

Universidade do Minho

Escola de Ciências

João Miguel Carvalho Nogueira

A Novel Lentivirus System to Decode the Non-Coding DNA Associated with Type-2 Diabetes Risk.

Outubro de 2023

A Novel Lentivirus System to Decode the Non-Coding DNA Associated with Type-2 Diabetes Risk.

João Miguel Carvalho Nogueira

UMinho | 2023

Universidade do Minho

Escola de Ciências

João Miguel Carvalho Nogueira

A Novel Lentivirus System to Decode the Non-Coding DNA Associated with Type-2 Diabetes Risk.

Dissertação de Mestrado

Mestrado em Genética Molecular

Trabalho efetuado sob a orientação de

Doutor Fábio Júnio Veríssimo Ferreira

e de

Doutora Maria Manuela Ribeiro Costa

Outubro de 2023

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial-SemDerivações

CC BY-NC-ND

https://creativecommons.org/licenses/by/4.0/

ACKNOWLEDGEMENTS

Ao Doutor José Bessa, não apenas pelo seu apoio e inestimável orientação científica, mas também pela oportunidade única de integrar neste projeto e na sua fantástica equipa de investigação.

Ao meu orientador, Doutor Fábio Ferreira. Além de ser um exemplar investigador, é uma pessoa excecional que tornou os dias no laboratório não apenas produtivos, mas também repletos de entusiasmo e animação.

Aos restantes membros do VDR – Ana Eufrásio, Francisca Ribeiro, Francisco Cadete, Isabel Guedes, Joana Marques, Joana Teixeira, João Amorim, Jovel Jose, Marta Duque e Sofia Borges – expresso a minha gratidão pela constante boa disposição e apoio que sempre proporcionaram.

Ao grupo da Doutora Elsa Logarinho, em especial ao "apêndice" Ana Rita Araújo, pelo seu companheirismo e empatia.

Aos membros das plataformas científicas do i3S e a todos os membros de grupos vizinhos que contribuíram para a realização deste trabalho, o meu sincero agradecimento.

Por fim, não posso deixar de agradecer à minha família e amigos, em particular aos meus pais, cujo esforço e investimento ao longo destes 5 anos de percurso académico tornaram a realização deste trabalho possível.

The work described in this dissertation is part of a manuscript currently being prepared for publication in a peer-reviewed journal.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

RESUMO

Diabetes tipo 2 (T2D) é considerado uma pandemia dos tempos modernos e está associado a complicações graves, como doenças cardiovasculares e uma maior probabilidade de mortalidade prematura. A suscetibilidade genética a T2D está ligada a polimorfismos de nucleótido único (SNPs) em regiões de possíveis potenciadores, o que implica que alterações na regulação transcricional desempenham um papel no desenvolvimento da doença. Neste trabalho, o nosso objetivo é: 1) desenvolver um ensaio repórter de potenciadores adequado para testes em larga escala, 2) validar possíveis potenciadores de células beta humanas potencialmente associados ao desenvolvimento de T2D e 3) decifrar funcionalmente o código nucleotídico dos potenciadores validados.

Construímos um lentivírus compatível com a clonagem em larga escala de bibliotecas de potenciadores, contendo um promotor mínimo e EGFP como gene repórter, para avaliar a atividade dos potenciadores clonados. Utilizando um potenciador do pâncreas endócrino previamente validado, testámos vários promotores mínimos e diferentes arquiteturas do lentivírus. Clonámos este potenciador a montante do promotor mínimo, a abordagem mais simplificada para testar a atividade do potenciador, ou a jusante, uma configuração que potencialmente recapitula a interação endógena potenciador-promotor através do looping da cromatina. Isto permitiu-nos encontrar a melhor solução que gera a maior relação sinal-ruído, mantendo simultaneamente as condições endógenas relativas à funcionalidade do potenciador. Em seguida, utilizámos marcas bioquímicas associadas à atividade de potenciador e SNPs associados a T2D para definir uma lista de aproximadamente 300 sequências reguladoras candidatas (CRSs). Clonámos esta biblioteca de CRSs no lentivírus desenvolvido, infetámos células MIN6 e através de FACS-sorting, isolámos células portadoras de CRSs capazes de aumentar a expressão de GFP para análise por sequenciação de nova geração (NGS) e para estabelecer linhas repórteres contendo um único potenciador. Os resultados deste estudo são cruciais para descodificar os mecanismos subjacentes às sequências de nucleótidos dos potenciadores e o seu impacto na expressão genética e, em última análise, no desenvolvimento do T2D.

Palavras-chave: Cromatina, Diabetes tipo 2 (T2D), Lentivírus, Potenciadores

ABSTRACT

Type-2 diabetes (T2D) is considered a modern-day pandemic associated with severe complications like cardiovascular diseases and an increased likelihood of premature mortality. T2D's genetic susceptibility is tied to single nucleotide polymorphisms (SNPs) that overlap with predicted enhancers, implying that changes in transcriptional regulation play a role in the disease's development. In this work, we aimed to: 1) develop an enhancer reporter assay suitable for large-scale enhancer testing, 2) validate predicted human beta-cell enhancers potentially associated with T2D development and 3) functionally decipher the nucleotide code of the validated enhancers.

We have built a lentivirus compatible with the large-scale cloning of libraries of enhancers, containing a minimal promoter and EGFP as a reporter gene, to assess the activity of cloned enhancers. Using a previously validated endocrine pancreas enhancer, we tested several minimal promoters and different architectures of the lentivirus. We cloned this enhancer upstream of the minimal promoter, the most simplified approach to test enhancer activity, or downstream, a configuration that potentially recapitulates the endogenous enhancer-promoter interaction by chromatin looping. This allowed us to find the best solution that generates the highest signal-to-noise ratio, while simultaneously maintaining endogenous conditions regarding enhancer functionality. Next, we have used biochemical marks associated with enhancer activity and T2D-associated SNPs to define a list of approximately 300 candidate regulatory sequences (CRSs). We cloned this library of CRSs in the developed lentivirus, infected MIN6 cells, and through FACS-sorting, we isolated cells carrying CRSs capable of increasing GFP expression for next-generation sequencing (NGS) analysis and to establish single enhancer reporter lines. The findings of this study are crucial to decoding the underlying mechanisms of enhancer nucleotide sequences and their impact on gene expression and ultimately T2D development.

Keywords: Chromatin, Enhancers, Lentivirus, Type-2 Diabetes (T2D)

TABLE OF CONTENTS

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS ii

ACKNOWLEDGEMENTS..... iii

STATEMENT OF INTEGRITY..... iv

RESUMO v

ABSTRACT..... vi

TABLE OF CONTENTS vii

LIST OF ABBREVIATIONS AND ACRONYMS..... x

UNITS.....xiii

LIST OF FIGURES.....xiv

CHAPTER 1: INTRODUCTION 1

1. PANCREAS..... 2

2. PANCREATIC DISEASES 3

2.1. DIABETES MELLITUS..... 3

2.1.1. TYPE 2 DIABETES (T2D)..... 4

2.1.2. OTHER FORMS OF DIABETES AND THEIR AETIOLOGY..... 5

3. CIS-REGULATION..... 6

3.1. ENHANCERS AND SILENCERS 7

3.2. PROMOTERS..... 8

3.3. INSULATORS..... 10

4. CIS-REGULATION AND DISEASE 10

4.1. CIS-REGULATION AND T2D DEVELOPMENT 11

5. ENHANCER VALIDATIONS..... 13

5.1. PROMOTERS IN ENHNACER VALIDATIONS 16

6. MODELS TO TEST PANCREATIC ENHANCERS..... 17

CHAPTER 2: HYPOTHESIS AND MAIN OBJECTIVES 19

CHAPTER 3: MATERIALS AND METHODS 21

1. CELL CULTURE 22
2. CELL STAINING 22
3. MICROSCOPY 22
4. LIBRARY SEQUENCE DESIGN 23
5. PCR AMPLIFICATIONS 23
6. BACTERIAL TRANSFORMATIONS 24
7. ASSEMBLY OF VECTORS 24

7.1. pLS-mP-CRSEL 24

7.2. pLS-mP-C3 25

7.3. pLS-GATAmP-CRSEL/pLS-HSP70mP-CRSEL/pLS-FOSmP-CRSEL 26

7.4. pLS-GATAmP-C3/pLS-HSP70mP-C3/pLS-FOSmP-C3 27

7.5. pLS-CRSEL-mP 27

7.6. pLS-C3-mP/pLS-056-mP/pLS-510-mP 27

7.7. pLS-C3 28

7.8. pLS-C3-GATAmP 29

7.9. Cloning of enhancer library (pLS-Pool-mP/pLS-GATAmP-Pool) 29

8. LUCIFERASE ASSAYS 30
9. LENTIVIRAL PRODUCTIONS 31
10. INFECTIONS 32
11. VIRAL TITRATIONS AND MOI CALCULATIONS 32
12. FLOW CYTOMETRY 33
13. FLUORESCENT-ACTIVATED CELL SORTING (FACS) 34
14. REPORTER LINES 35
15. CRE RECOMBINASE ASSAY 35

16. GENOTYPING 35

17. NEXT GENERATION SEQUENCING 36

CHAPTER 4: RESULTS AND DISCUSSION 37

1. MORPHOLOGICAL ANALYSIS OF TWO MIN-6 CELL SUBLINES 38

2. DEVELOPMENT OF NEW LENTIVIRAL VECTORS FOR USE IN MPRAs 41

2.1. CLONING OF THE C3-9 ENHANCER DOWNSTREAM OF THE REPORTER GENE LIMITS DETECTION OF ENHANCER ACTIVITY 46

2.2. POSITION OF C3-9 ENHANCER IMPACTS EXPRESSION OF THE REPORTER GENE .. 50

2.3. VALIDATION OF THE ENHANCER DETECTING SYSTEM BY COMPARISON WITH LUCIFERASE-REPORTER ASSAY 52

3. SELECTION OF CANDIDATE REGULATORY SEQUENCES 55

4. PARALLEL REPORTER ASSAYS 58

5. VALIDATION OF ENHANCERS USING REPORTER LINES 63

CHAPTER 5: CONCLUSIONS AND FUTURE PERSPECTIVES 70

REFERENCES 74

LIST OF ABBREVIATIONS AND ACRONYMS

3C/4C/5C - Chromosome Conformation Capture techniques

3D - Three-Dimensional

APC-Cy7 - Allophycocyanin 7

ARE - Antirepressor

ATAC-seq - Assay for Transposase-Accessible Chromatin using sequencing

BB - Back Bone

ChIP-seq - Chromatin Immunoprecipitation followed by sequencing

CMV - Cytomegalovirus

CO_2 - Carbon Dioxide

CRE - Cis-Regulatory Element

CRS - Candidate Regulatory Sequence

CRSEL - Candidate Regulatory Sequence Entry Locus

CTCF - CCCTC-binding factor

DAPI - 4',6-diamidino-2-phenylindole

DNA - Deoxyribonucleic acid

dNTP - Deoxynucleotide triphosphate

DTT - Dithiothreitol

EDTA - Ethylenediaminetetraacetic acid

EGFP - Enhanced Green Fluorescent Protein

eRNA - Enhancer Ribonucleic acid

FACS - Fluorescence-Activated Cell Sorting

FAIRE-seq - Formaldehyde-Assisted Isolation of Regulatory Elements

FBS - Fetal Bovine Serum

FITC - Fluorescein isothiocyanate

FOSmP - *FOS* minimal promoter

FSC-A - Forward scatter area

FSC-H - Forward scatter height

GABPA - GA-Binding Protein Alpha chain

GATA2mP - GATA-binding factor 2 minimal promoter

GCK - Glucokinase

GDM - Gestational Diabetes Mellitus

gDNA - Genomic DNA

GFP - Green Fluorescent Protein

GsbE - Gooseberry enhancer

GsbnE - Goosberry neuro enhancer

GTF - General Transcription Factor

GWAS - Genome-Wide Association Study

H3K27ac - Acetylation of lysine 27 on histone H3 protein subunit

H3K4me1 - Monomethylation of lysine 4 on histone H3 protein subunit

HLA - Human Leukocyte Antigen

HEPES- 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HNF1A - Hepatocyte Nuclear Factor-1-Alpha

HNF4A - Hepatocyte Nuclear Factor-4-Alpha

HP - High Passage

HSP70mP - 70 kilodaltons heat shock protein minimal promoter

LB - Lysogeny Broth

LP - Low Passage

LTR - Long Terminal Repeat

MEM - Minimum Essential Media

mmCryaa - *Mus musculus* crystallin alpha a

MODY - Mature Onset Diabetes of the Young

MOI - Multiplicity Of Infection

mP - Minimal Promoter

MPRA - Massively Parallel Reporter Assay

mRNA - Messenger Ribonucleic acid

NGS - Next-Generation Sequencing

P2A - 2A self-cleaving peptide

PBS - Phosphate-Buffered Saline

PCR - Polymerase Chain Reaction

PDX1 - Pancreatic and Duodenal homeobox 1

pre-mRNA - Precursor mRNA

PROMPT - Promoter upstream transcripts

PTF1A - Pancreas-associated Transcription Factor 1A

RL - Reporter Line

RNA - Ribonucleic acid

RNAPII - RNA polymerase II

RT - Room Temperature

SNP - Single Nucleotide Polymorphism

SSC-A - Side scatter area

SSC-H - Side scatter height

STARR-seq - Self-Transcribing Active Regulatory Region sequencing

SYT8 - Synaptotagmin 8

T1D - Type 1 Diabetes

T2D - Type 2 Diabetes

TAD - Topologically Associating Domain

TF - Transcription Factor

TFBS - Transcription Factor Binding Site

Tris-EDTA - Tris Ethylenediaminetetraacetic acid

TSS - Transcription Start Site

UTR - Untranslated region

WHO - World Health Organization

WPRE - Woodchuck hepatitis virus posttranscriptional regulatory element

YY1 - Yin Yang 1

ZED - Zebrafish enhancer detection

UNITS

a.u. - Arbitrary unit

bp – Base pair

°C - Degrees Celsius

fmols - Fentomoles

h - Hour

Kb - Kilobase

µg - Microgram

µL - Microliter

µM - Micromolar

µm - Micrometer

mL - Mililiter

mM - Milimolar

Min - Minute

Pmols - Picomoles

Rpm - Rotations Per Minute

s - Second

LIST OF FIGURES

Figure 1. Morphology of LP and HP MIN-6 cells. 40

Figure 2. Schematic representation of lentiMPRA vectors used in this work. 43

Figure 3. Assessment of the enhancer detection ability of the constructed vectors. 50

Figure 4. Comparison of episomal and lentiviral enhancer reporter systems, using enhancer C3-9 and CRSs for the PDX1 gene. 55

Figure 5. Assembly of CRSs lentiviral libraries. 58

Figure 6. Flow cytometry analysis of parallel reporter assays. 61

Figure 7. Preliminary analysis of NGS results from parallel reporter assays. 63

Figure 8. Schematics of the mCherry-Cre and Cre-Reporter vectors. 65

Figure 9. Cre-mediated enhancer validations. 69

CHAPTER 1:
INTRODUCTION

1. PANCREAS

The human pancreas is a complex organ with diverse cell lineages, playing crucial roles in both the digestive and endocrine systems (Jennings et al., 2015; Matsuda, 2018). These two roles sustain the usual division of this organ into two distinct functional tissues, the exocrine and the endocrine pancreas.

The exocrine pancreas, associated with the digestive system, is constituted mainly by epithelial acinar cells that produce many enzymes that are able to digest proteins, fats, carbohydrates, and nucleic acids like trypsin, lipase, lactase, and ribonuclease as well as ductal cells that form the pancreatic ducts that transport these and many other enzymes to the duodenum (Habener et al., 2005; Karpińska & Czauderna, 2022). Acinar cell dysfunction or pancreatic duct obstruction can impair enzyme secretion and cause digestive problems, potentially leading to inflammation and tumors such as adenocarcinoma (Andersen et al., 2017).

The endocrine pancreas, a crucial part of the complex pancreatic anatomy, plays a critical function in glucose homeostasis (Bakhti et al., 2019). This dynamic process involves the coordinated production and release of several vital hormones into the bloodstream. These endocrine cells gather in the pancreas in specialized enclaves known as the islets of Langerhans. These islets contain a variety of different cell types, including, α, β, δ, ε, and γ-cells (Bastidas-Ponce et al., 2017).

These islet cells conduct a highly tuned symphony of hormones that intricately control glucose homeostasis, a critical component of metabolic regulation. β-cells take the front stage among these cell types by releasing insulin, a hormone essential for glucose uptake and utilization within cells. Contrastingly, α-cells secrete glucagon, which increases the release of stored glucose into the bloodstream when levels drop, assisting in the maintenance of a stable source of energy. Complementing this interaction, δ -cells emit somatostatin, which acts as a regulator to modulate the release of both insulin and glucagon. Lastly, ε - and γ -cells generate ghrelin and pancreatic polypeptide, which aid in appetite management and digestion (Bastidas-Ponce et al., 2017; Mastracci & Sussel, 2012).

The harmonic cooperation of these hormones is meticulously regulated through intricate feedback loops to sustain equilibrium. Variations from this balance caused by endocrine cell failure result in abnormal amounts of these pancreatic hormones, culminating in the breakdown of glucose homeostasis, a hallmark of diabetes mellitus (Bich et al., 2020; Wei et al., 2020).

2. PANCREATIC DISEASES

A variety of risk factors, including hereditary and environmental factors, are associated with many human pancreatic diseases. This is especially true for chronic pancreatitis, pancreatic cancer, and the main focus of this work, diabetes, in which a complex interplay of those variables contributes to the development of multiple types of the disease. These disease types, which are distinguished by their etiology and complexity, frequently occur as a result of a convergence of mutations across several genetic loci, resulting in a genetically heterogeneous background of vulnerability (Andersen et al., 2017; Mohlke & Boehnke, 2015).

2.1. DIABETES MELLITUS

Diabetes is a multifaceted disease whose underlying issue is the disruption of glucose homeostasis, which in turn develops as a result of complex interactions between multiple factors. These factors include genetic susceptibility, environmental factors, the natural aging process, and lifestyle. This combination results in diabetes mellitus, a group of metabolic disorders characterized by high blood glucose levels, commonly termed hyperglycemia (Dziewulska et al., 2018; Kleinberger & Pollin, 2015).

This group of diseases has a huge global impact, with alarming statistics illustrating its prevalence. The World Health Organization's (WHO) Global Report on Diabetes in 2016 reported that diabetes affects over 422 million adults, highlighting its significance as a global health concern (Standl et al., 2019). Diabetes, if undetected or mismanaged, can lead to a host of vascular problems, most notably cardiovascular disease (Kleinberger & Pollin, 2015).

Diabetes mellitus is often categorized into four main forms: maturity-onset diabetes of the young (MODY), gestational diabetes mellitus (GDM), type 1 diabetes (T1D), and type 2 diabetes (T2D), the central topic of this project, each characterized by distinct risk factors, age of onset, and underlying causes.

2.1.1. TYPE 2 DIABETES (T2D)

Type 2 diabetes (T2D), formerly known as non-insulin dependent diabetes is a multifaceted disorder characterized by insulin resistance in specific tissues like the muscle, liver, and brain and a deficit in insulin production due to pancreatic beta-cell dysfunction and overall dysregulation of the endocrine pancreas (Chatterjee et al., 2017). This complex imbalance disrupts glucose homeostasis, resulting in clinical symptoms such as hyperglycemia. Furthermore, T2D is frequently associated with obesity and cardiovascular issues, aggravating its influence on general health (Saxena et al., 2007).

Despite significant progress, the complete interplay of molecular pathways underlying T2D remains unknown and adds to the disease's complexity.

Many genes have been associated with T2D, but single genetic alterations do not seem to be the primary triggers for this disease. Environmental factors such as diet and the natural process of aging, which interact with hereditary features, add layers to the disease's complexity (Chatterjee et al., 2017). Moreover, T2D might be shaped by the cumulative effects of numerous genetic elements. Interestingly, a significant proportion of T2D-associated single-nucleotide polymorphisms (SNPs) are found in non-coding regions of the genome, notably overlapping with predicted regulatory elements in pancreatic islets (Flannick & Florez, 2016; Pasquali et al., 2014) and with impact in glucose secretion (Miguel-Escalada et al., 2019). These findings reinforce the complex genetic landscape of glucose homeostasis and suggest impaired cis-regulatory mechanisms might contribute to T2D development.

2.1.2. OTHER FORMS OF DIABETES AND THEIR AETIOLOGY

Other forms of diabetes include gestational diabetes mellitus (GDM) which is a specific type of diabetes that develops during pregnancy, characterized by disturbed glucose tolerance due to hormonal changes and increased insulin demand. The increased insulin resistance experienced during pregnancy places both mother and child at elevated susceptibility to type 2 diabetes (Sweeting et al., 2022). Recent studies have also indicated that genetic and epigenetic factors may play a role in the development of GDM and its associated maternal and fetal complications (Dalfrà et al., 2020; Elliott et al., 2019).

Type 1 diabetes (T1D) is a chronic autoimmune disorder characterized by the breakdown of pancreatic beta cells, resulting in insulin deficiency. (Katsarou et al., 2017). The initial trigger for the formation of the autoantibodies targeting β-cells is still unclear but their presence is associated with the destruction of β-cells by autoantigen-specific CD4+ and CD8+ -(Battaglia, 2014; Roep & Peakman, 2012). T1D's genetic basis has been extensively studied for over five decades, revealing over 40 genetic loci linked to the disease, with particular emphasis on the HLA genes (Noble & Erlich, 2012).

Lastly, maturity-onset diabetes of the young (MODY) is a monogenic and autosomal form of diabetes and it is currently characterized by onset before the age of 25 in one family member, diabetes in two generations in a row, no β-cell autoantibodies, and persistent but often reduced endogenous insulin secretion (Ellard et al., 2008; Urakami, 2019). There are currently 14 genes whose mutations have been associated with MODY. They include hepatocyte nuclear factor-1-alpha gene (HNF1A), hepatocyte nuclear factor-4-alpha gene (HNF4A), glucokinase (GCK), which are the most prevalent, and PDX1. Broadly, these genes are required for the normal development of the endocrine pancreas and/or the production of endocrine enzymes. Thus, mutations in these genes are mainly associated with β-cell dysfunction, due to a lack of insulin production and/or secretion, which compromises glucose homeostasis (Urakami, 2019).

While MODY is frequently linked to a single gene mutation, its susceptibility, and emergence seem to be influenced by a variety of environmental and genetic factors (Antosik & Borowiec,

2016). Additionally, understanding how different genetic variations contribute to the development of MODY can be useful for understanding different multifactorial diabetes types (Horikawa, 2018).

3. CIS-REGULATION

Cell differentiation and tissue formation rely heavily on the establishment of precise transcriptional profiles. These profiles dictate which genes are expressed, the extent of their expression, and, consequently, the specialized functions of each cell (Roy & Kundu, 2014). While RNA Pol II binding to a gene's promoter is required for transcription to begin, the regulatory processes at the transcriptional level are considerably more complex and dependent on a variety of factors (Hager et al., 2009; Hahn, 2004). The patterns of gene expression that occur in specific locations and timeframes are dependent on the combined influence of various regulatory components that form the regulatory landscape of a gene (Bessa et al., 2014).

Non-coding DNA, which makes up the majority of the genome (Piovesan et al., 2019), plays an important function in gene regulation (Barrett et al., 2012). Notably, non-coding DNA is complex; in some cases, it can be transcribed into non-coding RNAs while staying silent in others (Maston et al., 2006; Park et al., 2022). Furthermore, both coding and non-coding DNA sequences can have regulatory functions by facilitating the binding of regulatory proteins such as transcription factors (TFs). These interactions between DNA and regulatory proteins are critical for gene expression regulation (Lambert et al., 2018; Latchman, 1997). While gene regulation can occur "in trans" by the binding of regulatory proteins to DNA, it is also clear that gene regulation also happens "in cis”. DNA, whether coding or non-coding, can influence the transcriptional activity of neighboring genes in the "in cis" mode by interacting with other sequences within the genome (Goncalves et al., 2012; Wittkopp & Kalay, 2012). This cis-regulatory mechanism adds a higher degree of complexity to gene expression regulation, ensuring that each cell's transcriptional profile is highly adjusted and tailored to its specific purpose within the organism (Barrett et al., 2012). This cis-regulation is ensured by cis-regulatory elements (CREs) which are DNA sequences, mainly non-coding, that can activate, sustain, repress, or prevent transcription (Wittkopp & Kalay, 2012). CREs are typically divided

into 4 classes: promoters, insulators, enhancers, and silencers (Noonan & McCallion, 2010), and, when active, CREs typically exist as accessible chromatin regions, devoid of nucleosomes (Klemm et al., 2019). A human gene's regulatory landscape often includes numerous cis-regulatory elements spread over large genomic distances (Hager et al., 2009; Ragvin et al., 2010).

3.1. ENHANCERS AND SILENCERS

Transcriptional enhancers are vital cis-regulatory elements that dictate when, where and how much genes are transcribed (Pennacchio et al., 2013). These non-coding regions increase gene transcription for specific genes or gene sets and they can also regulate how the transcript (pre-mRNA) is processed (Kwon et al., 2022). Enhancers can be situated upstream, downstream, or within target genes' introns and exons. Their influence is independent of their distance or orientation relative to the target gene, making it difficult to anticipate enhancer-gene control purely based on sequence structure analysis (Maston et al., 2006; Pennacchio et al., 2013). Enhancers have transcription factor binding sites (TFBSs) that interact cooperatively, recruiting co-activators and co-repressors. This interaction activates the promoter of the target gene (Mora et al., 2016).

The function of an enhancer is dependent on the binding of specific TFs capable of attracting additional factors and chromatin remodelers, which assist in the physical interaction between enhancer and promoter. Enhancer specificity is determined by unique TF combinations, and multiple tissue-specific enhancers can interact with the same gene promoter, contributing to the gene's expression pattern (Furlong & Levine, 2018.; Reményi et al., 2004).

The mechanisms underlying enhancers' long-range activity are unclear. However, the dominant theory indicates that enhancers may be located near their target gene promoters, mainly via chromatin loops (Andersson & Sandelin, 2020; Pennacchio et al., 2013). This theory is primarily supported by evidence from techniques based on chromosome conformation like 3C, 4C, and 5C where loci placed together in the 3D space are detected via PCR-based techniques or next-generation sequencing (Dekker et al., 2001; Dostie et al., 2006; Kadauke & Blobel, 2009; Simonis et al., 2006). These techniques have also revealed that the genome is

organized into topological associating domains (TADs) (Dekker & Heard, 2015). These TADs are specific genomic regions where promoters and distant enhancers preferentially interact while remaining insulated from loci in neighboring TADs (Dixon et al., 2012). The boundaries of TADs are marked by the presence of clusters of binding sites for proteins like CTCFs and cohesins, which are involved in shaping chromatin into loop structures (Hansen et al., 2017, 2018). TADs and chromatin loops are dynamic structures, and the establishment of TADs, driven by CTCF and cohesin, plays a critical role in facilitating specific interactions between enhancers and promoters (Hansen et al., 2018; Hnisz et al., 2016).

In contrast, silencers are DNA sequences where trans-acting proteins, often referred to as repressors bind. These repressors then lead to transcription repression (Maston et al., 2006). Remarkably, there are dual-function regulatory elements in which the same DNA sequence can act as an enhancer or a silencer, contingent upon the cell's context or the genes it targets (Huang & Ovcharenko, 2022; Vaz et al., 2021). This underscores the notion that the activity of these cis-regulatory sequences is fundamentally dependent on the interaction and cooperation among all the TFs bound to these sequences and the promoter (Spitz & Furlong, 2012).

3.2. PROMOTERS

Gene promoters are typically defined by the location of transcription start sites (TSSs), defined as a transcript's first transcribed genomic nucleotide (Andersson & Sandelin, 2020). The process of TSS selection is influenced by the presence of general transcription factors (GTFs) which together with RNAPII form the RNAPII pre-initiation complex (Smale & Kadonaga, 2003). They are also canonically defined by a conserved sequence that composes the TATA box, recognized by the TATA-binding protein which is another essential protein of the pre-initiation complex (Breathnach & Chambon, 1981).

Although promoters provide enough information for the transcriptional machinery to select TSSs, the rate of RNAPII initiation can be controlled by integrating additional signals. TFs can either directly or indirectly influence promoter RNAPII recruitment, initiation, or elongation.

Their input may come from either local (DNA regions close to the TSS) or distal elements like enhancers (Andersson & Sandelin, 2020; Lambert et al., 2018).

Recently it's been proposed that enhancer and promoter activity aren't mutually exclusive but rather a spectrum where a cis-regulatory sequence can have different degrees of these capabilities (Andersson & Sandelin, 2020; Henriques et al., 2018; Mikhaylichenko et al., 2018). This is explained due to the similar binding of TFs, co-factors, and even RNAPII to enhancers and promoters, and the dual role of those regions, depending on the cell type and environmental cues. Moreover, while most promoters direct transcription in an orientation-dependent manner, generating mRNAs in the sense strand and PROMPTs in the other antisense strand, enhancers can be transcribed into enhancer RNAs (eRNAs) bidirectionally, leading some authors to suggest these elements share a similar basic promoter architecture (Andersson & Sandelin, 2020; Ding et al., 2018; Scruggs et al., 2015). Regulatory elements with enhancer activity often exhibit some promoter activity, mainly in vitro but also in vivo, initiating local transcription (Van Arensbergen et al., 2017). However, the resulting eRNAs are transcribed at lower levels and are more prone to degradation within the nucleus, indicating that overall distal enhancers generate less promoter activity than gene promoters (Andersson & Sandelin, 2020). In a particular context, it was demonstrated that robust promoter activity can stem from sequences devoid of core promoter elements but with specific transcription factor binding motifs. This suggests that to understand promoter activity we need to take into consideration the specific transcriptional activators bound to these sequences which in turn could be used to distinguish promoters from distal enhancers (Nguyen et al., 2016).

It was also demonstrated that many promoters can exhibit enhancer activity (Dao et al., 2017; Engreitz et al., 2016). Chromatin conformation data confirms this duality, revealing interactions between distal gene promoters with other gene promoters, suggesting enhancer-like functions (Mifsud et al., 2015). For instance, the insulin gene promoter interacts with the *SYT8* gene promoter, potentially regulating its expression in human islets (Xu et al., 2011).

3.3. INSULATORS

Lastly, insulators function as boundary elements that hinder the activity of nearby regulatory elements, ensuring accurate gene activation. Insulators primarily restrict regulatory interactions by preventing the formation of chromatin loops (Ali et al., 2016; Maston et al., 2006). Insulators usually harbor binding sites for insulator proteins like the CTCF protein and therefore usually are located in the boundaries of TADs (Furlong & Levine, 2018).

Insulator elements fall into two categories: enhancer-blocking, which hinders communication between a promoter and enhancer when positioned in between, and barrier-insulators, which shield genes from silencing by preventing nearby heterochromatin from spreading to that gene (Gaszner & Felsenfeld, 2006). Interestingly, some authors suggest that distinct cell types have unique insulator-mediated nuclear arrangements and cell identity is shaped by gene expression patterns influenced by this organization (Dorman et al., 2007).

4. CIS-REGULATION AND DISEASE

Complex diseases like cancer and diabetes arise due to a combination of environmental and genetic influences. In fact, frequently genetic traits from various loci contribute to susceptibility to these diseases (Frank, 2004; Miguel-Escalada et al., 2019). Human complex diseases often involve disrupted gene transcription regulation, which together with some examples of the absence of coding sequence mutations in disease-associated genes hints at the significance of genetic variants in non-coding sequences in these diseases' development (Albert & Kruglyak, 2015).

Mutations in enhancer-associated TFBSs, for example, can disrupt the regulation of target genes, potentially leading to the loss of specific cell types or tissues during development (Lee & Young, 2013). Recent advancements in understanding human genetic diversity and cis-regulation of transcription have improved our comprehension of gene expression dysregulation in various human diseases. An illustration of the natural genetic variation is found in single nucleotide polymorphisms, which have been identified through SNP arrays, whole-genome sequencing projects, and computational analysis (Altshuler et al., 2012;

Winchester et al., 2009). These variations are mainly situated in non-coding regions and some have been directly linked to human traits and complex diseases (Fadason et al., 2022).

Genome-wide association studies (GWAS) enable the connection between SNPs and diseases or traits. These studies involve evaluating thousands of SNPs in a sizable group of individuals to establish reliable associations between common genetic variants and specific biological conditions or traits (Hardy & Singleton, 2009; Schaid et al., 2018). It's worth noting that while the idea of assigning functions to SNPs to decipher the connections between genetic variations, phenotypes, and diseases seems straightforward, it's exceedingly intricate. This complexity arises from the vast number of SNPs associated with a single phenotype and the contextual dependencies such as development, environment, and the combined effects they exert — factors often not easily decoded in bioinformatic analyses (Abdellaoui et al., 2023).

Nevertheless, there's a well-established connection between disease-associated non-coding polymorphisms, identified through GWAS analysis, and recognized regulatory elements like enhancers. This connection suggests an influence of non-coding sequences on the disease through the modulation of transcriptional regulation (Albert & Kruglyak, 2015; Maurano et al., 2012). In particular, many disease-associated SNPs were found to not be in linkage disequilibrium with coding genes, suggesting that polymorphisms in non-coding DNA can impact physiology (Pallejà et al., 2012). Despite the acknowledged importance of cis-regulation in human disease development, we still lack a complete understanding of the mechanisms that disrupt normal cis-regulatory networks. (M. J. Li et al., 2014).

4.1. CIS-REGULATION AND T2D DEVELOPMENT

Cis-regulatory sequences and epigenetic factors play crucial roles in shaping pancreatic cell lineage, preserving cellular differentiation, and maintaining the proper function of adult pancreatic endocrine cells (Sandovici et al., 2013). Moreover, as mentioned before, GWAS have revealed numerous SNPs connected to hereditary diseases like T2D. Around 90% of T2D-associated GWAS SNPs are non-coding and many coincide with predicted enhancers, some of which show activity in pancreatic islets (Khetan et al., 2021; Miguel-Escalada et al., 2019). This

suggests that disrupted transcriptional regulation is a prominent molecular mechanism underlying disease risk in these regions.

Pasquali and colleagues (Pasquali et al., 2014) identified active enhancers and binding sites for human islet-specific transcription factors and also observed an enrichment of T2D-associated SNPs in these regulatory regions. These findings support the idea that these SNPs might interfere with crucial transcription factor interactions, possibly impacting islet function. This underscores the significance of epigenetic mechanisms in orchestrating gene transcription during the complex progression of T2D. The contribution of cis-regulatory mutations to pancreatic diseases was also supported in another study where the authors demonstrated that recessive mutations within an enhancer downstream of the *PTF1A* gene, a pancreas-specific transcription factor, entirely nullify the activity of this CRE and result in pancreatic agenesis (Weedon et al., 2014).

A more recent study from Miguel Escalada and colleagues (Miguel-Escalada et al., 2019) paved the path to understanding human islet 3D chromatin architecture. In this work, the authors provided insights into the glucose-dependent activity of 3D hubs of promoters and enhancers and how genetic variations in these hubs impact the heritability of insulin secretion. Importantly, the study facilitated the connection between diabetes-associated enhancers and their target genes, even when these genes were located hundreds of kilobases apart. Ramos-Rodriguez and colleagues (Ramos-Rodríguez et al., 2019) further enlarged the collection of human islet regulatory elements. They accomplished this by mapping enhancers responsive to stimuli, which are associated with modifications in the β-cell transcriptome, proteome, and 3D chromatin structure.

Recent publications by our group also demonstrated *in vivo* how important enhancers are for proper pancreas function and how single and combinatorial nucleotide alterations in enhancer sequences can have a significant impact on enhancer activity. These data also demonstrated the functional conservation of non-coding sequences across species despite different nucleotide sequences. (Bordeira-Carriço et al., 2022; Eufrásio et al., 2020).

Thus, studying the non-coding DNA mutations' involvement in T2D susceptibility constitutes a new research opportunity. Particularly, the specific nucleotides essential for enhancer function and how their disruption leads to target gene misregulation and T2D susceptibility

remain unknown (Mularoni et al., 2017), so decoding these nucleotide sequences is a vital step in comprehending the genetic basis of T2D vulnerability. It is worth noting that although SNPs can affect the function of a regulatory sequence by disrupting a binding site in this DNA sequence, linking functions of SNPs to polygenic diseases like T2D is remarkably challenging due to their potential small or combinatorial effects. Despite this, SNP associations with diseases mark significant genomic regions relevant to these conditions, even if they aren't directly causal (Fadason et al., 2022).

Accordingly, validating predicted enhancers containing or nearby T2D-associated SNPs holds the potential to expand our understanding of the disease, as their functionality could be tied to crucial biological mechanisms in its development.

5. ENHANCER VALIDATIONS

The first step in studying enhancers is to identify potential enhancer sequences. Taking into consideration that transcription factors bind to these regulatory sequences, enhancers must be accessible, often residing within nucleosome-free regions (Klemm et al., 2019). Because of this, many techniques developed to identify accessible or open regions of chromatin are crucial for the identification of possible enhancers. These techniques include i) DNase-seq, which exploits the tendency of these regions to be uniquely sensitive to cleavage by DNase I, allowing for the digestion of these regions followed by sequencing (Sullivan et al., 2015), ii) Formaldehyde-Assisted Isolation of Regulatory Elements coupled to high-throughput sequencing (FAIRE-Seq) which allows the phase separation of nucleosome-free DNA from crosslinked, chromatin-bound DNA through sample crosslinking followed by chromatin shearing with sonication (Giresi & Lieb, 2009) and iii) Assay for Transposase Accessible Chromatin followed by high-throughput sequencing (ATAC-Seq) which requires a much lower amount of chromatin when compared with the other two techniques. ATAC-Seq employs a highly active transposase (Tn5) to specifically assess the accessibility of all genomic regions. The transposase binds to open sites, fragments DNA, and introduces sequencing adapters. Analysis of the sequencing outcome generates open chromatin maps that indicate nucleosome positioning and DNA accessibility (Buenrostro et al., 2015).

These techniques can't be applied solely to identify enhancers, as other cis-regulatory elements also exhibit accessibility. Transcriptional regulatory sequences' function is also characterized by interactions between DNA and DNA-binding proteins and post-translation modifications in histones (Wittkopp & Kalay, 2012). Several studies have associated distinct transcription factor bindings and, significantly, histone modifications with various cis-regulatory classes. These findings define unique chromatin signatures for each cis-regulatory class, permitting the recognition of potential enhancers through epigenetic marks (Kang et al., 2020). The marks associated with active enhancers include histone H3 acetylated at lysine 27 (H3K27ac) (Creyghton et al., 2010) and histone H3 monomethylated at lysine 4 (H3K4me1) (Heintzman et al., 2007).

Chromatin Immunoprecipitation coupled with high-throughput sequencing (ChIP-Seq) is a primary method for pinpointing chromatin regions marked by histone modifications or transcription factor binding in a cell. This method involves fragmenting chromatin, immunoprecipitating it using specific antibodies, and then sequencing. To employ this technique, prior understanding of potential transcription factors binding to the tested regulatory sequences is necessary for selecting the appropriate antibodies (Furey, 2012; Mundade et al., 2014).

Some enhancers can also be identified through comparative genomics using a technique known as phylogenetic footprinting. This method takes into consideration that certain non-coding enhancers are highly conserved across species. The rationale behind this strategy is that sequence conservation suggests DNA function, making conserved non-coding sequences attractive candidates for functional enhancers (Z. Zhang & Gerstein, 2003)

Thus, biochemical approaches like as ChIP-seq, which identifies DNA-bound proteins, ATAC-seq, which detects accessible chromatin regions, and others can be used to identify enhancers across the genome. These techniques, however, are only helpful in annotating candidate regulatory sequences (CRSs), and further experimental assays are necessary to confirm their expected activity (Skene et al., 2018). This is especially true when one takes into consideration that a significant portion of CRSs derived from ChIP-seq for H3K27ac are false positives in *in vivo* validations (Gorkin et al., 2020). It's worth noting that recent studies also suggest the existence of tens of thousands of enhancers dispersed throughout the genome that are

undetected by current chromatin profiling methods but demonstrate *in vivo* activity (Mannion et al., 2022).

Enhancer assays often consist of the use of vectors carrying a minimal promoter regulating the expression of a reporter gene or a barcode sequence. Enhancers potentiate the basal activity of the minimal promoter, *in vitro* or *in vivo*, as shown before (Bordeira-Carriço et al., 2022; Eufrásio et al., 2020). An example of this approach is the luciferase reporter assay where the candidate enhancer is cloned in an episomal vector upstream of a minimal promoter driving the expression of luciferase in this particular technique (Pasquali et al., 2014).

This standard episomal reporter assay has been the primary method for functionally confirming enhancers for decades. More recently, the development of massively parallel reporter assays (MPRAs) allowed for the quantification of the function of thousands of candidate enhancers at the same time. In this case, a library of CRSs is cloned in reporter vectors and introduced into cells. If a given CRS is a functional regulatory element, it impacts the expression of the reporter gene or barcode, which can be detected in the cell population via cytometry and DNA or RNA sequencing (Gordon et al., 2020; Inoue et al., 2017). In conventional MPRAs, distinct barcodes are assigned to individual enhancer sequences, which are then included in each reporter gene transcript. The activity of each candidate enhancer is quantified by comparing the relative abundance of its RNA barcode to its DNA-based representation (Gasperini et al., 2016).

A significant advantage of MPRAs lies in their capacity to efficiently assess numerous sequences for regulatory activity using standard methods in molecular biology, cell culture, and sequencing (Gasperini et al., 2016). CRSs can be synthesized DNA sequences chosen based on biochemical enhancer-associated marks, as previously discussed, or obtained through genome-wide 'shotgun MPRAs' that don't rely on predetermined assumptions (Arnold et al., 2013). Both are associated with limitations of MPRAs, like the relatively high cost of DNA synthesis and sequence size limitations or the high complexity of shotgun libraries, respectively (Gasperini et al., 2020).

One instance of an MPRA is STARR-seq (self-transcribing active regulatory region sequencing). Here, DNA fragments are inserted downstream of a core promoter and in the 3’UTR of a reporter gene. Active enhancer sequences promote their own transcription, being included in

the resulting reporter transcript, and therefore can be detected through RNA sequencing (Muerdter et al., 2015). Thus, in STARR-seq, the barcode for identifying a specific CRS is the CRS itself, transcribed with the reporter gene. However, this setup means certain sequences can influence mRNA abundance and stability, potentially leading to false negatives or false positives (D. Lee et al., 2020). MPRAs using sequence-specific barcodes or less biased techniques like microscopy or flow cytometry to quantify reporter gene expression driven by a CRS are less susceptible to such artifacts.

Another limitation of assays like STARR-seq and MPRAs, in general, is that they often involve transient transfection with non-integrated episomes. These lack the physiological chromatin context, testing cis-regulatory sequences outside a biologically relevant setting where they interact with chromatin structure. Lentiviral vectors have been developed to integrate MPRA vectors into the genome of cells addressing this concern. On top of that, it's been shown that the activity of sequences tested on episomes doesn't fully correlate with the same sequences tested using integrative lentivirus (Inoue et al., 2017).

Additionally, MPRAs fail to capture how CRSs affect and are affected by other cis-regulatory elements nearby, and the majority of MPRAs only test for enhancer activity using a single promoter, both factors could lead to a high false negative rate (Gasperini et al., 2020; Zabidi et al., 2015)

5.1. PROMOTERS IN ENHNACER VALIDATIONS

MPRAs and enhancer-testing systems often assume that enhancers act in a generic manner on promoters, regardless of promoter sequence. However, it's been suggested that promoters have specific preferences for certain enhancers due to interacting sets of TFs (van Arensbergen et al., 2014).

Like enhancers, promoters are also sometimes truncated in enhancer assays, losing genomic content and possibly compatibility with certain enhancers. Enhancer assays typically use minimal or basal promoters, the smallest functional part of the promoter retaining activity, usually containing the core promoter. Although enhancers are able to regulate the activity of multiple promoters, either simultaneous in enhancer hubs (Ferreira et al., 2023) or in response

to different stimuli (Ramos-Rodríguez et al., 2019), studies in Drosophila embryos have shown that two specific Drosophila enhancers preferentially activated promoters with the TATA-box compared to those without this core promoter element (Ohtsuki et al., 1998). In an earlier study, Li and Noll highlighted one of the first cases of specific compatibility between promoters and enhancers. They demonstrated that two enhancers (GsbE and GsbnE) of neighboring genes (gooseberry and gooseberry neuro) could only activate their respective promoter (X. Li & Noll, 1994). A more recent study looked at the compatibilities of 1000 enhancers and 1000 promoters. It was demonstrated that housekeeping gene promoters had activating motifs for GABPA and YY1. On the one hand, this limits their responsiveness to distal enhancers. On the other hand, promoters lacking these motifs were more responsive to enhancers (Bergman et al., 2022). These findings emphasize that a CRS potential enhancer activity depends on the specific promoter it can act on, hence, promoters play a critical role in enhancer-reporter assays, a facet sometimes overlooked.

On top of that, enhancer-reporter assays usually test enhancer activity by placing candidate enhancer sequences upstream of a promoter or, in the case of STARR-seq, in the 3'UTR of the reporter gene (D. Lee et al., 2020; Warren et al., 2022). However, this setup doesn't fully mirror how enhancers can interact with promoters separated by large genomic distances through chromatin loops (Andersson & Sandelin, 2020). Considering these factors, it's reasonable to interpret results from enhancer assays with some uncertainty, acknowledging that further exploration is required.

6. MODELS TO TEST PANCREATIC ENHANCERS

Enhancer reporter assays can be performed in vitro using cell lines, which are animal-free, easily manipulatable systems in which reporter vectors can be introduced through transfection or infection (Heintzman et al., 2007; Skelin, 2010). Testing enhancers in cell lines offers the benefits of consistent and reproducible results. However, the cell culture of differentiated cells lacks the developmental and physiological contexts (Gasperini et al., 2020). Moreover, cell lines' virtual unlimited cell divisions are a result of their tumorigenic origin or transformation, and therefore cell lines can be associated with abnormal chromosomal content and other genetic mutations (Skelin, 2010).

MIN-6 is one of the few cell lines with pancreatic beta-cell characteristics, like insulin secretion in response to glucose and other secretagogues. The cell line is derived from insulinomas from transgenic C57BL/6 mice that expressed an insulin/T-antigen construct (Ishihara et al., 1993). The resulting cells are islet-like making the model a suitable choice to test pancreatic enhancers. Regarding testing of enhancers linked to T2D development, it's important to note that these cells only provide a partial representation, because this disease also involves insulin resistance in non-pancreatic cells (Chatterjee et al., 2017). Hence, examining enhancers in other cell types can be beneficial and complementary. Other frequently used beta-cell lines in diabetes and pancreatic research include INS-1 (Asfari et al., 1992), βTC-6 (Poitout et al., 1995), and the human beta-cell line EndoC-βH1 (Ravassard et al., 2011).

In vivo, transgenic reporter assays remain a powerful tool to showcase tissue-specific enhancer activity and access CRS activity in broader biological contexts compared to in vitro cell lines (Visel et al., 2007). However, scaling up these assays is challenging, limiting the number of sequences tested in a single assay (Gasperini et al., 2016).

CHAPTER 2:
HYPOTHESIS AND MAIN OBJECTIVES

In recent years, the field of genomics has witnessed an exponential increase in the availability of diverse datasets predicting potentially significant enhancer sequences within the human genome based on histone modifications, chromatin accessibility, and TF binding. Particularly, in the context of complex diseases like T2D, these predicted enhancers can shed light on the molecular intricacies underlying the disease. While such datasets offer valuable insights, experimental validation is essential to confirm the functionality and relevance of these predicted enhancer sequences. MPRAs and similar approaches allowed us to test significant amounts of candidate sequences. However, we also wish to explore current problems and limitations of these types of assays to overcome the risk of oversimplification and asserting bias, which may produce misleading results.

On top of that, understanding how T2D-associated SNPs affect gene expression is critical not only to better predict susceptibility to the disease but also to prevent it and expand treatment options. Furthermore, deciphering the nucleotide code of non-coding cis-regulatory regions is an intriguing goal that will contribute to dissecting the fundamental rules of transcription factor-binding and gene expression regulation.

With this in mind, the current project aimed to:

i. Optimize a lentiviral vector capable of detecting enhancer activity;
ii. Explore vector architecture and promoter suitability in current enhancer-testing vectors;
iii. Select and test a library of candidate enhancers, derived from loci with T2D-associated SNPs.

CHAPTER 3:
MATERIALS AND METHODS

1. CELL CULTURE

MIN-6 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) containing 4.5g/L D-glucose and 4mM L-alanyl-glutamine (GlutaMAX), and supplemented with 15% fetal bovine serum (FBS), 5% antibiotic–antimycotic and 50 nM 2-mercaptoethanol (all from Gibco, Thermo Fisher Scientific, CA, USA). HEK293T cells were cultured in the same DMEM, supplemented with 10% FBS and 5% antibiotic–antimycotic. All cells were grown in ventilated flasks or cell culture plates at 37 °C and humidified atmosphere with 5% CO2.

2. CELL STAINING

For staining MIN-6 cells with DAPI and phalloidin, cells were first washed with PBS and fixed with 4% Paraformaldehyde Aqueous Solution (Electron Microscopy Sciences) in PBS for 10 minutes at room temperature with agitation (50rpm). Following fixation, cells were washed three times with PBS at room temperature for 5 minutes, with agitation (50rpm. Subsequently, cells were incubated with PureBlu DAPI (BioRad; 1:100) and phalloidin-FITC (1:300) in PBS at room temperature for 20 minutes. After this incubation, cells were washed twice using the same procedure as before.

3. MICROSCOPY

Pictures of non-confluent MIN-6 cells (DAPI/phalloidin-FITC staining and mCherry-Cre infections) in 24-well plates were acquired using a Leica DMI6000b (Leica Microsystems) equipped with a Hamamatsu FLASH4.0 camera (Hamamatsu). All experiments were done under controlled temperature, atmosphere, and humidity. Images were analyzed using Fiji-ImageJ software.

4. LIBRARY SEQUENCE DESIGN

200-bp candidate enhancer sequences were selected by crossing T2D-associated SNPs from DisGeNET with cis-regulatory classes (active enhancers) identified by Miguel-Escalada and colleagues. For this, we defined 2kb regions centered on ATAC-seq peak summits (adult islets) of these defined "active enhancers" sequences and intersected them with the positions of T2D-associated SNPs. From this selection, we obtained two classes of candidate enhancers: (1) CRSs that contained an SNP in the center 200bp of the selected sequence; and (2) CRSs that contained an SNP outside the center 200bp but within the 2kb selected region. We selected a total of 354 CRSs from this analysis. In our final library, we included an additional 11 CRSs near crucial beta-cell function genes, 14 CRSs with T2D-associated SNPs previously validated in zebrafish embryos by Eufrásio et al., and 7 control sequences encompassing described silencers, a promoter region, the CMV enhancer, and a validated bifunctional sequence (enhancer/silencer) (Vaz et al., 2021) culminating in a library of 386 sequences.

5. PCR AMPLIFICATIONS

PCR reactions that involved amplifying sequences that would be inserted in our vectors (enhancers and promoters) were performed using the proofreading i-MAX II Taq DNA Polymerase (iNtRON Biotechnology, Inc) in a final volume of 20µL, containing 2µL of 10X PCR Buffer, 0,75 µL of Forward and Reverse Primers (10 µM), 2µL of dNTP mix (10mM/each dNTP), 0.5 µL of i-MAX II DNA Taq polymerase and nuclease-free water to complete the final reaction volume. Amplifications were performed at 94ºC for 3 minutes, followed by 35 cycles at 94ºC for 30 secs, 58-65ºC (depending on the melting temperature) for 30 seconds, 72ºC for 1 minute/kb, and a final extension at 72ºC for 10 minutes.

PCR reactions to confirm the insertions/deletions of sequences in our vectors were performed by directly collecting plasmid DNA by picking isolated colonies from the plate and "mixing" in a PCR mix containing 5µL of NZYTaq II 2× Green Master Mix (NZYTech), 0,5µL of forward and reverse primers (10µM) and 4µL of nuclease-free water. Amplifications were performed at

94ºC for 4min, followed by 30 cycles at 94ºC for 20s, 58ºC-65ºC for 20s, 72ºC for 30s/1kb, and a final extension at 72ºC for 5min.

All PCR amplifications were confirmed by analyzing the PCR product in a 1-3% agarose gel (depending on the size of the expected amplicon) stained with SYBR Safe (NZYTech). The ladder used was the NZYDNA Ladder VI (NZYTech) when observing amplicons with less than 1kb size and the GeneRuler 1kb DNA Ladder (Thermo Fisher Scientific) for larger than 1kb amplicons.

6. BACTERIAL TRANSFORMATIONS

For bacterial transformations, plasmid DNA was added to 50µL of chemically competent bacteria (Mach1) and incubated in ice for 30 min. Cells were heat-shocked for 30s at 42ºC and rapidly transferred into ice for 2 min. 750µL of Lysogeny broth (LB) was added to the cells followed by an incubation at 37ºC for 1h with shaking (250rpm). Cells were then plated in LB agar plates containing 100µg/mL ampicillin and incubated overnight at 37ºC.

7. ASSEMBLY OF VECTORS

7.1. pLS-mP-CRSEL

For the assembly of the pLS-mP-CRSEL vector, 200ng of the pLS-mP vector (Inoue et al., 2017) was digested with 1µL of the restriction enzyme *KpnI* (ThremoFisher) by incubating at 37ºC overnight in a solution with (*KpnI*), 2µL of FD Buffer, 0,2µL of DTT (100mM) and H2O to complete 20µL of final reaction volume. The digestion was then confirmed by running the product in a 1% agarose gel. The rest of the digestion product was again run in an agarose gel and the band corresponding with the linearized vector was subsequently isolated and purified using the NZYGelpure kit (NZYTech) following the manufacturer's protocol.

The oligo-generated CRSEL cassette, which contained overhangs complementary with the digested vector, was PCR amplified using primers (Forward: 5'-GTTTTCCAGTCACACCTCAG-3'; Reverse: 5'-AGTACTTTACCCCATGCTTGG-3') and then then cloned into the linearized pLS-mP

vector via Gibson assembly (NEBuilder, New England Biolabs). For this, a mix containing 0,06 pmols of insert (CRSEL amplicon), 0,03pmols of linearized vector (pLS-mP), and 10µL of NEBuilder HiFi DNA Assembly Master Mix (2X) (NEB) was incubated at 50ºC for 1h. 5µL of the Gibson solution was used to transform competent bacteria. The insertion of the CRSEL cassette in the pLS-mP vector was confirmed by PCR as mentioned in section 5. All confirmations of insertions/deletions downstream of the EGFP gene were done using primers (Forward: 5'-CGAGTCGGATCTCCCTTTGG-3'; Reverse: 5'-GAGAACAGTACTTTACCCCATGC-3') and following the protocol described in section 5.

Colonies that presented bands with the correct size were selected and picked to grow overday in 3mL of liquid LB media with ampicillin (100µL/mg). From that overday culture, 7,5µL was then added to 100mL of liquid LB media with ampicillin (1:500 dilution) for overnight culture. Plasmid DNA was then extracted using QIAGEN Plasmid Midi Kit (Qiagen), according to the standard kit protocol, and confirmed by digestion.

7.2. pLS-mP-C3

For the assembly of the pLS-mP-C3 vector, the pLS-mP-CRSEL vector was digested with *Esp3I* FD (Thermo Fisher Scientific) by incubating 200ng of the vector at 37ºC overnight in a solution with 1µL of the restriction enzyme (*Esp3I*), 2µL of FD Buffer, 0,2µL of DTT (100mM) and H2O to complete 20µL of final reaction volume. The digestion was then confirmed by running the product in a 1% agarose gel. The rest of the digestion product was again run in an agarose gel and the band corresponding with the linearized vector was subsequently isolated and purified using the NZYGelpure kit (NZYTech) following the standard protocol.

The C3-9 enhancer was PCR amplified with primers (Forward: 5'-TATGGATCCAGGACCGGATCAACTCACCACATCCCTTACCCTTACTGGA-3'; Reverse: 5'-TATGGATCCTCGGTTCACGCAATGGGCAATGCGGGCTCTTTTT-3´) using 50ng of human gDNA (Promega G3041) as a template, adding the necessary overhangs, and subsequently cloned in the digested plS-mP-CRSEL vector by Gibson assembly (NEBuilder, New England Biolabs). The resulting Gibson solution was used to transform competent bacteria using the same protocols described before. The insertion of the enhancer in the vector was confirmed by PCR as done

in section 7.1. Colonies that presented bands with the correct size were selected and picked to grow as described before. Plasmid DNA was extracted using QIAGEN Plasmid Midi Kit (Qiagen), according to the manufacturer's kit protocol followed by confirmation by digestion.

7.3. pLS-GATAmP-CRSEL/pLS-HSP70mP-CRSEL/pLS-FOSmP-CRSEL

The mP in pLS-mP-CRSEL was removed by digesting this vector with the *SbfI* and *AgeI* restriction enzymes. To accomplish this, 200ng of the pLS-mP-CRSEL was incubated at 37ºC overnight in a solution containing 2µL of each restriction enzyme, *AgeI* (New England Biolabs), and *SbfI* (New England Biolabs), 2µL of the rCutSmart Buffer and H2O to complete 20µL of final reaction volume. The restriction reaction was then run in a 1% agarose gel and the linearized vector without the mP was purified using the NZYGelpure kit (NZYTech).

The GATA2mP (Forward: 5'-TTTCTGCCAGGGCCCGCTCTAGACCTGCAGGAGCTTGCGGCGATATT CATTAAA-3'; Reverse: 5'-CTCGCCCTTGCTCACCATGGTGGCGACCGGTAGTGTCCGCGCTTAGAAA ATGC-3') and HSP70mP (Forward: 5'-GCCAGGGCCCGCTCTAGACCTGCAGGGAGCGCCGGAGTA TAAATAGAGGCGCTTCGTCTACGGAGCGACAATTCAATTCAAACAAGCAAAGTGAACACGT-3') were PCR amplified using the ZED vector (Bessa et al., 2009), and *Drosophila melanogaster* gDNA (kindly provided by André Carvalho from Morais-de-Sá group, i3S) as templates, respectively. The FOSmP was obtained as complementary oligonucleotides (Sense: 5'-TTTCTGCCAGGGCCCGCTCTAGACCTGCAGGGCACTCATTCATAAAACGCTTGTTATAAAAGCAGTG GCTGCGGCGCCTCGTACTCCAACC-3'; Antisense: 5'-CTCGCCCTTGCTCACCATGGTGGCGACCGGT TCGTCTTGGCTTCTCAGTTGCTCGCTGCAGATGCGGTTGGAGTACGAGGCGCCGCA-3').

All these promoters were then cloned in the digested plS-mP-CRSEL vector by Gibson assembly (NEBuilder, New England Biolabs), and the resulting Gibson solution was used to transform competent bacteria using the same protocols described before. The insertion of the promoters in the vector was confirmed by PCR. Confirming insertions/deletions of minimal promoters and insertions/deletions upstream of the minimal promoter were performed using primers (Forward: 5'-CCATGCTTCTTGTGTATATCCG-3'; Reverse: 5'-ACTTGTGGCCGTTTACGTCG -3'). Colonies that presented bands with the correct size were selected and picked to initiate bacterial cultures as described before. Plasmid DNA was extracted using a QIAGEN Plasmid

Midi Kit (Qiagen), according to the manufacturer's protocol followed by confirmation by digestion.

7.4. pLS-GATAmP-C3/pLS-HSP70mP-C3/pLS-FOSmP-C3

The C3-9 enhancer was cloned into the pLS-GATAmP-CRSEL, pLS-HSP70mP-CRSEL, and pLS-FOSmP-CRSEL following the same procedure and protocols for the assembly of the pLS-mP-C3 vector described above (7.2.).

7.5. pLS-CRSEL-mP

pLS-mP was digested with *XbaI* (Anza, Thermo Fisher Scientific) by incubating 200ng of the vector at 37ºC, overnight, in a solution with 1µL of the restriction enzyme (*XbaI*), 2µL of Anza 10x Red Buffer, and H2O to complete 20µL of final reaction volume. The digestion was then confirmed in an agarose gel and the linearized vector was isolated and purified using the NZYGelpure kit (NZYTech).

The CRSEL cassette was subsequently cloned into the linearized vector by Gibson Assembly (NEBuilder, New England Biolabs) as explained before. The resulting Gibson solution was also used to transform competent bacteria following the same protocol and the insertion was confirmed by PCR as done in section 7.3. Colonies that presented bands with the correct size were selected, picked, and grown as done before. Plasmid DNA was then extracted using QIAGEN Plasmid Midi Kit (Qiagen) and confirmed by digestion.

7.6. pLS-C3-mP/pLS-056-mP/pLS-510-mP

pLS-CRSEL-mP was digested using the *Esp3I* FD (Thermo Fisher Scientific) and the linearized vector was isolated and purified following the same steps as done for the assembly of the pLS-mP-C3. As before, the C3-9 enhancer, the CRS056 (Forward: 5'-GCACTCAGCCAACGCCATC-3'; Reverse: 5'-AGGTAATGTAGGCTATCATGGGAAC-3') and CRS510 (Forward: 5'-

ACTTGATGAGTGATTCAGGTGC-3'; Reverse: 5'-TGGCCCCATCTAAGGAAGCC-3') were PCR-amplified using human DNA as a template and adding overhangs complementary to the linearized vector. The amplicons were cloned using Gibson Assembly (NEBuilder, New England Biolabs). The Gibson solution was used to transform competent bacteria, plasmid DNA was extracted and the insertion was confirmed by PCR as done in section 7.3. Plasmid DNA was then extracted using QIAGEN Plasmid Midi Kit (Qiagen) and confirmed by digestion

7.7. pLS-C3

pLS-C3-mP was digested to remove the mP by incubating 1200ng of this vector at 37ºC, overnight, in a mix containing 2µL of the restriction enzyme *AgeI* (New England Biolabs), 2µL of the restriction enzyme *SbfI* (New England Biolabs), 2µL of the rCutSmart Buffer and H2O to complete 20µL of final reaction volume. The restriction reaction was then run in a 1% agarose gel and the linearized vector without the mP was removed and purified using the NZYGelpure kit (NZYTech). To allow the re-circularization of the digested vector we first removed the sticky ends left by the restriction enzymes. For that, we mixed 20ng of the digested and purified vector with 4µL of Klenow Fragment of DNA Polymerase I (NZYTech), 4µL of Klenow reaction buffer (10x), 1µL of dNTPs (2mM) and nuclease-free H2O to complete 50µL of final reaction. This mix was then incubated for 15 minutes at 25ºC. Then, EDTA was added to the solution to a final concentration of 10mM, and the solution was again incubated for 10min at 75ºC. To perform the blunt-end ligation, 20ng of the Klenow reaction product (linearized vector with blunt ends) was incubated for 1h at 22ºC in a reaction containing 5µL of T4 DNA Ligase (Thermo Fisher Scientific), 2µL of the corresponding 10x T4 DNA Ligase buffer, 2µL of 50% PEG 4000 solution and nuclease-free water to complete 20µL of reaction volume.

Following the aforementioned protocol, 5µL of the T4 ligation mix was used to transform chemically competent bacteria. To confirm the removal of the mP in the vector, plasmid DNA was directly collected by picking isolated colonies from the plate and a PCR reaction was performed as done in section 7.3. Colonies corresponding to PCR products with the right amplification size were grown and plasmid DNA was extracted using a QIAGEN Plasmid Midi Kit (Qiagen). Lastly, a final confirmation was done by digestion of the plasmid DNA.

7.8. pLS-C3-GATAmP

To generate pLS-C3-GATAmP, the mP in the pLS-C3-mP was replaced with the GATA2mP, following the same procedure and protocols applied for the assembly of the pLS-GATAmP-CRSEL from pLS-mP-CRSEL, described in section 7.3.

7.9. Cloning of enhancer library (pLS-Pool-mP/pLS-GATAmP-Pool)

The oligonucleotide pools containing our candidate enhancer library were synthesized by Twist Bioscience in two batches (Pool 1 - 99 oligos and Pool 2 - 286 oligos). The libraries were PCR amplified using primers (Forward: 5'-GGATCCAGGACCGGATCAACT-3'; Reverse: 5'-GGATCCTCGGTTCACGCAATG-3') in a PCR mix comprising 12.5µL of 2x KAPA HiFi HotStart ReadyMix, 0.75µL of forward and reverse primers (10µM), and nuclease-free H2O to reach a final reaction volume of 25µL. The amplification protocol consisted of an initial denaturation at 95ºC for 3 minutes, followed by 20 cycles of denaturation at 98ºC for 20 seconds, annealing at 58ºC for 15 seconds, extension at 72ºC for 20 seconds, and a final extension at 72ºC for 1 minute. The PCR product was then separated on a 2% agarose gel, and the band of interest was excised and purified using the NZYGelpure kit (NZYTech).

Both pLS-CRSEL-mP and pLS-GATAmP-CRSEL were digested with the restriction enzyme *Esp3I* FD (Thermo Fisher Scientific) as described before (sections 7.2, 7.4, and 7.6) and the digestion reaction was run in a 1% agarose gel. Linearized vectors were isolated and purified using the NZYGelpure kit (NZYTech).

The PCR amplified oligos libraries were cloned into the linearized vectors using Gibson Assembly (NEBuilder, New England Biolabs) by incubating at 50ºC for 1h a mix containing 25µL of NEBuilder HiFi DNA Assembly Master Mix, 240ng of the linearized vectors and the maximum possible volume of the PCR product to complete 50µL as a final reaction volume.

The aforementioned protocol was used to transform 100µl of MachI bacteria with 5µL of the Gibson solution (10 times for each Pool). To control the presence of undigested vector in our

libraries, the digested vector used in the Gibson reaction was also used to transform bacteria; none or very few colonies grew. Moreover, to confirm the insertion of the oligos in the vector, 30 colonies were tested by collecting plasmid DNA and performing a PCR reaction, as mentioned previously (section 5).

To ensure sufficient representation of each CRS in the library, about 30 000 colonies for pool 1 and 85 000 colonies for pool 2 were used to generate the CRSs lentiviral library, so that a ratio of about 300 colonies/enhancer would be achieved. A necessary number of plates to ensure those number of colonies were selected, and colonies were removed by adding 5mL of LB to each plate and scraping the colonies out of the medium using a cell scrapper. The liquid was then transferred to a 50mL Falcon and another 5mL was added, to the plates to collect any colonies left, and also collected. After that, plasmid DNA was extracted using the QIAGEN Plasmid Midi Kit (Qiagen) and following the manufacturer's protocol.

8. LUCIFERASE ASSAYS

Two candidate enhancer sequences, CRS056 and CRS510, were selected based on their epigenetic marks indicating enhancer activity, such as accessible chromatin (ATAC-seq signal) and detection of the histone modification H3K27ac (ChIP-seq signal). These sequences were PCR-amplified using human gDNA (Promega G3041) as a template and following the same protocol described before. Initially, we TA-cloned these PCR products into pCR8/GW/TOPO (Invitrogen – Thermo Fisher Scientific), which was then used for recombination of the PCR product into the luciferase destination vector, pGL4.23-GW (Pasquali et al., 2014), through Gateway cloning using Gateway LR Clonase II (Invitrogen, Thermo Fisher Scientific). The empty pGL4.23-GW vector and non-transfected cells were used as negative controls, and the TK promoter from pGL4.54[luc2_TK], cloned into the pGL4.23-GW, was used as a positive control.

MIN-6 cells were transfected in a 6-well plate with 1 pmol of the pGL4.23-GW and 0.61 fmol of pNL1.1PGK[Nluc/PGK] per well with Lipofectamine 3000 (Invitrogen), according to manufacturers' instructions. Cells were seeded about 48 hours before transfection. Luciferase activity was assessed 48 h post-transfection using the NanoGlo Dual-Luciferase Reporter Assay System (Promega) in a Synergy 2 Multi-Mode Microplate Reader (BioTek). The enhancer

activity is expressed as the ratio of luc2/Nluc (pGL4.23-GW/pNL1.1PGK [Nluc/PGK]) and was presented as a relative response ratio, in which the activity of non-transfected cells and positive control vector were normalized to 0 and 1, respectively.

9. LENTIVIRAL PRODUCTIONS

Lentivirus was produced by transfecting HEK293T cells at 60/70% confluency. For that, a transfection mix was prepared by combining Lipofectamine 3000, P3000, the packaging vectors (pRSV-Rev (3 µg), pMD2.G (3 µg), pMDLg/pRRE (5 µg)), and the vector of interest (10 µg) in 2ml of Opti-MEM, according to manufacturers' instructions. The components were mixed gently by pipetting, to ensure thorough homogenization. The transfection mix was then incubated at room temperature for 20 minutes to allow for complex formation. Prior to transfection, the culture medium was aspirated and replaced with 13 mL of fresh complete medium (DMEM, 10% FBS). Following the 20-minute incubation, 2 mL of the transfection mix was slowly added to each T75 flask containing the HEK cells. The mix was gently swirled in the flask to ensure an even distribution of droplets and proper contact with the cells.

48 hours after transfection, the medium from the transfected HEK cells was collected and processed as follows: the medium (15 mL) was carefully removed from the cells and transferred to a 50 mL tube. The tube was then centrifuged at 1200 rpm for 5 minutes at room temperature (RT); after centrifugation, the supernatant was filtered through a 0.45 µm filter into a new 50 mL tube, to remove any remaining cell debris.

To concentrate the virus, a 4x Lentivirus Concentrator Solution was used. The Lentivirus Concentrator was prepared by dissolving 160 g of PEG-8000 and 28 g of NaCl in 160 mL of Milli-Q water and 40 mL of 10X PBS (pH 7.4). The solution was mixed gently with stirring and low heating until all solids were dissolved. The pH of the solution was then adjusted to 7.0-7.2, and the final volume was completed to 400 mL. The solution was sterilized by filtering it through a 0.2 µm PES filter. It was ensured that the solution was agitated for over an hour before filtering to ensure proper dissolution of any precipitate formed. 5 mL of the 4x Lentivirus Concentrator Solution, which constituted one-fourth of the final volume, was added to the filtered supernatant. The contents of the 50mL tube were gently mixed by inversion.

Subsequently, the tube was securely closed and incubated for 48h with continuous agitation at 25rpm, in a closed box, at 4ºC.

After incubation, the 50mL tubes were subjected to centrifugation at 4000 rpm for 1 hour at 4ºC. This centrifugation step allowed the viral particles to pellet (and concentrate) at the bottom of the tube. The supernatant was then carefully removed by aspiration. To resuspend the viral pellet, for each T75 flask of HEK cells used for transfection, 250 µL of ice-cold PBS was added. The resuspension was performed gently to ensure thorough mixing of the viral particles with the PBS solution. The resuspended viruses were aliquoted into appropriate volumes and subsequently frozen at -80ºC. This aliquoting step aimed to facilitate future use without repeated freeze-thaw cycles that could potentially affect the viability of the viruses.

10. INFECTIONS

MIN-6 cells were infected with the lentivirus using the appropriate volume necessary for each experiment, in medium containing 8µg/ml polybrene. After 24h, the medium was then replaced with fresh medium without polybrene. Phenotypes were analysed at least 72h post-infection.

11. VIRAL TITRATIONS AND MOI CALCULATIONS

For the titration of the virus solutions, cells were infected with 0.5, 1, 2, 4, and 8µL/ml of each virus solution in MIN-6 medium containing 8µg/ml polybrene. After 24h, the media was replaced with fresh media not containing polybrene. After 72h, genomic DNA was extracted from each well/condition. For this, cells were washed with PBS, trypsinized, and recovered in 1.5 mL tubes. This solution was then submitted to centrifugation and the cell pellet was washed with PBS, re-suspended in 10mM Tris-HCl, 0.1mM EDTA, and treated with 1µl of proteinase K at 56°C for 30 min, followed by proteinase K inactivation at 95°C for 10 min. qPCR was performed with iTaq Universal SYBR Green Supermix (Bio-Rad) in a CFX96 Real-Time PCR Detection System (Bio-Rad) according to manufacturers' instructions. 30ng of the extracted gDNA and H2O were used as template controls. Shortly, a sequence near the *Cryaa* gene was

used as a reference of the MIN-6 genome (mmCryaa- Forward: 5'-GCCGTGGTAGCTGAAGAGAC-3'; Reverse: 5'-GCCAGCTAAGATGCACATGA-3'); a sequence in the non-integrating part of the lentiviral vector (BB- Forward: 5'-TGCCGCATAGTTAAGCCAGTA-3'; Reverse: 5'-TCAAGCCTTGCCTTGTTGTAG-3') was used as a measurement of background vector contamination; and a sequence in the integrated lentiviral cassette (WPRE- Forward: 5'-TACGCTGCTTTAATGCCTTTG-3'; Reverse: 5'-GGGCCACAACTCCTCATAAAG-3') was used to detect the genome-integrated virus. 3 technical replicates were used per target sequence. Results were analyzed using the CFX Maestro software (Bio-Rad) and each point of the titration curve was calculated using the following equation:

$$\text{MOI} = \left(2 \times 2^{Cq(mmCryaa)-Cq(WPRE)}\right) - \left(2 \times 2^{Cq(mmCryaa)-Cq(BB)}\right)$$

12. FLOW CYTOMETRY

For flow cytometry analysis, cells were previously infected in 24-well plates and processed at least 72h post-infection. Cells were first washed with PBS. Then, 200µL of Tryple was added to each well and incubated for 10min at 37ºC. Cells were then collected in 1000 µL of PBS, resulting in a final volume of 1200 µL, and transferred into a 1.5mL tube. After centrifugation at 1200 rpm for 3 minutes, the supernatant was removed and the pellet was re-suspended in FACS Buffer (PBS 1X with HEPES 1M and EDTA 0.5M at a final concentration of 25mM and 2mM, respectively) containing 1:1000 Fixable Viability Dye eFluor 780 (eBioscience), incubated for 15min at R.T. while covered from direct light exposure. After incubation, cells were centrifuged at 1200rpm for 3 min, the supernatant was removed and the cells were fixed by re-suspending the pellet in PBS with 4% Paraformaldehyde (Electron Microscopy Sciences) and incubating at R.T for 10min. Lastly, cells were centrifuged and processed as before, and the pellet was re-suspended in FACS Buffer and transferred to FACS tubes.

Cells were then analyzed in a BD FACSCanto II or BD LSRFortessa flow cytometer (BD Biosciences, San Diego, CA, USA) using the blue laser (488 nm), the red laser (633 nm) and, in the BD LSRFortessa, also the yellow-green laser (561 nm). Cells were gated by forward scatter area (FSC-A) vs. side scatter area (SSA-A), FSC-A vs. FSC-height (FCS-H), and FSC-A vs. red laser

(vital dye) plots to exclude doublets/clumps and dead cells, respectively. To access GFP and mCherry, non-infected cells were gated by FCS-H vs blue laser and blue laser vs yellow-green laser, respectively, to define intensity thresholds.

Flow cytometry data was analyzed using the FlowJo v10 Software (BD Biosciences). The statistical analysis was performed using Microsoft Office Excel and GraphPad Prism 8 software. The average percentage of GFP-positive cells obtained in experiments using different vectors was compared using an unpaired Mann-Whitney t-test when comparing two conditions and a Kruskal-Wallis One-way ANOVA test for multiple comparisons. For the GFP intensity of GFP-positive cells, individual GFP intensity values were obtained and outliers were removed (ROUT; Q=1%). After this, the GFP intensity values are normalized by dividing each value of GFP intensity by the mean GFP intensity obtained in the GFP-positive cells from the control group (empty vector). Comparisons were performed using an unpaired Kolmogorov t-test. ns, $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, and ****$p \leq 0.0001$. Values are shown as mean ± S.D.

13. FLUORESCENT-ACTIVATED CELL SORTING (FACS)

For the sorting of cells, cells were processed as mentioned for flow cytometry, excluding the fixation step. Cells were then analyzed in a BD FACSARIA II cell sorter (BD Biosciences, San Diego, CA, USA) using the blue laser (488 nm) and the red laser (633 nm). Cells were gated by forward scatter area (FSC-A) vs. side scatter area (SSA-A), FSC-A vs. FSC-height (FCS-H), and FSC-A vs. APC-Cy7 (vital dye) plots to exclude dead cells and doublets/clumps. The gates were established based on the autofluorescence of non-infected cells. Single GFP-positive cells were sorted directly into 96-well plates containing 200 µl/well of MIN-6 medium to establish monoclonal reporter cell lines. On top of that, pools of GFP-negative, GFP-positive, and “Non-Sorted” cells were sorted and recovered in 700 µl of Buffer AL (Qiagen) and genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen), according to manufacturers’ instructions.

14. REPORTER LINES

After seeding individual GFP+ cells in 96-well plates, these cells were maintained and grown in MIN-6 medium and kept at 37 °C and a humidified atmosphere with 5% CO2. As these cells tend to aggregate in 3D clusters, which compromises their growth and proliferation, after two weeks, cells were trypsinized twice a week to break down these cell-to-cell interactions and allow for more homogeneous and efficient cell proliferation.

15. CRE RECOMBINASE ASSAY

Cells were treated with the Cre-mCherry lentivirus following the standard infection protocol described above. After a week, we confirmed the integration of the Cre Recombinase gene in our cells by microscopy, by assessing the co-expressed red fluorescent mCherry protein (Operetta CLS (Perkin Elmer) or Leica DMI6000 (Leica)). The phenotypes associated with the resulting activity of the Cre Recombinase were also analyzed, at least one-week post-treatment, via microscopy or by flow cytometry, following the protocols described before.

16. GENOTYPING

Established reporter cell lines and CRE recombinase-mediated deletions were validated. Cells were resuspended in 10mM Tris-HCl, 0.1mM EDTA, and treated with 1µl of proteinase K at 56°C for 30 min, followed by proteinase K inactivation at 95°C for 10 min. The extracted DNA was used for genotyping with NZYTaq II DNA polymerase (NZYTech) following the manufacturer's instructions. The insertion of a predicted enhancer in our cell lines and the excision of this sequence after being treated with the CRE Recombinase protein were confirmed in an agarose gel by the presence of an amplicon with the expected size.

17. NEXT GENERATION SEQUENCING

Cells transduced with either Pool-mP or GATAmP-Pool constructs were FACS-sorted into GFP-negative, GFP-positive, and “Non-sorted” groups. Subsequently, their gDNA was extracted using the DNeasy Blood and Tissue kit (Qiagen), according to the manufacturer’s instructions. Candidate enhancers were PCR amplified from these samples and the PCR products were sequenced with the NovaSeq 6000 System (Illumina). The sequencing data was processed using standard bioinformatic tools by our lab (Ferreira et al., 2023). The normalized number of reads, from three technical replicates, of a given CRS present in the group of GFP-positive cells was plotted against the normalized number of reads of the same CRS present in the group of GFP-negative cells. Sequences with a p-value<0.1 were considered significantly enriched in one of the populations.

CHAPTER 4:
RESULTS AND DISCUSSION

1. MORPHOLOGICAL ANALYSIS OF TWO MIN-6 CELL SUBLINES

The MIN-6 cell line is one of the few cell lines that exhibit characteristics similar to pancreatic beta-cells, including insulin production and secretion in response to glucose. These characteristics are unstable and compromised when these cells are kept in culture for long periods of time (Ishihara et al., 1993; O'Driscoll et al., 2006). Numerous studies have shown that high passage (HP) MIN-6 cells had impaired insulin production in response to glucose stimulation and changed gene expression profiles in comparison to their low passage (LP) counterparts, leading some authors to hypothesize that these differences are a result of HP cells losing some of their cell type-specific traits acquired during differentiation (Cheng et al., 2012; O'Driscoll et al., 2006).

As a preliminary experiment, we conducted a basic morphological analysis on two independent MIN-6 cell lines, obtained from two different sources: an HP line gently provided by collaborators (passage 70) and a new LP line (passage 20). These cells were seeded at low confluency to better visualize single-cell morphology and the formation and morphology of cell clusters sometimes referred to as pseudo islets (Drzazga et al., 2020; Hauge-Evans et al., 1999). We fixed cells and stained their DNA and actin filaments using DAPI and FITC-phalloidin, respectively, to properly examine their nuclei and cytoplasm. We then used fluorescence microscopy to acquire representative images of the two cell cultures. Clear differences were evident between the two groups. LP MIN-6 cells generally exhibited a mostly even, round shape (Fig. 1A), while HP cells displayed an irregular shape with membrane protrusions, similar to lamellipodia (Fig. 1B), a phenotype previously demonstrated by other authors (Cheng et al., 2012; O'Driscoll et al., 2006). Additionally, the images revealed the tendency of both cell sublines to form large 3D clusters. This structure is particularly relevant for following assays as it might impact the efficiency of transfection and infection of these cells, a topic further discussed later. It has also been shown that MIN-6 cells can be induced to form spherical pseudo islets, which structurally resemble primary islets (Hauge-Evans et al., 1999; Persaud et al., 2014). These structures are notably more similar to the large clusters observed in LP cells (Fig. 1A) rather than those formed in HP cells (Fig. 1B).

Given these considerations, it was crucial to account for the cell passage while conducting our assays. We aimed to test candidate enhancers in cells that could closely resemble the natural

phenotypes of beta cells, and the notable variations in these cells' morphology and organization could also potentially compromise the reproducibility of our enhancer assays. Therefore, in downstream assays, we consistently ensured that the enhancer testing was carried out in LP MIN-6 cells that shared morphological characteristics between them, namely in cultures composed mainly of small cell clusters.

It's crucial to note that while these cells were initially described as transformed differentiated beta-cells (Miyazaki et al., 1990), it has been documented that these cells express differentiation factors related to other endocrine cells. Moreover, these cells produce multiple pancreatic hormones, indicating that they possess not only beta cell features but also islet characteristics (Nakashima et al., 2009). Given this context, it's important to consider that the heterogeneity of these cells might also influence our outcomes when assessing human islet enhancer activity.

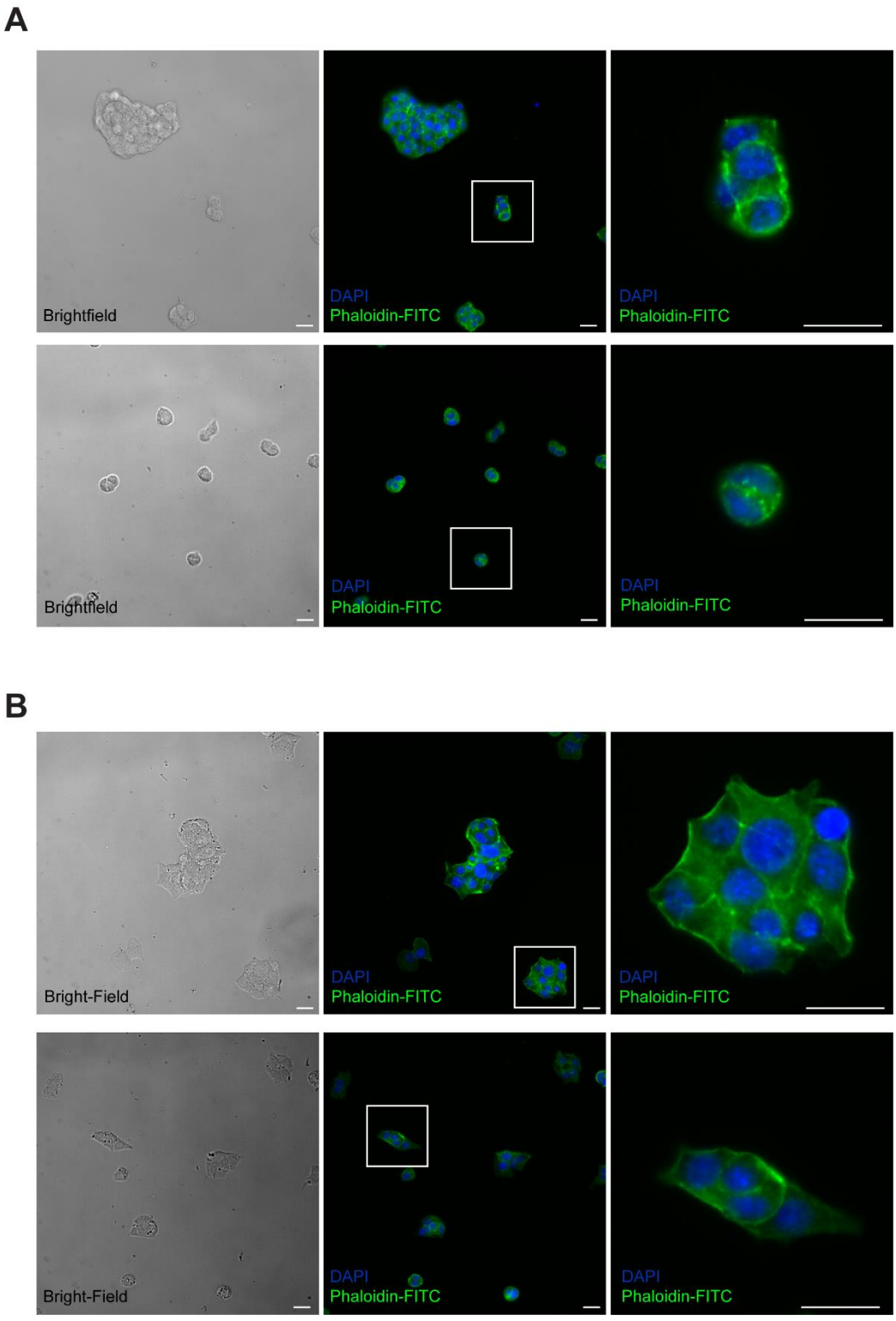


Figure 1. Morphology of LP and HP MIN-6 cells. Representative fluorescence microscopy images of Low Passage MIN-6 cells **(A)** and High Passage MIN-6 cells **(B)** at low confluency. Top images show bigger cell clusters and bottom pictures show smaller cell clusters. Cells were stained with DAPI (blue) and phalloidin (green). Pictures were acquired with a 40X objective, 400x amplification. Scale: 10µm

2. DEVELOPMENT OF NEW LENTIVIRAL VECTORS FOR USE IN MPRAs

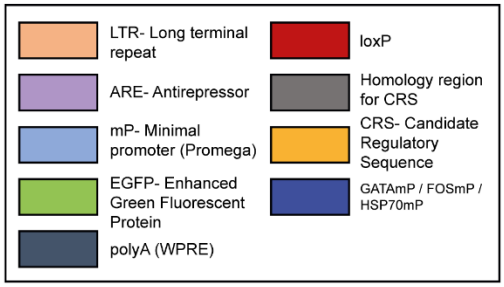
In order to obtain the most reliable and comprehensive results from our experiments, we decided to employ a lentiviral vector to assess the enhancer activity of the candidate enhancer sequences. This approach introduces a relevant layer of complexity to the assay, as the vectors' cassettes integrate into the genome of the cells and are, thereby, evaluated in the presence of chromatin. Testing enhancers within chromatin is crucial as it allows us to examine these regulatory elements in a biologically relevant setting, where their interplay with chromatin structure is considered. Furthermore, Inoue and colleagues have shown that chromosomally integrated sequences exhibit greater reproducibility and significant differences in activity than the same sequences assessed on episomes (Inoue et al., 2017).

The foundational design of all our vectors is based on the pLS-mP vector (Inoue et al., 2017), which serves as the blueprint for our constructs. This vector consists of a 32bp synthetical minimal promoter (mP) and the enhanced green fluorescent protein (EGFP) flanked by two distinct antirepressors (Fig. 2A). These antirepressors (ARE) have been shown to decrease the interaction between our vector and nearby cis-regulatory elements after integration in the genome (Inoue et al., 2017). Previous to this work, our group inserted a cassette downstream of the EGFP gene, designated CRSEL, for Candidate Regulatory Sequence Entry Locus, to provide a platform for the insertion of the candidate enhancers (Fig. 2B). To allow this insertion via Gibson assembly, the synthetic cassette included two homology arms that were compatible with the candidate enhancers we intended to clone into the vector. Additionally, this cassette also includes two loxP sites flanking the enhancers insertion site, which would be crucial for removing these sequences when exposed to the Cre recombinase protein in the following experiments.

To explore a novel approach, in this work we conducted tests to determine whether our newly constructed vector (pLS-mP-CRSEL) could effectively detect enhancers downstream of the reporter gene, as opposed to the conventional method of placing enhancers upstream of the minimal promoter. This approach had not been previously fully demonstrated, particularly through the detection of GFP expression and intensity. To assess the vector's suitability in this context, we employed Gibson assembly to clone the C3-9 enhancer (911bp), a previously validated enhancer in the MIN-6 cell line (Pasquali et al., 2014), into the pLS-mP-CRSEL vector

(Fig. 2C and Fig. 2D). In addition, we examined the functionality of three other minimal promoters selected from the literature within this framework: GATA2mP, a 1040bp promoter region from the zebrafish *gata2* gene previously used to assess enhancer activity of sequences placed downstream the reporter gene (Bessa et al., 2009), FOSmP, a 95bp sequence previously defined as the basal promoter of the human *FOS* and used in MPRA assays (Nguyen et al., 2016) and HSP70mP, a 133bp minimal promoter from the from Drosophila melanogasters *hsp70* gene previously used in fly for enhancer assays (Barolo et al., 2000; Mikhaylichenko et al., 2018) (Fig. 2E).

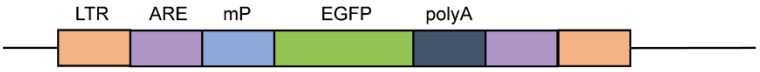
On top of that, we also wanted to compare this approach with the conventional strategy of placing enhancers upstream of the promoter. Thus, we used different cloning strategies, namely digestion/ligation and Gibson Assembly, to insert the CRSEL cassette upstream of the synthetical minimal promoter, as well as to clone the C3-9 enhancer into the same vector. These strategies were successful and resulted in the creation of the pLS-CRSEL-mP (Fig. 2F) and pLS-C3-mP (Fig. 2G) vectors, respectively. In parallel, we also employed this approach using the GATA2mP to further explore promoter-enhancer specificity (Fig. 2G). These vectors later allowed us to compare the capacity of our system to detect the activity of candidate enhancers placed downstream of the reporter gene with the more commonly used method of placing candidate enhancers upstream of the promoter. Lastly, as a control, we removed the minimal promoter from the pLS-C3-mP, originating the pLS-C3, to assess the promoter activity of the C3-9 enhancer (Fig. 2H).



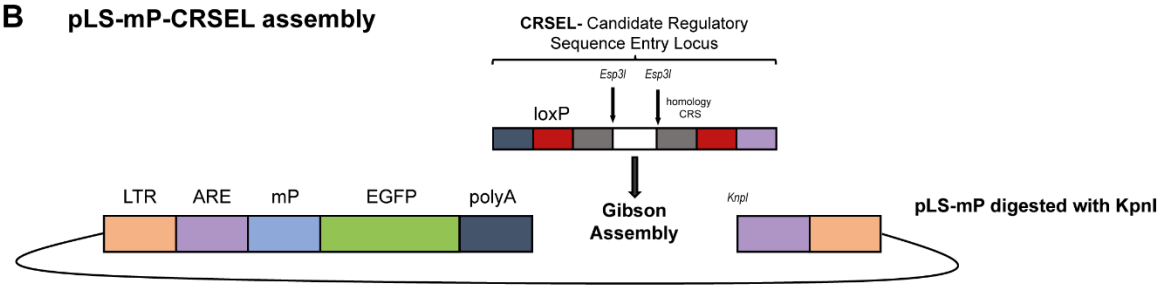
mP* - mP(Promega), GATAmP, HSP70mP or FOSmP

mP** - mP(Promega) or GATAmP

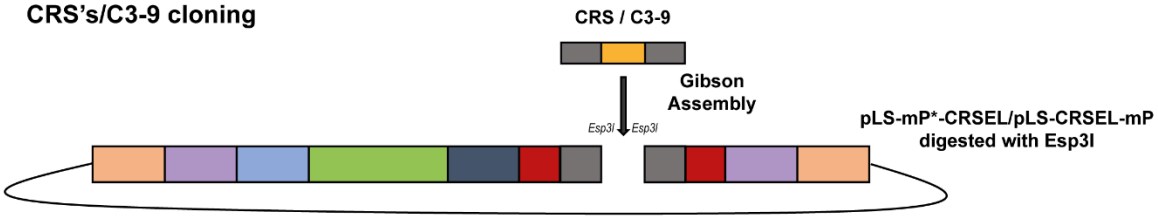
A **pLS-mP**



B **pLS-mP-CRSEL assembly**



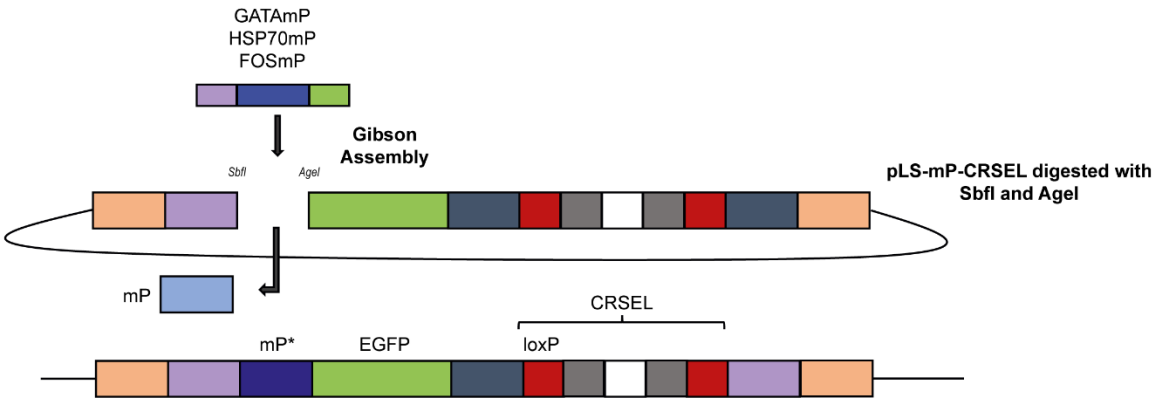
C **CRS's/C3-9 cloning**



D **pLS-(mP*)-C3**



E **pLS-(GATAmP/FOSmP/HSP70mP)-CRSEL assembly**



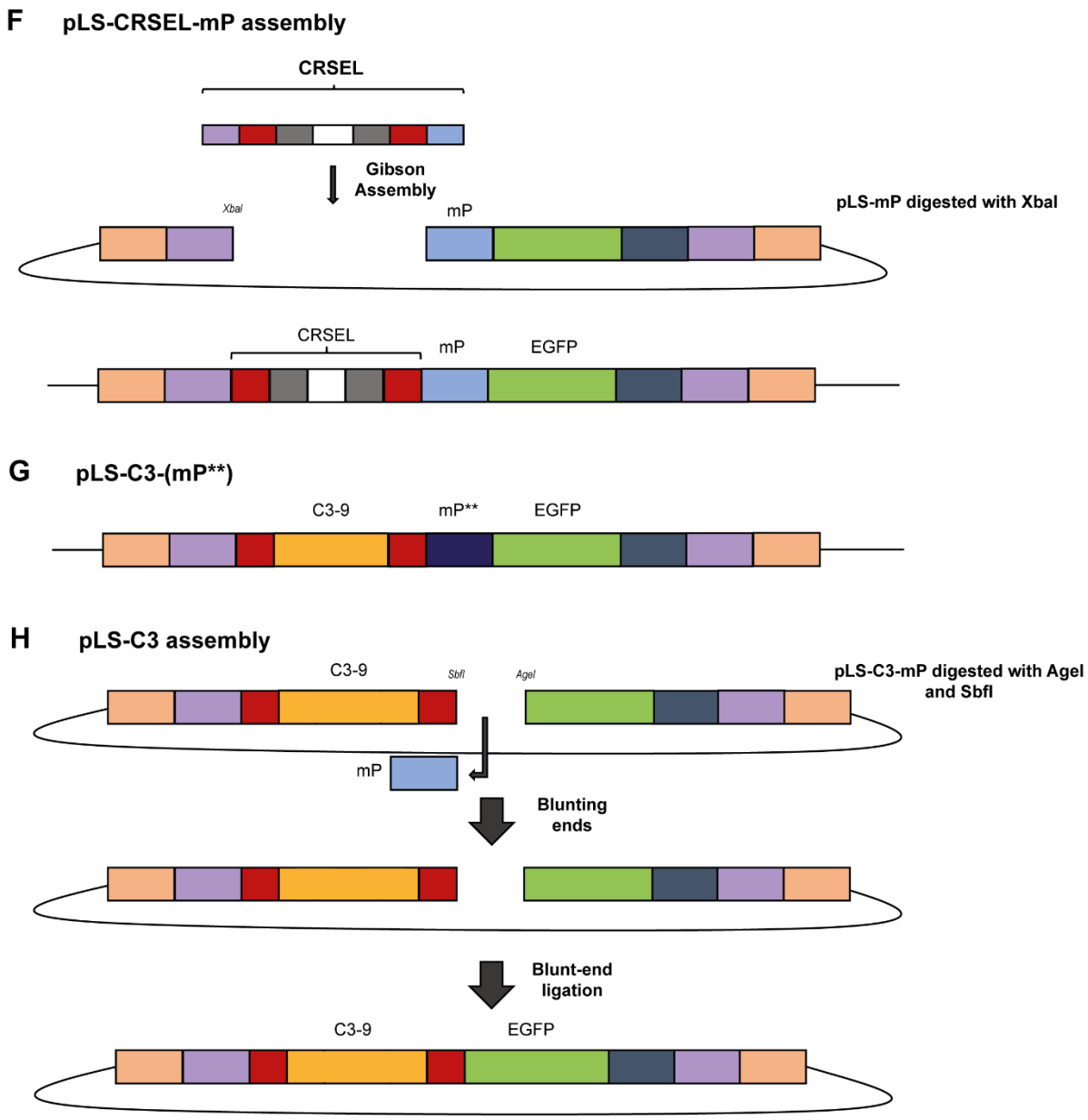


Figure legend in the next page

Figure 2. Schematic representation of lentiMPRA vectors used in this work. (A) Original pLS-mP vector containing a synthetic mP, the EGFP gene, and two ARE flanking these sequences. **(B)** Assembly of the pLS-mP-CRSEL vector. pLS-mP was digested using the *KpnI* restriction enzyme. The CRSEL cassette, which contained two loxP sites flanking the homology regions for the insertion of CRSs, was inserted in the linearized vector via Gibson assembly. **(C)** Introduction of the C3-9 enhancer or other CRS in the CRSEL cassette. Vectors containing this cassette were digested with the *Esp3I* restriction enzyme and the sequences were inserted in the linearized vector via Gibson assembly. **(D)** pLS-mP-C3, pLS-GATAmP-C3, pLS-FOSmP-C3 and pLS-HSP70mP-C3 vectors. **(E)** Assembly of the pLS-GATAmP-CRSEL, pLS-FOSmP-CRSEL, and pLS-HSP70mP-CRSEL vectors. The synthetic mP in pLS-mP-CRSEL was removed by digesting this vector with the *SbfI* and *AgeI* restriction enzymes and the other promoters were inserted via Gibson assembly. **(F)** Assembly of pLS-CRSEL-mP vector. pLS-mP was digested with *XbaI* and the CRSEL cassette was inserted via Gibson assembly. **(G)** pLS-C3-mP and pLS-C3-GATAmP vectors. **(H)** Assembly of the pLS-C3 vector. pLS-C3-mP was digested using *AgeI* and *SbfI*, removing the synthetic mP. The sticky ends from this restriction enzyme reaction were blunted and the vector was re-circularized via ligation.

In our analysis, we focused on two key parameters that would be most indicative of our vector's ability to detect enhancer activity. These parameters were selected to thoroughly assess the performance of our vectors. The first parameter involved measuring the percentage of GFP-positive cells in comparison to the corresponding control. This allowed us to quantify the proportion of cells that expressed GFP as an indicator of enhancer activity. The second parameter was the normalized GFP intensity of those GFP-positive cells. To achieve this, we divided the GFP intensity of each GFP-positive cell by the mean GFP intensity of GFP-positive cells in the control group (empty vector). Normalizing the GFP intensity helped account for variations in overall GFP expression levels in response to each minimal promoter and allowed us to compare the enhancer activity more accurately across different experiments.

By considering both the percentage of GFP-positive cells and the normalized GFP intensity, we hoped to gain a comprehensive understanding of how efficiently our vectors were able to read enhancer activity and, consequently, draw meaningful conclusions from our experiments. In fact, overall, the analysis of these parameters provided valuable insights into the impact of the enhancers on GFP expression and allowed the assessment of our vector's performance in a robust and reliable manner, as described below.

These results were obtained by infecting low passage MIN-6 cells with the lentiviral vectors mentioned above and performing flow cytometry analysis on these samples, 72 hours later.

All these vectors were titrated before transduction in these experiments to prevent multiple integrations of the viral cassette per cell and to ensure a similar percentage of infected cells with each vector, enabling a correct comparison. This involved infecting cells with each lentivirus at various virus solution concentrations and then performing qPCR analysis to determine the ratio of integrated vectors to the number of cells at each tested concentration (MOI). Using these values, we constructed titration curves that represented the relationship between the lentivirus solution volume (x) and MOI (y). These curves provided the equation needed to predict the volume to use for a specific MOI.

2.1. CLONING OF THE C3-9 ENHANCER DOWNSTREAM OF THE REPORTER GENE LIMITS DETECTION OF ENHANCER ACTIVITY

To assess the impact of different vector architectures on the ability of our system to detect enhancers, we first placed the C3-9 enhancer downstream of the EGFP gene as described before. This design provides a system that requires enhancers to interact with their target promoter through the formation of chromatin loops to promote transcription, thus resembling their likely natural genomic context and potential dynamics. Considering that the promoter nucleotide sequence likely limits the protein-DNA and protein-protein interactions required for loop formation (Kadauke & Blobel, 2009; Mahmoudi, 2002), to test this vector architecture we evaluated four minimal promoters, as previously mentioned. We used these constructs, produced lentivirus, and infected MIN-6 cells with a MOI of 1. We assessed enhancer activity by cytometry, in particular by quantifying the number of GFP-positive cells and the GFP intensity.

When employing the synthetic Promega mP, both the empty vector and the vector containing the C3-9 enhancer yielded 0.2% GFP-positive cells. Meanwhile, using the GATA2mP and the FOSmP, both empty vectors and vectors with the enhancer resulted in approximately 5% GFP-positive cells, and vectors with the HSP70mP produced around 10% GFP-positive cells (Fig. 3A). When examining the normalized GFP intensity of the GFP-positive cells, which is normalized using the mean intensity associated with their corresponding empty vectors, all conditions involving the C3-9 enhancer are associated with no increase in the mean GFP intensity (Fig. 3B).

Our findings indicate that this system does not effectively detect the activity of enhancers placed downstream of the GFP gene, regardless of the promoter used (Fig. 3A and Fig. 3B). Although some differences from the vector with the HSP70mP are observed when comparing the vector containing the C3-9 enhancer with the empty vector (an increase of around 2% of GFP-positive cells), we also detected a small decrease in the percentage of GFP-positive cells associated with the enhancer C3-9 in the vectors with GATAmP and FOSmP, which was not expected. The discrepancy between these results may suggest that these subtle differences are not significant and that the GFP expression we observed might be primarily driven by the promoter activity alone, with the enhancer contribution being undetected. Technical inconsistencies, such as differences in the MOI (Multiplicity of Infection) could be one of the reasons for these small changes between the empty vector and vector with the enhancer throughout replicates.

Another factor that can contribute to the small differences in the percentage of GFP-positive cells observed in our experiments is the way these cells tend to grow in 3D clusters. This cellular organization may difficult the contact between cells positioned inside of the cluster and the viral particles, decreasing infection efficiency. Differences in the number and size of the clusters of cells could lead to different efficiencies of infection, which then could lead to differences in the percentage of GFP-positive cells.

Although enhancers and promoters can interact thousands of base pairs apart via chromatin loops in endogenous conditions (Andersson & Sandelin, 2020; Pennacchio et al., 2013), this interaction may be compromised in our system due to a lack of chromatin loop formation. A hypothesis for the absence of chromatin of loops is the small distance between the promoter and the enhancer, which are separated only by the EGFP gene and the polyA element WPRE (woodchuck hepatitis virus post-transcriptional regulatory element), totalizing less than 1000bp. In vertebrates, enhancers are, on average, located about 20-50kb away from their target promoters (Furlong & Levine, 2018). However, determining the minimal distance needed for a chromatin loop to form and for these regulatory elements to interact is challenging due to the limited resolution (≈10kb) of chromatin conformation techniques like Hi-C (Y. Zhang et al., 2018).

Another factor that can potentially impair the formation of loops is the fact that antirepressors flank the gene cassette composed of the promoter, reporter gene, and enhancer. Although it

is not clear how these specific antirepressor sequences exert their insulator activity, these non-coding sequences might also work via chromatin loops (Ali et al., 2016) which, in turn, can interfere with the formation of new loops, preventing the interaction between the promoter and the enhancer.

Minimal promoters are typically defined as the smallest functional part of the promoter that maintains promoter activity, which usually includes the core promoter, a 50-100bp region where transcription starts (Maston et al., 2006). Our experiments also suggest that the GATAmP, FOSmP, and HSP70mP promoter sequences exhibit considerably higher basal activity than the synthetic mP sequence. Thus, the relatively high activity of these 3 minimal promoters could difficult the detection of an increase in GFP expression driven by an enhancer.

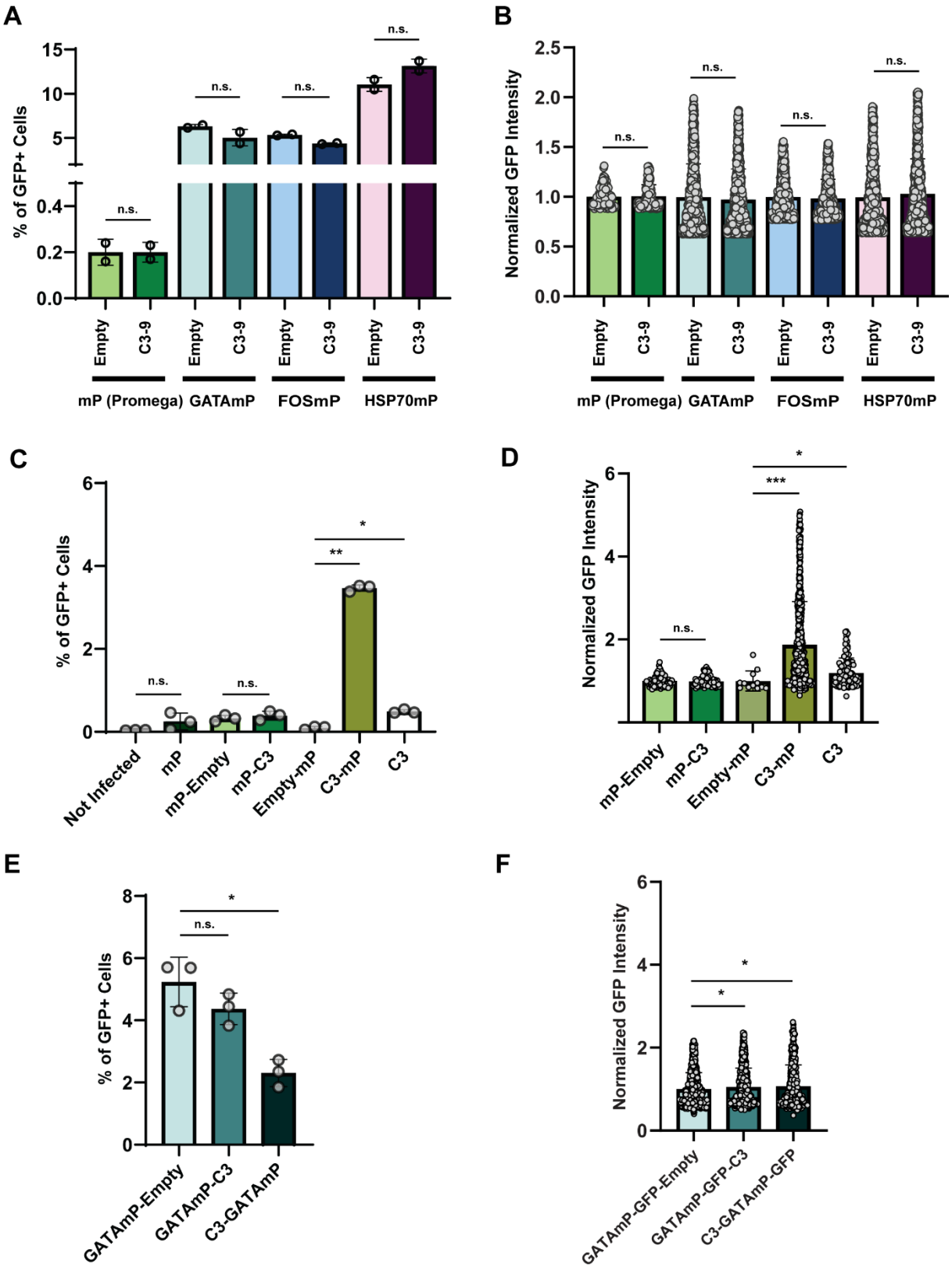


Figure legend in the next page

Figure 3. Assessment of the enhancer detection ability of the constructed vectors. (A) Average percentage of GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-mP-C3, pLS-GATAmP-C3, pLS-FOSmP-C3, pLS-HSP70mP-C3, and their respective empty vectors. n.s.: not statistically significant from unpaired t-test (non-parametric- Mann-Whitney test). **(B)** Normalized intensity of GFP from GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-mP-C3, pLS-GATAmP-C3, pLS-FOSmP-C3, pLS-HSP70mP-C3, and their respective empty vectors (n=1000 except mP-Empty and mP-C3 where n≈200). Normalization was performed by dividing each value of GFP intensity by the mean GFP intensity obtained in the GFP-positive cells from the respective empty vector. n.s.: not statistically significant from unpaired t-test (non-parametric Kolmogorov test). **(C)** Average percentage of GFP-positive cells from infections (MOI=1) with lentiviruses produced using the pLS-mP, pLS-mP-CRSEL, pLS-mP-C3, pLS-CRSEL-mP, pLS-C3-mP and pLS-C3 vectors. n.s.: not statistically significant; * p value≤0.05; ** p value≤0.01 by non-parametric ANOVA test (Kruskal Wallis test). **(D)** Normalized GFP intensity of GFP from GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-mP, pLS-mP-CRSEL, pLS-mP-C3, pLS-CRSEL-mP, pLS-C3-mP and pLS-C3 vectors (n≈200 except pLS-C3-mP where n=1000). Normalization was performed by dividing each value of GFP intensity by the mean GFP intensity obtained in the GFP-positive cells from the respective empty vector. n.s.: not statistically significant; * p value≤0.05; *** p value≤ 0.001 from unpaired t-test (non-parametric Kolmogorov test). **(E)** Average percentage of GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-GATAmP-CRSEL, pLS-GATAmP-C3, and pLS-C3-GATAmP vectors. n.s.: not statistically significant; * p value≤0.05 from unpaired t-test (non-parametric Mann-Whitney test). **(F)** Normalized intensity of GFP from GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-GATAmP-CRSEL, pLS-GATAmP-C3 and pLS-C3-GATAmP vectors (n=1000). Normalization was performed by dividing each value of GFP intensity by the mean GFP intensity obtained in the GFP-positive cells from the empty vector. n.s.: not statistically significant; * p value≤0.05 from unpaired t-test (non-parametric Kolmogorov test).

2.2. POSITION OF C3-9 ENHANCER IMPACTS EXPRESSION OF THE REPORTER GENE

To explore the impact of the enhancer position in relation to the minimal promoter in the expression of our reporter gene, we devised a strategy to compare two vector architectures. Accordingly, we used the abovementioned vector with the C3-9 enhancer cloned downstream of the EGFP and we built another version of the vector in which we can insert C3-9 or other candidate enhancers upstream of the minimal promoter. We used these two designs, produced lentivirus, infected MIN-6 cells with an MOI of 1, and assessed enhancer activity, as previously done.

We found that placing the C3-9 enhancer upstream of the synthetical minimal promoter (C3-mP) shows a clear and statistically significant increase in the percentage of GFP-positive cells and in GFP intensity compared to empty vectors (Empty-mP and mP-Empty; Fig. 3C). All control groups, which included non-infected cells, the original vector without any cassette (mP), and the vectors with the cassette placed downstream (mP-Empty) and upstream (Empty-mP) produced less than 1% GFP-positive cells. In contrast, when the C3-9 enhancer is placed upstream of the mP (C3-mP) we quantified approximately 4% of GFP-positive cells (Fig. 3C). On top of this, this vector (C3-mP) also outperformed the mP-C3 vector, in which the C3-9 enhancer is cloned downstream of the EGFP gene, having resulted in less than 1% GFP-positive cells, on average, similar to the controls. Regarding normalized GFP intensity, mP-C3 had no impact on normalized mean GFP intensity, while the C3-mP condition showed a two-fold increase in normalized GFP intensity compared to the control conditions (Fig. 3D).

The C3-9 enhancer alone (C3), without a minimal promoter, induced a small GFP expression increase compared with the controls. In particular, using this vector we obtained slightly more GFP-positive cells when compared with the controls (from 0,2% to 0,5%) and a normalized GFP intensity of approximately 1.2x bigger than controls (Fig. 3C and Fig. 3D). This indicates that the C3-9 enhancer may have a very small, but detectable, activity as a promoter. However, the GFP expression resulting from the C3 enhancer functioning as a promoter is significantly lower than when the sequence is upstream of a minimal promoter (C3-mP). The C3 alone resulted in around 0,5% of GFP-positive cells while C3-mP resulted in approximately 4% (Fig. 3C). Moreover, when compared with cells transduced with the empty vector, the normalized GFP intensity of the C3-mP-transduced cells doubled, while cells transduced with the C3 without a minimal promoter resulted in an increase of just around 20% (Fig. 3D). This distinction allows us to infer that most of the expression in C3-mP arises from the C3-9 enhancer's activity.

Surprisingly, the empty vector GATAmP-Empty resulted in the highest percentage of GFP-positive cells, compared with other conditions containing this promoter. On average, GATAmP-Empty resulted in 5% of GFP-positive cells, while GATA-C3 and C3-GATA resulted in approximately 4% and 2%, respectively (Fig. 3E). Regarding GFP intensity in these positive cells, those containing GATAmP with the C3-9 enhancer either upstream or downstream of the reporter gene (C3-GATAmP and GATAmP-C3) exhibited a slightly greater intensity than

cells with the empty vector (Fig. 3F). This implies that the predominant activity we observed stems from GATAmP itself, rather than significant input from the C3-9 enhancer. In this specific experiment, the GATAmP-C3 condition showed a statistically significant increase in normalized GFP intensity compared to GATAmP-Empty. This contrasted with the earlier experiment, indicating minor technical variations might cause such slight shifts in GFP intensity. Further replicates should be performed to validate these observations.

Although potentially impacted by putative promoter activity, currently, cloning enhancers upstream of the minimal promoter (synthetic mP) appears to be a preferable approach for detecting enhancer activity. Proteins binding to enhancers located upstream of the minimal promoter appear capable of interacting and augmenting the activity of RNA polymerase II. This interaction is likely less pronounced when these enhancers are physically separated from the promoter by the EGFP gene and polyA. However, it's important to note that cloning candidate enhancers directly upstream of minimal promoters might hinder the study of some important aspects impacting enhancer function, particularly for analyzing single nucleotide function in enhancer activity, as this architecture could potentially obscure information about loop formation and consequently long-range enhancer activity.

2.3. VALIDATION OF THE ENHANCER DETECTING SYSTEM BY COMPARISON WITH LUCIFERASE-REPORTER ASSAY

To validate the enhancer activity detection capability of our chosen lentiviral vector (pLS-CRSEL-mP) we compared it to the commonly used enhancer luciferase reporter assay, based on the use of a non-integrative vector. In addition to the C3-9 enhancer, we assessed two other candidate enhancer sequences for the human MODY gene PDX1, which encodes a critical transcription factor in pancreatic development (Ahlgren et al., 1998; Fujimoto & Polonsky, 2009). These sequences, CRS056 (Fig. 4A) and CRS510 (Fig. 4B) were selected based on their epigenetic marks indicating enhancer activity, such as accessible chromatin (ATAC-seq signal) and the histone modification H3K27ac (ChIP-seq signal). We PCR-amplified these sequences, cloned them into our pLS-CRSEL-mP vector (similar to what was done for C3-9 enhancer), produced lentivirus, infected MIN-6 cells, and conducted flow cytometry as previously described. For the luciferase assay, we initially TA-cloned the PCR products into

pCR8/GW/TOPO (Invitrogen, Thermo Fisher Scientific), which was then used for Gateway cloning of the PCR product into the luciferase destination vector, pGL4.23-GW (Pasquali et al., 2014). The success of this process was validated by Sanger sequencing of the final vectors. We then transfected MIN-6 cells with the luciferase vectors containing our sequences and measured the luminescent signal 48 hours later.

The firefly luciferase activity was normalized to the activity of the transfection control, consisting of a second vector coding for an alternative luciferase, the nanoluc, from pNL1.1PGK[Nluc/PGK]. The results were presented as a relative response ratio in which responses in non-transfected cells were normalized to 0 and the positive control (Tk promoter) to 1. For both CRS056 and CRS510, their normalized luciferase activity was close to 0, even lower than the empty luciferase vector. In contrast, the C3-9 enhancer resulted in a value of approximately 1.3. This high value indicates a higher activity than the positive control (Fig. 4D).

These results were corroborated in the lentiviral assay. When these sequences (CRS056 and CRS510) were cloned upstream of the mP (CRS056-mP and CRS510-mP), they resulted in nearly 0% of GFP-positive cells, akin to the empty vector (Empty-mP). Conversely, the C3 enhancer (C3-mP) averaged slightly more than 1.5% GFP-positive cells (Fig. 4E). Examining normalized GFP intensity, both CRS056-mP and CRS510-mP showed values close to 1 a.u., mirroring the control (Empty-mP), while C3-mP averaged around 1.4 a.u. (Fig. 4F).

These findings suggest a correlation between the enhancer activity observed in luciferase-reporter assays and in lentiviral vector assays. Interestingly, the two sequences we selected based on epigenetic marks for enhancer activity showed no enhancer activity in both assays, underscoring the importance of rigorously testing and validating candidate regulatory sequences. Additionally, these results demonstrate that not every sequence inserted upstream of the mP in pLS-CRSEL-mP leads to GFP expression in MIN-6 cells, something we had not previously demonstrated, thus further validating the lentiviral system we established.

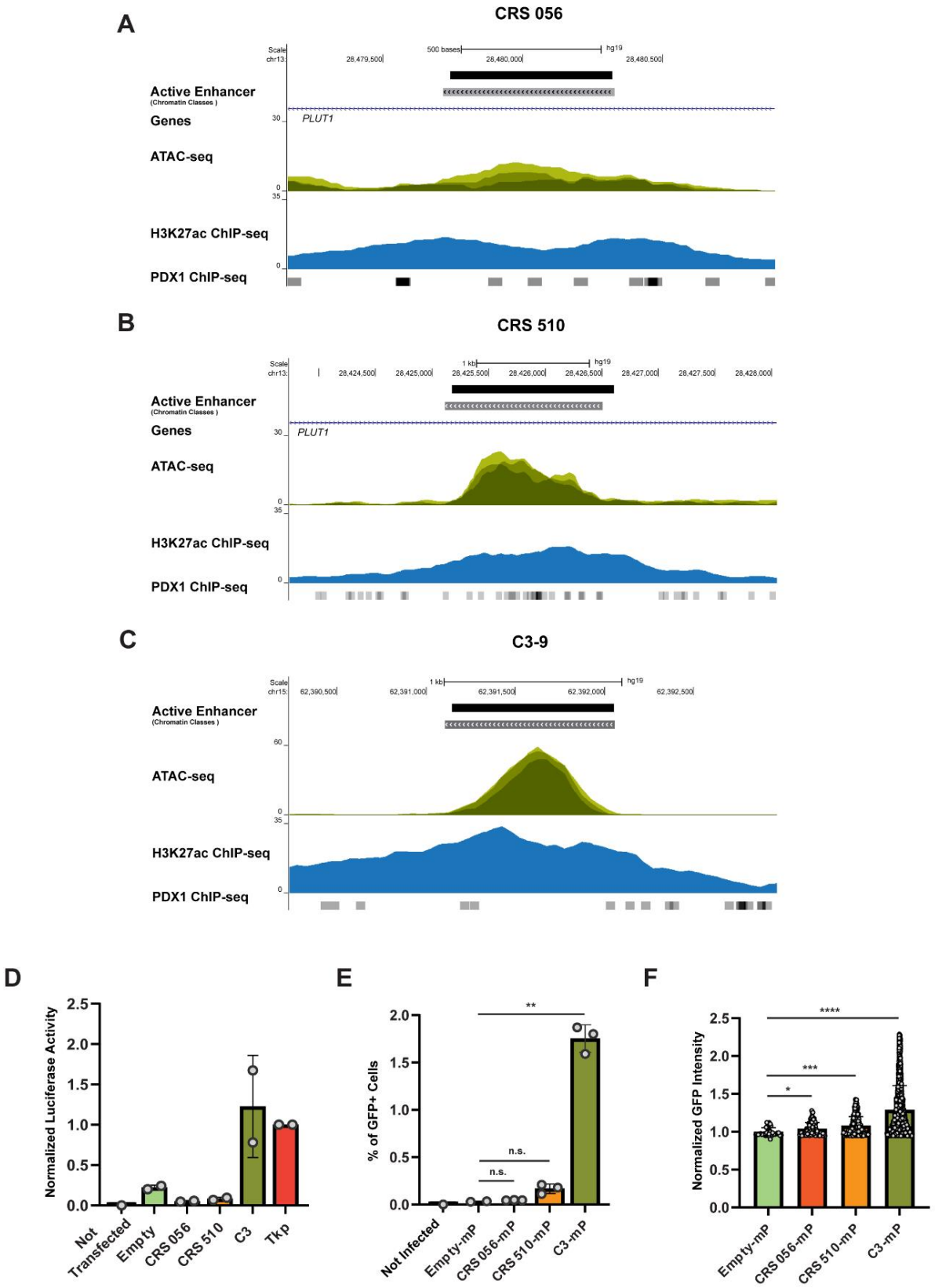


Figure legend in the next page

Figure 4. Comparison of episomal and lentiviral enhancer reporter systems, using enhancer C3-9 and CRSs for the *PDX1* gene. **(A-C)** Genomic landscapes of CRS056 **(A)**, CRS510 **(B)**, and C3-9 **(C)** in the human genome. The black bar denotes the PCR-amplified sequence cloned in our vectors, while the grey bar represents chromatin classes, specifically "active enhancers" defined by Miguel-Escalada and colleagues (Miguel-Escalada et al., 2019). Green indicates ATAC-seq signal (open chromatin), blue shows peaks of histone acetylation (H3K27ac) from ChIP-seq, and bottom grey/black bars represent TFBS for PDX1 as determined by ChIP-seq in pancreatic progenitor cells (Cebola et al., 2015). **(D)** Normalized luciferase activity derived from empty vector, CRS056, CRS510, C3-9, and Tkp (positive control). Results are presented as a relative response ratio, in which responses in non-transfected cells were normalized to 0 and the positive control to 1. **(E)** Average percentage of GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-CRSEL-mP, pLS-CRS056-mP, pLS-CRS510-mP, and pLS-C3-mP. n.s.: not statistically significant, ** p value ≤ 0.01 by non-parametric ANOVA test (Kruskal Wallis test). **(F)** Normalized intensity of GFP from GFP-positive cells from infections (MOI=1) with lentiviruses produced using pLS-CRSEL-mP, pLS-CRS056-mP, pLS-CRS510-mP, and pLS-C3-mP. Normalization was performed by dividing each GFP intensity value by the mean GFP intensity obtained in the GFP-positive cells from the empty vector. * p value ≤ 0.05, *** p value ≤ 0.001, **** p value ≤ 0.0001 from unpaired t-test (non-parametric Kolmogorov test).

3. SELECTION OF CANDIDATE REGULATORY SEQUENCES

Having established a lentiviral enhancer activity detection system, we aimed to scale it up and simultaneously test numerous candidate enhancer sequences with a putative role in T2D. To achieve this, our group integrated T2D-associated SNPs listed in DisGeNET with cis-regulatory classes (mainly active enhancers) identified by Miguel-Escalada and team (Miguel-Escalada et al., 2019). We defined 2kb regions centered on ATAC-seq peak summits and intersected them with the positions of T2D-associated SNPs. The 2kb regions overlapping an SNP were then cropped to a 200bp sequence, again centered in the ATAC-seq peak summits. This process yielded 19 CRSs with an SNP linked to T2D within the 200bp selected region, as illustrated in Fig. 5A, and 335 CRSs with a T2D-associated SNP outside the chosen 200bp region but within the 2kb selected region, as demonstrated in Fig. 5B. While SNPs might not align precisely with the predicted location of candidate enhancers, their proximity supports the significance of regulatory elements in these loci to T2D development. Alongside these 354 CRSs, we added 11 CRSs near crucial beta-cell function genes, 14 CRSs with T2D-associated SNPs previously validated in zebrafish embryos by our group (Eufrásio et al., 2020), and 6 control sequences

(see Materials and Methods for further details), culminating in a library of 386 sequences, as depicted in Fig. 5C.

The pool of candidate enhancers was synthesized as a pool of oligonucleotides, PCR amplified, and cloned into the pLS-CRSEL-mP and pLS-GATAmP-CRSEL vectors through Gibson assembly, originating the “Pool-mP” and “GATAmP-Pool” plasmid pools, respectively. Cloning of these pools was validated by colony PCR, which allowed the analysis of the frequency of correct cloning of a CRS in the final vector (Fig. 5D). When the frequency of the empty vector was below 10%, all colonies were polled together and the vectors carrying all different CRSs were extracted. These generated plasmid libraries are later used to generate lentivirus libraries and perform parallel enhancer reporter assays.

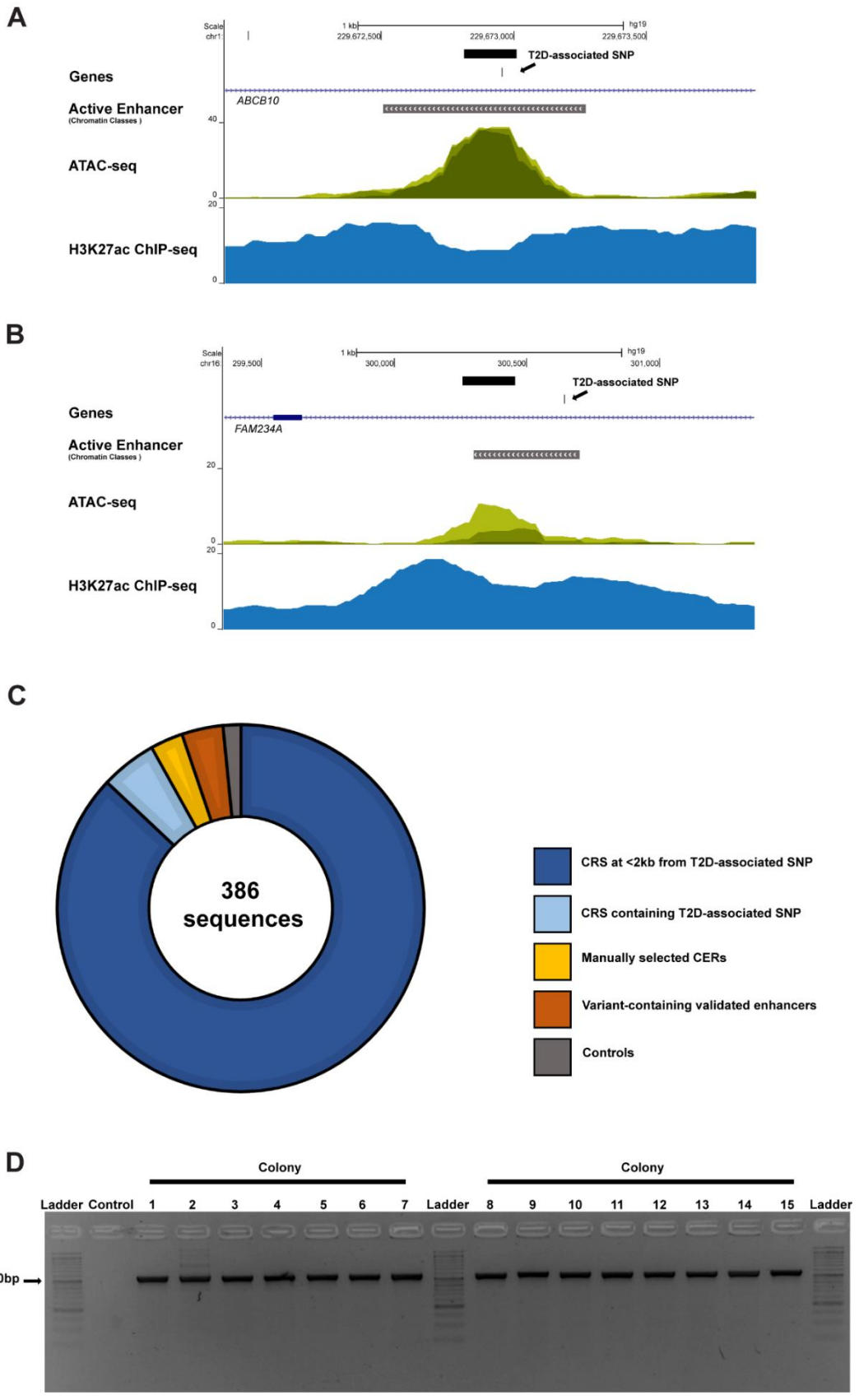


Figure legend in the next page

Figure 5. Assembly of CRSs lentiviral libraries. Representative landscapes of candidate enhancer regions containing an SNP in the selected 200bp sequence **(A)** or outside the selected 200bp sequence but within the 2kb selected region **(B)**. The black bar denotes the PCR-amplified sequence cloned in our vectors, while the grey bar represents chromatin classes, specifically "active enhancers" defined by Miguel-Escalada and colleagues (Miguel-Escalada et al., 2019). Green indicates ATAC-seq signal (open chromatin), blue shows peaks of histone acetylation (H3K27ac) from ChIP-seq. **(C)** Pie chart showing the composition of the lentiMPRA library. 386 CRSs in total in which 335 were CRSs with T2D-associated SNPs outside the tested 200bp sequence but within the 2kb selected region (dark blue), 19 CRSs with T2D-associated SNPs in the tested 200bp sequence (light blue), 11 manually selected CRSs located nearby important genes for beta-cell function (yellow), 14 CRSs with T2D-associated SNPs previously validated (orange) and an additional 7 control sequences (grey). **(D)** Agarose gel demonstrating the PCR product amplified through colony PCR. The primers used allowed the amplification of a ≈500bp sequence containing the selected 200bp CRSs.

4. PARALLEL REPORTER ASSAYS

To determine the enhancer activity of the cloned CRSs, we performed parallel reporter assays. Using the Pool-mP and GATAmP-Pool libraries, we generated lentiviruses and infected MIN-6 cells, following the procedures mentioned before, measuring enhancer activity through flow cytometry (Fig. A-H). Since we previously validated our vectors with a small number of candidate enhancers, which had more than 200bp (see chapter 2.3), it remained uncertain whether our systems would respond to our CRSs library.

The transduction with the empty vector (Empty-mP), as shown in previous experiments, consistently resulted in less than 1% of GFP-positive cells, while the vector with the CRSs library cloned upstream of this promoter (Pool-mP) averaged 10% GFP-positive cells (Fig. 6I). These cytometry results demonstrate that the CRSs library included sequences capable of enhancing GFP expression when cloned upstream Promega's synthetical mP. Moreover, since the empty vector (Empty-mP) consistently yielded a very low percentage of fluorescent cells - less than 1% - we can conclude that most cells infected with the lentiviral enhancer library (Pool-mP) showing a positive GFP signal likely contained an enhancer sequence that drives GFP expression and is not expressing GFP due to enhancer trapping (discussed below) or other technical issues. Results also show that GFP-positive cells from the Pool-mP exhibited a normalized GFP intensity of 1.2 a.u., a statistically significant increase from the few GFP-

positive cells obtained with the Empty-mP construct (Fig. 6J and Fig. 6K), and consistently across multiple replicates (Fig. 6K).

In contrast, the GATAmP-Pool library resulted in the same percentage of GFP-positive cells (≈7%) as the respective control (GATAmP-Empty) (Fig. 6L). Since the empty vector consistently produced similar percentages of GFP-positive cells, detecting sequences with enhancer activity using this vector would be challenging. These data corroborate previously mentioned data (see chapter 2.2), as GATAmP seems to be unable to consistently and/or efficiently respond to enhancer sequences placed downstream of the expression cassette. Moreover, some replicates showed an increase in normalized GFP intensity while others displayed a decrease, reaffirming the inconsistency previously observed with this construct. On average, across the replicates, the normalized GFP intensity equaled 1 a.u. (Fig. 6M and Fig. 6N).

Nevertheless, we hypothesize that enrichment of specific enhancer sequences could be found among fluorescent cells when compared to non-fluorescent cells. Therefore, we collected GFP-positive, GFP-negative, and non-sorted cells from both Pool-mP and GATAmP-Pool infected cell samples for subsequent next-generation sequencing and analysis. These data enabled us to confirm the presence of our sequences in non-sorted cells and to distinguish which sequences were more prevalent in the GFP-negative and GFP-positive populations. This differentiation would allow the identification of sequences with enhancer activity. Additionally, if enrichment of sequences in the GFP-positive cell population derived from the GATAmP-Pool vector-transduced cells was observed, it would be interesting to compare them to the sequences identified as enhancers by Pool-mP. This would allow the discovery of sequences able to enhance GFP expression specifically in either one, or both, cassette configurations, allowing further analysis of the regarding nucleotides and TFBSs necessary for loop formation or promoter specificity.

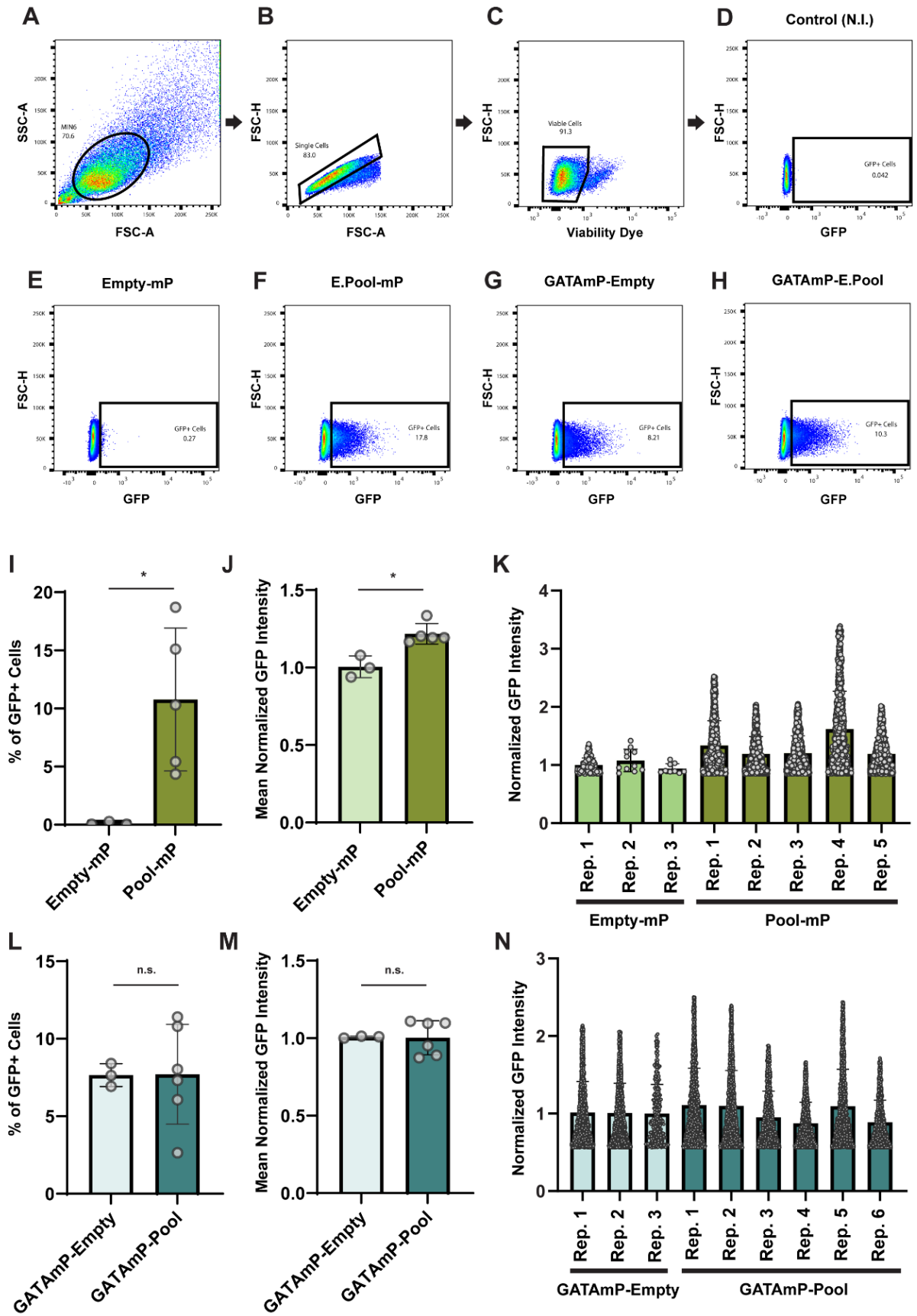


Figure legend in the next page

Figure 6. Flow cytometry analysis of parallel reporter assays. For this analysis, MIN-6 cells were selected using an FSC-A versus SSC-A plot **(A)** prior to single cell (singlets) selection performed on an FSC-A versus FSC-H plot **(B)**. From the singlets, viable cells were selected using a far-red (viability dye) versus FSC-H plot where cells marked with this dye were excluded **(C)**. The GFP-positive gate was defined on a GFP versus FSC-H plot based on viable cells from non-infected MIN-6 cells **(D)**. Representative plots of GFP versus FSC-H resulted from infections (MOI≈0.3) with lentiviruses produced using the Empty-mP **(E)**, Pool-mP **(F)**, GATAmP-Empty **(G)**, and GATAmP-Pool **(H)** vectors. Percentage of GFP-positive cells from infections (MOI≈0.3) using the Empty-mP and Pool-mP vectors **(I)** and the GATAmP-Empty and GATAmP-Pool vectors **(L)**. n.s.: not statistically significant, * p value≤0.05 from unpaired t-test (non-parametric Mann-Whitney test). Mean normalized GFP intensity **(J)** and normalized GFP intensity **(K)** from infections (MOI≈0.3) with lentiviruses produced using the Empty-mP and Pool-mP constructs (n=1000 for Pool-mP and n≈250 for Empty-mP). Mean normalized GFP intensity **(M)** and normalized GFP intensity **(N)** from infections (MOI≈0.3) with lentiviruses produced using the GATAmP-Empty and GATAmP-Pool constructs (n=4000). Normalizations were performed by dividing each value of GFP intensity by the mean GFP intensity obtained in the GFP-positive cells from their respective empty vector. n.s.: not statistically significant, * p value≤0.05 from unpaired t-test (non-parametric- Kolmogorov test).

Besides analysis of GFP expression and intensity of the different conditions, cells were also sorted and collected based on their GFP intensity. Their genomic DNA was extracted and the CRS-containing cassette was PCR-amplified, generating an amplicon library that our group sent for next-generation sequencing. From our next-generation sequencing data, we derived preliminary insights by plotting normalized read counts from three technical replicates for a particular CRS in the GFP-negative cell population (x-axis) against those from the GFP-positive cell population (y-axis). This simplistic approach enabled us to visualize the distribution of our sequences based on their enrichment in GFP-negative and GFP-positive cells. CRSs plotted above the black line (y=x) present a higher likelihood of enhancer activity, while those plotted below the line have a lower probability of being an active enhancer.

The results showed that the majority of CRSs in our library were plotted on or above the line, indicating enhancer activity. The red dots represented CRSs with a statistically significant difference in the number of normalized reads between the two populations (p-value <0.1). From these, we highlighted 4 CRSs (Fig. 7A). CRS7, depicted by a dot below the line indicating no enhancer activity, displayed epigenetic marks associated with enhancer activity (Fig. 7B) but it was classified by Miguel-Escalada and colleagues as "Strong CTCF" due to a high CTCF ChIP-seq signal, suggesting abundant CTCF binding in this region. Given that CTCF is a crucial

protein binding to insulator elements in vertebrates (Furlong & Levine, 2018), this sequence might act as an insulator, explaining its lack of enhancer activity in our dataset.

Notably, CRS283 and CRS284, representing two sequences lacking enhancer activity, are part of the same 2kb enhancer region. Eufrásio et al. illustrated that this region is functionally divided into distinct domains required for enhancer function. While the ≈2kb sequence displayed enhancer activity in the pancreatic domain of zebrafish embryos, its fragmentation into two parts overlapping with our CRSs resulted in regions that exhibited significantly reduced or no enhancer activity. In our assay, CRS283 and CRS284, both 200bp fragments from the same region, did not display enhancer activity. Thus, it would be interesting to test the complete enhancer region and the combined presence of both CRSs in our specific system. This would allow the comparison of the enhancer activity of these regions between the endocrine pancreas of zebrafish, previously published, and in MIN-6 cells. These results highlight limitations in current MPRAs, including length constraints on tested sequences and the dissociation of the tested sequence from its genomic context, potentially resulting in the loss of important DNA regions necessary for proper enhancer function.

Lastly, CRS217, one of the CRSs exhibiting higher enhancer activity in this analysis, is a previously validated enhancer in MIN6 cells, and a single nucleotide modification has been demonstrated to influence its activity (Pasquali et al., 2014).

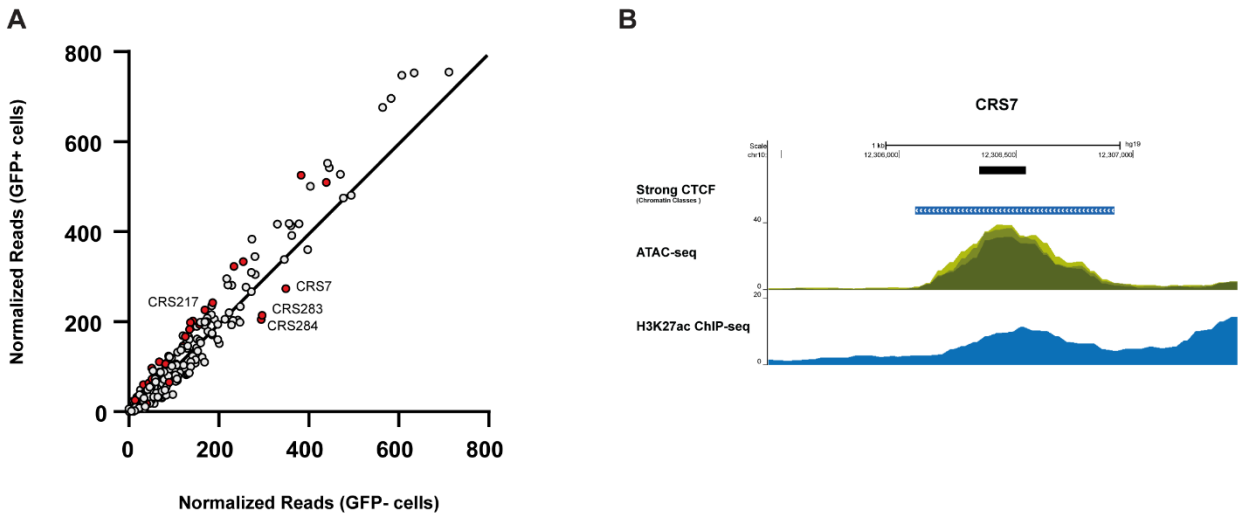


Figure 7. Preliminary analysis of NGS results from parallel reporter assays. (A) Normalized number of reads from three technical replicates of a given candidate regulatory sequence present in GFP-positive cells (y-axis) plotted against the normalized number of reads of the same candidate regulatory sequence present in GFP-negative cells (x-axis). Sequences with a p-value<0.1 were considered significantly enriched in one of the populations (red dots). **(B)** Genomic landscape of CRS7 in the human genome. The black bar denotes the PCR-amplified and cloned sequence in our vectors, while the blue bar represents chromatin classes, specifically "strong CTCF" defined by Miguel-Escalada and colleagues (Miguel-Escalada et al., 2019). Green indicates ATAC-seq signal (open chromatin), blue indicates ChIP-seq signal of histone acetylation (H3K27ac).

5. VALIDATION OF ENHANCERS USING REPORTER LINES

Besides the collection of pools of cells for next-generation sequencing, GFP-positive cells were also single-cell sorted and seeded in 96-well plates to establish enhancer reporter lines. Once GFP-positive enhancer reporter lines are established, validation becomes crucial to confirm that CRSs are driving GFP expression in those cells, rather than neighboring cis-regulatory elements near the integration site of the lentiviral cassette. This phenomenon is a type of position effect, termed enhancer trapping, in which the transcription levels of a given reporter gene depend on its genomic integration site, particularly in response to nearby active enhancers (Lewis, 1950.; Wilson et al., 1990).

Just as we questioned whether GATAmP actually functions as a minimal promoter due to a high background signal from the empty vector, we also speculate that this promoter, which is

significantly larger (1040bp) than Promega's minimal promoter (32bp), might excel at detecting nearby genomic landscapes, potentially reading multiple enhancers near the cassette integration site. A way to test this hypothesis would involve transfecting the empty vector with GATAmP as a non-integrating lentiviral vector (episome) to determine if the same elevated basal GFP expression is achieved. If not, it would suggest that the integrated cassette's high signal results from the effects of nearby enhancers. Although this vector's potential for other applications is intriguing, our primary goal is to accurately read the enhancer sequences that we introduce. To counteract this "position effect", our constructs are designed with antirepressors bordering the integrated reporter cassette, as mentioned earlier. It's worth considering that these antirepressor sequences may not effectively inhibit the interaction of larger promoters like GATAmP (1040bp), FOSmP (95bp), and HSP70mP (133bp) with nearby cis-regulatory elements. This is suggested by the high EGFP expression observed in empty vectors containing these promoters compared to mP (Promega), as demonstrated previously.

Our enhancer sequences are flanked in our vector by two loxP sites. When treated with Cre recombinase, this enzyme triggers the recombination of these loxP sites, leading to the excision of the enhancer sequence while preserving other elements of the lentiviral cassette, such as the minimal promoter and the EGFP gene. Thus, a reduction in EGFP expression in cells treated with Cre would indicate that the enhancer was the main responsible for driving the GFP expression.

We introduced the Cre recombinase protein into our cells through the transduction of a lentiviral vector that carries the Cre recombinase linked to mCherry via a P2A self-cleaving peptide (pL-mCherry-Cre, Fig. 8A). This arrangement allows us to identify the infected cells by their expression of mCherry. To set our experimental pipeline, we first established a Cre-reporter cell line and we used it to test the pL-mCherry-Cre vector. We accomplished this by initially transducing MIN-6 cells with the Cre-reporter vector that contained a mKate gene and a stop codon flanked by loxP sites, as well as a downstream GFP gene (fig. 8B). Then, to establish this Cre-reporter cell line, we sorted and collected cells expressing mKate by FACS. Upon treatment with the Cre recombinase, the mKate gene and stop codon are excised, causing the cells to start expressing GFP.

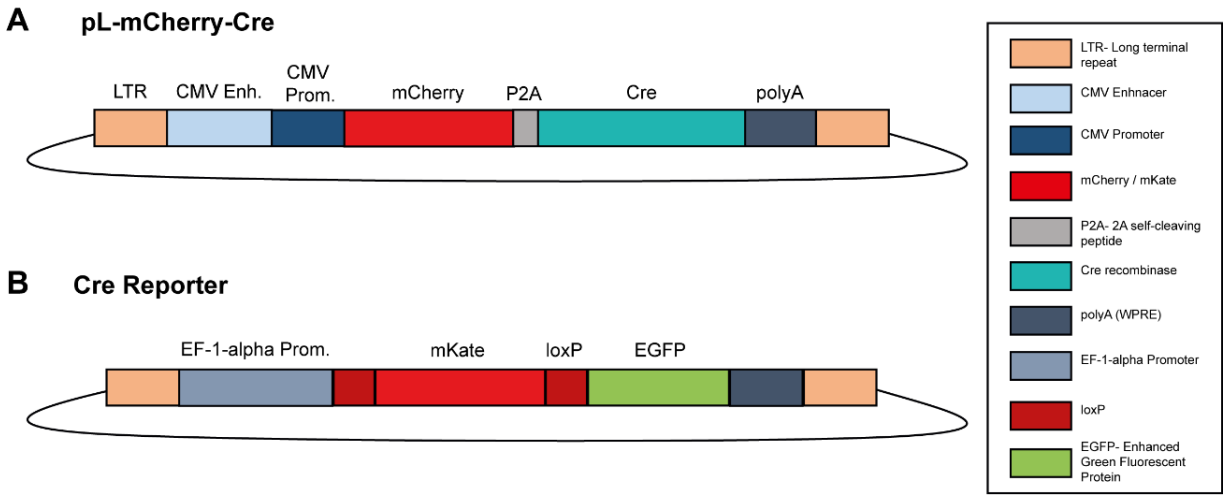


Figure 8. Schematics of the mCherry-Cre and Cre-Reporter vectors. (A)- pL-mCherry-Cre vector containing the Cre and mCherry genes linked with a P2A. **(B)-** Cre Reporter vector containing the mKate gene and a stop codon flanked by two loxP sites followed by the EGFP gene downstream.

Our observations indicated that, under control conditions, this reporter line exhibited solely red fluorescence (Fig. 9A). When these cells were subsequently transduced with the mCherry-Cre vector, we noticed some cells expressing GFP (Fig. 9B), indicating the functional activity of Cre-recombinase. It's worth noting that not every cell in this line expressed red fluorescence, as shown in Fig. 9A and 9B. Although these cells were initially sorted based on mKate positivity, further sorting would have been necessary to further enrich the population carrying the Cre-reporter cassette. An additional challenge in this experiment was the indistinguishability of mKate and mCherry fluorescence, preventing us from confirming the transduction with the Cre-mCherry vector even in the absence of GFP expression.

With the optimized mCherry-Cre virus, we proceeded to test it on pools of GFP-positive cells. To achieve this, we first transduced cells with Pool-mP and GATAmP-Pool constructs, as we did in previous experiments, FACS-sorted the resulting GFP-positive cells, and then seeded these sorted cells as polyclonal populations instead of single-cells. Subsequently, we transduced these populations of GFP-positive cells with the Cre recombinase (Fig. 9C-9J).

Interestingly, even though we had initially sorted these cells as GFP-positive, both the Pool-mP and GATAmP-Pool cell populations exhibited a small subpopulation of GFP-negative cells in the untreated controls (Fig. 9G and 9I). This occurrence might be attributed to some of the inserted sequences being epigenetically repressed, such as through DNA methylation, which would inhibit enhancer activity by preventing the binding of TFs to the DNA, leading to a loss of GFP expression over time (Héberlé & Bardet, 2019).

In this experiment, we wanted to guarantee that as many cells as possible were infected with Cre recombinase, and therefore a MOI>1 was used. We achieved around 25% infection rate with the mCherry-Cre construct. Within these infected cells (mCherry-positive cells), a noticeable increase in the percentage of GFP-negative cells was observed for the Pool-mP condition (57% GFP-negative cells among mCherry-positive cells) compared to the non-infected control (19% GFP-negative cells among mCherry-negative cells) (Fig. 9G, H). As anticipated, the GFP-positive to GFP-negative cell ratio in the mCherry-negative populations (not transduced with the mCherry-Cre vector) remained consistent in both conditions (Pool-mP control: 4,26; Pool-mP Cre treated: 3). In contrast, in the mCherry-positive population, that was successfully transduced with the Cre and potentially lost the enhancers, the GFP-positive to GFP-negative cell ratio is 0,75 (Fig. 9G, H). Since more than half of the mCherry-positive population is made up of GFP-negative cells (57%), and taking into consideration that only a small percentage of these cells was GFP-negative in the population not infected with mCherry-Cre, we can conclude that almost half of our cells respond to the Cre recombinase, indicating the presence of an enhancer accountable for GFP expression.

When examining the GATAmP-Pool populations, the percentage of GFP-positive cells within the mCherry-positive population remains similar to the percentage of GFP-positive cells detected in the mCherry-negative populations. This indicates no clear loss of GFP expression in the population of cells transduced with the Cre recombinase. In non-infected cells, we have roughly 8% GFP-negative cells and 92% GFP-positive cells. However, when these cells are treated with Cre (mCherry-positive population), we observe approximately 12% (2.9% out of 24.4%) GFP-negative cells and 88% (21.5% out of 24.4%) GFP-positive cells (Fig. 9I-, J). This suggests that the majority of, but not all, cells in this condition do not respond to the Cre recombinase, indicating that the GFP expression in these cells is, in most cells, is not driven by the candidate enhancers but is likely the result of position effect.

Nonetheless, we managed to FACS-sort cells and establish reporter lines that respond to Cre-recombinase from the GATAmP-C3 infections previously done, suggesting the activity of an enhancer downstream of the EGFP gene (Fig. 9K, L). The figures demonstrate that these 10 Cre-treated reporter lines (R.L.1 to 10) displayed lower mean GFP intensity compared to their respective controls, confirming that although the majority of the GFP expression is driven by the promoter, we can detect some activity from an enhancer downstream of the promoter and reporter gene. Moreover, for this specific experiment, the impact of enhancer excision could have been underestimated since we did not control the infection success through the detection of mCherry expression, so at this point, one cannot exclude that some of the cells retaining GFP expression were not transduced with the mCherry-Cre cassette.

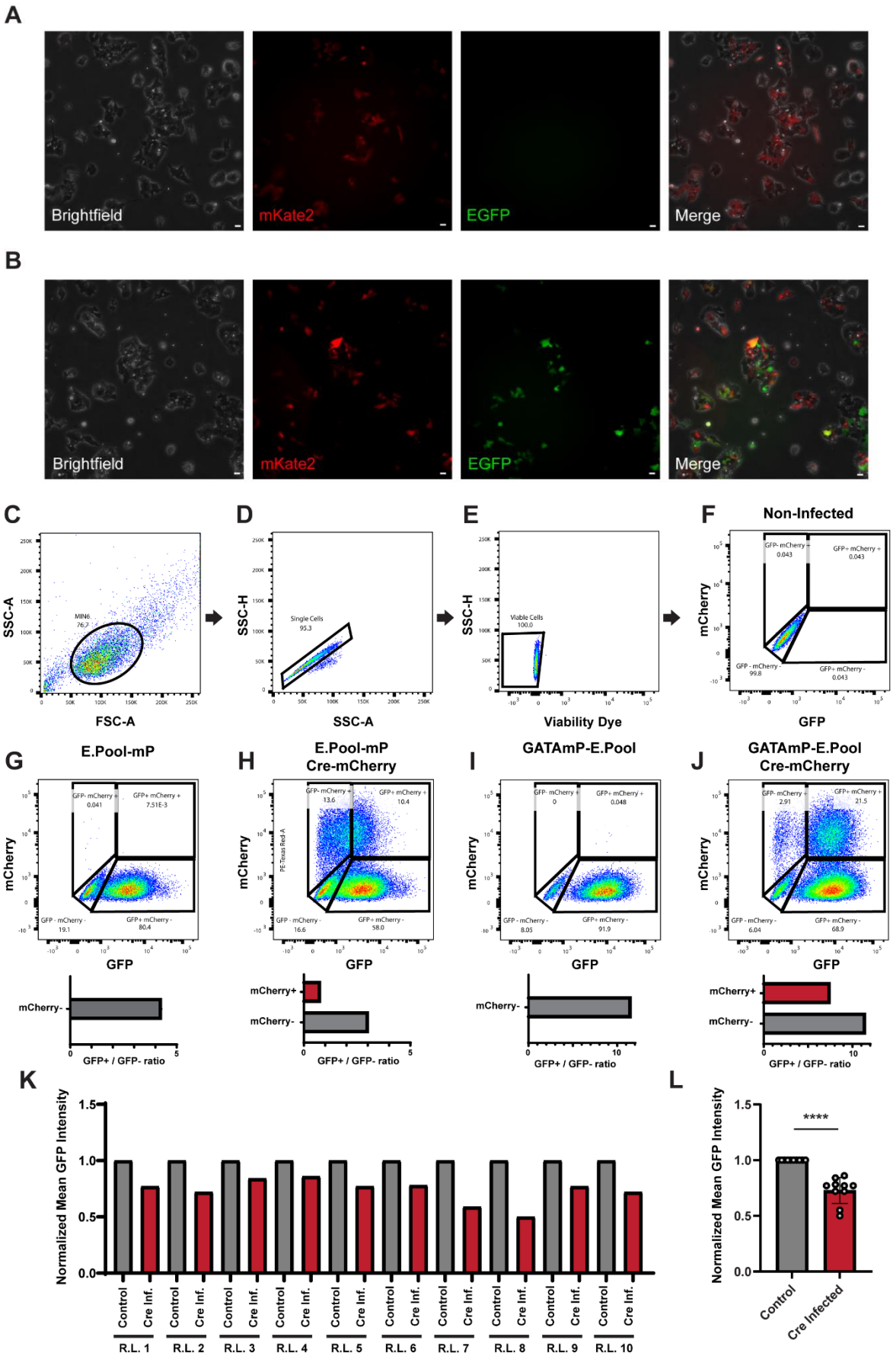


Figure legend in the next page

Figure 9. Cre-mediated enhancer validations. **(A, B)** Representative fluorescence microscopy images of the Cre Reporter cell line not infected **(A)** and infected with lentiviruses produced using the Cre-mCherry vector **(B)**. The top images show cells only expressing mKate, a red fluorescent protein, derived from the Cre Reporter cassette, while the bottom images also show cells expressing GFP, a result of the Cre-mediated excision of the mKate gene and stop codon. Pictures were taken with a 40X objective. Scale: 10µm. **(C-J)** For flow cytometry analysis, MIN-6 cells were selected using an FSC-A versus SSC-A plot **(C)** prior to single cell (singlets) selection performed on an SSC-A versus SSC-H plot **(D)**. From the singlets, viable cells were selected using a far-red (viability dye) versus SSC-H plot where cells marked with this dye were excluded **(E)**. The GFP-positive/negative and mCherry-positive/negative gates were selected on a green (GFP) versus red (mCherry) plots using viable cells from non-infected MIN-6 cells **(F)**. Plots of GFP versus mCherry from sorted GFP-positive cells transduced with the Pool-mP cassette **(G)**, from sorted GPF-positive cells transduced with the Pool-mP vector and subsequently infected with lentiviruses produced using the Cre-mCherry vector **(H)**, from sorted GFP-positive cells transduced with the GATAmP-Pool cassette **(I)** and from sorted GPF-positive cells transduced with the GATAmP-Pool cassette and subsequently infected with lentiviruses produced using the Cre-mCherry vector **(J)**. The bars below represent the ratio between GFP-positive cells and GFP-negative cells in the mCherry-negative population (grey bar) and in the mCherry-positive population (red bar). **(K, L)** Normalized GFP intensity of 10 enhancer reporter lines established from single GFP-positive cells transduced with the GATAmP-C3 construct and infected (red bar) and not infected (grey bar) with lentiviruses produced using the Cre-mCherry vector.

CHAPTER 5:
CONCLUSIONS AND FUTURE PERSPECTIVES

In this study, we successfully developed and assessed lentiviral-based enhancer detection vectors for Massively Parallel Reporter Assays. Our goal was to validate predicted enhancers linked to T2D on a large scale. This validation would allow us to explore the functionality and influence of these sequences, especially in the context of pancreatic β-cell function, which is pivotal to understanding diabetes etiology.

We first examined the impact of vector architecture, demonstrating that, in our current system, positioning a known enhancer (C3-9) downstream of a reporter gene was ineffective at detecting enhancer activity. Despite this, we were able to establish reporter lines with the GATAmP-C3 construct, in which GFP expression decreased upon removal of the C3 enhancer using Cre recombinase. This reduction indicates that GFP expression translated enhancer activity, suggesting that the system works but emphasizing the need for optimizations in this construct. Conversely, placing the enhancer upstream of the synthetic minimal promoter led to a notable increase in GFP-positive cells and normalized GFP intensity, which indicates this is a more effective approach for detecting enhancer activity.

However, there is a common challenge with this standard approach – placing enhancers upstream of minimal promoters can inadvertently result in the detection of promoter activity rather than enhancer activity, potentially yielding misleading results. To mitigate this, we are developing a new vector enabling the cloning of candidate enhancer sequences in both orientations (negative and positive strands). This approach is rooted in the understanding that most enhancers function regardless of their orientation, unlike promoters.

MPRAs are often incapable of fully confirming enhancer sequences that meet the original Banerji definition – active in both orientations and from multiple positions in relation to the promoter (Banerji et al., 1981). Developing vectors capable of this would enhance the representation of natural genomic conditions, bolstering result reliability. Additionally, we attempted to address issues concerning promoter-enhancer specificity in conventional enhancer assays. However, our assays with different promoters in vectors with the enhancers cloned downstream of the reporter gene were mostly inconclusive, because of the inability to fully distinguish C3-9 enhancer activity from promoter activity alone. Hence, few conclusions can be drawn regarding the C3 enhancer's preference for a specific promoter before further optimizations and validations are developed.

We then compiled a diverse candidate enhancer library, integrating T2D-associated SNPs and cis-regulatory elements. Through next-generation sequencing, we successfully identified multiple sequences demonstrating enhancer activity, confirming the effectiveness of our system and our library. It would be interesting in the future to evaluate this candidate enhancer library in human beta-cell lines, such as EndoC, and muscle cell lines. This would demonstrate how the results in mouse and human cells correlate, and offer a more comprehensive understanding of T2D-associated cis-regulatory elements in other cellular contexts whose potential dysregulation impacts disease progression. Given T2D's association with insulin resistance and beta-cell dysfunction, identifying genes regulated by the enhancers validated in those different cell types could help determine their relation to dysregulation of either the pancreas or muscle.

Moreover, interesting enhancer sequences validated in our system could be tested in vivo using models like zebrafish embryos to understand the cell-type specificity of their activity. Additionally, conducting deletions of these selected sequences in human cell lines could shed light on their necessity for normal expression of target genes.

Validating predicted enhancers associated with beta-cell function is crucial to confirm their relevance and functionality in T2D development. Our lab is particularly driven by the fascinating objective of unraveling the nucleotide code within these non-coding cis-regulatory regions, aiming to decode fundamental rules governing transcription factor binding and gene expression regulation. As emphasized earlier, a significant portion of T2D-associated sequence variants resides in these non-coding cis-regulatory elements. Some of these variants disrupt gene expression, heightening T2D susceptibility. However, understanding why certain nucleotide changes disrupt these regions, while others do not, remains a critical gap in the scientific knowledge. The next phase of this project involves mutagenesis screens in validated enhancers to gauge the impact of mutations on enhancer function, thereby establishing rules for disruption of transcription factor binding sites and its effects on specific enhancer sequences. To facilitate this, we will continue to grow individual enhancer reporter lines and validate them for the required assays.

However, conducting empirical experiments to gather evidence is resource-intensive, time-consuming, and often hinges on predefined hypotheses or predictions. Therefore, leveraging our results in predictive models such as machine learning holds immense potential for

obtaining quicker and cost-effective results. The insights gained from this approach can significantly enhance our collective understanding of T2D and serve as a valuable resource for fellow researchers investigating the genetic mechanisms underpinning this disease.

REFERENCES

Abdellaoui, A., Yengo, L., Verweij, K. J., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*. Cell Press. https://doi.org/10.1016/j.ajhg.2022.12.011

Ahlgren, U., Jonsson, J., Jonsson, L., Simu, K., & Edlund, H. (1998). β-Cell-specific inactivation of the mouseIpf1/Pdx1 gene results in loss of the β-cell phenotype and maturity onset diabetes. *Genes & development*, *12*(12), 1763-1768. https://doi.org/10.1101/gad.12.12.1763

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4), 197-212. https://doi.org/10.1038/nrg3891

Ali, T., Renkawitz, R., & Bartkuhn, M. (2016). Insulators and domains of gene expression. *Current opinion in genetics & development*, 37, 17-26. https://doi.org/10.1016/j.gde.2015.11.009

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56. https://doi.org/10.1038/nature11632

Andersen, D. K., Korc, M., Petersen, G. M., Eibl, G., Li, D., Rickels, M. R., Chari, S. T., & Abbruzzese, J. L. (2017). Diabetes, pancreatogenic diabetes, and pancreatic cancer. *Diabetes*, *66*(5), 1103–1110. https://doi.org/10.2337/db16-1477

Antosik, K., & Borowiec, M. (2016). Genetic factors of diabetes. *Archivum Immunologiae et Therapiae Experimentalis, 64*, 157-160 https://doi.org/10.1007/s00005-016-0432-8

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, *339*(6123), 1074-1077. https://doi.org/10.1126/science.1232542

Asfari, M., Janjic, D., Meda, P., Li, G., Halban, P. A., Wollheim, C. B., De Reeherche, L., & Jeantet, L. (1992). Establishment of 2-mercaptoethanol-dependent differentiated insulin-secreting cell lines. *Endocrinology*, *130*(1), 167-178. https://doi.org/ 10.1210/endo.130.1.1370150

Bakhti, M., Böttcher, A., & Lickert, H. (2019). Modelling the endocrine pancreas in health and disease. *Nature Reviews Endocrinology*, *15*(3), 155-171. https://doi.org/10.1038/s41574-018-0132-z

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

Barolo, S., Carver, L. A., & Posakony, J. W. (2000). GFP and β-galactosidase transformation vectors for promoter/enhancer analysis in Drosophila. *Biotechniques*, *29*(4), 726-732. https://doi.org/10.2144/00294bm10

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

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

Battaglia, M. (2014). Experiments by nature: lessons on type 1 diabetes. *Tissue Antigens*, *83*(1), 1-9. https://doi.org/10.1111/tan.12280

Bergman, D. T., Jones, T. R., Liu, V., Ray, J., Jagoda, E., Siraj, L., ... & Engreitz, J. M. (2022). Compatibility rules of human enhancer and promoter sequences. *Nature*, *607*(7917), 176-184. https://doi.org/10.1038/s41586-022-04877-w

Bessa, J., Luengo, M., Rivero-Gil, S., Ariza-Cosano, A., Maia, A. H., Ruiz-Ruano, F. J., ... & Gómez-Skarmeta, J. L. (2014). A mobile insulator system to detect and disrupt cis-regulatory landscapes in vertebrates. *Genome Research*, *24*(3), 487-495. https://doi.org/10.1101/gr.165654.113

Bessa, J., Tena, J. J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., ... & Gómez-Skarmeta, J. L. (2009). Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Developmental dynamics: an official publication of the American Association of Anatomists*, *238*(9), 2409-2417. https://doi.org/10.1002/dvdy.22051

Bich, L., Mossio, M., & Soto, A. M. (2020). Glycemia regulation: from feedback loops to organizational closure. *Frontiers in Physiology*, *11*, 69. https://doi.org/10.3389/fphys.2020.00069

Bordeira-Carriço, R., Teixeira, J., Duque, M., Galhardo, M., Ribeiro, D., Acemel, R. D., ... & 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

Breathnach, R., & Chambon, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. *Annual review of biochemistry*, *50*(1), 349-383. https://doi.org/10.1146/annurev.bi.50.070181.002025

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, *109*(1), 21-29. https://doi.org/10.1002/0471142727.mb2129s109

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

Cheng, K., Delghingaro-Augusto, V., Nolan, C. J., Turner, N., Hallahan, N., Andrikopoulos, S., & Gunton, J. E. (2012). High passage MIN6 cells have impaired insulin secretion with impaired glucose and lipid oxidation. *PloS one*, *7*(7), e40868. https://doi.org/10.1371/journal.pone.0040868

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

Dalfrà, M. G., Burlina, S., Del Vescovo, G. G., & Lapolla, A. (2020). Genetics and epigenetics: new insight on gestational diabetes mellitus. *Frontiers in endocrinology*, *11*, 602477. https://doi.org/10.3389/fendo.2020.602477

Dao, L. T., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., ... & Spicuglia, S. (2017). Genome-wide characterization of mammalian

promoters with distal enhancer functions. *Nature genetics*, *49*(7), 1073-1081. https://doi.org/10.1038/ng.3884

Dekker, J., & Heard, E. (2015). Structural and functional diversity of topologically associating domains. *FEBS letters*, *589*(20), 2877-2884. https://doi.org/10.1016/j.febslet.2015.08.044

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *science*, *295*(5558), 1306-1311. Htpps://doi.org/ 10.1126/science.1067799

Ding, M., Liu, Y., Liao, X., Zhan, H., Liu, Y., & Huang, W. (2018). Enhancer RNAs (eRNAs): new insights into gene transcription and disease treatment. *Journal of Cancer*, *9*(13), 2334. https://doi.org/10.7150/jca.25829

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., ... & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376-380. https://doi.org/10.1038/nature11082

Dorman, E. R., Bushey, A. M., & Corces, V. G. (2007, October). The role of insulator elements in large-scale chromatin structure in interphase. *In Seminars in cell & developmental biology*, *18*(5), 682-690. https://doi.org/ 10.1016/j.semcdb.2007.08.009

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., ... & Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, *16*(10), 1299-1309. https://doi.org/10.1101/gr.5571506

Drzazga, A., Cichońska, E., Koziołkiewicz, M., & Gendaszewska-Darmach, E. (2020). Formation of βTC3 and MIN6 Pseudoislets Changes the Expression Pattern of Gpr40, Gpr55, and Gpr119 receptors and improves lysophosphatidylcholines-potentiated glucose-stimulated insulin secretion. *Cells, 9*(9), 2062. https://doi.org/10.3390/cells9092062

Dziewulska, A., Dobosz, A. M., & Dobrzyn, A. (2018). High-throughput approaches onto uncover (Epi) genomic architecture of type 2 diabetes. *Genes*, *9*(8), 374. https://doi.org/10.3390/genes9080374

Ellard, S., Bellanné-Chantelot, C., Hattersley, A. T., Carette, C., Castano Gonzalez, L., De Nanclares Leal, G., Elles, R., Gaspar, G., Gasperikova, D., Hansen, T., Herr, M., Kamarainen, O., Kannengiesser, C., Klimes, I., Lacape, G., Losekoot, M., Malecki, M., Meyer, P., Njolstad, P., … Wuyts, W. (2008). Best practice guidelines for the molecular genetic diagnosis of maturity-onset diabetes of the young. *Diabetologia*, *51*(4), 546–553. https://doi.org/10.1007/s00125-008-0942-y

Elliott, H. R., Sharp, G. C., Relton, C. L., & Lawlor, D. A. (2019). Epigenetics and gestational diabetes: a review of epigenetic epidemiology studies and their use to explore epigenetic mediation and improve prediction. *Diabetologia, 62*(12), 2171–2178. https://doi.org/10.1007/s00125-019-05011-8

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, *539*(7629), 452–455. https://doi.org/10.1038/nature20149

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

Fadason, T., Farrow, S., Gokuladhas, S., Golovina, E., Nyaga, D., O'Sullivan, J. M., & Schierding, W. (2022). Assigning function to SNPs: Considerations when interpreting genetic variation. *Seminars in Cell and Developmental Biology*, *121*, 135–142. https://doi.org/10.1016/j.semcdb.2021.08.008

Ferreira, F. J., Galhardo, M., Teixeira, J., Logarinho, E., & Salazar, A. (2023). FOXM1 expression reverts aging chromatin profiles through repression of the senescence-associated pioneer factor AP-1. *bioRxiv*. https://doi.org/10.1101/2023.05.04.539315

Flannick, J., & Florez, J. C. (2016). Type 2 diabetes: Genetic data sharing to advance complex disease research. *Nature Reviews Genetics*, *17*(9), 535–549. https://doi.org/10.1038/nrg.2016.56

Frank, S. A. (2004). Genetic predisposition to cancer - Insights from population genetics. *Nature Reviews Genetics*, *5*(10), 764–772. https://doi.org/10.1038/nrg1450

Fujimoto, K., & Polonsky, K. S. (2009). Pdx1 and other factors that regulate pancreatic β-cell survival. *Diabetes, Obesity and Metabolism*, *11*(4), 30–37. https://doi.org/10.1111/j.1463-1326.2009.01121.x

Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, *13*(12), 840–852. https://doi.org/10.1038/nrg3306

Furlong, E. E. M., & Levine, M. (2018). Developmental enhancers and chromosome topology. *Science*, *361*(6409), 1341–1345. https://doi.org/10.1126/science.aau0320

Gasperini, M., Starita, L., & Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. *Nature Protocols*, *11*(10), 1782–1787. https://doi.org/10.1038/nprot.2016.135

Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, *21*(5), 292–310. https://doi.org/10.1038/s41576-019-0209-0

Gaszner, M., & Felsenfeld, G. (2006). Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, *7*(9) 703–713. https://doi.org/10.1038/nrg1925

Giresi, P. G., & Lieb, J. D. (2009). Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods*, *48*(3), 233–239. https://doi.org/10.1016/j.ymeth.2009.03.003

Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D. T., & Marioni, J. C. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research*, *22*(12), 2376–2384. https://doi.org/10.1101/gr.142281.112

Gordon, M. G., Inoue, F., Martin, B., Schubach, M., Agarwal, V., Whalen, S., Feng, S., Zhao, J., Ashuach, T., Ziffra, R., Kreimer, A., Georgakopoulous-Soares, I., Yosef, N., Ye, C. J., Pollard, K. S., Shendure, J., Kircher, M., & Ahituv, N. (2020). lentiMPRA and MPRAflow for high-

throughput functional characterization of gene regulatory elements. *Nature Protocols*, *15*(8), 2387–2412. https://doi.org/10.1038/s41596-020-0333-5

Gorkin, D. U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A. Y., Li, B., Chiou, J., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J. S., Davidson, J. M., Qiu, Y., Afzal, V., Akiyama, J. A., Plajzer-Frick, I., Novak, C. S., … Ren, B. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature*, *583*(7818), 744–751. https://doi.org/10.1038/s41586-020-2093-3

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

Hager, G. L., McNally, J. G., & Misteli, T. (2009). Transcription Dynamics. *Molecular Cell*, 35(6), 741–753. https://doi.org/10.1016/j.molcel.2009.09.005

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

Hansen, A. S., Cattoglio, C., Darzacq, X., & Tjian, R. (2018). Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus*, *9*(1), 20–32. https://doi.org/10.1080/19491034.2017.1389365

Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R., & Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *ELife*, *6*. https://doi.org/10.7554/eLife.25776.001

Hardy, J., & Singleton, A. (2009). Genomewide Association Studies and Human Disease. *New England Journal of Medicine*, *360*(17), 1759–1768. https://doi.org/10.1056/nejmra0808700

Hauge-Evans, A. C., Squires, P. E., Persaud, S. J., & Jones, P. M. (1999). Pancreatic-C e l l-t o-Cell Interactions Are Required for Integrated Responses to Nutrient Stimuli Enhanced Ca 2 + and Insulin Secretory Responses of MIN6 Pseudoislets. *Diabetes*, *48*. http://doi.org/10.2337/diabetes.48.7.1402

Héberlé, É., & Bardet, A. F. (2019). Sensitivity of transcription factors to DNA methylation. *Essays in Biochemistry*, *63*(6), 727–741. https://doi.org/10.1042/EBC20190033

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., & Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. https://doi.org/10.1038/ng1966

Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C., & Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes and Development*, *32*(1), 26–41. https://doi.org/10.1101/gad.309351.117

Hnisz, D., Day, D. S., & Young, R. A. (2016). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*, *167*(5), 1188–1200. https://doi.org/10.1016/j.cell.2016.10.024

Horikawa, Y. (2018). Maturity-onset diabetes of the young as a model for elucidating the multifactorial origin of type 2 diabetes mellitus. *Journal of Diabetes Investigation*, *9*(4), 704–712). Blackwell Publishing. https://doi.org/10.1111/jdi.12812

Huang, D., & Ovcharenko, I. (2022). Enhancer-silencer transitions in the human genome. *Genome Research*, *32*(3), 437–448. https://doi.org/10.1101/gr.275992.121

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., Ahituv, N., & Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research*, *27*(1), 38–52. https://doi.org/10.1101/gr.212092.116

Ishihara, H., Asano, T., Tsukuda, K., Katagirii, ’ H, Inukai, K., Anai, M., Kikuchi, M., Yazaki, Y., Miyazaki, J.-I., & Oka, Y. (1993). Pancreatic beta cell line MIN6 exhibits characteristics of glucose metabolism and glucose-stimulated insulin secretion similar to those of normal islets. *Diabetologia, 36*. https://doi.org/ 10.1007/BF00401058

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

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

Kang, H., Shokhirev, M. N., Xu, Z., Chandran, S., Dixon, J. R., & Hetzer, M. W. (2020). Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. *Genes and Development*, *34*(13–14), 1–18. https://doi.org/10.1101/GAD.335794.119

Karpińska, M., & Czauderna, M. (2022). Pancreas—Its Functions, Disorders, and Physiological Impact on the Mammals' Organism. *Frontiers in Physiology, 13*. https://doi.org/10.3389/fphys.2022.807632

Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J., Jacobsen, L. M., Schatz, D. A., & Lernmark, A. (2017). Type 1 diabetes mellitus. *Nature Reviews Disease Primers*, *3*. https://doi.org/10.1038/nrdp.2017.16

Khetan, S., Kales, S., Kursawe, R., Jillette, A., Ulirsch, J. C., Reilly, S. K., Ucar, D., Tewhey, R., & Stitzel, M. L. (2021). Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive β cell transcriptional activation. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-25514-6

Kleinberger, J. W., & Pollin, T. I. (2015). Personalized medicine in diabetes mellitus: Current opportunities and future prospects. *Annals of the New York Academy of Sciences*, *1346*(1), 45–56. https://doi.org/10.1111/nyas.12757

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207–220. https://doi.org/10.1038/s41576-018-0089-8

Kwon, B., Fansler, M. M., Patel, N. D., Lee, J., Ma, W., & Mayr, C. (2022). Enhancers regulate 3' end processing activity to control expression of alternative 3'UTR isoforms. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-30525-y

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, *172*(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology, 29*(12), 1305-1312. https://doi.org/ 10.1016/s1357-2725(97)00085-x

Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., Fitzgerald, D., Kyono, Y., Ma, L., White, K. P., & Gerstein, M. (2020). STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02194-x

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

Lewis, E. B. (1950). The phenomenon of position effect. *Advances in genetics*, *3*, 73-115. https://doi.org/10.1016/S0065-2660(08)60083-8

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

Li, X., & Noll, M. (1994). Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. *The EMBO Journal*, *13*(2). https://doi.org/ 10.1002/j.1460-2075.1994.tb06274.x

Mahmoudi, T. (2002). GAGA can mediate enhancer function in trans by linking two separate DNA molecules. *The EMBO Journal*, *21*(7), 1775–1781. https://doi.org/10.1093/emboj/21.7.1775

Mannion, B. J., Osterwalder, M., Tran, S., Plajzer-Frick, I., Novak, C. S., Afzal, V., Akiyama, J. A., Barton, S., Beckman, E., Garvin, T. H., Godfrey, P., Godoy, J., Hunter, R. D., Kato, M., Kosicki, M., Kronshage, A. N., Lee, E. A., Meky, E. M., Pham, Q. T., … Pennacchio, L. A. (2022). Uncovering Hidden Enhancers Through Unbiased In Vivo Testing. *bioRxiv* https://doi.org/10.1101/2022.05.29.493901

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, *7*, 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623

Mastracci, T. L., & Sussel, L. (2012). The endocrine pancreas: Insights into development, differentiation, and diabetes. *Wiley Interdisciplinary Reviews: Developmental Biology, 1*(5), 609–628. https://doi.org/10.1002/wdev.44

Matsuda, H. (2018). Zebrafish as a model for studying functional pancreatic β cells development and regeneration. *Development Growth and Differentiation*, *60*(6), 393–399. https://doi.org/10.1111/dgd.12565

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., ... & Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190-1195. https://doi.org/ 10.1126/science.1222794

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., Leproust, E., Follows, G. A., Fraser, P., Luscombe, N. M., & Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*(6), 598–606. https://doi.org/10.1038/ng.3286

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B. M., Rolando, D. M. Y., Farabella, I., Morgan, C. C., García-Hurtado, J., Beucher, A., Morán, I., Pasquali, L., Ramos-Rodríguez, M., Appel, E. V. R., Linneberg, A., Gjesing, A. P., Witte, D. R., … Ferrer, J. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics*, *51*(7), 1137–1148. https://doi.org/10.1038/s41588-019-0457-0

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes and Development*, *32*(1), 42–57. https://doi.org/10.1101/gad.308619.117

Miyazaki, J. I., Araki, K., Yamato, E., Ikegami, H., Asano, T., Shibasaki, Y., ... & Yamamura, K. I. (1990). Establishment of a pancreatic β cell line that retains glucose-inducible insulin secretion: special reference to expression of glucose transporter isoforms. *Endocrinology, 127*(1), 126-132. https://doi.org/10.1210/endo-127-1-126

Mohlke, K. L., & Boehnke, M. (2015). Recent advances in understanding the genetic architecture of type 2 diabetes. *Human Molecular Genetics, 24*(1), 85-92. https://doi.org/10.1093/hmg/ddv264

Mora, A., Sandve, G. K., Gabrielsen, O. S., & Eskeland, R. (2016). In the loop: promoter-enhancer interactions and bioinformatics. *Briefings in bioinformatics*, *17*(6), 980–995. https://doi.org/10.1093/bib/bbv097

Muerdter, F., Boryń, Ł. M., & Arnold, C. D. (2015). STARR-seq - Principles and applications. *Genomics*, *106*(3), 145–150. https://doi.org/10.1016/j.ygeno.2015.06.001

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

Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., & Lu, T. (2014). Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, *13*(18), 2847–2852. https://doi.org/10.4161/15384101.2014.949201

Nakashima, K., Kanda, Y., Hirokawa, Y., Kawasaki, F., Matsuki, M., & Kaku, K. (2009). MIN6 Is Not a Pure Beta Cell Line but a Mixed Cell Line with Other Pancreatic Endocrine Hormones. *Endocrine Journal*, *56*(1). https://doi.org/10.1507/endocrj.k08e-172

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, *26*(8), 1023–1033. https://doi.org/10.1101/gr.204834.116

Noble, J. A., & Erlich, H. A. (2012). Genetics of type 1 diabetes. *Cold Spring Harbor Perspectives in Medicine*, *2*(1). https://doi.org/10.1101/cshperspect.a007732

Noonan, J. P., & McCallion, A. S. (2010). Genomics of long-range regulatory elements. *Annual Review of Genomics and Human Genetics*, *11*, 1–23. https://doi.org/10.1146/annurev-genom-082509-141651

O'Driscoll, L., Gammell, P., McKiernan, E., Ryan, E., Jeppesen, P. B., Rani, S., & Clynes, M. (2006). Phenotypic and global gene expression profile changes between low passage and high passage MIN-6 cells. *Journal of Endocrinology*, *191*(3), 665–676. https://doi.org/10.1677/joe.1.06894

Ohtsuki, S., Levine, M., & Cai, H. N. (1998). Different core promoters possess distinct regulatory activities in the Drosophila embryo. Genes & development, *12*(4), 547-556. https://doi.org/10.1101/gad.12.4.547

Pallejà, A., Horn, H., Eliasson, S., & Jensen, L. J. (2012). DistiLD database: Diseases and traits in linkage disequilibrium blocks. *Nucleic Acids Research*, *40*(1). https://doi.org/10.1093/nar/gkr899

Park, E. G., Ha, H., Lee, D. H., Kim, W. R., Lee, Y. J., Bae, W. H., & Kim, H. S. (2022). Genomic Analyses of Non-Coding RNAs Overlapping Transposable Elements and Its Implication to Human Diseases. *International Journal of Molecular Sciences*, *23*(16). https://doi.org/10.3390/ijms23168950

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J. J., Morán, I., Gómez-Marín, C., Van De Bunt, M., Ponsa-Cobas, J., Castro, N., Nammo, T., Cebola, I., García-Hurtado, J., Maestro, M. A., Pattou, F., Piemonti, L., Berney, T., … Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, *46*(2), 136–143. https://doi.org/10.1038/ng.2870

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

Persaud, S. J., Hauge-Evans, A. C., & Jones, P. M. (2014). Insulin-Secreting Cell Lines. Potential for Research and Diabetes Therapy. *Cellular Endocrinology in Health and Disease*, 239–256. https://doi.org/10.1016/B978-0-12-408134-5.00015-9

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C., & Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, *12*(1). https://doi.org/10.1186/s13104-019-4343-8

Poitout, V., Stout, L. E., Armstrong, M. B., Walseth, T. F., Sorenson, R. L., & Robertson, R. P. (1995). Morphological and functional characterization of βTC-6 cells—an insulin-secreting cell line derived from transgenic mice. *Diabetes, 44*(3), 306-313. https://doi.org/10.2337/diab.44.3.306

Ragvin, A., Moro, E., Fredman, D., Navratilova, P., Drivenes, Ø., Engström, P. G., Alonso, M. E., De La Calle Mustienes, E., Gómez Skarmeta, J. L., Tavares, M. J., Casares, F., Manzanares, M., Van Heyningen, V., Molven, A., Njølstad, P. R., Argenton, F., Lenhard, B., & Becker, T. S. (2010). Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(2), 775–780. https://doi.org/10.1073/pnas.0911591107

Ramos-Rodríguez, M., Raurell-Vila, H., Colli, M. L., Alvelos, M. I., Subirana-Granés, M., Juan-Mateu, J., Norris, R., Turatsinze, J. V., Nakayasu, E. S., Webb-Robertson, B. J. M., Inshaw, J. R. J., Marchetti, P., Piemonti, L., Esteller, M., Todd, J. A., Metz, T. O., Eizirik, D. L., & Pasquali, L. (2019). The impact of proinflammatory cytokines on the β-cell regulatory landscape provides insights into the genetics of type 1 diabetes. *Nature Genetics*, *51*(11), 1588–1595. https://doi.org/10.1038/s41588-019-0524-6

Ravassard, P., Hazhouz, Y., Pechberty, S., Bricout-Neveu, E., Armanet, M., Czernichow, P., & Scharfmann, R. (2011). A genetically engineered human pancreatic β cell line exhibiting glucose-inducible insulin secretion. *Journal of Clinical Investigation*, *121*(9), 3589–3597. https://doi.org/10.1172/JCI58447

Reményi, A., Schöler, H. R., & Wilmanns, M. (2004). Combinatorial control of gene expression. *Nature Structural and Molecular Biology*, 11(9), 812–815. https://doi.org/10.1038/nsmb820

Roep, B. O., & Peakman, M. (2012). Antigen targets of type 1 diabetes autoimmunity. *Cold Spring Harbor Perspectives in Medicine*, *2*(4). https://doi.org/10.1101/cshperspect.a007781

Roy, S., & Kundu, T. K. (2014). Gene regulatory networks and epigenetic modifications in cell differentiation. *IUBMB Life*, *66*(2), 100–109. https://doi.org/10.1002/iub.1249

Sandovici, I., Hammerle, C. M., Ozanne, S. E., & Constância, M. (2013). Developmental and environmental epigenetic programming of the endocrine pancreas: Consequences for type 2 diabetes. *Cellular and Molecular Life Sciences*, *70*(9), 1575–1595. https://doi.org/10.1007/s00018-013-1297-1

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., De Bakker, P. I. W., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., & Altshuler, D. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, *316*(5829), 1331–1336. https://doi.org/10.1126/science.1142358

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

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., & Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell*, *58*(6), 1101–1112. https://doi.org/10.1016/j.molcel.2015.04.006

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

Skelin, M. (2010). Pancreatic beta cell lines and their applications in diabetes mellitus research. *ALTEX*, 105–113. https://doi.org/10.14573/altex.2010.2.105

Skene, P. J., Henikoff, J. G., & Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols*, *13*(5), 1006–1019. https://doi.org/10.1038/nprot.2018.015

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*, 449–479). https://doi.org/10.1146/annurev.biochem.72.121801.161520

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), 613–626). https://doi.org/10.1038/nrg3207

Standl, E., Khunti, K., Hansen, T. B., & Schnell, O. (2019). The global epidemics of diabetes in the 21st century: Current situation and perspectives. *European Journal of Preventive Cardiology*, *26*(2), 7–14. https://doi.org/10.1177/2047487319881021

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

Sweeting, A., Wong, J., Murphy, H. R., & Ross, G. P. (2022). A Clinical Update on Gestational Diabetes Mellitus. *Endocrine Reviews*, *43*(5), 763–793. https://doi.org/10.1210/endrev/bnac003

Urakami, T. (2019). Maturity-onset diabetes of the young (MODY): Current perspectives on diagnosis and treatment. *Diabetes, Metabolic Syndrome and Obesity*, 12, 1047–1056). https://doi.org/10.2147/DMSO.S179793

Van Arensbergen, J., Fitzpatrick, V. D., De Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., & Van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology*, *35*(2), 145–153. https://doi.org/10.1038/nbt.3754

van Arensbergen, J., van Steensel, B., & Bussemaker, H. J. (2014). In search of the determinants of enhancer-promoter interaction specificity. *Trends in Cell Biology*, *24*(11) 695–702. https://doi.org/10.1016/j.tcb.2014.07.004

Vaz, S., Ferreira, F. J., Macedo, J. C., Leor, G., Ben-David, U., Bessa, J., & Logarinho, E. (2021). FOXM1 repression increases mitotic death upon antimitotic chemotherapy through BMF upregulation. *Cell Death and Disease*, *12*(6). https://doi.org/10.1038/s41419-021-03822-5

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research*, *35*(1). https://doi.org/10.1093/nar/gkl822

Warren, T. L., Lambert, J. T., & Nord, A. S. (2022). AAV Deployment of Enhancer-Based Expression Constructs In Vivo in Mouse Brain. *Journal of Visualized Experiments*, *2022*(181). https://doi.org/10.3791/62650

Weedon, M. N., Cebola, I., Patch, A. M., Flanagan, S. E., De Franco, E., Caswell, R., Rodríguez-Seguí, S. A., Shaw-Smith, C., Cho, C. H. H., Allen, H. L., Houghton, J. A. L., Roth, C. L., Chen, R., Hussain, K., Marsh, P., Vallier, L., Murray, A., Ellard, S., Ferrer, J., & Hattersley, A. T. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature Genetics*, *46*(1), 61–64. https://doi.org/10.1038/ng.2826

Wei, Q., Qi, L., Lin, H., Liu, D., Zhu, X., Dai, Y., Waldron, R. T., Lugea, A., Goodarzi, M. O., Pandol, S. J., & Li, L. (2020). Pathological Mechanisms in Diabetes of the Exocrine Pancreas: What's Known and What's to Know. *Frontiers in Physiology*, *11*. https://doi.org/10.3389/fphys.2020.570276

Roglic, G. and World Health Organization, Global report on diabetes. 2016, Geneva, Switzerland: World Health Organization. 86 pages.

Wilson, C., Bellen, H. J., & Gehring, W. J. (1990). Position effects on eukaryotic gene expression. *Annual Review of Cell Biology*, *6,* 679–714. https://doi.org/10.1146/annurev.cb.06.110190.003335

Winchester, L., Yau, C., & Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics*, *8*(5), 353–366. https://doi.org/10.1093/bfgp/elp017

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. https://doi.org/10.1038/nrg3095

Xu, Z., Wei, G., Chepelev, I., Zhao, K., & Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nature Structural and Molecular Biology*, *18*(3), 372–378. https://doi.org/10.1038/nsmb.1993

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, *518*(7540), 556–559. https://doi.org/10.1038/nature13994

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W. J., Hu, M., Tang, J., & Yue, F. (2018). Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03113-2

Zhang, Z., & Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. Journal of Biology, 2(2), 1-4. https://doi.org/10.1186/1475-4924-2-11

Work performed at i3S, University of Porto in the VDR group, led by Doctor José Bessa.





Work funded by:

"la Caixa" Foundation