Universidade de Aveiro Departamento de Biologia
Ano 2018

Ana Catarina Macedo Eufrásio

Testing the *cis*-regulatory potential of type 2 diabetes associated non-coding sequences

O potencial regulatório em *cis* de sequências não codificantes associadas a diabetes tipo 2

DECLARAÇÃO

Declaro que este relatório é integralmente da minha autoria, estando devidamente referenciadas as fontes e obras consultadas, bem como identificadas de modo claro as citações dessas obras. Não contém, por isso, qualquer tipo de plágio quer de textos publicados, qualquer que seja o meio dessa publicação, incluindo meios eletrónicos, quer de trabalhos académicos.

Universidade de Aveiro Departamento de Biologia
Ano 2018

Ana Catarina Macedo Eufrásio

Testing the *cis*-regulatory potential of type 2 diabetes associated non-coding sequences

O potencial regulatório em *cis* de sequências não codificantes associadas a diabetes tipo 2

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia Molecular e Celular, realizada sob a orientação científica do Doutor José Bessa, investigador principal do Instituto de Investigação e Inovação em Saúde do Porto e do Doutor Luís Souto de Miranda, Professor auxiliar convidado do Departamento de Biologia da Universidade de Aveiro

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No 680156 - ZPR)

O JÚRI

Presidente

Prof. Doutor Mário Pacheco
Professor auxiliar com agregação do Departamento de Biologia da Universidade de Aveiro

Doutor José Bessa
Investigador auxiliar do Instituto de Investigação e Inovação em saúde da Universidade do Porto

Doutora Renata Freitas
Investigadora auxiliar do Instituto de Investigação e Inovação em saúde da Universidade do Porto

Dedico este trabalho à minha mãe, irmão e namorado pelo incansável apoio.

Agradecimentos

Ao meu orientador externo, Dr. José Bessa, pela orientação científica essencial à concretização deste projeto, pelo apoio incansável, atenção, disponibilidade e por todos os ensinamentos durante este último ano. Ao meu orientador interno da UA, Dr. Luís Souto pelo apoio e força prestada. A toda a equipa do laboratório VDR, por todas as gargalhadas, ensinamentos e os múltiplos "é normal, a ciência é assim" em especial à Joana Teixeira, a minha "madrinha" sempre pronta a ajudar. A todos os que contribuíram fora do meu laboratório. À minha família, por todo o apoio constante, compreensão e paciência. Ao meu namorado, que mesmo estando longe, que nunca me deixou querer menos que a lua.

Palavras-chave
Diabetes tipo 2; Ilhota endócrina; Células β; Estudos de associação em larga escala genómica, Polimorfismo de um nucleótido; Potenciador; Fatores de transcrição

Resumo

A diabetes tipo 2 (DT2) afeta mais de 300 milhões de pessoas em todo o mundo, causando complicações severas e morte prematura. Contudo, os mecanismos moleculares associados a esta doença são, atualmente, pouco conhecidos. DT2 é caracterizada, em parte, pela disfunção de ilhotas endócrinas pancreáticas, não havendo produção suficiente de insulina. Os recentes avanços em estudos de associação em larga escala genómica têm demonstrado uma clara associação entre polimorfismos de um só nucleótido (PSN) e D2T. Uma grande parte destas variantes estão localizadas em sequências não codificantes que coincidem com marcas epigenéticas de potenciadores e de sítios de ligação de fatores de transcrição essenciais para uma boa função e organização das ilhotas endócrinas. Os potenciadores são sequências não-codificantes que regulam a expressão dos seus genes-alvo, interagindo com os promotores em *cis.* A hipótese do presente projeto científico é demonstrar que os PSNs associados a D2T podem afetar os sítios de ligação dos fatores de transcrição e consequentemente, a atividade das sequências regulatórias potenciadoras, traduzindo-se em diferenças transcricionais do gene. A primeira abordagem para testar a hipótese centralizou-se em ensaios de transgénese em peixe-zebra. Cinco das dez sequências testadas foram consideradas potenciadoras em pâncreas endócrino. A segunda abordagem baseou-se no impacto das variantes nas sequências potenciadoras. Numa sequência, a atividade potenciadora foi afetada pela presença de uma variante num sítio de ligação de PDX1, um fator de transcrição importante no desenvolvimento do pâncreas. Como perspetivas futuras, irão ser testadas as sequências em células β humanas em cultura e identificar-se-ão os genes-alvo das sequências, por 4C, captura de conformação cromossómica circularizada. Este trabalho ajudará a compreender melhor a importância da presença de variantes em genoma não-codificante no desenvolvimento de DT2.

Keywords

Type 2 diabetes; Endocrine islet; β-cells; Genome wide association studies; Single nucleotide polymorphism; Enhancer; Transcription factor

Abstract

Type 2 diabetes (T2D) affects over 300 million people, causing severe complications and premature death, yet the underlying molecular mechanisms are largely unknown. This condition is partially characterized by endocrine pancreatic islet dysfunction, leading to insufficient insulin production. By now, genome-wide association studies have shown that some single nucleotide polymorphisms (SNPs) are associated to T2D. Part of these variants are located in non-coding sequences with marks for enhancer activity, and some of them overlap with binding sites of transcription factors (TFs) known to be required for proper endocrine pancreas function. Enhancers are non-coding sequences that regulate the expression of their target genes by interacting with their promoters in *cis*. Our working hypothesis is that T2D associated SNPs might impair TF binding, affecting the enhancer activity of the sequence, ultimately translating into transcriptional changes of the downstream genes. At first, to approach this hypothesis, we have performed *in vivo* transgenesis assays in zebrafish to test if sequences overlapping with T2D associated *loci* were enhancers. We found that five out of ten tested sequences are endocrine pancreas enhancers. Secondly, we analyzed the impact of the risk associated variant in the enhancer activity. We found that in one out of three sequences, the enhancer activity was disrupted by the presence of a single nucleotide modification in a putative binding site for PDX1, an important transcription factor in pancreas development. We further analyzed this sequence by dividing it in fragments, testing them for endocrine enhancer activity. These results lead us to conclude that most likely the loss of the PDX1 binding site is accompanied by the gain of a repressor binding site that might contribute to the inactivation of the tested enhancer. As future approaches, we will test the enhancer activity of the selected sequences in human beta cell lines and perform Circularized Chromosome Conformation Capture (4C-seq) to identify the enhancer's target genes. Overall this project will help to better understand the importance of non-coding variants in the development of T2D.

TABLE INDEX

FIGURE INDEX

List of abreviations

3C – chromosome conformation capture

bp – base pairs

ChIP – Chromatin immunoprecipitation

CREs – *Cis*-regulatory elements

CTCF - CCCTC-binding factor

ENCODE – Encyclopedia of DNA elements

FAIRE – Formaldehyde-assisted isolation of regulatory elements

FISH – Fluorescence *in situ* hybridization

FOXA2 – Hepatocyte nuclear factor 3-beta

GFP – green fluorescent protein

GTEx – Genotype tissue expression dataset

GTF – General transcription factors

GWAS – Genome Wide Association Studies

H2A.Z - Histone H2A.Z

H3K27ac – Histone H3 acetylated at lysine 27

H3K4me1 – Histone H3 monomethylated at lysine 4

H3K4me2 - Histone H3 dimethylated at lysine 4

HNF1β – Hepatocyte nuclear factor 1β

hpf – hours post-fertilization

ISL-1 – Insulin gene enhancer protein

LD – linkage disequilibrium

mRNA – messenger RNA

NEUROD1 - Neuronal differentiation 1 protein

NGS – Next-generation sequencing

NKX2.1 – NK2 homeobox 1 protein

NKX2.2 – NK2 homeobox 2 protein

NKX6.1 – NK6 homeobox 1 protein

NGN3 – Neurogenin 3

PAX4 – Paired box protein 4

PCR – polymerase chain reaction

PDX1 – Insulin promoter factor

PIC – Preinitiation complex

Pol II – RNA polymerase II

Ptf1a - Pancreas associated transcription factor 1a

RBPJ – Recombination signal binding protein for immunoglobulin kappa J region

RFX6 – Regulatory factor X6

SNPs – Single nucleotide polymorphisms

SOX9 – SRY-box 9

SST - Somatostatin

TFBS – Transcription factors biding sites

TF – Transcription factors

T2D – Type 2 Diabetes

TSS – Transcription start site

WT – wild-type

I. Introduction

1. Eucaryotic transcription and *cis*-regulatory elements

Eukaryotic transcription is an important process in which a RNA molecule is synthetized using DNA as template, in order to carry the information transcribed outside the cell nucleus (Hahn, 2004). Outside of the nucleus, the newly synthetized molecule, messenger RNA (mRNA), is translated into proteins, essential to cellular viability and function. All the cellular biological processes require a spatial and temporal regulation of gene expression and one of the key players for a proper protein-coding genes transcription is RNA polymerase II (Pol II), an enzyme that synthesizes mRNA (Butler & Kadonaga, 2002; Hahn, 2004), dependent from other DNA-specific-binding proteins (*Trans*-regulatory elements). These *trans*-regulatory elements bind to the promoter region and to *cis*-regulatory elements (CREs) of DNA, able to interact to each other by chromatin loops (Butler & Kadonaga, 2002; Hahn, 2004 ; Levine et al., 2014) and allowing the Pol II activity.

The promoter region is responsible for the transcription initiation, by Pol II (Juven-Gershon & Kadonaga, 2010). The core promoter allows the preinitiation complex formation (PIC) (Maston et al., 2006), being a decisive target, near the transcription start site (TSS) containing the TATA box, initiator element, downstream promoter element and motif ten element (Maston et al., 2006). These elements recruit the general transcription factors (GTF) for the PIC formation (Butler & Kadonaga, 2002; Frith et al., 2008; Lim et al., 2004). The proximal promoter is upstream of the core promoter and contains binding sites required for activators to initiate the gene transcription (Butler & Kadonaga, 2002; Maston et al., 2006). An important transcriptional activator is the mediator complex, which forms larger complexes with other structural proteins as cohesins (Poss et al., 2013), allowing the interaction between the promoter and enhancers by chromatin loops (Kagey et al., 2011).

At the promoter region, the eukaryotic transcription mediated by Pol II includes three phases: initiation, elongation and termination (Nechaev & Adelman, 2012).

Transcription is started when the initiation complex is recruited to the promoter region (Nechaev & Adelman, 2012). The GTFs, transcription factor (TF) IIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, bind to the core promoter and the PIC is formed (Cosma, 2002; Hahn, 2004). The binding of TFs will allow changes in the chromatin state. When PIC is formed, Pol II is recruited to TSS and mRNA synthesis begins (Hahn, 2004; Hampsey, 1998; Lee & Young, 2000 (Butler & Kadonaga, 2002).

After transcription initiation, most of the GTFs are released (Kaphingst et al.,2010) and the elongation factors are recruited. Pol II will add to the 3`end of the nascent mRNA one nucleotide at a time until the termination factors bind to the transcription complex. Then, Pol II is released and mRNA is processed (Hirose & Ohkuma, 2007; Ni et al., 2004). Although transcription is mostly centered at the promoter, promoters do not contain all the information required for the proper spatial and temporal regulation of transcription, being part of this information present in CREs. CREs are DNA sequences that contain specific recognition sites for TFs, repressing or enhancing the transcription of one specific gene, controlling gene expression (Butler & Kadonaga, 2002; Maston et al., 2006; Shibata et al., 2015). In addition, epigenetic modifications that alter chromatin structure, also contribute to gene transcriptional regulation, increasing its complexity (Müller & Stelling, 2009).

There are two types of CREs, the proximal and the distal. The proximal CREs are composed by promoters and their proximal regulatory elements. The distal CREs includes enhancers, silencers and insulators (Fig.1) (Bulger & Groudine, 1999; Maston et al., 2006; reviewed in Blackwood & Kadonaga, 2016).

CREs sequences can act long range, being located hundreds of kilobases (kb) away from the promoters that they interact with (Butler & Kadonaga, 2002).

Silencers are distal target binding sequences for *trans*-acting repressors resulting in transcription repression (Maston et al., 2006; Chen & Widom, 2005; Harris et al., 2005). Silencers can remodel chromatin (Heinzel et al., 1997) interfering with PIC assembly (Maston et al., 2006).

Insulators are boundary elements that block the action of neighbor regulatory elements of a specific gene, preventing the activation of the incorrect gene, often limiting regulatory landscapes.(Butler & Kadonaga, 2002; Maston et al., 2006). Insulators can disrupt enhancer-promoter interactions, inhibiting chromatin loops as described by Ali and and co-workers (2016). Besides enhancer blocking, they can act as a heterochromatin barriers, preventing a transcriptionally active euchromatin turn into inactive heterochromatin (Mutskov et al., 2002).

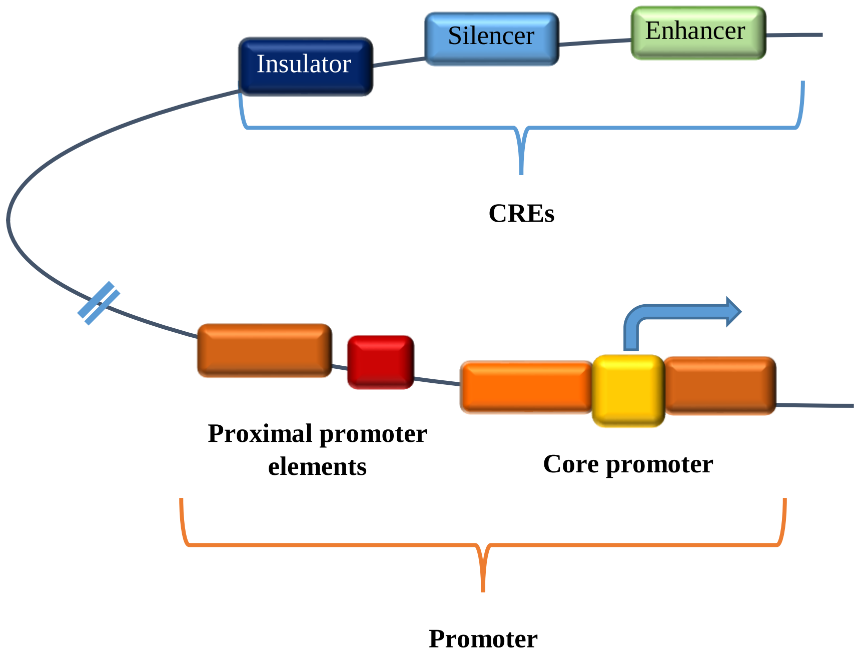


Figure 1. Representative gene regulatory region (Maston et al., 2006).

The human genome is composed by coding and non-coding DNA, both crucial to a proper function of cells and tissues. It is estimated that only 2% of the human genome corresponds to protein-coding regions, while 43% are transcribed non-coding regions and 55% are untranscribed regions (Fig. 2). The non-coding regions of the genome comprises CREs, contributing to several arrangements in transcriptional regulation, increasing the number and complexity of expression patterns. This complexity in expression patterns is an important factor in the appearance of new cellular functions (Barrett et al., 2012). This is one of the current explanations to why the increase of the complexity of organisms is accompanied by a lower protein-coding *per* DNA ratio (Shabalina & Spiridonov, 2004).

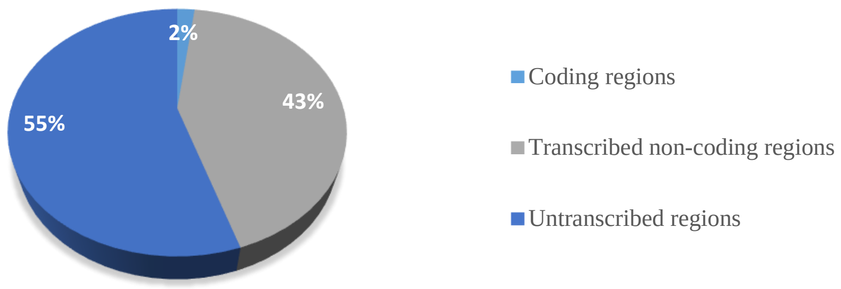


Figure 2. Proportions of the coding and non-coding sequences in the human genome (Shabalina & Spiridonov, 2004)

1.1.Enhancers

Enhancers are CREs that can increase the transcription level of a specific or a set of genes (Istrail & Davidson, 2005; Maston et al., 2006). They can be located downstream, upstream or within introns and exons of their target genes (Maston et al., 2006; Pennacchio et al., 2013). Their function is independent of their distance and/or orientation to the target gene, being difficult to predict which gene is controlled by an enhancer simply by sequence analysis (Atchison, 1988).

Enhancers contain specific transcription factors binding sites (TFBS) that interact cooperatively, recruiting co-activators and co-repressors, activating the promoter of the target gene (Maston et al., 2006; Mora et al., 2015). Different combinations of TFs determine the specificity of the enhancer. Additionally, different specific tissue enhancers can interact with the same gene promoter, composing the expression pattern of the gene. (Remenyi et al., 2004¸ Delic et al., 1991).

Being an important and fundamental DNA regulatory sequence, whose activity defines specific timings and locations for the transcriptional activity of genes, enhancers have arisen as elements of great potential and interest, being one of the best functional elements of the non-coding part of DNA described (Narlikar & Ovcharenko, 2009; Pennacchio et al., 2013). Currently, there are more than 80,000 putative enhancers identified in the human genome, using several genome-wide approaches and different techniques such as DNase I hypersensitivity, TFBS and chromatin marks assessment (Coppola et al., 2016).

One example of a long-range enhancer is the ZRS enhancer in the *LMBR1* gene that controls besides *LMBR1*, *SHH*, at one megabase (Mb) of distance from its promoter. *SHH* is expressed during limbs development, in the zone of polarizing activity. This specific zone is required for pattering and development of limbs. When ZRS presents single nucleotide variations, it acquires a gain of function, causing an ectopic expression of *SHH*, which leads to a congenital disease characterized by additional digits (preaxial polydactyl) (Lettice et al., 2002). This is also an excellent example showing that single nucleotide variations in regulatory elements might cause congenital abnormalities. (Lettice et al., 2002; Lettice et al., 2003) (Fig.4 - Andrey et al., 2017).

How enhancers can act at long range is poorly understood, however the most prevalent hypothesis is that enhancers can be placed near the promoter of their target genes by chromatin loops (Pennacchio et al., 2013; Vilar & Saiz, 2005; Andersson et al.,2015) (Fig 3).

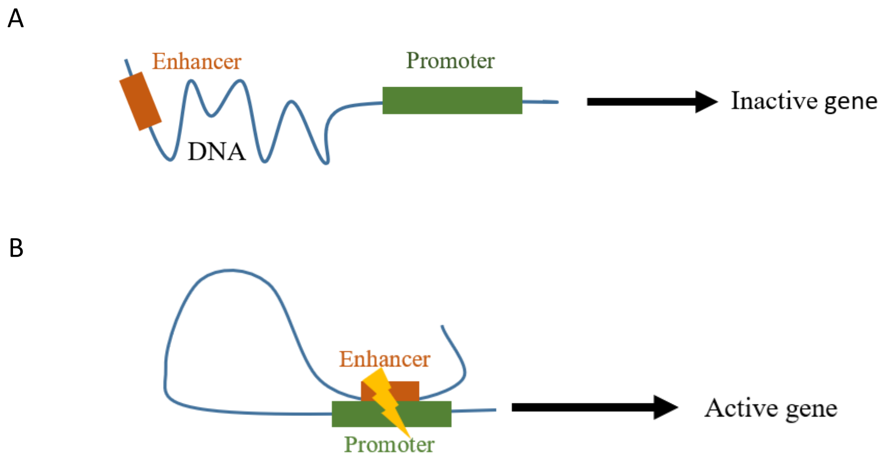


Figure 3 – The promoter-enhancer interaction regulates the gene expression. A. When physically distant, the promoter has no possibility to interact with the enhancer, resulting in a silent mode of gene expression; B. After a stimulus, the chromatin reorganizes and allows the interaction between promoter and enhancer, by proximity, by chromatin loops. Adapted from Andersson et al., 2015.

This was originally described in *Escherichia coli (E.Coli)* lactose operon, which is a regulatory bacterial element. In this particular case, a repressor binds in two *loci* by chromatin looping, blocking Pol II assessment in the DNA sequence and consequently the transcription (Mandal et al., 1990).

The chromatin loop theory settles in two main evidences. The first evidence came from techniques based on chromosome conformation capture (3C) (Dekker et al, 2002; (Kadauke & Blobel, 2009). This methodology is applied to determine the spatial organization of chromatin in a cell. Cells are fixed with formaldehyde, maintaining their nuclear structure, including physical interactions of genomic *loci*. DNA is cut by restriction enzymes and subsequently re-ligated. Fragments that are in the three-dimensional arrangement of the nucleus, close together, will be more frequently ligated, in contrast to fragments that remain faraway. Therefore, the distance of two *loci*, in the 3D distribution of DNA in the nucleous might be calculated by PCR based techniques. Two *loci* that are together in the 3D space will have a higher probability to be ligated and therefore will generate a higher PCR product when using quantitative PCR primers for these genomic locations (Kadauke & Blobel, 2009). There are other varieties that use next generation sequencing (NGS): 4C and 5C (Dostie et al., 2006; Simonis et al., 2006). The second evidence is based on the close proximity of enhancer and promoter regions in the cell nucleus, visualized by fluorescence *in situ* hybridization (FISH) (Pennacchio et al., 2013). This technique relies in probes against primary transcripts or DNA, detecting the proximal association of genomic regions (Kadauke & Blobel, 2009).

1.1.1. Enhancers identification and prediction

One of the main challenges of the enhancers study is their identification in the genome. NGS allied to computational biology has emerged as a good strategy, in part, to overcome this challenge (Pennacchio et al., 2013; Wang et al., 2013). Several approaches, such as chromatin immunoprecipitation (ChIP), DNaseI-digested chromatin (DNase hypersensitivity) and Formaldehyde-assisted isolation of regulatory elements (FAIRE) followed by NGS are now used as enhancer prediction tools (Bu et al., 2017).

Alternatively, enhancer identification can be performed based on comparative genomics by phylogenetic footprintings, exploring the fact that some non-coding enhancers are highly conserved between different species (Zhang & Gerstein, 2003). This strategy assumes that sequence conservation is an indicator of DNA functionality, therefore, conserved non-coding sequences are good candidates to be functional enhancers. Additionally, sequence conservation might help to identify functional orthologous enhancers, shedding light on the molecular mechanisms that might operate in these sequences. (Chatterjee et al.,2011; Fisher et al., 2006; Hare et al., 2008; McGaughey et al.,2009; Swanson et al., 2011).

Other approaches to identify enhancers must be explored, since not all enhancers show a high degree of sequence conservation among divergent species (Yang et al., 2015). Recently, it has been shown that specific chromatin epigenetic marks have been associated to enhancers, proving that chromatin signatures can be specific identifiers of enhancers. Thus, the epigenetic marks can be used as a great tool for prediction of these regulatory elements of the transcription in the human genome (Heintzman et al., 2007).

The epigenetic marks used to recognize a putative enhancer (Fig.4) are histone H3 acetylated at lysine 27 (H3K27ac) (Creyghton et al., 2010) and histone H3 monomethylated at lysine 4 (H3K4me1) (Heintzman et al., 2007). H3K4me1 is present in active and primed enhancers, allowing to distinguish enhancers and promoters (Heintzman et al., 2009). In contrast, H3K27ac is present when the enhancer is active, making the distinction between active and primed enhancers (Creyghton et al., 2010; Heintzman et al., 2009; Rada-Iglesias, 2018). The ENCODE project (Dunham et al., 2012), a consortium of many laboratories worldwide has described chromatin epigenetic marks in several tissues and cells lines genome wide. The available data from ENCODE have been extensively explained to predict regulatory functional elements, such as enhancers (Rosenbloom et al., 2012).

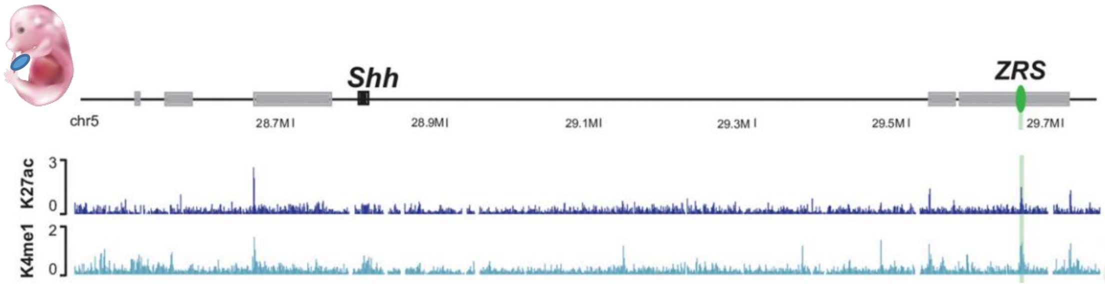


Figure 4 – ZRS enhancer in Lmbr1 locus. H3K27ac and H3K4me1 epigenetic marks profile showing a peak in enhancer locus. (Andrey et al., 2017).

One of the first associations between H3K4me1 and H3K27ac and enhancers was done by ChIP (Heintzman et al., 2007). ChIP is a technique based on crosslinking of DNA and proteins, followed by a specific antibody enrichment for a DNA-binding protein. The resultant DNA fragments are sequenced by NGS, being possible the identification of putative enhancers and TFs that might bind with enhancers, genome wide (Cuddapah et al., 2009; Hubner & Spector, 2010; Robertson et al., 2008; Robertson et al., 2007; Valouev et al., 2008). Additionally, Heintzman and colleagues have shown that sequences enriched for

H3K4me1 and H3K27ac function as enhancers, when tested for enhancer activity by reporter assays (Heintzman et al., 2007).

Besides addressing epigenetic marks, there are chromatin regions that are DNase I hypersensitive, that can also be an alternative strategy to identify enhancers (Dorschner et al., 2004). DNase I hypersensitivity assessment is based on the property of active CREs to be hypersensitive to cleavage by the endonuclease DNase I (Sullivan et al., 2015).

Another approach to predict enhancers is FAIRE. FAIRE, similar to DNA I hypersensitivity, detects open chromatin sites.(Song et al., 2011). This technique is based in biochemical differences between nucleosome bound DNA and nucleosome free DNA. Cells are crosslinked with formaldehyde, then they are lysed, sonicated and it is performed a phenol-chlorophorm DNA extraction. Crosslinking will fix DNA to nucleosomes, allowing that during phenol-chlorophorm DNA extraction, nucleosome bound and nucleosome free DNA will have different affinities to organic and aqueous phases, respectively (Giresi et al.,2007).

2. The vertebrate pancreas

The vertebrate pancreas forms from two different primordia from the foregut endoderm, the dorsal and the ventral bud (Pan & Brissova, 2016). The pancreas is constituted by an endocrine and exocrine/acinar component, having important roles in digestion and metabolism (Jennings et al., 2015). The endocrine compartment comprehends the hormone-expressing-cells (islets of Langerhans). These hormones are responsible for maintaining glucose homeostasis, controlling carbohydrate, lipid and protein metabolism. The exocrine compartment has a gastrointestinal function, containing digestive enzymes expressing-cells, that aid digestion by secreting these enzymes into the digestive tract (Habener et al., 2005; Jennings et al., 2015).

2.1.The endocrine pancreatic islet – islet of Langerhans

The endocrine pancreas is composed by small islets of hormone-expressing-cells scattered in the acinar tissue (Jennings et al., 2015). These hormone-expressing-cells are beta (β), alpha (α), epsilon (ε), delta (δ) and pancreatic polypeptide (PP) cells. β-cells produce insulin. α-cells are responsible for glucagon secretion and ε – cells ghrelin. Finally, the δ-cells secrets somatostatin and PP- cells pancreatic polypeptides (Sussel & Mastraci, 2013).

Each of these type of endocrine cells has its own precursor cell, that express a specific combination of TFs (Herrera et al., 2002) and their differentiation occurs during embryogenesis (Kulkarni, 2004).

2.2.Vertebrate pancreas development and transcriptional networks

Most of the knowledge about vertebrate pancreas development has been reached by knockout studies in mice, disrupting transcription factors involved in endocrine and exocrine pancreas formation (Habener et al., 2005) (Table 1).

Table 1. TFs knockout studies in mice showing consequences in pancreas development (Ahlgren et al, 1996; Ahlgren et al., 1998; Gittes et al., 1996; Lee et al., 1995; Naya et al., 1997; Murtaugh & Melton, 2003; Gu et al., 2011)

Transcription factor disrupted	Consequences
(Insulin promoter factor) PDX1	Pancreas agenesis
(Insulin gene enhancer protein) ISL-1	Death; lack of exocrine and islet cells differentiation
(NK2 homeobox 2 protein) NKX2.2	β-cells absence and α-cells reduction
(NK6 homeobox 2 protein) NKX6.1	β-cells inhibition
(Neuronal differentiation 1 protein) NEUROD1	Immature β-cells

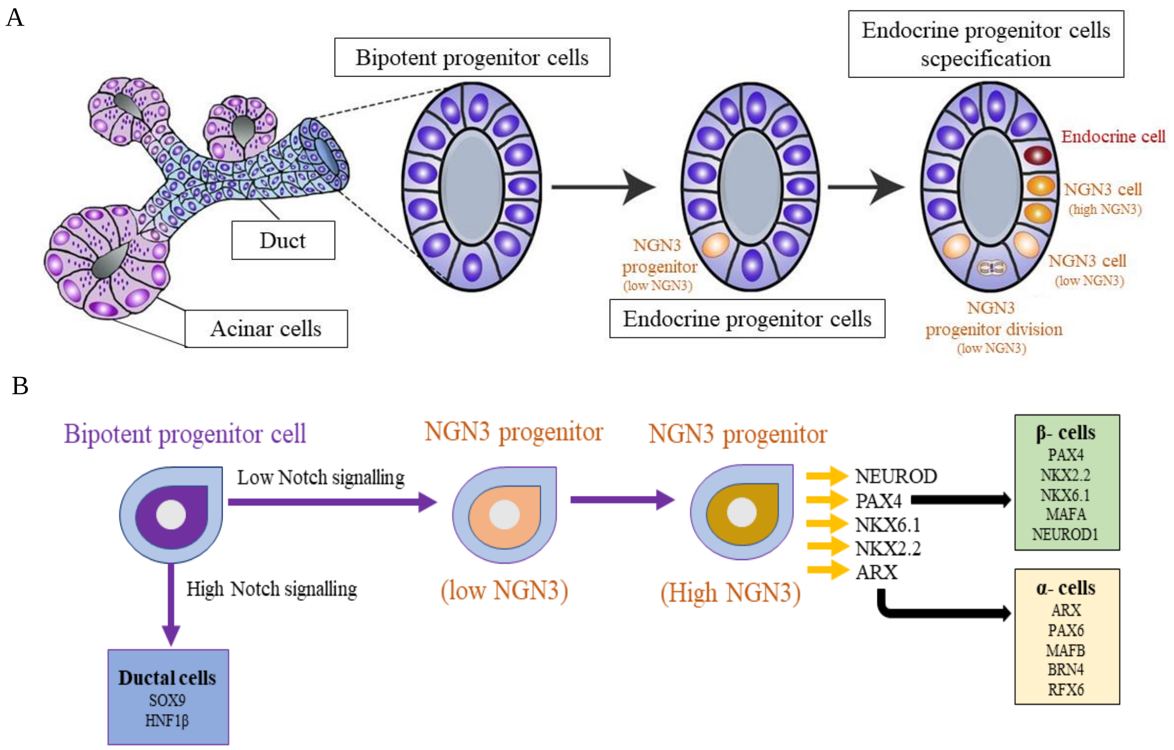


Figure 5 – Genetic lineage networks of pancreas development. (Adapted from Bastidas-Ponce et al., 2017)

During pancreas development, neurogenin 3 (NGN3) determines endocrine and exocrine fate, being expressed in a biphasic way in two different temporal waves of embryonic endocrine differentiation. In the first period of NGN3 expression (Fig.5 - A) occurs the primary transition of endocrine lineage and the second period of expression initiates right before the second wave (Fig.5 – B). The regulation of these levels is complex and not well established. However, currently, it is believed that the emerging expression of neurogenin 3 (NGN3) in bipotent progenitors (Fig.5-B) in the pancreatic epithelium inhibits Notch signaling and determines the fate of these cells as endocrine pancreas (Pan & Brissova, 2016; Habener et al., 2005; Murtaugh & Melton, 2003), while (Pan & Brissova, 2016; Habener et al., 2005; Murtaugh & Melton, 2003) high Notch signaling, in part, mediated by SRY-Box 9 (SOX9) and hepatocyte nuclear factor 1β (HNF1β) will determine a exocrine pancreas fate. (Bastidas-Ponce et al., 2017).

Two of the principal TF involved in endocrine pancreas development are homeobox protein ARX (ARX) and paired box protein 4 (PAX4). They start being co-expressed in NGN3 positive cells being more specific through its differential expression, during endocrine cells development. Cells that express a higher level of PAX4 will differentiate into β and δ-cells. In the other hand, the cells that express higher levels of ARX will be differentiated into α-cells (Bastidas-Ponce et al., 2017; Collombat, 2005).

The differentiation of α-cells will rely in multipotent pancreatic progenitor cells that express important TFs such as paired box protein 6 (PAX6), regulatory factor X6 (RFX6), POU Class 3 Homeobox 4 (POU3F4), hepatocyte nuclear factor 3-beta (FOXA2) and TF MafB (MAFB) (reviewed in Bramswig and Kaestner, 2011). For β-cell differentiation, PDX1 and NKX6.1 are the most important TF involved. They have as a direct target the *INSULIN* gene, being important not only for β-cell differentiation but also for a proper β-cell function (Ahlgren et al., 1996).

In NGN3 positive cells, NKX2.2 represses NEUROD, a TF present in pancreatic progenitor cells, generating α-cells and activates NEUROD to give rise to β-cells (Mastracci et al., 2013) .

Cell differentiation is followed by a functional maturation step, where cells acquire their function, the responsiveness to glucose. The two main required TFs for α and β-cells maturation are MAFA and MAFB. TF MafA (MAFA) expression is regulated by β-cells specific TFs NEUROD1, NKX6.1,NKX2.2, FOXA2, PAX6, RFX6 and GLIS Family Zinc Finger 3 (GLIS3) (Arda, Benitez, & Kim, 2013). This cluster of TFs together with PDX1 regulates the expression of *INSULIN* (Palanker et al., 2006; Taylor et al., 2013; Zhang et al., 2005).

After β-cells maturation, cells synthetize and secret insulin in response to glucose levels in blood plasma (Kulkarni, 2004). Insulin is a hypoglycemic agent, having the capacity to lower blood glucose levels, while glucagon counteracts the insulin action, stimulating glycogenolysis and gluconeogenesis (Jennings et al., 2015).

3. Cis-regulation and diseases

The sequencing of human genome has demonstrated that approximately 98% of total non-coding DNA presents marks for enhancer activity, suggesting that many of these sequences might be enhancers (Edalat, 2012; Hindorff et al., 2009; Pennacchio et al., 2013; Venter et al., 2009). It is reasonable to believe that variations in the sequences of these regulatory elements can result in transcriptional dysregulation of genes, phenotypic alterations and disease (Maston et al., 2006; Pennacchio et al., 2013). One example is the translocation in *β-GLOBIN* gene, with consequential thalassemias (Kleinjan & Coutinho., 2009). Thalassemias are caused by a disequilibrium of the levels of β-GLOBIN chains that transports hemoglobin in erythrocytes, due to mutations in one or more *GLOBIN* genes. A translocation in these genes removes *cis*-regulatory sequences, affecting their expression and consequently a disequilibrium in the expression of *β-GLOBIN* genes (Pennacchio et al.,2013).

Mutations in TFBSs within enhancers can result in misregulation of target genes having as consequence the loss of a normal cell type or tissue (Lee & Young, 2013). The recent advances in the study of transcriptional *cis*-regulation have led to a better understanding of dysregulation of gene expression in several human diseases (Lee & Young, 2013). One example of genetic variations are single nucleotide polymorphisms (SNPs), that have been identified in several whole-genome sequencing projects and computational analysis (Altshuler et al., 2012; Peters et al., 2012; Yngvadottir et al., 2009). These kind of variations are mostly located in non-coding regions and some can be specifically associated to human traits and complex diseases (Fig 6) (Lee & Young, 2014; Ernst, 2011; Hindorff et al., 2009; Maurano et al., 2012; Zhang & Lupski, 2015).

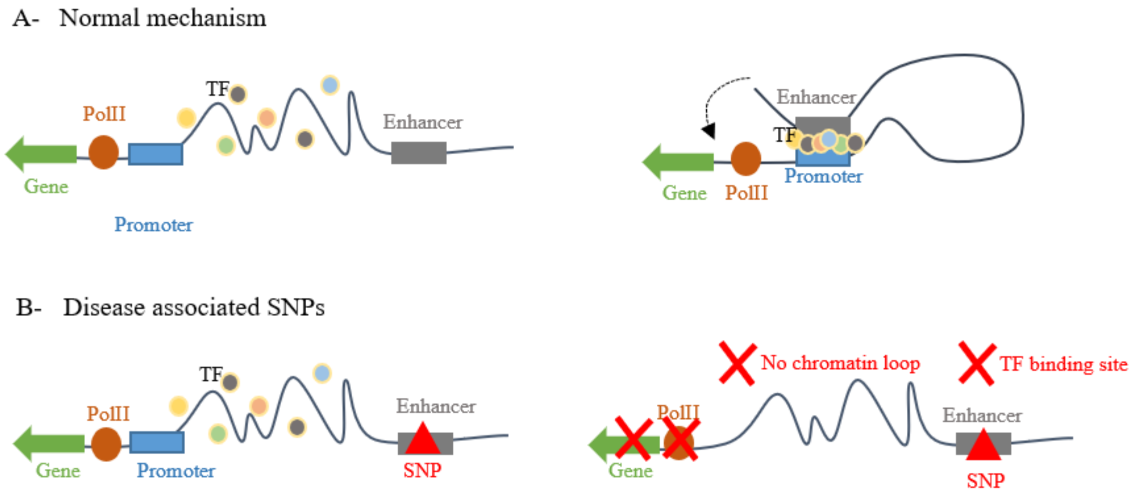


Figure 6 – A. Normal situation – The chromatin loop is formed and the distally TF bind to enhancer in order to activate the transcription, sideways with Pol II; B. Disease associated SNPs – The chromatin loop is impaired, the TF binding site disrupted, and the transcription is affected. Adapted from Heuvel et al., 2015

The association between SNPs and diseases or traits is possible by performing genome-wide-association studies (GWAS). This type of studies access thousands of SNPs in a large sample of individuals to establish an association reliable between common genetic variants with diseases and traits (Schaid et al., 2018). Therefore, GWAS provide statistical evidences that the presence of certain SNPs in non-coding DNA can increase disease susceptibility (Hindorff et al., 2009; Li et al., 2014; Pennacchio et al., 2013; Zhang & Lupski, 2015).

A combined analysis of GWAS and marks for enhancer prediction, such as DNase hypersensitivity, chromatin epigenetic marks, FAIRE and ChIP, many of them explored in large consortiums as the ENCODE project, allowed to infer that SNPs associated to disease may be often located in predicted enhancers. (Ahonen et al., 2009; Degner et al., 2012; Trynka et al., 2013). Maurano and coworkers observed that within 5134 SNPs associated with 654 phenotypes, 77% overlap with DNase hypersensitivity region (Maurano et al., 2012). In addition, Hindorff and co-workers and Li and colleagues have detected that 88% of disease associated variants are located in non-coding regions (Hindorff et al., 2009; Li et al., 2012).

One example of a disease associated SNP is preaxial polydactyly, as referred before. It is described that this disease is caused by mutations in one distal enhancer, ZRS, of the target gene *SHH.* (Lettice et al., 2002; Lettice et al., 2003). The analysis of the putative TFBS conserved in ZRS sequence showed consensus binding sites for homeobox protein CDX-1 (CDX), meis homeobox 1 (MEIS1) and SOX9, which are TFBS involved in limb development (van den Akker et al., 2002).Three of the six mutations showed to be the cause of disruption of CDX binding site, contributing to the disease (Evans, 2007; Lettice et al., 2003).

Another example is Hirschsprung disease, where the *RET* gene is affected, by the presence of three SNPs in MCS enhancer, in intron 1. Interestingly, one of the three mutations in *RET* reduces their expression directly by affecting SRY box 10 (SOX10) binding, being the other two mutations an indirect contribution (Emison et al., 2005; Fisher et al., 2006; Grice et al., 2005; Sribudiani et al., 2011).

Apart from the evidences from GWAS in the association of genetic variants to diseases, it is necessary to validate the putative functional impact of these variants on biological processes, the inherent molecular function and the pathways that can connect the variants to the disease (Li et al., 2014). To reach this aim, it is imperative the development of suitable assays including the use of *in vivo* models (Pennacchio et al., 2013; Zhang & Lupski, 2015).

3.1.Type 2 diabetes and *cis*-regulation

Type 2 diabetes (T2D) is a complex disease and one of the most common complex traits worldwide, affecting more than 300 million people. T2D is characterized mostly by the dysfunction of the endocrine pancreas (Fig 6), leading to insulin deficiency and loss of glucose homeostasis. However, the underlying molecular mechanisms are poorly understood (Alejandro et al., 2014; Pasquali et al., 2014; Sara, 2009; Chatterjee et al., 2017).

T2D has been associated to obesity, cardiovascular risk and hyperglycemia, caused by genetic susceptibility and environmental factors (Saxena et al., 2007). The environmental factors englobe lack of exercise, diet and aging. The aging factor is related to the β-cells decrease in proliferation capacity (Avrahami & Kaestner, 2012; Bhushan et al., 2013; Teta et al., 2005). The genetic mutations related to insulin insufficiency are rare and single genetic alterations does not seem to be the main cause of T2D, however, a considerable number of affected genes might contribute to the disease.

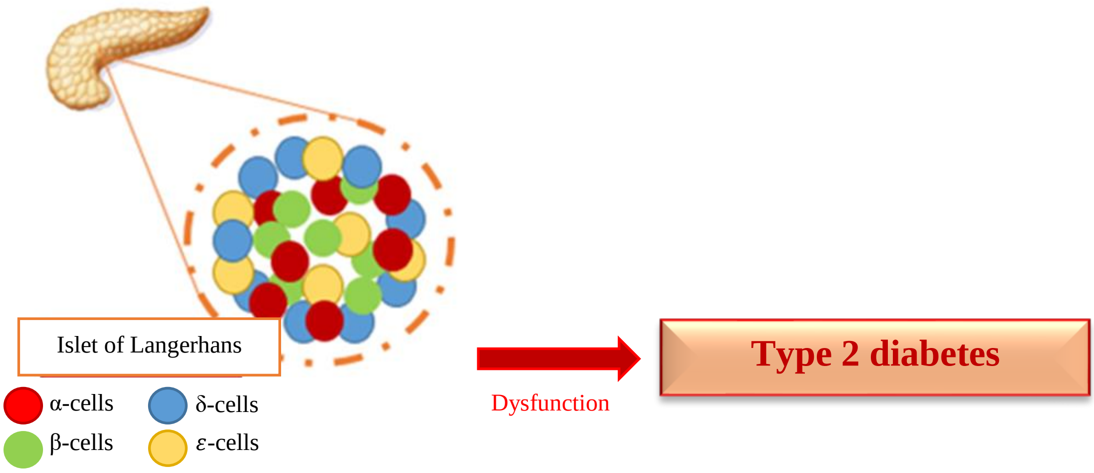


Figure 7 – Islet of Langerhans – endocrine cells. The endocrine pancreas dysfunction leads to T2D.

One of the most recent hypotheses is that the presence of SNPs in non-coding *cis*-regulatory sequences, such as enhancers of genes required for proper β-cell function can cause susceptibility to the disease. Supporting this hypothesis, several variants associated to T2D have been identified in non-coding *cis*-regulatory sequences in the past recent years (Morris et al.,2012).

3.1.1. Type 2 diabetes associated SNPs by genome wide association studies

Currently, several studies have shown that SNPs associated to a large number of diseases are enriched in non-coding *cis*-regulatory enhancers (Dunham et al., 2012; Hindorff et al., 2009; Maurano et al., 2012; Trynka et al., 2013). GWAS have been extremely important to identify and determine the frequency of these SNPs (Human Genome Sequencing Consortium ,2004), which can be analyzed by the presence of allelic variants in linkage disequilibrium (LD). It is assumed that two allelic variants are in LD when there is a non-random association of alleles at different genome locations (Mohlke and Scott, 2012).

Nowadays, there are approximately 88 established and published *loci* associated to T2D and 83 for glycemic traits (Mohlke & Boehnke, 2015).

Additionally, Pasquali and colleagues (Pasquali et al., 2014) have done a recent important contribution for the study of T2D genetics. In a large-scale study, the authors performed FAIRE-seq and ChIP-seq for epigenetic marks for enhancers activity (H3K4me1 and H3K27ac) to identify pancreatic enhancers. In addition, the authors performed ChIP-seq for islet TFs to predict their corresponding TFBS. The result was the identification of genomic sequences with *cis*-regulatory enhancer function active in the endocrine pancreas and targeted by specific islet TFs. Interestingly, the authors showed that SNPs associated to T2D were enriched in these enhancers. Thus, the T2D associated variants might have the potential to disrupt TFBS and islet enhancer activity, potentially causing a dysregulation of target genes. This way it was possible to integrate all the maps of epigenetic marks and TFBSs and create a complete and detailed dataset regarding the transcriptional regulation in pancreatic islets (Fig.8) (Pasquali et al., 2014).

As presented in the example above, there are genomic approaches that allow to predict active enhancers genome wide for a specific tissue like the endocrine pancreas. However, predictions should be validated by *in vivo* and *in vitro* assays. Combining the development of better predictions and sensitive and accurate methods of validation of enhancer activity it will be possible to better understand how genetic variants might impact in islet enhancer activity and consequently in islet function, resulting in human T2D.

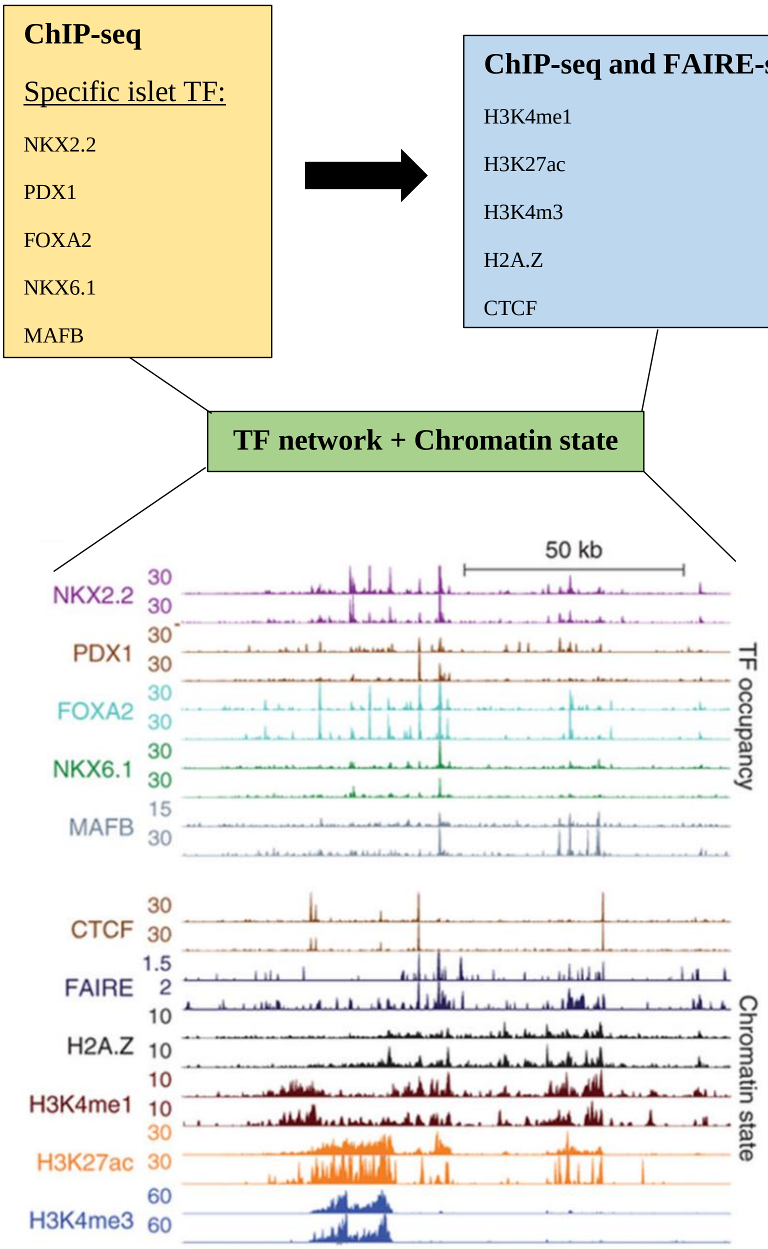


Figure 8 – TFBSs, active chromatin and histone modifications profile maps, showing the signals and the relation between the peaks. The islet specific TF showed ta pattern in binding in accessible chromatin sites. Adapted from Pasquali et al., 2014.

3.1.1.1.The case of rs163184 and rs13266634 T2D associated SNPs

a) rs163184

One of the many SNPs identified to be associated to T2D is the rs163184. This SNP is located in an intronic region of *KCNQ1* (potassium voltage-gated channel subfamily Q member 1) gene. Th*e* wild-type (*WT)* allele allows the binding of SP3 TF, directly, and LSDK1/KDM1A (lysine-specific histone demethylase) molecule, indirectly, via formation with SP3 TF complexes, stimulating the transcriptional activity of the gene. These bindings affect *CDKN1C* gene expression, being overexpressed, as demonstrated by Hiramoto and colleagues (Hiramoto et al., 2018).

CDKN1C is a negative regulator β-cells proliferation. Therefore, it can lead to a reduced insulin production, causing susceptibility to T2D (Hiramoto et al., 2018).

b) rs13266634

The SNP rs13266634 was identified as an established *locus* for T2D. This SNP is located in chromosome 8, in *SLC30A8* gene (Mohlke & Boehnke, 2015) that encodes for a zinc transporter, known to be required for zinc transport through the cell membranes and extracellular matrix (Faghih et al., 2014). The zinc flux is necessary to insulin secretion (Rutter, 2010; Xiang et al., 2008).

Many studies with single nucleotide variants in this gene have been done, however, the results are contradictory.

Flannick and colleagues demonstrated that 65% of the single nucleotide variants in *SLC30A8* resulted in a truncated protein and a reduced T2D risk (Flannick et al., 2014).

Other evidences supported the hypothesis that when the rs13266634 SNP is not present, the expression of *SLC30A8* increases the susceptibility to T2D (Mohlke & Boehnke, 2015; Xu et al., 2011). It has also been described that rs13266634 presence reduces the activity of the zinc transporter (Nicolson et al., 2009; Xiang et al., 2008). Other authors also demonstrated that the risk allele associated variant was associated to a lower insulin secretion and response (Horikoshi et al., 2007; Kirchhoff et al., 2008).

Studies in *Slc30a8* knockout mice also showed intriguing results. The phenotype was variable depending the gender and genetic background, suggesting that a perturbed zinc transporter will result in different biological responses (Flannick et al., 2014).

Further studies should be done to clarify the impact and the mechanism behind associated to this particular variant.

4. Models to test pancreatic enhancers

Putative enhancers can be validated by *in vivo* (Dorschner et al., 2004) or *in vitro* (Heintzman et al., 2007) reporter assays.

4.1. *In vitro* - cell lines

Cell lines are an animal-free tool that allow to study several physiological processes and pathologies. It is also possible to manipulate cell lines, by transfection, introducing reporter constructs, which contain reporter genes like luciferase, to test enhancer activity (Skelin et al., 2010).

However, cell lines can change their characteristics over time, showing chromosomal, genetic and protein expression abnormalities. (Skelin et al., 2010).

The main challenge of β-cell lines creation relies in the difficulty to mimic the same characteristics of the parental tissue, the insulin secretion and cell-to-cell interaction. One good example of a β-cell line is MIN-6, a transgenic mouse insulinoma cell line. It derives from transgenic C57BL/6 mice insulinomas that express an insulin-promoter/T-antigen construct, forming islet-like cells (Ishihara et al., 1993).

Many attempts were made to create a stable human β-cell line, however, the human lines created were not capable to secrete insulin, grow and were not stable in their function. Recently, it was established a promisor human β-cell line, from targeted oncogenesis in fetal pancreatic tissue that reproduces, in part, all the characteristics inherent of normal β-cells (Andersson et al., 2015; Ravassard et al., 2011; Scharfmann & Pechberty, 2014; Weir & Bonner-weir, 2011).

Enhancers have a tissue specific activity, however, cell lines are not a good system to demonstrate this tissue specificity, in contrast to *in vivo* models. Thus, the *in vitro* enhancer assays may not represent accurately the molecular and physiological cell mechanisms that might be active *in vivo*.

4.2. *In vivo* - Zebrafish

Zebrafish (*Danio rerio*) is a freshwater and a small bony fish. This species occupies shallow and highly vegetated regions and being omnivores, they feed on small insects, zooplankton and phytoplankton (Engeszer et al., 2007). Currently, the zebrafish is one of the most used model organisms in biomedicine and developmental biology, since genetic and embryological methods can be easily applied. It is cheap and easy to maintain in the laboratory, reproduces widely all year and it is possible to collect hundreds of eggs in one week. Furthermore, zebrafish reach the sexual maturity at 2-3 months, being appropriate for selection experiments and the creation of stable transgenic lines. Besides these advantages, the zebrafish embryos are transparent, which allow to follow the embryo development through time (Fig 9 - (Kimmel et al., 1995) (Grunwald & Eisen, 2002 ; Amsterdam & Hopkins, 2006) .

In the zebrafish genome, it has been described more than 20,000 genes and 69% of the these genes have orthologues in human (Howe et al., 2013; Tiso et al., 2009) . Besides the public availability of the zebrafish genome sequence, there are many other tools that are also available such as transgenic and mutant lines (https://zfin.org/) (Howe et al., 2013).

All these characteristics makes the zebrafish a perfect model organism to study several diseases in laboratory (Engeszer et al., 2007; Seth et al., 2013).

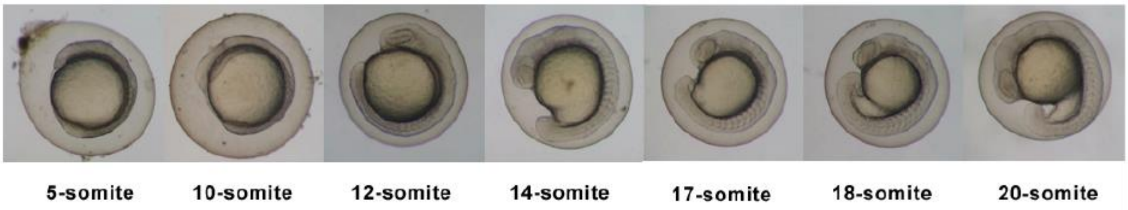


Figure 9 –Zebrafish developmental stages in segmentation period (10-24h). The embryos are transparent, being possible follow all the developmental stages. Based on Kimmel et al., 1995.

4.2.1. Zebrafish pancreas development

The zebrafish pancreas has two important compartments: an exocrine and an endocrine compartment (Tiso et al., 2009). The exocrine compartment comprises the acinar cells that produce digestive enzymes and the endocrine compartment corresponds to the islets of Langerhans, where the hormones are secreted to the plasma, regulating the blood glucose levels (Prince et al., 2017).

The formation of the zebrafish endocrine islet begins at 24hpf (Hours post-fertilization) and it is positioned dorsally to the yolk (Fig 10). At 48hpf, the zebrafish larvae have already an endocrine islet composed by insulin and somatostatin cells, bounded to glucagon and PP cells (Biemar et al., 2001; Huang et al., 2001; Tiso et al., 2009) (Fig.10).

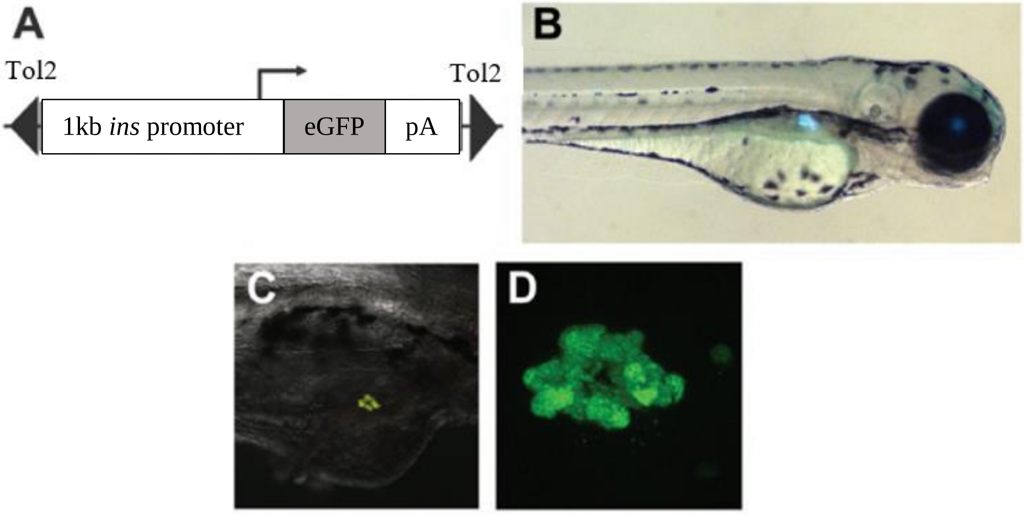


Figure 10 – A. It was constructed an plasmid vector containing an insulin promoter and Tol2 elements being possible the GFP expression in β-cells, when integrated in the genome; B. GFP expression in β-cells in 3 day old embryo; C. Confocal image (bright field) showing GFP expression in β-cells in 10 day old larvae; C and D. Zoomed confocal image of the endocrine islet at 10 year old showing the endocrine islet domain, regarding β-cells (Huang et al., 2001).

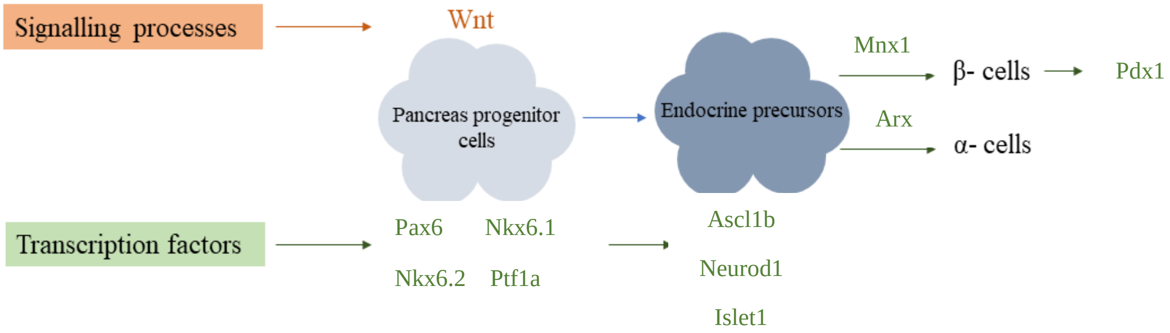


Figure 11 – Zebrafish endocrine pancreas principal development lineages. Adapted from Prince et al., 2017

Before β-cells differentiation, there is expression of mRNA from four crucial transcription factors, Pax6, Nkx6.1, NK6 homebox 2 protein (Nkx6.2), and pancreas associated transcription factor 1A (Ptf1A), being considered the cell progenitors of the pancreas, giving rise to all differentiated endocrine cells (Prince et al., 2017). The specific endocrine precursors are Isl-1, Neurod1 and achaete-scute family bHLH transcription factor

1b (Ascl1b), being responsible for differentiation of endocrine pancreatic islets (Biemar et al., 2001; Delporte et al., 2008; Parsons et al., 2009).

Pdx1 and Nkx6.1 have a crucial function in pancreas development and β-cells maturation, being expressed in differentiated β-cells (Kimmel et al., 2015). Recently, it has been described that Pdx1 is important for glucose metabolism (Jörgens et al., 2015) and activation of *INSULIN* expression in specific reporter assays (Menting et al., 2014).

The understanding of molecular pathways and the morphological changes in endocrine islet in zebrafish is possible due to the discover of specific biomarkers (Schiavone et al., 2014) allowing the development of reporter genes constructs and the creation of transgenic reporter lines, which allows the *in vivo* visualization of the pancreas. One example is the Tg(sst:mCherry*)* line, which labels *in vivo* the somatostatin (*sst)* cells, allowing the *in vivo* visualization of the endocrine islet (Fig.12).

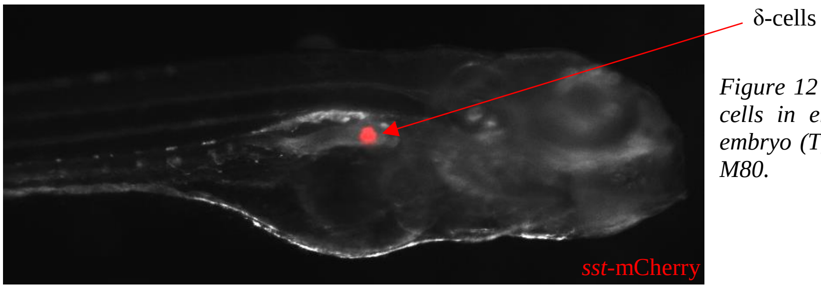


Figure 12 – mCherry *protein expression in δ-cells in endocrine pancreas of a 48hpf old embryo (Tg(sst:*mCherry*) reporter line). Leica M80.*

4.3. Zebrafish as a model to study type 2 diabetes non-coding variants

The zebrafish pancreas presents several similarities with human pancreas. The resemblances in function and structure allow the possibility to study the molecular mechanisms and the phenotypes associated to T2D (Kinkel and Prince, 2009).

T2D, is one of the most prevalent and challenging diseases in which genetics remains to be understood (Lu et al., 2018). Zebrafish has emerged as a potential tool to study T2D associated variants through different approaches, such as transgenesis assays.

4.3.1. Transgenesis assays

In zebrafish transgenesis assays to test enhancers, the sequence of interest to be tested is cloned in a vector, usually a transposable element to facilitate transgenesis, containing a reporter gene. The reporter gene usually encodes a fluorescent protein whose expression can be visualized *in vivo*. The vector is constructed in order to locate the sequence to be tested upstream of a minimal promoter and the reporter gene, which will allow the expression of the reporter gene when the cloned sequence acts as an enhancer (Narlikar & Ovcharenko, 2009). The vector is microinjected in one-cell stage zebrafish embryos and integrated randomly in to the genome. The expression pattern generated by the *in vivo* reporter gene allows the characterization of the tissue specific activity of the tested enhancer (Bessa et al., 2009; Soboleski et al., 2005; Kawakami, 2007) (Fig 13).

Transposons are mobile DNA sequences flanked by terminal repeats and an encoded enzyme, transposase, that when recognize specific DNA sequences in the genome, can activate the capacity of replication in the same genome (Plasterk, 1993). Tol2 is a transposon, being the most used system to create transgenic zebrafish lines, once has a high percentage of integration and transmission through generations (Kawakami et al., 2004).

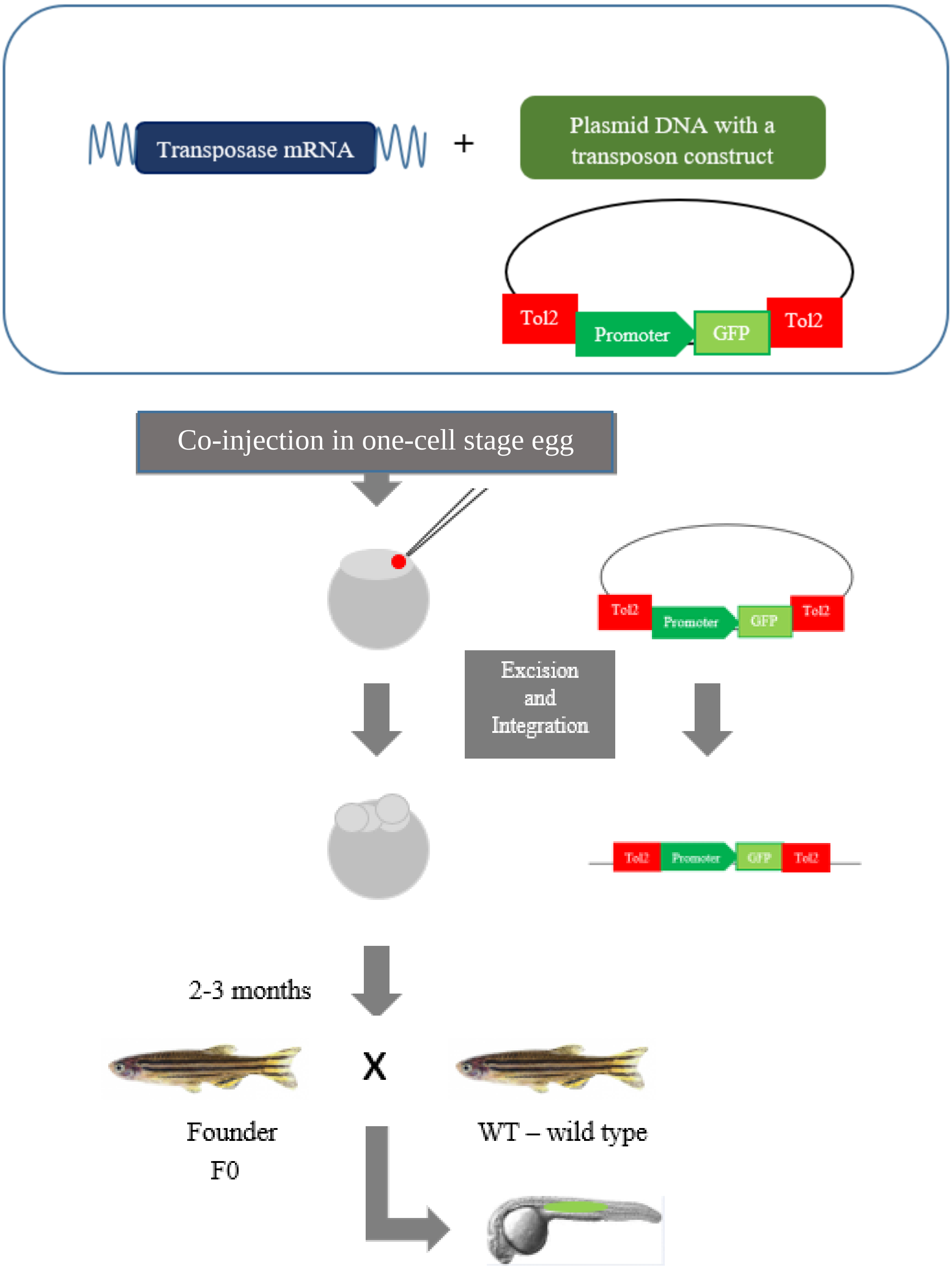


Figure 13 – Transgenesis in zebrafish. The transposase mRNA and a DNA plasmid containing Tol2, a promoter and a reporter gene (GFP – green fluorescent protein) are co-injected in one-cell stage egg. The construct is excised from de plasmid, allowing the integration in the zebrafish genome. This insertion is transmitted to the next generation (F1), when the original injected generation is crossed with WT (wild-type) fish. In this specific case, the promoter/enhancer is specific for spinal cord. Adapted from Kawakami, 2007. Images from http://www.zf-health.org/information/factsheet.html and http://dgallery.s3.amazonaws.com/zebrafish.png.

5. Hypothesis and main objectives

The emerging of large-scale studies and the total sequencing of human genome has provided important biological tools to identify *cis-* regulatory regions and to predict putative enhancers that might be fundamental for a good function of specific genes associated to several diseases. SNPs located in these regulatory regions can cause susceptibility to that kind of diseases, such as T2D. However, the study of putative enhancers is poorly known, due to the lack of investigations and validations *in vivo*.

The working hypothesis of this study is that T2D associated SNPs might impair TFBS important for endocrine enhancer activity in endocrine islets, contributing for disease susceptibility.

The aims are:

a) Identify human putative enhancers that overlap with described SNPs associated to T2D;
b) Test the enhancer potential of the identified sequences using *in vivo* reporter assays in zebrafish;
c) Analyze the impact of the presence of SNPs in the enhancer activity, for sequences validated as enhancers.

Overall, this work will allow us to better understand the impact of non-coding variants in endocrine enhancers, which can help to understand the molecular and genetic mechanism behind the development of T2D.

II. Materials and methods

1. SNPs and putative enhancers selection

The putative enhancers were selected based on Mohlke & Boehnke (2015) and Pasquali et al. (2014) GWAS bioinformatics data. The 176 sequences overlapping with T2D associated SNPs (Mohlke & Boehnke, 2015) were analyzed in UCSC Genome Browser (GRCh36/hg18) (https://genome.ucsc.edu/), using Pasquali and colleagues data and ENCODE data (Rosenbloom et al., 2012; Pasquali et al., 2014) for the presence of epigenetic marks in histones (H3K4me1 and H3K27ac). ENCODE data was used for different cell lines: Gm12878, a lymphoblastoid cell line, H1ES, a human embryonic stem cell line, HMEC, human mammary epithelial cells, HSMM, human skeletal muscle cells and myoblasts, HUVEC, a human umbilical vein endothelial cell culture, K562, lymphoblasts from bone marrow, NHEK, normal human epidermal keratinocytes and NHLF, human lung fibroblasts (Rosenbloom et al., 2012). The analysis was refined using data obtained from human pancreatic islets (Pasquali et al., 2014). The islet samples were analyzed by FAIRE and ChIP using the following marks: H3K4me1, H3K4m3, H3K27ac, CTCF (CCCTC-binding factor), H2A.Z (Histone H2A.Z), PDX1, MAFB, NKX6.1, FOXA2 and NKX2.2 (Pasquali et al., 2014).

Additionally , the selected sequences were also explored in Islet Regulome Browser, (http://www.isletregulome.org/isletregulome/) (Mularoni et al, 2017).

The analysis of these data resulted in a list of ten putative enhancers and overlapping SNPs associated to T2D (Table 2).

2. Primers design and PCR amplification

The PCR reactions were performed using the proofreading i-MAX II *Taq* DNA Polymerase (iNtRON Biotechnology, Inc) in a final volume of 20µL, containing 2µL of 10X PCR Buffer, 0,75 µL of Forward and Reverse Primers (10 µM), 2µL of dNTP mix (10mM/each NTP), 13µL of nuclease-free water and 0.5 µL of i-MAX II DNA *Taq* polymerase. Amplifications were performed at 94ºC for 3 minutes, followed by 35 cycles at 94ºC for 30 secs, 63-65ºC (depending on the melting temperature; Table 3) for 40 seconds, 72ºC for 1 minute/kb (depending of the size of the amplicon; Table 3) and a final extension at 72ºC for 10 minutes.

The PCR amplification was confirmed by analyzing the PCR product in 1% agarose gels stained with SYBR Safe (NZYtech). The ladder used was Gene Ruler 1kb (Thermo Scientific). The amplified product bands were excised from the agarose gel and purified using NZYGelpure kit (Nzytech), according to the standard protocol.

3. pCR8/GW/TOPO vector cloning and chemically competent bacteria transformation

The PCR products were TA cloned into pCR8/GW/TOPO (Invitrogen – Thermo Fisher Scientific). The vector is Gateway-adapted to provide an easy recombination of the PCR product of interest into any Gateway destination vector (Fig. 14).

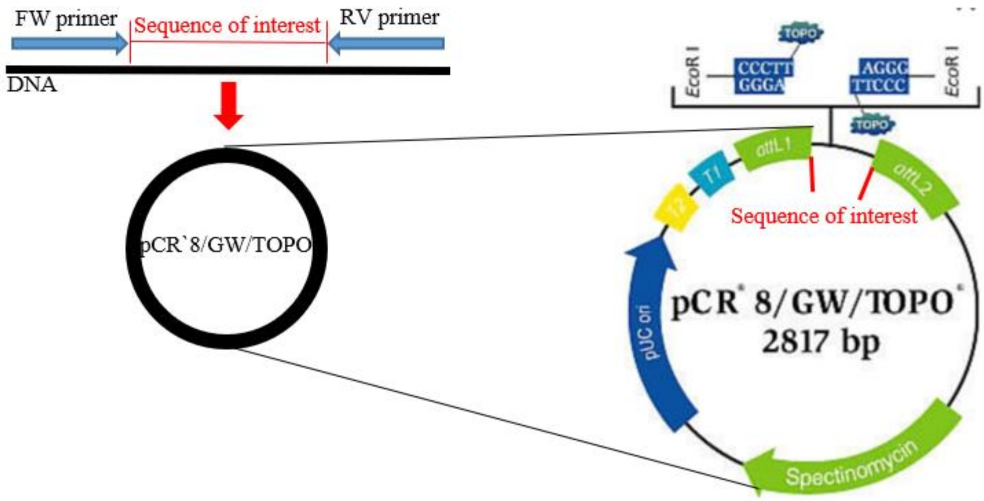


Figure 14 – The sequence of interest was amplified by PCR, using a set of specific primers. The amplified sequence was cloned into the commercial vector pCR8/GW/TOPO. The attL1 and attL2 flanks where the sequence is inserted, by TA cloning. The EcoRI enzyme restriction site is represented in the vector.

The cloning reaction consisted in a mix of purified PCR product (3 µL), salt solution, the commercial solution from the kit (1 µL), pCR8/GW/TOPO vector, diluted to 1/5 in dilution buffer (50& glycerol, 50mM Tris-HCl (ph=7.4), 1mM EDTA, 1mM DTT, 0,1% Triton X-100, 100 µg/ml bovine serum albumin (BSA) in ddH$_2$O) (1 µL), to a final volume of 6 µL, followed by a 30 minute of incubation at room temperature.

After incubation, 3 µL of each reaction were added into to 50 µL chemically competent bacteria and incubated in ice for 30 min. Cells were heat-shocked for 30 seconds at 42ºC and immediately transferred into ice for 2 min. 700 µL of Lysogeny broth (LB) were added the cells and incubated at 37ºC, for 1 hour with shaking. Cells were plated in LB agar plates containing 100 µg/mL spectinomycin and incubated overnight at 37ºC. The pCR8/GW/TOPO vector has spectinomycin resistance (Fig.14) to an efficient selection of the colonies.

Isolated colonies were picked from the plate and were transferred to LB containing spectinomycin and incubated overnight at 37ºC. Plasmid DNA was extracted using NZYMiniprep (Nzytech), according to the standard kit protocol. Plasmids DNA were digested with *EcoRI* enzyme (Anza; 8000 units) to confirm the insertion of the sequence of interest. This reaction consisted in 3 µL of miniprep product, 0,3 µL of the restriction enzyme to a final volume of 20 µL, followed by 2 hours of incubation at 37ºC. The digestion product was analyzed by 1% agarose gel electrophoresis and plasmids containing the PCR amplified sequence were sequenced for confirmation.

4. Z48 based vector recombination

The PCR product sequence contained in the pCR8/GW/TOPO was recombined into a destination vector, the Z48 vector (Fig 15). This vector has a minimal promoter upstream of the GFP (green fluorescent protein) reporter gene and a Z48 enhancer, that drives expression in the midbrain (Cebola et al, 2015). The vector contains a Tol2 transposon and ampicillin resistance.

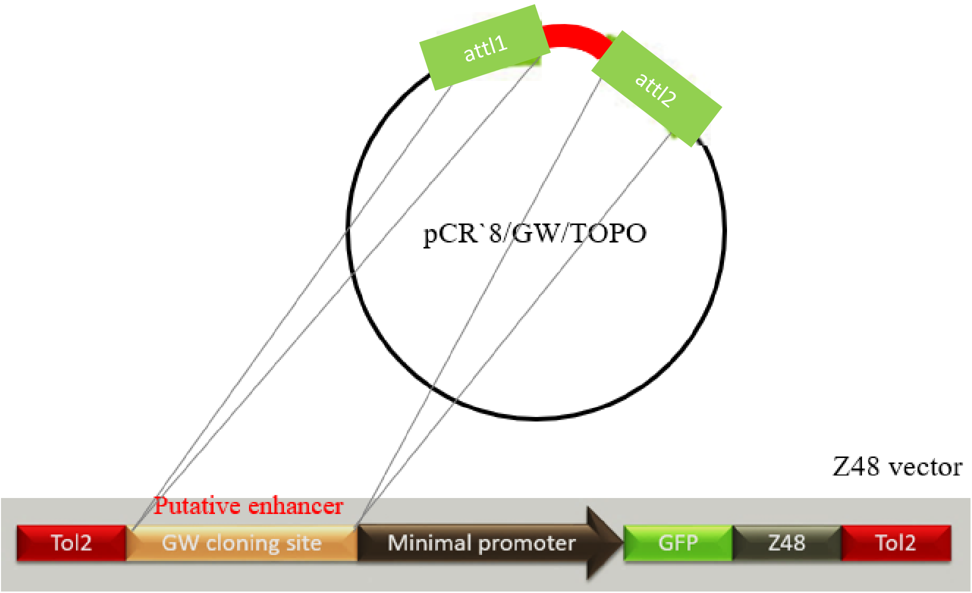


Figure 15 – The DNA fragment, after TOPO vector cloning, the entry vector, is recombined to a Z48 based vector, the destination vector, to test the enhancer activity of the sequence.

The recombination reaction was performed using a Gateway™ LR Clonase™ II Enzyme (Invitrogen - Thermo fisher scientific) in a final volume of 2,5µl, containing 1 µl of pCR8/GW/TOPO plasmid with the cloned sequence, 1 µl of Z48 destination vector, both at 50 ng/ µl and 0,5 µl of clonase enzyme. The reaction was incubated overnight, at 25ºC. To end the reaction, 0,25 µl of Proteinase K (Thermo fisher scientific) was added and incubate at 37ºC, for 10 min. The transformation was performed using chemically competent cells. Transformed bacteria were selected with ampicillin and isolated colonies were picked and grown to extract plasmid DNA.

5. Phenol/Chloroform DNA purification

Z48-plasmids were purified by a phenol/chloroform purification protocol: phenol/chloroform with isoamyl alcohol was added to the sample diluted in RNAse free water treated with DEPC (70 µl to a final volume of 100 µl), followed by vortex and centrifugation (13000 rpm), for phase separation (the DNA stayed in the aqueous upper phase).The aqueous phase was transferred to a RNAse free eppendorf and 100 µl of chlorophorm was added, vortexed and centrifuged as mentioned before. The aqueous phase was collected; 10 µl of sodium acetate (AcNa 3M) was added 100 µl of aqueous phase and 200 µl of ethanol (100%) to precipitate the sample, during 1h at -20ºC. Next, the sample was centrifuged 15 min (13000 rpm) at 4ºC and ethanol was removed. The pellet was diluted in 15 µl of DEPC treated water and quantified in a nanodrop (NanoDrop ND-1000 Spectophometer -Thermo Scientific).

6. Zebrafish maintenance and microinjection

6.1. Zebrafish husbandry

Zebrafish embryos were maintained according to standard protocols (Westerfield, 2000) at 28ºC in E3 medium (NaCl, KCl, $CaCl.2H_2O$ and $MgCl.6H_2O$). Zebrafish adults were maintained in a 14/10h photoperiod (light/dark), water temperature was kept at 26/27ºC and adults were fed three times a day.

6.2. Microinjection

6.2.1. Tol2 transposase mRNA synthesis

A Tol2 cDNA (complementary DNA) containing plasmid (Chien lab, 2007) was transformed using chemically competent cells and plated in LB agar plates with ampicillin (100ng/ µl). Three isolated colonies were transferred to LB broth with ampicillin and incubated overnight at 37ºC. Plasmid DNA of each colony was extracted and digested with *NotI* (Anza), to linearize the vector. The digestion reaction consisted in 1µl of enzyme, 4,5µl of DNA plasmid, 2 µl of buffer 10x and 12,5 µl of high pure water. *NotI* enzyme was used to linearize Tol2 cDNA vector, to transcribe the product, before purification by phenol/chloroform. The digestion product was analyzed by 1% agarose gel electrophoresis, extracted and purified by phenol/chloroform.

Tol2 RNA was transcribed *in vitro* using a mix reaction with 10 µl of transcription buffer 5x, 5 µl of DTT - dithiothreitol (50mM) and 5 µl of NTP mix (10mM A, 10mM U, 10mM C and 5mM G), followed by a 5 min incubation at 37ºC, then 5 µl of 5`CAP (25mM) were added. After 1 min of incubation, 12 µl of phenol/chloroform purified DNA was added, followed by an incubation of 1min. NZY Ribonuclease inhibitor was added in the reaction (2 µl), with 1min of incubation, then 1 µl of RNA polymerase (SP6), followed by 1h of incubation. This last step was repeated. All the incubations were made at 37ºC.

6.2.2. Sephadex and phenol/chlorophorm purification

RNA was purified using an adapted sephadex protocol. The piston of 1ml sterilized syringe was removed, in order to insert sterilized and DEPC treated aquarium filter 0,1mm. Next, sepahdex (Sigma-Aldrich) solution (1ml) (diluted in Tris- EDTA) was added to the top. Then, the syringe was placed in a 15ml falcon to discard liquid and centrifuged (4000 rpm) for 5 minutes, at 4ºC. The flow-through was discarded. To an efficient purification, the sephadex column, after the first centrifugation, has to reach at 0,6mm. Next, 50 µl of high pure water were added to column with an 0,5 eppendorf attached in the syringe, and the column was centrifuged again (4000 rpm) in the same conditions. It was checked if the volume in the 0,5 eppendorf was 50 µl. If so, the synthetized RNA was loaded in the column and was placed a new 0,5 eppendorf in the syringe. It was followed a new centrifugation (4000 rpm), with the same conditions. The RNA was collected to a new RNAse free eppendorf and placed in ice. To certify the total collection of the RNA, another 50 µl of high pure water was added in the column and centrifuged (4000 rpm). The water was added to the RNA and placed in ice. Next, phenol/chloroform purification was performed, the RNA quantified and stored at -80ºC.

6.2.3. Zebrafish breeding and embryos collection

One male and two females were placed in a breeding tank overnight separated by a divider. In the next morning, when the lights turned on, the divider was removed, and the fish started to reproduce. The breeding tanks have a net at the bottom, so the eggs can fall through the net and not be eaten by the adult fishes. Eggs were collected to proceed to microinjection. An *in vivo* reporter line of the endocrine pancreas was used, Tg (*sst*: mCherry). This reporter line has a *sst* promoter that drives expression of mCherry in δ-cells.

6.2.4. Microinjection at one-cell stage embryos

Microinjections were performed using a Narishige microinjector. One cell-stage embryos were injected with 2-5 nanoliters containing transposase mRNA, at 25ng/ µl, the Z48 enhancer assay vector (25 ng/ µl) and 0,05% phenol red. After microinjection, embryos were maintained in E3 (embryo medium) medium with 0,2mM 1-phenyl-2-thiourea (PTU 1x), to avoid pigmentation development.

7. Fixation and DAPI staining

At 48 hours for fertilization (hpf), microinjected embryos were selected by expression of GFP in the midbrain, dechorionated and fixed overnight in formaldehyde (4%, in PBS – phosphate-buffered saline – 1x). Then, embryos were washed with 500 µl of PBS-T (0,1% Triton in PBS-1x) 5 min, at room temperature, followed by permeabilization, with 500 µl of PBS-T 0,5%, during 30 min. After permeabilization, embryos were washed for the second time with PBS-T 0,1% and incubated with DAPI (1: 1000) in 200 µl of PBS-T 0,1%., for 4 hours at room temperature. After the incubation, embryos were washed 6 times (10 min each and the sixth, 30 min), with PBS-T 0,1%, at room temperature. PBS-T 0,1% was removed and 50% of glycerol in PBS 1x (NaCl, KCl, Na_2HPO4 and KH_2PO_4 in ddH_2O) was added. Microscopy slides were prepared using 50% of glycerol in PBS 1x. Embryos were analyzed in a confocal microscope (Leica - SP5II).

8. Immunohistochemistry

Embryos were fixed at 48 hpf, in formaldehyde (4%, in PBS – 1x) as previously described. Embryos were washed with 500 µl of PBS-T (0,1% Triton in PBS-1x) 5 min, two times. Permeabilization was performed with 500 µl of PBS-T 1%, for 2 hours, followed by a wash of 5 min with PBS-T 0,1%. Block with 200 µl of bovine serum albumin BSA in PBS-T (5%), for 1 hour. Embryos were incubated with an anti-Nkx6.1(F55A10; Hybridoma bank) primary antibody (1:50) in 200 µl of BSA+PBS-T (5%), for 48 hours, followed by 6 washes with PBS-T 0,1% (5 washes, 10 min each and 1 wash for 30 minutes). The embryos were incubated with DAPI (1: 1000) and an anti-mouse secondary antibody (Alexa fluor 647 nm – 1:800) in PBS-T at 4ºC, overnight. After overnight incubation, embryos were washed, as previously described and glycerol, at 50% in PBS 1x, was added. Microscopy slides were prepared using mounting medium (50% of glycerol in PBS 1x). Embryos were analyzed in confocal microscope (Leica - SP5II).

9. Predicting the impact of T2D risk variants in TFBS - JASPAR analysis

To predict how T2D risk variants could impact in the ability of TFs to bind to the respective sequences, JASPAR software was used. JASPAR uses a set of annotated position weight matrices for TFBS (Khan et al., 2018).

The *WT* and risk variant sequences were analyzed using the 719 specific position weight matrices for vertebrates, available in JASPAR. Sequences were analyzed and ranked by position-specific score matrix (Tables 5, 6 and Supplementary data). The relative score is a threshold score between 0 and 1 and is calculated by (score - min_score) / (max_score - min_score), being more accurate than the total score (http://jaspar.genereg.net/) (Sandelin, 2004).

III. Results and discussion

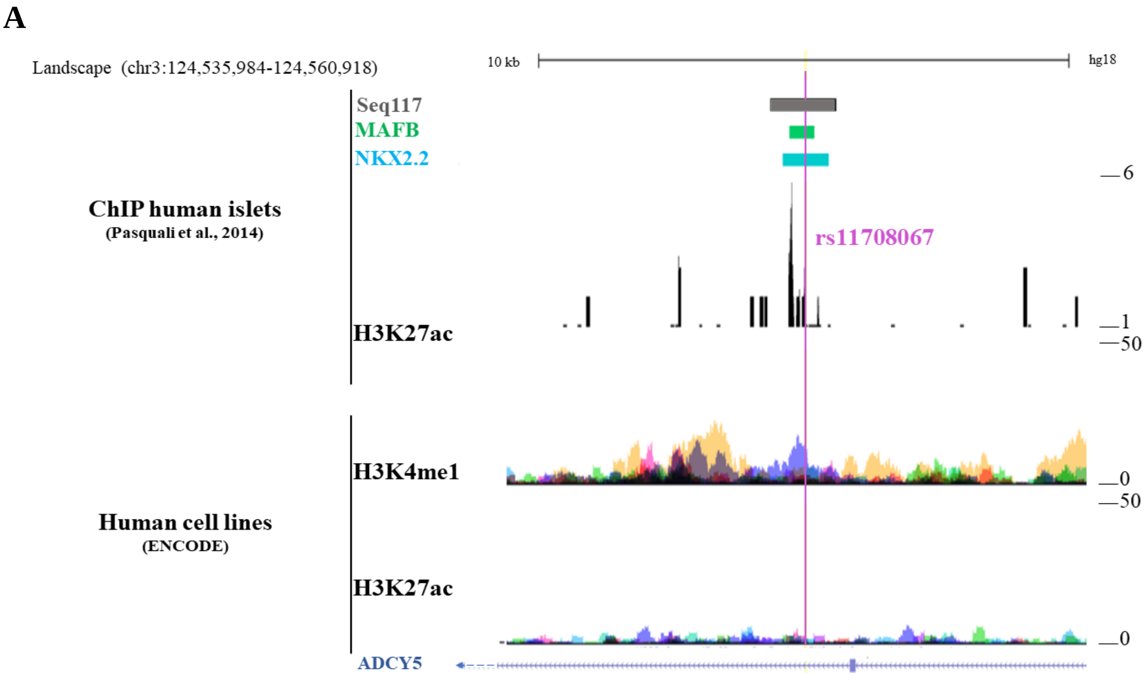
1. Selection of putative enhancers that overlap with SNPs associated to T2D

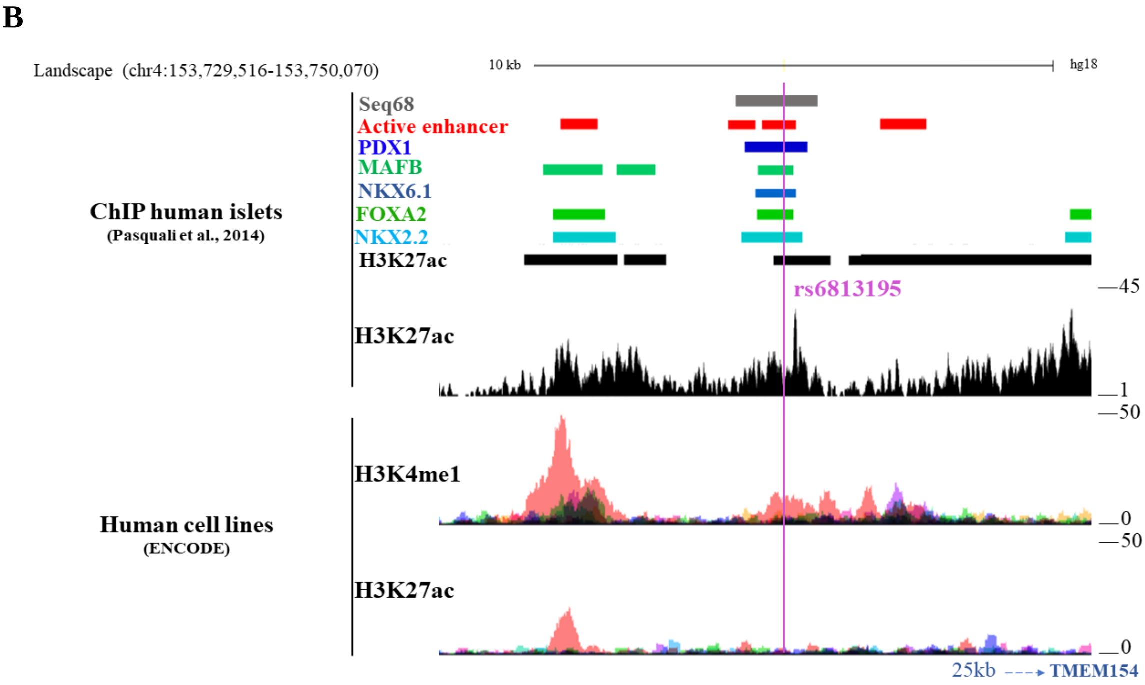
To identify putative enhancers that overlap with T2D associated SNPs we have combined datasets of ChIP - seq for epigenetic marks of enhancer activity, from different cell lines (H3K4me1 and H3K27ac; ENCODE PROJECT) (Rosenbloom et al., 2012) with a dataset of 176 described T2D associated SNPs (Mohlke & Boehnke, 2015). Once cell lines from different tissues (see materials and methods) might not be the best biological samples for the prediction of endocrine enhancers, we improved this analysis, using a ChIP-seq dataset from human endocrine islets (Pasquali et al., 2014). This dataset consisted in prediction of active endocrine enhancers by the presence of epigenetic marks for enhancer activity (H3K4me1, H3K4m3, H3K27ac, CTCF, H2A.Z). Additionally, an islet TFBS (PDX1, MAFB, NKX6.1, FOXA2 and NKX2.2) were identified by ChIP-seq. (Pasquali et al., 2014). The islet dataset can be consulted in Islet regulome browser (http://www.isletregulome.org/isletregulome/) (Mularoni et al., 2017).

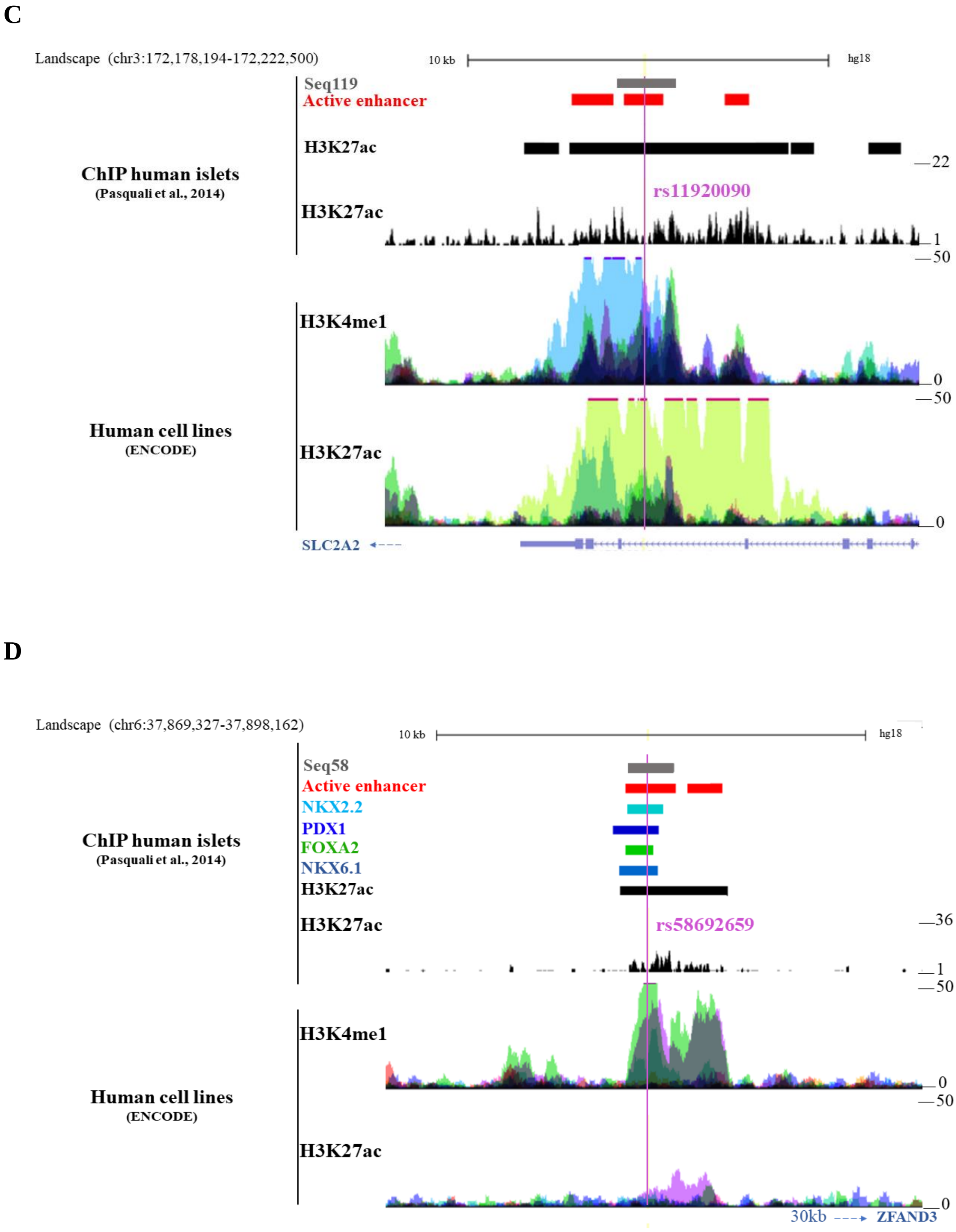
Ten sequences were selected by the consistent overlap of signal from the different mentioned datasets (Table 2).

Table 2 – Sequences selection, localization, TF associated and nearby genes. (Data from ENCODE UCSC Genome Browser and Islet Regulome Browser)

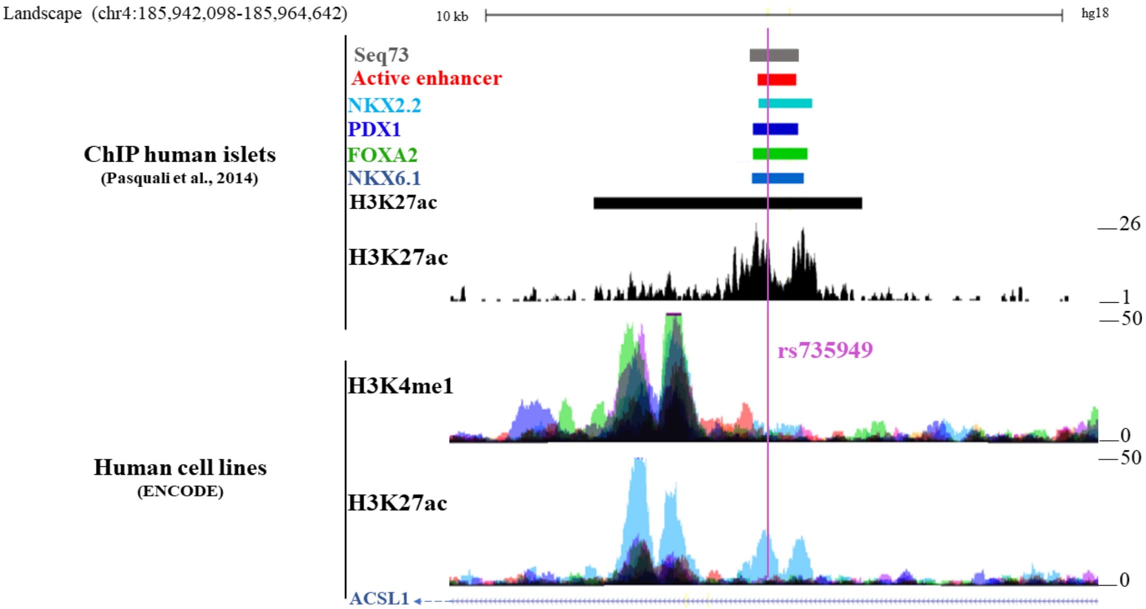
Sequemce	SNP	Coordinates (GRCh36/hg18)	TF binding sites	Nearby genes
Seq790	rs7903146 C > T	chr10:114,747,755-114,749,047	NKX2.2, FOXA2, NKX6.1, MAFB	TCF7L2
Seq219	rs2191349 G > T	chr7:15,030,232-15,031,383	NKX2.2, MAFB, PDX1, NKX6.1	DGKB
Seq117	rs11708067 G > A	chr3:124,547,836-124,549,066	NKX2.2, MAFB	ADCY5
Seq68	rs6813195 G > C	chr4:153,739,117-153,740,469	PDX1, NKX2.2, FOXA2, NKX6.1, MAFB	TMEM154
Seq119	rs11920090 A > T	chr3:172,199,576-172,201,167	FOXA2	SLC2A2
Seq132	rs13266634 C > T	chr8:118,251,516-118,256,576	NKX2.2, FOXA2	SLC30A8
Seq58	rs58692659 C > A	chr6:37,883,211-37,884,278	PDX1, NKX2.2, FOXA2, NKX6.1	ZFAND3
Seq72	rs72695654 G > T	chr4:185,953,124-185,953,774	PDX1, NKX2.2, FOXA2, NKX6.1	ACSL1
Seq73	rs735949 T > C	chr4:185,952,953-185,953,787	PDX1, NKX2.2, FOXA2, NKX6.1	ACSL1
Seq460	rs4607517 G > A	chr7:44,201,930-44,202,603	NKX2.2	YKT6; GCK


A
Landscape (chr3:124,535,984-124,560,918)
10 kb
hg18
Seq117
MAFB
NKX2.2
ChIP human islets
(Pasquali et al., 2014)
rs11708067
H3K27ac
6
1
50
H3K4me1
Human cell lines
(ENCODE)
0
50
H3K27ac
0
ADCY5

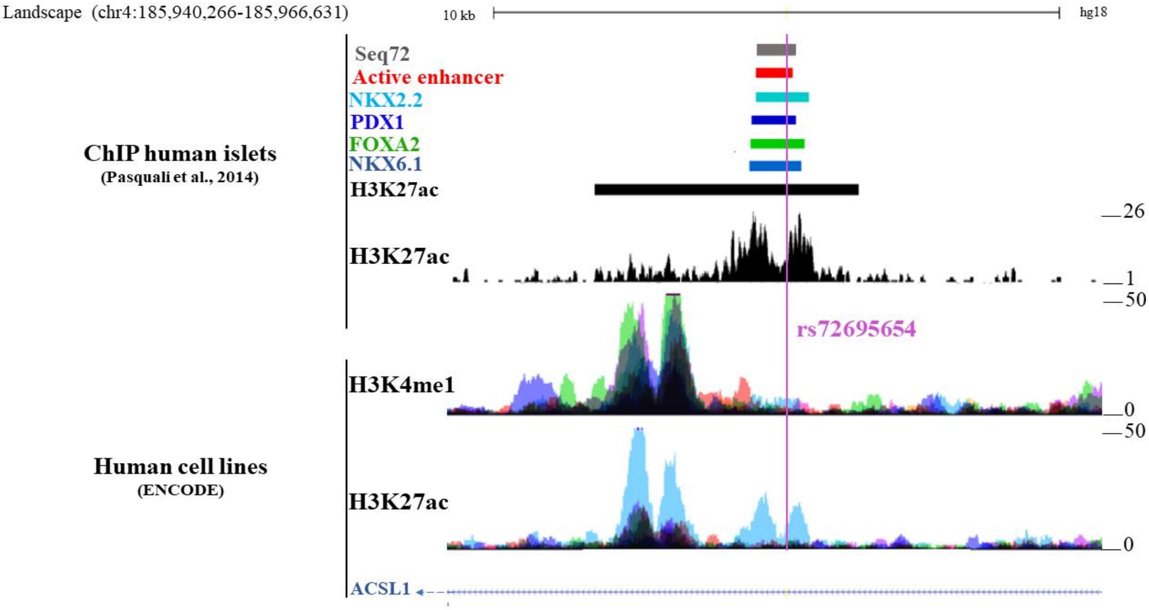

B
Landscape (chr4:153,729,516-153,750,070)
10 kb
hg18
Seq68
Active enhancer
PDX1
MAFB
NKX6.1
FOXA2
NKX2.2
H3K27ac
ChIP human islets
(Pasquali et al., 2014)
rs6813195
45
H3K27ac
1
50
H3K4me1
Human cell lines
(ENCODE)
0
50
H3K27ac
0
25kb
TMEM154


C
Landscape (chr3:172,178,194-172,222,500)
10 kb
hg18
Seq119
Active enhancer
H3K27ac
ChIP human islets
(Pasquali et al., 2014)
22
rs11920090
H3K27ac
1
50
H3K4me1
0
50
Human cell lines
(ENCODE)
H3K27ac
0
SLC2A2
D
Landscape (chr6:37,869,327-37,898,162)
10 kb
hg18
Seq58
Active enhancer
NKX2.2
PDX1
FOXA2
NKX6.1
H3K27ac
ChIP human islets
(Pasquali et al., 2014)
H3K27ac
rs58692659
36
1
50
H3K4me1
0
50
Human cell lines
(ENCODE)
H3K27ac
0
30kb
ZFAND3

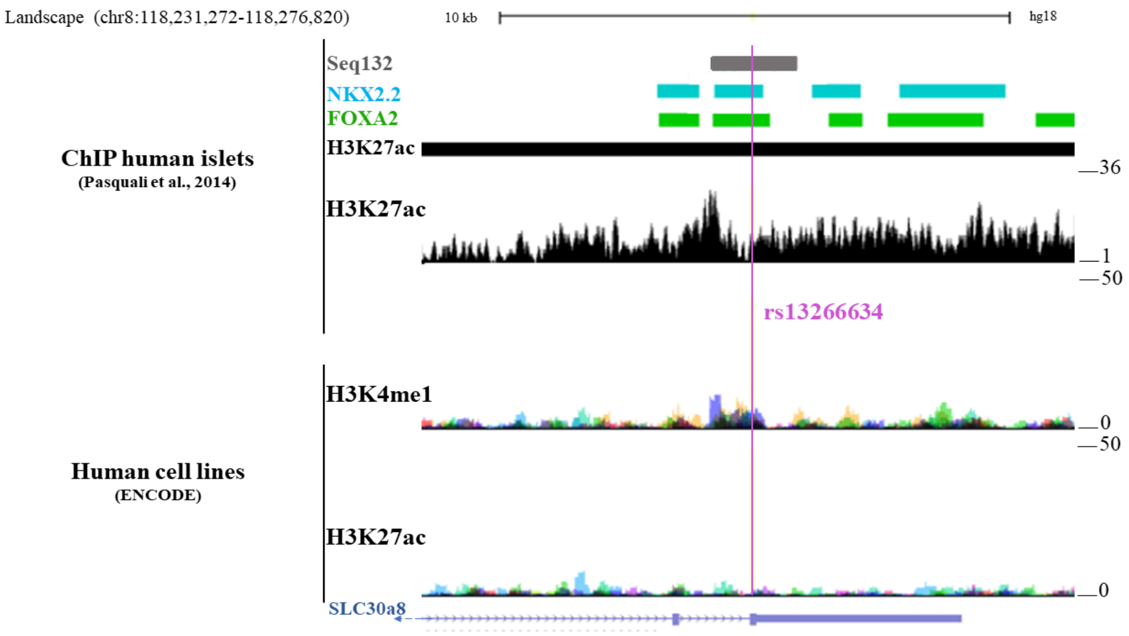
E


Landscape (chr4:185,942,098-185,964,642)
10 kb
hg18
ChIP human islets
(Pasquali et al., 2014)
Seq73
Active enhancer
NKX2.2
PDX1
FOXA2
NKX6.1
H3K27ac
H3K27ac
26
1
50
rs735949
Human cell lines
(ENCODE)
H3K4me1
0
50
H3K27ac
0
ACSL1

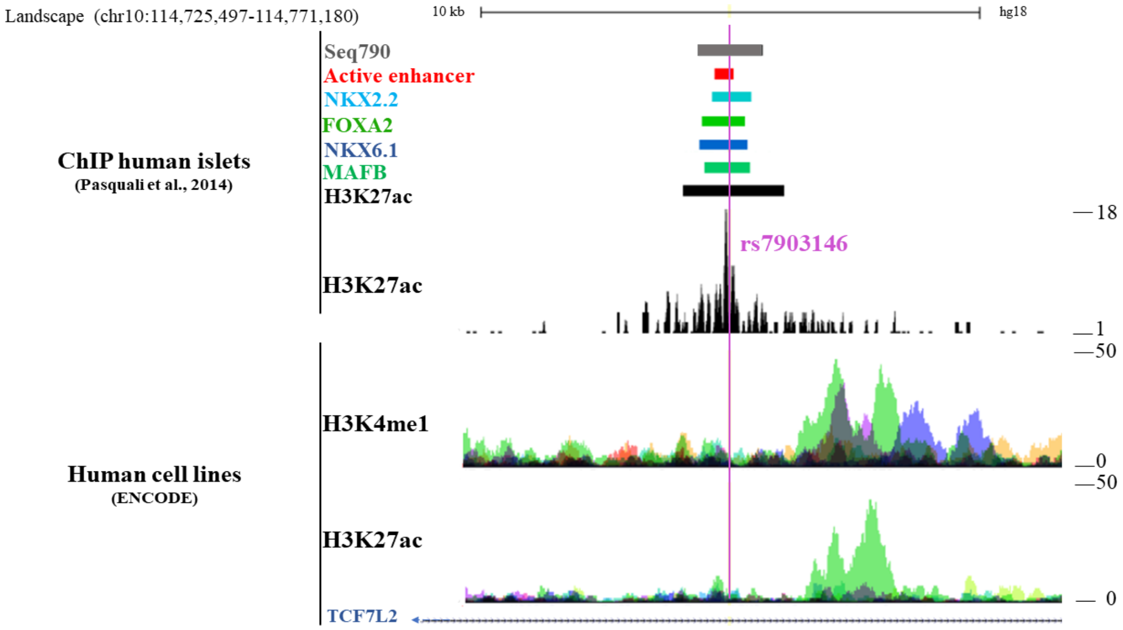
F


Landscape (chr4:185,940,266-185,966,631)
10 kb
hg18
ChIP human islets
(Pasquali et al., 2014)
Seq72
Active enhancer
NKX2.2
PDX1
FOXA2
NKX6.1
H3K27ac
H3K27ac
26
1
50
rs72695654
Human cell lines
(ENCODE)
H3K4me1
0
50
H3K27ac
0
ACSL1

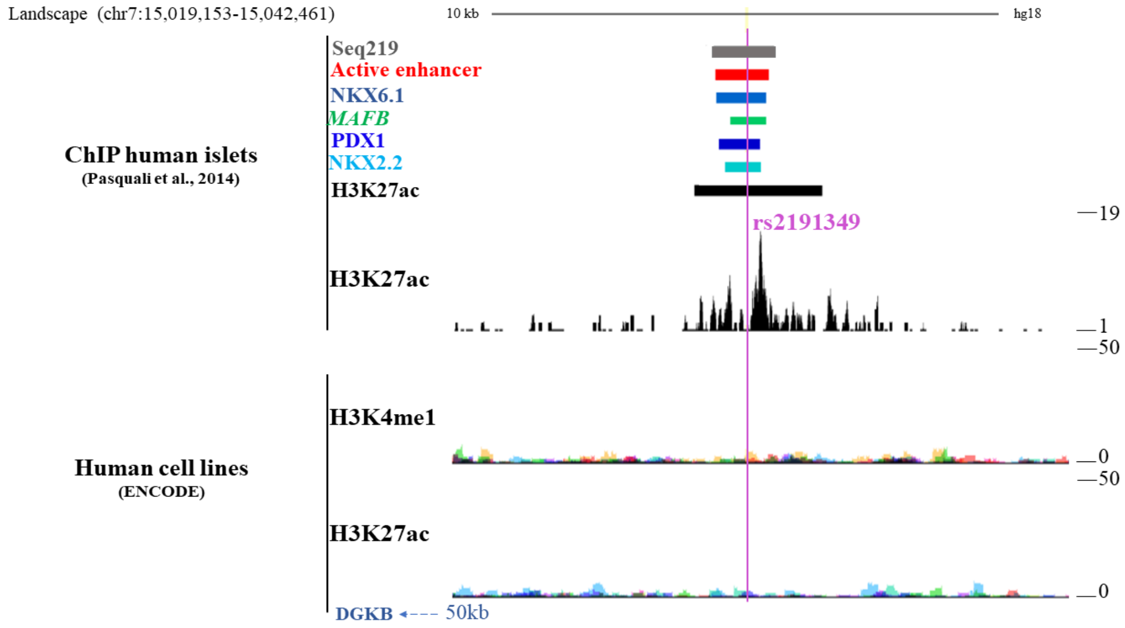
G


Landscape (chr8:118,231,272-118,276,820)
10 kb
hg18
ChIP human islets
(Pasquali et al., 2014)
Seq132
NKX2.2
FOXA2
H3K27ac
H3K27ac
36
1
50
rs13266634
Human cell lines
(ENCODE)
H3K4me1
0
50
H3K27ac
0
SLC30a8

H


Landscape (chr10:114,725,497-114,771,180)
10 kb
hg18
ChIP human islets
(Pasquali et al., 2014)
Seq790
Active enhancer
NKX2.2
FOXA2
NKX6.1
MAFB
H3K27ac
18
rs7903146
H3K27ac
1
50
Human cell lines
(ENCODE)
H3K4me1
0
50
H3K27ac
0
TCF7L2

I



J

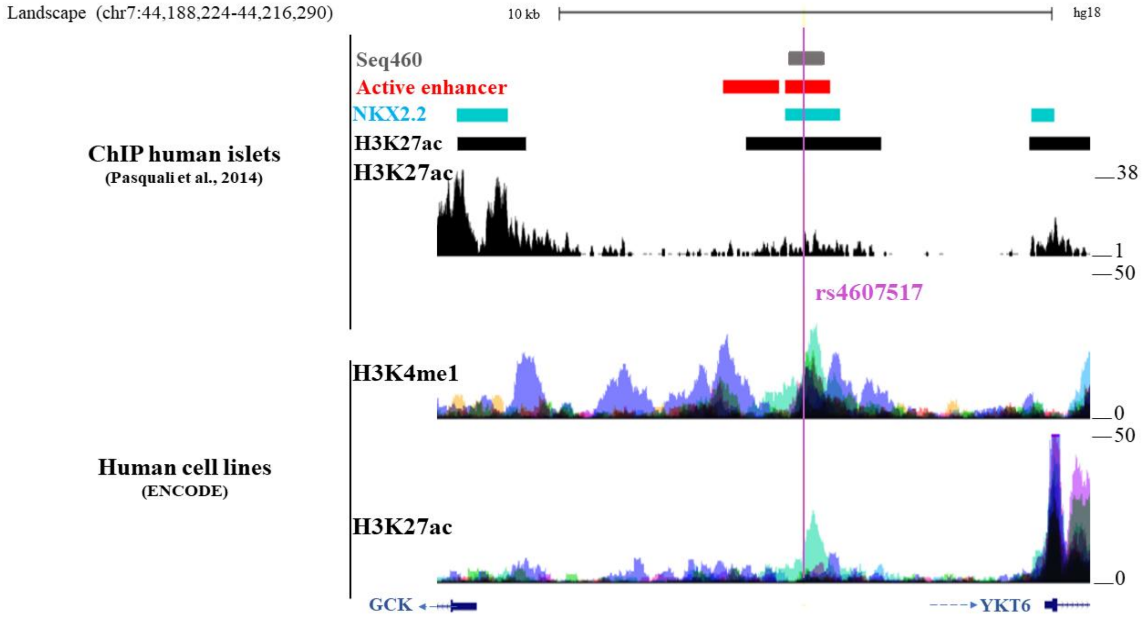


Figure 16 (A-J) – Genomic landscape of the selected sequences. The relative size of the sequence in the genome is represented in grey. The TFBS determined by ChIP-seq of human islets are represented in the respective colors (PDX1 – dark blue, MAFB – dark green, NKX6.1 – blue, FOXA2 – light green, NKX2.2 – light blue). The peak of acetylation determined in ChIP-seq is represented in black. Above, the profiles of acetylation and methylation in different cell lines from ENCODE data. In blue the nearby genes of the putative enhancer. In purple the SNP associated to T2D.

For sequences seq117, seq68, seq119 and seq58 (Fig.16 – A, B, C and D), high levels of H3K4me1 were detected overlapping the T2D associated SNP or nearby regions. Also, for sequences seq73, seq72 and seq119, high levels of H3K27ac mark were detected (Fig.16 – E, F and C). These results made us hypothesize that seq117, seq68, seq73, seq72 and seq58 could be enhancers, although it is not known if seq117, seq58 and seq68 could be active or primed. Seq119 could be more robustly predicted as an enhancer since this sequence overlap broadly with H3K27ac and H3K4me1. To improve this prediction of putative pancreas enhancers, apart from ENCODE data, we have analyzed data from Pasquali and co-workers (Pasquali et al, 2014). Using pancreatic islets, this data allowed to predict enhancers by combining different epigenetic marks, H3K27ac, H3K4me1, H3K4m3, CTCF and H2A.Z, labeled as "active enhancers" (Fig.16). All the sequences overlapped with this cluster of "active enhancers", except seq117 and seq132. Besides predicting enhancers, Pasquali and co-workers have also used ChIP-seq to identify the binding site of islet TFs (FOXA2, MAFB, NKX6.1, NKX2.2 and PDX1) in human endocrine islets (Pasquali et al., 2014).

Seq790 shows binding sites for MAFB and FOXA2 (Fig16- H). FOXA2 is expressed in multipotent pancreatic progenitor cells and posteriorly expressed in differentiated β-cells. The dual role of FOXA2 and MAFB suggests that seq790 could be a putative enhancer with an important function in endocrine islet differentiation and maturation, as seq132 and seq117 (Fig.16 – G and A). Seq790 also overlap with TFs present in differentiated cells: NKX6.1 and NKX2.2. NKX6.1 is important for β-cell differentiation and has as target gene *INSULIN*, being also critical in β-cell function (Ahlgren et al., 1996; Taylor et al., 2013).

Seq219 (Fig.16- I) shows binding sites for NKX6.1, MAFB, PDX1 and NKX2.2. PDX1 is required for pancreas development, β-cell differentiation and maturation, having, like NKX6.1, *INSULIN* as target gene (Ahlgren et al., 1996). Therefore, seq219 is a strong candidate to endocrine enhancer.

Seq117 (Fig.16- A) and seq132 (Fig.16- G), showed a coincident binding profile for MAFB and NKX2.2. MAFB is expressed in multipotent pancreatic progenitor cells of α and β-cells, being important for its inherent differentiation (Qiu et al., 2017). In addition, MAFB is also required for β-cell maturation (Qiu et al., 2017), suggesting that seq117 could be a putative enhancer with an important function in endocrine islet differentiation β -cell maturation. NKX2.2 is a transcriptional activator of NEUROD1 expression, being also important for the maturation of β-cells (Gu et al., 2011; Mastracci et al., 2013).

Seq58, seq73 and seq72 (Fig.16-D ,E and F) show binding sites for all islet TFs, except MAFB, while seq68 (Fig.16 – B) show binding sites for all specific islet TFs: NKX6.1, NKX2.2, FOXA2, MAFB and PDX1.

Seq460 (Fig16- J) only overlaps with NKX2.2 binding, while seq119 (Fig.16-E) did not show any binding site of the analyzed TFs.

All these data together allowed to build a list of high confidence putative enhancers (Table 2).

Although the target genes of these putative enhancers remain unknown, nearby genes might be good candidates, therefore, we have analyzed which genes are in closest vicinity to the respective putative enhancers (Table 2). For seq790 (Fig.16-H), the nearest gene is *TCF7L2,* a gene for which its loss-of-function has been described to affect *INSULIN* expression. In addition, its loss-of-function has been associated to T2D, since the presence of some genetic variants in the coding region of *TCF7L2* affect the levels of the protein and show T2D related phenotypes (Gloyn, et al.,2009). Seq219 (Fig.16-I) is in the genomic vicinity of *DGKB.* Interestingly, genetic variants in the coding region of this gene have been associated with a lower insulin release in the initial phase of the response to glucose from β-cells (Billings & Florez, 2010). Seq117 (Fig.16-A) is located in an intron of the *ADCY5* gene. This gene encodes an enzyme that helps to convert adenosine triphosphate to cyclic adenosine monophosphate being involved in signaling processes and in β-insulin secretion (Roman et al., 2017). SNPs located in the coding region of this gene are associated to T2D (Roman et al., 2017). Seq68 (Fig.16- B) has *TMEM154* as the most nearby gene, that encodes a transmembrane protein. Variants in the coding region of this gene might have an effect in secretion of intestinal hormones that can affect pancreatic β-cells (Harder et al., 2015), which could be indirectly associated to T2D. The *SLC2A2* gene is located nearby seq119 (Fig.16-C) and encodes a transmembrane carrier protein, also known by GLUT2 (glucose transporter 2), that has been shown to be important for proper insulin secretion (Laukkanen et al., 2005). Seq132 (Fig.16-G) is located in a *SLC30A8* intron that encodes for a zinc transporter, that is necessary for insulin crystallization and secretion (Rutter, 2010; Xiang et al., 2008). Surprisingly, studies with this gene have shown that 65% of the coding variants present in in *SLC30A8* resulted in a truncated protein and a reduced T2D risk (Flannick et al., 2014).

The *ZFAND3* gene is the nearest gene of the seq58 (Fig.16-D), which encodes a zinc finger protein. It is suggested that the variants located in these gene have a sex specificity in American Indian population (Muller et al., 2017), since coding SNPs in *ZFAND3* are associated to T2D in woman (Muller et al., 2017). Sequences seq72 (Fig.16-F) and seq73 (Fig.16-E) are both located in *ACSL1* gene intronic regions. This gene encodes a long-chain acyl-CoA synthetase 1, that has the capacity to covert fatty acids into acyl-CoAs. The specific role of this gene in the pancreas domain it is not known in humans. However, the study of common variants in the coding region of *ACSL1* by meta-analysis show an association with fasting glucose and diabetes (Manichaikul et al., 2016). Lastly, seq460 (Fig.16-J) has two very nearby genes: *GCK* and *YKT6*. *GCK* encodes glucokinase, necessary in glucose metabolisms pathways by β-cells, modulating insulin secretion. Mutations in *GCK* gene are, consequently, directly related to T2D, by altering this enzyme activity (Gloyn, 2003). *YKT6* gene encodes a protein receptor and it is involved in vesicular transport between membranes. It has roles in exosomes production and release in lung cancer cells. Due to its role, SNPs in the coding region of this gene are related to cancer cells survival (Ruiz-Martinez et al., 2016), but how the variants can be related to glucose homeostasis is unknown (Choi et al., 2017),

In summary, we have selected genomic sequences that overlap with SNPs associated to T2D and with epigenetic marks for enhancer activity. In addition, most of these sequences have binding sites for islet TFs and the nearby genes are associated to endocrine pancreas dysregulation. These characteristics make these sequences as good candidates to be enhancers, which activity might change in the presence of T2D associated variants. To determine if indeed these predictions are robust, sequences must be tested for enhancer activity.

2. PCR amplification of putative endocrine enhancers overlapping with T2D associated SNPs

To amplify the selected putative enhancer sequences, we have designed primers flanking these genomic sequences (Table 3). Fragments sizes varied from 651 bp to 1687 bp (base-pairs). PCR amplification was performed using human genomic DNA as a template and a *Taq* DNA polymerase with proofreading activity, as described in materials and methods. Resulting PCR products were run in an 1% agarose gel, bands were confirmed to have the expected molecular weight and then were excised from the agarose gel and purified using a gel pure kit as described in material and method section. Purified PCR products are show in figure 17.

Table 3 – Primers used in PCR amplification and characteristics for each sequence SNP associated.

Sequence	**Forward primer 5`- 3`(bp)**	**Reverse primer 3`- 5`(bp)**	**Tm (Cº)**	**Product size (bp)**
Seq790	AGGTGTGGGGGTATATGGTATCC	CACCAGGTCATGGAAACTTAGCC	65	1293
Seq219	CTACTGACATCAGCCAATGAGTCTAATACC	GTCCTCCAGGGCCTCTATATTCATGG	65	1152
Seq117	CTTCCCGGATGTGGAGATTCAGCC	GGAGGAGAAAGGAGGAAGCAACACC	65	1231
Seq68	CCTGGAGATTGTCTTCTAAGCTGC	GCAACTCAGATTGCATCTAGAGCC	65	1353
Seq119	ATGGCACAAACAAACATCCCACTCATTCC	ACTAATGGGCTGGTAGAAGAGGGCC	65	1592
Seq132	GCATTTACTGCCTCAAGAGAAAGC	GTGGCAACAACTTGGTGGGG	63	1687
Seq58	CTCTGAGAAGGAAATTGAACGC	AAAACCTCACATTAAAGCCATCCC	59	1068
Seq72	TTCGCAAAACATCTCATCACC	CAGGGTGAGAACTGAAGGC	63	651
Seq73	TCACCTGTGCCTGGCTGGG	GTGGGGTGGCCTGCAGGG	63	835
Seq460	GCCTATCTTCAAATCTCTACTTCCC	GATCAGGAAGACAGCGCTTGG	63	674

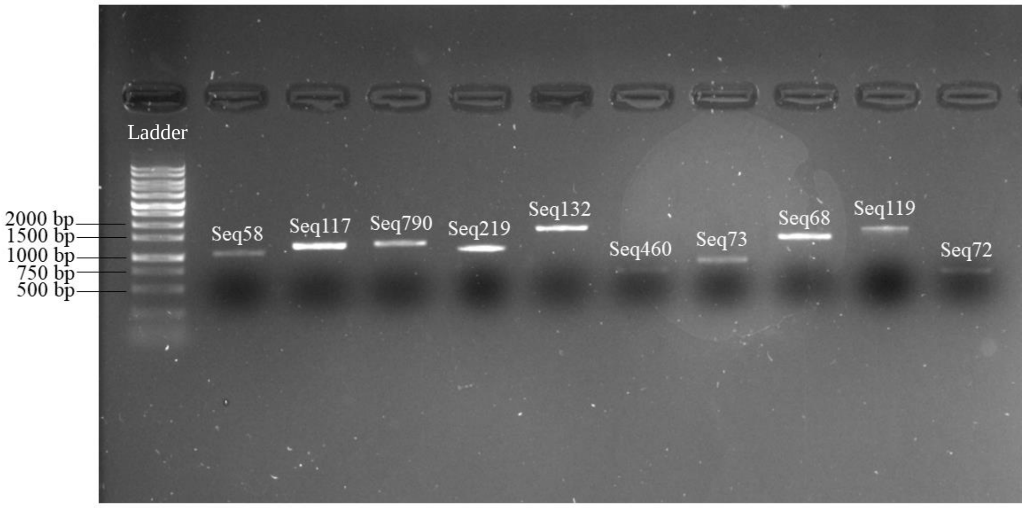


Figure 17 – Gel resulted from electrophoresis containing the product of the sequences PCR amplification (Seq58 – 1068bp; Seq117 – 1231 bp; Seq790 – 1293 bp; Seq219 – 1152 bp; Seq132 – 1687 bp; Seq460 – 674 bp; Seq73 – 835 bp; Seq68 – 1353 bp; Seq119 – 1592 bp; Seq72 – 651 bp) Ladder Gene Ruler 1kb.

3. Cloning of the PCR amplified genomic fragments in PCR8/ GW/ TOPO vector

After purification of the PCR products, DNA fragments were cloned in the PCR8/GW/TOPO vector. PCR8/GW/TOPO vector is a commercial TA compatible vector that is diluted in a TOPO isomerase mixture, facilitating the ligation of the amplified sequences. After the cloning reaction, DNA was transformed in chemically competent bacteria and plated in LB agar plates containing spectinomycin. Several colonies were selected to grow overnight in liquid media with spectinomycin and then plasmid extraction was performed. After extraction, plasmid DNA was cut with *EcoRI* to confirm the successful cloning of the sequences. *EcoRI* flanks the PCR8/GW/TOPO cloning site (Fig.18 – A and B). The digestion product had the DNA band of 2799 bp containing the vector backbone and a second band with the sequence of interest correct size (Fig.18-B).

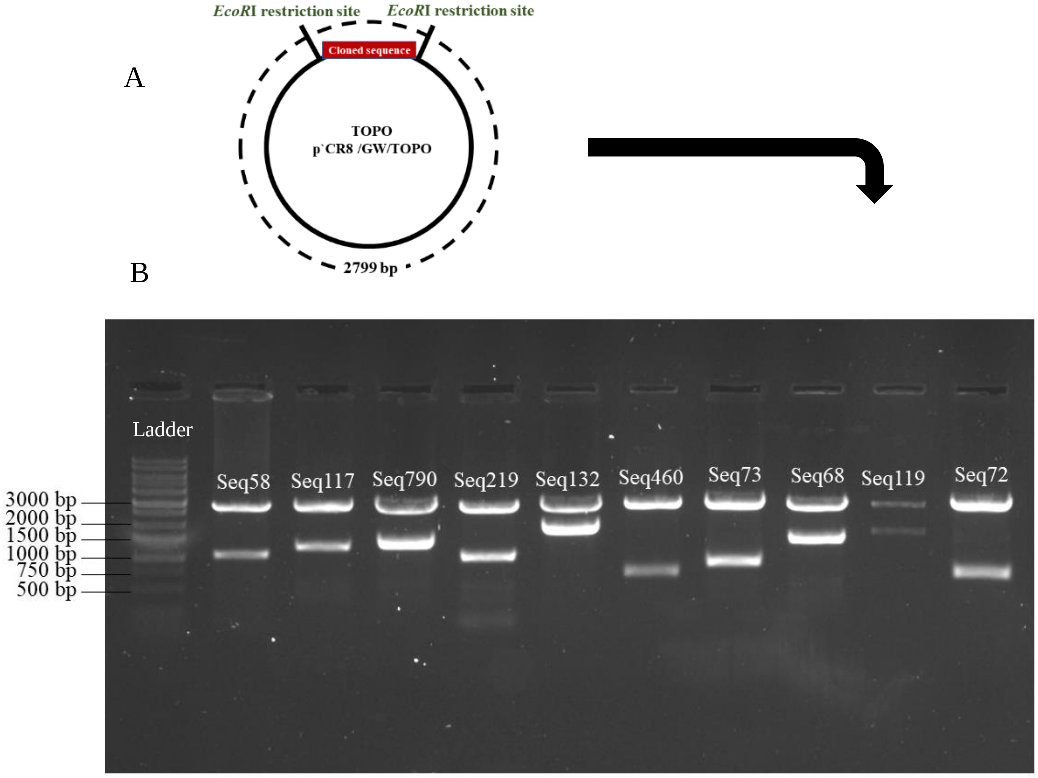


Figure 18 – A- Graphic scheme of TOPO and the cloned sequence; B - Gel electrophoresis resulted from electrophoresis containing the product of EcoRI enzyme digestion. (Seq58; Seq117; Seq790; Seq219; Seq132; Seq460; Seq73; Seq68; Seq119 and Seq72). A -The digestion product will have a DNA band of 2799 bp containing the vector backbone and an extra DNA band of the size of the cloned fragment. Ladder Gene Ruler 1kb.

4. Recombination of sequences to test for enhancer activity into Z48 transposon

The cloned sequences in PCR8/GW/TOPO were recombined in Z48 destination vector (Fig. 19 - B). The sequences were cloned between two *EcoRI* enzyme restriction sites (Fig.19 – A). In this specific vector, there is a third *EcoRI* restriction site. Therefore, the digestion with this enzyme allowed to confirm the successful insertion of the sequences by the visualization of three fragments in electrophoresis gel (Fig.19-B and C). The digestion product had the DNA bands of 4279 bp and 1835 bp from the vector backbone and a third band with the sequence of interest correct size (Fig.19-B).

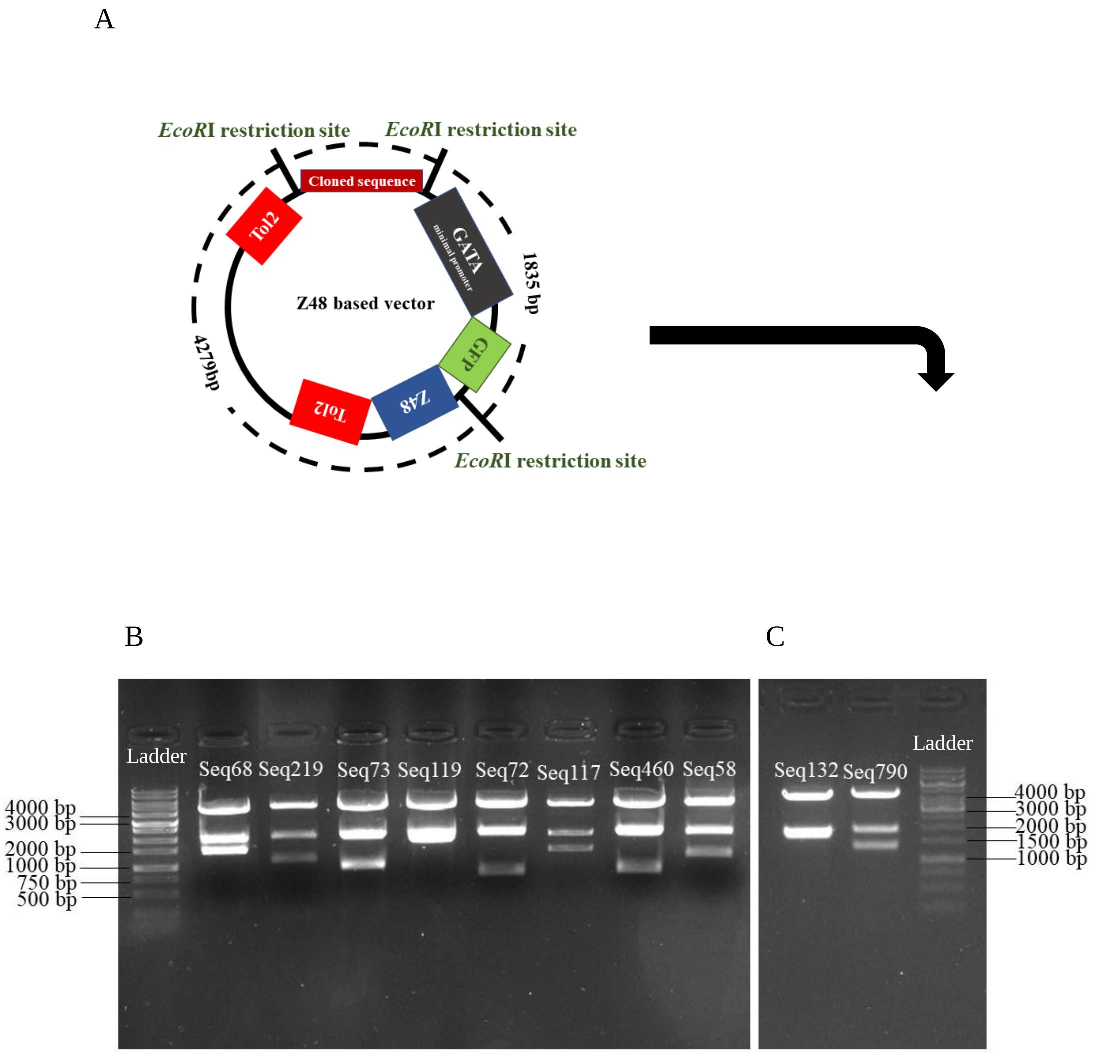


Figure 19 – A – Graphic scheme of Z48 based vector and the cloned sequene; B - Gel resulted from electrophoresis containing the product of EcoRI enzyme digestion. (Seq68; Seq219; Seq73; Seq119; Seq72; Seq117; Seq460; Seq58; C - Seq132; Seq790. The digestion product will be 4279 bp, 1835 bp and the size of the fragment. Ladder Gene Ruler 1kb.

5. *In vivo* transgenesis assays for endocrine pancreas enhancer activity in zebrafish

5.1. Defining the endocrine pancreas domain of zebrafish using *in vivo* reporter lines

To visualize zebrafish endocrine domain, we have used an *in vivo* transgenic reporter line Tg(*sst*:mCherry) that shows expression of mCherry in δ-cells. This reporter line contains the promoter of *Somatostatin* gene (*sst*) upstream of the mCherry reporter gene. To verify the position of β-cells relative to δ-cells, we crossed the *sst*-mCherry line with an *in vivo* reporter line for *Insulin*, containing the promoter of *Insulin* upstream of the *in vivo* reporter gene *GFP*. Embryos were grown up to 48hpf, a developmental time when the endocrine cells of zebrafish pancreas are already differentiated, and we analyzed the embryos by confocal microscopy (Fig.20). In all cases observed, the expression of *GFP* remained inside the expression domain of mCherry (Fig.20), showing that the *sst*-mCherry reporter line can be used to localize the zebrafish endocrine pancreas domain in further transgenesis assays to detect endocrine enhancers.

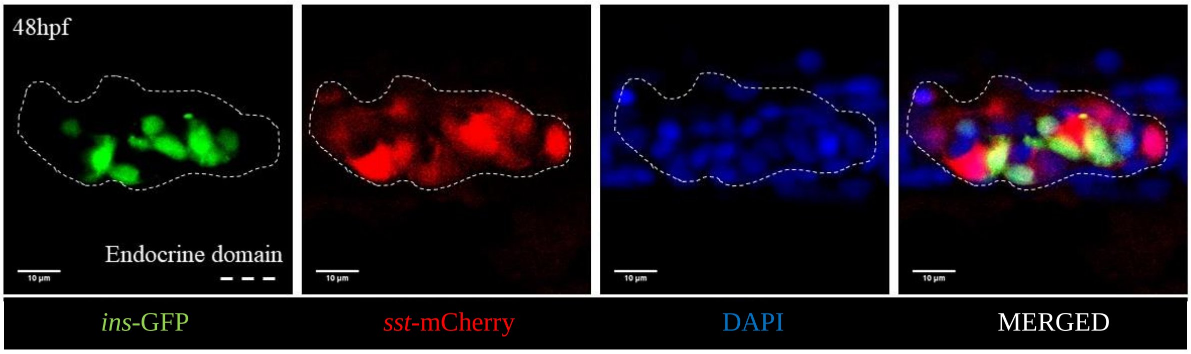


Figure 20 – Representative image showing the pancreas endocrine domain (dash line), regarding two reporter lines Tg:sst:mCherry and Tg:ins:GFP. Leica confocal SP5II; Zoom 2,91; Magnification 40x.

5.2. Zebrafish transgenesis using the Z48 transposon: controls

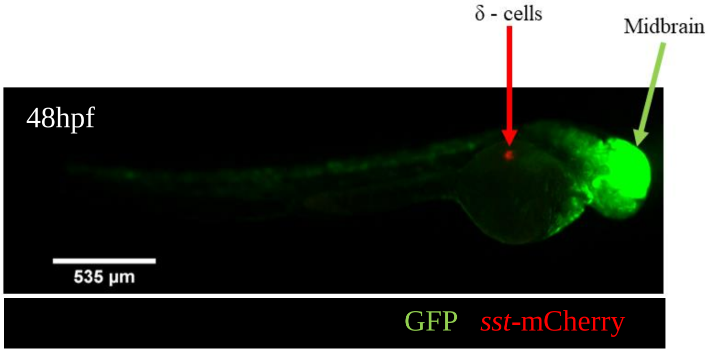


Figure 21 – Representative image of the GFP expression pattern when the Z48 vector is correctly injected and integrated in the zebrafish genome. Tg:sst:mCherry; Leica M205.

To test if the selected sequences (Table 2), cloned in Z48, are endocrine pancreatic enhancers, each Z48 vector was injected in one-cell stage zebrafish embryos from the *sst*-mCherry reporter line (Fig.21; red arrow). The Z48 transposable element contains a minimal promoter upstream of GFP and a downstream enhancer that activates expression of GFP in the midbrain, which can be used as a control for the transgenesis (Fig.21; green arrow). After microinjection of each Z48 transposable element, if the mobilization of the Z48 transposable element into the zebrafish genome was efficient, GFP expression is detected in the midbrain of 48hpf embryos (Fig.21). At 48hpf, pancreas endocrine cells are already differentiated (Fig.20), being this the adequate developmental time selected to perform the current assay.

The negative control for the current enhancer activity assay corresponded to a microinjection of the Z48 vector lacking a cloned sequence upstream of the minimal promoter, being denominated as Z48 empty vector. Upon injection and selection of embryos that presented GFP expression in the midbrain, embryos were analyzed in the confocal microscope to determine if GFP expression was detected in the pancreatic domain defined by the expression of mCherry in the *sst*-mCherry transgenic background. In forty-three embryos, positive for GFP expression in the midbrain, none has shown expression of GFP in the endocrine pancreas domain (Fig.22 – A and Fig.23). This negative control allowed to access the noise associated to random integrations of the transposable element in the zebrafish genome, also named position effect (Chung et al.,1993), establishing a minimal threshold to be compared with the results obtained with the tested sequences. Because noise was not observed in the negative control, it was important to determine the sensibility of the assay, otherwise false negative sequences could be identified. For that we have selected a positive control for the experiment (Fig.22 – B and Fig.23), which was the microinjection of a vector containing GFP as reporter gene under the control of the insulin promoter, known to drive robust expression in endocrine pancreas. Five out of nine embryos injected with the positive control showed expression of GFP in pancreas endocrine cells (Fig. 22- B and Fig.23). This result allowed to understand the level of integration and activity that it would be possible to expect injecting strong and robust enhancer.

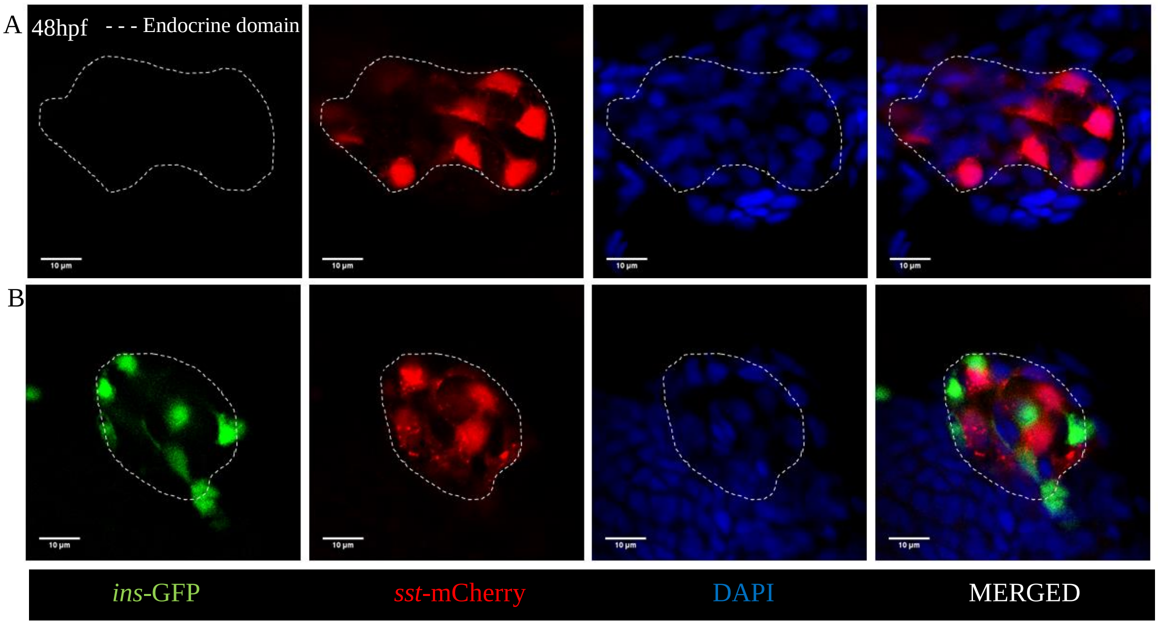


Figure 22 – A – Representative confocal images showing the negative control with no GFP in endocrine pancreas. B – Representative image of a positive control showing GFP expression in endocrine pancreas domain.; The dashed line represents the endocrine pancreatic domain. Leica confocal SP5II; Zoom 2,91 x; Magnification 40x.

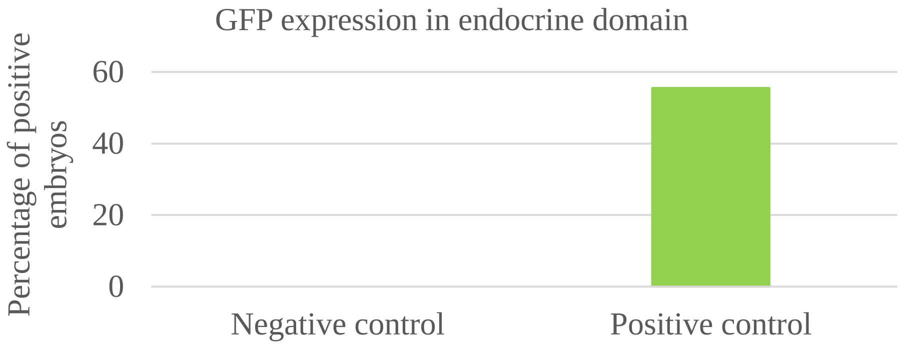


Figure 23 – Graph showing the total percentage of embryos with endocrine GFP expression

5.3. Zebrafish transgenesis using Z48 transposon: Endocrine pancreas enhancer assays

After testing the ten selected sequences for enhancer activity in the endocrine pancreas, five have shown GFP expression in the endocrine pancreas, being clearly above the threshold set by the negative control (seq219, seq132, seq58, seq73 and seq460) (Fig.24), while the remaining five sequences were not able to drive expression of GFP in endocrine cells (seq117, seq790, seq72, seq68 and seq119) (Fig.24 and 25).

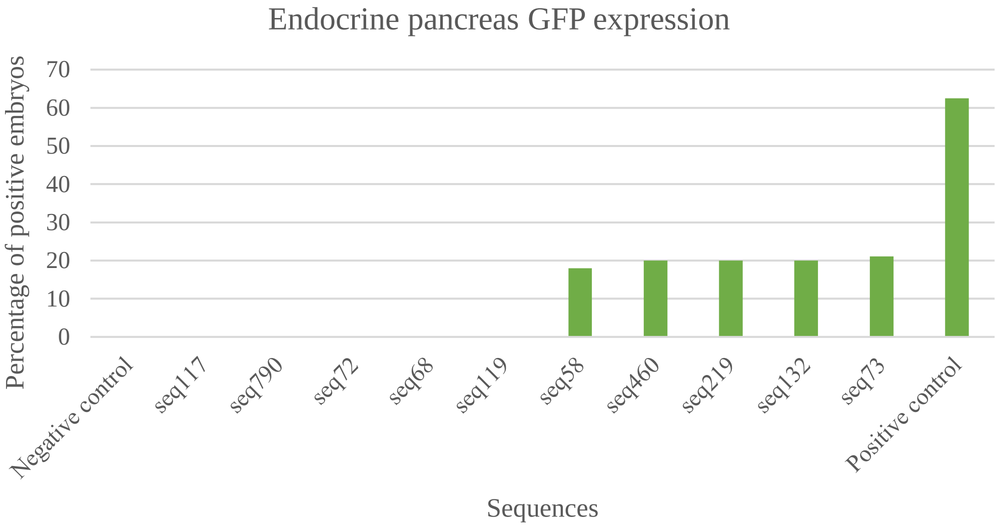


Figure 24 – Representative graph showing the total percentage of positive embryos with GFP expression in endocrine pancreas domain from each sequence analyzed.

5.3.1. Endocrine pancreas enhancers

Seq219 (n=20), seq132 (n=20), seq58 (n=22), seq73 (n=19) and seq460 (n=20) were able to drive expression of GFP in pancreas endocrine cells (Fig.25 – A to E and fig 26). Interestingly, seq219 and seq132 did not overlap with any significant signal of H3K27ac and H3K4me1 present in cell lines derived from different tissues, not including the endocrine pancreas (ENCODE data; Fig.16). However, when analyzing by ChIP-seq results from endocrine islets, it is possible to detect an enrichment for H3K27ac and binding of TFs important for endocrine proper function. The validation of these sequences as endocrine enhancers by *in vivo* reporter assays underline the relevance and accurateness of the predictions for enhancer activity when using endocrine pancreatic islets. Seq58, seq73 and seq460 in contrast, overlap with significant signals of H3K27ac and H3K4me1 derived from not endocrine cell lines, together with high levels of H3K27ac and binding sites of islet TFBS derived from pancreas endocrine islets (Fig.16).

In summary, from the sequences that have shown to be pancreatic endocrine enhancers, all of them have shown high signal of H3K27ac mark in pancreas endocrine cells, together with binding sites of TFs important for endocrine pancreas function. Presence of H3K27ac and H3K4me1 in cell lines not derived from endocrine cells was not required for at least two sequences (seq219 and seq132). This is in agreement with the observation that different enhancers might be active or inactive in different tissues (Remenyi et al., 2004¸ Delic et al., 1991), suggesting that predictions based in epigenetic marks for enhancer activity will be more accurate when analyzing datasets from the tissue to be studied, in this case, the endocrine pancreas.

Although we were able to identify five sequences with enhancer activity on the endocrine pancreas, it is yet to be determined the exact expression pattern that these enhancers drive within this tissue. To overcome this problem, the positive enhancers were recombined in a ZED (Zebrafish Enhancer Detector) vector (Bessa et al., 2009), in the same way as Z48 recombination protocol, that has two insulators flanking the cloned sequence to avoid the "position effect" (Chung et al.,1993) from the activity of nearby genomic regulatory regions. Injected embryos are being reared to adults to generate stable transgenic lines containing the validated enhancers, allowing to determine a consistent expression pattern of GFP in the endocrine pancreas.

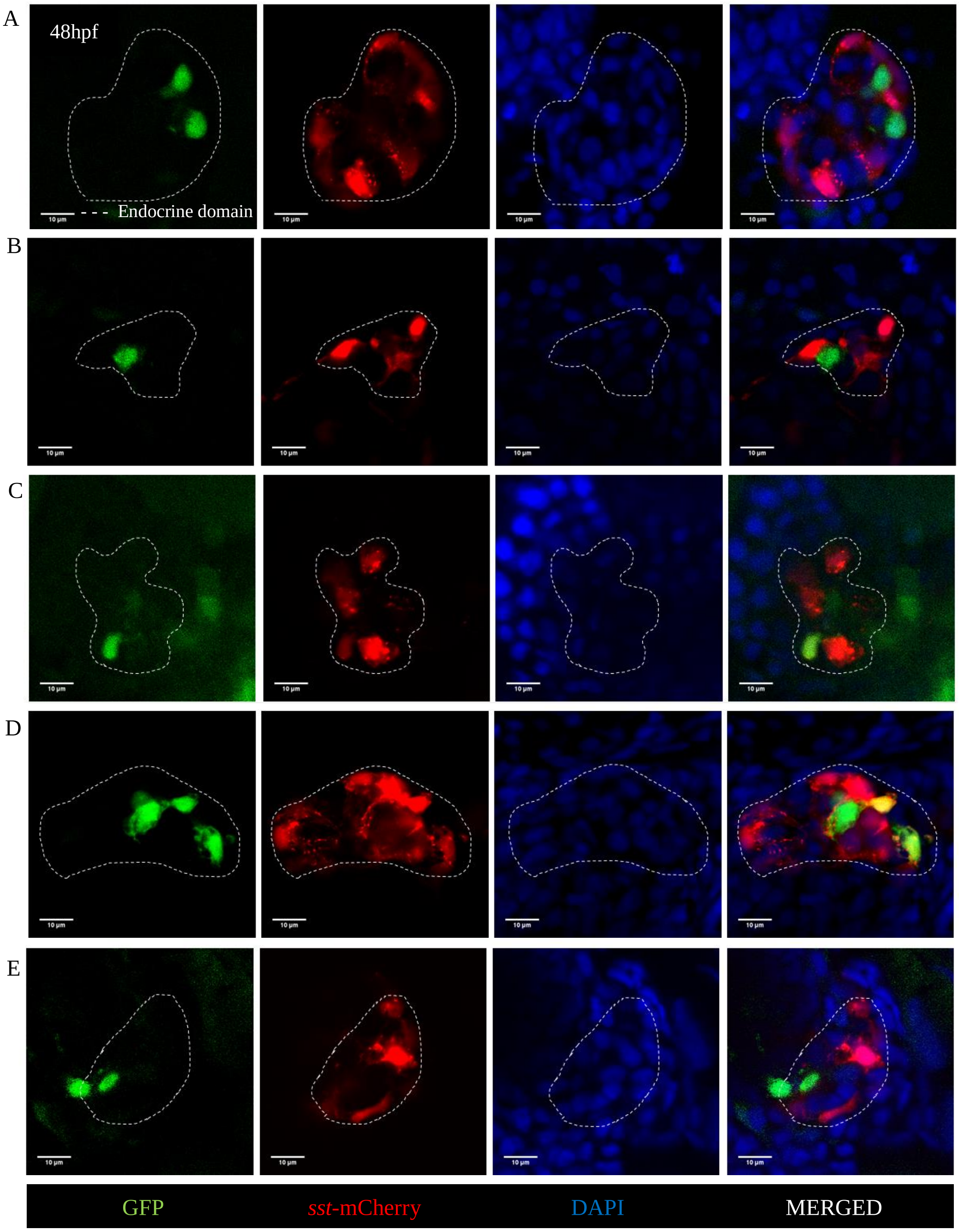


Figure 25 – Representative confocal images of seq219 (A), seq132 (B), seq58 (C), seq73 (D) and seq460 (E) analysis. All these five sequences showed GFP positive cells. It was used Tg:sst:mCherry as reporter line; The dashed line represents the endocrine pancreatic domain .Leica confocal SP5II; Zoom 2,91 x; Magnification 40x.

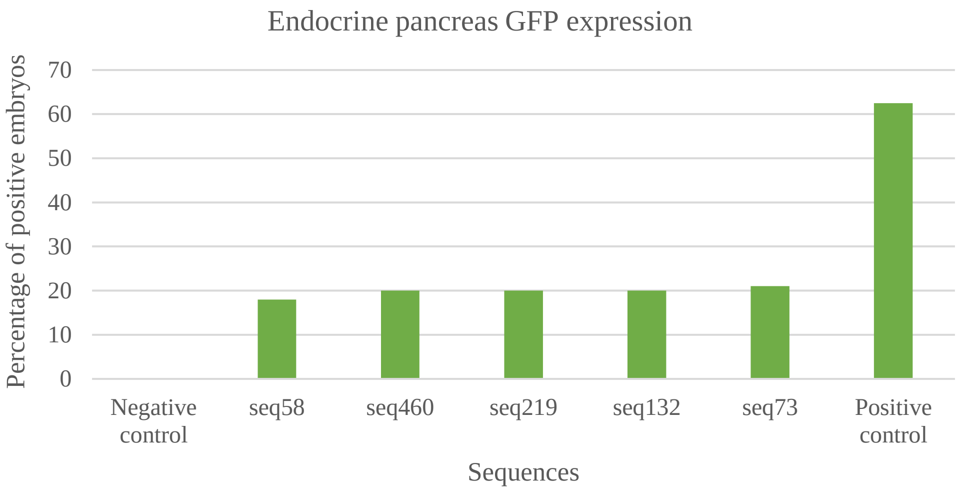


Figure 26 – Graph showing the total percentage of embryos expressing GFP in endocrine domain.

5.3.2. Sequences with no endocrine pancreas enhancer activity

The remaining five sequences tested for endocrine enhancer activity were not able to drive expression of GFP in endocrine cells, having 0% of embryos with GFP expression in the *sst*-mCherry domain, namely: seq117 (n=21), seq790 (n=20), seq72 (n=27), seq68 (n=19) and seq119 (n=18) (Fig.27 – A-E and Fig.28).

Seq117 shows an overlap with H3K4me1 signal but reduced H3K27ac signal derived from cell lines not related with endocrine pancreas (Fig.16). Regarding the data derived from endocrine pancreas islets, this sequence shows little overlap with H3K27ac mark, however binding sites for MAFB and NKX2.2 were identified.

The remaining four sequences, seq790, seq72, seq68 and seq119, regardless of their signal for H3K27ac and H3K4me1 derived from not endocrine cell lines, all of them presented high levels of H3K27ac in endocrine cells (Fig.16). These results suggest that, although sequences present histone marks associated to enhancer activity, namely H3K27ac in cells from the tissue where enhancer activity is being evaluated, presence of H3K27ac is not sufficient to determine these sequences as enhancers. An alternative explanation could be related with the sensitiveness of our assay, since random integrations were expected to generate at least some noise in the negative control, described as position effect (Chung et al.,1993). Therefore, endocrine expression could be very restrictive in our assay, compromising the detection of very week enhancers. A third explanation for the absence of enhancer activity in the endocrine pancreas could be related with interspecies specific response, since the analyzed sequences are from the human genome. Nevertheless the model organism used for the reporter assays was the zebrafish, that could lack the proper combination of transcription factors required for the activity of some human endocrine enhancers (Davis et al., 2014).

.

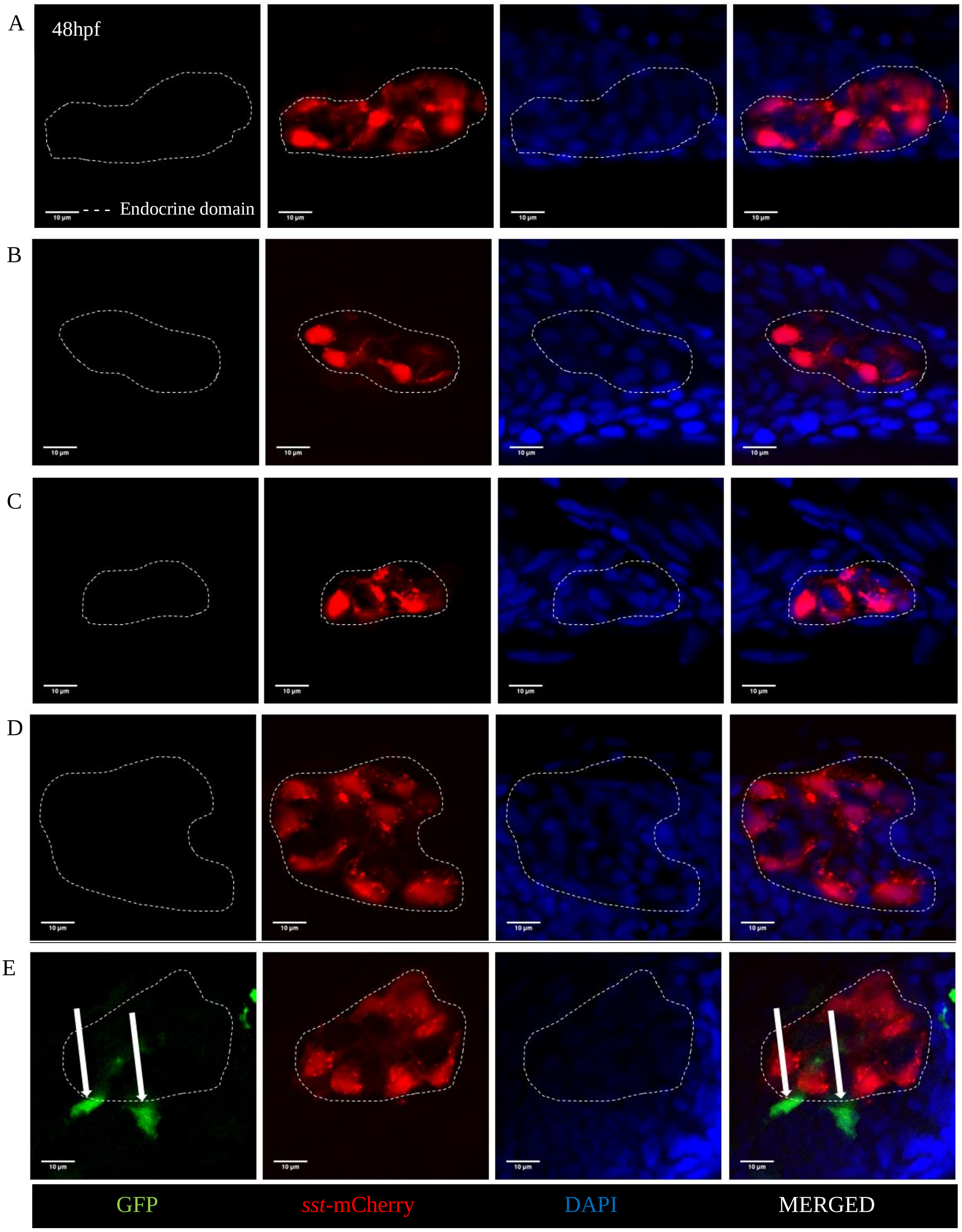


Figure 27 – Representative confocal images of seq117 (A), seq790 (B), seq72 (C), seq119 (D) and seq68(E) analysis. 0% of the embryos expressed GFP in the endocrine domain. In E 15% of the embryos showed GFP expression in the adjacent area (white arrows). It was used Tg:sst:mCherry as reporter line; The dashed line represents the endocrine pancreatic domain. Leica confocal SP5II; Zoom 2,91 x; Magnification 40x.

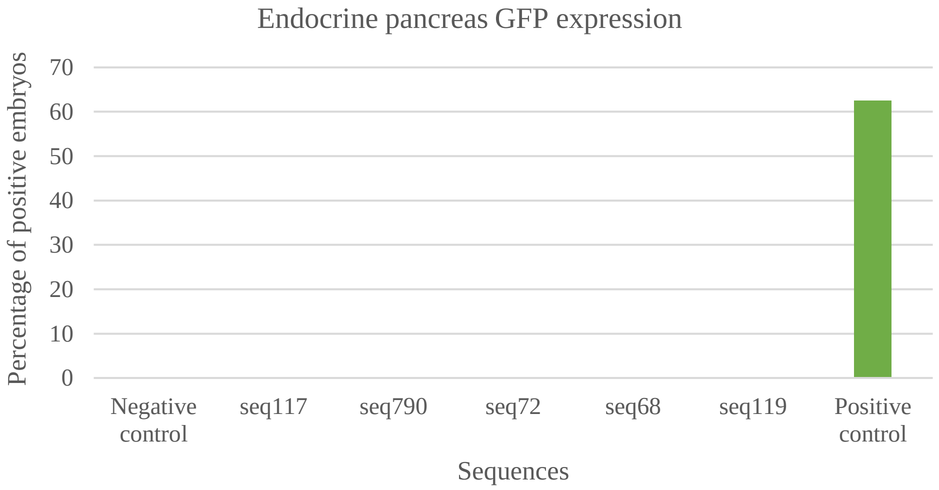


Figure 28 – Graph showing the total percentage of embryos with GFP expression in endocrine domain in seq117, seq790, seq71, seq68 and seq119.

5.3.3. Putative enhancers of endocrine progenitor cells

During the course of the enhancer activity assays, we observed the presence of GFP positive cells in the adjacent area of the endocrine domain in assays for seq68, seq58 and seq73 (Fig.27- E (white arrows), Fig.28 and fig.30). One possible identity for the GFP labeled cells could be pancreatic progenitor cells, since they are described to be adjacent to the endocrine differentiated domain. Indeed, at 24hpf in zebrafish embryos, the pancreatic progenitor domain defined by Nkx6.1 TF do not co-localize within the domain of endocrine pancreatic differentiated cells, but ventrally to these hormone-producing cells (Binot et al., 2010) (Fig.31). At 48hpf, Nkx6.1 is expressed at the base of the endocrine islet, in the ventral bud (Ghaye et al., 2015) (Fig.31).

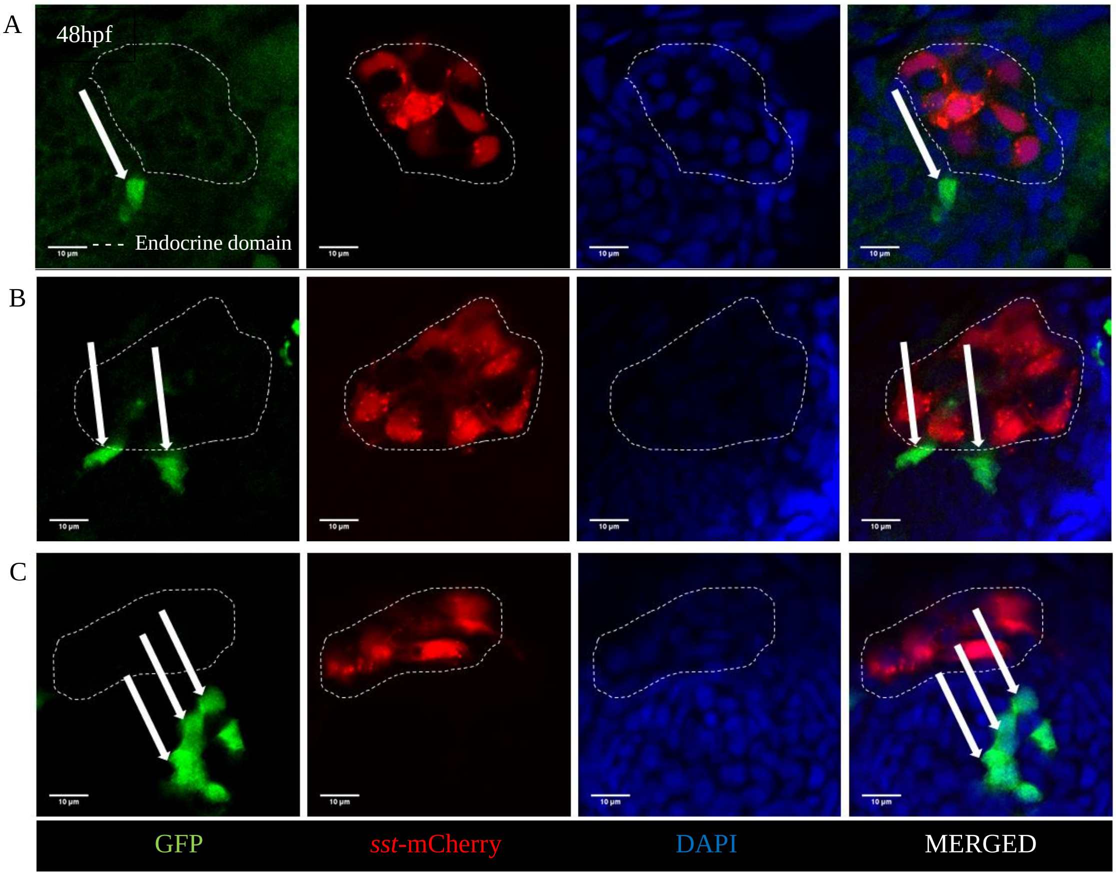


Figure 29 – Representative images of the sequences with potential to be enhancers for pancreatic progenitors A – Seq58; B – Seq68; C – Seq73. The dashed line represents the endocrine pancreatic domain. Leica confocal SP5II; Zoom 2,91 x; Magnification 40x

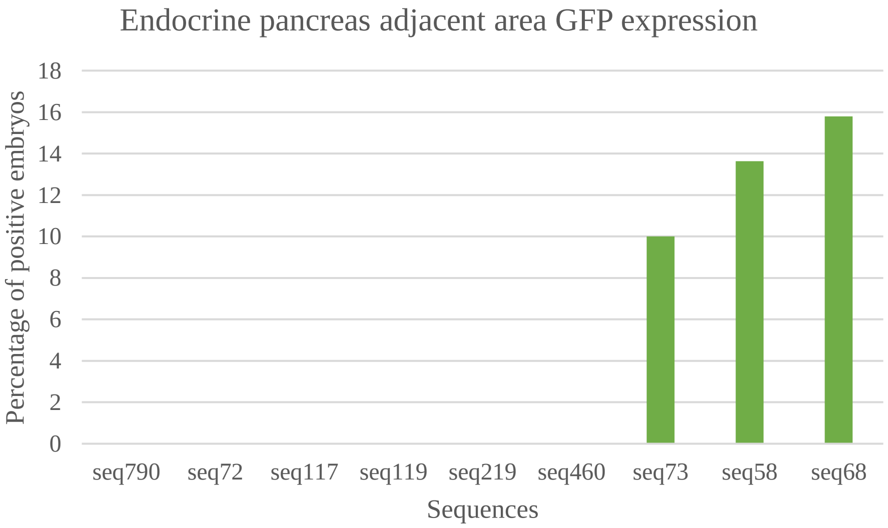


Figure 30 – Graph showing the total percentage embryos with GFP positive cells in endocrine pancreas adjacent domain from each sequence analyzed.

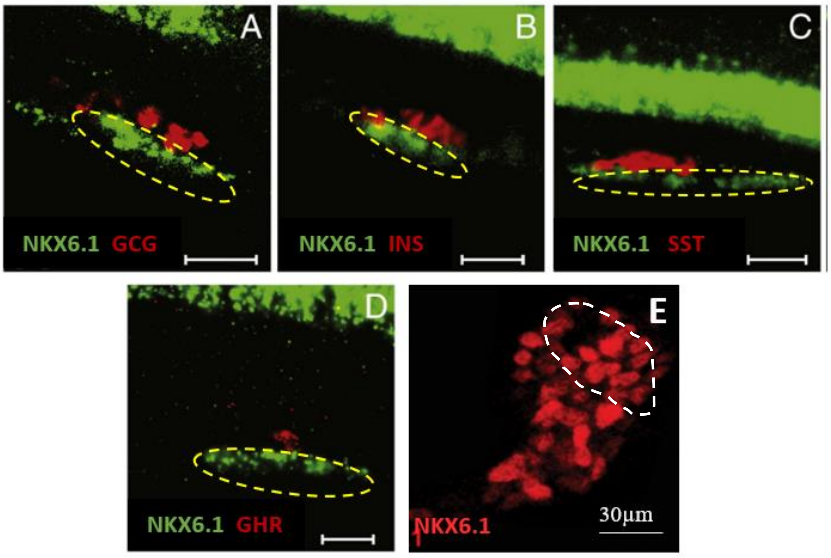


Figure 31 – NKX6.1 pancreatic expression at 24h, regarding: A- glucagon; B- insulin; C- somatostatin; D – ghrelin. Scale: 20 µm. E - NKX6.1 expression at 48h in zebrafish. The endocrine domain is highlighted with white dash. Adapted from Binot et al., 2010 and Ghaye et al., 2015)

Therefore, to define the endocrine pancreatic progenitor cells domain, regarding the endocrine domain, we have crossed Tg(*sst*:mCherry) and Tg(*ins*:GFP) reporter lines and we stained 48hpf embryos with Nkx6.1 antibody.

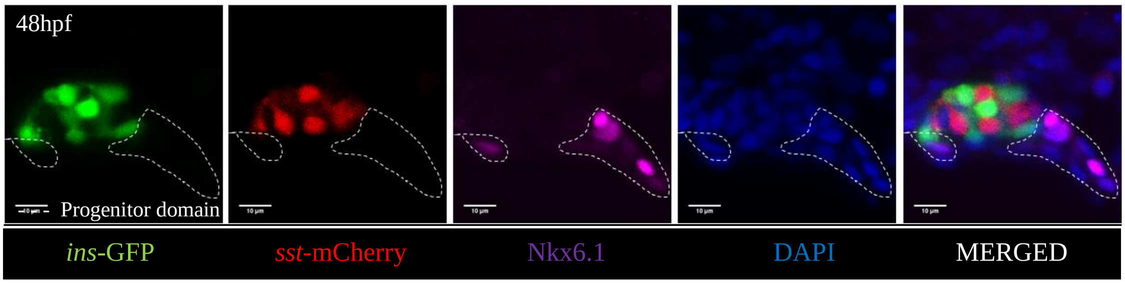


Figure 32 – NKX6.1 pancreatic expression at 48hpf,in zebrafish endocrine domain.. The dashed line represents the pancreatic progenitor domain defined by Nkx6.1 antibody. Leica confocal SP5II; Zoom 2,91 x; Magnification 40x

After analyzing these embryos by confocal microscopy, we found, as previously described (Ghaye et al., 2015), that the Nkx6.1 expression domain is ventral and adjacent to the endocrine differentiated islet (Fig. 32). Yet, the question still remains if sequences seq68, seq58 and seq73 are able to drive expression in endocrine progenitor cells, being therefore endocrine progenitor enhancers. Indeed, the same way that SNPs associated to T2D could impair the proper function of enhancers active in differentiated cells, resulting in pancreatic malfunction, SNPs in progenitor endocrine enhancers could affect developmental processes required for proper pancreas differentiation, potentially being as well a source of pancreatic mal function. To address this question, we have performed enhancer assays, in a *sst*-mCherry reporter line background in embryos stained with anti-Nkx6.1 at 48hpf (Fig.34, 35 and 36).

The negative control was obtained by the microinjection of Z48 empty vector, showing zero embryos with expression of GFP co-localized with anti-Nkx6.1 (Fig.33).

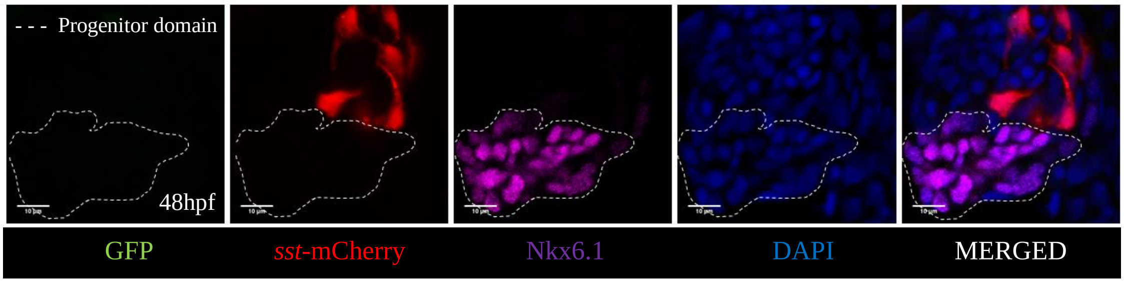


Figure 33 – Representative images of the negative control, with GFP negative cells co-localizing with NKX6.1 positive cells. The dashed line represents the pancreatic progenitor domain defined by Nkx6.1 antibody. Leica confocal SP5II; Zoom 2,91 x; Magnification 40x

The three sequences, seq68, seq58 and seq73 were then tested for enhancer activity in endocrine progenitor cells that were considered as putative progenitors enhancers by immunohistochemistry using the NKX6.1 antibody.

a) Seq68 is an enhancer of endocrine pancreas progenitor

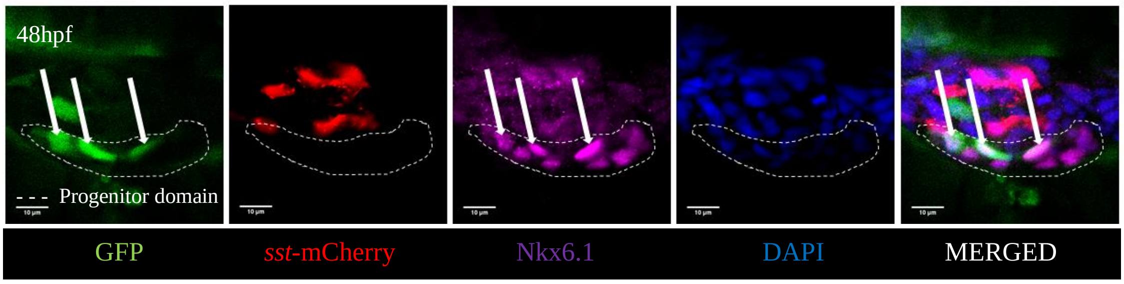


Figure 34 – NKX6.1 pancreatic expression at 48h, in zebrafish endocrine domain. The dashed line represents the pancreatic progenitor domain defined by Nkx6.1 antibody Leica confocal SP5II; Zoom 2,91 x; Magnification 40x

After testing seq68 for enhancer activity, it was found that 75% of the analyzed embryos (n=4) presented co-expression of GFP with the anti-Nkx6.1 antibody (Fig.34 and 36). These results indicate that the seq68 could be an enhancer of endocrine progenitor cells. Interestingly, when we look at the genomic landscape of seq68 (Fig.16), a NKX6.1 binding site overlap with this sequence, supporting the progenitor identity of the uncovered enhancer.

b) Seq73 and seq58 are not enhancers of endocrine progenitor cells

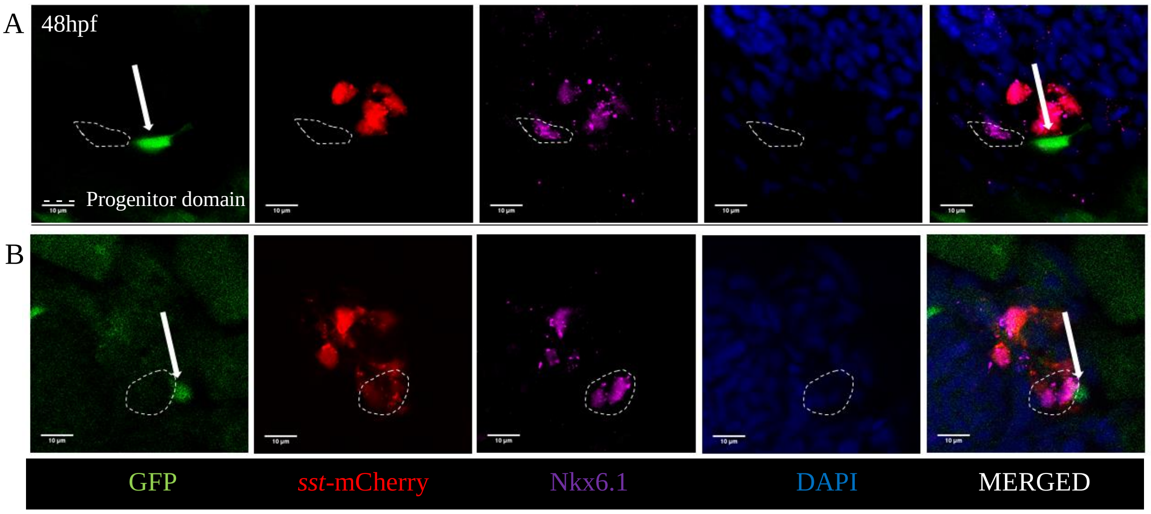


Figure 35 – Representative image of the seq58 (A) and seq73(B) with GFP positive cells not co-localizing with NKX6.1 positive cells. The dashed line represents the pancreatic progenitor domain defined by Nkx6.1 antibody Leica confocal SP5II; Zoom 2,91 x; Magnification 40x

For sequences seq58 and seq73, none of the analyzed embryos showed expression of GFP co-localized with Nkx6.1 (n=3) (Fig.35 and 36), suggesting that these sequences are not enhancers of endocrine progenitors cells. The presence of GFP positive cells outside of the differentiated endocrine domain could be explained by the "position effect", already referred, due to the random integration of the Z48 transposon in the zebrafish genome. Alternatively, these sequences could be enhancers of pancreatic progenitor cells that are in a developmental state previous to Nkx6.1 expression and therefore, previous to endocrine fate determination. To access the possibility, other markers should be used, such as Pdx1 or Ptf1a. A third possibility could be that these sequences are enhancers active in cells located nearby the endocrine domain, but not related to this tissue.

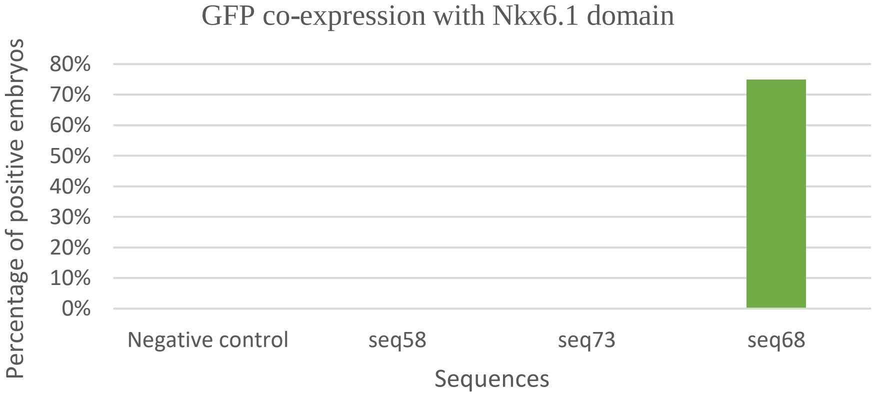


Figure 36 – Graph showing the total percentage embryos with GFP positive cells that colocalized with NKX6.1 antibody in endocrine pancreas adjacent domain.

5.4. Impact of the T2D risk variant in uncovered enhancers

To determine to what extent genetic variants associated T2D could impact in enhancer activity, we focused in three sequences, seq460, seq73 and seq132, which previously we determined to be endocrine pancreas enhancers We amplified by PCR exactly the same sequence but containing the single nucleotide variant associated to T2D, and we compared the ability of the three sequences to drive expression of GFP in the endocrine pancreas. Seq460 did not show differences in the percentage of embryos with GFP expression in the endocrine pancreas, when comparing to *WT* sequence (Fig. 37 and Table 4). However, sequences seq73 and seq132 showed differential enhancer activity when comparing the *WT* and T2D variants. For seq73, WT variant has shown 21% (n=19) of embryos with GFP expression in the endocrine pancreas, while in T2D associated variant it showed 0% (n=20) (Fig. 39 and Table 4).

Sequence	**WT allele** **% endocrine GFP positive cells**	**Risk allele** **% endocrine GFP positive cells**
seq132	20%	50%
seq73	21%	0%
seq460	20%	20%

Table 4 – Percentage of endocrine GFP positive cells relative to WT and Risk allele.

Although sequences were selected after PCR amplification for having exactly the same sequence with the exception of the T2D associated SNP, after analyzing in detail the *WT* seq73 sequence, we found a small deletion that is not present in the T2D associated variant. To completely exclude that the identified deletion could be causing the differential enhancer activity between the two tested sequences, we must amplify again the *WT* sequence, not containing this deletion, and perform the enhancer assay.

Regarding seq132, we observed that *WT* variant has 20% (n=20) of embryos with GFP expression in the endocrine pancreas, while T2D associated variant showed 50% (n=20). These results suggest that the T2D associated variant has a gain of function, when comparing to the *WT* variant. This could be a consequence of either an increase of the transcriptional output of GFP or an increase of the expression pattern driven by the T2D variant. To distinguish both causes, as future perspectives we will repeat these assays in human cell lines using luciferase as a reporter gene, that will allow to assess the transcriptional output for each variant in a quantitative manner. To access the possibility that the T2D associated variant results in an increased expression pattern, we will generate stable transgenic lines as described previously. To better understand how the different variants could be affecting the binding of TFs, therefore explaining the differences in the enhancer activity we have observed, we performed an *in silico* analysis of TFBS using JASPAR (http://jaspar.genereg.net/) (Sandelin, 2004).

a) Seq460

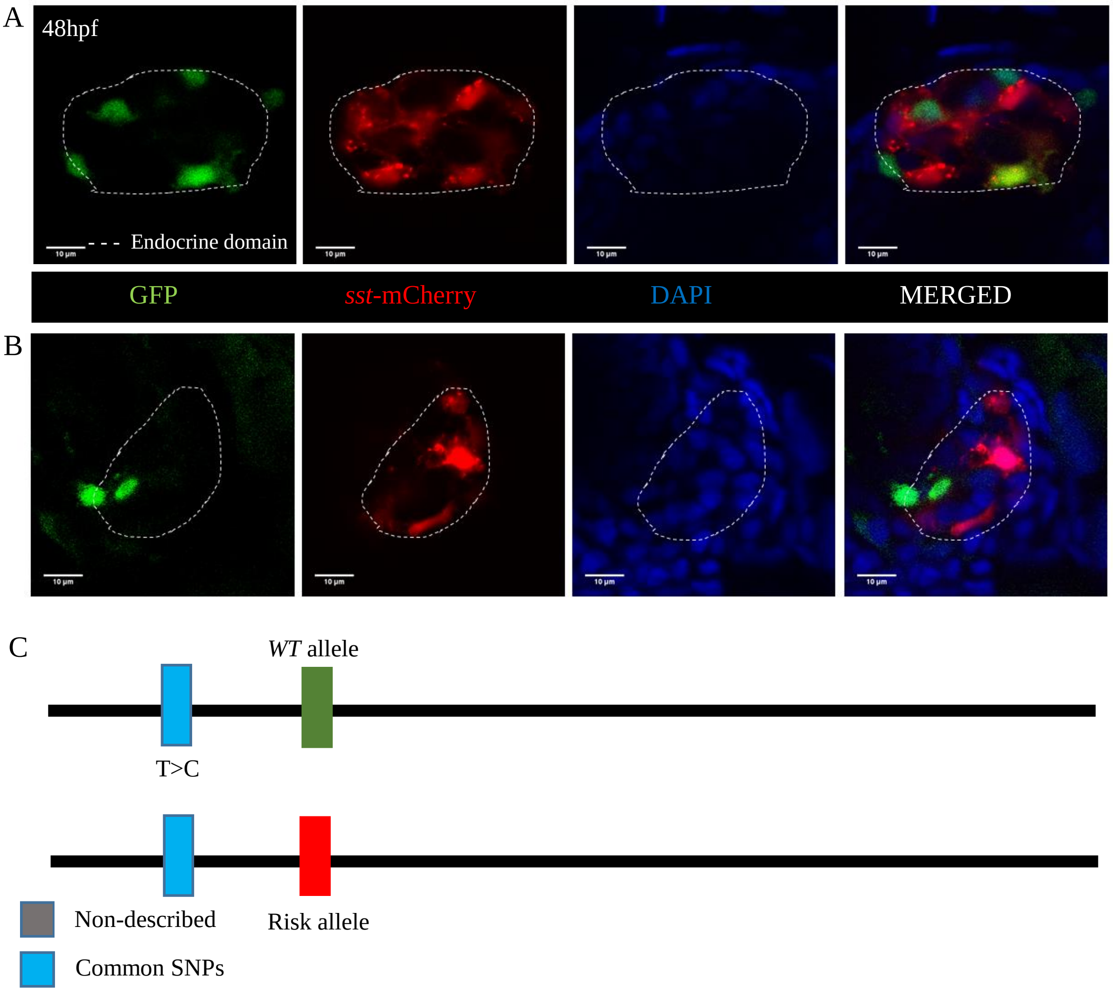


Figure 37 – A and B - Representative image of the seq460 without the risk associated variant present (WT sequence) (A) and seq460 with the risk associated variant. The dashed line represents the endocrine pancreatic domain Leica confocal SP5II; Zoom 2,91 x; Magnification 40x C- Representative squeme of the seq460 without the risk associated variant present (WT sequence) and seq460 with the risk associated variant.

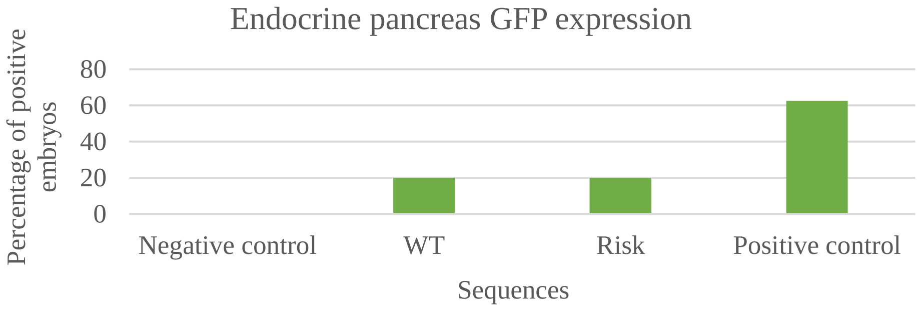


Figure 38 – Representative graph showing the total percentage of positive embryos in WT and Risk allele associated in seq460.

b) Seq73

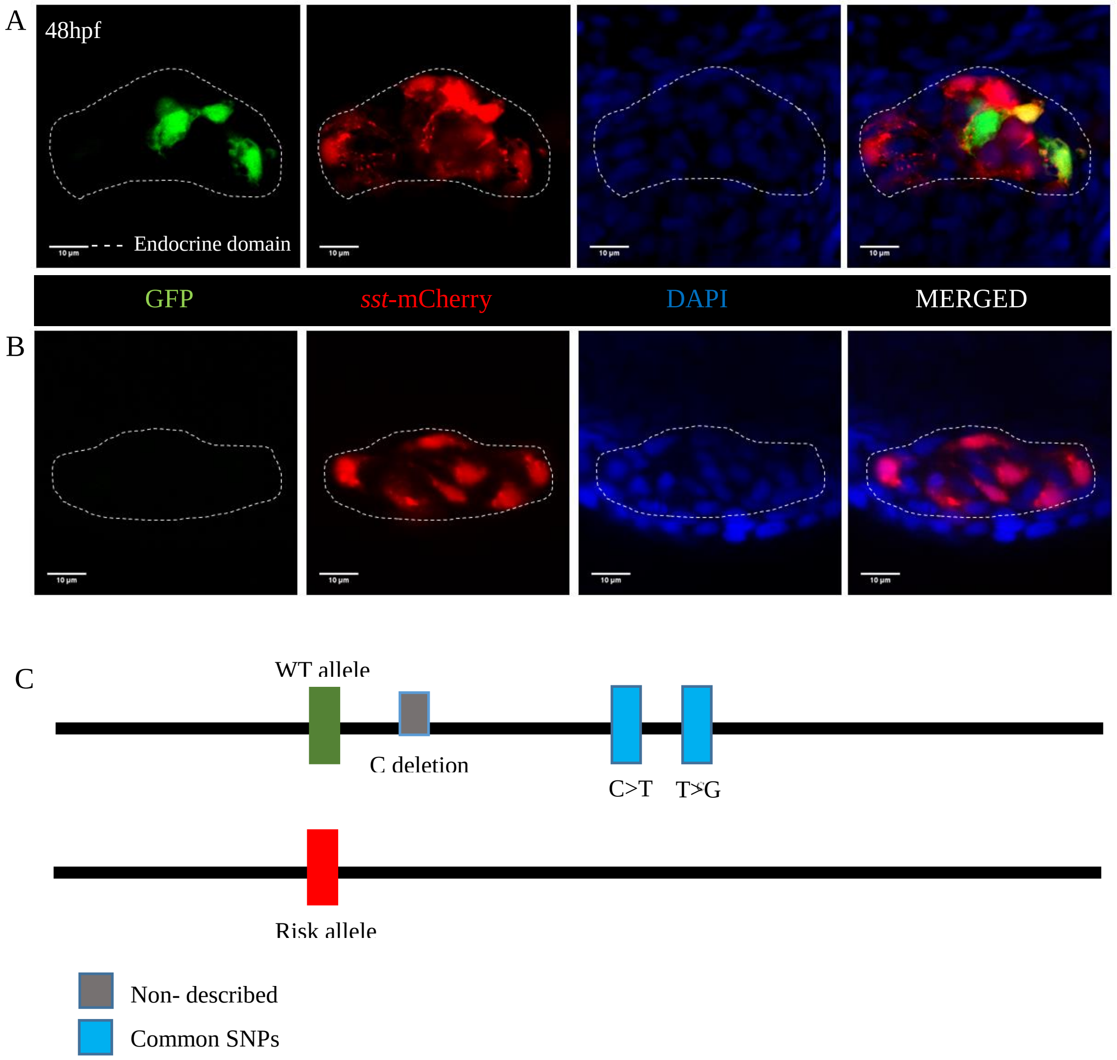


Figure 39 – A and B - Representative image of the seq73 without the risk associated variant present (WT sequence) (A) and seq73 with the risk associated variant. The dashed line represents the endocrine pancreatic domain .Leica confocal SP5II; Zoom 2,91 x; Magnification 40x C- Representative squeme of the seq73 without the risk associated variant present (WT sequence) and seq73 with the risk associated variant.

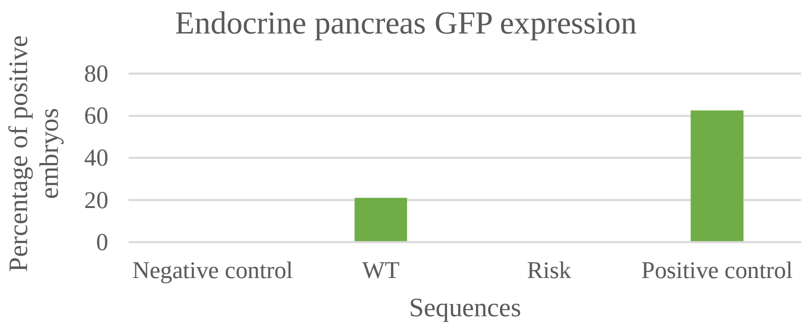


Figure 40 – Representative graph showing the total percentage of positive embryos in WT and Risk allele associated in seq73.

c) Seq132

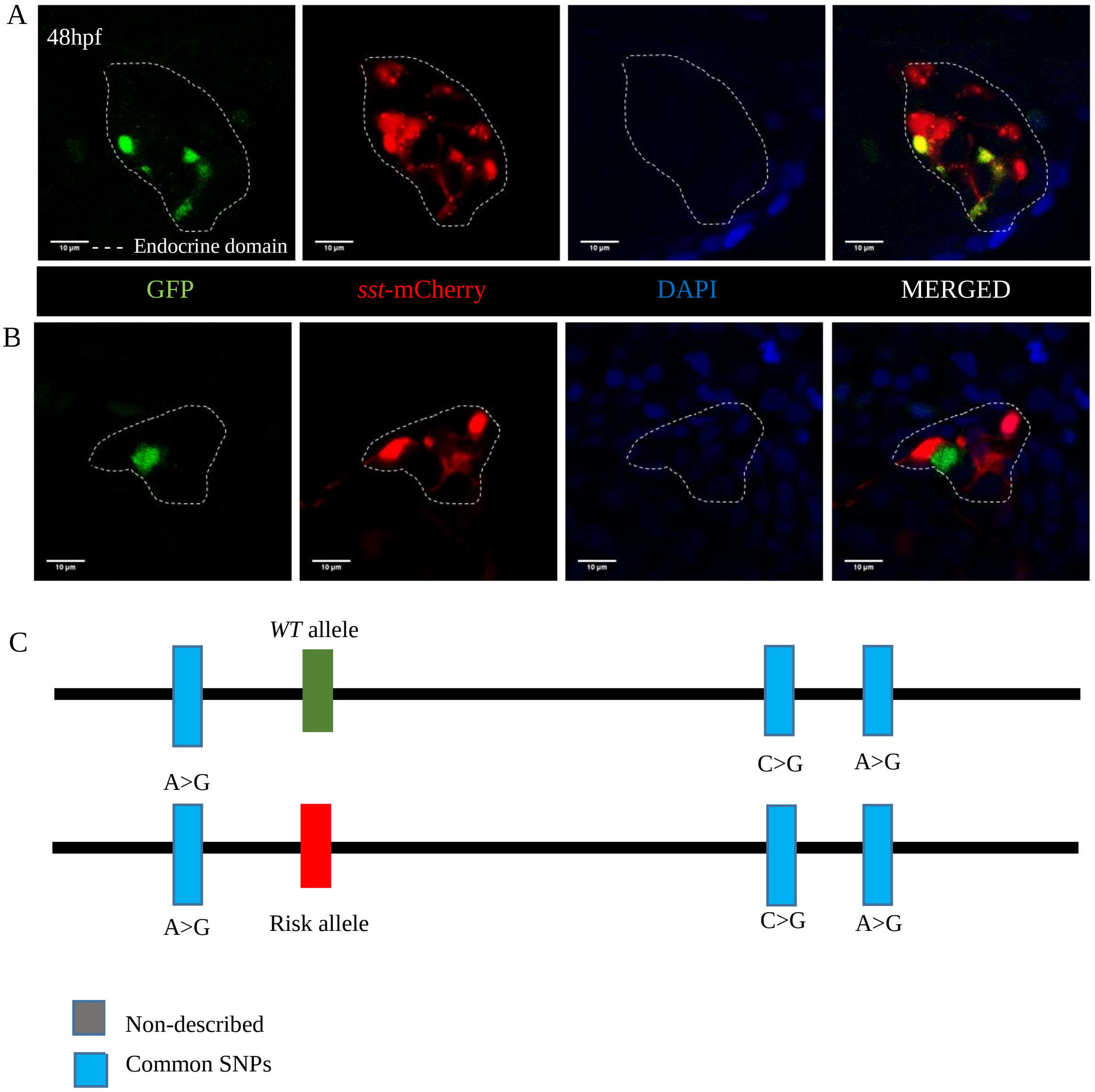


Figure 41 – A and B- Representative image of the seq132 without the risk associated variant present (WT sequence) and seq132 with the risk associated variant. The dashed line represents the endocrine pancreatic domain .Leica confocal SP5II; Zoom 2,91 x; Magnification 40x; C- Representative squeme of the seq132 without the risk associated variant present (WT sequence) and seq132 with the risk associated variant.

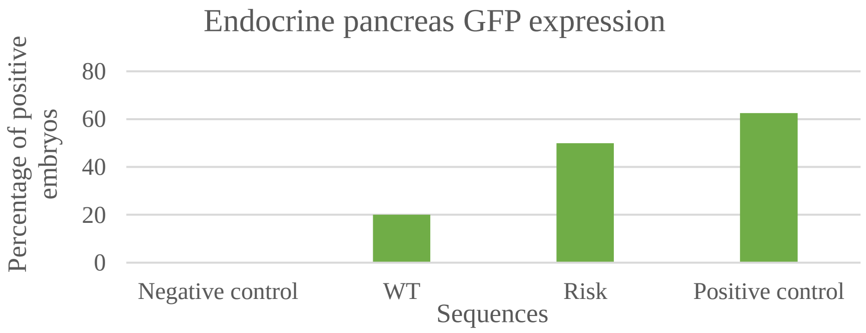


Figure 42 – Graph showing the total percentage of positive embryos in WT and Risk allele associated in seq132..

d) JASPAR analysis – TFBS prediction

Table 5 – JASPAR analysis from the seq132, where the only differential SNP in the two sequences is the risk and WT allele. In grey: TFBS motif; In green: WT allele; In red: Risk allele; # Complementary chain

TF	Allele	Affinity score (0-low; 1 – high)		Sequence
TCF3	*WT*	0.952250390723	#	**CCGAACCACTTGGCTGTCCCAGCTGGCTGCTGTTGATAAAG**
TCF3	Risk	0	#	**CCGAACCACTTGGCTGTCCCGGCTGGCTGCTGTTGATAAAG**
PAX5	*WT*	0.858808679485		**CTTTATCAACAGCAGCCAGCTGGGACAGCCAAGTGGTTCGG**
PAX5	Risk	0.888219182514		**CTTTATCAACAGCAGCCAGCCGGGACAGCCAAGTGGTTCGG**
PAX2	*WT*	0	#	**CCGAACCACTTGGCTGTCCCAGCTGGCTGCTGTTGATAAAG**
PAX2	Risk	0.856667239833	#	**CCGAACCACTTGGCTGTCCCGGCTGGCTGCTGTTGATAAAG**
NEUROD1	*WT*	0.872118037809		**CTTTATCAACAGCAGCCAGCTGGGACAGCCAAGTGGTTCGG**
NEUROD1	Risk	0		**CTTTATCAACAGCAGCCAGCCGGGACAGCCAAGTGGTTCGG**
RBPJ	*WT*	0.855183014761		**CTTTATCAACAGCAGCCAGCTGGGACAGCCAAGTGGTTCGG**
RBPJ	Risk	0		**CTTTATCAACAGCAGCCAGCCGGGACAGCCAAGTGGTTCGG**
ASCL1	*WT*	0.840690936086		**CTTTATCAACAGCAGCCAGCTGGGACAGCCAAGTGGTTCGG**
ASCL1	Risk	0		**CTTTATCAACAGCAGCCAGCCGGGACAGCCAAGTGGTTCGG**
SP1	*WT*	0	#	**CCGAACCACTTGGCTGTCCCAGCTGGCTGCTGTTGATAAAG**
SP1	Risk	0.801216594697	#	**CCGAACCACTTGGCTGTCCCGGCTGGCTGCTGTTGATAAAG**

The table 5 shows a selected group of TFs that bind differentially to *WT* sequence and T2D risk variants, in the seq132 (See supplementary data – table 1). This group of TFs was selected by their potential function in the pancreas. We observed that the *WT* variant has a predicted increased affinity to bind TF 3 (TCF3), NEUROD1, (Recombination Signal Binding Protein for Immunoglobulin Kappa J Region) (RBPJ) and ASCL1, while the T2D Risk variant lose the affinity to bind these TFs, gaining affinity to bind PAX2 and Specificity protein 1 (SP1). In addition, although not being an example of gain or loss, the binding site of paired homeobox 5 (PAX5) is predicted to be more stable in the T2D risk variant than in the *WT* variant. These TFs have different characteristics as following:

TCF3 is a TF related to neuronal differentiation. More importantly, it can bind to short regulatory DNA sequences in *INSULIN* gene, acting as a transcriptional activator (Uniprot dataset, 2018), however, its role is not yet fully understood in endocrine pancreas function (Cristancho et al., 2011).

PAX2, as referred before, has a key role in pancreas development, since it controls the relative proportion of endocrine and exocrine pancreas tissues (Zaiko et al., 2004). In addition, since this transcription factor works as an activator, the presence of its binding site in the T2D risk variant could explain the gain of function observed in this variant (Fig.41 and 42).

NEUROD1 regulates *INSULIN* gene expression is important for pancreas cell fate determination (Itkin-Ansari et al., 2005) and the absence of this TF may result in T2D (RefSeq, Jul, 2008).

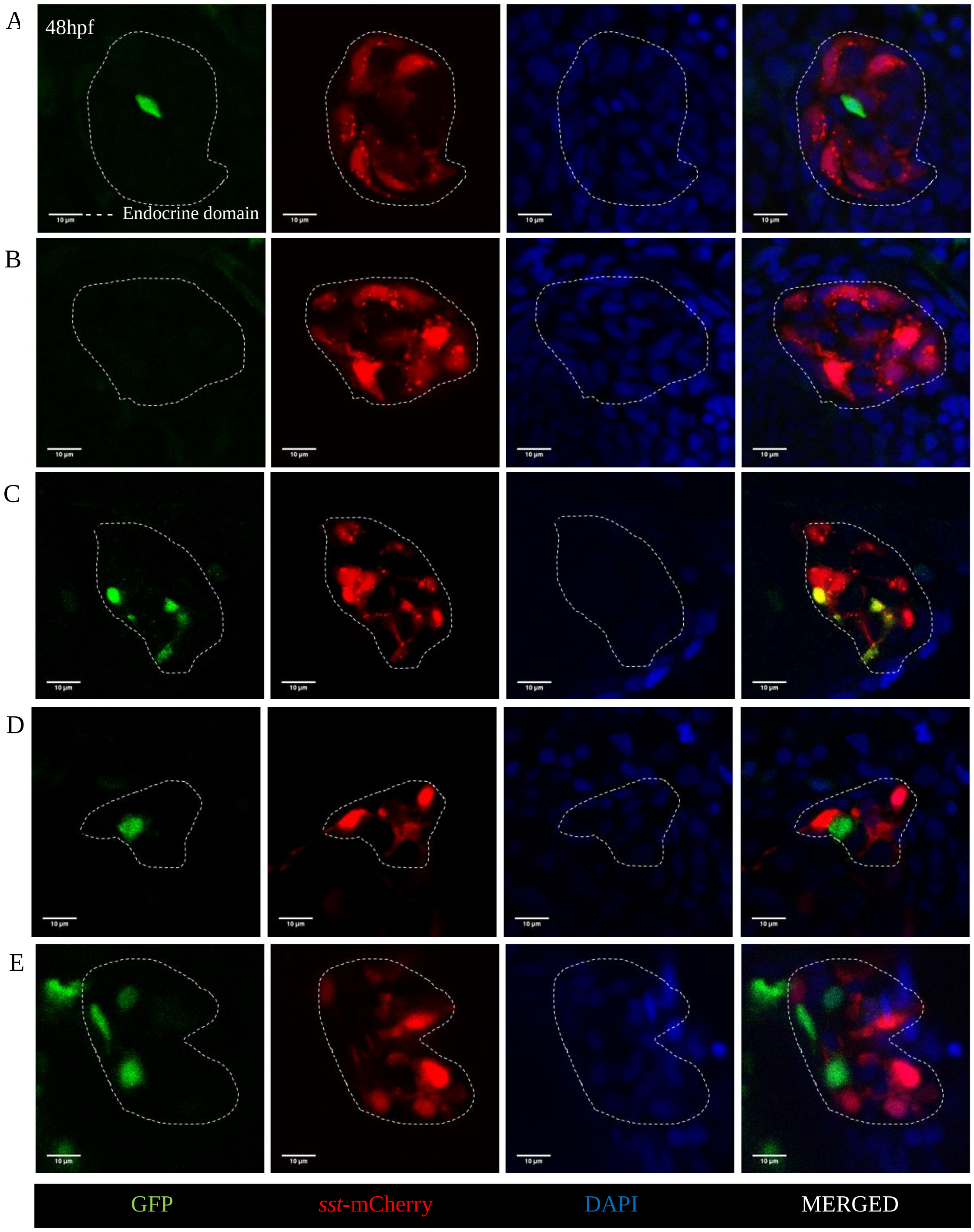
RBPJ have been associated to Notch signaling (Lake et al., 2014), a pathway that works as a key regulator of pancreas embryonic development and homeostasis (Kim et al.,2010). This TF is not functioning as activator in the sequence *WT*. Interestingly, RBPJ can work as a transcriptional repressor when it is not binding to Notch proteins. RBPJ repressive activity could explain the decreased enhancer activity observed associated to the *WT* variant (Kim et al.,2010).

ASCL1 controls neuronal differentiation, being described as a transcriptional activator. The only association with diabetes resides in the consequences of high glucose levels, that alters the expression of (Fu et al.,2006)..

SP1 is a zinc finger TF. Interestingly, post-translational modifications such as phosphorylation, acetylation and glycosylation significantly affect the activity of this protein, which can operate as an activator or a repressor of transcription (Solomon et al., 2008) (Pan et al, 2001).

In summary, a single nucleotide modification has the potential to change the binding of several TFs that eventually might impact in the transcriptional output of the enhancer. This should be further addressed by performing ChIP-PCR in the corresponding sequences, to determine which of the proposed TFs are effectively binding to the different variants, allowing us to build a better molecular explanation for the differential enhancer activity observed when comparing the *WT* and T2D risk associated variants.

e) A new single nucleotide variant can disrupt the enhancer activity in seq132



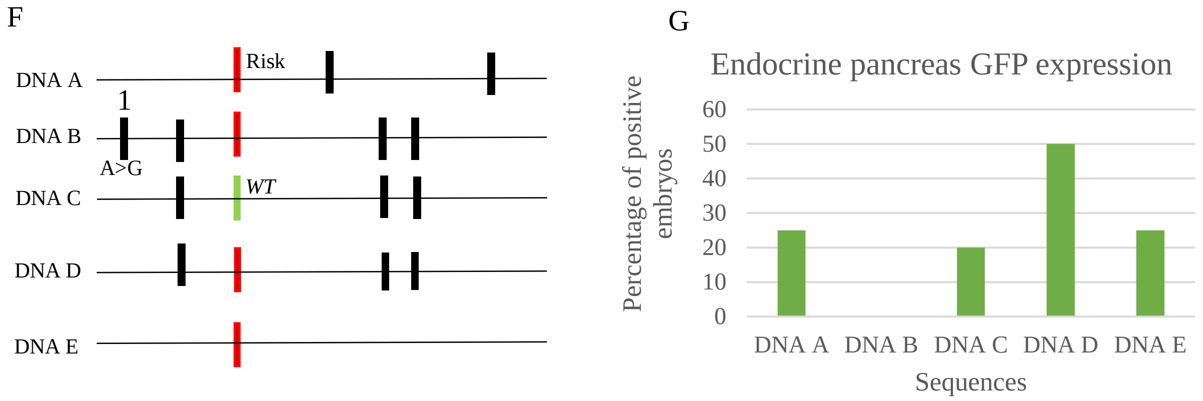


Figure 43 – A-E - Representative image of the different sequences of seq132 with risk allele. The sequence B didn't show GFP expression in endocrine domain. The dashed line represents the endocrine pancreatic domain. Leica confocal SP5II; Zoom 2,91x; Magnification 40x F – Graphic squeme of the sequences correspondent to the different DNAs, including the present SNPs. G – Representative graph showing the total percentage and the number of analyzed embryos in the different sequences injected.

For the case of seq132, we tested five different sequences for enhancer activity in the endocrine pancreas. Each of these sequences, A to E (Fig 43), were exactly the same among each other, with the exception for the SNPs annotated in figure 43 - F. Interestingly, when testing sequences C and D, that only vary in the *WT* and T2D risk allele respectively, it was observed that DNA C (*WT* variant) presented a decreased percentage of embryos with GFP expression in the endocrine pancreas (20%, n=20) when comparing to DNA D (T2D risk variant; 50%, n=20; Fig.43). When testing a new DNA (DNA B) containing a new sequence that contains an extra single nucleotide modification, not described as a common SNP, we observed that the enhancer activity was completely lost (0%, n=20). To better understand molecularly what could be causing the ablation of the enhancer by the single nucleotide modification present in DNA B, we explored differentially putative binding sites of TFs using JASPAR, as previously described, analyzing the DNA A against DNA B (Table 6). Strikingly, the *in silico* prediction showed a putative binding site for PDX1 TF, whose binding site is lost by the presence of the single nucleotide modification present in DNA B (See supplementary data – table 2).

Table6 - JASPAR analysis from the seq132 – DNA B, showing a putative binding site for PDX1. In grey: TFBS motif; In green: WT allele; In purple: New variant; # Complementary chain

TF	Sequence	Affinity score (0-low;1 - high)	Sequence
PDX1	A	0.965432863673	# **GAGTCATTTTTTAGCAGCCTAATGTGTTATCCTTGGCCTGA**
PDX1	B	0	# **GAGTCATTTTTTAGCAGCCCAATGTGTTATCCTTGGCCTGA**
HOXB3	A	0.925700528633	**TCAGGCCAAGGATAACACATTAGGCTGCTAAAAAATGACTC**
HOXB3	B	0	**TCAGGCCAAGGATAACACATTGGGCTGCTAAAAAATGACTC**
HOXB2	A	0.903052079754	**TCAGGCCAAGGATAACACATTAGGCTGCTAAAAAATGACTC**
HOXB2	B	0	**TCAGGCCAAGGATAACACATTGGGCTGCTAAAAAATGACTC**
LHX9	A	0.864562264679	**TCAGGCCAAGGATAACACATTAGGCTGCTAAAAAATGACTC**
LHX9	B	0	**TCAGGCCAAGGATAACACATTGGGCTGCTAAAAAATGACTC**
SOX17	A	0.86255929928	**TCAGGCCAAGGATAACACATTAGGCTGCTAAAAAATGACTC**
SOX17	B	0	**TCAGGCCAAGGATAACACATTGGGCTGCTAAAAAATGACTC**
GSC	A	0.852756075706	# **GAGTCATTTTTTAGCAGCCTAATGTGTTATCCTTGGCCTGA**
GSC	B	0	# **GAGTCATTTTTTAGCAGCCCAATGTGTTATCCTTGGCCTGA**
PRRX2	A	0.843137571821	**TCAGGCCAAGGATAACACATTAGGCTGCTAAAAAATGACTC**
PRRX2	B	0	**TCAGGCCAAGGATAACACATTGGGCTGCTAAAAAATGACTC**
SOX13	A	0	**TCAGGCCAAGGATAACACATTAGGCTGCTAAAAAATGACTC**
SOX13	B	0.829650813997	**TCAGGCCAAGGATAACACATTGGGCTGCTAAAAAATGACTC**

As referred before, PDX1 is an important TF in pancreas development, β-cell differentiation and function. The presence of a single modification disrupted the putative binding site of PDX1, which could be causing the disruption of the enhancer activity of the sequence.

Apart from the differentially affinity for the binding of PDX1 in DNA B, other transcription factors were also identified, namely:

Homeobox 3 (HOXB3), homeobox 2 HOXB2 and lim homeobox 9(LHX9) are homeobox TF directly related to development. Variants that allow to create new binding sites for these 3 homeobox TF are associated to an increase in NAD-dependent deacetylase sirtuin 2 (SIRT2) promoter activity in beta cells, contributing to T2D through diverse pathways as a risk factor. (Liu et al.,2018).

SRY-Box 17 (SOX17) is a key transcriptional regulator that can act by regulating other transcription factors including HNF1β and FOXA2, which are known to regulate postnatal β-cell function. SOX17 has a critical role in regulating insulin trafficking and secretion (Jonatan et al., 2014).

Paired mesoderm homeobox protein 2 (PRRX2) is a TF that is involved in adipocyte differentiation. An impaired adipogenesis may underlie the development of diabetes (Du et al., 2013).

Because PDX1 has the highest relative score and because of its very well-known function in the pancreas, this is the best candidate to explain the loss of the enhancer activity in DNA B. In addition, looking to ChIP-seq data from endocrine pancreas (Pasquali et al, 2014), we are able to identify a clear enrichment for the binding of PDX1 in this sequence, further supporting the *in silico* prediction for the binding of PDX1 (Figure 44). Further studies of this new single nucleotide modification may give interesting insights about new genetic modifications that might impact in T2D. For this, it would be interesting to: 1) analyze if this single nucleotide modification is present in the human population and with which frequency, 2) determine is this single nucleotide modification is more or less prevalent in T2D patients.

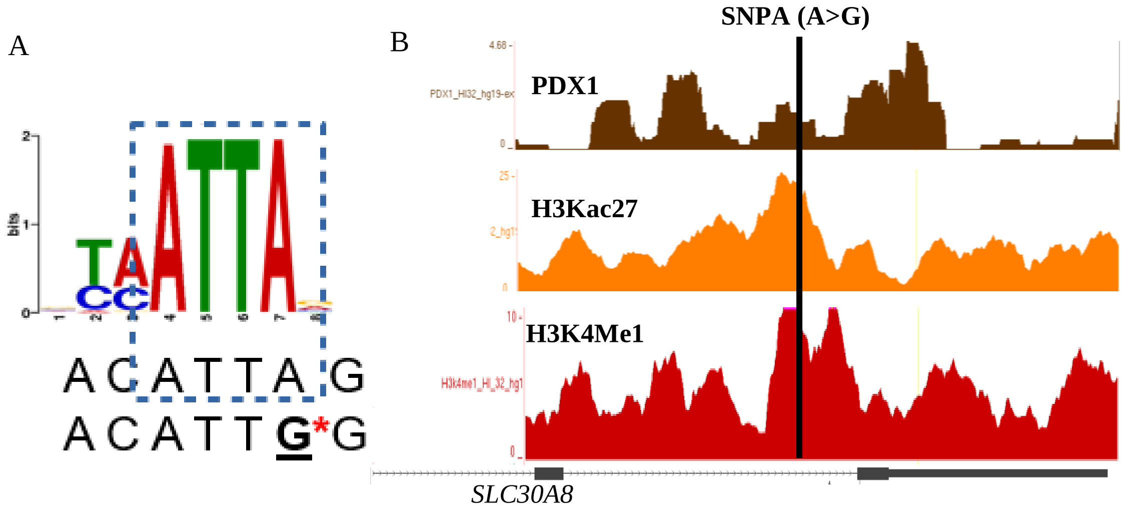


Figure 44 – A- Resulted prediction by JASPAR analysis, showing the new variant (*) located in PDX1 binding site. B- ChIP analysis including the newly discovered variant, overlapping with acetylation and methylation signals, as PDX1 signal.

The interesting results obtained by the analysis of different variants of seq132 lead us to further explore this sequence. We hypothesize that seq132 might have different topological regions that confer the enhancer activity. To evaluate these topological regions, we have fragmented the sequence and tested the different fragments for enhancer activity in endocrine pancreas (Fig.45). Curiously, fragments 1 (9,8%; n=43), 2 (6,5%; n=31) and 1 extended (10%; n=30) showed a decreased enhancer activity when comparing to the total fragment (20%; n=30; Fig45). Nevertheless, these results demonstrate that these fragments can work independently of each other. In addition, it was surprising to find that fragment 2 was able to drive enhancer activity on its own. This is surprising due to previous experiments with the total fragment with the mutated putative binding site of PDX1, that suggested that this binding site, present in fragment 1, is necessary to the activity of the total fragment. This hypothesis is contradicted by the results obtained by fragment 2, that do not contain the binding site of PDX1. Therefore, the putative binding site of PDX1 should not be necessary for the activity of the enhancer. As an alternative explanation that is coherent with all the presented results, is that the single nucleotide modification might ablate the putative binding of PDX1, generating another putative binding of a transcriptional repressor, which could explain the loss of the enhancer activity observed in DNA B.

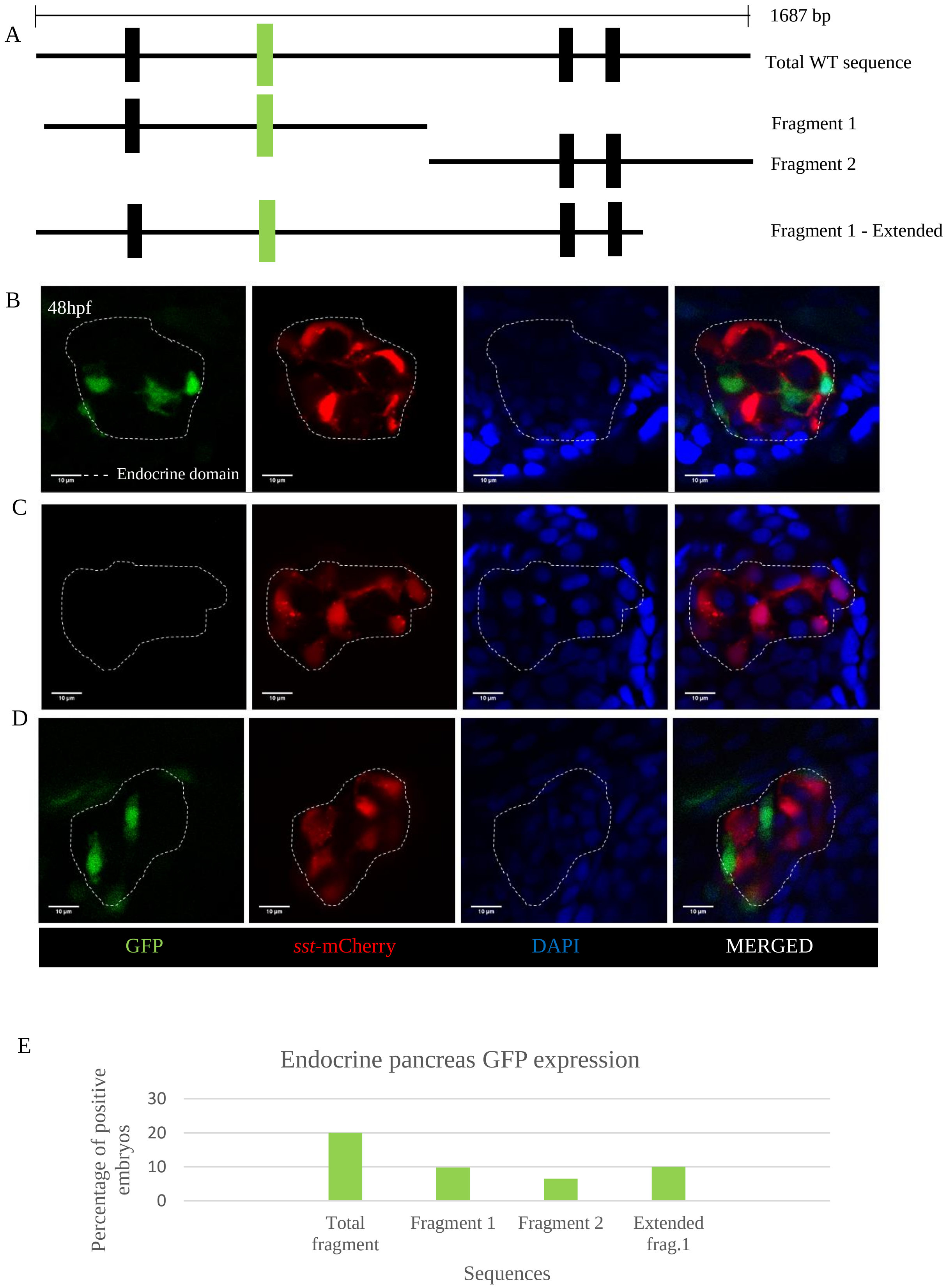


Figure 45 – A- Graphic squeme of the total sequence and the fragments. B-D – Representative confocal images of the different fragments (B – fragment 1; C – Fragment 2; D – Extended fragment 1). The dashed line represents the endocrine pancreatic domain .Leica confocal SP5II; Zoom 2,91x; Magnification 40x. E– Representative graph showing the total percentage in the different fragments and in total sequence.

IV. General conclusions and future perspectives

The transcriptional regulatory mechanism of gene expression is required for a proper cell and tissue function (Alberts, 2002). In this sense, a disruption in this mechanism might result in several diseases (Kleinjan & Coutinho, 2009), such as T2D. The elements that control this transcriptional mechanism are located mostly in non-coding genome and are named as CREs. Among these elements, there are endocrine pancreas enhancers that can serve as target site for TFBS (Pennacchio et al., 2013) and might be important for endocrine pancreas islets function and development. Enhancers can act in long range, by chromatin loops, allowing the interaction between the TF with the promoters of the enhancer target gene (Mora et al., 2015).

The progressive findings in enhancers function and associated fingerprints allow their identification in the genome. This identification and the study of such regulatory regions have become important due to the association of variations in enhancers to transcriptional dysregulation of genes, phenotypic alterations and disease (Maston et al., 2006; Pennacchio et al., 2013).

Currently, several studies have shown that enhancers are enriched in SNPs associated to several diseases (Dunham et al., 2012; Hindorff et al., 2009; Maurano et al., 2012; Trynka et al., 2013), such as T2D (Mohlke & Boehnke, 2015; Pasquali et al., 2014). The study of putative enhancers that contains T2D associated SNPs is poorly known due to the lack of investigations and validations *in vivo*.

The main hypothesis of this work was that SNPs might impair TFBS resulting in a modulation of enhancer activity in endocrine islets impacting in their function, having the potential to contribute for disease susceptibility. The first question that we have addressed in this work was to understand to what extend sequences that overlap with SNPs associated to T2D could be endocrine enhancers. We selected 10 of such sequences based on epigenetic marks of enhancer activity and TFBS. Then we tested if these 10 sequences were or not pancreas endocrine enhancers, having found that this was the case for 5 out of the 10 tested sequences. Then, we wanted to analyze the impact of the presence of the T2D associated variant in enhancer activity. For one sequence, we were able to demonstrate the impact of a T2D associated variant in enhancer activity by transgenesis assays, supporting the observations by analyzing *in silico* differentially affinities for the binding of TFs.

Furthermore, we were capable to discover a new single nucleotide modification, located in a PDX1 binding site, that ablated the enhancer activity of one of the tested sequences.

Overall, this project helped us to understand the importance of non-coding regions in transcriptional regulation and the impact in these machinery by the presence of SNPs in T2D.

As future steps, we will aim to continue studying the impact of the associated SNPs to T2D and other putative SNPs not described yet, identify the enhancer`s target genes by 4C and test the enhancer activity in human β-cell lines.

V. Bibliographic references

Ahlgren, U., Jonsson, J., & Edlund, H. (1996). The morphogenesis of the pancreatic mesenchyme is uncoupled from that of the pancreatic epithelium in IPF1/PDX1-deficient mice. *Development (Cambridge, England)*, *122*(5), 1409–16. https://doi.org/10.1016/0092-8674(88)90391-1

Ahlgren, U., Jonsson, J., & Jonsson, L. (1998). β -Cell-specific inactivation of the mouse Ipf1 / Pdx1 gene results in loss of the β -cell phenotype and maturity onset ? diabetes service of the mouse Ipf1 / Pdx1 gene results in loss of the β -cell phenotype and maturity onset diabetes. *Genes & Development*, 1763–1768. https://doi.org/10.1101/gad.12.12.1763

Alberts B, Johnson A. (2002). An Overview of Gene Control. *Molecular Biology of the Cell*. (4th ed). New York: Garland Science

Aerts, S. (2012). *Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets*. *Current Topics in Developmental Biology* (1st ed., Vol. 98). Elsevier Inc. https://doi.org/10.1016/B978-0-12-386499-4.00005-7

Ahlgren, U., Jonsson, J., & Edlund, H. (1996). The morphogenesis of the pancreatic mesenchyme is uncoupled from that of the pancreatic epithelium in IPF1/PDX1-deficient mice. *Development (Cambridge, England)*, *122*(5), 1409–16. https://doi.org/10.1016/0092-8674(88)90391-1

Ahonen, T., Saltevo, J., Laakso, M., Kautiainen, H., Kumpusalo, E., & Vanhala, M. (2009). Gender differences relating to metabolic syndrome and proinflammation in finnish subjects with elevated blood pressure. *Mediators of Inflammation*, *2009*. https://doi.org/10.1155/2009/959281

Alejandro, E. U., Gregg, B., Blandino-Rosano, M., Cras-Meneur, C., & Bernal-Mizrachi, E. (2014). Natural history of beta-cell adaptation and failure in type 2 diabetes. *Mol Aspects Med*, *1*(734), 19–41. https://doi.org/10.1016/j.mam.2014.12.002

Ali, T., Renkawitz, R., & Bartkuhn, M. (2016). Insulators and domains of gene expression. *Current Opinion in Genetics and Development*, *37*, 17–26.

https://doi.org/10.1016/j.gde.2015.11.009

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Amsterdam, A., & Hopkins, N. (2006). Mutagenesis strategies in zebrafish for identifying genes involved in development and disease. *Trends in Genetics*, *22*(9), 473–478. https://doi.org/10.1016/j.tig.2006.06.011

Andersson, L. E., Valtat, B., Bagge, A., Sharoyko, V. V., Nicholls, D. G., Ravassard, P., … Mulder, H. (2015). Characterization of stimulus-secretion coupling in the human pancreatic EndoC-βH1 beta cell line. *PLoS ONE*, *10*(3), 1–18. https://doi.org/10.1371/journal.pone.0120879

Andrey, G., Schï¿½pflin, R., Jerković, I., Heinrich, V., Ibrahim, D. M., Paliou, C., … Mundlos, S. (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Research*, *27*(2), 223–233. https://doi.org/10.1101/gr.213066.116

Arda, H. E., Benitez, C. M., & Kim, S. K. (2013). Gene regulatory networks governing pancreas development. *Developmental Cell*, *25*(1), 5–13. https://doi.org/10.1016/j.devcel.2013.03.016

Atchison, M. L. (1988). Enhancers: Mechanisms of Action and Cell Specificity. *Annual Review of Cell Biology*, *4*(1), 127–153. https://doi.org/10.1146/annurev.cb.04.110188.001015

Avrahami, D., & Kaestner, K. H. (2012). Epigenetic regulation of pancreas development and function. *Seminars in Cell and Developmental Biology*, *23*(6), 693–700. https://doi.org/10.1016/j.semcdb.2012.06.002

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 PART 1), 299–308. https://doi.org/10.1016/0092-8674(81)90413-X

Barrett, L. W., Fletcher, S., & Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular* *and Molecular Life Sciences*, *69*(21), 3613–3634. https://doi.org/10.1007/s00018-012-0990-9

Bastidas-Ponce, A., Scheibner, K., Lickert, H., & Bakhti, M. (2017). Cellular and molecular mechanisms coordinating pancreas development. *Development*, *144*(16), 2873–2888. https://doi.org/10.1242/dev.140756

Bessa, J., Tena, J. J., De La Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., … Gómez-Skarmeta, J. L. (2009). Zebrafish Enhancer Detection (ZED) vector: A new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Developmental Dynamics*, *238*(9), 2409–2417. https://doi.org/10.1002/dvdy.22051

Bhushan, A., Rizza, R. A., & Butler, P. C. (2013). Postnatal Expansion of β -Cell Mass in Humans, *57*(6), 1584–1594. https://doi.org/10.2337/db07-1369.

Biemar, F., Argenton, F., Schmidtke, R., Epperlein, S., Peers, B., & Driever, W. (2001). Pancreas development in zebrafish: Early dispersed appearance of endocrine hormone expressing cells and their convergence to form the definitive islet. *Developmental Biology*, *230*(2), 189–203. https://doi.org/10.1006/dbio.2000.0103

Billings, L. K., & Florez, J. C. (2010). The genetics of type 2 diabetes: What have we learned from GWAS? *Annals of the New York Academy of Sciences*, *1212*, 59–77. https://doi.org/10.1111/j.1749-6632.2010.05838.x

Binot, A. C., Manfroid, I., Flasse, L., Winandy, M., Motte, P., Martial, J. A., … Voz, M. L. (2010). Nkx6.1 and nkx6.2 regulate α- and β-cell formation in zebrafish by acting on pancreatic endocrine progenitor cells. *Developmental Biology*, *340*(2), 397–407. https://doi.org/10.1016/j.ydbio.2010.01.025

Blackwood, E. M., Kadonaga, J. T., Blackwood, E. M., & Kadonaga, J. T. (2016). Going the Distance : A Current View of Enhancer Action Linked references are available on JSTOR for this article : Going the Distance : A Current View of Enhancer Action, *281*(5373), 60–63.

Bramswig NC., Kaestner, KH. (2011). Transcriptional regulation of α-cell differentiation. https://doi.org/10.1111/j.1463-1326.2011.01440.x

Bu, H., Gan, Y., Wang, Y., Zhou, S., & Guan, J. (2017). A new method for enhancer prediction based on deep belief network. *BMC Bioinformatics*, *18*(Suppl 12). https://doi.org/10.1186/s12859-017-1828-0

Bulger, M., & Groudine, M. (1999). Looping versus linking: Toward a model for long-distance gene activation. *Genes and Development*, *13*(19), 2465–2477. https://doi.org/10.1101/gad.13.19.2465

Butler, J. E. F., & Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, *16*(20), 2583–2592. https://doi.org/10.1101/gad.1026202.

Cebola, I., Rodríguez-seguí, S. A., Cho, C. H., Bessa, J., Maestro, M. A., Jennings, R. E., & Pasquali, L. (2015). TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors, *17*(5), 615–626. https://doi.org/10.1038/ncb3160.TEAD

Chatterjee, S., Bourque, G., & Lufkin, T. (2011). Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes. *BMC Developmental Biology*, *11*(1), 63. https://doi.org/10.1186/1471-213X-11-63

Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. *The Lancet*, *389*(10085), 2239–2251. https://doi.org/10.1016/S0140-6736(17)30058-2

Chen, L., & Widom, J. (2005). Mechanism of transcriptional silencing in yeast. *Cell*, *120*(1), 37–48. https://doi.org/10.1016/j.cell.2004.11.030

Choi, J. W., Moon, S., Jang, E. J., Lee, C. H., & Park, J. S. (2017). Association of prediabetes-associated single nucleotide polymorphisms with microalbuminuria. *PLoS ONE*, *12*(2), 1–13. https://doi.org/10.1371/journal.pone.0171367

Chung, J. H., Whiteley, M., & Felsenfeld, G. (1993). A 5′ element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, *74*(3), 505–514. https://doi.org/10.1016/0092-8674(93)80052-G

Collins, F. S., Lander, E. S., Rogers, J., & Waterson, R. H. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Collombat, P. (2005). The simultaneous loss of Arx and Pax4 genes promotes a somatostatin-producing cell fate specification at the expense of the - and -cell lineages in the mouse endocrine pancreas. *Development*, *132*(13), 2969–2980. https://doi.org/10.1242/dev.01870

Coppola, C. J., Ramaker, R. C., & Mendenhall, E. M. (2016). Identification and function of enhancers in the human genome. *Human Molecular Genetics*, *25*(R2), R190–R197. https://doi.org/10.1093/hmg/ddw216

Cosma, M. P. (2002). Ordered recruitment: gene-specific mechanism of transcription activation. *Molecular Cell*, *10*(2), 227–36. https://doi.org/10.1016/S1097-2765(02)00604-4

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., … Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, *107*(50), 21931–21936. https://doi.org/10.1073/pnas.1016071107

Cristancho, A. G., Schupp, M., Lefterova, M. I., Cao, S., Cohen, D. M., Chen, C. S., … Lazar, M. A. (2011). Repressor transcription factor 7-like 1 promotes adipogenic competency in precursor cells. *Proceedings of the National Academy of Sciences*, *108*(39), 16271–16276. https://doi.org/10.1073/pnas.1109409108

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T. Y., Cui, K., & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, *19*(1), 24–32. https://doi.org/10.1101/gr.082800.108

Dandona, P., & Dhindsa, S. (2011). Update: Hypogonadotropic Hypogonadism in Type 2 Diabetes and Obesity. *The Journal of Clinical Endocrinology & Metabolism*, *96*(9), 2643–2651. https://doi.org/10.1210/jc.2010-2724

Davis, E. E., Frangakis, S., & Katsanis, N. (2014). Interpreting human genetic variation with in vivo zebrafish assays. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1842*(10), 1960–1970. https://doi.org/10.1016/j.bbadis.2014.05.024

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., … Pritchard, J. K. (2012). DNase-I sensitivity QTLs are a major determinant of human expression variation. *Nature, 482*(7385), 390–394. https://doi.org/10.1038/nature10808

Delic J., Onclercq R., Moisan-Coppey M.. (1991) Inhibition and enhancement of eucaryotic gene expression by potential non-B DNA sequences., *180*(3), 1273–1283.

Delporte, F. M., Pasque, V., Devos, N., Manfroid, I., Voz, M. L., Motte, P., … Peers, B. (2008). Expression of zebrafish pax6b in pancreas is regulated by two enhancers containing highly conserved cis-elements bound by PDX1, PBX and PREP factors. *BMC Developmental Biology, 8,* 1–19. https://doi.org/10.1186/1471-213X-8-53

Dekker J., Rippe K., Dekker M., Kleckner N. (2002) Capturing chromossome conformation. https://doi.org/10.1126/science.1067799.

Dorschner, M. O., Hawrylycz, M., Humbert, R., Wallace, J. C., Shafer, A., Kawamoto, J., … Stamatoyannopoulos, J. A. (2004). High-throughput localization of functional elements by quantitative chromatin profiling. *Nature Methods, 1*(3), 219–225. https://doi.org/10.1038/nmeth721

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., … Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements, 1299–1309. https://doi.org/10.1101/gr.5571506.1

Du, B., Cawthorn, W. P., Su, A., Doucette, C. R., Yao, Y., Hemati, N., … Macdougald, O. A. (2013). The transcription factor paired-related homeobox 1 (Prrx1) inhibits adipogenesis by activating transforming growth factor-β (TGFβ) signaling. *Journal of Biological Chemistry, 288*(5), 3036–3047. https://doi.org/10.1074/jbc.M112.440370

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57–74. https://doi.org/10.1038/nature11247

Edalat, F. (2012). Engineering approaches toward deconstructing and controlling the stem cell environment, *40*(6), 1301–1315. https://doi.org/10.1007/s10439-011-0452-9.

Emison, E. S., McCallion, A. S., Kashuk, C. S., Bush, R. T., Grice, E., Lin, S., … Chakravarti, A. (2005). A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, *434*(7035), 857–863. https://doi.org/10.1038/nature03467

Engeszer, R. E., Patterson, L. B., Rao, A. A., & Parichy, D. M. (2007). Zebrafish in The Wild: A Review of Natural History And New Notes from The Field. *Zebrafish*, *4*(1), 21–40. https://doi.org/10.1089/zeb.2006.9997

Ernst, J. et al. (2011). Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types. *Nature*, *473*(7345), 43–49. https://doi.org/10.1038/nature09906.Systematic

Evans, J. a. (2007). Diaphragmatic defects and limb deficiencies - taking sides. *American Journal of Medical Genetics. Part A*, *143A*(18), 2106–2112. https://doi.org/10.1002/ajmg.a

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L., & McCallion, A. S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science*, *312*(5771), 276–279. https://doi.org/10.1126/science.1124070

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B. R., Grarup, N., Burtt, N. P., … Altshuler, D. (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nature Genetics*, *46*(4), 357–363. https://doi.org/10.1038/ng.2915

Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., & Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. *Genome Research*, *18*(1), 1–12. https://doi.org/10.1101/gr.6831208

Fu, J., Tay, S. S. W., Ling, E. a, & Dheen, S. T. (2006). High glucose alters the expression of genes involved in proliferation and cell-fate specification of embryonic neural stem cells. *Diabetologia*, *49*(5), 1027–38. https://doi.org/10.1007/s00125-006-0153-3

Ghaye, A. P., Bergemann, D., Tarifeño-Saldivia, E., Flasse, L. C., Von Berg, V., Peers, B., … Manfroid, I. (2015). Progenitor potential of nkx6.1-expressing cells throughout zebrafish life and during beta cell regeneration. *BMC Biology*, *13*(1), 1–24. https://doi.org/10.1186/s12915-015-0179-4

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE

(Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research, 17*(6), 877–885. https://doi.org/10.1101/gr.5533506

Giresi, P. G., & Lieb, J. D. (2012). Isolation of Active Regulatory Elements from Eukaryotic Chromatin Using FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements). *Tag-Based Next Generation Sequencing, 48*(3), 243–255. https://doi.org/10.1002/9783527644582.ch14

Gittes, G. K., Galante, P. E., Hanahan, D., Rutter, W. J., & Debase, H. T. (1996). Lineage-specific morphogenesis in the developing pancreas: role of mesenchymal factors. *Development (Cambridge, England), 122*(2), 439–447.

Gloyn, A. L. (2003). Glucokinase (GCK) Mutations in Hyper- and Hypoglycemia: Maturity-Onset Diabetes of the Young, Permanent Neonatal Diabetes, and Hyperinsulinemia of Infancy. *Human Mutation, 22*(5), 353–362. https://doi.org/10.1002/humu.10277

Gloyn, A. L., Braun, M., & Rorsman, P. (2009). Type 2 diabetes susceptibility gene TCF7L2 and its role in ??-cell function. *Diabetes, 58*(4), 800–802. https://doi.org/10.2337/db09-0099

Grice, E. A., Rochelle, E. S., Green, E. D., Chakravarti, A., & McCallion, A. S. (2005). Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. *Human Molecular Genetics, 14*(24), 3837–3845. https://doi.org/10.1093/hmg/ddi408

Grunwald, D. J., & Eisen, J. S. (2002). Headwaters of the zebrafish — emergence of a new model vertebrate. Nature Reviews Genetics, 3(9), 717–724. doi:10.1038/nrg892

Habener, J. F., Kemp, D. M., & Thomas, M. K. (2005). Minireview: Transcriptional regulation in pancreatic development. *Endocrinology, 146*(3), 1025–1034. https://doi.org/10.1210/en.2004-1576

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural and Molecular Biology, 11*(5), 394–403. https://doi.org/10.1038/nsmb763

Haldeman, S. D. (n.d.). An Atlas of, 13–16.

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews : MMBR*, *62*(2), 465–503. https://doi.org/10.1158/0008-5472.CAN-07-2721

Harder, M. N., Appel, E. V. R., Grarup, N., Gjesing, A. P., Ahluwalia, T. S., Jørgensen, T., … Hansen, T. (2015). The type 2 diabetes risk allele of TMEM154-rs6813195 associates with decreased beta cell function in a study of 6,486 danes. *PLoS ONE*, *10*(3), 1–13. https://doi.org/10.1371/journal.pone.0120890

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., & Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics*, *4*(6). https://doi.org/10.1371/journal.pgen.1000106

Harris, M. B., Mostecki, J., & Rothman, P. B. (2005). Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *Journal of Biological Chemistry*, *280*(13), 13114–13121. https://doi.org/10.1074/jbc.M412649200

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–112. https://doi.org/10.1038/nature07829

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., … Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. https://doi.org/10.1038/ng1966

Heinzel, T., Lavinsky, R. M., Mullen, T. M., Söderstrom, M., Laherty, C. D., Torchia, J., … Rosenfeld, M. G. (1997). A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. *Nature*. https://doi.org/10.1038/387043a0

Herrera, P. L. (2002). Defining the cell lineages of the islets of Langerhans using transgenic mice. *Int J Dev Biol*, *46*(1), 97–103.

Herrera, P. L., Nepote, V., & Delacour, A. (2002). Pancreatic cell lineage analyses in mice. *Endocrine*, *19*(3), 267–277. https://doi.org/10.1385/ENDO:19:3:267

Heuvel, A. Van Den, Stadhouders, R., Andrieu-soler, C., Grosveld, F., & Soler, E. (2015).

Blood Spotlight Long-range gene regulation and novel therapeutic applications. *Blood Spotlight*, *125*(10), 1521–1525. https://doi.org/10.1182/blood-2014-11-567925.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Hiramoto, M., Udagawa, H., Ishibashi, N., Takahashi, E., Kaburagi, Y., Miyazawa, K., … Yasuda, K. (2018). A type 2 diabetes-associated SNP in KCNQ1 (rs163184) modulates the binding activity of the locus for Sp3 and Lsd1/Kdm1a, potentially affecting CDKN1C expression. *International Journal of Molecular Medicine*, *41*(2), 717–728. https://doi.org/10.3892/ijmm.2017.3273

Hirose, Y., & Ohkuma, Y. (2007). Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. *Journal of Biochemistry*, *141*(5), 601–608. https://doi.org/10.1093/jb/mvm090

Horikoshi, M., Hara, K., Ito, C., Shojima, N., Nagai, R., Ueki, K., … Kadowaki, T. (2007). Variations in the HHEX gene are associated with increased risk of type 2 diabetes in the Japanese population. *Diabetologia*, *50*(12), 2461–2466. https://doi.org/10.1007/s00125-007-0827-5

Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K., … Westerfield, M. (2013). ZFIN, the Zebrafish Model Organism Database: Increased support for mutants and transgenics. *Nucleic Acids Research*, *41*(D1), 854–860. https://doi.org/10.1093/nar/gks938

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Huang, H., Vogel, S. S., Liu, N., Melton, D. A., & Lin, S. (2001). Analysis of pancreatic development in living transgenic zebrafish embryos. *Molecular and Cellular Endocrinology*, *177*(1–2), 117–124. https://doi.org/10.1016/S0303-7207(01)00408-7

Hubner, M. R., & Spector, D. L. (2010). Chromatin Dynamics. *Annual Review of Biophysics*, *448*(7153), 471–489. https://doi.org/10.1038/nature06008.Genome-wide

Ishihara, H., Asano, T., Tsukuda, K., Katagiri, H., Inukai, K., Anai, M., … Oka, Y. (1993). Pancreatic beta cell line MIN6 exhibits characteristics of glucose metabolism and glucose-stimulated insulin secretion similar to those of normal islets. *Diabetologia*, *36*(11), 1139–1145. https://doi.org/10.1007/BF00401058

Istrail, S., & Davidson, E. H. (2005). Logic functions of the genomic cis-regulatory code. *Proceedings of the National Academy of Sciences*, *102*(14), 4954–4959. https://doi.org/10.1073/pnas.0409624102

Itkin-Ansari, P., Marcora, E., Geron, I., Tyrberg, B., Demeterco, C., Hao, E., … Levine, F. (2005). NeuroD1 in the endocrine pancreas: Localization and dual function as an activator and repressor. *Developmental Dynamics*, *233*(3), 946–953. https://doi.org/10.1002/dvdy.20443

Jennings, R. E., Berry, A. A., Strutt, J. P., Gerrard, D. T., & Hanley, N. A. (2015). Human pancreas development. *Development*, *142*(18), 3126–3137. https://doi.org/10.1242/dev.120063

Jonatan, D., Spence, J. R., Method, A. M., Kofron, M., Sinagoga, K., Haataja, L., … Wells, J. M. (2014). Sox17 regulates insulin secretion in the normal and pathologic mouse β cell. *PLoS ONE*, *9*(8), 1–16. https://doi.org/10.1371/journal.pone.0104675

Jörgens, K., Stoll, S. J., Pohl, J., Fleming, T. H., Sticht, C., Nawroth, P. P., … Kroll, J. (2015). High tissue glucose alters intersomitic blood vessels in zebrafish via methylglyoxal targeting the VEGF receptor signaling cascade. *Diabetes*, *64*(1), 213–225. https://doi.org/10.2337/db14-0352

Juven-Gershon, T., & Kadonaga, J. T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology*, *339*(2), 225–229. https://doi.org/10.1016/j.ydbio.2009.08.009

Kadauke, S., & Blobel, G. A. (2009). Chromatin loops in gene regulation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1789*(1), 17–25. https://doi.org/10.1016/j.bbagrm.2008.07.002

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., Berkum, N. L. Van, … Young, R. A. (2011). Mediator and cohesin connect gene expression and chromatin architecture , *467*(7314), 430–435. https://doi.org/10.1038/nature09380.

Kaphingst, K. A., Persky, S., & Lachance, C. (2010). Testing communication strategies to convey genomic concepts using virtual reality techonology, *14*(4), 384–399. https://doi.org/10.1080/10810730902873927.

Kawakami, K. (2007). Tol2: A versatile gene transfer vector in vertebrates. *Genome Biology*, *8*(SUPPL. 1), 1–10. https://doi.org/10.1186/gb-2007-8-s1-s7

Kawakami, K., Takeda, H., Kawakami, N., Kobayashi, M., Matsuda, N., & Mishina, M. (2004). A Transposon-Mediated Gene Trap Approach Identifies Developmentally Regulated Genes in Zebrafish, *7*, 133–144. https://doi.org/10.1016/j.devcel.2004.06.005

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., … Mathelier, A. (2018). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, *46*(D1), D260–D266. https://doi.org/10.1093/nar/gkx1126

Kim, W., Shin, Y. K., Kim, B. J., & Egan, J. M. (2010). Notch signaling in pancreatic endocrine cell and diabetes. *Biochemical and Biophysical Research Communications*, *392*(3), 247–251. https://doi.org/10.1016/j.bbrc.2009.12.115

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., & Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics : An Official Public*, *203*(3), 253–310. https://doi.org/10.1002/aja.1002030302

Kimmel, R. A., Dobler, S., Schmitner, N., Walsen, T., Freudenblum, J., & Meyer, D. (2015). Diabetic pdx1-mutant zebrafish show conserved responses to nutrient overload and anti-glycemic treatment. *Scientific Reports*, *5*(September), 1–14. https://doi.org/10.1038/srep14241

Kinkel, M. D., & Prince, V. E. (2009). On the diabetic menu: Zebrafish as a model for pancreas development and function. *BioEssays*, *31*(2), 139–152. https://doi.org/10.1002/bies.200800123

Kirchhoff, K., Machicao, F., Haupt, A., Schäfer, S. A., Tschritter, O., Staiger, H., … Fritsche, A. (2008). Polymorphisms in the TCF7L2, CDKAL1 and SLC30A8 genes are associated with impaired proinsulin conversion. *Diabetologia*, *51*(4), 597–601. https://doi.org/10.1007/s00125-008-0926-y

Kleinjan, D. J., & Coutinho, P. (2009). Cis-ruption mechanisms: Disruption of cis-regulatory control as a cause of human genetic disease. *Briefings in Functional Genomics and Proteomics*, *8*(4), 317–332. https://doi.org/10.1093/bfgp/elp022

Kulkarni, R. N. (2004). The islet β-cell. *International Journal of Biochemistry and Cell Biology*, *36*(3), 365–371. https://doi.org/10.1016/j.biocel.2003.08.010

Lake, R. J., Tsai, P. F., Choi, I., Won, K. J., & Fan, H. Y. (2014). RBPJ, the Major Transcriptional Effector of Notch Signaling, Remains Associated with Chromatin throughout Mitosis, Suggesting a Role in Mitotic Bookmarking. *PLoS Genetics*, *10*(3). https://doi.org/10.1371/journal.pgen.1004204

Laukkanen, O., Lindström, J., Eriksson, J., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., … Laakso, M. (2005). Polymorphisms in the SLC2A2 (GLUT2) gene are associated with the conversion from impaired glucose tolerance to type 2 diabetes: The Finnish Diabetes Prevention Study. *Diabetes*, *54*(7), 2256–2260. https://doi.org/10.2337/diabetes.54.7.2256

Lee, T. I., & Young, R. a. (2000). Transcription of Eukaryotic Protein- Coding Genes. *Annual Review of Genetics*, 77–137.

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Lee JE, Hollenberg SM, Snider L, Turner DL, Lipnick N, Weintraub H. (1995) Conversion of Xenopus ectoderm into neurons by NeuroD, a basic helix-loop-helix protein. Science 1995;268:836–844. [PubMed: 7754368]

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., … de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–1735. https://doi.org/10.1093/hmg/ddg180

Lettice, L. A., Horikoshi, T., Heaney, S. J. H., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., … Noji, S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences*, *99*(11), 7548–7553. https://doi.org/10.1073/pnas.112212199

Levine, M., Cattoglio, C., & Tjian, R. (2014). Looping back to leap forward: Transcription enters a new era. *Cell, 157*(1), 13–25. https://doi.org/10.1016/j.cell.2014.02.009

Li, M. J., Wang, P., Liu, X., Lim, E. L., Wang, Z., Yeager, M., … Wang, J. (2012). GWASdb: A database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Research*, *40*(D1), 1047–1054. https://doi.org/10.1093/nar/gkr1182

Li, M. J., Yan, B., Sham, P. C., & Wang, J. (2014). Exploring the function of genetic variants in the non-coding genomic regions: Approaches for identifying human regulatory variants affecting gene expression. *Briefings in Bioinformatics*, *16*(3), 393–412. https://doi.org/10.1093/bib/bbu018

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., & Kadonaga, J. T. (2004). The MTE , a new core promoter element for transcription by RNA polymerase II. *Genes and Development*, *32*, 1606–1617. https://doi.org/10.1101/gad.1193404.interactions

Liu, T., Yang, W., Pang, S., Yu, S., & Yan, B. (2018). Functional genetic variants within the SIRT2 gene promoter in type 2 diabetes mellitus. *Diabetes Research and Clinical Practice*, *137*, 200–207. https://doi.org/10.1016/j.diabres.2018.01.012

Lu, T. T. H., Heyne, S., Dror, E., Casas, E., Leonhardt, L., Boenke, T., … Pospisilik, J. A. (2018). The Polycomb-Dependent Epigenome Controls β Cell Dysfunction, Dedifferentiation, and Diabetes. *Cell Metabolism*, 1294–1308. https://doi.org/10.1016/j.cmet.2018.04.013

Mandal, N., Su, W., Haber, R., Adhya, S., & Echols, H. (1990). DNA looping in cellular repression of transcription of the galactose operon. *Genes and Development*, *4*(3), 410–418. https://doi.org/10.1101/gad.4.3.410

Manichaikul, A., Wang, X.-Q., Zhao, W., Wojczynski, M. K., Siebenthall, K., Stamatoyannopoulos, J. A., … Bornfeldt, K. E. (2016). Genetic association of long-chain acyl-CoA synthetase 1 variants with fasting glucose, diabetes, and subclinical atherosclerosis. *Journal of Lipid Research*, *57*(3), 433–442. https://doi.org/10.1194/jlr.M064592

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, *7*(1), 29–59.

https://doi.org/10.1146/annurev.genom.7.080505.115623

Mastracci, T. L., Anderson, K. R., Papizan, J. B., & Sussel, L. (2013). Regulation of Neurod1 Contributes to the Lineage Potential of Neurogenin3+ Endocrine Precursor Cells in the Pancreas. *PLoS Genetics*, *9*(2), 1–14. https://doi.org/10.1371/journal.pgen.1003278

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associate Variation in Regulatorty DNA. *Science*, *337*(September), 1190–1195. https://doi.org/10.1126/science.1222794

McGaughey, D. M., Stine, Z. E., Huynh, J. L., Vinton, R. M., & McCallion, A. S. (2009). Asymmetrical distribution of non-conserved regulatory sequences at PHOX2B is reflected at the ENCODE loci and illuminates a possible genome-wide trend. *BMC Genomics*, *10*, 1–14. https://doi.org/10.1186/1471-2164-10-8

Menting, J. G., Yang, Y., Chan, S. J., Phillips, N. B., Smith, B. J., Whittaker, J., … Lawrence, M. C. (2014). Protective hinge in insulin opens to enable its receptor engagement. *Proceedings of the National Academy of Sciences*, *111*(33), E3395–E3404. https://doi.org/10.1073/pnas.1412897111

Milewski, W. M., Duguay, S. J., Chan, S. J., & Steiner, D. F. (2014). Expression in Zebrafish *, *139*(3), 1440–1449.

Mohlke, K. L., & Boehnke, M. (2015). Recent advances in understanding the genetic architecture of type 2 diabetes. *Human Molecular Genetics*, *24*(R1), R85–R92. https://doi.org/10.1093/hmg/ddv264

Mohlke, K. L., & Scott, L. J. (2012). What will diabetes genomes tell us? *Current Diabetes Reports*, *12*(6), 643–650. https://doi.org/10.1007/s11892-012-0321-4

Mora, A., Sandve, G. K., Gabrielsen, O. S., & Eskeland, R. (2015a). In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, *17*(April), bbv097. https://doi.org/10.1093/bib/bbv097

Mora, A., Sandve, G. K., Gabrielsen, O. S., & Eskeland, R. (2015b). In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, *17*(November 2015), bbv097. https://doi.org/10.1093/bib/bbv097

Morris, A., Voight, B., & Teslovich, T. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990. https://doi.org/10.1038/ng.2383.Large-scale

Mularoni, L., Ramos-Rodríguez, M., & Pasquali, L. (2017). The pancreatic islet regulome browser. *Frontiers in Genetics*, *8*(FEB), 1–8. https://doi.org/10.3389/fgene.2017.00013

Müller, D., & Stelling, J. (2009). Precise regulation of gene expression dynamics favors complex promoter architectures. *PLoS Computational Biology*, *5*(1). https://doi.org/10.1371/journal.pcbi.1000279

Muller, Y. L., Piaggi, P., Chen, P., Wiessner, G., Okani, C., Kobes, S., … Baier, L. J. (2017). Assessing variation across 8 established East Asian loci for type 2 diabetes mellitus in American Indians: Suggestive evidence for new sex-specific diabetes signals in GLIS3 and ZFAND3. *Diabetes/Metabolism Research and Reviews*, *33*(4). https://doi.org/10.1002/dmrr.2869

Murtaugh, L. C., & Melton, D. A. (2003). Genes, Signals, and Lineages in Pancreas Development. *Annual Review of Cell and Developmental Biology*, *19*(1), 71–89. https://doi.org/10.1146/annurev.cellbio.19.111301.144752

Mutskov, V. J., Farrell, C. M., Wade, P. A., Wolffe, A. P., & Felsenfeld, G. (2002). The barrier function of an insulator couples high histone acetylation levels with specific protection of promoter DNA from methylation. *Genes and Development*, *16*(12), 1540–1554. https://doi.org/10.1101/gad.988502

Narlikar, L., & Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Briefings in Functional Genomics and Proteomics*, *8*(4), 215–230. https://doi.org/10.1093/bfgp/elp014

Naya, F. J., Huang, H. P., Qiu, Y., Mutoh, H., DeMayo, F. J., Leiter, A. B., & Tsai, M. J. (1997). Diabetes, defective pancreatic morphogenesis, and abnormal enteroendocrine differentiation in BETA2/NeuroD-deficient mice. *Genes and Development*, *11*(18), 2323–2334. https://doi.org/10.1101/gad.11.18.2323

Nechaev, S., & Adelman, K. (2012). Transcription Initiation Into Productive Elongation, *1809*(1), 34–45. https://doi.org/10.1016/j.bbagrm.2010.11.001.Pol

Ni, Z., Schwartz, B. E., Werner, J., Suarez, J. R., & Lis, J. T. (2004). Coordination of Transcription, RNA Processing, and Surveillance by P-TEFb Kinase on Heat Shock Genes. *Molecular Cell*, *13*(1), 55–65. https://doi.org/10.1016/S1097-2765(03)00526-4

Nicolson, T. J., Bellomo, E. A., Wijesekara, N., Loder, M. K., Baldwin, J. M., Gyulkhandanyan, A. V., … Rutter, G. A. (2009). Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes*, *58*(9), 2070–2083. https://doi.org/10.2337/db09-0551

Ohneda, K., Ee, H., & German, M. (2000). Regulation of insulin gene transcription. *Online*, *11*, 227–233. https://doi.org/10.1006/10.1006/scdb.2000.0171

Palanker, L., Necakov, A. S., Sampson, H. M., Ni, R., Hu, C., Thummel, C. S., & Krause, H. M. (2006). Dynamic regulation of. *Development*, *3562*, 3549–3562. https://doi.org/10.1101/gad.1752608.lineages

Pan, F. C., & Brissova, M. (2016). Pancreas development in humans, 1–10. https://doi.org/10.1097/MED.0000000000000047.

Pan, X., Solomon, S. S., Borromeo, D. M., Martinez-Hernandez, A., & Raghow, R. (2001). Insulin deprivation leads to deficiency of Sp1 transcription factor in H-411E hepatoma cells and in streptozotocin-induced diabetic ketoacidosis in the rat. *Endocrinology*, *142*(4), 1635–1642. https://doi.org/10.1210/endo.142.4.8083

Parsons, M. J., Pisharath, H., Yusuff, S., Moore, J. C., Siekmann, A. F., Lawson, N., & Leach, S. D. (2009). Notch-responsive cells initiate the secondary transition in larval zebrafish pancreas. *Mechanisms of Development*, *126*(10), 898–912. https://doi.org/10.1016/j.mod.2009.07.002

Pasquali, L., Gaulton, K. J., Rodríguez-seguí, S. A., Gómez-marín, C., Bunt, M. Van De, Ponsa-cobas, J., … Gloyn, A. L. (2014). Europe PMC Funders Group Pancreatic islet enhancer clusters enriched in type 2 diabetes risk – associated variants, *46*(2), 136–143. https://doi.org/10.1038/ng.2870.Pancreatic

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews Genetics*, *14*(4), 288–295. https://doi.org/10.1038/nrg3458

Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., … Drmanac, R. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, *487*(7406), 190–195. https://doi.org/10.1038/nature11236

Plasterk, R. H. A. (1993). Molecular mechanisms of transposition and its control. *Cell*, *74*(5), 781–786. https://doi.org/10.1016/0092-8674(93)90458-3

Poss, Z. C., Ebmeier, C. C., & Taatjes, D. J. (2013). The Mediator complex and transcription regulation. *Critical Reviews in Biochemistry and Molecular Biology*, *48*(6), 575–608. https://doi.org/10.3109/10409238.2013.840259

Prince, V. E., Anderson, R. M., & Dalgin, G. (2017). *Zebrafish Pancreas Development and Regeneration: Fishing for Diabetes Therapies*. *Current Topics in Developmental Biology* (1st ed., Vol. 124). Elsevier Inc. https://doi.org/10.1016/bs.ctdb.2016.10.005

Qiu, W. L., Zhang, Y. W., Feng, Y., Li, L. C., Yang, L., & Xu, C. R. (2017). Deciphering Pancreatic Islet β Cell and α Cell Maturation Pathways and Characteristic Features at the Single-Cell Level. *Cell Metabolism*, *25*(5), 1194–1205.e4. https://doi.org/10.1016/j.cmet.2017.04.003

Rada-Iglesias, A. (2018). Is H3K4me1 at enhancers correlative or causative? *Nature Genetics*, *50*(1), 4–5. https://doi.org/10.1038/s41588-017-0018-3

Ravassard, P., Hazhouz, Y., Pechberty, S., Bricout-neveu, E., Armanet, M., Czernichow, P., & Scharfmann, R. (2011). Technical advance A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. *The Journal of Clinical Investigation*, *121*(9), 3589–3597. https://doi.org/10.1172/JCI58447DS1

Remenyi, A., Scholer, H. R., & Wilmanns, M. (2004). Combinatorial control of gene expression. *Nat Struct Mol Biol*, *11*(9), 812–815. https://doi.org/10.1038/nsmb820

Robertson, A. G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., … others. (2008). Genome-wide relationship between histone H3 lysine 4 mono-and tri-methylation and transcription factor binding. *Genome Research*, *18*(12), 1906–1917. https://doi.org/10.1101/gr.078519.108.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., … Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657. https://doi.org/10.1038/nmeth1068

Roman, T. S., Cannon, M. E., Vadlamudi, S., Buchkovich, M. L., Wolford, B. N., Welch, R. P., … Mohlke, K. L. (2017). A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the ADCY5 locus. *Diabetes*, *66*(9), 2521–2530. https://doi.org/10.2337/db17-0464

Rosenbloom, K. R., Dreszer, T. R., Long, J. C., Malladi, V. S., Sloan, C. A., Raney, B. J., … Kent, W. J. (2012). ENCODE whole-genome data in the UCSC Genome Browser: Update 2012. *Nucleic Acids Research*, *40*(D1), 620–625. https://doi.org/10.1093/nar/gkr1012

Ruiz-Martinez, M., Navarro, A., Marrades, R. M., Viñolas, N., Santasusagna, S., Muñoz, C., … Monzo, M. (2016). YKT6 expression, exosome release, and survival in non-small cell lung cancer. *Oncotarget*, *7*(32). https://doi.org/10.18632/oncotarget.9862

Rutter, G. A. (2010). Think zinc: New roles for zinc in the control of insulin secretion. *Islets*, *2*(1), 49–50. https://doi.org/10.4161/isl.2.1.10259

Sandelin, A. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, *32*(90001), 91D–94. https://doi.org/10.1093/nar/gkh012

Sara, B. (2009). Type 2 Diabetes, *560*. https://doi.org/10.1007/978-1-59745-448-3

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., De Bakker, P. I. W., Chen, H., … Altshuler, D. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, *316*(5829), 1331–1336. https://doi.org/10.1126/science.1142358

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Scharfmann, R., & Pechberty, S. (2014). Development of a conditionally immortalized human pancreatic β cell line. *The Journal of Clinical Investigation*, *124*(5), 1–12. https://doi.org/10.1172/JCI72674.very

Schiavone, M., Rampazzo, E., Casari, A., Battilana, G., Persano, L., Moro, E., … Argenton, F. (2014). Zebrafish reporter lines reveal in vivo signaling pathway activities involved in pancreatic cancer. *Disease Models & Mechanisms*, *7*(7), 883–894. https://doi.org/10.1242/dmm.014969

Seth, A., Stemple, D. L., & Barroso, I. (2013). The emerging use of zebrafish to model metabolic disease. *Disease Models & Mechanisms*, *6*(5), 1080–1088. https://doi.org/10.1242/dmm.011346

Shabalina, S. A., & Spiridonov, N. A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, *5*(4). https://doi.org/10.1186/gb-2004-5-4-105

Shibata, M., Gulden, F. O., & Sestan, N. (2015). From trans to cis: Transcriptional regulatory networks in neocortical development. *Trends in Genetics*, *31*(2), 77–87. https://doi.org/10.1016/j.tig.2014.12.004

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., Wit, E. De, … Laat, W. De. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture – on-chip (4C), *38*(11), 1348–1354. https://doi.org/10.1038/ng1896

Skelin, M., Rupnik, M., & Cencic, A. (2010). Pancreatic beta cell lines and their applications in diabetes mellitus research. *Altex*, *27*(2), 105–113. https://doi.org/10.14573/altex.2010.2.105

Soboleski, M. R., Oaks, J., & Halford, W. P. (2005). Green fluorescent protein is a quantitative reporter of gene expression in individual eukaryotic cells. *Federation of American Societies for Experimental Biology*, *19*(3), 440–442. https://doi.org/10.1096/fj.04-3180fje

Solomon, S. S., Majumdar, G., Martinez-Hernandez, A., & Raghow, R. (2008). A critical role of Sp1 transcription factor in regulating gene expression in response to insulin and other hormones. *Life Sciences*, *83*(9–10), 305–312. https://doi.org/10.1016/j.lfs.2008.06.024

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., … Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, *21*(10), 1757–1767. https://doi.org/10.1101/gr.121541.111

Sribudiani, Y., Metzger, M., Osinga, J., Rey, A., Burns, A. J., Thapar, N., & Hofstra, R. M. W. (2011). Variants in RET associated with hirschsprung’s disease affect binding of transcription factors and gene expression. *Gastroenterology*, *140*(2), 572–582. https://doi.org/10.1053/j.gastro.2010.10.044

Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyannopoulos, J. A., & Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology*, *3–4*, 40–47. https://doi.org/10.1016/j.cpb.2015.10.001

Sussel, L. & Mastraci, T. L. (2013). The Endocrine Pancreas: insights into development, differentiation and diabetes. *National Institutes of Health*, *392*(3), 685–705. https://doi.org/10.1002/wdev.44.

Swanson, C. I., Evans, N. C., & Barolo, S. (2011). Rapid evolutionary rewiring of a structurally constrained eye enhancer, *18*(3), 359–370. https://doi.org/10.1016/j.devcel.2009.12.026.Structural

Taylor, B. L., Liu, F. F., & Sander, M. (2013). Nkx6.1 Is Essential for Maintaining the Functional State of Pancreatic Beta Cells. *Cell Reports*, *4*(6), 1262–1275. https://doi.org/10.1016/j.celrep.2013.08.010

Teta, M., Long, S. Y., Wartschow, L. M., Rankin, M. M., & Kushner, J. A. (2005). Very slow turnover of beta-cells in aged adult mice. *Diabetes*, *54*(September), 2557–2567. https://doi.org/10.2337/diabetes.54.9.2557

Tiso, N., Moro, E., & Argenton, F. (2009). Zebrafish pancreas development. *Molecular and Cellular Endocrinology*, *312*(1–2), 24–30. https://doi.org/10.1016/j.mce.2009.04.018

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, *45*(2), 124–130. https://doi.org/10.1038/ng.2504

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., … Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-

Seq data. *Nature Methods*, *5*(9), 829–834. https://doi.org/10.1038/nmeth.1246

van den Akker, E., Forlani, S., Chawengsaksophak, K., de Graaff, W., Beck, F., Meyer, B. I., & Deschamps, J. (2002). Cdx1 and Cdx2 have overlapping functions in anteroposterior patterning and posterior axis elongation. *Development*, *129*(9), 2181–93. https://doi.org/Unsp dev2786

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2009). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Vilar, J. M. G., & Saiz, L. (2005). DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Current Opinion in Genetics and Development*, *15*(2), 136–144. https://doi.org/10.1016/j.gde.2005.02.005

Wang, C., Zhang, M. Q., & Zhang, Z. (2013). Computational Identification of Active Enhancers in Model Organisms. *Genomics, Proteomics and Bioinformatics*, *11*(3), 142–150. https://doi.org/10.1016/j.gpb.2013.04.002

Weir, G. C., & Bonner-weir, S. (2011). Finally! A human pancreatic β cell line. *J Clin Invest*, *121*(9), 3395–3397. https://doi.org/10.1172/JCI58899.

Westerfield M.(2000) *The zebrafish book. A Guide for the laboratory use of zebrafish (Danio Rerio*), 4th edition, University of Oregon

Xiang, J., Li, X. Y., Xu, M., Hong, J., Huang, Y., Tan, J. R., … Ning, G. (2008). Zinc transporter-8 gene (SLC30A8) is associated with type 2 diabetes in Chinese. *Journal of Clinical Endocrinology and Metabolism*, *93*(10), 4107–4112. https://doi.org/10.1210/jc.2008-0161

Xu, K., Zha, M., Wu, X., Yu, Z., Yu, R., Xu, X., … Yang, T. (2011). Association between rs13266634 C/T polymorphisms of solute carrier family 30 member 8 (SLC30A8) and type 2 diabetes, impaired glucose tolerance, type 1 diabetes-A meta-analysis. *Diabetes Research and Clinical Practice*, *91*(2), 195–202. https://doi.org/10.1016/j.diabres.2010.11.012

Yang, S., Oksenberg, N., Takayama, S., Heo, S. J., Poliakov, A., Ahituv, N., … Boffelli, D. (2015). Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics*, *16*(1), 1–13. https://doi.org/10.1186/s12864-015-2070-7

Yngvadottir, B., Macarthur, D. G., Jin, H., & Tyler-Smith, C. (2009). The promise and reality of personal genomics. *Genome Biology, 10,* 237. https://doi.org/10.1186/gb-2009-10-9-237

Zhang, C., Moriguchi, T., Kajihara, M., Harada, A., Shimohata, H., Oishi, H., … Takahashi, S. (2005). MafA Is a Key Regulator of Glucose-Stimulated Insulin Secretion MafA Is a Key Regulator of Glucose-Stimulated Insulin Secretion. *Molecular and Cellular Biology*, *25*(12), 4969–76. https://doi.org/10.1128/MCB.25.12.4969

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, *24*(R1), R102–R110. https://doi.org/10.1093/hmg/ddv259

Zhang, Z., & Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *Journal of Biology,* *2*(2), 11. https://doi.org/10.1186/1475-4924-2-11

VI. Supplementary data

Table 1 – JASPAR complementary analysis. From the 719 matrices available from vertebrates in JASPAR database, 393 matrices were identified associated to WT/Risk seq132. Then, these results were refined by eliminating all the TFBS motifs that not included the WT7Risk locus, resulting in the table below. (A). Regarding the risk associated variant sequence and based on the relative scores obtained, it was created a resume table with the number of TFs that had differential affinity with the risk sequence, as the gain/loss of binding (B).

A)

Name	Score	Relative score	Sequence ID	Start	End	Strand	Predicted sequence
PAX5	13,7514	0,888219	risk	11	29	+	AACAGCAGCCAGCCGGGAC
PAX5	11,2341	0,858809	wt	11	29	+	AACAGCAGCCAGCTGGGAC
Tcfcp2l1	6,51404	0,826334	wt	13	26	-	CCAGCTGGCTGCTG
Tcfcp2l1	5,41685	0,810356	risk	13	26	-	CCGGCTGGCTGCTG
SMAD2::SMAD3::SMAD4	8,42222	0,822335	risk	14	26	-	CCGGCTGGCTGCT
PAX1	8,67591	0,844119	risk	14	30	-	TGTCCCGGCTGGCTGCT
PAX9	8,50939	0,831403	risk	14	30	-	TGTCCCGGCTGGCTGCT
Myod1	8,713	0,90222	wt	15	27	-	CCCAGCTGGCTGC
THAP1	5,21293	0,827175	risk	16	24	+	CAGCCAGCC
THAP1	5,13392	0,824757	wt	16	24	+	CAGCCAGCT
Hand1::Tcf3	6,75256	0,822532	risk	16	25	-	CGGCTGGCTG
Hand1::Tcf3	6,22582	0,807013	wt	16	25	-	CAGCTGGCTG
TAL1::TCF3	7,37002	0,805375	wt	16	27	+	CAGCCAGCTGGG
ASCL1	8,27687	0,840691	wt	16	28	+	CAGCCAGCTGGGA
NEUROD1	8,88346	0,855731	wt	16	28	-	TCCCAGCTGGCTG
TWIST1	8,03054	0,845006	wt	16	28	+	CAGCCAGCTGGGA
ZBTB18	4,742	0,812299	wt	16	28	+	CAGCCAGCTGGGA
Myb	5,54864	0,839174	wt	17	26	-	CCAGCTGGCT
SP1	6,14518	0,801217	risk	17	26	-	CCGGCTGGCT
Myog	6,4557	0,880515	wt	17	27	-	CCCAGCTGGCT
Tcf12	5,5395	0,85822	wt	17	27	-	CCCAGCTGGCT
Tcf3	12,555	0,95225	wt	17	27	-	CCCAGCTGGCT
Tcf3	3,5146	0,831997	risk	17	27	-	CCCGGCTGGCT
USF1	3,66816	0,821319	wt	17	27	-	CCCAGCTGGCT
ZEB1	5,18344	0,832761	wt	17	27	-	CCCAGCTGGCT
TFAP2A(var.2)	3,40101	0,805747	risk	17	28	-	TCCCGGCTGGCT
ASCL1	7,47043	0,827144	wt	17	29	-	GTCCCAGCTGGCT
NEUROD1	9,74202	0,872118	wt	17	29	+	AGCCAGCTGGGAC
TWIST1	7,92463	0,842924	wt	17	29	-	GTCCCAGCTGGCT
Atoh1	13,7404	0,995554	wt	18	25	-	CAGCTGGC
Atoh1	5,27962	0,867004	risk	18	25	-	CGGCTGGC
TFAP2A	6,77176	0,881869	risk	18	26	+	GCCAGCCGG
TFAP2A	5,558	0,842262	wt	18	26	+	GCCAGCTGG
Ascl2	7,13469	0,823026	wt	18	27	-	CCCAGCTGGC
Ascl2	6,17312	0,803928	wt	18	27	+	GCCAGCTGGG

Atoh1	4,88617	0,814033	wt	18	27	-	CCCAGCTGGC
Atoh1	4,69235	0,810347	wt	18	27	+	GCCAGCTGGG
FIGLA	8,92208	0,893885	wt	18	27	-	CCCAGCTGGC
FIGLA	8,35565	0,882604	wt	18	27	+	GCCAGCTGGG
ID4	6,01732	0,840488	wt	18	27	-	CCCAGCTGGC
ID4	5,08781	0,822668	wt	18	27	+	GCCAGCTGGG
MYB	3,78195	0,808536	wt	18	27	-	CCCAGCTGGC
Neurog1	6,37768	0,850801	wt	18	27	+	GCCAGCTGGG
Neurog1	5,56536	0,831939	wt	18	27	-	CCCAGCTGGC
NHLH1	7,53104	0,842895	wt	18	27	-	CCCAGCTGGC
NHLH1	5,58268	0,809935	wt	18	27	+	GCCAGCTGGG
TCF3	8,0431	0,904707	wt	18	27	-	CCCAGCTGGC
TCF3	7,57315	0,896806	wt	18	27	+	GCCAGCTGGG
TCF4	8,45918	0,916803	wt	18	27	-	CCCAGCTGGC
TCF4	6,52403	0,887207	wt	18	27	+	GCCAGCTGGG
TFAP4	7,49354	0,848832	wt	18	27	-	CCCAGCTGGC
TFAP4	6,30787	0,829728	wt	18	27	+	GCCAGCTGGG
EBF1	0,968359	0,818794	wt	18	28	+	GCCAGCTGGGA
Myog	8,92548	0,913604	wt	18	28	+	GCCAGCTGGGA
Tcf12	7,86887	0,89074	wt	18	28	+	GCCAGCTGGGA
Tcf3	5,84578	0,863006	wt	18	28	+	GCCAGCTGGGA
ZEB1	4,54018	0,820895	wt	18	28	+	GCCAGCTGGGA
Myod1	7,37601	0,885033	wt	18	30	+	GCCAGCTGGGACA
Bhlha15	7,74005	0,890508	wt	19	26	+	CCAGCTGG
Bhlha15	7,74005	0,890508	wt	19	26	-	CCAGCTGG
Tcfcp2l1	7,01468	0,833625	wt	19	32	-	GCTGTCCCAGCTGG
Tcfcp2l1	5,30934	0,808791	risk	19	32	-	GCTGTCCCGGCTGG
Tcfcp2l1	5,10924	0,805877	wt	19	32	+	CCAGCTGGGACAGC
ZEB1	5,26234	0,841225	wt	20	25	+	CAGCTG
ZEB1	5,26234	0,841225	wt	20	25	-	CAGCTG
Atoh1	5,96758	0,877457	wt	20	27	+	CAGCTGGG
STAT3	1,32493	0,813606	wt	20	30	+	CAGCTGGGACA
STAT3	0,586987	0,804666	risk	20	30	+	CAGCCGGGACA
Hic1	6,92242	0,857816	wt	21	29	-	GTCCCAGCT
HIC2	5,81832	0,858157	wt	21	29	-	GTCCCAGCT
TFDP1	6,17637	0,803288	risk	21	31	+	AGCCGGGACAG
RBPJ	6,29924	0,855183	wt	22	31	+	GCTGGGACAG
MZF1	5,26193	0,825553	wt	23	28	+	CTGGGA
Pax2	5,41577	0,856667	risk	23	30	-	TGTCCCGG
MEIS1	4,4281	0,874519	wt	25	31	+	GGGACAG
MEIS1	4,4281	0,874519	risk	25	31	+	GGGACAG

B)

Binding classification	**Nr. of TFs**
Gain of binding	7
Loss of binding	49
Equally binding	1
Differential affinity	9
More affinity with *WT*	9
More affinity with Risk	0
Total	66

Table 2– JASPAR complementary analysis. From the 719 matrices available from vertebrates in JASPAR database, 395 matrices were identified associated to the new variant locus of seq132 (Sequence A and B). Then, these results were refined by eliminating all the TFBS motifs that not included the new variant locus, resulting in the table below. (A). Regarding the new variant sequence and based on the relative scores obtained, it was created a resume table with the number of TFs that had differential affinity with the risk sequence, as the gain/loss of binding (B).

A)

Barhl1	5,78458	0,857933	A	15	24	-	CCTAATGTGT
BARX1	7,03086	0,908928	A	16	23	+	CACATTAG
BSX	7,087	0,914781	A	16	23	+	CACATTAG
Dlx1	5,79912	0,850705	A	15	24	+	ACACATTAGG
Dlx1	4,44089	0,816334	A	15	24	-	CCTAATGTGT
Dlx3	4,19829	0,806643	A	16	23	+	CACATTAG
Dlx4	3,95614	0,812796	A	16	23	+	CACATTAG
DLX6	3,83269	0,812547	A	16	23	+	CACATTAG
EMX1	6,84	0,855305	A	15	24	-	CCTAATGTGT
EMX2	4,97426	0,854383	A	15	24	+	ACACATTAGG
EN1	3,52958	0,818276	A	16	23	-	CTAATGTG
EN1	3,46397	0,81685	A	16	23	+	CACATTAG
EN1	3,29767	0,813236	B	16	23	-	CCAATGTG
EN2	3,54739	0,802258	A	15	24	+	ACACATTAGG
ESX1	4,37769	0,804111	A	15	24	+	ACACATTAGG
EVX1	6,87292	0,87899	A	15	24	+	ACACATTAGG
EVX1	3,46919	0,801535	B	15	24	+	ACACATTGGG
EVX2	6,71104	0,879358	A	15	24	+	ACACATTAGG
GBX2	2,87318	0,800436	A	15	24	+	ACACATTAGG
GSC	6,79121	0,852756	A	15	24	-	CCTAATGTGT
GSC2	6,44675	0,84727	A	15	24	-	CCTAATGTGT
GSX1	6,43485	0,868316	A	15	24	+	ACACATTAGG
GSX2	6,33842	0,854244	A	15	24	+	ACACATTAGG
HLTF	5,99807	0,891561	B	15	24	+	ACACATTGGG
HOXA2	7,13602	0,891768	A	15	24	+	ACACATTAGG
HOXB2	7,39411	0,903052	A	15	24	+	ACACATTAGG
HOXB3	8,23982	0,925701	A	15	24	+	ACACATTAGG
LBX2	5,10636	0,81309	A	15	24	+	ACACATTAGG

LHX2	4,13569	0,83228	A	15	24	+	ACACATTAGG
Lhx4	4,50603	0,827991	A	16	23	+	CACATTAG
LHX9	5,96023	0,864562	A	16	23	+	CACATTAG
LHX9	4,35348	0,820612	B	16	23	-	CCAATGTG
LHX9	4,24138	0,817546	A	16	23	-	CTAATGTG
MAX	3,37815	0,822678	B	13	22	+	TAACACATTG
MEOX1	3,98444	0,807705	A	15	24	+	ACACATTAGG
MEOX2	4,64834	0,807508	A	15	24	+	ACACATTAGG
MIXL1	4,19813	0,814476	A	15	24	+	ACACATTAGG
MSX1	4,10645	0,811173	A	16	23	+	CACATTAG
Myb	5,49038	0,838215	A	22	31	-	TTAGCAGCCT
NEUROD2	3,71909	0,825543	B	14	23	+	AACACATTGG
Neurog1	5,87923	0,839227	B	14	23	-	CCAATGTGTT
NFIC	4,18303	0,815303	B	21	26	+	TGGGCT
NFIX	4,44073	0,850052	B	20	28	-	GCAGCCCAA
NKX6-2	4,52884	0,825328	A	16	23	+	CACATTAG
Nobox	5,55365	0,808411	A	15	22	-	TAATGTGT
NOTO	9,595	0,937407	A	15	24	+	ACACATTAGG
NOTO	6,23956	0,87807	B	15	24	+	ACACATTGGG
NOTO	3,03128	0,821335	A	15	24	-	CCTAATGTGT
OTX1	4,08116	0,838898	A	16	23	-	CTAATGTG
OTX2	4,27851	0,835533	A	16	23	-	CTAATGTG
PDX1	6,65422	0,897612	A	16	23	+	CACATTAG
Pdx1	8,72988	0,965433	A	18	23	-	CTAATG
POU6F2	6,11604	0,821316	A	14	23	+	AACACATTAG
PRRX1	4,10662	0,80646	A	16	23	+	CACATTAG
Prrx2	5,34257	0,843138	A	18	22	+	CATTA
Prrx2	4,75481	0,811404	A	16	23	+	CACATTAG
RAX2	4,7258	0,819227	A	16	23	+	CACATTAG
RUNX1	6,66952	0,802185	A	12	22	-	TAATGTGTTAT
SHOX	5,6106	0,833115	A	16	23	-	CTAATGTG
Shox2	3,7445	0,820281	A	16	23	+	CACATTAG
Shox2	3,08584	0,805505	A	16	23	-	CTAATGTG
Shox2	3,06607	0,805062	B	16	23	-	CCAATGTG
SOX10	4,67125	0,81183	B	17	22	-	CAATGT
SOX10	5,27395	0,838587	B	18	23	+	CATTGG
SOX13	6,61122	0,829651	B	15	25	+	ACACATTGGGC
SOX15	4,98147	0,809948	A	14	23	-	CTAATGTGTT
SOX15	4,70413	0,803721	B	14	23	-	CCAATGTGTT
Sox17	7,40552	0,862559	B	16	24	+	CACATTGGG
Sox17	6,0738	0,819538	B	16	24	-	CCCAATGTG
Sox2	5,09819	0,842691	B	16	23	-	CCAATGTG
Sox3	3,50202	0,817477	B	14	23	-	CCAATGTGTT
Sox6	5,08245	0,81139	B	14	23	-	CCAATGTGTT

TFEC	5,78176	0,806881	A	14	23	+	AACACATTAG
THAP1	4,49031	0,805059	B	19	27	-	CAGCCCAAT
Twist2	5,20841	0,80929	B	14	23	-	CCAATGTGTT
UNCX	4,38507	0,807021	A	16	23	+	CACATTAG
VAX1	5,17372	0,826976	A	16	23	-	CTAATGTG
VAX1	4,75703	0,815147	A	16	23	+	CACATTAG
VAX2	4,94313	0,824748	A	16	23	+	CACATTAG
VAX2	4,68486	0,817921	A	16	23	-	CTAATGTG
VENTX	7,26487	0,886751	A	15	23	+	ACACATTAG
VSX1	4,72882	0,801896	A	16	23	-	CTAATGTG

B)

Binding classification	**Nr. of TFs**
Gain of function	54
Loss of function	16
Differential affinity	6
More affinity with A	5
More affinity with B	1
Total	76