U.PORTO

DOUTORAMENTO

BIOLOGIA MOLECULAR E CELULAR

Understanding the impact of cis-regulatory disruption in the development of type 2 diabetes

Ana Catarina Macedo Eufrásio

D

2024

U.PORTO Ana Catarina Macedo Eufrásio .Understanding the impact of cis-regulatory disruption in the development of type 2 diabetes

Understanding the impact of cis-regulatory disruption in the development of type 2 diabetes
Ana Catarina Macedo Eufrásio

D .ICBAS 2024

SEDE ADMINISTRATIVA INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
FACULDADE DE CIÊNCIAS

ANA CATARINA MACEDO EUFRÁSIO

UNDERSTANDING THE IMPACT OF CIS-REGULATORY DISRUPTION IN THE DEVELOPMENT OF TYPE 2 DIABETES

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

Orientador – Doutor José Carlos Ribeiro Bessa
Categoria – Investigador Principal
Afiliação – Instituto de Biologia Molecular e Celular (IBMC), Instituto de Investigação e Inovação em Saúde (i3S) da Universidade do Porto.

“The future is in the hands of those who explore... and from all the beauty they discover while crossing perpetually receding frontiers, they develop for nature and for humankind an infinite love”

Jacques Yves Cousteau, Oceanographer

Financial support

This work was financially supported by:

- PhD fellowships SFRH/BD/147762/2019 and COVID/BD/153562/2024, Fundação para a Ciência e Tecnologia (FCT).
- Programa Graduado em Biologia Molecular e Celular, Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto (ICBAS-UP), Faculdade de Ciências da Universidade do Porto (FCUP).
- European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG-680156-ZPR).
- “La Caixa” Foundation (under the grant agreement HR21-01212).
- Portuguese funds through FCT in the framework of the project PTDC/BIA-MOL/3834/2021.

Acknowledgements/Agradecimentos

Tal como um grande vocalista dizia "…all we need is just a little patience…". Talvez este tenha sido o meu lema nos meus últimos tempos como estudante de doutoramento. Tal como qualquer aluno de doutoramento, houve momentos de angústia, de desespero e de, claro, muita paciência. Apesar de tudo, o meu doutoramento foi das experiências mais bonitas da minha carreira científica. Convivi com pessoas incríveis, aprendi com os melhores e, mais importante que tudo, conheci uma Ana que sempre sonhei conhecer. É claro que toda esta minha jornada não teria sido possível sem todas essas pessoas e, por esse motivo, tenho de prestar o meu grande agradecimento.

Ao meu orientador José Bessa, que sempre acreditou em mim e foi cúmplice das minhas ideias mais fora da caixa. Obrigada pela grande e inesquecível oportunidade, por todos os conselhos, e mais importante que tudo, obrigada por nunca me deixar esquecer no que realmente a ciência é feita, curiosidade e entusiasmo pelo que é novo, fascinante e impactante. Sem essa força, seria muito difícil continuar quando tudo corria mal.

Ao meu grupo, que será sempre meu, VDR. Às Joanas, que se tornaram mais do que meras colegas, mas família, obrigada por toda a amizade, carinho e apoio em tudo. Sem ordem, obrigada Marta, Fábio, João A, João N, Francisco C, Francisco M, Sofia (a minha compincha dos clones), Miguel, Inês, Diogo, Nerea e Isabel por todas as gargalhadas, discussões superprodutivas e por criarem um ambiente no lab tão bom! Aos membros passados do VDR, Chiara, Francisca, Inês C., Mafalda, Pablo, Ana G. e Leonor, que também contribuíram tanto para a minha jornada ter sido tão bonita. Não posso deixar de agradecer a muitas pessoas fora do VDR, especialmente à minha orientadora emprestada, Alexandra Moreira, por confiar em mim. Quero também agradecer à Ana Jesus, Joana Azevedo, Filipa Castro, Ana Moutinho, Ana Rita Araújo, Monika Vilares, Sara Vaz, Ana Rita Castro, Renata Freitas, Isabel Silveira e Joana Loureiro por todo o apoio. A todo o pessoal técnico que também me amparou em muitas horas difíceis, Emília Cardoso, Catarina Meireles, Rob Mensink, Ana Mafalda Rocha, Maria Azevedo, Paula Sampaio, Tânia Meireles, Paula Magalhães e André Maia.

À minha família, especialmente ao meu irmão, mãe, avós, padrinhos e primos, que sempre acreditaram em mim e que nunca me deixaram cair. Um obrigada à minha afilhada que, apesar de ser tão pequenina, já é uma das minhas maiores forças, quer para continuar, como para perceber que a vida pode ser tão simples e bonita, quando está por perto.

Um obrigada aos meus 7, contando comigo, que a Biologia Marinha me deu: João Cruz, Pedro Fonseca, Catarina Faísca, Diogo Soares, João Pereira e Daniela Teixeira, a

minha “miga” do coração, que nunca me deixaram de apoiar, mesmo depois de tantos desencontros.

Ao meu marido, por ser a minha maior inspiração e nunca me deixar querer menos que o universo. Obrigada por acreditares em mim, por me dares confiança, por toda a paciência e por seres o meu melhor amigo.

List of publications

Original research article, published

Eufrásio, A., Perrod, C., Ferreira, F. J., Duque, M., Galhardo, M., & Bessa, J. (2020). In Vivo Reporter Assays Uncover Changes in Enhancer Activity Caused by Type 2 Diabetes–Associated Single Nucleotide Polymorphisms. Diabetes, 69(12), 2794–2805. https://doi.org/10.2337/db19-1049

Original research article, published as preprint in bioRxiv, under revision in Nucleic Acids Research

Eufrásio, A., Azevedo, J., Machado, J., Ferreira, A., Moutinho, A., Henriques, F., Jesus, A., Tavares, J., Pereira-Castro, I., Teixeira, J., Pinto, P. A., Bessa, J., & Moreira, A. (2023). A short 3´UTR motif regulates gene expression in bilaterians (p. 2023.11.15.567165). bioRxiv. https://doi.org/10.1101/2023.11.15.567165

Original research article, published

Bordeira-Carriço, R., Teixeira, J., Duque, M., Galhardo, M., Ribeiro, D., Acemel, R. D., Firbas, P. N., Tena, J. J., **Eufrásio, A**., Marques, J., Ferreira, F. J., Freitas, T., Carneiro, F., Goméz-Skarmeta, J. L., & Bessa, J. (2022). Multidimensional chromatin profiling of zebrafish pancreas to uncover and investigate disease-relevant enhancers. Nature Communications, 13(1), 1945. https://doi.org/10.1038/s41467-022-29551-7

Original research article, published

Diogo, P., Martins, G., Simão, M., Marreiros, A., **Eufrásio, A**. C., Cabrita, E., & Gavaia, P. J. (2023). Type I Diabetes in Zebrafish Reduces Sperm Quality and Increases Insulin and Glucose Transporter Transcripts. International Journal of Molecular Sciences, 24(8), 7035. https://doi.org/10.3390/ijms24087035

Original research article, published

Matinha-Cardoso, J., Coutinho, F., Lima, S., **Eufrásio, A**., Carvalho, A. P., Oliva-Teles, A., Bessa, J., Tamagnini, P., Serra, C. R., & Oliveira, P. (2022). Novel protein carrier system based on cyanobacterial nano-sized extracellular vesicles for application in fish. Microbial Biotechnology, 15(8), 2191–2207. https://doi.org/10.1111/1751-7915.14057

Original research article, in preparation

Loureiro, JR., Castro AF., Figueiredo, AS., **Eufrásio, A**., Bessa, J. and Silveira I. The ATTTC Pentanucleotide Repeat in Spinocerebellar Ataxia Type 37 Disrupts a Multifunctional Cis-Regulatory Region.

Resumo

Durante o desenvolvimento de um organismo, a regulação da transcrição genética é crucial para estabelecer diversos tipos celulares com funções distintas. Ao compreender toda a maquinaria da transcrição, em específico, a função dos elementos cis-regulatórios (ECRs), incluindo dos potenciadores, pode fornecer-nos pistas importantes na área das doenças genéticas. A região 3' não traduzida (3´UTR) também desempenha um papel fundamental na regulação genética, contendo um conjunto distinto de ECRs, incluindo elementos de sequência a jusante (USE). Variantes genéticas nestas sequências regulatórias podem perturbar a transcrição genética, resultando em alterações fenotípicas e, consequentemente, em doenças. O quão estas variantes perturbam a função das ECRs ainda é subexplorado. Estudos de grande escala têm vindo a associar polimorfismos de um nucleótido (SNPs) a diabetes tipo 2 (DT2). A grande maioria destes SNPs estão localizados em possíveis potenciadores endócrinos pancreáticos, sugerindo que SNPs podem prejudicar a atividade do potenciador e, consequentemente a expressão genética, resultando num aumento de risco do desenvolvimento de DT2. No capítulo 2, usando ensaios de transgénese em mosaico, em peixes-zebra, verificámos que SNPs são capazes de modular a atividade de potenciadores endócrinos pancreáticos. Um SNP específico (rs13266634), localizado num exão de *SLC30A8*, geralmente associado à diminuição da função de SLC30A8 e ao aumento do risco de DT2, aumentou a atividade do seu potenciador. O ganho de função de *SLC30A8* pode ser uma possível causa para o aumento do risco de DT2, dando novas pistas no que respeita à relação complexa entre variações genéticas, atividade dos potenciadores e suscetibilidade a DT2. No capítulo 3, expandimos o nosso estudo para 33 pares de sequências de risco e não-risco. Destas, uma variante (rs5215), no gene *KCNJ11,* resulta numa substituição do aminoácido isoleucina por valina, e aumentou a atividade do potenciador. Ao explorar o potenciador ortólogo em murganho, mostrámos que uma substituição de um simples nucleótido, não afetando o aminoácido, modula a atividade do potenciador. Estes resultados fornecem uma nova perspetiva numa área de potenciadores codificantes pouco explorada que contêm dupla função, onde um SNP codificante tem o potencial de modular a atividade de um potenciador. Finalmente, no capítulo 4, estudámos um USE do gene *polo* da Drosófila em peixes-zebra e linhas celulares humanas e demonstrámos a sua capacidade de ativar a expressão genética e de atuar como uma esponja para proteínas de ligação a RNA. A presença de um SNP no USE, associado ao cancro do colón, resultou num ganho de função do mRNA do gene, sugerindo o seu envolvimento no desenvolvimento do cancro do colón. Concluindo, os resultados apresentados nesta tese doutoral fornecem evidências de que a perturbação de ECRs podem desempenhar um papel no desenvolvimento de doenças.

Abstract

During development, precise spatial and temporal regulation of gene transcription is crucial for establishing diverse cell types in our organism. Understanding the role of cis-regulatory elements (CREs), including enhancers, can give us important insights on how genetic diseases are developed. Post-transcriptionally, the 3´prime untranslated region (3´UTR), also plays a role in gene regulation, comprising its distinct set of CREs, including the upstream sequence element (USE). Variations in CREs can disrupt gene transcription, leading to phenotypic changes and diseases. To what extent these genetic variants disrupt CREs function is yet to be fully explored. Genome-wide association studies have been associating several single nucleotide polymorphisms (SNPs) with type 2 diabetes (T2D). The vast majority overlap with putative endocrine enhancers, suggesting that these variants might impair enhancer activity and consequently, gene expression, resulting in an increased risk of T2D development. Chapter 2 demonstrates, through in vivo mosaic transgenesis assays in zebrafish, that SNPs can modulate endocrine pancreatic enhancer activity. Notably, SNP rs13266634 in the *SLC30A8* gene, traditionally associated with increased T2D risk, unexpectedly boosted the activity of an overlapping enhancer, suggesting a gain of function in SLC30A8 as a potential cause for T2D susceptibility. In chapter 3, we expanded our study to test 33 pairs of T2D risk and non-risk sequences. Of these, in one enhancer, its risk variant (rs5215), in gene *KCNJ11*, results in an isoleucine-to-valine aminoacid, and enhanced the enhancer activity. Exploring the orthologous enhancer in mice revealed that a single nucleotide substitution, without affecting the amino acid, modulates enhancer activity. These findings shed light on dual-function coding enhancers, where a coding SNP can modulate enhancer activity. In chapter 4, we investigated a 3´UTR upstream sequence element (USE) from Drosophila's *polo* gene in zebrafish and human cell lines. The USE was found to activate gene expression across species and act as an RNA binding protein sponge. A colon cancer-associated SNP in the USE resulted in a gain of function of the mRNA gene, suggesting its involvement in colon cancer development. Overall, this doctoral thesis provides compelling evidence that the disruption of CREs may play a significant role in disease development.

Table of contents

Financial support vi
Acknowledgments viii
List of publications x
Resumo xii
Abstract xiv
Table of contents xvi
Abbreviations xx
CHAPTER 1 – GENERAL INTRODUCTION 1
1. *Cis*-regulation 3
1.1. Transcriptional *cis*-regulation 3
1.1.1. Enhancers 4
1.1.2. Enhancers prediction 5
1.1.3. Enhancers validation 6
1.2. Cis-regulation in the 3´prime untranslated region (3´UTR) 7
1.2.1. Polyadenylation signals (PAS) 7
1.2.2. Upstream sequence elements (USEs) 7
1.3. Cis-regulation and diseases 8
1.3.1. Disease-associated single nucleotide polymorphisms (SNPs): Genome-wide association studies (GWAS) 9
1.4. Type 2 diabetes (T2D) genetics 9
1.4.1. T2D disease – general aspects 9
1.4.2. The human endocrine pancreas 10
1.4.3. T2D and *cis*-regulation 10
1.4.3.1. T2D-associated SNPs by GWAS 11
1.5. Polyadenylation mutations in disease 12
2. Zebrafish as a disease model 13
2.1. General characteristics 13
2.2. Transgenesis assays for the study of *cis*-regulation 14
3. Research aims 14
4. References 15
CHAPTER 2 – IN VIVO REPORTER ASSAYS UNCOVER CHANGES IN ENHANCER ACTIVITY CAUSED BY TYPE 2 DIABETES -ASSOCIATED SINGLE NULCEOTIDE POLYMORPHISMS 27
1. Introduction 31
2. Research design and methods 32

2.1. Zebrafish husbandry and embryo culture 32
2.2. Putative enhancer selection 32
2.3. In vivo mosaic transgenesis assays 33
2.4. *sst*:mcherry reporter line....33
2.5. Immunohistochemistry 34
2.6. Assessment of enhancer activity 34
2.7. Prediction of TFBSs affected by type 2 diabetes risk variants....34
2.8. ChIP on plasmid....34
2.9. Real-time quantitative expression analysis....35
2.10. 4C sequencing 35
2.11. CRISPR inactivation and CRISPR activation targeting 35
2.12. Statistical analysis....36
2.13. Data and resource availability 36
3. Results 36
3.1. Identification of endocrine pancreatic enhancers in vivo by mosaic transgenesis in zebrafish....36
3.2. Identification of endocrine pancreatic enhancers overlapping with type 2 diabetes-associated SNPs 38
3.3. SNPs associated with increased risk of type 2 diabetes modulate enhancer activity....40
3.4. The enhancer seq132mm belongs to the Slc30a8 regulatory landscape 41
3.5. The SLC30A8 seq132 enhancer is divided into different functional domains .42
3.6. A single nucleotide mutation impairs seq132 enhancer 44
4. Discussion....45
5. Article information....48
6. References 48
7. Supplemental material....53

CHAPTER 3 – THE SEQ5215 CODING ENHANCER IN KCNJ11 GENE 97
1. Introduction....99
2. Research design and methods 101
2.1. Zebrafish transgenic lines and husbandry 101
2.2. Putative enhancer identification 102
2.3. In vivo mosaic transgenesis assays 102
2.4. Immunohistochemistry 103
2.5. Cell culture....103
2.6. Luciferase reporter assays 103

2.7. Statistical analysis 103
3. Results 104
3.1. Testing the functionality of putative human endocrine enhancers and the impact of T2D SNPs in their regulatory output 104
3.2. The human seq5215 107
3.2.1. Seq5215 in vivo and in vitro reporter assays 107
3.3. The seq5215mm is an ortholog enhancer of seq5215 in mice 110
3.3.1. Studying the impact of a silent mutation within an enhancer 111
4. Discussion and future perspectives 112
5. References 114
6. Supplementary material 117

CHAPTER 4 – A SHORT 3´UTR MOTIF REGULATES GENE EXPRESSION IN BILATERIANS 119
1. Introduction 123
2. Materials and methods 124
2.1. Bioinformatic analysis 124
2.2. Disease-association diseases and DplUSE vs 3´UTR sequence conservation analysis 125
2.3. Oligonucleotides and Antibodies 125
2.4. Plasmid constructs 125
2.5. In vitro transcription 126
2.6. UV cross-linking assays 126
2.7. Immunoprecipitation 127
2.8. Zebrafish husbandry and embryo culture 127
2.9. Microinjections procedures, zebrafish GFP-USE transgenic line establishment, GFP screening and quantification 128
2.10. Immunohistochemistry 128
2.11. Cell culture and transfection 129
2.12. RNA extraction and RT-qPCR 129
2.13. Statistical analysis 130
3. Results 130
3.1. A fruit fly upstream sequence element is conserved in the 3´UTR of humans, mouse and zebrafish genes 130
3.2. The DplUSE sequence increases protein expression in a vertebrate model system 131
3.3. DplUSE RNA recruits protein factors essential for the correct expression of genes involved in zebrafish development and in human cells function 134

3.4. A bilaterian ultraconserved molecular mechanism operates in the DplUSE RNA 139
3.5. PTBP1 is part of the molecular mechanism modulating the expression of DplUSE-containing genes .....
3.6. A single-nucleotide polymorphism in the DplUSE sequence is associated with increased cancer risk 144
4. Discussion 145
5. Acknowledgements 147
6. References 148
7. Supplementary material 154

CHAPTER 5 – CONCLUDING REMARKS 161
1. References 166

Abbreviations

3C	Chromosome conformation
4C-seq	Circular Chromatin Conformation Capture followed by next-generation sequencing
ABCC8	ATP Binding Cassette Subfamily C Member 8
APA	Alternative polyadenylation
ATAC-seq	Assay for transposase-accessible chromatin followed by next-generation sequencing
ChIP-seq	Chromatin immunoprecipitation followed by next-generation sequencing
CREs	*Cis*-regulatory elements
CRISPRa	CRISPR/Cas9 with dCas9 fused to a transcriptional activation domain
CRISPRi	CRISPR/Cas9 with dCas9 fused to a transcriptional repressor domain
dpf	Days-post-fertilization
FISH	Fluorescence in situ hybridization
GWASs	Genome Wide Association Studies
hpf	Hours-post-fertilization
KATP	ATP-sensitive potassium
KCNJ11	Potassium inwardly rectifying channel subfamily J member 11
MODY	Maturity-Onset Diabetes of the Young
MPRAs	Massively parallel reporter assays
mRNA	Messenger RNA
NC	Negative Control
NGS	Next-generation sequencing
PAS	Polyadenylation signals
Pol II	RNA polymerase II
poly (A)	Polyadenylation
PP	Pancreatic polypeptide
PTV	Promoter test vector
RBPs	RNA binding proteins
SDM	Site-directed-mutagenesis
SNPs	Single nucleotide polymorphisms
SNVs	Single nulceotide variations
SUR1	Sulphonylurea receptor 1
T2D	Type 2 Diabetes
TFBS	Transcription factor binding sites

TFs	Transcription factors
TREs	Trans-regulatory elements
USE	Upstream sequence element
UTR	Untranslated region

CHAPTER 1

GENERAL INTRODUCTION

1. C*is*-regulation

1.1. Transcriptional *cis*-regulation

Gene transcription is an important process in which an RNA molecule is synthetized using DNA as template, in order to carry the information transcribed outside the cell nucleus [1]. Outside the nucleus, the newly synthetized molecule, messenger RNA (mRNA) is translated into proteins, essential to cellular viability and function [1,2]. Throughout the development of an organism, the cellular biological processes require a spatial and temporal regulation of gene expression, in order to establish different cell types [1–3]. To understand how genes can be regulated, is it imperative to know how the transcription machinery works in detail. One of the key players for a proper transcription is RNA polymerase II (Pol II), an enzyme that synthetizes the primary RNA, pre-mRNA [1,2], dependent from DNA-specific-binding proteins and chromatin regulators, apart from the general transcription machinery [4–7], named trans-regulatory elements (TREs). The binding of TREs in cis-regulatory elements (CREs), including promoters and enhancers, disrupt the nucleosome organization, initiating gene transcription[1–3,6,8–11] . This way, Pol II can start pre-mRNA synthesis, extending it until reach a specific termination signal, where the pre-mRNA is released [3]and ready to be processed in order to be translated. The CREs can, therefore, repress or enhance the transcription of specific genes, controlling gene expression [8,9,11], contributing for several arrangements in transcriptional regulation, increasing the number and diversity of expression patterns and consequently, the organism complexity. Therefore, a high degree of complexity in organisms is generally accompanied with an increased amount of non-coding DNA[12,13]. In addition, epigenetic modifications that alter chromatin structure also contribute to gene transcriptional regulation[14].

CREs can act in a proximal or distal manner, depending on their type. Proximal CREs are composed by promoters and their proximal regulatory elements, while distal CREs includes enhancers, silencers and insulators[15–17] (Figure 1). The latter can act in long range, by chromatin loops[1,2,16], being located hundreds of kilobases away from the promoters that they interact with[2].

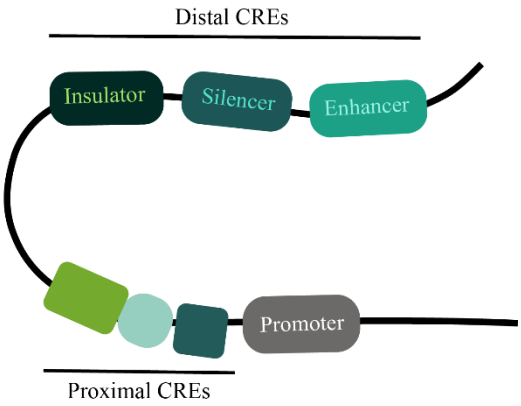


Figure 1- The regulatory sequences involved in transcriptional regulation. (Adapted from [15]).

Silencers are distal target binding sequences for TREs, namely repressors, resulting in a transcription repression [15,18,19]. The binding of repressors can modify the chromatin state[20] or even block activating transcription factors (TFs), necessary for transcription initiation[15].

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[2,15]. Insulators can block the action of neighbor regulatory elements in two distinct ways: disrupting enhancer-promoter interactions or inhibiting chromatin loops, as described by Ali et al [21]. Insulators can also act as heterochromatin barriers, preventing transcriptionally active euchromatin from turning into inactive heterochromatin[22].

1.1.1. Enhancers

Enhancers are CREs that can increase the transcription level of a specific or a set of genes[11,17,23,24], being located downstream, upstream or within introns and exons of their target genes [15,24]. Their function is independent of their distance and/or orientation to the target genes, being difficult to predict which gene is controlled by an enhancer simply by sequence analysis[25]. Enhancers contain specific transcription factor binding sites (TFBS) that interact cooperatively, recruiting co-activators and triggering target gene expression[15,26]. Additionally, different specific tissue enhancers can interact with the same gene promoter, composing the expression pattern of a given gene[27,28]. Being a fundamental CRE, whose activity defines specific timings and locations for the transcriptional activity of genes, enhancers have arisen as elements of great potential and interest, becoming one of the best described regulatory functional element [24,29]. Currently, there are more than 1 million 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[30,31]. As referred before, enhancers have the capability of acting in long-range, by chromatin loops[10,16,24,26,30,32,33], however, how the formation of this chromatin loops

operate, is poorly understood. Nevertheless, the chromatin loop theory is settled in two main evidences. The first evidence came from techniques based on chromosome conformation capture (3C)[34–38]. This methodology is applied to determine the spatial organization of chromatin within cell. The second evidence is based on the proximity of enhancer and promoter regions in the cell nucleus, visualized by fluorescence in situ hybridization (FISH). This technique relies in probes against primary transcripts or DNA, detecting the proximal association of genomic regions[39].

1.1.2. Enhancers prediction

One of the main challenges when studying enhancers is their identification in the genome. Next generation sequencing (NGS) allied to computational biology has emerged as a good strategy, in part, to overcome this challenge[24,30,40]. When TFs bind to enhancers, protein complexes are recruited, including chromatin remodelers and histone modifiers[41]. The recruitment of such complexes can modify histones and chromatin structure, allowing the opening of the chromatin and enhancer-promoter interactions[42]. The assessment of these chromatin modifications by using techniques such as Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) can help to identify regions that function as CREs, including enhancers[43]. This method relies on the construction of an NGS library using Tn5 transposase. NGS adapters are integrated in the transposase, allowing the fragmentation and the incorporation of those adaptors in open or accessible chromatin regions (Figure 2-A). The resulted library is sequenced, and the open chromatin regions can be analyzed by bioinformatic tools[43]. The chromatin accessibility can be also accessed based on chromatin regions that are DNase I hypersensitive[44]. DNase hypersensitivity assessment is based on the property of active CREs to be hypersensitive to cleavage by the endonuclease DNase I[44,45](Figure 2 – B). Another way to predict enhancers is based on DNA-Protein interactions. Chromatin immunoprecipitation (ChIP) is a technique based on crosslinking of DNA and proteins, followed by a specific antibody enrichment for a DNA-binding protein. The resulted fragments are sequenced by NGS, allowing the identification of putative enhancers and TFs that might bind enhancers, genome-wide [46–50] (Figure 2 -C). Specific chromatin epigenetic modifications have been associated to enhancers, proving that chromatin signatures can be a reliable tool for prediction of regulatory elements in the human genome[51–55](Figure 2 – D). It was demonstrated that using antibodies against H3K27ac [56]and H3K4me1[51], that there is an accurate correlation between enhancers and these specific epigenetic marks [51,55]. H3K4me1[4,57] is present in active and primed enhancers, allowing to distinguish enhancers and promoters. In contrast, H3K27ac is present when the enhancer is active, making the

distinction between active and primed enhancers[56]. While the epigenetic marks imply the potential of one sequence be an enhancer, it is imperative to confirm their functionality.

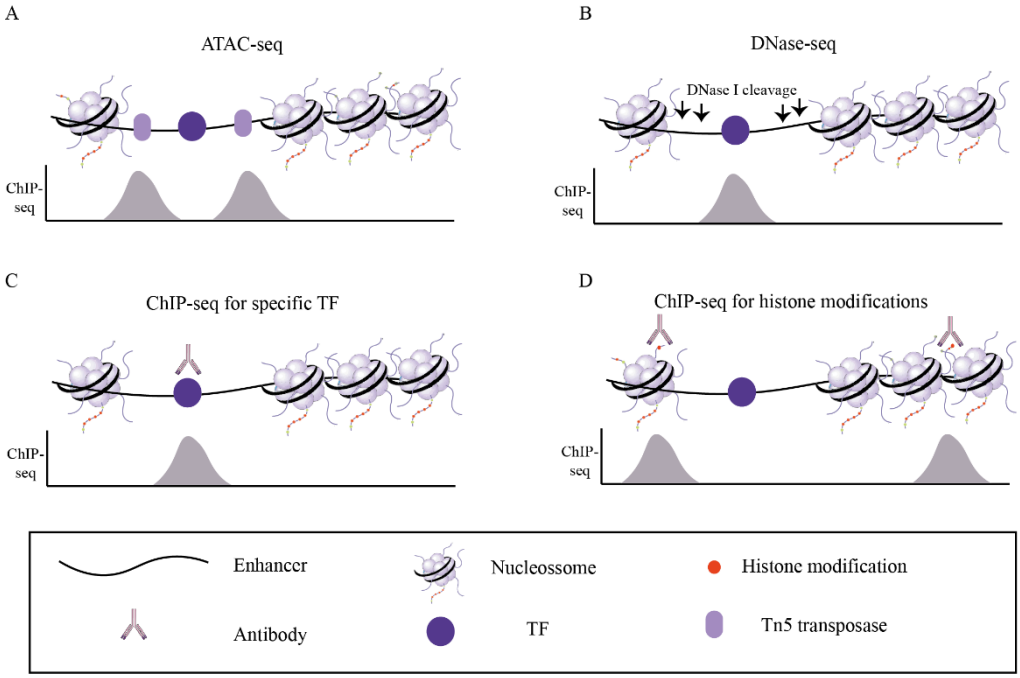


Figure 2 - Genome-wide techniques for enhancer prediction. A – ATAC-seq; B-DNase-seq, C-ChIP-seq for a specific TF; D – ChIP-seq for histone modifications. Adapted from[60].

1.1.3. Enhancers validation

Although more than 1 million putative enhancers have been identified through several biochemical annotations, only a small part of these are validated and even a smaller part are linked to their target gene[58]. This is due, in part, to the high complex nature of the enhancer, that generally presents a heterogeneous landscape[58]. Nevertheless, several studies have demonstrated that sequences enriched for H3K27ac can function as enhancers, when tested for enhancer activity by in vivo and in vitro reporter assays[55,59–63]. In these type of reporter assays, the putative enhancers are inserted in a construct containing a minimal promoter and a reporter gene or a barcode. Then, the constructs are introduced into cell lines or living organisms, allowing to measure their gene expression level. Although important, this kind of time-consuming reporter assays stands as the other factor for a low rate of validated enhancers, posing a limitation for this field. Currently, advanced high-throughput technologies, known as massively parallel reporter assays (MPRAs), enable the simultaneous validation of thousands of genes[64–66], in vitro, in one single experiment. However, for a thorough understanding of the functionality of a particular enhancer, it is essential to evaluate it through in vivo validation [55,59].

1.2. Cis-regulation in the 3´prime untranslated region (3´UTR)

After transcription, the newly pre-mRNA undergoes processing steps prior to translation step[67–69], in the cytoplasm. The pre-mRNA processing involves three important steps: the addition of a 5´cap structure, splicing and cleavage of the 3´end[70]. This pre-mRNA 3´end cleavage it is followed by the addition of a poly(A) tail (polyadenylation), defining the 3´UTR[70]. Along with the aforementioned CREs, the 3´UTR also plays a crucial role in gene regulation, post-transcriptionally, comprising its distinct set of CREs. These specific CREs serve as a binding for RNA binding proteins (RBPs), controlling subcellular localization, translation efficiency and mRNA stability[71]. Additionally, the 3´UTR it is also involved in protein-protein interactions, influencing the formation of different protein complexes[70,72,73].

1.2.1. Polyadenylation signals (PAS)

The cleavage of the 3´end of pre-mRNA and the following polyadenylation are controlled by PAS. Upstream of PASs, the hexameric conserved sequence AAUAAA, named canonical poly(A) signal, is considered the most important signal to initiate cleavage and polyadenylation processes[69,74,75]. The vast majority of genes have multiple PASs, being possible the formation of several transcript isoforms, differing in their coding region and 3´UTRs length[76]- alternative polyadenylation (APA)[75]. The formation of several transcript isoforms can alter binding sites within the 3´UTR region, contributing for gene regulation[75,77].

1.2.2. Upstream sequence elements (USEs)

The conserved canonical poly(A) signal has been described as the most crucial sequence to initiate cleavage and polyadenylation, in the 3´UTR. However, in some genes, the canonical poly(A) can include other variations or can even not be present. Nevertheless, the genes are still cleaved and polyadenylated, suggesting an important role of other sequences, such as USEs, in the use of other PAS in cleavage and polyadenylation processes[78–80]. USEs are pyrimidine or U-rich sequences that, by the recruitment of specific RBPs, can modulate PAS selection and 3´end formation efficiency [78,79,81,82]. Furthermore, it was demonstrated that USEs are more present near of weak PAS, when comparing with the canonical poly(A) signal, highlighting its regulatory role[83].

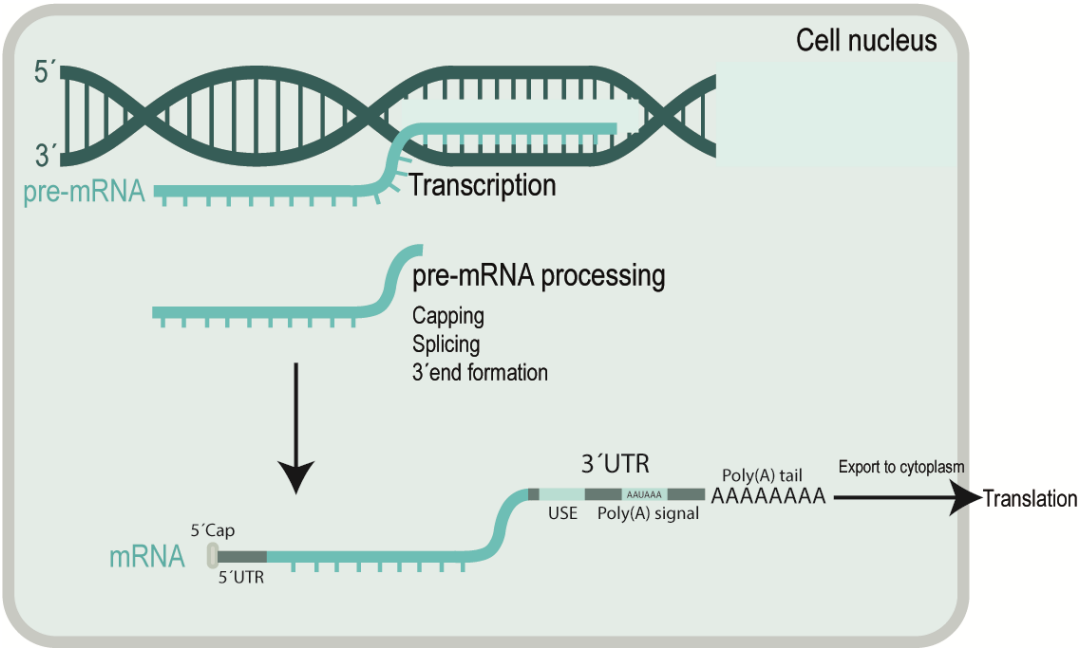


Figure 3 – RNA transcription following pre-mRNA processing.

1.3. *Cis*-regulation and diseases

Until recently, the vast majority of studies on diseases were focused on the protein-coding part of the genome, given its potential for a direct impact that they could have in gene expression and protein translation. However, since a significant portion of the genome is regulating the spatial and temporal control of gene expression, the field of gene regulation has been evolved into one of the most intriguing areas of study in the context of diseases. Variations in such sequences can result in transcriptional dysregulation of genes, phenotypic alterations and disease[15,24,84,85]. For example, regulatory mutations can result in a significant downregulation of relevant gene/s, causing diseases such as beta-thalassemia[86], hemophilia[87] and atherosclerosis[88]. Regulatory mutations can also have the potential to contribute to disease susceptibility by increasing gene transcription levels, by the presence of single-nucleotide variations in regulatory regions[89,90]. For example, individuals carrying a single nucleotide variation (G to T) within an Sp1-binding site located in the initial Col1A1 intron, have an increased risk to develop osteoporosis[85]. Both situations, downregulation or upregulation, are caused by TFBS disruption that impairs the recruitment of key TFs[84,85,88].

1.3.1. Disease-associated single nucleotide polymorphism (SNPs): Genome-wide association studies (GWAS)

Several studies on TFBS disruption highlights how the presence of a SNP can increase the susceptibility of individuals to a disease by altering the expression level of a specific gene. Over the past few years, SNPs have been identified in several whole-genome sequencing projects and computational analysis[91–95]. These kind of point mutations are mostly located in CREs and can be associated to human traits and complex diseases[84,96–99], performing GWAS. GWAS identify associations between genetic *loci* and traits and/or diseases, scanning genomes (using genome-wide SNP array) from a large number of individuals with and without a common trait, in order to find SNPs that can be used to predict the presence of a disease. These studies are used to understand how genes contribute for a specific disease and develop possible treatment strategies[93]. Once the variants associated with the disease are found in higher frequency in disease cases that in controls, statistical analysis is performed to indicate the probability that the SNP is associated with a particular disease[94]. All the GWAS are published in GWAS catalog, an organized and structed online repository[95]. The identification of disease-associated SNPs allowed a more comprehensive study of the complexities of the gene regulatory landscapes[84] and a combined analysis of GWAS allied with enhancers prediction techniques allowed to infer that SNPs associated to disease may be often located in enhancers[100,101]. For example, SNPs located in *SHH* distal enhancer, ZRS, are directly associated with preaxial polydactyly, a congenital disease characterized by additional digits[102]. The analysis of the putative TFBS in ZRS enhancer showed a consensus binding sites for CDX-1, MEIS1 and SOX9, which are TFs involved in limb development. Three of the six polymorphisms described showed to be the cause of disruption of CDX1 binding site, contributing to the disease[102,103]. Another similar example is described in *RET* gene, where the presence of three mutations in MCS enhancer, in intron 1, lead to reduced gene expression. One mutation affects directly SOX10 binding site, contributing for Hirschsprung disease. The other two mutations can also increase the risk for the same disease disrupting GATA2 and RARB TFBS[84,104,105].

1.4. Type 2 Diabetes (T2D) genetics

1.4.1. T2D disease – general aspects

T2D, the most prevalent form of diabetes[106], is a chronic and multifactorial disease, characterized by insulin impairment/resistance in peripheral tissues and insufficient insulin secretion by pancreatic β-cells[107–110], leading to an impairment of the glucose levels in the organism, resulting in hyperglycemia. The incidence and prevalence of diabetes are

increasing, with a global estimate of 783 million adults being diabetic by 2045, representing more than 1 in 10 adults worldwide[108,111]. As chronic traits, T2D may present cardiovascular complications, visual loss and limb amputations, leading to a high rate of mortality in affected people[106,108].

1.4.2. The human endocrine pancreas

One of the principal traits of T2D is the inability of regulating blood glucose levels, due to endocrine pancreas dysfunction. The human pancreas is composed by two main components: the endocrine and exocrine[112,113], having important roles in digestion and metabolism[113]. The exocrine compartment has a gastrointestinal function, containing digestive enzymes expressing-cells, that aid digestion by secreting these enzymes into the digestive tract[113,114]. The endocrine compartment comprehends the hormone-expressing-cells (islets of Langerhans), responsible for maintaining glucose homeostasis, controlling carbohydrate, lipid, and protein metabolism. These hormone-expressing-cells are β, α, ε, 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[115]. Each type of endocrine cells has its own precursor cell, expressing a specific combination of TFs[116] and their differentiation occurs during embryogenesis[117]. The current knowledge suggests that the endocrine dysfunction observed in T2D may rely in beta-cell *cis*-regulation disruption.

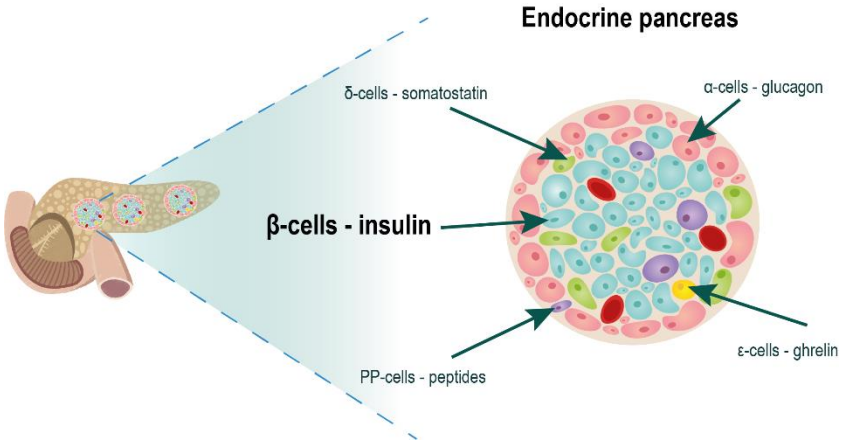


Figure 4 - The human endocrine pancreas and the constituent cell types. In T2D, β-cells cannot produce insulin to keep blood glucose in check, leading to the development of T2D.

1.4.3. T2D and *cis*-regulation

Being a multifactorial disease, it is widely accepted that T2D is a result from a combination of population aging and several environmental factors, such as diet, absence of physical activity and sedentary lifestyle, being the excess body weight the most prevalent

risk factor for T2D development[118–120]. However, there are several cases of T2D patients that do not present excess body weight and individuals that present excess body weight that do not develop T2D, suggesting that T2D may have other hidden causes, such as genetic contributions[108]. The underlying molecular mechanism that might contribute to T2D is poorly understood[121,122]. Nevertheless, there are evidences that show an important genetic contribution in T2D, such as the high-prevalence of the disease in families and studies in twins, where it is shown that there is a higher risk for T2D in monozygotic, when compared with dizygotic twins[123–125]. Another important evidence is the described causality of variants for a specific type of diabetes (Maturity-Onset Diabetes of the Young - MODY)[126,127].

T2D is a polygenic disease, meaning that can be caused by multiple genetic mutations, in several genes[108], that together, will increase the susceptibility to the disease. Currently, one of the most recent hypotheses is that the presence of SNPs in CREs, such as enhancers of genes required for proper β-cell/endocrine pancreas function can cause susceptibility to the disease. Supporting this hypothesis, several variants associated to T2D have been identified in CREs in the past recent years[96,98,100,128]. GWAS have been extremely important to identify and determine the frequency of these SNPs.

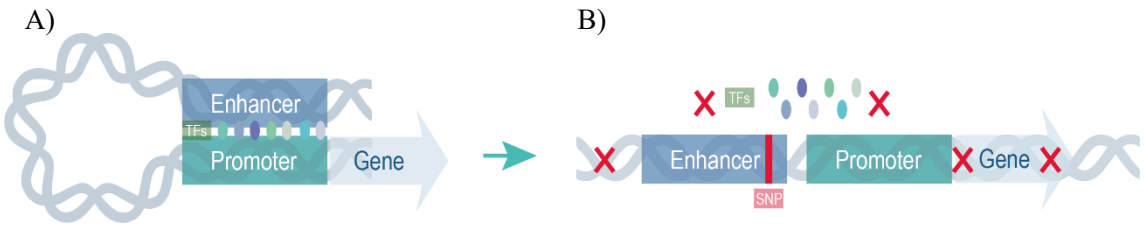


Figure 5 - A. Normal situation – The chromatin loop is formed and the distally TF bind to enhancer in order to activate the transcription; B. Disease associated SNPs – The chromatin loop is impaired, the TF binding site disrupted, and the transcription is affected.

1.4.3.1. T2D-associated SNPs by GWAS

Currently, there are more than 2244 T2D-associated SNPs by GWAS, distributed by more than 139 publications[95]. Additionally, a relatively recent and major contribution for the study of T2D genetics, in a large-scale study, it was performed ChIP-seq for specific epigenetic marks for enhancer activity (H3K4me1 and H3K27ac) were performed to identify human pancreatic enhancers[121], in human pancreatic islets. Then, ChIP-seq for islet important TFs- PDX1, NKX6.1, FOXA2, MAFB and NKX2.2- was used 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. This study allowed to infer that T2D associated variants might have the potential to disrupt TFBS and consequently enhancer activity, causing a dysregulation of target gene expressions[121]. Recently, these datasets were updated by adding ATAC-seq profile and HI-C, a method based on 3C, providing a more complete and a spatial context of the landscapes in study[129–131]. This way it was possible to integrate all the maps of epigenetic marks, ATAC-seq and TFBSs and create a complete and detailed dataset regarding the 3D transcriptional regulation in pancreatic islets[121,132].

One SNP, rs163184, identified as been associated with T2D, is located in an intronic region at *KCNQ1* gene. The presence of rs163184 risk variant disrupts SP3 binding site which is transcriptional regulator, functionating as an activator or repressor, depending on the promoter and cellular context[133]. Using a rat pancreatic cell line containing the risk variant, the authors showed an overexpression of another gene, in close proximity with *KCNQ1*, the *CDKN1C*. It has been demonstrated that the loss of *CDKN1C* leads to an increase of β-cell proliferation[134]. Therefore, these results suggest that the risk-associated variant of rs163184 may induce *CDKN1C* overexpression because of a repressor domain of SP3 binding site disruption in *KCNQ1* locus increasing the risk for T2D, by decreasing the number of β-cells [134] . Several genome-wide studies have showed that T2D associated SNPs are enriched in non-coding regions such as enhancers[101,121,128,135]. However, coding regions, including exons, can also function as enhancers[136]. Variations in such coding regions can result in several phenotypes, although these types of variations are not so explored. Such variants can disrupt the translated protein and/or the overlapping enhancer, suggesting that a phenotype resulting from a coding mutation can be explained not only by protein loss or gain of function but also by the dysregulation of regulatory mechanisms[136].

1.5. Polyadenylation mutations in disease

The cleavage and polyadenylation processes regulate several pre-mRNA processing functions, including stability of mRNA, subcellular localization and nuclear-cytoplasmatic transport[75,78,137]. Consequently, perturbations in 3´UTRs can lead to several pathologies, highlighting their essential regulatory role in cell function[71,138]. Generally, shorter 3´UTRs are more stable than normal or longer 3´UTRs, a characteristic that it should be taken to account when studying disease-related genes[139,140]. For example, genes involved in cell growth are upregulated by having shorter 3´UTRs, being demonstrated that APA is essential

for oncogene activation[139]. Accordingly with these results, in tumor tissues, it was demonstrated a higher number of shorter 3´UTRs, comparing with normal tissues[141,142], in several cancers. Recently, it was demonstrated that SNPs can disrupt or even generate new PAS[143], affecting the mRNA length and expression and APA[144]. Interestingly, the majority of APA-affected SNPs are disease-associated, by GWAS[143]. Mutations in the canonical poly(A) signal were also shown as causal for some diseases, such as neonatal diabetes, where a single nucleotide variation, in the poly(A) signal of *INS* gene, reduced insulin mRNA expression, caused by a reduction of the polyadenylation efficiency [145]. On the other hand, other mutations in poly(A) sites can result in an enhanced polyadenylation efficiency. For example, it is hypothesized that the presence of intronic T2D-associated SNPs in *TCF7L2* gene may enhance the selection of one specific proximal poly(A) site, leading to a production of a truncated protein isoform that repress T-cell factor/lymphoid-enhancer factor, contributing for T2D development[146]. However, further in vivo experiments should be conducted to better understand how SNPs can be contributing to the selection of one poly(A) over other. In summary, the underexplored mutations in the 3´UTR affecting polyadenylation process should not be dismissed, as they could significantly contribute to the development of the disease.

2. Zebrafish as a disease model

2.1. General characteristics

Zebrafish (Danio Rerio) is a freshwater and a small bony fish[147]and, currently, 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 possible to reproduce them and to create stable transgenic and mutant lines. Besides these advantages, the zebrafish embryos are transparent, which allow to follow the embryo development through time[148–150]. In the zebrafish genome, it has been described more than 20,000 genes and 69% of the these genes have orthologues in human[151,152]. Due to this, it is not surprising that zebrafish presents several molecular mechanisms that are similar to humans. The resemblances in function and structure between several zebrafish and human organs allow the possibility to study disease-related molecular mechanisms and the phenotypes that can be associated to the disease. Besides the public availability of the zebrafish genome sequence, there are many other tools that are also available such as fluorescent transgenic and mutant lines (https://zfin.org/)[151]. These transgenic lines that express fluorescent proteins in different molecules, organelles, cells and tissues allow their

in vivo visualization and tracking. allowing a better understanding of molecular pathways and the morphological changes in those organs or tissues[153]. All these characteristics makes the zebrafish a perfect model organism to study several diseases in laboratory [154]. Over the past few years, zebrafish has emerged as a potential tool to study the *cis*-regulation and its disruption in several diseases, like T2D, through transgenesis assays.

2.2. Transgenesis assays for the study of *cis*-regulation

In zebrafish transgenesis assays to test a *cis*-regulatory activity, the sequence of interest to be tested is cloned in a Tol2 transposon vector, containing a reporter gene.

The Tol2 transposon vector contains 2 Tol2 transposon arms that serves as recognition site for the Tol2 transposase enzyme, that will facilitate the vector integration in zebrafish genome and the transmission through generations[157]. 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 a CRE[29].

The vector and the Tol2 mRNA are microinjected in one-cell stage zebrafish embryos and integrated randomly into the genome. The expression pattern generated by the in vivo reporter gene allows the characterization of the tissue specific activity of the tested CRE[155,156].

3. Research aims

This work aims to understand how variants in CREs might contribute to the development of disease. Therefore, four main research aims were established:

1. Identify putative CREs that might contribute to the development of disease and their putative target genes (Chapter 2, 3 and 4).
2. Test the functional potential of the identified sequences and the impact of binding sites disruption, using in vivo and in vitro reporter assays (Chapter 2, 3 and 4).
3. Genomic deletion/disruption of CREs to determine their impact in its target genes transcription and its association with disease (Chapter 2, 3 and 4).
4. Prediction of new alleles associated specifically to T2D by high-throughput CRIPSR mediated mutagenesis targeting CREs (Chapter 3).

4. References

1. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat. Struct. Mol. Biol.* **11**, 394–403 (2004).

2. Butler, J. E. F. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).

3. Organization and regulation of gene transcription | Nature. https://www.nature.com/articles/s41586-019-1517-4.

4. Sterner, D. E. & Berger, S. L. Acetylation of histones and transcription-related factors. *Microbiol Mol Biol Rev* **64**, 435–459 (2000).

5. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707–719 (2007).

6. Orphanides, G. & Reinberg, D. A unified theory of gene expression. *Cell* **108**, 439–451 (2002).

7. Pugh, B. F. Control of gene expression through regulation of the TATA-binding protein. *Gene* **255**, 1–14 (2000).

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

9. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13–25 (2014).

10. Kadauke, S. & Blobel, G. A. Chromatin loops in gene regulation. *Biochim. Biophys. Acta* **1789**, 17–25 (2009).

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

12. Barrett, L. W., Fletcher, S. & Wilton, S. D. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* **69**, 3613–3634 (2012).

13. Shabalina, S. A. & Spiridonov, N. A. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol.* **5**, 105 (2004).

14. Müller, D. & Stelling, J. Precise Regulation of Gene Expression Dynamics Favors Complex Promoter Architectures. *PLoS Comput Biol* **5**, (2009).

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

16. Bulger, M. & Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev.* **13**, 2465–2477 (1999).

17. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 60–63 (1998).

18. Chen, L. & Widom, J. Mechanism of transcriptional silencing in yeast. *Cell* **120**, 37–48 (2005).

19. Harris, M. B., Mostecki, J. & Rothman, P. B. Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *J. Biol. Chem.* **280**, 13114–13121 (2005).

20. Heinzel, T. *et al.* A complex containing N-CoR, mSln3 and histone deacetylase mediates transcriptional repression. *Nature* **387**, 43–48 (1997).

21. Ali, T., Renkawitz, R. & Bartkuhn, M. Insulators and domains of gene expression. *Curr. Opin. Genet. Dev.* **37**, 17–26 (2016).

22. Mutskov, V. J., Farrell, C. M., Wade, P. A., Wolffe, A. P. & Felsenfeld, G. The barrier function of an insulator couples high histone acetylation levels with specific protection of promoter DNA from methylation. *Genes Dev* **16**, 1540–1554 (2002).

23. Levine, M. Transcriptional Enhancers in Animal Development and Evolution. *Current Biology* **20**, R754–R763 (2010).

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

25. Atchison, M. L. Enhancers: mechanisms of action and cell specificity. *Annu. Rev. Cell Biol.* **4**, 127–153 (1988).

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

27. Bhattacharjee, S., Renganaath, K., Mehrotra, R. & Mehrotra, S. Combinatorial Control of Gene Expression. *Biomed Res Int* **2013**, (2013).

28. Delic, J., Onclercq, R. & Moisan-Coppey, M. Inhibition and enhancement of eukaryotic gene expression by potential non-B DNA sequences. *Biochem. Biophys. Res. Commun.* **181**, 818–826 (1991).

29. Narlikar, L. & Ovcharenko, I. Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic* **8**, 215–230 (2009).

30. Coppola, C. J., C Ramaker, R. & Mendenhall, E. M. Identification and function of enhancers in the human genome. *Hum. Mol. Genet.* **25**, R190–R197 (2016).

31. Mills, C. *et al.* PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence. *PLOS ONE* **15**, e0243791 (2020).

32. Doyle, B., Fudenberg, G., Imakaev, M. & Mirny, L. A. Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. *PLoS Comput Biol* **10**, e1003867 (2014).

33. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96–100 (2014).

34. Akgol Oksuz, B. *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* **18**, 1046–1055 (2021).

35. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).

36. Kubiak, M. *et al.* Chromosome Conformation Capture Reveals Two Elements That Interact with the PTBP3 (ROD1) Transcription Start Site. *Int J Mol Sci* **20**, 242 (2019).

37. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).

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

39. Shakoori, A. R. Fluorescence In Situ Hybridization (FISH) and Its Applications. *Chromosome Structure and Aberrations* 343–367 (2017) doi:10.1007/978-81-322-3673-3_16.

40. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell* **155**, 27–38 (2013).

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

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

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

44. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).

45. Sullivan, A. M. *et al.* Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Rep* **8**, 2015–2030 (2014).

46. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).

47. Hübner, M. R. & Spector, D. L. Chromatin dynamics. *Annu Rev Biophys* **39**, 471–489 (2010).

48. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834 (2008).

49. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).

50. Robertson, A. G. *et al.* Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res.* **18**, 1906–1917 (2008).

51. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

52. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488–496 (2016).

53. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).

54. Chromatin signatures of active enhancers - PubMed. https://pubmed.ncbi.nlm.nih.gov/22555596/.

55. Bordeira-Carriço, R. *et al.* Multidimensional chromatin profiling of zebrafish pancreas to uncover and investigate disease-relevant enhancers. *Nat Commun* **13**, 1945 (2022).

56. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010).

57. Heintzman, N. D. *et al.* Histone Modifications at Human Enhancers Reflect Global Cell Type-Specific Gene Expression. *Nature* **459**, 108–112 (2009).

58. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**, 292–310 (2020).

59. Eufrásio, A. *et al.* In Vivo Reporter Assays Uncover Changes in Enhancer Activity Caused by Type 2 Diabetes–Associated Single Nucleotide Polymorphisms. *Diabetes* **69**, 2794–2805 (2020).

60. Ryan, G. E. & Farley, E. K. Functional genomic approaches to elucidate the role of enhancers during development. *WIREs Systems Biology and Medicine* **12**, e1467 (2020).

61. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521–530 (2012).

62. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

63. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**, 38–52 (2017).

64. Cm, V. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome research* **25**, (2015).

65. Deng, C. *et al.* Massively parallel characterization of psychiatric disorder-associated and cell-type-specific regulatory elements in the developing human cortex. *bioRxiv* 2023.02.15.528663 (2023) doi:10.1101/2023.02.15.528663.

66. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

67. Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).

68. Regadas, I. *et al.* Several cis-regulatory elements control mRNA stability, translation efficiency, and expression pattern of Prrxl1 (paired related homeobox protein-like 1). *J Biol Chem* **288**, 36285–36301 (2013).

69. Rodríguez-Molina, J. B. & Turtola, M. Birth of a poly(A) tail: mechanisms and control of mRNA polyadenylation. *FEBS Open Bio* **13**, 1140–1153 (2023).

70. Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biology* **9**, 563–576 (2012).

71. Lawless, C. *et al.* Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics* **10**, 7 (2009).

72. Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol* **11**, a034728 (2019).

73. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171–194 (2017).

74. Balbo, P. B. & Bohm, A. Mechanism of Poly(A) Polymerase: Structure of the enzyme-MgATP–RNA ternary complex and kinetic analysis. *Structure* **15**, 1117–1131 (2007).

75. Zhang, Y. *et al.* Alternative polyadenylation: methods, mechanism, function, and role in cancer. *Journal of Experimental & Clinical Cancer Research* **40**, 51 (2021).

76. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**, 201–212 (2005).

77. The Role of Alternative Polyadenylation in Disease. *News-Medical.net* https://www.news-medical.net/health/The-Role-of-Alternative-Polyadenylation-in-Disease.aspx (2022).

78. Moreira, A., Wollerton, M., Monks, J. & Proudfoot, N. J. Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *The EMBO Journal* **14**, 3809 (1995).

79. Eufrásio, A. *et al.* Unravelling the regulatory function of a short 3'UTR sequence in zebrafish and in human cells. 2023.11.15.567165 Preprint at https://doi.org/10.1101/2023.11.15.567165 (2023).

80. Nunes, N. M., Li, W., Tian, B. & Furger, A. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J* **29**, 1523–1536 (2010).

81. Alternative Polyadenylation in Human Diseases. https://www.e-enm.org/journal/view.php?doi=10.3803/EnM.2017.32.4.413.

82. Lutz, C. S. & Moreira, A. Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression. *Wiley Interdiscip Rev RNA* **2**, 23–31 (2011).

83. Oliveira, M. S. *et al.* Cell Cycle Kinase Polo Is Controlled by a Widespread 3' Untranslated Region Regulatory Sequence in Drosophila melanogaster. *Mol Cell Biol* **39**, e00581-18 (2019).

84. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

85. Epstein, D. J. Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic* **8**, 310–316 (2009).

86. Thein, S. L. Molecular basis of β thalassemia and potential therapeutic targets. *Blood Cells Mol Dis* **70**, 54–65 (2018).

87. Miller, C. H. The Clinical Genetics of Hemophilia B (Factor IX Deficiency). *Appl Clin Genet* **14**, 445–454 (2021).

88. de Vooght, K. M. K., van Wijk, R. & van Solinge, W. W. Management of gene promoter mutations in molecular diagnostics. *Clin Chem* **55**, 698–708 (2009).

89. Degtyareva, A. O., Antontseva, E. V. & Merkulova, T. I. Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *Int J Mol Sci* **22**, 6454 (2021).

90. Shastry, B. S. SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet* **52**, 871–880 (2007).

91. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

92. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am J Hum Genet* **93**, 779–797 (2013).

93. Genome-Wide Association Studies (GWAS). *Genome.gov* https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies.

94. What are genome wide association studies (GWAS)? *EMBL-EBI Train online* https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations-2019/what-gwas-catalog/what-are-genome-wide (2019).

95. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. - PubMed - NCBI. https://www.ncbi.nlm.nih.gov/pubmed/30445434.

96. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).

97. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

98. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).

99. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102–R110 (2015).

100. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).

101. Huang, D. & Ovcharenko, I. Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res* **43**, 225–236 (2015).

102. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).

103. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7548–7553 (2002).

104. Emison, E. S. *et al.* A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* **434**, 857–863 (2005).

105. Sribudiani, Y. *et al.* Variants in RET associated with Hirschsprung's disease affect binding of transcription factors and gene expression. *Gastroenterology* **140**, 572-582.e2 (2011).

106. Henning, R. J. Type-2 diabetes mellitus and cardiovascular disease. *Future Cardiology* **14**, 491–509 (2018).

107. Barbagallo, M. & Dominguez, L. J. Type 2 diabetes mellitus and Alzheimer's disease. *World J Diabetes* **5**, 889–893 (2014).

108. Staiger, H., Machicao, F., Fritsche, A. & Häring, H.-U. Pathomechanisms of Type 2 Diabetes Genes. *Endocrine Reviews* **30**, 557–585 (2009).

109. Pathophysiology and pharmacological treatment of insulin resistance - PubMed. https://pubmed.ncbi.nlm.nih.gov/11133066/.

110. β-Cell Failure in Diabetes and Preservation by Clinical Treatment | Endocrine Reviews | Oxford Academic. https://academic.oup.com/edrv/article/28/2/187/2354993.

111. Home *et al.* IDF Diabetes Atlas | Tenth Edition. https://diabetesatlas.org/.

112. Jeon, J., Correa-Medina, M., Ricordi, C., Edlund, H. & Diez, J. A. Endocrine Cell Clustering During Human Pancreas Development. *J Histochem Cytochem* **57**, 811–824 (2009).

113. Jennings, R. E., Berry, A. A., Strutt, J. P., Gerrard, D. T. & Hanley, N. A. Human pancreas development. *Development* **142**, 3126–3137 (2015).

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

115. Mastracci, T. L. & Sussel, L. The Endocrine Pancreas: insights into development, differentiation and diabetes. *Wiley Interdiscip Rev Membr Transp Signal* **1**, 609–628 (2012).

116. Herrera, P. L., Nepote, V. & Delacour, A. Pancreatic cell lineage analyses in mice. *Endocrine* **19**, 267–278 (2002).

117. Kulkarni, R. N. The islet beta-cell. *Int. J. Biochem. Cell Biol.* **36**, 365–371 (2004).

118. Hossain, P., Kawar, B. & El Nahas, M. Obesity and diabetes in the developing world--a growing challenge. *N Engl J Med* **356**, 213–215 (2007).

119. Tuomilehto, J. *et al.* Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* **344**, 1343–1350 (2001).

120. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. *N Engl J Med* **346**, 393–403 (2002).

121. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics* **46**, 136–143 (2014).

122. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. *Lancet* **389**, 2239–2251 (2017).

123. Ali, O. Genetics of type 2 diabetes. *World J Diabetes* **4**, 114–123 (2013).

124. Jali, M. V., Kambar, S., Jali, S. M. & Gowda, S. Familial early onset of type-2 diabetes mellitus and its complications. *N Am J Med Sci* **1**, 377–380 (2009).

125. Elbein, S. C. The Genetics of Human Noninsulin-Dependent (Type 2) Diabetes Mellitus1. *The Journal of Nutrition* **127**, 1891S-1896S (1997).

126. Wang, J. *et al.* A mutation in the insulin 2 gene induces diabetes with severe pancreatic beta-cell dysfunction in the Mody mouse. *J Clin Invest* **103**, 27–37 (1999).

127. Gloyn, A. L. The search for type 2 diabetes genes. *Ageing Research Reviews* **2**, 111–127 (2003).

128. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981–990 (2012).

129. Miguel-Escalada, I. *et al.* Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics* **51**, 1137–1148 (2019).

130. Belton, J.-M. *et al.* Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

131. Bysani, M. *et al.* ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes. *Sci Rep* **9**, 7785 (2019).

132. Mularoni, L., Ramos-Rodríguez, M. & Pasquali, L. The Pancreatic Islet Regulome Browser. *Front Genet* **8**, (2017).

133. Majello, B., De Luca, P. & Lania, L. Sp3 Is a Bifunctional Transcription Regulator with Modular Independent Activation and Repression Domains*. *Journal of Biological Chemistry* **272**, 4021–4026 (1997).

134. Hiramoto, M. *et al.* 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. *Int. J. Mol. Med.* **41**, 717–728 (2018).

135. Bonnycastle, L. L. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).

136. Hirsch, N. & Birnbaum, R. Y. Dual Function of DNA Sequences: Protein-Coding Sequences Function as Transcriptional Enhancers. *Perspect. Biol. Med.* **58**, 182–195 (2015).

137. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev* **25**, 1770–1782 (2011).

138. Danckwardt, S., Hentze, M. W. & Kulozik, A. E. 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J* **27**, 482–498 (2008).

139. Implications of polyadenylation in health and disease - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4615197/.

140. Lin, Y. *et al.* An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res* **40**, 8460–8471 (2012).

141. Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684 (2009).

142. Shortening of 3′UTRs Correlates with Poor Prognosis in Breast and Lung Cancer | PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031129.

143. Thomas, L. F. & Sætrom, P. Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microRNA-regulation. *PLoS Comput Biol* **8**, e1002621 (2012).

144. Rehfeld, A., Plass, M., Krogh, A. & Friis-Hansen, L. Alterations in Polyadenylation and Its Implications for Endocrine Disease. *Frontiers in Endocrinology* **4**, (2013).

145. Garin, I. *et al.* Recessive mutations in the INS gene result in neonatal diabetes through reduced insulin biosynthesis. *Proc Natl Acad Sci U S A* **107**, 3105–3110 (2010).

146. del Bosque-Plata, L., Martínez-Martínez, E., Espinoza-Camacho, M. Á. & Gragnoli, C. The Role of TCF7L2 in Type 2 Diabetes. *Diabetes* **70**, 1220–1228 (2021).

147. Engeszer, R. E., Patterson, L. B., Rao, A. A. & Parichy, D. M. Zebrafish in the wild: a review of natural history and new notes from the field. *Zebrafish* **4**, 21–40 (2007).

148. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev Dyn* **203**, 253–310 (1995).

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

150. Amsterdam, A. & Hopkins, N. Mutagenesis strategies in zebrafish for identifying genes involved in development and disease. *Trends Genet* **22**, 473–478 (2006).

151. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).

152. Tiso, N., Moro, E. & Argenton, F. Zebrafish pancreas development. *Mol. Cell. Endocrinol.* **312**, 24–30 (2009).

153. Schiavone, M. *et al.* Zebrafish reporter lines reveal in vivo signaling pathway activities involved in pancreatic cancer. *Dis Model Mech* **7**, 883–894 (2014).

154. Zang, L., Torraca, V., Shimada, Y. & Nishimura, N. Editorial: Zebrafish Models for Human Disease Studies. *Frontiers in Cell and Developmental Biology* **10**, (2022).

155. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).

156. Chen, X., Gays, D. & Santoro, M. M. Transgenic Zebrafish. *Methods Mol Biol* **1464**, 107–114 (2016).

157. Mobilization of quiet, endogenous Tc3 transposons of Caenorhabditis elegans by forced expression of Tc3 transposase - PubMed. https://pubmed.ncbi.nlm.nih.gov/8389698/.

CHAPTER 2

IN VIVO REPORTER ASSAYS UNCOVER CHANGES IN ENHANCER ACTIVITY CAUSED BY TYPE 2 DIABETES -ASSOCIATED SINGLE NUCLEOTIDE POLYMORPHISMS

This chapter was published as:

Ana Eufrásio, Chiara Perrod, Fábio J. Ferreira, Marta Duque, Mafalda Galhardo and José Bessa*. In vivo Reporter Assays Uncover Changes in Enhancer Activity Caused by Type 2 Diabetes–Associated Single Nucleotide Polymorphisms. Diabetes

* co-corresponding author

Diabetes 2020;69(12):2794–2805

https://doi.org/10.2337/db19-1049

Author Contributions. A.E. carried out the experiments. A.E., M.D., and J.B. wrote the manuscript. A.E. and J.B. conceived, designed, and analyzed the data. C.P. performed the ChIP-seq experiment and CRISPRa and CRISPRi assays. F.J.F. performed the 4C-seq assay. M.D. contributed to the development of the sst:mCherry reporter line. M.G. performed the bioinformatic analysis of the 4C-seq experiment. J.B. designed and supervised the study. All authors revised the manuscript. J.B. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

1. Introduction

Type 2 diabetes affects >300 million people, causing severe complications and premature death[1], yet the underlying molecular mechanisms are largely unknown. This disease is highly complex, multifactorial, and partially characterized by endocrine pancreatic dysfunction, leading to insufficient insulin production[1]. Genome-wide association studies (GWASs) have identified several single nucleotide polymorphisms (SNPs) associated with an increased risk of type 2 diabetes[2–4]. Part of these variants are located in noncoding sequences with epigenetic marks associated to enhancer activity known to regulate the expression of their target genes by interacting with their promoters[5] and some overlap with transcription factor (TF) binding sites (TFBSs) required for proper endocrine pancreatic function[6–10]. In this way, type 2 diabetes–associated SNPs may ultimately translate into transcriptional changes of the target genes[6–10]. Methods to predict enhancers include profiling chromatin accessibility[11] and histone modifications (e.g., H3K27ac, H3K4me1)[12]. The majority of the enhancer testing assays are performed in vitro in specific cell lines, missing cellular diversity and physiological contexts. To overcome these limitations, animal models have been used[13,14]. The zebrafish has been successfully used for the study of pancreas development and function[15], having an endocrine compartment with the same cell types (α-, β-, δ-, and ε-cells) and functions as in mammal pancreas[16,17]. Additionally, orthologous TFs operate in the zebrafish pancreas during early development, some of which are also important for adult pancreas maintenance. As in humans, Pdx1 plays a crucial role in zebrafish pancreas development and β-cell maturation and function[18], and Nkx6.1 is required for the identity of endocrine pancreatic progenitors[19,20].

Some coding mutations are associated with the development of type 2 diabetes, while others might confer a protective effect[2,21]. Interestingly, *SLC30A8*, a zinc transporter–encoding gene, shows contradicting results. The coding SNP rs13266634 located in the *SLC30A8* gene is associated with an increased risk for type 2 diabetes[3] because of a tryptophan-to-arginine switch at protein position 325, causing reduced zinc transport activity[22,23]. Zinc is essential for insulin packaging and maturation and secretion in β-cells[24]; thus, the decrease of SLC30A8 activity is the simplest explanation for the increased risk for the disease. Surprisingly, recently identified protein-truncating SNPs in *SLC30A8* have been associated with a protective effect[2,25] by enhanced insulin secretion[26]. Further work is needed to clarify the type 2 diabetes association with *SLC30A8* SNPs. An unexplored explanation for these apparently contradicting results could be that the coding SNP rs13266634 exerts a specific impact on adjacent or overlapping cis-regulatory sequences.

In this work, we investigated the impact that SNPs located in putative enhancer regions have in enhancer activity, using an in vivo approach. Ten sequences that overlap with type 2 diabetes–associated loci and with marks for enhancer activity were tested, one overlapping with an exon of *SLC30A8* (seq132wt). To test sequences for enhancer activity in the endocrine pancreas, we performed in vivo mosaic transgenesis assays in zebrafish embryos. We show that this strategy is sensitive, has low noise, and can be quantitative to address enhancer activity. Using this method, we observed that 6 out of 10 tested sequences are endocrine pancreatic enhancers, including the *SLC30A8* exon-containing sequence. In addition, two sequences were found to be pancreatic progenitor enhancers. We also found that the type 2 diabetes–associated SNP (henceforth referred to as risk allele) decreased the enhancer activity of two enhancers, while the risk allele of two sequences resulted in a gain of enhancer activity. Interestingly, the *SLC30A8* coding risk allele (seq132risk) showed an increase in enhancer activity, demonstrating that coding SNPs have the potential to modulate the target gene activity at both the transcriptional and the protein level. To better understand how SNPs can affect enhancer activity, we focused on the seq132 enhancer. We divided seq132wt into different fragments, observing that all are necessary for a robust pancreatic enhancer activity. We also show that common SNPs in seq132 modulate its pancreatic enhancer activity and that a single nucleotide mutation ablates completely the endocrine pancreatic enhancer activity of seq132. Additionally, we observed a chromatin interaction between seq132 and the promoter of Slc30a8 gene in murine cells and noted that targeting transcriptional modulators to seq132 using CRISPR affects the transcription of *Slc30a8*, strongly suggesting that the seq132 enhancer belongs to the regulatory landscape of *Slc30a8*. Overall, in this work, we use an in vivo system to validate enhancers that overlap with type 2 diabetes–associated SNPs, showing several cases where nucleotide variations result in complex changes in enhancer activity, including a classical and poorly understood coding SNP.

2. Research design and methods

2.1. Zebrafish husbandry and embryo culture

Zebrafish (Danio rerio) were handled according to European animal welfare regulations and standard protocols. Embryos were cultured at 28°C in Petri dishes containing E3 medium supplemented with 1-phenyl-2-thiourea to delay pigmentation formation[27].

2.2. Putative enhancer selection

Putative enhancer sequences were selected on the basis of GWAS data that uncovered 163 SNPs associated with type 2 diabetes or glycemic traits (P < 5e-8),

considering all variants in high linkage disequilibrium (1000 Genomes Project Utah residents with ancestry from northern and western Europe r2 > 0.8), with lead GWAS SNPs being the exception for rs735949 (P < 3.70e-6)[6,28–30]. Risk alleles associated with seq132, seq117, and seq790 are part of the previously described GWAS credible set SNPs[31]. Sequences were analyzed using the Islet Regulome Browser[32], which identifies active enhancers by the presence of H3K4me1, H3K27ac, and H2A.Z epigenetic marks in adult human endocrine pancreatic samples. The analysis was further refined by the presence of PDX1, MAFB, NKX6.1, FOXA2, and NKX2.2 binding obtained by chromatin immunoprecipitation sequencing (ChIP-seq) profiles[6,28], resulting in a list of 10 putative enhancer sequences, each overlapping with one SNP associated with type 2 diabetes (Supplementary Table 1).

2.3. In vivo mosaic transgenesis assays

Zebrafish transgenesis was performed using the Tol2 transposon system[33]. One-cell embryos from the Tg(sst:mCherry) zebrafish reporter line were microinjected with 3 nL containing 25 ng/μL Tol2 transposase mRNA and 25 ng/μL phenol/chloroform-purified reporter vector. Injections were performed at least two times. Ins:GFP reporter was built by isolating the insulin promoter from the ins-CFP-NTR vector (SacI and BamHI)[34], cloning it in a pEM-MCS vector[35] and recombining it to a Tol2-based transposon containing a Gateway cassette and a GFP reporter gene. Human sequences were PCR amplified from human genomic DNA using specific primers (Supplementary Table 2), cloned in TOPO vector (pCR8/GW/TOPO TA Cloning KIT; Invitrogen) and recombined in vitro to the Z48 transgenesis vector[36] by the Gateway system. Risk SNPs from seq58, seq68, seq73, seq219, and seq460 were inserted by site-directed mutagenesis using specific primers containing the risk variant (Supplementary Table 2). Injected embryos showing expression of GFP in the midbrain were selected for immunohistochemistry at 48 hours post fertilization (hpf).

2.4. *sst*:mCherry reporter Line

To identify the zebrafish endocrine pancreatic domain, we developed an in vivo reporter line that drives expression of mCherry in δ-cells. Primers for the somatostatin (sst) promoter amplification were designed (Supplementary Table 2) and the amplified fragment cloned in a Tol2 transposon containing mCherry as reporter gene. The vector was microinjected in one-cell embryos using the Tol2 transposon system. Embryos were selected for sst:mCherry-positive cells and raised until adulthood and a stable transgenic line was isolated.

2.5. Immunohistochemistry

The 48-hpf microinjected embryos were dechorionated and fixed in 4% formaldehyde in PBS (PBS1×) overnight at 4°C and then washed in PBS-T (0.5% Triton X-100 in PBS1×) with 1% Triton X-100 in PBS1× (2 h) and 5% BSA in PBS-T (0.1%). Embryos were incubated with anti-Nkx6.1 (1:75) (F55A12; Developmental Studies Hybridoma Bank) and anti-insulin (1:50) (ab210560; Abcam) diluted in 5% BSA-PBS-T followed by washing. Embryos were then incubated with anti-mouse Alexa Fluor 647 (1:800), anti-rabbit Alexa Fluor 647 (1:800) (Thermo Fisher Scientific), and DAPI (1:1,000) (Invitrogen) diluted in 5% BSA-PBS-T. Embryos were washed and stored in 50% glycerol in PBS1×. Microscopy slides were prepared using 50% glycerol in PBS1×. Confocal imaging was performed using a Leica SP5II confocal microscope.

2.6. Assessment of enhancer activity

Embryos were analyzed, using confocal microscopy, for the presence of GFP-positive cells in the endocrine pancreatic domain (*sst*:mCherry reporter domain) or in the endocrine progenitor domain (anti-Nkx6.1). One embryo was considered positive if at least one GFP-positive cell was detected within the endocrine pancreatic domain or progenitor domain. Quantifications are presented as percentages of positive embryos to ensure the quantification of different transposon integrations.

2.7. Prediction of TFBSs affected by type 2 Diabetes risk variants

The wild-type (wt) and risk variant sequences were analyzed using 719 specific position weight matrices for vertebrate TFs using JASPAR software[37]. TFs were ranked by position-specific score matrix. The relative score is a threshold score between 0 and 1 and is calculated by (score − minimum score) / (maximum score − minimum score), meaning 1 is the highest affinity and 0 is no affinity of binding[37]. TFs that showed differential binding affinity in wt and risk variants were selected and filtered by presence of H3K4me3 at their promoters[6,32].

2.8. ChIP on plasmid

Seq119wt and risk were cloned into a pLVX lentiviral backbone (#125839; Addgene) as KpnI-ApaI (Anza) fragments. Lentiviral particles were produced in HEK-293 cells (packaging plasmids psPAX2, #12260, and pCMV-VSV-G, #8454; Addgene) and used to infect MIN6 cells (a gift from Lorenzo Pasquali) according to standard procedures. Infected cells were selected with puromycin (1 µg/mL; Sigma-Aldrich) for 12 days, starting 48 h after infection. Three to 10 million cells were used for ChIP[12] with 4 µg of Nkx6.1 antibody (F5510; Developmental Studies Hybridoma Bank) and magnetic Dynabeads (Thermo

Fisher Scientific). Eluted chromatin was purified with a MiniElute Kit (QIAGEN). Immunoprecipitated DNA was dissolved in water and further analyzed by real-time PCR (iTaq Universal SYBR Green Supermix, CFX 384; Bio-Rad).

2.9. Real-time quantitative expression analysis

MIN6 cells were harvested for RNA extraction with TRIzol (Ambion) and treated with DNase (Thermo Fisher Scientific). Five hundred nanograms to 1 µg of DNA-free RNA was retrotranscribed with iScript cDNA Synthesis Kit (Bio-Rad). cDNA was used for quantitative PCR (iTaq Universal SYBR Green Supermix, CFX 384). *Slc30a8* expression was calculated by the ΔCt method to actb housekeeping mRNA.

2.10. 4C sequencing

The 4C sequencing (4C-seq) was performed on 10 million MIN6 cells using sequential DpnII and Csp6I as previously described[38], with minor modifications. The 4C template was purified using an Amicon Ultra-15 Centrifugal Filter Unit (Millipore). Two libraries were independently prepared with the Expand Long Template PCR System (Roche) using specific primers (Supplementary Table 2). Libraries were purified with QIAquick PCR Purification Kit (QIAGEN) followed by the Agencourt AMPure XP reagent (Beckman Coulter). Libraries were sequenced on an Ion S5 XL System (Ion 540 Chip, Ion Torrent; Thermo Fisher Scientific). Previously described processing[39,40] was used with a custom Perl script. More than 3.5 million reads were aligned to the mouse genome (mm10) using Bowtie2 (default parameters, global alignment mode)[41]. Reads within fragments flanked by restriction sites of the same enzyme (checked with bedtools) or fragments <40 base pairs (bp) were filtered out. Mapped reads were then converted to reads-per-first-enzyme-fragment-end units and smoothed using a 30-fragment mean running window algorithm.

2.11. CRISPR inactivation and CRISPR activation targeting

Twenty-nucleotide single guide RNAs (sgRNAs) targeting the murine *Slc30a8* enhancer with high predicted cleavage were selected from the UCSC Genome Browser CRISPR target track and cloned into the lentiviral backbone for enhancer inactivation (CRISPRi) [in Lenti-(BB)-hPGK-KRAB-dCas9-2A-BlastR, #118155; Addgene] or activation (CRISPRa) (in lentiSAMv2, #75112; Addgene) as previously described[42]. Lentiviral particles were produced in HEK-293 cells (packaging plasmids: pRSV-rev, #12253; pMDLg/pRRE, #12251; and pMD2G, #12259; Addgene) and used to infect MIN6 cells according to standard procedures. Infected cells were selected by blasticidin (8 µg/mL) (Sigma-Aldrich) for 12 days, starting 48 h after infection.

2.12. Statistical analysis

Statistical analyses were performed by using the χ2 test with Fisher correction and unpaired t test, applying a significance level of $P \leq 0.05$. For real-time expression experiments, statistical analysis was performed with the Mann-Whitney test.

2.13. Data and resource availability

All data generated or analyzed during this study are included in the published article (and its Supplementary Material), with the exception of 4C-seq sequencing data. The 4C-seq data sets have been deposited in the European Nucleotide Archive at EMBL-EBI under accession number PRJEB39688 (https://www.ebi.ac.uk/ena/browser/view/PRJEB39688). The *sst*:mCherry reporter line generated during the current study is available from the corresponding author upon reasonable request.

3. Results

3.1. Identification of endocrine pancreatic enhancers in vivo by mosaic transgenesis in zebrafish

In vivo enhancer reporter assays can be performed in zebrafish either by generating stable transgenic lines, a time-consuming approach, or by mosaic transgenesis[43] on the basis of the analysis of many independent integration events. To test sequences for endocrine pancreatic enhancer activity, we used a Tol2 transposon[44] containing a minimal promoter, a GFP reporter gene, and a midbrain-specific enhancer (Z48) acting as an internal control of transgenesis[36] (Fig. 1A). As an endocrine marker, we developed an in vivo reporter line that drives expression of mCherry in δ-cells (*sst*:mCherry). To validate sst:mCherry as an endocrine pancreatic reporter, we generated double-positive embryos for the sst:mCherry and insulin (ins:GFP) reporter transgenes (Fig. 1B). At 48 hpf, all GFP-positive cells (ins:GFP) were located within the *sst:*mCherry expression pattern (Fig. 1C), indicating that the *sst*:mCherry reporter line can be used to define the endocrine pancreatic domain. Next, to understand whether the mosaic strategy to identify endocrine pancreatic regulatory elements was sensitive enough, we mobilized a Tol2 transposon containing the insulin promoter upstream of GFP. The mosaic analysis of 48-hpf–injected embryos, using confocal microscopy, revealed that 69% (n = 23) showed GFP expression in the endocrine pancreatic domain (Fig. 1D and Supplementary Fig. 1). Random integrations of the Z48 transgenesis vector can generate noise as a result of the influence of regulatory elements located in the genomic landscapes of each integration, termed position effect[43]. To test whether the Z48 transgenesis vector was prone to position effect, we mobilized this vector without a sequence to test (negative control [NC]). These injected embryos did not show

expression of GFP in the endocrine pancreas (0%, n = 43) at 48 hpf. In total, these results show that the use of mosaic transgenic embryos is sensitive enough to identify endocrine pancreatic regulatory sequences and that the associated noise as a result of the position effect is very low (Fig. 1D and Supplementary Fig. 1). Because endocrine pancreatic enhancers might also be active in pancreatic progenitor cells, we asked whether mosaic transgenesis assays could also be applied for these cell types. To test this hypothesis, we used an anti-Nkx6.1 antibody to define the endocrine pancreatic progenitor domain (Fig. 1E), and we mobilized the Z48 vector containing a known progenitor enhancer from the human SOX9 locus[36]. Twenty-seven percent (n = 11) of embryos showed expression of GFP within the pancreatic progenitor domain labeled by Nkx6.1, contrasting with the NC, for which GFP was not detected (0%, n = 23) (Fig. 1F and Supplementary Fig. 1).

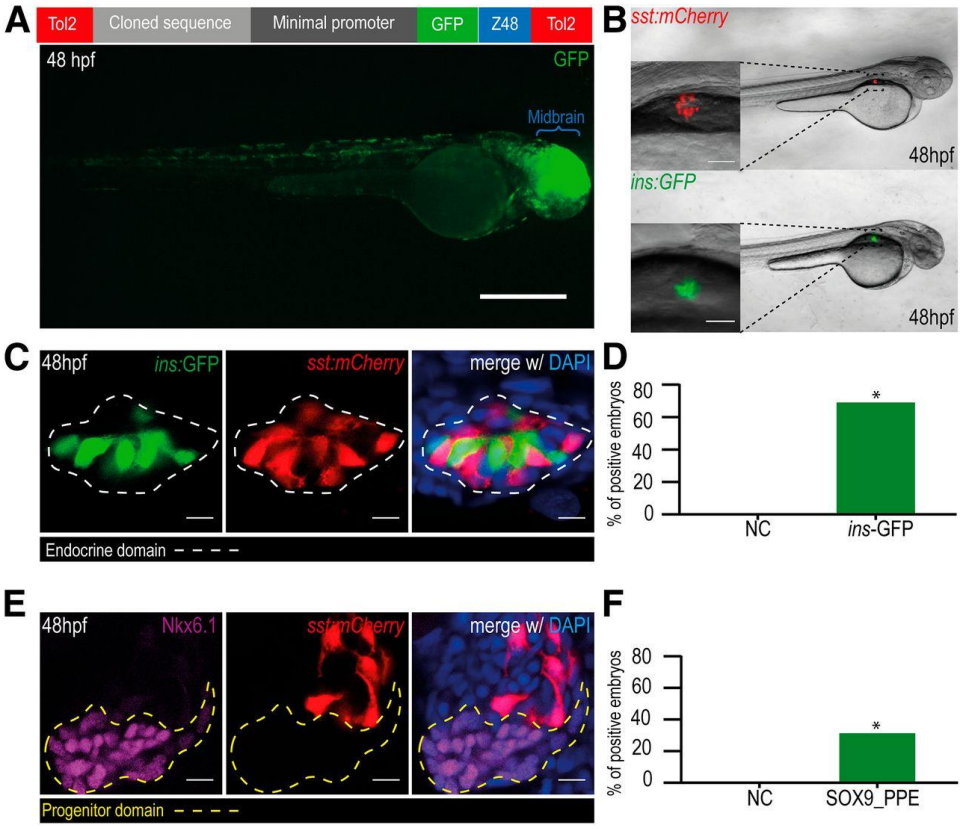


Figure 1. A*: Schematic representation of the Z48 vector (top) and a representative image of a Z48-injected embryo (bottom) showing GFP expression in the midbrain, mediated by the Z48 enhancer (blue), at 48 hpf. The expression of GFP in the midbrain functions as an internal control of transgenesis. Scale bar = 200 µm.* B*: Representative images of* sst:mCherry *(top) and* ins:*GFP (bottom) reporter lines at 48 hpf, showing mCherry and GFP expression in δ- and β-cells, respectively. Scale bars = 50 µm.* C*: Confocal images showing the endocrine pancreatic domain (dashed line), defined by the cross of the* sst:mCherry *and* ins:GFP *reporter lines. The 48-hpf embryos were counterstained with the nuclear marker DAPI. Scale bars = 10 µm.* D*: Percentage of embryos showing GFP-positive cells within the endocrine domain when injected with a vector containing GFP as reporter gene under the control of the insulin promoter (*ins:GFP*) (69%,* n = *23) or with the Z48 vector without an endocrine enhancer (NC).* E*: Confocal images from 48-hpf embryos stained with anti-Nkx6.1 antibody (purple) to define the progenitor pancreatic domain (dashed line) in the* sst:mCherry *reporter line and the nuclear marker DAPI. Scale bars = 10 µm.* F*: Graph representing the percentage of embryos with GFP expression in progenitor cells when injected with a pancreatic progenitor enhancer (SOX9_PPE) (27%,* n = *11) or the NC (0%,* n = *43). *P < 0.05, by* χ^2 *test.*

3.2. Identification of endocrine pancreatic enhancers overlapping with type 2 diabetes–associated SNPs

We selected 10 sequences (Supplementary Table 1) that overlap with SNPs previously associated with type 2 diabetes that are enriched for enhancer marks (H3K4me1, H3K27ac, and H2A.Z) and for TFBSs of endocrine pancreas TFs (FOXA2, NKX2.2, NKX6.1, MAFB, and PDX1)[6] (Fig. 2A and Supplementary Fig. 2A–H). Nine of these sequences are noncoding while one, seq132, partially overlaps with a coding exon of the *SLC30A8* gene. The respective sequences that do not contain the type 2 diabetes–associated variant (wt alleles) were cloned in Z48 transgenesis vector, and in vivo enhancer assays were performed by mosaic transgenesis in zebrafish embryos (36). Out of the 10 tested sequences, 6 showed a consistent expression of GFP in the endocrine pancreatic domain, therefore being endocrine pancreatic enhancers (seq58wt, seq68wt, seq73wt, seq132wt, seq219wt, and seq460wt) (Fig. 2B and D and Supplementary Figs. 3–5). Stable transgenic lines were generated for three of these sequences (seq132wt, seq460wt, and seq58wt) to confirm their endocrine pancreatic enhancer activity (Supplementary Fig. 6). Interestingly, for at least three of the tested sequences (seq58, seq68, and seq73), the GFP signal was also detected adjacent to the endocrine pancreatic domain (Supplementary Fig. 7), suggesting that these sequences may be pancreatic progenitor enhancers. To address this hypothesis, we labeled embryos injected with reporters of seq58, seq68, and seq73 with anti-Nkx6.1 antibody, showing that seq68 and seq73 drive GFP expression in the Nkx6.1-positive progenitor domain (45% [n = 13] and 25% [n = 12], respectively) (Fig. 2C and E), being therefore identified as pancreatic progenitor enhancers. To further characterize the identified enhancers, we determined in which endocrine pancreatic cell types they drive expression. For that, we injected the Z48 transgenesis vector containing the respective enhancers in a *gcga*:mCherry reporter line (α-cells), counterstaining these embryos with anti-insulin to label β-cells. We found that the majority of tested enhancers were able to drive expression in β-cells and that most of them were able to drive expression in more than one cell type (Supplementary Fig. 8).

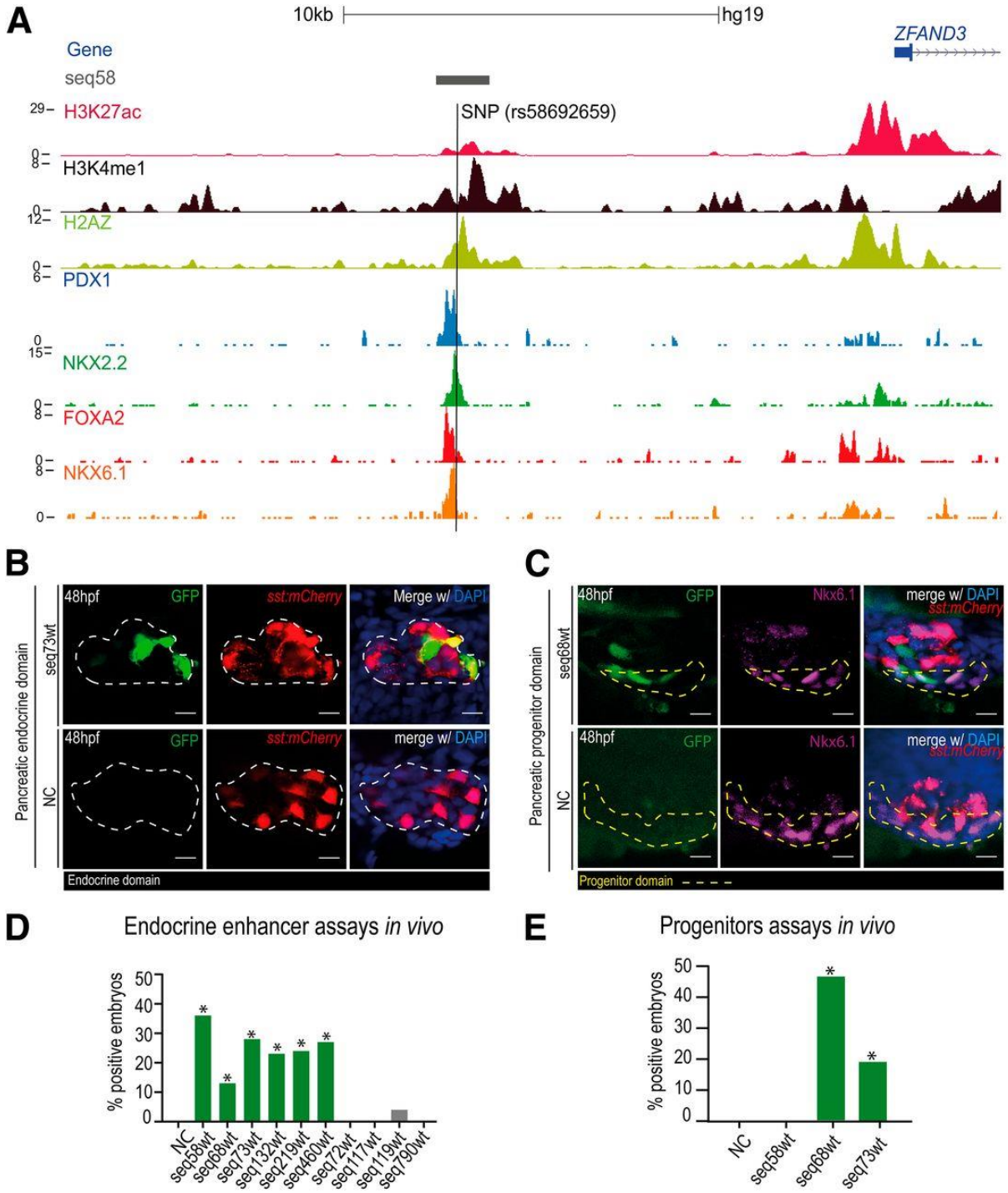


*Figure 2. A: Genomic landscapes of the putative enhancer seq58. Tracks represent H3K27ac, H3K4me1, histone variant H2A.Z, and TF binding (PDX1, NKX2.2, FOXA2, and NKX6.1) from ChIP-seq data of human endocrine pancreatic samples. Human ZFAND3 is the nearest gene to the putative enhancer (blue). The location of the type 2 diabetes–associated SNP (rs58692659) is represented as a vertical black line. B: In vivo reporter assay for endocrine pancreatic enhancers. Top panels show a representative confocal image a sst:mCherry zebrafish embryo injected with the Z48 enhancer reporter vector containing the seq73wt sequence, showing GFP-positive cells within the endocrine pancreatic domain (dashed line), contrasting that absent in embryos injected with NC (bottom). Scale bars = 10 μm. C: In vivo reporter assay for pancreatic progenitor enhancers. Confocal analysis of seq68 reporter assay shows colocalization of GFP-positive cells with Nkx6.1 progenitor marker, contrasting with NC for the pancreatic progenitor domain (dashed line) that did not show GFP-positive cells. All vectors were injected in the sst:mCherry reporter line and embryos analyzed at 48 hpf and stained with DAPI. Scale bars = 10 μm. D: Graph representing the percentage of embryos with GFP expression in endocrine pancreatic domain at 48 hpf for each sequence analyzed: seq58wt (36%, n = 56), seq68wt (13%, n = 47), seq73wt (28%, n = 47), seq132wt (23%, n = 34), seq219wt (24%, n = 38), seq460wt (27%, n = 36), seq72wt (0%, n = 27), seq117wt (0%, n = 21), seq119wt (4%, n = 27), seq790wt (0%, n = 20), and NC (0%, n = 43). E: Graph representing the number of embryos with GFP expression in pancreatic progenitor domain, defined by Nkx6.1 staining at 48 hpf, for each sequence analyzed: seq58wt (0%, n = 12), seq68wt (46%, n = 13), seq73wt (25%, n = 12), and NC (0%, n = 43). *P < 0.05, by χ2 test.*

3.3. SNPs associated with increased risk of type 2 diabetes modulate enhancer activity

To address the possible impact that type 2 diabetes–associated SNPs have in overlapping enhancers, we tested the corresponding variants (risk alleles), performing enhancer assays for endocrine pancreas. Of the six previously identified endocrine pancreatic enhancers, two (seq58risk and seq219risk) showed a decreased enhancer activity for the respective risk allele and two an increase (seq68 and seq132) compared with the wt allele (Fig. 3A and B and Supplementary Figs. 9 and 10). Strikingly, for seq132, the risk allele is in a coding exon of *SLC30A8* (seq132risk 56%, n = 36; seq132wt 23%, n = 34) (Fig. 3B). For one sequence, the risk allele was able to drive GFP expression in the endocrine pancreas above the established threshold, while the wt allele did not (seq119risk 14%, n = 28; seq119wt 4%, n = 27) (Fig. 3B). Overall, these results demonstrate that type 2 diabetes–associated SNPs have the potential to modulate the activity of enhancers in a sequence-specific manner. Differential binding of TFs to wt and risk alleles could explain the observed differential enhancer activity. To test this hypothesis, we generated a stable mouse MIN6 β-cell line containing human wt and risk sequences of the seq119 enhancer. Seq119 risk showed both an increased predicted affinity to Nkx6.1 binding (Supplementary Fig. 11A) and increased enhancer activity (Fig. 3B). Performing ChIP-PCR, we demonstrated that Nkx6.1 binds with higher affinity to the seq119 risk variant (Supplementary Fig. 11B). Additionally, we predicted bioinformatically TFBSs for wt and risk alleles of each sequence (Supplementary Table 3). Sequences were then clustered in two groups: sequences that had shown differential enhancer activity between wt and risk alleles (seq58, seq68, seq119, seq132, and seq219) and sequences that

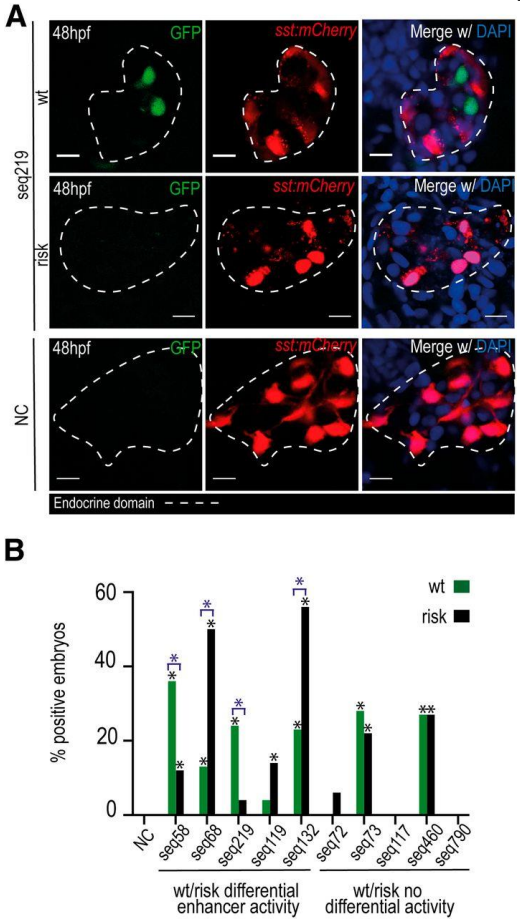


*Figure 3. A: Representative confocal images for seq219wt, seq219risk, and NC. Seq219wt showed GFP expression in endocrine pancreatic domain (dashed line), defined by the sst:mCherry reporter line. The 48-hpf embryos were stained with DAPI. Scale bars = 10 μm. B: Graph showing the total percentage of positive embryos in wt and risk alleles for each of the 10 sequences analyzed: seq58wt/risk (36%, n = 56; 12%, n = 43, respectively), seq68wt/risk (13%, n = 47; 50%, n = 32), seq73wt/risk (28%, n = 47; 22%, n = 36), seq132wt/risk (23%, n = 34; 56%, n = 36), seq219wt/risk (24%, n = 38; 6%, n = 35), seq119wt/risk (4%, n = 27; 14%, n = 28), seq72wt/risk (0%, n = 27; 6%, n = 30), seq117wt/risk (0%, n = 21; 0%, n = 8), seq460wt/risk (27%, n = 36; 27%, n = 30), and seq790wt/risk (0%, n = 20; 0%, n = 14). Six sequences showed differential enhancer activity between wt and risk allele. *P < 0.05, by χ2 test.*

did not (seq72, seq73, seq117, seq460, and seq790). Although both groups showed a similar number of predicted TFBSs in the wt allele, the differential activity enhancers group showed a higher number of predicted differential binding between wt and risk alleles (Supplementary Fig. 11C). These results suggest that differential binding of TFs might control the regulatory output of wt and risk variants.

3.4. The enhancer seq132mm belongs to the *Slc30a8* regulatory landscape

Among the detected enhancers, we found that seq132, a sequence that partially overlaps with an exon of *SLC30A8*, is an endocrine pancreatic enhancer. Additionally, we showed that the type 2 diabetes–associated risk allele (seq132risk), which encodes a tryptophan-to-arginine substitution causing a decrease in the function of SLC30A8 (24), has increased enhancer activity compared with the wt allele (seq132wt). To determine whether seq132 belongs to the regulatory landscape of *SLC30A8*, we used the MIN6 cell line to detect chromatin interaction points, since enhancers contact the promotor of the genes that they control. First, we performed enhancer assays in zebrafish, demonstrating that the mouse orthologous sequence (seq132mm) of the seq132 human enhancer is also an endocrine pancreas enhancer (Fig. 4A). Then, using 4C-seq[38], we observed the existence of an interaction between the Slc30a8 promoter and seq132mm (Fig. 4B and Supplementary Fig. 12). To further validate that seq132mm belongs to the regulatory landscape of *Slc30a8*, we targeted seq132mm using the CRISPR/Cas9 system with a dCas9 fused to a transcriptional activation domain (CRISPRa) and another to a repressor domain (CRISPRi), observing a significant increase and decrease of *Slc30a8* expression levels, respectively (Fig. 4C). These results strongly suggest that seq132mm belongs to the regulatory landscape of *Slc30a8*, and because of the remarkable conservation in the activity and sequence of this enhancer, we propose that this regulatory mechanism is conserved in humans.

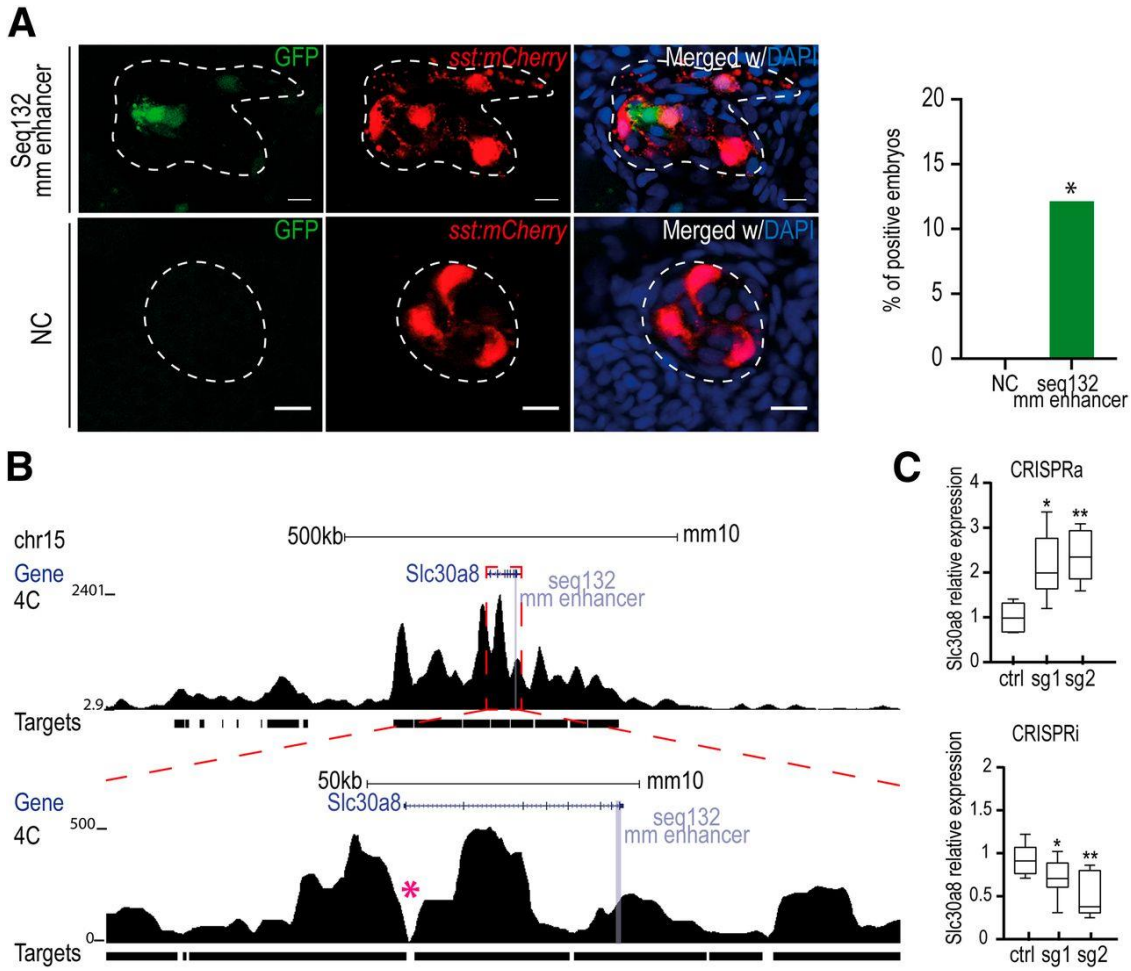


*Figure 4. A: Representative confocal images for the mouse seq132 (seq132mm enhancer chr15:52334298 + 52335281). Seq132mm enhancer showed GFP-positive cells within the endocrine pancreatic domain (dashed line) defined by the sst:mCherry reporter line. The 48-hpf embryos were stained with DAPI. Scale bars = 10 μm. Representative graph showing the total percentage of positive embryos for the sequence analyzed (13%, n = 35). *P < 0.05, by χ2 test. B: Genomic landscape of the mouse Slc30a8 gene (blue), showing 4C-seq profiles (black), with view point in Slc30a8 promoter (pink asterisk) in the Min6 cell line; zoom-out (top) and zoom-in (bottom) of the Slc30a8 gene and seq132mm enhancer. The targets line represents the regions where the interaction is significant. C: sgRNAs targeting murine Slc30a8 enhancer in CRISPRa or CRISPR assay in Min6 cells. Slc30a8 expression was calculated relative to the β-actin housekeeping gene by quantitative PCR. Dot blots represent the 10–90% quantile of six biological replicates. *P < 0.05, **P < 0.01. ctrl, control.*

3.5. The *SLC30A8* seq132 enhancer is divided into different functional domains

Next, we wanted to understand whether seq132 is divided into different functional domains. For that, we divided seq132wt into four fragments (Fig. 5A) and performed enhancer assays for each (Fig. 5B). Fragments seq132wt1 (872 bp) and seq132wt2 (967 bp) showed a milder endocrine pancreatic enhancer activity than the seq132wt total fragment (seq132wt 23%, n = 34; seq132wt1 9.8%, n = 41; seq132wt2 6.5%, n = 31). We also tested another fragment, seq132wt3 (899 bp), that contains seq132wt1 and extends to the end of the coding sequence of *SLC30A8*. Seq132wt3 was able to drive GFP expression in endocrine pancreatic cells (10%, n = 30), as was the remaining fragment seq132wt4 (788 bp), although with a decreased efficiency (4%, n = 23) (Fig. 5C). From these results, we

conclude that seq132 has several functional domains spread through this sequence, and the sum of these parts is necessary for this enhancer to be fully functional. These results also suggest that other SNPs could potentially affect the output of this enhancer. To test this, we have performed enhancer assays with seq132risk, which contains three other common SNPs with no known association to type 2 diabetes (seq132risk#: rs2466296, rs2466295, and rs2466294) (Fig. 6A). Interestingly, seq132risk# was a less active enhancer than seq132risk, showing an activity similar to seq132wt (Fig. 6B and C) and demonstrating that the impact of disease risk alleles in the activity of enhancers might be modulated by other adjacent polymorphisms.

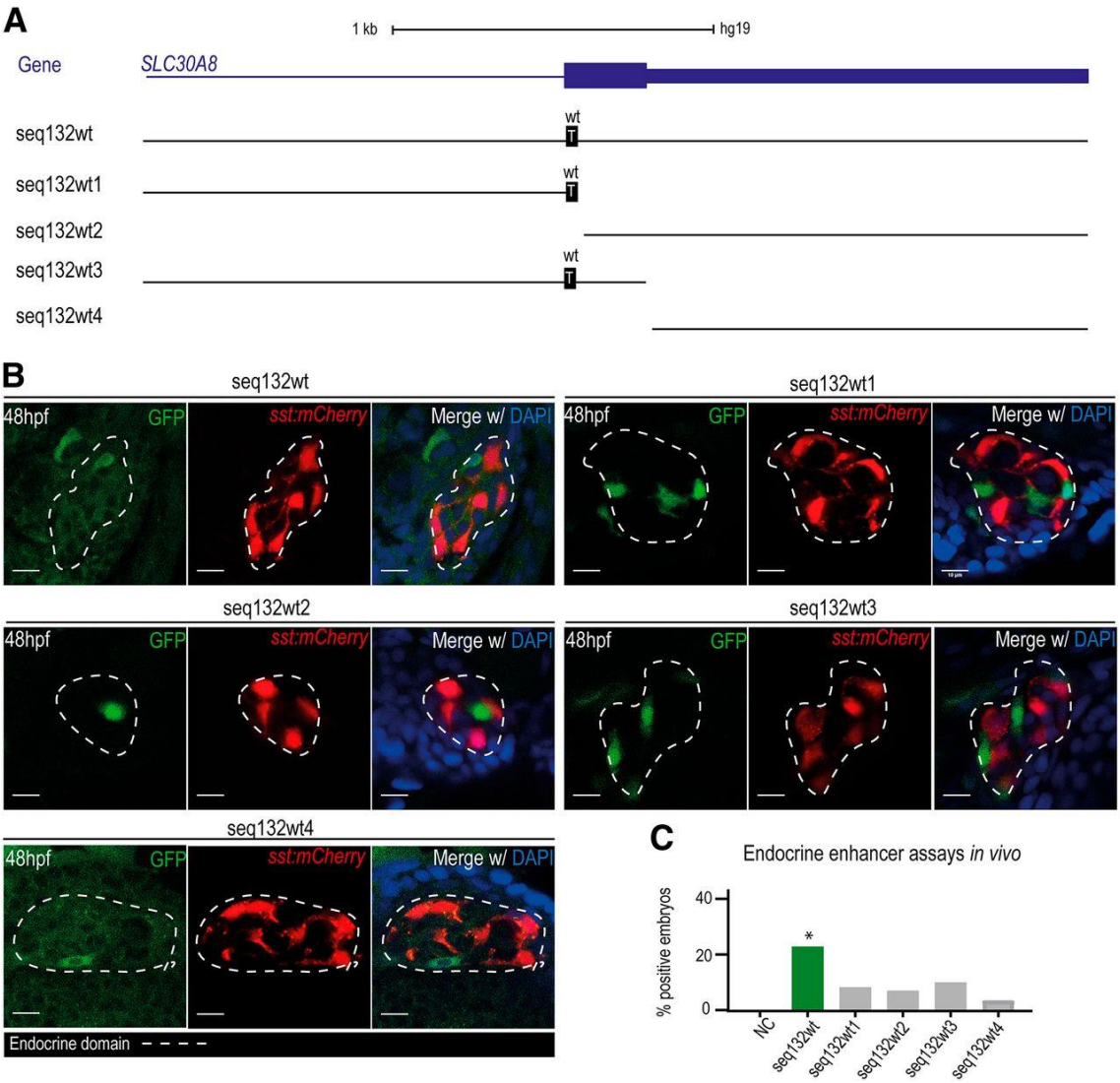


*Figure 5. A: Schematic representation of the four analyzed fragments derived from seq132wt: seq132wt1, seq132wt2, seq132wt3, and seq132wt4. The gene SLC30A08 is represented in blue. The wt allele is discriminated in black boxes. B: Representative confocal images for the total sequence, seq132wt, and the four analyzed fragments showing GFP-positive cells in the endocrine pancreatic domain (dashed line) defined by the sst:mCherry reporter line. The 48-hpf embryos were stained with DAPI. Scale bars = 10 µm. C: Graph representing the total percentages of positive embryos for seq132wt (23%, n = 34), seq132wt1 (9.8%, n = 41), seq132wt2 (6.5%, n = 31), seq132wt3 (10%, n = 30), and seq132wt4 (4%, n = 23). *P < 0.05, by χ2 test.*

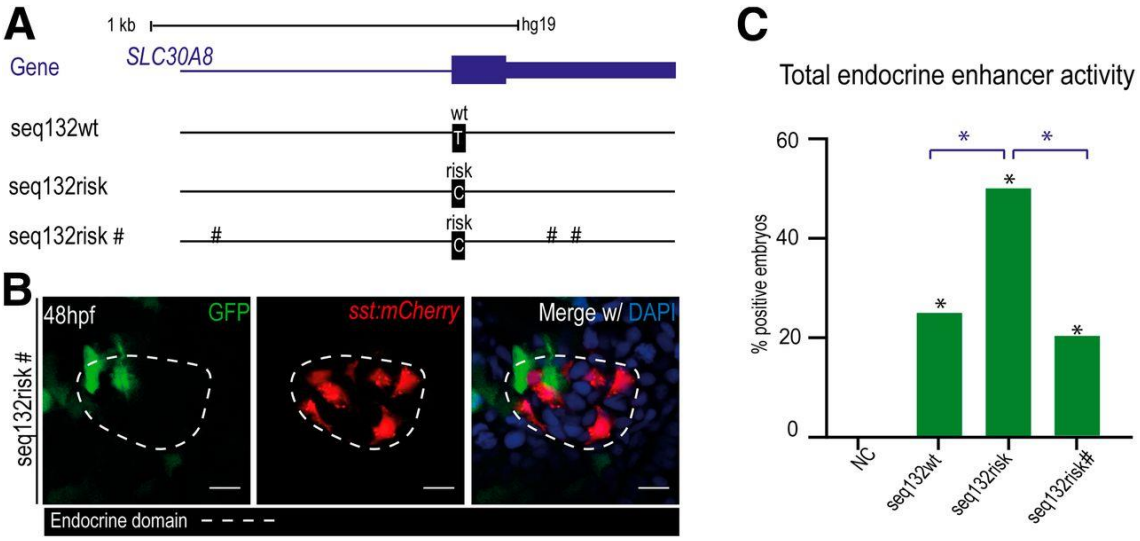


*Figure 6..A:Schematic representation of the three different analyzed versions of seq132: seq132wt, seq132risk, and seq132risk#. The gene SLC30A08 is represented in blue. The symbol # represents common variants with no association with type 2 diabetes. The wt and risk alleles are represented in black boxes. B: Representative confocal images for the seq132risk#, showing GFP expression in the endocrine pancreatic domain (dashed line) defined by the sst:mCherry reporter line. The 48-hpf embryos were stained with DAPI. Scale bars = 10 µm. C: Graph representing the total percentages of positive embryos for the seq132wt (23%, n = 34), seq132risk (56%, n = 36), and seq132risk# (20%, n = 20). *P < 0.05, by χ2 test.*

3.6. A Single nucleotide mutation impairs seq132 enhancer

Focusing on seq132, we wanted to further determine whether a single nucleotide mutation could lead to the complete ablation of the activity of this enhancer. Previous results have shown that Pdx1, an important TF required for proper pancreatic function, controls the activity of one endocrine pancreatic enhancer located in the second intron of the mouse *Slc30a8* gene[45]. On the basis of this, we hypothesized that seq132 could also be controlled by PDX1 binding. After performing TFBS analysis (Supplementary Table 3), we found that within seq132, there is a high score–predicted binding site for PDX1 (JASPAR score 0.9654) (Fig. 7A). To test whether this binding site is required for enhancer activity, we performed transgenesis assays using seq132wt containing a mutation in the predicted binding site of PDX1 (seq132wtPDX1) (Fig. 7B and C). This is an adenine-to-guanine substitution in the PDX1 consensus binding site, resulting in a predicted ablation of the binding of PDX1 (PDX1 predicted binding score: seq132wt 0.9654; seq132wtPDX1 0). Comparing endocrine pancreatic enhancer activity between seq132wt and seq132wtPDX1, we found that the seq132wtPDX1 sequence is unable to drive GFP expression in the endocrine pancreas (0%, n = 20) (Fig. 7D). Because we previously observed that seq132risk had increased endocrine pancreatic enhancer activity, we explored the possibility of the risk SNP rescuing the loss of function observed for the seq132wtPDX1 sequence. We observed that

seq132riskPDX1 also showed no endocrine pancreatic enhancer activity (0%, n = 23); thus, the risk SNP is not sufficient to rescue the loss of the PDX1 binding site (Fig. 7D).

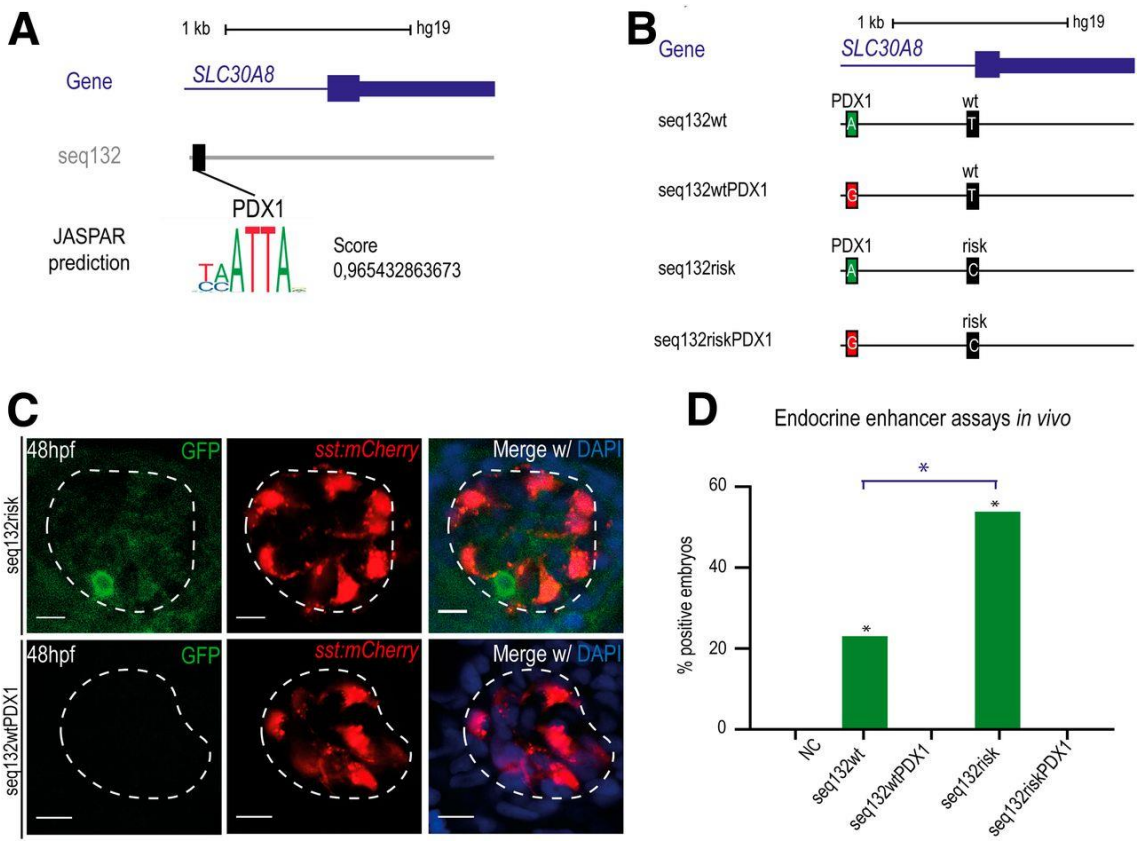


*Figure 7. A: PDX1 binding site prediction by JASPAR software. The human SLC30A8 is represented in blue. The vertical black line represents the single nucleotide mutation that overlaps with a PDX1 putative binding site. B: Schematic representation of the analyzed sequences: seq132wt, seq132wtPDX1, seq132risk, and seq132riskPDX1. The wt and risk alleles are discriminated in black boxes and the mutation in red boxes. The green boxes represent the binding site for PDX1. C: Representative confocal images for the analyzed sequence containing the risk variant in the absence (top) of the mutation in the putative binding site for PDX1, showing GFP expression in endocrine pancreatic domain (dashed line) defined by the sst:mCherry reporter line. The sequence containing the wt allele and the mutation (bottom) was not able to drive GFP expression. The 48-hpf embryos were stained with DAPI. Scale bars = 10 μm. D: Graph representing the total percentages of positive embryos for the different sequences represented in panel B: seq132wt (23%, n = 34), seq132wtPDX1 (0%, n = 21), seq132risk (56%, n = 36), and seq132riskPDX1 (0%, n = 20). *P < 0.05, by χ2 test.*

4. Discussion

In this work, we demonstrate the feasibility to perform enhancer assays using mosaic transgenesis in zebrafish. The zebrafish pancreas, as its mammal counterpart, is composed mainly by α-, β-, and δ-cells that secrete the hormones glucagon, insulin, and somatostatin, respectively. The malfunction of these cell types can contribute to type 2 diabetes development. Therefore, enhancer assays evaluating type 2 diabetes–associated alleles should consider all these cell types. In the current work, we used an *sst*:mCherry reporter construct to determine the zebrafish endocrine pancreatic domain in vivo, making available the inherent cellular complexity of a fully functional pancreas to define the activity of enhancers. In contrast, most of in vitro assays are limited to only one endocrine

pancreatic cell type, in many cases not fully functional[46]. In vivo assays have, however, their own limitations. This is particularly important when studying SNPs that might modulate the activity of enhancers and, therefore, transcriptional levels rather than binary activation or inactivation states. In this work, we demonstrate that a mosaic transgenesis method is sensitive enough and has low levels of noise, making it possible to assay quantitatively the activity of enhancers. In the current study, variation of enhancer activity comprises at least three parameters predicted to have an impact on the number of embryos with GFP expression in the endocrine pancreas: 1) transcriptional output, affecting the amount of GFP expression per cell and its detection potential defined by the GFP detection threshold; 2) expression domain, defined by the potential of GFP expression in different endocrine pancreatic cell types; and 3) robustness, defined by the stability of GFP expression. Although the present assay does not discriminate these three sources of differential enhancer activity, the introduction of further improvements can offer this possibility. Expression domain can be discriminated if specific endocrine pancreatic cell markers are used, as shown for wt alleles (Supplementary Fig. 7), and differences in transcriptional output can be observed if an internal control is used and GFP expression per cell is quantified.

Using mosaic transgenesis in zebrafish, we validated 6 out of 10 tested sequences as human endocrine pancreatic enhancers, two of which are also pancreatic progenitor enhancers, underlining the accuracy of enhancers prediction. Although species-specific regulatory outputs cannot be completely excluded, genetic networks and TFs that operate in enhancers are usually highly conserved between distantly related vertebrates as zebrafish and human, making interspecies enhancers assays reliable.

Exploring the impact of the risk-associated SNPs in enhancer activity, we found that for seq58 and seq219, the risk allele dramatically decreased the enhancer activity while for seq68, seq119, and seq132 the opposite result was observed, suggesting that disease-associated SNPs have the potential to be translated into a loss or gain of function of target genes[47]. For one case, seq119, we further demonstrated the increased binding affinity of Nkx6.1 in the risk allele, suggesting that the enhancer activity outputs might be explained by the differential binding of TFs to wt and risk alleles, consistent with previous works[48,49].

Seq132 overlaps with exon 8 of *SLC30A8*. Strikingly, the risk allele rs13266634[4], which results in an amino acid substitution impairing the function of SLC30A8[22], showed a significantly higher level of enhancer activity than the wt sequence. These results demonstrate that coding SNPs have the potential to modulate the activity of overlapping enhancers, a mechanism poorly explored but compatible with described overlapping

exon/enhancer functions[50]. *SLC30A8* encodes a zinc transporter, an ion essential for insulin maturation and secretion in β-cells[22,23]. Therefore, the rs13266634-associated increased risk for type 2 diabetes is commonly attributed to the decrease of the zinc transporter activity. In opposition to this hypothesis, the loss of function of SLC30A8 was shown to enhance insulin secretion[26], and 12 identified truncating SNPs in the *SLC30A8* gene have been associated with a protective effect against type 2 diabetes[2,25]. This incongruence could be explained if the rs13266634 association with type 2 diabetes results, not from the reduced SLC30A8 zinc transporter activity, but from an increase of its transcription caused by a more active enhancer.

Further exploring the seq132 enhancer, we observed that the increase of its activity when containing the risk allele can be reverted by the presence of three other common SNPs, highlighting how combinatorial variations in single nucleotides can alter enhancer activity.

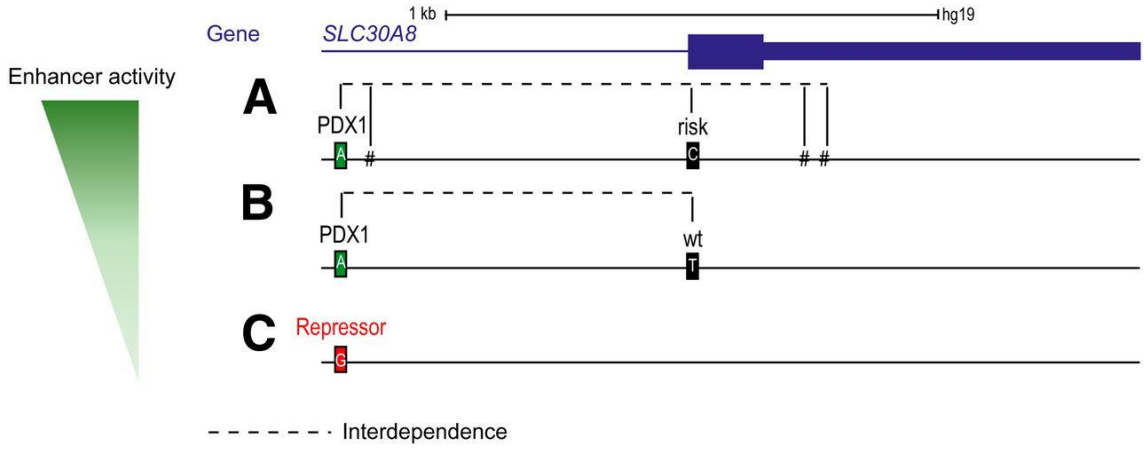


Figure 8. Theoretical model of the functional domains of seq132. Schematic representation from the three possible sequence versions and the respective associated enhancer activity level: high (A), low (B), and no enhancer activity (C). The gene SLC30A08 is represented in blue. The symbol # represents common variants with no association with type 2 diabetes. The wt and risk alleles are represented in black boxes, the putative repressor in red, and the binding site for PDX1 in green. The dashed line shows interdependence.

Furthermore, a mutation that disrupts the predicted binding site of PDX1 in seq132 results in the complete ablation of the enhancer. This suggests that the loss of the binding of PDX1 might coincide with the gain of a transcriptional repressor, since this mutation ablates the activity of nonoverlapping seq132 subfragments that have mild autonomous enhancer activity (Fig. 8). Overall, in this work, we open new avenues on the understanding of the code embedded in noncoding regulatory sequences and show the complexity of effects that one single nucleotide variation might have in the activity of enhancers and its possible impact on human disease.

5. Article Information

Acknowledgments. The authors thank Lorenzo Pasquali (Josep Carreras Leukaemia Research Institute, Barcelona, Spain) for helpful suggestions and critical reading of the manuscript. The authors acknowledge the contribution of Joana Teixeira (i3S–Instituto de Investigação e Inovação em Saúde, Universidade do Porto, and IBMC–Instituto de Biologia Celular e Molecular, Porto, Portugal) for the ChIP-seq data plotting, Silvia Naranjo (CABD - Centro Andaluz de Biología del Desarrollo, Universidad Pablo de Olavide, Seville, Spain) for the sst:mCherry vector, Ana Maia (i3S–Instituto de Investigação e Inovação em Saúde, Universidade do Porto, and IBMC–Instituto de Biologia Celular e Molecular, Porto, Portugal) for the ins:GFP construct, the support of i3S Scientific Platform Advanced Light Microscopy, members of the national infrastructure Portuguese Platform of BioImaging (supported by POCI010145FEDER022122), and the assistance of the Genomics i3S Scientific Platform (supported by POCI-01-0145-FEDER-022184).

Funding. This study was supported by the H2020 European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. ERC-2015-StG-680156-ZPR), the Fundação para a Ciência e a Tecnologia (FCT) (IF/00654/2013), and the European Regional Development Fund (Norte-01-0145-FEDER-000029). A.E., M.D., and F.J.F. are PhD fellows from FCT (grants SFRH/BD/147762/2019 to A.E., SFRH/BD/135957/2018 to M.D., and PD/BD/105745/2014 to F.J.F.). J.B. acknowledges FCT for a scientific stimulus grant (CEECIND/03482/2018).

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

6. References

1. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. *Lancet* **389**, 2239–2251 (2017).

2. Bonnycastle, L. L. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).

3. Carlson, J. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).

4. Balkau, B. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).

5. Ong, C.-T. & Corces, V. G. Enhancers: emerging roles in cell fate specification. *EMBO Rep.* **13**, 423–430 (2012).

6. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics* **46**, 136–143 (2014).

7. Stitzel, M. L. *et al.* Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* **12**, 443–455 (2010).

8. R, B. *et al.* Genome-wide analysis of histone modifications in human pancreatic islets. *Genome Res* **20**, 428–433 (2010).

9. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17921–17926 (2013).

10. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).

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

12. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).

13. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88-92 (2007).

14. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).

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

16. Youson, J. H. *et al.* The fish endocrine pancreas: review, new data, and future research directions in ontogeny and phylogeny. *Gen. Comp. Endocrinol.* **148**, 105–115 (2006).

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

18. Kimmel, R. A. *et al.* Diabetic pdx1-mutant zebrafish show conserved responses to nutrient overload and anti-glycemic treatment. *Sci Rep* **5**, 14241 (2015).

19. Nelson, S. B., Schaffer, A. E. & Sander, M. The transcription factors Nkx6.1 and Nkx6.2 possess equivalent activities in promoting beta-cell fate specification in Pdx1+ pancreatic progenitor cells. *Development* **134**, 2491–2500 (2007).

20. Binot, A.-C. *et al.* Nkx6.1 and nkx6.2 regulate alpha- and beta-cell formation in zebrafish by acting on pancreatic endocrine progenitor cells. *Dev. Biol.* **340**, 397–407 (2010).

21. Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).

22. Nicolson, T. J. *et al.* 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**, 2070–2083 (2009).

23. Kim, I. *et al.* A low-risk ZnT-8 allele (W325) for post-transplantation diabetes mellitus is protective against cyclosporin A-induced impairment of insulin secretion. *Pharmacogenomics J.* **11**, 191–198 (2011).

24. Rutter, G. A. & Chimienti, F. SLC30A8 mutations in type 2 diabetes. *Diabetologia* **58**, 31–36 (2015).

25. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).

26. Dwivedi, O. P. *et al.* Loss of ZnT8 function protects against diabetes by enhanced insulin secretion. *Nat. Genet.* **51**, 1596–1606 (2019).

27. Karlsson, J., von Hofsten, J. & Olsson, P. E. Generating transparent zebrafish: a refined method to improve detection of gene expression during embryonic development. *Mar. Biotechnol.* **3**, 522–527 (2001).

28. Mohlke, K. L. & Boehnke, M. Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum. Mol. Genet.* **24**, R85-92 (2015).

29. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981–990 (2012).

30. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).

31. M, T. *et al.* Integration of Human Pancreatic Islet Genomic Data Refines Regulatory Mechanisms at Type 2 Diabetes Susceptibility Loci. *eLife* vol. 7 https://pubmed.ncbi.nlm.nih.gov/29412141/ (2018).

32. Mularoni, L., Ramos-Rodríguez, M. & Pasquali, L. The Pancreatic Islet Regulome Browser. *Front Genet* **8**, (2017).

33. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage - PubMed. https://pubmed.ncbi.nlm.nih.gov/11027340/.

34. Curado, S. *et al.* Conditional targeted cell ablation in zebrafish: A new tool for regeneration studies. *Developmental Dynamics* **236**, 1025–1035 (2007).

35. PME-MCS - Tol2kit. http://tol2kit.genetics.utah.edu/index.php/PME-MCS.

36. Cebola, I. *et al.* TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. *Nat. Cell Biol.* **17**, 615–626 (2015).

37. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).

38. van de Werken, H. J. G. *et al.* 4C technology: protocols and data analysis. *Meth. Enzymol.* **513**, 89–112 (2012).

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

40. Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334**, 222–225 (2011).

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

42. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes | Nature Genetics. https://www.nature.com/articles/s41588-019-0457-0.

43. Chung, J. H., Whiteley, M. & Felsenfeld, G. A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell* **74**, 505–514 (1993).

44. The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs - PubMed. https://pubmed.ncbi.nlm.nih.gov/17937395/.

45. Pound, L. D. *et al.* The pancreatic islet β-cell-enriched transcription factor Pdx-1 regulates Slc30a8 gene transcription through an intronic enhancer. *Biochemical Journal* **433**, 95–105 (2011).

46. Skelin, M., Rupnik, M. & Cencic, A. Pancreatic beta cell lines and their applications in diabetes mellitus research. *ALTEX* **27**, 105–113 (2010).

47. Vezzoli, G. *et al.* R990G polymorphism of calcium-sensing receptor does produce a gain-of-function and predispose to primary hypercalciuria. *Kidney International* **71**, 1155–1162 (2007).

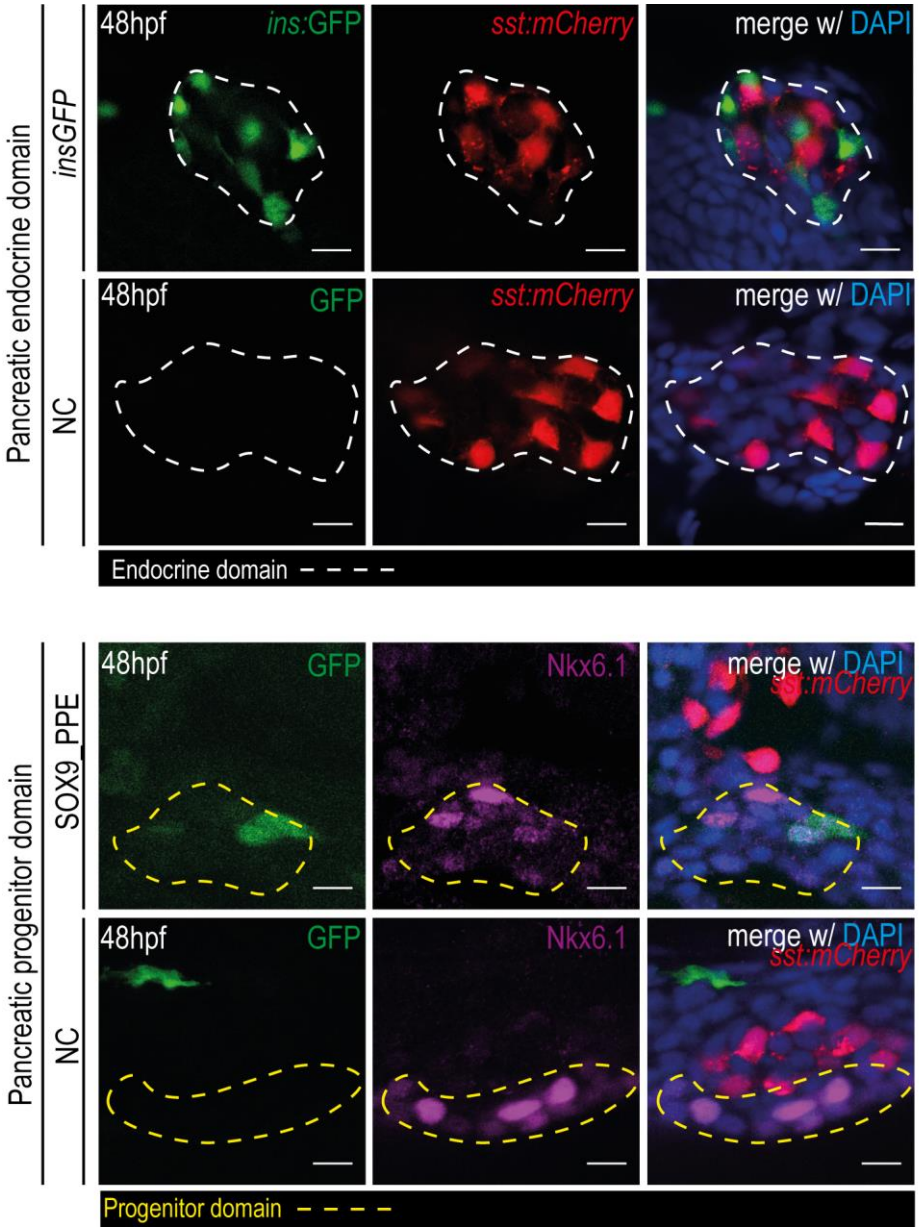
48. Tjian, R. & Maniatis, T. Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**, 5–8 (1994).

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

50. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* **22**, 1059–1068 (2012).

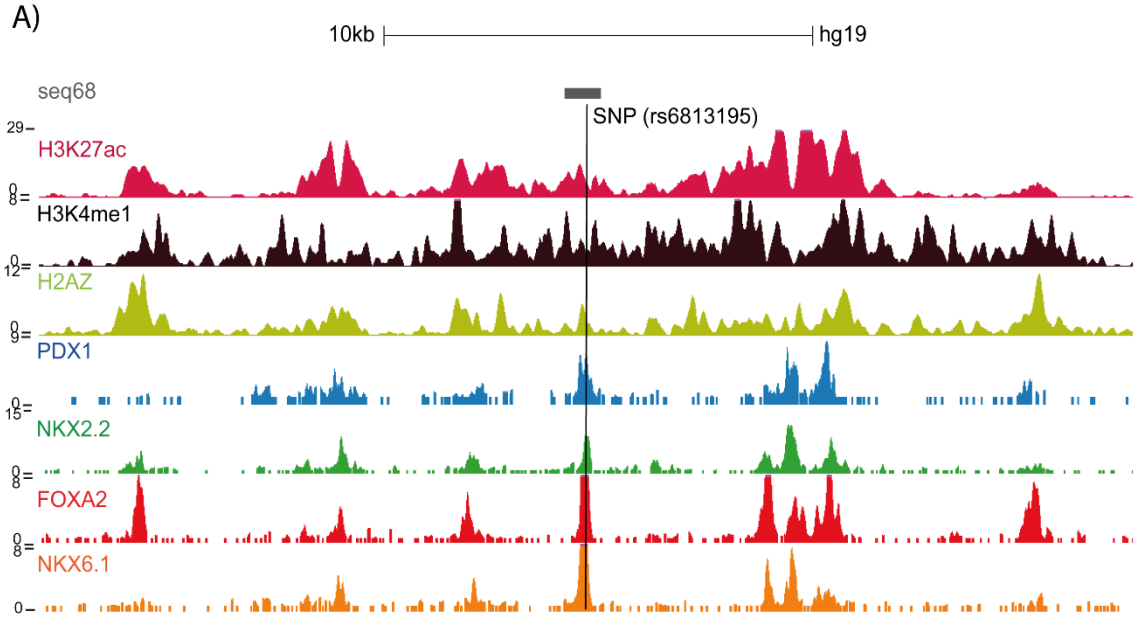
7. Supplemental material

Online Supplemental Material

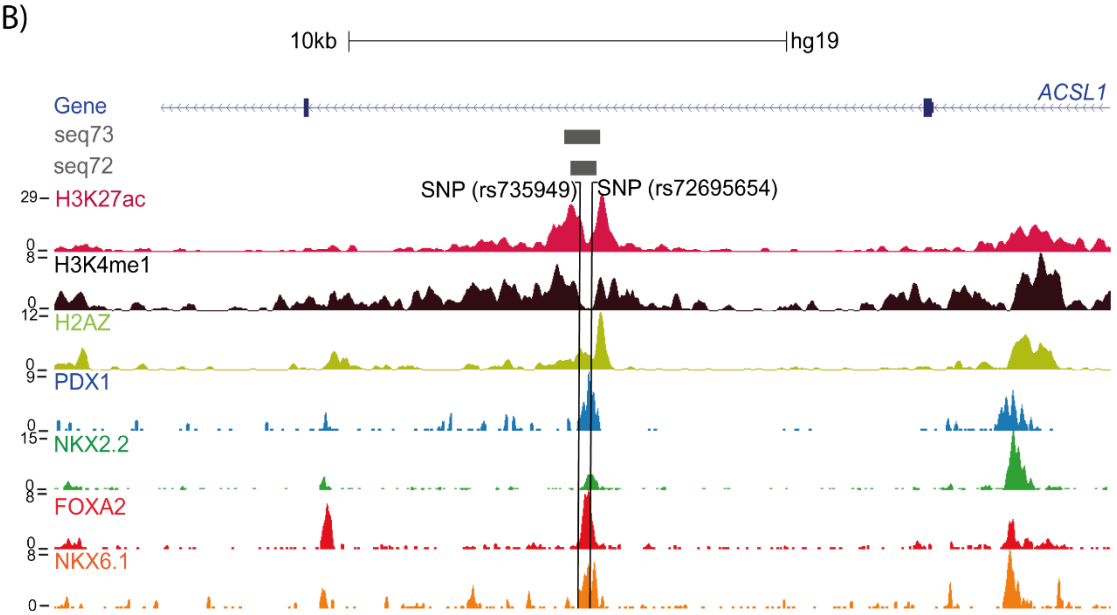


Supplementary Fig.1 - *Top: Representative confocal images resulted from the injection of ins:GFP vector, showing GFP expression in endocrine pancreatic domain (white dashed line) defined by the sst:mCherry reporter line. The NC did not show GFP positive cells in endocrine pancreatic domain. The 48hpf embryos were stained with DAPI (blue). Scale bar, 10 µm. Bottom: Confocal images from Z48 vector containing a pancreatic progenitor enhancer (SOX9_PPE), showing GFP positive cells co-localization with the progenitor domain (yellow dashed line) defined by the Nkx6.1 staining (purple) The NC was not able to drive GFP expression in the pancreatic progenitor domain. The 48 hpf embryos were stained with DAPI (blue) and Nkx6.1 (purple). Scale bar, 10 µm.*

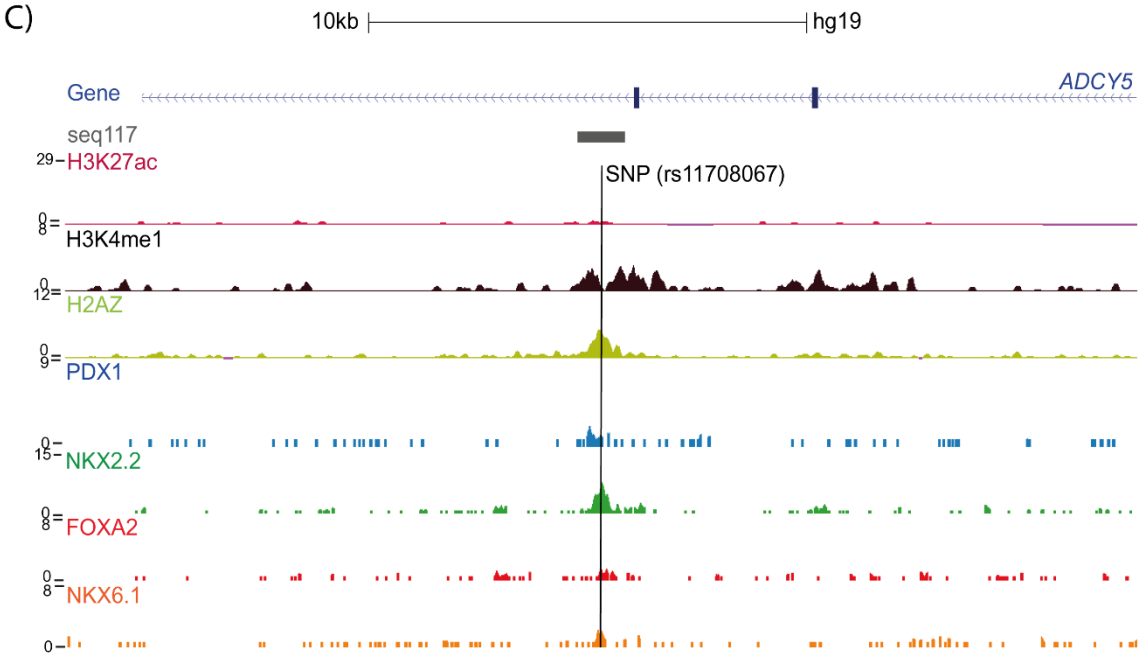
A)


10kb
hg19
seq68
SNP (rs6813195)
H3K27ac
H3K4me1
H2AZ
PDX1
NKX2.2
FOXA2
NKX6.1

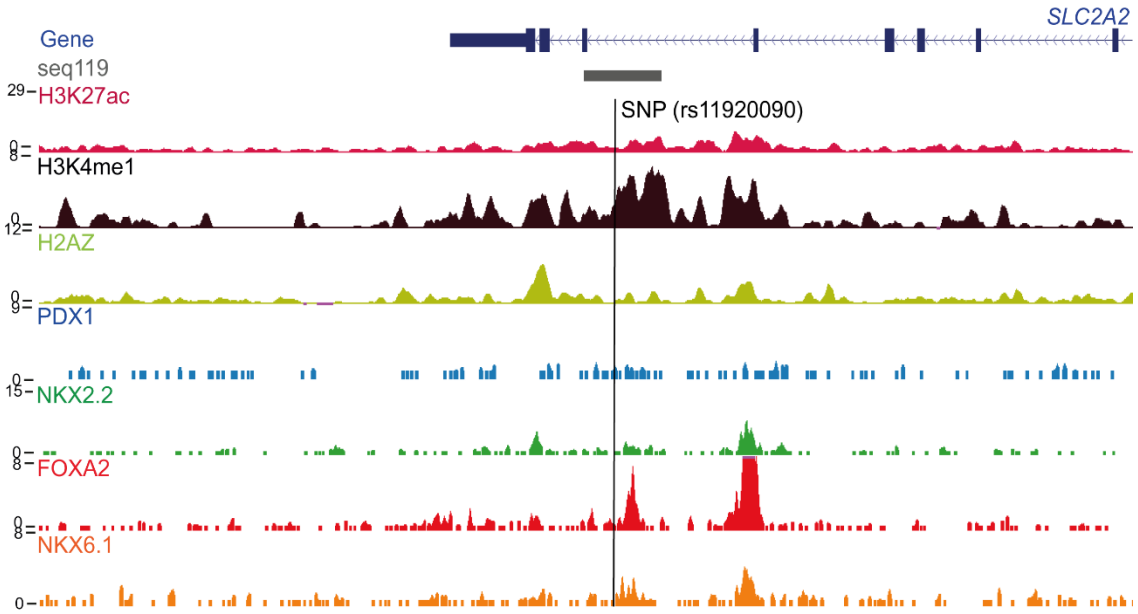
B)

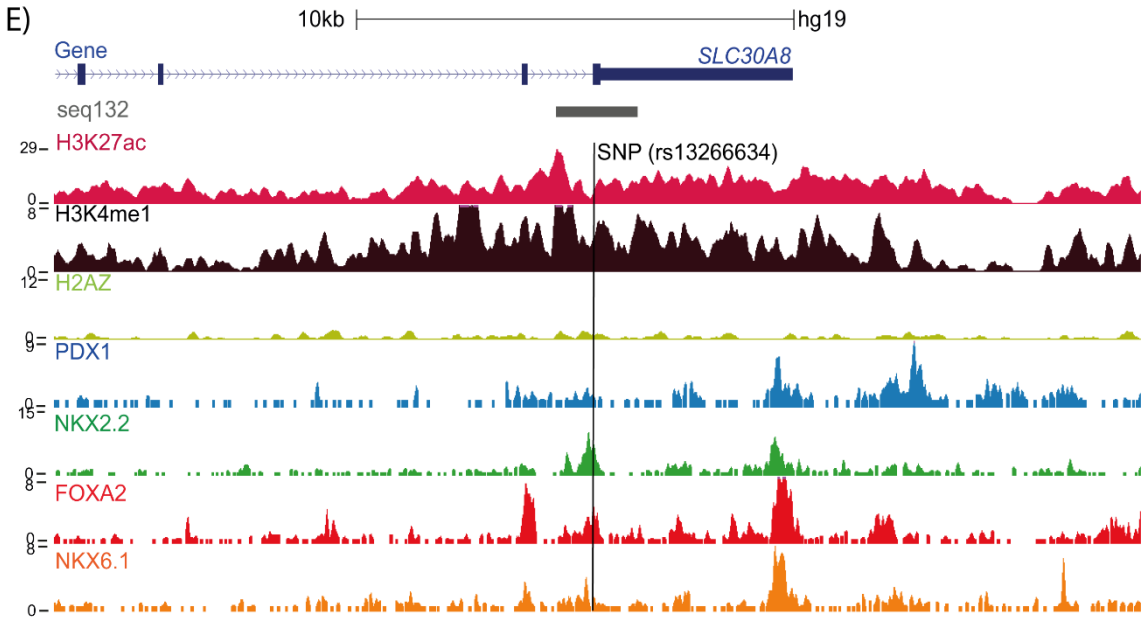

10kb
hg19
Gene
ACSL1
seq73
seq72
SNP (rs735949)
SNP (rs72695654)
H3K27ac
H3K4me1
H2AZ
PDX1
NKX2.2
FOXA2
NKX6.1

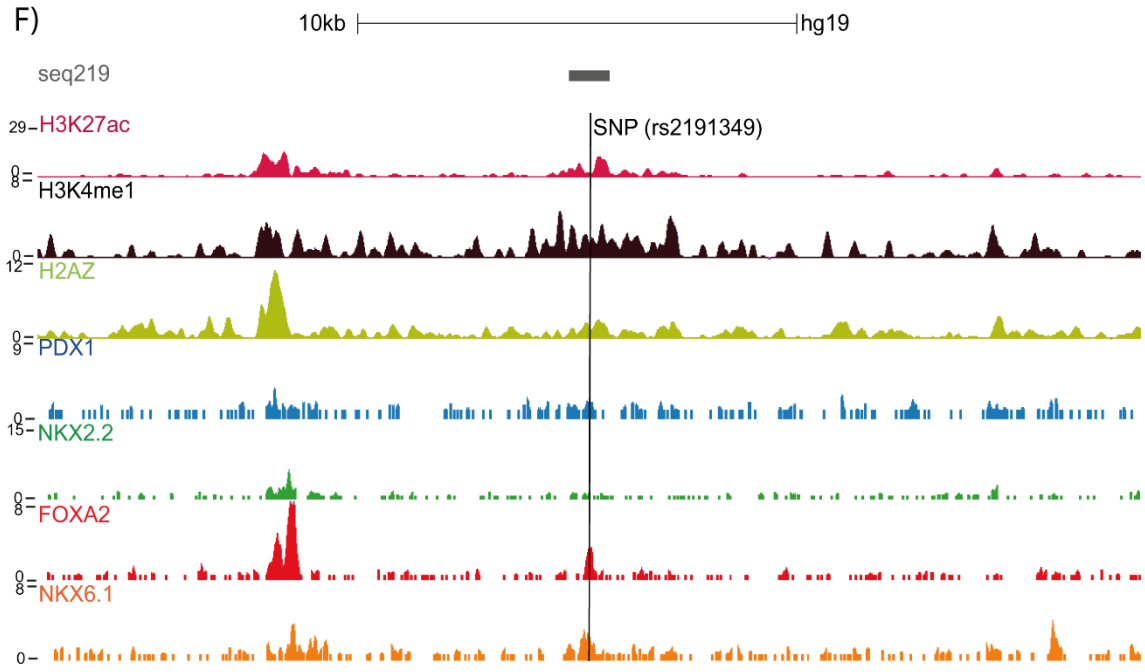
C)

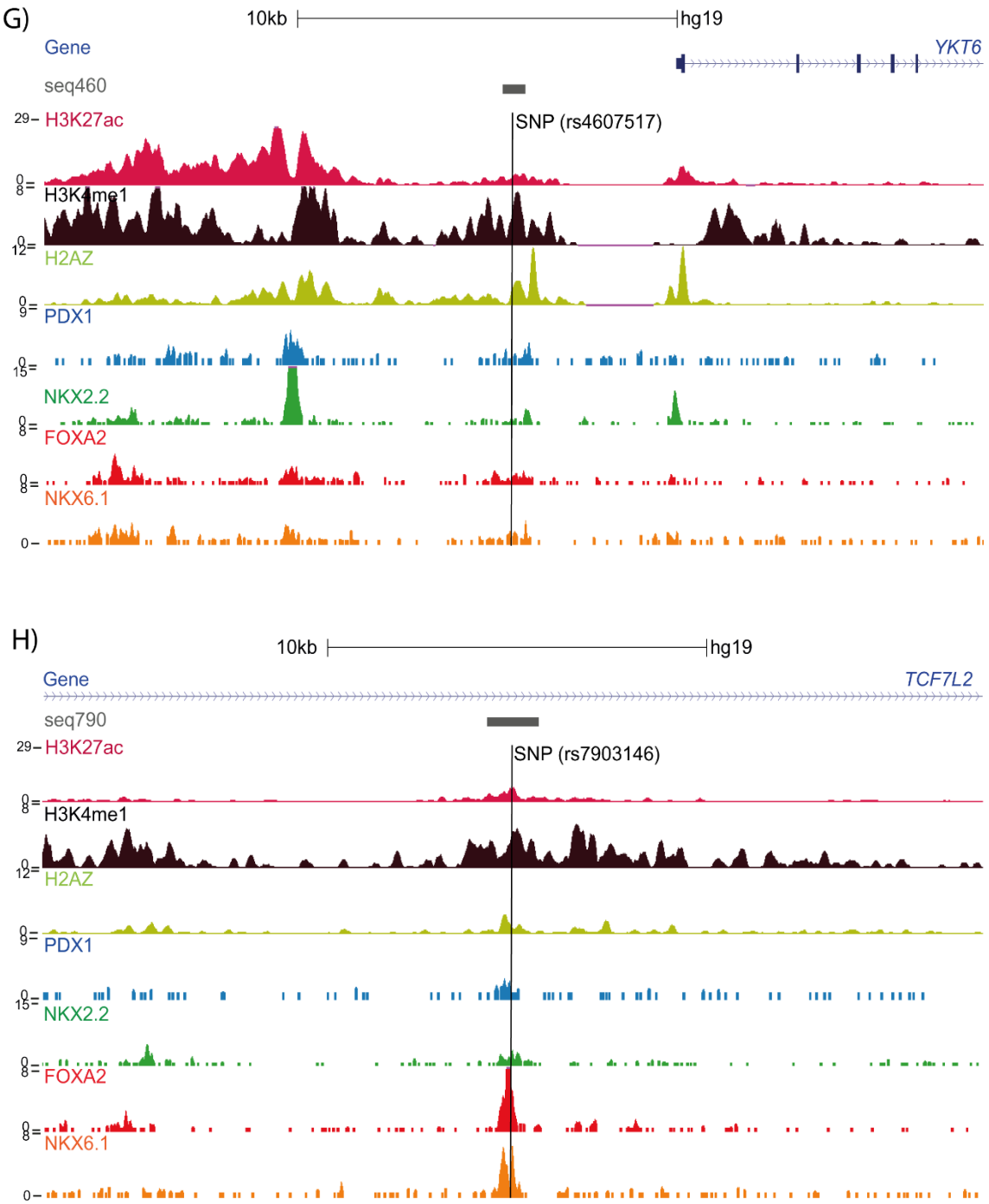

10kb
hg19
Gene
ADCY5
seq117
29
H3K27ac
SNP (rs11708067)
0
8
H3K4me1
0
12
H2AZ
0
9
PDX1
0
15
NKX2.2
0
8
FOXA2
0
8
NKX6.1
0

D)

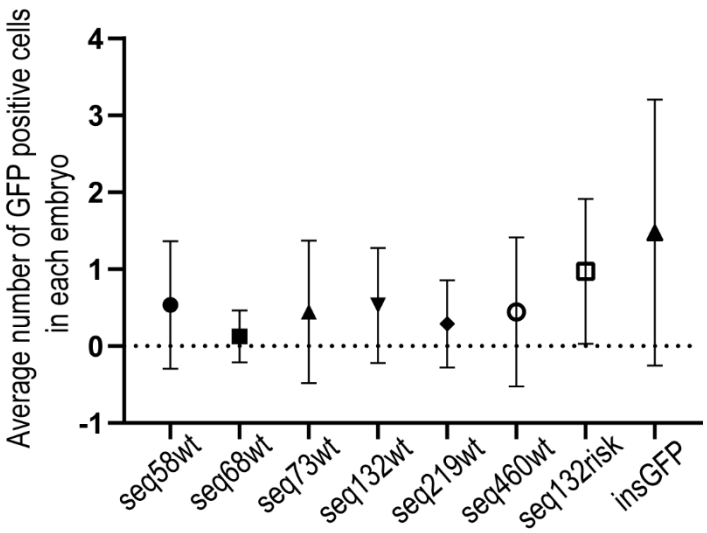

SLC2A2
Gene
seq119
29
H3K27ac
SNP (rs11920090)
0
8
H3K4me1
0
12
H2AZ
0
9
PDX1
0
15
NKX2.2
0
8
FOXA2
0
8
NKX6.1
0


E)
10kb
hg19
Gene
SLC30A8
seq132
SNP (rs13266634)
H3K27ac
H3K4me1
H2AZ
PDX1
NKX2.2
FOXA2
NKX6.1

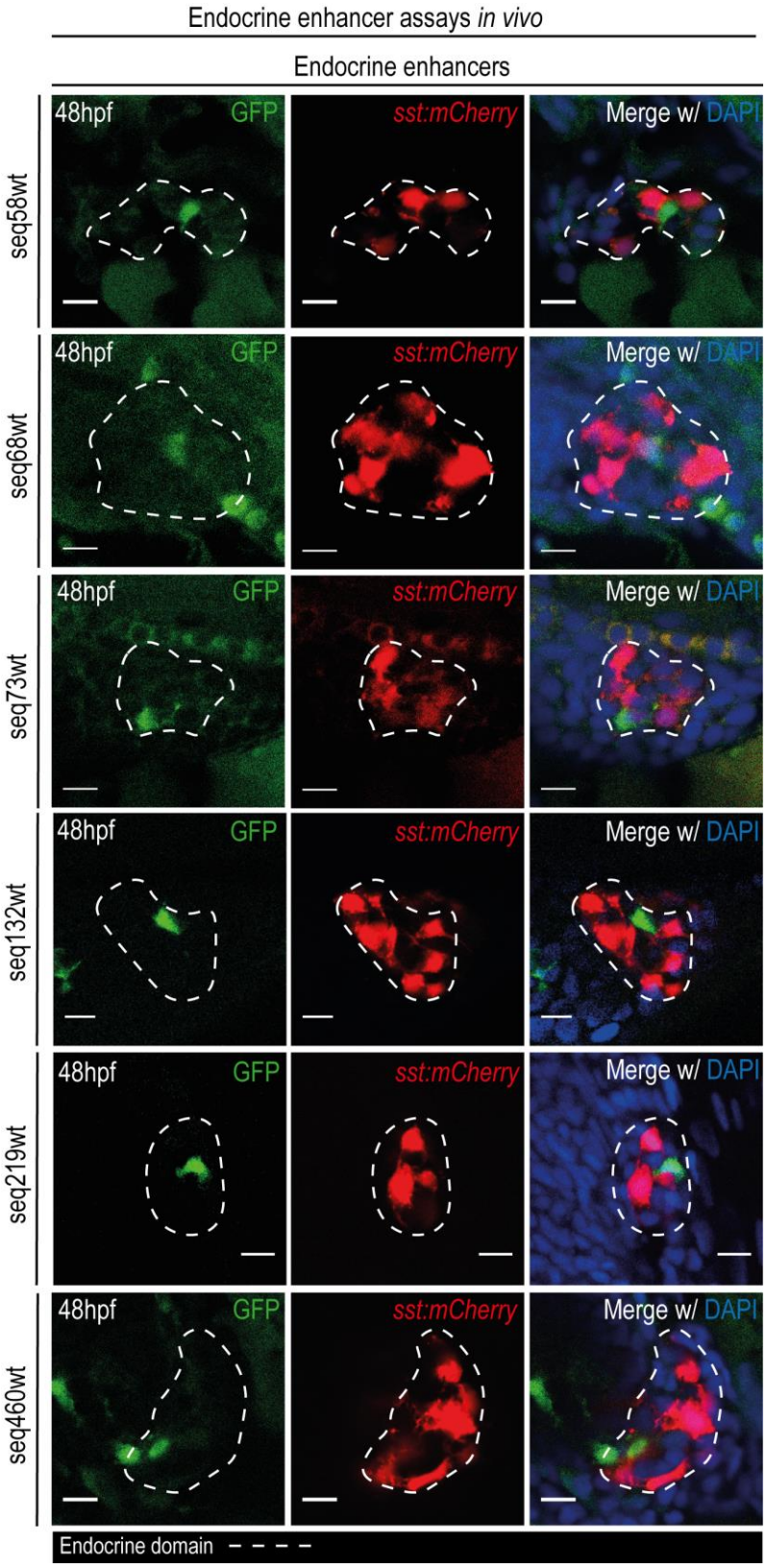

F)
10kb
hg19
seq219
SNP (rs2191349)
H3K27ac
H3K4me1
H2AZ
PDX1
NKX2.2
FOXA2
NKX6.1



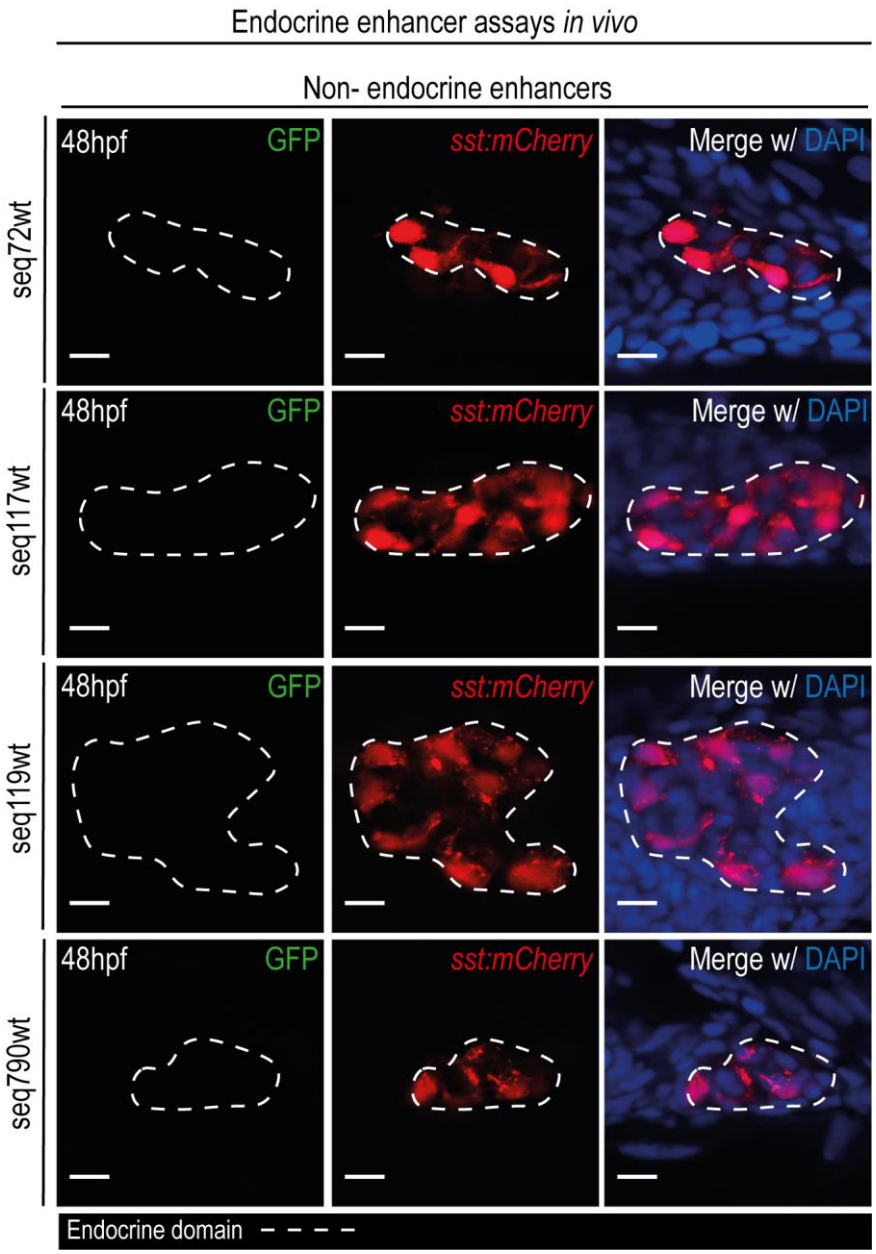
Supplementary Fig.2 *Genomic landscapes of the putative enhancers seq68 (A), seq72 (B), seq73 (B), seq117 (C), seq119 (D), seq132 (E) seq219 (F), seq460 (G) and seq790 (H) (represented in grey). Tracks represent H3K27ac (pink), H3K4me1 (black), histone variant H2A.Z (light green) and TF binding (PDX1 - blue, NKX2.2 - dark green, FOXA2 - red and NKX6.1 - orange) from ChIP-seq data for human endocrine pancreas samples. The location of the type 2 diabetes associated SNPs (rs6813195, rs735949, rs72695654, rs11708067, rs11920090, rs13266634, rs2191349, rs4607517 and rs7903146) is represented as a vertical black line.*



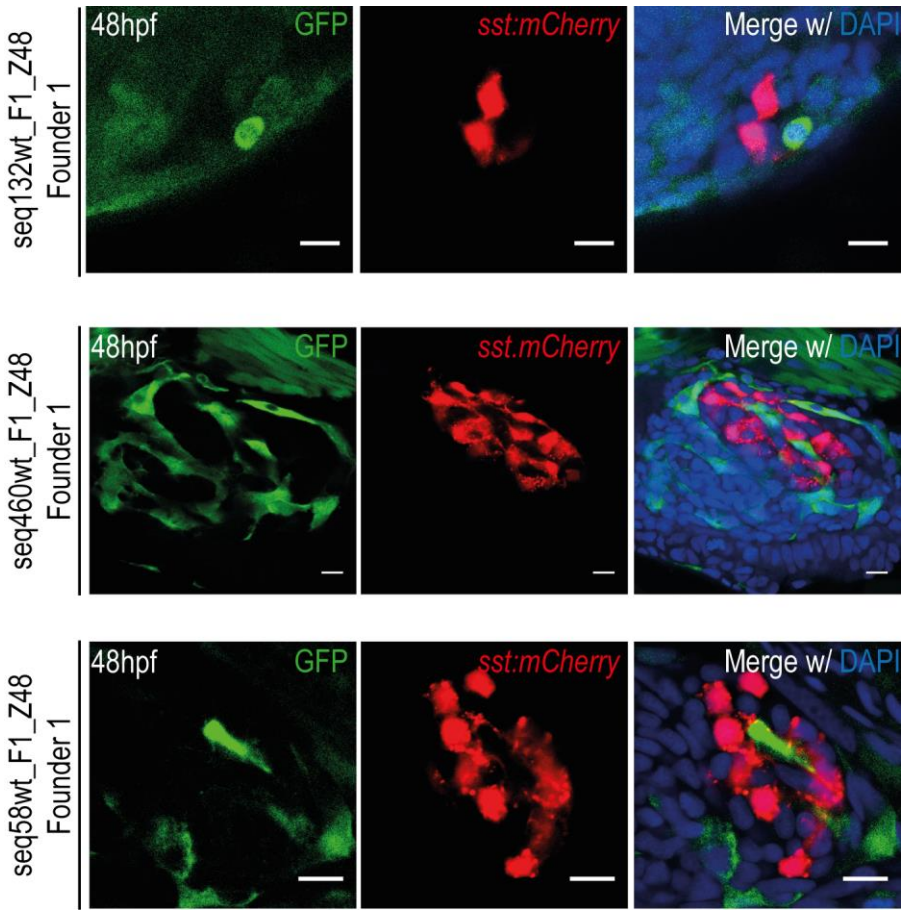
Supplementary Fig. 3 *– Graph representing the average number of GFP positive cells per embryo per sequence and in insGFP promoter vector. Data are mean ±SD.*



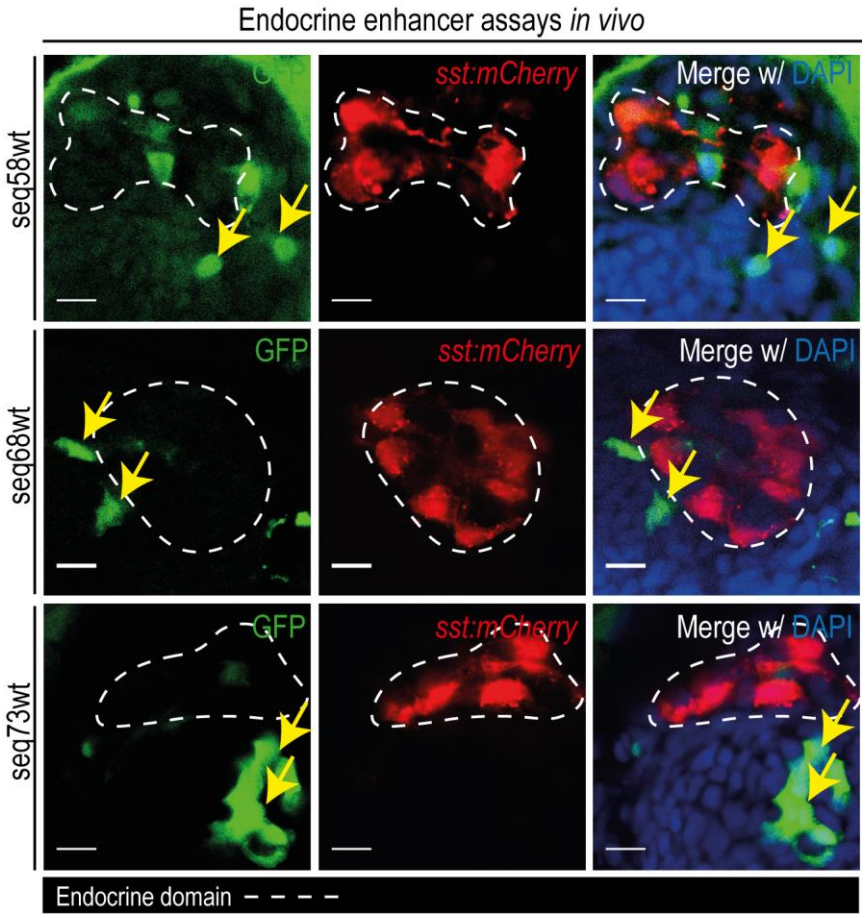
Supplementary Fig.4 - *Representative confocal images for seq58wt, seq68wt, seq73wt, seq132wt, seq219wt and seq460wt. All six sequences showed GFP positive cells in endocrine pancreatic domain (white dashed line) defined by the sst:mCherry reporter line. The 48 hpf embryos were stained with DAPI (blue). Scale bar, 10 µm.*



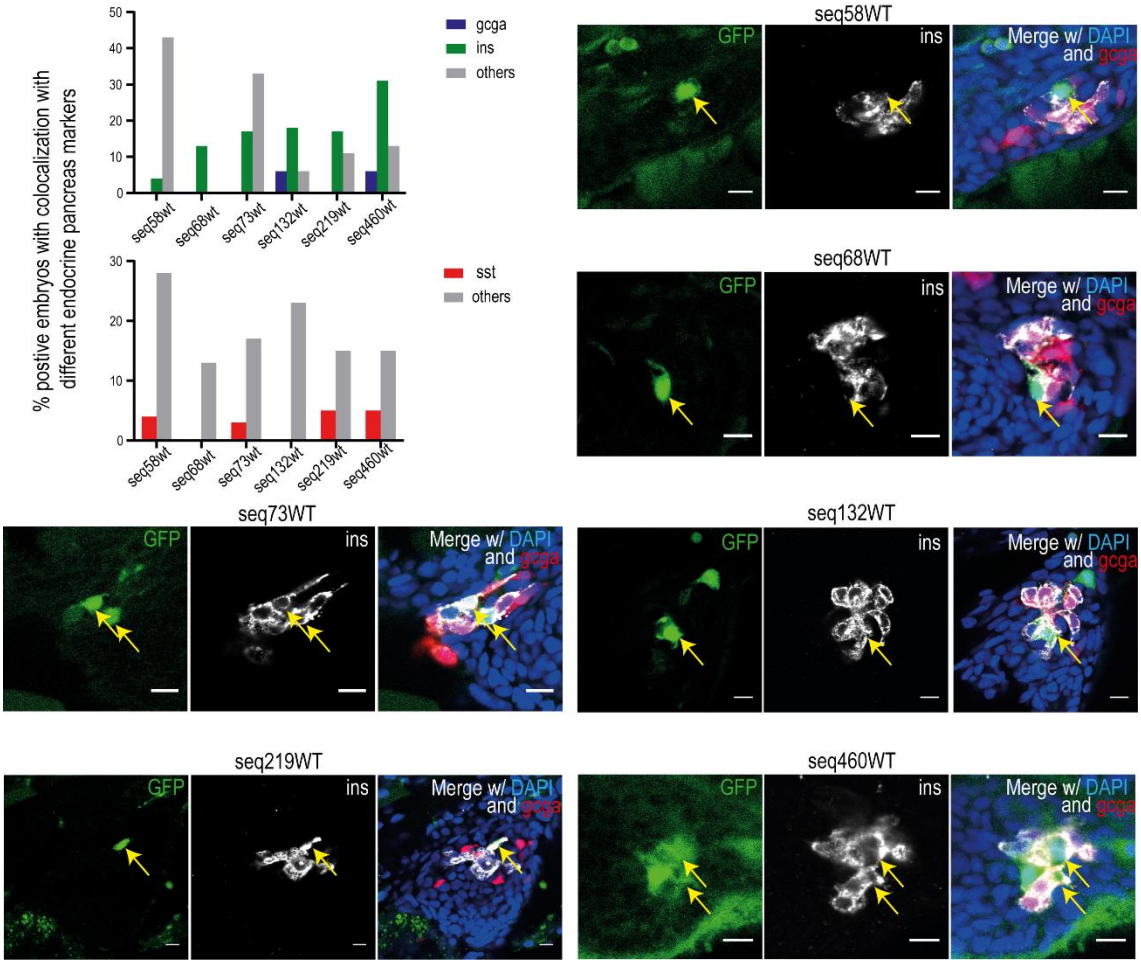
Supplementary Fig.5 - *Representative confocal images of seq72wt, seq117wt, seq119wt and seq790wt. All four sequences did not show GFP positive cells in endocrine pancreatic domain (white dashed line) defined by the sst:mCherry reporter line. The 48 hpf embryos were stained with DAPI (blue). Scale bar, 10 µm.*



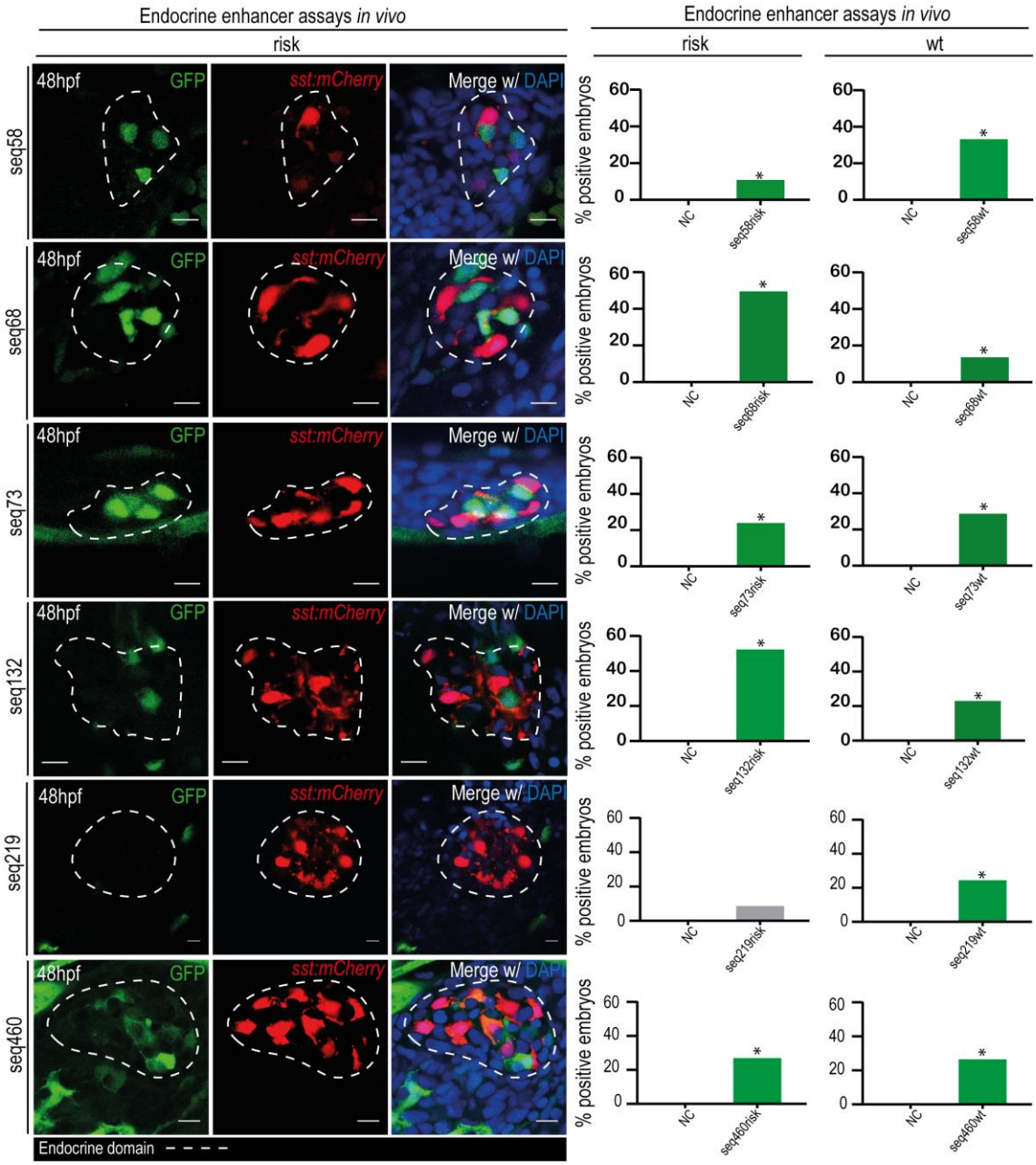
Supplementary Fig.6 - *Representative confocal images of stable transgenic lines from seq132wt, seq460wt and seq58wt. All 3 sequences presented GFP positive cells within endocrine pancreatic domain, defined by sst:mCherry reporter line. The 48hpf embryos were stained with DAPI. Scale bar: 10 µm.*



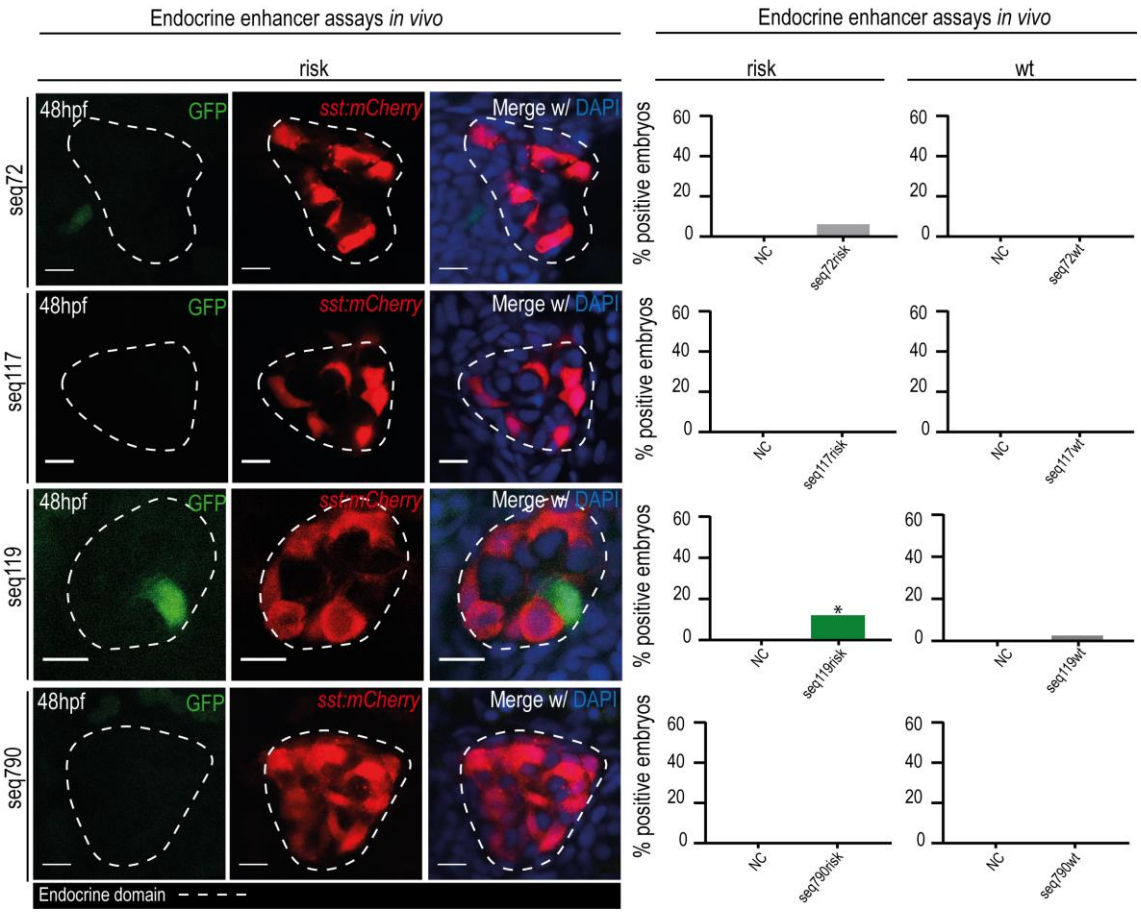
Supplementary Fig.7 - *Representative confocal images of seq58wt, seq68wt and seq73wt. These three sequences presented GFP positive cells in adjacent area to the endocrine domain defined by the sst:mCherry reporter line. The 48 hpf embryos were stained with DAPI (blue). Scale bar, 10 µm.*



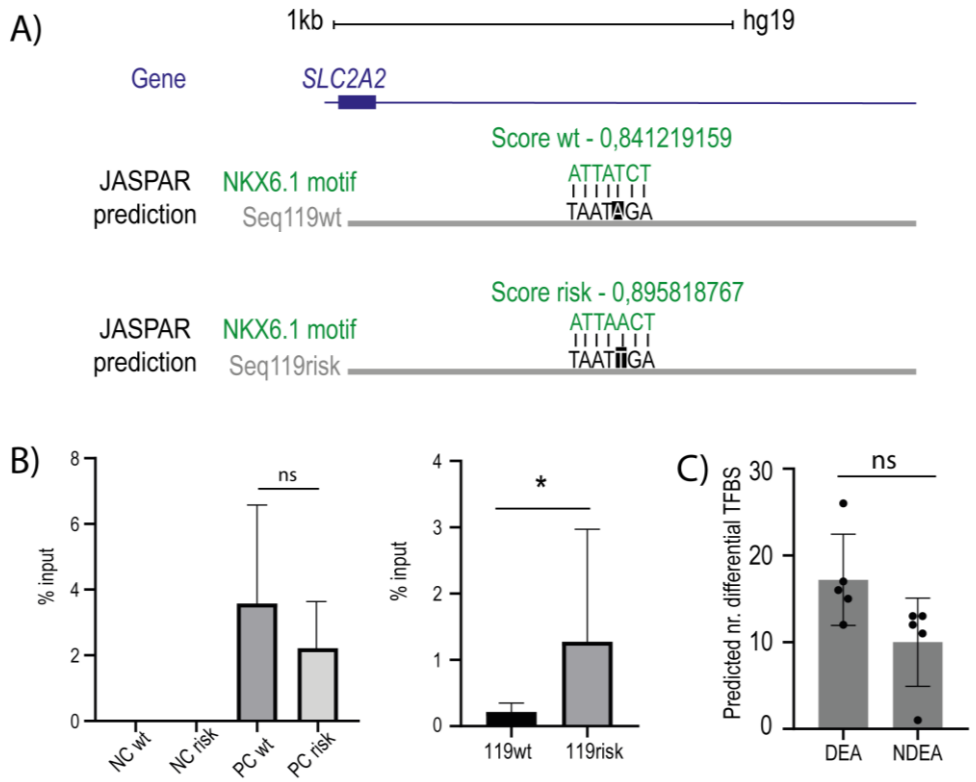
Supplementary Fig.8 - *Graphs representing the percentage of the number of embryos with GFP positive cells colocalizing with somatostatin (sst), glucagon (gcga) and insulin (ins) secreting cells. Representative confocal images resulted from the injection of seq58wt (n=24), seq68wt (n=15), seq73wt (n=12), seq132wt (n=16), seq219wt (n=18) and seq460wt (n=16). The endocrine pancreatic domain (white dashed line) is defined by the gluc:mCherry reporter line. The yellow arrows indicate GFP colocalizing with insulin. The 48 hpf embryos were stained with DAPI (blue) and insulin (white). Scale bar, 10 µm.*



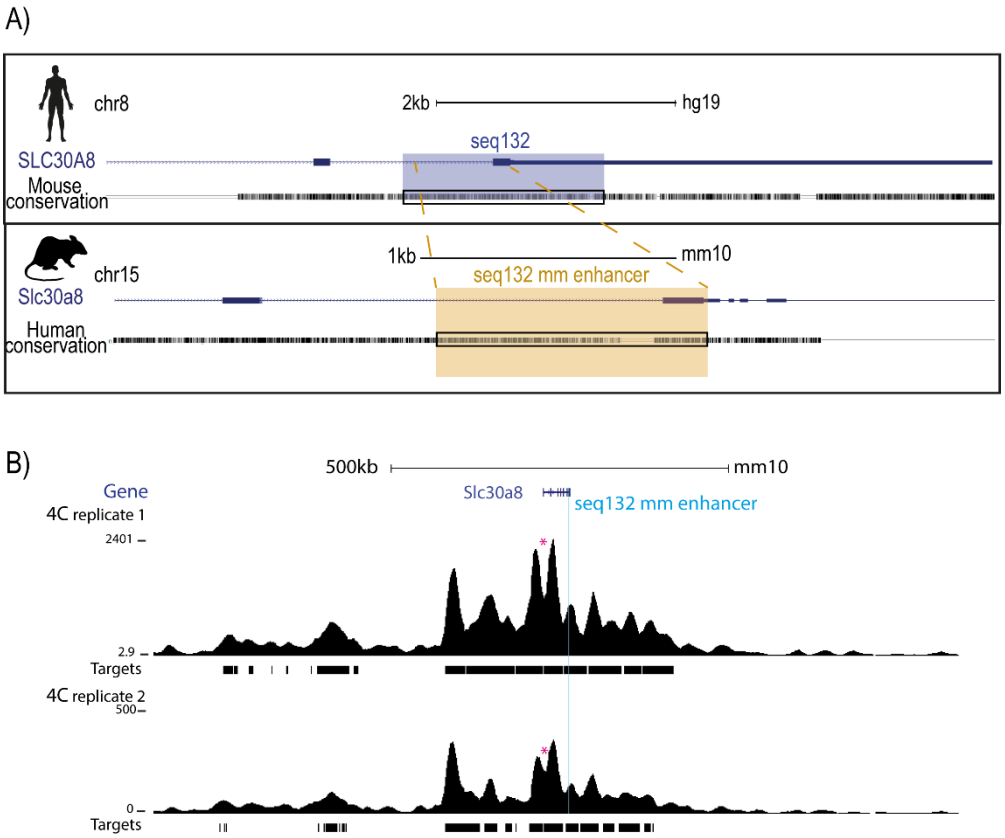
Supplementary Fig.9 - *Left: Representative confocal images for all the risk alleles of the identified endocrine pancreatic enhancers: seq58risk, seq68risk, seq73risk, seq132risk, seq219risk and seq460risk. The endocrine pancreatic domain is represented by the white dashed line and it is defined by the sst:mCherry reporter line. The 48hpf embryos were stained with DAPI (blue). Scale bar, 10 µm. Right: Graphs showing the total percentage of positive embryos for wt and risk alleles of the aforementioned sequences seq58 wt/risk (36%, n=56 / 12%, n=43), seq68 wt/risk (13%, n=47 / 50%, n=32), seq73 wt/risk (28%, n=47 / 22%, n=36), seq132 wt/risk (23%, n=34 / 56%, n=36), seq219wt//risk (24%, n=38 / 6%, n=35) seq460 wt/risk (27%, n=36 / 27%, n=30) Chi-square test: p<0,05 (*).*



Supplementary Fig.10 - *Left: Representative confocal images from all the risk alleles of the wt sequences that previously tested negative for enhancer activity (non-endocrine pancreatic enhancers): seq72risk, seq117risk, seq119risk and seq790risk. The endocrine pancreatic domain is represented by the white dashed line and it is defined by the sst:mCherry reporter line. The 48hpf embryos were stained with DAPI (blue). Scale bar, 10 µm. Right: Graphs showing the total percentage of positive embryos in wt and risk alleles of the identified endocrine pancreatic enhancers: seq72 wt/risk (0%, n=27 / 6%, n=30), seq117 wt/risk (0%, n=21 / 0%, n=8), seq119 wt/risk (4%, n=27 / 14%, n=28) and seq790 wt/risk (0%, n=20 / 0%, n=14). Chi-square test: p<0,05 (*).*



Supplementary Fig.11 –(A) *Jaspar relative score prediction of differential Nkx6.1 binding in wt and risk variants (Score wt: 0,841219159; Score risk: 0,895818767).* ***(B) Left:*** *Graph representing the negative (cryaa promoter) and positive (hnf1**α** promoter) control regions used in ChIP, showing no enrichment and a higher enrichment in both alleles, respectively.* ***Right****: ChIP enrichment for Nkx6.1 in MIN6 stable lines with human seq119wt and seq119risk. Data are mean ±SD. Mann Whitney test: p<0,05 (*).* ***(C)*** *Graph representing the number of differential predicted TFBS within the group of enhancers with (DEA) and without differential enhancer activity (NDEA). The two groups were compared by unpaired t-test. p<0,05.*



Supplementary Fig.12 –(A) *Sequence conservation of seq132 between human (top) and mouse (bottom). The seq132 and seq132mm are highlighted.* ***(B)*** *Genomic landscape of mouse gene slc30a8 (blue) showing the 2 replicates from 4C performed in MIN6 cell line (black). The viewpoint, Slc30a8 promoter, is represented in pink (asterisk).*

Sequence	Coordinates (GRCh37/hg19)	SNP	wt/risk allele	P-value	Disease related	r2 from the lead SNP	Credible set SNPs (<80%PPA)(34)	Putative enhancer state (5)	Chromatin state (5)	TFBS (5)
seq58	chr6:37,775,193-37,776,330	rs5869265	C /A	2,00E-10 (Lead SNP)	Type 2 diabetes	1000 Genomes Project CEU r2 = 0.96 (LD with rs9470794)	no	Weak	H3K27ac, H3K4me1, H2AZ	PDX1, NKX2.2, FOXA2, NKX6.1
seq68	chr4:153,519,496-153,521,069	rs6813195	T / C	4,00E-14	Type 2 diabetes	Lead SNP	no	Weak	H3K27ac, H3K4me1, H2AZ	PDX1, NKX2.2, FOXA2, NKX6.1, MAFB
seq72	chr4:185,716,034-185,716,803	rs72695654	G /T	3,70E-06 (Lead SNP)	Fasting glucose and type 2 diabetes	1000 Genomes Project CEU r2 = 0.94 (LD with rs735949)	no	Strong	H3K27ac, H3K4me1, H2AZ	PDX1, NKX2.2, FOXA2, NKX6.1
seq73	chr4:185,715,954-185,716,970	rs735949	C/ T	3,70E-06	Fasting glucose and type 2 diabetes	Lead SNP	no	Strong	H3K27ac, H3K4me1, H2AZ	PDX1, NKX2.2, FOXA2, NKX6.1
seq117	chr3:123,065,139-123,066,383	rs11708067	G / A	5,00E-32	Type 2 diabetes	Lead SNP	yes	Weak	H3K27ac, H3K4me1, H2AZ	NKX2.2, MAFB
seq119	chr3:170,716,833-170,718,552	rs11920090	A/ T	8,00E-13	Homeostasis model assessment of beta-cell function	Lead SNP	no	Strong	H3K27ac, H3K4me1	FOXA2
seq132	chr8:118,183,990-118,185,784	rs13266634	T / C	1,00E-47	Type 2 diabetes	Lead SNP	yes	Weak	H3K27ac, H3K4me1	NKX2.2, FOXA2
seq219	chr7:15,063,701-15,064,872	rs2191349	G / T	3,00E-44	Fasting blood glucose and type 2 diabetes	Lead SNP	no	Other	H3K27ac, H3K4me1, H2AZ	NKX2.2, MAFB, PDX1, NKX6.1
seq460	chr7:44,235,330-44,236,228	rs4607517	G / A	7,00E-92	Type 2 diabetes	Lead SNP	no	Strong	H3K27ac, H3K4me1, H2AZ	NKX2.2
seq790	chr10:114,757,655-114,759,128	rs7903146	C / T	6,00E-447	Peak insulin response	Lead SNP	yes	Strong	H3K27ac, H3K4me1	FOXA2,

PPA: Posterior probability of association

Supplementary Table 1 *–List of the selected putative enhancer sequences. The respective coordinates, overlapping type 2 diabetes associated SNPs, wt and risk alleles, p-values, disease related, r2 from the lead SNP, chromatin states, and the overlapping TFBSs with strong ChIP-seq peak, are discriminated for each sequence. The SNPs that are part of the credible set are discriminated (31). The rs5869265 and rs72695654 are in high linkage disequilibrium (100 Genomes Project CEU r^2>0,8) with lead SNPs reported in GWAS catalog, being considered the risk alleles A and T respectively.*

Oligoname	Sequence	Application	Specie	References
seq58_FW	CTCTGAGAAGGAAATTGAACGC	Reporter enhancer assay	Human	
seq58_RV	AAAACCTCACATTAAAGCCATCCC	Reporter enhancer assay	Human	
seq68FW	CCTGGAGATTGTCTTCTAAGCTGC	Reporter enhancer assay	Human	
seq68RV	GCAACTCAGATTGCATCTAGAGCC	Reporter enhancer assay	Human	
seq73_FW	TCACCTGTGCCTGGCTGGG	Reporter enhancer assay	Human	
seq73_RV	GTGGGGTGGCCTGCAGGG	Reporter enhancer assay	Human	
seq132_FW	GCATTTACTGCCTCAAGAGAAAGC	Reporter enhancer assay	Human	
seq132_RV	GTGGCAACAACTTGGTGGGG	Reporter enhancer assay	Human	
seq219_FW	CTACTGACATCAGCCAATGAGTCTAATACC	Reporter enhancer assay	Human	
seq219_RV	GTCCTCCAGGGCCTCTATATTCATGG	Reporter enhancer assay	Human	
seq460_FW	GCCTATCTTCAAATCTCTACTTCCC	Reporter enhancer assay	Human	
seq460_RV	GATCAGGAAGACAGCGCTTGG	Reporter enhancer assay	Human	
seq72_FW	TTCGCAAAACATCTCATCACC	Reporter enhancer assay	Human	
seq72_RV	CAGGGTGAGAACTGAAGGC	Reporter enhancer assay	Human	
seq117_FW	CTTCCCGGATGTGGAGATTCAGCC	Reporter enhancer assay	Human	
seq117_RV	GGAGGAGAAAGGAGGAAGCAACACC	Reporter enhancer assay	Human	
seq119_FW	ATGGCACAAACAAACATCCCACTCATTCC	Reporter enhancer assay	Human	
seq119_RV	ACTAATGGGCTGGTAGAAGAGGGCC	Reporter enhancer assay	Human	
seq790_FW	AGGTGTGGGGGTATATGGTATCC	Reporter enhancer assay	Human	
seq790_RV	CACCAGGTCATGGAAACTTAGCC	Reporter enhancer assay	Human	
seq132wt1_FW	GCATTTACTGCCTCAAGAGAAAGC	Reporter enhancer assay	Human	
seq132wt1_RV	GCAGTCGGGGTCCTGGTCAACTGG	Reporter enhancer assay	Human	
seq132wt2_FW	CCAGTTGACCAGGACCCCGACTGC	Reporter enhancer assay	Human	
seq132wt2_RV	GTGGCAACAACTTGGTGGGG	Reporter enhancer assay	Human	
seq132wt3_FW	GCATTTACTGCCTCAAGAGAAAGC	Reporter enhancer assay	Human	
seq132wt3_RV	GAGCTAGTCACAGGGGTCTTCACAG	Reporter enhancer assay	Human	
seq132wt4_FW	ACATGCTGCTATGCAGTTTCTGC	Reporter enhancer assay	Human	
seq132wt4_RV	GTGGCAACAACTTGGTGGGG	Reporter enhancer assay	Human	
Seq119 ChIP Nkx6BS F	TGCACTATTGACACCGGAAA	ChIP on plasmid	human	
Seq119 ChIP Nkx6BS R	CCAGAAAAATGCAGGCCTAT	ChIP on plasmid	human	
mm Nkx6.1 +ChIP F	CAGAGAAACCCTGTCTTGGAA	ChIP on plasmid, positive control	mouse	doi: 10.1074/jbc.M109.064238
mm Nkx6.1 +ChIP R	GGAAGATAAGGGCCAAGCTC	ChIP on plasmid, positive control	mouse	doi: 10.1074/jbc.M109.064238
mm cryaa pm F	GCCGTGGTAGCTGAAGAGAC	ChIP on plasmid, negative control	mouse	
mm cryaa pm R	GCCAGCTAAGATGCACATGA	ChIP on plasmid, negative control	mouse	
mm Slc30a8 F	GAGGCATCTATCACTGCCACA	RT-qPCR	mouse	
mm Slc30a8 R	TACCACCCAAATGTCAGCCG	RT-qPCR	mouse	
mm actb F	tgacatccgtaaagacctcta	RT-qPCR, housekeeping	mouse	DOI: 10.1182/blood-2010-08-302976
mm actb R	caggaggagcaatgatcttga	RT-qPCR, housekeeping	mouse	DOI: 10.1182/blood-2010-08-302976
mm sg1	AGCAAGAGGGAGATCCCTGGCACC	CRISPRa and CRISPRi	mouse	
mm sg2	ATTCTGTGCAATCAGCTCAGCACC	CRISPRa and CRISPRi	mouse	
seq132 mm FW	TGACTCAAGATGTTACTTAGAGGC	Reporter enhancer assay	mouse	
seq132mm RV	CTGACAGTGTGACCGAGC	Reporter enhancer assay	mouse	
4C_primer1	CCATCTCATCCCTGCGTGTCTCCGACTCAGAGCACGAATCGATCCTCTTTACTATATTTGATG	4C	mouse	
4C_primer2	CCTCTCTATGGGCAGTCGGTGATCAACATGCTGTTAGTCTTAGAAG	4C	mouse	
sg_NegCTRL_mm_66895_rv	CACCGCGAGGTATTCGGCTCCGCG	CRISPRa and CRISPRi	mouse	DOI: https://doi.org/10.1038/s41588-019-0457-0
sg_NegCTRL_mm_66895_fw	AAACCGCGGAGCCGAATACCTCGC	CRISPRa and CRISPRi	mouse	DOI: https://doi.org/10.1038/s41588-019-0457-0
sst2_FW	AGCCTCTATGTCCTTCGTC	sst2 promoter amplification	zebrafish	DOI: https://doi.org/10.1016/j.diff.2008.09.014
sst2_RV	CTGCTGCTTCTTTAACTCAG	sst2 promoter amplification	zebrafish	DOI: https://doi.org/10.1016/j.diff.2008.09.014
seq58_SDM_FW	GTTGAGCCGGGAGAAAAGAT	Site directed mutagenesis	human	
seq58_SDM_RV	ATCTTTTCTCCCGGCTCAAC	Site directed mutagenesis	human	
seq68_SDM_FW	GTAAATACTGGCCTTGCCGC	Site directed mutagenesis	human	
seq68_SDM_RV	GCGGCAAGGCCAGTATTTAC	Site directed mutagenesis	human	
seq73_SDM_FW	AGAAAATAGGACTGCCCG	Site directed mutagenesis	human	
seq73_SDM_RV	GGAACAATGTGTATTTGTAGGA	Site directed mutagenesis	human	
seq219_SDM_FW	ggatcatttaagaccagtt	Site directed mutagenesis	human	
seq219_SDM_RV	aactggtcttaaatgatcc	Site directed mutagenesis	human	
seq460_SDM_FW	GGCGATGCAGCAGAAGGC	Site directed mutagenesis	human	
seq460_SDM_RV	GCTGTCACTCAACTTGGATTC	Site directed mutagenesis	human	

Supplementary Table 2 – *Total list of the primers used in this study.*

Seq58

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
NEUROD1	GAACAGATGGCAG	16,1826	9,90078	0,995050964	0,875148484	wt	1
Bhlha15	CCATCTGT	11,6381	4,64482	0,963409531	0,832620322	wt	1
YY1	AAAGATGGCAGC	6,93956	15,3079	0,825561041	0,946979791	risk	1
OLIG3	AACAGATGGC	9,28684	6,4711	0,898049606	0,851225106	wt	0
Tcf12	AACAGATGGCA	8,04759	0	0,893235401	0	wt	1
Gata4	TCTTTTCTCCC	0	8,17506	0	0,889325094	risk	1
TCF4	GCCATCTGTT	5,89183	0	0,877538534	0	wt	1
Tcf3	GCCATCTGTTC	6,89191	0	0,87692178	0	wt	0
GATA3	AGAAAAGA	0	6,5876	0	0,872653708	risk	1
GATA6	GGGAGAAAAGATG	0	8,71767	0	0,859319004	risk	1
Myc	CCATCTGTTC	5,37231	0	0,853670959		wt	1
MAX	GAACAGATGG	4,81313	0	0,842786913	0	wt	1
HNF4G	GGAGAACAGATGGCA	6,94403	0	0,838750612	0	wt	0
Ascl2	GCCATCTGTT	7,3974	0	0,828244302	0	wt	1
Rxra	CAGATGGCAGC	4,30832	0	0,826386987	0	wt	0
TFEC	AACAGATGGC	6,42627		0,82243229		wt	0
RORC	GAAAAGATGGCA	0	6,31808	0	0,82119829	risk	1
TFAP4	AACAGATGGC	5,62338	0	0,818700149	0	wt	1
NFAT5	CTTTTCTCCC	0	5,26747	0	0,817415415	risk	1
ZEB1	CATCTG	4,56217	4,38815	0,813430428	0,806522192	wt	0
HNF4A	CTGCCATCTGTTCTC	5,29646	0	0,805371839	0	wt	1
FOXO3	GGAGAACA	6,25847	0	0,80235975	0	wt	1
ASCL1	AGAACAGATGGCA	5,9011	0	0,800782542	0	wt	1

Seq68

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
NFIX	TGCGGCAAG	0	5,48411	0	0,874615693	risk	1
NFE2L2	GTGATGCAGCA	10,8334	8,37657	0,870973071	0,80990749	wt	1
Rxra	GCAAGGCCAGT	6,28014	6,28014	0,854789085	0,854789085	wt	0
CEBPA	GCTGCATCACT	2,70475	0,579997	0,830392604	0,804937845	wt	1
Mafb	GCTGCATC	5,51425	0	0,827417498	0	wt	1
Esrrg	GCAAGGCCAG	4,14226	4,14226	0,817111519	0,817111519	wt	1
MEIS1	TTGCCGC	1,59862	0	0,814599111	0	risk	1
JUNB	CAGTGATGCAG	1,88779	0	0,814286261	0	wt	1
JUND	AGTGATGCAGC	0,248792	0	0,80933893	0	wt	1
ETS1	CTTGCT	3,65816	0	0,809079322	0	wt	1
HOXA2	CCGCATCACT	3,35648	0	0,806317854	0	risk	1
CEBPB	CCTTGCTGCAT	-1,54594	0	0,803734637	0	wt	1
NR2F2	GCAAGGCCAGT	5,97343	5,97343	0,802707635	0,802707635	wt	1

Seq73

TF	Binding site	Score WT	Score risk	Relative score WT	Relative score risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0
FOXH1	ACAAATACACA	11,5433	3,03764	0,93501129	0,815358964	wt	0
FOXD2	ACAAACA	6,94322	8,35322	0,900266554	0,930169815	risk	1
FOXP3	ACAAACA	6,1809	8,5041	0,8711956	0,929646506	risk	0
NR3C1	GGAACAATGTGTATT	13,3344	10,1148	0,91454486	0,874464368	wt	0
FOXL1	TACAAATA	6,4764	0	0,898546012	0	wt	0
FOXO6	ATACACA	5,96965	0	0,875240106	0	wt	1
Foxa2	CAATGTGTATTT	9,70117	9,54966	0,872638676	0,868758253	wt	1
NR3C2	TGGAACAATGTGTATTT	10,3632	7,05619	0,87035699	0,839105275	wt	1
FOXO4	ATACACA	6,0096	0	0,867463112	0	wt	1
FOXC1	AATGTGTA	5,08571	0	0,866539251	0	wt	1
Neurog1	ACAATGTGTG	5,67924	6,96099	0,834583127	0,864345402	risk	0
Foxd3	CAATGTGTGTTT	8,6765	9,20803	0,842668731	0,853272957	risk	0
Foxq1	CAATGTGTATT	9,6916	0	0,850163839	0	wt	1
FOXG1	ATACACAT	5,92499	0	0,846162563	0	wt	0
HOXA5	CACACATT	0	5,42911	0	0,844085516	risk	1
Myc	CAATGTGTAT	4,1004	3,5073	0,835002487	0,826297332	wt	1
FOXK2	AAATACACATT	7,11292	0	0,831751967	0	wt	1
MEF2C	TTCCTACAAATACAC	7,22312	0	0,829959064	0	wt	0
NEUROD2	TACACATTGT	3,81787	0	0,827157388	0	wt	0
ZNF354C	ATACAC	4,82832	0	0,821096857	0	wt	1
Foxj2	ATACACAT	4,5094	0	0,81784419	0	wt	1
NKX2-8	GTATTTGTA	2,43163	0	0,814510374	0	wt	0
STAT1	AAATTCCTACAAACA	7,88368	7,97238	0,808911597	0,81034354	risk	1

Seq132

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
Tcf3	CCCAGCTGGCT	12,555	0	0,952250391	0	wt	0
TCF4	CCCAGCTGGC	8,45918	0	0,916803022	0	wt	1
Myog	GCCAGCTGGGA	8,92548	0	0,913604247	0	wt	0
Myod1	CCCAGCTGGCTGC	8,713	0	0,902220172	0	wt	1
FIGLA	CCCAGCTGGC	8,92208	0	0,893885428	0	wt	0
Tcf12	GCCAGCTGGGA	7,86887	0	0,890740198	0	wt	1
PAX5	ACAGCAGCCAGCCGGGA	11,2341	13,7514	0,858808679	0,888219183	risk	1
TFAP2A	GCCAGCCGG	0	6,77176	0	0,881869392	risk	1
NEUROD1	AGCCAGCTGGGAC	9,74202	0	0,872118038	0	wt	1
Pax2	TGTCCCGG	0	5,41577	0	0,85666724	risk	1
RBPJ	GCTGGGACAG	6,29924	0	0,855183015	0	wt	1
Neurog1	GCCAGCTGGG	6,37768	0	0,850800832	0	wt	0
ASCL1	CAGCCAGCTGGGA	8,27687	0	0,840690936	0	wt	1
PAX9	TGTCCCGGCTGGCTGCT	0	8,50939	0	0,831403104	risk	0
TFAP4	GCCAGCTGGG	6,30787	0	0,829728315	0	wt	1
STAT3	CAGCTGGGACA	1,32493	0,586987	0,813605668	0,804665821	wt	1
ZBTB18	CAGCCAGCTGGGA	4,742	0	0,812298861	0	wt	1
Atoh1	GCCAGCTGGG	4,69235	0	0,810347344	0	wt	0
NHLH1	GCCAGCTGGG	5,58268	0	0,809935213	0	wt	0
MYB	CCCAGCTGGC	3,78195	0	0,808536243	0	wt	1
Ascl2	GCCAGCTGGG	6,17312	0	0,803928059	0	wt	0
TFDP1	AGCCGGGACAG	0	6,17637	0	0,803288255	risk	1
SP1	CCGGCTGGCT	0	6,14518	0	0,801216595	risk	1

Seq219

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
FOXC1	atcatgta	5,4601	0	0,887481988	0	WT	1
GRHL2	gtaagaccagttaaa	12,4502	10,9803	0,879806296	0,853364284	WT	1
Pax2	gatcatgt	5,84664	5,00666	0,87358725	0,840602054	WT	1
Lhx4	tcatttaa	0	6,2302	0	0,867647007	Risk	0
HLF	tcttacatgatc	7,1943	0	0,863221714	0	WT	1
DBP	gatcatgtaaga	6,09006	0	0,860332187	0	WT	1
TEF	tcttacatgatc	6,74804	0	0,857210797	0	WT	1
JUND(var,2	aaggatcatgtaaga	4,87603	0	0,847713259	0	WT	1
VAX2	taaatgat	0	5,7904	0	0,84714633	Risk	1
HOXB3	gatcatttaa	0	4,82217	0	0,846110076	Risk	1
ISL2	tcatttaa	0	5,93316	0	0,841852634	Risk	1
HOXB2	gatcatttaa	0	5,00057	0	0,840874055	Risk	1
JUN	aggatcatgtaag	4,85199	0	0,840563509	0	WT	1
Dlx4	tcatttaa	0	4,9254	0	0,839239544	Risk	1
NFIL3	tcatgtaagac	8,83467	0	0,83897551	0	WT	1
Pax6	tttaaggatcattt	0	10,9967	0	0,837270028	Risk	1
GSC	catgatcctt	5,76966	5,16334	0,830352106	0,817054565	WT	1
NKX6-1	tcatttaa	0	4,26911	0	0,829643585	Risk	1
Arid3a	atttaa	0	5,399	0	0,826577908	Risk	1
Arid3b	gtcttaaatga	0	6,72959	0	0,822961343	Risk	1
Bhlhe40	cttacatgatc	2,91117	0	0,822486649		WT	1
LIN54	tcttaaatg	0	4,2698	0	0,822270435	Risk	1
LBX2	atcatttaag	0	5,41535	0	0,821581912	Risk	1
VENTX	atcatttaa	0	3,33065	0	0,821411748	Risk	0
ATF4	ggatcatgtaaga	6,64264	0	0,813569591	0	WT	1
HOXA2	gatcatttaa	0	3,5678	0	0,811095483	Risk	1
Prrx2	tcatttaa	0	4,52469	0	0,804040681	Risk	0
FOXD2	ttaaatg	0	2,29556	0	0,801698956	Risk	1
HLTF	gatcatttaa	0	3,40649	0	0,800223744	Risk	1

Seq460

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
Pax2	Tgtcactc	5,06043	8,09328	0,842713818	0,96180982	risk	1
ISL2	tcactcaa	0	6,74948	0	0,863090898	risk	0

Seq72

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
NR2F2	TAAAGGTCACT	0	13,5291	0	0,966059123	risk	1
Esrrg	TAAAGGTCAC	4,04628	11,8595	0,815699532	0,930637668	risk	1
Rxra	TAAAGGTCACT	2,70904	10,9064	0,803350977	0,921424841	risk	1
ESRRA	TAAAGGTCACT	7,2383	12,6831	0,806915622	0,912014985	risk	1
RORA	CTAAAGGTCA	0	10,643	0	0,89723706	risk	1
NR2F1	TAAAGGTCACTAC	0	10,2106	0	0,884776523	risk	0
HNF4A	GTGACCTTTAGCTTT	6,26925	10,1082	0,818320526	0,8694205	risk	1
NR1H4	GTAGTGACCTT	0	9,93038	0	0,869246518	risk	1
Nr5a2	AAGCTAAAGGTCACT	0	11,4836	0	0,859419351	risk	1
HNF4G	AAAAGCTAAAGGTCC	8,22637	0	0,855273743	0	wt	0
Pax2	GGTCACTA	0	5,01899	0	0,841086371	risk	1
Nr2f6	AAAGCTAAAGGTCA	0	11,6441	0	0,834649752	risk	1
MNX1	GGTCACTACA	0	4,34715	0	0,808772596	risk	1
FOXA1	TCTGTAGTGACCTTT	0	0,795672	0	0,800124752	risk	1

Seq117

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
VDR	AGAGTGCA	0	7,57674	0	0,879535949	risk	1
CEBPA	AGTGCAAAATC	0	6,16935	0	0,871898783	risk	1
ISL2	GCACTCTA	0	7,04119	0	0,870680299	risk	0
CEBPB	GAGTGCAAAAT	0	3,95144	0	0,861358913	risk	1
NFAT5	ATTTTGCACT	0	6,443	0	0,841377938	risk	1
NFATC2	TTTTGCA	0	6,9943	0	0,838994417	risk	1
NFIA	AGTGCAAAAT	0	2,84446	0	0,835788317	risk	1
E2F3	AATAGAGCGCAAAATC	8,24254	0	0,81806109	0	wt	1
ETS1	TTTGCG	3,82245	0	0,816657782	0	wt	1
Gmeb1	GATTTTGCGCTCTATTA	6,63759	0	0,815722909	0	wt	0
NFATC1	ATTTTGCACT	5,28505	5,80683	0,800501023	0,815372467	risk	1
NFATC3	ATTTTGCACT	0	5,89827	0	0,808852865	risk	1
E2F4	GAGCGCAAAAT	3,48749	0	0,800038631	0	wt	1

Seq119

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
Prrx2	AATTA	5.58756	9.12437	0,838048459	0,999999981	Risk	0
Dlx4	TCAATTAT	5.68314	9.4114	0,859912623	0,961628714	Risk	1
HESX1	GATAATTGAA	3.34856	11.3415	0,805953711	0,959513688	Risk	0
MIXL1	TTCAATTATC	4.90977	9.55679	0,832996443	0,953935619	Risk	1
LBX2	TTCAATTATC	0	9.14439	0	0,924059323	Risk	1
VSX1	TCAATTAT	0	8.95335	0	0,920612325	Risk	1
HOXB2	TTCAATTATC	3.45078	7.73966	0,800614667	0,912028513	Risk	1
HOXB3	TTCAATTATC	4.07366	7.50279	0,828678707	0,908536693	Risk	1
VAX2	TCAATTAT	0	8.10382	0	0,908303252	Risk	1
HOXA2	TTCAATTATC	0	7.57297	0	0,901646588	Risk	1
NKX6-1	ATAATTGA	4.8019	7.31493	0,841219159	0,895818767	Risk	1
Lhx4	ATAATTGA	4.70094	7.41615	0,832474118	0,894923879	Risk	1
VENTX	TTCAATTAT	3.25782	6.81217	0,820202239	0,879232284	Risk	0
VDR	TTGGTTCA	0	7.43784	0	0,875575509	Risk	1
CEBPA	ATTTGATAATA	5.53487	4.87453	0,864297673	0,856386755	WT	1
HNF4G	AATTATCAAATTCCA	6.645	7.40642	0,834897627	0,844708587	Risk	0
LHX6	GATAATTGAA	0	5.571	0	0,844078596	Risk	1
RHOXF1	TTGAACCA		1.90526	0,803733662	0,841008099	Risk	0
SRF	AGAACCAAAATAGGCC	10.146	9.09013	0,840880715	0,828168717	WT	1
BARHL2	AATTATCAAA	0	2.50809	0	0,813281927	Risk	0
PDX1	ATAATTGA	2.24364	2.26688	0,809382437	0,809847369	Risk	1
MNX1	AATTATCAAA	0	4.32515	0	0,808268525	Risk	1

Seq790

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity allele	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
DBP	aattatataata	5,55267	13,7444	0,853285763	0,960698744	risk	1
HLF	aattatataata	0	12,8054	0	0,9512252	risk	1
TEF	aattatataata	5,72107	12,4399	0,84390711	0,93094498	risk	1
NFIL3	ttatataatat	8,52783	11,927	0,831126032	0,918084591	risk	0
CEBPB	attatataat	0	6,37874	0	0,871215647	risk	1
CEBPD	attatataat	0	6,29609	0	0,861146811	risk	1
ZEB1	gcacagctg	5,88416	0	0,845306775	0	wt	1
CEBPA	attatataata	0	3,35271	0	0,838155202	risk	1
Arid3a	attata	0	5,399	0	0,826577908	risk	1
Prrx2	tatta	0	4,91772	0	0,825515567	risk	1
HOXB3	agatattata	0	3,83659	0	0,823157828	risk	1
CEBPG	attatataat	0	3,26846	0	0,817915184	risk	1
TEAD1	tagatactat	4,39783	0	0,812735256	0	wt	1
Dlx4	gatattat	0	3,50408	0	0,800462443	risk	1

Seq132wt vs seq132PDX1

TF	Binding site	Score WT	Score Risk	Relative score WT	Relative score Risk	Highest affinity binidig site	Expression in endocrine pancreas (active promoter - 1; inactive promoter -0)
Pdx1	CTAATG	8,72988	0	0,965432864	0	PDX1 A	1
HOXB3	ACACATTAGG	8,23982	0	0,925700529	0	PDX1 A	1
HOXB2	ACACATTAGG	7,39411	0	0,90305208	0	PDX1 A	1
HOXA2	ACACATTAGG	7,13602	0	0,891767719	0	PDX1 A	1
HLTF	ACACATTGGG	0	5,99807	0	0,891560658	No PDX1 G	1
VENTX	ACACATTAG	7,26487	0	0,886750634	0	PDX1 A	0
Sox17	CACATTGGG	0	7,40552	0	0,862559299	No PDX1 G	0
NFIX	GCAGCCCAA	0	4,44073	0	0,850051973	No PDX1 G	1
Prrx2	CATTA	5,34257	0	0,843137572	0	PDX1 A	1
Neurog1	CCAATGTGTT	0	5,87923	0	0,839227048	No PDX1 G	0
Lhx4	CACATTAG	4,50603	0	0,827991314	0	PDX1 A	1
VAX2	CACATTAG	4,94313	0	0,824748192	0	PDX1 A	1
MAX	TAACACATTG	0	3,37815	0	0,822677822	No PDX1 G	1
NFIC	TGGGCT	0	4,18303	0	0,81530306	No PDX1 G	1
Dlx4	CACATTAG	3,95614	0	0,812795645	0	PDX1 A	1
Sox6	CCAATGTGTT	0	5,08245	0	0,811390272	No PDX1 G	1
TFEC	AACACATTAG	5,78176	0	0,806881079	0	PDX1 A	0
THAP1	CAGCCCAAT	0	4,49031	0	0,805059037	No PDX1 G	1
RUNX1	TAATGTGTTAT	6,66952	0	0,802184621	0	PDX1 A	0

Supplementary Table 3 – *List of the predicted TFBS for pancreatic TFs. The wt and risk alleles were analyzed using 719 specific position weight matrices for TFs in vertebrates. The score and relative score are represented for each allele pair. In each row, the allele for which the TF is predicted to bind preferentially is discriminated. Included also, are the seq132wt and seq132wtPDX1, where the nucleotide change resides in a putative binding site for PDX1.*

Wt /risk sequences alignments

Seq58

Query = wt

Sbjct = risk

```
Query  1     CTCTGAGAAGGAAATTGAACGCAATAGTTTTGAATTCTCAATGTTCATGTTTCCCCCGTA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     CTCTGAGAAGGAAATTGAACGCAATAGTTTTGAATTCTCAATGTTCATGTTTCCCCCGTA  60

Query  61    TGTGCGTGCGCATGCACATATATAAACACATCTAGCTTCACACCTGTGTACTCACACCTG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    TGTGCGTGCGCATGCACATATATAAACACATCTAGCTTCACACCTGTGTACTCACACCTG  120

Query  121   TGTGTACACACTCCTCCTGTACATGTGTATACACGCACACCCTCCCACCAGCATGAGCAG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   TGTGTACACACTCCTCCTGTACATGTGTATACACGCACACCCTCCCACCAGCATGAGCAG  180

Query  181   TATGTGGATTTTTTCTTCCAGCAGCCCTTTCAAGAACATACACCAGATGGGATTGGGTCT  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   TATGTGGATTTTTTCTTCCAGCAGCCCTTTCAAGAACATACACCAGATGGGATTGGGTCT  240

Query  241   GCTACATGACTGAGCAGCCCCCACATCAGAAAGTCATCATCATTGAATCCATCATGTGGT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   GCTACATGACTGAGCAGCCCCCACATCAGAAAGTCATCATCATTGAATCCATCATGTGGT  300

Query  301   GCAGCCTTGGCAAGGGCACAGCTCCTCGTCCTGCCCAATGGGCCGTGGAGGAATGCCTCC  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GCAGCCTTGGCAAGGGCACAGCTCCTCGTCCTGCCCAATGGGCCGTGGAGGAATGCCTCC  360

Query  361   TCCTCCAGAGCAGAGAGCAGATATGGGCAGGCAGTTGCCAGGCAGTTGAGCCGGGAGAAC  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   TCCTCCAGAGCAGAGAGCAGATATGGGCAGGCAGTTGCCAGGCAGTTGAGCCGGGAGAAA  420

Query  421   AGATGGCAGCTGCACAGCCTCCTTCTAGCCCAGTTTCCTGGCTGCCGGGGAAGGAGTTGC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   AGATGGCAGCTGCACAGCCTCCTTCTAGCCCAGTTTCCTGGCTGCCGGGGAAGGAGTTGC  480

Query  481   AGGCGGCAAGGATCCTGCAGAAACCGCCCTCCCTATAACAGGCCATCTTGCCCTCTATCC  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  481   AGGCGGCAAGGATCCTGCAGAAACCGCCCTCCCTATAACAGGCCATCTTGCCCTCTATCC  540

Query  541   TCTTACTGGGCTTTGGAGGTCCCAATCCTGCCCTGGTTCACCACCTGTGTGGCTTACCAA  600
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   TCTTACTGGGCTTTGGAGGTCCCAATCCTGCCCTGGTTCACCACCTGTGTGGCTTACCAA  600

Query  601   GCTGTGCAACTTGGGGCAGGATGAATATCCTGAGTGCTGGTACTAGAACTCTAATTCTCT  660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   GCTGTGCAACTTGGGGCAGGATGAATATCCTGAGTGCTGGTACTAGAACTCTAATTCTCT  660

Query  661   AATCTCAGTTATAACCTCAAGTCTGCTCCTTTACTCCTAATCAGAAGACAACTAGCCTTA  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   AATCTCAGTTATAACCTCAAGTCTGCTCCTTTACTCCTAATCAGAAGACAACTAGCCTTA  720

Query  721   CCTGATTAATCATGACCTTAGGAAGAAACCCCTGGATAATTTTAGCAAATATCACATCTC  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   CCTGATTAATCATGACCTTAGGAAGAAACCCCTGGATAATTTTAGCAAATATCACATCTC  780

Query  781   TCAAATCTACACACACACATTGTCTCCAGGTATAGGAGACACTGTAGATGCCAGTCTGGT  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   TCAAATCTACACACACACATTGTCTCCAGGTATAGGAGACACTGTAGATGCCAGTCTGGT  840

Query  841   CCGTATCCCTCAGCCCTTACCACTTCAGTGATCCCACCAAATTCCAATGGCCCGTCTCTG  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   CCGTATCCCTCAGCCCTTACCACTTCAGTGATCCCACCAAATTCCAATGGCCCGTCTCTG  900

Query  901   CATGTCTTTGCCTGAGGGTTTTCTCTGAAACCAGAGAAGCCTGTTTTGCCTGCAAAGGCA  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   CATGTCTTTGCCTGAGGGTTTTCTCTGAAACCAGAGAAGCCTGTTTTGCCTGCAAAGGCA  960

Query  961   GGCTGAAGTGCAATGATCTATGTCAACTATTGATGGAAGTGGGTATATAAAACTCAGTGT  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GGCTGAAGTGCAATGATCTATGTCAACTATTGATGGAAGTGGGTATATAAAACTCAGTGT  1020

Query  1021  GCCACCTCCCCCGACTCCTCAGATGGGATGGCTTTAATGTGAGGTTTT  1068
              ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  GCCACCTCCCCCGACTCCTCAGATGGGATGGCTTTAATGTGAGGTTTT  1068
```

Seq68

Query = wt

Sbjct = risk

```
Query  1     CCTGGAGATTGTCTTCTAAGCTGCCAAAATCTTGTCCATTGTTCTCTAGGCACCGTCTTA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     CCTGGAGATTGTCTTCTAAGCTGCCAAAATCTTGTCCATTGTTCTCTAGGCACCGTCTTA  60

Query  61    CTGTGTACCAGATTTTTTATGATTATGATTCTATGGTAGCAGTAAAGTGCTAAGTCTGAG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CTGTGTACCAGATTTTTTATGATTATGATTCTATGGTAGCAGTAAAGTGCTAAGTCTGAG  120

Query  121   GAGATACCAGCATGCAACCTACATACTTCTAAAGGTTGCTTAACTAACAATACTTCTTGT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GAGATACCAGCATGCAACCTACATACTTCTAAAGGTTGCTTAACTAACAATACTTCTTGT  180

Query  181   GCATATCTCCCAAGGTGTACAAATACAGCTGTCCTGAAGTAGAATGTCAAAAGCCACACT  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   GCATATCTCCCAAGGTGTACAAATACAGCTGTCCTGAAGTAGAATGTCAAAAGCCACACT  240

Query  241   AGATAGCATTTCCATTCCTTTCTGGCAAACTGGCTGCAAATTTGAATAGTATTCAGAGAT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   AGATAGCATTTCCATTCCTTTCTGGCAAACTGGCTGCAAATTTGAATAGTATTCAGAGAT  300

Query  301   GAGTGGGGAGCTGAAAGGCGATATAAAAAAACAGACTTTACCTCAATATTTTGCCAAAGG  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GAGTGGGGAGCTGAAAGGCGATATAAAAAAACAGACTTTACCTCAATATTTTGCCAAAGG  360

Query  361   CAACTTAATTCAGAGTATTACATAATTGTTTGCTTATTACTTCATTTTCATTATCAACCC  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   CAACTTAATTCAGAGTATTACATAATTGTTTGCTTATTACTTCATTTTCATTATCAACCC  420

Query  421   TGGGAGGTGGTAGATTTGAGTAATTCATTTAACTTCACTGATCTTCAATTTCTTCATCTT  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   TGGGAGGTGGTAGATTTGAGTAATTCATTTAACTTCACTGATCTTCAATTTCTTCATCTT  480

Query  481   CAAAAAAGAGAGACAGACAGAAAATATTTTGCTGTACAAAAAAGCATCCAAGATCTTATC  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   CAAAAAAGAGAGACAGACAGAAAATATTTTGCTGTACAAAAAAGCATCCAAGATCTTATC  540
```

```
Query  541   AAATTAAAAATTCTTGGTTCATGATTTCTATATTGCAAATGAGGAAATGAGGTGCAAAGG  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   AAATTAAAAATTCTTGGTTCATGATTTCTATATTGCAAATGAGGAAATGAGGTGCAAAGG  600

Query  601   CGTTTAAATGCCAACTCCAAGTTCAAAAAGTGATATTACAGCAGGATTAAGCTCAAATCG  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   CGTTTAAATGCCAACTCCAAGTTCAAAAAGTGATATTACAGCAGGATTAAGCTCAAATCG  660

Query  661   AAGCCTCCTGGTTCCTAGCATAGTACGTAGGCCAAAACAAAGCCAGTCAGAAGGTTAAGG  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   AAGCCTCCTGGTTCCTAGCATAGTACGTAGGCCAAAACAAAGCCAGTCAGAAGGTTAAGG  720

Query  721   GGAATTTAATTGGCAATAAAATACATCAACAGCAAAAAGGGAGAGGAAGCAAGTAAACAA  780
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   GGAATTTAATTGGCAATAAAATACATCAACAGCAAAAAGGGAGAGGAAGCAAGTAAACAA  780

Query  781   GAACTGATGCTGTAAATACTGGCCTTGCTGCATCACTGGCTAGACAGAGCAGGAGTGTGG  840
             |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  781   GAACTGATGCTGTAAATACTGGCCTTGCCGCATCACTGGCTAGACAGAGCAGGAGTGTGG  840

Query  841   TTGCCATGGTTAATGCAACCACTGAGTCCTGAGGTTTTTAAAAAATATTGCAGATCAGGT  900
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   TTGCCATGGTTAATGCAACCACTGAGTCCTGAGGTTTTTAAAAAATATTGCAGATCAGGT  900

Query  901   GCAGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGACCGAGGTGGACAGATTGCCTGA  960
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   GCAGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGACCGAGGTGGACAGATTGCCTGA  960

Query  961   GTTCAGGAGTTCAAGACCAGCCTGGGCAACATGACAAAACTCTGTCTCCACAAAAAATAC  1020
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GTTCAGGAGTTCAAGACCAGCCTGGGCAACATGACAAAACTCTGTCTCCACAAAAAATAC  1020

Query  1021  AAAAATTAACCGGGCATGGTGGTACATGCCTGTAGTCCCAGCTACTAGGGAGGCTGAGGC  1080
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  AAAAATTAACCGGGCATGGTGGTACATGCCTGTAGTCCCAGCTACTAGGGAGGCTGAGGC  1080

Query  1081  ACGAGAATCACTTGAACCCGGGAGGCAGAGGTTGCACTGAGCTGAGATCATGCCACTGCA  1140
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  ACGAGAATCACTTGAACCCGGGAGGCAGAGGTTGCACTGAGCTGAGATCATGCCACTGCA  1140
```

```
Query  1141  TCCTGGACAGGGTGACAGGGTAAGAGACCCTGCCTCCAAAAGAAAAAAAAAAAAAAGAAA  1200
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  TCCTGGACAGGGTGACAGGGTAAGAGACCCTGCCTCCAAAAGAAAAAAAAAAAAAAGAAA  1200

Query  1201  AAGCAAAAAATGTTAAATATCTCTAACCTTAAGTGAACTGGATGGGAGGTGTTAAAAATG  1260
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  AAGCAAAAAATGTTAAATATCTCTAACCTTAAGTGAACTGGATGGGAGGTGTTAAAAATG  1260

Query  1261  GCCAAGAGGCCCTTGTCTGATTCACTGTGGCATTCATTCAGTTCAGGACAGCCACTAGAA  1320
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GCCAAGAGGCCCTTGTCTGATTCACTGTGGCATTCATTCAGTTCAGGACAGCCACTAGAA  1320

Query  1321  ATAAGCCATGGCTCTAGATGCAATCTGAGTTGC  1353
             ||||||||||||||||||||||||||||||||
Sbjct  1321  ATAAGCCATGGCTCTAGATGCAATCTGAGTTGC  1353
```

Seq72

Query = wt

Sbjct = risk

```
Query  1    TTCGCAAAACATCTCATCACCAGAAAAAAAAATCAGCTGACCACAAACCCGCCTCAATGAA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TTCGCAAAACATCTCATCACCAGAAAAAAAAATCAGCTGACCACAAACCCGCCTCAATGAA  60

Query  61   AGGGAAGAAAATTTGAAGTCACAATGTCCAAATTCCTACAAACACACATTGTTCCAGAAA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   AGGGAAGAAAATTTGAAGTCACAATGTCCAAATTCCTACAAACACACATTGTTCCAGAAA  120

Query  121  ATAGGACTGCCCGTAGAAATTTACATGCCAACCATTTGAATATACAAGGCATCTGCTTCA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  ATAGGACTGCCCGTAGAAATTTACATGCCAACCATTTGAATATACAAGGCATCTGCTTCA  180

Query  181  CTTCTGCCTAGTTCAAAGGGAGCCACATTCCCATCTACTACCTTTGGTCAGCTTTTGGAC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CTTCTGCCTAGTTCAAAGGGAGCCACATTCCCATCTACTACCTTTGGTCAGCTTTTGGAC  240

Query  241  AATGGGTGCTGAAGGTGTTTTGTCAAGAAACACAGCCTACGCTGGTCATTTTTCATTTGG  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  AATGGGTGCTGAAGGTGTTTTGTCAAGAAACACAGCCTACGCTGGTCATTTTTCATTTGG  300

Query  301  CACCAGTTGGTTCCATCTGGCTATTTTTAAACATAAGTAAATATTACAGCACGCATTTCT  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  301  CACCAGTTGGTTCCATCTGGCTATTTTTAAACATAAGTAAATATTACAGCACGCATTTCT  360

Query  361  GTCCTTTATGCGTTTCTGTCATCAAACACGTTCTCTGTTTAATGAAGCCTGTGAGAAGCG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  GTCCTTTATGCGTTTCTGTCATCAAACACGTTCTCTGTTTAATGAAGCCTGTGAGAAGCG  420

Query  421  ACATCATTTGTACAAAATCAAATTGTGTTTTCAGACCTGCTTTATCATATTCCCTATGTA  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  ACATCATTTGTACAAAATCAAATTGTGTTTTCAGACCTGCTTTATCATATTCCCTATGTA  480

Query  481  TCTGTAGGGACCTTTAGCTTTTTCTTCCTGGGTAAAGAGAGCCTGGAAATAGTTTTTTAA  540
            ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  TCTGTAGTGACCTTTAGCTTTTTCTTCCTGGGTAAAGAGAGCCTGGAAATAGTTTTTTAA  540

Query  541  ATATAACAATCAACCTCTTTCTTTCTTTAAATTCCTAAGCCAGGCACTTAATACTCCTGT  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  ATATAACAATCAACCTCTTTCTTTCTTTAAATTCCTAAGCCAGGCACTTAATACTCCTGT  600

Query  601  GGCCTGGTCTTGGCATAAAGATCAGAAACTCTGCCTTCAGTTCTCACCCTG  651
            |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GGCCTGGTCTTGGCATAAAGATCAGAAACTCTGCCTTCAGTTCTCACCCTG  651
```

Seq73

Query = wt

Sbjct = risk

```
Query  1    TCACCTGTGCCTGGCTGGGTTTTCTGATTTTTCAACACTTGCTCCTAACCCTCAGAACCT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    TCACCTGTGCCTGGCTGGGTTTTCTGATTTTTCAACACTTGCTCCTAACCCTCAGAACCT  60

Query  61   TTTTTCTTTTCATTCCTAGTCCATCTCCAGTTTCCAACTTCCCGGCTCCCCACTATCCCA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TTTTTCTTTTCATTCCTAGTCCATCTCCAGTTTCCAACTTCCCGGCTCCCCACTATCCCA  120

Query  121  GAGGATTTGCTGAAGAGGGACGGGGGTAGGGGTAGGAGGCTGACACGGTACTTCGCAAAA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GAGGATTTGCTGAAGAGGGACGGGGGTAGGGGTAGGAGGCTGACACGGTACTTCGCAAAA  180

Query  181  CATCTCATCACCAGAAAAAAAATCAGCTGACCACAAACCCGCCTCAATGAAAGGGAAGAA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CATCTCATCACCAGAAAAAAAATCAGCTGACCACAAACCCGCCTCAATGAAAGGGAAGAA  240

Query  241  AATTTGAAGTCACAATGTCCAAATTCCTACAAACACACATTGTTCCAGAAAATAGGACTG  300
            |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  241  AATTTGAAGTCACAATGTCCAAATTCCTACAAATACACATTGTTCCAGAAAATAGGACTG  300
```

```
Query  301  CCCGTAGAAATTTACATGCCAACCATTTGAATATACAAGGCATCTGCTTCACTTCTGCCT  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  CCCGTAGAAATTTACATGCCAACCATTTGAATATACAAGGCATCTGCTTCACTTCTGCCT  360

Query  361  AGTTCAAAGGGAGCCACATTCCCATCTACTACCTTTGGTCAGCTTTTGGACAATGGGTGC  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  AGTTCAAAGGGAGCCACATTCCCATCTACTACCTTTGGTCAGCTTTTGGACAATGGGTGC  420

Query  421  TGAAGGTGTTTTGTCAAGAAACACAGCCTACGCTGGTCATTTTTCATTTGGCACCAGTTG  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  TGAAGGTGTTTTGTCAAGAAACACAGCCTACGCTGGTCATTTTTCATTTGGCACCAGTTG  480

Query  481  GTTCCATCTGGCTATTTTTAAACATAAGTAAATATTACAGCACGCATTTCCGTCCTTTAT  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  GTTCCATCTGGCTATTTTTAAACATAAGTAAATATTACAGCACGCATTTCCGTCCTTTAT  540

Query  541  GCGTTTCTGTCATCAAACACGTTCTCTGTTTAATGAAGCCTGTGAGAAGCGACATCATTT  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  GCGTTTCTGTCATCAAACACGTTCTCTGTTTAATGAAGCCTGTGAGAAGCGACATCATTT  600

Query  601  GTACAAAATCAAATTGTGTTTTCAGACCTGCTTTATCATATTCCCTATGTATCTGTAGGG  660
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GTACAAAATCAAATTGTGTTTTCAGACCTGCTTTATCATATTCCCTATGTATCTGTAGGG  660

Query  661  ACCTTTAGCTTTTTCTTCCTGGGTAAAGAGAGCCCGGAAATAGTTTTTTAAATATAACAA  720
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661  ACCTTTAGCTTTTTCTTCCTGGGTAAAGAGAGCCCGGAAATAGTTTTTTAAATATAACAA  720

Query  721  TCAACCTCTTTCTTTCTTTAAATTCCTAAGCCAGGCACTTAATACTCCTGTGGCCTGGTC  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721  TCAACCTCTTTCTTTCTTTAAATTCCTAAGCCAGGCACTTAATACTCCTGTGGCCTGGTC  780

Query  781  TTGGCATAAAGATCAGAAACTCTGCCTTCAGTTCTCACCCTGCAGGCCACCCCAC  835
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781  TTGGCATAAAGATCAGAAACTCTGCCTTCAGTTCTCACCCTGCAGGCCACCCCAC  835
```

Seq117

Query = wt

Sbjct = risk

```
Query  1     CTTCCCGGATGTGGAGATTCAGCCAGGGGCAGCCTTCTTCCTGGGAGACCAAGGTGACTG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     CTTCCCGGATGTGGAGATTCAGCCAGGGGCAGCCTTCTTCCTGGGAGACCAAGGTGACTG  60
```

```
Query  61    CTGCCTACACTGTCCCTACCTGACAGAGAAGCATGGGGGGTTGGGAGACCCCCAAAACTG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CTGCCTACACTGTCCCTACCTGACAGAGAAGCATGGGGGGTTGGGAGACCCCCAAAACTG  120

Query  121   GGGGCCAATGTGTTATGCAGATACATGTCTCTCCTAAAGGCTTAGAGTCTCAGTTTTGAG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GGGGCCAATGTGTTATGCAGATACATGTCTCTCCTAAAGGCTTAGAGTCTCAGTTTTGAG  180

Query  181   AGACAGATGGAAGACATGCAGTCAGGATCCACAGTCTCGGATCCCAGGGCAGACTTGGGA  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   AGACAGATGGAAGACATGCAGTCAGGATCCACAGTCTCGGATCCCAGGGCAGACTTGGGA  240

Query  241   ACTGGGCAGTTCTCATTCTGCCTGAAGAAAACAGTGCTTCCCTCACCCCTACTGCTGTCA  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   ACTGGGCAGTTCTCATTCTGCCTGAAGAAAACAGTGCTTCCCTCACCCCTACTGCTGTCA  300

Query  301   GCCTCACCTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTGTGTGTATGTGTGTGT  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GCCTCACCTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTGTGTGTATGTGTGTGT  360

Query  361   GTGTGTGAGAGAGAGAGATAATTTGCACTATTTCACTTAATCACTTTCTTGAAAACGCAG  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   GTGTGTGAGAGAGAGAGATAATTTGCACTATTTCACTTAATCACTTTCTTGAAAACGCAG  420

Query  421   GCCAAGCCTCCAAAGAGTCCTCTTTACACCAACCTCAGAGGACTCGTGAGTCCTAGGGAA  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   GCCAAGCCTCCAAAGAGTCCTCTTTACACCAACCTCAGAGGACTCGTGAGTCCTAGGGAA  480

Query  481   ACTTGCCTGGGGAGAAGGAACTCTCCGGAAGGCAAACACAGCAGCTCCTTTTCTGGGCAT  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   ACTTGCCTGGGGAGAAGGAACTCTCCGGAAGGCAAACACAGCAGCTCCTTTTCTGGGCAT  540

Query  541   ATGGGAGGATTATCAGGCTCTCCAGGAAGCCATGCTGGATGAAGGCCACTGTATCGCTTC  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   ATGGGAGGATTATCAGGCTCTCCAGGAAGCCATGCTGGATGAAGGCCACTGTATCGCTTC  600

Query  601   GTGTCCCCGTGGAACTCATAAGCAGATTTTGCCCTCTATTAATCTACATCTGTTTGCACG  660
             |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
Sbjct  601   GTGTCCCCGTGGAACTCATAAGCAGATTTTGCACTCTATTAATCTACATCTGTTTGCACG  660

Query  661   TCCCTGCTGTCAGCAGCTTCTGTCTAGGGAGGCGGCAAACACGCAGTGAAAGGAGCGTGG  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   TCCCTGCTGTCAGCAGCTTCTGTCTAGGGAGGCGGCAAACACGCAGTGAAAGGAGCGTGG  720

Query  721   GCTCTGGAGCGGCCACTTGGGTGCAGCCTGGCTGTACCACTGCTGGCTCGGGGACCTTGA  780
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  721   GCTCTGGAGCGGCCACTTGGGTGCAGCCTGGCTGTACCACTGCTGGCTCGGGGACCTTGA  780

Query  781   GCCCTTGAGTCTTTTAATCTAGAATGCATTCCCCACCCTTTTTTCTTTCTCCATGACTGG  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   GCCCTTGAGTCTTTTAATCTAGAATGCATTCCCCACCCTTTTTTCTTTCTCCATGACTGG  840

Query  841   CTTTTTAAAGAGTACACAATGTAGACTGTCCCACGTTCTGGATTTGTCCAGTTGTTTCTT  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   CTTTTTAAAGAGTACACAATGTAGACTGTCCCACGTTCTGGATTTGTCCAGTTGTTTCTT  900

Query  901   GGTGATGTGTTTTTTTCTTTAAACTTTTTATTACAGAAAATTTCAAATAGAGTATACATG  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   GGTGATGTGTTTTTTTCTTTAAACTTTTTATTACAGAAAATTTCAAATAGAGTATACATG  960

Query  961   AAGTAAATAGCATACATGAGCCATCTATCAACCGGGTGTGTTGATAAAGGTAGCTGGCTT  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   AAGTAAATAGCATACATGAGCCATCTATCAACCGGGTGTGTTGATAAAGGTAGCTGGCTT  1020

Query  1021  GCAGTGCTGTCAGGAGGACTGGTGATAAGGTGGGCAGAGCCTGGCCCAGTGCCTGGGACG  1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  GCAGTGCTGTCAGGAGGACTGGTGATAAGGTGGGCAGAGCCTGGCCCAGTGCCTGGGACG  1080

Query  1081  CACTGGCTTAGGAGGCATGGAAGCTGCCAGATGGTGGTGAAGATGAGATGATGGTGGTGA  1140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  CACTGGCTTAGGAGGCATGGAAGCTGCCAGATGGTGGTGAAGATGAGATGATGGTGGTGA  1140

Query  1141  CATTCGGGTCCCAGTCTATGCTTTCCATCTCTCTTGGTAGGCTAGGAGCTCCTCCCCAAA  1200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  CATTCGGGTCCCAGTCTATGCTTTCCATCTCTCTTGGTAGGCTAGGAGCTCCTCCCCAAA  1200

Query  1201  GCGGCTGGTGTTGCTTCCTCCTTTCTCCTCC  1231
              ||||||||||||||||||||||||||||||
Sbjct  1201  GCGGCTGGTGTTGCTTCCTCCTTTCTCCTCC  1231
```

Seq119

Query = wt

Sbjct = risk

```
Query  1     ATGAAGATGGCACAAACAAACATCCCACTCATTCCAATTAGAAAGAGAGAACGTCGCCCT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGAAGATGGCACAAACAAACATCCCACTCATTCCAATTAGAAAGAGAGAACGTCGCCCT  60

Query  61    GCCTTCTCCACAAGGAATACCTAGGACAATTTTAAAGAAATTATCTCAGTTTTGTGAGTC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  61    GCCTTCTCCACAAGGAATACCTAGGACAATTTTAAAGAAATTATCTCAGTTTTGTGAGTC  120

Query  121   ACAAAATATTGACTGTTTACTTAAATCATTTTTACTGCCAAAGAGGTCATTTTAATTTAT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   ACAAAATATTGACTGTTTACTTAAATCATTTTTACTGCCAAAGAGGTCATTTTAATTTAT  180

Query  181   ATAACAGATAAGAGGCAATTCCATCCTTCCATTTTACAGGAGTGAGATCAAGGTGGGGAG  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   ATAACAGATAAGAGGCAATTCCATCCTTCCATTTTACAGGAGTGAGATCAAGGTGGGGAG  240

Query  241   TGACTCTAGCCCCAGGTGGCAGAGCTAGTTTCAGTGGACAGCGGGTAATGATCAAATATC  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   TGACTCTAGCCCCAGGTGGCAGAGCTAGTTTCAGTGGACAGCGGGTAATGATCAAATATC  300

Query  301   TGCTTAATTTAGTCACATTGGTTTGAATAGACCATTATAGAGATAGAACCATAATGTTAC  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   TGCTTAATTTAGTCACATTGGTTTGAATAGACCATTATAGAGATAGAACCATAATGTTAC  360

Query  361   CACTACTCCAGGATTATCACTTTGGGTACAATATTCACAGACTGTACAGCTGGGGTCTCT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   CACTACTCCAGGATTATCACTTTGGGTACAATATTCACAGACTGTACAGCTGGGGTCTCT  420

Query  421   TCCAATAGATGTCCTGTACATTTAAAATTTGAAGGTTCCACATTTAAACATGACATTAAC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   TCCAATAGATGTCCTGTACATTTAAAATTTGAAGGTTCCACATTTAAACATGACATTAAC  480

Query  481   TTATTTTCTAATAGTTATCCTAAGTTATAATTACAAGGGTCTGTGCACTATTGACACCGG  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   TTATTTTCTAATAGTTATCCTAAGTTATAATTACAAGGGTCTGTGCACTATTGACACCGG  540

Query  541   AAACAGTACAACTGTTCAGGTGTCAACTATTCTCATCCAATTTTCTCACTGCTTAATAAA  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   AAACAGTACAACTGTTCAGGTGTCAACTATTCTCATCCAATTTTCTCACTGCTTAATAAA  600

Query  601   GTCCTATTATGACCACTCACCTAGCAACATGTGGAATTTGATAATAGAACCAAAATAGGC  660
             |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct  601   GTCCTATTATGACCACTCACCTAGCAACATGTGGAATTTGATAATTGAACCAAAATAGGC  660

Query  661   CTGCATTTTTCTGGTATGCTTAATTGTATCTTGTGCTGGGAGTGAACTATTCGTTTTAAA  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   CTGCATTTTTCTGGTATGCTTAATTGTATCTTGTGCTGGGAGTGAACTATTCGTTTTAAA  720

Query  721   AGTTAGCACTTTAAAGCAAAGGGTTAAAATGCTTTCTTATCTTAACTGAGGAGATACGTC  780
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   AGTTAGCACTTTAAAGCAAAGGGTTAAAATGCTTTCTTATCTTAACTGAGGAGATACGTC  780
```

```
Query  781   ATTGATTCCTTTAGTTTTGTGGACCTTTGCATTAATGGTTAAGTTTTACAATTAGCCAAG  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   ATTGATTCCTTTAGTTTTGTGGACCTTTGCATTAATGGTTAAGTTTTACAATTAGCCAAG  840

Query  841   ACCAGAGTTTGTTATACTCTTGCCTGATCCTTGGAATATTACCAAGGCTGGTGAGTCACT  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   ACCAGAGTTTGTTATACTCTTGCCTGATCCTTGGAATATTACCAAGGCTGGTGAGTCACT  900

Query  901   CTCCCCCTCACCCCCCACAAAACTTCCTGTCATAATTTCTCACTTTGGCACCTGGTGAGC  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   CTCCCCCTCACCCCCCACAAAACTTCCTGTCATAATTTCTCACTTTGGCACCTGGTGAGC  960

Query  961   AAGTTGGCTGGCTCATCAGCCTATCAGTAAATCAATAGCAATCCAATGTCTTTAGTATAG  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   AAGTTGGCTGGCTCATCAGCCTATCAGTAAATCAATAGCAATCCAATGTCTTTAGTATAG  1020

Query  1021  TGAGCTCTGGCTCTGAAGACATCTTAGAAACATTTCACCTACAGTGAAAGCCTGTTTCTC  1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  TGAGCTCTGGCTCTGAAGACATCTTAGAAACATTTCACCTACAGTGAAAGCCTGTTTCTC  1080

Query  1081  AGAGGCACACTCATATCAGAGTTTTAAATCACTTCAGATAAGATGATAAGCTAAGAGCAC  1140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  AGAGGCACACTCATATCAGAGTTTTAAATCACTTCAGATAAGATGATAAGCTAAGAGCAC  1140

Query  1141  ATATTTAATTATTGAGAAGGAAGTTGGAAAAAACAGACTTTTCTCAAGTTCATTAAATTT  1200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  ATATTTAATTATTGAGAAGGAAGTTGGAAAAAACAGACTTTTCTCAAGTTCATTAAATTT  1200

Query  1201  CTCTCTATATTTTAACTCTTTTATTATGATGTTCAAAACTGGCTTGGTGCCACCACGTCA  1260
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  CTCTCTATATTTTAACTCTTTTATTATGATGTTCAAAACTGGCTTGGTGCCACCACGTCA  1260

Query  1261  AGCTACCAAGAAAATAGAAGCATTTCTTTGTGAATAAGCTCCCTTACTATAAACTCTAAA  1320
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  AGCTACCAAGAAAATAGAAGCATTTCTTTGTGAATAAGCTCCCTTACTATAAACTCTAAA  1320

Query  1321  ACTGAACATCTTGAAAATTGAGTTGGCCTCTGTCCCTGAGGCCTGGTGTACACTCAGGAT  1380
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  ACTGAACATCTTGAAAATTGAGTTGGCCTCTGTCCCTGAGGCCTGGTGTACACTCAGGAT  1380

Query  1381  GAGTGTGACAAGGAGTGAGACATGGTCAGTGGCTGGAGATGTCTTCCCAATTAGAACGCT  1440
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  GAGTGTGACAAGGAGTGAGACATGGTCAGTGGCTGGAGATGTCTTCCCAATTAGAACGCT  1440

Query  1441  CATCTTTGCCCCAGGGCCTCACGCAAATCCTGGGCTCTTAGTTCCAATTTGGTAGTCACC  1500
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  1441  CATCTTTGCCCCAGGGCCTCACGCAAATCCTGGGCTCTTAGTTCCAATTTGGTAGTCACC  1500

Query  1501  GCTTCTTTCTCAGTGTCATCCTGCTTCCTAGATCCCTCCATTTTCTTTCTGCTTTTTTTC  1560
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  GCTTCTTTCTCAGTGTCATCCTGCTTCCTAGATCCCTCCATTTTCTTTCTGCTTTTTTTC  1560

Query  1561  TCTCAAGATTACAGGCCCTCTTCTACCAGCCCATTAGT  1598
             |||||||||||||||||||||||||||||||||||||
Sbjct  1561  TCTCAAGATTACAGGCCCTCTTCTACCAGCCCATTAGT  1598
```

Seq132

Query = wt

Sbjct = risk

```
Query  1     GCATTTACTGCCTCAAGAGAAAGCTCAGGCCAAGGATAACACATTAGGCTGCTAAAAAAT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GCATTTACTGCCTCAAGAGAAAGCTCAGGCCAAGGATAACACATTAGGCTGCTAAAAAAT  60

Query  61    GACTCAAGGAGTTACTTAGGGGCAGGGGACAGGAACACTAGTGACGGTCATTTGAAAACT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GACTCAAGGAGTTACTTAGGGGCAGGGGACAGGAACACTAGTGACGGTCATTTGAAAACT  120

Query  121   GTGCTTTCTCAGACATCGTCATCTTCTCGAAGGGGCTTAGGTCTGACTTGAGGGAATTTA  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GTGCTTTCTCAGACATCGTCATCTTCTCGAAGGGGCTTAGGTCTGACTTGAGGGAATTTA  180

Query  181   CAGCATTCACTATGCCAATATTTCACTGCAAATGGAAGAATCTAGTAAACACTCAGGTTG  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   CAGCATTCACTATGCCAATATTTCACTGCAAATGGAAGAATCTAGTAAACACTCAGGTTG  240

Query  241   CTATGGAGGAGACTTTACAGCCTGCTGATAGCATTTGGGACAGGAAAAAAATGGCAGTGA  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   CTATGGAGGAGACTTTACAGCCTGCTGATAGCATTTGGGACAGGAAAAAAATGGCAGTGA  300

Query  301   GGGTTGCTGCTCTAAGAAAGGTCTGTGGGGAGCTCTAAACGCTTCCTGGTTCCTGTCACT  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GGGTTGCTGCTCTAAGAAAGGTCTGTGGGGAGCTCTAAACGCTTCCTGGTTCCTGTCACT  360

Query  361   ATCCTTGATATAAATTCTGATTACATACAAAGCCTGCAACCCAGAGATCCCTGGGAGCTG  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   ATCCTTGATATAAATTCTGATTACATACAAAGCCTGCAACCCAGAGATCCCTGGGAGCTG  420

Query  421   GTGAGCAGGGCTGTATAAGAAATAATAGTTCTGTCTTGGCTTTTTCCAGGGCTTGTCTCC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  421   GTGAGCAGGGCTGTATAAGAAATAATAGTTCTGTCTTGGCTTTTTCCAGGGCTTGTCTCC  480

Query  481   CCTTCCATAGTAAGCTCCTAGGAATGCCAGACTCCAGAGATAACAGTGGACAGAAAGAGT  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   CCTTCCATAGTAAGCTCCTAGGAATGCCAGACTCCAGAGATAACAGTGGACAGAAAGAGT  540

Query  541   TCCCATAGCGACAGGGCACTTTGCTGCACTAGAGTTTCCCCTGCCTTGTCTGTGTGAATG  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   TCCCATAGCGACAGGGCACTTTGCTGCACTAGAGTTTCCCCTGCCTTGTCTGTGTGAATG  600

Query  601   TAGCTGATTATCAGAGCAAACGTGGCTTCCTCTGAGTGCCCTGCCTCTGCCCCACCCCAG  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   TAGCTGATTATCAGAGCAAACGTGGCTTCCTCTGAGTGCCCTGCCTCTGCCCCACCCCAG  660

Query  661   CAGGTCAAAGACAAAGTACTTGAAGTTGGAGTCAGAGCAGTCGCCCATGCGTGTGCAATC  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   CAGGTCAAAGACAAAGTACTTGAAGTTGGAGTCAGAGCAGTCGCCCATGCGTGTGCAATC  720

Query  721   AGTGCTAATCTCCCTGTGCTTCTTTATCAACAGCAGCCAGCTGGGACAGCCAAGTGGTTC  780
             |||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  721   AGTGCTAATCTCCCTGTGCTTCTTTATCAACAGCAGCCAGCCGGGACAGCCAAGTGGTTC  780

Query  781   GGAGAGAAATTGCTAAAGCCCTTAGCAAAAGCTTTACGATGCACTCACTCACCATTCAGA  840
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   GGAGAGAAATTGCTAAAGCCCTTAGCAAAAGCTTTACGATGCACTCACTCACCATTCAGA  840

Query  841   TGGAATCTCCAGTTGACCAGGACCCCGACTGCCTTTTCTGTGAAGACCCCTGTGACTAGC  900
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   TGGAATCTCCAGTTGACCAGGACCCCGACTGCCTTTTCTGTGAAGACCCCTGTGACTAGC  900

Query  901   TCAGTCACACCGTCAGTTTCCCAAATTTGACAGGCCACCTTCAAACATGCTGCTATGCAG  960
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   TCAGTCACACCGTCAGTTTCCCAAATTTGACAGGCCACCTTCAAACATGCTGCTATGCAG  960

Query  961   TTTCTGCATCATAGAAAATAAGGAACCAAAGGAAGAAATTCATGTCATGGTGCAGTGCAC  1020
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   TTTCTGCATCATAGAAAATAAGGAACCAAAGGAAGAAATTCATGTCATGGTGCAGTGCAC  1020

Query  1021  ATTTTATCTATTTATTTAGTTGCATTCACCATGAAGGAAGAGGCACTGAGATCCATCAAT  1080
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  ATTTTATCTATTTATTTAGTTGCATTCACCATGAAGGAAGAGGCACTGAGATCCATCAAT  1080

Query  1081  CAATTGGATTATATACTGATCAGTAGCTGTGTTCAATTGCAGGAATGTGTATATAGATTA  1140
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  CAATTGGATTATATACTGATCAGTAGCTGTGTTCAATTGCAGGAATGTGTATATAGATTA  1140
```

```
Query  1141  TTCCTGAGTGGAGCCGAAGTAACAGCTGTTTGTAACTATCGGCAATACCAAATTCATCTC  1200
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  TTCCTGAGTGGAGCCGAAGTAACAGCTGTTTGTAACTATCGGCAATACCAAATTCATCTC  1200

Query  1201  CCTTCCAATAATGCATCTTGAGAACACATAGGTAAATTTGAACTCAGGAAAGTCTTACTA  1260
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  CCTTCCAATAATGCATCTTGAGAACACATAGGTAAATTTGAACTCAGGAAAGTCTTACTA  1260

Query  1261  GAAATCAGTGGAAGGGACAAATAGTCACAAAATTTTACCAAAACATTAGAAACAAAAAAT  1320
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GAAATCAGTGGAAGGGACAAATAGTCACAAAATTTTACCAAAACATTAGAAACAAAAAAT  1320

Query  1321  AAGGAGAGCCAAGTCAGGAATAAAAGTGACTCTGTATGCTAACGCCACATTAGAACTTGG  1380
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  AAGGAGAGCCAAGTCAGGAATAAAAGTGACTCTGTATGCTAACGCCACATTAGAACTTGG  1380

Query  1381  TTCTCTCACCAAGCTGTAATGTGATTTTTTTTTCTACTCTGAATTGGAAATATGTATGAA  1440
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  TTCTCTCACCAAGCTGTAATGTGATTTTTTTTTCTACTCTGAATTGGAAATATGTATGAA  1440

Query  1441  TATACAGAGAAGTGCTTACAACTAATTTTTATTTACTTGTCACATTTTGGCAATAAATCC  1500
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  TATACAGAGAAGTGCTTACAACTAATTTTTATTTACTTGTCACATTTTGGCAATAAATCC  1500

Query  1501  CTCTTATTTCTAAATTCTAACTTGTTTATTTCAAAACTTTATATAATCACTGTTCAAAAG  1560
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  CTCTTATTTCTAAATTCTAACTTGTTTATTTCAAAACTTTATATAATCACTGTTCAAAAG  1560

Query  1561  GAAATATTTTCACCTACCAGAGTGCTTAAACACTGGCACCAGCCAAAGAATGTGGTTGTA  1620
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  GAAATATTTTCACCTACCAGAGTGCTTAAACACTGGCACCAGCCAAAGAATGTGGTTGTA  1620

Query  1621  GAGACCCAGAAGTCTTCAAGAACAGCCGACAAAAACATTCGAGTTGACCCCACCAAGTTG  1680
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  GAGACCCAGAAGTCTTCAAGAACAGCCGACAAAAACATTCGAGTTGACCCCACCAAGTTG  1680

Query  1681  TTGCCAC  1687
             |||||||
Sbjct  1681  TTGCCAC  1687
```

Seq219

Query = wt

Sbjct = risk

```
Query  1     CTACTGACATCAGCCAATGAGTCTAATACCTCTCAGTGGATATTGGAATCATAGCTCCCC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     CTACTGACATCAGCCAATGAGTCTAATACCTCTCAGTGGATATTGGAATCATAGCTCCCC  60

Query  61    AATTGCCTGTCAGCTACAGAAGCCAGATACCACTGAGAAATCATAATTATTATGCCCCAA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    AATTGCCTGTCAGCTACAGAAGCCAGATACCACTGAGAAATCATAATTATTATGCCCCAA  120

Query  121   GTGCCCTAGCTCACATTAGTTGTTAGTAGAGCTATCCTTAAAACAAAGACAGGAGTAGTT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GTGCCCTAGCTCACATTAGTTGTTAGTAGAGCTATCCTTAAAACAAAGACAGGAGTAGTT  180

Query  181   GAAATTATTATATATGTATTTCTATTACTGAATATTTGACGTAGTAGAGTTCTAGAGACC  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   GAAATTATTATATATGTATTTCTATTACTGAATATTTGACGTAGTAGAGTTCTAGAGACC  240

Query  241   AAAGCCACTAAAGCCAATTTAAATTTATTTATAGAAAAACGAGCCAAATTGTGGGTATAT  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   AAAGCCACTAAAGCCAATTTAAATTTATTTATAGAAAAACGAGCCAAATTGTGGGTATAT  300

Query  301   AAATGGACTGAATATATATTCTATGCTAAGGCTAACCATAAACCAAAGGAGTAATCCTCA  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   AAATGGACTGAATATATATTCTATGCTAAGGCTAACCATAAACCAAAGGAGTAATCCTCA  360

Query  361   GTCAAATAACTGGCATCAATTATTTTTCTCATATTTTTAGATATAATGATTATGAGATCT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   GTCAAATAACTGGCATCAATTATTTTTCTCATATTTTTAGATATAATGATTATGAGATCT  420

Query  421   ACCTTGTCATGGGTTTGCTTTGCTGTGAAATCAAGATCATAGCAAAGGTAAAGCAACTGG  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   ACCTTGTCATGGGTTTGCTTTGCTGTGAAATCAAGATCATAGCAAAGGTAAAGCAACTGG  480

Query  481   GTTCTACCTTAAGGTATCACTTGGCAACCAGTCTCCTACCCAGCTGGTAAACATTTCTGA  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   GTTCTACCTTAAGGTATCACTTGGCAACCAGTCTCCTACCCAGCTGGTAAACATTTCTGA  540

Query  541   TTACCAATGCAATACCAGTAGGCCTTAACTTTGGCTGGAAATTAGAGTCATTTAAGGATC  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   TTACCAATGCAATACCAGTAGGCCTTAACTTTGGCTGGAAATTAGAGTCATTTAAGGATC  600

Query  601   ATGTAAGACCGGTTAAATCAAACATCTAGCGGTGGGGTCTGAGCGTCAGTAAAATGAAAA  660
             || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  601   ATTTAAGACCGGTTAAATCAAACATCTAGCGGTGGGGTCTGAGCCTCAGTAAAATGAAAA  660

Query  661   AGCTTCCCAGCTGAGTCAGGGCCAGACCTGCGGAAGCCATTCTCATTCCATTATGGAAGG  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   AGCTTCCCAGCTGAGTCAGGGCCAGACCTGCGGAAGCCATTCTCATTCCATTATGGAAGG  720

Query  721   GCCAGGTGTTGCCAAGGCCACACACTTTGACATTCTAGTCCCCTGGATTCAATATATAAA  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   GCCAGGTGTTGCCAAGGCCACACACTTTGACATTCTAGTCCCCTGGATTCAATATATAAA  780

Query  781   ACTTAGTGATATGTCCCTCAGCTGCTATAGTGAATGGTATGTAGCTGTGCAAGGATTACA  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   ACTTAGTGATATGTCCCTCAGCTGCTATAGTGAATGGTATGTAGCTGTGCAAGGATTACA  840

Query  841   AACCTGAAACAAAATAATGATGCAAGACAATCCTTTTGAGCTTGTGAAACCCAACTCTAG  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   AACCTGAAACAAAATAATGATGCAAGACAATCCTTTTGAGCTTGTGAAACCCAACTCTAG  900

Query  901   CTTTCAGCTGGCTGAATGTGTTTCAAGCCAGGTATGGAGGTGCCAAAGGGTTAGCAGCCA  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   CTTTCAGCTGGCTGAATGTGTTTCAAGCCAGGTATGGAGGTGCCAAAGGGTTAGCAGCCA  960

Query  961   CAAGGAACAACTTTCCCTTCATATATTGGTTCAACACAGTAATCACAATGAATAGAACTC  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   CAAGGAACAACTTTCCCTTCATATATTGGTTCAACACAGTAATCACAATGAATAGAACTC  1020

Query  1021  AAGAGTAAAATGAATTCTGTCAATCAACCAGTCCTGTTGTCAGATAAATCATCATCACCT  1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  AAGAGTAAAATGAATTCTGTCAATCAACCAGTCCTGTTGTCAGATAAATCATCATCACCT  1080

Query  1081  TTGAGGACAATAGCTTGGCCTTAATGGATCTTGAAGGGAAGCCAGACCATGAATATAGAG  1140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  TTGAGGACAATAGCTTGGCCTTAATGGATCTTGAAGGGAAGCCAGACCATGAATATAGAG  1140

Query  1141  GCCCTGGAGGA  1151
              ||||||||||| 
Sbjct  1141  GCCCTGGAGGA  1151
```

Seq460

Query = wt

Sbjct = risk

```
Query  1    GCCTATCTTCAAATCTCTACTTCCCATAGATGCCTTCCACCCTCTATGATTAATAATCCT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GCCTATCTTCAAATCTCTACTTCCCATAGATGCCTTCCACCCTCTATGATTAATAATCCT  60
```

```
Query  61   TTTCTTCATTAATGAGAAGACAAATAATTATCACTGTTACCTGTTAAACATCAACAATGA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TTTCTTCATTAATGAGAAGACAAATAATTATCACTGTTACCTGTTAAACATCAACAATGA  120

Query  121  CTAGAGCTTGTAATATAATGTTTTATTATTCCCAACAACATTCCTATGGAAAGAATATTT  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  CTAGAGCTTGTAATATAATGTTTTATTATTCCCAACAACATTCCTATGGAAAGAATATTT  180

Query  181  TTCTAGTTTTATAGATAAAAACAAAAGGAGACTCAGGGAGGTAATGGTTTGCTCACTTTT  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  TTCTAGTTTTATAGATAAAAACAAAAGGAGACTCAGGGAGGTAATGGTTTGCTCACTTTT  240

Query  241  GTGATTTTGTGATGTGTCAGTGCTGGGACTGAATCCAAGTTGGGTGACAGCTGGGGCGAT  300
            |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||
Sbjct  241  GTGATTTTGTGATGTGTCAGTGCTGGGACTGAATCCAAGTTGAGTGACAGCTGGGGCGAT  300

Query  301  GCAGCAGAAGGCAGGTCTTGCTTTTTGGTAACACAGTGAGCTTGGGAAGTCTGCAACCAC  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  GCAGCAGAAGGCAGGTCTTGCTTTTTGGTAACACAGTGAGCTTGGGAAGTCTGCAACCAC  360

Query  361  ATGACTATGAATGAGGCTTCCTGGAACAGTAGAGGCTGAGTGAGGCCAGAGAGGAGCAGG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  ATGACTATGAATGAGGCTTCCTGGAACAGTAGAGGCTGAGTGAGGCCAGAGAGGAGCAGG  420

Query  421  AAGAGTTGCCAGGAGTGGGCTGCATGACAGTTTGTTAAGCATTTCACTGTTGACATCTGA  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  AAGAGTTGCCAGGAGTGGGCTGCATGACAGTTTGTTAAGCATTTCACTGTTGACATCTGA  480

Query  481  ACCCCAACTCTGTCCCTTACCAACCATCTCCTTAAATTGAATCTCAGTCTCTATGTCTGT  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  ACCCCAACTCTGTCCCTTACCAACCATCTCCTTAAATTGAATCTCAGTCTCTATGTCTGT  540

Query  541  AGGGTGGCAATAATATTTTCCCCTGTTTTTGTGAGGATTAAATGATGTGGCACATGGAAA  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  AGGGTGGCAATAATATTTTCCCCTGTTTTTGTGAGGATTAAATGATGTGGCACATGGAAA  600

Query  601  GTACTTGGCATAAGGTAAACTTTGGGACAGCCACTATAATGTGTCAAGTGGATACACTGA  660
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GTACTTGGCATAAGGTAAACTTTGGGACAGCCACTATAATGTGTCAAGTGGATACACTGA  660

Query  661  AATTTATTCCATCCAAGCGCTGTCTTCCTGATC  693
            ||||||||||||||||||||||||||||||||
Sbjct  661  AATTTATTCCATCCAAGCGCTGTCTTCCTGATC  693
```

Seq790

Query = wt

Sbjct = risk

```
Query  1     AAAGGTGTGGGGGTATATGGTATCCCTGATTTACTATAGAAGATGCAGAGAGTGAAGGGA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     AAAGGTGTGGGGGTATATGGTATCCCTGATTTACTATAGAAGATGCAGAGAGTGAAGGGA  60

Query  61    GATGAGGTGGGGAGGAGGGGCCCAGGTTCTGGTCCTACTTTTTTTTTTTTTTTCTAAAG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GATGAGGTGGGGAGGAGGGGCCCAGGTTCTGGTCCTACTTTTTTTTTTTTTTTCTAAAG  120

Query  121   AGATGGAGTCTTACCATGTTGGCCAGTCTAGGCTTGAACTCCTGGCCTCAAGAGGTGCTC  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   AGATGGAGTCTTACCATGTTGGCCAGTCTAGGCTTGAACTCCTGGCCTCAAGAGGTGCTC  180

Query  181   TCACCTCAGCCTCCCAAAGTGCTGGGATTATAGGCGTGAGCCACCGAGTTTAGCCCAGGT  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   TCACCTCAGCCTCCCAAAGTGCTGGGATTATAGGCGTGAGCCACCGAGTTTAGCCCAGGT  240

Query  241   TCTGTTTCTTGCTTAGTCACTTTCTGTTTGAACAAAATTGGAATTTCCTTTTTGGATCTG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   TCTGTTTCTTGCTTAGTCACTTTCTGTTTGAACAAAATTGGAATTTCCTTTTTGGATCTG  300

Query  301   TTTCTTTAATTGTAAATTGAATCGGACTAAAACCTTTCCAATTTTTTCACATGTGAAGAC  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   TTTCTTTAATTGTAAATTGAATCGGACTAAAACCTTTCCAATTTTTTCACATGTGAAGAC  360

Query  361   ATACACAAAAGTTTTATTGGAGGGTTGCACATGTGAAAGAAAAAGGGAGAAAGCAGGATT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   ATACACAAAAGTTTTATTGGAGGGTTGCACATGTGAAAGAAAAAGGGAGAAAGCAGGATT  420

Query  421   GAGCAGGGGGAGCCGTCAGATGGTAATGCAGATGTGATGAGATCTCTGCCGGACCAAAGA  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   GAGCAGGGGGAGCCGTCAGATGGTAATGCAGATGTGATGAGATCTCTGCCGGACCAAAGA  480

Query  481   GAAGATTCCTTTTTAAATGGTGACAAATTCATGGGCTTTCTCTGCCTCAAAACCTAGCAC  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   GAAGATTCCTTTTTAAATGGTGACAAATTCATGGGCTTTCTCTGCCTCAAAACCTAGCAC  540

Query  541   AGCTGTTATTTACTGAACAATTAGAGAGCTAAGCACTTTTTAGATACTATATAATTTAAT  600
             |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  541   AGCTGTTATTTACTGAACAATTAGAGAGCTAAGCACTTTTTAGATATTATATAATTTAAT  600

Query  601   TGCCGTATGAGGCACCCTTAGTTTTCAGACGAGAAACCACAGTTACAGGGAAGGCAAGTA  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  601   TGCCGTATGAGGCACCCTTAGTTTTCAGACGAGAAACCACAGTTACAGGGAAGGCAAGTA  660

Query  661   ACTTAGTCAATGTCAGATAACTAGGAAAAGGTTAGAGGGGCCCTGGACACAGGCCTGTGT  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   ACTTAGTCAATGTCAGATAACTAGGAAAAGGTTAGAGGGGCCCTGGACACAGGCCTGTGT  720

Query  721   GACTGAGAAGCTTGGGCACTTCACTGCTACATTTCATCTCTTCGCTATAAACATTTTAGC  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   GACTGAGAAGCTTGGGCACTTCACTGCTACATTTCATCTCTTCGCTATAAACATTTTAGC  780

Query  781   TTTTTGTGTTTGCTGACTGGCAACAATACATAGTGAAAGTTCTAATAATTTGTAATGCTT  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   TTTTTGTGTTTGCTGACTGGCAACAATACATAGTGAAAGTTCTAATAATTTGTAATGCTT  840

Query  841   TTGCATGTCTTTGTATTTTTCTTGGTTATCACATCACATCAAATTAAGATACTGATCAGC  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   TTGCATGTCTTTGTATTTTTCTTGGTTATCACATCACATCAAATTAAGATACTGATCAGC  900

Query  901   AGTGTGAGAGGTTATTTTTCCATGTCCTCTTCATTAGTGTTAGCTTGTGGATGGATTTGA  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   AGTGTGAGAGGTTATTTTTCCATGTCCTCTTCATTAGTGTTAGCTTGTGGATGGATTTGA  960

Query  961   GGCTCTCTGTGCTTTCCCCCCAGCAAAGTGAATACCAGACTTTCCTATTAAAAAAAGTAT  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GGCTCTCTGTGCTTTCCCCCCAGCAAAGTGAATACCAGACTTTCCTATTAAAAAAAGTAT  1020

Query  1021  TTTATTTTTCAGAGACAGGGTCTCATTCTGTCTCCCAGGCTGGAGTGCAGTGGCACAATC  1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  TTTATTTTTCAGAGACAGGGTCTCATTCTGTCTCCCAGGCTGGAGTGCAGTGGCACAATC  1080

Query  1081  ATAGCCCACTGCAGCCTCCAACTCTTGGGTTCAAATGATCCTCCTGCCTCAGCCTCTCTT  1140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  ATAGCCCACTGCAGCCTCCAACTCTTGGGTTCAAATGATCCTCCTGCCTCAGCCTCTCTT  1140

Query  1141  AAGCAGTGCCTTTCCCCATTCTCATGGGACTTTCCAATCCATGAGATACTTTGCTGCAGG  1200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  AAGCAGTGCCTTTCCCCATTCTCATGGGACTTTCCAATCCATGAGATACTTTGCTGCAGG  1200

Query  1201  GAAGCCCTGTCTGTCCAGGCCTGTGTAATAGACGACTTCACATGGTCCTGTGTTGTTGTT  1260
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  GAAGCCCTGTCTGTCCAGGCCTGTGTAATAGACGACTTCACATGGTCCTGTGTTGTTGTT  1260

Query  1261  TGCCTTCTGTGTGGCTAAGTTTCCATGACCTGGTG  1295
              ||||||||||||||||||||||||||||||||||
Sbjct  1261  TGCCTTCTGTGTGGCTAAGTTTCCATGACCTGGTG  1295
```

wt allele

risk variant

common variant

CHAPTER 3

THE SEQ5215 CODING ENHANCER IN KCNJ11 GENE

1. Introduction

Type 2 diabetes (T2D) is a multifactorial and polygenic disease that affects more than 400 millions of people, worldwide[1,2] and it is partially characterized by endocrine pancreas dysfunction leading to hyperglycemia[3]. The molecular mechanisms involved in endocrine pancreas dysfunction are poorly understood, however, recent studies have shown that a vast part of the studied T2D-associated single-nucleotide polymorphisms (SNPs), identified by genome-wide association studies (GWAS), are present within genes demonstrated to affect β-cell function[4]. The majority of T2D-associated-SNPs are located in regions that overlap with epigenetic marks specifically associated with enhancers[5,6], and some of them might coincide with transcription factor binding sites (TFBS) important for a proper endocrine pancreas function, resulting in different transcriptional outputs[6–8]. Enhancers can be predicted by profiling chromatin accessibility and histone modifications[9–11], such as H3K27ac and H3K4me1. A recent study to explore T2D genetics was performed to integrate all the maps of epigenetic marks, chromatin accessibility and TFBS, offering a detailed dataset of all putative enhancers, in human pancreatic islets[6,12,13]. While enhancers are commonly located in non-coding regions, in vivo reporter assays have shown that sequences that overlap with coding regions have also the capability to show enhancer activity [7,14,15], demonstrating that DNA sequences can have dual – coding and regulatory-functions, even in different target genes, [14]. Due to the accumulation of functions of these, mostly underexplored, coding enhancers mutations in such sequences can result in multiple outcomes[16] and can explain contradicting results of certain cases, such as the *SLC30A8* gene case. [7]. Here, the common T2D-associated SNP rs13266634, located in the *SLC30A8* gene, is commonly associated with a decrease in β-cell function, due to a reduced protein function (less zinc transporter activity). However, in recent studies, the loss-of-function mutations in *SLC30A8* are associated with a protective effect against T2D[17]. Importantly, the aminoacid where the SNP is located is predicted to not affect the protein structure, specifically in dimerization or in appearing an active zinc binding site[18]. The different outcomes observed in these studies and the fact that the resulted protein does not alter significantly, often overlook the potential dysregulation of the gene as a possibility and the hypothesis that the rs13266634 may be modulating an enhancer. It is important to consider the regulatory role of the coding sequence in addition to its protein-coding function when investigating such cases. Considering the regulatory functions of the coding sequence alongside its protein-coding role could provide insights into the underlying mechanisms of the observed phenotypes, potentially elucidating the true nature of the outcomes observed in these studies.

Massively parallel reporter assays (MPRAs) can be a useful tool to explore the complexity of the enhancer activity of coding enhancers. Using this large-scale technique, it has been demonstrated that single-nucleotide variations (SNVs) in liver enhancers leading to the aminoacid changes and SNVs that maintain the aminoacid unaltered might have similar impact in the regulatory output of these enhancers [19]. These results further support that SNVs in sequences with overlapping coding and regulatory functions might affect both these functions. In this same study, it was observed that most of SNVs in coding enhancers overlap with liver-associated TFBS, highlighting the regulatory role of those coding enhancers[19].

One example of a relevant coding SNP is the rs5215, located in *KCNJ11* (potassium inwardly rectifying channel subfamily J member 11) gene, associated with an increased risk of hyperglycemia and T2D, in several GWAS[20–24]. *KCNJ11* encodes for a subunit (KIR6.2) of the ATP-sensitive potassium (KATP) channel, in β-cells membrane[25]. KATP channels are important for glucose homeostasis, by controlling the electrical activity of the cell membrane. When the glucose levels increase, the KATP channels close, allowing the insulin release into the bloodstream[26,27](Figure 1). Along with KIR6.2, the KATP channels are composed by a second subunit, SUR1 (sulphonylurea receptor 1), which is encoded by the *ABCC8 (* ATP Binding Cassette Subfamily C Member 8) gene[26], a neighboring gene locate in the same genomic landscape of *KCNJ11*.

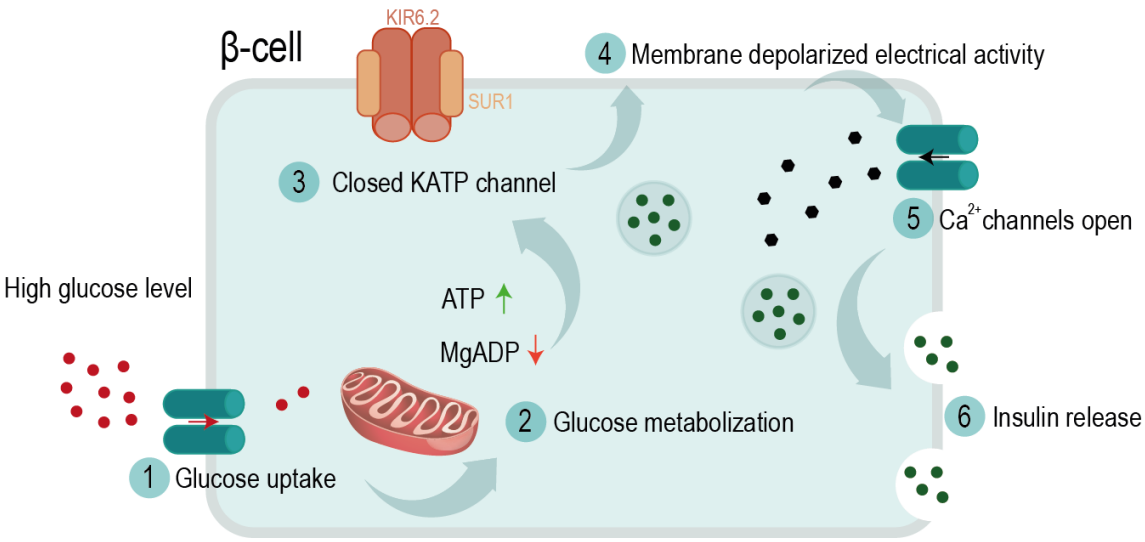


Figure 1 – The role of KATP channel in insulin secretion upon high glucose levels in the bloodstream. KATP channels are inhibited by ATP and activated by MgADP. Adapted from [21])

Mutations occurring in either *KCNJ11* or *ABCC8* can result in hypoglycemia or hyperglycemia, depending on the nature of these genetic alterations. Until now, it has been reported more than 900 mutations located in these two genes that cause congenital

hyperinsulinism and monogenic diabetes[21]. Gain-of-function mutations are linked to diabetes traits while loss-of-function mutations are related with hyperinsulinism[21,27,28].

GWAS have identified several SNPs located in *KCNJ11* and *ABCC8* genes that are associated with T2D. However, how these SNPs might contribute to T2D is not well established, mainly due to T2D be a polygenic disease. Nevertheless, even small effects in key regulators of glucose metabolism, such as KATP channels can result in severe outcomes and should be explored[21]. Rs5215 is located in the only exon of *KCNJ11* gene, it is a coding SNP (T – non risk allele to C – risk allele) and results in an aminoacid substitution from isoleucine to valine[20]. The resultant impact in the protein function due to this SNP has yet to be investigated.

In this chapter, and as an update of our previous report[7], we expanded the number of putative endocrine pancreatic enhancers and tested 33 pairs of T2D risk (containing the T2D associated variant) and non-risk sequences (not containing the T2D associated variant), to understand the impact of T2D-associated variants in the activity of enhancers. Then we focused our attention in seq5215, one of the validated enhancers that overlaps with the coding SNP rs5215. Regarding seq5215, we aim first to address the enhancer activity of seq5215, using in vivo (zebrafish) and in vitro (Min6 cell line) reporter assays. Secondly, understand the impact of the rs5215 SNP in the regulatory output of seq5215 enhancer. Thirdly, decouple the regulatory and coding functions, by studying a seq5215 enhancer sequence that disturbs the seq5215 regulatory output, but at the same time not disrupting the coding sequence of *KCNJ11*, by introduction of a coding synonymous single nucleotide modification. Lastly, we aim to dissect the seq5215, through a CRISPR/Cas9 based large-scale technique (MPRA), in order to predict new alleles that can disrupt its regulatory functions, using Min6 cell line.

2. Research design and methods

2.1. Zebrafish transgenic lines and husbandry

sst:mCherry transgenic reporter line was used for in vivo mosaic transgenesis assays[7]. Zebrafish were handled according to European animal welfare regulations and standard protocols. 48 hours post-fertilization (hpf) embryos were cultured at 28°C in Petri dishes containing E3 medium (sodium chloride, potassium chloride, calcium chloride, magnesium sulfate buffered and methylene blue, as fungicide) supplemented with 1-phenyl-2-thiourea to delay pigmentation formation[29].

2.2. Putative enhancer identification

As an update from the previous selection[7], the putative enhancer identification was based on GWAS data that identified 163 SNPs associated with T2D or glycemic traits (P-value < 5×10^{-8})[30],SNPs in high linkage disequilibrium with lead SNPs, and a dataset that integrated chromatin immunoprecipitation (ChIP-seq) and RNA-seq profiles in human pancreatic islets, by Parker and co-workers[31] (Figure 2A). The latter dataset includes the prediction of active/poised/weak promoters, strong/weak enhancers, insulators and repressed/transcribed regions[31] and can be consulted in the Islet Regulome Browser[12]. The two datasets were bioinformatically crossed, using BedTools[32], resulting in a list of 80 sequences that overlap with T2D-associated SNPs and Parker et al dataset[31] (Figure 2A). From the 80 sequences, 51 were predicted to be strong or weak enhancers (Figure 2A). The final 33 human putative endocrine pancreatic enhancers are described in supplementary table 1. Next, these sequences were tested, using in vivo reporter assays, as described previously[7].

Recently, an updated dataset was included in Islet Regulome browser, by Miguel-Escalada and co-workers [13]. As an additional and confirmatory result, we expanded the SNPs dataset for 1330 T2D-associated SNPs from DisGeNET[33](P-value of $<1\times10^{-6}$). Then, the SNPs dataset was bioinformatically crossed with the defined chromatin classes by Miguel-Escalada and co-workers[13] (Figure 2B). The T2D-associated SNPs included SNPs associated with the DisGeNET categories: Diabetes Mellitus non-insulin dependent, Permanent Neonatal, Late-Onset diabetes, Type II Diabetes and Maturity-Onset Diabetes of the Young (MODY)[33]. The chromatin classes dataset was defined by the presence of ChIP-seq and ATAC-seq profiles in human endocrine islets[13]. From the 33 ready to test sequences, 14 overlap with the category of active enhancers from Miguel-Escalada and colleagues dataset, including the seq5215 and the overlapping SNP rs5215 (Figure 2B).

2.3. In vivo mosaic transgenesis assays

The human putative enhancers were amplified by PCR from human genomic DNA using the primers described in supplementary table 2. Next, the sequences were cloned in TOPO vector (pCR8/GW/TOPO TA Cloning KIT, Invitrogen) and recombined, through Gateway system to Z48 enhancer transgenesis reporter vector[34] or to promoter test vector (PTV)[34]. Both systems contained GFP as reporter gene, having the enhancer vector a minimal promoter and the promoter vector only the reporter gene. Risk variants were inserted by site-directed mutagenesis using primers containing the risk variant (Supplementary table 2). Zebrafish transgenesis was conducted using Tol2 transposon system[34,35]. One-cell embryos from the transgenic line *sst:mCherry* were microinjected with

3nL containing 25ng/ µL *Tol2* transposase mRNA and 25 ng/µL phenol/chloroform-purified reporter vector. The microinjected embryos were selected by the presence of GFP in midbrain (internal transgenesis control), at 48hpf.

2.4. Immunohistochemistry

The selected embryos were dechorionated and fixed in 4% formaldehyde in PBS (PBS1×) overnight at 4°C and then washed in PBS-T (0.5% Triton X-100 in PBS1×), followed by a permeabilization step with PBS-T for 30 minutes. Embryos were then incubated with DAPI (1:100, BioRad) diluted in PBS-T overnight, at 4ºC, washed and stored in 50% glycerol in PBS1× (140 mM NaCl. 2.7 mM KCl. 10 mM Na_2HPO). Confocal imaging was performed using Leica SP5II confocal microscope.

2.5. Cell culture

Min6 cells were cultured in Modified Eagle Medium (DMEM, Gibco) with GlutaMAX, supplemented with 15% fetal bovine serum (FBS, Gibco) and 1x Antibiotic-Antimycotic (Gibco). Cells were maintained in a 5% CO_2 humidified chamber at 37ºC.

2.6. Luciferase reporter assays

The sequences to be tested were cloned in the pGL4.23-GW[luc2/minP] vector (Addgene #603232), using the previous gateway cloning system, and co-transfected with pNL1.1PGK[Nluc/PGK] (Promega #N1441) in Min6 cell line, using Metafectene Pro (Biontex) as transfection reagent. As negative control (NC), it was used a sequence (seq117), where it was validated as negative, not showing enhancer activity in vivo in the zebrafish pancreas and in vitro in Min6 cells reporter assays. Cells were transfected in a 6-well plate with 1,5µg of pGL4.23-GW and 0.6 fmol of pNL1.1PGK[Nluc/PGK] per well, accordingly with manufactures instructions. Luciferase activity was assessed 48 hours post-transfection using the Nano-Glo-Dual-Luciferase reporter assay system (Promega), in Synergy 2 Multi-Mode microplate reader (BioTek). The assessment of enhancer activity relies on the ratios between luc2 and Nluc levels.

2.7. Statistical analysis

For the in vivo mosaic transgenesis assays, statistical analysis was performed by using the chi-square test with Fisher correction and unpaired t-test, applying a significance level of $P \leq 0.05$.

For the in vitro luciferase assays, two-sided t-test was used. At least 3 independent replicates of the transfection were performed.

All the statistical tests were conducted in GraphPad Prism 8.

3. Results

3.1. Testing the functionality of putative human endocrine enhancers and the impact of T2D SNPs in their regulatory output

Thirty-three (33) sequences were selected based on the overlapping of SNP associated to T2D and glycemic traits, along with sequences predicted to belong to strong and weak enhancers classes, in pancreatic endocrine cells, defined by Parker and co-workers, in 2013 (Figure 2A and supplementary table 1).

Ten (10) of the 33 sequences coincide with 10 sequences tested previously[7]. The remaining 23 that do not contain the T2D-associated risk allele (non-risk) were cloned in Z48 transgenesis vector and in vivo mosaic transgenesis assays was performed, using zebrafish as organism model[7], as previously done in chapter 2 . Out of the new 23 tested sequences, 7 where able to drive GFP expression in the endocrine pancreatic domain with higher frequency (seq704, seq180, seq479, seq770, seq142, seq116 and seq150) than the empty Z48 vector (negative control; NC), being, therefore, endocrine pancreatic enhancers (Figure 3).

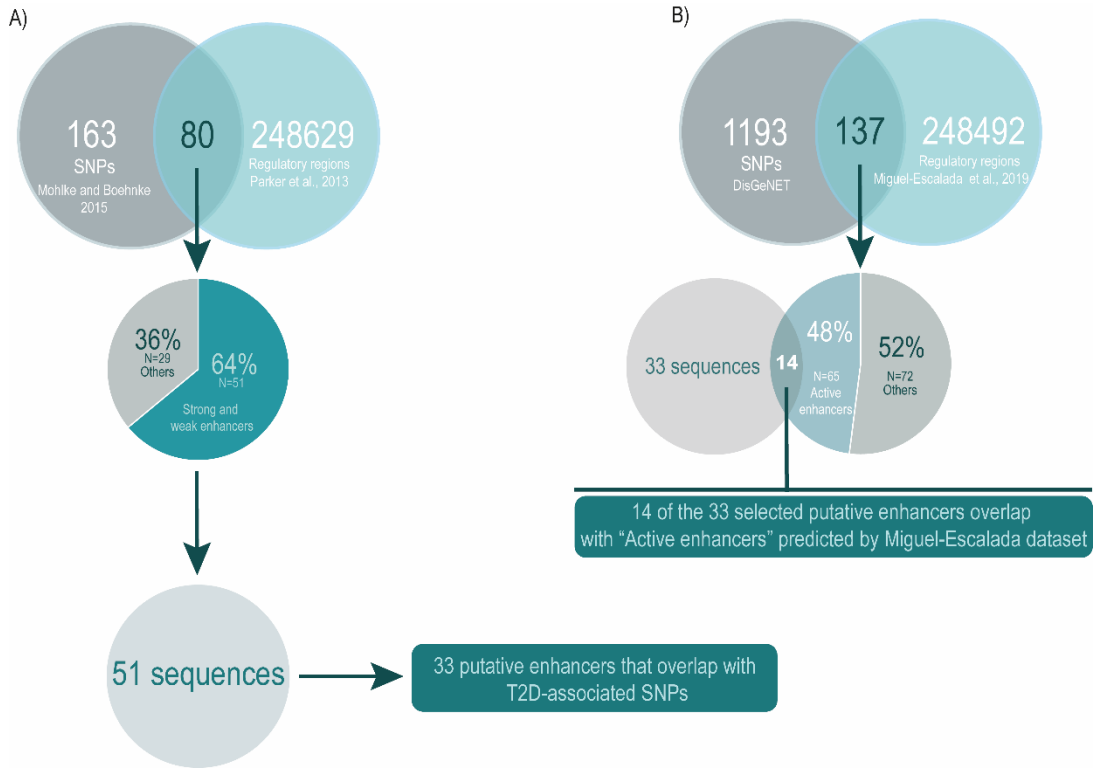


Figure 2 – A) Representative diagram of the selection of the 33 putative enhancers to test for enhancer activity, including 163 SNPs, selected based on Mohlke and Boehnke (2015) and the predicted regulatory regions define by Parker et al. (2013). B) Representative diagram of the analysis including the 1193 SNPs from DisGeNET database and Miguel-Escalada et al (2019) predicted regulatory regions dataset. From the 33 putative enhancers previously selected, 14 overlapped with the "Active enhancers" category from Miguel-Escalada (2019) dataset.

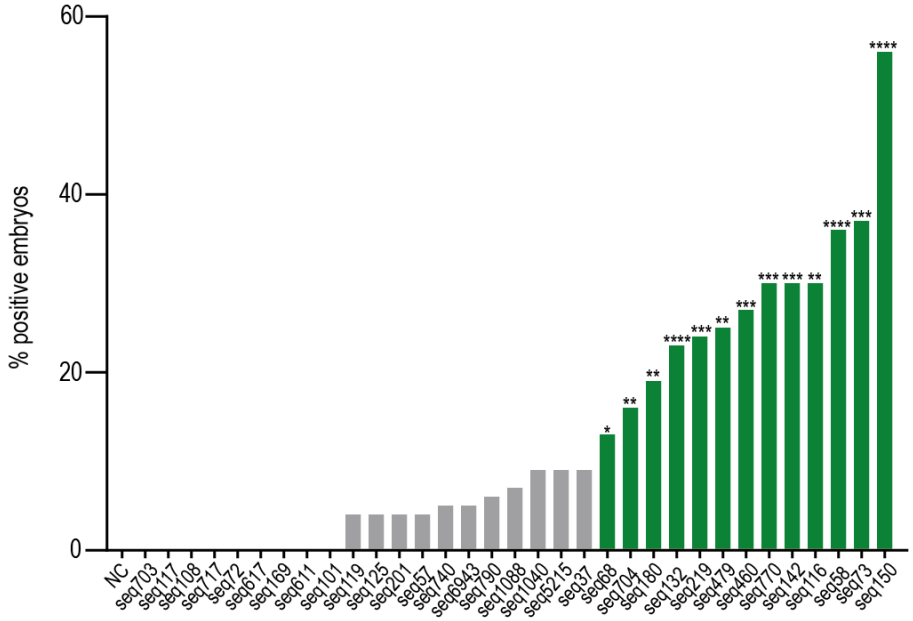


Figure 3- Graph representing the percentage of embryos with GFP expression in pancreatic endocrine domain at 48hpf for each non-risk sequence analyzed. Chi-*square test and Fisher`s exact test comparing with NC.*

To study the impact of T2D-associated SNPs within enhancer regions, we cloned the new 23 sequences in Z48 transgenesis vector, harboring the T2D-associated risk allele, introduced by site directed mutagenesis (SDM). Subsequently, we performed in vivo mosaic transgenesis assays using the same strategy as in the previous experiment. Considering a difference of >10% of enhancer activity, between the non-risk and risk sequence, for 7 pairs, the T2D risk variant has shown a decreased enhancer activity, while for 9 pairs, the T2D risk variants have demonstrated an increased enhancer activity, including the seq611 (Figure 4, 5 and supplementary table 3). These results suggests that T2D-associated SNPs can modulate enhancer activity.

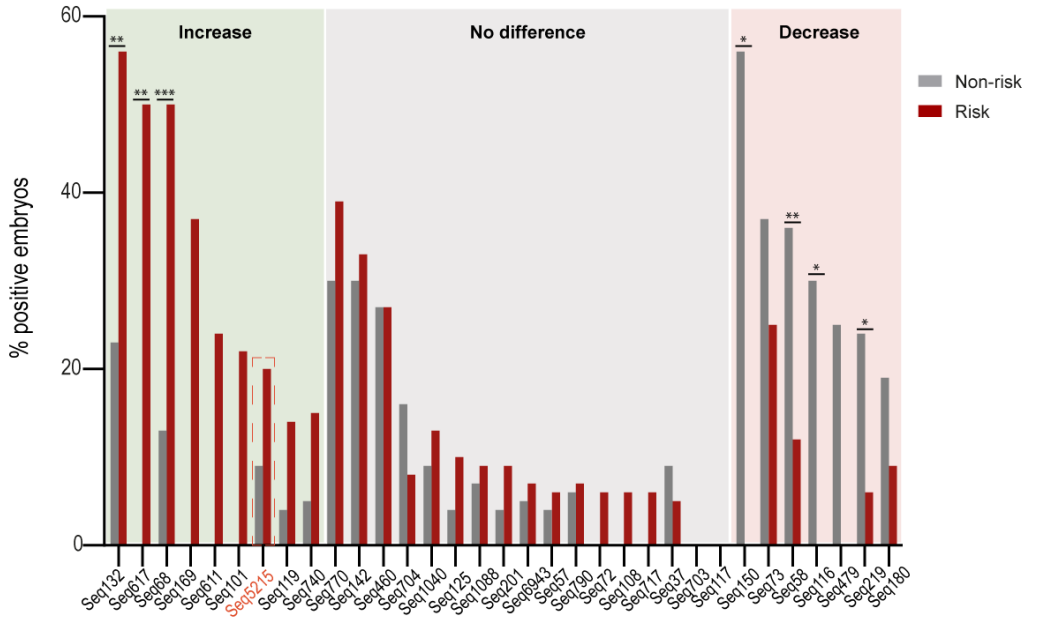


Figure 4 - Graph showing the total percentage of positive embryos in non-risk and risk alleles for each of the 33 sequences analyzed. In 33 sequences, the risk version incremented the enhancer activity for 9, 17 didn't show any difference between non-risk and risk sequence and 7 decrease the enhancer activity. Chi-square test and Fisher`s exact test comparing risk and non-risk alleles.

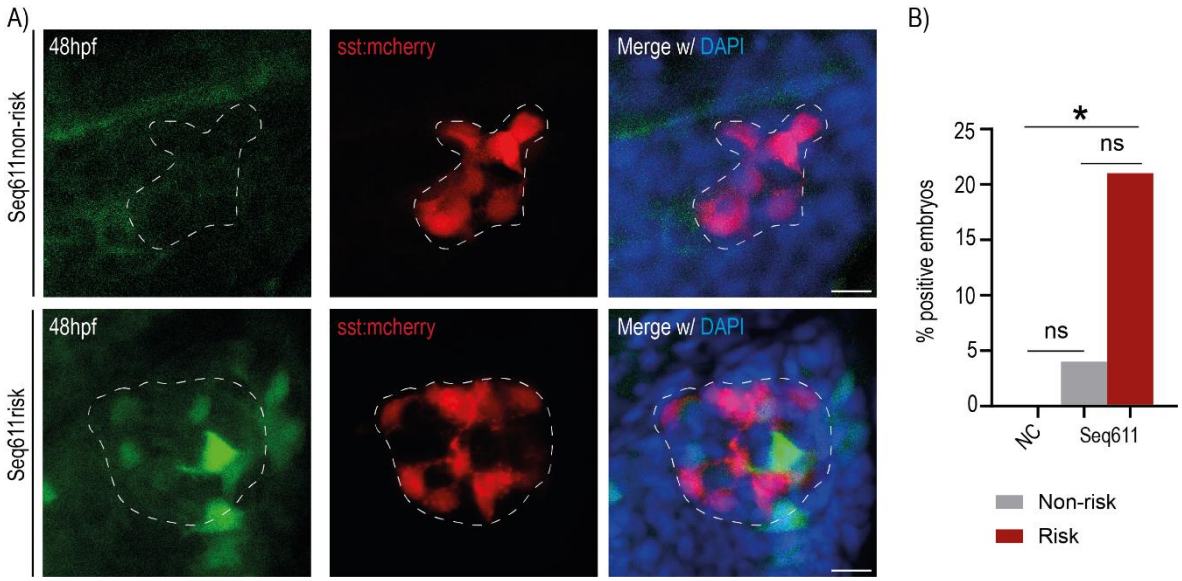


Figure 5- A) Representative confocal image with a sst:mCherry zebrafish embryo (red) injected with the Z48 enhancer reporter vector containing the seq611non-risk sequence (above), showing absent GFP+ cells (green) in the endocrine pancreas, contrasting with present GFP+ cells in embryos injected with seq611risk (below). B) Graph representing the percentage of embryos with GFP expression in pancreatic endocrine domain at 48hpf for seq611risk and seq611non-risk. Chi-square test and Fisher`s exact test.

3.2. The human seq5215

The seq5215 partially overlaps with the *KCNJ11* exon, being the overlapping T2D-associated SNP, rs5215, a coding SNP. This sequence intersects with the chromatin classes "strong" and "active enhancer", defined by Parker et al (2013) and Miguel-Escalada (2019), respectively, which means that the same sequence emerged in two distinct and different analysis (Figure 2B). These chromatin classes were defined by the presence of ChIP-seq peaks for H3K27ac histone [31] and ATAC-seq[13] profiles, as demonstrated in figure 6, in human endocrine islets. To notice, within the vicinity of the *KCNJ11* genomic landscape, the *ABCC8* gene is present, suggesting that enhancers within this landscape might co-regulate both genes (Figure 6).

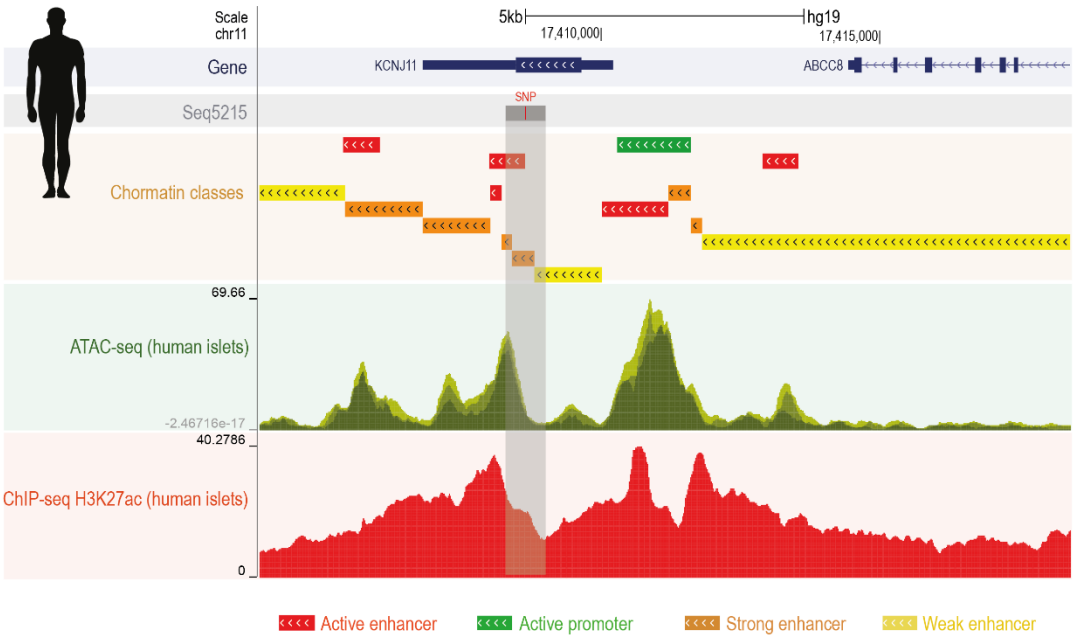


Figure 6 - Genomic landscapes of the human KCNJ11 gene, showing profiles for ATAC-seq (green), HEK27ac (red), and the chromatin classes defined and predicted by Miguel-Escalada et al (2019) and Parker et al (2013), in human adult islets. The enhancer seq5215 is highlighted in grey and the overlapping SNP (rs5215) in red.

3.2.1. Seq5215 in vivo and in vitro reporter assays

Performing in vivo mosaic transgenesis assay, one of the sequences that showed an increase in enhancer activity, in the presence of the T2D-associated SNP, was seq5215. Although the difference between non-risk and risk sequence is not statistically significant, the non-risk sequence it is not an endocrine pancreatic enhancer, while the risk sequence is an endocrine pancreatic enhancer (seq5215non-risk 9%, n= 56 and seq5215 20%, n=25) (Figure 7). Heterozygous activating mutations in *KCNJ11* were found in 4 out of 10 patients with permanent neonatal diabetes[28] and, in mice, overexpressing a mutant KATP channel led to the development of diabetes and, consequently, death[36]. These observations suggest

that the observed effect within the presence of the rs5215 SNP might be linked to an increased risk for T2D due to an increased activity of an endocrine enhancer located in the coding region of *KCNJ11*.

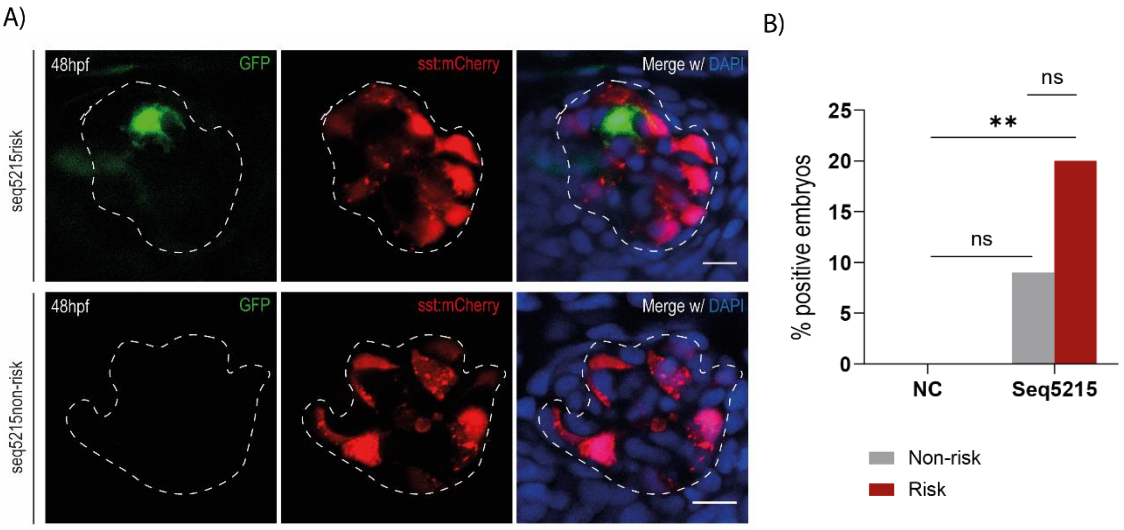


Figure 7- A) Representative confocal image with a sst:mCherry zebrafish embryo (red) injected with the Z48 enhancer reporter vector containing the seq5215risk sequence (above), showing GFP+ cells (green) in the endocrine pancreas, contrasting absent in embryos injected with seq5215non-risk (below). B) Graph representing the percentage of embryos with GFP expression in pancreatic endocrine domain at 48hpf for seq5215risk and seq5215non-risk. Chi-square test and Fisher`s exact test.

Performing luciferase assays in vitro, using Min6 cell line, the seq5215 non-risk had a significant reduction of the enhancer activity, when comparing with seq5215 risk enhancer (Figure 8). These results align with the findings from in vivo experiments (Figure 7).

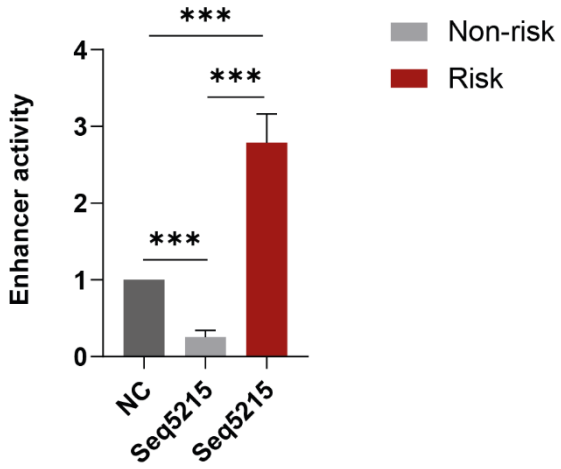


Figure 8- Luciferase expression assay mediated by seq5215non-risk and seq5215risk sequences. T-test; 3 replicates

Similar to enhancers, promoters can also be predicted by profiling histone modifications. Specifically, H3K4me3 serves as a distinctive marker for promoters. The H3K4me3 profile in *KCNJ11* gene reveals 2 distinct peaks, as shown in figure 9.

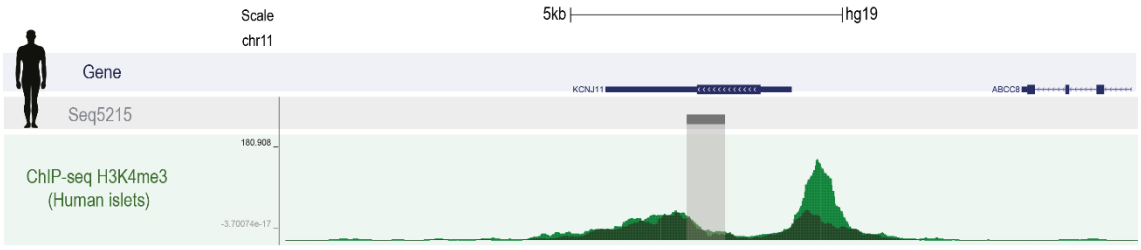


Figure 9- Genomic landscapes of the human KCNJ11 gene, showing profiles for H3K4me3 (green), in human adult islets. The enhancer seq5215 is highlighted in grey.

One peak partially overlaps with seq5215, suggesting that this sequence can have the potential to behave as a promoter. To test this hypothesis, we cloned, by gateway system, the seq5215 in a PTV that contained GFP as reporter gene and performed in vivo mosaic transgenesis assays in *sst:*mCherry transgenic line. We cloned the seq5215 in two possible directions regarding the GFP reporter gene (GW1GW2 and GW2GW1). As an internal control of transgenesis, we used *mylz*:mCherry vector, that drives mCherry expression specifically in muscle fibers of the zebrafish (Figure 10 - left). As a positive control, we mobilized a Tol2 transposon containing the insulin promoter upstream of GFP and as a negative control, we microinjected the empty PTV. If seq5215 functions as an endocrine pancreatic promoter, we expect to observe GFP expression in endocrine pancreas, at 48hpf.

The injected embryos with the empty PTV did not show GFP expression (0%, n=16), while in the insulin promoter microinjected embryos, 65% of the embryos showed GFP expression, in the endocrine pancreas (n=20) (Figure 10). The seq5215 did not show promoter activity in both directions, GW1GW2 (0%, n=14) and GW2GW1 (0%, n=14) (Figure 10). These results allow to conclude that the seq5215 is not sufficient to drive GFP expression in endocrine pancreatic cells as a promoter.

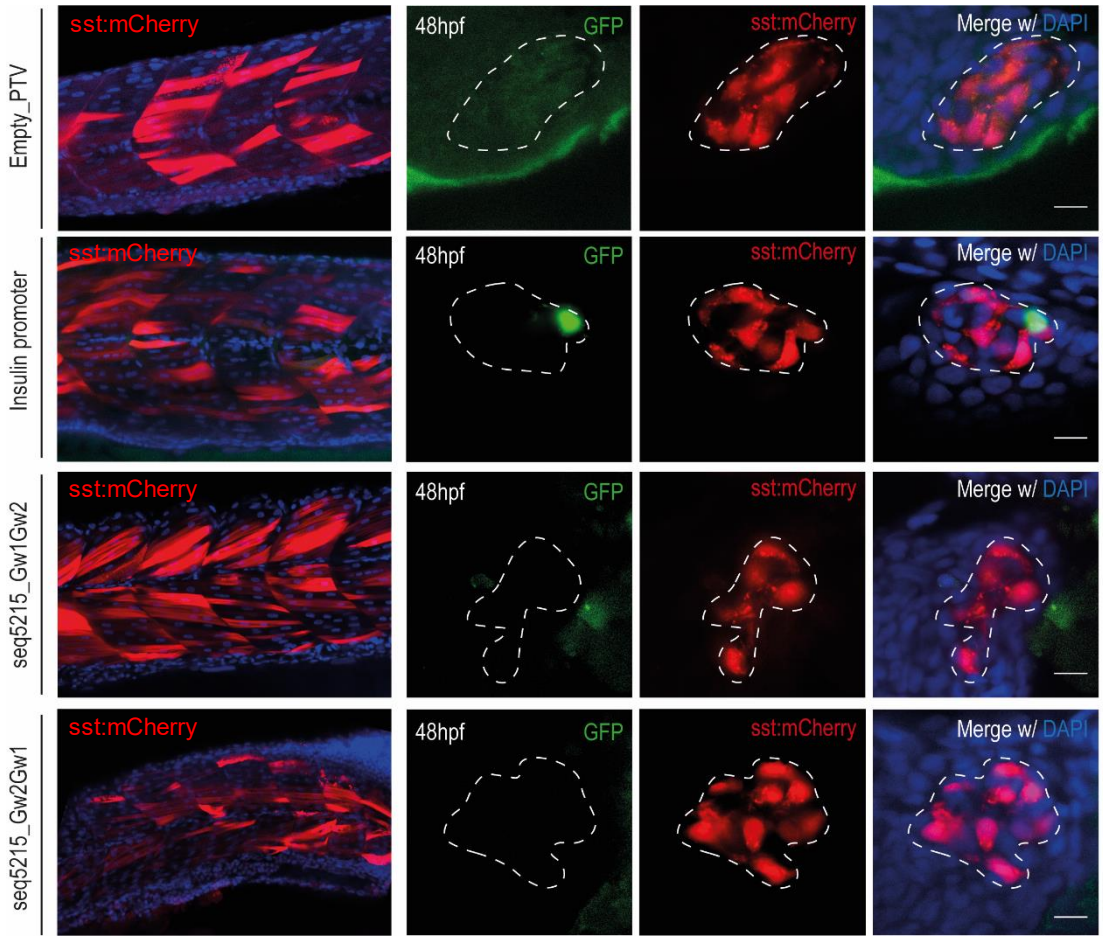


Figure 10 - Representative confocal images with a sst:mCherry zebrafish embryo (red) co-injected with the mylz:mcherry (left images) and empty PTV reporter vector and containing the seq5215 sequence, in both directions (Gw1Gw2 and Gw2Gw1), showing no GFP+ cells (green) in the endocrine pancreas.

3.3. The seq5215mm is an ortholog enhancer of seq5215 in mice

To better explore the seq5215 enhancer and the rs5215 coding SNP impact, we asked if we could find a sequence that could behave as seq5215 in the *Kcnj11* orthologue mouse gene. Similar to humans, the *Kcnj11* landscape includes the *Abcc8* gene, further supporting the hypothesis that shared regulatory mechanisms might control the expression of these two genes (Figure 11). We predicted the seq5215mm to be the ortholog putative enhancer of seq5215, taking into account the sequence conservation and the epigenetic marks associated to enhancer activity, such as DNase hypersensitive, histone modifications and CTCF binding data[37] (Figure 11), found in the ENCODE dataset and overlapping with seq5215mm (Figure 11).

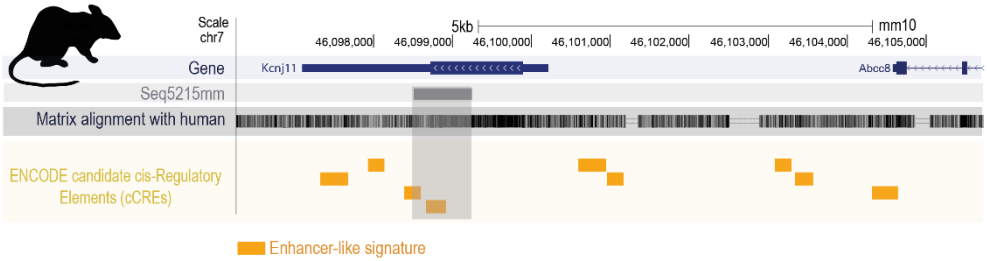


Figure 11 - Genomic landscape of the mouse Kcnj11 gene, showing the sequence conservation between human and mice in grey and the candidate cis-regulatory elements defined by ENCODE project in yellow. The enhancer seq5215mm is highlighted in grey.

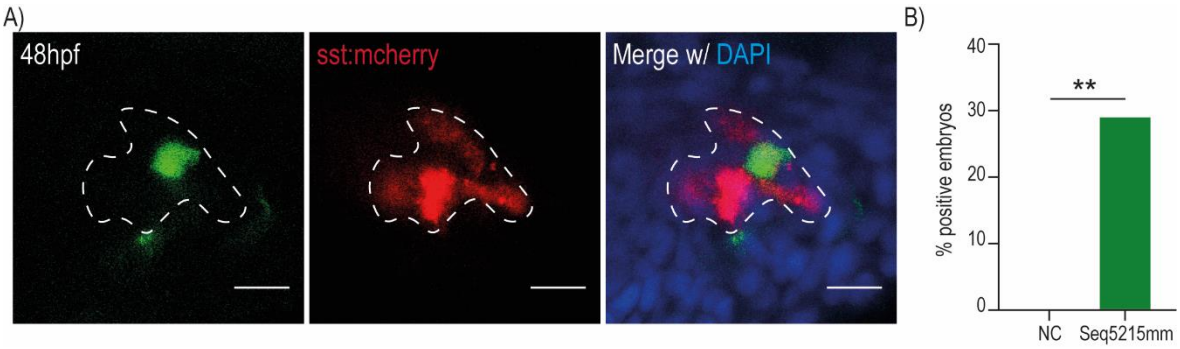


Figure 12 - A) Representative confocal image with a sst:mCherry zebrafish embryo (red) injected with the Z48 enhancer reporter vector containing the seq5215mm sequence, showing GFP+ cells (green) in the endocrine pancreas. B) Graph representing the percentage of embryos with GFP expression in pancreatic endocrine domain at 48hpf for seq5215mm. Chi-square test and Fisher`s exact test.

By performing in vivo mosaic transgenesis assays, we observed that seq5215mm was able to express GFP in endocrine pancreas (29%, n=14), being therefore an endocrine pancreatic enhancer (Figure 12), suggesting a similar function as seq5215 in human. However, there is still a crucial issue that has not been thoroughly investigated: the decoupling of the dual function, coding and regulatory, of the enhancer.

3.3.1. Studying the impact of a silent mutation within an enhancer

To address the impact of a silent mutation within the seq5215mm enhancer, we generated 2 single-point mutations, sequentially, in the seq5215mm, precisely within the same aminoacid where the rs5215 coding SNP is located in seq5215 (Figure 13A), but leaving the resultant amonoacid, isoleucine (Iso), unaltered, resulting in the same coding protein. Next, we performed luciferase enhancer assays for the 3 sequence versions (ATT – reference genome, ATA and ATC- Figure 13A), using Min6 cell line (DNA1). The results showed that the enhancer activity was significantly incremented in the ATC sequence version, comparing with the reference sequence, ATT (Figure 13B). Although not significant,

the enhancer activity was decreased in ATA sequence version, when compared with the reference sequence, ATT (Figure13B). These experiments demonstrate that a single nucleotide substitution, while leaving the aminoacid unaffected, modulates enhancer activity, showing that nucleotide alterations in sequences with overlapping coding and regulatory functions might affect only one of these functions without potentially affecting the other. As a proof of principle, we next mutated the enhancer of DNA1, in order to have other 3 additional single nucleotide mutations. The enhancer activity was, again, modulated in a way that there were not significant differences between the 3 versions (Figure 13C), highlighting the importance of the genomic context, where an impact of a SNP can be modulated by the presence of other SNPs.

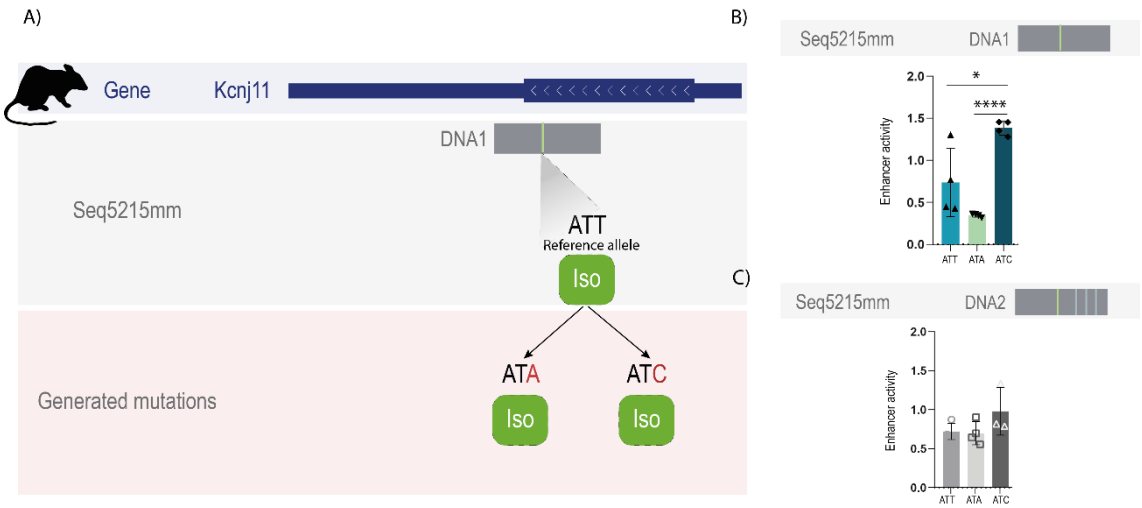


Figure 13 - A) Representative diagram of seq5215mm enhancer in Kcnj11 and the 2 generated single nucleotide mutations (represented in red) that resulted in more 2 sequence versions – ATA and ATC- maintaining the aminoacid unaltered (Iso). B) Representative graph of the luciferase assays performed with the 3 sequence versions (ATT, ATA and ATC), with the DNA1 background. C) Representative graph of the luciferase assays performed with the 3 sequence versions (ATT, ATA and ATC), with the DNA2 background. T-test; 3 replicates.

4. Discussion and future perspectives

In this chapter, we emphasized the importance and viability of using zebrafish as an in vivo reporter model in enhancer assays, specifically for the molecular study of T2D. Until now, we were able to validate, at least, 26 enhancers in more than 75 tested sequences, which represents, approximately, to a validation of 40% of efficiency in the prediction of endocrine pancreatic enhancers. However, in vivo reporter assays have its own limitations, being considered slow, expensive and usually not compatible with high throughput strategies. Additionally, there are several other relevant sequences to test, that due to the

limitations, we could not test. MPRAs have emerged as one of the most efficient tools to overcome those problems, due to its capability of test thousands of putative enhancers simultaneously, using next-generation sequencing as readout [38]. This way, we can identify and explore only the most striking enhancers in vivo, saving financial resources and time. Additionally, when studying SNPs, using MPRAs, we can easily quantify the differences between non-risk and risk sequences. Currently, we are implementing this strategy in the lab to identify endocrine pancreatic enhancers.

Here, we found an enhancer, seq5215, that overlaps with the *KCNJ11* exon. *KCNJ11* encodes for a subunit of the KATPs channels, present in β-cells membrane, that regulate insulin secretion. Therefore, mutations in this gene have been strongly linked to T2D-related phenotypes, being gain-of-function related with hyperglycemia and low levels of insulin secretion. Ou results demonstrate that the presence of the coding SNP, rs5215, results in an increment of the enhancer activity, both in vivo and in vitro, suggesting that this SNP might contribute for T2D predisposition, by enhancing the activity of the overlapping enhancer. To better understand how this SNP can increment the enhancer activity and at the same time, being part of the coding-protein region, we were able to find the ortholog enhancer in mice, seq5215mm. We demonstrated that a single nucleotide substitution, while leaving the aminoacid unaffected, modulates enhancer activity, a mechanism poorly explored but already described[14,19]. These results show that nucleotide alterations might impact in one function, maintained the other unaltered, further highlighting the importance of accessing enhancers and coding sequences function at different levels. Indeed, recent studies shown that mutations in coding enhancers can result in several and complex phenotypes[39] and might explain cases that are not explained only by the coding mutations itself.

Overall, these findings can give new insights from the underexplored field of dual function coding enhancers, where a coding SNP has the potential to modulate the activity of an enhancer and the possible impact on human disease.

As on-going experiments, we are performing, using CRISPR-Cas9 system, the deletion of the seq5215mm, in Min6 cell line, in order to understand to which extent, seq5215 is controlling not only *Kcnj11,* but the neighboring gene *Abcc8*. If so, those results will give us new insights regarding the regulatory landscape of these crucial genes in T2D development. We are also using MPRAs, first, validating the seq5215 and secondly, performing a large mutagenesis screen to explore the impact of new mutations in the enhancer function, including coding mutations. This way, we will be able to create a full map of all the possible mutations that can disrupt the enhancer activity of seq5215 to further explore with more

detail the ones that will present more striking results. As future experiments, we aim to use a human β-cell line to validate our results and to proceed to a more complex context, as mice, to fully understand the impact of single nucleotide variations can have in the development of T2D-like-phenotypes.

5. References

1. Ali, O. Genetics of type 2 diabetes. *World J Diabetes* **4**, 114–123 (2013).

2. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. *Lancet* **389**, 2239–2251 (2017).

3. Prasanna Kumar, H. R., Gowdappa, H. B., Hosmani, T. & Urs, T. Exocrine Dysfunction Correlates with Endocrinal Impairment of Pancreas in Type 2 Diabetes Mellitus. *Indian J Endocrinol Metab* **22**, 121–125 (2018).

4. Staiger, H., Machicao, F., Fritsche, A. & Häring, H.-U. Pathomechanisms of Type 2 Diabetes Genes. *Endocrine Reviews* **30**, 557–585 (2009).

5. Ong, C.-T. & Corces, V. G. Enhancers: emerging roles in cell fate specification. *EMBO Rep.* **13**, 423–430 (2012).

6. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics* **46**, 136–143 (2014).

7. Eufrásio, A. *et al.* In Vivo Reporter Assays Uncover Changes in Enhancer Activity Caused by Type 2 Diabetes–Associated Single Nucleotide Polymorphisms. *Diabetes* **69**, 2794–2805 (2020).

8. Wu, Y. *et al.* Functional annotation of sixty-five type-2 diabetes risk SNPs and its application in risk prediction. *Scientific Reports* **7**, 1–11 (2017).

9. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010).

10. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

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

12. Mularoni, L., Ramos-Rodríguez, M. & Pasquali, L. The Pancreatic Islet Regulome Browser. *Front Genet* **8**, (2017).

13. Miguel-Escalada, I. *et al.* Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics* **51**, 1137–1148 (2019).

14. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* **22**, 1059–1068 (2012).

15. Transcriptional enhancers in protein-coding exons of vertebrate developmental genes - PubMed. https://pubmed.ncbi.nlm.nih.gov/22567096/.

16. Hirsch, N. & Birnbaum, R. Y. Dual Function of DNA Sequences: Protein-Coding Sequences Function as Transcriptional Enhancers. *Perspect. Biol. Med.* **58**, 182–195 (2015).

17. Dwivedi, O. P. *et al.* Loss of ZnT8 function protects against diabetes by enhanced insulin secretion. *Nat. Genet.* **51**, 1596–1606 (2019).

18. Nicolson, T. J. *et al.* 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**, 2070–2083 (2009).

19. Birnbaum, R. Y. *et al.* Systematic Dissection of Coding Exons at Single Nucleotide Resolution Supports an Additional Role in Cell-Specific Transcriptional Regulation. *PLoS Genet* **10**, e1004592 (2014).

20. Althwanay, A., Turkistani, L., Alarfaj, B., Alhwaider, E. & Albagshi, F. A pilot study on the association of KCNJ11 genetic polymorphism (rs5215 Val250Ile) with Type 2 diabetes mellitus among the Saudi Arabian population (Eastern province). *IJMDC* 1062–1069 (2020) doi:10.24911/IJMDC.51-1587353314.

21. De Franco, E. *et al.* Update of variants identified in the pancreatic β-cell KATP channel genes KCNJ11 and ABCC8 in individuals with congenital hyperinsulinism and diabetes. *Human Mutation* **41**, 884–905 (2020).

22. Chatterjee, R. *et al.* KCNJ11 variants and their effect on the association between serum potassium and diabetes risk in the Atherosclerosis Risk in Communities (ARIC) Study and Jackson Heart Study (JHS) cohorts. *PLOS ONE* **13**, e0203213 (2018).

23. Haghvirdizadeh, P. *et al.* KCNJ11: Genetic Polymorphisms and Risk of Diabetes Mellitus. *J Diabetes Res* **2015**, 908152 (2015).

24. Malekizadeh, A. *et al.* Association of common genetic variants of KCNJ11 gene with the risk of type 2 diabetes mellitus. *Nucleosides, Nucleotides & Nucleic Acids* **40**, 530–541 (2021).

25. KCNJ11 gene: MedlinePlus Genetics. https://medlineplus.gov/genetics/gene/kcnj11/.

26. Smith, A. J., Taneja, T. K., Mankouri, J. & Sivaprasadarao, A. Molecular cell biology of KATP channels: implications for neonatal diabetes. *Expert Reviews in Molecular Medicine* **9**, 1–17 (2007).

27. Ashcroft, F. M. & Rorsman, P. KATP channels and islet hormone secretion: new insights and controversies. *Nat Rev Endocrinol* **9**, 660–669 (2013).

28. Babenko, A. P. *et al.* Activating Mutations in the ABCC8 Gene in Neonatal Diabetes Mellitus. *New England Journal of Medicine* **355**, 456–466 (2006).

29. Aleström, P. *et al.* Zebrafish: Housing and husbandry recommendations. *Lab Anim* **54**, 213–224 (2020).

30. Mohlke, K. L. & Boehnke, M. Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum. Mol. Genet.* **24**, R85-92 (2015).

31. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17921–17926 (2013).

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

33. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* **48**, D845–D855 (2020).

34. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).

35. The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs - PubMed. https://pubmed.ncbi.nlm.nih.gov/17937395/.

36. Expression of an activating mutation in the gene encoding the KATP channel subunit Kir6.2 in mouse pancreatic beta cells recapitulates neonatal diabetes - PubMed. https://pubmed.ncbi.nlm.nih.gov/19065048/.

37. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

38. Zheng, Y. & VanDusen, N. J. Massively Parallel Reporter Assays for High-Throughput In Vivo Analysis of Cis-Regulatory Elements. *J Cardiovasc Dev Dis* **10**, 144 (2023).

39. Ahituv, N. Exonic enhancers: proceed with caution in exome and genome sequencing studies. *Genome Med* **8**, 14 (2016).

6. Supplementary material

	Location							
Sequence	chr	start	end	Overlapping SNP	p-value	Associated trait	Gene/nearby gene	Parker et al (2013)	Miguel-Escalada et al (2019)
Seq101	2	165512270	165513422	rs10195252	4x10E-34	T2D	COBLL1	weak_enhancer	NA
Seq117	3	123065146	123066376	rs11708067	5x10E-32	T2D	ADCY5	weak_enhancer	Inactive enhancer
Seq119	3	170716882	170718473	rs11920090	8x10E-13	Fasting glucose-related traits	SLC2A2	strong_enhancer	Active_enhancer_I
Seq180	3	12392081	12393390	rs1801282	6x10E-10	T2D	PPARG	weak_enhancer	NA
Seq68	4	153519667	153521019	rs6813195	4x10E-14	T2D	TMEM154	weak_enhancer	Active_enhancer_I
Seq72	4	185716130	185716780	rs72695654	3x10E-6 (Lead SNP)	Pasquali - LD with rs735949	ACSL1	strong_enhancer	Active_enhancer_I
Seq73	4	185715959	185716793	rs735949	3.70x10E-6	T2D	ACSL1	strong_enhancer	Active_enhancer_I
Seq770	5	76425686	76426788	rs7708285	1x10E-9	T2D	ZBED3-AS1	strong_enhancer	NA
Seq58	6	37775233	37776300	rs58692659	2x10E-10 (Lead SNP)	Pasquali	ZFAND3	weak_enhancer	Active_enhancer_II
Seq460	7	44235386	44236078	rs4607517	7x10E-92	Fasting plasma glucose	YKT6	strong_enhancer	Active_enhancer_II
Seq219	7	15063701	15064872	rs2191349	3x10E-44	T2D	DGKB	NA	Active_enhancer_I
Seq6943	7	50791371	50792520	rs6943153	2x10E-13	Fasting plasma glucose	GRB10	strong_enhancer	NA
Seq108	9	22133720	22134465	rs10811661	8x10E-15	T2D	CDKN2B-AS1	strong_enhancer	Active_enhancer_II
Seq132	8	118184022	118185708	rs13266634	1x10E-47	T2D	SLC30A8	weak_enhancer	Active_enhancer_I
Seq703	9	4288789	4289921	rs7034200	1x10E-12	T2D	GLIS3	weak_enhancer	NA
Seq704	9	4287090	4287789	rs7041847	2x10E-14	T2D	GLIS3	strong_enhancer	NA
Seq1088	10	113041678	1,13E+08	rs10885122	3x10E-16	impaired fasting glucose	ADRA2A	weak_enhancer	NA
Seq125	10	80942221	80943036	rs12571751	2x10E-12	T2D	ZMIZ1	strong_enhancer	NA
Seq169	10	71093198	71094345	rs16926246	3x10E-54	Glycated hemoglobin levels	HK1	weak_enhancer	NA
Seq790	10	114757762	114759057	rs7903146	6x10E-447	T2D	TCF7L2	strong_enhancer	Active_enhancer_I
Seq116	11	45872729	45873451	rs11605924	1x10E-26	Fasting glucose-related traits	CRY2	strong_enhancer	NA
Seq5215	11	17408287	17408983	rs5215	5x10E-11	T2D	KCNJ11	strong_enhancer	Active_enhancer_I
Seq150	12	65223054	65224406	rs150781447	1x10E-16	Glycemic traits	TBC1D30	weak_enhancer	NA
Seq617	13	75898197	75899246	rs61736969	3.6x10E-25	Fasting glucose-related traits	TBC1D4	weak_enhancer	NA
Seq57	13	33554168	33555444	rs576674	1x10E-23	T2D	KL	weak_enhancer	Active_enhancer_II
Seq37	14	100838814	100839927	rs3783347	2x10E-11	Fasting glucose-related traits	WARS1	weak_enhancer	NA
Seq201	15	99258314	99259105	rs2018860	3x10E-8	Fasting glucose-related traits	IGF1R	strong_enhancer	NA
Seq717	15	62396062	62396926	rs7172432	8x10E-22	T2D	CDCD4A	weak_enhancer	NA
Seq740	15	38822782	38823911	rs7403531	4x10E-9	T2D	RASGRP1	weak_enhancer	NA
Seq142	16	53800626	53801740	rs1421085	6x10E-39	T2D	FTO	weak_enhancer	NA
Seq479	17	2262362	2263078	rs4790333	3x10E-9	Fasting proinsulin levels	SGSM2	strong_enhancer	Active_enhancer_I
Seq1040	19	19407499	19408752	rs10401969	3x10E-38	T2D	SUGP1	weak_enhancer	Active_enhancer_II
Seq611	20	22556724	22557434	rs6113722	2x10E-12	Glycemic traits	LINC00261	strong_enhancer	Active_promoter

Supplementary table 1 – List of the 33 putative enhancers selected for enhancer activity assays. The location of the sequences, the overlapping SNP, the p-value of association, the associated trait, the gene/nearby gene where is located the sequence and overlapping SNP and the overlapping with the datasets from Parker et al (2013) and Miguel-Escalada et al (2019) are described.

Sequence	Fw primer	Rv primer	SDM Fw primer	SDM Rv primer
Seq101	GCTCACACAATGGAAAATTAGATGCAGAAGC	GGCTAAATAACCACGCTCAGGAAATGGG	GGAATGGGAAAGAACATTTAACC	GGTTAAATGTTCTTTCCCATTCC
Seq180	GGTAATTTGGCACAGCTAGTATTTCTCC	GCACAGAGATAAAATAGATTAGGTTAGG	CCTATTGACCCAGAAAGC	GCTTTCTGGGTCAATAGG
Seq770	GGCTAAAGATTCCCAGACTTATTAACG	GCTCACTGTTTTCCATCTCGGTTCTGG	GCATAGAGAAAATGTGGACTGC	GCAGTCCACATTTTCTCTATGC
Seq6943	CCTGGAGGCTCTCATTTCATAACAGTGC	CCTGTAACATTGACTGTGCTCCCTTCTGG	GGAGGTAGGTCACACTGAGAGC	GCTCTCAGTGTGACCTACCTCC
Seq108	GAGGTGTGTCTATTCTGGC	CAGGCTAGAATGCAGTGGC	GCTTTAGTTTTCCCATGACAGTAAG	CTTACTGTCATGGGAAAACTAAAGC
Seq703	CCAGAAAGTAAAGTGATTAATACC	TTTTGTGCCTTAGTTTCCTTAGG	TCCTAAGTCTACTTGGTAAC	GTTACCAAGTAGACTTAGGA
Seq704	TGGCAGCATTAGAGTCTTGG	AAGTGTTTCTCCTTGATAGCC	CCTTGACCAGTAGCTCCCCAAAATGTATGC	GCATACATTTTGGGGAGCTACTGGTCAAGG
Seq1088	GGTTTATGATGTGGATGAGAAGCAGG	GGGAACAACATACACTATTCTTGG	CACAAGTTAGGGGGATATGCACC	GGTGCATATCCCCCTAACTTGTG
Seq125	GTGTCAGCTGGACTCAGAGTGC	GCAAATCTCACCCACAAATCTTCC	CGTATCCATTAGATGTGCACTGTTTTCC	GGAAAACAGTGCACATCTAATGGATACG
Seq169	GGCTGTTCTGTTCTCCCC	GGTTTGCTGGCTCAGATAAGG	GGAGTGCTGCTCTGCTTGCTGG	CCAGCAAGCAGAGCAGCACTCC
Seq116	TGACCTCTGTGGTTTTTCATTACC	CTTTTGGGTCCAGGCTGTCC	CCTGCGAGAGGCACTTTACTGTTCTGC	GCAGAACAGTAAAGTGCCTCTCGCAGG
Seq5215	AGTACAGGTTCCTGCTGAGGC	GGTACGCAAGACCACCAGC	GATGGTGTTGCC	CCATCAAAGTGCCC
Seq150	AGACCAGTAGGAGACTTGCAGTAGC	GGCAGAAAATATACAGACATTCAAACC	CCATGTGCTTCACTTTCAATGA	TCATTGAAAGTGAAGCACATGG
Seq617	CCAATATATTAGAAACAGATTAGAACAAGC	AGACTATAAATATCTCAATCCTTCC	GTACATGCGTCAAACTGACG	CGTCAGTTTGACGCATGTAC
Seq57	CGTGATGGTTAGATGGATTGC	GGCCTTCTATTTCCTTACTCTGG	CTGCTAACGTTGAACCC	GGGTTCAACGTTAGCAG
Seq37	GGGAGGAATGCCGGAGTACCC	GGGATCACTCAGGCATTCTACC	GCCTGAACTCTGCTTATTG	CAATAAGCAGAGTTCAGGC
Seq201	GGTCACATGCCCCCAGGC	CCCTGGAACAAACACCTTGGC	TTAATATTCCCAAAGAAACTGTCTGG	CCAGACAGTTTCTTTGGGAATATTAA
Seq717	GCAACTTGGCGTACTTGGTGTACTGG	GCCTTTCACCTTTCACCAATGCCTGGGG	TTAATATTCCCAAAGAAACTGTCTGG	CCAGACAGTTTCTTTGGGAATATTAA
Seq740	GGAACTGTGACATATACTTGTAGG	CCTTGGCTGAGTGAGTCATC	CCTTTATGGCTGGTTAGC	GCTAACCAGCCATAAAGG
Seq142	CCACTATATAGGTGATGGGTTAAAGC	GGGAACTCTTTCTATACTTCAGG	GGCATGACATTGATTAAGTGT	ACACTTAATCAATGTCATGCC
Seq479	GGGATTTCACCGTGTTAGC	AGGCGACCGCAAAAGCAGC	GTCAGTAATGCCTGGTCTAATGG	CCATTAGACCAGGCATTACTGAC
Seq1040	GGTATACCGACAAGAGAACTGAAAACG	GCGGAACATGGCTCTGAACCC	CTACACTGCGGATGAACC	GGTTCATCCGCAGTGTAG
Seq611	TCAAAGTTTTACATTTTACTCG	AAACGCCCAATCTTTATCGC	GCGCTCTCTTGCAGGGTGAAGGTGC	GCACCTTCACCCTGCAAGAGAGCGC
Seq5215mm	GGTACAGGCTCCTGCTGG	GCGCAAGACCACCAGCC		
Seq5215mm_ATA			GGTAACACCATAAAAGTGCCCACAC	GTGTGGGCACTTTTATGGTGTTACC
Seq5215mm_ATC			GGTAACACCATCAAAGTGCCCACAC	GTGTGGGCACTTTGATGGTGTTACC

Supplementary table 2 – List of primers used in this study.

Sequence	Non-risk (%)	N	Risk (%)	N
Seq101	0	17	22	18
Seq180	19	21	9	11
Seq770	30	20	39	23
Seq6943	5	16	7	12
Seq108	0	19	6	20
Seq703	0	17	0	12
Seq704	16	63	8	26
Seq1088	7	24	9	16
Seq125	4	21	10	21
Seq169	0	6	37	8
Seq116	30	10	0	26
Seq5215	9	56	20	25
Seq150	56	18	0	5
Seq617	0	17	50	19
Seq57	4	21	6	20
Seq37	9	26	5	11
Seq201	4	13	9	11
Seq717	0	22	6	20
Seq740	5	20	15	47
Seq142	30	10	33	4
Seq479	25	20	0	15
Seq1040	9	22	13	14
Seq611	0	25	24	63

Supplementary table 3 – Total percentage of embryos that express GFP expression in endocrine pancreas and the number of analyzed embryos for each tested sequence.

CHAPTER 4

A SHORT 3´UTR MOTIF REGULATES GENE EXPRESSION IN BILATERIANS

This chapter was published as:

Ana Eufrásio*, Joana Azevedo*, Joana Machado*, Alexandre Ferreira, Ana Moutinho, Filipe Henriques, Ana Jesus, Joana Tavares, Isabel Pereira-Castro, Joana Teixeira, Mafalda Araújo, Pedro B.P. Pinto, José Bessa#, Alexandra Moreira#. A short 3´UTR motif regulates gene expression in bilaterians. bioRxiv

Under revision in Nucleic Acids Research

* These authors contributed equally for this work

co-corresponding authors

doi: https://doi.org/10.1101/2023.11.15.567165

Author contribuitions. AEufrásio, AMoutinho, JAzevedo and AJesus performed all the zebrafish experiments; JMachado performed all the experiments in human cell lines and IP-C supervised the work; JTavares designed the bioinformatic script, AFerreira, FHenriques, JTeixeira and AEufrásio performed all the bioinformatic analyses; PBPP performed the in vitro transcription and UV crosslinking assays; JBessa, AMoutinho, AEufrásio, PBPP and AMoreira wrote the paper; JBessa and AMoreira designed the project and supervised the team.

1.Introduction

The 3' untranslated region (3'UTR) of the mRNA is defined by cleavage and polyadenylation of the pre-mRNA and has a central function in regulation of gene expression. This level of regulation is achieved by the existence of specific sequences in the 3'UTR that bind RNA binding proteins (RBPs), or that are targeted by miRNAs, controlling the fate of the mRNA (reviewed in [1–8]). One of the most important regulatory mechanisms found in the 3'UTR are the polyadenylation signals (PAS), that are conserved among fish, flies, mammals, and worms[9]. Similarly to mammals, the most common PAS in zebrafish (Danio rerio) is the canonical AAUAAA, which is present in 62.3% of all genes, and also contains the GU/U-rich downstream sequence[10]. However, about half of pre-mRNAs do not contain a canonical PAS (AAUAAA)[11,12] and yet, they are efficiently cleaved and polyadenylated. Usage of non-canonical PAS is enhanced and fine-tuned by the presence of auxiliary cis-elements in the pre-mRNA, such as the upstream sequence elements (USEs) that are localized upstream of the PAS. Therefore, USEs are fundamental for the proper control of gene activity. In humans, USEs have been characterized as sequences that activate PAS selection and 3' end formation efficiency by acting as an extra-platform for the recruitment of trans-acting factors, such as RBPs, with an important function in diseases[6,13–26]. Other examples of USEs have been identified in other species such as the fruit fly (Drosophila melanogaster). In the 3'UTR of the Drosophila's gene polo, which codes for a key cell cycle kinase[27], two polyadenylation sites originate two distinct mRNAs[28]. Upstream of polo's proximal PAS there is a highly conserved 28-nucleotides USE, which has a function in increasing Polo protein levels at the kinetochores[14]. Previous bioinformatic analysis has shown that the most conserved region of polo's USE (TTGTTTTT; DplUSE) is present upstream of a PAS in 2.7% of all human genes[14]. Additionally, we have previously shown that human DplUSEs are more prevalent upstream of weaker PAS than the canonical AATAAA, including in the human genome, suggesting their role in enhancing PAS efficiency in human genes, as it occurs in fly genes. Reinforcing this hypothesis, a sequence similar to DplUSE, having one nucleotide mismatch (TTGT[G/T]TTT) has been found in the 3´UTR of the human gene prothrombin F2, being demonstrated its role as a regulatory sequence of gene activity and its potential link to human disease[29]. All these results suggest that the DplUSE might be present in vertebrate genomes and its genes' regulatory function activity might be highly conserved, a question that is yet to be addressed.

Here we show that the DplUSE sequence in present in the 3´ UTR of several genes of zebrafish, mouse and humans, suggesting a conserved role of the DplUSE sequence in the regulation of genes activity. To further understand the DplUSE role, we have performed in vivo reporter assays in zebrafish using the DplUSE in the 3´UTR of the GFP reporter

gene, showing that DplUSE increases the expression of GFP in several different tissues. Additionally, we demonstrated that microinjection of DplUSE RNA in one-cell stage zebrafish embryos leads to several abnormalities during embryonic development, possibly by sequestering the molecular machinery that operates in the DplUSE, as might be the case of specific RBPs. Indeed, when the DplUSE is over-expressed in human cells, genes that contain the DplUSE sequence are downregulated, further supporting the dominant negative effect of the ectopic expression of DplUSE. These results suggest that overall, the molecular machinery that operates in the DplUSE is required for the proper early embryonic development of a vertebrate species as the zebrafish. To better understand the molecular mechanism of the DplUSE , we have explored the binding of candidate RBPs to the DplUSE sequence, showing that HuR/ELAVL1[30–35], hnRNPC[20,36] and PTBP1[37–40] RBPs bind to DplUSE . Importantly, we further show in human cells that when depleting PTBP1, the orthologue RBP that in the fruit fly was observed to be required for the function of the DplUSE, several genes that contain the DplUSE sequence are downregulated, suggesting that there is an ultra-conserved molecular mechanism controlling gene expression via the DplUSE sequence. Finally, we searched for SNPs that might affect the DplUSE sequence and are associated to human disease. We found that the rs3087967 SNP associated with malignant tumor of colon is located in the 3´UTR of the oncogene C11orf53/POU2AF2[41] and creates an ectopic consensus of DplUSE, causing a gain of function of the mRNA and suggesting its involvement in colon cancer development. Collectively, we provide in vivo evidence for the function of DplUSE in zebrafish, its conserved mechanism of action in zebrafish and humans as a regulator of gene expression and the potential implications of the DplUSE consensus in human disease.

2. Materials and Methods

2.1. Bioinformatic analysis

The NCBI RefSeq transcripts and the respective genes of Danio rerio (assembly danRer10), Mus musculus (assembly mm10) and Homo sapiens (assembly hg19) genomes were retrieved from UCSC browser [56]. The 3'UTR regions were selected and queried for the presence of the most conserved USE sequence (TTGTTTT) followed by pA1 (ATTAAA) spaced at a maximum of 450 bp, using an R-based script. Subsequently, the zebrafish genes were converted to human orthologs using BioMart[57].

The Venn diagrams were made by using BioVenn. The Gene Ontology terms from human, zebrafish and mouse USE-containing gene lists were obtained using the GO Enrichment Analysis using Panther tools[58]. Only GO terms related to biological processes with a false discovery rate (FDR) <0.05 were considered. The terms were ordered by the

fold enrichment (FE) and then by the FDR. The USE sequence data was intercepted by a dataset of SNPs described as linked to human diseases, available at the DisGeNET database[59]. Single-nucleotide polymorphisms (SNPs) that occurred in the USE sequence and that had a clinical significance were identified.

2.2. Disease-association diseases and DplUSE vs 3´UTR sequence conservation analysis

After obtaining the coordinates for the most conserved USE sequence, we, subsequently extracted associated disease information from DisGeNet datasets. The disease data was sourced from human-curated information, combining data from UniProt, ClinVar, Orphanet, the GWAS Catalog, and CTD, using the Python programming language.

The 3´UTR coordinates regions were obtained from USCS genome browser and the phyloP values for each nucleotide from 100 vertebrate dataset track. This dataset represents multiple alignments of 100 vertebrates species and the measurement of evolutionary conservation using phyloP values These phyloP values serve as a metric for measuring evolutionary conservation at individual alignment sites. Positive scores indicate higher conservation, while negative values suggest lower conservation, relative to the evolution predicted under neutral drift.

2.3. Oligonucleotides and Antibodies

The list of oligonucleotides and antibodies used in this work is in Supplementary Table 1.

2.4. Plasmid constructs

pUC19mini-TOL-USEmt constructs contain the DplUSE/DplUSEmt sequences, USE normal variant/USE risk variant (from C11orf53) and the PAS (ATTAAA) located between both Tol2 arms[43]. GFP-DplUSE and GFP-DplUSEmt pre-mRNAs were synthesized by in vitro transcription using pCS2-GFP-DplUSE and pCS2-GFP-DplUSEmt constructs, respectively. pCS2-GFP-DplUSE/pCS2-GFP-DplUSEmt contains a CMV promoter located upstream of the GFP coding sequence and DplUSE/DplUSEmt sequence. In order to microinject GFP-DplUSE and GFP-DplUSEmt RNAs in the transgenic GFP-USE zebrafish embryos, in vitro transcription using pCS2-USE construct was performed. This construct contains the SP6 promoter upstream of the USE sequence.

pSPT19-*polo* wt, pGEM7-*polo* wt, pGEM7-*polo* wt-DSE, pSPT19-*polo* mt, pGEM7-*polo* mt or the pGEM7-USE1mt-DSE contain the DplUSE/DplUSEmt sequence located

downstream of T7 or SP6 promoters, and they were constructed using the pBS-polo and pxb7 plasmids previously described[14], and pGEM-7Zf(+) and pSPT19 as vector backbones.

2.5. In vitro transcription

For synthesis of capped mRNA to microinject in zebrafish embryos, the template plasmids were initially linearized with appropriate restriction enzymes. In vitro transcription was performed incubating 1x transcription buffer, 5 mM DTT, 1 mM ATP, 1 mM UTP, 1 mM CTP, 0.5 mM GTP, 2.5 mM G(5')ppp(5')G RNA Cap Analogue (New England Biolabs), 4 µg DNA, 100 U Ribonuclease Inhibitor (NZYtech) and 60U of SP6 RNA polymerase (Thermo Fisher) for 2 h at 37ºC. The DNA was digested by adding 10 U DNase I recombinant (Sigma) and incubating for one additional hour. Transcripts were purified by performing phenol/chloroform purification. For UV cross-linking assays, the DplUSE pre-mRNAs were synthesized by performing in vitro transcription using pSPT19-*polo* wt, the pGEM7-*polo* wt or the pGEM7-*polo* wt-DSE constructs as templates. The DplUSEmt pre-mRNAs were synthesized by in vitro transcription using the pSPT19-*polo* mt, pGEM7-*polo* mt or the pGEM7-USE1mt-DSE. The DplUSE-(a), DplUSE-(b) and DplUSE-(c) pre-mRNAs were synthesized by in vitro transcription using constructs with pGEM-7Zf(+) as the parental vector. In vitro transcription was performed as previously described[40], with minor modifications, by incubating 1 µg of linearized plasmids with 1x transcription buffer (Roche Applied Science), 10 mM DTT, 28.79 U of RNAguardTM Ribonuclease Inhibitor, 0.1 mM GTP, 0.5 mM CTP, 0.05 mM ATP and UTP, 1 mM Cap Analogue (Ambion), 8 µCi of α-^{32}P-[ATP] and α-^{32}P-[UTP] and 20 U of T7 RNA Polymerase or SP6 RNA Polymerase (Roche Applied Sicences). The reaction was incubated for 1 h at 37ºC. The DNA was digested by adding 5 U of DNase I and incubating at 37ºC for 15 min. The transcripts were purified using Illustra ProbeQuantTM G-50 Micro Columns according with the manufacturer's instructions.

2.6. UV cross-linking assays

These assays were performed as previously described[40] with minor modifications. Briefly, a mixture containing 2 mM EDTA, 1 mM ATP, 20 mM CP, 11.5 U RNAguard™ Ribonuclease Inhibitor, 2.5% (w/v) PVA and 40 µg/mL of tRNA was incubated with 100 cps of radiolabeled RNA and 5 µL of HeLa cell nuclear extracts. Reactions were incubated for 10 min at 30ºC then 1 µL of 2.5 µg/mL tRNA was added. For competition assays 1, 50 and 150 pmoles of unlabeled competitor RNAs were added with labelled RNA. The reaction mixtures were irradiated twice with 96x10^4 µJ/cm^3 of UV light for 3 min, using the Hoefer UVC 500 Ultraviolet Crosslinker (GE Healthcare Life Sciences). RNase A was added to the samples, for 30 min at 37ºC, to degrade unprotected RNA. Samples were boiled at 95ºC for 5 min in a 2x SDS gel-loading buffer, in order to denature proteins, and separated by

electrophoresis in a 10% SDS-PAGE. The gel was fixed a 10% (v/v) acetic acid, 10% (v/v) glycerol solution for 30 min at room temperature and then dried at 80ºC under vacuum for 2 hours. The radiolabeled proteins bands were visualized by autoradiography.

2.7. Immunoprecipitations

For immunoprecipitation of UV crossed-linked proteins with monoclonal antibodies 4F4 anti-hnRNP C (mouse) (kind gift from Gideon Dreyfuss, Howard Hughes Medical Institute, University of Pennsylvania School of Medicine) [61] and HuR (19F12, Thermo) 400 µL of 10% (v/v) protein A Sepharose CL-4B beads (GE Healthcare Life Sciences) in IP-2 buffer [50 Mm Tris pH 7.9, 50 mM NaCl, 0.1% (v/v) NP-40] was incubated with 40 µg of rabbit anti-mouse antibody (DakoCytomation) in a vertical wheel for 90 min at 4ºC. The beads were washed 3 times with ice-cold IP-2 buffer. One UV cross-linking reaction and 2 µL of anti-hnRNP C or HuR was added to the beads and the mixture was rotated for 1 h at 4ºC. The beads were subsequently washed 3 times and dried. For PTBP1 immunoprecipitation, a rabbit anti-PTBP1[60] serum was used (generous gift from Chris W. J. Smith, Department of Biochemistry, University of Cambridge). To 100 µl of 50% (v/v) protein A sepharose-PBS, 20 µl of anti-PTBP1 and 600 µl of PBS were added, and the mix incubated for 1 hour at 4º C with gentle mixing. The beads were then washed with PBS and an UV crosslinking reaction was added and incubated overnight at 4º C. After incubation, the beads were washed two times with 800 µl of binding buffer I (20 mM Hepes pH 7.9, 150 mM NaCl, 0.05% (v/v) triton X-100) and then two times with binding buffer II (20 mM Hepes pH 7.9, 150 mM NaCl, 1% (v/v) triton X-100). As a non-specific control, an immunoprecipitation was performed using a rabbit pre-immune serum. 20 µL 2x SDS gel-loading buffer was added to the beads and the proteins were denatured at 95ºC for 5 min and separated by gel electrophoresis in a 10% SDS-PAGE. The gel was fixed and dried at 80ºC under vacuum for 2 h and the radiolabeled protein bands were visualized by autoradiography.

2.8. Zebrafish husbandry and embryo culture

Adult wild-type Tuebingen (TU) zebrafish (*Danio rerio*) were maintained in a recirculating system under conditions approved by the i3S Animal Welfare and Ethics Committee and the Portuguese National Authority for Animal Health (DGAV). Fertilized eggs were kept at 28 ºC in E3 medium supplemented with 0.001% 1-phenyl-2-thiourea to prevent pigmentation development[61].

2.9. Microinjection procedures, zebrafish GFP-DplUSE transgenic line establishment, GFP screening and quantification

One-cell stage AB WT embryos were microinjected with 3nL containing 25ng/ul *Tol2* transposase mRNA and 25ng/ul phenol/chloroform-purified reporter vectors: pUC19-miniTOL-GFP-DplUSE, pUC19-miniTOL-GFP-DplUSEmt and mylz:mCherry, being the latter used as internal control of transgenesis, driving mCherry expression specifically in muscle fibers of the zebrafish. Embryos were screened at 24hpf for GFP expression, using a fluorescence stereomicroscope. The images were acquired using Leica M205 and analyzed using ImageJ software[62]. For the GFP-RNAs microinjection, one-cell stage AB WT embryos were microinjected with 3nL containing 300ng/ul GFP-DplUSE and GFP-DplUSEmt RNAs. Embryos were screened and quantified at 24hpf for GFP expression as described before. Generation of the GFP-DplUSE zebrafish transgenic line was performed using the *Tol2* transposon system. One-cell stage AB wild-type embryos were microinjected with 3nL containing 25ng/ul *Tol2* transposase mRNA and 25ng/ul phenol/chloroform-purified reporter vector (pUC19-miniTOL-GFP-DplUSE). Embryos were screened at 48hpf for GFP expression, using a fluorescence stereomicroscope, and raised until adulthood. Potential founders were then crossed with WT zebrafish and the progeny with consistent GFP expression was selected and raised until adulthood, establishing this way the GFP-DplUSE transgenic line. For the USE-RNAs microinjection, one-cell stage AB WT embryos were microinjected with 3nL containing 300ng/ul DplUSE and DplUSEmt RNAs. Embryos were screened and quantified at 24hpf and 5dpf for GFP expression as described before. The fluorescence was calculated using the following formula: GFP Fluorescence = Integrated Density – (Area of selected area X Mean fluorescence of background readings).

For the assessment of the impact of the rs3087967 risk variant, we performed in vivo mosaic transgenesis assay. One-cell stage AB WT embryos were microinjected with 3nL containing 25ng/ul Tol2 transposase mRNA and 25ng(ul phenol/chloroform-purified reporter vectors (pUC19-miniTOL-GFP normal variant and pUC19-miniTOL-GFP risk variant and mylz:mCherry). Embryos were raised until reach the 4dpf larval stage, fixed and stained with DAPI (PUREBLU™ DAPI (1351303; Bio-Rad). The images were acquired using SP5II confocal microscope.

2.10. Immunohistochemistry

Transgenic GFP-USE zebrafish embryos were fixed in 4% formaldehyde in PBS1x, overnight at 4ºC, and then washed in PBS-T (0.1% Triton X-100 in PBS1x), permeabilized in PBS-T 0.5%, rewashed and incubated with DAPI (1:1000) (Invitrogen) diluted in PBS-T 0.1%, for 4h. Embryos were extensively washed and mounted in microscopic slides in 50%

glycerol in PBS1x. Images were acquired using Leica SP5II confocal microscope and analyzed in ImageJ. DplUSE and DplUSEmt RNAs microinjected embryos were fixed and washed as described before. For permeabilization, the embryos were incubated with 100% cold MeOH and blocked with in situ hybridization buffer. The embryos were then incubated with anti-cleaved caspase 3 pAb (1:100) (Anti-caspase-3, cleaved (Ab-2) Rabbit pAB (PC679; Merck) followed by a wash and a second incubation with anti-rabbit 568. Finally, the embryos were washed, the images acquired in Leica M205.and analyzed in ImageJ.

pUC19-miniTOL-GFP normal variant and pUC19-miniTOL-GFP risk variant microinjected embryos were fixed, washed and incubated with DAPI as described before. Images were acquired using Leica SP5II confocal microscope and analyzed in IMARIS software (Imaris (RRID:SCR_007370)).

2.11. Cell culture and transfection

HeLa adherent cell line was grown and maintained in Dulbecco's Modified Eagle Medium (DMEM) with GlutaMAX, supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin solution (10000 U/mL) (Gibco, Thermo Fisher Scientific). The cell line was maintained in a humidified incubator at 37º C with 5% CO_2. HeLa cells were transfected with 500 ng of plasmid DNA using Lipofectamine 2000 (Thermo Fisher Scientific) and following manufacturer's instructions. To knockdown *PTBP1*, two siRNAs targeting this RBP (siPTBP1) were transfected into HeLa cells (siPTBP1_1: 5'-GCACAGUGUUGAAGAUCAU-3'; siPTBP1_2: 5'- AACUUCCAUCAUUCCAGAGAA-3'; Sigma-Aldrich), using the jetPRIME® (Polyplus Transfection) reagent, following the manufacturer's guidelines. A non-target siRNA pool (siNTC: 5'-UGGUUUACAUGUCGACUAA-3'; 5'- UGGUUUACAUGUUGUGUGA-3'; 5'-UGGUUUACAUGUUUUCUGA-3'; 5'- UGGUUUACAUGUUUUCCUA-3'; Dharmacon) was also transfected, to serve as a negative control. Cells were transfected at 50% confluency with 50 nM of siPTBP1 (25 nM siPTBP1_1 + 25 nM siPTBP1_2) or 50 nM of siNTC. Cells were collected for total RNA extraction 48h post-transfection.

2.12. RNA extraction and RT-qPCR

RNA was extracted using TRIzol (Ambion), according to the manufacturer protocol. 1 µg of RNA were treated with Recombinant DNase I (Roche) and cDNA was synthesized using Superscript IV reverse transcriptase (Invitrogen) and random hexamers. RT-qPCR reactions were performed using SYBR Select Master Mix (Applied Biosystems) in the 7500 Fast Real-Time PCR System thermocycler (Applied Biosystems). The list of primers used are listed in Supplementary Table 1 and the results were analysed using the ΔΔCt method using 18S as the housekeeping gene.

2.13. Statistical Analysis

Statistical analyses were performed using unpaired Student's t-test and by using χ^2 test with Fisher correction. p-values <0.05 were considered as statistically significant. *, p<0.05; **, p<0.01; ***, p<0.001; ****, p<0.0001.

3. Results

3.1. A fruit fly upstream sequence element is conserved in the 3' UTR of humans, mouse and zebrafish genes.

In a previous study, we have identified a USE sequence (DplUSE) located in the 3'UTR of the cell cycle gene polo from the fruit fly (Drosophila melanogaster)[14]. We found that this sequence is essential to maintain Polo protein levels required for cell cycle progression[14], suggesting that this USE could represent a regulatory mechanism for gene activity. To understand if the DplUSE sequence could be present in vertebrate species we have developed a bioinformatic script to identify all the genes that contain in their 3'UTR the DplUSE sequence (TTGTTTT) located at a maximum distance of 450 nucleotides (nt) upstream of the weak ATTAAA polyadenylation signal (Figure 1A).Our bioinformatic search identified the presence of the DplUSE in 2110 human genes, 1648 mouse genes and 714 zebrafish genes, with subgroups sharing orthologues between species (Figure 1B). From the 2110 human genes that contain DplUSE, we observed a significant enrichment for diseases such as congenital abnormalities, colorectal cancer and malignant neoplasms (MNeoplasms), comparing with the control genes (Figure1C). Importantly, 31 orthologue genes in human, mouse and zebrafish share the presence of the DplUSE in their 3'UTR (Figure 1B and D). From these 31 orthologue genes, we verified a significant enrichment for congenital abnormalities, MNeoplasms, and for nervous system diseases, comparing with the control genes (Figure S1A). Altogether, these results highlight the potential of the regulatory function of DplUSE in human diseases. Considering this, we then explored whether the DplUSE sequence could show a higher level of conservation than the respective 3´UTR region, when aligned with multiple vertebrate genomes. The results showed that indeed the DplUSE motif is significantly more conserved, when compared to the remaining 3´UTR, both in the 31 orthologue genes dataset (Figure 1E), as in the 2110 human genes (Figure S1B and C). Next, we evaluated the degree of conservation of each nucleotide of the DplUSE motif in vertebrates based on the multiple alignments of 100 vertebrate species and measurements of evolutionary conservation using phyloP[42]. For this analysis we extended 3 nucleotides (N) upstream and downstream of DplUSE. Using the DplUSE motifs identified in the 2110 human DplUSE containing genes, we observed that

the TTT located in the 5th to 7th position of the DplUSE motif (TTGTTTTT) present highest values of conservation while the T located in the 4th position DplUSE motif (TTGTTTTT) presented the lowest value of conservation (Figure1F). Similar results were observed when analysing the DplUSE motif present in the 31 genes subset (Sup Fig1D). These results suggest that the DplUSE motif might accommodate some nucleotide variants. Nevertheless, using the multiple DplUSE human sequences aligned with the genome of key vertebrate species (mouse, alligator and zebrafish), we found that the most constant consensus was the TTGTTTTT motif, with the exception of the alignments with zebrafish, which motif was TTCTTTC (Sup.Fig.1E and F). Analyses of the gene ontology (GO) terms related to biological processes demonstrated that for Homo sapiens DplUSE-containing genes are involved in dynamic processes such as cellular migration, adhesion and communication, and mRNA processing (Figure S1G). For Mus musculus, DplUSE-containing genes GO terms include regulation of epithelial to mesenchymal transition and cell response to glucose (Figure S1H), and for Danio rerio, DplUSE-containing genes are involved in mitotic cell cycle process, protein ubiquitination and several developmental processes (Figure S1I). We then asked if there were common GO terms that are highly enriched (F.E ≥ 1.5) between the three organisms (Figure S1J). Although there are no common GO terms for the three species all together, we found that 10 out of 22 human GO terms are common to mouse, 2 are common to zebrafish GO terms; and 4 out of 14 zebrafish GO terms are common to mouse (Figure S1K). These GO terms include cell cycle related functions such as “cell division” (human and mouse), “chromosome organization” (mouse and zebrafish), “mitotic cell cycle process” and “mitotic cell cycle” (human and zebrafish), which is consistent with the role of the fruit fly gene polo, from where the DplUSE has been originally identified, suggesting that DplUSE might have conserved regulatory functions in cell cycle control. GO terms for other important cell biology terms as “positive regulation of cell migration” and “mRNA processing” have been also identified.

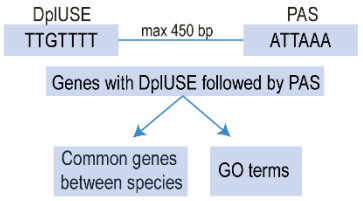
3.2. The DplUSE sequence increases protein expression in a vertebrate model system

To understand if the DplUSE might act as a gene’s activity regulatory sequence in vertebrates, increasing gene’s activity as observed in the fruit fly, we have developed two expression vectors, pUC19miniTOL-GFP-DplUSE and pUC19miniTOL-GFP-DplUSEmt. These constructs are ubiquitous expression vectors containing the DplUSE sequence, or the DplUSE sequence mutated in 16 nucleotides (DplUSEmt), subcloned downstream of the GFP coding sequence and assembled in a Tol2 transposon (Figure 1G), suitable to perform transgenesis reporter assays in zebrafish embryos. Microinjection of these constructs in one-cell stage wild-type (WT) embryos, simultaneously with Tol2 mRNA and

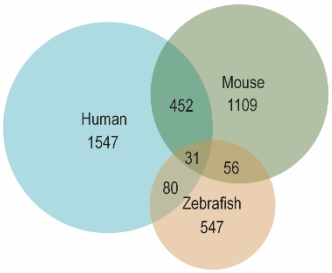
another transposon that drives expression of mCherry in the muscle (control of transgenesis), showed that, upon 24 hours post-fertilization (24 hpf), embryos microinjected with GFP-DplUSE DNA exhibited more intense GFP fluorescence, while embryos microinjected with GFP-DplUSEmt DNA demonstrated a more diffused and faded fluorescence (Figure 1H). Quantification of GFP intensity (Figure 1J left) revealed that embryos microinjected with GFP-DplUSE DNA showed a significantly higher expression than when microinjected with GFP-DplUSEmt DNA. These results demonstrate that the DplUSE is able to enhance GFP expression in a vertebrate species as in the fruit fly, demonstrating its conserved function from arthropods to vertebrates.

To better understand if this activating effect takes place at a transcriptional or post-transcriptional level, GFP-DplUSE and GFP-DplUSEmt RNA were in vitro transcribed and microinjected in one-cell stage zebrafish embryos. Despite the fact that the fluorescence in these embryos was far more diffuse than the one observed for embryos microinjected with the expression vectors (Figure 1H), it is still possible to notice and quantify a higher GFP expression in embryos microinjected with GFP-DplUSE RNA than with GFP-DplUSEmt RNA (Figure 1I and 1J, right), in a similar manner to the results obtained upon the microinjection of GFP-DplUSE expression vector (Figure 1H and 1J, left). These results indicate that in zebrafish the DplUSE sequence increases gene expression in a post-transcriptional manner, which is compatible with the mechanistic action of the DplUSE sequence found in the fruit fly, where it was observed to be dependent on the binding of RBPs[14].

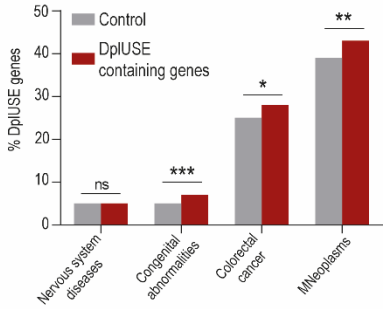
A)



B)



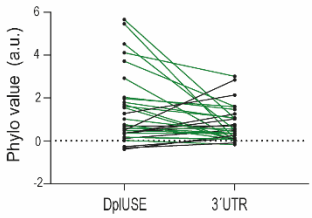
C)



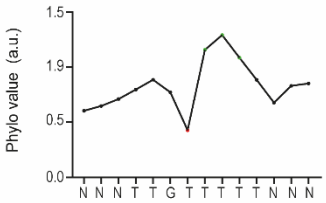
D)

31 DplUSE-containing genes common in Human, Mouse and Zebrafish			
ACBD3	*MED18*	*DCLRE1B*	*RNF180*
CAND1	*NEGR1*	*DKC1*	*SOD2*
CDKN2C	*NPEPPS*	*DYNLT3*	*STRN*
CHUK	*NR5A2*	*EXT2*	*STX6*
CLDN12	*PPM1K*	*KIF20A*	*UBE3C*
CNPY1	*PRDM10*	*LUC7L*	*YES1*
CPSF6	*RAB27A*	*MAP4K5*	*ZMAT2*
CSNK1A1	*RGMB*	*MBNL2*	

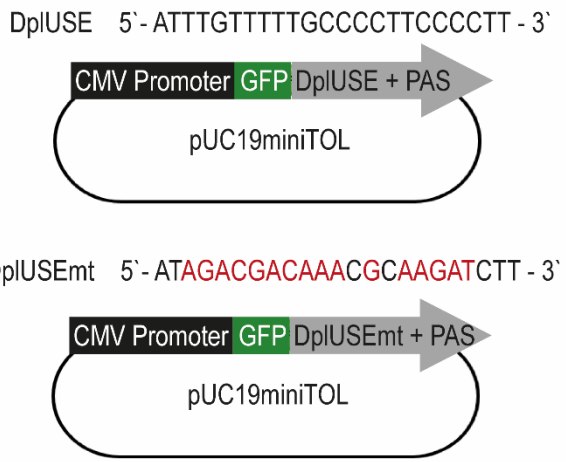
E)



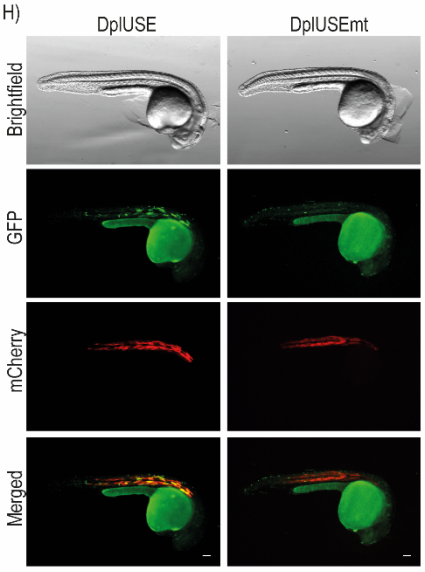
F)



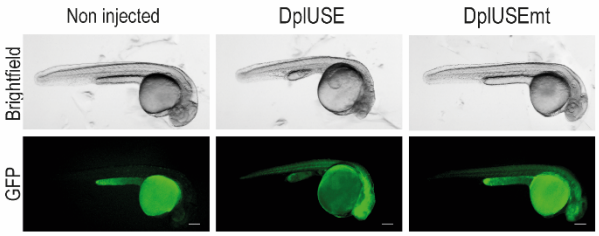
G)



H)



I)



J)

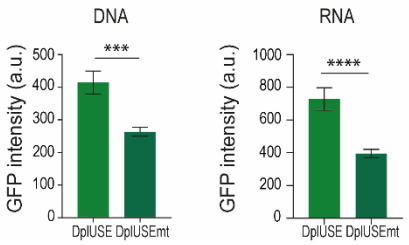


Figure 1 - The DplUSE sequence is conserved across human, zebrafish and mouse genes and increases GFP expression in a post-transcriptional manner in zebrafish. *(A) Workflow used to obtain the Gene Ontology (GO) terms related to biological processes of the common DplUSE-containing genes between the human and zebrafish transcriptomes obtained from an R-based bioinformatic script, which identified all the genes that contained, in the 3' UTR, the most conserved region of the USE sequence (TTGTTTT) located upstream of the polyadenylation signal (ATTAAA), at a maximum separating distance of 450 base pairs (bp). (B) Venn Diagram representing: the number of DplUSE-containing genes found in human (2110), zebrafish (714) and mouse (1648). 80 USE-containing genes are common between human and zebrafish while 31 are common between the 3 species. (C) Representative graph showing the percentage of the 2110 DplUSE-containing human genes that are associated with the respective disease (DisGenet data source). (D) The 31 USE-containing genes that are common between the 3 species: human, mouse and zebrafish. (E) Representative graph of the average value of the extracted conservation values from each nucleotide within the USE and 3´UTR sequence, for each one of the 31 ortholog genes (data source:100vertebrates). (F) Representative graph showing the average conservation value for each nucleotide within the 2110 genes. (G) Schematic representation of pUC19miniTOL-GFP-DplUSE and pUC19miniTOL-GFP-DplUSEmt plasmids used to microinject one-cell stage zebrafish embryos. The black boxes correspond to the cytomegalovirus (CMV) promoter region; the green boxes correspond to the GFP reporter gene; and the grey boxes represent the DplUSE sequence, or the DplUSE sequence mutated in 16 nucleotides, followed by the polyadenylation signal (PAS)- ATTAAA. (H) Representative images from 24hpf zebrafish embryos microinjected with pUC19miniTOL-GFP-DplUSE/pUC19miniTOL-GFP-DplUSEmt with Tol2 mRNA and mylz:mCherry vector. (I) Representative images from 24hpf zebrafish embryos microinjected with GFP-DplUSE RNA and GFP-USEmt RNA. (J) Representative graph for the GFP expression quantification from microinjected embryos with pUC19miniTOL-GFP-DplUSE and pUC19miniTOL-GFP-DplUSEmt (left); Representative graph for the GFP expression quantification from microinjected embryos with GFP-USE and GFP-USEmt RNAs; Statistical significance was determined by $\chi 2$ test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05. Images were acquired with Leica M205(scale bar = 100 µm) and the GFP expression was quantified for at least 25 embryos in each.*

3.3. DplUSE RNA recruits protein factors essential for the correct expression of genes involved in zebrafish development and in human cells function

To better characterize the function of the DplUSE sequence in controlling gene's activity, a transgenic GFP-DplUSE zebrafish line was generated using the Tol2 transposon system[43]. Although the expression of GFP is driven by a ubiquitous promoter (CMV), the stable transgenic line exhibits GFP expression mainly in the notochord, muscle, pituitary gland and eye (Figure 2A and 2B), suggesting a possible role of the DplUSE in regulation of gene expression in a tissue and cell-type specific manner. Specifically, confocal imaging demonstrates the clear and intense pattern of GFP expression in the notochord and muscle (Figure 2B), suggesting that the DplUSE sequence might have tissue specific regulatory functions. To address the function of the DplUSE and the respective molecular machinery that operates in this sequence, we considered a dominant negative approach, induced by ectopic introduction of DplUSE RNA. In this case we postulate that factors such as RBPs might be sequestered by the ectopic DplUSE RNA, affecting therefore the expression of DplUSE controlled genes. To test this hypothesis, we have microinjected DplUSE RNA in transgenic GFP-DplUSE embryos and we quantified the GFP expression, in comparison with the microinjection of DplUSEmt RNA (DplUSE sequence mutated in 17 nt) (Figure 2C). We found that the microinjection of DplUSE RNA was able to decrease GFP expression, contrasting with the microinjection of USEmt RNA at 5 dpf, particularly in the notochord and

in the muscle (Figure 2D to F), therefore corroborating the dominant negative effect hypothesis. These results were consistent at 48 hpf, where it was also observed a decrease in GFP expression in the notochord (Figure S2A and B). Further observation of the imaging revealed that the microinjection of DplUSE RNA also affected the consistency of the GFP expression, leading to the development of patchy GFP expression, a phenomenon that occurred less frequently upon microinjection of DplUSEmt RNA at 5 dpf (Figure 2D and G) and at 48 hpf (Figure S2A and S2C).

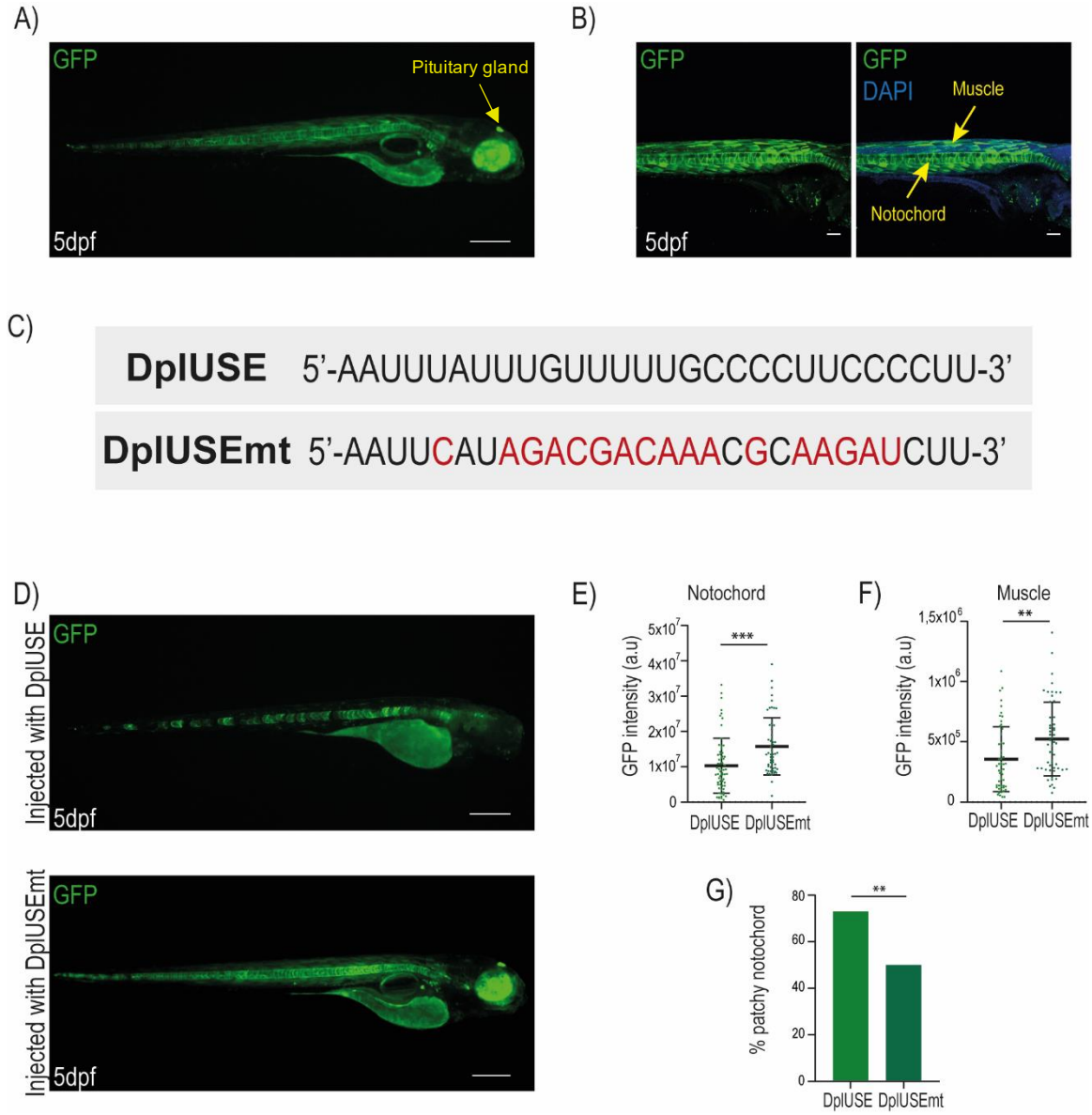
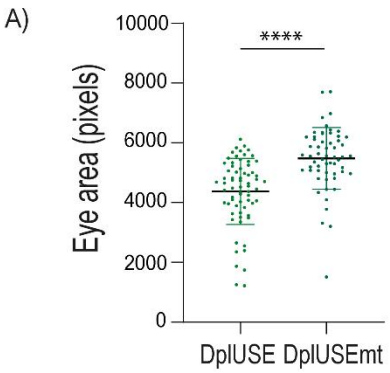
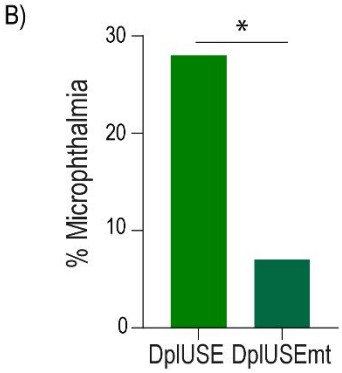


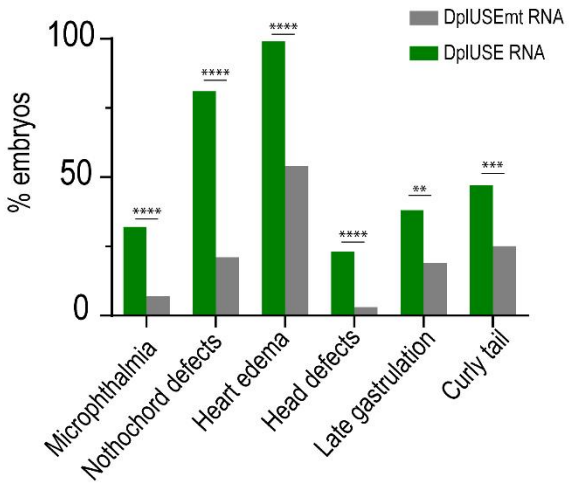
Figure 2 - Microinjection of DplUSE RNA causes a decrease in GFP expression in transgenic GFP-USE animals. *(A) Representative image of a 5dpf GFP-DplUSE transgenic line. (B) Confocal images of transgenic GFP-USE zebrafish, stained with the nuclear marker DAPI, showing GFP expression in the notochord and muscle (Scale bar= 10 µm). (C) DplUSE RNA and DplUSEmt RNA sequences used to microinject transgenic GFP-DplUSE zebrafish embryos. (D) Representative image of GFP expression in transgenic GFP-DplUSE zebrafish embryos microinjected with DplUSE RNA or USEmt RNA, at 5 dpf. Images were acquired using Leica M205. (E) Quantification of GFP expression in notochord (n=64 DplUSE; n=54 DplUSEmt) and (F) in muscle (n=53 DplUSE; n=54 DplUSEmt). (G) Quantification of the percentage of transgenic GFP-DplUSE embryos that showed a patchy expression of GFP in the notochord (73%,n=69 DplUSE;50%,n=.59 DplUSEmt) Statistical significance was determined by χ2 test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05. (Scale bar = 100 µm).*

Next, we have injected DplUSE RNA and focused on developed phenotypes. At 5 dpf, zebrafish larvae microinjected with DplUSE RNA exhibited microphthalmia, a condition where either one or both eyes are abnormally small or absent[44]. When the eye size was quantified, a significant decrease was observed, when compared to larvae microinjected with DplUSEmt RNA (Figure 3A and B). This phenotype was also observed at 48 hpf (Figure S2D and S2E). Apart from these eye phenotypes, a careful evaluation of 24 hpf larvae allowed us to identify other important phenotypes affecting the notochord, heart, head, late gastrulation and tail, that are more prevalent in DplUSE RNA than with DplUSEmt RNA injected embryos (Figure 3C-E). At 5 dpf, malformations in the notochord, head, tail and swim bladder where also detected (Figure S3A, B and C). To better understand the mechanisms that might be causing these phenotypes, we inquired if apoptosis was increased by performing immunohistochemistry staining with anti-cleaved caspase-3 polyclonal antibody (pAb) and analyzed the presence of cleaved caspase-3 fluorescence signal in 48hpf embryos microinjected with DplGFP-USE and DplGFP-USEmt RNAs. We observed a more prevalent presence of cleaved caspase-3 staining in embryos microinjected with DplGFP-USE RNA, when comparing with not injected embryos and embryos injected with DplGFP-USEmt RNA (Figure 3F and G), suggesting that, in part, these phenotypes are caused by apoptosis. Overall, these results show that ectopic DplUSE RNA might work as a dominant negative, interfering with the normal DplUSE regulatory functions, as show using the GFP-DplUSE reporter line. In addition, ectopic DplUSE RNA causes several developmental abnormalities, possibly by sequestering RBPs that bind to the DplUSE mRNA of endogenous USE-containing genes, interfering with the expression of these genes and others that might also depend on the function of these RBPs and which expression is necessary for correct embryo development.

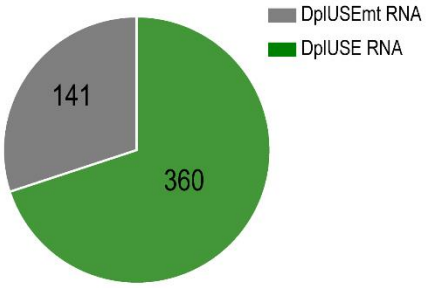




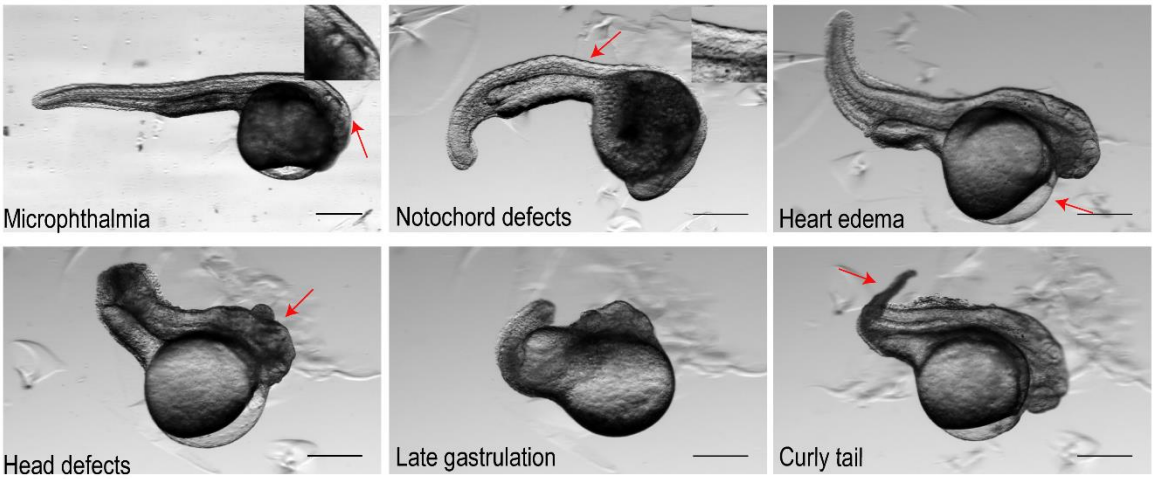
C) Defects on 24hpf zebrafish embryos



D) Total number of defects in 24hpf zebrafish embryos



E)



F)

G)

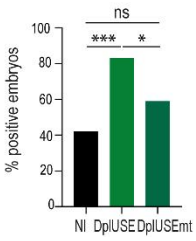


Figure 3 - Microinjection of DplUSE RNA in one-cell stage zebrafish embryos induces the development of several defects. *(A) Representative graph of the quantification of eye size (pixels) in microinjected embryos with DplUSE and DplUSEmt RNAs, at 5dpf. (n=69 DplUSE; n=58 DplUSEmt) (B) Representative graph of que quantification of the percentage of microphthamia in microinjected embryos with DplUSE and DplUSEmt RNAs, at 5dpf (28%, n=68 DplUSE; 7%,n=58 DplUSEmt. (C) Quantification of the several defects developed 24h upon the microinjection of GFP-DplUSE RNA (in green) and GFP-DplUSEmt RNA (in grey) in one-cell stage zebrafish embryos. (D) Pie chart representing the total number of defects developed upon the microinjection of DplGFP-USE RNA and GFP-DplUSEmt RNA. (E) Representative images showing the various abnormalities upon microinjection of GFP-DplUSE RNA at 24 hpf, indicated with red arrows. (F) Representative images of not injected embryos and in embryos injected with DplUSE and DplUSE RNAs, at 48hpf, showing in purple, the cleaved anti-caspase-3 staining. (G) Representative graph showing the percentage of embryos with consistent appearance of apoptotic cells, characterized by the presence of anti-caspase-3 staining (42%,n=39 Not injected; 83%,n=33 DplUSE;59%, n=33 DplUSEmt). Statistical significance was determined by χ^2 test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05 Images acquired with Leica M205.Scale bar = 100 µm.*

To understand if the DplUSE sequence shares similar regulatory mechanisms in human cells as observed in the fruit fly and zebrafish, we started by asking whether the DplUSE would have an activating effect in GFP expression in HeLa cells, using reporter constructs. Transfection with the pUC19miniTOL-GFP-DplUSE and pUC19miniTOL-GFP-DplUSEmt constructs show that DplUSE is also able to enhance the expression of GFP, showing a ~2-fold increase in GFP expression with the DplUSE-containing construct in comparison to the DplUSEmt construct (Figure 4A). These results recapitulate the ones observed in zebrafish and suggest that the DplUSE gene regulatory mechanisms are conserved in vertebrates. Furthermore, similar to what we observed in zebrafish (Figure 2D), we reasoned that the DplUSE transcribed from the transfected plasmid may also act as a dominant negative, titrating the molecular machinery that operates at the DplUSE, resulting in a decrease of expression of genes that are under the control of DplUSE. Thus, after transfection, the mRNA levels of nine DplUSE-containing genes (KIF20A, EXT2, CLDN12, MBNL2, MED18, NR5A2, CSNK1A1, DKC1 and MAP4K5), selected from the 31 genes identified by the bioinformatic script that are common to human, mouse and zebrafish (listed in Figure 1D) were quantified by RT-qPCR (Figure 4B-J). Our show that the transfection of GFP- DplUSE leads to a significant decrease in the expression of most of these genes in comparison with GFP- DplUSEmt (Figure 4B-J). Thus, our results show that the ectopic expression of DplUSE might act as a dominant negative, downregulating the expression of genes that contain the DplUSE and that are associated with cancer, such as KIF20A which is a tumour-associated antigen involved in the glioma cell growth and cell survival, Claudin12, which is overexpressed in colorectal carcinomas, NR5A2, which negatively regulates cell proliferation and tumour growth in nervous system malignancies, CSNK1A, which is a negative regulator of canonical Wnt signalling pathway and of oncogenic RAS-induced autophagy and Cdkn2c, which suppresses tumorigenesis, further

supporting the conserved role of the DplUSE sequence as regulator of gene expression in vertebrates.

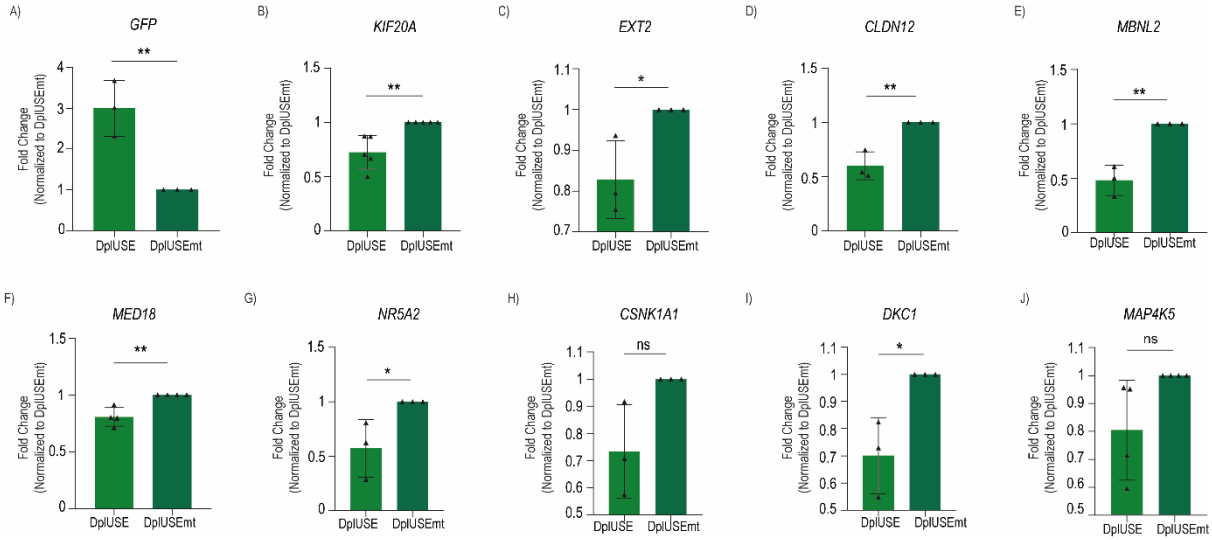


Figure 4 -The DplUSE increases GFP expression and modulates DplUSE-containing genes in human cells. *(A) RT-qPCR showing that the mRNA levels of GFP increase by 3-fold upon transfection with the pUC19miniTOL-GFP-DplUSE plasmid (DplUSE) when comparing with the transfection with the pUC19miniTOL-GFP-DplUSEmt plasmid (USE mut). mRNA levels of KIF20A (B), EXT2 (C), CLDN12 (D), MBNL2 (E), MED18 (F), NR5A2 (G), CSNK1A1 (H), DKC1 (I) and MAP4K5 (J) were assessed by RT-qPCR. Apart from CSNK1A1 and MAP4K5, transfection with the pUC19miniTOL-GFP-DplUSE plasmid decreases the mRNA levels of the USE-containing human genes. The data is presented as the mean ± SD of at least three independent experiments and analysed by two-tailed unpaired Student's t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05.*

3.4. A bilaterian ultraconserved molecular mechanism operates in the DplUSE RNA

Next, we aimed to better understand the molecular mechanism of the DplUSE sequence in controlling gene's activity. We reasoned that, because of its nature as a USE located at the 3'UTR of genes, DplUSE could be binding RBPs, that have been shown to be players of USEs function[22,26,40,45,46]To test this hypothesis, and to determine if human proteins bind to the DplUSE, constructs containing the DplUSE or DplUSEmt were used to perform in vitro transcription generating radioactively labelled pre-mRNAs that were incubated with human HeLa cells nuclear extracts to perform UV cross-linking assays. The results show that multiple proteins assemble onto the DplUSE pre-mRNA (Figure 5B, lane 1) and that mutation of the USE sequence leads to a decrease in the binding (Figure 5B, lane 2). To determine the binding specificity of these proteins to the DplUSE RNA motif, UV cross-linking competition assays were performed, where increasing amounts of unlabeled DplUSE or DplUSEmt RNA were added to the reaction mix (Figure 5C). The results showed that DplUSE RNA unlabeled competitor (1 pmol, 50 pmol and 150 pmol- Figure 5C, lane 2, 3 and 4, respectively) significantly inhibits the assembly of the ~35 kDa, ~40 kDa and ~60 kDa proteins onto the radioactive labelled DplUSE pre-mRNA. This competition is DplUSE

sequence specific, as addition of the same amounts of unlabeled DplUSEmt RNA (1 pmol, 50 pmol and 150 pmol – Figure 5C, lane 5, 6 and 7, respectively) did not cause any noticeable changes in the proteins bound to DplUSE pre-mRNA. These results also reinforce the dominant negative mechanism observed when overexpressing the DplUSE in zebrafish embryos (Figure 2, 3 and 4) and human cells (Figure 4).

In the fruit fly, the RBP Heph was previously shown to bind to DplUSE RNA of the fly's gene polo[14] being required for the DplUSE regulatory functions. For this reason, we tested if the human orthologue of Heph, the RBP PTBP1, could bind to the DplUSE RNA in human cells, similarly to the fruit fly, supporting an ultra-conserved molecular mechanism operating at the DplUSE. To test this hypothesis, we performed immunoprecipitation assays after UV cross-linking. Results show that similarly to Heph, the human orthologue PTBP1 binds specifically to the DplUSE and not to the DplUSEmt control RNA (Fig. 5F). PTBP1 is a 57 kDa protein with multiple functions[47] that acts as a regulator of alternative splicing[48,49] with neuronal functions[39,50,51] and as a polyadenylation regulator in virus, humans and Drosophila[14,20,37,40,45].

To identify other RBPs that are binding to the DplUSE we analysed with detail the molecular weights of the proteins assembled onto the DplUSE pre-mRNA (Figure 6B, lane 1 and Figure 6C). We hypothesized that HuR/ELAVL1 and hnRNPC could be binding to DplUSE . HuR is a 32 kDa RBP, ubiquitously expressed in all human tissues that recognizes and binds to U-rich and AU-rich elements present in the 3'UTR of target mRNAs modulating their stability[30,31,34,52]. Furthermore, its fruit fly orthologue ElavL1 has been shown to promote 3'UTR extensions observed in the nervous system[35]. In addition, HnRNP C1/C2 isoforms (41 and 43 kDa, respectively) bind to polyU tracts and assemble on nascent transcripts and modulate several mRNA functions, including mRNA stability[36,53,54]. Using antibodies against these proteins it was possible to immunoprecipitate these RBPs after UV cross-linking using the DplUSE pre-mRNA and HeLa cell nuclear extracts (Figure 5D-F). In contrast, when the UV cross-linking assays were performed with DplUSEmt pre-mRNA, the proteins were not as efficiently immunoprecipitated (Figure 5D). Overall, these results show that PTBP1, HuR and hnRNPC bind specifically to the DplUSE RNA in human cells, likely making part of the molecular regulatory mechanism operating at the DplUSE sequence.

To further demonstrate that the fruit fly's polo sequence TTGTTTT is responsible for the ultra-conserved binding of RBPs to the pre-mRNA in human cells, we tested an expanded version of the fruits fly's DplUSE sequence (AATTTATTTGTTTTTGCCCCTTCCCCTT). We also included specific mutations at the 5´ half of the sequence (USE-5'mt; AATT(C)AT(A)(G)(A)(C)(G)(A)(C)(A)GCCCCTTCCCCTT),

at the 3´ half (USE-3'mt; AATTTATTTGTTTTT(A)(A)C(G)C(A)(A)(G)(A)(T)CTT) and both 5´ and 3´ together (DplUSEmt-2; AATT(C)AT(A)(G)(A)(C)(G)(A)(C)(A)(A)(A)C(G)C(A)(A)(G)(A)(T)CTT; Figure S4A) and performed in vitro transcription and UV cross-linking assays using nuclear extracts from human Hela cells. These results show that proteins bind strongly only to the non-mutated extended DplUSE sequence and to the USE-3'mt sequence, therefore suggesting that the ultra-conserved bilaterian binding of RBPs is located at the DplUSE sequence (TTGTTTT; Figure S4A). To further validate these results, we performed similar assays splitting the expanded DplUSE sequence in three different fragments: DplUSE -(a), DplUSE -(b) and DplUSE -(c) (Figure S4A). Results show that the USE-(a) fragment and the expanded DplUSE sequence, the only tested fragments that contain the DplUSE sequence (TTGTTTT), bind strongly to proteins. Reinforcing the role of HuR and hnRNPC RBPs in the regulatory function of the DplUSE sequence, in both assays it was observed that the proteins which binding depends of the DplUSE sequence (TTGTTTT) have MWs similar to HuR and hnRNPC (Figure S4A).

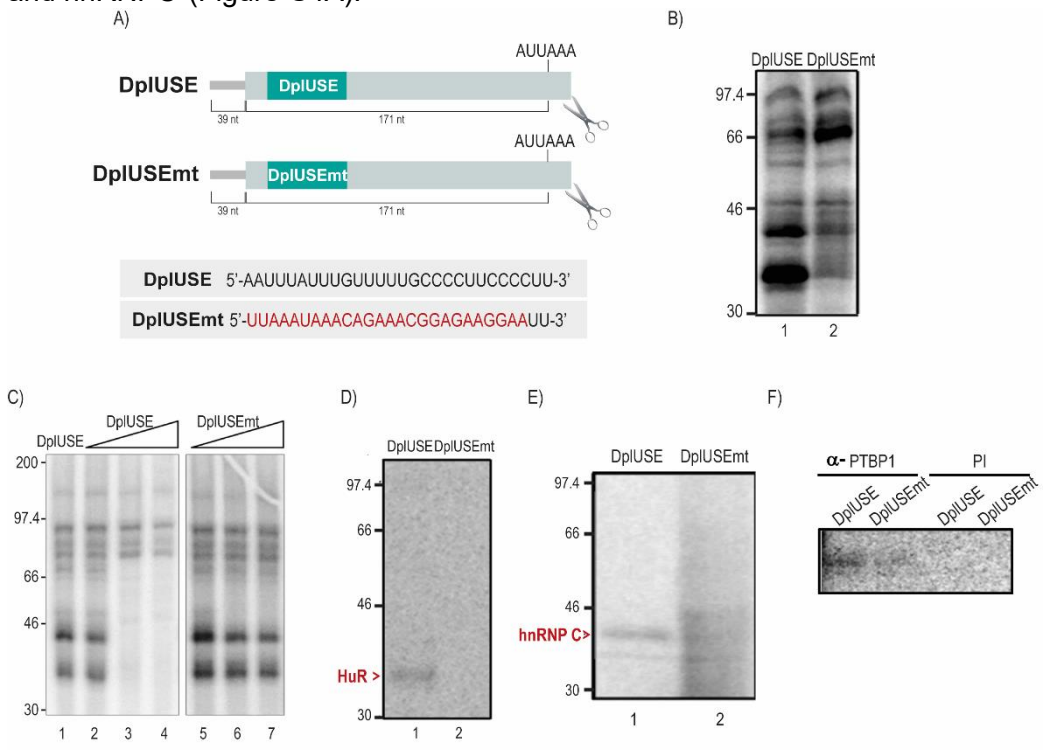


Figure 5 - HuR, hnRNP C and PTBP1 bind to the DplUSE sequence in a specific manner in human cells. *(A) Schematic representation of the pre-cleaved pre-mRNAs used to perform the UV cross-linking assays. The DplUSE sequence and the DplUSE sequence mutated in 26 nucleotides (DplUSEmt) is presented. The cleaved pre-mRNAs were obtained by linearizing the plasmids with restriction enzymes in proximity to the cleavage site (portrayed by the pair of scissors). (B) UV cross-linking assay using the pre-cleaved pre-mRNAs DplUSE and DplUSEmt incubated with HeLa cells nuclear extracts. Asterisks define the bands correspondent to the proteins that undergo a decrease in the binding upon the mutation of the DplUSE sequence to DplUSEmt. (C) UV cross-linking competition assay using pre-cleaved pre-mRNAs DplUSE and DplUSEmt with HeLa cell nuclear extracts (lane 1) and an increasing amount of unlabeled DplUSE (lane 2-4) and DplUSEmt (lane 5-7) pre-mRNAs. One (lane 2, 5), 50 (lane 3,6) and 150 pmols (lane 4,7) of unlabeled oligonucleotides were used to compete with the labeled pre-mRNAs. (D-F) Immunoprecipitations assays performed subsequently to UV cross-linking of pre-cleaved pre-mRNAs DplUSE and DplUSEmt with HeLa cell nuclear extracts. Immunoprecipitations were performed using HuR, hnRNP C and PTBP1 antibodies.*

3.5. PTBP1 is part of the molecular mechanism modulating the expression of DplUSE-containing genes

To investigate if PTBP1 is required for the control of the expression of DplUSE-containing genes, we performed PTBP1 depletion using siRNAs in HeLa cells. A knockdown of approximately 75% was obtained (Figure S5) and the expression of nine USE-containing genes (KIF20A, EXT2, CLDN12, MBNL2, MED18, NR5A2, DKC1) was quantified by RT-qPCR. These genes were selected by their potential to respond to the dominant negative effect of the ectopic expression of DplUSE (Figure 5). We observed that EXT2, CLDN12, MED18 and DKC1 mRNA levels decreased significantly upon PTBP1 knockdown, while KIF20A and NR5A2 were not clearly affected and MBNL2 mRNA levels increased (Figure 6). Our results indicate that PTBP1, as its fruit fly´s orthologue Heph, modulate the expression of DplUSE-containing genes, supporting the existence of a regulatory mechanism that operates at DplUSE and that is conserved in bilaterians.

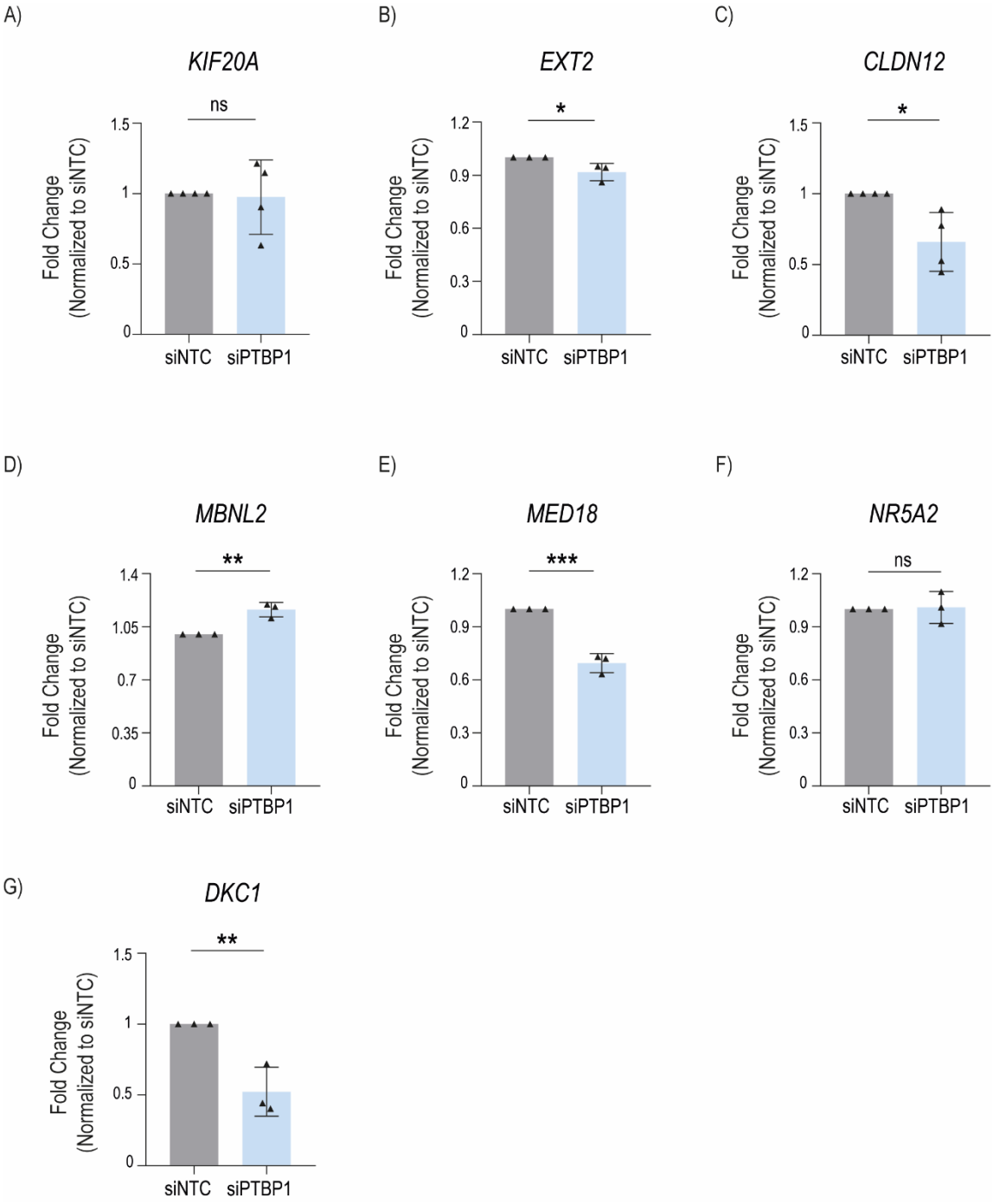
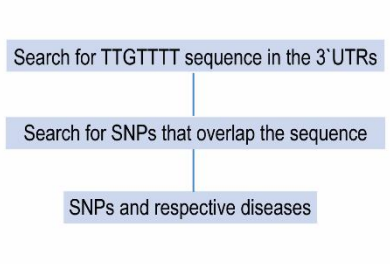


Figure 6 - PTBP1 knockdown affects the expression of human DplUSE-containing genes. *(A) PTBP1 knockdown was performed using siPTBP1 and DplUSE-containing genes mRNA levels were assessed by RT-qPCR. A siNTC pool was also transfected to be used as a control. EXT2 (B), CLDN12 (C), MED18 (E) and DKC1 (H) mRNA levels decreased upon PTBP1 knockdown, like expected, but MBNL2 (D) mRNA levels increased. For KIF20A (A) and NR5A2 (F) the difference was not statistically significant. The data is presented as the mean ± SD of at least three independent experiments analysed by two-tailed unpaired Student's t-test, *p < 0.05, **p < 0.01, ***p < 0.001 and ns – not significant.*

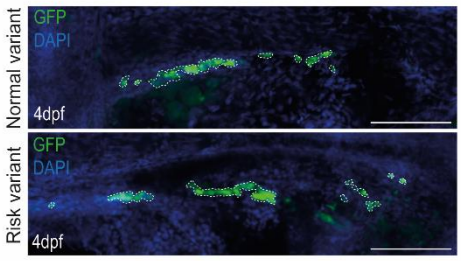
3.6. A single nucleotide polymorphism in the DplUSE sequence is associated with increased cancer risk.

In agreement with the mechanism of action of previously described USEs and the results of this study, the DplUSE RNA sequence acts as a platform for the recruitment of RBPs to regulate gene expression. Thus, modifications in the USE sequences may be involved in the deregulation of gene expression in the context of human disease. To investigate the potential impact of the DplUSE sequence in human diseases, we searched for the occurrence of single nucleotide polymorphisms (SNPs) overlapping with the DplUSE sequence (TTGTTTT) (Figure 7A). We found two genes containing a SNP associated with human diseases in the DplUSE sequence. One of these SNPs has been associated to the increased risk of development of Malignant tumor of colon (MTC) (rs3087967; POU2AF2/C11orf53[55]) and the other is associated with alanine aminotransferase blood levels after remission induction therapy (rs149940960; GREB1[56]) (Figure 7B). Of note, one of these genes, POU2AF2/C11orf53 has been recently described to act as a coactivator of POU2F3 to maintain chromatin accessibility, therefore having the potential to control genetic programs[41]. Additionally, it has been also shown that depletion of POU2AF2/C11orf53 using an in vivo xenograft model of small cell lung cancer, repressed tumour growth and delayed progression of disease in mice[41], showing the tumorigenic potential of these gene in specific cellular contexts. Importantly, the MTC risk-associated variant (rs3087967) reflects the acquisition of an ectopic DplUSE consensus, compared with the respective non-risk variant. Therefore, in agreement with our results that show increased gene expression in the presence of DplUSE consensus in the 3'UTR of genes, we asked if the rs3087967 variant, in the context of the 3'UTR of the tumorigenic POU2AF2/C11orf53 gene, has the potential to increase its expression levels. To test this, we cloned the 3`UTR of POU2AF2/C11orf53, containing the two variants (risk and non-risk) downstream of GFP, under the control of a CMV promoter. Using the Tol2 system, we performed mosaic transgenesis assays in zebrafish and analyzed the larvae at 4 dpf. Quantifying the GFP levels in gut cells, we observed that the risk variant showed higher levels of GFP expression, when comparing with the non-risk variant (Figure 7C and D). These results suggest that the acquisition of an ectopic USE in the 3`UTR of the tumorigenic POU2AF2/C11orf53 might lead to an ectopic expression of this gene and consequently, contribute to the development of MTC. These results indicate that the DplUSE may have important unidentified roles in the clinical context.

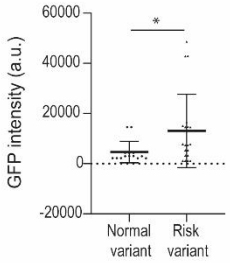
A)



C)



D)



B)

Gene	Disease	SNP	Risk variant	Sequence impact	Normal sequence
GREB1	Alanine aminotransferase measurement	rs149940960	T	TTGTTTT	TTGCTTT
C11orf53	Malignant tumor of colon	rs3087967	T	TTGTTTT	TCGTTTT

Figure 7 - A SNP in C11orf53 generates an ectopic consensus of the DplUSE which is associated with malignant tumour of colon *(A) Workflow used to identify SNPs that overlap with USE sequence and that had an associated disease. (B) DplUSE related SNPs with the respective gene, disease, sequence impact and normal sequence. (C) Representative confocal images of microinjected embryos with GFP-USE of C11orf53 with and without the presence of the risk variant, showing GFP expression in gut cells of 48hpf zebrafish embryos. The dashed lines represent different GFP expression domain within an embryo(D) Representative graph of the quantification of the GFP expression in the defined GFP expression domains (n=14, normal variant; n=20 risk variant). The images were acquired using the confocal microscope LeicaSP5II and the GFP expression quantified using IMARIS software. The statistical analysis was done by two-tailed unpaired Student's t-test, *p < 0.05, **p < 0.01, ***p < 0.001.*

4. Discussion

USEs are characterized as U-rich *cis*-auxiliary elements capable of modulating mRNAs 3'-end formation by acting as an extra-platform for the recruitment of *trans*-acting factors, such as RBPs[5,13,17,18,20,26,37,40,45,63–65]. As a result, USEs are considered to be essential to ensure a proper control of gene activity. These elements were initially described in retrovirus, such as the HIV-1, in the selective use of the PAS at the 3' of the duplicated viral transcripts[21,24,66]. It was proposed that the existence of USEs in HIV-1 revealed the need for a level of regulation that is not normally required by other PAS. However, this concept was challenged when USEs were identified in cellular genes[14,15,19,22,26,29] and it is now widely accepted that the general arrangement of mRNA 3'-end formation signals in eukaryotes contains an USE associated with a PAS[67].

In this work we found that a short motif (DplUSE) located in the USE of the *Drosophila melanogaster's polo* gene[14] is also present in highly divergent vertebrate species as zebrafish, mouse and human. Importantly, we show that this motif has conserved functions in gene regulation in the fruit fly, zebrafish and humans, being required to sustain high levels of gene activity in distantly related bilaterian species. Further supporting that this motif may participate in an ultra-conserved mechanism, when analyzing the molecular

machinery that operates at the DplUSE sequence, we found that several RBPs bind specifically to this motif in human cells and that one of these RBPs is the same orthologue as the one found to bind to and to be required for the DplUSE function in the fruit fly[14]. These results support the existence of a very ancestral mechanism common to highly divergent bilaterians that might have important biological roles. Indeed, this seems to be the case, since the DplUSE sequence on average shows higher levels of conservation in vertebrates than their respective 3'UTRs, suggesting a selective pressure to keep this motive or alternative ones with similar functions in the genome. Nevertheless, as only 31 orthologue genes common to zebrafish, mouse and human have the canonical DplUSE, it is likely the existence of alternative motifs. Supporting this hypothesis, previous studies have identified a similar functional motif in the human gene *prothrombin F2*, diverging in only one nucleotide, comparing to the DplUSE motif (TTGT[G/T]TTT)[29], and when analyzing the conservation of the DplUSE motif across vertebrate species, we observed that some nucleotides are more conserved than others.

DplUSE motif seems to be essential in many biological contexts. In *Drosophila* it was demonstrated that its absence dysregulates the cell cycle gene *polo*, causing severe phenotypes in mitosis in neuroblasts and abdominal segments of the adult fly[14]. In the current work we show that using a dominant negative for the molecular players that operate at the DplUSE sequence, causes severe developmental phenotypes in zebrafish embryos, suggesting that DplUSE controlled genes and the molecular players that operate at the DplUSE are important for vertebrate development. Concomitantly, a similar dominant negative strategy applied to human cells lead to downregulation of 9 out of 9 tested genes (2 of them showing a tendency), from the pool of the 31 DplUSE-containing genes that are common to zebrafish, mouse and humans and that are enriched for human diseases such as Nervous system diseases, Congenital abnormalities, and Malignant Neoplasms. EXT2, for instance, is a glycosyltransferase deregulated in the EXT2-related syndrome where one of the manifestations of the disease is microcephaly [73], a phenotype partially recapitulated in zebrafish embryos microinjected with DplUSE RNA. *CLDN12* is another DplUSE-containing gene, which is overexpressed in colorectal carcinomas[68]. Further suggesting that DplUSE controlled genes have important relevant functions to human health, we found that the 2110 human DplUSE containing genes are enriched for several diseases, including Congenital Abnormalities, Colorectal Cancer and Malignant Neoplasm. These results lead us to postulate that alterations in the DplUSE sequence might be at the basis of the development of human disease. This is the case of the colorectal cancer associated SNP rs3087967, located at the 3'UTR of the gene POU2AF2/C11orf53 known to have a tumorigenic potential, and which mechanistic association to colorectal cancer development

has not yet been understood. We show that rs3087967 generates an ectopic DplUSE motif that enhances mRNA activity in intestine cells in vivo, being a reasonable mechanistic explanation for the involvement of rs3087967 in colorectal cancer development.

Overall, our study has unveiled the emerging importance of the DplUSE motif, identified in *Drosophila*, and that now we show to have a conserved function across vertebrates and that likely represents an ultra-conserved mechanism of gene expression control. We also show that this motif and the associated ultra-conserved mechanism of gene regulation is of extreme importance to understand several human diseases. The data underlying this article are available in the article and in its online supplementary material.

5. Acknowledgements

We thank all the current and past members of the Gene Regulation and the Vertebrate Development and Regeneration research groups for their contributions and critical discussions. We are very grateful to Gideon Dreyfuss (Howard Hughes Medical Institute, University of Pennsylvania School of Medicine, USA) and Chris Smith (Cambridge University, UK) for kind gifts of antibodies. The laboratory of AMoreira was funded by National Funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the project UIDB/04293/2020, by Programa Gilead Génese, ref 13854, by FCT under the project EXPL/SAU-PUB/1073/2021 (http://doi.org/10.54499/EXPL/SAU-PUB/1073/2021) and by Programa Operacional Regional do Norte and co-funded by European Regional Development Fund under the project “The Porto Comprehensive Cancer Center” with the reference NORTE-01–0145-FEDER-072678—Consórcio PORTO.CCC—Porto Comprehensive Cancer Center. The laboratory of JBessa was supported by: European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-2015-StG1077 680156-ZPR); ”La Caixa” Foundation (under the grant agreement HR21-01212); Portuguese funds through Fundação para a Ciência e a Tecnologia (FCT) in the framework of the project PTDC/BIA-MOL/3834/2021. JB is supported by the FCT grant CEECIND/03482/2018. Ana Eufrásio and Joana Azevedo are PhD students (FCT studentships SFRH/BD/147762/2019 to AEufrásio and 2022.14339.BD to JAzevedo). IP-C is supported by a FCT Researcher contract (DL 57/2016/CP1355/CT0016).

Conflict of interest: The authors declare that they have no potential conflicts of interest.

6. References

1. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171–194 (2017).

2. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).

3. Alternative polyadenylation of mRNA precursors | Nature Reviews Molecular Cell Biology. https://www.nature.com/articles/nrm.2016.116.

4. Szostak, E. & Gebauer, F. Translational control by 3'-UTR-binding proteins. *Brief Funct Genomics* **12**, 58–65 (2013).

5. Pereira-Castro, I. & Moreira, A. On the function and relevance of alternative 3'-UTRs in gene expression regulation. *Wiley Interdiscip Rev RNA* **12**, e1653 (2021).

6. Gebauer, F., Schwarzl, T., Valcárcel, J. & Hentze, M. W. RNA-binding proteins in human genetic disease. *Nat Rev Genet* **22**, 185–198 (2021).

7. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).

8. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**, 853–866 (2011).

9. Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3**, 385–396 (2012).

10. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**, 2054–2066 (2012).

11. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J.-M. & Gautheret, D. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Res* **10**, 1001–1010 (2000).

12. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**, 201–212 (2005).

13. Darmon, S. K. & Lutz, C. S. Novel upstream and downstream sequence elements contribute to polyadenylation efficiency. *RNA Biol* **9**, 1255–1265 (2012).

14. Oliveira, M. S. *et al.* Cell Cycle Kinase Polo Is Controlled by a Widespread 3' Untranslated Region Regulatory Sequence in Drosophila melanogaster. *Mol Cell Biol* **39**, e00581-18 (2019).

15. Moreira, A., Wollerton, M., Monks, J. & Proudfoot, N. J. Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *The EMBO Journal* **14**, 3809 (1995).

16. Kaufmann, I., Martin, G., Friedlein, A., Langen, H. & Keller, W. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23**, 616–626 (2004).

17. Danckwardt, S. *et al.* p38 MAPK controls prothrombin expression by regulated RNA 3' end processing. *Mol Cell* **41**, 298–310 (2011).

18. Lutz, C. S. & Alwine, J. C. Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal. *Genes Dev* **8**, 576–586 (1994).

19. Natalizio, B. J., Muniz, L. C., Arhin, G. K., Wilusz, J. & Lutz, C. S. Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. *J Biol Chem* **277**, 42733–42740 (2002).

20. Wilusz, J. & Shenk, T. A uridylate tract mediates efficient heterogeneous nuclear ribonucleoprotein C protein-RNA cross-linking and functionally substitutes for the downstream element of the polyadenylation signal. *Mol Cell Biol* **10**, 6397–6407 (1990).

21. Gilmartin, G. M., Fleming, E. S. & Oetjen, J. Activation of HIV-1 pre-mRNA 3' processing in vitro requires both an upstream element and TAR. *EMBO J* **11**, 4419–4428 (1992).

22. Hall-Pogar, T., Zhang, H., Tian, B. & Lutz, C. S. Alternative polyadenylation of cyclooxygenase-2. *Nucleic Acids Research* **33**, 2565–2579 (2005).

23. Nedeljkovic, M., Costessi, L., Iaconcig, A., Porro, F. & Muro, A. F. Long-distance regulation of Add2 pre-mRNA3'end processing. *RNA Biol* **10**, 516–527 (2013).

24. DeZazzo, J. D., Kilpatrick, J. E. & Imperiale, M. J. Involvement of long terminal repeat U3 sequences overlapping the transcription control region in human immunodeficiency virus type 1 mRNA 3' end formation. *Mol Cell Biol* **11**, 1624–1630 (1991).

25. Hinzman, E. E., Barr, J. N. & Wertz, G. W. Identification of an Upstream Sequence Element Required for Vesicular Stomatitis Virus mRNA Transcription. *J Virol* **76**, 7632–7641 (2002).

26. Brackenridge, S. & Proudfoot, N. J. Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. *Mol Cell Biol* **20**, 2660–2669 (2000).

27. Llamazares, S. *et al.* polo encodes a protein kinase homolog required for mitosis in Drosophila. *Genes Dev* **5**, 2153–2165 (1991).

28. Pinto, P. A. B. *et al.* RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J* **30**, 2431–2444 (2011).

29. Danckwardt, S. *et al.* The prothrombin 3'end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. *Blood* **104**, 428–435 (2004).

30. Ma, W. J., Cheng, S., Campbell, C., Wright, A. & Furneaux, H. Cloning and characterization of HuR, a ubiquitously expressed Elav-like protein. *J Biol Chem* **271**, 8144–8151 (1996).

31. Hinman, M. N. & Lou, H. Diverse molecular functions of Hu proteins. *Cell Mol Life Sci* **65**, 3168–3181 (2008).

32. Rothamel, K. *et al.* ELAVL1 primarily couples mRNA stability with the 3' UTRs of interferon-stimulated genes. *Cell Rep* **35**, 109178 (2021).

33. Brennan, C. M. & Steitz, J. A. HuR and mRNA stability. *Cell Mol Life Sci* **58**, 266–277 (2001).

34. López de Silanes, I., Zhan, M., Lal, A., Yang, X. & Gorospe, M. Identification of a target RNA motif for RNA-binding protein HuR. *Proc Natl Acad Sci U S A* **101**, 2987–2992 (2004).

35. Hilgers, V., Lemke, S. B. & Levine, M. ELAV mediates 3' UTR extension in the Drosophila nervous system. *Genes Dev* **26**, 2259–2264 (2012).

36. Görlach, M., Burd, C. G. & Dreyfuss, G. The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. *J Biol Chem* **269**, 23074–23078 (1994).

37. Lou, H., Gagel, R. F. & Berget, S. M. An intron enhancer recognized by splicing factors activates polyadenylation. *Genes Dev* **10**, 208–219 (1996).

38. Sawicka, K., Bushell, M., Spriggs, K. A. & Willis, A. E. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* **36**, 641–647 (2008).

39. Linares, A. J. *et al.* The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife* **4**, e09268 (2015).

40. Moreira, A. *et al.* The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3′ end formation by two distinct mechanisms. *Genes Dev.* **12**, 2522–2534 (1998).

41. Szczepanski, A. P. *et al.* POU2AF2/C11orf53 functions as a coactivator of POU2F3 by maintaining chromatin accessibility and enhancer activity. *Sci Adv* **8**, eabq2403 (2022).

42. Arneson, A. & Ernst, J. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun Biol* **2**, 1–14 (2019).

43. Kawakami, K. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biology* **8**, S7 (2007).

44. Richardson, R., Tracey-White, D., Webster, A. & Moosajee, M. The zebrafish eye-a paradigm for investigating human ocular genetics. *Eye (Lond)* **31**, 68–86 (2017).

45. Castelo-Branco, P. *et al.* Polypyrimidine Tract Binding Protein Modulates Efficiency of Polyadenylation. *Mol Cell Biol* **24**, 4174–4183 (2004).

46. Danckwardt, S. *et al.* Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals. *The EMBO Journal* **26**, 2658–2669 (2007).

47. Singh, R., Valcárcel, J. & Green, M. R. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science* **268**, 1173–1176 (1995).

48. Southby, J., Gooding, C. & Smith, C. W. Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons. *Mol Cell Biol* **19**, 2699–2711 (1999).

49. Wagner, E. J. & Garcia-Blanco, M. A. RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol Cell* **10**, 943–949 (2002).

50. Boutz, P. L. *et al.* A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21**, 1636–1652 (2007).

51. Zhang, L., Liu, W. & Grabowski, P. J. Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB. *RNA* **5**, 117–130 (1999).

52. Srikantan, S. & Gorospe, M. HuR function in disease. *Front Biosci (Landmark Ed)* **17**, 189–205 (2012).

53. Hamilton, B. J., Nagy, E., Malter, J. S., Arrick, B. A. & Rigby, W. F. Association of heterogeneous nuclear ribonucleoprotein A1 and C proteins with reiterated AUUUA sequences. *J Biol Chem* **268**, 8881–8887 (1993).

54. Barnett, S. F., Friedman, D. L. & LeStourgeon, W. M. The C proteins of HeLa 40S nuclear ribonucleoprotein particles exist as anisotropic tetramers of (C1)3 C2. *Mol Cell Biol* **9**, 492–498 (1989).

55. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

56. Liu, Y. *et al.* Genome-Wide Study Links PNPLA3 Variant With Elevated Hepatic Transaminase After Acute Lymphoblastic Leukemia Therapy. *Clin Pharmacol Ther* **102**, 131–140 (2017).

57. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**, bar030 (2011).

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

59. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* **48**, D845–D855 (2020).

60. Kaminski, A., Hunt, S. L., Patton, J. G. & Jackson, R. J. Direct evidence that polypyrimidine tract binding protein (PTB) is essential for internal initiation of translation of encephalomyocarditis virus RNA. *RNA* **1**, 924–938 (1995).

61. Karlsson, J., von Hofsten, J. & Olsson, P. E. Generating transparent zebrafish: a refined method to improve detection of gene expression during embryonic development. *Mar. Biotechnol.* **3**, 522–527 (2001).

62. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676–682 (2012).

63. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171–194 (2017).

64. Graber, J. H., Cantor, C. R., Mohr, S. C. & Smith, T. F. In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species. *Proceedings of the National Academy of Sciences* **96**, 14055–14060 (1999).

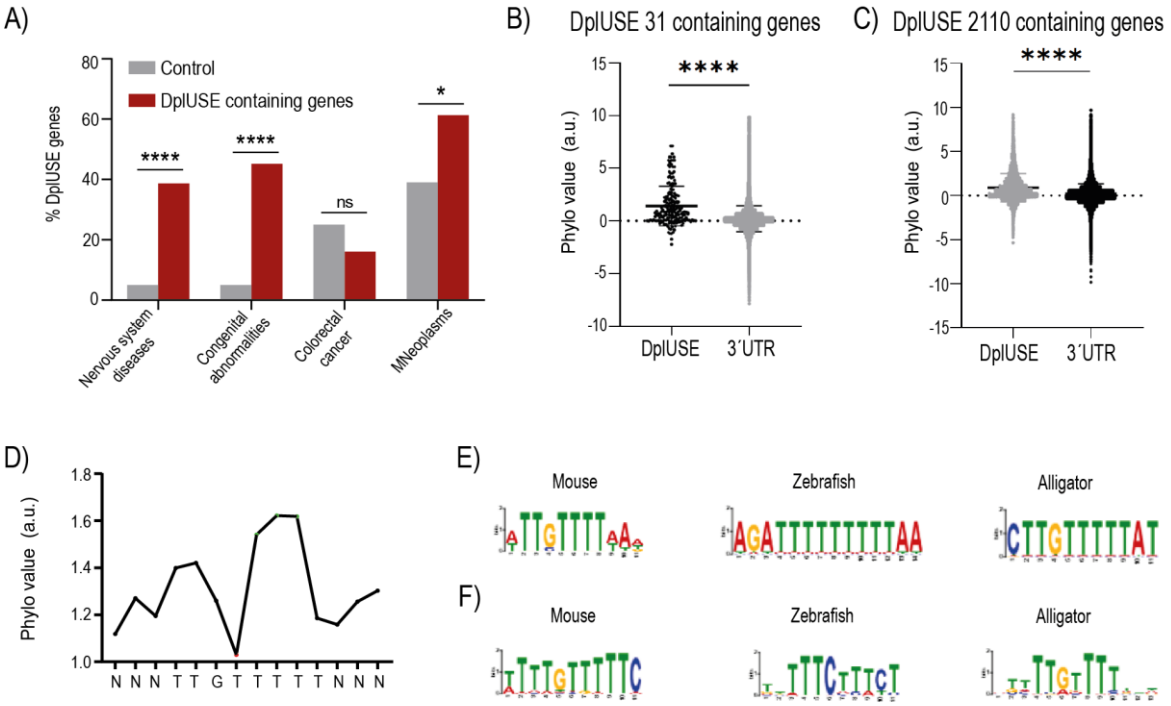
65. Legendre, M. & Gautheret, D. Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4**, 7 (2003).

66. Weichs an der Glon, C., Monks, J. & Proudfoot, N. J. Occlusion of the HIV poly(A) site. *Genes Dev* **5**, 244–253 (1991).

67. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev* **25**, 1770–1782 (2011).

68. Alghamdi, R. A. & Al-Zahrani, M. H. Identification of key claudin genes associated with survival prognosis and diagnosis in colon cancer through integrated bioinformatic analysis. *Front Genet* **14**, 1221815 (2023).

7. Supplementary material



G)

Homo sapiens DplUSE-containing genes GO Terms	Number of genes	Fold enrichment	FDR
import into nucleous (GO:0051170)	24	2.33	4.53E-02
homophilic cell adhesion via plasma membrane adhesion molecules (GO: 0007156)	35	2.14	1.65E-02
regulation of mRNA processing (GO:0050684)	29	2.13	4.79E-02
regulation of organelle assembly (GO:1902115)	39	2.11	1.12E-02
cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742)	48	1.9	1.29E-02
positive regulation of cell migration (GO:0030335)	89	1.74	7.50E-04
regulation of Wnt signalling pathway (GO:0030111)	54	1.71	4.42E-02
response to growth factor (GO:0070848)	80	1.69	5.06E-03
regulation of cell migration (GO:0030334)	147	1.67	1.42E-05
positive regulation of cell motility (GO:2000147)	89	1.66	3.29E-03
cell-cell adhesion (GO:0098609)	84	1.66	4.74E-03
cellular response to growth factor stimulus (GO:0071363)	74	1.66	1.24E-02
covalent chromatin modifications (GO:0016569)	59	1.66	4.61E-02
mRNA processing (GO:0006397)	70	1.64	2.52E-02
positive regulation of locomotion (GO:0040017)	90	1.63	5.10E-03

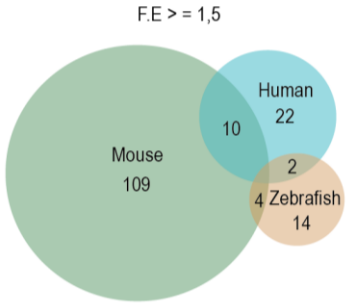
H)

Mus musculus DplUSE-containing genes GO Terms	Number of genes	Fold enrichment	FDR
multivesicular body sorting pathway (GO:0071985)	11	3.79	3.59E-02
cellular response to glucose stimulus (GO:0071333)	17	3.75	2.28E-03
cellular response to hexose stimulus (GO:0071331)	17	3.62	3.11E-03
cellular response to monosaccharide stimulus(GO:0071326)	17	3.55	3.65E-03
cellular response to carbohydrate stimulus (GO:0071322)	18	3.29	4.74E-03
Golgi to plasma membrane transport (GO:0006893)	14	3.15	3.35E-02
cellular glucose homeostasis (GO:0001678)	21	3.11	2.74E-03
regulation of epithelial to mesenchymal transition (GO:0010717)	22	2.77	6.58E-03
response to glucose (GO:0009749)	22	2.63	1.50E-02
post-Golgi vesicle-mediated transport (GO:0006892)	20	2.6	2.97E-02
response to hexose (GO:0009746)	22	2.55	1.87E-02
response to monosaccharide (GO:0034284)	22	2.52	2.03E-02
response to carbohydrate (GO:0009743)	23	2.32	4.03E-02
vesicle-mediated transport to the plasma membrane (GO:0098876)	23	2.26	4.89E-02
establishment of vesicle localization (GO:0051650)	27	2.23	2.84E-02

I)

Danio rerio DplUSE-containing genes GO Terms	Number of genes	Fold enrichment	FDR
positive regulation of protein ubiquitination (GO:0031398)	10	6.74	2.80E-03
ribonucleoside metabolic process (GO:0009119)	7	6.44	3.63E-02
positive regulation of protein modification by small protein conjugation or removal (GO:1903322)	10	6.13	4.20E-03
regulation of protein ubiquitination (GO:0031396)	12	5.17	3.28E-03
regulation of protein modification by small protein conugation or removal (GO:1903320)	13	5.06	2.27E-03
cytosolic transport (GO:0016482)	12	3.62	3.62E-02
Notch signalling pathway (GO:0007219)	13	3.25	4.72E-02
endocrine system development (GO:0035270)	18	3.09	1.10E-02
connective tissue development (GO:0061448)	20	2.83	1.29E-02
cartilage development (GO:0051216)	19	2.83	1.79E-02
protein polyubiquitination (GO:0000209)	20	2.56	4.76E-02
organelle fission (GO:0048285)	22	2.5	3.13E-02
small molecule biosynthetic process (GO:0044283)	34	2.21	8.38E-02
mitotic cell cycle process (GO:1903047)	30	2.17	2.60E-02
nucleobase-containing compound biosynthetic process (GO:0034654)	49	2.13	1.37E-03

J)



K)

	Human	Mouse	Zebrafish
positive regulation of cell migration (GO:0030335)	•	•	
mRNA processing (GO:0006397)	•	•	
positive regulation of cell motility (GO:2000147)	•	•	
response to growth factor (GO:0070848)	•	•	
positive regulation of locomotion (GO:0040017)	•	•	
cell division (GO:0051301)	•	•	
positive regulation of cellular component movement(GO:0051272)	•	•	
cellular response to growth factor stimulus (GO:0071363)	•	•	
chromatin organization (GO:0006325)	•	•	
mRNA metabolic process (GO:0016071)	•	•	
chromosome organization (GO:0051276)		•	•
establishment of localization in cell (GO:0051649)		•	•
cellular protein localization (GO:0034613)		•	•
cellular macromolecule localization (GO:0070727)		•	•
mitotic cell cycle (GO:0000278)	•		•
mitotic cell cycle process (GO:1903047)	•		•

Figure S1 - *(A) Representative graph showing the percentage of the 31 ortholog DplUSE-containing genes that are associated with the respective disease (DisGenet data source). (B) Representative graph of the extracted conservation values from each nucleotide within the USE and 3´UTR sequence, for the 31 ortholog genes (data source: 100vertebrates) (C) Representative grpah of the extracted conservation values from each nucleotide within the USE and 3´UTR sequence, for the 2110 human genes (data source: 100vertebrates) (D) Representative graph showing the average conservation value for each nucleotide within the 31 DplUSE containing genes. (E) Discovered motifs found from the alignment of the converted DplUSEs from the 31 ortholog genes, in mouse, zebrafish and american alligator. (F) Discovered motifs found from the alignment of the converted DplUSEs from the 2110 human genes, in mouse, zebrafish and american alligator. (G) DplUSE-containing genes GO terms from human, mouse (H) and zebrafish (I). (J) Venn diagram representing the common GOs highly enriched between human, zebrafish and mouse (F.E≥1.5). A total of 10 common GO terms were found between human and mouse, 4 between zebrafish and mouse and 2 between human and zebrafish. (K)Table representing the GO terms found to be highly enriched in human, zebrafish and mouse.*

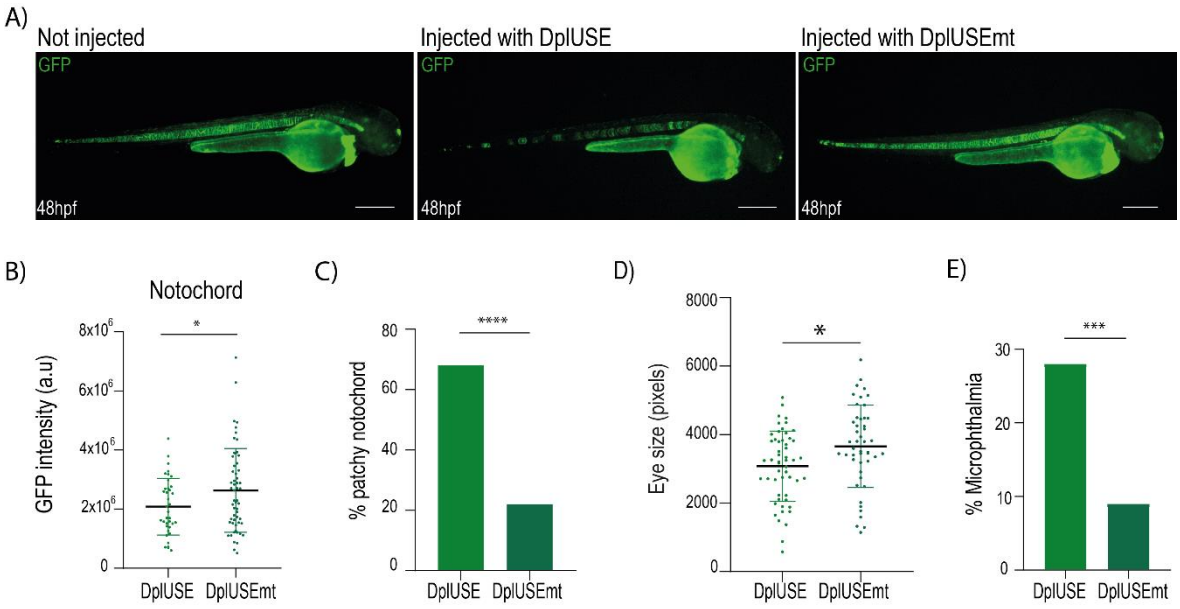


Figure S2 - *Microinjection of DplUSE RNA causes a decrease in GFP expression and microphthalmia in transgenic GFP-DplUSE animals, at 48hpf. (A) Representative image of GFP expression in transgenic GFP-DplUSE zebrafish embryos not microinjected and microinjected with USE RNA or USEmt, at 48hpf. Images were acquired using Leica M205.(B) Quantification of GFP expression in notochord (n=37 DplUSE; n=57 DplUSEmt).(C) Quantification of the percentage of transgenic GFP-USE embryos that showed a patchy expression of GFP in the notochord (68%,n=79 DplUSE; 22%,n=61 DplUSEmt). (D) Quantiffication of the eye size (n=53 DplUSE and n=47 DplUSEmt) and (E) microphthalmia (28%, n=83 DplUSE; 9%, n=91 DplUSEmt) of transgenic GFP-USE embryos upon the microinjection of DplUSE RNA and DplUSEmt RNA, at 48hpf. Statistical significance was determined by χ2 test with Fisher correction or by two tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05. (Scale bar = 100μm).*

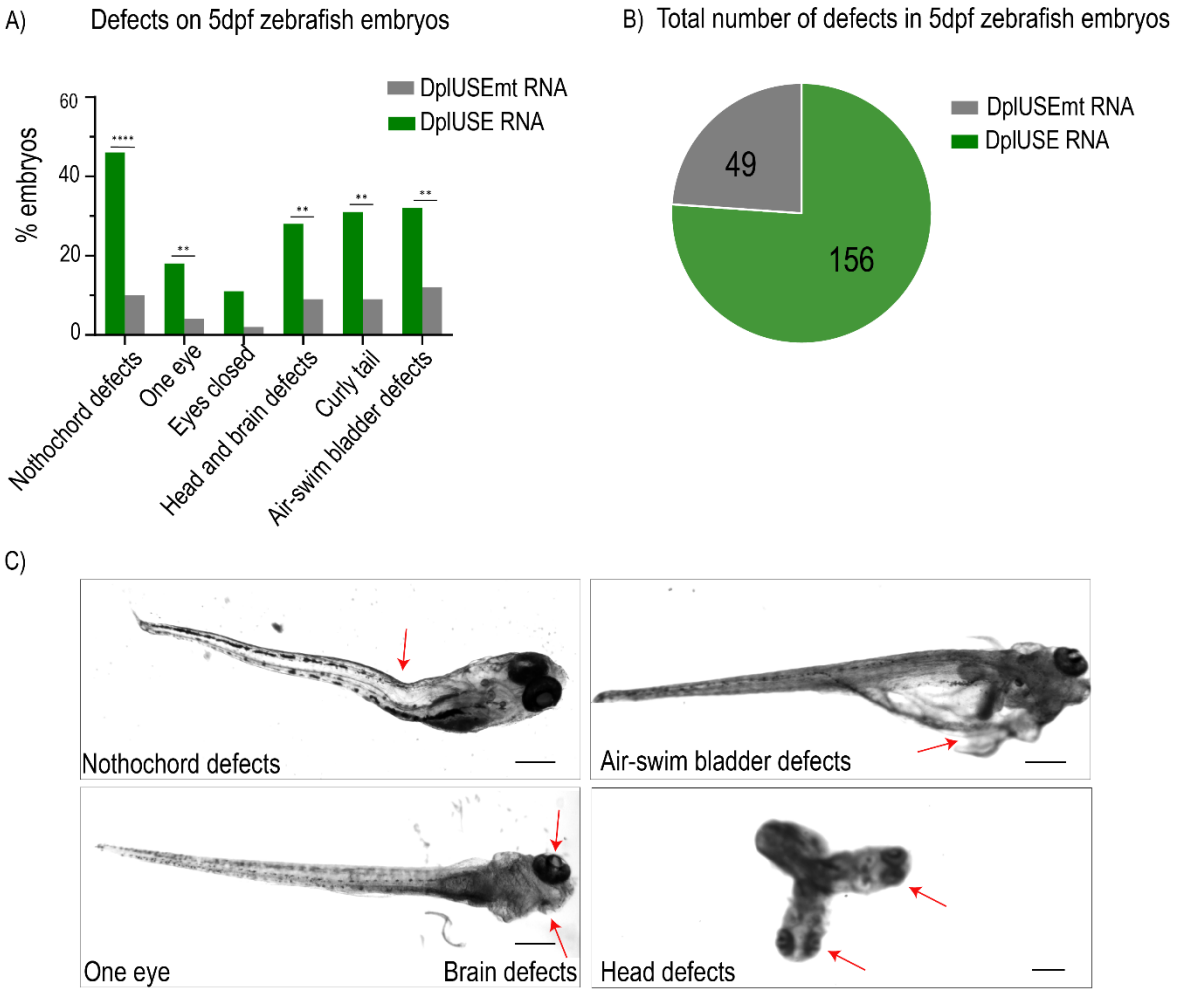


Figure S3 - *Microinjection of GFP-DplUSE RNA in one-cell stage embryos leads to the development of various abnormalities 5 days upon the microinjection procedure.(A) Quantification of the several defects developed 5 days upon the microinjection of GFP-DplUSE RNA (in green) and GFP-DplUSEmt RNA (in grey) in one-cell stage zebrafish embryos. Statistical significance was determined using χ2 with Fisher correction. (B) Pie chart representing the total number of defects developed upon the microinjection of GFP-DplUSE RNA and GFP-DplUSEmt RNA. (C) Representative images showing the various abnormalities upon microinjection of GFP-DplUSE RNA at 5 dpf, indicated with red arrows. Scale bar = 100 µm. Statistical significance was determined by χ2 test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05 Images acquired with Leica M205.Scale bar = 100 µm.*

A)

DplUSE	AAUUUAUUUGUUUUUGCCCCUUCCCCUU
DplUSEmt-2	AAUUCAUAGACGACAAACGCAAGAUCUU
DplUSE-5`mt	AAUUCAUAGACGACAGCCCCUUCCCCUU
DplUSE-3`mt	AAUUUAUUUGUUUUUAACGCAAGAUCUU
DplUSE-(a)	AAUUUAUUUGUUUUU
DplUSE-(b)	GUUUUUGCCCC
DplUSE-(c)	GCCCCUUCCCCUU

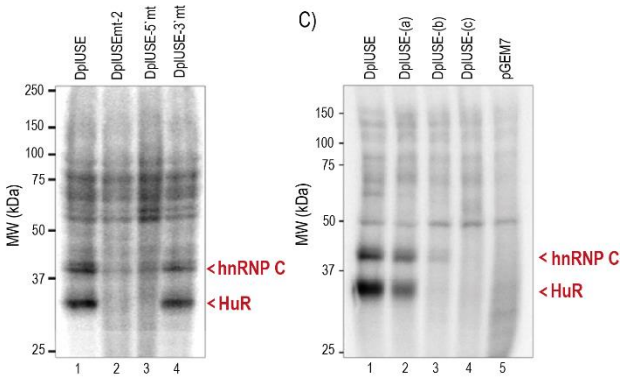


Figure S4 - *To map the binding sites in the DplUSE, several RNAs were in vitro transcribed: A) DplUSEmt-2, DplUSE-5'mt, DplUSE-3'mt, DplUSE-(a), DplUSE-(b) and DplUSE-(c). B-C) UV. cross-linking assays were performed using these RNAs with HeLa cell nuclear extracts. As a positive and negative controls an DplUSE RNA, containing the wild-type DplUSE sequence (lanes 1) and a RNA transcribed from a linearized pGEM7 polylinker (lane 5) were used, respectively.*

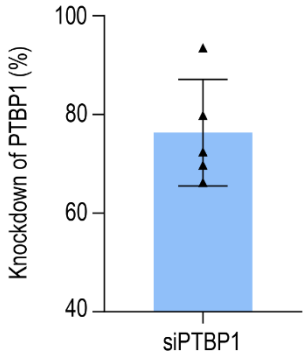


Figure S5 - *PTBP1 gene expression was efficiently silenced in HeLa cells. In order to knockdown PTBP1, HeLa cells were transfected with two siRNAs targeting PTBP1. A siNTC was also transfected, to act as a negative control. mRNA levels of PTBP1 were assessed by RT-qPCR, showing a mean knockdown of ~75%.*

A)

Oligonucleotides	Sequence
FW_useORF	CCGCTCGAGTAATAAGGCCAAATGACTTCTGGG
RV_useORF	CGGGGTACCACAAAGAGCAGCATTGAGG
FW_useORF_SiteDirectMut	CCTAAAACAATCTTTGGAATACACC
RV_useORF_SiteDirectMut	GGTGTATTCCAAAGATCGTTTTAGG
CSNK1A1_F	AAAGCAGAAAGCAGCACAGCAG
CSNK1A1_R	TTGTCAGTTTGCTTGCCTGTGG
KIF20A_F	GCAATGAACGGGGAATTGGC
KIF20A_R	GGATGCCTGTCCCACTTCTG
MAP4K5_F	CAGACAGGGTTGTCGTTTTGG
MAP4K5_R	CATGTCCAGCCAAGATGTAGAGA
CLDN12_F	TGGCAACCATTGTACTCCCA
CLDN12_R	TTCAATGGCAGAGAGGCGAG
EXT2_F	GACCTCTGACGAGCTGCAAT
EXT2_R	GTCCCAGAGATGCAGACGAC
MED18_F	CAAGCCAGCCCATTTGTTCTC
MED18_R	TGTCTCCCATTTCTGGCTGTC
DKC1_F	TATGTGGGGATTGTCCGGCT
DKC1_R	CTGCAGCAATAAGTGGGGGT
MBNL2_F	GGCCGTTGTTCGAGAGAGAA
MBNL2_R	AAGCATTGCTGCTGCAGTTT
NR5A2_F	CCCTATACCAGCTCACCCGA
NR5A2_R	AGATGTGGGATGCTTGCTGG
eGFP_F	GAGCTGAAGGGCATCGACTT
eGFP_R	TTCTGCTTGTCGGCCATGAT
PTBP1_F	TCCAGAAGGACCGCAAGATG
PTBP1_R	TCGTGGTTGTGCAGGTCAAT

B)

Antibodies
Anti-caspase-3, cleaved (Ab-2) Rabbit pAB (PC679 - Merck)
PUREBLU™DAPI (1351303 - Bio-Rad)

Supplementary table 1 - *(A) Oligonucleotides and (B) antibodies used in this study.*

CHAPTER 5

CONCLUDING REMARKS

Currently, the study of human genetic diseases is starting to be centered on unraveling the regulatory regions embedded within the genome, as a result of 3 main factors: the increasing number of diseases with no mutations in protein-coding genes[1], the strong evidence of the enrichment of disease-associated single nucleotide polymorphisms (SNPs) in regulatory regions, through genome-wide-association studies (GWAS)[2] and the development of new and reliable techniques to identify *cis*-regulatory elements (CREs), in parallel with recent advances in high-throughput sequencing methods.

Type 2 diabetes (T2D) is a complex and multifactorial disease that in part affects endocrine pancreatic function[3–5]. Several evidences point to a strong genetic predisposition for T2D, including GWAS, that identified various T2D-associated SNPs in the past few years[3,6–8]. The vast majority of these SNPs are located in regions that overlap with enhancer regions[9], predicted by techniques such as chromatin immunoprecipitation with sequencing (ChIP-seq) and assay for transposase-accessible chromatin with sequencing (ATAC-seq). These findings arise the hypothesis that SNPs, being located in enhancers, might disrupt transcription binding sites (TFBS) and, consequently, impair the enhancer activity and the transcription of their target gene. However, in which extent a SNP can impact the enhancer activity and gene transcription is underexplored. In chapter II, we addressed the possible impact of T2D-associated SNPs in enhancer regions. Using mosaic transgenesis in zebrafish, we confirmed the functionality of 6 out of 10 tested sequences as human endocrine pancreas enhancers. Among these, 2 also exhibited enhanced activity for pancreatic progenitors, highlighting the precision of enhancer predictions. When in the presence of T2D-associated SNPs, we observed a significant decrease in enhancer activity in 2 sequences, whereas in 3 showed the opposite effect, suggesting that a disease-associated SNP is capable to induce loss or gain-of-function in target genes. In one case, we provided additional evidence of increased binding affinity of Nkx6.1 with the risk variant sequence, indicating that the differential binding of transcription factors (TFs) may explain the enhancer activity variations. In another specific case, we were able to find one coding SNP (rs13266634) that causes an aminoacid substitution that resulted in a higher enhancer activity when comparing with the non-risk sequence, suggesting that coding SNPs have also the capability to modulate the activity of overlapping enhancers. This SNP is associated with a higher increased risk for T2D due to a reduced protein (SLC30A8) activity, however, it has been described a protective effect of SNPs in this particular gene (*SLC30A8*) for T2D and that the loss-of-function of SLC30A8 protein results in higher level of insulin secretion. In light of these results, our findings propose an alternative explanation: an increase in gene transcription, due to a more active enhancer, resulted for the presence of rs13266634 SNP. Furthermore, the observed increase in enhancer activity can be reversed by the presence

of 3 other common SNPs and a mutation disrupting a PDX1 predicted binding site results in a complete absence of enhancer activity, proving the complexity and the impact that single nucleotide variations (SNVs) can have in enhancer activity and gene transcription. Overall, in chapter II, we demonstrated that T2D-associated- SNPs can modulate enhancer activity and we open new avenues for understanding the code within regulatory regions, highlighting the potential impact on human disease. Additionally, this chapter highlights the robustness of zebrafish as an in vivo model to study *cis*-regulation in human diseases.

Next, we reinforced and updated the tested dataset of sequences for enhancer activity, in chapter III. We were able to validate approximately, 40% of more than 70 tested sequences and from 33 pair of non-risk and risk sequences, we identified 7 pairs where the T2D risk variant has shown a decreased enhancer activity, while for 9 pairs, the T2D risk variants have demonstrated an increased enhancer activity. The chapter II arose one important question and led us to explore a field that remains to be completely understood, coding enhancers, through seq5215 and seq5215mm, located in *KCNJ11* and *kcnj11* genes, respectively. We demonstrated that a SNV, while leaving the aminoacid unaltered, can modulate enhancer activity. The dual function of coding enhancers, although already described [10,11], is still an area lacking exploration, due to the existing gap in studies regarding the impact of disease-associated SNPs, leading to silent mutations, in the regulatory function. Our results can give new insights in this field and understand new mechanisms that might have impact in human disease.

Genes are also regulated post-transcriptionally, through CREs located within the 3´ untranslated region (3´UTR), that serve as binding platforms of RNA binding proteins (RBPs), described to be crucial in human diseases[12–14]. Together, CREs and RBPs allow the usage of alternative polyadenylation signals (APAS), contributing for the differential expression of target genes and consequently for cell-specific expression patterns and different cellular functions[15]. One of these CREs is the upstream sequence element (USE), that has been well characterized in humans. Although 3´UTR genetic mutations in human disease are well described, mutations specifically in USE are underexplored. Given the importance of USE, we hypothesize that this 28-nucleotides sequence could be conserved across species and that its disruption could contribute for disease. In chapter IV, we identified, for the first time, a USE in zebrafish, that it is conserved across humans, mice and fruit fly, suggesting a crucial biological role. Exploring this USE, we found that it acts in a post-transcriptionally manner, due to its capability of increase GFP expression in zebrafish. By introducing exogenous USE RNA in zebrafish, we observed several defects in organism development, suggesting its function in zebrafish development and as a sponge, by recruiting RBPs from the endogenous USE RNA. Accordingly with these results,

we also observed a decrease in GFP expression, when microinjecting USE RNA in USE-GFP transgenic line. Overexpressing USE in human cells, we verified that the expression of 9 out of 31 USE-containing genes shared with humans was decreased, including genes that are associated with serious human diseases, such cancer. Next, we asked which were the RBPs that were binding in the USE. Through UV-cross linking, we found 2 RBPs (HuR and hnRNP C) that bind to the USE, known for their roles in mRNA stability in zebrafish and human cells. We ended up this chapter by uncovering the clinical relevance by identifying a SNP (rs3087967) in the USE of the POU2AF2/C11orf53 gene, associated with colorectal cancer. This SNP creates an ectopic USE consensus, leading to increased gene expression in zebrafish gut cells, demonstrated in a mosaic transgenic reporter assay. Given the tumorigenic potential of POU2AF2/C11orf53, the findings suggested that rs3087967 may contribute to colorectal cancer development, shedding light on the novel role of USEs in human disease mechanisms.

Until now, the vast majority of human genetic diseases studies predominantly focused on the coding region, neglecting the regulatory potential inherent in both the coding and non-coding parts of the genome. The gene regulation field is now getting the proper importance and we hope that this doctoral thesis can contribute for a better understanding of how mutations in CREs can be implicated in human disease. Overall, we explored the transcriptional gene regulation by identified human endocrine pancreatic enhancers and the impact of disease-associated SNPs within enhancers in T2D (Figure 1). In addition, we also explored the post-transcriptional gene regulation by identifying, for the first time, a USE in zebrafish, and we demonstrated its role in gene regulation and in human disease (Figure 1). The findings here show that the regulatory regions of the genome cannot be disregarded and that the regulatory code represent a new world of hidden messages to discover!

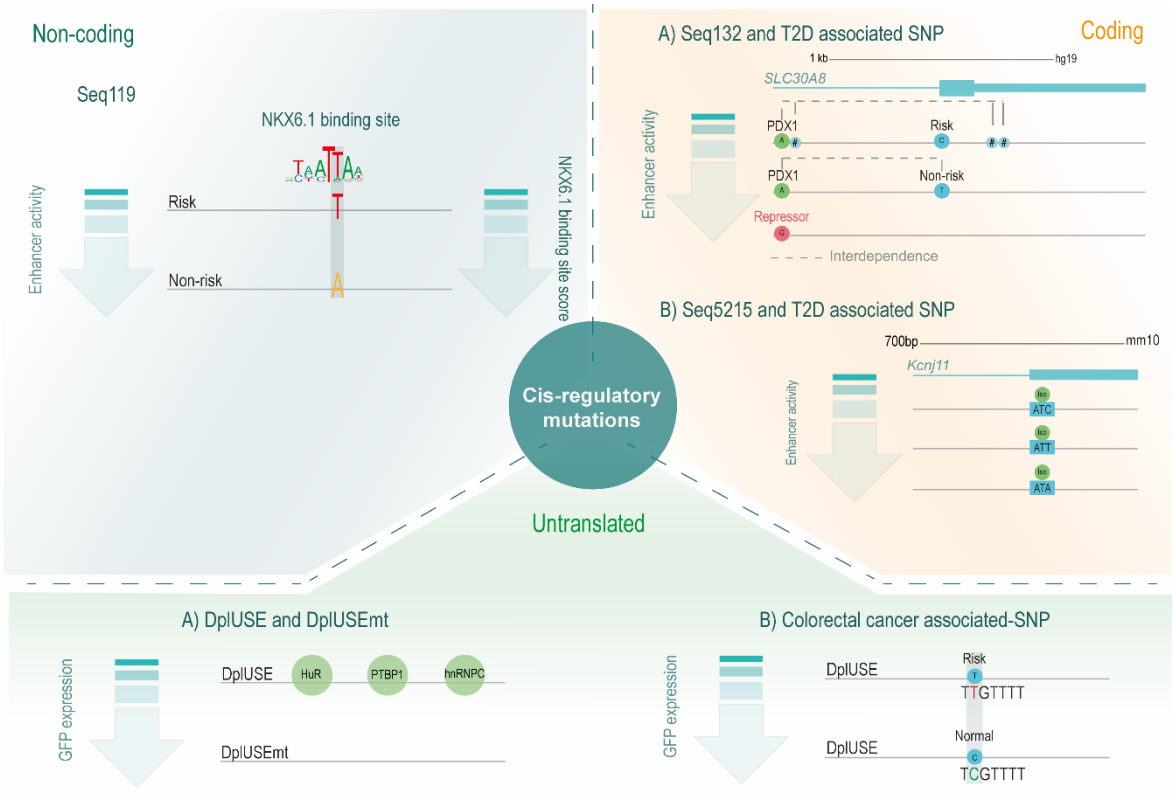


Figure 1 – Three types of possible cis-regulatory mutations in human diseases: Non-coding (above, left) – the seq119 showed a modulation in enhancer activity, when the T2D-associated SNP was present, due to a predicted higher affinity of NKX6.1 binding site.; Coding (above, right)- A) The seq132 enhancer activity is modulated through the presence of the T2D-associated SNP, common SNPs and a single nucleotide variation in a predicted PDX1 binding site; and untranslated (below) – a short motif in the 3´UTR is able to modulate the gene expression when disrupted (left) or with the presence of a colorectal cancer associated SNP (right) .

1. References

1. Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics* **31**, 67–76 (2015).

2. Epstein, D. J. Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic* **8**, 310–316 (2009).

3. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics* **46**, 136–143 (2014).

4. Ali, O. Genetics of type 2 diabetes. *World J Diabetes* **4**, 114–123 (2013).

5. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. *Lancet* **389**, 2239–2251 (2017).

6. Wu, Y. *et al.* Functional annotation of sixty-five type-2 diabetes risk SNPs and its application in risk prediction. *Scientific Reports* **7**, 1–11 (2017).

7. Mohlke, K. L. & Boehnke, M. Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum. Mol. Genet.* **24**, R85-92 (2015).

8. Sun, W. *et al.* Integrative analysis of super enhancer SNPs for type 2 diabetes. *PLoS One* **13**, e0192105 (2018).

9. Degtyareva, A. O., Antontseva, E. V. & Merkulova, T. I. Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *Int J Mol Sci* **22**, 6454 (2021).

10. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* **22**, 1059–1068 (2012).

11. Birnbaum, R. Y. *et al.* Systematic Dissection of Coding Exons at Single Nucleotide Resolution Supports an Additional Role in Cell-Specific Transcriptional Regulation. *PLoS Genet* **10**, e1004592 (2014).

12. Boutz, P. L. *et al.* A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21**, 1636–1652 (2007).

13. Sawicka, K., Bushell, M., Spriggs, K. A. & Willis, A. E. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* **36**, 641–647 (2008).

14. RNA-binding proteins in human genetic disease | Nature Reviews Genetics. https://www.nature.com/articles/s41576-020-00302-y.

15. Corbett, A. H. Post-transcriptional regulation of gene expression and human disease. *Current Opinion in Cell Biology* **52**, 96–104 (2018).