expression of the predicted target genes of different clusters of pancreatic enhancers. a) Left Panel: number of sequences contained in each of the four clusters observed in DevE: C1, Cluster 1; C2, Cluster 2; C3, Cluster 3; and C4, Cluster 4. Right Panels: mean density of H3K27Ac signal for each cluster, centered in its summit and expanded ±2kb. ChIP-seq data for H3K27ac obtained from adult pancreas and embryos at different developmental stages (dome, 80% epiboly, 24hpf and 48hpf). **b)** Left panel: Schematic representation of gene-to-enhancer association by genomic proximity with GREAT[45]. Right Panels: Tissue specific expression enrichment for genes associated to PsE and PsE+DevE. **c)** Upper Panel: Schematic representation of gene-to-enhancer association by chromatin interaction points defined by HiChIP for H3K4me3 (HC). Lower Panels: Average expression of genes interacting with different enhancer sets detected by HC in adult zebrafish pancreas (PsE+DevE,n=8840 genes, DevE, n=5449 genes, C1, n=1917 genes, C2, n=1402 genes, C3, n=1888 genes and C4, n=2531 genes), compared to the average expression of all genes (AllG, n=33737 genes). Gene expression was determined by RNA-seq from pancreatic cells (acinar n=4; duct n=3; endocrine n=4), whole pancreas (n=2) and muscle (control, n=2). One-sided Wilcoxon tests (≥), *p*-values<0.05 considered significant. PsE+DevE,*****p*<2E-16; DevE and C1-C4, *****p*<0.0001. Error bars represent the 95% confidence interval. **d)** Ratio between the average expression of genes interacting with pancreas-specific enhancers (PsE, C1, C2, C3 and C4 clusters; HC) and the average expression of all genes throughout different pancreatic zebrafish tissues (exocrine, endocrine, acinar, duct; AllG). The muscle was used as control. **e)** At the left side of each panel, the average gene expression (transcripts per million, TPM) detected from RNA-seq from zebrafish embryos at different developmental stages (0 to 120hpf;[115]) is plotted. The top bar of each color is the average expression of genes associated with pancreas enhancers (PsE+DevE,n=8840 genes, DevE, n=5449 genes, C1, n=1917 genes, C2, n=1402 genes, C3, n=1888 genes and C4, n=2531 genes) and the bottom bar of each color is the average expression of all genes (AllG, n=33737 genes), with the respective value depicted for each bar. On the right side of each panel is depicted the ratio between the average expression of all genes (HC/AllG) and the average expression of genes associated with pancreas enhancers, maintaining the same color code. Error bars represent the 95% confidence interval. Data included in Source Data file for (**a**-**e**).



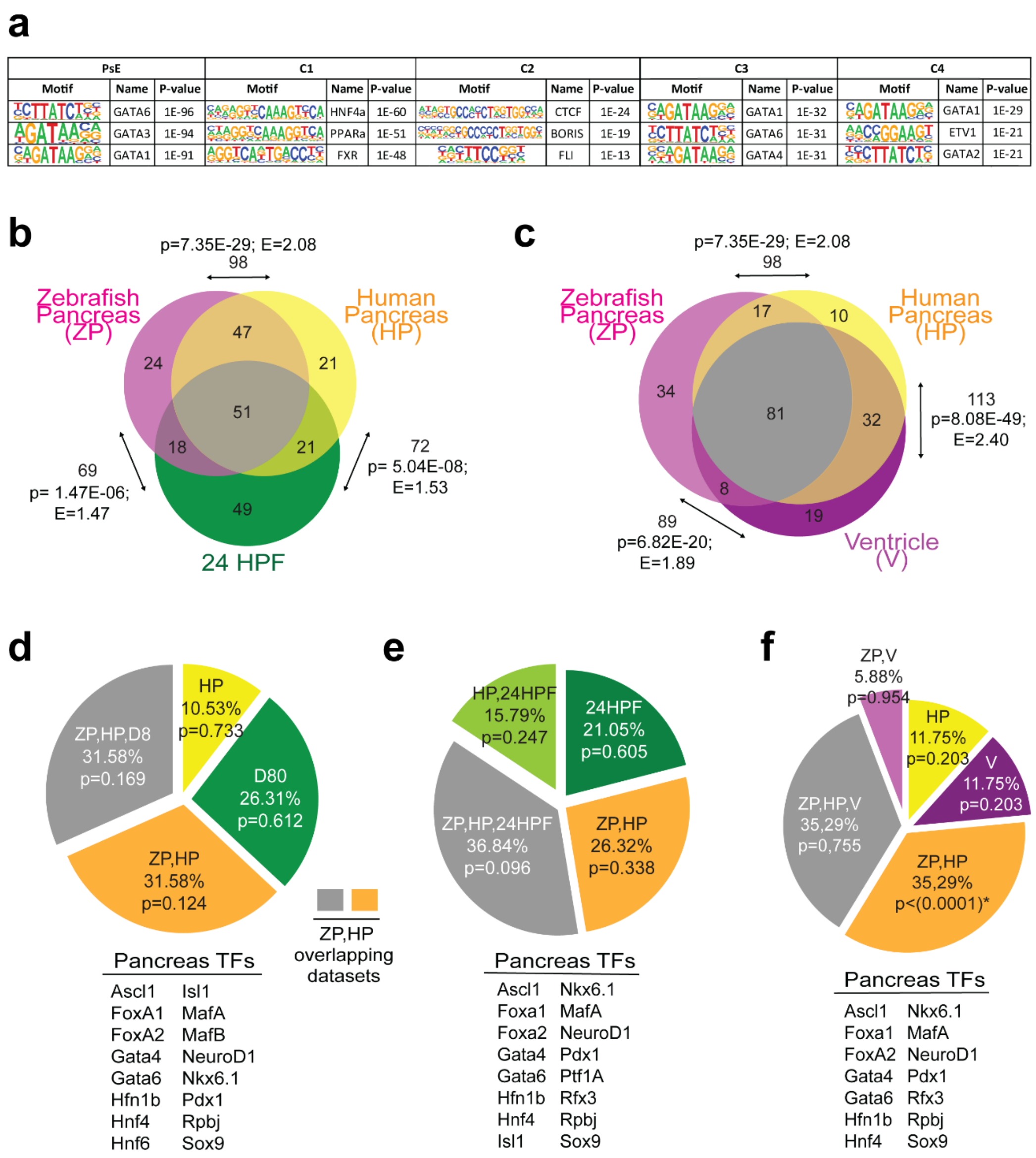
Gene ontology for zebrafish genes associated to super-enhancers

Molecular Function	Danio rerio (REF) #	Zebrafish genes associated to super-enhancers #	expected	Fold Enrichment	+/-	P value
transcription factor binding	203	10	2,14	4,68	+	3,88E-02
binding	5971	93	62,81	1,48	+	1,50E-02
transcription coregulator activity	229	11	2,41	4,57	+	2,14E-02
transcription regulator activity	1167	41	12,27	3,34	+	1,12E-08
RNA polymerase II proximal promoter sequence-specific DNA binding	243	11	2,56	4,30	+	3,57E-02
transcription regulatory region DNA binding	390	16	4,10	3,90	+	2,88E-03
DNA binding	837	23	8,80	2,61	+	1,72E-02
nucleic acid binding	1865	47	19,62	2,40	+	1,38E-05
organic cyclic compound binding	1959	47	20,61	2,28	+	5,99E-05
regulatory region nucleic acid binding	391	16	4,11	3,89	+	2,97E-03
RNA polymerase II regulatory region sequence-specific DNA binding	334	13	3,51	3,70	+	3,57E-02
RNA polymerase II regulatory region DNA binding	336	14	3,53	3,96	+	9,30E-03
RNA polymerase II transcription factor activity, sequence-specific DNA binding	461	16	4,85	3,30	+	2,06E-02
DNA-binding transcription factor activity	1067	35	11,22	3,12	+	2,12E-06
Unclassified	13833	111	145,50	0,76	-	0,00E00

Gene ontology for human genes associated to super-enhancers

Molecular Function	Homo sapiens (REF) #	Human genes associated to super-enhancers #	expected	Fold Enrichment	+/-	P value
transcriptional repressor activity	68	15	4,18	3,59	+	3,48E-02
RNA polymerase II transcription factor activity, sequence-specific DNA binding	365	58	22,44	2,58	+	5,35E-07
DNA-binding transcription factor activity	983	105	60,44	1,74	+	1,11E-04
transcription regulator activity	1100	115	67,64	1,70	+	8,32E-05
RNA polymerase II proximal promoter sequence-specific DNA binding	193	32	11,87	2,70	+	1,74E-03
transcription regulatory region DNA binding	330	46	20,29	2,27	+	8,14E-04
regulatory region nucleic acid binding	331	46	20,35	2,26	+	8,54E-04
RNA polymerase II regulatory region sequence-specific DNA binding	279	40	17,16	2,33	+	2,16E-03
RNA polymerase II regulatory region DNA binding	282	41	17,34	2,36	+	1,32E-03
Unclassified	11259	606	692,29	0,88	-	0,00E00
G-protein coupled receptor activity	563	9	34,62	0,26	-	3,15E-04
transmembrane signaling receptor activity	826	20	50,79	0,39	-	8,59E-04

Supplementary Figure 2.3 In vivo enhancer validation and comparisons between Human and Zebrafish putative pancreatic enhancers. a-c In vivo validation of selected putative pancreatic enhancers by transgenic zebrafish reporter assays, identified by ATAC and H3k27ac ChIP-seq data. **a)** Representative confocal images of F0 transgenic zebrafish larvae for all validated enhancers (E1-E6 and E11, n values are discriminated in Supplementary Dataset 4); whole mount immunohistochemistry of 10-12 dpf zebrafish larvae showing GFP reporter expression (green), within the pancreatic domain (dashed white line). The empty vector was used as negative control (NC). The exocrine pancreas is stained with anti-Alcam (white) antibody and nuclei are stained with DAPI. Scale bar: 60 µm. **b)** Percentage of F0 transgenic zebrafish larvae with GFP expression within the exocrine pancreas for sequences with low H3K27ac ChIP signal value: (($-\log_{10}$(*p*-value)<35) (sequences E11 to 17). Two-sided chi-square test with Yates' correction, *p*-value<0.05 were considered significant (*****p*<0.0001). The empty vector was used as negative control (NC). The exact *p*-value are discriminated in Supplementary Dataset 4. **c)** H3K27ac ChIP-seq signal [H3K27ac ChIP-seq -log10(p-value)] of validated pancreatic enhancer sequences (validated enhancers) versus tested sequences without enhancer activity in the differentiated pancreas (non-enhancers; centre, median; box, upper and lower quartile; whiskers, minimum and maximum value). Prior to enhancer validation we divided the sequences into two groups based on their H3K27ac ChIP-seq signal; “high H3K27ac” [sequences corresponding to the top 10 higher values of H3K27ac ChIP-seq -log10(p-value)] and “low H3K27ac” (the remaining 7 sequences). The dashed grey line represents the threshold between the “high” [-log10(p-value) from 36.5 to 92.1] and “low H3K27ac" groups [-log10(p-value) from 18.5 to 28.4] **d)** Distribution of the median PhastCons scores for each zebrafish putative enhancer sequence active in adult pancreas (14301), adult pancreas only (PsE, 6918), adult pancreas and embryo (DevE, 8368) and in embryo (Embryo, 65871). Putative active enhancer sequences converted from DanRer10 to DanRer7 (liftOver) to match the available conservation scores for zebrafish and 7 vertebrates. The median value is zero for the 4 enhancer sets (lower bar of the boxplots) since most sequences are not conserved, while the average is, respectively, 0.31, 0.33, 0.32, 0.30 (back diamond inside the boxplot). **e)** PhastCons scores (99 vertebrate genomes against hg38) for human sequences converted from zebrafish putative enhancers filtered by ATAC-seq peaks. Grey dots: conserved sequences not overlapping the H3K27ac mark in human pancreas (512 pancreas, 258 PsE, 254 DevE and 334 embryo). Blue dots: conserved sequences also showing H3K27ac signal in human pancreas (ENCODE data; 73 pancreas, 33 PsE, 40 DevE). Green diamonds: average (grey dots: 0.50, 0.55, 0.46, 0.51; blue dots: 0.38, 0.27, 0.47, respectively for pancreas, PsE, DevE and embryo). Red horizontal line: median (grey dots: 0.6, 0.7, 0.4, 0.6; blue dots: 0.09, 0.03, 0.34, respectively for pancreas, PsE, DevE and embryo). **f)** Zebrafish and human genes associated with super-enhancers, identified by ROSE[56]. Statistical significance was calculated by hypergeometric enrichment test, *p*-value (p) and the enrichment are represented. **g)** Gene ontology for genes associated with super-enhancers in zebrafish (above) and human (below). **h)** H3K27ac profile of the landscape of a gene important in pancreatic development in zebrafish (*gata6*; left) and human (*GATA6*; right); super-enhancers are highlighted in purple (zebrafish) or blue (human). Data included in Source Data file for (**b**-**g**).



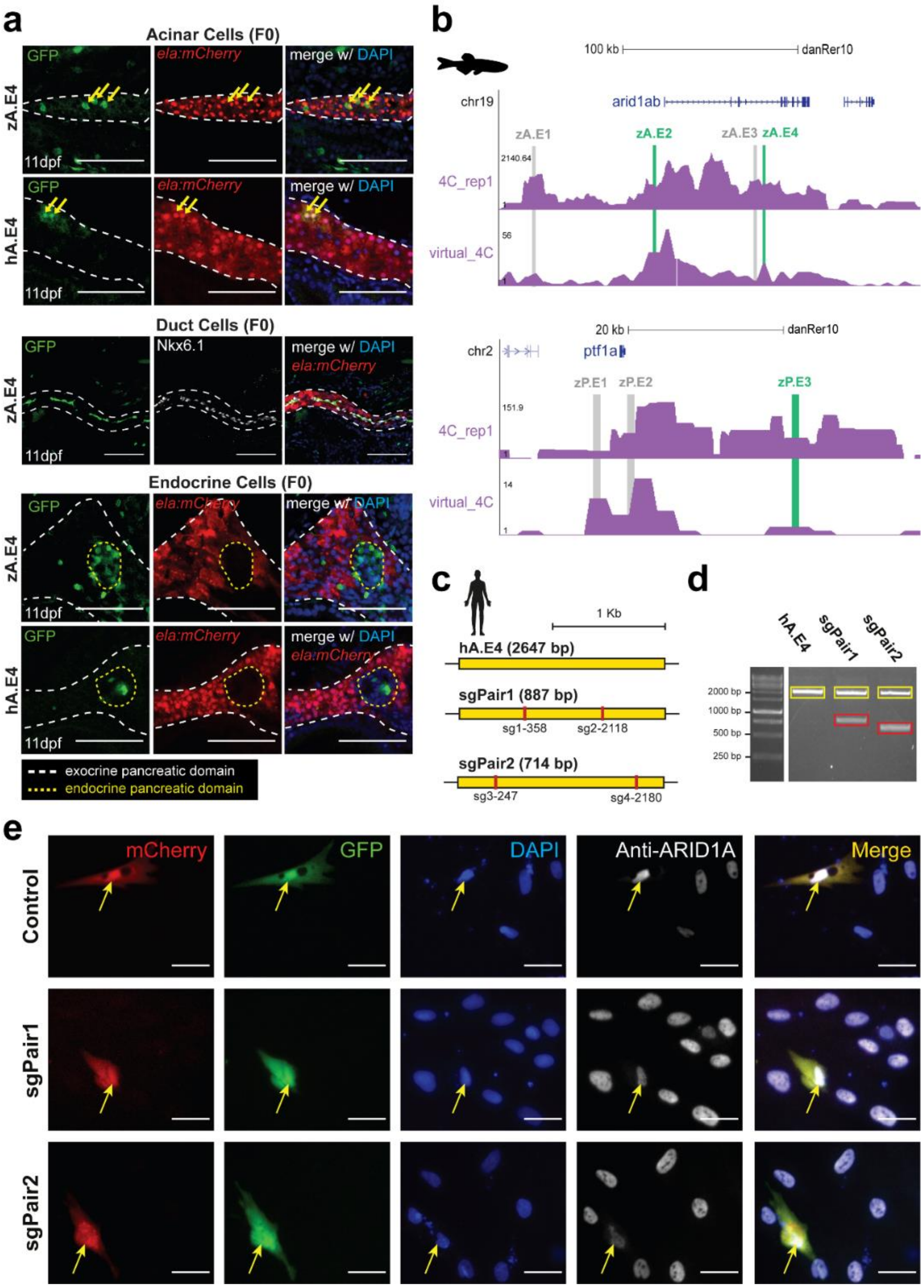
PsE			C1			C2			C3			C4		
Motif	Name	P-value	Motif	Name	P-value	Motif	Name	P-value	Motif	Name	P-value	Motif	Name	P-value
	GATA6	1E-96		HNF4a	1E-60		CTCF	1E-24		GATA1	1E-32		GATA1	1E-29
	GATA3	1E-94		PPARa	1E-51		BORIS	1E-19		GATA6	1E-31		ETV1	1E-21
	GATA1	1E-91		FXR	1E-48		FLI	1E-13		GATA4	1E-31		GATA2	1E-21

Pancreas TFs	
Ascl1	Isl1
FoxA1	MafA
FoxA2	MafB
Gata4	NeuroD1
Gata6	Nkx6.1
Hfn1b	Pdx1
Hnf4	Rpbj
Hnf6	Sox9

Pancreas TFs	
Ascl1	Nkx6.1
Foxa1	MafA
Foxa2	NeuroD1
Gata4	Pdx1
Gata6	Ptf1A
Hfn1b	Rfx3
Hnf4	Rpbj
Isl1	Sox9

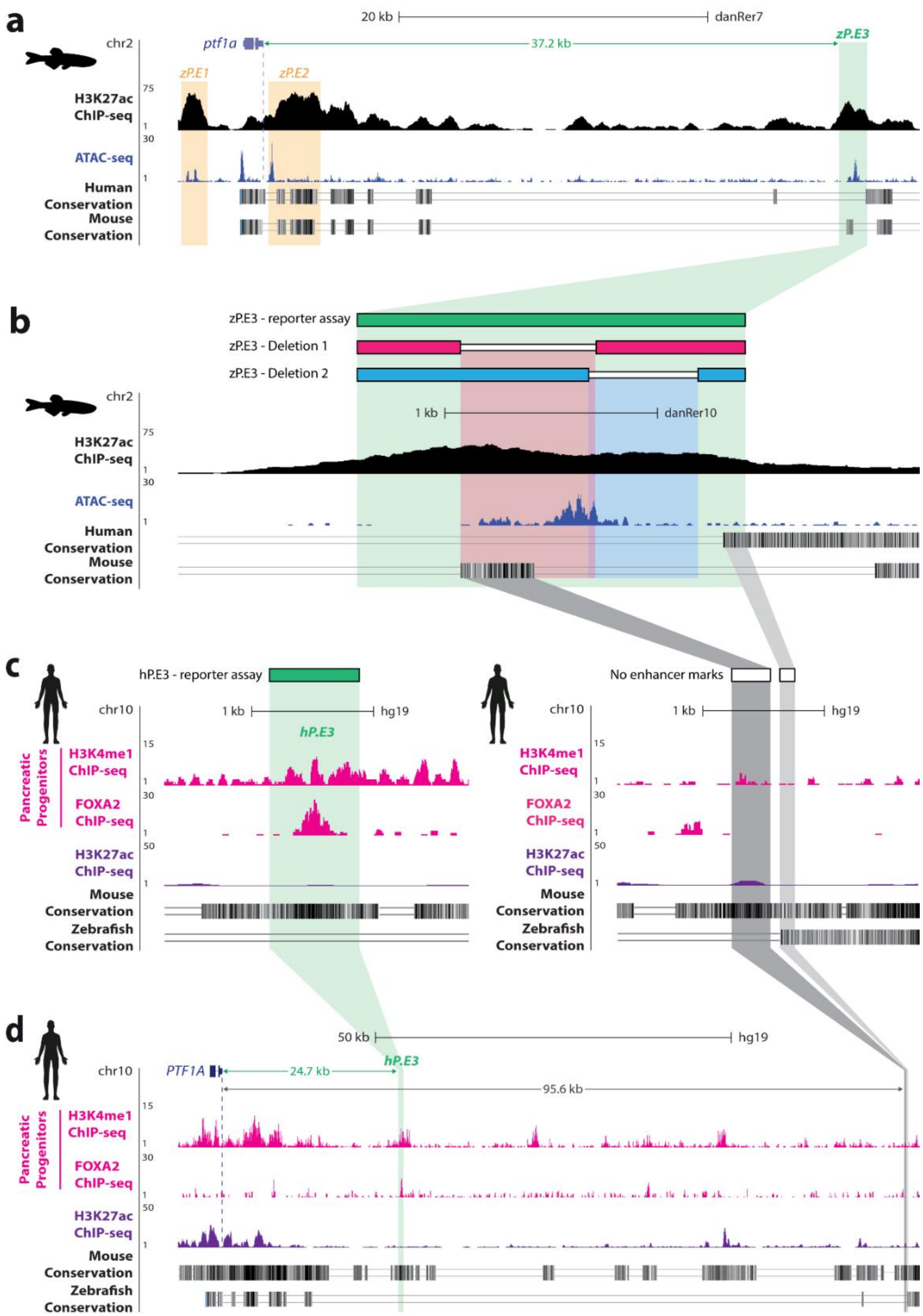
Pancreas TFs	
Ascl1	Nkx6.1
Foxa1	MafA
FoxA2	NeuroD1
Gata4	Pdx1
Gata6	Rfx3
Hfn1b	Rpbj
Hnf4	Sox9

Supplementary Figure 2.4 Zebrafish and human pancreatic enhancers share TFBS. a) List of top three TFBS motifs enriched in H3K27ac ChIP-seq data, for the different enhancer sets: pancreas specific enhancers (PsE) and clusters of developmental enhancers C1, C2, C3 and C4, with the respective *p*-value calculated by HOMER (Two-sided hypergeometric enrichment test)[100]. **b-c** Regions enriched in ATAC-seq and H3K27ac signal from zebrafish pancreas (ZP), human pancreas (HP), 24hpf zebrafish embryos (24HPF) and human ventricle (V) were investigated for TFBS motifs (Supplementary Dataset 3t-u). The top 140 enriched TFBS motifs from each tissue were selected and the overlap of those sets was analyzed. Arrows: number of TFBS motifs shared between two different groups, the enrichment of TFBS motifs and respective *p*-value for each intersection. Statistical significance was determined by hypergeometric enrichment test. Number of TFBS motifs identified in each group/intersection. The exact *p*-value and enrichment are described in the figure. **b)** ZP, HP and 24HPF; **c)** ZP, HP, V. **d-f** The motifs corresponding to 25 pancreas transcription factors (Pancreas TFs) selected from literature were analyzed for their representation in H3K27ac peaks filtered with ATAC-seq peaks among the different zebrafish (ZP, 24HPF, D80) and human tissues (HP, V), and tissue-intersections shown in b-c). The distribution of these TFBS motifs through the different tissues is shown along with the percentage and the respective *p-value* indicated for each group: **d)** ZP, HP and dome+80%epiboly (D80); **e)** ZP, HP and 24hpf; **f)** ZP, HP and V. The list of TFs presents in the tissues, for each graph, is depicted below. Statistical significance was assessed by two-sided Fisher exact test, *p-values*<0.007 were considered significant (Bonferroni correction). The exact *p*-values are discriminated in the graph. Data included in Source Data file for (**a**-**f**).

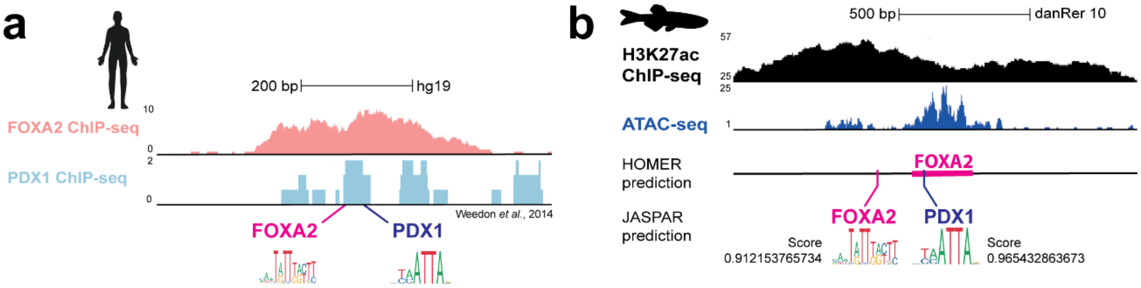


Supplementary Figure 2.5 Human and zebrafish ARID1A/arid1ab enhancer reporter assays and CRISPR-Cas9-mediated deletion of the zA.E4 in a human ductal cell line. **a)** Human and zebrafish *ARID1A/arid1ab* enhancers drive expression in various pancreatic cell types. Representative confocal images of 11dpf

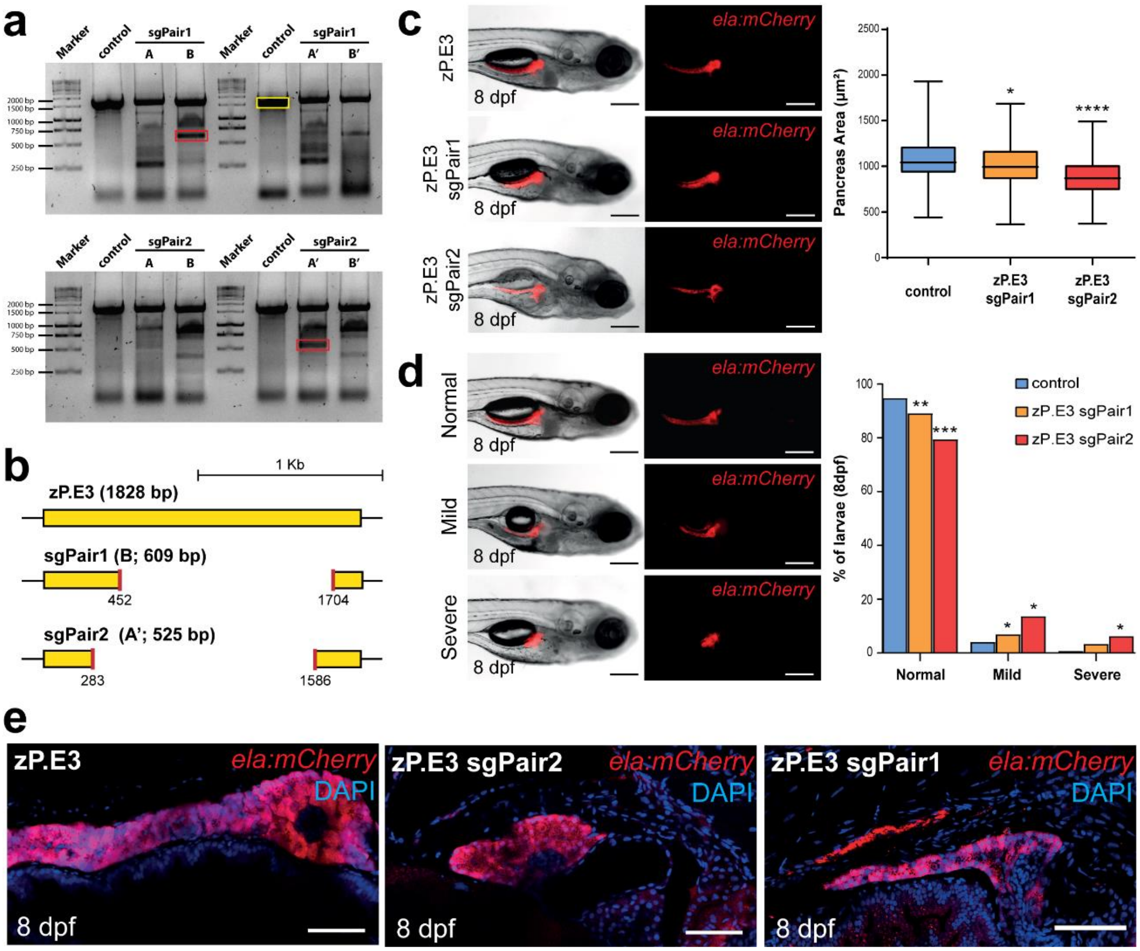
Tg(*ela:mCherry*) zebrafish larvae injected with Z48 vector carrying the zebrafish zA.E4 or human hA.E4 enhancer. The top panel shows elastase-producing acinar cells (red; zA.E4 n=28 ; hA.E4 n=39), the middle panel shows Nkx6.1-positive duct cells (white; zA.E4 n=10 ; hA.E4 n=22), and the bottom panel shows the endocrine pancreas domain (yellow dashed line; zA.E4 n=26 ; hA.E4 n=39). In all panels, the exocrine pancreas domain is indicated with a white dashed line and nuclei are stained with DAPI (blue). Scale bar: 60 µm. **b)** Genomic landscape of zebrafish *arid1ab* gene (top) and *ptf1a* gene (bottom), showing *arid1ab* 4C interactions (purple) from 4C-seq replicates and virtual 4C from HiChIP for H3K4me3 in adult zebrafish pancreas. Tested putative enhancers are highlighted by the colored boxes (grey and green). **c)** Schematic depiction of the targeting strategy for deletion of the hA.E4 locus. The CRISPR sgRNA target sites are depicted in red. **d)** Agarose gel showing the wild-type (yellow) and deleted (red) PCR amplified hA.E4 sequence after gene editing with each respective sgRNA pair (n=3). **e)** Representative fluorescent microscopy images of transfected hTERT-HPNE human cells (n=3), defined by the co-expression of GFP (green) and mCherry (green). Nuclei are stained with DAPI (blue) and anti-ARID1A antibody (white). The yellow arrows indicate the double transfected cell. Images were captured with Leica DMI6000 FFW microscope. Scale bar: 40 µm. Abbreviations: ela, elastase**.**



Supplementary Figure 2.6 Regions of sequence conservation within the zP.E3 enhancer sequence. a) UCSC Genome Browser view of zP.E3 showing H3K27ac ChIP-seq (black) and ATAC-seq (blue) from whole zebrafish pancreas (above), and the conservation tracks (below) displaying where the human and mouse genomes aligns to the zebrafish sequence (darker shading indicates higher BLASTZ scores; white indicates no alignment). The validated distal *ptf1a* enhancer (zP.E3) is indicated by the green box and other validated enhancers by the orange boxes. **b)** Zoom-in of the zP.E3 region depicted in a) and schematic depiction of the generated zP.E3 deletion alleles; Deletion1 and Deletion2. The zP.E3 sequence contains a 332bp region conserved with mouse (dark grey) and a 120bp region conserved with human (light grey). **c)** UCSC Genome Browser view of the validated hP.E3 (left panel) and the regions depicted in b) [the region conserved between zebrafish and mouse (dark grey) and the region conserved between zebrafish and human (light gray)] (right panel), showing H3K4me1 and FOXA2 ChIP-seq from pancreatic progenitor cells (pink), H3K27ac ChIP-seq from adult pancreatic tissue (purple) and conservation tracks for mouse and zebrafish (below). **d)** Zoom-out of the regions depicted in c), showing the full *PTF1A* landscape.

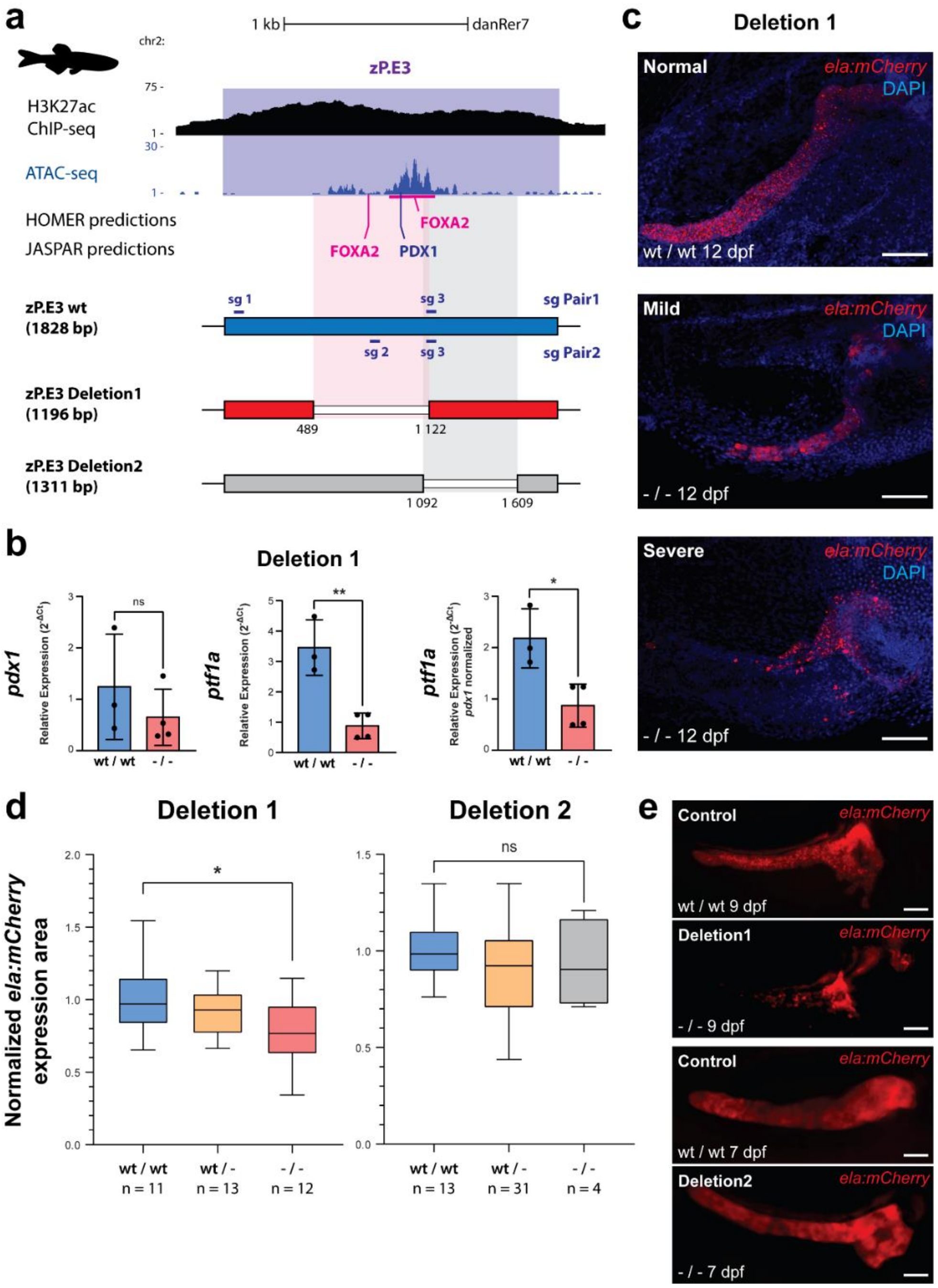


Supplementary Figure 2.7 In vivo enhancer validation and comparisons between Human and Zebrafish putative pancreatic enhancers. a) ChIP-seq density plots at the hP.E3 locus showing FOXA2 (pink) and PDX1 (blue) ChIP-seq peaks generated from human endocrine islets[2]. The location of the respective predicted binding sites is depicted below. **b)** H3K27ac ChIP-seq (black) and ATAC-seq (blue) read density plots at the zP.E3 locus, and putative FOXA2 and PDX1 transcription factor binding sites predicted by JASPAR[116] and HOMER[100] with respective score.

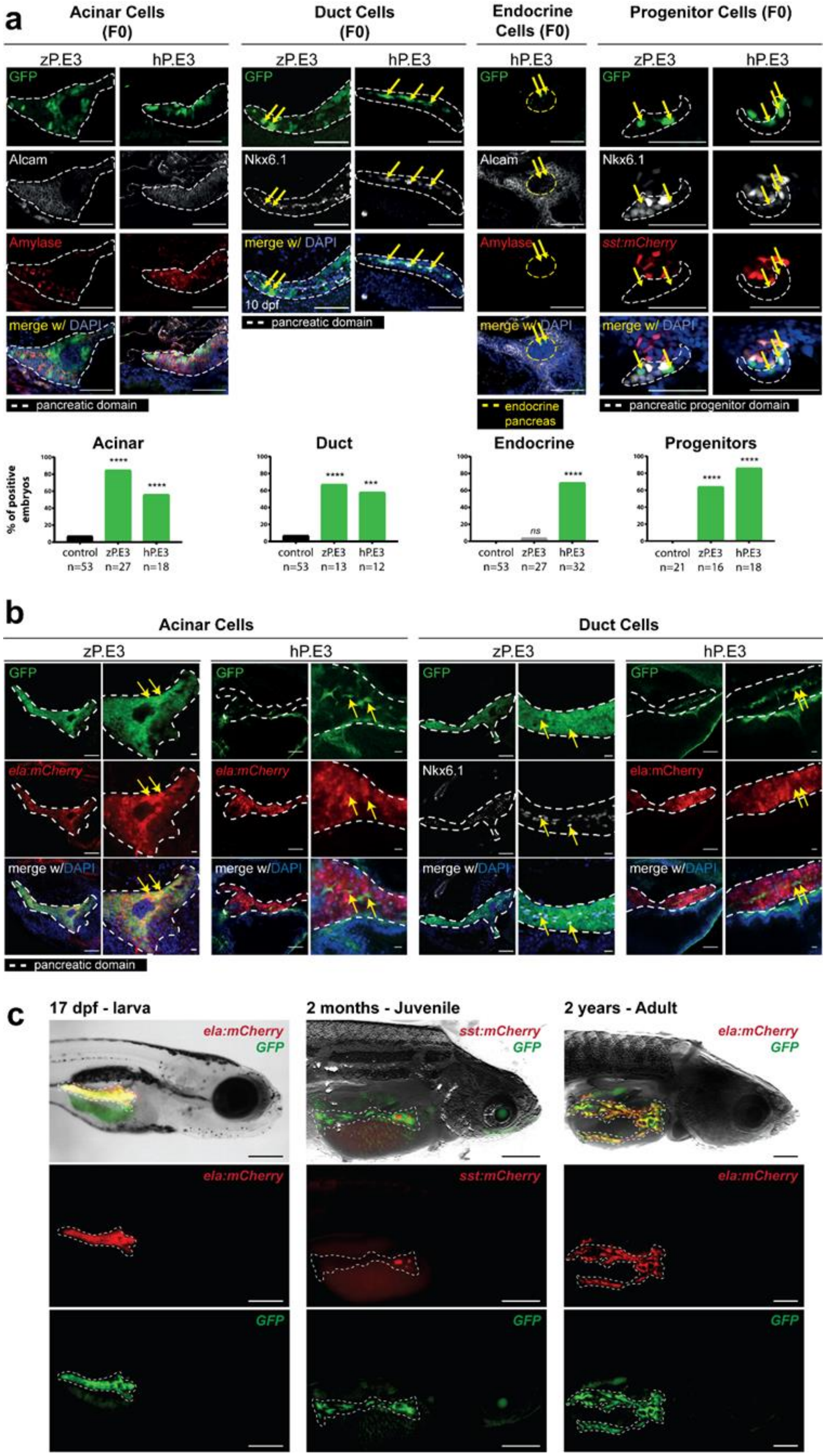


Supplementary Figure 2.8 CRISPR-Cas9-mediated deletion of the zP.E3 enhancer impairs pancreas development. a) PCR screening of zP.E3 deletion after co-injection of zebrafish embryos with Cas9 and different combinations of sgRNAs: successful deletions appear as truncated PCR products (red box), in comparison with the wild-type sequence from non-injected embryos (control, yellow box). A, B; A', B' represent different batches of embryos injected with each pair of sgRNAs (n=5 independent injections per sgRNA pair). **b)** Schematic representation of the deletions induced by CRISPR-Cas9 depicted in a) (yellow and red boxes). **c)** Tg(*ela:mCherry*) embryos were injected with Cas9 alone (zP.E3) or co-injected with Cas9 and a pair of sgRNAs (zP.E3 sgPair1 or zP.E3 sgPair2) and monitored at 8dpf (Cas9 alone, n=140; zP.E3 sgPair1, n=110; zP.E3 sgPair2, n=108 larvae, pooled from 3 independent experiments each). Representative live images are shown in the left panels. Scale bar: 250 µm. The quantification of total pancreas area is represented in the right panel (centre, median; box, upper and lower quartile; whiskers, minimum and maximum value). Unpaired student's t-test (two-tailed), *p-values*<0.05 were considered significant ((***p*= 0.0107, *****p*=1.1486×10E-11). **d)** The injected F0 8dpf larvae from c) were classified as either normal or as one of the two following classes: mild pancreatic defect characterized by significantly reduced pancreas size (mild), or severe pancreatic defects characterized by a reduced pancreas with absence of the pancreatic tail (severe). Representative live images of each pancreatic phenotype are shown in the left panels. Scale bar: 250 µm. The percentage of larvae in each phenotypic class is represented in the right panel and the n described in **c**. Fisher's exact test (two-sided), *p*-values<0.05 were considered significant (*p*-values, left to right: ***p*=0.0083,

****p*=0.0002, **p*=0.0184, *p*=0.0102, and *p*=0.0229). **e)** Representative confocal fluorescent images of 8dpf Tg(*ela:mCherry*) larvae showing impaired development of pancreas upon injection of Cas9 and the respective sgPairs targeting the zP.E3 enhancer (zP.E3 sgPair1 or zP.E3 sgPair2), in comparison to the control, injected with Cas9 alone (zP.E3). This experiment, independent of the 3 replicates presented in c-d), produced the following results: zP.E3 (100% normal pancreas in a total of n=18 larvae), zP.E3 sgPair1 (8.33% of pancreatic phenotypes in a total of n=12 larvae) and zP.E3 sgPair2 (18.75% of pancreatic phenotypes in a total of n=16 larvae). Nuclei are stained with DAPI (blue) and acinar cells are labeled with mCherry (red). Scale bar: 60 µm. Abbreviations: ela, elastase. Data included in Source Data file for (**c**).



Supplementary Figure 2.9 Independent deletions of the zP.E3 enhancer have distinct phenotypic outcomes. a) UCSC Genome Browser view of zP.E3 showing H3K27ac ChIP-seq (black) and ATAC-seq (blue) from whole zebrafish pancreas samples (upper panels), along with the location of predicted TFBS for FOXA2 and PDX1 (middle panel), schematic depiction of the sgRNA pairs and the generated zP.E3 deletion alleles: Deletion1 (generated with sgPair1) and Deletion2 (generated with sgPair2). **b)** Relative expression of *pdx1* (left panel) and *ptf1a* (middle panel) in pancreatic progenitor cells of 48hpf embryos, and corresponding *pdx1*-normalized *ptf1a* expression (right panel). Each biological replicate was obtained from a batch of 30 embryos. Unpaired student's t-test (two-tailed), *p-values*<0.05 were considered significant (***p*=0.0039, **p*=0.0172). **c)** Representative confocal images (maximum intensity projections) of 12dpf Tg(*ela:mCherry*) larvae showing impaired development of pancreas in Deletion1 homozygous larva (-/-) compared to the control (wt/wt sibling). Larvae resulted from a single incross of heterozygous animals and only the homozygous larvae (wt/wt and -/-) were selected for confocal imaging: wt/wt, n=3 (100% normal phenotypes); -/-, n=7 (57.14% normal, 14.29% mild, and 28.57% severe phenotypes). Nuclei are stained with DAPI (blue) and acinar cells with mCherry (red). Scale bar: 60 µm. **d)** Normalized area of *ela:mCherry* expression of Tg(*ela:mCherry*) Deletion1 and Deletion2 homozygous (-/-) and heterozygous (wt/-) larva (9 and 7dpf, respectively), compared to the respective control (wt/wt siblings; centre, median; box, upper and lower quartile; whiskers, minimum and maximum value). Individual values were normalized to the mean of their respective control group. Unpaired student's t-test (two-tailed), *p-values*<0.05 were considered significant (**p*=0.0353). **e)** Representative live image of the pancreas of Tg(*ela:mCherry*) Deletion1 and Deletion2 homozygous larva (-/-) (9 and 7dpf, respectively), compared to a control larva (wt/wt sibling). From top to bottom: 9dpf wt/wt, n= 11 (100% normal phenotypes); 9dpf Deletion1 -/-, n= 12 (33.33 normal, 33.33 mild, and 33.33 severe phenotypes); 7dpf wt/wt, n=13 (100% normal phenotypes); 7dpf Deletion1 -/-, n=4 (100% normal phenotypes). Scale bar: 60 µm. Abbreviations: ela, elastase; ins, insulin. Data included in Source Data file for (**b**) and (**d**).



Supplementary Figure 2.10 Human and zebrafish distal *PTF1a/ptf1a* enhancers drive similar reporter expression in various pancreatic cell types. a) Left panels: Representative confocal images of GFP expression (green) driven by zebrafish zP.E3 or human hP.E3 sequences in F0 10dpf zebrafish larvae in pancreatic acinar cells, stained with anti-Alcam (white) and anti-Amylase (red), duct cells, stained with anti-Nkx6.1 (white) and cells of the principal islet of the endocrine pancreas, which lack staining of acinar cell markers. Right panel: Representative confocal images of 48hpf F0 zebrafish embryos showing zP.E3 and hP.E3-driven GFP expression within the pancreatic progenitor domain. Pancreatic progenitor cells are identified by anti-Nkx6.1 staining (white), adjacent to the principal islet marked by somatostatin expression in differentiated delta-cells (red). Nuclei are stained with DAPI (blue). Yellow arrows point to examples of GFP expression in each pancreatic cell type. Scale bar: 60 µm. The corresponding percentage of zebrafish embryos showing zP.E3 or hP.E3-mediated GFP expression in transient transgenesis assays are depicted below. Statistical significance determined by Fisher's exact test (two-sided), *p*-values<0.05 were considered significant (*p*-values, left to right: ****p=3.387×10-12; ****p=5.643×10-5; ****p=8.00×10-5; ***p=0.0005; ****p=5.10×10-13; ****p=2.299×10-5; ****p=2.586×10-5). **b)** Left Panels: Representative confocal images of Tg(*zP.E3:GFP*, *ela:mCherry*) and Tg(*hP.E3:GFP, ela:mCherry*) F1 larvae, showing co-localization of zP.E3 and hP.E3 mediated GFP expression (green) and acinar cell-specific mCherry expression (red). Right Panels: Representative confocal images of Tg(*zP.E3:GFP*) and Tg(*hP.E3:GFP, ela:mCherry*) F1 larvae showing GFP expression in the duct cells of the exocrine pancreas. In Tg(*zP.E3:GFP*) larvae duct cells are stained with anti-Nkx6.1 (white), while in Tg(*hP.E3:GFP, ela:mCherry*) larvae duct cells appear as mCherry-negative cells within the exocrine pancreatic tissue. Tg(*zP.E3:GFP, ela:mCherry*) larvae: 100% of larvae show co-localization of GFP with acinar specific mCherry expression (n=8 pooled from two impendent experiments) or with duct specific Nkx6.1 staining (n=8). Tg(*hP.E3:GFP, ela:mCherry*) larvae: of 2 analyzed larvae (n=2) 100% show GFP expression in duct cells and acinar cells, although, in the case of acinar cells, in a very reduced number of cells. Nuclei are labeled with DAPI (blue). Yellow arrows point to examples of GFP-positive duct cells. Scale bar 60 µm. **c)** Zebrafish zP.E3 reporter transgenic line drives GFP expression in the exocrine pancreas, from larva to adult. Representative images of zP.E3-driven GFP expression (green) in 17dpf Tg(*ela:mCherry*) larvae (n=1; scale bar: 100 µm), 2 months Tg(*sst:mCherry*) juvenile (n=1), and 2 year old Tg(*ela:mCherry*) adult (scale bar: 500 µm), showing sustained enhancer activity in the exocrine pancreatic tissue (n=1;delimited by the white dashed line). mCherry is represented in red. Abbreviations: ela, elastase; sst, somatostatin. Data included in Source Data file for (**a**).

2.8.2 Supplementary datasets

The supplementary dataset described in this chapter are available online, as supplementary material of the publication: https://www.nature.com/articles/s41467-022-29551-7#Sec40

2.9 References

Akerman, I., Maestro, M.A., De Franco, E., Grau, V., Flanagan, S., García-Hurtado, J., Mittler, G., Ravassard, P., Piemonti, L., Ellard, S., et al. (2021). Neonatal diabetes mutations disrupt a chromatin pioneering function that activates the human insulin gene. Cell Rep *35*, 108981.

Alvarsson, A., Jimenez-Gonzalez, M., Li, R., Rosselot, C., Tzavaras, N., Wu, Z., Stewart, A.F., Garcia-Ocaña, A., and Stanley, S.A. (2020). A 3D atlas of the dynamic and regional variation of pancreatic innervation in diabetes. Sci Adv *6*, eaaz9124.

Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep *9*, 9354.

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

Andrews, S (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

van Arensbergen, J., Dussaud, S., Pardanaud-Glavieux, C., García-Hurtado, J., Sauty, C., Guerci, A., Ferrer, J., and Ravassard, P. (2017). A distal intergenic region controls pancreatic endocrine differentiation by acting as a transcriptional enhancer and as a polycomb response element. PLoS One *12*, e0171508.

Ariza-Cosano, A., Visel, A., Pennacchio, L.A., Fraser, H.B., Gómez-Skarmeta, J.L., Irimia, M., and Bessa, J. (2012). Differences in enhancer activity in mouse and zebrafish reporter assays are often associated with changes in gene expression. BMC Genomics *13*, 713.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem *94*, 890–898.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol *9*, e1003326.

Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F., et al. (2009). Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. Dev Dyn *238*, 2409–2417.

Bessa, J., Luengo, M., Rivero-Gil, S., Ariza-Cosano, A., Maia, A.H.F., Ruiz-Ruano, F.J., Caballero, P., Naranjo, S., Carvajal, J.J., and Gómez-Skarmeta, J.L. (2014). A mobile insulator system to detect and disrupt cis-regulatory landscapes in vertebrates. Genome Res *24*, 487–495.

Bogdanovic, O., Fernandez-Miñán, A., Tena, J.J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S.J., Veenstra, G.J.C., and Gómez-Skarmeta, J.L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res *22*, 2043–2053.

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

Buffry, A.D., Mendes, C.C., and McGregor, A.P. (2016). The Functionality and Evolution of Eukaryotic Transcriptional Enhancers. Adv Genet *96*, 143–206.

de la Calle-Mustienes, E., Feijóo, C.G., Manzanares, M., Tena, J.J., Rodríguez-Seguel, E., Letizia, A., Allende, M.L., and Gómez-Skarmeta, J.L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res *15*, 1061–1072.

Cebola, I., Rodríguez-Seguí, S.A., Cho, C.H.-H., Bessa, J., Rovira, M., Luengo, M., Chhatriwala, M., Berry, A., Ponsa-Cobas, J., Maestro, M.A., et al. (2015). TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. Nat Cell Biol *17*, 615–626.

Cooper, G.M., and Brown, C.D. (2008). Qualifying the relationship between sequence conservation and molecular function. Genome Res *18*, 201–205.

Demirbilek, H., Cayir, A., Flanagan, S.E., Yıldırım, R., Kor, Y., Gurbuz, F., Haliloğlu, B., Yıldız, M., Baran, R.T., Akbas, E.D., et al. (2020). Clinical Characteristics and Long-term Follow-up of Patients with Diabetes Due To PTF1A Enhancer Mutations. The Journal of Clinical Endocrinology & Metabolism *105*, e4351–e4359.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. Cell *166*, 538–554.

Duque, M., Amorim, J.P., and Bessa, J. (2021). Ptf1a function and transcriptional cis-regulation, a cornerstone in vertebrate pancreas development. FEBS J.

Eichenlaub, M.P., and Ettwiller, L. (2011). De novo genesis of enhancers in vertebrates. PLoS Biol *9*, e1001188.

Elgar, G., and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet *24*, 344–352.

Emera, D., Yin, J., Reilly, S.K., Gockley, J., and Noonan, J.P. (2016). Origin and evolution of developmental enhancers in the mammalian neocortex. Proc Natl Acad Sci U S A *113*, E2617-2626.

ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

Eufrásio, A., Perrod, C., Ferreira, F.J., Duque, M., Galhardo, M., and Bessa, J. (2020). In Vivo Reporter Assays Uncover Changes in Enhancer Activity Caused by Type 2 Diabetes-Associated Single Nucleotide Polymorphisms. Diabetes *69*, 2794–2805.

Evliyaoğlu, O., Ercan, O., Ataloğlu, E., Zübarioğlu, Ü., Özcabı, B., Dağdeviren, A., Erdoğan, H., De Franco, E., and Ellard, S. (2018). Neonatal Diabetes: Two Cases with Isolated Pancreas Agenesis due to Homozygous PTF1A Enhancer Mutations and One with Developmental Delay, Epilepsy, and Neonatal Diabetes Syndrome due to KCNJ11 Mutation. J Clin Res Pediatr Endocrinol *10*, 168–174.

Fernández-Miñán, A., Bessa, J., Tena, J.J., and Gómez-Skarmeta, J.L. (2016). Assay for transposase-accessible chromatin and circularized chromosome conformation capture, two methods to explore the regulatory landscapes of genes in zebrafish. Methods Cell Biol *135*, 413–430.

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science *312*, 276–279.

Fujitani, Y., Fujitani, S., Boyer, D.F., Gannon, M., Kawaguchi, Y., Ray, M., Shiota, M., Stein, R.W., Magnuson, M.A., and Wright, C.V.E. (2006). Targeted deletion of a cis-regulatory region reveals differential gene dosage requirements for Pdx1 in foregut organ differentiation and pancreas formation. Genes Dev *20*, 253–266.

Furlong, E.E.M., and Levine, M. (2018a). Developmental enhancers and chromosome topology. Science *361*, 1341–1345.

Furlong, E.E.M., and Levine, M. (2018b). Developmental enhancers and chromosome topology. Science *361*, 1341–1345.

Gabbay, M., Ellard, S., De Franco, E., and Moisés, R.S. (2017). Pancreatic Agenesis due to Compound Heterozygosity for a Novel Enhancer and Truncating Mutation in the PTF1A Gene. J Clin Res Pediatr Endocrinol *9*, 274–277.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat Genet *42*, 255–259.

GBD 2017 Pancreatic Cancer Collaborators (2019). The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol *4*, 934–947.

Gordon A, Hannon G. (2003). Fastx-Toolkit. Fastq/a Short-Reads Pre-Processing Tools. (Unpublished work).

Gorkin, D.U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., et al. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. Nature *583*, 744–751.

Greenwald, W.W., Chiou, J., Yan, J., Qiu, Y., Dai, N., Wang, A., Nariai, N., Aylward, A., Han, J.Y., Kadakia, N., et al. (2019). Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. Nat Commun *10*, 2078.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88.

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

Hiller, M., Agarwal, S., Notwell, J.H., Parikh, R., Guturu, H., Wenger, A.M., and Bejerano, G. (2013). Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. Nucleic Acids Res *41*, e151.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res *34*, D590-598.

Huang, J., Lok, V., Ngai, C.H., Zhang, L., Yuan, J., Lao, X.Q., Ng, K., Chong, C., Zheng, Z.-J., and Wong, M.C.S. (2021). Worldwide Burden of, Risk Factors for, and Trends in Pancreatic Cancer. Gastroenterology *160*, 744–754.

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol *31*, 227–229.

Ishibashi, M., Mechaly, A.S., Becker, T.S., and Rinkwitz, S. (2013). Using zebrafish transgenesis to test human genomic sequences for specific enhancer activity. Methods *62*, 216–225.

Jennings, R.E., Scharfmann, R., and Staels, W. (2020). Transcription factors that shape the mammalian pancreas. Diabetologia *63*, 1974–1980.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182.

Jin, K., and Xiang, M. (2019). Transcription factor Ptf1a in development, diseases and reprogramming. Cell Mol Life Sci *76*, 921–940.

Jones, S., Li, M., Parsons, D.W., Zhang, X., Wesseling, J., Kristel, P., Schmidt, M.K., Markowitz, S., Yan, H., Bigner, D., et al. (2012). Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. Hum Mutat *33*, 100–103.

Kawakami, K., Takeda, H., Kawakami, N., Kobayashi, M., Matsuda, N., and Mishina, M. (2004). A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. Dev Cell *7*, 133–144.

Khetan, S., Kursawe, R., Youn, A., Lawlor, N., Jillette, A., Marquez, E.J., Ucar, D., and Stitzel, M.L. (2018). Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. Diabetes *67*, 2466–2477.

Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.-C., Gustafson, E.H., Sinha, S., and Furlong, E.E. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. ELife *6*, e28440.

Kimura, Y., Fukuda, A., Ogawa, S., Maruno, T., Takada, Y., Tsuda, M., Hiramatsu, Y., Araki, O., Nagao, M., Yoshikawa, T., et al. (2018). ARID1A Maintains Differentiation of Pancreatic Ductal Cells and Inhibits Development of Pancreatic Ductal Adenocarcinoma in Mice. Gastroenterology *155*, 194-209.e2.

Kinkel, M.D., and Prince, V.E. (2009). On the diabetic menu: Zebrafish as a model for pancreas development and function. Bioessays *31*, 139–152.

Klein, A.P., Wolpin, B.M., Risch, H.A., Stolzenberg-Solomon, R.Z., Mocci, E., Zhang, M., Canzian, F., Childs, E.J., Hoskins, J.W., Jermusyk, A., et al. (2018). Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. Nat Commun *9*, 556.

Klein, F.A., Pakozdi, T., Anders, S., Ghavi-Helm, Y., Furlong, E.E.M., and Huber, W. (2015). FourCSeq: analysis of 4C sequencing data. Bioinformatics *31*, 3085–3091.

Kvon, E.Z., Waymack, R., Gad, M., and Wunderlich, Z. (2021). Enhancer redundancy in development and disease. Nat Rev Genet *22*, 324–336.

Kycia, I., Wolford, B.N., Huyghe, J.R., Fuchsberger, C., Vadlamudi, S., Kursawe, R., Welch, R.P., Albanus, R. d’Oliveira, Uyar, A., Khetan, S., et al. (2018). A Common Type 2 Diabetes Risk Variant Potentiates Activity of an Evolutionarily Conserved Islet Stretch Enhancer and Increases C2CD4A and C2CD4B Expression. Am J Hum Genet *102*, 620–635.

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

Lascar, N., Brown, J., Pattison, H., Barnett, A.H., Bailey, C.J., and Bellary, S. (2018). Type 2 diabetes in adolescents and young adults. Lancet Diabetes Endocrinol *6*, 69–80.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

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

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics *5*, 1752–1779.

Lippi, G., and Mattiuzzi, C. (2020). The global burden of pancreatic cancer. Arch Med Sci *16*, 820–824.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

MacDonald, P.W., Liang, K., and Janssen, A. (2019). Dynamic adaptive procedures that control the false discovery rate. Electronic Journal of Statistics *13*, 3009–3024.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Steinthorsdottir, V., Scott, R.A., Grarup, N., Cook, J.P., et al. (2018). Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet *50*, 1505–1513.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods *9*, 1185–1188.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495–501.

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B.M., Rolando, D.M.Y., Farabella, I., Morgan, C.C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet *51*, 1137–1148.

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.-D., Fernandez, J.P., Mis, E.K., Khokha, M.K., and Giraldez, A.J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods *12*, 982–988.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet *44*, 981–990.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods *13*, 919–922.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011). The dynamic architecture of Hox gene clusters. Science *334*, 222–225.

Nord, A.S., Blow, M.J., Attanasio, C., Akiyama, J.A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., et al. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell *155*, 1521–1531.

Park, J.T., and Leach, S.D. (2018). Zebrafish model of KRAS-initiated pancreatic cancer. Anim Cells Syst (Seoul) *22*, 353–359.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., NISC Comparative Sequencing Program, et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci U S A *110*, 17921–17926.

Pashos, E., Park, J.T., Leach, S., and Fisher, S. (2013). Distinct enhancers of ptf1a mediate specification and expansion of ventral pancreas in zebrafish. Developmental Biology *381*, 471–481.

Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat Genet *46*, 136–143.

Pennacchio, L.A., and Visel, A. (2010). Limits of sequence and functional conservation. Nat Genet *42*, 557–558.

Pérez-Rico, Y.A., Boeva, V., Mallory, A.C., Bitetti, A., Majello, S., Barillot, E., and Shkumatava, A. (2017). Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes. Genome Res *27*, 259–268.

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L.I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford) *2015*, bav028.

Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell *163*, 68–83.

Prince, V.E., Anderson, R.M., and Dalgin, G. (2017). Zebrafish Pancreas Development and Regeneration: Fishing for Diabetes Therapies. Curr Top Dev Biol *124*, 235–276.

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

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rahier, J., Wallon, J., and Henquin, J.C. (1981). Cell populations in the endocrine pancreas of human neonates and infants. Diabetologia *20*, 540–546.

Roman, T.S., Cannon, M.E., Vadlamudi, S., Buchkovich, M.L., Wolford, B.N., Welch, R.P., Morken, M.A., Kwon, G.J., Varshney, A., Kursawe, R., et al. (2017). A Type 2 Diabetes-Associated Functional Regulatory Variant in a Pancreatic Islet Enhancer at the ADCY5 Locus. Diabetes *66*, 2521–2530.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A.A., Ogurtsova, K., et al. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract *157*, 107843.

Saito, K., Iwama, N., and Takahashi, T. (1978). Morphometrical analysis on topographical difference in size distribution, number and volume of islets in the human pancreas. Tohoku J Exp Med *124*, 177–186.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G.J., and Marti-Renom, M.A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol *13*, e1005665.

Shen, J., Peng, Y., Wei, L., Zhang, W., Yang, L., Lan, L., Kapoor, P., Ju, Z., Mo, Q., Shih, I.-M., et al. (2015). ARID1A Deficiency Impairs the DNA Damage Checkpoint and Sensitizes Cells to PARP Inhibitors. Cancer Discov *5*, 752–767.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Shirakawa, J., Fernandez, M., Takatani, T., El Ouaamari, A., Jungtrakoon, P., Okawa, E.R., Zhang, W., Yi, P., Doria, A., and Kulkarni, R.N. (2017). Insulin Signaling Regulates the FoxM1/PLK1/CENP-A Pathway to Promote Adaptive Pancreatic β Cell Proliferation. Cell Metab *25*, 868-882.e5.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034–1050.

Sinclair, A., Saeedi, P., Kaundal, A., Karuranga, S., Malanda, B., and Williams, R. (2020). Diabetes and global ageing among 65-99-year-old adults: Findings from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract *162*, 108078.

Snetkova, V., Ypsilanti, A.R., Akiyama, J.A., Mannion, B.J., Plajzer-Frick, I., Novak, C.S., Harrington, A.N., Pham, Q.T., Kato, M., Zhu, Y., et al. (2021). Ultraconserved enhancer function does not require perfect sequence conservation. Nat Genet *53*, 521–528.

Splinter, E., de Wit, E., van de Werken, H.J.G., Klous, P., and de Laat, W. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods *58*, 221–230.

Tiyaboonchai, A., Cardenas-Diaz, F.L., Ying, L., Maguire, J.A., Sim, X., Jobaliya, C., Gagne, A.L., Kishore, S., Stanescu, D.E., Hughes, N., et al. (2017). GATA6 Plays an Important Role in the Induction of Human Definitive Endoderm, Development of the Pancreas, and Functionality of Pancreatic β Cells. Stem Cell Reports *8*, 589–604.

Vaz, S., Ferreira, F.J., Macedo, J.C., Leor, G., Ben-David, U., Bessa, J., and Logarinho, E. (2021). FOXM1 repression increases mitotic death upon antimitotic chemotherapy through BMF upregulation. Cell Death Dis *12*, 1–14.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science *346*, 1007–1012.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wang, S.C., Nassour, I., Xiao, S., Zhang, S., Luo, X., Lee, J., Li, L., Sun, X., Nguyen, L.H., Chuang, J.-C., et al. (2019a). SWI/SNF component ARID1A restrains pancreatic neoplasia formation. Gut *68*, 1259–1270.

Wang, W., Friedland, S.C., Guo, B., O'Dell, M.R., Alexander, W.B., Whitney-Miller, C.L., Agostini-Vulaj, D., Huber, A.R., Myers, J.R., Ashton, J.M., et al. (2019b). ARID1A, a SWI/SNF subunit, is critical to acinar cell homeostasis and regeneration and is a barrier to transformation and epithelial-mesenchymal transition in the pancreas. Gut *68*, 1245–1258.

Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., De Franco, E., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., Allen, H.L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet *46*, 61–64.

Westerfield, M (2000). The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio). (Univ. of Oregon Press).

White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T., et al. (2017). A high-resolution mRNA expression time course of embryonic development in zebrafish. ELife *6*, e30860.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wittkopp, P.J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet *13*, 59–69.

Wolpin, B.M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G.M., Wang, Z., Arslan, A.A., Beane-Freeman, L., Bracci, P.M., Buring, J., et al. (2014). Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nat Genet *46*, 994–1000.

Wong, E.S., Zheng, D., Tan, S.Z., Bower, N.L., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B.M., Kikuchi, K., et al. (2020). Deep conservation of the enhancer regulatory code in animals. Science *370*, eaax8137.

Wu, J.N., and Roberts, C.W.M. (2013). ARID1A mutations in cancer: another epigenetic tumor suppressor? Cancer Discov *3*, 35–43.

Yang, S., Oksenberg, N., Takayama, S., Heo, S.-J., Poliakov, A., Ahituv, N., Dubchak, I., and Boffelli, D. (2015). Functionally conserved enhancers with divergent sequences in distant vertebrates. BMC Genomics *16*, 882.

Zhan, X., and Liu, D.J. (2015). SEQMINER: An R-Package to Facilitate the Functional Interpretation of Sequence-Based Associations. Genet Epidemiol *39*, 619–623.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9,* 137.

Chapter III

Identification and in vivo functional assessment of a pancreatic enhancer of the tumour suppressor gene *arid1a*

3.1 Introduction

Pancreatic cancer (PC) is one of the most lethal forms of cancer, with a rising incidence, prevalence, and mortality in developed countries (Saiki et al., 2021). Many efforts have been made to develop early detection methods, improve the poor prognosis, or even find more effective treatments. However, until now, only limited improvements were observed in patient outcomes (Singhi et al., 2019; Mizrahi et al., 2020).

Most of PC analysis has been largely focused on the identification of driver mutations within the protein-coding regions, where the most-well characterized pathogenic alterations are known to occur, being the contributions of non-coding regions poorly analyzed (Scarpa and Mafficini, 2018; Sondka et al., 2018). More recently, some studies described that mutations in the non-coding genome can disrupt regulatory functions, affecting the expression of genes known to be involved in the initiation and progression of PC (Scarpa and Mafficini, 2018). However, few studies have explored the impact of these non-coding alterations in vivo (Ren et al., 2021).

The *ARID1A* mutations are frequently described in PC, being found in 10% of the intraductal papillary mucinous neoplasms (IPMN; Suenaga et al., 2018). Nevertheless, the effects of the loss of ARID1A expression in pancreas have been poorly well-characterized (Wang et al., 2019a). A study of Zhu and colleagues showed that the pancreatic-specific loss *Arid1a* in mice is enough to initiate a pancreatic inflammation and generate precursor lesions, however, *Arid1a* loss-of-function alone is not sufficient for further progression to higher grades of PC (Wang et al., 2019a). Additionally, in combination with *Kras* activating mutations, *Arid1a* mutations accelerates the progression of pancreatic lesions, leading to the development of more aggressive forms of PC (Wang et al., 2019a). Zhu and colleagues also showed that by incorporating a loss of one *Trp53* allele, one of the most frequently mutated tumor suppressor genes in human cancers, in a *Arid1a* mutant mice, a subgroup of mice is able to develop PC (Wang et al., 2019a). These results clarify the role of *Arid1a* coding mutations in PC development. Additionally, in the work described in Chapter II of this doctoral thesis, we have identified a pancreatic enhancer of the human *ARID1A* gene, which deletion showed to decrease the levels of ARID1A expression in a human pancreatic duct cell line, along with a zebrafish putative functional equivalent enhancer (Bordeira-Carriço et al., 2022). These results suggested that the loss of this enhancer may interfere with the DNA-damage response, with possible implications in the increased risk for PC development (Bordeira-Carriço et al., 2022).

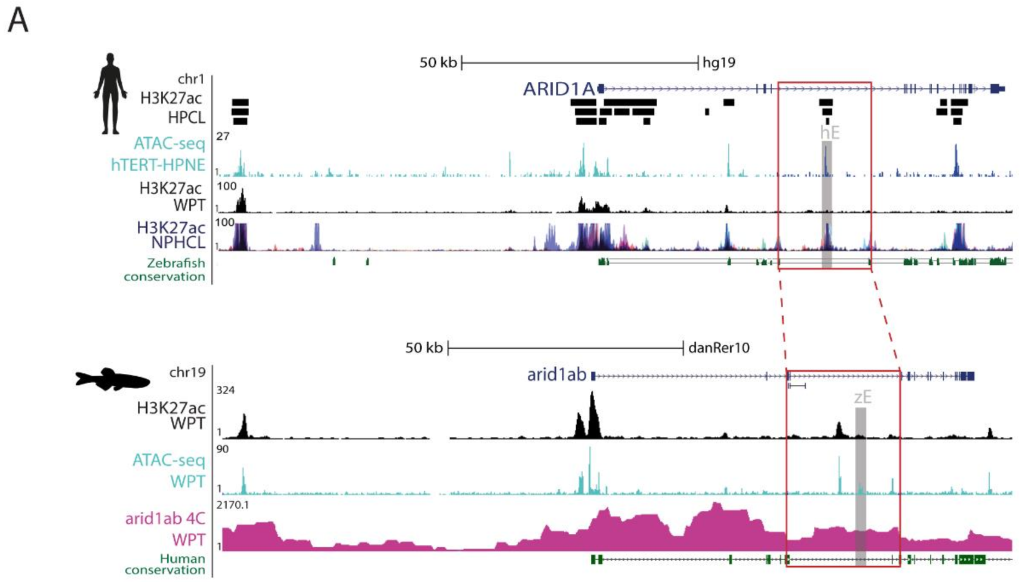
In this chapter, we aimed to address this hypothesis in vivo by targeting a pancreatic enhancer of the zebrafish *arid1ab*.

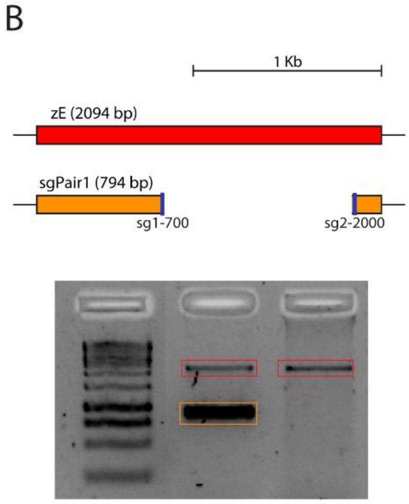
3.2 Results and discussion

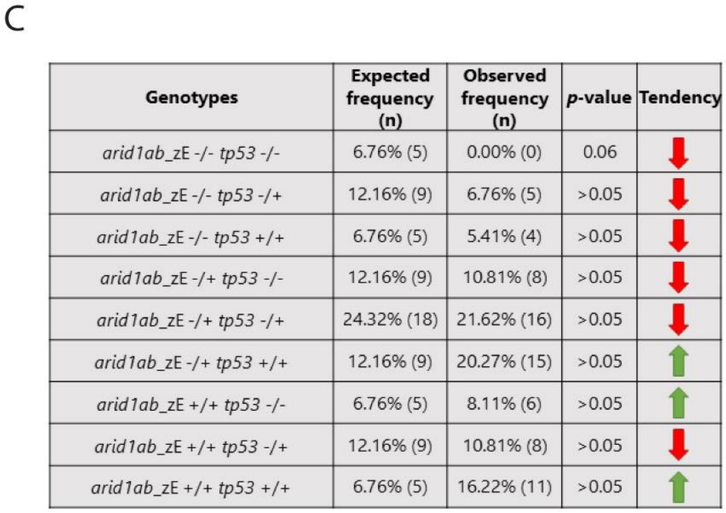
Previously, we have identified a human/zebrafish syntenic block containing a human pancreatic enhancer of the gene *ARID1A* (Fig.3.1a; hE), that upon deletion in human pancreatic duct cells, caused a downregulation of ARID1A expression. This syntenic block also contains a zebrafish arid1ab pancreatic enhancer (zE), with similar regulatory information to its human counterpart (Fig.2.4 from chapter II of this doctoral thesis). To better understand the phenotypic consequences of the loss-of-function of the arid1ab zE enhancer, in particular in the context of PC development, we have generated genomic deletions targeting the zE enhancer sequence, through the CRISPR-Cas9 system (Amorim et al., 2020). We designed and synthesized a pair of sgRNAs (sgPair1; Fig.3.1b) targeting the arid1ab zE enhancer, allowing us to isolate a genomic deletion for this enhancer (Fig.3.1b).

As previous mentioned, several studies have been described that *Arid1a* mutations alone are not able to further progress to higher grades of PC (Wang et al., 2019a; Wang et al., 2019b), but in combination with other mutations, the *Arid1a*-deficiency is potentiated, leading to the acceleration of tumor formation (Wang et al., 2019a; Wang et al., 2019b). To explore the potential role of zE loss-of-function in pancreatic tumor formation, we generated a zebrafish line containing the arid1ab enhancer mutation in a *tp53* mutant background (Berghmans et al., 2005). Then, we incrossed the heterozygous fish for these mutations (arid1ab_zE -/+; *tp53* +/-) and genotyped the respective progeny between 3 to 6 months post-fertilization. When comparing with the expected mendelian segregation, we found some differences in the genotypes obtained (Fig.3.1c). In arid1ab_zE -/-; *tp53* -/-, arid1ab_zE -/-; *tp53* -/+, arid1ab_zE -/-; *tp53* +/+; arid1ab_zE -/+; *tp53* -/-, arid1ab_zE -/+; *tp53* -/+ and arid1ab_zE +/+; *tp53* -/- fish, it was observed a decrease between the expected genotypic frequency and the observed genotypic frequency. In the particular case of arid1ab_zE -/-; *tp53* -/-, we expected to have 6.76% of the progeny with this genotype. However, in a total of 73 genotyped fish, we did not find any fish with it, meaning that the presence of homozygous mutations of arid1ab enhancer and *tp53* gene in the same fish might lead to the lethality. Additionally, we also observed a huge decreasing in the expected frequency (-5.4%) in arid1ab_zE -/-; *tp53* -/+ genotype in relation to the observed frequency, indicating that the effect caused by homozygous mutations of arid1ab enhancer could be potentiated by the presence of one mutant allele of *tp53*. This possible potentiating effect is also observed in heterozygous fish (arid1ab_zE -/+; *tp53* -/+), however with a less impact in the observed genotypic frequency (-2.7%). The opposite tendency, an increased between the expected genotypic frequency and the observed genotypic frequency, was also observed in some of the genotypes: arid1ab_zE -/+; *tp53* +/+, arid1ab_zE +/+; *tp53* -/- and arid1ab_zE +/+; *tp53* +/+. All these genotypes are in homozygosity for one of the mutants, indicating that the effect that we observed in previous results is only observed when both mutations are present.

Although, we did not find statistical significance in these results, likely because of the limited number of the genotyped animals, they are indicating that there might be a genetic interaction between arid1ab_zE and *tp53* mutations. This point needs to be further addressed in the future, increasing the number of genotyped animals.







Genotypes	Expected frequency (n)	Observed frequency (n)	*p*-value	Tendency
*arid1ab*_zE -/- *tp53* -/-	6.76% (5)	0.00% (0)	0.06	↓
*arid1ab*_zE -/- *tp53* -/+	12.16% (9)	6.76% (5)	>0.05	↓
*arid1ab*_zE -/- *tp53* +/+	6.76% (5)	5.41% (4)	>0.05	↓
*arid1ab*_zE -/+ *tp53* -/-	12.16% (9)	10.81% (8)	>0.05	↓
*arid1ab*_zE -/+ *tp53* -/+	24.32% (18)	21.62% (16)	>0.05	↓
*arid1ab*_zE -/+ *tp53* +/+	12.16% (9)	20.27% (15)	>0.05	↑
*arid1ab*_zE +/+ *tp53* -/-	6.76% (5)	8.11% (6)	>0.05	↑
*arid1ab*_zE +/+ *tp53* -/+	12.16% (9)	10.81% (8)	>0.05	↓
*arid1ab*_zE +/+ *tp53* +/+	6.76% (5)	16.22% (11)	>0.05	↑

Figure 3.1. The human and zebrafish *ARID1A/arid1ab* regulatory landscapes contain an equivalent pancreatic enhancer. a) Human *ARID1A* genomic landscape with H3K27ac enriched intervals from human pancreatic cell lines (HPCL, black bars, top-to-bottom: PT-45-P1, CFPAC-1 and HPAF-II), H3K27ac profile from human pancreas (WPT, black) and from non-pancreatic human cell lines (NPHCL; GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK and NHLF; Data from ENCODE). ATAC-seq data from human ductal cell line (hTERT-HPNE; light blue). Human/zebrafish sequence conservation (dark green). Genomic landscape of the zebrafish *arid1ab* gene (bottom), showing profiles for H3K27ac ChIP-seq (black), ATAC-seq (light blue) and 4C with viewpoint in the arid1ab promoter (pink) in adult zebrafish pancreas. Pancreatic enhancers are highlighted in grey (hE and zE) and zebrafish/human syntenic box is highlighted with red box. **b)** Schematic depiction of the targeting strategy for deletion of the zE locus. The CRISPR sgRNA target sites are depicted in blue. Agarose gel showing the wild-type (wt; red) and deleted (orange) PCR amplified zE sequence after gene editing with each respective sgRNA pair (sPair1). **c)** The expected and observed genotype frequencies, and the respective number of animals, resulting from a cross between fish with arid1ab enhancer mutant and *tp53* mutant (arid1ab_zE -/+; *tp*53 +/-). The statistical analysis was performed by unpaired Student's t-test. ns=no significant, values represent mean ± SD.

3.3 Conclusion and future perspectives

As previous mentioned, several studies have been described that *Arid1a* mutations can have a role in the development of PC (Wang et al., 2019a; Wang et al., 2019b). However, some of these studies only explore the implication of coding mutations (Wang et al., 2019a; Wang et al., 2019b). Additionally, in the work described in Chapter II of this doctoral thesis, we have identified a pancreatic enhancer of the human *ARID1A* gene, which deletion showed to decrease in the levels of ARID1A expression in a human pancreatic duct cell line, along with a zebrafish putative functional equivalent enhancer, suggesting that the loss of this enhancer may interfere with the DNA-damage response, with possible consequences in the increased risk for PC development (Bordeira-Carriço et al., 2022). Thus, in this chapter, we aimed to address this hypothesis in vivo by targeting a pancreatic enhancer of the zebrafish *arid1ab*. Since it is well described that *Arid1a* mutations alone are not able to further progress to higher grades of PC (Wang et al., 2019a; Wang et al., 2019b), we generated a zebrafish line containing the arid1ab enhancer mutation in a *tp53* mutant background (Berghmans et al., 2005). Analyzing the progeny of an incross of heterozygous fish for these mutations (arid1ab_zE -/+; tp53 +/-), we identified some differences in the genotypes obtained (Fig.3.1c), especially in the ones that carry both genomic mutations. Although, we did not find statistical significance in these observations, they are indicating that there might be a genetic interaction between arid1ab_zE and *tp53* mutations, especially in lethality observed in arid1ab_zE -/-; tp53 -/- fish. Thus, in future experiments, it would be important to further addressed this problem, increasing the number of animals analyzed. Additionally, if we confirm that exist a genetic interaction between arid1ab_zE and *tp53* mutations, next we need to understand the impact of these arid1ab_zE enhancer mutation in the transcription of the gene. To access that we need to measure and compare the amount of arid1ab that is transcribed in fish carrying the mutation or not (e.g. RT-PCR). Moreover, it would be also essential to identify the specific timepoint of death of arid1ab_zE -/+; *tp53* +/+ fish and the specific pancreatic cell type, where this mutation have more impact. We need to observe and genotype the animals since birth, to determine when the double mutans start to die. We can also cross these mutant fish with endocrine and exocrine pancreas reporter lines (e.g Insulin:GFP and Elastase:mcherry) to identify the specific pancreatic cell type affected by the enhancer mutation. This type of information is relevant because it will give us a context where the arid1ab_zE enhancer mutation are acting and will help us to understand the biological mechanisms affected. After clarifying these points, it will be also relevant to understand if these genetic interactions are able to trigger the development of PC in the pancreas of the fish. We can assess that by monitoring the formation of pancreatic tumors in fish or searching for pancreatic tissue abnormalities (e.g tissue dedifferentiation, disruption of normal borders between pancreatic tissue and other adjacent tissues; Park et al., 2008; Lodestijn et al., 2021).

3.4 Materials and Methods

3.4.1 Zebrafish stocks, husbandry, breeding and embryo rearing

Adult zebrafish AB/TU WT strains were obtained from the Gomez-Skarmeta's laboratory in Seville (CABD). WT and mutant lines were maintained at 26-28ºC under a 10h dark/14h light cycle in a recirculating housing system according to standard protocols (Westerfield, M, 2000). Embryos were grown at 28ºC in E3 medium or E3 supplemented with 0.01% PTU (1-phenyl-2-thiourea; (Ishibashi et al., 2013). For the establishment of transgenic and mutant zebrafish lines, embryos were microinjected, selected, bleached and grown until adulthood. Adult F0s were outcrossed with WT adults and the offspring screened for the internal control of transgenesis and the pattern of expression of the regulatory element, or for the respective mutations, by genotyping. The i3S animal facility and this project were licensed by Direcção Geral de Alimentação e Veterinária (DGAV) and all the protocols used for the experiments were approved by the i3S Animal Welfare and Ethics Review Body.

3.4.2 Cas9 target design, sgRNA synthesis and mutant generation

Small guide RNAS (sgRNAs) targeting regions flanking zE were designed using the CRISPRscan algorithm (Moreno-Mateos et al., 2015). Oligonucleotides (1.5μL at 100 μM each) were annealed in vitro by incubation at 95ºC for 5 min in 2x Annealing Buffer (10mM Tris-HCl, pH7.5-8.0, 50mM NaCL, 1mM EDTA) followed by slow cooling at RT, and inserted into 100ng of pDR274 vector (#42250, Addgene) previously cut with BsaI (1:10). The pDR274 vectors carrying sgRNA sequences were linearized with HindIII (1:10), purified with phenol/chloroform and transcribed with T7 RNA polymerase. Final sgRNAs were purified as described previously (Bessa et al., 2009). One cell-stage zebrafish embryos were co-injected with two sgRNAs (40 ng/µl each) and Cas9 protein (300 ng/µl). Zebrafish mutant lines for zE deletion were generated using the combinations sgRNA1+sgRNA2 (sgPair1; Table S1). Enhancer deletions in zebrafish were detected with PCR using HOT FIREPol DNA Polymerase with the flanking primers used to amplify the enhancers (Table S1). PCR products were visualized by electrophoresis in 2% agarose gel and confirmed by Sanger sequencing.

3.4.3 Zebrafish genotyping

Adult fish were genotyped by fin clipping and genomic DNA was used as template for PCR amplification (Table S1).

3.4.3 Statistical Analysis

The statistical analysis was performed by unpaired Student's t-test. In all analyses, p-value<0.05 was required for statistical significance and calculated in GraphPad Prism 5 (San Diego, CA, USA).

3.5 Supplementary table

Table S1. List of primers used in this study

Name	Sequence	Application
arid1ab_zE_Fw	AAGCAATGAAGGCTGTTTTGTTTTC	Genotyping
arid1ab_zE_Rv	TTTAGCACAGAGTGTGTTCTTGC	Genotyping
arid1ab_sgRNA1_Fw	TAGGAGAGCGTGAAGAAATCAG	Crispr-cas9
arid1ab_sgRNA1_Rv	AAACCTGATTTCTTCACGCTCTCC	Crispr-cas9
arid1ab_sgRNA2_Fw	TAGGTGAGCACAGAGCCAACAC	Crispr-cas9
arid1ab_sgRNA2_Rv	AAACGTGTTGGCTCTGTGCTCACC	Crispr-cas9

3.6 References

Amorim, J.P., Bordeira-Carriço, R., Gali-Macedo, A., Perrod, C., and Bessa, J. (2020). CRISPR-Cas9-Mediated Genomic Deletions Protocol in Zebrafish. STAR Protoc *1*, 100208. https://doi.org/10.1016/j.xpro.2020.100208.

Berghmans, S., Murphey, R.D., Wienholds, E., Neuberg, D., Kutok, J.L., Fletcher, C.D.M., Morris, J.P., Liu, T.X., Schulte-Merker, S., Kanki, J.P., et al. (2005). tp53 mutant zebrafish develop malignant peripheral nerve sheath tumors. Proc Natl Acad Sci U S A *102*, 407–412. https://doi.org/10.1073/pnas.0406252102.

Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F., et al. (2009). Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. Dev Dyn *238*, 2409–2417. https://doi.org/10.1002/dvdy.22051.

Bordeira-Carriço, R., Teixeira, J., Duque, M., Galhardo, M., Ribeiro, D., Acemel, R.D., Firbas, P.N., Tena, J.J., Eufrásio, A., Marques, J., et al. (2022). Multidimensional chromatin profiling of zebrafish pancreas to uncover and investigate disease-relevant enhancers. Nat Commun *13*, 1945. https://doi.org/10.1038/s41467-022-29551-7.

Ishibashi, M., Mechaly, A.S., Becker, T.S., and Rinkwitz, S. (2013). Using zebrafish transgenesis to test human genomic sequences for specific enhancer activity. Methods *62*, 216–225. https://doi.org/10.1016/j.ymeth.2013.03.018.

Lodestijn, S.C., van Neerven, S.M., Vermeulen, L., Bijlsma, M.F. (2021). Stem cells in the exocrine pancreas during homeostasis, injury, and cancer. Cancers 13, 3295. https://doi.org/10.3390/cancers13133295

Mizrahi, J.D., Surana, R., Valle, J.W., and Shroff, R.T. (2020). Pancreatic cancer. Lancet *395*, 2008–2020. https://doi.org/10.1016/S0140-6736(20)30974-0.

Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.-D., Fernandez, J.P., Mis, E.K., Khokha, M.K., and Giraldez, A.J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods *12*, 982–988. https://doi.org/10.1038/nmeth.3543.

Park, S.W., Davison, J.M., Rhee, J., Hruban, R.H., Maitra, A., Leach, S.D. (2008). Oncogenic KRAS induces progenitor cell expansion and malignant transformation in zebrafish exocrine pancreas. Gastroenterology *134*, 2080-2090.

Ren, B., Yang, J., Wang, C., Yang, G., Wang, H., Chen, Y., Xu, R., Fan, X., You, L., Zhang, T., et al. (2021). High-resolution Hi-C maps highlight multiscale 3D epigenome reprogramming during pancreatic cancer metastasis. J Hematol Oncol *14*, 120. https://doi.org/10.1186/s13045-021-01131-0.

Saiki, Y., Jiang, C., Ohmuraya, M., and Furukawa, T. (2021). Genetic Mutations of Pancreatic Cancer and Genetically Engineered Mouse Models. Cancers (Basel) *14*, 71. https://doi.org/10.3390/cancers14010071.

Scarpa, A., and Mafficini, A. (2018). Non-coding regulatory variations: the dark matter of pancreatic cancer genomics. Gut *67*, 399–400. https://doi.org/10.1136/gutjnl-2017-314310.

Singhi, A.D., Koay, E.J., Chari, S.T., and Maitra, A. (2019). Early Detection of Pancreatic Cancer: Opportunities and Challenges. Gastroenterology *156*, 2024–2040. https://doi.org/10.1053/j.gastro.2019.01.259.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

Suenaga, M., Yu, J., Shindo, K., Tamura, K., Almario, J.A., Zaykoski, C., Witmer, P.D., Fesharakizadeh, S., Borges, M., Lennon, A.-M., et al. (2018). Pancreatic Juice Mutation Concentrations Can Help Predict the Grade of Dysplasia in Patients Undergoing Pancreatic Surveillance. Clin Cancer Res *24*, 2963–2974. https://doi.org/10.1158/1078-0432.CCR-17-2463.

Wang, S.C., Nassour, I., Xiao, S., Zhang, S., Luo, X., Lee, J., Li, L., Sun, X., Nguyen, L.H., Chuang, J.-C., et al. (2019a). SWI/SNF component ARID1A restrains pancreatic neoplasia formation. Gut *68*, 1259–1270. https://doi.org/10.1136/gutjnl-2017-315490.

Wang, W., Friedland, S.C., Guo, B., O'Dell, M.R., Alexander, W.B., Whitney-Miller, C.L., Agostini-Vulaj, D., Huber, A.R., Myers, J.R., Ashton, J.M., et al. (2019b). ARID1A, a SWI/SNF subunit, is critical to acinar cell homeostasis and regeneration and is a barrier to transformation and epithelial-mesenchymal transition in the pancreas. Gut *68*, 1245–1258. https://doi.org/10.1136/gutjnl-2017-315541.

Westerfield, M (2000). The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio). (Univ. of Oregon Press).

Chapter IV

The importance of human pancreatic enhancers in pancreatic cancer development

This chapter includes preliminary data in preparation for publication as:
J. Teixeira, M. Galhardo, J. Bessa*. Nucleotide polymorphisms associated to increased risk of pancreatic cancer modulate pancreatic enhancer’s function.
* Corresponding author

Author contributions:
J.T. performed all wet lab experiments and respective analyses. J.T. and M.G. performed computational analyses and data interpretation. J.B. supervised the project and contributed to all analyses. J.T. and J.B. wrote the manuscript with input from all authors. All authors contributed to the development and discussion of the work. J.T and J.B wrote the manuscript.

4.1 Introduction

Pancreatic cancer (PC) is the fourth leading cause of cancer death in the western countries and the seventh leading cause of cancer-related deaths worldwide (Rawla et al., 2019; Luo et al., 2020). The incidence, prevalence, and mortality of PC have been increased by 55.0%, 63.0% and 53.0%, respectively, during the last twenty-five years and future projections propose that its burden may double during the next forty years (Lippi and Mattiuzzi, 2020; GBD 2017 Pancreatic Cancer Collaborators, 2019). Moreover, PC has one of the worst survival rates in comparison with other common cancers, with only 9.0% of patients with advanced disease surviving more than five years after diagnosis (Rawla et al., 2019; Tiriac et al., 2019). Pancreatic ductal adenocarcinoma (PDAC) is responsible for 95.0% of PC cases, being the most common form, compared with neuroendocrine or islet cell tumours which are extremely rare (Becker et al., 2014; McKenna and Edil, 2014). Although this disease is devastating, currently there are no early detection methods or effective treatments available (Xu et al., 2019). Therefore, understanding the mechanisms involved in the initiation and progression of this deathly disease is critical to the breakthrough of novel strategies for the early diagnosis and treatment.

PC is a complex disease involving genetic and non-genetic factors. Several studies have approached the causes of genetic susceptibility for PC, many focusing on the coding genome (Felsenstein et al., 2018; Scarpa and Mafficini, 2018). Several genes have been associated to the development of PC, among them, the *nuclear receptor subfamily 5 group A member 2 (NR5A2)*, a member of the orphan nuclear hormone receptors family (Luo et al., 2017). This gene is highly expressed in several organs such as the ovaries, intestine, liver, and pancreas, being part of liver and pancreas early development and exocrine differentiation in adulthood (Lazarus et al., 2012; Lee and Moore, 2008; Lin et al., 2014). However, the precise role of *NR5A2* in PC is still unclear (Guo et al., 2021). Several studies have been described that *NR5A2* overexpression in PC cell lines promotes the cell migration and invasion, leading to epithelial-to-mesenchymal transition (Lin et al., 2014). In contrast, other studies have shown that in mice, heterozygous mutations in the *Nr5a2* gene makes the pancreas more susceptible to damage, and in cooperation with other mutations, can lead to pancreatic tumorigenesis (Flandez et al., 2014).

Far less studies on the genetic susceptibility of PC have been done using more unbiased genome-wide approaches, including genome-wide association studies (GWAS; Klein et al., 2018; Campa et al., 2020). In these GWAS, it has been observed that many PC associated variants are non-coding (López de Maturana et al., 2021; Maurano et al., 2012; Arnes et al., 2019), however, only few studies have functionally approached the role of non-coding sequences in PC (Diaferia et al., 2016; Feigin et al., 2017).

During many years, non-coding regions of the DNA have been considered as “junk DNA”. However, these regions start to gain relevance since they have been shown to have transcriptional regulatory functions that can be easily predicted due to the technological advancement in high-throughput sequencing and chromatin profiling (Scacheri and Scacheri, 2015; Alexander et al., 2010). These non-coding regions of the DNA can be cis-regulatory elements (CREs), that precisely control the transcription of target genes. Among different CREs, enhancers interact with the promoter region of target genes, controlling and increasing the tissue-specific gene’s expression (Pennacchio et al., 2007). Thus, the identification of active cis-regulatory regions in the human genome is crucial for understanding gene’s activity and assessing the impact of genetic alterations in the development of human diseases (Worsley-Hunt et al., 2011; Coppola et al., 2016). Many genome-wide assays have been developed to identify CREs, being chromatin immunoprecipitation followed by sequencing (ChIP-seq) one of the most widely used methods. This assay allows a genome-wide prediction of active CREs by profiling epigenetic modifications of histones (Barski et al., 2007). The presence of H3K27ac and H3K4me1 marks active enhancers, while active promoters are characterized by the presence of H3K27ac and H3K4me3 (Creyghton et al., 2010; Ernst and Kellis, 2017). On the other hand, regions marked by the presence of H3K27me3 are identified as regions containing repressed chromatin (Cai et al., 2021; Ernst and Kellis, 2017). ChromHMM is a useful software that combines numerous genome-wide epigenomic profiles and applies combinatorial and spatial mark patterns to infer a complete annotation of CREs for each tissue or cell type (Ernst and Kellis, 2017). Arda and colleagues (Arda et al., 2018) generated several chromatin maps for different cell populations of pancreas that allows the identification of putative active CREs in a cell-type-specific manner (Arda et al., 2018). Thus, the deep analysis of these pancreatic chromatin profiles could be useful to identify pancreatic CREs and understand if alterations in these non-coding sequences might contribute to gene’s transcriptional changes that could be the trigger to an increased risk in the development of PC.

In the present study, we showed that PC risk variants are enriched in genomic locations with epigenetic marks associated to enhancer activity in pancreatic duct and acinar cells. Furthermore, the target genes of the putative pancreatic enhancers that overlap with PC associated SNPs, are enriched for pancreatic development and cis-regulatory functions, suggesting that PC associated SNPs might dysregulate important pancreatic genes. Next, we validated in vitro some of these sequences as enhancer, performing luciferase enhancer reporter assays in a human duct cell line and in vivo using zebrafish. Furthermore, we also showed that for some of the tested sequences, the PC risk allele impacts significantly in the regulatory output of the enhancer, when compared with the non-risk allele. Focusing on the genomic landscape of the *NR5A2* gene, we found 4 enhancers that we validated in vitro. One of these enhancers, seq44, showed a dramatic decrease in its enhancer activity when harbouring PC risk allele, compared with the non-risk allele, suggesting that it could dysregulate *NR5A2* expression. Overall, this study shows that genetic variation in pancreatic enhancers may be a contributing factor to PC genetic risk.

4.2 Results and Discussion

4.2.1 Pancreatic cancer risk variants are enriched in human putative pancreatic enhancers of duct and acinar cells

To determine if PC risk variants are enriched in enhancers active in human pancreatic cells, we first searched for risk variants associated with PC in the DisGeNET database (Piñero et al., 2020), which contains a comprehensive collection of genetic variants associated with human diseases. We found 278 PC-associated variants, 115 (41.4%) of which were found in non-coding regions and 163 (58.6%) of them were in coding regions (Table S1). Additionally, we also searched for human pancreatic ChIP-seq datasets that allow the identification of putative CREs active in the different cell populations of the pancreas. To perform a more complete analysis, we used the datasets generated by ChromHMM software, that through the combination of several ChIP-seq datasets for different histone marks, allows the annotation of different regulatory categories: “enhancers”, “promoters”, “repressed chromatin” and “no signal” (Table S2 and Table S3) in each pancreatic cell type (Arda et al., 2018). Additional information about the chromatin profiles included in each regulatory category are described in material and methods section. Because the vast majority of PC types derive from acinar and duct cells (Backx et al., 2022), including the ones used in this study, we selected chromatin profiles from acinar and duct cells. In addition, and because we did not include in our study risk variants associated to PC types derived from endocrine cells, we selected chromatin profiles from pancreatic endocrine cells as a control, and as less related tissues, colon, and heart ventricle. Comparing the percentage of base pairs (bp) of each category in each pancreatic cell type, we observed that enhancer regions are present in similar proportions in the three pancreatic cell types (Supplementary Fig.4.1a, Table S4; Duct – 14.4%; Acinar – 13.6%; Endocrine – 18.7%). In contrast, “repressed chromatin” and “no signal” categories vary greatly between the different pancreatic cell types (Supplementary Fig.4.1a and Table S4; Repressed chromatin: Duct – 15.0%; Acinar – 4.22%; Endocrine – 41.86%. No signal: Duct – 49.0%; Acinar – 65.0%; Endocrine – 34.2%).

To understand if there is an enrichment of PC risk variants in any chromatin category described in the different cell types used, we calculated the percentage of overlap of non-coding PC-associated variants, designated as “PC SNPs”, within human pancreatic chromatin states categories from the different pancreatic cell types (Fig.4.1a-c; Table S2-S4). As controls, we have performed a similar assay using a set of random SNPs (15175044 SNPs) from the 1000 genomes annotations (1000 Genomes Project Consortium et al., 2015), designated as “Control SNPs” (Fig.4.1a-c; Table S4). Regarding the pancreatic duct cells, we found that 31.9% of “PC SNPs” overlap with enhancer category comparing with 14.1% of “Control SNPs” (Fig.4.1a; $p<0.0001$; Table S4). Additionally, we observed that 20.7% of “PC SNPs” overlap with repressed chromatin category comparing with 37.5% of “Control SNPs” (Fig.4.1a; $p<0.001$; Table S4). Similar results were obtained when using chromatin profiles of pancreatic acinar cells, where we observed that 28.7% of “PC SNPs” overlap with enhancer category in comparison with 13.0% of “Control SNPs” (Fig.4.1b; $p<0.0001$; Table S4). In contrast, we observed that 13.9% of “PC SNPs” overlap with repressed chromatin category in pancreatic acinar cells in comparison with 23.3% of “Control SNPs” (Fig.4.1b; $p<0.05$; Table S4).

These results contrasted with the ones observed in pancreatic endocrine cells (Fig.4.1c and Table S4), where the percentage of “PC SNPs” and “Control SNPs” that overlap with the different regulatory categories showed similar values. We observed that 21.1% of “PC SNPs” overlap with enhancer category in comparison with 18.2% of “Control SNPs”. Additionally, we observed that 47.4% of “PC SNPs” overlap with repressed chromatin category in comparison with 48.2% of “Control SNPs” in endocrine cells. Regarding the colon and heart ventricle tissues, we also observed similar values in the overlap with “PC SNPs” and “Control SNPs” (Supplementary Fig.4.3). We observed that 7.0% (colon) and 4.4% (heart ventricle) of “PC SNPs” overlap with the enhancer category, in comparison to 4.0% (colon) and 3.7% (heart ventricle) of “Control SNPs” (Table S4). Overall, we found an enrichment in “PC SNPs” in putative enhancers active in pancreatic duct and acinar cells, suggesting that alterations in duct and acinar enhancer sequences might affect their cis-regulatory functions, increasing the risk for PC development. Additionally, we observed that “PC SNPs” are depleted from putative enhancers active in tissues distantly related to the pancreas, as the right ventricle and colon, which can be explained by the known tissue specificity property of enhancers (Tobias et al., 2021).

To understand if the observed enrichment of “PC SNPs” in pancreatic duct and acinar enhancers is specific of PC-associated variants, we have performed a similar analyses using variants associated with other types of cancer. This set of risk variants was designated as “Other Cancer SNPs”. Regarding the pancreatic duct cells, we found that 20.9% of “Other Cancer SNPs” overlap with enhancer category (*p*<0.0001, comparing to the “Control SNPs” category), a smaller percentage in comparison to the 31.9% of “PC SNPs” (Supplementary Fig4.2 and Table S4; *p*<0.01). Regarding the pancreatic acinar cells, we found that 19.5% of “Other Cancer SNPs” (*p*<0.0001, comparing to the “Control SNPs” category), also a smaller percentage in comparison to the 28.7% of “PC SNPs” (Supplementary Fig.4.2 and Table S4; *p*<0.05). Importantly, for the endocrine, right ventricle and colon tissues, we observed that 25.4%, 6.8% and 6.1% of “Other Cancer SNPs” overlap with enhancers, comparing with 18.2%, 4.0% and 3.7% of “Control SNPs”, respectively (Supplementary Fig.4.2 and Table S4; *p*<0.0001). These results suggest that variants associated to other types of cancer not related to pancreatic tissues tend to be located in active pancreatic enhancers. These results might be explained by the accumulation of functions observed in some enhancers. Although enhancers tend to have tissue and cell-type specific activities, they might also be active in several and different cell types (Andersson et al., 2014), some of them even showing a complete non-tissue specific characteristic, being described as ubiquitous enhancers (Andersson et al., 2014). Having this in mind, next we asked how many duct and acinar putative enhancers are shared and how many are exclusively active in each of these pancreatic cell types. It is well described that the exocrine part of the pancreas, which constitutes 95% of the total pancreatic mass, is mainly composed by acinar and duct cells (Habener et al., 2005; Jennings et al., 2020). However, it is not known if these two different cell types shared or not many putative active enhancers.

We found 122608 enhancers to be only active in pancreatic acinar cells (OnlyA), 138548 only active in pancreatic duct cells (OnlyD) and 303083 to be active in these two pancreatic cell types (Shared; Fig4.1d). Then, we calculated the overlap of “PC SNPs” and "Control SNPs” with the three sets of putative enhancers (OnlyA, OnlyD and Shared). We found that the biggest enrichment of PC risk variants was detected in the Shared group of enhancers (Fig.4.1e; *p*<0.0001), suggesting that these variants that affect enhancers active in these two cell types, duct and acinar cells, might have a higher impact in the dysregulation of genes that could contribute to the development of PC.

Overall, we detected an enrichment in “PC SNPs” in putative enhancers active in pancreatic duct and acinar cells, contrasting with the results that we observed in endocrine cells or in other cell types not related to pancreas. These results are indicating that the dysregulation of gene transcription in these two pancreatic cell types might have a relevant contribution in the development of PC. In fact, several studies have been intensively explored the cell origin of PDAC, being the exocrine pancreas, a tissue composed mainly for acinar and duct cells, proposed as the principal source of PDAC genesis (Stanger and Dor, 2006; Backx et al., 2022). As previously described, PDAC is the more aggressive and lethal form of PC. However, the specific cell origin of this type of PC still until now to clarify (Backx et al., 2021; Wood and Maitra, 2021). Curiously, with our analysis, we observed a huge enrichment of these variants in the group of putative enhancers shared between these two pancreatic cell types. Thus, we can speculate that the origin of this pancreatic disease may reside in both pancreatic cells types and being the reason why is so difficult to distinguish the acinar and duct contribution to PC genesis. Thus, when the activity of these shared enhancers is affected, the transcriptional gene regulation of both cell types might be also modified, compromising the function of whole exocrine pancreas.

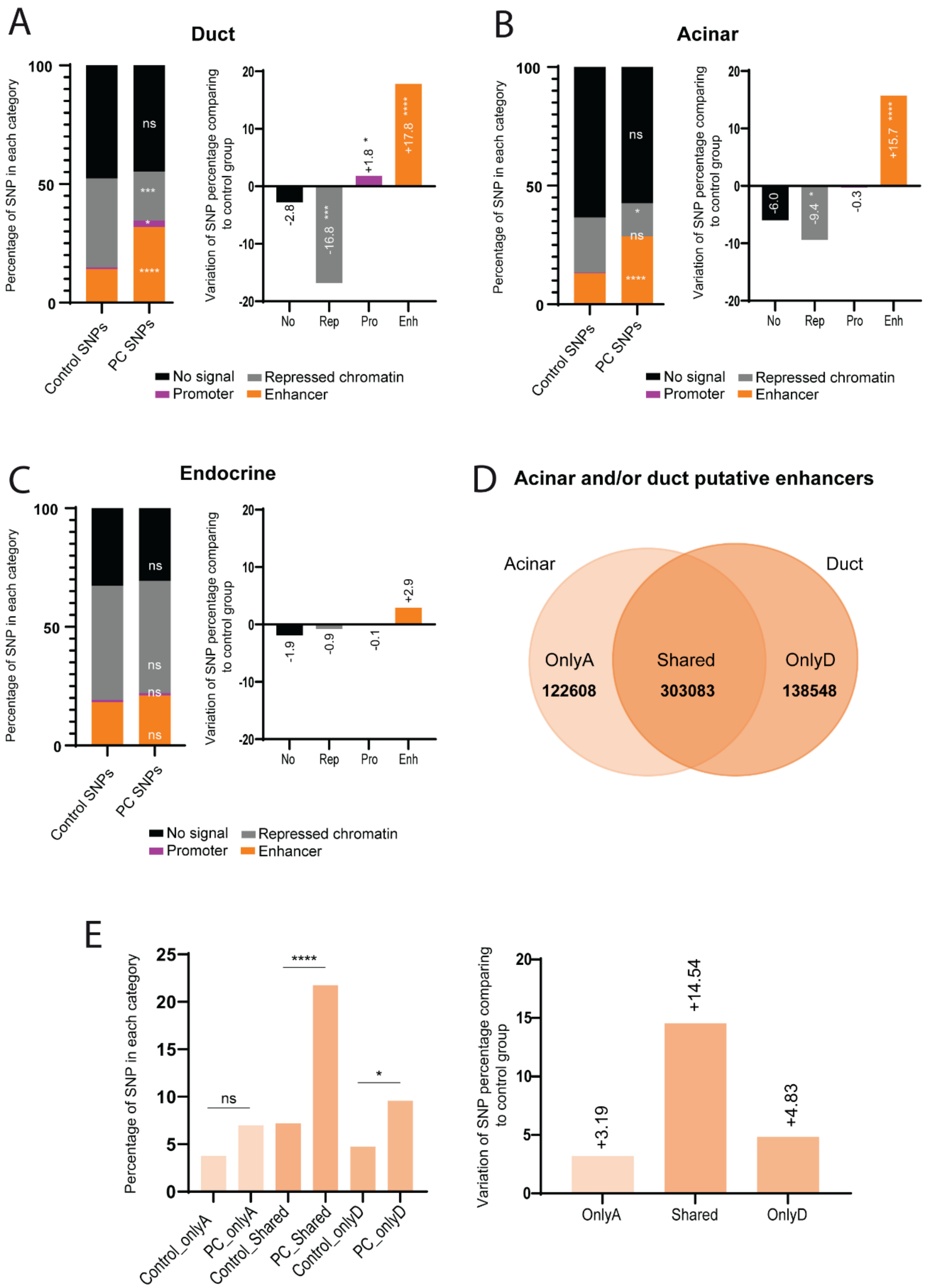

A
Duct
Percentage of SNP in each category
100
50
0
ns
***
*
****
Control SNPs
PC SNPs
Variation of SNP percentage comparing to control group
20
10
0
-10
-20
-2.8
-16.8 ***
+1.8 *
+17.8 ****
No
Rep
Pro
Enh
No signal
Repressed chromatin
Promoter
Enhancer
B
Acinar
Percentage of SNP in each category
100
50
0
ns
*
ns
****
Control SNPs
PC SNPs
Variation of SNP percentage comparing to control group
20
10
0
-10
-20
-6.0
-9.4 *
-0.3
+15.7 ****
No
Rep
Pro
Enh
No signal
Repressed chromatin
Promoter
Enhancer
C
Endocrine
Percentage of SNP in each category
100
50
0
ns
ns
ns
ns
Control SNPs
PC SNPs
Variation of SNP percentage comparing to control group
20
10
0
-10
-20
-1.9
-0.9
-0.1
+2.9
No
Rep
Pro
Enh
No signal
Repressed chromatin
Promoter
Enhancer
D
Acinar and/or duct putative enhancers
Acinar
Duct
OnlyA
122608
Shared
303083
OnlyD
138548
E
Percentage of SNP in each category
25
20
15
10
5
0
ns
****
*
Control_onlyA
PC_onlyA
Control_Shared
PC_Shared
Control_onlyD
PC_onlyD
Variation of SNP percentage comparing to control group
20
15
10
5
0
+3.19
+14.54
+4.83
OnlyA
Shared
OnlyD

Figure 4.1. Human pancreatic chromatin state categories (duct, acinar and endocrine cell types) and its overlapping with SNPs. **a)** The percentage of variants that overlap with chromatin stage categories in duct cells and the respective variation between groups. **b)** The percentage of variants that overlap with chromatin stage categories in acinar cells and the respective variation between groups. **c)** The percentage of variants that overlap with chromatin stage categories in endocrine cells and the respective variation between groups. The random set of variants is labelled as "Control SNPs" and the variants associated to pancreatic cancer is labelled as "PC SNPs". **d)** The putative enhancer regions present in duct and acinar cells (OnlyA=putative enhancer regions only present in acinar cells; Shared=putative enhancer regions shared between duct and acinar cells; OnlyD=putative enhancer regions only present in duct cells). **e)** The percentage of PC and "Control SNPs" in each set of putative enhancers and the respective variation between them. Control_OnlyA = The percentage of "Control SNPs" that overlap with the set of putative enhancer only present in acinar cells; Pc_OnlyA = The percentage of "PC SNPs" that overlap with the set of putative enhancer only present in acinar cells; Control_Shared = the percentage of "Control SNPs" that overlap with the set of putative enhancer shared between duct and acinar cells; Pc_Shared = The percentage of "PC SNPs" that overlap with the set of putative enhancer shared between duct and acinar cells; Control_OnlyD=The percentage of "Control SNPs" that overlap with the set of putative enhancer only present in duct cells; Pc_OnlyD=The percentage of "PC SNPs" that overlap with the set of putative enhancer only present in duct cells. Values are represented as percentages and compared by two-sided Chi-square with Yates' correction test. *p*-values<0.05 were considered significant (*****p*<0.0001, **p*<0.05, ns, no statistical significance). Chromatin categories: No – No signal; Rep – Repressed chromatin; Pro – Promoter; Enh – Enhancer.

4.2.2 Duct and acinar enhancers that overlap with PC SNPs are enriched in genes involved in pancreatic development and transcriptional regulation

Since we observed that "PC SNPs" might be affecting enhancer functions in pancreatic duct and acinar cells, it is important to identify which are the genes that are controlled by these putative enhancer regions as well as their biological functions. Using GREAT software, we identified the nearest genes to each putative pancreatic enhancer that overlap with "PC SNPs" (McLean et al., 2010), as the best candidate genes to be controlled by these enhancers. Regarding the putative enhancers active in pancreatic duct cells that overlap with "PC SNPs", we identified 52 different genes associated by proximity to them, and 48 different genes associated to equivalent acinar enhancers group (Table S5 and Table S6). Using this list of genes, we performed a gene ontology enrichment assay using PANTHER (Mi et al., 2019). We observed that the genes from the duct cells group are enriched for pancreatic developmental functions (*p*<1.76E-5; Fig.4.2a and Table S7) and the genes from the acinar cells group are enriched for cis-regulatory functions (*p*<6.09E-3; Fig.3.2b and Table S8). As a control, we performed a gene ontology enrichment analysis for the genes identified by GREAT (McLean et al., 2010) that are nearby all the putative enhancers active in pancreatic duct cells and acinar cells, observing a lack of GO terms enrichment for these two groups of genes (Table S9 and

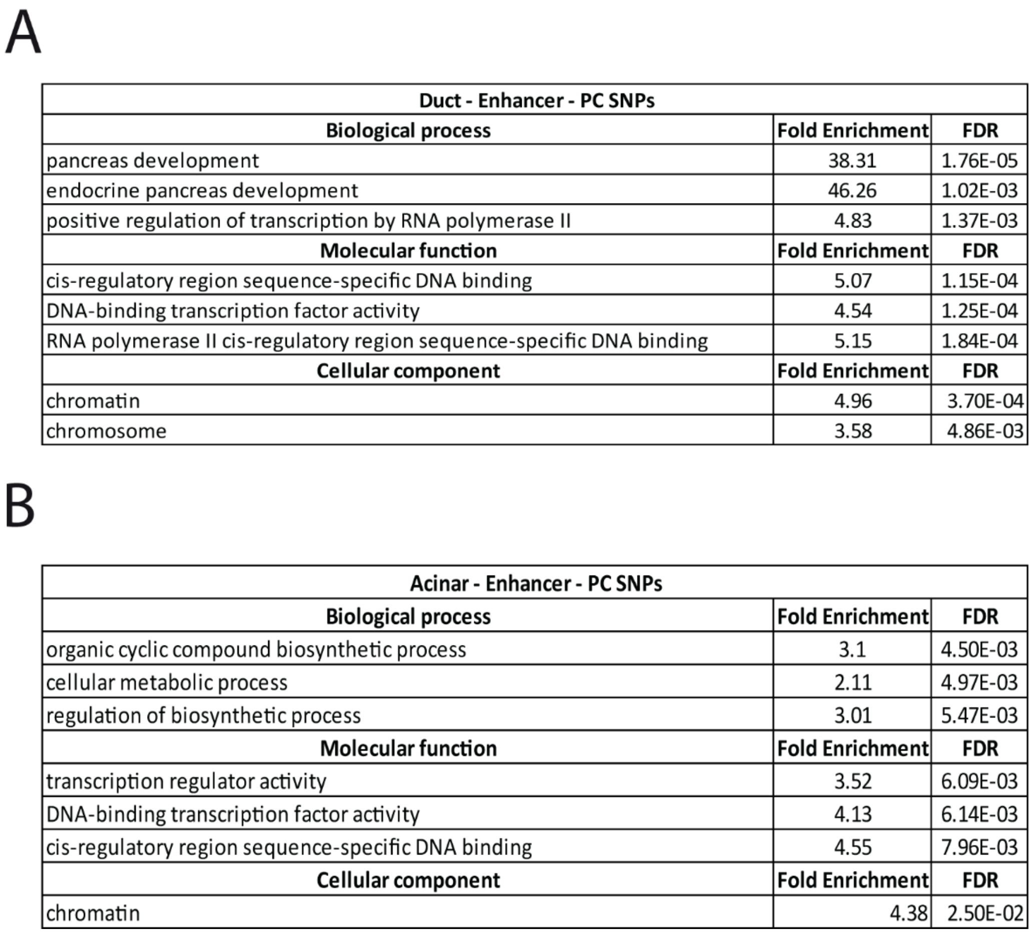
Table S10). These results suggest that enhancers that overlap with SNPs associated to PC control genes associated with cis-regulatory functions, with the potential to affect complex genetic networks, and pancreatic developmental functions.

To further explore the function of the different genes associated by proximity to putative duct and acinar enhancers that overlap with “PC SNPs”, we went to CancerGenetics website (http://www.cancerindex.org), a database that provide a comprehensive and reliable information about genes and proteins, and genetic variations associated with cancer, in order to identify the number of genes associated to PC and with other types of cancer. Regarding the genes associated to enhancers active in duct cells, we observed that 9.62% (5/52) of them are linked to PC and 36.65% (19/52) of them are described as having a role in other types of cancer. In the case of the genes associated to enhancers active in acinar cells, we found that 8.70% (4/46) of them are already linked to PC and 32.60% (15/46) of them are described as having an association with other types of cancer. It is known that the human genome contains around 19.116 nuclear protein-coding genes (Piovesan et al., 2019). Additionally, in CancerGenetics website, 229 genes are described as having a role in PC and around 2178 genes are linked to other types of cancer. Based on this, in the list of genes associated to duct and acinar enhancers, we found an enrichment in genes associated to PC, in comparison with the nuclear protein-coding genes (duct: 8.33 of fold enrichment with p<0.00001; acinar: 7.27 of fold enrichment with p<0.00001). In addition, we also found an enrichment in genes related with other types of cancer, in these two lists of genes, however with smaller values of fold enrichment (duct: 3.15 of fold enrichment with p<0.00001; acinar: 2.86 of fold enrichment with p<0.00001). Next, we want to evaluate what are the genes with more enhancers nearby. Looking to the lists of genes that we obtained from the previously analysis, we found that the gene with more duct putative enhancers associated was the *NR5A2* gene (5 enhancers; Table S5), already studied in the context of PC (Cobo et al., 2018). In the case of acinar cells, although the *NR5A2* is not the gene with more putative enhancers associated, it is present in the top 3 and it is the first top gene associated PC (4 enhancers, Table S6).

In this section, we observed that the putative enhancers active in pancreatic duct cells that overlap with “PC SNPs”, are enriched in genes with pancreatic developmental functions. This class of genes is really important during the pancreas organogenesis, controlling the formation of a mature organ. Furthermore, in adulthood some of these genes still having a relevant function, keeping the cells in a differentiated state (Bastidas-Ponce et al., 2017; Pan and Wright, 2011). Thus, our results are indicating that, when the activity of duct enhancers are affected, this could alter the function of their target genes that are responsible for the maintenance of the pancreatic differentiated state. And this alteration could lead to the transformation of mature pancreatic cells into a more progenitor and proliferative cells, a typical phenotype of PC cells (Kong et al., 2011). Previous studies already described that genetic alterations in genes involved in pancreatic development might contribute to PC development.

Additionally, we also observed that the putative enhancers active in pancreatic acinar cells that overlap with “PC SNPs”, are enriched in genes with cis-regulatory functions, particularly genes that encode for transcription factors (TFs). This result is particularly interesting because it is well described that TFs are the main regulators of gene expression and they are involved in several and complex genetic networks (Mitsis et al., 2020; Wilkinson et al., 2017). Thus, when the activity of acinar enhancers is altered, this could have a huge impact in acinar cells, because it might affect the transcription of thousands of genes, some of them with important roles in this pancreatic cell type.

Furthermore, with these analysis, we also detected that the *NR5A2* is one of the genes with more putative enhancer regions nearby, indicating that this gene could have important functions in these two pancreatic cells types. Based on the results that we observed the previous section (section 4.2.2), it is expected that both cell types shared the same top genes, because we found that a huge number of enhancers that overlap with “PC SNPs” are shared between pancreatic acinar and duct cells. It is known that this gene is important in early pancreatic development and have a particular function in acinar differentiation (von Figura et al., 2014; Cobo et al., 2018). However, in the pancreatic duct cells any important function of this genes was until now described. Additionally, the precise contribution of *NR5A2* gene in PC context is also still unclear (Guo et al., 2021). It is well described that the loss of acinar identity is one of the first stages in PC initiation (von Figura et al., 2014). Based on this information and regarding the results presented here, we can propose that in pancreatic acinar cells, alterations in the enhancer sequences that control the expression of *NR5A2* gene could have an impact its transcription, causing a downregulation of gene, that can lead to the de-differentiation of acinar cells and consequent loss of its identity.



A

Duct - Enhancer - PC SNPs		
Biological process	**Fold Enrichment**	**FDR**
pancreas development	38.31	1.76E-05
endocrine pancreas development	46.26	1.02E-03
positive regulation of transcription by RNA polymerase II	4.83	1.37E-03
Molecular function	**Fold Enrichment**	**FDR**
cis-regulatory region sequence-specific DNA binding	5.07	1.15E-04
DNA-binding transcription factor activity	4.54	1.25E-04
RNA polymerase II cis-regulatory region sequence-specific DNA binding	5.15	1.84E-04
Cellular component	**Fold Enrichment**	**FDR**
chromatin	4.96	3.70E-04
chromosome	3.58	4.86E-03

B

Acinar - Enhancer - PC SNPs		
Biological process	**Fold Enrichment**	**FDR**
organic cyclic compound biosynthetic process	3.1	4.50E-03
cellular metabolic process	2.11	4.97E-03
regulation of biosynthetic process	3.01	5.47E-03
Molecular function	**Fold Enrichment**	**FDR**
transcription regulator activity	3.52	6.09E-03
DNA-binding transcription factor activity	4.13	6.14E-03
cis-regulatory region sequence-specific DNA binding	4.55	7.96E-03
Cellular component	**Fold Enrichment**	**FDR**
chromatin	4.38	2.50E-02

Figure 4.2. Gene ontology enrichment analysis in pancreatic duct and acinar cells with the respective fold enrichment and false discovery rate (FDR). a) Gene ontology enrichment analysis for the genes associated by proximity to putative enhancer regions in duct cells that overlap with PC risk SNPs. **b)** Gene ontology enrichment analysis for the genes associated by proximity to putative enhancer regions in acinar cells that overlap with PC risk SNPs**.**

4.2.3 Pancreatic risk alleles modulate enhancer activity in vitro

We then wanted to functionally test if the duct and acinar putative cis-regulatory sequences, that overlap with PC risk variants, have enhancer activity in pancreas. We focused our analysis on the landscape of *NR5A2* gene since it was one of the gene with more putative pancreatic enhancers associated and it was already described that mutations in this gene contribute to PC development. We selected 4 sequences from this genomic landscape, that overlap with PC risk SNPs and with different pancreatic chromatin state categories in pancreatic duct and acinar cells. Seq41 and seq38 overlap with "enhancer" category in duct cells and with "no signal" category in acinar cells, while seq44 and seq67 overlap with “enhancer” category in both pancreatic cell types (Fig.4.3a). The respective sequences, containing the non-risk allele (wt), were cloned in p.GL4.23GW vector and luciferase reporter assays were performed in a human pancreatic duct cell line (hTERT-HPNE) to test their enhancer activity. We performed the enhancer reporter assays in a pancreatic duct cell line, since no human acinar cell line is reported and studies with primary cells isolated from human healthy donors are scarce (Backx et al., 2022). Out of the 4 tested sequences, all induced significant luciferase activity in comparison to the control (seq41wt, *p*<0.01; seq38wt, *p*<0.05; seq44wt, *p*<0.0001 and seq67wt, *p*<0.0001), demonstrating that these 4 sequences are duct enhancers (Fig.4.3c).

Then, to understand the possible impact that PC risk SNPs have in the enhancer activity of these sequences, we performed luciferase assays in hTERT-HPNE cell lines with vector containing the PC risk allele. Of the 4 previously identified duct enhancers, only 1, seq44, showed a significant change in enhancer activity for the respective risk allele, corresponding to a 4.7-fold decrease in luciferase activity comparing to the non-risk allele (Fig.4.3c; *p*<0.01). These results suggest that the risk allele could be affecting the enhancer function, consequently decreasing the transcription of *NR5A2*. As previously described, several studies have been exploring the role of *NR5A2* in the development of PC. However, until now there is not a clear function for this gene in the development of this pancreatic disease, as explained in section 2.22 (Guo et al., 2021; Cobo et al., 2018). Several studies have been observed that mutations in the *Nr5a2* gene makes the pancreas more prone to damage, leading to the development of acute pancreatitis. This inflammatory state increases the risk of developing cancer (Flandez et al., 2014; Cobo et al., 2018). Based on this and regarding the decrease in luciferase activity that we observed in seq44risk, we can speculate that alterations in this enhancer sequence, that downregulate the expression of *NR5A2*, could make the pancreas more prone to damage, increasing the change to form pancreatic tumours.

Then, we explored another regulatory landscape that contains several PC risk alleles, although the genes observed in this landscape have not yet been extensively associated to the PC (*MEIS1*, *LINC01829* and *ETAA1*; Fig.4.3b). We selected 3 sequences from this landscape, that overlap with PC risk SNPs and with different pancreatic chromatin state categories in pancreatic duct and acinar cells. Seq56 overlap with “no signal” category in duct and acinar cells. Seq65 and seq34 overlap with “no signal” category in acinar cells and with “enhancer” category in duct cells (Fig.4.3b). Out of the 3 tested sequences containing the non-risk allele, none was able to induce significant luciferase activity in comparison to the control. However, when we performed luciferase assays with the sequences containing the risk alleles, 2 sequences (seq56risk and seq34risk) showed enhancer activity (Fig.4.3e). Comparing the enhancer activity of wt and risk alelles, we found that in the seq56, the risk allele shows a significant increase in enhancer activity (Fig.4.3e; $p<0.01$).

Comparing the results obtained in the two genomic landscapes explored in this study, we observed that the ChromHMM data for pancreatic duct cells is robust for enhancer prediction, because 66.7% (4/6) of the predicated enhancer regions showed enhancer activity in luciferase assays. Although we have selected a small number of sequences to test, our percentage is similar to other studies (70.5%; Yue et al., 2014). Additionally, we can also observe that the impact of the risk SNP in the enhancer activity of the sequences can differ. In the case of seq44 present in the landscape of *NR5A2* gene, we observed a decrease in the enhancer activity, when the risk SNP is present. However, in the case of seq56, we observed an increase in enhancer activity with the risk SNP. Thus, these results suggest that PC SNPs have the potential to be translated into a loss or a gain of function of the target genes, because they can have different impacts in enhancer activity. Similar results were found by Eufrásio and colleagues work, where they tested several putative endocrine pancreatic enhancers that overlap with SNPs associated to type 2 diabetes (Eufrásio et al., 2020).

Moreover, we also note differences in enhancer activity in the same sequence containing the wt or risk SNP. These differences could be explained by the binding of different TFs in the enhancer sequence. The presence of risk or wt SNP in sequence change the nucleotide and this alteration could modify the recognition site of a specific TF. So, the alteration of this recognition site, could decrease or increase the binding affinity of specific TFs, affecting type of TFs that binds, and consequently changing the enhancer activity of the sequence (Eufrásio et al., 2020; Tjian and Maniatis, 1994; Ong and Corces, 2011).

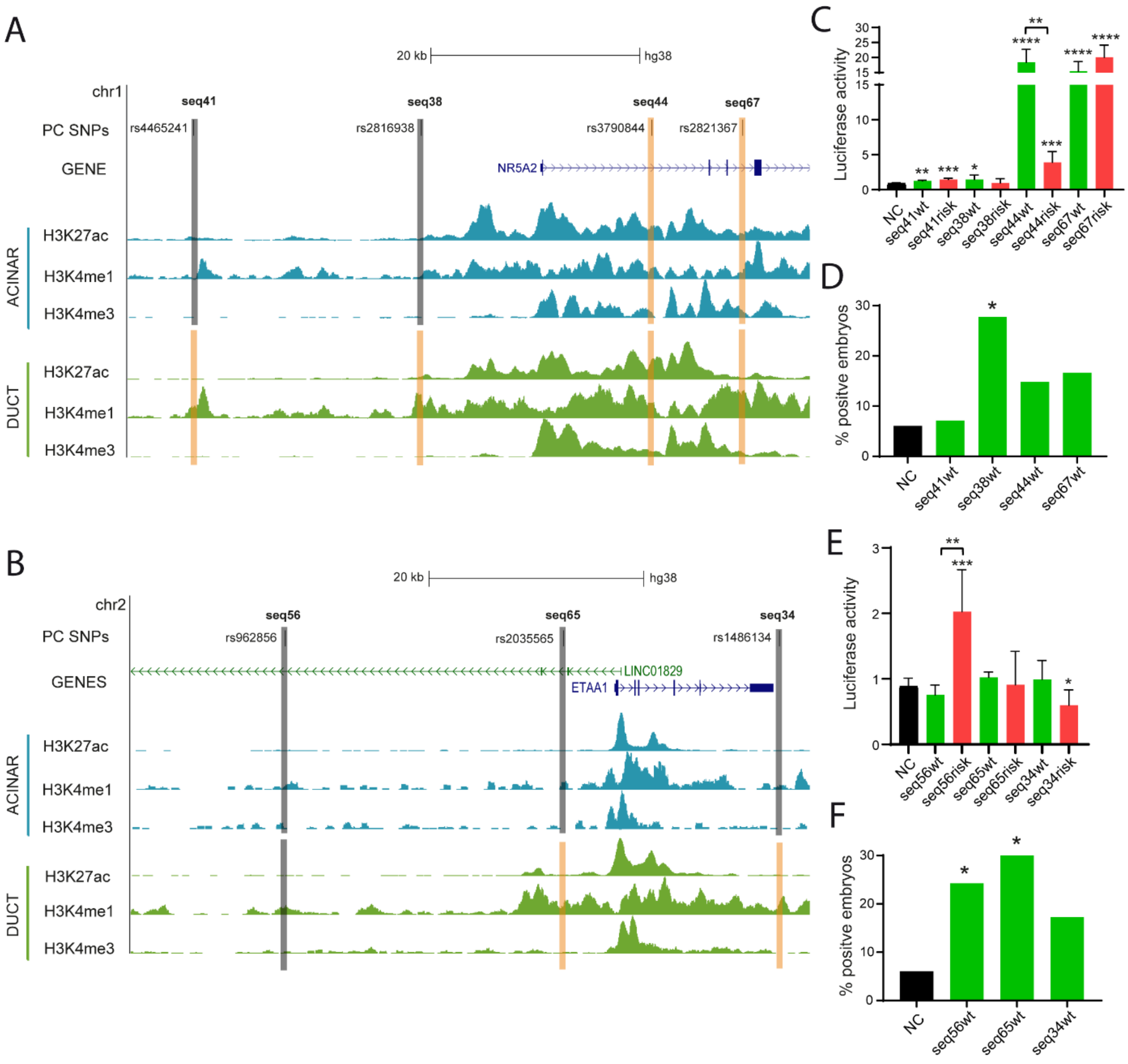


Figure 4.3 a) The *NR5A2* regulatory landscape in pancreatic acinar and duct cell types containing several putative cis-regulatory regions (in orange are labelled the putative enhancer regions and in grey are labelled the repressed chromatin regions). The genomic landscape shows the PC risk SNPs present and the epigenetic profile of H3K27ac (active enhancers), H3K4me1(active enhancers), H3K4me3 (active promoters) ChIP-seq data from pancreatic acinar and duct cells **b)** The *ETAA1*, *LINC01829* and *MEIS1* regulatory landscape in pancreatic acinar and duct cell types containing several putative cis-regulatory regions (in orange are labelled the putative enhancer regions and in grey are labelled the repressed chromatin regions). The genomic landscape shows the PC risk SNPs present and the profile of H3K27ac (active enhancers), H3K4me1(active enhancers), H3K4me3 (active promoters) ChIP-seq from pancreatic acinar and duct cells **c)** Luciferase enhancer reporter assays performed in human hTERT-HPNE cells for seq41wt, seq38wt, seq44wt and seq67wt showing luc2/Nluc ratios, relative to the negative control (NC; two-sided t-test; *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001) **d)** Percentage of F0 zebrafish larvae with GFP expression in the exocrine pancreas following in vivo transient transgenesis reporter assays for seq41wt, seq41risk, seq38wt, seq38risk, seq44wt, seq44risk, seq67wt and seq67risk. The empty enhancer reporter vector was used as the

negative control (NC). Values are represented as percentages and compared by two-sided Chi-square with Yates' correction test. p-values<0.05 were considered significant (*****p*<0.0001, **p*<0.05). The exact *p*-value and n are discriminated in Table S11. **e)** Luciferase enhancer reporter assays performed in human hTERT-HPNE cells for seq56wt, seq65wt, and seq34wt showing luc2/Nluc ratios, relative to the negative control (NC; two-sided t-test; **p*<0.05, ***p*<0.01, ****p*<0.001, *****p*<0.0001). **f)** Percentage of F0 zebrafish larvae with GFP expression in the exocrine pancreas following in vivo transient transgenesis reporter assays for seq56wt, seq65wt and seq34wt. The empty enhancer reporter vector was used as the negative control (NC). Values are represented as percentages and compared by two-sided Chi-square with Yates' correction test. p-values<0.05 were considered significant (*****p*<0.0001, **p*<0.05). The exact *p*-value and n are discriminated in Table S11.

4.2.4 Pancreatic risk alleles modulate enhancer activity in vivo

We also assessed the enhancer activity of the in vitro tested sequences in an in vivo model, the zebrafish. The respective sequences, containing the non-risk allele, were cloned in Z48 transgenesis vector, and in vivo enhancer assays were performed by mosaic transgenesis in zebrafish embryos. Then, we searched for colocalization of GFP cells with anti-Alcam, an exocrine marker and counted the number embryos where this colocalization is present (Fig.4.4 and Table S11). Out of the 4 tested sequences containing the non-risk allele in the landscape of *NR5A2* gene, 1 showed a consistent expression of GFP in the exocrine pancreatic domain, therefore being an exocrine pancreatic enhancer [seq38wt (n=36); *p*<0.05; Fig.4.3d]. On the other hand, out of the 3 tested sequences present in the landscape of *MEIS1*, *LINC01829* and *ETAA1* genes, 2 showed a consistent expression of GFP in the exocrine pancreatic domain, therefore being exocrine pancreatic enhancers [seq56wt (n=33) and seq65wt (n=30); *p*<0.05; Fig.4.3f]. Although the remaining 4 tested sequences did not show a statistical significance regarding their enhancer activity, all of them show higher value in comparison to the control.

Comparing the results obtained by in vitro and in vivo assays, we can observe differences in the number of sequences that showed enhancer activity (in vitro: 66.7% vs in vivo: 33.3%). These differences could be explained by the cell type that we are analysing in each experiment. In the case of in vitro, we performed luciferase assays in a duct cell line, however, in the in vivo assay, we are analysing whole exocrine domain, that is composed by several types of cells.

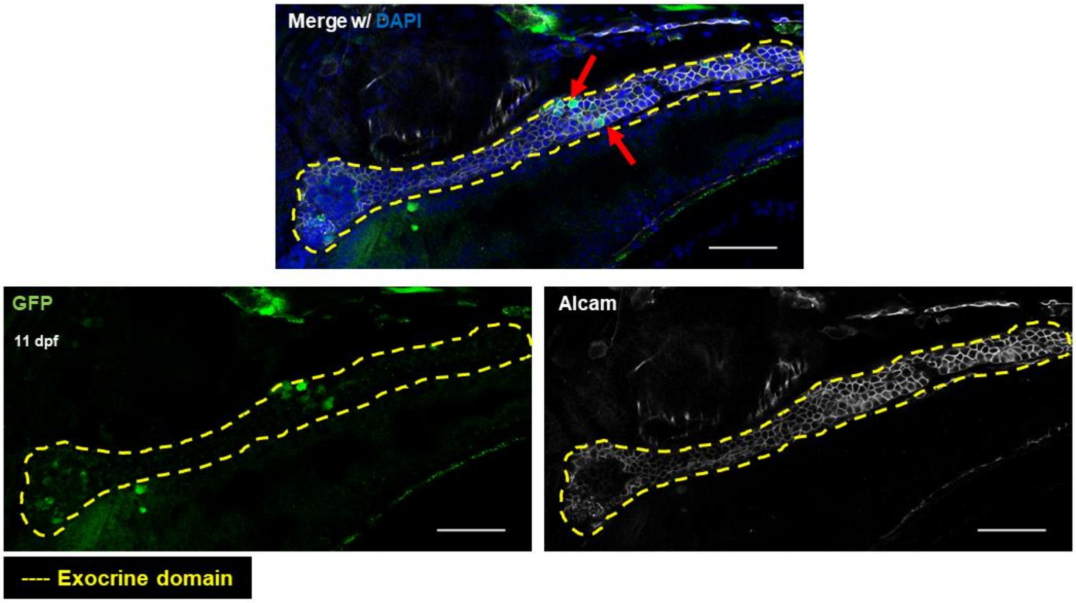


Figure 4.4 In vivo reporter assay for exocrine pancreatic enhancers. Representative confocal image of a 11dpf zebrafish pancreas injected with the Z48 enhancer reporter vector containing the seq38wt sequence, showing GFP-positive cells (green, red arrows) within the exocrine pancreatic domain (yellow dashed line), labelled by anti-Alcam staining (white). Nuclei were stained with DAPI (blue). Scale bars: 5 um.

4.3 Conclusion and future perspectives

In the present study, combining a set of PC risk alleles with different human pancreatic chromatin states information, we found a significant enrichment of PC risk alleles in putative enhancer regions in pancreatic duct and acinar cell types. These findings suggest that alterations in these sequences can dysregulate its cis-regulatory output and consequently the proper transcription of target genes. Analyzing the genomic landscape of *NR5A2*, already linked to PC (Lin et al., 2014; Flandez et al., 2014), we selected a set of duct and acinar putative cis-regulatory sequences that overlap with PC risk SNPs and we validated these sequences as enhancers. Additionally, we performed this assay using these sequences containing the wt and the risk alelle, in order to investigate the impact of the PC risk SNP in the activity of these sequences. We found that some of these regions are pancreatic duct enhancers. We found a particular case, seq44, where the presence of risk allele was able to significantly alter the enhancer activity, suggesting that it could decrease the *NR5A2* expression. As previously said, the precise role of *NR5A2* in PC is still unclear (Guo et al., 2021; Flandez et al., 2014). However, several studies have been described that heterozygous mutations in the *Nr5a2* gene makes the pancreas more susceptible to damage, and in cooperation with other mutations, can lead to pancreatic tumorigenesis (Flandez et al., 2014). Based on this, we can speculate that alterations in this particular sequence could affect the transcription of this gene, making the pancreas more prone to damage, triggering the development of PC. Additionally, we also observed that in some cases the presence of the risk did not show an impact in enhancer activity. We can explain that by the redundancy in the consensus sequence where the TFs bind (Khan et al., 2018). A single TF is able to recognize a multitude of similar DNA sequences, which are usually called as binding site motifs using models such as position weight matrices. Thus, when a change of a nucleotide occurs in a sequence, the recognition site of a particular TF might not be changed, allowing the bind of the same TF and consequently maintaining the enhancer activity.

In this study, we applied a software in order to identify the gene that was nearby our putative enhancer regions. However, it is well described that enhancers do not necessarily play a role in the nearest promoter region but can circumvent neighbouring genes to control genes placed more distantly (Arnold et al., 2019; Laverré et al., 2022). So, in future experiments, it will be important to also identify the distant target gene of these enhancer sequences. Appling chromosome conformation capture and its derivative methods, such as 4C or HI-C (Belton et al., 2012; Dekker et al., 2002) to pancreatic tissues, it would be straightforward to identify the putative target genes that are interacting with these enhancer sequences. In the particular case of the seq44, it will be externally relevant to clearly understand if this enhancer sequence is interacting with the promoter region of *NR5A2* gene and if this sequence is also interacting with other important pancreatic target genes. To clarify this point, we can perform 4C assay in duct cells, using as viewpoint the seq44. Then, it will be also important to determine if this enhancer is indeed controlling the transcription of *NR5A2* gene. To address this, it would be possible to delete the enhancer region through CRISPR-Cas9 system, as we previously described for *ARID1A* gene in chapter II, and then evaluate the *NR5A2* expression levels. Finally, it will be pertinent to evaluate the contribution of this gene dysregulation in the development of PC. We can address this issue, performing in vitro assays in the seq44 deletion genetic background, including cell proliferation, colony formation and spheroid formation assays (Kim et al., 2021; Roe et al., 2017; Somerville et al., 2018). In addition, this problem can also be assessed in vivo using zebrafish as a model. But first, we need to identify if the zebrafish regulatory landscape of *nr5a2* gene has a functional equivalent pancreatic enhancer for the seq44 identified in human genome. Then, we can delete the enhancer region through CRISPR-Cas9 system, as we previously described for *ptf1a* gene in chapter II and generate stable transgenic lines. With the establishment of this fish line, we can evaluate in vivo if the deregulation of *nr5a2* gene contribute for the formation of pancreatic tumours.

4.4 Material and Methods

4.4.1 Experimental procedures

4.4.1.1 Cell culture

hTERT-HPNE (ATCC CRL-4023) cells were cultured in a 5% CO2-humidified chamber at 37ºC in DMEM (1x, 4.5 g/L D-glucose with pyruvate; #D6429, Gibco, ThermoFisher Scientific), supplemented with 10% fetal bovine serum (#BCS0615, biotecnomica), 10ng/mL human recombinant EGF (#11343406, Immunotools) and 750ng/mL puromycin (#P8833-25MG, Sigma-Aldrich) in TC Dish 100 (SARSTEDT). When cells reached 90% of confluence, they were split using TrypLE Express (#12604-021, Gibco, ThermoFisher Scientific; approximately 0.5 mL per 10 cm2).

4.4.1.2 Luciferase reporter assay

The selected enhancer sequences were cloned in the pGL4.23GW[luc2/minP] vector (Addgene #603232) and co-transfected along with pNL1.1PGK[Nluc/PGK] (Promega #N1441) in hTERT-HPNE cells using Lipofectamine 3000 (Thermo Fisher), following manufacturer's instructions. The promoter of tyrosine kinase was cloned into the pGL4.23GW[luc2/minP] vector and used as positive control (pGL4.23GW[luc2/Tkp]; Vaz et al., 2021). As negative control, a region without marks of enhancer activity (H3K27ac) was cloned into the pGL4.23GW [luc2/minP] vector. The luciferase activity was measured 48 hours post transfection with the Nano-Glo Luciferase Assay System (Promega, #N1610) on a Synergy 2 microplate reader (BioTek). Results were presented as luc2/Nluc ratios, relative to the negative control. Two-sided t-test was used to calculate statistical significance. At least three independent replicates of the transfection were performed.

4.4.1.3 Immunohistochemistry in zebrafish

Zebrafish larvae with 48hpf were euthanized by prolonged immersion in 200-300 mg/L tricaine (MS222; ethyl-3-aminobenzoate methanesulfonate, #E10521-10G, Sigma-Aldrich). Whenever necessary the chorion was removed, and the zebrafish were fixed in formaldehyde 4% (#F1635-500ML, Sigma-Aldrich) for 1h at RT (8-12dpf larvae). Permeabilization was carried out by incubation with 1% Triton X-100 in PBS for 1h at room temperature (RT) followed by blocking with 5% bovine serum albumin (BSA) in 0.1% Triton X-100 for 1h at RT. Zebrafish were incubated with the primary antibody diluted in blocking solution (5% bovine serum albumin

(BSA) in 0.1% Triton X-100) at 4ºC overnight (O.N) and then incubated with the secondary antibody plus DAPI (1:1000, D1306 Invitrogen, ThermoFisher Scientific) diluted in blocking solution for 4 hours at RT. After each antibody incubation, embryos were washed 6 times in PBS-T (0.5 % Triton X-100 in PBS-1x) 5 minutes at RT. Embryos were stored in 50% Glycerol/PBS at 4ºC before microscopy slides preparation in the mounting medium (50% Glycerol/PBS). Images were acquired with a Leica TCS SP5 II confocal microscope (Leica Microsystems, Germany; LAS AF software (v.2.6.3.8173) and processed by ImageJ software (v.1.8.0). Primary antibodies: mouse anti-Alcam (1:50, #ZN-8, DSHB) and mouse anti-Nkx6.1 (1:50, #F55A10, DSHB). Secondary antibodies: goat anti-mouse AlexaFluor647 (1:800, #A-21236 Invitrogen, ThermoFisher Scientific).

4.4.1.4 Zebrafish husbandry and breeding and embryo rearing

Zebrafish (*Danio rerio*) were handled according to European animal welfare regulations and standard protocols. Embryos were grown at 28ºC in E3 medium or E3 supplemented with 0.01% PTU (Karlsson et al., 2001).

4.4.1.5 In vivo mosaic transgenesis assays

Sequences were amplified by PCR from human genomic DNA using the primers in Table S12 (designed to span the ChIP-seq signals; Sigma-Aldrich), with the proof-reading iMax TM II DNA polymerase (INtRON Biotechnology) following the manufacturer's instructions. PCR products were visualized by electrophoresis on an 1% agarose gel, the bands excised, purified with NZYGelpure kit (NZYTech) and cloned into the entry vector pCR®8/GW/TOPO (#250020 Invitrogen, ThermoFisher Scientific) according to manufacturer's instructions. All the sequences were confirmed by sanger sequencing. The vectors were then recombined into the destination vectors Z48 (de la Calle-Mustienes et al., 2005) using Gateway® LR Clonase® II Enzyme mix (Invitrogen, ThermoFisher Scientific), following manufacturer's instructions.

Standard chemical transformation was performed with MultiShotTM FlexPLate Mach1TM T1R (Invitrogen, ThermoFisher Scientific), grown O.N. at 37ºC. Vector selection was performed with 100 μg/ml Spectinomycin (Sigma-Aldrich) in the growth medium for the pCR®8/GW/TOPO vectors, or 100 μg/ml Ampicillin (Normon) for the Z48 vector. Plasmids were purified with NZYMiniprep kit (NZYTech) and confirmed by Sanger sequencing using the primers in Table S12. Final plasmids were purified with phenol/chloroform and concentration was determined by NanoDrop 1000 Spectrophotometer (ThermoFisher Scientific).

Zebrafish transgenesis was performed using the tol2 transposon system (Kawakami et al., 2000). Tol2 cDNA was transcribed by Sp6 RNA polymerase (ThermoFisher Scientific) after Tol2-pCS2FA vector linearization with NotI restriction enzyme (Anza, Invitrogen, ThermoFisher Scientific). Tol2 mRNA was purified as previously described (Bessa et al., 2009). One-cell stage embryos were injected with 1nL solution containing 25ng/µL of transposase mRNA, 25ng/µL of phenol/chloroform purified plasmid, and 0.05% phenol red. Injections were performed at least three times.

Risk SNPs from seq41, seq38, seq44, seq67, seq56, seq65 and seq34 were inserted by site-directed mutagenesis using specific primers containing the risk allele (Table S12). All the sequences were confirmed by sanger sequencing. Injected embryos showing expression of GFP in the midbrain were selected for immunohistochemistry at 48 hours post fertilization (hpf) and maintained in 28ºC in E3 medium with PTU until 8-12 dpf because in this timepoint, the pancreas is already fully developed.

4.4.1.6 Assessment of enhancer activity

Embryos were analysed, using confocal microscopy, for the presence of GFP-positive cells in the exocrine pancreatic domain (anti-Alcam). One embryo was considered positive if at least one GFP-positive cell was detected within the exocrine pancreatic domain. Quantifications are presented as percentages of positive embryos to ensure the quantification of different transposon integrations. In the table S11 is described the number of larvae injected in each condition.

4.4.1.7 Statistical analysis

Two-sided t-test was used to calculate statistical significance in in vitro assays in order to compare the luciferase activity between wt, risk and control sequences. At least three independent replicates of the transfection were performed in in vitro assays. Two-sided chi-square test with Yates's correction was applied to the in vivo validation in order to compare the number of embryos that showed GFP expression. To calculate the statistical significance of the enrichments observed between the different regulatory categories in the several cell types, Two-sided chi-square test with Yates's correction was applied. In all analyses, p<0.05 was required for statistical significance and calculated in GraphPad Prism 5 (San Diego, CA, USA).

4.4.2 Bioinformatic analysis

4.4.2.1 Selection and design of the SNPs datasets

The human pancreatic cancer risk alleles dataset (designated in the graphs as “PC SNPs”) was created based on risk alleles associated to pancreatic cancer available the DisGeNET database (Piñero et al., 2020). The selection of SNPs was performed based on the described disease (adenocarcinoma of pancreas, malignant neoplasm of pancreas, pancreatic cancer, pancreatic carcinoma, pancreatic ductal adenocarcinoma and pancreatic neoplasm) and based on DisGeNET score for variant-disease associations (>=0.7). The score is calculated based in the number of curated sources (UniProt, ClinVar, the GWAS Catalog, and GWASdb) where the variation was described (Piñero et al., 2020).

Then, we applied the annotatePeaks.pl module of HOMER (v.4.11.1;(Heinz et al., 2010)) to identify the genomic distribution of these SNPs and we classified them as “coding” and “non-coding” SNPs. The SNPs classified as “coding” were deleted from the dataset, since we are focused on the study of non-coding regions. To create human non-coding risk SNPs associated with several cancer types, but not related with PC dataset (designated in the graphs as “no PC SNPs”), we used the same database and parameters, but we just selected risk alleles associated to any type of cancer, excluding the risk alleles present in the “PC SNPs” dataset. We also selected the SNPs that are in non-coding regions, using the method previously described. Control SNPs dataset was created based on the SNPs described in The International Genome Sample Resource (IGSR) and the 1000 Genomes Project browser (https://www.internationalgenome.org; Fev.2020; (1000 Genomes Project Consortium et al., 2015). Basically, this dataset contains all the variants described until February 2020 in this database (15175044 SNPs). We also selected the SNPs that are in non-coding regions, using the method previously described.

4.4.2.2 Selection and creation of cis-regulatory category based on ChIP-seq datasets

The human pancreatic ChromHMM datasets (duct, acinar and endocrine cell types) used in this study were the same from the study of Arda and colleagues (2018; (Arda et al., 2018) and are accessible on Gene Expression Omnibus repository with accession number GSE79468. The ChromHMM software integrate multiple Chip-seq datasets from different histone marks (H3K27me3, H3K4me3, H3K27ac and H3K4me1), generating genome-wide maps of chromatin state annotations for each cell type. In the original publication, ten different chromatin states were identified. However, to simplify our analysis, we regrouped the chromatin stages, creating four different categories. 1) Enhancer category: includes the weak and active enhancer category; 2) Promoter category: includes active transcription start site (TSS), bivalent TSS and flanking TSS categories; 3) Repressed chromatin category: includes repressed chromatin category and 4) no signal category: includes low or no signal categories.

The information of chromatin state annotations used in control tissues (colon and right ventricle tissues) was from the Roadmap epigenomics project (http://www.roadmapepigenomics.org/; E076 – Colon smooth muscle and E105 – Right ventricle). In these datasets, we also regrouped the chromatin stages, creating the four categories previously described.

4.4.2.3 The overlapping between SNPs and human cis-regulatory elements

To calculate the overlap of variants (risk and control SNPs) with human chromatin states categories from the different cell types/tissues (pancreatic and control tissues), we intercepted each set of SNPs with human chromatin states categories for the different tissues. These interceptions were performed using Bedtools “intersect” (v.2.27; (Quinlan and Hall, 2010). Statistical significance was determined by two-sided chi-square test with Yates correction. The *p*-values<0.05 were considered significant.

4.5 Acknowledgements

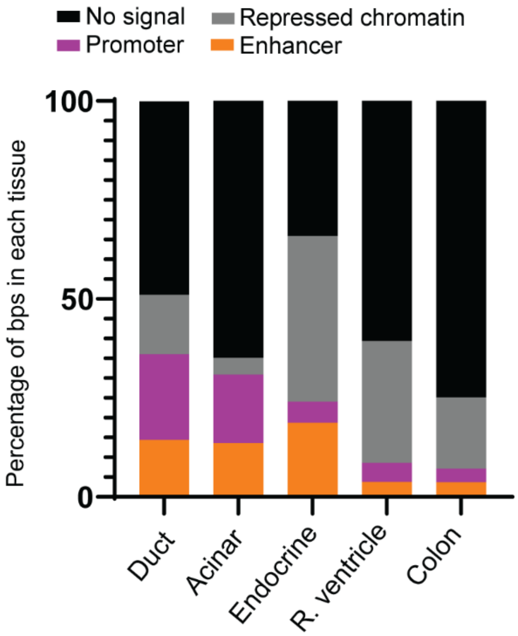
We thank to the i3S Scientific Platform: Advanced Light Microscopy, member of PPBI (POCI-01-0145-FEDER-022122); Translational Cytometry; Cell Culture and Genotyping and i3s HPC facility. We also thank to Mafalda Sousa for the automated cell analysis and Isabel Guedes for maintenance of the zebrafish lines.

4.6 Funding

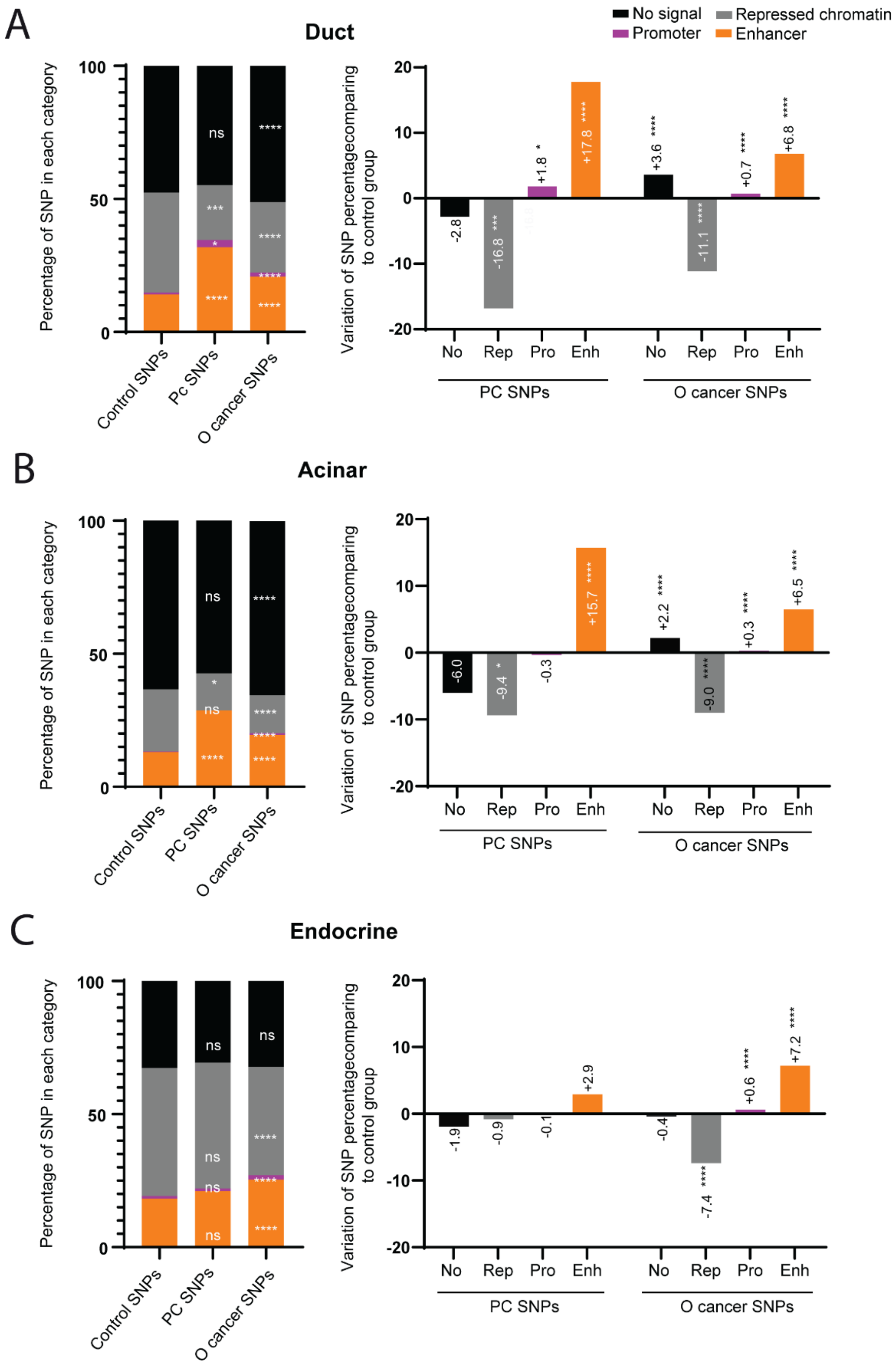
This study was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-2015-StG-680156-ZPR). JB is supported by an FCT CEEC grant (CEECIND/03482/2018). JTx (SFRH/BD/126467/2016) is PhD fellow from FCT. MG was supported by the EnvMetaGen project via the European Union’s Horizon 2020 research and innovation programme (grant 668981). This work was funded by National Funds through FCT—Fundação para a Ciência e a Tecnologia, I.P., under the project UIDB/04293/2020.

4.7 Supplementary information

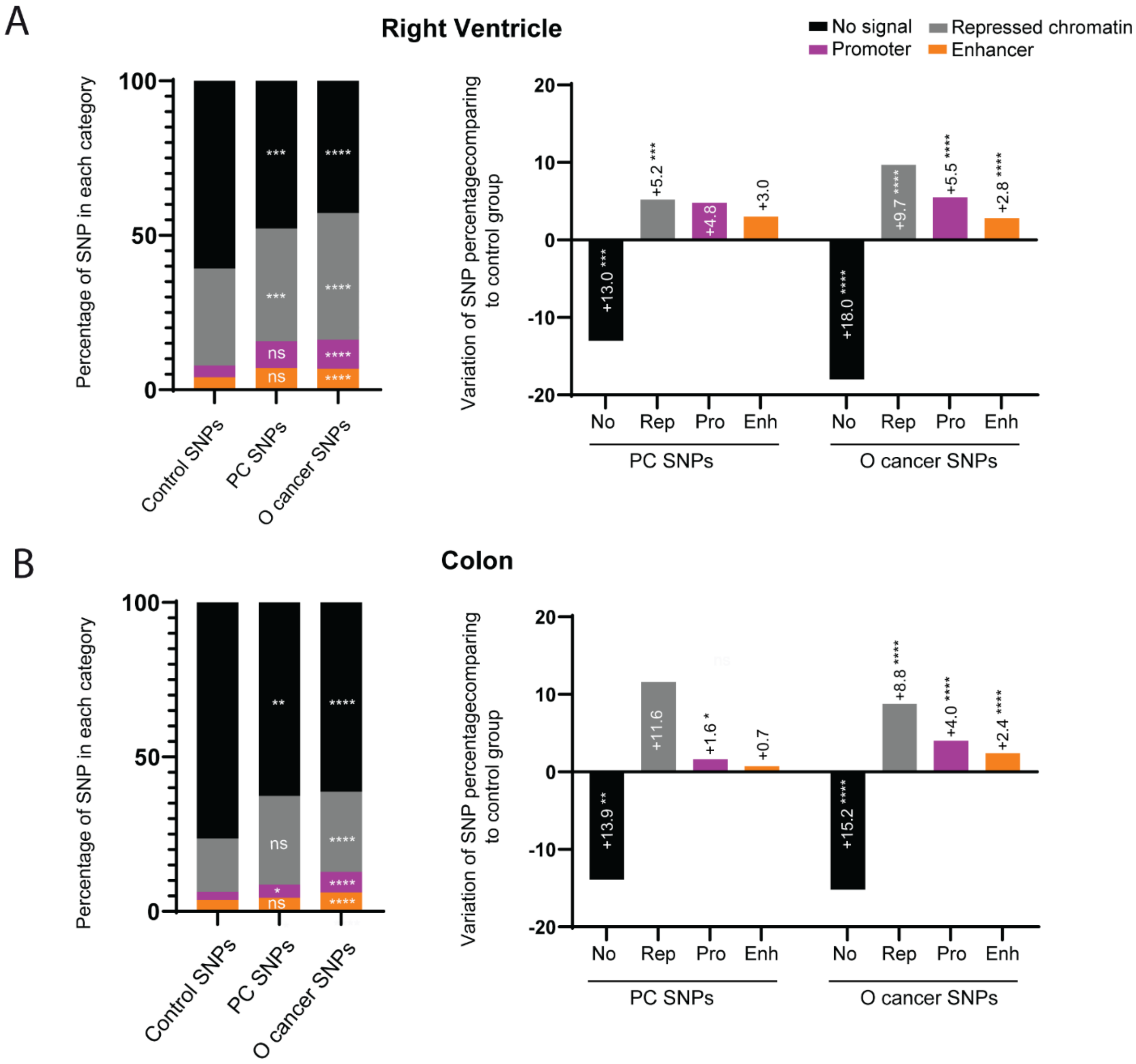
4.7.1 Supplementary figures



Supplementary Figure 4.1. Percentage of base pairs present in human pancreatic chromatin state category in each cell type/tissue. Pancreatic cell types/tissues: duct, acinar and endocrine. Control tissues: Right ventricle and colon.



Supplementary Figure 4.2. Human pancreatic chromatin states information (duct, acinar and endocrine cell types) and the respective overlapping of SNPs. **a)** The percentage of variants that overlap with chromatin stage categories in duct cells and the respective variation between groups. **b)** The percentage of variants that overlap with chromatin stage categories in acinar cells and the respective variation between groups. **c)** The percentage of variants that overlap with chromatin stage categories in endocrine cells and the respective variation between groups. The random selection of common variants is labelled as “Control SNPs”, the group SNPs associated to pancreatic cancer is labelled as “PC SNPs” and the set of SNPs associated to several other types of cancer, excluding PC is labelled as “O cancer SNPs”. Chromatin categories: No – No signal; Rep – Repressed chromatin; Pro – Promoter; Enh – Enhancer.



Supplementary Figure 4.3. Human chromatin states information (right ventricle and colon tissues) and the respective overlapping of SNPs. **a)** The percentage of variants that overlap with chromatin stage categories in Right ventricle and the respective variation between groups. **b)** The percentage of variants that overlap with chromatin stage categories in Colon and the respective variation between groups. The random selection of common SNPs is labelled as “Control SNPs”, the set of variants associated to several other types of cancer, excluding PC is labelled as “O cancer SNPs”. Chromatin categories: No – No signal; Rep – Repressed chromatin; Pro – Promoter; Enh – Enhancer.

4.7.2 Supplementary tables

Table S1. List of non-coding variants associated to pancreatic cancer and the respective coordinates.

Chromosome	Coordinates (hg38)	Ref_SNP	Disease
chr7	40827064	rs17688601	Pancreatic carcinoma
chr10	118519432	rs12413624	Pancreatic carcinoma
chr5	1319565	rs451360	Pancreatic carcinoma
chr21	29305751	rs1153280	Malignant neoplasm of pancreas
chr12	32283475	rs708224	Malignant neoplasm of pancreas
chr11	9907995	rs12362504	Malignant neoplasm of pancreas
chr3	189790682	rs9854771	Pancreatic carcinoma
chr6	160403722	rs2504938	Pancreatic carcinoma
chr1	199936700	rs12029406	Malignant neoplasm of pancreas
chr8	38611785	rs7832232	Pancreatic carcinoma
chr4	79043433	rs1455311	Malignant neoplasm of pancreas
chr7	47448971	rs73328514	Pancreatic carcinoma
chr17	18850557	rs4924935	Pancreatic carcinoma
chr2	234706553	rs6736997	Pancreatic carcinoma
chr11	96947703	rs1944788	Pancreatic carcinoma
chr3	3314490	rs9874556	Pancreatic carcinoma
chr5	1286401	rs2736100	Pancreatic carcinoma
chr2	67392524	rs2035565	Malignant neoplasm of pancreas
chr2	101305708	rs6711606	Pancreatic carcinoma
chr6	155876368	rs4269383	Malignant neoplasm of pancreas
chr11	9951515	rs10500715	Malignant neoplasm of pancreas
chr2	67366671	rs962856	Pancreatic carcinoma
chr9	133273813	rs505922	Pancreatic carcinoma
chr13	65907683	rs1585440	Pancreatic carcinoma
chr9	4426631	rs10974531	Pancreatic carcinoma
chr13	27902841	rs9554197	Pancreatic carcinoma
chr8	128555832	rs1561927	Pancreatic carcinoma
chr1	200041696	rs3790843	Malignant neoplasm of pancreas
chr6	161815043	rs3016539	Pancreatic carcinoma
chr15	36363821	rs4459505	Pancreatic carcinoma
chr7	18798993	rs12531908	Pancreatic carcinoma
chr15	36359637	rs4130461	Pancreatic carcinoma
chr21	29313290	rs1153287	Malignant neoplasm of pancreas
chr1	64073289	rs1747924	Pancreatic carcinoma
chr1	18351676	rs16861827	Pancreatic carcinoma
chr2	104762499	rs12615966	Pancreatic carcinoma
chr5	39394887	rs2255280	Malignant neoplasm of pancreas

Continuation of the previous table			
chr5	1248932	rs4583925	Pancreatic carcinoma
chr5	1287079	rs2853677	Pancreatic carcinoma
chr9	133263862	rs657152	Pancreatic carcinoma
chr3	196024759	rs4927850	Malignant neoplasm of pancreas
chr5	1344343	rs31490	Pancreatic carcinoma
chr12	27583053	rs1975920	Malignant neoplasm of pancreas
chr22	28904318	rs16986825	Pancreatic carcinoma
chr21	29328775	rs1153294	Malignant neoplasm of pancreas
chr2	71459496	rs112493246	Pancreatic carcinoma
chr6	160412632	rs9364554	Pancreatic carcinoma
chr6	68432116	rs9363918	Malignant neoplasm of pancreas
chr1	112503773	rs351365	Pancreatic carcinoma
chr11	125644678	rs521102	Pancreatic carcinoma
chr3	13029299	rs361052	Pancreatic carcinoma
chr13	73322084	rs9564966	Malignant neoplasm of pancreas
chr5	1299098	rs2735948	Pancreatic carcinoma
chr1	200038304	rs3790844	Pancreatic carcinoma
chr11	18363391	rs9783347	Pancreatic carcinoma
chr7	155813978	rs288746	Pancreatic carcinoma
chr8	75558169	rs2941471	Pancreatic carcinoma
chr21	29354452	rs2027605	Malignant neoplasm of pancreas
chr2	152798206	rs12478462	Pancreatic carcinoma
chr17	72405335	rs7214041	Pancreatic carcinoma
chr6	160413796	rs2457571	Pancreatic carcinoma
chr18	13357201	rs981621	Pancreatic carcinoma
chr5	2109787	rs6879627	Pancreatic carcinoma
chr9	104125300	rs2417487	Pancreatic carcinoma
chr9	133279294	rs495828	Pancreatic carcinoma
chr13	73357977	rs1886449	Pancreatic carcinoma
chr1	199994494	rs4465241	Pancreatic carcinoma
chr6	170019278	rs2172905	Malignant neoplasm of pancreas
chr21	29348513	rs117214	Malignant neoplasm of pancreas
chr17	72404025	rs11655237	Pancreatic carcinoma
chr2	136797654	rs1427593	Pancreatic carcinoma
chr15	36362396	rs8028529	Pancreatic carcinoma
chr7	130995762	rs6971499	Pancreatic carcinoma
chr2	133680388	rs1901440	Pancreatic carcinoma
chr1	200016240	rs2816938	Pancreatic carcinoma
chr21	42358786	rs1547374	Malignant neoplasm of pancreas
chr8	124864572	rs7015626	Malignant neoplasm of pancreas

Continuation of the previous table			
chr8	123753462	rs10088262	Pancreatic carcinoma
chr1	236276616	rs6662005	Pancreatic carcinoma
chr5	7893008	rs162049	Pancreatic carcinoma
chr21	45499218	rs2236479	Malignant neoplasm of pancreas
chr11	9956424	rs7106914	Malignant neoplasm of pancreas
chr5	1321972	rs401681	Pancreatic carcinoma
chr7	153928758	rs6464375	Malignant neoplasm of pancreas
chr2	71461486	rs138529893	Pancreatic carcinoma
chr13	79725587	rs2039553	Pancreatic carcinoma
chr7	153941163	rs6973850	Pancreatic carcinoma
chr17	32550640	rs225190	Pancreatic carcinoma
chr6	1339954	rs9502893	Malignant neoplasm of pancreas
chr20	44458008	rs6073450	Pancreatic carcinoma
chr1	200047018	rs2821367	Pancreatic carcinoma
chr21	46120286	rs4458293	Pancreatic carcinoma
chr18	59211042	rs1517037	Pancreatic carcinoma
chr21	29356542	rs2832290	Malignant neoplasm of pancreas
chr7	153925577	rs7779540	Pancreatic carcinoma
chr9	117306874	rs10983614	Pancreatic carcinoma
chr16	23629276	rs587776417	Pancreatic carcinoma
chr9	133261737	rs2073828	Pancreatic carcinoma
chr1	238745053	rs2689154	Pancreatic carcinoma
chr17	6238357	rs7503953	Malignant neoplasm of pancreas
chr13	73334709	rs9573163	Malignant neoplasm of pancreas
chr9	133261703	rs687289	Pancreatic carcinoma
chr10	85980996	rs10788473	Pancreatic carcinoma
chr12	120987058	rs7310409	Pancreatic carcinoma
chr17	37718512	rs4795218	Pancreatic carcinoma
chr22	48533757	rs5768709	Malignant neoplasm of pancreas
chr3	74669607	rs1447826	Malignant neoplasm of pancreas
chr5	124688588	rs4285214	Pancreatic carcinoma
chr20	2674279	rs1810636	Malignant neoplasm of pancreas
chr18	13366863	rs12456874	Malignant neoplasm of pancreas
chr19	29164379	rs2903018	Malignant neoplasm of pancreas
chr8	127707639	rs10094872	Pancreatic carcinoma
chr1	199996040	rs10919791	Pancreatic carcinoma
chr11	96970814	rs17275283	Pancreatic carcinoma
chr13	73342491	rs9543325	Pancreatic carcinoma

Table S2. List of human chromatin states information in pancreatic duct cells that overlap with PC risk variants (Categories: No – No signal; Rep – Repressed chromatin; Pro – Promoter; Enh - Enhancer).

Duct regions				**PC SNPs**		
Chromosome	**Coordinates (hg38)**		**Chromatin category**	**Chromosome**	**Coordinates**	**Ref_SNP**
chr11	9907653	9922653	No	chr11	9907995	rs12362504
chr11	9947053	9952053	No	chr11	9951515	rs10500715
chr11	9952653	9972853	No	chr11	9956424	rs7106914
chr11	18332453	18382053	No	chr11	18363391	rs9783347
chr11	96887600	96949000	No	chr11	96947703	rs1944788
chr11	96966800	96980600	No	chr11	96970814	rs17275283
chr11	125627905	125660705	No	chr11	125644678	rs521102
chr12	32274666	32287466	No	chr12	32283475	rs708224
chr12	120986997	120999797	No	chr12	120987058	rs7310409
chr15	36359799	36362599	No	chr15	36362396	rs8028529
chr15	36363599	36365799	No	chr15	36363821	rs4459505
chr16	23597079	23638679	No	chr16	23629276	rs587776417
chr17	18722087	18855287	No	chr17	18850557	rs4924935
chr17	32520782	32580382	No	chr17	32550640	rs225190
chr18	13355401	13362401	No	chr18	13357201	rs981621
chr18	59210368	59211568	No	chr18	59211042	rs1517037
chr1	112503178	112507378	No	chr1	112503773	rs351365
chr1	199935472	199941272	No	chr1	199936700	rs12029406
chr1	199995872	200004872	No	chr1	199996040	rs10919791
chr20	2665154	2680554	No	chr20	2674279	rs1810636
chr21	29322479	29332279	No	chr21	29328775	rs1153294
chr21	29355879	29360879	No	chr21	29356542	rs2832290
chr21	45472286	45512686	No	chr21	45499218	rs2236479
chr21	46110486	46143086	No	chr21	46120286	rs4458293
chr22	48492788	48533788	No	chr22	48533757	rs5768709
chr2	67340468	67367268	No	chr2	67366671	rs962856
chr2	71459670	71461670	No	chr2	71461486	rs138529893
chr2	133676829	133681029	No	chr2	133680388	rs1901440
chr2	234695556	234724356	No	chr2	234706553	rs6736997
chr3	196012929	196066929	No	chr3	196024759	rs4927850
chr4	79035846	79048846	No	chr4	79043433	rs1455311
chr5	1240485	1260885	No	chr5	1248932	rs4583925
chr5	1265485	1314885	No	chr5	1286401	rs2736100
chr5	1265485	1314885	No	chr5	1287079	rs2853677
chr5	1265485	1314885	No	chr5	1299098	rs2735948
chr5	1317285	1328285	No	chr5	1319565	rs451360
chr5	1317285	1328285	No	chr5	1321972	rs401681

Continuation of the previous table						
chr5	7887887	7903887	No	chr5	7893008	rs162049
chr6	1338165	1340765	No	chr6	1339954	rs9502893
chr6	155875266	155880666	No	chr6	155876368	rs4269383
chr6	161813968	161835768	No	chr6	161815043	rs3016539
chr7	18788377	18800577	No	chr7	18798993	rs12531908
chr8	38584482	38626282	No	chr8	38611785	rs7832232
chr8	75545765	75569165	No	chr8	75558169	rs2941471
chr8	123752960	123758160	No	chr8	123753462	rs10088262
chr8	127701955	127717354	No	chr8	127707639	rs10094872
chr9	4423400	4429600	No	chr9	4426631	rs10974531
chr9	104096919	104132519	No	chr9	104125300	rs2417487
chr9	133253213	133268388	No	chr9	133263862	rs657152
chr9	133253213	133268388	No	chr9	133261737	rs2073828
chr9	133253213	133268388	No	chr9	133261703	rs687289
chr9	133274584	133282027	No	chr9	133279294	rs495828
chr10	85975243	85994843	Rep	chr10	85980996	rs10788473
chr13	65884468	65908468	Rep	chr13	65907683	rs1585440
chr17	6223080	6240080	Rep	chr17	6238357	rs7503953
chr19	29164093	29167493	Rep	chr19	29164379	rs2903018
chr1	18343706	18353506	Rep	chr1	18351676	rs16861827
chr3	13028300	13039100	Rep	chr3	13029299	rs361052
chr3	74669049	74678449	Rep	chr3	74669607	rs1447826
chr6	160400768	160413368	Rep	chr6	160403722	rs2504938
chr6	160400768	160413368	Rep	chr6	160412632	rs9364554
chr7	153926915	153933515	Rep	chr7	153928758	rs6464375
chr7	153940915	153943515	Rep	chr7	153941163	rs6973850
chr7	155813106	155819906	Rep	chr7	155813978	rs288746
chr9	117297721	117335321	Rep	chr9	117306874	rs10983614
chr10	118515488	118521688	Rep	chr10	118519432	rs12413624
chr2	104759542	104765542	Rep	chr2	104762499	rs12615966
chr6	160413368	160414568	Rep	chr6	160413796	rs2457571
chr7	153924515	153926915	Rep	chr7	153925577	rs7779540
chr1	238736300	238854300	Pro	chr1	238745053	rs2689154
chr21	29353879	29355879	Pro	chr21	29354452	rs2027605
chr2	136796430	136809630	Pro	chr2	136797654	rs1427593
chr2	152782686	152858486	Pro	chr2	152798206	rs12478462
chr3	3223916	3342116	Pro	chr3	3314490	rs9874556
chr6	68303308	68450508	Pro	chr6	68432116	rs9363918
chr7	40782001	40834601	Pro	chr7	40827064	rs17688601
chr5	2109486	2110086	Pro	chr5	2109787	rs6879627

Continuation of the previous table						
chr12	27582067	27589667	Enh	chr12	27583053	rs1975920
chr15	36358199	36359799	Enh	chr15	36359637	rs4130461
chr1	199994272	199995872	Enh	chr1	199994494	rs4465241
chr1	236269500	236280300	Enh	chr1	236276616	rs6662005
chr20	44457760	44458160	Enh	chr20	44458008	rs6073450
chr21	29308479	29314279	Enh	chr21	29313290	rs1153287
chr21	29346479	29348679	Enh	chr21	29348513	rs117214
chr21	42358491	42359091	Enh	chr21	42358786	rs1547374
chr2	67390668	67393468	Enh	chr2	67392524	rs2035565
chr13	73322063	73322463	Enh	chr13	73322084	rs9564966
chr13	73333863	73337863	Enh	chr13	73334709	rs9573163
chr13	73341463	73343063	Enh	chr13	73342491	rs9543325
chr13	73357863	73359463	Enh	chr13	73357977	rs1886449
chr13	79725265	79727265	Enh	chr13	79725587	rs2039553
chr17	72404859	72407059	Enh	chr17	72405335	rs7214041
chr1	64071728	64073728	Enh	chr1	64073289	rs1747924
chr1	200016072	200017472	Enh	chr1	200016240	rs2816938
chr5	124687707	124688707	Enh	chr5	124688588	rs4285214
chr6	170016576	170020376	Enh	chr6	170019278	rs2172905
chr7	130995641	130996241	Enh	chr7	130995762	rs6971499
chr13	27902463	27903263	Enh	chr13	27902841	rs9554197
chr17	72403859	72404859	Enh	chr17	72404025	rs11655237
chr18	13366801	13368601	Enh	chr18	13366863	rs12456874
chr2	71459470	71459670	Enh	chr2	71459496	rs112493246
chr5	39393098	39395098	Enh	chr5	39394887	rs2255280
chr7	47448202	47450002	Enh	chr7	47448971	rs73328514
chr8	124863158	124867958	Enh	chr8	124864572	rs7015626
chr9	133273385	133274584	Enh	chr9	133273813	rs505922
chr17	37717202	37718802	Enh	chr17	37718512	rs4795218
chr1	200031872	200038872	Enh	chr1	200038304	rs3790844
chr1	200043272	200047672	Enh	chr1	200047018	rs2821367
chr21	29305479	29305879	Enh	chr21	29305751	rs1153280
chr22	28904012	28904612	Enh	chr22	28904318	rs16986825
chr2	101305138	101306938	Enh	chr2	101305708	rs6711606
chr5	1343685	1344685	Enh	chr5	1344343	rs31490
chr8	128554954	128556154	Enh	chr8	128555832	rs1561927
chr1	200040472	200041872	Pro	chr1	200041696	rs3790843
chr3	189790611	189790811	Pro	chr3	189790682	rs9854771

Table S3. List of human chromatin states information in pancreatic acinar cells that overlap with PC risk variants. (Categories: No – No signal; Rep – Repressed chromatin; Pro – Promoter; Enh - Enhancer).

Acinar regions				Pc SNPs		
Chromosome	**Coordinates**		**Chromatin category**	**Chromosome**	**Coordinates**	**Ref_SNP**
chr10	85971443	85994243	No	chr10	85980996	rs10788473
chr11	9907053	9950653	No	chr11	9907995	rs12362504
chr11	9951853	10005853	No	chr11	9956424	rs7106914
chr11	18336053	18382853	No	chr11	18363391	rs9783347
chr11	96944400	96950200	No	chr11	96947703	rs1944788
chr11	125627505	125684905	No	chr11	125644678	rs521102
chr12	27581467	27583067	No	chr12	27583053	rs1975920
chr12	32273466	32301666	No	chr12	32283475	rs708224
chr13	27876463	27905063	No	chr13	27902841	rs9554197
chr13	73313263	73324063	No	chr13	73322084	rs9564966
chr13	73338863	73351463	No	chr13	73342491	rs9543325
chr13	73353063	73365063	No	chr13	73357977	rs1886449
chr15	36357599	36360199	No	chr15	36359637	rs4130461
chr15	36363799	36378199	No	chr15	36363821	rs4459505
chr16	23597479	23638479	No	chr16	23629276	rs587776417
chr17	6225880	6257480	No	chr17	6238357	rs7503953
chr17	32534982	32558582	No	chr17	32550640	rs225190
chr18	13343601	13361001	No	chr18	13357201	rs981621
chr18	13365801	13369001	No	chr18	13366863	rs12456874
chr18	59210568	59213168	No	chr18	59211042	rs1517037
chr1	18344706	18353306	No	chr1	18351676	rs16861827
chr1	112502778	112507778	No	chr1	112503773	rs351365
chr1	199929672	199940672	No	chr1	199936700	rs12029406
chr1	199982272	199994672	No	chr1	199994494	rs4465241
chr1	200016072	200016672	No	chr1	200016240	rs2816938
chr20	2665354	2687154	No	chr20	2674279	rs1810636
chr20	44451160	44460760	No	chr20	44458008	rs6073450
chr21	29356279	29359279	No	chr21	29356542	rs2832290
chr21	45495886	45512686	No	chr21	45499218	rs2236479
chr21	46109286	46133286	No	chr21	46120286	rs4458293
chr22	48485588	48565188	No	chr22	48533757	rs5768709
chr2	67326268	67366868	No	chr2	67366671	rs962856
chr2	67389468	67394668	No	chr2	67392524	rs2035565
chr2	71455270	71463870	No	chr2	71459496	rs112493246
chr2	71455270	71463870	No	chr2	71461486	rs138529893
chr2	234695156	234782556	No	chr2	234706553	rs6736997
chr3	3307316	3319716	No	chr3	3314490	rs9874556

Continuation of the previous table						
chr3	13024300	13056100	No	chr3	13029299	rs361052
chr3	195915529	196045729	No	chr3	196024759	rs4927850
chr4	79028646	79045846	No	chr4	79043433	rs1455311
chr5	1228485	1257285	No	chr5	1248932	rs4583925
chr5	1258685	1314885	No	chr5	1286401	rs2736100
chr5	1258685	1314885	No	chr5	1287079	rs2853677
chr5	1258685	1314885	No	chr5	1299098	rs2735948
chr5	1318285	1330685	No	chr5	1319565	rs451360
chr5	1318285	1330685	No	chr5	1321972	rs401681
chr5	7882887	7906287	No	chr5	7893008	rs162049
chr5	39391898	39395298	No	chr5	39394887	rs2255280
chr6	1339165	1342565	No	chr6	1339954	rs9502893
chr6	155872066	155879666	No	chr6	155876368	rs4269383
chr6	160400968	160413768	No	chr6	160403722	rs2504938
chr6	160400968	160413768	No	chr6	160412632	rs9364554
chr6	170018776	170026176	No	chr6	170019278	rs2172905
chr7	40818001	40836401	No	chr7	40827064	rs17688601
chr7	153919515	153929915	No	chr7	153925577	rs7779540
chr7	153919515	153929915	No	chr7	153928758	rs6464375
chr7	155811306	155850506	No	chr7	155813978	rs288746
chr8	75556565	75576165	No	chr8	75558169	rs2941471
chr8	123744760	123759360	No	chr8	123753462	rs10088262
chr8	127684555	127708155	No	chr8	127707639	rs10094872
chr9	4407200	4438600	No	chr9	4426631	rs10974531
chr9	104095919	104154919	No	chr9	104125300	rs2417487
chr9	133261397	133264197	No	chr9	133263862	rs657152
chr9	133261397	133264197	No	chr9	133261737	rs2073828
chr9	133261397	133264197	No	chr9	133261703	rs687289
chr9	133276748	133288028	No	chr9	133279294	rs495828
chr15	36361799	36363599	Rep	chr15	36362396	rs8028529
chr21	42357691	42359291	Rep	chr21	42358786	rs1547374
chr2	104761742	104763542	Rep	chr2	104762499	rs12615966
chr2	133675829	133691429	Rep	chr2	133680388	rs1901440
chr5	2108086	2110486	Rep	chr5	2109787	rs6879627
chr7	153940115	153952515	Rep	chr7	153941163	rs6973850
chr10	118517288	118520488	Pro	chr10	118519432	rs12413624
chr11	96950200	96978800	Pro	chr11	96970814	rs17275283
chr13	65899068	65911468	Pro	chr13	65907683	rs1585440
chr19	29137493	29190493	Pro	chr19	29164379	rs2903018
chr1	238743300	238770300	Pro	chr1	238745053	rs2689154

Continuation of the previous table						
chr2	152775686	152819686	Pro	chr2	152798206	rs12478462
chr3	74664649	74688049	Pro	chr3	74669607	rs1447826
chr6	68375508	68446908	Pro	chr6	68432116	rs9363918
chr6	161810368	161815768	Pro	chr6	161815043	rs3016539
chr7	18796377	18858777	Pro	chr7	18798993	rs12531908
chr11	9950653	9951853	Enh	chr11	9951515	rs10500715
chr12	120984997	120987797	Enh	chr12	120987058	rs7310409
chr1	199994672	199996072	Enh	chr1	199996040	rs10919791
chr21	29308479	29313679	Enh	chr21	29313290	rs1153287
chr21	29352079	29354679	Enh	chr21	29354452	rs2027605
chr2	136796230	136798030	Enh	chr2	136797654	rs1427593
chr3	189790611	189791411	Enh	chr3	189790682	rs9854771
chr9	117306521	117309521	Enh	chr9	117306874	rs10983614
chr9	133268988	133273984	Enh	chr9	133273813	rs505922
chr13	79725265	79726265	Enh	chr13	79725587	rs2039553
chr17	37716192	37719805	Enh	chr17	37718512	rs4795218
chr1	64071928	64073328	Enh	chr1	64073289	rs1747924
chr21	29305679	29307479	Enh	chr21	29305751	rs1153280
chr21	29346679	29352079	Enh	chr21	29348513	rs117214
chr22	28903212	28906212	Enh	chr22	28904318	rs16986825
chr7	47445202	47449202	Enh	chr7	47448971	rs73328514
chr8	38610482	38614082	Enh	chr8	38611785	rs7832232
chr13	73333463	73337063	Enh	chr13	73334709	rs9573163
chr17	18850287	18851687	Enh	chr17	18850557	rs4924935
chr17	72403059	72406859	Enh	chr17	72405335	rs7214041
chr17	72403059	72406859	Enh	chr17	72404025	rs11655237
chr21	29328479	29331479	Enh	chr21	29328775	rs1153294
chr5	124686307	124689507	Enh	chr5	124688588	rs4285214
chr6	160413768	160416968	Enh	chr6	160413796	rs2457571
chr8	124861958	124869958	Enh	chr8	124864572	rs7015626
chr1	200035072	200038872	Enh	chr1	200038304	rs3790844
chr1	200041472	200042472	Enh	chr1	200041696	rs3790843
chr1	200044672	200047672	Enh	chr1	200047018	rs2821367
chr1	236275100	236277100	Enh	chr1	236276616	rs6662005
chr2	101305138	101307938	Enh	chr2	101305708	rs6711606
chr5	1343285	1344685	Enh	chr5	1344343	rs31490
chr7	130995241	130995841	Enh	chr7	130995762	rs6971499
chr8	128555354	128556954	Enh	chr8	128555832	rs1561927

Table S4. Percentage of variants in each chromatin category in the different tissues

(Categories: No – No signal; Rep – Repressed chromatin; Pro – Promoter; Enh - Enhancer).

Percentages of bp		**Percentages of overlapping w/ control SNPs**		**Percentages of overlapping w/ PC SNPs**		**Percentages of overlapping w/ No PC SNPs**	
Duct	%	**Duct**	%	**Duct**	%	**Duct**	%
Enh	14.4	Enh	14.1	Enh	31.9	Enh	20.9
Pro	21.6	Pro	0.8	Pro	2.6	Pro	1.5
Rep	15.0	Rep	37.5	Rep	20.7	Rep	26.4
No	48.9	No	47.6	No	44.8	No	51.2
Acinar	%	**Acinar**	%	**Acinar**	%	**Acinar**	%
Enh	13.6	Enh	13	Enh	28.7	Enh	19.5
Pro	17.3	Pro	0.3	Pro	0	Pro	0.6
Rep	4.2	Rep	23.3	Rep	13.9	Rep	14.3
No	64.9	No	63.4	No	57.4	No	65.6
Endocrine	**%**	**Endocrine**	%	**Endocrine**	%	**Endocrine**	%
Enh	18.7	Enh	18.2	Enh	21.1	Enh	25.4
Pro	5.3	Pro	1	Pro	0.9	Pro	1.6
Rep	41.9	Rep	48.2	Rep	47.4	Rep	40.8
No	34.2	No	32.6	No	30.6	No	32.2
Right ventricle	**%**	**Right ventricle**	%	**Right ventricle**	%	**Right ventricle**	%
Enh	3.8	Enh	4.0	Enh	7.0	Enh	6.8
Pro	4.8	Pro	3.9	Pro	8.7	Pro	9.4
Rep	30.8	Rep	31.3	Rep	36.5	Rep	41.0
No	60.6	No	60.8	No	47.8	No	42.8
Colon	**%**	**Colon**	%	**Colon**	%	**Colon**	%
Enh	3.7	Enh	3.7	Enh	4.4	Enh	6.1
Pro	3.4	Pro	2.7	Pro	4.3	Pro	6.7
Rep	18.0	Rep	17.1	Rep	28.7	Rep	25.9
No	74.9	No	76.5	No	62.6	No	61.3

Table S5. List of genes associated to duct enhancer and their association to PC and other types of cancer. The genes are ranked by the number of enhancers associated to them.

Genes	Number of putative enhancers associated	PC (n=229)	Other types of cancer (n=2178)
NR5A2	5	Yes	Yes
KLF12	4	No	No
KLF5	4	No	Yes
BACH1	3	No	Yes
GRIK1	3	No	No
ZNF281	3	No	No
SLC39A11	2	No	No
SOX9	2	Yes	Yes
GPR137B	1	No	No
ABO	1	No	No
C9	1	No	Yes
CLPTM1L	1	No	Yes
CNOT11	1	No	No
DAB2	1	No	Yes
DDX52	1	No	No
DLL1	1	No	No
DPH6	1	No	No
DYSF	1	No	No
ERMARD	1	No	No
ERO1B	1	No	No
ETAA1	1	No	No
FAM210A	1	No	No
GSX1	1	No	No
HNF1B	1	Yes	Yes
HNF4A	1	No	Yes
KLF14	1	No	No
KREMEN1	1	No	No
LDLRAD4	1	No	No
MEIS1	1	No	Yes
MEIS2	1	No	No
MKLN1	1	No	No
MTSS1	1	No	Yes
MYC	1	Yes	Yes
NDFIP2	1	No	No
OBP2B	1	No	No
P3H2	1	No	No

Continuation of the previous table			
PDX1	1	Yes	Yes
PPFIBP1	1	No	No
REP15	1	No	No
RNF149	1	No	No
ROR1	1	No	Yes
SPRY2	1	No	Yes
TFF1	1	No	Yes
TFF2	1	Yes	Yes
TNS3	1	No	No
TP63	1	No	No
TTPAL	1	No	No
UBE2U	1	No	No
ZNF572	1	No	No
ZNF608	1	No	No
ZNF638	1	No	No
ZNRF3	1	No	Yes

Table S6. List of genes associated to acinar enhancer and their association to PC and other types of cancer. The genes are ranked by the number of enhancers associated to them.

Genes	Number of putative enhancers associated	PC (n=229)	Other types of cancer (n=2178)
BACH1	5	No	Yes
GRIK1	5	No	No
NR5A2	4	Yes	Yes
ZNF281	3	No	No
SLC39A11	2	No	No
SOX9	2	Yes	Yes
ABO	1	No	No
ASTN2	1	No	No
C12orf43	1	No	No
CLPTM1L	1	No	Yes
CNOT11	1	No	No
DDX52	1	No	No
ERO1B	1	No	No
FGFR1	1	No	Yes
GPR137B	1	No	No
HNF1A	1	Yes	Yes
HNF1B	1	No	Yes
KLF12	1	No	No
KLF14	1	No	No
KLF5	1	No	Yes
KREMEN1	1	No	No
LPA	1	No	No
MKLN1	1	No	No
MTSS1	1	No	Yes
MYC	1	Yes	Yes
NDFIP2	1	No	No
OBP2B	1	No	No
P3H2	1	No	No
PRPSAP2	1	No	No
RNF149	1	No	No
ROR1	1	No	Yes
SBF2	1	No	No
SLC22A3	1	No	No
SPRY2	1	No	Yes
SWAP70	1	No	No
TACC1	1	No	No
TERT	1	No	Yes

Continuation of the previous table			
THSD7B	1	No	No
TNS3	1	No	No
TP63	1	No	Yes
TRIM32	1	No	No
TVP23B	1	No	No
UBE2U	1	No	No
ZNF572	1	No	No
ZNF608	1	No	No
ZNRF3	1	No	Yes

Table S7. Gene ontology results for duct enhancers that overlap with PC SNPs

Biological process	Fold Enrichment	FDR
pancreas development	38.31	1.76E-05
endocrine pancreas development	46.26	1.02E-03
positive regulation of transcription by RNA polymerase II	4.83	1.37E-03
endocrine system development	18.96	3.76E-03
positive regulation of RNA biosynthetic process	3.83	4.03E-03
renal tubule development	24.29	4.52E-03
nephron tubule development	25.56	4.57E-03
negative regulation of myeloid cell differentiation	23.13	4.64E-03
regulation of transcription by RNA polymerase II	3.01	4.85E-03
positive regulation of transcription, DNA-templated	3.83	5.34E-03
positive regulation of RNA metabolic process	3.62	5.60E-03
positive regulation of nucleic acid-templated transcription	3.83	6.40E-03
nephron epithelium development	20.03	7.69E-03
negative regulation of cell population proliferation	5.64	1.09E-02
positive regulation of macromolecule biosynthetic process	3.35	1.17E-02
positive regulation of nucleobase-containing compound metabolic process	3.29	1.41E-02
negative regulation of transcription by RNA polymerase II	4.73	1.54E-02
kidney epithelium development	15.54	1.71E-02
positive regulation of cellular biosynthetic process	3.18	1.72E-02
nephron development	15.67	1.74E-02
maintenance of gastrointestinal epithelium	58.29	1.96E-02
positive regulation of biosynthetic process	3.12	2.03E-02
type B pancreatic cell differentiation	55.51	2.13E-02
negative regulation of gene expression	3.06	2.30E-02
enteroendocrine cell differentiation	50.68	2.51E-02
regulation of cellular macromolecule biosynthetic process	2.28	3.05E-02
epithelial cell differentiation	5.19	3.05E-02
columnar/cuboidal epithelial cell differentiation	20.45	3.09E-02
positive regulation of gene expression	2.77	3.11E-02
tube development	4.63	3.14E-02
epithelial structure maintenance	43.18	3.14E-02
regulation of macromolecule biosynthetic process	2.26	3.19E-02
regulation of RNA biosynthetic process	2.35	3.23E-02
regulation of nucleic acid-templated transcription	2.36	3.25E-02
regulation of transcription, DNA-templated	2.36	3.33E-02
negative regulation of RNA biosynthetic process	3.61	3.70E-02
negative regulation of nucleic acid-templated transcription	3.61	3.74E-02
negative regulation of transcription, DNA-templated	3.62	3.79E-02
epithelium development	3.9	3.80E-02

Continuation of the previous table		
regulation of cellular biosynthetic process	2.17	3.85E-02
positive regulation of cellular metabolic process	2.4	3.86E-02
regulation of animal organ morphogenesis	10.73	4.27E-02
columnar/cuboidal epithelial cell development	33.31	4.64E-02
positive regulation of nitrogen compound metabolic process	2.41	4.78E-02
Molecular function		
cis-regulatory region sequence-specific DNA binding	5.07	1.15E-04
DNA-binding transcription factor activity	4.54	1.25E-04
RNA polymerase II cis-regulatory region sequence-specific DNA binding	5.15	1.84E-04
DNA-binding transcription factor activity, RNA polymerase II-specific	4.44	3.44E-04
RNA polymerase II transcription regulatory region sequence-specific DNA binding	4.34	3.80E-04
double-stranded DNA binding	3.92	5.12E-04
regulatory region nucleic acid binding	4.06	5.70E-04
transcription regulatory region sequence-specific DNA binding	4.06	6.40E-04
sequence-specific double-stranded DNA binding	3.9	8.45E-04
DNA-binding transcription activator activity	7.59	1.20E-03
DNA-binding transcription activator activity, RNA polymerase II-specific	7.65	1.23E-03
sequence-specific DNA binding	3.66	1.44E-03
transcription regulator activity	3.38	1.74E-03
DNA binding	2.82	9.67E-03
Cellular component		
chromatin	4.96	3.70E-04
chromosome	3.58	4.86E-03

Table S8. Gene ontology results for acinar enhancers that overlap with PC SNPs

Biological process	Fold Enrichment	FDR
organic cyclic compound biosynthetic process	3.1	4.50E-03
cellular metabolic process	2.11	4.97E-03
regulation of biosynthetic process	3.01	5.47E-03
transcription by RNA polymerase II	3.41	5.61E-03
organic cyclic compound metabolic process	2.5	5.68E-03
regulation of macromolecule biosynthetic process	3.04	5.70E-03
cellular aromatic compound metabolic process	2.54	5.72E-03
regulation of RNA metabolic process	3.05	5.73E-03
aromatic compound biosynthetic process	3.14	5.79E-03
regulation of nitrogen compound metabolic process	2.69	5.83E-03
cellular nitrogen compound biosynthetic process	2.81	5.85E-03
regulation of nucleobase-containing compound metabolic process	2.98	5.88E-03
regulation of cellular biosynthetic process	3.02	5.90E-03
RNA biosynthetic process	3.13	5.94E-03
nucleic acid metabolic process	2.65	5.94E-03
regulation of cellular macromolecule biosynthetic process	3.05	5.96E-03
regulation of transcription by RNA polymerase II	3.51	6.03E-03
transcription, DNA-templated	3.14	6.09E-03
heterocycle metabolic process	2.55	6.22E-03
regulation of RNA biosynthetic process	3.24	6.34E-03
regulation of metabolic process	2.64	6.36E-03
metabolic process	2.01	6.39E-03
regulation of primary metabolic process	2.66	6.42E-03
nucleic acid-templated transcription	3.14	6.52E-03
cellular macromolecule metabolic process	2.33	6.67E-03
nucleobase-containing compound metabolic process	2.59	6.97E-03
regulation of nucleic acid-templated transcription	3.24	7.04E-03
heterocycle biosynthetic process	3.14	7.67E-03
primary metabolic process	1.99	7.76E-03
regulation of transcription, DNA-templated	3.24	7.92E-03
nucleobase-containing compound biosynthetic process	3.18	9.96E-03
macromolecule biosynthetic process	2.65	1.04E-02
cellular macromolecule biosynthetic process	2.66	1.05E-02
cellular nitrogen compound metabolic process	2.34	1.07E-02
regulation of macromolecule metabolic process	2.49	1.26E-02
cellular biosynthetic process	2.46	1.42E-02
regulation of gene expression	2.67	1.52E-02

Continuation of the previous table		
organic substance biosynthetic process	2.43	1.57E-02
biosynthetic process	2.42	1.57E-02
organic substance metabolic process	1.89	1.62E-02
regulation of cellular metabolic process	2.88	1.99E-02
nitrogen compound metabolic process	1.92	2.07E-02
macromolecule metabolic process	1.95	2.81E-02
cellular process	1.51	3.63E-02
RNA metabolic process	2.42	3.70E-02
Molecular function		
transcription regulator activity	3.52	6.09E-03
DNA-binding transcription factor activity	4.13	6.14E-03
cis-regulatory region sequence-specific DNA binding	4.55	7.96E-03
DNA-binding transcription factor activity, RNA polymerase II-specific	3.98	1.32E-02
RNA polymerase II cis-regulatory region sequence-specific DNA binding	4.62	1.33E-02
RNA polymerase II transcription regulatory region sequence-specific DNA binding	3.89	1.41E-02
sequence-specific DNA binding	3.54	1.52E-02
regulatory region nucleic acid binding	3.64	1.88E-02
transcription regulatory region sequence-specific DNA binding	3.65	2.09E-02
sequence-specific double-stranded DNA binding	3.5	2.54E-02
DNA binding	2.77	3.41E-02
DNA-binding transcription activator activity	6.52	3.56E-02
double-stranded DNA binding	3.31	3.80E-02
DNA-binding transcription activator activity, RNA polymerase II-specific	6.57	3.99E-02
Cellular component		
chromatin	4.38	2.50E-02

Table S9. List of genes associated to putative enhancers active in pancreatic duct cells

This table are available in:
https://drive.google.com/drive/folders/1tT9dxHaJyxYCAFNSFuKrpfaT5q5r2l7v?usp=sharing

Table S10. List of genes associated to putative enhancers active in pancreatic duct cells

This table are available
in:https://drive.google.com/drive/folders/1tT9dxHaJyxYCAFNSFuKrpfaT5q5r2l7v?usp=sharing

Table S11. List of enhancer sequences tested in vivo.

Name	% GFP expression	P-value	N
seq41_wt	7.14	Ns	28
seq38_wt	27.78	0,017	36
seq44_wt	14.81	Ns	27
seq67_wt	16.67	Ns	30
seq56_wt	24.24	0,039	33
seq65_wt	30.00	0,012	30
seq34_wt	17.24	Ns	29

Table S12. List of primers used in this study.

Name	Sequence	Application
seq41_Fw	TACAAGCCATGGACCCTTTGC	Enhancer assay
seq41_Rv	GGAGTTGATGGTTAGGATGCC	Enhancer assay
seq38_Fw	GTAGTGGGCTATGATTCTGCC	Enhancer assay
seq38_Rv	GGCAGGTGACATTAACCAGG	Enhancer assay
seq44_Fw	GCTTGTCTCACTAGGTCAGCC	Enhancer assay
seq44_Rv	GGCTCAGGCTCCAGTCCC	Enhancer assay
seq67_Fw	AATAAAGCAATAACAGGGATACATATCACC	Enhancer assay
seq67_Rv	CGTTACAATAGCCCACAAAGATTTCC	Enhancer assay
seq56_Fw	GAAGTTGACATGCTCTGGTCC	Enhancer assay
seq56_Rv	AGTGGAAGTGAAGATTGACTGC	Enhancer assay
seq65_Fw	ATCATGACCTCTGCAGTTCC	Enhancer assay
seq65_Rv	CTAAGAATTTGCTAGAGGGCC	Enhancer assay
seq34_Fw	GCTTCCTTATTGTATCGG	Enhancer assay
seq34_Rv	TGGCTATCAGTATCAGG	Enhancer assay
seq41_SDM_Fw	GATGGTGCTGAACCTATCACTTAG	Site directed mutagenesis
seq41_SDM_Rv	CTAAGTGATAGGTTCAGCACCATC	Site directed mutagenesis
seq38_SDM_Fw	CTCACCTGTATACCCAGCAATTTGG	Site directed mutagenesis
seq38_SDM_Rv	CCAAATTGCTGGGTATACAGGTGAG	Site directed mutagenesis
seq44_SDM_Fw	CTAAAACTGGAGAGTCTGTCG	Site directed mutagenesis
seq44_SDM_Rv	CGACAGACTCTCCAGTTTTAG	Site directed mutagenesis
seq67_SDM_Fw	GTATACCTAGAATTTACAATAAATTT	Site directed mutagenesis
seq67_SDM_Rv	AAATTTATTGTAAATTCTAGGTATAC	Site directed mutagenesis
seq56_SDM_Fw	TCTATTTCCCACCCACTTTTTTCT	Site directed mutagenesis
seq56_SDM_Rv	AGAAAAAAGTGGGTGGGAAATAGA	Site directed mutagenesis
seq65_SDM_Fw	GTATATTGGTAGGTTCAGAGGGTAAG	Site directed mutagenesis
seq65_SDM_Rv	CTTACCCTCTGAACCTACCAATATAC	Site directed mutagenesis
seq34_SDM_Fw	CCAGGAACATTGGGGTTGC	Site directed mutagenesis
seq34_SDM_Rv	GCAACCCCAATGTTCCTGG	Site directed mutagenesis

.

4.8 References

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M.B. (2010). Annotating non-coding regions of the genome. Nat Rev Genet *11*, 559–571. https://doi.org/10.1038/nrg2814.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461. https://doi.org/10.1038/nature12787.

Arda, H.E., Tsai, J., Rosli, Y.R., Giresi, P., Bottino, R., Greenleaf, W.J., Chang, H.Y., and Kim, S.K. (2018). A Chromatin Basis for Cell Lineage and Disease Risk in the Human Pancreas. Cell Syst *7*, 310-322.e4. https://doi.org/10.1016/j.cels.2018.07.007.

Arnes, L., Liu, Z., Wang, J., Maurer, C., Sagalovskiy, I., Sanchez-Martin, M., Bommakanti, N., Garofalo, D.C., Balderes, D.A., Sussel, L., et al. (2019). Comprehensive characterisation of compartment-specific long non-coding RNAs associated with pancreatic ductal adenocarcinoma. Gut *68*, 499–511. https://doi.org/10.1136/gutjnl-2017-314353.

Arnold, P.R., Wells, A.D., and Li, X.C. (2019). Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. Front Cell Dev Biol *7*, 377.

Backx, E., Coolens, K., Van den Bossche, J.-L., Houbracken, I., Espinet, E., and Rooman, I. (2021). On the Origin of Pancreatic Cancer. Cell Mol Gastroenterol Hepatol S2352-345X(21)00248-4.

Backx, E., Coolens, K., Van den Bossche, J.-L., Houbracken, I., Espinet, E., and Rooman, I. (2022). On the Origin of Pancreatic Cancer: Molecular Tumor Subtypes in Perspective of Exocrine Cell Plasticity. Cell Mol Gastroenterol Hepatol *13*, 1243–1253. https://doi.org/10.1016/j.jcmgh.2021.11.010.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837. https://doi.org/10.1016/j.cell.2007.05.009.

Bastidas-Ponce, A., Scheibner, K., Lickert, H., and Bakhti, M. (2017). Cellular and molecular mechanisms coordinating pancreas development. Development *144*, 2873–2888. https://doi.org/10.1242/dev.140756.

Becker, A.E., Hernandez, Y.G., Frucht, H., and Lucas, A.L. (2014). Pancreatic ductal adenocarcinoma: risk factors, screening, and early detection. World J Gastroenterol *20*, 11182–11198. https://doi.org/10.3748/wjg.v20.i32.11182.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods *58*, 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001.

Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F., et al. (2009). Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. Dev Dyn *238*, 2409–2417. https://doi.org/10.1002/dvdy.22051.

Cai, Y., Zhang, Y., Loh, Y.P., Tng, J.Q., Lim, M.C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L., et al. (2021). H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun *12*, 719. https://doi.org/10.1038/s41467-021-20940-y.

de la Calle-Mustienes, E., Feijóo, C.G., Manzanares, M., Tena, J.J., Rodríguez-Seguel, E., Letizia, A., Allende, M.L., and Gómez-Skarmeta, J.L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res *15*, 1061–1072. https://doi.org/10.1101/gr.4004805.

Campa, D., Gentiluomo, M., Obazee, O., Ballerini, A., Vodickova, L., Hegyi, P., Soucek, P., Brenner, H., Milanetto, A.C., Landi, S., et al. (2020). Genome-wide association study identifies an early onset pancreatic cancer risk locus. Int J Cancer *147*, 2065–2074. https://doi.org/10.1002/ijc.33004.

Cobo, I., Martinelli, P., Flández, M., Bakiri, L., Zhang, M., Carrillo-de-Santa-Pau, E., Jia, J., Sánchez-Arévalo Lobo, V.J., Megías, D., Felipe, I., et al. (2018). Transcriptional regulation by NR5A2 links differentiation and inflammation in the pancreas. Nature *554*, 533–537. https://doi.org/10.1038/nature25751.

Coppola, C.J., C Ramaker, R., and Mendenhall, E.M. (2016). Identification and function of enhancers in the human genome. Hum Mol Genet *25*, R190–R197. https://doi.org/10.1093/hmg/ddw216.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931–21936. https://doi.org/10.1073/pnas.1016071107.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311. https://doi.org/10.1126/science.1067799.

Diaferia, G.R., Balestrieri, C., Prosperini, E., Nicoli, P., Spaggiari, P., Zerbi, A., and Natoli, G. (2016). Dissection of transcriptional and cis-regulatory control of differentiation in human pancreatic cancer. EMBO J *35*, 595–617. https://doi.org/10.15252/embj.201592404.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc *12*, 2478–2492. https://doi.org/10.1038/nprot.2017.124.

Eufrásio, A., Perrod, C., Ferreira, F.J., Duque, M., Galhardo, M., and Bessa, J. (2020). In Vivo Reporter Assays Uncover Changes in Enhancer Activity Caused by Type 2 Diabetes-Associated Single Nucleotide Polymorphisms. Diabetes *69*, 2794–2805. https://doi.org/10.2337/db19-1049.

Feigin, M.E., Garvin, T., Bailey, P., Waddell, N., Chang, D.K., Kelley, D.R., Shuai, S., Gallinger, S., McPherson, J.D., Grimmond, S.M., et al. (2017). Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma. Nat Genet *49*, 825–833. https://doi.org/10.1038/ng.3861.

Felsenstein, M., Hruban, R.H., and Wood, L.D. (2018). New Developments in the Molecular Mechanisms of Pancreatic Tumorigenesis. Adv Anat Pathol *25*, 131–142. https://doi.org/10.1097/PAP.0000000000000172.

von Figura, G., Morris, J.P., Wright, C.V.E., and Hebrok, M. (2014). Nr5a2 maintains acinar cell differentiation and constrains oncogenic Kras-mediated pancreatic neoplastic initiation. Gut *63*, 656–664. https://doi.org/10.1136/gutjnl-2012-304287.

Flandez, M., Cendrowski, J., Cañamero, M., Salas, A., del Pozo, N., Schoonjans, K., and Real, F.X. (2014). Nr5a2 heterozygosity sensitises to, and cooperates with, inflammation in KRas(G12V)-driven pancreatic tumourigenesis. Gut *63*, 647–655. https://doi.org/10.1136/gutjnl-2012-304381.

GBD 2017 Pancreatic Cancer Collaborators (2019). The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol *4*, 934–947. https://doi.org/10.1016/S2468-1253(19)30347-4.

Guo, F., Zhou, Y., Guo, H., Ren, D., Jin, X., and Wu, H. (2021). NR5A2 transcriptional activation by BRD4 promotes pancreatic cancer progression by upregulating GDF15. Cell Death Discov *7*, 78. https://doi.org/10.1038/s41420-021-00462-8.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

Karlsson, J., von Hofsten, J., and Olsson, P.E. (2001). Generating transparent zebrafish: a refined method to improve detection of gene expression during embryonic development. Mar Biotechnol (NY) *3*, 522–527. https://doi.org/10.1007/s1012601-0053-4.

Kawakami, K., Shima, A., and Kawakami, N. (2000). Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. Proc Natl Acad Sci U S A *97*, 11403–11408. https://doi.org/10.1073/pnas.97.21.11403.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res *46*, D260–D266. https://doi.org/10.1093/nar/gkx1126.

Kim, H.-R., Yim, J., Yoo, H.-B., Lee, S.E., Oh, S., Jung, S., Hwang, C.-I., Shin, D.-M., Kim, T., Yoo, K.H., et al. (2021). EVI1 activates tumor-promoting transcriptional enhancers in pancreatic cancer. NAR Cancer *3*, zcab023. https://doi.org/10.1093/narcan/zcab023.

Klein, A.P., Wolpin, B.M., Risch, H.A., Stolzenberg-Solomon, R.Z., Mocci, E., Zhang, M., Canzian, F., Childs, E.J., Hoskins, J.W., Jermusyk, A., et al. (2018). Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. Nat Commun *9*, 556. https://doi.org/10.1038/s41467-018-02942-5.

Kong, B., Michalski, C.W., Erkan, M., Friess, H., and Kleeff, J. (2011). From tissue turnover to the cell of origin for pancreatic cancer. Nat Rev Gastroenterol Hepatol *8*, 467–472. https://doi.org/10.1038/nrgastro.2011.114.

Lazarus, K.A., Wijayakumara, D., Chand, A.L., Simpson, E.R., and Clyne, C.D. (2012). Therapeutic potential of Liver Receptor Homolog-1 modulators. J Steroid Biochem Mol Biol *130*, 138–146. https://doi.org/10.1016/j.jsbmb.2011.12.017.

Laverré, A., Tannier, E., and Necsulea, A. (2022). Long-range promoter-enhancer contacts are conserved during evolution and contribute to gene expression robustness. Genome Res *32*, 280–296.

Lee, Y.-K., and Moore, D.D. (2008). Liver receptor homolog-1, an emerging metabolic modulator. Front Biosci *13*, 5950–5958. https://doi.org/10.2741/3128.

Lin, Q., Aihara, A., Chung, W., Li, Y., Chen, X., Huang, Z., Weng, S., Carlson, R.I., Nadolny, C., Wands, J.R., et al. (2014). LRH1 promotes pancreatic cancer metastasis. Cancer Lett *350*, 15–24. https://doi.org/10.1016/j.canlet.2014.04.017.

Lippi, G., and Mattiuzzi, C. (2020). The global burden of pancreatic cancer. Arch Med Sci *16*, 820–824. https://doi.org/10.5114/aoms.2020.94845.

López de Maturana, E., Rodríguez, J.A., Alonso, L., Lao, O., Molina-Montes, E., Martín-Antoniano, I.A., Gómez-Rubio, P., Lawlor, R., Carrato, A., Hidalgo, M., et al. (2021). A multilayered post-GWAS assessment on genetic susceptibility to pancreatic cancer. Genome Med *13*, 15. https://doi.org/10.1186/s13073-020-00816-4.

Luo, W., Tao, J., Zheng, L., and Zhang, T. (2020). Current epidemiology of pancreatic cancer: Challenges and opportunities. Chin J Cancer Res *32*, 705–719. https://doi.org/10.21147/j.issn.1000-9604.2020.06.04.

Luo, Z., Li, Y., Zuo, M., Liu, C., Tian, W., Yan, D., Wang, H., and Li, D. (2017). Effect of NR5A2 inhibition on pancreatic cancer stem cell (CSC) properties and epithelial-mesenchymal transition (EMT) markers. Mol Carcinog *56*, 1438–1448. https://doi.org/10.1002/mc.22604.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195. https://doi.org/10.1126/science.1222794.

McKenna, L.R., and Edil, B.H. (2014). Update on pancreatic neuroendocrine tumors. Gland Surg *3*, 258–275. https://doi.org/10.3978/j.issn.2227-684X.2014.06.03.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495–501. https://doi.org/10.1038/nbt.1630.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res *47*, D419–D426. https://doi.org/10.1093/nar/gky1038.

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G.P., and Eliopoulos, E. (2020). Transcription factors and evolution: An integral part of gene expression (Review). World Academy of Sciences Journal *2*, 3–8. https://doi.org/10.3892/wasj.2020.32.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet *12*, 283–293. https://doi.org/10.1038/nrg2957.

Pan, F.C., and Wright, C. (2011). Pancreas organogenesis: from bud to plexus to gland. Dev Dyn *240*, 530–565. https://doi.org/10.1002/dvdy.22584.

Pennacchio, L.A., Loots, G.G., Nobrega, M.A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. Genome Res *17*, 201–211. https://doi.org/10.1101/gr.5972507.

Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res *48*, D845–D855. https://doi.org/10.1093/nar/gkz1021.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., and Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes *12*, 315. https://doi.org/10.1186/s13104-019-4343-8.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rawla, P., Sunkara, T., and Gaduputi, V. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. World J Oncol *10*, 10–27. https://doi.org/10.14740/wjon1166.

Roe, J.-S., Hwang, C.-I., Somerville, T.D.D., Milazzo, J.P., Lee, E.J., Da Silva, B., Maiorino, L., Tiriac, H., Young, C.M., Miyabayashi, K., et al. (2017). Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. Cell *170*, 875-888.e20. https://doi.org/10.1016/j.cell.2017.07.007.

Scacheri, C.A., and Scacheri, P.C. (2015). Mutations in the noncoding genome. Curr Opin Pediatr *27*, 659–664. https://doi.org/10.1097/MOP.0000000000000283.

Scarpa, A., and Mafficini, A. (2018). Non-coding regulatory variations: the dark matter of pancreatic cancer genomics. Gut *67*, 399–400. https://doi.org/10.1136/gutjnl-2017-314310.

Somerville, T.D.D., Xu, Y., Miyabayashi, K., Tiriac, H., Cleary, C.R., Maia-Silva, D., Milazzo, J.P., Tuveson, D.A., and Vakoc, C.R. (2018). TP63-Mediated Enhancer Reprogramming Drives the Squamous Subtype of Pancreatic Ductal Adenocarcinoma. Cell Rep *25*, 1741-1755.e7. https://doi.org/10.1016/j.celrep.2018.10.051.

Stanger, B.Z., and Dor, Y. (2006). Dissecting the cellular origins of pancreatic cancer. Cell Cycle *5*, 43–46. https://doi.org/10.4161/cc.5.1.2291.

Tiriac, H., Plenker, D., Baker, L.A., and Tuveson, D.A. (2019). Organoid models for translational pancreatic cancer research. Curr Opin Genet Dev *54*, 7–11. https://doi.org/10.1016/j.gde.2019.02.003.

Tjian, R., and Maniatis, T. (1994). Transcriptional activation: a complex puzzle with few easy pieces. Cell *77*, 5–8. https://doi.org/10.1016/0092-8674(94)90227-5.

Tobias, I.C., Abatti, L.E., Moorthy, S.D., Mullany, S., Taylor, T., Khader, N., Filice, M.A., and Mitchell, J.A. (2021). Transcriptional enhancers: from prediction to functional assessment on a genome-wide scale. Genome *64*, 426–448. https://doi.org/10.1139/gen-2020-0104.

Vaz, S., Ferreira, F.J., Macedo, J.C., Leor, G., Ben-David, U., Bessa, J., and Logarinho, E. (2021). FOXM1 repression increases mitotic death upon antimitotic chemotherapy through BMF upregulation. Cell Death Dis *12*, 542. https://doi.org/10.1038/s41419-021-03822-5.

Wilkinson, A.C., Nakauchi, H., and Göttgens, B. (2017). Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. Cell Syst *5*, 319–331. https://doi.org/10.1016/j.cels.2017.07.004.

Wood, L.D., and Maitra, A. (2021). Insights into the origins of pancreatic cancer. Nature *597*, 641–642.

Worsley-Hunt, R., Bernard, V., and Wasserman, W.W. (2011). Identification of cis-regulatory sequence variations in individual genome sequences. Genome Med *3*, 65. https://doi.org/10.1186/gm281.

Xu, Y., Liu, J., Nipper, M., and Wang, P. (2019). Ductal vs. acinar? Recent insights into identifying cell lineage of pancreatic ductal adenocarcinoma. Ann Pancreat Cancer *2*, 11. https://doi.org/10.21037/apc.2019.06.03.

Yue F, Cheng Y, Breschi A et al (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515:355–364. doi:10.1038/nature13992

Chapter V – Concluding remarks

5.1 Concluding remarks

Pancreatic cancer (PC) is seriously social burden due to its late diagnosis, poor prognosis, and early metastasis, which leads to a continuous increasing rate of incidence and death worldwide (Klein, 2021; Hayashi et al., 2021). Therefore, there is an urgent need to understand the complicated biology behind this lethal disease, to find novel methods to early detect PC and to develop new treatments to improve the patient outcomes (Rucki and Zheng, 2014).

Medical genetics has been exhaustively investigating coding-regions of the genome to identify the origin of several human genetic diseases. However, the cause of many diseases is still not clarified, such as the case of PC. Thanks to the precious information of genome-wide association studies, it has been observed that many PC predisposition genomic variants are lying within the non-coding genome, potentially affecting the mechanisms of control of transcription (Maurya, 2021). To understand the regulatory mechanisms involved in the development of PC, the regulatory networks involved in the proper function and development of pancreas, needs to be understood first. In chapter II, we identified the zebrafish putative enhancers that operate in the differentiated pancreas, through the application of several next-generation sequencing assays, such as ATAC-seq (chromatin availability) and H3K27ac ChIP-seq (active promoter/enhancer regions). We found that most of the active putative enhancers are in intergenic regions, being in concordance with previous studies (Wang et al., 2017; Farber and Lane, 2019). Also, comparing the H3K27ac data from adult zebrafish pancreas to whole zebrafish embryos, we found that half of putative enhancers are active only in the differentiated adult pancreas, suggesting that their activity is not restricted to the pancreas and have also a role during development. Enhancer activity is widely flexible, with enhancers highly tissue specific, and others with a broad activity, being active in multiple tissues (Montefiori et al., 2018) (Jung et al., 2019). Thus, the results that we showed in chapter II are in line with what is expected to be the enhancer behaviour. We also found that, although most of the zebrafish pancreatic enhancers do not share significant sequence identity with human pancreatic enhancers, they share many transcription factors binding sites (TFBSs), and their target genes are enriched for human pancreas diseases, suggesting the existence of functionally equivalent enhancers in zebrafish and humans, as proposed for other tissues and species (Khoueiry et al., 2017; Yang et al., 2015). Additionally, these results also reinforce the robustness of zebrafish as an in vivo model to study cis-regulation in the context of human pancreatic diseases.

Then, we focused on the regulatory landscape of *arid1ab*, a tumour-suppressor gene active in the pancreas (Kimura et al., 2018; Wang et al., 2019) and we showed that within a microsyntenic region in the *ARID1A/arid1ab* locus in humans and zebrafish, there are pancreatic enhancers that share regulatory information, although not sharing significant sequence identity. We further showed that the deletion of the human *ARID1A* pancreatic enhancer impairs ARID1A expression, defining a locus for non-coding mutations that may increase the risk for PC. Additionally, in chapter III of this doctoral thesis, we generated a deletion in the zebrafish *arid1ab* pancreatic enhancer (arid1ab_zE), described in chapter II, and started to explore the in vivo impact of this mutation in the development of PC. In these experiments, we were able to identify adult zebrafish harbouring a deletion in arid1ab_zE, however allelic combinations of mutations in arid1ab_zE and *tp53* coding gene were less frequent than expected. Although the number of genotyped animals were not sufficient to obtain a statistically significant result, they are indicative that the arid1ab_zE pancreatic enhancer might be important for organism viability, in particular when in presence of a *tp53* mutant sensitized genetic background. Apart from increasing the number of genotyped animals, future experiment should determine if the viability is somehow related with disruption of pancreatic tissue, including exploring phenotypes related with PC.

In the chapter II, we also explored the potential of functional equivalency of an enhancer in the regulatory landscape of the human and zebrafish *PTF1A*/*ptf1a* genes (Jin and Xiang, 2019), showing that the zebrafish and human enhancers share regulatory information and biological functions during pancreas development. The loss-of-function assays of the zebrafish *ptf1a* enhancer shown a reduction in the pancreatic progenitor domain and pancreatic hypoplasia, a phenotype consistent with the impact of mutations described in the human *PTF1A* regulatory landscape, which are associated with pancreatic agenesis (Weedon et al., 2014). The decrease of the pancreatic progenitor domain in zebrafish may explain the phenotype observed in humans, contributing to the elucidation of its molecular and cellular origin. Overall, in chapter II, we described a strategy that allow us to identify and test in vivo enhancers that, when altered, can affect the expression of disease-associated genes. This strategy can be useful to identify where in the genome disease-causing non-coding mutations may occur by predicting disease relevant enhancers based on phenotypic description of enhancers' loss-of-function. Another challenge is the systematic identification of human/zebrafish putative functional equivalent enhancers, that can also help to identify disease relevant non-coding sequences. This challenge can likely be solved by applying computational methods to identify human/zebrafish enhancers that share similar chromatin profiles, TFBS motifs and chromatin looping with the promoter of corresponding orthologue genes.

In this thesis, we further explored the importance of human pancreatic enhancers in the development of PC, as described on chapter IV. We showed that PC risk alleles are enriched in genomic locations that overlap with epigenetic marks associated to enhancer activity, present in pancreatic duct and acinar cells, the group of cells suggested to give origin to pancreatic ductal adenocarcinoma (Gao et al., 2020), the most common form of PC (Backx et al., 2022). After predicting the target genes of putative pancreatic enhancers that overlap with PC associated SNPs, we have performed gene ontology enrichment analysis, finding an enrichment for pancreatic development and cis-regulatory functions. These results indicate the expected biological functions that might be affected by the sequence alteration in the PC associated enhancer regions. Then, we validated in vitro some of these sequences as enhancer, performing luciferase enhancer reporter assays in a human duct cell line and in vivo using zebrafish. With these assays, we observed that for some of the tested sequences, the risk allele impacts significantly in the regulatory output of the enhancer, when comparing with the non-risk allele, demonstrating that PC risk alleles have the potential to modulate the activity of enhancers in a sequence-specific manner. Focusing on the genomic landscape of the *NR5A2* gene, we found one duct enhancer, seq44, that showed a dramatic decrease in its enhancer activity in the PC associated allele, comparing with the non-risk allele. These results suggest that PC associated alleles have the potential to significantly affect the regulatory output of enhancers, contributing to the dysregulation of their target genes and consequently, contribute to PC development.

In the next decades it is predicted, due to an increment in incidence and mortality, that PC will become the third leading cause of cancer-related deaths by 2025 in Europe (GBD 2017 Pancreatic Cancer Collaborators, 2019). However, until now there are no early detection methods or effective treatments available to prevent these increasing numbers. Most of the studies focusing on the genetic causes of PC have focused on the coding part of genome, while the contribution of non-coding regions, already described as having important roles in the transcriptional regulation (Gahan, 2005), have been vastly disregarded. Overall, in this doctoral thesis, we studied the importance of pancreatic transcriptional CREs in the development of PC. We identified pancreatic enhancers in zebrafish genome and their functional equivalents in human, explored the impact of enhancer mutations in the development of PC in vivo and identified and functional assessed human pancreatic enhancers in PC development. The findings presented here show that the non-coding regions cannot be ignored and could be essential to discover novel genetic players that will be helpful to develop novel diagnostic and prognostic approaches, as well as novel therapeutics to combat this highly lethal and devastating disease.

5.2 References

Backx, E., Coolens, K., Van den Bossche, J.-L., Houbracken, I., Espinet, E., and Rooman, I. (2022). On the Origin of Pancreatic Cancer: Molecular Tumor Subtypes in Perspective of Exocrine Cell Plasticity. Cell Mol Gastroenterol Hepatol *13*, 1243–1253. https://doi.org/10.1016/j.jcmgh.2021.11.010.

Farber, J.E., and Lane, R.P. (2019). Bioinformatics Discovery of Putative Enhancers within Mouse Odorant Receptor Gene Clusters. Chem Senses *44*, 705–720. https://doi.org/10.1093/chemse/bjz043.

Gao, H.-L., Wang, W.-Q., Yu, X.-J., and Liu, L. (2020). Molecular drivers and cells of origin in pancreatic ductal adenocarcinoma and pancreatic neuroendocrine carcinoma. Exp Hematol Oncol *9*, 28. https://doi.org/10.1186/s40164-020-00184-0.

Gahan, P.B. (2005). Molecular biology of the cell (4th edn) B. Alberts, A. Johnson, J. Lewis, K. Roberts and P. Walter (eds), Garland Science, 1463 pp., ISBN 0-8153-4072-9 (paperback) (2002). Cell Biochemistry and Function *23*, 150–150.

GBD 2017 Pancreatic Cancer Collaborators (2019). The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol *4*, 934–947.

Hayashi, A., Hong, J., and Iacobuzio-Donahue, C.A. (2021). The pancreatic cancer genome revisited. Nat Rev Gastroenterol Hepatol *18*, 469–481. https://doi.org/10.1038/s41575-021-00463-z.

Jin, K., and Xiang, M. (2019). Transcription factor Ptf1a in development, diseases and reprogramming. Cell Mol Life Sci *76*, 921–940. https://doi.org/10.1007/s00018-018-2972-z.

Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet *51*, 1442–1449. https://doi.org/10.1038/s41588-019-0494-8.

Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.-C., Gustafson, E.H., Sinha, S., and Furlong, E.E. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. Elife *6*, e28440. https://doi.org/10.7554/eLife.28440.

Kimura, Y., Fukuda, A., Ogawa, S., Maruno, T., Takada, Y., Tsuda, M., Hiramatsu, Y., Araki, O., Nagao, M., Yoshikawa, T., et al. (2018). ARID1A Maintains Differentiation of Pancreatic Ductal Cells and Inhibits Development of Pancreatic Ductal Adenocarcinoma in Mice. Gastroenterology *155*, 194-209.e2. https://doi.org/10.1053/j.gastro.2018.03.039.

Klein, A.P. (2021). Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. Nat Rev Gastroenterol Hepatol *18*, 493–502. https://doi.org/10.1038/s41575-021-00457-x.

Maurya, S.S. (2021). Role of Enhancers in Development and Diseases. Epigenomes *5*, 21. https://doi.org/10.3390/epigenomes5040021.

Montefiori, L.E., Sobreira, D.R., Sakabe, N.J., Aneas, I., Joslin, A.C., Hansen, G.T., Bozek, G., Moskowitz, I.P., McNally, E.M., and Nóbrega, M.A. (2018). A promoter interaction map for cardiovascular disease genetics. Elife *7*, e35788. https://doi.org/10.7554/eLife.35788.

Rucki, A.A., and Zheng, L. (2014). Pancreatic cancer stroma: understanding biology leads to new therapeutic strategies. World J Gastroenterol *20*, 2237–2246. https://doi.org/10.3748/wjg.v20.i9.2237.

Wang, M., Hancock, T.P., MacLeod, I.M., Pryce, J.E., Cocks, B.G., and Hayes, B.J. (2017). Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. Genet Sel Evol *49*, 56. https://doi.org/10.1186/s12711-017-0331-4.

Wang, S.C., Nassour, I., Xiao, S., Zhang, S., Luo, X., Lee, J., Li, L., Sun, X., Nguyen, L.H., Chuang, J.-C., et al. (2019). SWI/SNF component ARID1A restrains pancreatic neoplasia formation. Gut *68*, 1259–1270. https://doi.org/10.1136/gutjnl-2017-315490.

Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., De Franco, E., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., Allen, H.L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet *46*, 61–64. https://doi.org/10.1038/ng.2826.

Yang, S., Oksenberg, N., Takayama, S., Heo, S.-J., Poliakov, A., Ahituv, N., Dubchak, I., and Boffelli, D. (2015). Functionally conserved enhancers with divergent sequences in distant vertebrates. BMC Genomics *16*, 882. https://doi.org/10.1186/s12864-015-2070-7.