Three-dimensional chromatin organisation in human pancreatic islets

by Joan Ponsà-Cobas

Department of Medicine

Division of Diabetes, Endocrinology and Metabolism Section of Epigenomics and Disease

Imperial College London

London, United Kingdom

Thesis submitted to Imperial College London for the degree of

Doctor of Philosophy

July 14, 2017

Copyright declaration

Abstract

Diabetes is a group of metabolic diseases that affects millions of people. Despite this, little is known about the underlying molecular mechanisms. Diabetes is characterised by an impaired blood-glucose regulation that can lead to severe consequences, such as kidney failure, and premature death. Pancreatic islets are one of the major tissues to understand diabetes pathogenesis as they produce insulin, a hormone central for blood-glucose homeostasis. Our previous work showed that studying epigenomic regulation is key to giving insight into the molecular mechanisms underlying diabetes, as risk-associated genomic variants are enriched at transcriptional regulatory regions named enhancers. To give further insight in pancreatic islet transcriptional regulation, I aimed to decipher the 3D chromatin organisation, an aspect of epigenomic regulation in human pancreatic islets that remained largely unexplored until now.

As part of my PhD project I have studied high-resolution chromatin interaction maps that characterise 3D chromatin organisation at different levels, from single interactions between specific pair of genomic loci to large genomic topological domains known as TADs. These high-resolution chromatin interaction maps, integrated with a large collection of epigenomic datasets, allowed me to describe several aspects of islet 3D chromatin organisation, such as the identification of islet-selective chromatin structures associated to islet-specific gene expression. Moreover, I identified groups of enhancers that gather in 3D space. These 3D enhancer clusters were frequently found in loci key for islet function and highly enriched in diabetes associated variants.

The results of this thesis allow us to have a more accurate picture of the epigenomic regulation in human pancreatic islets and how non-coding diabetes risk variants could be impairing enhancer-promoter communication.

Acknowledgements

I would like to thank my supervisor Professor Jorge Ferrer for giving me the opportunity to be part of his team and his guidance through my studentship.

Thanks to all current and past members of J. Ferrer's lab for their useful discussions and advice during this journey, this project would not have succeeded without them. It has been an honour to be able to work with such talented scientists and wonderful people.

Dennis Affram	Natalia Castro	Vanessa Grau	Nikolina Nakic
Ildem Akerman	Inês Cebola	Mark Kalisz	Lorenzo Pasquali
Mar Armengol	Miguel Correa	Miguel Angel Maestro	Delphine Rolando
Goutham Atla	Matías De Vas	Irene Miguel	Meritxell Rovira
Anthony Beucher	Berta Font	Ignasi Morán	Natalia Ruiz
Silvia Bonas	Xavier García	Claire C. Morgan	Ana Sancho
Leontien Bosch	Roser González	Loris Mularoni	

I would also like to acknowledge Dr. Biola Javierre and Professor Peter Fraser for their advice and our constructive discussions, which was led to a very productive collaboration. Also thanks to Dr. Thomas Thorne, Dr. Rodrigo Liberal Fernandes and Alvaro Perdones-Montero; who, despite not being involved in the project, dedicated part of their time to advise me on the machine learning analyses.

A special mention to my parents, Elvira and Josep, and my siblings, Laia and Sergi. Despite the distance, they have been an indispensable support over the years.

Finally, thanks to Roser for her unconditional support. I doubt anyone would be reading this thesis if it had not been for her.

Statement of contributions

I hereby declare that all work described in this dissertation has been performed by myself, unless otherwise explicitly stated, and that all elements derived from the work of others have been appropriately credited and referenced.

Table of contents

- Copyright declaration 2
- Abstract 3
- Acknowledgments 4
- Statement of contributions 5
- Table of contents 6
- List of figures 7
- List of tables 10
- List of abbreviations and terms 11

1. Introduction 13
   - **1.1.** Regulation of gene transcription 14
   - **1.2.** Chromatin and chromatin regulators 15
   - **1.3.** DNA binding transcription factors and transcription machinery 17
   - **1.4.** DNA regions that control gene transcription 22
   - **1.5.** Compartmentalisation of gene regulation 32
   - **1.6.** Tissue-specific transcriptional circuitries 47
   - **1.7.** Regulatory genomics and diseases 49
2. Rationale, hypotheses and aims 56
3. 3D chromatin organisation in human pancreatic islets 60
4. Promoter centric chromatin interaction domains 85
5. Experimental identification of non-coding functional variants 105
6. Discussion and prospects 111
7. Conclusions 122
8. Methods 125

- Bibliography 156
- Appendix A. Copyright permissions 168
- Appendix B. Publications 179

List of figures

Fig. 1: Chromatin modifying and remodelling factors 16
Fig. 2: Regulatory region activation through TF binding 19
Fig. 3: Assembly of the transcriptional machinery on gene promoters 21
Fig. 4: Enhancer-promoter communication through chromatin folding 24
Fig. 5: CRE-seq and STARR-seq reporter constructs 28
Fig. 6: Promoter capture Hi-C experimental design 36
Fig. 7: Nuclear chromatin compartmentalisation 38
Fig. 8: Extrusion model 42
Fig. 9: Tissue-specific gene regulation 47
Fig. 10: Systematic identification of causal genomic variants 51
Fig. 11: Cell heterogeneity in human pancreatic islets 52
Fig. 12: Pancreatic cell differentiation pathways 53
Fig. 13: Schematic of overall strategy 61
Fig. 14: Schematic representation of chromatin states at the *MAFB* locus 62
Fig. 15: Enrichments at non-baited promoter-interacting regions 67
Fig. 16: CTCF as chromatin interaction facilitator 68
Fig. 17: Tissues selective chromatin interactions in the *FOXA2* locus 71
Fig. 18: Presence of islet-selective interactions correlates with islet-specific gene expression 72
Fig. 19: Mediator bound enhancers are more frequently found at promoter-interacting regions of islet-selective chromatin interactions 73
Fig. 20: Islet-specific CTCF binding sites are more frequent in islet-selective chromatin interactions 74
Fig. 21: Tissue-selective chromatin interactions in the *ISL1* locus 75
Fig. 22: Islet TAD-like compartments 77
Fig. 23: Islet chromatin compartmentalisation exhibits known features of topological associating domains (TADs) 78
Fig. 24: Identification of chromatin contacts through re-ligation events 80
Fig. 25: Distance from an interacting site to the closest epigenomic factor 82
Fig. 26: Cooperative work between islet LDTFs, CTCF, MED1 and SMC1 84

Fig. 27: KCNJ11 promoter-associated domain (PAD) 86
Fig. 28: Overlap between TAD and PAD segmentation 86
Fig. 29: Epigenomic states at islet PADs are coherent with gene expression 87
Fig. 30: Comparative between epigenomic states at PADs and remaining TAD-like space 88
Fig. 31: Enhancer assignment 90
Fig. 32: Enhancer assignments considering chromatin interaction maps accentuate their association with islet-specific expressed genes 91
Fig. 33: Logit distribution as gene probability of being islet specific 93
Fig. 34: Feature's informativeness to identify islet-specific expressed genes among expressed genes 94
Fig. 35: PAD classification based on enhancer content 97
Fig. 36: Characterisation of PADs with different enhancer contents 98
Fig. 37: Enhancer-rich PADs showed similar structural features to PADs associated with enhancer clusters and super-enhancers 99
Fig. 38: Enhancer-rich PADs showed a similar enrichment for islet-specific expressed genes as PADs associated to enhancer clusters and super-enhancers 100
Fig. 39: Most enhancer clusters and super-enhancers were assigned to enhancer-rich PADs 101
Fig. 40: Enhancers forming enhancer rich PADs showed strong enrichment for T2D and FG risk associated variants, as previously observed for super-enhancers and enhancers forming enhancer clusters 102
Fig. 41: Overlap between enhancer-rich PADs 104
Fig. 42: Enhancer activity detected by STARR-seq 106
Fig. 43: Comparison between STARR-seq and Luciferase reporter assay 107
Fig. 44: Diagram of the different constructs used to measure enhancer activity 108
Fig. 45: Impact of different promoter types over enhancer activity luciferase reporter assays 109
Fig. 46: Islet regulome 126
Fig. 47: ChromHMM segmentation in 15 states 127
Fig. 48: ChromHMM genome coverage 127

Fig. 49: Virtual 4C around *KCNJ11's* locus 129
Fig. 50: Enrichments at baited promoters 130
Fig. 51: Distance from an interacting site to the closest interrogated epigenomic factor site 132
Fig. 52: Gene classification based on tissue-specificity expression patterns 134
Fig. 53: CTCF tissue-specificity 136
Fig. 54: DI domain interconnectivity distribution 137
Fig. 55: Selection of genes assigned to islet active enhancers and control genes 141
Fig. 56: Enhancer assignments considering chromatin interaction maps accentuate their association with islet-specific expressed genes 141
Fig. 57: Enhancer assignments considering chromatin interaction maps are coherent with enhancer-promoter H3K27ac correlations 143
Fig. 58: TSS length distribution 146
Fig. 59: Epigenomic characterisation of islet-specific expressed genes 147
Fig. 60: Correlation between epigenomic features 148
Fig. 61: Logistic Regression analysis to determine features associated to gene classes 150
Fig. 62: High correlation between STARR-seq biological replicates 154

List of tables

Table 1: Promoter classification based on TSS shape and CpG island content 23

Table 2: Comparative between CRE-seq and STARR-seq 28

Table 3: Effect on non-coding variants on *cis*-regulatory elements 49

Table 4: Descriptive analysis of pcHi-C interactions detected in human islets 65

Table 5: Description of TAD-like domains 77

Table 6: Description of islet PADs 85

Table 7: Epigenomic features interrogated in a logistic regression analysis to determine their association with tissue-specific gene expression 93

Table 8: Summary of the 4 different enhancer activity reporter assays 109

Table 9: Summary of CTCF ChIP-seq datasets used to determine tissue-specificity of CTCF-binding sites in human pancreatic islets 135

Table 10: A and B variables from the DI score formula adapted to pcHi-C 137

Table 11: Summary of H3K27ac ChIP-seq datasets used for enhance-promoter correlations 143

Table 12: Number of reads obtained per samples 154

List of abbreviations and terms

CGI – CpG island.

ChIP – chromatin immunoprecipitation.

ChIP-seq – chromatin immunoprecipitation sequencing.

CRE-seq – *cis*-regulatory element analysis by sequencing.

DNA – deoxyribonucleic acid.

EC – enhancer cluster.

ENCODE – encyclopedia of DNA elements.

eQTL – expression quantitative trait locus.

eRNA – enhancer RNA.

FG – fasting glycemia.

GWAS – genome-wide association study.

H3K27ac – histone 3 lysine 27 acetylation.

H3K27me3 – histone 3 lysine 27 trimethylation.

H3K36me3 – histone 3 lysine 36 trimethylation.

H3K4me1 – histone 3 lysine 4 monomethylation.

H3K4me3 – histone 3 lysine 4 trimethylation.

IQR – Interquartile range.

LD – linkage disequilibrium.

LDTFs – lineage-determining transcription factor.

MPRA – massively parallel reporter assay.

NGS – next-generation sequencing.

PAD – promoter-associated domain.

pcHi-C – promoter capture Hi-C.

PCR – Polymerase chain reaction.

PIC – pre-initiation complex.

PIR – promoter-interacting region.

qPCR – quantitative real-time PCR.

RNA – ribonucleic acid.

RNA Pol II – RNA polymerase II.

RNA-seq – RNA sequencing.

SDTFs – signal-dependent transcription factor.

SE – super-enhancer.

SNP – single-nucleotide polymorphism.

SNV – single-nucleotide variant.

STARR-seq – Self-transcribing active regulatory region sequencing.

T2D – type-2 diabetes.

TAD – topological associating domain.

TF – transcription factor.

TFBS – transcription factor binding site.

TSS – transcription start site.

Chapter 1

Introduction

The human body is formed by hundreds of cell types that are raised from a common pluripotent cellular lineage. The differentiation process from a pluripotent cell to any cell type is possible because the genome contains all the information required for all possible cell fates. As the genomic information is the same for all cell types, cell identity and proper cell function are determined by cell-specific transcriptional programs.

Precise gene regulation is managed by a broad collection of transcriptional regulatory elements located throughout the genome. These regulatory elements act as recruitment platforms for the transcriptional machinery and are formed by clusters of short DNA sequences named transcriptional factor binding sites (TFBS) (Cooper and Hausman, 2007). Transcription factors (TFs) are key proteins for transcription regulation as they bring other factors required for the proper assembly of the transcriptional machinery to specific loci. Therefore, it is not surprising that cell-specific transcriptional programs are managed by small groups of cell-specific TFs (Heinz et al., 2015).

It is widely known that gene expression misregulation can lead to abnormal cell functions and in ultimate instances to development of diseases. Several genome wide association studies (GWAS), conducted to determine the genetic factors behind a certain pathogenicity, have identified hundreds of non-coding genomic variants (McClellan and King, 2010). In many cases these non-coding variants tend to occur in genomic regulatory regions, affecting their functionality and provoking alterations in gene expression. Therefore, a better understanding of the molecular mechanisms involved in cell-specific gene expression and in the effect of non-coding variants could provide insight in the genetic factors behind major diseases such as cancer or diabetes.

1.1. Regulation of gene transcription

The human genome contains more than 40,000 coding and non-coding genes (Aken et al., 2016). Although the human body is formed by millions of cells grouped in different cell types that take care of a broad range of biological functions, all cells, except for the germ line, have virtually the same information encoded in their genome. Cell identity, therefore, is determined by gene expression programs or transcriptional circuitries that regulate specific groups of genes. The regulation of gene transcription is not only required to determine a specific cell stage but also for its proper biological function.

Eukaryotic gene expression is regulated through a set of events that need to occur in order to allow gene transcription. This process begins with the creation of an active epigenomic environment conducted by key proteins known as transcription factors (TFs) and chromatin modifiers. It is then followed by the assembly and activation of the transcriptional machinery at the gene promoter; and it proceeds with gene transcription, transcript stability, translation efficiency and peptide stability.

1.2. Chromatin and chromatin regulators

The biological information codified in the genome is localised in a cellular compartment called **nucleus**. The haploid human genome contains 3.2 billion nucleotides that need to be compacted 400,000-fold (Schneider and Grosschedl, 2007) to fit within the nucleus. To achieve this high degree of compactness, double-helix DNA is wrapped around a group of proteins called histones, forming a structure named nucleosome. The combination of DNA and proteins form a complex of macromolecules named **chromatin** (Cooper and Hausman, 2007).

Due to the extension of the genome, DNA is packed in nucleosomes creating a structure named "beads on a string". The distance between the nucleosomes determines the grade of compactness that is not homogenous throughout the genome. Highly compacted regions are named heterochromatin, while those more loose are known as euchromatin. **Heterochromatin** are loci in a highly condensed state, associated with gene repression or inactivation, and highly repeated DNA sequences such as centromeric or telomeric sequences. In contrast, 90% of the chromatin of non-dividing cells is in a chromatin state of low compactness, named **euchromatin**. About 10% of the euchromatin is in a specially decondensed state that allows the transcriptional machinery to access genomic information (Cooper and Hausman, 2007). Thus, it is clear that local chromatin compactness limits the reading of genomic information codified in a locus.

Chromatin accessibility is regulated through multiple mechanisms including post-translational histone modifications (PTMs) produced by **chromatin modifying factors**. These factors can be classified as "readers", "writers" or "erasers" depending on their functions. Chromatin factors known as "writers" and "erasers" have opposite functions incorporating or removing modifications such as histone post-translational acetylation or methylation. Histone post-translational modifiers can be grouped in families depending on the resultant modification, such as histone acetyltransferases (HATs), histone deacetylases (HDACs), lysine methyltransferases (KMTs) and lysine demethylases (KDMs). These modifications are recognised by the "readers". Some of these readers are **chromatin remodelling factors**, which are encompassed in 4 main families: SWI/SNF, ISWI, chromodomain-helicase DNA-

binding protein (CHD) and INO80 complexes (Ho and Crabtree, 2010). Chromatin remodelling factors use the energy of ATP hydrolysis to restructure nucleosomes and consequently alter chromatin accessibility (Chen and Dent, 2014) (Fig. 1).

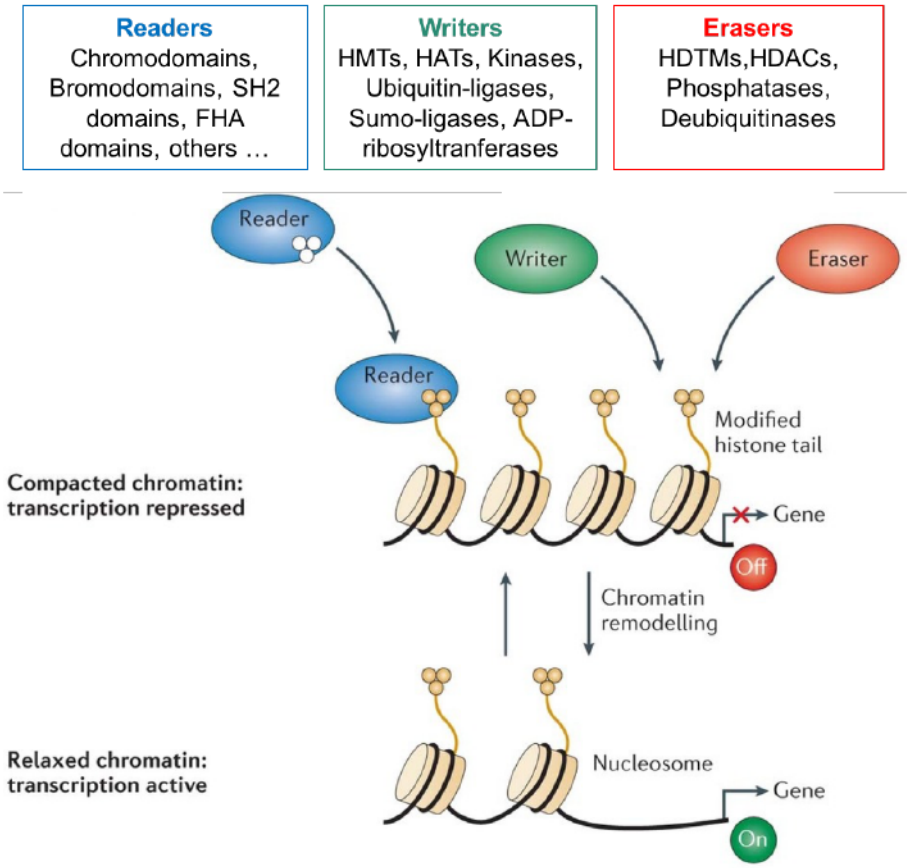


Fig. 1: Chromatin modifying and remodelling factors. Model illustrating how chromatin accessibility and gene regulation can be modulated through histone PTMs and the action of chromatin factors. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Drug Discovery, (Højfeldt et al., 2013), copyright (2013)

1.3. DNA binding transcription factors and the transcription machinery

In addition to chromatin and its modifiers, there is an additional transcription regulation mechanism formed by the transcription factors (TFs) and the transcription apparatus.

- **DNA binding transcription factors (TFs)**

In response to cellular stimuli there is an increase of specific DNA binding transcription factor (TF) proteins in the nucleus. TFs, as other proteins, are formed by multiple domains with different functions. TFs present DNA binding domains that recognise specific DNA sequences, or motifs, in accessible chromatin regions known as **TF binding sites (TFBS)** (Maston et al., 2006). In addition, TFs also present protein interacting domains used to attract other proteins to the same genomic locus.

It has been predicted that there are more than a thousand genes in the human genome that codify for DNA binding TFs (Vaquerizas et al., 2009), which proteins have been systemically studied to characterise (*in vitro*) TF binding models (Isakova et al., 2017; Jolma et al., 2010, 2013, 2015). TFs can be grouped in almost 200 TF families based on their DNA binding domains (e.g.: Zinc-finger, Homeodomain, Helix-loop-helix) (Wilson et al., 2007), and some of the major TF families are extensively reviewed in Luscombe, Austin, Berman, & Thornton, 2000; Pabo & Sauer, 1992; Rohs et al., 2010.

As part of the mechanism that regulates gene expression, TFBS DNA sequences are frequently located at **transcriptional regulatory genomic elements** (Maston et al., 2006). These elements serve as platforms to recruit the molecular machinery that will modulate gene expression. However, there is not perfect correlation between the presence of a TFBS motif and the degree of TF binding, which means that further mechanisms might contribute.

TF DNA sequence recognition is not exclusively driven by the linear DNA sequence but also by its local topography. Thus, at linear level, protein – DNA interaction is based on the complementary recognition between hydrogen bond acceptors and donors from the two macromolecules. However, this recognition must be possible through the 3D DNA-protein assembly (Rohs et al., 2010).

In addition to DNA sequence composition and its shape, DNA sequence recognition is also governed by **TF co-binding**. *In vitro* experiments have shown that many TFs need cooperative interactions to bind their target sequence in nucleosomal DNA (Adams and Workman, 1995; Zaret and Carroll, 2011). Moreover, there is experimental evidence of combinatorial epigenomic marks, such as histone methylation (Bartke et al., 2010) and histone acetylation (Shogren-Knaak, 2006), affecting TF binding.

Although most TFs are only able to recognise DNA binding motifs in euchromatin regions, there is a group of TFs, such as FOXA or GATA binding factors, known as **pioneer TFs** that do not have this limitation (Cirillo et al., 2002; Hatta and Cirillo, 2007; Zaret and Carroll, 2011). Although the full mechanism by which pioneer TF bind to compacted DNA remains to be elucidated, recent studies have shown that they are able to recognise partial DNA motifs that are exposed on the nucleosomes surface (Soufi et al., 2015; Ye et al., 2016). However, as previously mentioned for non-pioneer TFs, the presence of binding site sequences is not a good predictor of the TF occupancy and the recognition of these DNA motif may occur through cooperative binding.

It has been shown that pioneer TFs binding precedes the occupancy of other TFs during the activation of developmental regulatory regions (Zaret and Carroll, 2011). These results suggest that pioneer TFs form part of the triggering mechanism that activates silenced regulatory regions. In that sense, one main function of pioneer TF would be to increase local chromatin accessibility, by disrupting local internucleosomal interactions and destabilising the chromatin structure (Schalch et al., 2005). Thus, by increasing chromatin accessibility, pioneer TFs facilitate the binding of other proteins.

TFs are key elements on the regulation of transcriptional programs as their presence modulates the transcriptional machinery recruitment to target genes. However, not all expressed TFs are equally important to regulate a cell-specific transcriptional program. It has been shown that for each cell type there is a specific small set of TFs that governs and drives its transcriptional circuitry. These TFs are known as **master regulators** or **lineage-determining transcription factors (LDTFs)** (Heinz et al., 2015). Some LDTFs may be pioneer TFs, but not all pioneers are necessarily LDTF. Although LDTFs have only been recently

described (Heinz et al., 2010; Tronche and Yaniv, 1992) some characteristics are already well established. First, LDTFs are highly expressed in the relevant cell type. Second, a LDTF is under an auto-regulatory loop, meaning that it is able to modulate its own transcription. Third, all LDTFs modulate the expression of the other LDTFs in the relevant cell type, forming a core regulatory circuit (Lee and Young, 2013). Finally, LDTFs are strongly associated with tissue-specific distal transcriptional regulatory elements known as super-enhancers (SEs) (Whyte et al., 2013) or highly bound enhancer clusters (ECs) (Pasquali et al., 2014).

Part of a cell expression regulation occurs as a response to internal and external signals. Thus, many regulatory elements are also regulated by **signal-dependent transcription factors (SDTFs)**, such as nuclear receptor TFs. Whereas LDTFs expression tends to be cell-specific, SDTFs may be expressed in a broad collection of cell types. However, as LDTFs, the effect on expression regulation of SDTFs can be cell-specific (Heinz et al., 2015).

In summary, in addition to other mechanism, gene transcription is regulated through the cooperative binding of different types of TFs and other proteins. The cooperative action between TFs is governed through a hierarchy that allows cell-specific regulation during differentiation and in front of different stimuli (Fig. 2) (Heinz et al., 2015).

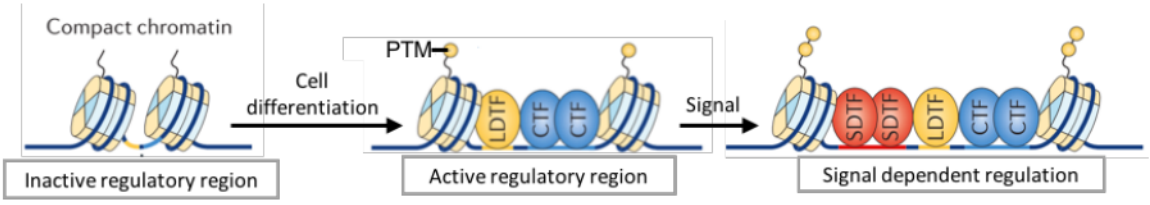


Fig. 2: Regulatory region activation through TF binding. Diagram illustrating the cooperative binding between LDTFs, SDTFs and cofactors (CTF, see following section) to activate transcriptional regulatory elements. The cooperative binding between LDTFs, some of them pioneer TFs, and cofactors drives the activation of regulatory regions during differentiation allowing cell-specific gene expression. Activity states of regulatory regions tend to correlate with the presence of histone PTMs. In response to different, intra or inter, cellular signals SDTFs bind to certain regulatory regions modulating their activity. Adapted by permission from Macmillan Publishers Ltd: Nature reviews. Molecular cell biology, Heinz et al., 2015; copyright (2015).

- **TF recruitment of regulatory proteins**

One of the main functions of TFs is to drive the genomic localisation of other proteins that do not have a DNA sequence binding domain. By interacting with TFs' protein binding domains, other proteins known as **cofactors** are recruited to specific genomic locations where their action is required. Cofactors have two main functions, act as chromatin modifiers or facilitate the recruitment and assembly of the transcriptional machinery.

A clear example of a chromatin modifier cofactor would be the histone acetyltransferase (HAT) p300. As a "writer", p300 has the capacity to acetylate the lysine 27 of the histone H3 (H3K27ac); a histone mark that correlates with the presence of transcriptional regulatory elements known as active enhancers (Heintzman et al., 2007, 2009). In contraposition to p300, the Polycomb repressive complex 2 (PRC2) contains the Enhancer of zeste methyltransferase (E(z)) able to methylate the lysine 27 of the histone H3 (H3K27me3). The presence of H3K27me3 prevents the deposition of H3K27ac by p300. PRC2 as part of the Polycomb group (PcG) has been associated to gene expression repression, although the exact mechanism is still unknown (Schwartz and Pirrotta, 2013).

Another relevant cofactor is the Mediator complex, a protein complex that is essential for gene transcription. One of its main functions is to act as a scaffold for the recruitment, assembly and activation of the RNA transcription machinery required for gene transcription (Hahn, 2004). Mediator is recruited to specific regulatory regions by TFs, which do not interact directly with the RNA polymerase. Additionally, it has been observed that the Mediator complex recruits chromatin remodelers, such as SWI/SNF, which are involved in nucleosome removal increasing chromatin accessibility (Allen and Taatjes, 2015; Lemieux and Gaudreau, 2004).

- **RNA polymerase II transcriptional machinery**

Transcription is carried out by an enzymatic class named RNA polymerase (RNA Pol). In eukaryotes, there are 3 types of RNA polymerases (RNA Pol I-III) that transcribe different classes of RNAs. However, the most relevant for this thesis is the **RNA Pol-II,** which is in

charge of gene transcription by producing messenger RNA (mRNA) (Cooper and Hausman, 2007). Eukaryotic gene transcription generally begins with the recruitment of the transcriptional machinery containing an RNA Pol-II into the promoter, a transcriptional regulatory region located just in front of genes. The promoter acts as a platform for the assembly of the **transcriptional preinitiation complex** (**PIC**) (Hahn, 2004) (Fig. 3). Although the RNA Pol-II transcriptional machinery may be stably assembled at the promoter, transcription initiation cannot occur without the formation of the **transcriptional bubble**. The transcriptional bubble is generated by an ATP-dependent DNA helicase able to separate the two DNA strands, which allows the RNA Pol-II to migrate through a single strand DNA template. Once the transcription start site (TSS) is recognised, transcription is initiated. After transcribing a few tens of base pairs (bp), the RNA Pol-II can leave the transcriptional machinery complex and the promoter to continue with the transcription elongation. At the end of this process, the polymerase will have produced a RNA copy of the genomic information (Cooper and Hausman, 2007).

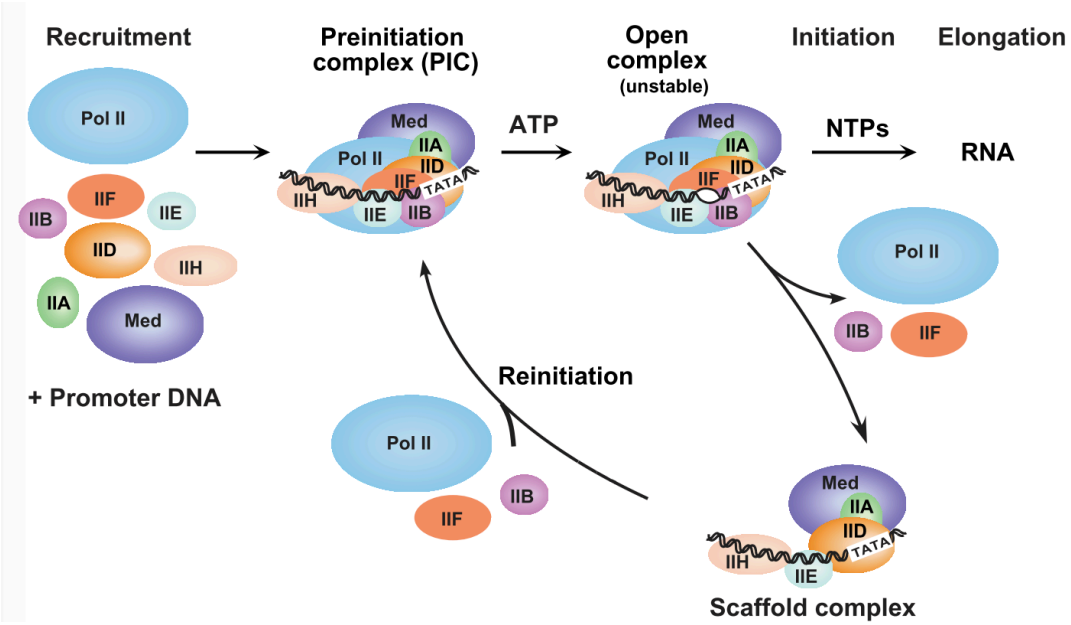


Fig. 3: Assembly of the transcriptional machinery on gene promoters. Diagram illustrating the molecular processes that occur at gene promoters, which lead to gene transcription. Transcriptional machinery formed by the RNA-pol II (Pol II), general TFs (II A-H) and Mediator (Med), is recruited to the core promoter, formed by DNA sequences such as the TATA-box (TATA). Reprinted by permission from Macmillan Publishers Ltd: Nature structural & molecular biology, Hahn, 2004; copyright (2004)

1.4. DNA regions that control gene transcription

Gene transcription regulation can occur in two different levels, ***trans*** or ***cis***. ***Trans*-acting factors** encompass all proteins that regulate gene expression by binding (directly or indirectly) to DNA. ***Cis*-regulatory elements** are DNA sequences embedded in the genome that affect gene expression by recruiting *trans*-acting factors. There are two main types of *cis*-regulatory regions, promoters and enhancers.

- **Promoters**

Promoters are essential transcriptional regulatory regions located at 5' end of genes. Promoters are required to define the RNA Pol transcription start site (TSS), and determine the direction of transcription through the presence of **core promoter DNA sequences;** such as the TATA box, the TFIIB-recognition element (BRE), the Initiator element (Inr) and downstream promoter element (DPE). Therefore, promoters ensure the correct location of the RNA Pol-II and proper transcription. Although these sequences are not essential for the functionality of the promoter, different combinations of these sequences are frequently found in different promoters. These sequences are recognised by different elements of the transcriptional machinery such as the ubiquitously expressed **general TFs** (TFIIA-H) (Hahn, 2004). These TFs bring the RNA-Pol II and cofactors required for the transcriptional preinitiation complex formation, such as Mediator (Fig. 3). Upstream of the core promoter there is a regulatory extension named **proximal promoter**. Proximal promoters are bound by cell-specific TFs, which modulate the cell-specific expression pattern of the gene.

A further layer of promoter regulation would be the presence of **DNA methylation**. This epigenomic modification typically consists on the addition of a methyl group on CpG dinucleotides (as reviewed in Jones, 2012). It has been determined that (approx.) 70% of all promoters contain CpG-rich regions of DNA, known as CpG islands or CGI (Saxonov et al., 2006). Although originally DNA methylation at *cis*-regulatory regions, such as promoters, was associated with silencing (Deaton and Bird, 2011), the latest reports indicate it may vary with context. Initial studies showed DNA methylation as a mechanism to block TF binding. However, recent publications have indicated that DNA methylation could also have the opposite effect, being required to allow TF binding (Zhu et al., 2016). It is important to

mention, that DNA methylation and CGI are not exclusively observed at promoters. In contraposition to its repressible role, CpG methylation is enriched at highly transcribed gene bodies. Despite the role of CpG methylation at gene bodies remains largely unknown, a recent study conducted in S. Oliviero's lab indicates that intragenic DNA methylation may prevent spurious transcription initiation (Neri et al., 2017).

It has been proposed that there are 3 **promoter classes** (type I-III) that are associated with different types of genes. They have been extensively reviewed in Lenhard, Sandelin, & Carninci, 2012 (Table 1). Briefly, they can be mainly differentiated by two features: TSS "shape" and CpG content. Next generation sequencing (NGS) techniques, such as cap analysis of gene expression (CAGE), have allowed the precise definition of gene TSSs (Carninci et al., 2006; Shiraki et al., 2003). It has been reported that some genes have a precise TSS that initiates from a single nucleotide position (henceforth named "sharp" TSS) while other genes have multiple clustered TSSs (henceforth named "broad" TSSs) (Forrest et al., 2014). Genes containing CpG islands can be further classified according to their extension. CpG islands can be found as just a few nucleotides in front of the TSS ("short" CpG islands) or appear as large regions that extend throughout the gene body ("large" CpG islands).

Table 1. Promoter classification based on TSS shape and CpG island content. *

Promoter class	Gene type	TSS	CpG content
Type I	Adult tissue-specific genes	"sharp"	absence of CpG islands
Type II	Ubiquitously expressed genes	"broad"	"short" CpG islands
Type III	Developmental genes	"broad"	"large" CpG islands

* Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Lenhard et al., 2012; copyright (2012).

- **Enhancers**

Spatiotemporal expression patterns of genes are determined by **enhancers**, regulatory regions that modulate promoter activity. Enhancer DNA sequences are typically composed by groups of clustered TFBS which recruit cell-specific TFs and cofactors that work cooperatively and determine the activity of the enhancer (Heinz et al., 2015). As opposed to

promoters, enhancers do not have a precise genomic position in relation to the genes they regulate. Enhancer activity has been shown to be independent of the relative linear distance and orientation to the target promoter (Banerji et al., 1981). This is due to their capacity to gain proximity with target promoters by looping over long genomic distances and skipping untargeted regions. The mechanism of chromatin loop formation will be extensively discussed in further sections but in general lines, the genome can be bent by topological proteins such as Cohesin due to its polymeric nature (Kagey et al., 2010; Tolhuis et al., 2002) (Fig. 4). Thus, allowing a physic interaction between two distal regulatory regions.

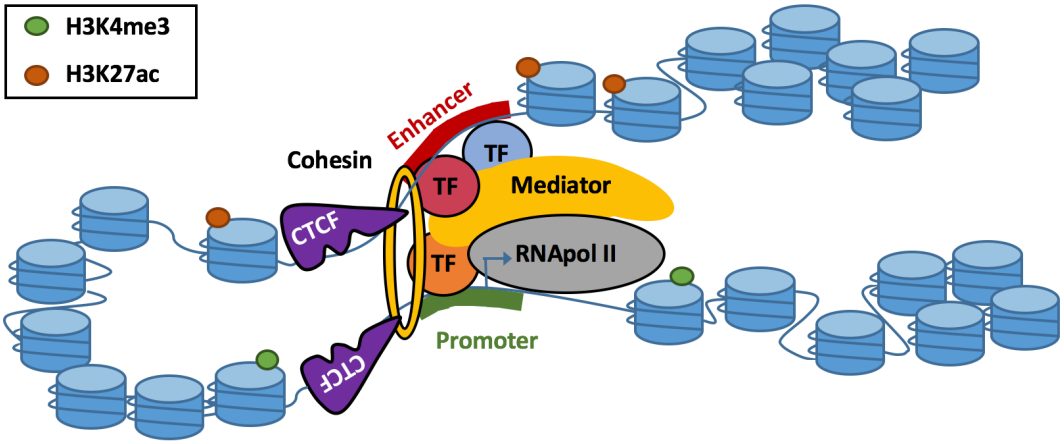


Fig. 4: Enhancer-promoter communication through chromatin folding. TFs at enhancers are able to interact with the transcriptional machinery assembled at promoters. Despite these elements being in large linear distance they can communicate due to the combined action of different interacting factors such as CTCF, Mediator and Cohesin.

Although none of the known histone modifications, individually or in combination, are perfect predictors of enhancer activity (Shlyueva et al., 2014a), there is some correlation. For that reason, enhancers can be sub-classified in categories such as active, poised or repressed based on the presence of histone marks.

- **Active enhancers** are typically defined as open chromatin regions containing H3K27ac (Heintzman et al., 2009). Active enhancers do also contain H3K4me1, a histone modification associated with different types of active *cis*-regulatory regions (Heintzman et al., 2007).

Different sub-classes of active enhancer have been reported suggesting that they may have different roles or relevance in tissue-specific gene transcription regulation. Enhancers are not evenly distributed through the genome; some are located in short linear proximity forming what has been named **enhancer clusters** (ECs). Among all ECs active in a specific tissue, a fraction of them tend to be highly bound by tissue specific TFs. These **highly bound ECs** are frequently located near tissue-specific expressed genes, suggesting that they may be key tissue-specific gene expression regulators (Pasquali et al., 2014). Studies have shown that highly bound active enhancers tend to be also bound by **Mediator** complex, a cofactor highly relevant for gene transcription regulation (Kagey et al., 2010; Whyte et al., 2013). R. Young's lab identified a sub-set of stitched active enhancers, that expand large genomic regions (median size 8.7 kb), which are highly bound by Mediator. These elements were called **super-enhancers (SEs)**. It has been observed that SEs have an exceptional enrichment for cell-specific key TFs and chromatin marks that can be also used to define them (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). In concordance with highly bound ECs (Pasquali et al., 2014), SEs are not only strongly associated with tissue specific expressed genes but also disease associated genomic variants (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013).

Currently, there is a debate regarding if all enhancers within an EC are equally important for gene regulation as contradictory result have been recently published (Dukler et al., 2017). Some studies suggest that clustered enhancers work in synergy within a hierarchical structure (Shin et al., 2016) while others seem to indicate that all enhancers within an EC have an independent effect and work in an additive manner (Hay et al., 2016). Although, it is possible that EC mechanism of action is different in several scenarios encompassing a broad variety of options. These contradictory results show that further work is needed to fully understand the regulatory role of enhancer clusters.

- **Primed enhancers** are characterised by the lack of H3K27ac and being functionally inactive, although they contain H3K4me1 and may be bound by several TFs. It is widely considered as the preliminary state before enhancers become fully active as a response to stimuli (Heinz et al., 2015).

- **Repressed enhancers** are distal *cis*-regulatory regions actively repressed by the PcG complex, which deposits the histone modification H3K27me3. It has been suggested that repressed enhancers facilitate gene transcription repression by interacting with its promoter, bringing the PcG complex and blocking RNA Pol-II elongation (Schwartz and Pirrotta, 2013; Simon and Kingston, 2009). However, further work is required to determine the PcG silencing mechanism.

  Despite H3K27me3 and H3K27ac are mutually exclusive (Rada-Iglesias et al., 2011), it has been reported that a sub-set of enhancers actively repressed by PcG, might also display features associated with active enhancers (e.g. p300 binding, H3K4m1 or H3K122ac; Pradeepa et al., 2016). This set of enhancers are commonly known as **bivalent** or **poised enhancers** (Rada-Iglesias et al., 2011). Poised enhancers are present in a broad variety of tissues, and they seem to be especially relevant during development (Calo and Wysocka, 2013; Rada-Iglesias et al., 2011). However, it is important not to overestimate the number of poised elements for a given sample as part of the data could be affected by sample heterogeneity. Thus, loci identified as bivalent could actually be monovalent in distinct cell sub-populations (Chen and Dent, 2014).

  Although primed and poised enhancers are not functionally active, the main view in the field is that upon specific stimuli they might be rapidly activated. This hypothesis is further supported by the fact that these enhancers are preferentially bound by signal dependent factors (Heinz et al., 2015; Shlyueva et al., 2014b).

- **Enhancer activity reporter assays**

As exposed in the previous sections, enhancers are highly relevant to understand gene expression regulation. Even though the co-localisation of histone marks is a powerful tool to create genome wide maps of *cis*-regulatory elements, the presence of these marks does not always correlate with the *cis*-regulatory activity of a specific locus. This is because genome wide maps are subjected to certain limitation such as sample heterogeneity, missing data for known or unknown histone marks and technical aspects such as arbitrary thresholds. Additionally, human genome-wide maps are frequently generated using a small collection of

samples from healthy individuals, therefore they cannot be used to assess the effect of genomic variants in gene regulation. For that reason, experimental evidences are required to support genome wide-maps of *cis*-regulatory regions and interrogate the effect of non-coding variants.

Enhancer activity can be directly determined by *in vivo* and *in vitro* reporter assays.

Enhancer activity *in vivo* reporter assays consist on transgenic constructs in which a reporter gene is cloned downstream of a minimal promoter. As the promoter barely has transcriptional activity, the reporter signal is mainly driven by the enhancer activity of the interrogated DNA fragment. *In vivo* assays are especially powerful to detect tissue specific enhancer activity patterns in an animal model of interest. However, these assays are also subject to the limitations and inconveniences associated with animal model assays; such as cost, necessity of ethical approval or difficulty to extrapolate results to human. Moreover, these assays tend to be semi-quantitative. Enhancer activity transgenic reporter assays are extensively reviewed in Kvon, 2015.

On the other hand, **enhancer activity *in vitro* reporter assays** frequently are more quantitative and cost-effective than *in vivo* assays. One of the most frequently used reporter assays is the **Luciferase reporter assay.** It is frequently employed due to its easy performance. As in *in vivo* assays, the reporter gene is cloned downstream of a minimal promoter, thus its expression is principle driven by the enhancer activity of the assessed DNA fragment. The main disadvantage of *in vitro* luciferase reporter assays is its difficulty to be scaled up, as each interrogated DNA fragment needs to be tested individually.

To overcome this limitation, during the last years some labs have developed enhancer activity reporter assays based on high-throughput detection methods. These methods are encompassed within the term **"massive parallel reporter assays" (MPRAs)**, recently reviewed in Inoue & Ahituv, 2015; Shlyueva, Stampfel, et al., 2014. MPRAs have been used during the last few years not only for massive validation of enhancers (Arnold et al., 2013; Kheradpour et al., 2013) but also to assess how enhancers may have a different behaviour depending on metabolic context (Shlyueva et al., 2014b) or sequence variation (Tewhey et

al., 2016; Ulirsch et al., 2016). Among all MPRAs, there are two methods based on NGS that are becoming predominant, named STARR-seq (Arnold et al., 2013; Vanhille et al., 2015) and CRE-seq (Inoue et al., 2016; Kheradpour et al., 2013; Shen et al., 2016) (Fig. 5 and Table 2). The main technical differences between them are: (i) the relative position of the enhancer to the promoter, (ii) the readout, (iii) the promoter type and (iv) the size of the interrogated genomic fragment.

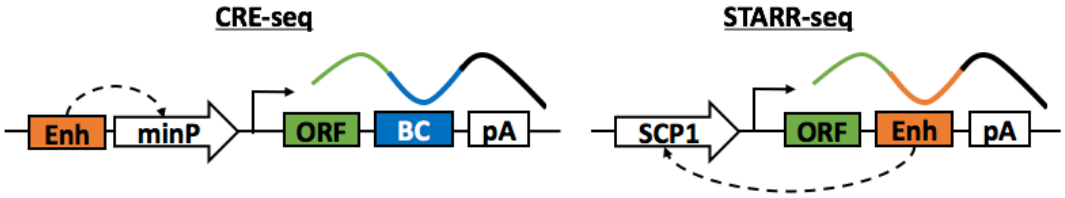


Fig. 5: CRE-seq and STARR-seq reporter constructs. Diagram illustrating the different elements and relative positions on two MPRAs systems, CRE-seq and STARR-seq. (left) In CRE-seq an enhancer (Enh) is cloned in front of a minimal promoter (minP) that drives the expression of an Open Reading Frame (ORF), frequently a fluorescent protein, and a barcode (BC) with a polyA (pA) signal to increase mRNA stability. (right) On STARR-seq, a synthetic core promoter (SCP1) drives the expression of a sgGFP, that acts as ORF, and enhancers.

Table 2: Comparative between CRE-seq and STARR-seq.

	CRE-seq	STARR-seq
Promoter	Minimal promoter	Synthetic core promoter
Enhancer position	Upstream promoter	Downstream promoter
Enhancer size	600 – few kb.	(approx.) 200 bp.
Read-out	Barcode (BC) sequence	Enhancer sequence
Scale	hundreds	thousands

In the CRE-seq assay, the transcription of a barcode sequence located downstream of enhancer and a minimal promoter is used as a readout of the enhancer activity. Thus, the main limitation is the number of barcode sequences used during the library preparation. The current number of barcodes that can be synthesised is subject to the desired length and sequence complexity. In that sense, library preparation is also a limitation. In some CRE-seq variants, enhancers and barcodes are synthesised together limiting the enhancer size to 100-200 bp (Inoue et al., 2016). In others, a sub-set of enhancers are selected from fragmented genomic DNA using complementary RNA probes (Shen et al., 2016). In that case, enhancers and promoters are cloned as a pool. Therefore, to ensure that a sub-set of

barcodes are uniquely associated with an enhancer, it is important to use a collection of barcodes several times larger than the number of interrogated enhancers.

In the **STARR-seq** method the enhancer is self-transcribed as a result of its own activity. This strategy provides more flexibility, increasing the number of enhancers that can be assessed and simplifying the cloning strategy (Arnold et al., 2013). Nevertheless, this method also has its drawbacks. It can only interrogate small DNA fragments (~200bp), as their size could affect transcript stability. Moreover, the STARR-seq construct contains a synthetic core promoter instead of the minimal promoter normally used in enhancer reporter assays. A minimal promoter is frequently preferred since its transcriptional activity per se is really low, while this synthetic core promoter exhibits higher rate of transcription that could disturb the assessment of the enhancer of interest. However, researchers from Alexander Stark's lab opted for this synthetic core promoter so that the same construct could also be used to assess repressive activity.

A general drawback of most in vitro reporter assays is that due to their episomal nature they are not able to reflect the chromosomal context of the enhancers. Although some researchers have tried to overcome this problem using integrative viral constructs (Inoue et al., 2016) it still does not represent the actual enhancer chromosomal context.

Fortunately, due to the recent development of highly efficient genome editing tools, such as **CRISPR–Cas9** (Mali et al., 2013) or CRISPR-Cpf1 (Zetsche et al., 2015), some work has been done to interrogate enhancers in their native chromosomal context. CRISPR-Cas9 based strategies can be used to mutate (Rajagopal et al., 2016; Tewhey et al., 2016), activate (CRISPRa) (Hilton et al., 2015) or repress (CRIPRi) (Qi et al., 2013) a specific enhancer as reviewed in Lopes, Korkmaz, & Agami, 2016. Moreover, this method has been adapted to perform genetic screenings interrogating several enhancers in a single experiment (Diao et al., 2016; Fulco et al., 2016; Korkmaz et al., 2016). However, to the best of my knowledge, it has not been design a CRISP-Cas9 based strategy that allows a systematic interrogation of single nucleotide variants (SNV) and their effect on enhancer activity.

- **Genome-wide maps of gene transcription regulatory regions**

In 2001, as part of the **human genome project,** the whole human genome was sequenced. This study revealed that despite the large number of protein-coding genes codified in the genome (~20,000), these only occupied a small part of it (1-2%). It opened a big debate regarding the functionality of the **non-coding DNA** regions that some researchers named "junk DNA" (Chi, 2016; Crow, 2016). It was argued that these regions must have a function, maybe as a protection for transposable elements or a role in evolution. But, it was not until 2012 with the Encyclopedia of DNA Elements (ENCODE) project when it was revealed that a large proportion of the non-coding genomic DNA has a function.

The **ENCODE project** aimed to identify all chromatin regulatory elements by generating more than 1,500 genomic datasets in almost 150 cell types (Dunham et al., 2012). To generate that large number of datasets, members of the ENCODE project took advantage of the different high-throughput methods available at the time. A big proportion of those datasets were generated using chromatin immunoprecipitation followed by (next-generation) sequencing (ChIP-seq). **ChIP-seq** consists on: fixation of protein-DNA interactions using a crosslinking reagent such as formaldehyde, selection of specific fragmented genomic regions by using an antibody that recognise a protein of interest (e.g. a TF or a histone with a specific post-translational modification) and finally, determination of the genomic positions that are frequently bound by the protein of interest via high-throughput next-generation sequencing. In addition to protein bound genomic regions, the ENCODE project also characterised chromatin accessibility (by FAIRE-seq, DNAse-seq), DNA methylation (by RRBS), gene transcription (by RNA-seq, CAGE, RNA-PET) and chromatin interacting regions (by 5C and ChIA-PET). The overlap of these datasets allowed the segmentation of the genome in cell-specific chromatin states, presumably with different functional properties.

In addition, the NIH Roadmap Epigenomics Consortium generated 111 epigenome maps encompassing a broad collection of human tissues and cell lines (Kundaje et al., 2015). An integrative analysis revealed that: (i) combinations of histone modifications correlate better with gene expression patterns than if these marks are interrogated separately. It also

showed that, (ii) enhancer activity patterns across tissues are concordant with the gene expression patterns, suggesting that these epigenome maps reflect cis-regulatory circuits. A broad collection of human epigenomic maps like this is especially interesting as some of the tissues are interrogated in their embryonic and adult stage, therefore it is a useful resource to understand epigenomic dynamics during tissue differentiation. Moreover, as it covers a broad collection of human tissues, it can be used to interrogate overlapping tissue regulatory circuits and to study tissue-specific transcriptional regulatory elements. Finally, it provides a powerful resource to elucidate the effect of non-coding disease associated variants.

These epigenomic regulatory maps show that chromatin regulatory elements are spread in different points of the linear genomic sequence, sometimes several kilobases (kb) away one from the other. However, despite their genomic linear distance, regulatory elements cluster in the 3D space allowing the synergy between them to modulate gene transcription. Therefore, although epigenome maps are powerful resources to interrogate cell-specific gene expression and the effect of non-coding variants by identifying putative *cis*-regulatory regions, they cannot be accurately interpreted without characterising the 3D chromatin organisation in the same tissue (Bonev Boyan and Cavalli Giacomo, 2016).

1.5. Compartmentalisation of gene regulation

Since 1903, when Ramon y Cajal reported sub-nuclear structures named coiled bodies or Cajal bodies (Gall, 2003), researchers have been studying the chromatin nuclear organisation. However, after more than a century, the scientific community still wonders about the tremendous organisational challenge presented by nuclear DNA packing.

Studies based on light microscopy, combined with hybridisation techniques or electron microscopy, revealed that chromatin is structured in compartments that occupy defined territories within the nucleus (Geyer et al., 2011) (see Fig. 7 in section 1.5). Different types of territories or compartments have different features with an impact over gene expression. An evidence of how 3D chromatin organisation affects gene expression is the presence of **transcription factories** (Iborra et al., 1996a, 1996b; Sutherland and Bickmore, 2009), loci with a focal accumulation of the transcriptional machinery on transcribing genes. It has been determined that although the presence of these structures is a common feature among cell types, the number of them vary from hundreds to thousands. Based on experimental evidences, a model has been proposed in which the transcriptional machinery is fixed in different nuclear loci at pre-assembled transcription factories where genes move to be transcribed (Iborra et al., 1996a). Moreover, it has been suggested that the function behind this 3D organisation in transcription factories may be the enhancement of transcription efficacy and facilitate gene co-regulation (Sutherland and Bickmore, 2009). Thus, despite the numerous questions still to be addressed, it seems clear that 3D chromatin organisation may be an additional layer of gene expression regulation that needs to be studied.

During the last decade, knowledge on chromatin structures and DNA packing has expanded rapidly thanks to the development of **chromosomal conformation techniques.**

- **Methods to study 3D chromatin organisation**

Chromatin organisation can be studied through two types of methods: (i) Microscopy-based technologies using fluorescence probes and (ii) NGS-based methods that capture chromatin conformation through DNA ligation products.

- **Fluorescence *in situ* conformational method**

DNA fluorescent *in situ* hybridisation (FISH) has allowed to visually interrogate chromatin interactions thanks to the co-localisation of two sets of fluorescent DNA probes targeting two different loci located in linear distant space. As each set of probes is label with a different fluorochrome, it is relatively easy to determine whether the two loci interact in the same focal space as the resultant colour would be different. During the last years, a few variants of this technique have been developed to increase its resolution like cryo-FISH (Branco et al., 2008), or to be performed in 3D-preserved nuclei (3D-FISH) (Cremer et al., 2008).

- **Chromosome conformation capture methods**

All chromosome conformation capture methods are founded in the same principle. This principle is the formation of DNA ligation products that contain the sequence of two interacting loci, which can be quantified. The process consists on formaldehyde cross-link a sample of interest in order to fix topological chromatin contacts. Chromatin is isolated and digested with a restriction enzyme, creating pairs of interacting chromatin fragments linked by formaldehyde covalent bonds. Free edges of two chromatin fragments are ligated into a piece and then crosslink reversed. As a result, the DNA sequence of two interacting genomic loci is contained in the same DNA molecule. The abundance of a sequence specific DNA molecule directly correlates to the interaction frequency of the two ligated regions.

Different topological techniques have been developed based on this principle in order to provide a broad spectrum of options in terms of resolution, coverage and cost. Moreover, those variations might answer different scientific questions (Davies et al., 2017; Dekker et al., 2013; Wit and Laat, 2012).

Chromosome conformation capture (3C). In 2002, Dekker et al. described a conformation technique that was used to demonstrate that chromatin interaction between regulatory elements and target genes exist *in vivo* (Dekker, 2002). At the same time, this technique established the principle previously mentioned in which most chromatin conformation techniques are based on. The technique was named chromosome conformation capture (3C), in which a specific chromatin interaction is detected and quantified by PCR. Primers are designed to map two loci of interest and pairwise interaction frequencies are computed by comparing the application efficiency of different primer pairs.

This technique allowed the experimental validation of key statements of current molecular biology such as that chromatin loops are driven and stabilised by TFs (Drissen, 2004) or CTCF (Splinter, 2006; Zhao et al., 2006).

Circularised chromosome conformation capture (4C). Few years after 3C was developed; it was combined with other quantification methods rather than PCR, such as microarrays (Simonis et al., 2006) or NGS (Splinter and de Laat, 2011), in order to interrogate a larger number of possible target loci interacting with a specific locus of interest. This technique was originally called chromosome conformation capture on-chip (4C) and its known as a "one vs all" strategy in contraposition to the "one vs one" strategy carried out by 3C.

Chromosome conformation capture carbon copy (5C). In order to increase the interrogated space, Dostie and colleagues reported a technique that follows a "many vs many" strategy and is called chromosome conformation capture carbon copy (5C) (Dostie et al., 2006). In 5C, 3C templates are hybridised with a collection of oligonucleotides that map a set of loci of interest. Pairs of nucleotides mapping two interacting loci are located in enough close proximity to be ligated. Taking advantage of a universal sequence present in the oligonucleotides, ligation products are used as template in a multiplex PCR. As in 4C, amplified ligation products are quantified by microarrays or NGS. Although it allows to interrogate many interactions in a single experiment, 5C presents a lower resolution than other chromatin conformation techniques as there can be loci to which no primers can be designed (Wit and Laat, 2012).

High-throughput conformation capture (Hi-C). In 2009, Lieberman-Aiden et al. developed a conformation technique that has an "all vs all" strategy and led to the discovery of TADs (Dixon et al., 2012; Nora et al., 2012) named Hi-C (Lieberman-Aiden and Berkum, 2009). This technique has a higher efficiency in comparison to the previously described 3C based techniques. This is achieved by the labelling with biotin of the DNA fragments resultants from a ligation, which can be selected by biotin pull-down before the DNA amplification. The biotin selected DNA fragments are used as starting material for a NGS library preparation. The resultant Hi-C data is used to generate a genome-wide matrix of interaction frequencies.

One major limitation of standard Hi-C is its resolution. Due to library complexity and cost, it is difficult to go lower than 10-20 kb of resolution (Wit and Laat, 2012). A Hi-C variant, named ***in-situ* Hi-C,** is capable to increase the resolution to 1-5 kb (Rao et al., 2014). This increase in resolution is achieved performing the DNA restriction and ligation in intact nuclei.

Chromatin interaction analysis by paired-end tag (ChIA-PET). ChIA-PET is the result of pairing chromatin immunoprecipitation (ChIP) with 3C. It allows the interrogation of chromatin interactions between loci bound by a protein of interest (Fullwood et al., 2009). In this technique, after the digestion and re-ligation step and before the NGS library preparation, a sub-set of ligation chromatin fragments are selected by pull-down using an antibody. This technique allowed the scientific community to give insight into how key chromatin structure proteins such as CTCF or Cohesin may act (Dowen et al., 2014; Ji et al., 2016; Tang et al., 2015).

Recently, a new method named **HiChIP** has appeared as an alternative to ChIA-PET (Mumbach et al., 2016). Following the same reasoning as *in-situ* HiC, sample fixation is performed in intact nucleus to reduce the signal-to-noise ratio and to improve chromatin interaction capture efficiency. These methodological differences allowed HiChIP a 10-fold increase in coverage and at the same time it required 100 times less starting material than ChIA-PET.

Promoter capture Hi-C (pcHi-C). Concerned that Hi-C methods may not be able to reflect all complexity involved in gene expression regulation, as consequence of their limitations on spatial resolution; in 2015 Peter Fraser's lab reported a new variant of Hi-C named **promoter capture Hi-C (pcHi-C)**. This variant achieves single HindIII restricted DNA fragment resolution by reducing NGS library complexity, allowing a higher coverage in a cost-efficient manner. As it happens with ChIA-PET, this reduction on library complexity is obtained by selecting a sub-set of chromatin interactions. In order to select a highly informative sub-set of interactions, Fraser and colleagues used a collection of RNA probes against almost 22,000 annotated promoters, which were used to pull-down promoter centred interactions before NGS library preparation (Mifsud et al., 2015) (Fig. 6). Thus, pcHi-C maps contain almost all gene regulatory circuitries in a specific cellular context, helping to understand how regulatory regions modulate gene expression and the possible effect of non-coding genomic variants (Javierre et al., 2016; Mifsud et al., 2015; Schoenfelder et al., 2015).

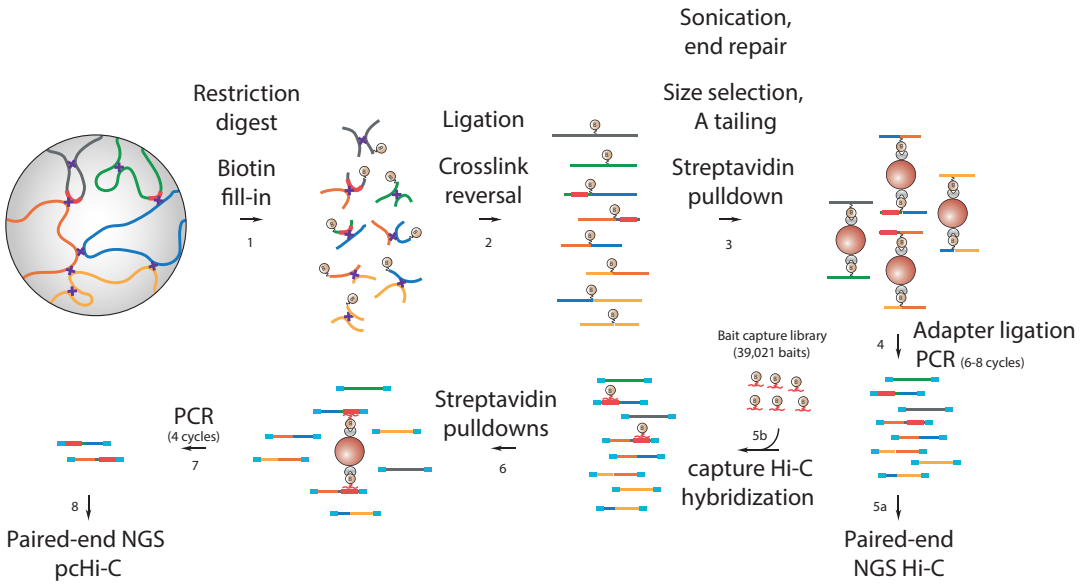


Fig. 6: Promoter capture Hi-C experimental design. The sample is fixed with formaldehyde, digested with HindIII and re-ligated to generate a Hi-C library (steps 1-4). The Hi-C library is hybridised using biotin-RNA "baits" against annotated promoters (step 5b). Chromatin interactions involving annotated promoters are selected by streptavidin pulldown, generating a pcHi-C library that is analysed by next-generation sequencing (NGS). Reprinted by permission from Genome Research, Schoenfelder et al., 2015; copyright (2015).

Following a similar reasoning and design as pcHi-C, there are 2 other methods that reduce library complexity using collections of probes against target regions named **Capture-C**

(Davies et al., 2015) and **HiCap** (Sahlén et al., 2015). These methods differ in which Capture-C uses DNA probes, theoretically with a higher efficiency than RNA probes; while HiCap implements a 4-cutter enzyme instead of a 6-cutter enzyme, that should give a higher resolution. However, to the best of my knowledge there is no published comparison between the 3 techniques to determine the benefits of subtle changes from the pcHi-C.

Genome architecture mapping (GAM). Despite 3C based methods are a powerful tool to study chromatin conformation, it is important to keep in mind their limitations due to technical aspects; such as bias due restriction sites density or limitations to quantify simultaneous contacts from >2 loci.

Recently, Ana Pombo and colleagues have combined ultrathin cryosectioning with laser microdissection and DNA sequencing to study chromatin conformation without the limitations associated to 3C methods. This method has been named **genome architecture mapping** (GAM) (Beagrie et al., 2017). Thus, genome wide chromatin contact frequencies are inferred by determining the presence of any genomic loci in a set of nuclei cryosections through NGS. Hence, loci in close 3D proximity are frequently detected in the same cryosection despite their linear genomic distance.

- **Chromatin compartments**

As previously mentioned, the knowledge on nucleus organisation has been extensively expanded during the last decade. The following sections summarise the current knowledge on **chromatin compartments** (Dekker et al., 2013) (Fig. 7) and **chromatin interacting factors**.

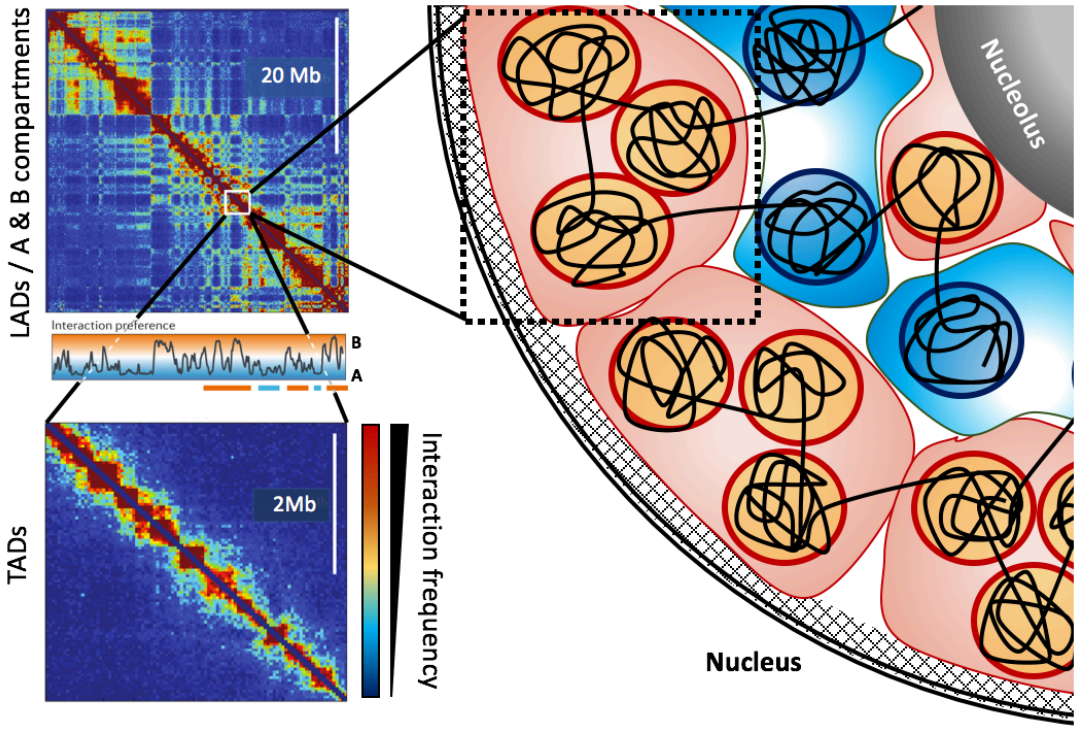


Fig. 7: Nuclear chromatin compartmentalisation. Diagram exemplifying how chromatin can be compartmentalised in LADs, A (blue) and B(red) compartments and TADs. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics, Job Dekker et al., 2013; copyright (2013).

The nucleolus and nucleolus-associated chromatin domains (NADs). The ribosome is a highly relevant protein complex required to translate the RNA to peptides, an essential process for cell functionality. An average growing mammalian cell contains several millions of ribosomes that must be synthesised in each cell division (Cooper and Hausman, 2007). Therefore, it is not surprising that to fulfil this high demand evolution has led to several strategies; such as the presence of hundreds of copies of the ribosomal genes, the combined work of two RNA polymerases (RNA pol I and III) or the formation of a specialised chromatin compartment named **nucleolus**.

This nuclear compartment can be observed by electron microscopy (Pederson, 2011) or light microscopy, due to its peripheral heterochromatin enrichment. The nucleolus is the 3D chromatin space where nucleolus organising regions (NORS) cluster. NORS are genomic regions that codify for several tandem copies of the ribosomal genes. So, the nucleolus is the nuclear spherical compartment where ribosomal DNA transcription (rDNA), ribosomal RNA (rRNA) processing and ribosome biogenesis occur.

Recent studies (van Koningsbruggen et al., 2010; Németh et al., 2010), in which DNA sequences bound to the nucleolus were characterised by NGS, reported the presence of **nucleolus-associated chromatin domains (NADs)**. These studies showed that NADs are not only enriched in rDNA but also centromere or telomere loci with a high density of repetitive sequences, gene deserts and heterochromatin. Thus, these studies proved that there is a specific 3D chromatin organisation around the nucleolus periphery with a likely effect on gene expression regulation (Padeken and Heun, 2014; Pombo and Dillon, 2015).

Lamin-associated domains (LADs). It has been observed, that as it happens around the nucleolus, there is an enrichment of highly condensed chromatin contacting the Lamin meshwork located at the internal side of the nuclear envelope. Thus, using a technique to determine interactions between DNA fragments and nuclear Lamin, named Lamin-DamID, researchers have defined **Lamin-associated domains (LADs)** (Guelen et al., 2008).

Studies done by Susan Gasser and colleagues showed the key role of LADs on gene repression during differentiation (Gruenbaum and Foisner, 2015; Mattout et al., 2015; Towbin et al., 2010) and B. van Steensel et al. reported that LADs can be divided in two sub-categories: (i) constitutive (cLADs) which are conserved among tissues and species, and (ii) facultative (fLADs) that are tissue-specific (Meuleman et al., 2013; Pombo and Dillon, 2015). Therefore, LADs are genomic regions with length size from 10kb to 100Mb located at nuclear periphery, frequently delimited by CTCF sites, that are involved in gene expression regulation through the PcG repression and heterochromatin formation.

A and B compartments. Based on Hi-C studies, Dekker's lab confirmed the presence of chromosome territories and described the segmentation of the nucleus in several mega-base compartments. This study reported that there are two types of compartments, A and B; and that chromatin interactions tend to occur between regions located in the same type of compartment and depleted between compartment A and B. Lieberman-Aiden et al. also reported that the A compartment is more frequently associated with active transcribed regions containing active epigenomic marks than compartment B. Moreover, although those compartments seem to be quite homogeneous between cell types (Lieberman-Aiden

and Berkum, 2009; Schmitt et al., 2016) discrepancies between cell types tend to be coherent with cell–specific epigenomic states and gene expression patterns.

A recent study that achieved very high resolution (Rao et al., 2014) reported that these two compartment types can be further divided in 6, A 1-2 and B 1-4. A1 and A2 compartments are still enriched in active chromatin marks and one of the main differences between them seems to be the replication timing. Regarding the B compartments, B1 is highly enriched in PcG repression, B2 and B3 are especially enriched in heterochromatic, NADs and LADs; while B4 encompassed many genes from the KRAB-ZNF family with bivalent marks.

Topological associating domains (TADs). Studying the genome 3D chromatin organisation in human and mouse cells using Hi-C, Dixon et al. observed that chromatin interactions tend to be contained within domains named **topological associating domains (TADs)** (Dixon et al., 2012). Therefore, TADs could delimitate the genomic space in which a gene promoter can communicate with distal regulatory regions.

As in LADs, the structural protein CTCF is highly enriched at TAD borders as it is present in >75% of them (Dixon et al., 2012). This indicates that CTCF is involved in TAD border formation (Dixon et al., 2012; Nora et al., 2012) although it is probably not the sole mechanism. Moreover, disruption of CTCF at TAD borders has confirmed its role as insulator, blocking inter-TAD chromatin interactions (Lupiáñez et al., 2015, 2016; Nora et al., 2012).

It has been observed that TADs borders are highly conserved between tissues (Schmitt et al., 2016) and species (Dixon et al., 2012) and that co-regulated genes tend to be contained in the same TAD (Le Dily et al., 2014; Nora et al., 2012). Additionally, several studies have revealed that intra-TAD chromatin interactions can be tissue-specific (Javierre et al., 2016; Phillips-Cremins et al., 2013; Rao et al., 2014; Schmitt et al., 2016), which may result in tissue-specific gene expression. This is coherent with the observation of a hierarchical chromatin organisation formed by TAD, sub-TADs (Phillips-Cremins et al., 2013) and meta-TADs (Fraser et al., 2015).

Meta-TAD organisation is defined by interactions between adjacent TADs and it extends to encompass the whole chromosome (Fraser et al., 2015). It has been observed that TADs within the same meta-TAD form part of the same (A or B) compartment. Moreover, meta-TAD hierarchy is also concordant with the localisation on LAD borders. As it happens with the A and B compartment transition, it has been reported that the organisation of the meta-TAD hierarchical tree changes during development correlates with changes on gene expression.

Studies that achieved higher resolution than standard Hi-C revealed the presence of **sub-TAD** structures, that can be tissue-specific or constitutive (Phillips-Cremins et al., 2013; Rao et al., 2014). Phillips-Cremins and colleagues performed high-resolution 5C in mouse neural progenitor (NPC) and mouse embryonic stem (mES) cells in seven genomic loci. In this study, Phillips-Cremins et al. identified 260 constitutive, 165 NPC and 86 mES specific interactions. A characterisation of those interactions revealed that the ones tissue-specific tend to be short (<300 kb), mediated by Cohesin and Mediator complex, and occur between enhancer and promoters, whereas constitutive interactions are longer (600bp-1kb) and mediated by CTCF and Cohesin.

Recently, a study carried out in Bing Ren's lab in which they performed Hi-C in 21 human samples, covering 14 tissues and 7 cell lines, reported the presence of tissue specific structures within TADs. Schmitt and colleagues detected the presence of frequently interacting regions, or FIREs, by identifying loci with an unexpectedly high contact frequency. FIREs are highly tissue specific, as 60% are only present in 1 or 2 of the 21 samples, and positioned near tissue specific genes. FIREs have a high overlap with annotated enhancers and especially super-enhancers. Thus, suggesting that the FIREs are the result of tissue-specific interactions, between enhancer and genes, conducted by two well-known structural proteins, Cohesin and CTCF. This study provides additional evidence on how tissue specific chromatin interactions are associated to tissue-specific gene expression, probably by allowing the co-localisation of promoters and enhancers in close 3D proximity (Schmitt et al., 2016). However, it is still limited by the low resolution obtained with standard Hi-C due to the low sequencing coverage achievable.

- **Chromatin loops and regulatory interactions**

The smallest structure in 3D chromatin organisation is the **chromatin loop**. Conceptually it is very basic but biologically it is becoming a very active and complex topic. As it could be done with a string, chromatin can be folded due to its polymeric nature; locating two loci (from the same chromosome) in closer tri-dimensional proximity than in linear proximity.

Chromatin loops can be defined as structural or regulatory (Krijger and de Laat, 2016). **Structural** chromatin loops are formed by key structural proteins, such as CTCF and Cohesin, and their role is to create the structures that sub-compartmentalise chromosomes into topological domains. As chromatin compartmentalisation is similar between cell type and species, it would not be surprising for these interactions to be highly ubiquitous. On the other hand, enhancer-promoter **regulatory** loops seem to be driven by the cooperative work of ubiquitous structural proteins and tissue-specific TFs. Moreover, these interactions frequently occur in a tissue-specific manner within tissue-invariant TAD structures. Therefore, it is reasonable to hypothesise that tissue-specific chromatin loops may be involved in tissue-specific gene expression regulation.

Many models have been proposed to determine the mechanism of chromatin loop formation, but recently the **extrusion model** is standing out from the rest. As it has been reviewed by Professor Matthias Merkenschlager and Elphège P. Nora, this model is mainly driven by CTCF and Cohesin (Sanborn et al., 2015). The chromatin folding starts with Cohesin, which acts as an extruding factor. The chromatin fibre is slid through the ring structure formed by the Cohesin complex creating a loop. This process continues until (a) the Cohesin complex dissociates, unfolding the chromatin, or (b) the loop is anchored by CTCF sites at both ends (Fig. 8) (Merkenschlager and Nora, 2016).

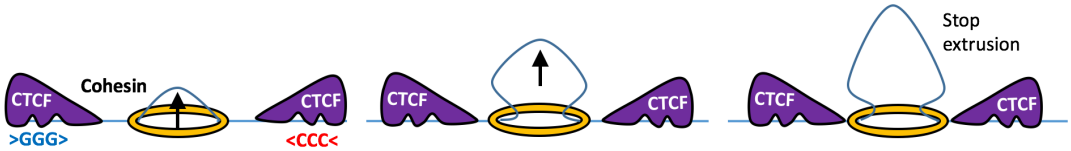


Fig. 8: Extrusion model. Diagram illustrating the main concepts of the extrusion model in which the chromatin fibres slid through the Cohesin complex, forming a loop. This process continues until the loop is stabilised by CTCF that acts as anchor points at both ends of the loop. Adapted by permission

from Annual Review of Genomics and Human Genetics, Merkenschlager & Nora, 2016; copyright (2016).

This model fits with experimental observations such as the fusion of two adjacent topological domains when a CTCF boundary is lost (Nora et al., 2012). However, there are questions that remain unanswered: (i) Which is the initiation mechanism of the extrusion loop formation? (ii) What drives it? (iii) Despite CTCF motifs in convergent orientation being preferentially used, which could be the read-out mechanism? (Merkenschlager and Nora, 2016).

An aspect of this model that needs to be considered is that it is based on ubiquitously expressed proteins which have ubiquitous genomic localisation among cell types (Cuddapah et al., 2008; Kim et al., 2008; Ong and Corces, 2014). This aspect is concordant with the observation that tri-dimensional chromatin conformation can be surprisingly similar among cell types (Schmitt et al., 2016). However, this model does not contemplate the role of other key actors, such as the Mediator complex or non-coding RNAs, which may have an important task on tissue-specific chromatin loop formation.

- **Role of interacting factors in 3D chromatin organisation**

Chromatin is bent in the 3D space thanks to a broad collection of **chromatin interacting factors,** such as CTCF, Mediator or non-coding RNAs; which are reviewed in this section.

Lamin. Lamin is a family of filament proteins that form a meshwork in the internal face of the nuclear envelope. These scaffold-like proteins have been associated with heterochromatin formation and localisation within the nucleus, tissue-specific gene expression and 3D chromatin organisation (Burke and Stewart, 2013; Gruenbaum and Foisner, 2015; Mattout et al., 2015).

CTCF. The CCCTC-binding factor or CTCF is one of the most studied chromatin structural protein in invertebrates. This factor is ubiquitously expressed and 40-70% of its chromatin binding sites are conserved among cell types (Cuddapah et al., 2008; Kim et al., 2008; Ong and Corces, 2014). Although originally it was described as a barrier or enhancer blocking

insulator (Bell et al., 1999; Kellum and Schedl, 1991); more recent studies suggest that, in fact, CTCF may act as a **chromatin looping facilitator** (Ong and Corces, 2014)**.** This is clearly exemplified in a work done by Felsenfeld and colleagues, in which they interrogated the 3D chromatin organisation around the insulin locus in human pancreatic islets. In this study the authors discovered a chromatin interaction between the insulin promoter and STY8, an important gene for insulin secretion. The authors also showed that the transcription of these two genes was regulated by their chromatin interaction; communication that was perturbed by CTCF depletion (Xu et al., 2011). Thus, this and other studies have provided clear evidences that CTCF is required for at least a sub-set of chromatin interactions (Phillips-Cremins and Corces, 2013).

It is also known that CTCF binds to a relatively long (~20bp) and well characterised consensus sequence (Kim et al., 2008). Although the precise mechanism is unknown, it has been proposed that two CTCF molecules are able to interact forming a homodimer, preferentially between sequences with convergent orientation (Merkenschlager and Nora, 2016; Tang et al., 2015; de Wit et al., 2015). Additionally, the CTCF consensus sequence contains a CpG dinucleotide, which can be differentially methylated affecting its binding and consequently its activity as structural protein (Engel et al., 2004).

Despite DNA methylation being able to modulate CTCF activity, less than 50% of the cell-type-specific CTCF binding sites seem to be DNA methylation dependent (Wang et al., 2012), proving that other factors must be involved in CTCF binding regulation. Among all the proteins interrogated on regulating CTCF activity only Cohesin seems to be essential for CTCF function. Therefore, it is not unexpected that more than 50% of CTCF binding sites are also co-occupied by it (Ong and Corces, 2014; Phillips-Cremins and Corces, 2013; Rubio et al., 2008; Wendt et al., 2008).

Cohesin. Cohesin is a protein complex that forms a ring-like structure, which was initially studied by its role in sister chromatin cohesion during cell division (Gause et al., 2008; Nasmyth and Haering, 2009). Recently it has been described that this structural protein complex may have an important role in chromatin loop formation and stabilization by its cooperative role with CTCF (Merkenschlager and Nora, 2016). Moreover, it has been

reported that the Cohesin complex, jointly with Mediator, is involved in tissue-specific gene expression through the formation of enhancer-promoter loops (Kagey et al., 2010).

Condensin. A recent publication showed that Condensin could be as important as other structural proteins such as Cohesin for (yeast) topological organisation (Kim et al., 2016). It was already known that Condensin frequently co-localises with Cohesin and it is involved in chromatin segmentation, but this study showed that their role may occur at different scales. The authors of this study observed that in *S. cerevisiae*, Condensin is involved in long range interactions creating domains with a 300 kb median size. In contrast, Cohesin is involved in shorter interactions creating domains with a 70 kb median sizes and contained within the Condensin mediated domains. Moreover, Condensin mediated boundaries frequently interact with centromeric regions, suggesting that this organisation may be especially relevant during mitosis.

Mediator. Mediator complex facilitates communication between enhancer and promoters, probably due its large volume. Mediator is used as a scaffold to accommodate the transcriptional machinery at the promoter and at the same time it interacts with the TFs of multiple enhancers at different chromatin loci. Although it is not essential, Mediator is highly relevant for chromatin loop formation and stability (Allen and Taatjes, 2015). In addition to Cohesin, it has been reported that Mediator complex binding and chromatin loop formation is modulated by the interaction with a class of enhancer-like long non-coding RNAs named activating lncRNAs (Lai et al., 2013).

Non-coding RNAs. It has been observed that a large proportion of the transcribed RNA is not translated to proteins (Fantom Consortium, 2005). Non-coding RNAs (ncRNAs) is a broad term that encompasses a large compendium of transcripts with different features and functions. Non-coding RNAs are involved in gene expression regulation through a wide spectrum of mechanisms, from interaction with TFs and chromatin remodelers (Rinn et al., 2007) to regulating mRNA stability (Kretz et al., 2012). Non-coding RNA have been extensively reviewed in Rinn & Chang, 2012. However, this section will illustrate how two well-known classes of ncRNAs, lncRNA and eRNAs, can regulate gene expression through 3D chromatin organisation.

- **lncRNAs.** Long non-coding RNAs (lncRNA) are >200bp non-coding RNA low-transcribed molecules preferentially with a nuclear localisation. Currently, more than 150,000 human lncRNAs have been identified (Zhao et al., 2016) and the function of the vast majority is still unknown. However, based on their characteristics it has been proposed that some of them may have a role in gene expression regulation and nuclear architecture organisation, with already a few examples (reviewed in Fatica & Bozzoni, 2014; Rinn & Chang, 2012).

  Prof. John Rinn, proposed the "Cat's Cradling" model in which cell-type specific lncRNA facilitate the formation of chromatin compartments by acting as landscape markers for architectural proteins (such as Lamin), which pull the DNA polymeric chromatin fibres. Therefore, by changing chromatin organisation, cell-type specific lncRNA may participate on wiring cell-specific transcriptional programs (Melé and Rinn, 2016).

- **eRNAs.** eRNA are short non-coding RNAs with a short lifetime and generated as a result of transcriptional process at active enhancers. Most eRNA present a 5' cap RNA, are not spliced or poly-adenylated and come from a bi-directional transcription. However, there is a small proportion of mono-directional poly-adenylated eRNAs (Andersson et al., 2014; Djebali et al., 2012; Lam et al., 2014).

  It has been hypothesised that eRNA may have 3 possible roles: (i) "noise" generated as consequence of an enhancer-promoter interaction, (ii) by-product of a transcriptional process in the enhancer required for its activation or (iii) active element required for enhancer activity. Despite not been mutually exclusive, the latest is the most likely as recent evidence show that eRNA can have an active role in the formation and stabilisation of enhancer-promoter interactions, potentially by interacting with the Cohesin complex (as reviewed by Lam et al., 2014).

1.6. Tissue-specific transcriptional circuitries

The development of NGS techniques to measure gene transcription has allowed the characterisation of gene expression profiles in a large collection of tissues and samples. The analysis of these gene expression datasets has confirmed that cell types and tissues have characteristic gene expression patterns. Thus, it suggests that gene expression is managed by tissue-specific transcriptional circuitries, which form gene pathways that are interconnected and regulate each other at different levels (Graf and Enver, 2009; Marbach et al., 2016; Neph et al., 2012; Saint-André et al., 2016) (Fig. 9).

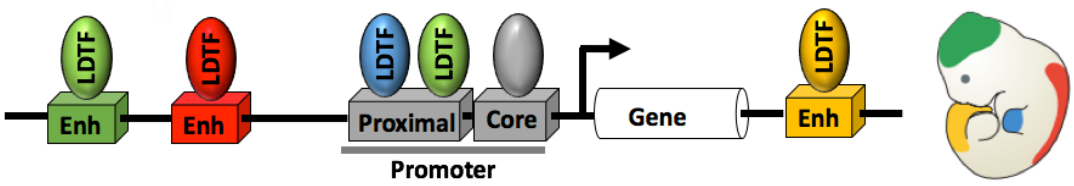


Fig. 9: Tissue-specific gene regulation. Tissue-specific gene expression is driven by the cooperative work of LDTFs and other proteins that bind to different *cis*-regulatory elements such as enhancers (enh.) or promoters. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology, Krijger & de Laat, 2016; copyright (2016).

Although complete transcriptional circuitries have not yet been described in most cell types, there are several lines of evidence (Graf and Enver, 2009; Lee and Young, 2013; Saint-André et al., 2016) that tissue-specific transcriptional circuitries are controlled by a small number of TF, frequently named master regulators or lineage-determining TFs (LDTFs). LDTFs bind and facilitate the binding of other proteins to different *cis*-regulatory elements. Among all types of *cis*-regulatory elements involved in transcriptional circuitries regulation, enhancers and specially enhancer clusters seem to have a key role modulating tissue-specific gene expression (Heinz et al., 2015). Enhancers, as most *cis*-regulatory elements, are used as recruiting platforms for different proteins that are able to module gene transcription. Additionally, it seems that there is certain degree of affinity between types of enhancer and promoters. A study using a massive parallel reporter assays (MPRA) with a collection of house-keeping and developmental gene promoters paired with multiple enhancers revealed that certain enhancers only act as such in presence of certain types of promoters (Zabidi et al., 2014). Thus, due to enhancer-promoter regulation, a gene may be expressed in different tissues and at different levels of expression.

Moreover, it seems that a subset of the active enhancers in a cell-type or tissue may have a stronger control on gene regulation than the rest. Although different labs have figured out different arbitrary definitions and names to define them (CORES in Gaulton et al., 2010, super-enhancers in Whyte et al., 2013, stretch enhancers in Parker et al., 2013, enhancer cluster in Pasquali et al., 2014) some features seem to appear repeatedly, which are: (i) close linear proximity between arrays of enhancers, (ii) located near tissue-specific expressed genes, frequently bound by (iii) LDTFs and (iv) chromatin interacting factors such as Mediator or Cohesin. Moreover, this subset of clustered enhancers is being actively characterised in different tissues not only by its role in gene regulation but also because it frequently contains genomic variants associated with diseases.

1.7. Regulatory genomics and diseases

Studies conducted to determine the genetic factors behind major human diseases observed that these can be caused by non-coding variants in genomic regulatory regions. It has been also shown that the molecular mechanism impaired and its severity may be different depending on the type of genomic variant (SNV or indel) and the affected *cis*-regulatory element (enhancer, promoter or insulator) (Table 3).

Table 3. Effect of non-coding variants on *cis*-regulatory elements. *

Effect	Mechanism	References
Modulate activity of *cis*-regulatory elements	SNPs and small indels at enhancer or promoters can create or disrupt TFBSs, modifying TFs affinity for a local DNA sequences. Therefore, TFs and co-TFs bound to a specific *cis*-regulatory element may change, affecting its activity.	Gaulton et al., 2010, 2015; Pasquali et al., 2014
Modify genes' *cis*-regulatory landscape	Indels may delete or insert *cis*-regulatory elements at a gene's vicinity altering its transcription regulation.	Van der Ploeg et al., 1980; Weedon et al., 2014; X. Zhang et al., 2015
Alter 3D chromatin local conformation	Genomic variants that affect the binding of structural proteins, such as CTCF, changing 3D chromatin organisation. If the genomic variant occurs within a topological domain, its effects will remain contained within it and creating or disrupting intra-domain interactions. However, if the genomic variant occurs at a domain boundary region, its effects can be broader merging or creating two topological domains. In any case, it modifies genes *cis*-regulatory landscape by rewiring the 3D chromatin organisation.	de Wit et al., 2015; Lupiáñez et al., 2016; Narendra et al., 2015; Nora et al., 2012

* Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology, Krijger & de Laat, 2016; copyright (2016).

- **Systematic strategy to identify causal non-coding genomic variants**

Genomic variants linked to a higher risk of developing a disease are frequently identified through genome wide association studies (**GWAS**). These are genetic studies in which the genome of two populations, one with the studied feature and a control population, are compared. Normally those studies are carried out using genotypic arrays that cover specific regions of the genome. However, due to the continuous cost reduction in NGS techniques, recent studies have been applying whole-genome sequencing (Flannick and Florez, 2016). A major limitation of these studies is the high number of non-causal variants detected due to linkage disequilibrium (LD).

Because of different factors, such as genomic recombination rates or genetic drifts, alleles of two genomic variants are in LD when their observed frequency is higher than randomly expected (Slatkin, 2008). This is frequently detected between genomic variants located in close linear proximity as the genomic information is inherited in haplotype blocks. Therefore, GWAS cannot differentiate between causal and non-causal variants in high LD. Additionally, the general trend is that GWAS variants are frequently detected in non-coding parts of the genome (Flannick and Florez, 2016; McClellan and King, 2010). Thus, it is highly challenging to hypothesise which variants are more likely to be causal without knowing their potential mechanistic effect.

In summary, lists of GWAS variants are a great resource to identify regions containing genomic signals associated with pathogenicity. However, due to their limitations, it is highly recommended to cross them with maps of regulatory regions to enrich for potential causal variants (Fig. 10).

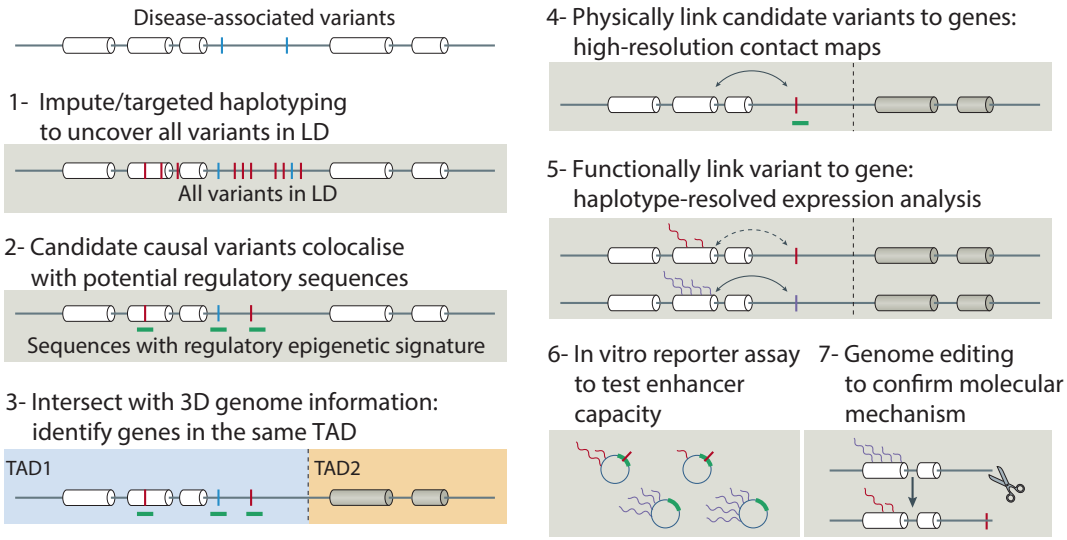


Fig. 10: Systematic identification of causal genomic variants. Diagram illustrating the different steps of systematic identification of causal variants. (1) Identification of genomic variants associated to disease. (2) Co-localisation of genomic variants with genomic regulatory elements. (3) Implement chromatin territory maps to associate non-coding variants and possible target genes, (4) Associations that can be further validated with high-resolution chromatin interaction maps. (5) Associate genomic variants and target genes interrogating haplotype-resolved expression datasets. (6) Experimentally validate non-coding variants by MPRAs and genome editing assays. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology, Krijger & de Laat, 2016; copyright (2016).

As exposed in previous sections, the genome is compartmentalised in domains that may confine the effect of regulatory regions and causal non-coding genomic variants. Therefore, maps of chromatin interactions and domains can be highly informative to associate genomic variants and target genes. If possible, this association can be further validated using expression datasets comparing samples with the risk-allele and samples with the non-risk allele.

A challenge that researchers frequently need to face is the selection of likely causal variants for experimental validation. Despite applying the previously mentioned steps, the list of likely causal variants may often contain too many to be interrogated in money and time consuming experimental assays. Thus, computational tools have been created to prioritise non-coding variants based on their probability of being disease-causal variants. These methods cover a broad variety of approaches, such as functional annotations, conservation and/or machine learning (as reviewed in Nishizaki & Boyle, 2016).

After this systematic process, a much shorter list of likely causal variants can be tested in *in vitro* or *in vivo* assays to determine the affected molecular mechanisms and their consequences.

- **Gene regulation in pancreatic islets**

As mentioned before, understanding gene expression regulation in disease-relevant tissues is especially important to decipher the perturbed molecular mechanisms. Among all human tissues, pancreatic islets are especially important to understand impaired glucose regulation associated with diabetes. Despite its relevance, this tissue is specially challenging to study as human pancreatic islets are difficult to be obtained for research purposes.

Pancreatic islets or islets of Langerhans are spheres of endocrine tissue embedded in the pancreas and surrounded by exocrine tissue. Pancreatic islets are formed by 5 cell-types: alpha, beta, delta, gamma or PP and epsilon. However, among all of them, beta-cells are especially relevant as they compose around 70% of the human islet mass and they produce insulin (Scharfmann et al., 2008) (Fig. 11).

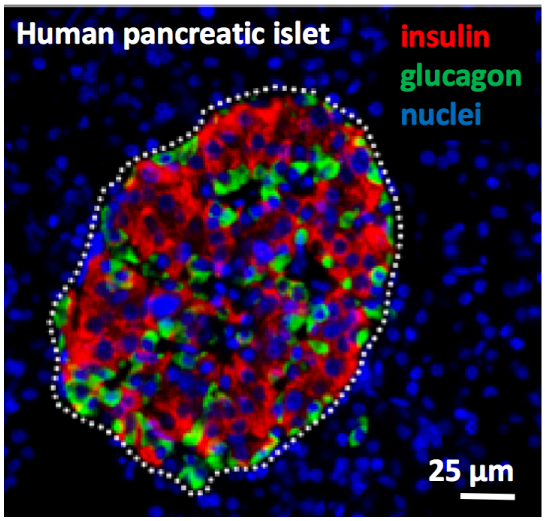


Fig. 11: Cell heterogeneity in human pancreatic islets. Section of human pancreatic islets stained for nuclear DNA (blue), insulin-producing beta-cells (red) and glucagon-producing alpha-cells (green). Reprinted by permission from PLoS ONE, Scharfmann et al., 2008; copyright (2008).

Insulin is a hormone involved in blood glucose regulation by promoting its absorption into liver, fat and skeletal muscle cells. Thus, impaired insulin production and secretion leads to accumulation of glucose in the blood stream. Constant high blood glucose levels can have severe consequences, such as kidney failure and blindness, and premature death (Wilcox, 2005). Therefore, understanding the genome regulation in pancreatic islets is essential to discover regulatory mechanisms that control insulin synthesis and secretion.

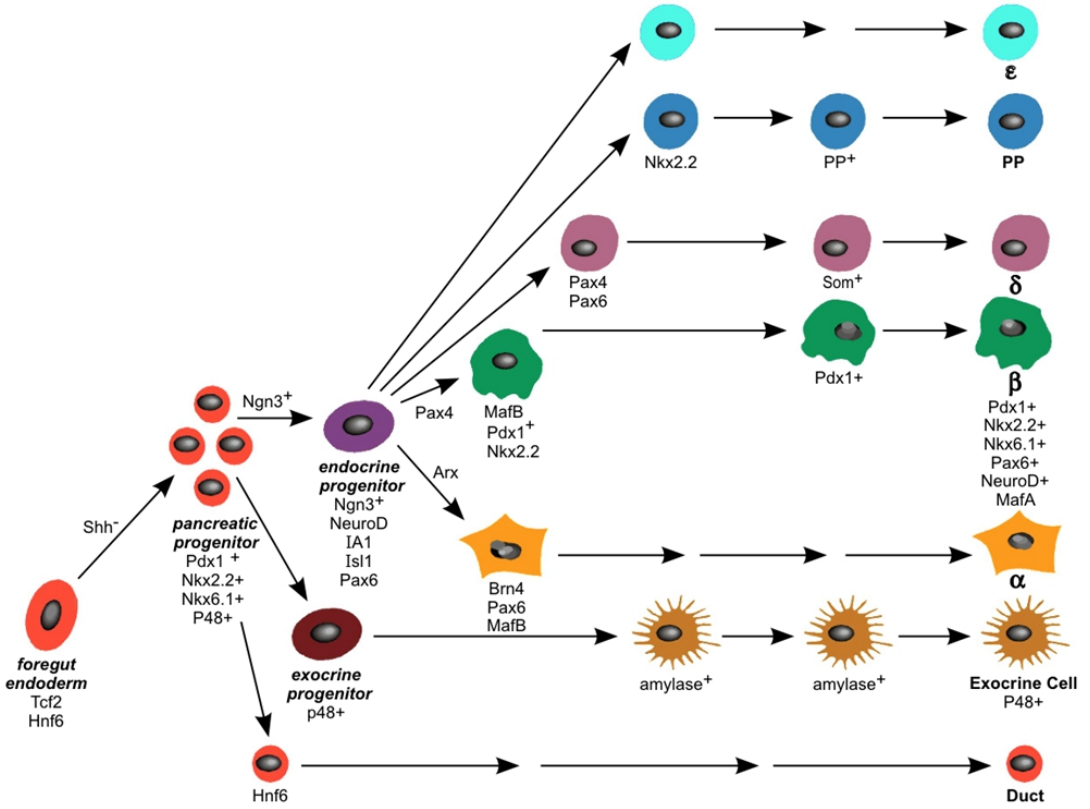


Fig. 12: Pancreatic cell differentiation pathways. Diagram summarising the different steps during cell differentiation from foregut endoderm to endocrine and exocrine pancreatic cells. Key LDTFs are specified in each step. Reprinted by permission from Stem Cells in Clinic and Research, Jiang & Morah, 2011; copyright (2011).

To better understand pancreatic islets gene regulation several groups have performed a big effort to identify key TFs for islet development and maintenance of cell identity (as recently reviewed by Romer & Sussel, 2015 and summarized in Fig. 12). This work is especially important, not just to identify LDTFs involved in tissue-specific gene expression but also to give insight into the molecular mechanisms perturbed by disease associated genomic variants. For example, a recent work carried out by Inês Cebola and colleagues revealed that genomic variants in a distal enhancer can alter expression of a key developmental TF, PTF1a

(also known as P48), being the most common cause of isolated pancreas agenesis (Weedon et al., 2014). Thus, illustrating how relevant can be the characterisation of LDTFs and transcriptional regulatory maps in combination with genetic studies of diseases.

Additionally, a transcriptional analysis done in beta-cells by I. Morán and colleagues has been able to identify more than a thousand lncRNAs. This study found that islet lncRNAs are frequently cell-specific and active during pancreatic islet development, suggesting they could be involved in islet tissue-specific gene expression (Morán et al., 2012). In fact, a more recent study showed that a subset of these beta-cell specific lncRNAs, in combination with some TFs, form a core regulatory circuitry involved in gene transcriptional regulation (Akerman et al., 2017). This is especially relevant as Morán et al. showed that some beta-cell lncRNA have a mis-regulated expression in T2D patients and are located near T2D susceptibility loci.

Therefore, this exemplifies how characterising gene expression regulation in human pancreatic islets can provide insight into the molecular mechanism perturbed in a major disease such as diabetes.

- **Role of gene regulation in Diabetes Mellitus**

Diabetes Mellitus is a group of metabolic diseases characterised by an impaired blood glucose regulation that can lead to serious health complications.

Currently, diabetes is sub-classified in 3 main categories: Type-1 Diabetes (T1D), Maturity Onset Diabetes of the Young (MODY) and Type-2 diabetes (T2D). This section will focus on T2D as it is the most frequent form of diabetes.

T1D and MODY are two forms of diabetes that follow a Mendelian inheritance. **T1D** is an autoimmune disorder characterised by a resultant reduction of the pancreatic islet mass and insulin production rates. Different forms of **MODY (1-11)** are due to loss-of-function mutations that affect key TFs (HNF1a, HNF4a, PDX1, …) or other key proteins such as Glucokinase (an enzyme involved in glycolysis), leading to impaired insulin secretion and fasting glycemia (Tallapragada et al., 2015).

T2D is a common disease, developed due to the combined effect of risk-associated genomic variants and environmental factors, that affects more that 400 million people in the word. In order to determine the genetic factors behind this pandemic, more than 50 GWAS have been conducted during the last 15 years (as reviewed in Flannick & Florez, 2016) collecting thousands of samples from different ethnic populations. This huge effort from the scientific community has allowed the identification of hundreds of common and rare (MAF < 5%) genomic variants associated with a higher probability of developing the disease, most of them in non-coding regions of the genome. However, as previously mentioned in section 1.7, due to different factors, it is reasonable to hypothesise that not all genomic variants identified in GWAS are causal variants. Therefore, it is important to cross these lists of risk-associated variants with transcriptional and epigenomic maps, to enrich them for likely causal genomic variants. Thereafter, a further experimental validation would be required to validate them and give insight into the impaired molecular mechanisms.

To fill the gap between the cataloguing of risk-associated variants and identification of causal variants with an effect on gene regulation, the group of prof. Jorge Ferrer generated a map of regulatory regions in human pancreatic islets based on epigenomic marks. This study observed that enhancers are not evenly distributed through the genome, but a sub-set of them are located in close proximity forming *cis*-regulatory regions known as enhancer clusters (EC). This study also showed that EC are enriched on type-2 diabetes (T2D) or impaired fasting glycemia (FG) associated SNPs. Pasquali et. al. also provided evidences that EC highly bound by key islet TFs are frequently located near islet specific expressed genes. These results indicate that EC are important for islet-specific gene expression and that disruption of their functionality by non-coding genomic variants is likely to increase the risk of developing T2D or impaired FG (Pasquali et al., 2014). However, at that time, 3D chromatin organisation data in human pancreatic islets was not available, therefore all associations between non-coding elements and genes were based on linear proximity with the limitations and inaccuracy that this implies.

Chapter 2

Rationale, hypotheses and aims

Linear *cis*-regulatory maps do not reflect gene transcription regulation through 3D chromatin organisation

Cis-regulatory elements, such as enhancers and promoters, are key for gene transcription regulation. During the last years, big efforts have been made to create *cis*-regulatory maps in several cell lines and tissues (Dunham et al., 2012; Kundaje et al., 2015). These maps have been able to identify thousands of *cis*-regulatory regions in the linear chromatin space, being a great source for better understanding gene transcription regulation. These *cis*-regulatory maps are especially important in disease-relevant tissues, such as pancreatic islets, since they could help to unveil the genetic factors underlying severe diseases like diabetes (Krijger and de Laat, 2016; Pasquali et al., 2014). However, such linear maps do not inform on how distal regulatory elements communicate with their targets. It is known that distal *cis*-regulatory elements and gene promoters establish proximity interactions in 3D space, thus modulating gene transcription (Bonev Boyan and Cavalli Giacomo, 2016).

Researchers have used several approaches to overcome the lack of interaction maps that relate regulatory elements and their target genes. For example, enhancers are frequently assigned to the closest (active) gene assuming this would be its most likely target. However, those approximations do not ensure accurate associations. Distal *cis*-regulatory elements gain proximity with target genes through structural proteins that bend the chromatin in 3D space. Thus, 3D chromatin conformation allows distal *cis*-regulatory elements not only to interact with multiple loci but also to skip untargeted genes (Sanyal et al., 2012). Therefore, *cis*-regulatory elements cannot be precisely associated to target genes relying exclusively on linear chromatin maps.

In the current thesis, I propose the following hypotheses:

A. Integration of chromatin interaction and epigenomic *cis*-regulatory maps can provide novel insight into tissue-specific gene regulation

Despite the big efforts to generate chromatin interaction maps in several human cell lines and tissues during the last decade (Dixon et al., 2012; Javierre et al., 2016; Nora et al., 2012; Schmitt et al., 2016), genome-wide chromatin conformation data in human pancreatic islets is still missing. I propose that 3D chromatin interaction maps in human pancreatic islets are necessary to accurately interpret gene transcription regulation in this disease-relevant tissue. Moreover, I hypothesise that the characterisation of 3D chromatin interaction maps in combination with epigenomic datasets (Pasquali et al., 2014) will allow me to make precise associations between *cis*-regulatory elements and target genes.

It has been shown that the 3D chromatin organisation leads to a compartmentalisation of genes and *cis*-regulatory elements in what is known as topological associating domains (TADs) (Dixon et al., 2012, 2015; Nora et al., 2012). Achieving higher resolution than standard high-throughput chromatin conformation assays allowed the identification of interactions within TADs. Tissue-specific intra-TAD interactions were reported to correlate with tissue-specific gene transcription (Javierre et al., 2016; Phillips-Cremins et al., 2013; Rao et al., 2014; Schmitt et al., 2016). However, the functional significance and the epigenomic factors underlying these tissue-specific chromatin structures remains poorly understood.

These considerations led me to propose that high-resolution techniques, such as promoter capture Hi-C (pcHi-C) (Mifsud et al., 2015), could be used to define chromatin domains and tissue-specific chromatin interactions in human pancreatic islets.

Furthermore, I hypothesise that a systematic analysis of human pancreatic islet-selective chromatin structures and their associated epigenomic features would give insight into epigenomic factors behind islet-specific gene expression.

B. High-resolution chromatin interaction maps can define 3D enhancer domains

Among the collection of active enhancers present in a specific tissue, there are many examples where several enhancers are located in close linear proximity. This feature has been used in definitions with arbitrary thresholds to identify stretch enhancers (Parker et al., 2013), enhancer clusters (Pasquali et al., 2014) or super-enhancers (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). These elements are emerging as a highly relevant type of *cis*-regulatory regions as they have been linked to tissue-specific gene expression and disease susceptibility (Hnisz et al., 2013; Lovén et al., 2013; Pasquali et al., 2014; Whyte et al., 2013). However, how 3D chromatin organisation could influence the definition and function of these *cis*-regulatory elements remains poorly understood.

As it is known that distal *cis*-regulatory elements communicate through 3D chromatin interactions, I consider that it would be coherent to group enhancers by 3D organisation rather than linear genomic proximity. Moreover, I hypothesise that large groups of enhancers contained within the same chromatin 3D domain could be especially relevant for both tissue-specific chromatin and transcription regulation. I speculate that high-resolution chromatin interaction maps will provide novel insight into how epigenomic factors drive enhancer gathering in 3D space. Furthermore, I propose that the interpretation of high-resolution chromatin interaction maps will allow us to picture enhancer clustering with more clarity.

C. High-resolution interaction maps could associate cis-regulatory elements, and disease-relevant non-coding variants, with their target genes more precisely than links based on linear proximity

Maps of *cis*-regulatory regions have been used to determine the molecular mechanism impaired by non-coding genomic variants associated with major diseases, such as diabetes. A *cis*-regulatory map in human pancreatic islets revealed that enhancer clusters are enriched in type-2 diabetes (T2D) risk and fasting glycemia (FG) variation associated variants (Pasquali et al., 2014). However, genomic variants and target genes cannot be precisely associated by linear proximity. Consequently, the target genes of T2D and FG risk genomic

variants will remain largely unknown until the characterisation of 3D chromatin structures in human pancreatic islets.

I propose that high-resolution chromatin interaction maps in a diabetes-relevant tissue, such as pancreatic islets, would be highly informative. These maps, in combination with the different epigenomic datasets generated by our lab and others (Bhandare et al., 2010; Pasquali et al., 2014; Stitzel et al., 2010), could help the scientific community on the field to hypothesise on the molecular mechanisms impaired by non-coding genomic variants and their possible effects on gene expression regulation. Furthermore, I hypothesise that these high-resolution chromatin interaction maps could provide an accurate list of genes linked to diabetes-associated non-coding genomic variants. This information will give further insight into the gene pathways impaired in diabetes.

To test these hypotheses I aim to:

- Determine chromatin organisation in human pancreatic islets by analysing high-resolution chromatin interaction maps.
- Elucidate novel features underlying the formation of tissue-specific chromatin interactions and interaction domains.
- Associate epigenomic features including 3D chromatin organisation with tissue-specific gene expression.
- Associate distal *cis*-regulatory elements with target gene promoters.
- Identify enhancers that cluster due to 3D chromatin organisation.
- Use islet interaction maps to associate diabetes-associated non-coding variants with likely target genes.

Chapter 3

3D chromatin organisation in human pancreatic islets

3.1. Integrative epigenomic analysis of human pancreatic islets

Understanding gene expression regulation is essential to interpret the effect of genomic variants associated with major diseases such as diabetes. For that reason, there has been a big effort to study gene expression regulation in human pancreatic islets and the effect of diabetes risk-associated variants (Flannick and Florez, 2016; Gaulton et al., 2010, 2015; Pasquali et al., 2014). However, an extensive characterisation of the 3D chromatin organisation in this tissue is still missing.

In order to fill this scientific gap, our group aimed to create a high-resolution chromatin interaction map in human pancreatic islets using promoter capture Hi-C (pcHi-C) (Mifsud et al., 2015; see section 1.5), in collaboration with Professor Peter Fraser's lab (Babraham Institute, Cambridge, UK).

This dataset was combined with published epigenomic (Pasquali et al., 2014) and gene-expression (Morán et al., 2012) maps, as well as newly generated ChIP-seq datasets for two key regulatory proteins (Mediator and Cohesin subunits) (Allen and Taatjes, 2015; Kim and Shiekhattar, 2015; Merkenschlager and Nora, 2016) (Fig. 13).

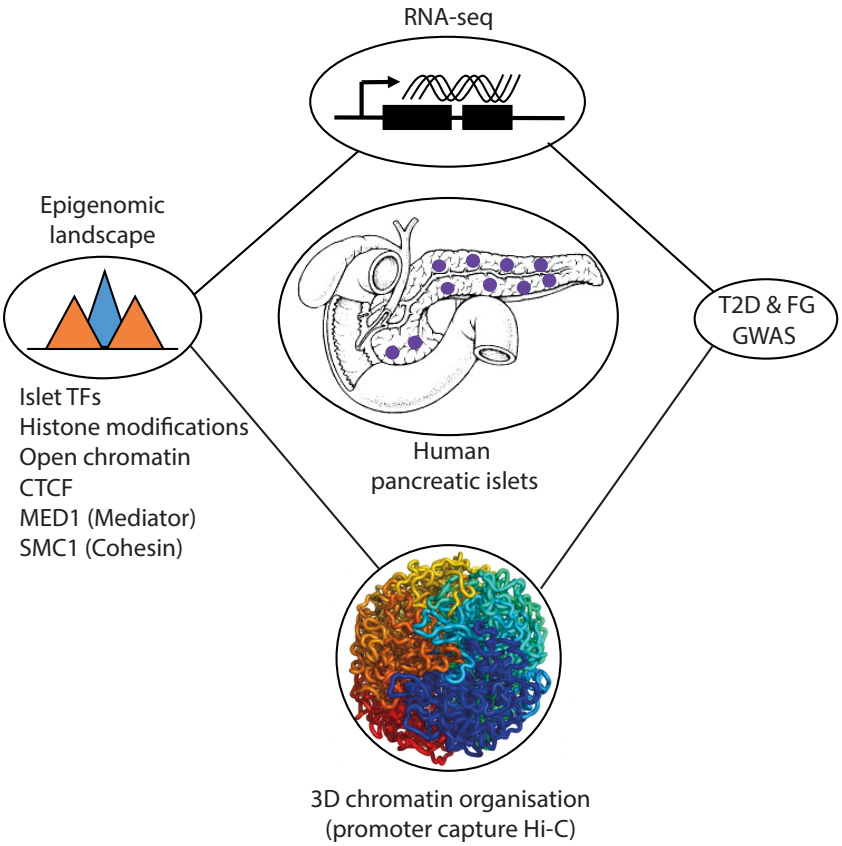


Fig. 13: Schematic of overall strategy. To increase our understanding of gene regulation in human pancreatic islets, I analysed the relationship between transcriptional and epigenomic landscapes, 3D chromatin organisation, and genomic variants associated with type-2 diabetes (T2D) and fasting glycemia (FG) variation.

Epigenomic maps in human pancreatic islets. A few years ago our group was able to identify candidate *cis*-regulatory elements in isolated pancreatic islets, including active enhancers and promoters, based on epigenomic marks (Pasquali et al., 2014). That work was centred on open chromatin regions (detected by FAIRE) that were characterised based on the binding of transcription factors, CTCF and the presence of histone modifications (H3K4me1, HK4me3, H3K27ac) in the adjacent nucleosomes (Pasquali et al., 2014). Recently, we incorporated new datasets including ATAC-seq to define open chromatin, ChIP-seq for H3K27me3 and H3K9me3 to define repressed chromatin, binding profiles for SMC1 (a Cohesin subunit) and MED1 (a Mediator complex subunit), as well CAGE maps of active transcription start sites. These new datasets were analysed by my colleagues (Irene Miguel-Escalada, Goutham Atla and Claire Morgan) following two independent strategies: clustering

of open chromatin regions (Pasquali et al., 2014) and chromatin segmentation through ChromHMM (Ernst and Kellis, 2012) (see section 8.1 for more details).

Clustering of open chromatin regions was based on genomic patterns of modified histones and DNA-binding proteins. This resulted in the annotation of 16,313 active promoters, 45,683 active enhancers, 66,029 inactive enhancers and 29,915 CTCF strongly bound regions. A subset of 13,635 enhancers (30% of all enhancers, which we called class I active enhancers), showed a very strong enrichment of H3K27ac and MED1 binding, compared to all other active (H3K27ac-enriched) enhancers (Fig. 46 in methods section 8.1). I note that this clustering approach to define regulatory elements aimed to select open chromatin regions but, unlike ChromHMM, it avoided long stretches of adjacent regions enriched on modified histones. I refer to the islet open chromatin maps as the **islet regulome**.

ChromHMM segmentation was used to characterise the co-occurrence of 12 epigenomic features in the entire genome, rather than only focusing on open chromatin regions. This analysis provided 15 states that were manually merged into 9 states based on similarities on their epigenomic profiles. These 9 ChromHMM states were: Polycomb repressed, heterochromatin, transcription, active enhancers, inactive *cis*-regulatory regions, bivalent promoters, active promoters, CTCF-rich regions and quiescent genomic regions (see section 8.1 for more information regarding the islet regulome and the islet ChromHMM).

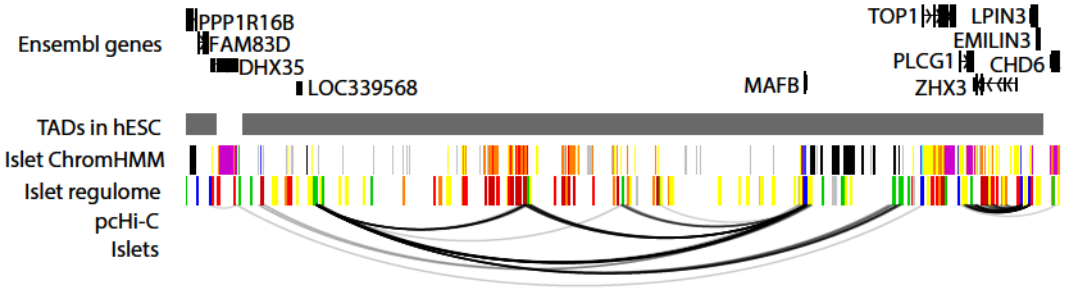


Fig. 14: Schematic representation of chromatin states at the *MAFB* locus. Screenshot around the MAFB locus exemplifying our collection of datasets. Tracks from top to bottom are: Ensembl gene annotation; TADs defined in human ESC (taken from Dixon et al., 2012); islet ChromHMM segmentation, where active promoters are indicated in blue, active enhancers are indicated in red or orange, inactive *cis*-regulatory elements in yellow, repressed regions in grey or black, CTCF binding sites in green and highly transcribed regions in purple; the islet regulome formed by different types of open chromatin sites categorised based on epigenomic marks, in which active promoters are

indicated in blue, active enhancers in red and orange, and inactive enhancers are shown in yellow; pcHi-C RNA probes targeting annotated promoters; pcHi-C interactions detected in human pancreatic islets. For more information regarding the islet ChromHMM and the islet regulome analyses see section 8.1

In summary, our integrative analysis allowed me to interrogate different aspects of gene expression regulation including *cis*-regulatory elements, chromatin states and chromatin interactions (Fig. 13, Fig. 14). These datasets were also combined with RNA-seq datasets in 15 human tissues, to determine gene expression patterns, and a list of type-2 diabetes (T2D) risk and fasting glycemia (FG) variation associated variants (Fig. 13). This creates a starting point to study the role of 3D chromatin organisation and non-coding variants on tissue-specific gene expression.

3.2. Generation of a high-resolution chromatin interaction map in human pancreatic islets

To identify long range chromatin interactions in pancreatic islets, four pancreatic islet samples obtained from cadaveric donors (with a purity > 70%) were used as starting material for Hi-C library preparation. Three Hi-C libraries were generated from each biological sample, and used as technical replicates. Then, the 12 Hi-C libraries were processed independently with SureSelect target enrichment, using 37,608 RNA probes against 21,177 human annotated promoters. The resultant pcHi-C libraries were sequenced on an Illumina platform. Raw reads from 3 technical replicates were pooled and mapped to the human genome (GRCh37/hg19) using the HiCUP pipeline (Wingett et al., 2015), filtering experimental artefacts such circularized reads and re-ligation products. This generated > 600M uniquely mapped pcHi-C paired-end sequence reads (ditags). Statistically significant chromatin interactions were determined using CHICAGO (Cairns et al., 2016), as in (Javierre et al., 2016), a method that builds a background model with an expected distribution of ligation signals for each bait, and identifies ligation fragments that exceed this expectation consistently in biological replicates. Chromatin and libraries were prepared by Xavier Garcia (IDIBAPS) and Dr. Biola Javierre (Babraham Institute, Cambridge, UK), respectively. For more information regarding pcHi-C library preparation see sections 1.5 and 8.3.

CHICAGO analysis of pcHi-C libraries yielded a total of 175,784 statistically significant high-confidence interactions (CHICAGO score ≥ 5) in human pancreatic islets. A descriptive analysis of human islet chromatin interactions map is shown in Table 4.

Table 4. Descriptive analysis of pcHi-C interactions in human islets.

Number of Hind III fragment baits	22,076
Baits with no annotated promoters (% of all baits)	484 (2.2 %)
Baits with one annotated promoter (% of all baits)	14,362 (65.0 %)
Baits with >1 annotated promoter (% of all baits)	7,230 (32.8 %)
Baits with at least one active promoter (% of all baits)	12,530 (56.7%)
Total number of interactions	175,784
Baits with ≥ 1 interaction (% of all baits)	16,030 (72.6 %)
Baits with active promoters with ≥ 1 interaction (% of all interactions)	9,796 (44.3 %)
Median number of interactions per bait (IQR*)	6 (2-14)
Median distance of *cis*-interactions (kb, IQR)	289.4 (166.7- 479.4)
Interactions in *cis*	175,122 (99.6 %)
Interactions in *trans*	662 (0.4 %)
Bait-to-bait interactions	13,386 (7.6 %)
Bait-to-non-bait interactions	162,398 (92.4 %)
Number of non-baited promoter-interacting regions	97,285

*IQR Interquartile range

In summary, we have been able to generate a high-resolution map of long range chromatin interactions in human pancreatic islets. This map is formed by 175,784 interactions that link annotated promoters with distal genomic regions.

3.3. Analysis of promoter-interacting regions (PIRs)

To understand the nature of regions that are in 3D proximity with annotated promoters, I analysed the epigenomic features of promoter-interacting regions. To this end, I analysed the overlap of interacting regions with the binding profiles of CTCF, Cohesin (SMC1), Mediator (MED1), as well as active enhancers and promoters that were annotated in our islet regulome.

To assess this overlap, I first determined the resolution of interacting regions that were identified through pcHi-C. I found that epigenomic factors such as CTCF-bound sites, which are expected to be enriched at promoter-interacting regions, were enriched not only at the latter, but also at the adjacent Hind III fragment. The results of this analysis are presented in greater detail in section 3.6. This is somewhat expected, because if an interacting fragment restriction site overhang is in proximity to a promoter fragment, the overhanging site from the immediately adjacent fragment should also be in proximity. I thus decided to extend non-baited interacting regions to include the adjacent Hind III fragments.

For all analyses, I distinguished interactions occurring between two pcHi-C baits, and those occurring between a baited fragment and a non-baited fragment.

I observed that promoter-interacting regions that did not contain baits showed a clear enrichment of elements known for driving structural and regulatory loops, such as CTCF, Cohesin and Mediator, as well as active enhancers (Fig. 15) (Krijger and de Laat, 2016). A similar scenario was true for baited interacting sites, although there was an expected much greater enrichment for active promoters rather than for active enhancers (Fig. 50 in methods section 8.5).

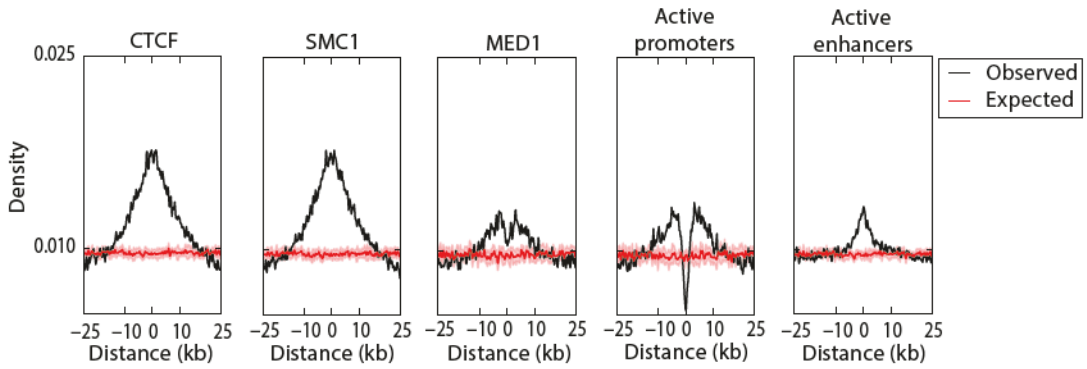


Fig. 15: Enrichments at non-baited promoter-interacting regions. Density plot showing the distribution of CTCF, SMC1, MED1, active promoters and enhancers around a -/+25 kb window of all promoter-interacting regions. The observed distribution is shown as a black line. The expected distribution was generated after randomising the positions of the interrogated epigenomic factor, and is shown as a red line, with values between interquartile ranges are shown as a red area. Note that the observed depletion at position 0 for active promoters and MED1 (a cofactor highly enriched at active promoters) was expected, as baits that contain annotated promoters were excluded from this analysis.

These findings were confirmed by my colleague Delphine Rolando, who interrogated the enrichments of islet ChromHMM states at islet pcHi-C interacting sites compared to an artificial set of interactions. This analysis replicated the enrichment of CTCF-bound sites, active enhancers, and promoters at interacting regions, and further showed enrichment of Polycomb-repressed chromatin (data not shown).

Characterisation of CTCF-associated interactions. CTCF is a key evolutionary conserved structural protein enriched at interacting regions determined by C-based methods (Dixon et al., 2012; Gómez-Marín et al., 2015; Ong and Corces, 2014; Zuin et al., 2014). Therefore, I decided to more accurately quantify the impact of CTCF binding regions on *cis* chromatin interactions detected by pcHi-C (Fig. 16).

In my analysis, I quantified the presence of islet CTCF binding sites at islet *cis*-interactions occurring between baited promoters and non-baited interacting regions. This quantification was done differentiating whether CTCF was observed at one specific interaction edge or both. I further characterised the mechanism by which CTCF could be acting as chromatin facilitator by interrogating the presence of a CTCF binding motif and its orientation when CTCF was binding both edges of an interaction.

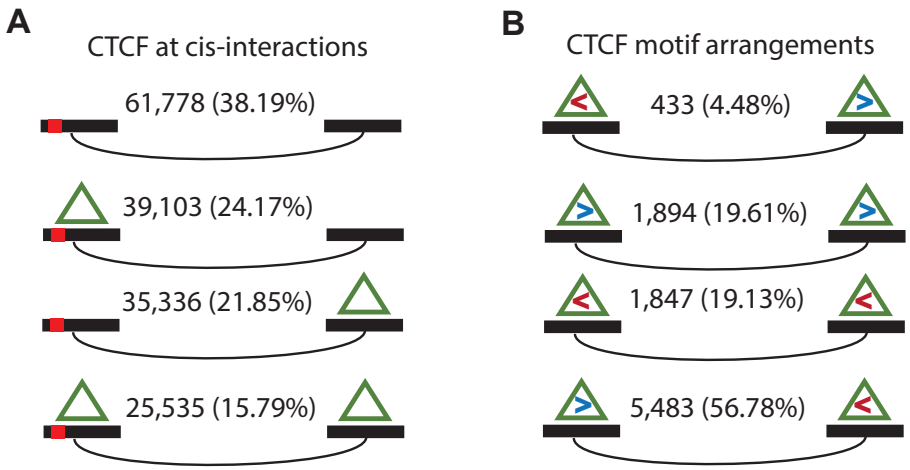


Fig. 16: CTCF as chromatin interaction facilitator. (A) Diagram illustrating the presence of CTCF among the 161,752 *cis*-interactions between baited and non-baited regions. Baited sites are indicated with a small red box. (B) Diagram representing the orientation of CTCF motif among the 9,657 interactions that showed a CTCF peak and motif in both interaction edges. Motifs with a forward orientation are indicated with a blue ">" symbol while motifs with reverse orientation are shown with a red "<" symbol.

Coherent with previous studies, I noticed that CTCF was present in a major fraction (99,974, 61.8%) of all pcHI-C interactions (Fig. 16A). Moreover, when CTCF was present in both edges of an interaction, it was frequently bound to sequence motifs that were arranged in convergent orientation, in keeping with previously reported findings (Rao et al., 2014; Sanborn et al., 2015; Tang et al., 2015; de Wit et al., 2015) (Fig. 16B). Thus, these results exemplify the importance of this protein in chromatin interaction formation. It also shows how pcHi-C maps can be used to confirm known aspects of topological chromatin conformation.

Enhancer-promoter interactions detected through pcHi-C maps. There were 35,286 promoter-enhancer interactions, which represent 20% out of all interactions. These interactions occurred between 7,149 baited regions (45.5% of all interacting baits) and 21,414 non-baited regions (22.0% of all bait-to-non-bait interactions). This observation shows that our pcHi-C map can associate promoters with distal *cis*-regulatory elements.

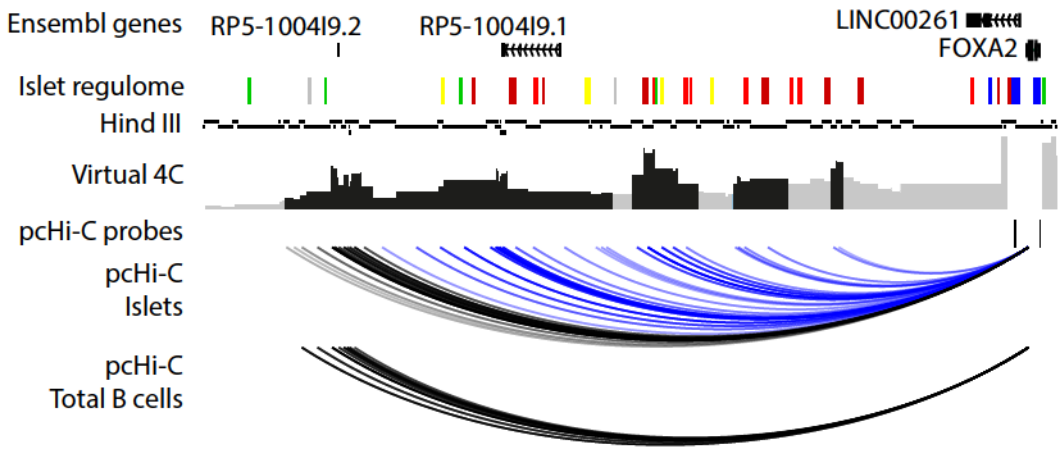
In summary, these results demonstrate that our high-resolution chromatin interaction maps in human pancreatic islets can recapitulate known features of 3D chromatin organisation. In addition, it illustrates how pcHi-C maps can associate distal regulatory regions, such as enhancers, with target promoters in human pancreatic islets.

3.4. Identification of tissue-specific chromatin structures

It has been suggested that chromatin organisation is formed by tissue-invariant structural interactions and tissue-selective regulatory interactions (Krijger and de Laat, 2016). In fact, previous high-resolution chromatin interaction maps have been able to associate the presence of tissue-selective interactions with tissue-specific gene expression (Javierre et al., 2016; Phillips-Cremins et al., 2013; Rao et al., 2014). However, the epigenomic factors driving the formation of tissue-selective interactions are still poorly understood.

Therefore, to define islet-selective chromatin structures we compared our pcHi-C map in human pancreatic islets with publicly available pcHi-C maps done in 4 hematopoietic cell types, namely Erythroblasts, Naïve CD4+ T lymphocytes, Total B lymphocytes, and M1 Macrophages (Javierre et al., 2016). Although this collection of samples does not cover all human tissue complexity, we considered it would reflect a representative picture of tissue-invariant chromatin organisation. Thus, through this comparison, we found that 59,672 islet interactions (34% of all interactions) were present in at least three out of the four hematopoietic tissues, highlighting that a significant proportion of pcHi-C interactions represent tissue-invariant structural interactions. By contrast, we identified 53,839 (31 %) interactions that were exclusively present in pancreatic islets, and henceforth named *islet-selective* interactions (Fig. 17). In conclusion, these results indicate that pcHi-C maps are composed of tissue-invariant interactions, from which probably a significant proportion are structural, and of tissue-selective interactions, which are potentially involved in tissue-specific gene regulation (Krijger and de Laat, 2016).

Fig. 17: Tissue-selective chromatin interactions in the *FOXA2* locus. Screenshot around the *FOXA2* gene. Tracks from top to bottom are: Ensembl gene annotation; islet regulome formed by different types of open chromatin sites classified through cluster analysis using epigenomic marks, in which active promoters are indicated in blue, active enhancers in red and orange, and inactive enhancers are shown in yellow; virtual digestion of the hg19 genome using HindIII restriction enzyme; virtual 4C showing the interaction frequencies, in human pancreatic islets, using *FOXA2* promoter as view point, statistically significant (CHICAGO score >= 5) promoter interacting regions are indicated in black; pcHi-C RNA probes targeting annotated promoters; and finally pcHi-C interactions detected in human pancreatic Islets and total B cells (Javierre et al., 2016). Islet-selective interactions are highlighted in blue. Figure shown in the following page.



Islet-selective chromatin interactions are associated with islet-specific gene expression. To systematically assess the relationship between islet-selective interactions and tissue-specific gene expression, I defined islet-specific expressed genes by comparing RNA-seq profiles in 18 human tissues. To do so, I assessed two aspects of gene expression: tissue-selectivity and relative expression in islets compared to all tissues. Overall selectivity of gene expression across tissues was measured computing the coefficient of variation (C.V.) among 18 human tissues. Relative expression in islets compared to the other tissues was assessed applying an islet specificity Z-score (Cebola et al., 2015). Thus, among 21,177 baited genes, 12,559 (59.3%) were defined as expressed and 8,618 (40.7%) non-expressed if their expression in human pancreatic islets was greater or lower than 1.5 transcripts per million (TPMs) respectively. Among all expressed genes, 983 (4.6%) had a C.V. as well as an islet-specificity Z-score greater than the 0.75 percentile of both values, and were defined as islet-specific expressed genes (Fig. 52 in methods section 8.8). The remaining set of 11,497 (54,3%) expressed genes were classified as expressed, non-islet-specific. Finally, as human pancreatic islets were surrounded by exocrine tissue before been collected and this can be a source of contamination, 79 (0.4%) genes with an expression 3 times higher in acinar cells than in pancreatic islets were considered as likely acinar contaminants.

I found that genes contained in baits showed islet-specific expression with increasing frequency as the number of islet-specific chromatin interactions increased (Fig. 18). Therefore, this result demonstrates a clear correlation between the presence of tissue-selective chromatin interactions and tissue-specific gene expression.

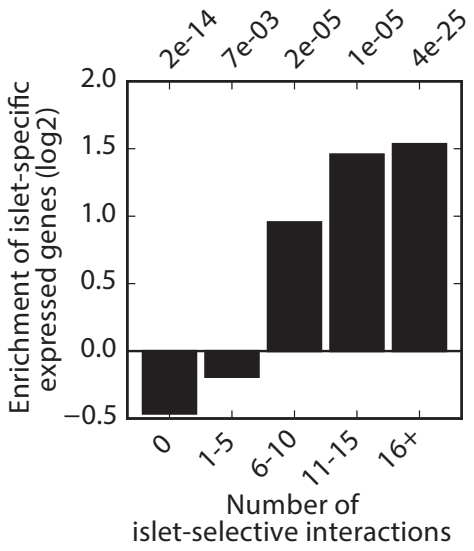


Fig. 18: Presence of islet-selective interactions correlates with islet-specific gene expression. Barplot indicating the enrichment, in a log2 scale, of islet-specific expressed genes among those genes associated with N islet-selective pcHi-C interactions. Statistical significance was measured in a hypergeometric test and the resultant p-value is indicated on top of each bar.

Islet-selective chromatin interactions are associated with co-binding of lineage-determining TFs, islet-specific CTCF binding sites and Mediator-bound enhancers.

Enhancer-promoter communications as a driving factor of islet-selective chromatin structures. As exposed before, little is known about the epigenomic factors involved in tissue-selective chromatin structure. However, it is well-stablished that enhancer-activity is associated with tissue-specific gene regulation (Javierre et al., 2016; Phillips-Cremins et al., 2013; Rao et al., 2014). This association is especially clear for a subset of enhancers that form enhancer clusters (Pasquali et al., 2014) and super-enhancers (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013), both of which show high binding of lineage-determining TFs (LDTFs) and Mediator. Thus, I decided to study the relationship between enhancers, their occupancy profiles, and tissue-selective chromatin structures.

By interrogating the co-occurrence between enhancers and the presence of islet-selective chromatin interactions, I could observe that enhancers, especially Mediator-bound enhancers, were enriched at islet-selective interactions compared to non-islet-selective interactions (Fig. 19). Notably, the occurrence of Mediator-bound enhancers was 1.3-fold

more frequent at non-baited interacting regions of islet-specific interactions than at non-islet-selective interactions ($p < 1e-20$, chi-square test). These results show that Mediator-bound enhancers were more specifically associated with islet-selective interactions, and suggest that islet-selective chromatin organisation could be partially driven by enhancer-promoter communication through the Mediator complex.

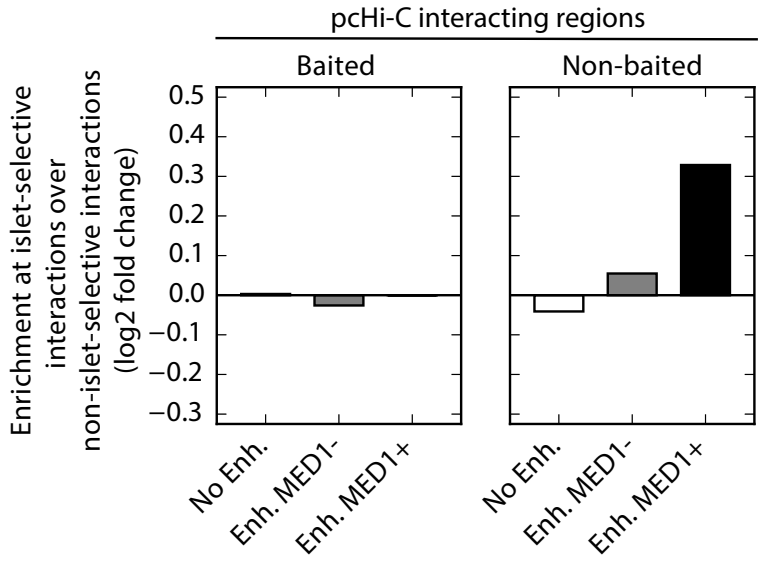


Fig. 19: Mediator bound enhancers are more frequently found at interacting regions of islet-selective chromatin interactions. Bar plots comparing the fraction of enhancers (Enh.) at islet selective vs. non-islet-selective interactions. Interacting sites were classified as "baited" (left) and "non-baited" interacting regions (right) and analysed separately. Interacting sites were further characterised and divided in 3 classes: (i) sites that did not overlap with enhancers (No Ehn.), (ii) sites overlapping enhancers not bound by Mediator (Enh. MED1-) and (iii) sites overlapping Mediator-bound enhancers (Enh. MED1+).

Islet-selective chromatin interactions are partially driven by islet-specific CTCF binding events. To gain further insight into the formation of tissue-specific chromatin structures I decided to examine the role of CTCF, as it has been proved to be a highly relevant chromatin interacting factor (Dixon et al., 2012; Gómez-Marín et al., 2015; Ong and Corces, 2014; Zuin et al., 2014). CTCF is a chromatin structural protein with a highly tissue-invariant binding profile (Cuddapah et al., 2008; Kim et al., 2008; Ong and Corces, 2014). However, by comparing CTCF binding with its profile in 14 non-pancreatic tissues and cell lines, I was able to identify > 2,000 highly tissue-specific sites that were only present in human islets and at most two non-pancreatic tissues (Fig. 53 in methods section 8.10).

To determine whether islet-specific CTCF binding sites could be involved in the formation of islet-selective chromatin interactions, I interrogated the co-occurrence of this structural protein at pcHi-C interacting points. I could observe that, similarly to Mediator-bound enhancers (Fig. 19), there was a 1.4-fold higher frequency of islet-specific CTCF binding sites at non-baited interacting regions of islet-selective chromatin interactions vs. non-islet-specific interactions (p-value <1e-9, chi-square test) (Fig. 20). As it could be anticipated, non-islet-specific CTCF binding sites were depleted at islet-selective chromatin interactions.

I noted that Mediator-bound enhancers at non-baited interacting regions of islet-selective interactions were 6 times more frequent than islet-specific CTCF binding sites at the same regions. Accordingly, Mediator-bound enhancers had a more significant p-value for enrichment in islet-selective interactions (Fig. 19, 20).

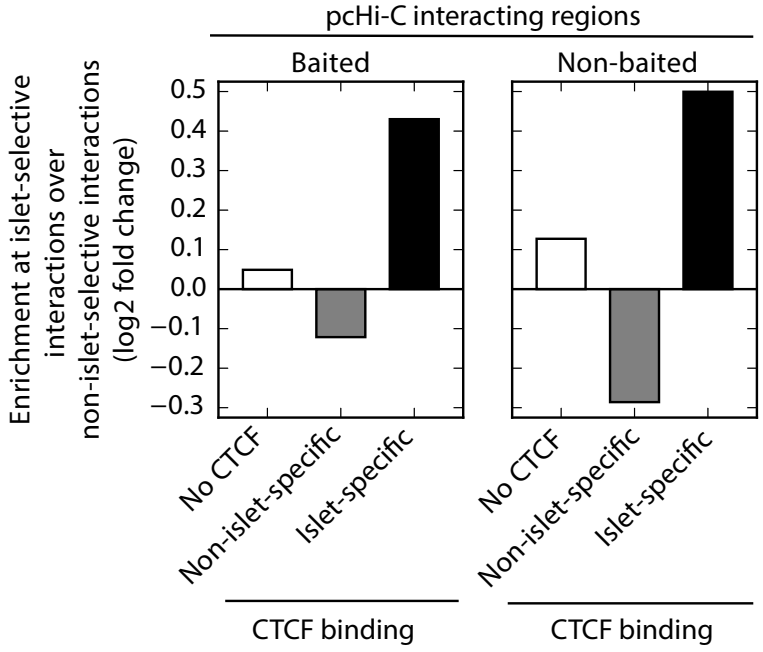


Fig. 20: Islet-specific CTCF binding sites are more frequent in islet-selective chromatin interactions. Bar plots comparing the fraction of CTCF binding sites at islet-selective over non-islet-selective interactions. Interacting sites were classified as "baited" (left) and "non-baited" interacting regions (right) and analysed separately. Interacting sites were further characterised and divided in 3 classes: (i) sites that did not overlap with CTCF (No CTCF) (ii) sites overlapping a non-islet-specific CTCF binding site and (iii) sites overlapping an islet-specific CTCF binding site.

In section 3.7 I present a more detailed analysis of the factors that could be driving islet-specific CTCF binding, and show that there is a correlation between the co-binding of LDTFs, Mediator, and the presence of islet-specific CTCF binding sites.

A clear example of the exposed result is the *ISL1* locus, which encodes for an islet TF that is essential for the beta-cells (Ediger et al., 2014) and shows an islet-specific gene expression pattern (Fig. 52 in methods section 8.8). I did not only observe that the *ISL1* locus presents an islet-selective chromatin organisation, but also that these islet-selective interactions occurred frequently between the *ISL1* promoter and Mediator-bound enhancers. Moreover, in other few cases, these interactions occurred between islet-specific CTCF binding sites (Fig. 21).

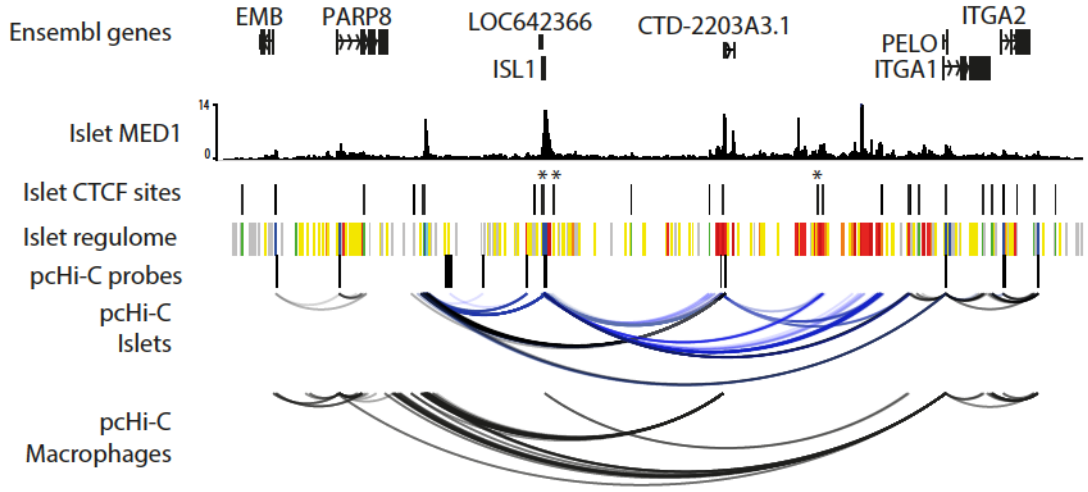


Fig. 21: Tissue-selective chromatin interactions in the *ISL1* locus. Screenshot around the ISL1 gene. Tracks from top to bottom are: Ensembl gene annotation; Mediator (MED1) ChIP-seq signal, position of CTCF binding sites; islet regulome formed by different types of open chromatin sites categorised based on epigenomic marks, in which active promoters are indicated in blue, active enhancers in red and orange, and inactive enhancers are shown in yellow; pcHi-C RNA probes targeting annotated promoters, and finally pcHi-C interactions detected in human pancreatic islets and Macrophages (Javierre et al., 2016). Islet-selective interactions are highlighted in blue. Islet-specific CTCF binding sites are indicated with an asterisk. For more information regarding the islet regulome see section 8.1.

Based on these results, it is reasonable to suggest that tissue-selective chromatin interactions could be partially driven by the collaborative binding of lineage-determining TFs and CTCF, as well as by active enhancers bound by the Mediator complex. These results provide insight into the epigenomic factors involved in the formation of tissue-selective chromatin structures.

3.5. Identification of TAD-like structures in human pancreatic islets

It is widely accepted that chromatin interactions are compartmentalised in topological associating domains (TADs) (Dixon et al., 2012) (reviewed in section 1.5). It has been observed that this compartmentalisation is important for gene expression regulation as its perturbation provokes aberrant gene expression patterns (Franke et al., 2016; Lupiáñez et al., 2015, 2016).

TADs have been defined in several tissues using Hi-C (Lieberman-Aiden and Berkum, 2009), including whole pancreas (Schmitt et al., 2016). However, to the best of my knowledge, there is not any published Hi-C dataset from human pancreatic endocrine islets, which represent less than 4% of pancreatic cells. For that reason, I assessed the possibility of defining TAD-like structures using pcHi-C data.

Dixon et al. noticed that interactions detected by Hi-C (Lieberman-Aiden and Berkum, 2009) tend to end abruptly at focal genomic regions, suggesting the presence of chromatin topological boundaries (Dixon et al., 2012). Most interactions tend to be contained between those boundaries that form the TADs. Dixon et al. also observed that at TAD boundaries there is an enrichment for interactions going toward the centre of the domain. Thus, Dixon et at. formulated a score that quantifies the degree of interaction bias for a given locus to systematically identify TADs. This score was coined as directionality index (DI) score.

As TAD organisation has been described in several tissues (Dixon et al., 2012; Schmitt et al., 2016) and islet pcHi-C interactions seem highly concordant with TAD compartmentalisation in other human cell types (Fig. 14), I hypothesised that the same DI score would be suitable to define TAD-like structures using this dataset. However, it was necessary to consider that pcHi-C achieves a higher resolution than standard Hi-C by reducing the catalogue of detectable interactions (Mifsud et al., 2015). Therefore, parameters in the DI formula were modified to account for it (as summarised in Table 10, section 8.13).

By implementing the DI formula, it was possible to identify loci with an interaction bias and define territories flanked by loci with opposite DI scores. Those territories were named DI domains.

A known feature of chromatin organisation in TADs is the low degree of interconnectivity between domains (Dixon et al., 2012). This feature was also recapitulated in most DI domains as observed by computing the ratio between the number of inter-domain and intra-domain interactions (Fig. 54 in methods section 8.13). However, a small fraction (approx. 10%) of DI domains showed a high degree of interconnectivity. Therefore, as a post-analytic correction, adjacent DI domains with high degree of interconnectivity (log2 ratio inter/intra-domain interactions > 0) were merged. Thus, I defined chromosomal territories that I named **islet TAD-like structures** (Table 5 and Fig. 22).

Table 5: Description of TAD-like domains.

Number	3,589
Size (kb, IQR)	472.5 (257.3-777.4)
Median number interactions per TAD (IQR)	26 (9-60)
Genome coverage	69.8 %

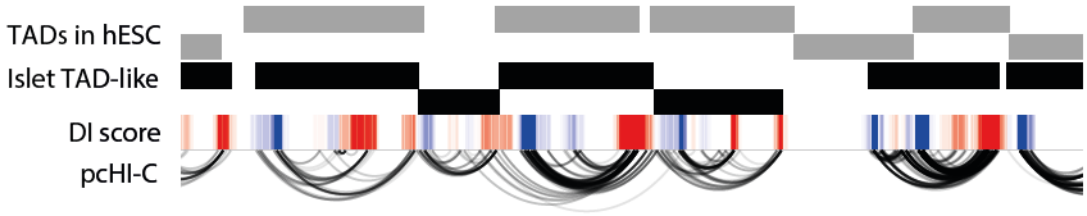


Fig. 22: Islet TAD-like compartments. Screenshot at chr11:1132582-4719948 genomic region exemplifying the identification of TAD-like structures based on islet pcHi-C interaction directionality index (DI) scores. Negative DI scores, indicating interactions looping to the left, are shown in blue and positive DI scores, indicating interactions looping to the right, are shown in red. For comparison, TAD defined in hESC using Hi-C data (Dixon et al., 2012) are represented in grey.

To ensure the biological importance of islet TAD-like domains, I confirmed that these structures recapitulate known TAD features, such as (i) the enrichment of CTCF at TAD borders (Dixon et al., 2012) with a convergent orientation (Rao et al., 2014; Sanborn et al., 2015) (Fig. 23A) or (ii) the fact that TAD boundary regions are broadly conserved among tissues (Schmitt et al., 2016) (Fig. 23B). Coherent with tissue-invariant TAD compartmentalisation, most islet pcHi-C interactions were contained within TADs defined

using Hi-C data in other tissues (Dixon et al., 2012) (Fig. 23C). Therefore, although our promoter centric high-resolution chromatin interaction maps are not reflecting all existent interactions, especially structural interactions where promoters may not be involved, it seems that these maps are able to reflect TAD compartmentalisation.

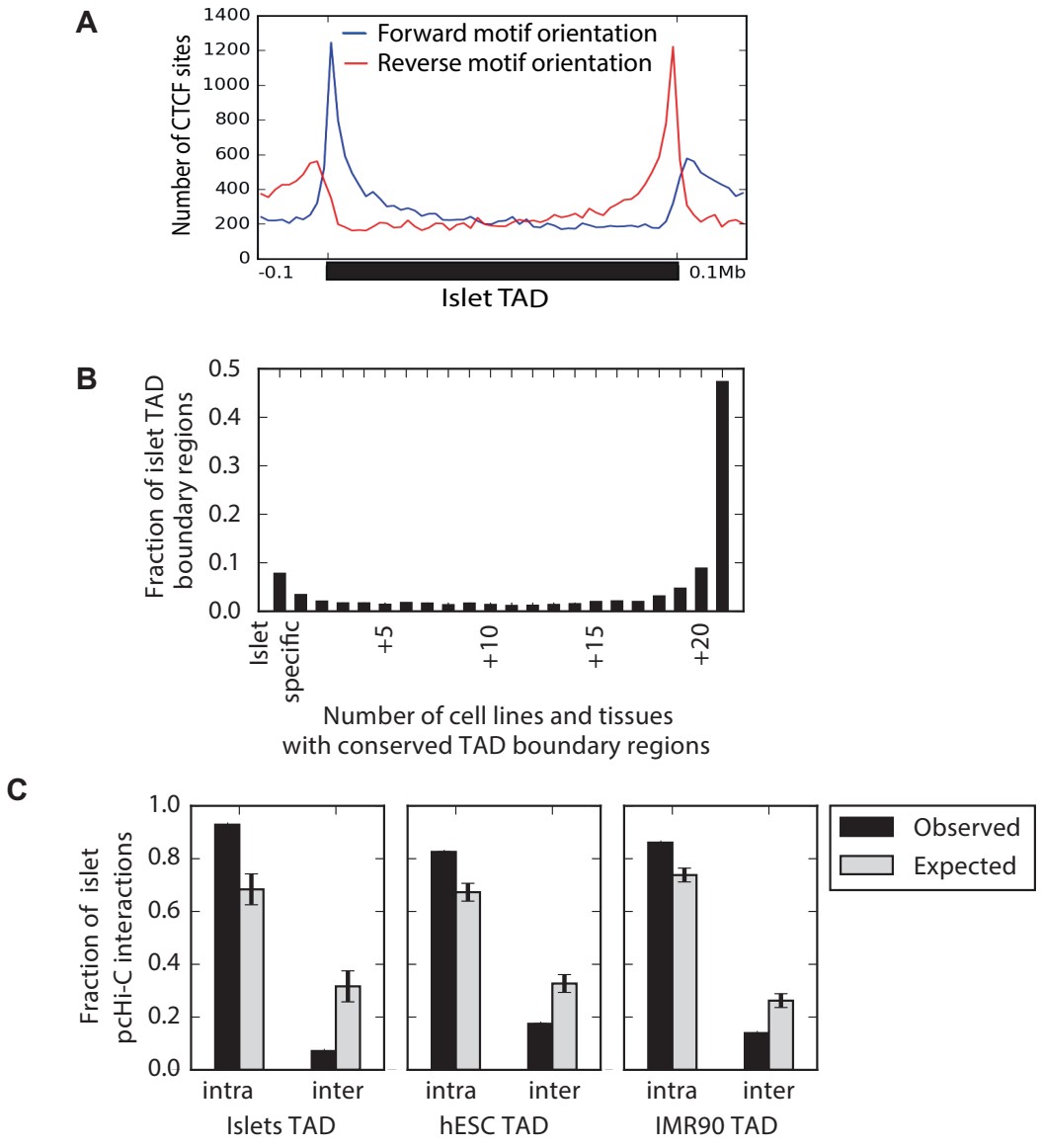


Fig. 23: Islet chromatin compartmentalisation exhibits known features of topological associating domains (TADs). (A) CTCF distribution towards islet TAD-like compartments accounting for CTCF motif orientation. CTCF binding sites with forward or reversed motif are shown in blue or red respectively. (B) Tissue-selectivity of islet TAD-like boundary regions. The degree of tissue-selectivity was determined by comparing TAD segmentation in human pancreatic islets against TADs defined by Hi-C in 21 tissues from Schmitt et al., 2016. (C) Intra and inter connectivity of islet pcHi-C interactions regarding islet TAD-like structures and TADs defined in human ESC and human IMR90 fibroblasts (from Dixon et al., 2012). Expected values were calculated after randomising TAD positions 5 times.

3.6. Resolution of pcHi-C interacting regions

To identify long-range chromatin interactions, C-based methods (like pcHi-C) quantify the frequency of re-ligation events between distal loci. To do so, the sample is fixed to keep its 3D chromatin conformation, then it is digested with a restriction enzyme (like HindIII), re-ligated and des-crosslinked (Fig. 24A). This generates linear genomic fragments that contain the sequence of two interacting loci that can be quantified by next-generation sequencing (Bonev Boyan and Cavalli Giacomo, 2016; Dekker et al., 2013).

Each restricted fragment has several potential ligation partners and their frequency of ligation is directly affected by their spatial proximity at the time of fixation. It has been determined that only ~5% of all genomic restriction fragments in a C-based library re-ligate with their adjacent partner in the linear template (Fig. 24B) (Davies et al., 2015). Therefore, it is assumed that a large proportion of ligation events reflect 3D chromatin contacts between distal loci at the time of fixation. Thus, re-ligation events are used to determine contact frequency between loci. Chromatin interactions are defined as pairs of loci with higher contact frequencies than expected (Fig. 24C). However, it is important to remember that the identification of confident chromatin interactions not only relies on contact frequency but also on sequencing coverage, so a specific re-ligation event is observed a minimum number of times (Davies et al., 2017).

To identify statistically significant interactions based on pcHi-C data, M. Spivakov ’s lab developed CHICAGO (Cairns et al., 2016). This algorithm identifies confident pcHi-C by considering expected frequency of random collisions between proximal genomic fragments (Brownian motion), observed contact frequencies and sequencing coverage. However, it is not difficult to envision the challenge to differentiate between the actual non-baited interacting region and its adjacent fragments. Due to 3D spatial proximity, overhanging edges of an interacting fragment may show similar re-ligation probability with a baited promoter than the closest overhanging edges from the adjacent non-interacting fragments with the same baited promoter (Fig. 24D). Nevertheless, the actual interacting fragment will generally show a high interaction frequency as both edges will re-ligated with the baited

fragment while the adjacent non-interacting fragments will only present re-ligation events for the closet edge to the interacting fragment.

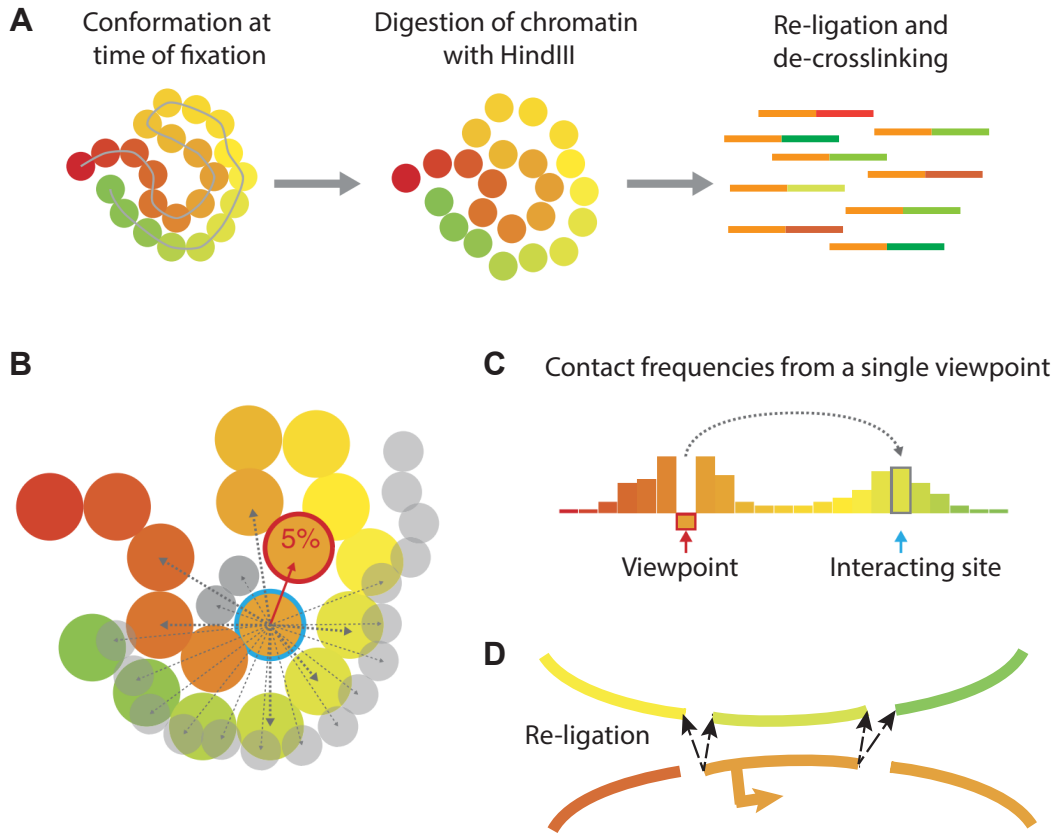


Fig. 24: Identification of chromatin contacts through re-ligation events. (A) Diagram illustrating the generation of re-ligated genomic fragments during the generation of a C-based library. (B) Diagram illustrating the computation of interaction frequencies from a specific viewpoint (highlighted with a blue circle) and all genomic fragments. It also shows that it has been determined that only 5% of re-ligation events occur between fragments that are adjacent in the linear template. (C) 4-C like representation of the contact frequencies from a given viewpoint. A dashed arrow indicates the association between the viewpoint and a distal interacting region. (D) Schematic indicating the possible ligation events between a promoter (viewpoint) and interacting site. Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: Nature Methods Davies et al., 2017; copyright (2017)

Moreover, there is another technical aspect that needs to be considered regarding the identification of confidential interacting sites. Although CHICAGO estimates the effect of "bait-specific" and "other end-specific" bias factors coming from technical sources such as library pull-down, to the best of my knowledge the current algorithm does not consider

Brownian motion at the non-baited interacting sites. This may be causing the identification of false positive interacting sites adjacent to actual interacting sites.

I considered that the two previously mentioned aspects could have an impact over CHICAGO's capacity to identify interacting sites with statistical confidence. To assess this question, I considered that most confident interacting sites would overlap with an epigenomic factor known for driving chromatin interactions. Therefore, I interrogated the overlap of interacting sites and a list of epigenomic factors that was formed by: CTCF, Cohesin, Mediator, active promoters and active enhancers. (See methods section 8.6 for more details).

It was noticeable that despite the clear enrichment of chromatin interacting factors at non-baited promoter-interacting regions (Fig. 15), less than 25 % of them directly overlapped any of the interrogated factors (Fig. 25A left, Fig. 51 in methods section 8.6). I also found that most of these "empty" non-baited promoter-interacting regions occur near known interacting factors more frequently than expected (Fig. 25A right). Moreover, this was only observed for non-baited interacting sites as most baited sites did overlap with an interrogated feature and the enrichment was restricted to the baited HindIII fragment (Fig. 25B, Fig. 51 methods section 8.6).

Based on these results, I considered that constraining our analysis to the non-baited interacting HindIII fragment could mask interactions between distal *cis*-regulatory regions. This would reduce our capacity to associate baited promoters and elements located in non-baited interacting regions. Therefore, I considered reasonable to extend non-baited promoter-interacting regions covering the adjacent HindIII fragment from each side when associating distal epigenomic factors and promoters.

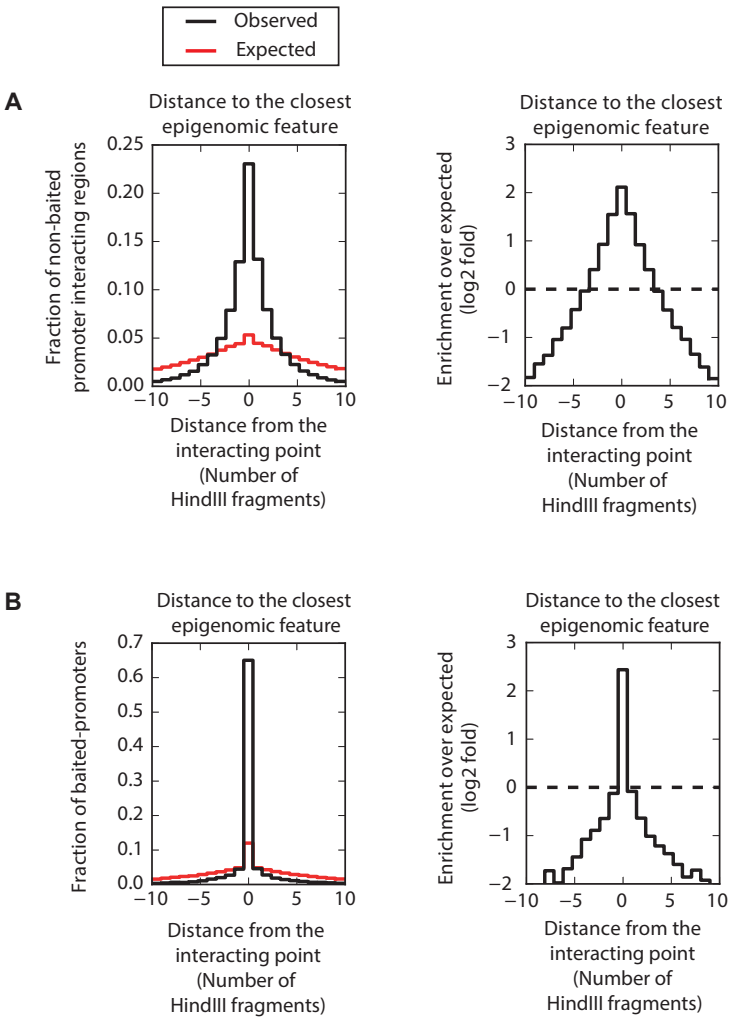


Fig. 25: Distance from an interacting site to the closest epigenomic factor. Histogram showing the distance from the closest epigenomic factor (any of them) to (A) non-baited promoter-interacting regions or (B) baited promoters. The list of interrogated epigenomic factors encompassed CTCF, MED1, SMC1, active promoters and active enhancers. Distance was computed as the number of HindIII fragments from the interrogated point. A distance equal to 0 means that the epigenomic factor overlaps with the interacting HindIII fragment. The expected distribution was generated after randomising the positions of the interrogated epigenomic factors in human pancreatic islets. Enrichments over expected distributions at non-baited sites (top) and baited promoters (bottom) are shown on the right panels.

3.7. Islet-specific CTCF binding driven by lineage-determining TFs

As previously mentioned, islet-specific interactions were not only associated with Mediator-bound enhancers, but also with CTCF binding sites that are only encountered in pancreatic islets, or in islets and a minority of tissues (Fig. 19, Fig. 20 in section 3.4). I therefore examined this subset of islet-specific CTCF binding sites. To reveal whether 5 islet lineage-determining TFs (FOXA2, PDX1, MAFB, NKX6.1 and NKX2.2) could be driving islet-specific CTCF binding, I interrogated their co-occurrence with CTCF. In addition, I also interrogated the co-occurrence of Mediator (MED1) and Cohesin (SMC1) to further understand if there is any relation between all these factors (Fig. 26A).

Even though only 8% of all CTCF binding sites co-occurred with MED1 (Fig. 26A); the results indicated that this overlap is likely to be driven by lineage-determining TFs (LDTFs), as their co-occurrence increased with the number of co-binding LDTFs (Fig. 26C). This correlation was especially clear at islet-specific CTCF binding sites (Fig. 26B-C). However, the same observation is not true for Cohesin (Fig. 26D), which frequently co-occurred with CTCF independently of LDTFs co-binding or CTCF tissue selectivity.

Therefore, it seems that despite the small overlap between them, CTCF and MED1 have a high degree of co-localisation at loci highly bound by LDTF.

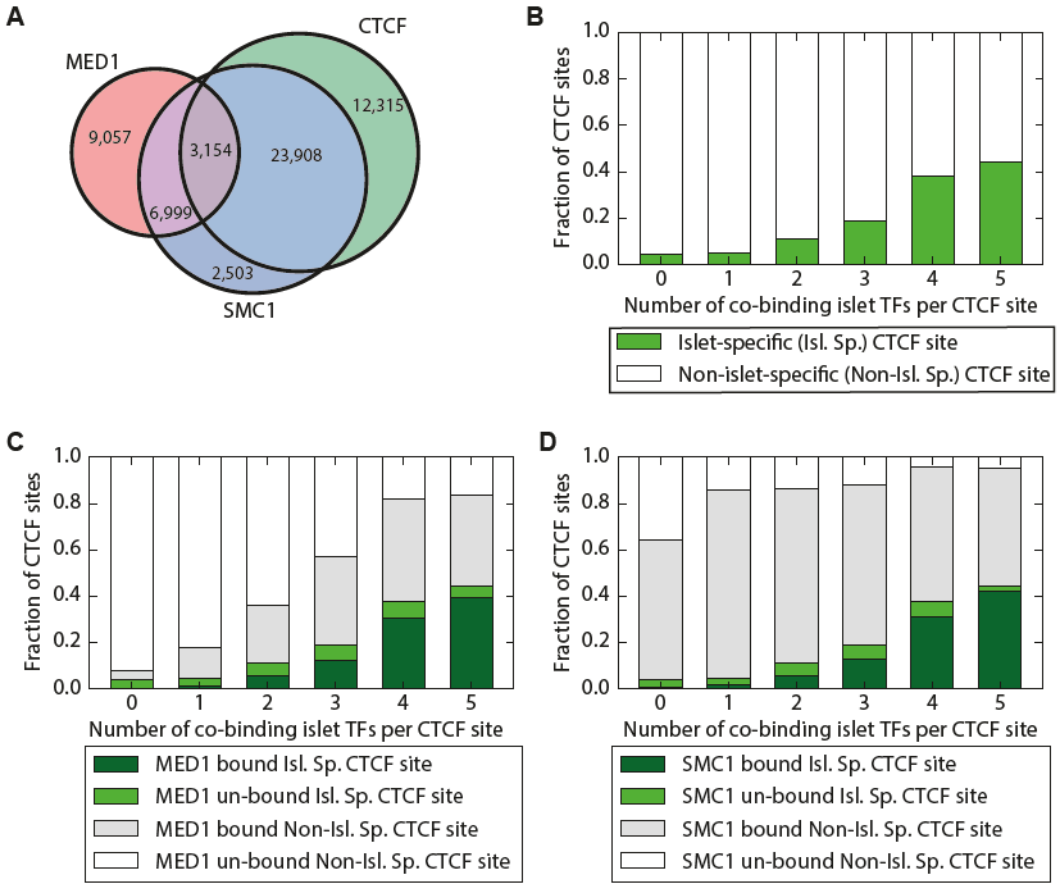


Fig. 26: Cooperative work between islet LDTFs, CTCF, MED1 and SMC1. (A) Venn-Diagram illustrating the overlap between Mediator (MED1), Cohesin (SMC1) and CTCF consistent binding sites in human pancreatic islets. (B) Stacked bar plots showing the fraction of CTCF sites in relation to the co-binding of lineage-determining TFs (LDTFs). CTCF and Mediator (C) or Cohesin (D) co-occupancy in relation to the binding of LDTFs. The interrogated islet LDTFs were: FOXA2, PDX1, MAFB, NKX2.2, NKX6.1. CTCF binding was differentiated between islet-specific (Isl.Sp.) and non-islet-specific (Non-Isl.Sp.).

Chapter 4

Promoter centric chromatin interaction domains

4.1. Identification of islet promoter-associated domains (PADs)

Although TAD compartmentalisation has been proved to affect gene expression regulation (Andrey et al., 2013; Gonzalez-sandoval and Gasser, 2016; Guo et al., 2015; Lupiáñez et al., 2015, 2016), it is not clear if TADs represent the smallest territorial unit that defines the genomic linear space able to modulate gene transcription. Indeed, studies that achieved a higher resolution than standard Hi-C revealed evidences of sub-structures within TADs (Phillips-Cremins et al., 2013; Rao et al., 2014).

To interrogate whether pcHi-C maps are able to detect intra-TAD chromatin organisation, I assessed the possibility that promoters have more constrained intra-TAD regulatory landscapes. Thus, per each baited promoter in pcHi-C, I grouped all intra-TAD interacting regions in a single linear genomic segment (Fig. 27). These genomic intervals were named **promoter-associated domains (PADs)** (Table 6).

Note that PADs are not necessarily meant to represent separate substructures of TADs (Phillips-Cremins et al., 2013), but simply define the genomic space that is likely to contain regulatory interactions with each annotated gene. Thus, different PAD segments can overlap and some PADs fully cover a genomic region delimited by a TAD (Table 6, Fig. 27, Fig. 49).

Table 6: Description of islet PADs.

Total number of PADs	16,030
Median size (kb, IQR)	376.58 (211.43 – 624)
Median number of interactions per PAD (IQR)	5 (2-14)
Median number of PADs per TAD (IQR)	3 (2-6)

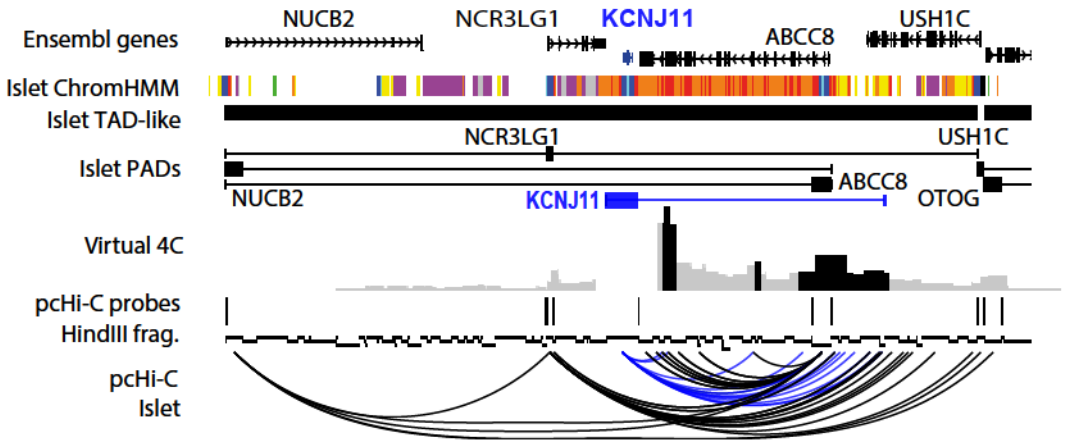


Fig. 27: *KCNJ11* promoter-associated domain (PAD). Screenshot around the *KCNJ11* gene locus exemplifying promoter-associated domain definition. Tracks from top to bottom are: Ensembl gene annotation; islet ChromHMM where active promoters are indicated in blue and active enhancer are indicated in orange and red; Islet TAD-like compartment;, PADs defined for all interacting promoters in the locus; virtual 4C showing the interaction frequencies, in human pancreatic islets, using *KCNJ11* promoter as view point, statistically significant (CHICAGO score >= 5) promoter interacting regions are indicated in black; pcHi-C RNA probes used to target annotated promoters; virtual digestion of the hg19 genome using HindIII restriction enzyme and finally islet chromatin interactions detected by pcHi-C. For PADs, thicker lines correspond to the HindIII fragment that contains the indicated gene promoter. PcHi-C chromatin interactions originated from the *KCNJ11* promoter are highlighted in blue and used to defined the *KCNJ11* PAD, also highlighted in blue. For more information regarding the islet ChromHMM and its colour code see section 8.1.

I observed that a large proportion of islet PADs were significally smaller than islet TAD-like structures, as 44% of islet PADs were >25% smaller than the TAD-like compartment where they were contained (Fig. 27, Fig. 28). Since a big proportion of PADs are smaller than their respective TAD-like compartment, this suggests that some promoters do not interact with any region within a TAD and their interactions are focalised in a smaller area.

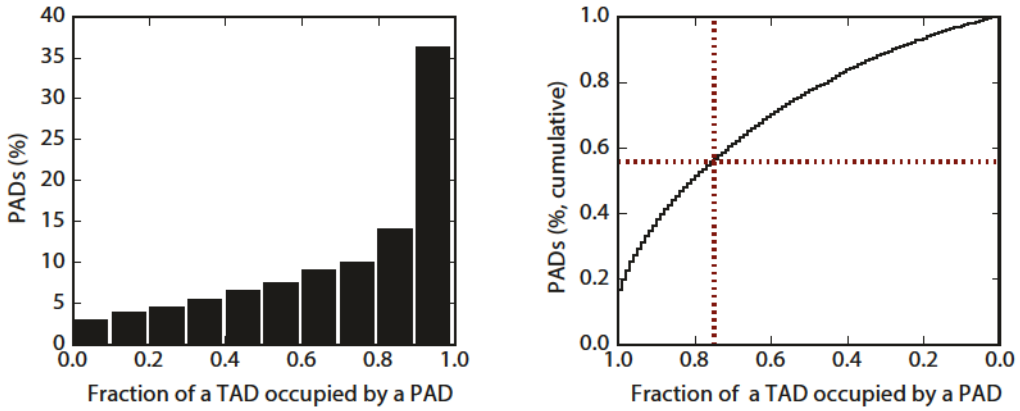


Fig. 28: Overlap between TAD and PAD segmentation. From left to right: Histogram indicating the percentage of PADs that did occupy a given fraction of TAD space. Cumulative distribution of the

percentage of PADs that did occupy a given fraction of TAD space. Dashed maroon lines indicate that 44% of PAD occupy 75% or less of a TAD space.

Regulatory states in islet PADs are concordant with gene expression. To understand whether islet PADs could encompass the *cis*-regulatory space of gene promoters, I interrogated whether the chromatin states contained within the PADs were coherent with the gene expression patterns. To do so, I took advantage of our islet ChromHMM map, a genome-wide segmentation of chromatin states in human pancreatic islets. This map was formed by 9 states, which were defined based on 12 epigenomic datasets generated in human pancreatic islets (see section 8.1. for more details). I computed the enrichment of each ChromHMM state within each islet PAD over its genomic distribution. I then grouped islet PADs in 5 bins based on the expression levels associated to the bait. This analysis revealed that active chromatin states, such as active enhancers, were enriched at islet PADs of highly expressed genes and depleted at those linked to lowly expressed genes. Therefore, ChromHMM states within PADs were coherent with gene expression patterns (Fig. 29).

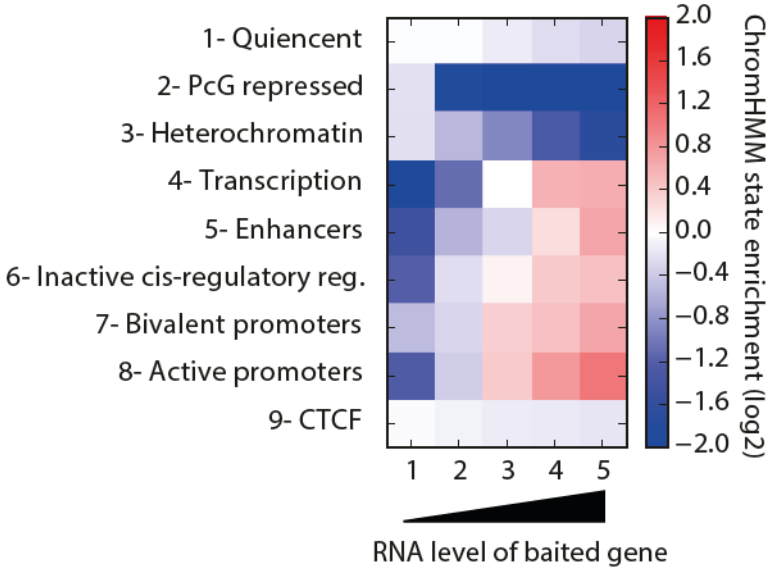


Fig. 29: Epigenomic states at islet PADs are coherent with gene expression. Heatmap indicating the median enrichment over genomic distribution of a given ChromHMM state in a PAD. PADs were separated in 5 bins based on bait gene expression levels in human pancreatic islets.

Next, I addressed if epigenomic states in PADs differ from non-PAD space within TAD compartments. To do so, I selected 7,085 PADs at least 25% smaller than the TAD in which they were contained (Fig. 28). I compared the ChromHMM states at PADs and the remaining TAD space in relation to gene expression. This comparison showed that islet PADs

associated with highly active genes were enriched in active enhancers and depleted in heterochromatin as compared to the remaining TAD space (Fig. 30).

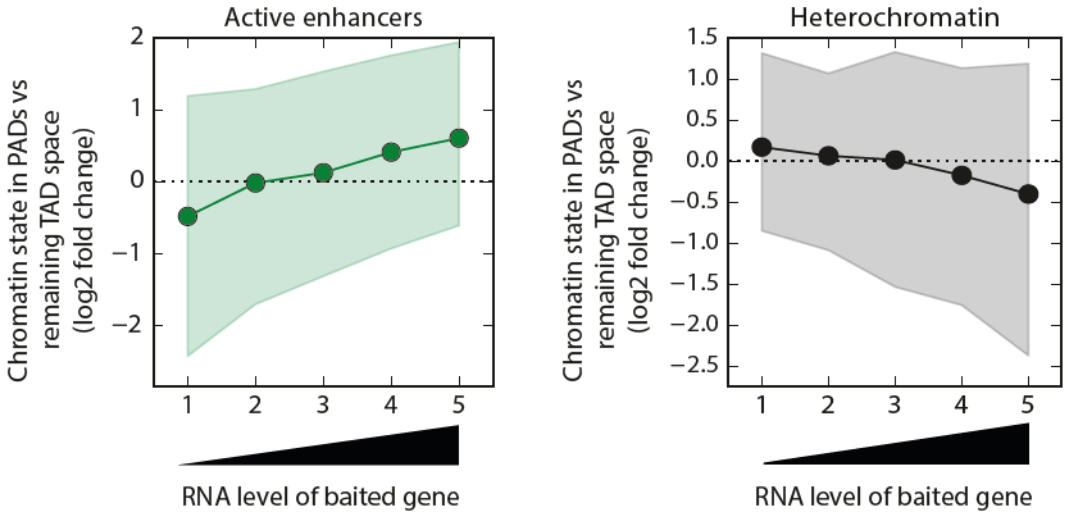


Fig. 30: Comparison between chromatin states at PADs and remaining TAD space. Graphs showing enrichment of active enhancers (left) or heterochromatin ChromHMM states (right) in PADs over remaining TAD-like space. This analysis was performed using PADs that were at least 25% smaller than the TAD in which they are contained (Fig. 28). Median enrichments are indicated with circles and interquartile ranges are showed as a coloured area. Data was separated in 5 bins based on the gene expression level of the baited gene in human pancreatic islets.

These results suggest that PADs represent a more restricted genomic linear space where *cis*-regulatory elements have a regulatory effect over a specific gene promoter.

Assignment of regulatory elements to target genes. Interrogating our pcHi-C map I could identify 35,286 interactions between enhancers and promoters, which corresponded to 40% of all 45,683 candidate active enhancers. This indicates that a large proportion of all active enhancers could be precisely associated to a promoter based on experimentally detected chromatin interactions.

Other candidate enhancers did not show physical interactions with any gene. This could be for several reasons. For example, it can be that enhancer-promoter regulatory interactions are dynamic, or only active in specific physiological contexts, and therefore difficult to capture with current experimental procedures. Moreover, due to background modelling and high rate of random collisions, short interactions require a very high sequencing coverage to be considered as statistically significant. Therefore, some interacting enhancers may be too

close to the baited promoter to detect confident interactions (CHICAGO score ≥ 5) between them. Finally, some candidate enhancers defined by chromatin marks may not represent bona fide functional regulatory elements.

Taking into account these considerations, I assessed the possibility that pcHI-C data could be used to predict likely targets of candidate enhancers that did not show high confident interactions to any gene promoter. My previous results indicated that the chromatin state in PADs was more coherent with gene expression patterns than simply considering TAD compartmentalisation (Fig. 30). Thus, it suggested that associations based on PADs would be more reliable than assumptions based on either linear proximity or TAD maps. However, very often enhancers map to more than one overlapping PAD, and it was therefore necessary to consider that maybe not all promoters in a genomic space with enhancers are under enhancer regulation. Therefore, I used pcHiC data to tentatively associate enhancers to promoters that were not linked by direct interactions, using the following criteria (Fig. 31):

1. I reasoned that an enhancer that does not show any promoter interaction does not necessarily regulate the gene(s) located in the bait of all PADs to which the enhancer maps. However, if a non-interacting enhancer is contained within one or more PADs whose baits have active promoters that do interact with other enhancers, I reasoned that these promoters are regulated by enhancers within their PADs. I therefore tentatively associated any non-interacting enhancers to their PADs if these had enhancer-interacting active gene promoters.

2. I assumed that a +/-10 kb window around any baited promoter encompasses a region where random collisions are too frequent to enable the identification of high confidence interactions above background noise. I also assumed that the close linear proximity was indicative of physical 3D proximity between putative enhancers and target genes. Therefore, non-interacting enhancers (not assigned in step 1) that reside within 10kb of a baited active promoter, were automatically assigned to them.

3. The remaining non-interacting enhancers were associated to a baited active promoter if they were exclusively contained within a single PAD.

All of these steps of our assignment strategy required an active promoter, identified either in the islet ChromHMM or the islet regulome approaches, to associate a non-interacting enhancer to a baited promoter.

This strategy allowed me to assign 80 % of all enhancers in human pancreatic islets to a promoter (Fig. 31).

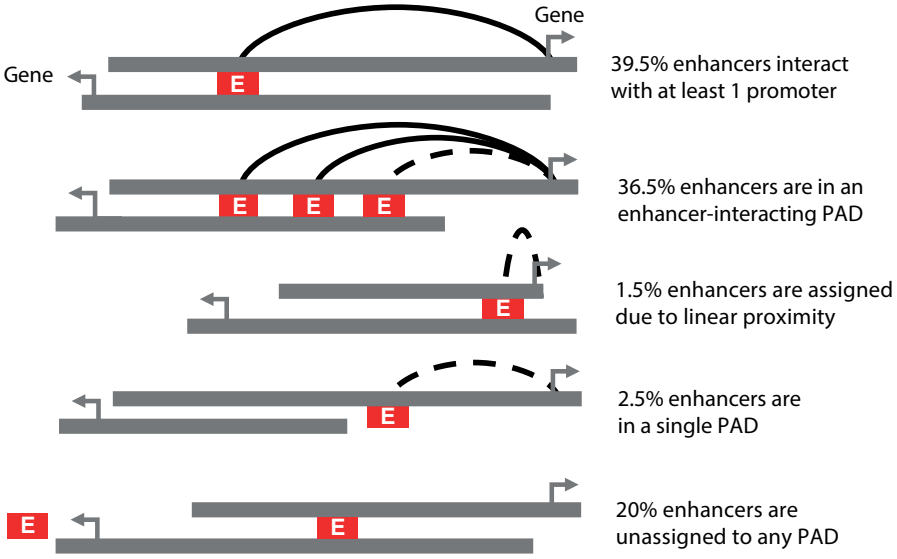


Fig. 31: Enhancer assignment. Diagram illustrating the different steps followed to assign enhancers to promoters. First, enhancers were linked to promoters based on pcHi-C interactions. Second, non-interacting enhancers were tentatively assigned to active promoters that showed interactions with other enhancers. Third, proximal (<10kb) enhancers and active promoters were linked based on linear proximity. Finally, enhancers were assigned to an active promoter if they did not overlap with other PADs. PcHi-C interactions are indicated as solid black arcs, inferred enhancer-promoter associations as dashed black lines, PADs are indicated as grey horizontal bars and enhancers (E) as red boxes.

To validate this strategy, I asked whether the assignment of any given enhancer to a gene A represented an improvement compared to a gene B whose PAD did overlap with the enhancer, but this enhancer was not assigned to the gene B (see Fig. 55 in section 8.22). I assumed that the real target of any islet enhancer is expected to be active in this particular tissue. Moreover, a large proportion of these associations should involve islet-specific expressed genes, because enhancers are often found near islet-specific expressed genes and are generally thought to be important for lineage-specific expression. Therefore, I

computed the enrichment of islet-specific genes (section 8.8) among genes with assigned enhancers. To do so, I generated two lists of genes. The first list was named “assigned genes” and contained all genes with an assigned enhancer (Fig. 31). The second list called “control genes” contained all genes whose PADs overlapped with active enhancers but these were not assigned to them (Fig. 55 in methods section 8.22). Per each gene expression class (non-expressed, islet-specific and non-islet-specific expressed), I computed the enrichment among “assigned genes” compared to “control genes”. The results indicated a clear enrichment of islet-specific expressed genes in genes with assigned enhancers (Fig. 32, Fig. 56). This is coherent with the notion that tissue-specific gene expression is driven through enhancer-promoter communication, and suggests that most assignments differentiate bona fide targets from other genes in the vicinity of enhancers (Heinz et al., 2015; Maston et al., 2006).

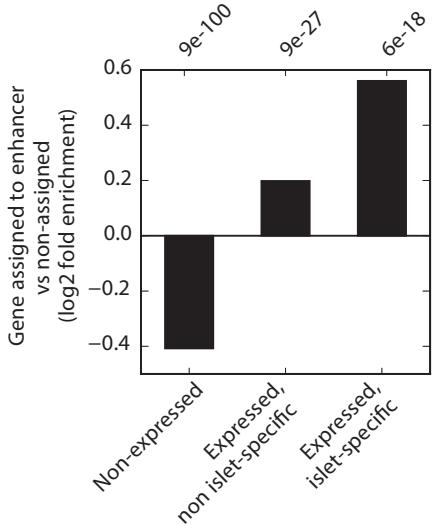


Fig. 32: Enhancer assignments considering chromatin interaction maps accentuate their association with islet-specific expressed genes. For the three different gene expression classes, I computed the log2 ratio for enhancer “assigned genes” vs. genes whose PADs also contained an enhancer but these were not assigned to them (“control genes”). Statistical significance was measured in a chi-square test comparing the frequencies of each gene expression class among “assigned genes” and “control genes” and the resultant p-value is indicated on top of each bar.

Moreover, a complementary analysis showed that assigned enhancer-promoter pairs present better H3K27ac correlations than pairs associated purely by linear proximity (see section 8.23). These results showed that our pcHi-C interaction maps allowed us to define the target genes of a large number islet enhancers based on direct physical interactions, and further allowed inference of an extended list of enhancer target genes.

Identification of PAD features that predict tissue-specific gene expression. To better understand the features of gene-specific regulatory landscapes that are associated with tissue-specific gene expression, I assessed the relationship of a compendium of 15 epigenomic features with the 3 previously mentioned gene expression classes (Fig. 32). These features tried to cover many known aspects of gene regulation, including (a) the presence of regulatory elements (e.g. assigned enhancers), (b) chromatin modifications that are functionally meaningful (e.g. H3K27me3 at TSS) (c) nature of chromatin interactions (e.g. fraction of islet-selective interactions). The interrogated features are listed in Table 7 (see section 8.24 for more details).

Just comparing the distribution of these regulatory features between the tree gene expression classes (non-expressed, expressed but not islet-specific, and specifically expressed in islets) already revealed that some features were especially characteristic of a particular gene class. As expected, promoter activity marked by H3K4me3 was clearly higher at expressed gene promoters, especially for non-islet-specific gene, and almost absent in non-expressed gene promoters (Heintzman et al., 2007). Moreover, the presence of repressive marks, such as H3K27me3 or H3K9me3 (Schwartz and Pirrotta, 2013), in non-pancreatic tissues was coherent with patterns of tissue-specificity promoter activity determined by gene expression (Fig. 58 in methods section 8.24).

I considered several machine learning approaches to identify the most informative features to predict gene expression classes. I specifically explored machine learning algorithms such as Gaussian mixture models, random forests and multiple logistic regression. However, I found that the model obtained with multiple logistic regression was the most informative and accurate one, as assessed with a confusion matrix and different metric scores (see section 8.24 and Fig. 61 for more details). I will therefore only discuss the results I obtained implementing logistic regression.

Logistic regression is a type of machine learning classifier. Therefore, using a compendium of given features, the algorithm tries to find a pattern that correlates with a given categorical classification. This machine learning classifier determines the coefficients (β) for a list of features (χ) that fits a logit function (Fig. 33). This function computes the probability

(p) of a given gene to belong to a certain class (k) (Formula 1). Thus, the higher the coefficient the bigger the weight of a given feature has in the logit function (Bewick et al., 2005).

$$logit(p_k) = \sum(\beta_i * \chi_i)$$

Formula 1: Logit function.

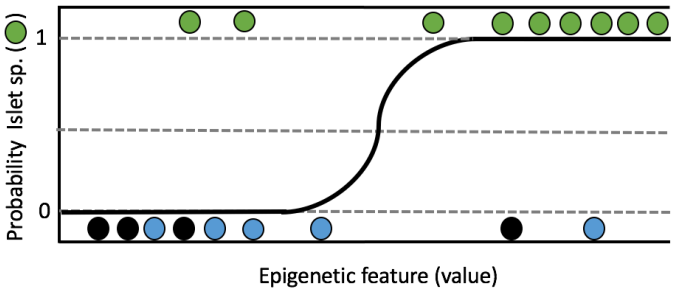


Fig. 33: Logit distribution as gene probability of being islet-specific expressed. Diagram illustrating how a logit distribution may represent the probability of a given gene of being classified as islet-specific expressed based on the values of a given epigenomic feature. Genes classified as islet-specific are indicated as green circles, while expressed non-islet specific are in blue and non-expressed are in black. Genes are sorted in the x-axis based on their value for a given epigenomic feature.

In a logistic regression analysis independence is assumed between the features and it is known that the use of highly correlated features can interfere with the estimation (Bewick et al., 2005). Thus, after computing pair-wise Pearson's correlation among the 15 interrogated epigenomic features, 12 lowly correlated (Pearson's correlation < 0.65) features were kept for a logistic regression analysis (Table 7, Fig. 59 in methods section 8.24).

Table 7: Epigenomic features interrogated in a logistic regression analysis to determine their association with tissue-specific gene expression.

H3K4me3 signal at TSS in human pancreatic islets	✓
H3K27me3 signal at TSS in human pancreatic islets	✓
H3K9me3 signal at TSS in human pancreatic islets	✓
Number of H3K4me3 peaks from (139) other tissues at TSS	✓
Number of H3K27me3 peaks from (139) other tissues at TSS	✓
Number of H3K9me3 peaks from (139) other tissues at TSS	
TSS length determined by CAGE in human islets (bp)	✓
CpG island (CGI) length (kb)	✓
Number of islet pcHi-C interactions	✓
Fraction of islet-selective interactions	✓
Fraction of promoter-enhancer interactions	✓
Distance to the closest TAD borders (kb)	✓

Number of islet enhancers overlapping the PAD	
Number of islet assigned class I-III enhancers	
Number of islet assigned class I enhancers	✓

✓**:** features used in a logistic regression analysis after filtering for correlating features.

An initial logistic regression analysis, comparing all 3 gene expression classes, confirmed that some features were highly characteristic of a particular gene class. However, it also revealed that some features were shared among expressed genes and provided little information to differentiate between islet-specific and non-islet-specific expressed genes (Fig. 61A in methods section 8.24). Therefore, I decided to perform a second analysis comparing exclusively islet-specific and non-islet-specific expressed genes to further characterise features associated with tissue-specific gene expression (Fig. 34).

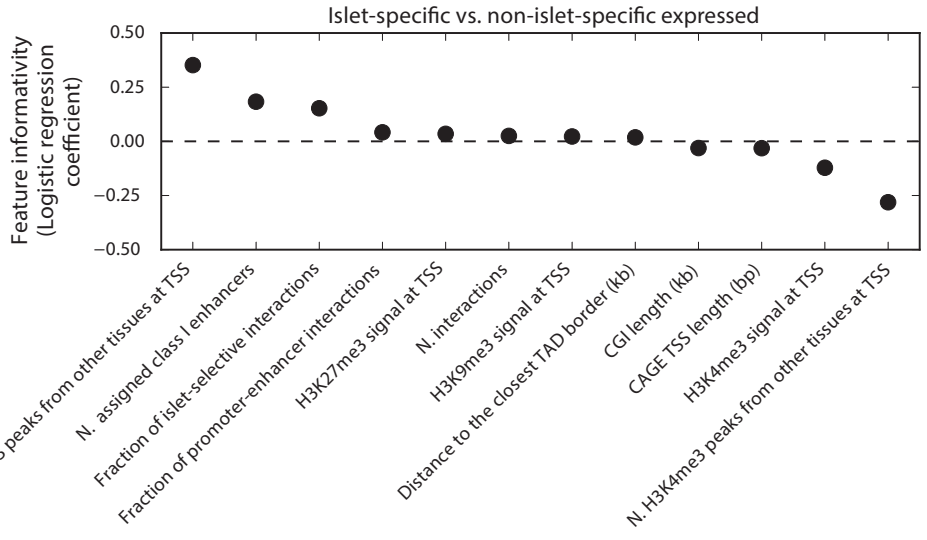


Fig. 34: Feature informativeness to identify islet-specific expressed genes among expressed genes. Dot plot showing the logistic regression coefficient as a measurement of feature's informativeness to differentiate between islet-specific and non-islet specific expressed genes.

This analysis revealed that among all interrogated features, tissue-specificity of repressive marks (namely H3K27me3) at promoters was the most informative feature to predict tissue-specific gene expression. This is interesting, but to some extent redundant, because for any gene that is expressed in islets, the existence of H3K27me3 repressive marks in many non-islet tissues is expected to lead to an islet-specific expression pattern. However, this analysis also showed that among the features that were not directly associated to promoter activity,

the number of assigned class I (H3K27ac and Mediator-rich) enhancers was the most informative one, followed by the fraction of islet-selective interactions (Fig. 34). This unbiased analysis therefore provides a confirmation that enhancer assignments are associated with tissue-specificity gene expression in human islets, and it further points that the tissue-specificity of interactions is also important.

4.2. Identification of enhancer-rich PADs

My observation that the number of assigned Mediator-rich enhancers as well as the fraction of tissue-specific interactions are informative features to predict tissue-specific gene expression (Fig. 34) extends previous associations of tissue-specific gene expression with tissue-specific chromatin structure (Javierre et al., 2016). Moreover, it is also coherent with the notion that tissue-specific genes are located in proximity to enhancer domains.

Previous studies have defined enhancer domains as genomic regions with a high enhancer density and strongly bound by TFs, Mediator or H3K27ac. These domains are known as super-enhancers (Whyte et al., 2013), stretch enhancers (Parker et al., 2013) or enhancer clusters (Pasquali et al., 2014). As mentioned, these studies linked the presence of enhancer domains to tissue-specific expression of nearby genes. Moreover, it has been reported that enhancer domains tend to be enriched in disease-associated genomic variants (Hnisz et al., 2013; Lovén et al., 2013; Pasquali et al., 2014; Whyte et al., 2013).

However, these definitions only group enhancers that are annotated in close linear proximity determined by arbitrary thresholds. Thus, a more accurate enhancer domain definition needs to consider the extent to which enhancer cluster in the 3D space of islet cell nuclei. For example, it is theoretically possible that enhancers that do not cluster in the linear genomic template do cluster in 3D space. Furthermore, previous studies could only attempt to assign enhancer domains to target genes through linear proximity, which recent studies have shown may often be incorrect (Sanyal et al., 2012). As I proposed, this limitation could be overcome through the interpretation of high-resolution chromatin interaction maps.

I therefore considered that our integrative analysis of high-resolution chromatin interaction and *cis*-regulatory maps has the potential to identify 3D enhancer domains by defining PADs with a high enhancer content. PADs were therefore classified in 3 categories based on their number of assigned enhancers: (i) PADs without assigned enhancers named *enhancer-less PADs*; (ii) PADs with at least one assigned class I-III enhancer but less than 3 class I

enhancers, named *enhancer-poor PADs*; (iii) PADs with 3 or more assigned class I enhancers named *enhancer-rich PADs* (Fig. 35).

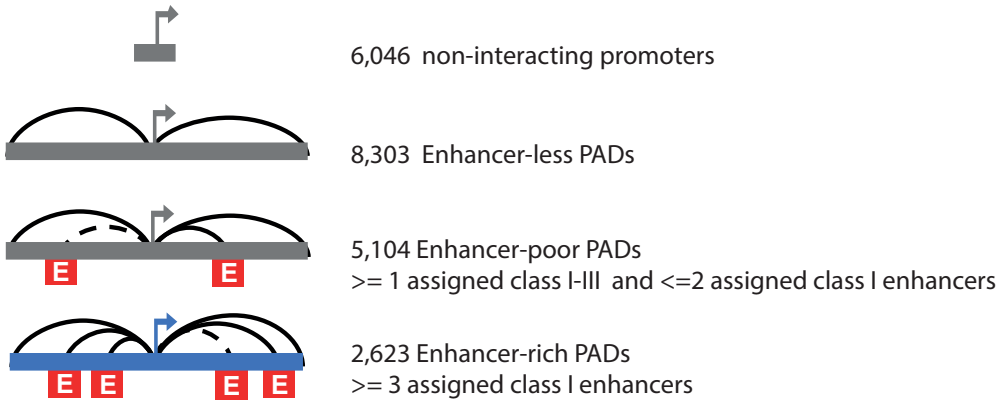


Fig. 35: PAD classification based on enhancer content. Diagram illustrating the PAD classification based on the number of assigned enhancers. Categories from top to bottom are: non-interacting promoter baits; PADs without assigned enhancers (*enhancer-less PADs*); PADs with at least one assigned class I-III enhancer, but less than 3 class I enhancers (*enhancer-poor PADs*), PADs with 3 or more assigned class I enhancers (*enhancer-rich PADs*). For more details regarding the different classes (I-III) of active enhancer see section 8.1. PcHi-C interactions are illustrated as solid black arcs, inferred enhancer-promoter associations as dashed black arcs, PADs are illustrated as grey horizontal bars and enhancers (E) as red boxes. Enhancer-rich PADs are highlighted in blue.

Note that this PAD classification was centred on the assignment of class I enhancers (Fig. 35). Among all 45,685 open chromatin regions that had typical enhancer signatures in pancreatic islet islets, 13,635 class I enhancers were strongly enriched in epigenomic features associated to enhancer activity (H3K27ac and MED1) (Heintzman et al., 2009; Kagey et al., 2010). Thus, I considered that class I enhancers as a highly-confident subset of active enhancers (see section 8.1). Nevertheless, as a control of the proposed categorisation, I interrogate the number of assigned class I-III enhancers per PAD class and their association with gene expression patterns. I determined that different enhancer PAD classes showed expected class I-III enhancer content distributions (Fig. 36A). Moreover, among the different PAD categories only enhancer-rich PADs were frequently associated with islet-specific expressed genes (Fig.36B).

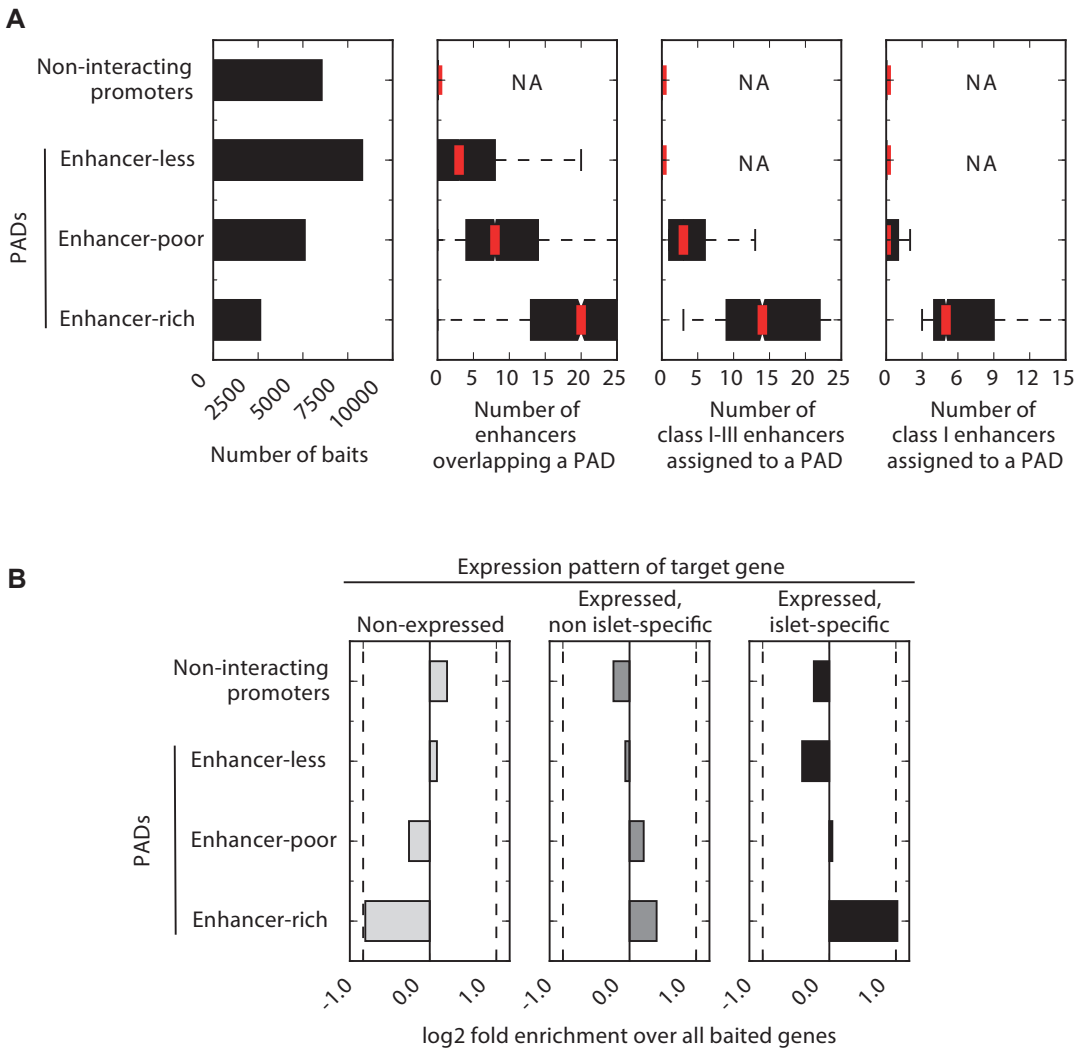


Fig. 36: Characterisation of PADs with different enhancer contents. (A) Panels showing different features per PAD class based on their number of assigned enhancers. (B) Gene expression pattern for different classes of PADs. For comparative purposes the same features are shown for non-interacting baits.

Then, in order to determine their informativeness, enhancer-rich PADs were systematically compared to existing definitions of enhancer domains, namely islet enhancer clusters (Pasquali et al., 2014) and super-enhancers (Whyte et al., 2013). After this analysis, I could determine that:

(i) Enhancer-rich PADs are associated with islet-selective chromatin structures.

Enhancer-rich PADs were enriched in tissue-selective chromatin interactions, as well as in direct interactions occurring between promoters and enhancers (Fig. 37, top panels). This enrichment was similar to the one observed for PADs with assigned enhancer clusters and super-enhancers (Fig. 37, bottom panels). These results were in concordance with the observation that super-enhancers are enriched at frequently interacting regions (FIREs), which are tissue-specific chromatin structures presenting an unexpected large number of chromatin interactions detected by Hi-C (Schmitt et al., 2016). However, by taking a gene-centric approach, our enhancer-rich PADs associate 3D enhancer domains to the likely target gene within the TAD, extending the current knowledge on gene regulation.

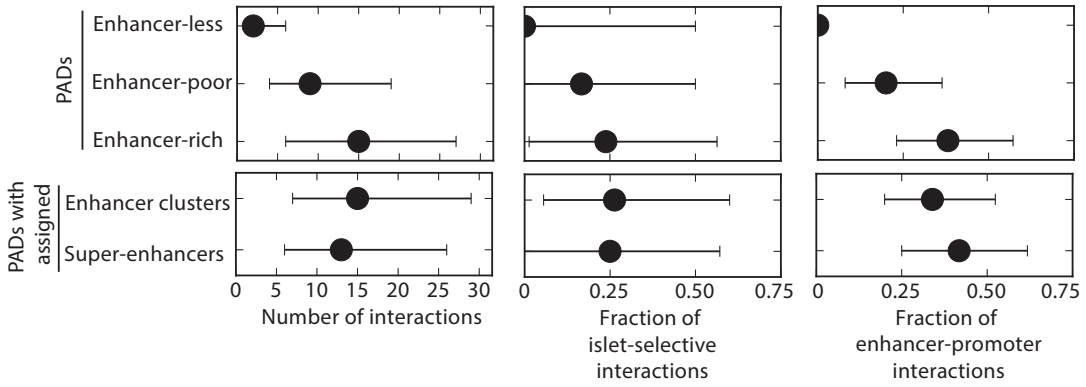


Fig. 37: Enhancer-rich PADs showed similar structural features to PADs associated with enhancer clusters and super-enhancers. Median values are presented as a dot, and interquartile ranges is indicated as a line.

(ii) Enhancer-rich PADs are frequently linked to islet-specific gene expression.

I also observed that, although there were nearly a two-fold increase in the number of genes associated with enhancer-rich PADs compared to the number of genes associated with PADs containing enhancer clusters and super-enhancers, the enrichment of islet-selective expressed genes was comparable in all three classes (Fig. 38). This could be because 3D chromatin interactions maps can capture enhancer clusters in 3D that are not visualised in linear maps, leading to the identification of a larger number of genes linked to enhancer domains without a significant reduction in the enrichment of islet-specific genes.

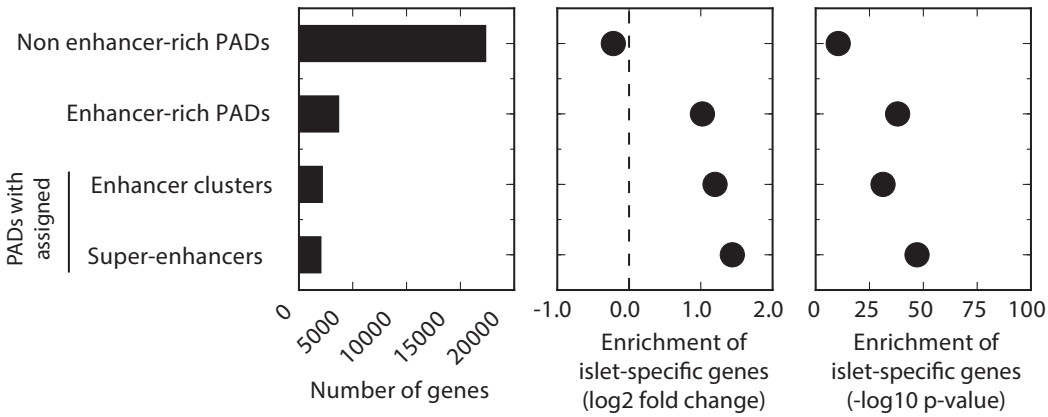


Fig. 38. Enhancer-rich PADs showed a similar enrichment for islet-specific expressed genes as PADs associated to enhancer clusters and super-enhancers. The panels show from left to right: number of genes associated to a given class of PADs, enrichment of islet specific genes computed as a log2 fold change over expected and enrichment statistical significance computed in a hypergeometric test.

(iii) Most enhancer clusters and super-enhancer are assigned to an enhancer-rich PAD.

I next examined the overlap between these regulatory domains and enhancer-rich PADs. I assigned enhancer clusters and super-enhancers to promoters following the strategy previously proposed (Fig. 35), and observed that 86% of enhancer-clusters and 91% of super-enhancers were assigned to enhancer-rich PADs (Fig. 39A). However, a significant proportion of enhancer-rich PADs do not contain an assigned enhancer cluster or super-enhancer (Fig. 39B).

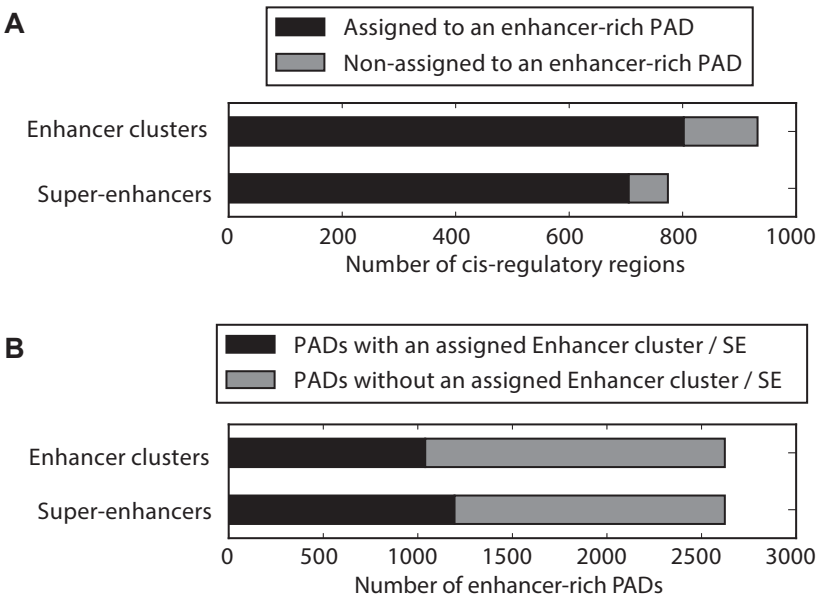


Fig. 39: Most enhancer clusters and super-enhancers were assigned to enhancer-rich PADs. (A) Horizontal bar plots indicating the number of enhancer clusters (EC) or super-enhancers (SE) assigned to enhancer-rich PADs. (B) Horizontal bar plots showing the number of enhancer-rich PADs with an assigned enhancer cluster or super-enhancers compared to those enhancer-rich PADs without an assigned element.

(iv) Enhancers at enhancer-rich PADs are enriched in type-2 diabetes (T2D) and fasting glycemia (FG) GWAS variants.

Our previous study showed that common variants associated with type-2 diabetes and fasting glycemia variation often map to islet enhancer clusters (Pasquali et al., 2014). I therefore performed variant set enrichment (VSE) analysis (Yang et al., 2011) to test if variants in haplotypes associated with type-2 diabetes and fasting glycemia were enriched in enhancers forming enhancer-rich PADs. The results showed that, although there were a much larger number of enhancers in enhancer-rich PADs than enhancers composing enhancer clusters ("clustered enhancers") or super-enhancers, enhancer-rich PADs had a highly significant enrichment for T2D and FG risk associated variants (Fig. 40).

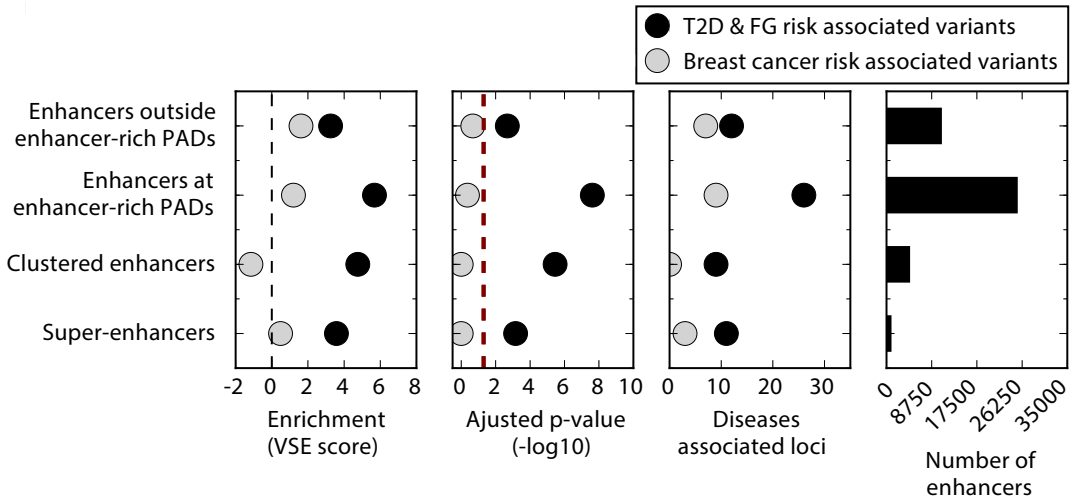


Fig. 40: Enhancers forming enhancer rich PADs showed strong enrichment for T2D and FG risk associated variants, as previously observed for super-enhancers and enhancers forming enhancer clusters. Results from a VSE analysis computing the enrichment of type-2 diabetes (T2D) risk and fasting glycemia (FG) variation associated variants in different sets of islet enhancers. Genomic variants associated to breast cancer were used as a negative control. Enrichment was measured as a VSE score and an adjusted p-value in a –log10 scale. The number of LD blocks with at least one genomic variant overlapping an enhancer is indicated as the number of disease associated loci. Finally, the total number of elements interrogated per enhancer set is showed as a bar plot.

These result show that despite enhancer-rich PAD encompass a larger set of enhancers than previous definitions (Fig. 40 right panel), these 3D enhancer domains present a statistically significant enrichment of disease-risk variants relevant for human pancreatic islets.

In summary, these results define enhancer-rich PADs as meaningful 3D *cis*-regulatory domains. Furthermore, it may suggest that the definition of enhancer clusters (Pasquali et al., 2014) or super-enhancers (Whyte et al., 2013) is not able to fully reflect enhancer gathering in the 3D chromatin space. Therefore, previous enhancer domain definitions based on linear proximity may mask relevant aspects of enhancer regulation.

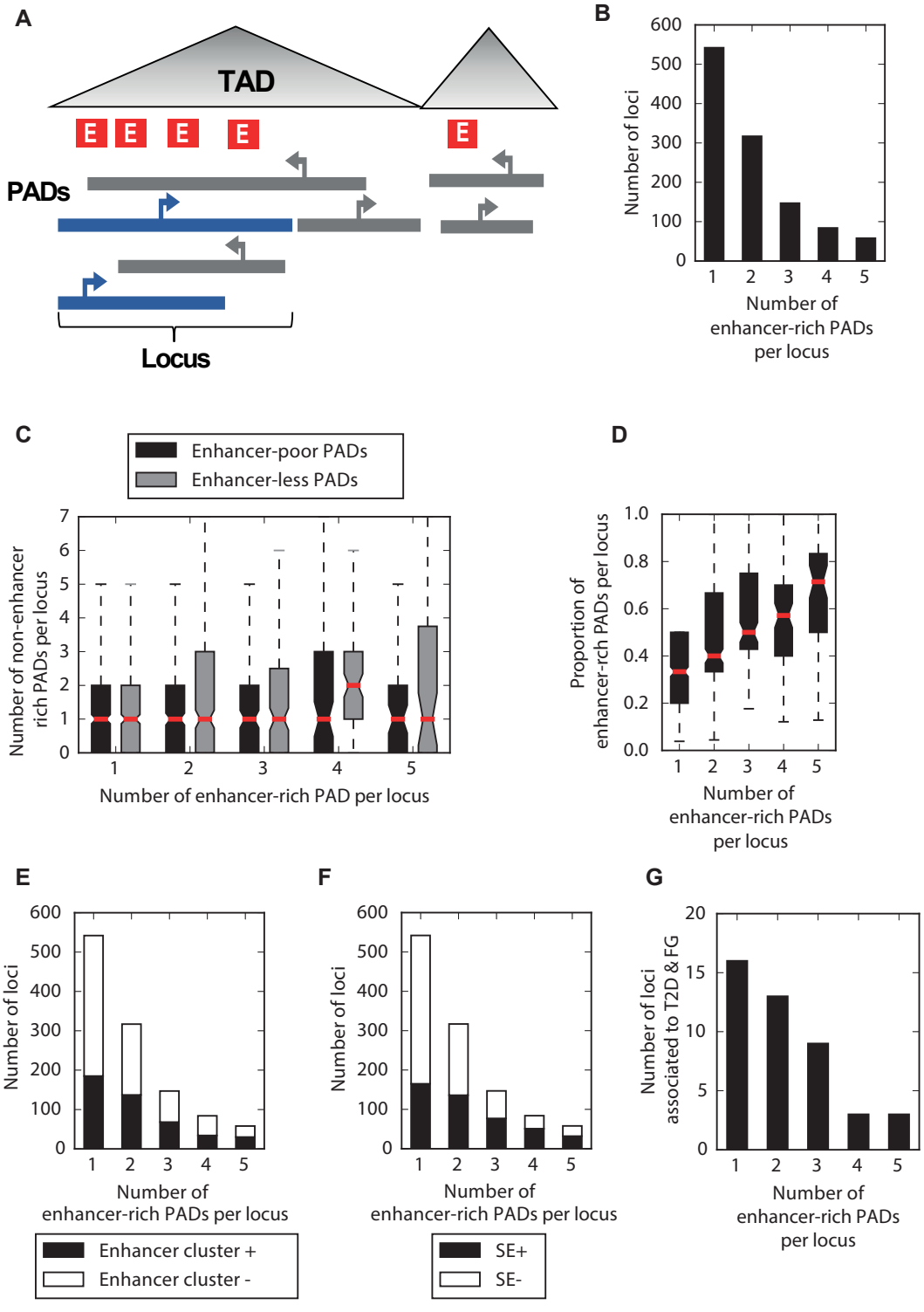
4.3. Clusters of enhancer-rich PADs

Something to take into consideration was the fact that PADs in the same region had a tendency to overlap with each other (Table 6, Fig. 27). Therefore, I wondered if in a locus with a high enhancer density all PADs were classified as enhancer-rich or the locus was composed by PADs with different enhancer contents (Fig. 41A).

To interrogate this question, I considered the regions containing at least one enhancer-rich PAD. In loci with overlapping enhancer-rich PADs, these were merged into a single genomic region (Fig. 41A-B). I could observe that although enhancer-rich PADs overlapped, they also overlapped with enhancer-poor or enhancer-less PADs (Fig. 41B-D). This observation occurred independently of the presence of enhancer clusters, super-enhancers or GWAS associated variants (Fig. 41E-G). Therefore overlapping PADs, or even enhancer-rich PADs that overlap, probably do not reflect topological compartments where all genes are regulated by all enhancers the enhancers present in the compartment.

Nevertheless, interconnectivity between enhancer-rich PADs, either by promoter-promoter interactions or through shared enhancers, is a highly relevant question that to the best of our knowledge has not been addressed. For that reason, we are currently addressing this question by defining enhancer-promoter connectivity networks in PADs that could potentially explain gene expression co-regulation and cell specificity.

Fig. 41: Overlap between enhancer-rich PADs. (A) Schematic illustrating the merge of overlapping enhancer-rich PADs into a single locus. Enhancer-rich PADs are highlighted in blue. (B) Bar plot of the number of loci regarding the number of enhancer-rich PADs per locus. (C) Boxplots showing the number of non-enhancer rich PADs regarding the number of enhancer-rich PADs per locus**.** Enhancer-less are shown with black boxes and enhancer-poor with grey boxes. (D) Boxplots showing the proportion of enhancer-rich PADs among all PADs in a locus regarding the number of enhancer-rich PADs per locus. (E,F,G) Bar plot showing the number of loci containing different the numbers of enhancer-rich PADs in each locus, as in B, differentiating if the locus overlapped with an enhancer cluster (E), a super-enhancer (F) or a LD block of variants associated to type-2 diabetes (T2D) and fasting glycemia (FG) (G). Figure shown in the following page.


A
TAD
E
PADs
Locus
B
Number of loci
Number of
enhancer-rich PADs
per locus
C
Enhancer-poor PADs
Enhancer-less PADs
Number of non-enhancer
rich PADs per locus
Number of enhancer-rich PAD per locus
D
Proportion of
enhancer-rch PADs per locus
Number of
enhancer-rich PADs
per locus
E
Number of loci
Number of
enhancer-rich PADs per locus
Enhancer cluster +
Enhancer cluster -
F
Number of loci
Number of
enhancer-rich PADs per locus
SE+
SE-
G
Number of loci
associated to T2D & FG
Number of
enhancer-rich PADs
per locus

Chapter 5

Experimental identification of non-coding functional variants

A major benefit of characterising the 3D chromatin organisation in a disease-relevant tissue, such as human pancreatic islets, is that it allows us to associate non-coding variants in distal *cis*-regulatory regions with target genes. The functional relevance of our pcHiC interaction maps and promoter-associated domains (PADs) is being further interrogated by my colleagues Irene Miguel-Escalada and Inês Cebola, by applying CRISPR genome editing technology in human beta cells and human pancreatic islets. One of the aims of these experiments is to prove that our 3D chromatin characterisation can be used to infer the effect of non-coding variants on gene expression regulation in an endogenous chromosomal context. However, as previously mentioned (in section 1.7), genetic studies conducted to unveil genomic sequence variations linked to complex diseases may generate large lists of variants with a large proportion of non-causal candidates due to several factors such as linkage disequilibrium. Therefore, I undertook an approach that has the potential of rapidly testing a large number of candidate regulatory variants, which can then be used to select putative causal variants with genome editing tools.

Theoretically, a STARR-seq library (Arnold et al., 2013) containing enhancers with T2D or FG risk alleles could be compared with a second library containing the non-risk alleles (see section 1.4-Enhancer activity reporter assays). This comparison would allow me to identify genomic variants able to modulate enhancer activity. Thus, based on this criterion, we could greatly reduce the number of likely functional variants.

To test these genomic variants in a human beta-cell line (EndoC-βH3, Benazra et al., 2015), the original human STARR-seq vector needs to be modified to overcome certain technical limitations. For example, EndoC-βH3 are highly difficult to be transfected by lipofectamine or nucleo-fection. Therefore, I determined that the best approach would be to combine STARR-seq with viral infection. I also designed an indexing method to enable quantitative quantification of non-clonal reads from independent pooled libraries. As STARR-seq was not implemented previously in our lab, I considered reasonable to perform a small-scale experiment in an easier cellular model as a proof of concept.

In our previous work, we found that a significant number of predicted human pancreatic islet active enhancers were functional in a mouse beta-cell line named MIN6 (Ishihara et al., 1993; Pasquali et al., 2014). Thus, I selected a small collection of regions (aprox. 650bp each) to be tested by STARR-seq in MIN6. This small collection was formed by 45 predicted active enhancers, 4 inactive enhancers, 5 open chromatin regions lacking other epigenomic marks. Moreover, I included 15 closed chromatin regions that are highly unlikely to have enhancer activity as negative controls. This small collection of fragments was cloned into the original human STARR-seq vector and tested in MIN6. The result of two independent biological replicates indicated that only 3 out of the 45 predicted active enhancers could be distinguished from the negative controls by STARR-seq (Fig. 42).

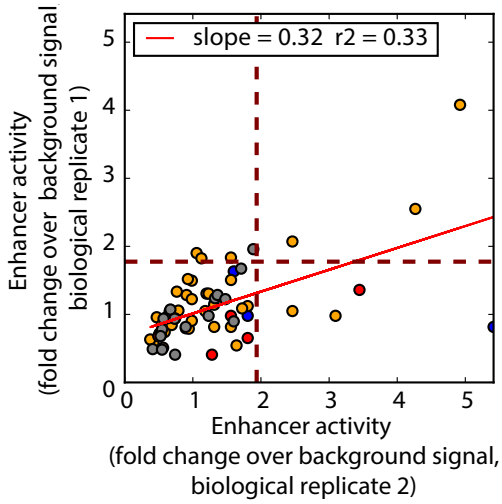


Fig. 42: Enhancer activity detected by STARR-seq. Dot plot showing reproducible enhancer activity detected in two STARR-seq assays. Predicted active enhancers are indicated in orange, inactive enhancers in blue, open chromatin regions lacking other epigenomic marks in red and close chromatin regions in grey. Background signal was determined as the mean signal from all negative control regions. Minimum enhancer activity was determined as two standard deviations of the mean background signal, indicated with dashed maroon lines. Correlation between biological replicates is indicated as r2 and slope of a linear correlation.

Surprised by the fact that few predicted enhancers acted such as by STARR-seq, contradicting our previous work (Pasquali et al., 2014), I decided to perform some validations by traditional luciferase reports assays. Thus, 12 predicted enhancers, including those 3 that were validated by STARR-seq (Fig. 42, Fig. 43A), were cloned separately in the pGL4.23-GW (Pasquali et al., 2014; addgene #60323) and tested in the same cell line. Additionally, I included 4 closed chromatin regions also tested by STARR-seq to determine

the background signal. In opposition to my STARR-seq results (Fig. 42, Fig. 43A), 10 out of 12 predicted enhancers acted such as in a traditional luciferase reporter assay (Fig. 43B).

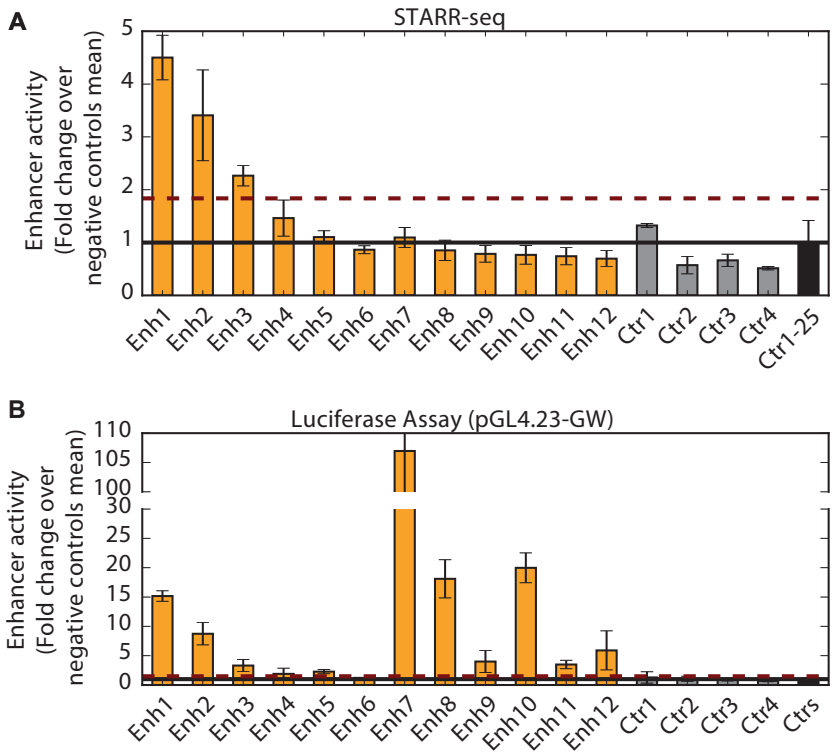


Fig. 43: Comparison between STARR-seq and Luciferase reporter assay. Bar plot showing mean and standard deviation for 2 biological replicates. Predicted active enhancers are indicated with orange bars. Close chromatin regions used as negative controls to determine the background signal are depicted as grey bars. The average negative control signal is represented as a black bar. Minimal enhancer activity was determined as two standards deviation (maroon dashed line) over the control mean (black line).

By comparing the two constructs, the human STARR-seq vector and the pGL4.23-GW, two features appeared as clear differences (Fig. 44). These features were (i) the promoter used in the construct and (ii) the relative position between the promoter and the tested enhancers. It has been shown by STARR-seq that enhancers behave differently with different types of promoters. A study done in Stark's lab showed that developmental enhancers only act such as in the presence of developmental promoters. The same was true for enhancers associated with house-keeping genes (Zabidi et al., 2014). On the other hand, it is well established that enhancer activity is independent on its orientation or relative position to the promoter (Banerji et al., 1981). Thus, I decided to assess the effect of both features.

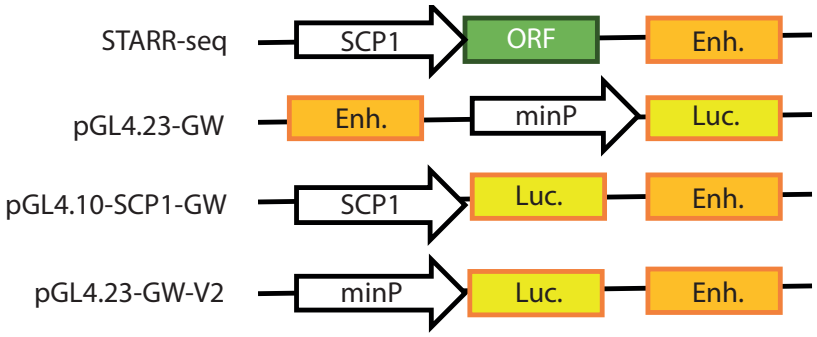


Fig. 44: Diagram of the different constructs used to measure enhancer activity. Diagram indicating the elements and their order in 4 enhancer activity reporter constructs. From top to bottom those constructs are: STARR-seq, pGL4.23-GW, pGL4.10-SPC1-GW and pGL4.23-GW-V2. The tested enhancer is illustrated as an orange box. Luciferase reporter gene is indicated as a yellow box and the open reading frame (ORF) present in the STARR-seq vector in a green box. The two types of promoters tested, SCP1 and minP, are indicated as a white arrow. SCP1 is a synthetic core promoter with basal activity and minP is a minimal promoter with a very weak basal activity.

To test the effect of the enhancer's relative distance to the promoter and the promoter type, I created two new vectors for traditional luciferase reporter assays. Both constructs contained a Gateway cloning cassette (GW) downstream of the reporter gene. Each construct contained a different promoter, the SCP1 synthetic core promoter used in the human STARR-seq vector or the minimal promoter (minP) present in the pGL4.23. These vectors were named pGL4.10-SCP1-GW and pGL4.23-GW-V2 respectively (Fig.44). As none of the two new vectors were used in previous assays, I validated their capacity of detecting enhancer activity using the CMV enhancer as a positive control. These two new constructs were used as backbones to assess the same 12 predicted active enhancers previously tested (Fig. 45).

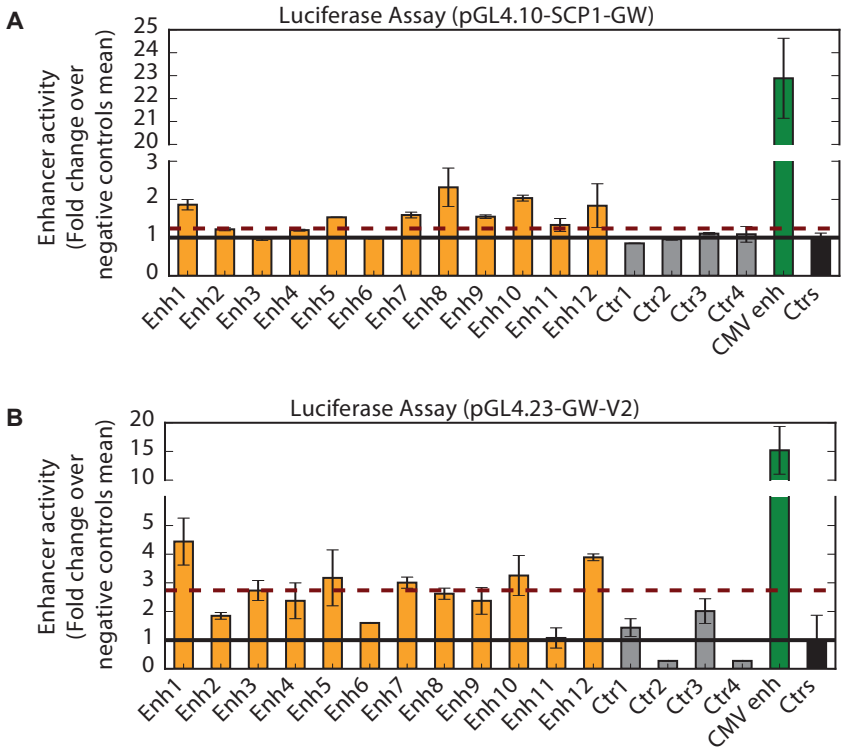


Fig. 45: Impact of different promoter types over enhancer activity luciferase reporter assays. Bar plot showing mean and standard deviation for 2 biological replicates. Predicted active enhancers are indicated with orange bars and closed chromatin regions used as negative controls to determine the background signal as grey bars. The average negative control signal is represented as a black bar. Minimal enhancer activity was determined as two standards deviations (maroon dashed line) over the control mean (black line). CMV enhancer (green) is used as positive control.

Table 8: Summary of the 4 different enhancer activity reporter assays.

Assay	Vector	Enh1	Enh2	Enh3	Enh4	Enh5	Enh6	Enh7	Enh8	Enh9	Enh10	Enh11	Enh12
STARR-seq	STARR-seq	■	■	■									
Luciferase	pGL4.23- GW	■	■	■		■		■	■	■	■	■	■
Luciferase	pGL4.10-SPC1-GW	■				■		■	■	■	■		
Luciferase	pGL4.23-GW-V2	■						■					■

(fragments which showed enhancer activity in a particular assay are indicated with grey boxes)

In summary, although a large proportion of tested enhancers (9 out of 12) were validated in at least 2 different assays (Table 8), each assay gave inconsistent results when compared to the others. Therefore, despite the difficulty of extracting a clear conclusion from them, these results may suggest that several factors including the read-out (luciferase activity or RNA-seq) or the promoter used (minP or SCP1) have a strong impact over the assay performance, at least in my hands. Due to the inconsistent results and time restrictions I abandoned the possibility of testing a large collection of enhancers and genomic variants by a MPRA in benefit of more promising analyses.

Chapter 6

Discussion and prospects

During my PhD project, I interrogated high-resolution chromatin interaction maps in human pancreatic islets. These maps characterise 3D chromatin organisation at different levels. First, our pcHi-C interaction map allowed me to observe interactions between gene promoters and distal genomic regions with high-resolution. A comparison between chromatin organisation in human pancreatic islets and distant cellular lineages led us to identify islet-selective chromatin structures and the epigenomic factors underlying them. Second, as pcHi-C maps reflect both tissue-invariant structural interactions and tissue-selective regulatory loops, I defined TAD-like structures in human pancreatic islets. These islet TAD-like structures recapitulated known aspects of tissue-invariant TAD compartmentalisation. Third, I defined promoter's *cis*-regulatory niches, that I named promoter-associated domains (PAD) (Fig. 27 in section 4.1). This was possible as pcHi-C interrogates chromatin folding from a promoter centric perspective at a very high resolution. Finally, an integrative analysis of these pcHi-C maps and PADs with islet *cis*-regulatory maps allowed me to identify genes under enhancer regulation (Fig. 35 in section 4.2). Enhancer regulation is especially interesting in disease-relevant tissues, like human pancreatic islets, not only because it has been associated to tissue-specific gene regulation (Heinz et al., 2015; Maston et al., 2006) but also because it has been shown that enhancers are enriched in disease associated genomic variants (Calo and Wysocka, 2013; Hnisz et al., 2013; Lovén et al., 2013; Pasquali et al., 2014). Therefore, picturing enhancer-promoter communication through 3D chromatin organisation can help us to better understand gene expression regulation and hypothesise on the impact of non-coding variants associated with complex traits.

- **Interpretation of pcHi-C maps**

Although promoter capture Hi-C (pcHi-C) (Mifsud et al., 2015) achieves a higher resolution than standard Hi-C (Lieberman-Aiden and Berkum, 2009), it is still subject to technical limitations generally associated to chromatin conformation assays. A major technical

limitation is the background correction to detect statistically significant interactions. To address this problem, researchers from M. Spivakov 's lab developed CHICAGO, an algorithm that based on a background modelling determines reliable chromatin interactions (Cairns et al., 2016).

To determine the resolution of pcHi-C maps generated using CHICAGO, I interrogated the distance between an interacting region and the closest epigenomic factors at HindIII fragment level, and its enrichment over the expected distribution. I found that while the interrogated epigenomic factors were precisely enriched only at baited HindIII fragments, this was not true for non-baited interacting regions. I determined that the closest interrogated epigenomic factors were enriched at non-baited interacting regions and this enrichment was kept on the adjacent genomic fragments following a decreasing distribution (Fig. 25 in section 3.6). Based on these results, I determined that it was reasonable to extend non-baited interacting regions encompassing the adjacent HindIII fragments. This would accentuate our capacity to associate promoters and distal cis-regulatory elements through chromatin interactions, providing a more accurate interpretation of pcHi-C maps.

- **Identification of factors underpinning islet-selective chromatin structures**

It is well established through Hi-C (Lieberman-Aiden and Berkum, 2009) that the human genome is folded in the 3D space in globular compartments, named topological associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012). It has been shown that TAD compartments are important for gene transcription regulation as co-regulated genes tend to be contained in the same TAD (Le Dily et al., 2014; Nora et al., 2012) and the disruption of TAD borders leads to aberrant gene expression (Franke et al., 2016; Lupiáñez et al., 2015, 2016). However, it has been determined that TAD compartmentalisation is highly conserved among cell types (Dixon et al., 2012; Schmitt et al., 2016). Therefore, it is likely that TAD compartmentalisation is mainly driven by tissue-invariant chromatin interactions. However, it is difficult to imagine how these tissue-invariant chromatin structures could be involved in tissue-specific gene expression regulation.

Recent advances on C-based methods allowed the generation of chromatin interaction maps with higher-resolution, which are able to characterise intra-TAD organisation. This gain on resolution led to the identification of tissue-selective chromatin loops and their association to tissue-specific expressed genes (Javierre et al., 2016; Phillips-Cremins et al., 2013; Rao et al., 2014). These studies suggested that tissue-specific gene regulation could be partially driven by tissue-selective chromatin interactions. However, little is known about the epigenomic factors underlying these tissue-selective chromatin structures.

As part of my PhD project, I identified some epigenomic factors that could be involved in the formation of tissue-selective chromatin structures. First, we identified islet-selective chromatin interactions through a comparison between our pcHi-C map in human pancreatic islets and distal cellular linages (Fig. 17 and Fig. 21 in section 3.4). As occurred in previous studies, these islet-selective chromatin structures were frequently associated to islet-specific expressed genes (Fig. 18 in section 3.4). Later, I integrated this information with epigenomic datasets that were also generated in human pancreatic islets. I was then able to show that islet-specific CTCF binding sites and especially Mediator-bound enhancers were enriched at islet-selective chromatin structures (Fig. 19, Fig. 20 and Fig. 21 section 3.4). Both epigenomic factors have been associated to the binding of lineage-determining TFs (LDTFs) (Fig. 26 in section 3.6) (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). This lead me to hypothesise that the co-occurrence of LDTFs, CTCF and/or Mediator at *cis*-regulatory regions, especially at enhancers, could be involved in the formation of tissue-selective chromatin interactions. Additionally, previous studies have associated the presence of enhancers or tissue-selective structures with tissue-specific gene expression. Therefore, my results may act as a bridge between the two-previous independent observation showing that enhancer communication may partially occur through tissue-selective interactions.

- **Promoter-associated domains, gene-specific niches of cis-regulatory elements**

As already mentioned, previous studies characterised the chromatin conformation in different tissues by Hi-C. These studies reported that the genome is compartmentalised in tissue-invariant domains named TADs. However, there is no published Hi-C dataset generated in human pancreatic islets. Therefore, although pcHi-C maps only reflect

promoter centric chromatin interactions, I attempted to define chromatin compartments using our islet pcHi-C map. I was able to determine that pcHi-C maps reflected TAD-like domains, which exhibit known features of tissue-invariant compartmentalisation (Fig. 22, 23 in section 3.5).

Moreover, due to their high-resolution, pcHi-C maps allowed to study 3D chromatin organisation within TAD compartments. I saw that promoters had their own chromatin interaction landscape, and that although they sometimes covered an entire TAD, they could often span considerably shorter distances (Fig. 27, 28 in section 4.1). I defined these genomic regions as promoter-associated domains (PADs). I also observed that chromatin states at PAD space were coherent with gene expression patterns and more informative than at TADs (Fig. 29, 30 in section 4.1). These results support my hypothesis that PADs encompass the cis-regulatory 3D space associated to a given promoter and that these territories do not necessary correspond to tissue-invariant TADs, although sometimes they can cover the same genomic space. These results are clearly exemplified in the *KCNJ11* locus (Fig. 27 in section 4.1), for which the PAD is clearly smaller than the corresponding TAD. *KCNJ11* PAD encompasses most of the enhancers in the locus and at the same time that it trims other intra-TAD genomic segments that, based on their epigenomic "flavour", are less likely to have a regulatory effect of the on *KCNJ11*.

In summary, characterisation of high-resolution pcHi-C maps led me to identify promoter-associated domains. These domains reflect promoter centric cis-regulatory landscapes defined by both tissue-invariant and tissue-selective chromatin interactions. The identification of promoter-associated domains provides a highly informative framework to identify target genes of distal cis-regulatory elements and non-coding genomic sequence variants with a likely impact over gene expression regulation.

- **Enhancer-promoter associations based on chromatin organisation differentiate bona fide enhancers' targets from other genes in the vicinity of enhancers**

I observed that 40% of all mapped active enhancers were present in promoter-interacting regions. This allowed a direct association between enhancers and promoters through

experimentally detected chromatin interactions. However, it also implied that the remaining 60% of non-interacting enhancers could not be associated to a promoter based on statistically significant interactions. Moreover, it raised the question of whether the lack of possible enhancer-promoter interactions could be due to technical and/or biological reasons.

There is evidence indicating that enhancer-promoter communication occurs in bursts (Fukaya et al., 2016), which could fit in a "hit-and-run" model (Banerji et al., 1981; Freire-Pritchett et al., 2017; Varala et al., 2015). Therefore, it suggests that enhancer-promoter interactions may be highly dynamic and in a constant process of formation and disassociation. Although I studied a population of cells, it is possible that these interactions are not permanent, and very dependent on physiological states which can increase the likelihood that specific sets of enhancers establish frequent interactions. Furthermore, some enhancer-promoter interactions could be only present in a small proportion of the cells at any given time. On the other hand, it has been described recently that different chromatin interaction factors (CTCF and Cohesin) have different residence and re-binding times (Hansen et al., 2016). Thus, it suggests that interactions driven by different interacting factors may have different dynamics. Additionally, it is known that structural proteins, such as CTCF, bind in a much more stable way than TFs (Chen et al., 2014; Hansen et al., 2016; Mazza et al., 2012). These observations led me to hypothesise that structural interactions are probably more stable while enhancer-promoter interactions may be more dynamic. This implies that it is practically impossible to detect all enhancer-promoter interactions in a "fixed snapshot" (obtained with most conformational techniques), due their high dynamism and dependence on the physiological state.

The inability to detect dynamic interactions could be further accentuated due to technical reasons, such as the requirement of a minimum number of reads to detect statistically significant interactions. An interaction present in a small subset of cells would get a lower sequencing coverage than a constitutive structural interaction present in most cells, thus enhancing the detection of constitutive structural interactions over dynamic regulatory interactions.

Nevertheless, the integration of our high-resolution chromatin interaction and *cis*-regulatory maps in human pancreatic islets allowed us to define an enhancer-promoter association strategy. Considering islet 3D chromatin organisation determined by pcHi-C, the identification of PADs and published knowledge regarding enhancer-promoter communication, I attempted to associate enhancers and promoters that were not linked by direct interactions. Based on this strategy I could associate 80% of all annotated enhancers in our islet-regulome with at least one promoter (Fig. 31 in section 4.1). Furthermore, I could determine that our assignment strategy accentuated the association between enhancer and tissue-specific expressed genes (Fig. 32 in section 4.1), which is coherent with the notion that tissue-specific gene regulation is partially driven by enhancer regulation (Heinz et al., 2015; Maston et al., 2006). These results suggest that high-resolution chromatin interaction maps allow better enhancer-promoter associations than the assignments based on linear proximity.

Although my attempt to assigning enhancer to promoter may provide an improvement over associations based on linear proximity (Fig. 32 in section 4.1), this type of assignments requires further validation. My colleagues are interrogating our high-resolution chromatin interaction maps and enhancer-promoter assignments further, through computational and experimental analyses. To do so, they are computing enhancer-promoter correlations, such as human islet allelic-specific enhancer/gene expression correlations (Cowles et al., 2002; Gaur et al., 2013; Yan, 2002), linkage of SNPs in regulatory elements with expression quantitative trait loci (eQTL) (Ardlie et al., 2015; Fadista et al., 2014) or correlation between enhancer-associated histone marks and gene expression among different human pancreatic samples or human tissues. Moreover, my colleagues are experimentally establishing the effect of sequence modification at cis-regulatory elements on gene transcriptional levels of predicted target genes. Preliminary results from Ignasi Morán already highlighted the fact that PAD organisation is coherent with enhancer-promoter regulation detected through allelic-specific gene expression and that it is more informative than TAD segmentation (data not shown). Furthermore, Inês Cebola has already carried out a first set of experiments showing that CRISPR deletion of islet enhancers results in deranged expression of target genes predicted by pcHi-C (data not shown).

We expect to observe that the enhancer-promoter assignments that I have computed considering pcHi-C interaction maps truly reflect functional enhancer-promoter communication and that these assignments provide a clear improvement over associations based on genomic linear proximity and/or TAD compartmentalisation. I also expect that the results from both computational and experimental studies will provide further validation to my hypotheses that PADs reflect enhancer-promoter communication and that our high-resolution chromatin interaction maps can be used to precisely assign distal cis-regulatory elements and regulatory T2D variants to target genes.

- **Enhancer-rich PADs reflect 3D enhancer gathering and are relevant *cis*-regulatory domains for tissue-specific gene transcription**

It has been shown that among all active enhancers in a specific cellular context some of them may have a particularly relevant role driving tissue-specific gene expression regulation. These *cis*-regulatory elements have been named enhancer clusters and super-enhancers. Both, enhancer clusters and super-enhancers, have been identified based on definitions that consider enhancer gathering based on linear proximity and abundancy of key regulatory proteins such as lineage-determining TFs or Mediator (Gaulton et al., 2010; Parker et al., 2013b; Pasquali et al., 2014; Whyte et al., 2013).

In order to identify large groups of enhancers that gather in 3D space, I searched for PADs with a high enhancer content, which were named enhancer-rich PADs (see section 4.2). Interestingly, enhancer-rich PADs exhibited features associated to enhancer clusters and super-enhancers. Those features were: (i) association with tissue-specific expressed genes, (ii) formation of tissue-selective chromatin structures and (iii) enrichment for disease associated genomic variants (Fig. 37-40 in section 4.2). This indicates that enhancer-rich PADs may reflect a 3D chromatin structure relevant for gene expression regulation.

Moreover, most enhancer clusters and super-enhancers were assigned to enhancer-rich PADs. However, our high-resolution chromatin maps allow us to group enhancer clusters and super-enhancers with additional enhancers that they were not previously associated with through linear proximity. Furthermore, a significant proportion of enhancer-rich PADs

do not contain enhancer clusters or super-enhancers (Fig. 39 in section 4.2). Thus, although enhancer-rich PADs have similar properties to enhancer clusters and super-enhancers, enhancer-rich PADs group a larger number of enhancers than previous definitions based on linear proximity. This suggests that as enhancer-rich PADs were defined based on 3D organisation, they may reflect enhancer gathering more accurately than enhancer-clusters and super-enhancers, which rely exclusively on linear proximity.

I showed that there was a high overlap between PADs located in the same locus (Table 6 and Fig. 27 in section 4.1). I also showed that enhancer-rich PADs overlap with other enhancer-rich PADs. Furthermore, I noticed that this overlap did not reflect topological compartments in which all genes are regulated by many enhancers as those loci also contained enhancer-less and enhancer-poor PADs (Fig. 41 in section 4.3). However, whether overlapping-enhancer rich PADs are interconnected though promoter-promoter interactions or shared enhancers remains to be interrogated. This is going to be answered through a network analysis, conducted by Delphine Rolando, which could provide novel knowledge on how 3D enhancer-promoter regulatory circuitries modulate gene expression.

In summary, the integration and interpretation of high-resolution chromatin interaction and cis-regulatory maps allowed me to define 3D enhancer domains associated to gene promoters, named enhancer-rich PADs. These enhancer domains seem to be especially important for tissue-specific gene transcription regulation and to understand the molecular mechanism underlying major disease such as diabetes.

- **Tissue-selective chromatin structures associated to tissue-specific gene repression**

Although there is a clear correlation between the presence of islet-selective chromatin structures and islet-specific expressed genes, it does not mean that tissue-selective chromatin interactions were exclusively observed near islet-specific expressed genes. In fact, I observed the presence of islet-selective chromatin interactions in Polycomb-rich areas (data not shown). Based on these observations, it is likely that tissue-selective chromatin structures could be also generated through tissue-selective gene repression. Other tissue-

selective interactions were not associated with discernible epigenomic features, and warrant further attention in the future.

- **Association of diabetes susceptibility regulatory variants with target genes**

The identification of target genes of GWAS signals represents one of the most challenging goals of current human genetics. The high-resolution chromatin interaction maps characterised in this PhD project give me the opportunity to systematically identify likely target genes of diabetes associated non-coding variants annotated in distal cis-regulatory regions. This will supply a highly informative piece of information for the scientific community devoted to give insight into the genetic factors underlying diabetes.

- **Dynamic nature of chromatin organisation**

So far, I have examined chromatin interactions in islets cultured under standardised conditions. However, it is possible that many components of 3D chromatin structure are dynamic. Characterising chromatin organisation in pancreatic islets exposed to different metabolic stimuli, such as different glucose levels, could help us to better understand the dynamic changes on gene transcription occurring in this tissue. Based on previous work (Le Dily et al., 2014), I hypothesise that chromatin organisation in TADs (Dixon et al., 2012) would not present a drastic change under metabolic stress. However, the epigenomic state of invariant TADs could be different. Therefore these TADs would be switching between what it is known as A and B compartments (Lieberman-Aiden and Berkum, 2009). I have shown that epigenomic states, especially the presence of active enhancers in combination with the presence of tissue-selective chromatin interactions are associated with tissue-specific gene expression. Therefore, I would expect that epigenomic changes driven by metabolic stimuli will concur with the presence of stimulus-selective intra-TAD structures associated to relevant stimulus-responsive genes.

Supporting this reasoning, a recent piece of work was published comparing pcHi-C interactions maps in hESC and in ESC-derived neuroectodermal cells (NESC) (Freire-Pritchett et al., 2017). This work revealed that pcHi-C interactions, in combination with *cis*-regulatory

maps defined based on epigenomic marks are able to reflect chromatin conformation changes occurring during cell differentiation. These results already provide an improvement over similar work done using Hi-C (Dixon et al., 2015; Schmitt et al., 2016) which struggled to identify major changes on 3D chromatin organisation. Nevertheless, most of the work done until now relies on identifying presence or absence of statistically significant interactions among fairly different cellular states. Therefore, it is still unknown if the current interpretation of high-resolution chromatin maps can reflect the more subtle changes that may occur in a single cell type under different stimuli.

- **Discrepancies between episomal enhancer reporter assays**

One of the components of my thesis was intended to implement STARR-seq to test enhancer variants (Arnold et al., 2013; Vockley et al., 2015). I observed a low concordance with conventional enhancer reporter assays, which led me to question the biological and/or technical factors behind this observation (see chapter 5).

It has been observed that not all promoters respond equally to the activity of a given enhancer. There is evidence showing that some promoters only respond to the regulation of a given type of enhancer. For example, work interrogating the relationship between housekeeping and developmental *cis*-regulatory elements revealed that promoters only respond to enhancers that belong to the same category of *cis*-regulatory elements (Zabidi et al., 2014). Therefore, in concordance with my observations, it is possible that enhancer reporter assays based on different promoters provide different results.

Moreover, it is possible that due to technical reasons, enhancer activity reporter assays' sensibility depends on several factors such as the read-out system or the promoter's basal activity.

Additionally, it could be that enhancer reporter assays in which an enhancer is cloned few bp upstream of a target promoter measure slightly different aspects of enhancer activity than those in which the enhancer is cloned a few kbs downstream (Fig. 44 in chapter 5). In the first scenario, where the enhancer is cloned in close linear proximity, it is probable that

as far as the tested enhancer keeps its capacity to recruit TFs this would be enough to modulate the tested promoter. Meanwhile, if the enhancer is in a distal position, it also needs to retain its capacity to loop and communicate with the promoter. Therefore, although reporter assays with distal enhancer cloning sites may be more faithful with enhancer biology, they may also have a higher false negative rate. However, to the best of my knowledge there is not any experimental evidence that supports this reasoning.

Nevertheless, it is important to remember that most enhancer reporter assays, especially episomal assays, do not test enhancer activity under their chromosomal context. Therefore, it is difficult to determine whether a method is more reliable than another. For that reason, I would not be surprised if in a near future enhancer activity studies mainly rely on techniques such as CRISPR genome editing (Mali et al., 2013; Xie et al., 2017) or large studies interrogating the correlation between histone marks at *cis*-regulatory elements and gene expression.

Chapter 7

Conclusions

- **Tissue-selective chromatin interactions are associated with tissue-specific gene transcription.** A systematic analysis revealed a clear correlation between tissue-specific gene expression patterns and the formation of promoter-centric islet-selective 3D chromatin structures (Fig. 18 in section 3.4).

- **A subset of tissue-selective chromatin interactions is probably formed due to collaborative binding of lineage-determining TFs (LDTFs) and other regulatory proteins, such as Mediator and CTCF, at *cis*-regulatory elements.** It has been suggested that chromatin organisation is formed by tissue-invariant structural interactions and tissue-selective regulatory interactions (Krijger and de Laat, 2016). An extensive characterisation of high-resolution chromatin interaction maps in human pancreatic islet, in combination to other epigenomic datasets, allowed me not only to identify islet-selective chromatin interactions but also to elucidate novel features regarding their formation. My results indicate that tissue-selective chromatin interactions are likely partially driven by Mediator-bound enhancers, as well tissue-specific CTCF bound regions associated with binding sites of LDTFs (Fig. 19, 20 in section 3.4 and Fig. 26 in section 3.7). These tissue-selective chromatin structures may be allowing communication between distal cis-regulatory elements, such as enhancers, and promoters.

- **Promoter-associated domains (PADs) are likely to encompass most of promoters' *cis*-regulatory space.** Chromatin organisation is known to be compartmentalised in tissue-invariant topological associating domains (TADs) (Dixon et al., 2012; Schmitt et al., 2016). Our high-resolution chromatin interaction maps allowed me to define promoter-centric intra-TAD chromatin territories, named promoter-associated domains (PADs) (Fig. 30 in section 4.1). My analysis suggests that PAD organisation reflects the 3D chromatin conformation through which promoters communicate with other *cis*-regulatory elements (Fig. 29, 30 in section 4.1). I also noticed that PADs tend to overlap

(Fig. 27 in section 4.1), indicating that they do not represent closed topological compartments as it could occur with TADs.

- **High-resolution chromatin interactions enable linkage of islet enhancers to islet expressed genes.** The integration of islet pcHi-C interactions, PADs and islet cis-regulatory maps allowed the association of a significant proportion of annotated enhancers (Fig. 31 in section 4.1). My result suggests that enhancer-promoter assignments considering high-resolution chromatin interaction maps provide an improvement over previous associations that rely on genomic linear proximity (Fig. 32 in section 4.1).

- **Enhancer-rich PADs as tissue-specific regulatory domains.** I could identify promoter-associated domains (PADs) with a high enhancer content. These domains were named enhancer-rich PADs (Fig. 36 in section 4.2). A systematic comparison revealed that enhancer-rich PADs exhibited features associated to enhancer clusters (Pasquali2014) and super-enhancers (Whyte et al., 2013). These features were: (i) association with tissue-specific gene expression, (ii) correlation with tissue-selective chromatin structures (Schmitt et al., 2016) and (iii) enrichment for disease-associated genetic variants (Hnisz et al., 2013; Pasquali et al., 2014) (Fig. 37-40 in section 4.2). However, it was noticeable that enhancer-rich PADs were composed by a larger set of enhancers than using previous definitions that rely on linear proximity. Therefore, as enhancer-rich PADs were defined considering high-resolution chromatin interaction maps and do present the same features as enhancer clusters or super-enhancers, it is reasonable to hypothesise that these enhancer-rich PADs do reflect enhancer-promoter regulation more faithfully than previous definitions based on linear proximity. Thus, providing a more precise picture of the epigenomic regulation in a disease-relevant tissue.

In summary, the results of this thesis not only provide a compendium of useful resources but also novel knowledge regarding tissue-specific chromatin organisation associated with gene expression epigenomic regulation in human pancreatic islets. This is especially interesting for a disease-relevant tissue, as it enables the identification of likely target genes

for disease risk genomic variants annotated in *cis*-regulatory regions beyond previous assumptions based on linear proximity

Chapter 8
Methods

8.1. Refining *cis*-regulatory annotation in human pancreatic islets

- Islet regulome

Chromatin accessibility in human pancreatic islets was determined by Xavi Garcia and Nikolina Nakic using ATAC-seq (Buenrostro et al., 2013). Open chromatin sites in human pancreatic islets were in turn characterised based on a compendium of histone marks associated with active *cis*-regulatory regions (H3K4me3, H3K4me1, H3K27ac) and chromatin-bound factors (MED1/Mediator, SMC1/Cohesin, CTCF) by Irene Miguel-Escalada and Goutham Atla. Briefly, 6 kb windows divided in 100bp bins and centred around 157,940 open chromatin sites were characterised based on their ChIP-seq signal distribution of the 6 previously mentioned epigenomic factors. ChIP-seq signal (-log10 p-value) was determined using MACS2 (Zhang et al., 2008). The open chromatin sites were categorised in 7 groups by applying a k-median clustering using flexClust (Leisch, 2006). These 7 clusters were manually merged into 4 major categories that were named: active promoters, (class I-III) active enhancers, inactive enhancers and CTCF-enriched sites. Goutham Atla also found that the active enhancers category was formed by 3 clusters (class I-III) differentiated based on their enrichment for H3K4me1, H3K27ac and Mediator (MED1) binding, in which class I showed the strongest enrichments (Fig. 46).

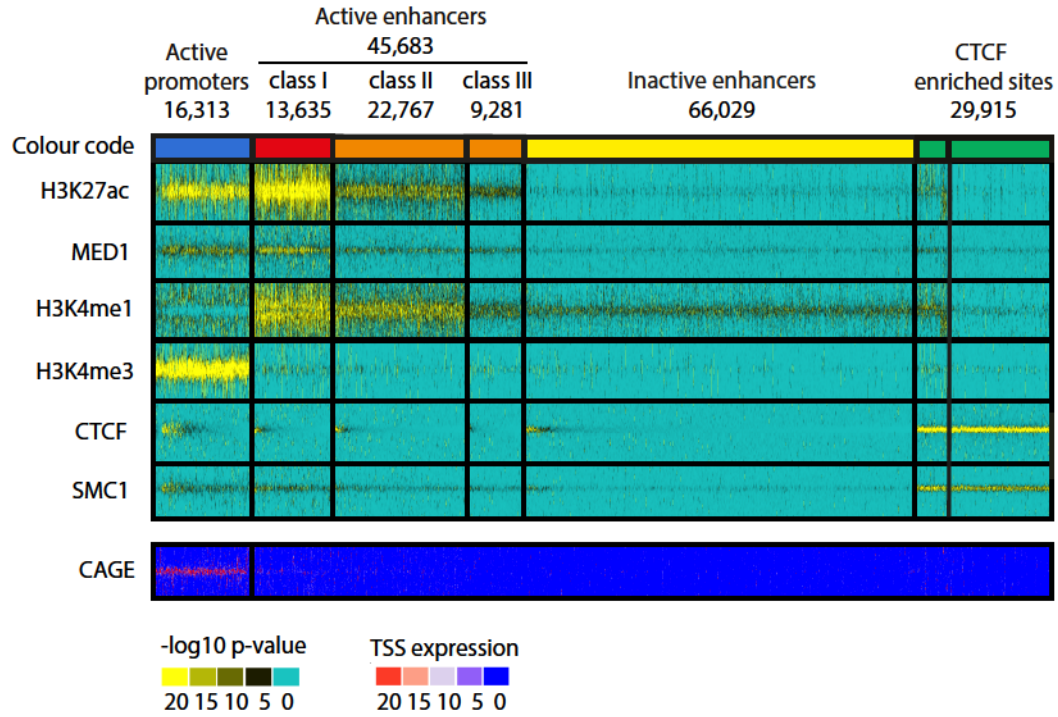


Fig. 46: Islet regulome. Clustering of 174,253 open chromatin sites based on 6 epigenomic datasets. These epigenome datasets, from top to bottom, are: H3K27ac, MED1, H3K4me1, H3K4me3, CTCF, SMC1. This analysis provided 7 clusters that were manually merged in 4 categories: active promoter, active enhancers, inactive enhancers and CTCF enriched sites. Transcription initiation was determined by CAGE-seq (Kanamori-Katayama et al., 2011; Shiraki et al., 2003) and used as a validation of our definition of active promoters. The colour code used in WashU browser screenshots is indicated on top of each cluster. Note that a subset of open chromatin regions that did not show distinct chromatin mark enrichments and did not fall in any of these clusters.

- **Islet ChromHMM**

The epigenomic landscape in human pancreatic islets was independently characterised by my colleague Claire Morgan. ChromHMM (Ernst and Kellis, 2012) was used to segment the genome based on the co-occurrence of 12 epigenomic marks in human pancreatic islets. These included 7 histone modifications associated with either active *cis*-regulatory regions (H3K4me3, H3K9ac, H3K4me1, H3K27ac) (Heintzman et al., 2007, 2009; Karmodiya et al., 2012), repression (H3K27me3, H3K9me3), (Schwartz and Pirrotta, 2013) or transcriptional elongation (H3K36me3) (Barski et al., 2007). It also incorporated a histone variant (H2A.Z) associated to accessible chromatin regions such as active promoters and enhancers (Barski et al., 2007), and three chromatin interacting factors (MED1/Mediator, SMC1/Cohesin, CTCF) (Allen and Taatjes, 2015; Dekker and Mirny, 2016; Merkenschlager and Nora, 2016; Ong and Corces, 2014).

To facilitate the interpretation, the original 15 ChromHMM states were merged into 9 states based on similarities in their epigenomic profiles (Fig. 47, Fig. 48). Thus, for example 4 original ChromHMM states (state 7-10) representing different flavours of active enhancers were combined in one single category.

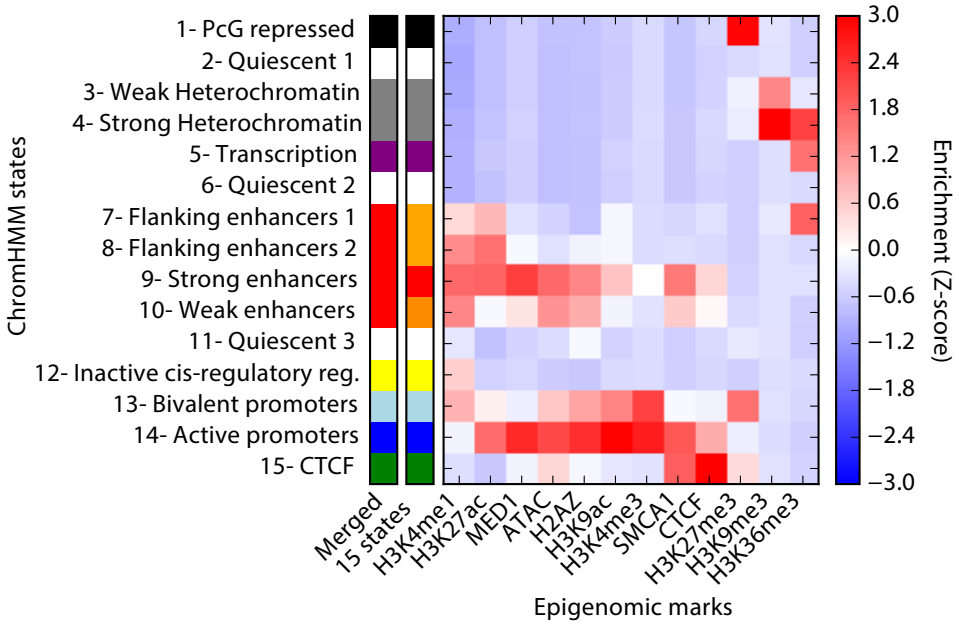


Fig. 47: ChromHMM segmentation in 15 states. Heatmap showing the enrichment of each interrogated epigenomic feature for a given ChromHMM state (data generated by Claire Morgan). Colour codes for the original 15 ChromHMM states defined by Claire Morgan are showed in column "15 states". Colour code for the merged 9 ChromHMM state are showed in column "Merged".

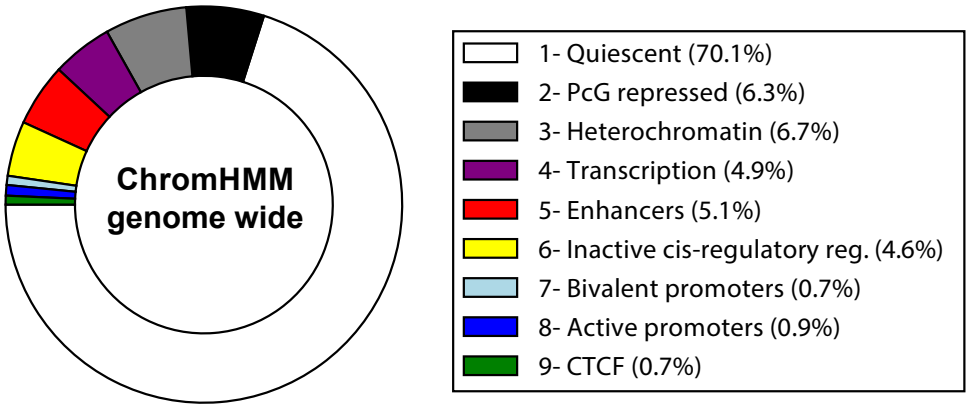


Fig. 48: ChromHMM genome coverage. Pie chart illustrating the genomic space covered by a given ChromHMM state, where active promoters are indicated in blue, active enhancers are indicated in red, inactive *cis*-regulatory regions in yellow, repressed regions in grey or black, CTCF binding sites in green and highly transcribed regions in purple. The ChromHMM states and percentages of the genome covered by each state are shown in the adjacent box.

8.2. Definition of super-enhancers (SE) based on Mediator signal

Active enhancers defined in the islet regulome (section 8.1) were used to define super-enhancers by Irene Miguel-Escalada, based on Mediator (MED1) occupancy. In summary, using the algorithm ROSE (Lovén et al., 2013) as previously described (Hnisz et al., 2013; Lovén et al., 2013), active enhancers were stitched based on linear proximity (12.5 kb) and ranked based on MED1 occupancy. The elbow of the MED1 occupancy distribution was used as a threshold to define 774 super-enhancers due to their high occupancy.

8.3. Generation of pcHi-C interactions maps

Human pancreatic islet preparations from four cadaveric donors were cultured for three days and formaldehyde crosslinked nuclei were prepared by Xavi Garcia. PcHi-C libraries were made by Dr. Biola Javierre from Professor Peter Fraser's lab (Babraham Institute, Cambridge, UK) as previously described (Javierre et al., 2016). Hi-C library generation was carried with in-nucleus digestion with HindIII and re-ligation (Nagano et al., 2015), followed by chromatin de-crosslinking and purification by phenol-chloroform extraction. DNA was sheared to an average size of 400 bp by mechanic DNA fragmentation (Covaris). The DNA fragments were end-repaired, adenine-tailed and size-selected ranging from 250 to 550 bp DNA fragments. Ligation events marked by biotin were selected with Streptavidin DynaBeads and ligated to paired-end adaptors for Illumina sequencing. Hi-C libraries were amplified 7–8 PCR amplification cycles. 3 Hi-C libraries were made per pancreatic islet preparation, generating 12 Hi-C libraries. The resultant product was used for promoter capture using a custom RNA library composed of 37,608 RNA baits against 21,177 human annotated gene promoters. After library enrichment, a post-capture PCR amplification was performed with 4 PCR amplification cycles (Fig. 6 in section 1.5).

Each pcHi-C library was sequenced in 3 lanes from Illumina HiSeq2500 platform obtaining a total number of 694,826,673 paired-end reads. Raw reads from 3 technical replicates generated from each of the four islet samples were pooled and mapped to the human genome (GRCh37/hg19). Additionally, reads were filtered out from experimental artefacts such as circularised reads and re-ligation products by using the HiCUP pipeline (Wingett et al., 2015), generating 600,112,182 uniquely mapped pcHi-C paired-end sequence reads (ditags). Statistically significant chromatin interactions (score ≥ 5) were determined using

CHICAGO (Cairns et al., 2016) accounting for consistence between biological replicates. This generated a chromatin interaction map formed by 175,784 interactions.

8.4. Virtual 4C using pcHi-C data

In order to better understand local 3D chromatin conformation, pcHi-C data involving a given promoter bait (or view point) was represented as a "virtual" 4C by my colleague Delphine Rolando. Merged read counts, computed using CHICAGO (Cairns et al., 2016), were visualised in a histogram centered on a viewpoint, showing the frequency in which a given locus was observed in the same read as the interrogated baited promoter. Promoter interacting regions were coloured based on CHICAGO scores for visualisation purposes. Statistically significant interacting regions (CHICAGO score >= 5) were indicated in black, whereas non-statistically significant interacting regions (CHICAGO score < 5) were shown in grey (Fig. 49).

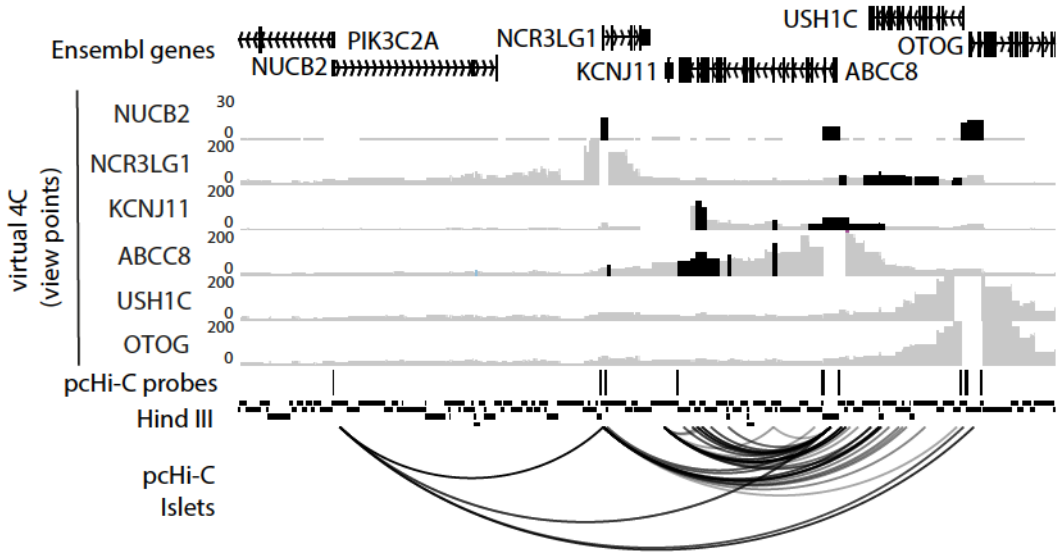


Fig. 49: Virtual 4C around *KCNJ11's* locus. Screenshot around the *KCNJ11* gene locus characterising its 3D chromatin organisation. Tracks from top to bottom are: Ensembl gene annotation; collection of virtual 4C from different view points (*NUCB2*, *NCR3LG1*, *KCNJ11*, *ABCC8*, *USH1C* and *OTOG* promoter) representing the number of merged reads, where statistically significant interacting regions (CHICAGO score >= 5) were indicated in black; pcHi-C RNA probes used to target annotated promoters; virtual digestion of the hg19 genome using HindIII restriction enzyme and finally islet chromatin interactions detected by pcHi-C.

8.5. Enrichment of epigenomic features at promoter-interacting regions

To assess the epigenomic factors that could be involved in the formation of chromatin interactions, I interrogated the presence of CTCF, Cohesin (SMC1), Mediator (MED1) islet binding sites, as well as active promoters and enhancers define in the islet regulome at pcHi-C integrating sites. This characterisation was done distinguishing for baited fragments and non-baited interacting fragments.

For each pcHi-C interacting point I computed the distance to any interrogated epigenomic feature (e.g. MED1 peaks or active enhancers) within a +/- 50 kb window. The distance density was estimated using Gaussian kernels (python 2.7 function scipy.stats.gaussian_kde) (Oliphant, 2007).

The expected distribution was computed using 10 randomizations of the given list of coordinates. Each randomization was generated using shuffleBed from BedTools (Quinlan and Hall, 2010) avoiding non-overlapping random coordinates and excluding blacklisted genomic regions (Kundaje, 2013) (Fig. 50).

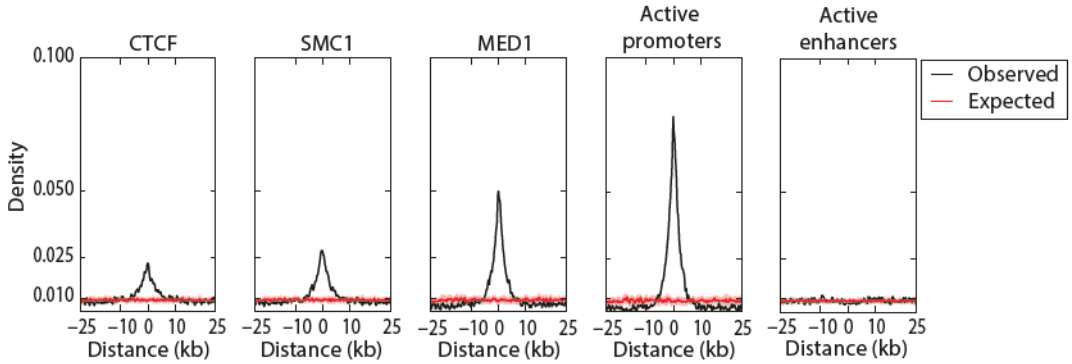
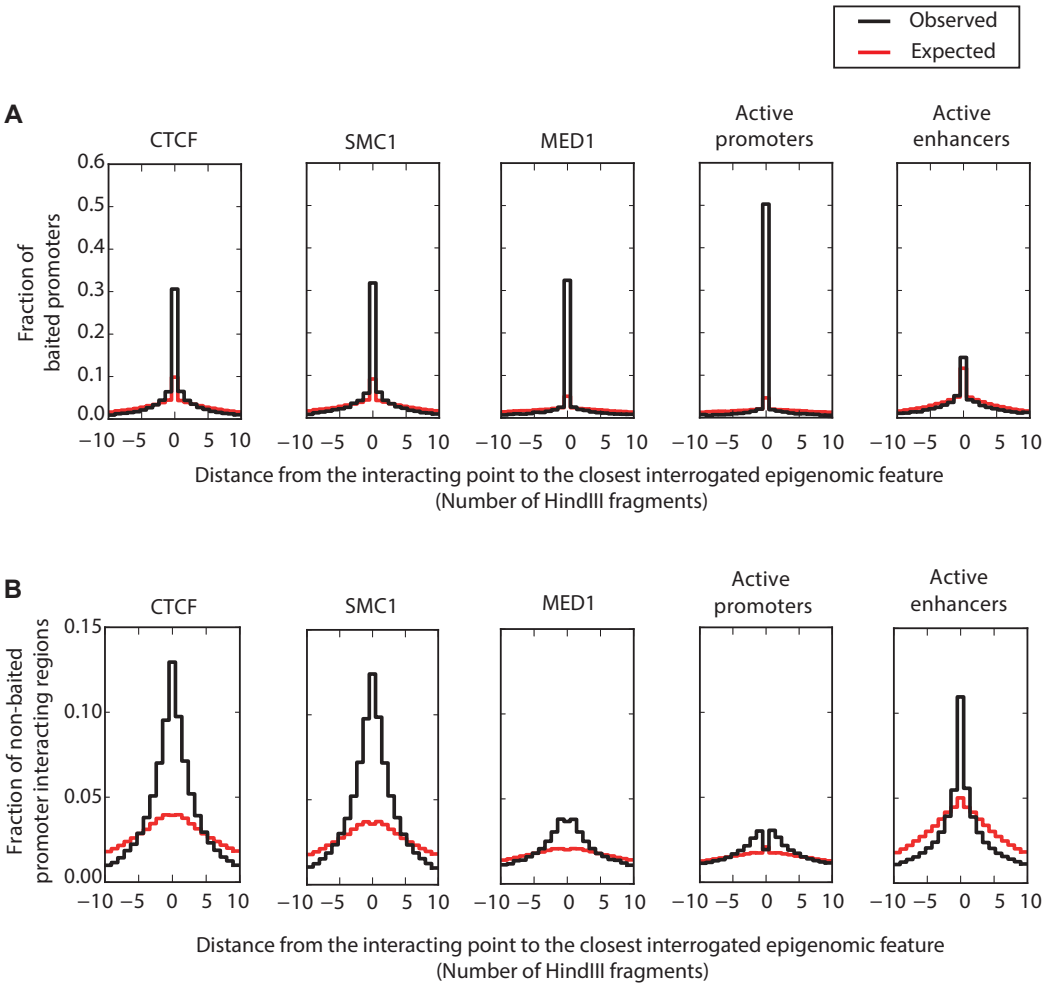


Fig. 50: Enrichments at baited promoters. Density plot showing the distribution of interacting factors (CTCF, SMC1 and MED1), active promoters and enhancers in a -/+25 kb window around all promoter-interacting regions. Expected distribution is generated after randomising the positions of the interrogated epigenomic factor. Observed distribution is shown as a black line. Median expected distribution is shown as a red line, and values between interquartile ranges are shown as red area.

8.6. Association between epigenomic factors and pcHi-C interacting points

As exposed in section 3.6, I considered that different technical aspects could be influencing CHICAGO's resolution to determine HindIII genomic interacting fragments, an aspect which needed to be considered for an accurate interpretation of pcHi-C interaction maps. To determine the resolution of our pcHi-C map I considered that the majority of confident interacting sites would overlap with an epigenomic factor known for driving chromatin interactions. Therefore, I interrogated the presence of 5 epigenomic factors (active promoters, active enhancers, MED1, SMCA1 or CTCF binding sites) at baited (Fig. 51A) and non-baited promoter-interacting regions (Fig. 51B). For each epigenomic factor (Fig. 51) or all of them at the same time (Fig. 15), I computed the closest site to each interacting point using ClosestBed (BedTools) and the distance between two elements was computed as the number of HindIII fragments using a custom script. The expected distribution was computed from 10 randomisations per element using shuffleBed from BedTools (Quinlan and Hall, 2010) as described in section 8.5.

Fig. 51: Distance from an interacting site to the closest interrogated epigenomic factor site. Histogram showing the distance from the closest interrogated epigenomic factor site to (A) a baited or (B) a non-baited interacting region. The list of interrogated epigenomic factors encompassed CTCF, MED1, SMC1, active promoters and active enhancers. Distance was computed as the number of HindIII fragments from the interrogated interacting point. A distance equal to 0 means that the epigenomic factor overlapped with the interacting HindIII fragment. Note that because baited regions were analysed separately (A) active promoters are depleted at position 0 of the non-baited interacting loci (B), as occurred in the previous analysis (Fig.15 in section 3.3). Figure shown in the following page.



For baited sites, any of the overlapping epigenomic factors were assigned exclusively to the baited HindIII fragment. Non-baited promoter-interacting regions were extended +/- 1 HindIII fragment and associated to any overlapping epigenomic factor within these 3 HindIII fragment window.

8.7. Identification of islet-specific interactions

Islet-specific chromatin interactions were defined as those consistent pcHi-C interactions (CHICAGO score ≥ 5) exclusively present in human pancreatic islets in comparison to 4 hematopoietic cell types (Erythroblasts, Naïve CD4+, Total B and Macrophages M1 cells) (Javierre et al., 2016). Based on this criterion, 53,839 (31% of all islet interactions) were classified as *islet-selective*.

A set of 59,672 (34%) interactions that were consistently detected (CHICAGO score ≥ 5) in human pancreatic tissue and 3 out of 4 hematopoietic cell types was also defined as *ubiquitous interactions*.

8.8. Gene classification based on expression selectivity in human tissues

A publicly available collection of RNA-seq datasets generated in 18 human tissues were obtained from The Human BodyMap 2 Project and ENCODE/LICR Project (Birney et al., 2007; Dunham et al., 2012) as well as human pancreatic islets and acinar tissue (Morán et al., 2012). Reads were aligned and processed by my colleagues Ignasi Morán and Delphine Rolando. In addition to islets this collection included RNA-seq datasets generated in: pancreatic acinar, adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, muscle, ovary, prostate, testes, thyroid tissues and white blood cells.

The reads were aligned using STAR aligner version 2.3.0 (Dobin et al., 2013). Because the data was also used for allelic studies, the reads were aligned against a modified version of the hg19 genome in which common SNPs (Global Minor Allele Frequency> 1%) from the dbSNP database 142 were masked (Sherry et al., 2001). A maximum mismatch of 10 nucleotides was used, and non-uniquely aligned reads were removed. Quantification of the raw read count was done using HTseq-Count version 0.6.1 with python 2.6.6 and Pysam version 0.8.3 (Anders et al., 2015). Counts were then converted into TPMs using the formula described in (Wagner et al., 2012).

I measured overall tissue selectivity of expression of genes across tissues as a coefficient of variation (C.V.) among all 18 samples. Additionally, gene expression enrichment in pancreatic islets was computed as a Z-score, comparing the gene expression level in human pancreatic islet against the mean value in all 18 tissues (Fig. 52A) (Cebola et al., 2015). I also defined expressed/non-expressed status among all 21,117 baited genes. 12,559 (59.3%) were defined as expressed, and 8,618 (40.7%) as non-expressed if their expression in human pancreatic islets was greater or lower than 1.5 transcripts per million (TPMs), respectively. Among all expressed genes, 983 (4.6%) that had both a coefficient of variation accros tissues and an islet Z-score greater than the 75th percentile were defined as “islet–specific” expressed genes (Fig. 52B). Therefore, the remaining set of 11,497 (54,3%) expressed genes

were classified as expressed, non-islet-specific. Finally, as human pancreatic islets are surrounded by exocrine tissue before being collected and this can be a source of contamination, 79 (0.4%) genes with an expression 3 times higher in acinar cells than in human pancreatic islets were considered as likely acinar contaminants.

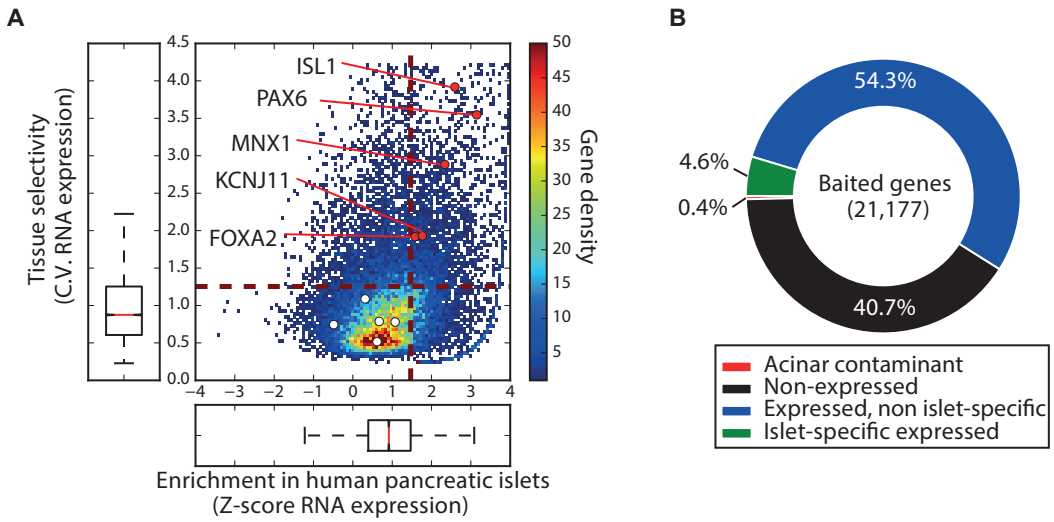


Fig. 52: Gene classification based on tissue-specificity expression patterns. (A) Heatmap showing gene expression enrichment in human pancreatic islets (Z-score) and tissue selectivity (coefficient of variation) for all islet expressed genes (≥ 1.5 TMPs). Those genes with an islet Z-score and a coefficient of variation (C.V.) greater that the 75th percentile (indicated with maroon dashed lines) were defined as "islet specific" expressed genes (top-right quadrant). The Z-score and CV distributions are shown in box plots parallel to their respective axes. As reference, Z-score and coefficient of variation (C.V.) values of 5 islet key genes (ISL1, PAX6, MNX1, KCNJ11, FOXA2) are indicated with red dots. Values for 5 conventional "house-keeping" genes (GAPDH, TBP, B2M, RPLP0, ACTB) are shown with white dots. (B) Pie chart showing the proportion of each gene expression class among the 21,177 baited genes.

8.9. Association between islet-selective interaction and islet-specific gene expression

I classified genes in baited fragments based on the number of islet-selective interactions. Later I computed the enrichment of 3 gene classes based on their expression in human pancreatic islets (non-expressed, islet specific-expressed and expressed, non-islet-specific) (see section 8.8). The enrichment of the 3 gene classes was computed as a log2 fold difference between genes with ≥ N islet-selective interactions and baited genes with < N islet-selective interactions. A hypergeometric test was computed to determine if the given enrichment was statistically significant.

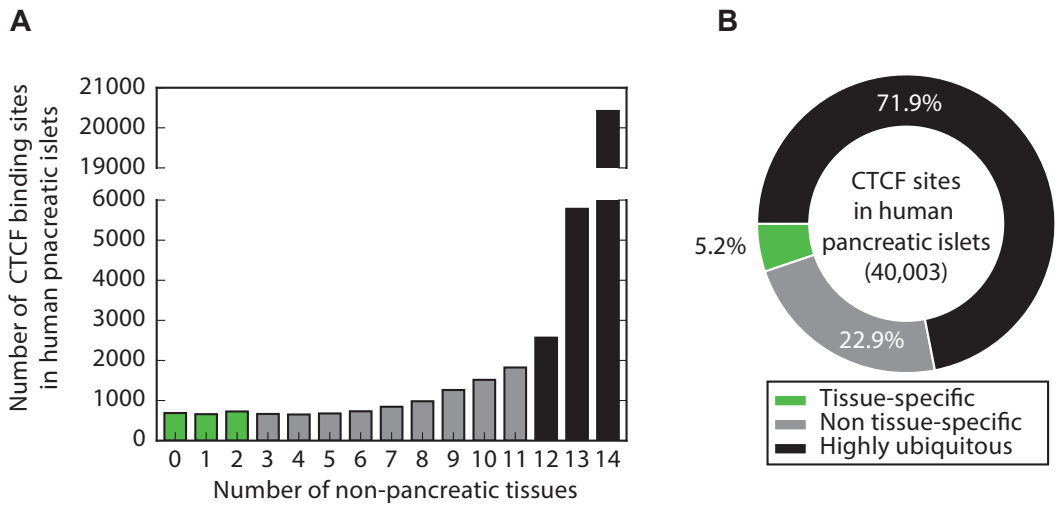
8.10. Islet-specific CTCF binding sites

A collection of 14 ChIP-seq datasets for CTCF from the ENCODE project (Dunham et al., 2012) covering a broad list of human tissues (Table 9) was used to interrogate CTCF binding tissue-specificity. I defined as islet-specific those CTCF binding sites present in pancreatic islets and up to 2 non-pancreatic tissues (Fig. 53). I also defined a set of 28,770 peaks as highly ubiquitous as they were observed in pancreatic islets and at least 12 non-pancreatic tissues. Intersections between datasets were done using intersectBed (BedTools; Quinlan & Hall, 2010).

Table 9: Summary of CTCF ChIP-seq datasets used to determine tissue-specificity of CTCF-binding sites in human pancreatic islets.

Sample ID	Human tissue	Source	Number of CTCF peaks
--	Pancreatic islets	J. Ferrer's lab	40,003
GSM733765	Astrocytes	ENCODE	63,295
GSM1003474	B-cells	ENCODE	52,783
GSM1022677	Cardiac myocytes	ENCODE	67,750
GSM1003508	CD14	ENCODE	44,999
GSM822281	Fibroblast	ENCODE	69,303
GSM733672	H1 hESC	ENCODE	104,538
GSM733645	HEPG2	ENCODE	69,097
GSM733724	HMEC	ENCODE	54,380
GSM733762	HSMM	ENCODE	78,134
GSM1006886	Kidney	ENCODE	73,464
GSM1006882	Lung	ENCODE	64,202
GSM733636	NHEK	ENCODE	73,625
GSM733784	Osteoblast	ENCODE	91,918
GSM1006883	Spleen	ENCODE	70,124

Fig. 53: CTCF tissue-specificity. (A) Bar plot showing the number of islet CTCF sites present in other 14 non-pancreatic tissues. The value 0 represents CTCF binding sites exclusively observed in human pancreatic islets. (B) Chart showing the percentage of tissue-specific (green), non-tissue-specific (grey) and highly ubiquitous (black) CTCF sites observed in human pancreatic islets). Figure shown in the following page.



8.11. Chromatin binding factor co-occupancy at CTCF binding sites

Coordinates of 5 islet lineage-determining TFs (FOXA2, PDX1, MAFB, NK6.1, NKX2.2), MED1 (Mediator) and SMC1 (Cohesin) were crossed with active enhancers or CTCF binding sites using intersectBed (BedTools; Quinlan & Hall, 2010).

8.12. Enrichment of Mediator-bound enhancers and islet-specific CTCF binding sites at islet-selective interacting points.

The presence of enhancers or CTCF binding sites at pcHi-C interacting points was computed grouping them by MED1 binding or tissue-specificity respectively. Enrichments of these elements at islet-specific interactions were computed over non-islet-specific interactions as a fold change in a log2 scale.

8.13. Definition of TAD-like structures

- **DI domains**

Directionality Index (DI) score was computed genome wide using the formula proposed by Dixon et al., 2012.

$$DI = \left(\frac{B-A}{|B-A|}\right)\left(\frac{(A-E)^2}{E}+\frac{(B-E)^2}{E}\right) \qquad E = (A+B)/2$$

Formula 2: Directionality index (DI) score. Formula for DI score as originally described in Dixon et al., 2012 .

Table 10: A and B variables from the DI score formula adapted to pcHi-C.

Variable	Hi-C (Dixon et al. 2012)	pcHi-C
A	The number of reads that map from a given 40kb bin to the upstream 2Mb.	The number of intra-chromosome interactions going upstream from a given sliding window of 5 HindIII fragments.
B	The number of reads that map from the same 40kb bin to the downstream 2Mb.	The number of intra-chromosome interactions going downstream from a given sliding window of 5 HindIII fragments.

The original definition of variables used for calculating the DI score was adapted to pcHi-C as shown in Table 10. DI domains were defined as genomic territories flanked by regions with a negative DI score on the 5' edge and a positive DI score on the 3' edge.

- **Chromatin domains interconnectivity**

Interconnectivity between DI domains was computed as a log2 ratio between the number of inter-domain and intra-domain interactions. Therefore, ratios lower than 0 correspond to domains with more intra-domain interaction than inter-domain interactions.

Adjacent DI domains with interconnectivity ratios greater than 0 were merged. These merged DI domains defined 3,598 islet TAD-like compartments (Fig. 54).

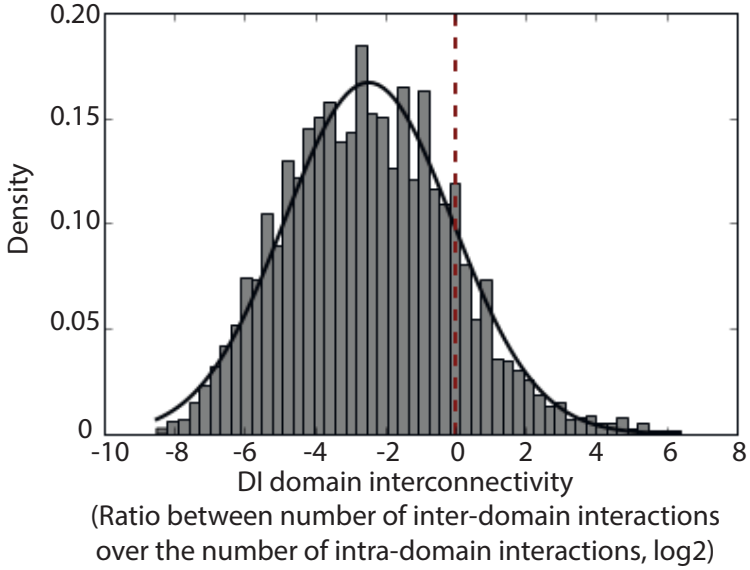


Fig. 54: DI domain interconnectivity distribution. Histogram showing interconnectivity ratios (number of inter-domain interactions over the number of intra-domain interactions) for adjacent DI domains in a log2 scale. Interconnectivity threshold (> 0) is indicated as maroon dashed line.

8.14. CTCF motif mapping

A *de novo* motif analysis was conducted using HOMER (Heinz et al., 2010) on a list of consistent CTCF peaks providing a highly accurate position weight matrix (PWM) for the CTCF binding motif. My colleague Irene Miguel-Escalada defined consistent CTCF binding sites in human islets as those reproduced in at least 2 out 3 biological replicates.

Using annotatePeaks.pl from HOMER, the CTCF binding sequence was mapped at consistent CTCF peaks requiring a minimum log score of 5. In cases where multiple motifs were mapped within the same CTCF peak, only the motif with the highest log score was kept. I thus determined the position and the orientation of the CTCF binding motif within CTCF consistent binding sites.

8.15. CTCF occupancy in TAD-like compartments

The location of CTCF occupancy sites relative to TAD-like compartments was computed using computeMatrix scale-regions and plotProfile from DeepTools 2 (Ramirez et al., 2014) grouping CTCF sites based on the orientation of their binding motif.

8.16. Tissue-specificity of TAD boundary regions

The degree of TAD boundary tissue-specificity was determined as in Schmitt et al., 2016. In summary, the genomic space was divided 40 kb bins. Thus, islet TAD boundaries were defined as 40 kb binned genomic regions overlapping an islet TAD edge. This was done to define TAD boundaries similar to those previously defined by Schmitt et al. using Hi-C. Later, islet TAD boundaries were combined with TAD boundaries from 21 additional tissues into a single putative TAD boundaries reference file. Reference TAD boundaries within 200 kb window were merged into a single TAD boundary “region” using mergeBed (BedTools).

Merging of adjacent boundary bins into TAD boundary regions was performed because TAD boundaries defined in different tissues may be slightly shifted (by a few bins). Therefore,

although TAD boundaries do not directly overlap, both borders may be within the same boundary region (Schmitt et al., 2016). Finally, the number of islet TAD boundary regions overlapping TAD boundaries from the remaining 21 human samples was computed using intersectBed (BedTools).

8.17. Degree of inter-TAD connectivity

The degree of inter-TAD connectivity through pcHi-C interactions was determined using islet TAD-like structures (see section 8.13) and TADs previously defined in human ESC and human IMR90 fibroblasts using Hi-C (Dixon et al., 2012). The percentage of inter-TAD and intra-TAD interactions was computed among all *cis* pcHi-C interactions detected in human pancreatic islets. Expected values were computed after shuffling TAD genomic positions 5 times using shuffleBed (BedTools, Quinlan & Hall, 2010).

8.18. Definition of promoter-associated domains

Promoter-associated domains (PADs) were defined as the linear space covered by all the interactions starting from a specific promoter bait and the most distant promoter-interacting regions within the same islet TAD-like compartment. A total number of 16,030 PADs were defined in this manner.

8.19. Fraction of TAD spaces occupied by PADs

As PADs were limited by TAD-like boundaries, I computed the fraction of a TAD space occupied by a given PAD. Therefore, if both elements occupied exactly the same genomic space, the fraction of occupied space was equal to 1.

8.20. ChromHMM enrichment in PADs

ChromHMM segmentation (Ernst and Kellis, 2012) is described in section 8.1. The contribution of each ChromHMM state was computed as a fraction of the genomic PAD space. The enrichment of ChromHMM in the PAD was calculated as a fold difference over genomic distribution in a log2 scale (Fig. 29 in section 4.1, Fig. 48 in section 8.1). The enrichment of each ChromHMM state in an islet PAD versus the remaining islet TAD space was also computed as a ratio in log2. This was only computed for 7,085 PADs that were at least 25% smaller than their corresponding TAD (Fig. 28, Fig. 30 in section 4.1).

8.21. Enhancer-promoter assignment

Enhancers were assigned to promoters based on a list of consecutive steps. Each step prevailed over the following steps. Therefore, each step was only performed on unassigned enhancers. Steps 2-4 were only applied for baited promoters that had a typical active promoter state in the islet regulome or islet ChromHMM (see section 8.1). The enhancer-promoter assignment steps were:

1. Interacting enhancers were associated to promoters based on the presence of pcHi-C chromatin interactions. In cases where the enhancer formed multiple loops with multiple baited promoters, all transcriptional targets were considered.
2. If a non-interacting enhancer was contained within a PAD of a baited active promoter that did show interactions with other enhancers, the non-interacting enhancer was tentatively associated to that PAD's promoter.
3. I assumed that a +/-10 kb window around any baited promoter would encompass a region where random collisions are too frequent to enable the identification of high confidence interactions above background noise. On the other hand, I reasoned that this linear distance is likely to provide sufficient 3D proximity to establish functional enhancer – promoter communication. Therefore, non-interacting enhancers (not assigned in step 2) residing within 10kb of a baited active promoter were tentatively assigned to them.
4. The remaining non-interacting enhancers were associated to a baited active promoter if they were only contained within a single PAD.

8.22. Enhancer assignment validation based on gene expression class enrichments

To validate enhancer-promoter assignments from section 8.21, I computed the enrichment of islet-specific genes (section 8.8) among all genes with assigned enhancers. To do so, I generated two lists of genes. A first list, named "assigned genes" that contained all genes with an assigned enhancer in section 8.21. The second list was named "control genes", and it contained all genes to which PADs overlapped with active enhancers but those active enhancers were not assigned to those genes (Fig. 55).

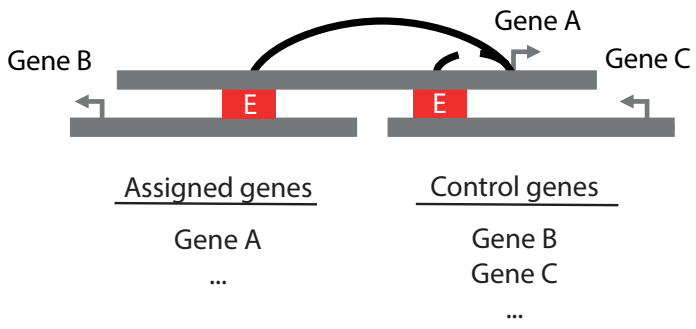


Fig. 55: Selection of genes assigned to islet active enhancers and control genes. Diagram illustrating the generation of two gene sets, assigned genes and control genes, to validate our enhancer assignment strategy. PcHi-C interactions are indicated as solid black arcs, inferred enhancer-promoter associations as dashed black lines, PADs are indicated as grey horizontal bars and enhancers (E) as red boxes.

Genes were classified in 3 categories based on their gene expression pattern as: non-expressed, islet-specific and non-islet-specific expressed (see section 8.8). Per each gene class, I computed the enrichment among "assigned genes" compared to "control genes" as log2 of the ratio. The statistical significance of this enrichment was assessed performing a chi-square test (python 2.7 function scipy.stats.chi2_contingency) comparing the frequency of each gene class among the two lists, "assigned genes" and "control genes" (Fig. 32 and Fig. 56).

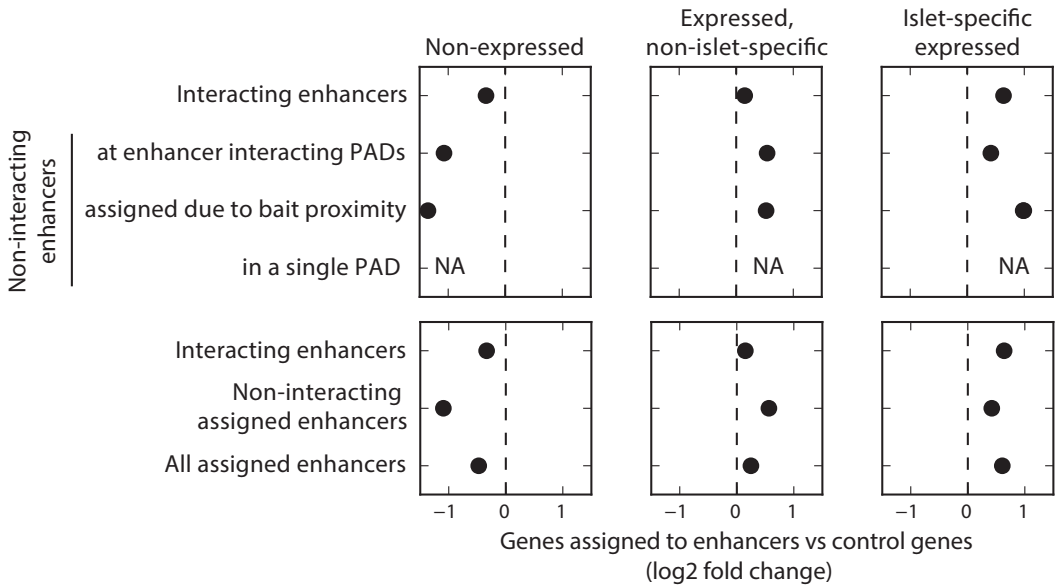


Fig. 56: Enhancer assignments considering chromatin interaction maps accentuate their association with islet-specific expressed genes. For the three different gene expression classes, I

computed the log2 ratio for enhancer "assigned genes" vs. genes whose PADs also contained an enhancer but these were not assigned to them ("control genes"). This analysis was done per each step of the enhancer-promoter assignment strategy (top) or grouped in interacting enhancers, non-interacting assigned enhancers (step 2-4 of the assignment strategy) or all assigned enhancers (step 1-4 of the assignment strategy) (bottom).

8.23. Enhancer assignment validation based on H3K27ac correlations.

To validate enhancer-promoter assignments from section 8.21, G. Atla and S. Bonas computed H3K27ac correlations between loci forming each possible enhancer-promoter intra-TAD pair. To do so, they selected a set of H3K27ac ChIP-seq datasets with at least 15M of mappable non-clonal reads (Table 11) covering abroad collection of human tissues. Human islet samples were down samples to 30M reads, to facilitate the comparison with the epigenome roadmap samples. Normalised read counts were computed per each enhancer (+/- 750bp centred window) and promoter defined in the islet regulome (see section 8.1). Finally, spearman's rho values (scipy.stats.spearmanr) among ChiP-seq samples were computed as metric of intra-TAD enhancer-promoter correlation per each possible pair defined using hESC TADs (Dixon et al., 2012).

Based on the enhancer-promoter strategy proposed in section 8.21 enhancers were classified in 4 categories: interacting enhancers, assigned non-interacting enhancers at enhancer interacting PADs, non-interacting enhancers assigned due to bait proximity, assigned non-interacting PAD specific enhancers (yellow) and non-assigned enhancers (Fig. 57A). Enhancer-promoter pairs based on pcHi-C data ("Assignment") were compared against enhancer-promoter pairs found in the same PAD but that not assigned ("Control") (Fig. 57B). Enhancer-promoter correlations rho values per each enhancer type were used to compare the informativity of "assigned" enhancer-promoter pairs versus "control" pairs. The results showed that "assigned" pairs present a better correlation than "control" pairs, suggesting that enhancer-promoter assignments (section 8.21) based on pcHi-C interactions supposes and improvement over associations based on linear proximity.

Fig. 57: Enhancer assignments considering chromatin interaction maps are coherent with enhancer-promoter H3K27ac correlations. (A) Based on the enhancer-promoter strategy proposed in section 8.21 enhancers were categorised as interacting enhancers (maroon), assigned non-interacting enhancers at enhancer interacting PADs (red), non-interacting enhancers assigned due to bait proximity (orange), assigned non-interacting PAD specific enhancers (yellow) and non-assigned enhancers (grey). (B) Enhancer-promoter assignments were assessed based on H3K27ac correlation among samples. Enhancer-promoter pairs based on pcHi-C data ("Assignment") are compared against enhancer-promoter pairs found in the same PAD but that not assigned ("Control") as

illustrated in the schematics. An array of density plots show enhancer-promoter correlations (rho values) per each enhancer type specified in Fig. 57A. “Assignment” values are show with a solid line and “control” values with a dashed line. Figure shown in the following page.

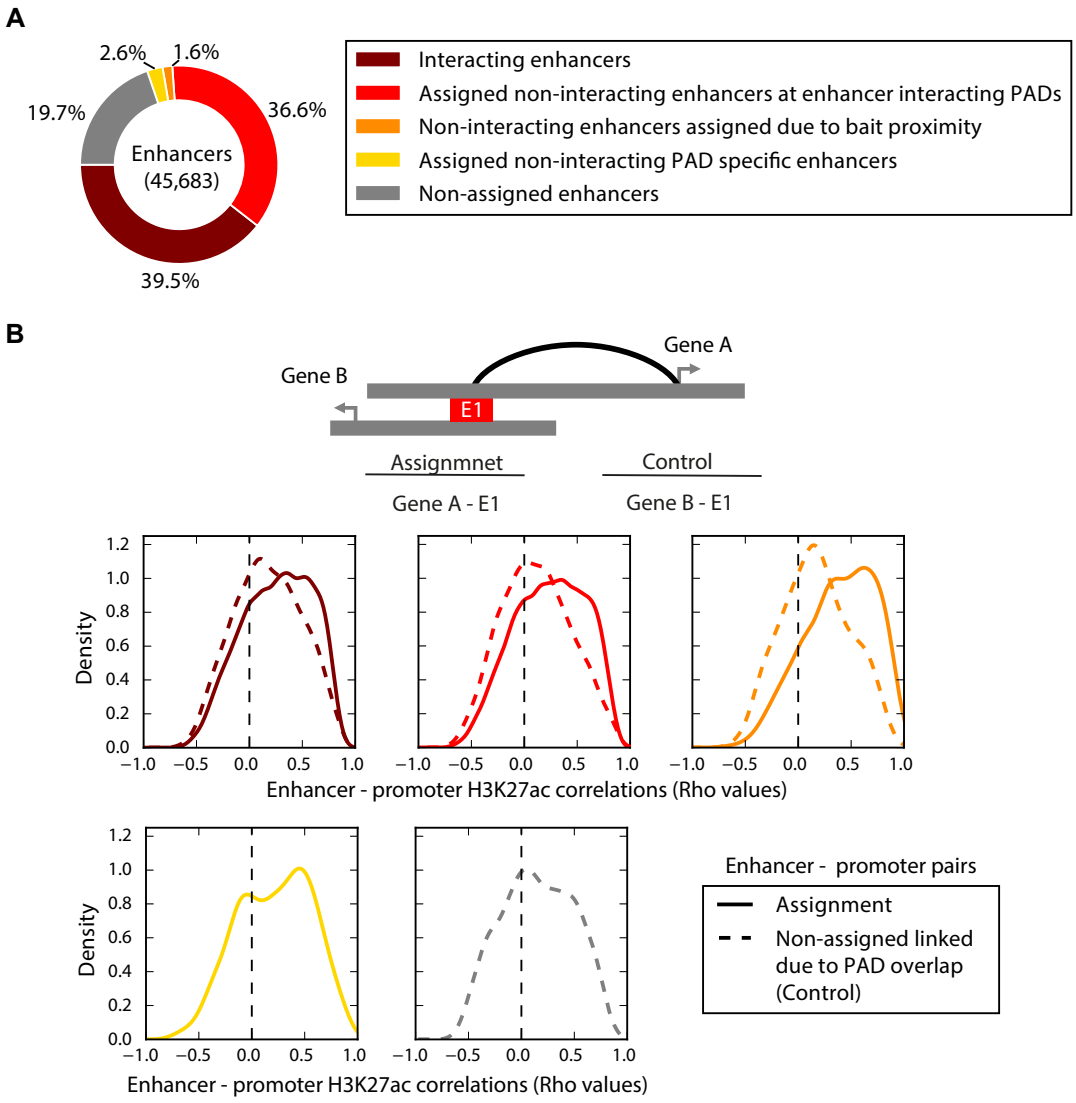


Table 11: Summary of H3K27ac ChIP-seq datasets used to determine enhance-promoter correlations.

Sample ID	Human tissue	Source	Number of reads
E034	Primary T cells	Epigenome Roadmap	30
E037	Primary T helper memory cells	Epigenome Roadmap	27.7
E041	Primary T helper cells PMA-I stimulated	Epigenome Roadmap	30
E042	Primary T helper 17 cells PMA-I stimulated	Epigenome Roadmap	18.7
E043	Primary T helper cells	Epigenome Roadmap	16.7
E045	Primary T cells effector/memory	Epigenome Roadmap	16.1
E047	Primary T CD8+ naive cells	Epigenome Roadmap	30
E055	Penis foreskin fibroblast primary cells	Epigenome Roadmap	27.1
E056	Penis foreskin fibroblast primary cells	Epigenome Roadmap	28.4
E058	Penis foreskin keratinocyte primary cells	Epigenome Roadmap	30

E059	Penis foreskin melanocyte primary cells	Epigenome Roadmap	30
E061	Penis foreskin melanocyte primary cells	Epigenome Roadmap	30
E062	Primary mononuclear cells from peripheral blood	Epigenome Roadmap	30
E063	Adipose Nuclei	Epigenome Roadmap	16.6
E065	Aorta	Epigenome Roadmap	21.2
E066	Liver	Epigenome Roadmap	30
E067	Brain angular gyrus	Epigenome Roadmap	30
E068	Brain anterior caudate	Epigenome Roadmap	30
E069	Brain cingulate gyrus	Epigenome Roadmap	30
E071	Brain hippocampus middle	Epigenome Roadmap	30
E072	Brain inferior temporal lobe	Epigenome Roadmap	30
E073	Brain dorsolateral prefrontal cortex	Epigenome Roadmap	30
E074	Brain substantia nigra	Epigenome Roadmap	30
E075	Colonic mucosa	Epigenome Roadmap	16.2
E078	Duodenum smooth muscle	Epigenome Roadmap	30
E079	Esophagus	Epigenome Roadmap	19.7
E080	Fetal adrenal gland	Epigenome Roadmap	19.5
E084	Fetal intestine large	Epigenome Roadmap	28.6
E085	Fetal intestine small	Epigenome Roadmap	29.7
E087	Pancreatic islets	Epigenome Roadmap	16.7
E089	Fetal muscle trunk	Epigenome Roadmap	17.5
E090	Fetal muscle leg	Epigenome Roadmap	26.1
E091	Placenta	Epigenome Roadmap	21.6
E092	Fetal stomach	Epigenome Roadmap	30
E093	Fetal thymus	Epigenome Roadmap	30
E094	Gastric	Epigenome Roadmap	16.8
E095	Left ventricle	Epigenome Roadmap	30
E096	Lung	Epigenome Roadmap	29.5
E097	Ovary	Epigenome Roadmap	16.6
E099	Placenta amnion	Epigenome Roadmap	25.9
E100	Psoas muscle	Epigenome Roadmap	15.8
E101	Rectal mucosa	Epigenome Roadmap	17.7
E102	Rectal mucosa	Epigenome Roadmap	15.9
E103	Rectal smooth muscle	Epigenome Roadmap	18.6
E104	Right atrium	Epigenome Roadmap	24.6
E106	Sigmoid colon	Epigenome Roadmap	30
E108	Skeletal muscle female	Epigenome Roadmap	18.9
E109	Small intestine	Epigenome Roadmap	30
E111	Stomach smooth muscle	Epigenome Roadmap	22.2
E112	Thymus	Epigenome Roadmap	16.9
E113	Spleen	Epigenome Roadmap	16.9
HI 129 H	Pancreatic islets	J. Ferrer's lab	30
HI 129 L	Pancreatic islets	J. Ferrer's lab	30
HI 130 H	Pancreatic islets	J. Ferrer's lab	30
HI 130 L	Pancreatic islets	J. Ferrer's lab	30
HI 131 H	Pancreatic islets	J. Ferrer's lab	30
HI 131 L	Pancreatic islets	J. Ferrer's lab	30
HI 132 H	Pancreatic islets	J. Ferrer's lab	30
HI 132 L	Pancreatic islets	J. Ferrer's lab	30
HI 135 H	Pancreatic islets	J. Ferrer's lab	30
HI 135 L	Pancreatic islets	J. Ferrer's lab	30
HI 137 H	Pancreatic islets	J. Ferrer's lab	30
HI 137 L	Pancreatic islets	J. Ferrer's lab	30
HI 149	Pancreatic islets	J. Ferrer's lab	21.6
HI 152 H	Pancreatic islets	J. Ferrer's lab	30
HI 152 L	Pancreatic islets	J. Ferrer's lab	30
HI HI26	Pancreatic islets	J. Ferrer's lab	30
HI HI32	Pancreatic islets	J. Ferrer's lab	15.6
HI HI34	Pancreatic islets	J. Ferrer's lab	30
HI HI76	Pancreatic islets	J. Ferrer's lab	15.8
HI HI82	Pancreatic islets	J. Ferrer's lab	30
HI HI153	Pancreatic islets	J. Ferrer's lab	26.6

8.24. Promoter characterisation

To carry discover features that predict islet-specific gene expression through machine learning I characterised different types of gene promoters based on the following 15 features:

- Number of pcHi-C interactions in human pancreatic islets.
- Fraction of islet-selective interactions originated from the promoter's bait (see section 8.7).
- Fraction of promoter-enhancer interactions originated from the promoter's bait (see section 8.6).
- Number of enhancers overlapping the promoter-associated domain (PAD). The number of overlapping enhancer defined in islet regulome (see section 8.1) were computed using intersectBed (BedTools,(Quinlan and Hall, 2010)).Number of enhancers overlapping the promoter-associated domain (PAD). The number of overlapping enhancers defined in the islet regulome (see section 8.1) was computed using intersectBed (BedTools; Quinlan & Hall, 2010).
- Number of active enhancers assigned to the promoter's bait (see section 8.21).
- Number of class I active enhancers assigned to the promoter's bait (see section 8.1).
- Width of the TSS determined by CAGE (bp). CAGE shape was determined by my colleague Goutham Atla as previously described (Forrest et al., 2014). In summary, CAGE is a NGS based technique to identify Transcription Start Sites (TSS) at base-pair (bp) resolution (Kanamori-Katayama et al., 2011; Shiraki et al., 2003). CAGE tags starting sites (CTSS were clustered based on proximity (< 20 bp) and characterised based on interquartile width (q0.1 – q0.9). CTSS can be classified as "sharp" or "broad" (≥ 11 bp) based on their interquartile width (Forrest et al., 2014) (Fig. 58).
- Length of the overlapping CpG island. A list of bona fide CpG islands (Bock et al., 2007) was downloaded from UCSC web browser. CGI were associated to Ensembl annotated TSS based in overlap using intersectBed (BedTools).
- H3K4me3, H3K9me3 or H3K27me3 signal in human pancreatic islets was computed as area under the curve within a +/- 1kb window around the annotated TSS. This data was scaled from 0 to 1 only for representative purposes.

- Number of H3K4me3, H3K9me3 or H3K27me3 peaks from other tissues overlapping the TSS. Per each histone mark I downloaded a collection of peaks called in 139 human tissues as part of the Roadmap Epigenome project (Kundaje et al., 2015). These datasets were intersected with annotated TSS using intersected (BedTools).

To assess the quality of our map of CAGE TSSs in human pancreatic islets, I interrogated whether previously reported features were recapitulated in this dataset. I found that the CAGE TSS length presents a bimodal distribution that could be used to define "sharp" (<11 bp) and "broad" active promoters in human pancreatic islets. Moreover, I also noticed that islet-specific expressed frequently present a "sharp" promoter rather than a "broad" promoters (Fig. 58). Both observations were coherent with previous reports (Forrest et al., 2014; Lenhard et al., 2012), validating CAGE TSS length as likely informative feature for gene expression classification.

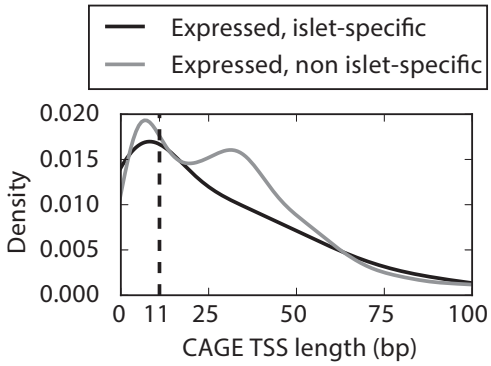


Fig. 58: TSS length distribution. Density plot showing the TSS length distribution for islet-specific (black) and non-islet specific expressed genes (grey). TSS length was determined as 0.1-0.9 interquartile width determined by Goutham Atla using CAGE. Length equal to 11 bp is highlighted with a dashed line, and it has been previously used as threshold to define "sharp" and "broad" promoters (Forrest et al., 2014).

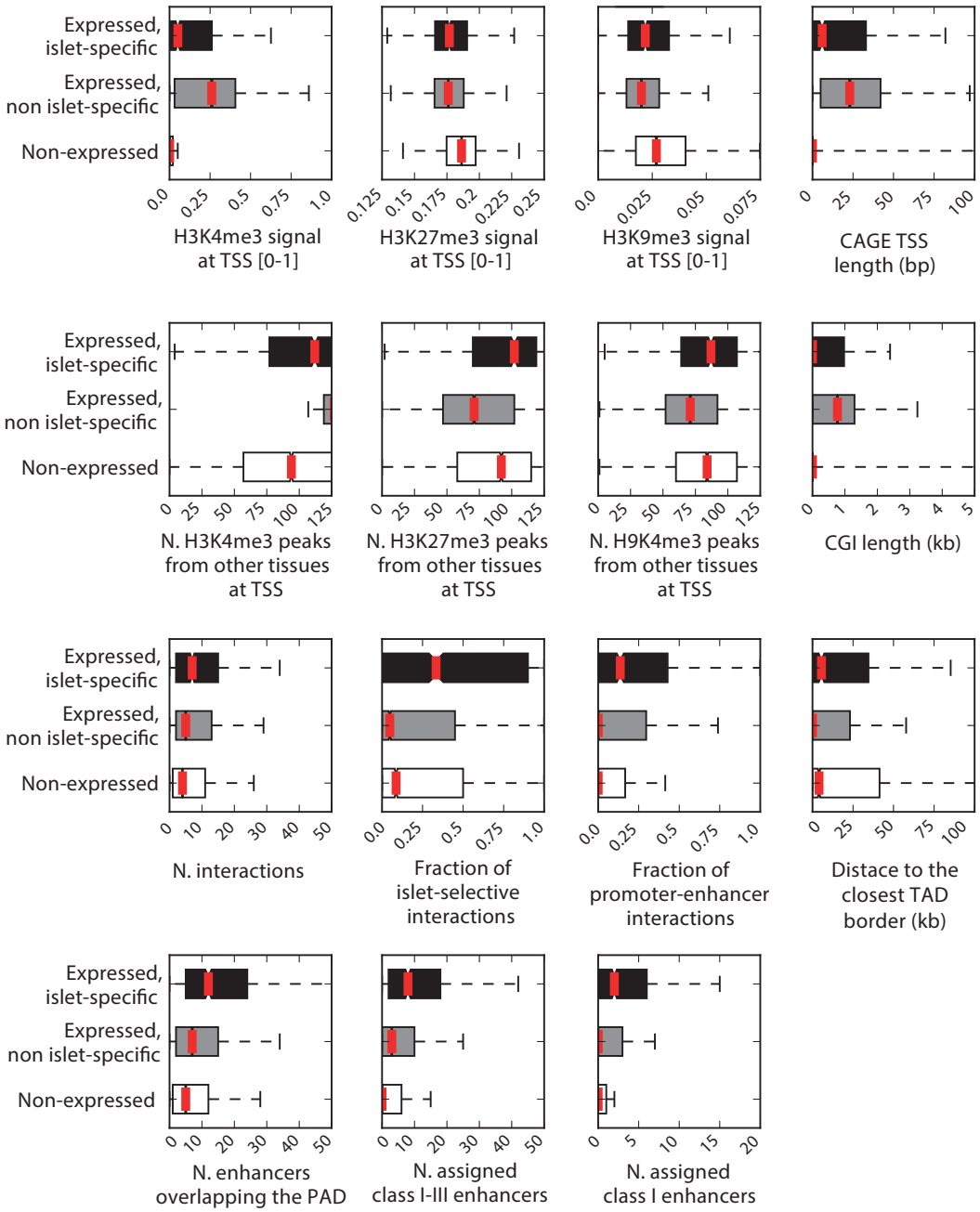


Fig. 59. Epigenomic characterisation of islet-specific expressed genes. Array of boxplots interrogating epigenomic features listed in section in 8.23 differentiating by gene expression class (see section 8.8). Each panel interrogates an epigenomic feature and contains a boxplot per gene class (non-expressed; expressed, non-islet-specific and expressed, islet-specific)

Correlation between the 15 epigenomic features was determined by Pearson's correlation (computed using python 2.7 function scipy.stats.pearsonr; Oliphant, 2007). A list of 12 non-highly correlated features (Pearson's score < 0.65) (Fig. 60, Table 7) was used for a logistic regression analysis. In logistic regression independency among the variables is assumed. Therefore, was important to use a compendium of lowly correlated features.

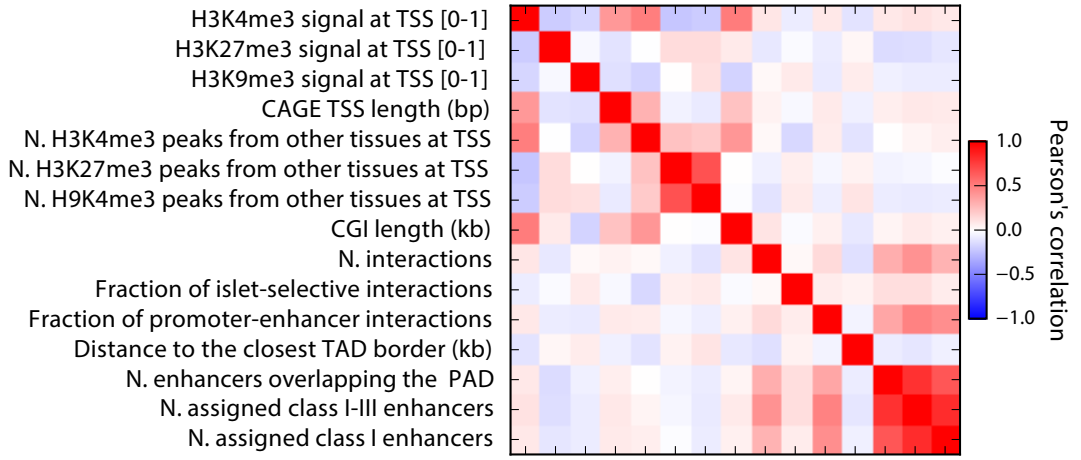


Fig. 60: Correlation between epigenomic features. Heatmap showing pair-wise Pearson's correlation scores between 15 epigenomic features.

Logistic regression

A logistic regression (LR) analysis was conducted to determine which epigenomic factors could be used to predict gene expression patterns (see section 8.8) using python 2.7 and scikit-learn (Pedregosa and Varoquaux, 2011). The model was generated using balanced class weights, thus accounting for the fact that islet-specific expressed genes were much less frequent than other classes of genes (Fig. 52). The machine learning (ML) classifier was trained 50 times with a random sampling containing 70% of the full data set and the remaining 30% was kept for validations. In each of the 50 rounds, LR coefficients were computed as a metric of feature importance for each gene class. In summary, the ML classifier computes the coefficients (β) for a list of features (χ) that fits a logit function (Fig. 33 in section 4.1). This function computes the probability (p) that a given gene belongs to a gene expression class (k) (Formula 1 in section 4.1). Thus, the higher the coefficient the bigger the weight of a given feature in the decision function (Bewick et al., 2005).

The ML model obtained in each of the 50 permutations was assessed computing its confusion matrix and 4 different metric scores. The confusion matrix reveals the model's

accuracy per gene category (Fig. 61B, D). An ideal confusion matrix would be formed by 1 at the diagonal and 0 on the remaining positions of the matrix indicating a perfect prediction. The 4 metric scores assess the general performance of the model, independently of the sample categories, each score ranges from 0 to 1, being 1 the best value (Fig. 61C, E). The different computed metric scores were:

- Accuracy: fraction of corrected predictions.
- Recall: the model's ability to identify all positive samples, independently of the number of false negative. True positive/ (true positive + false negative)
- Precision: the model's ability to identify all negative samples, independently of the number of false positive. True positive/ (true positive + false positive)
- F1 score: Weighted average of the precision and recall. (2 * (precision * recall)) / (precision + recall)

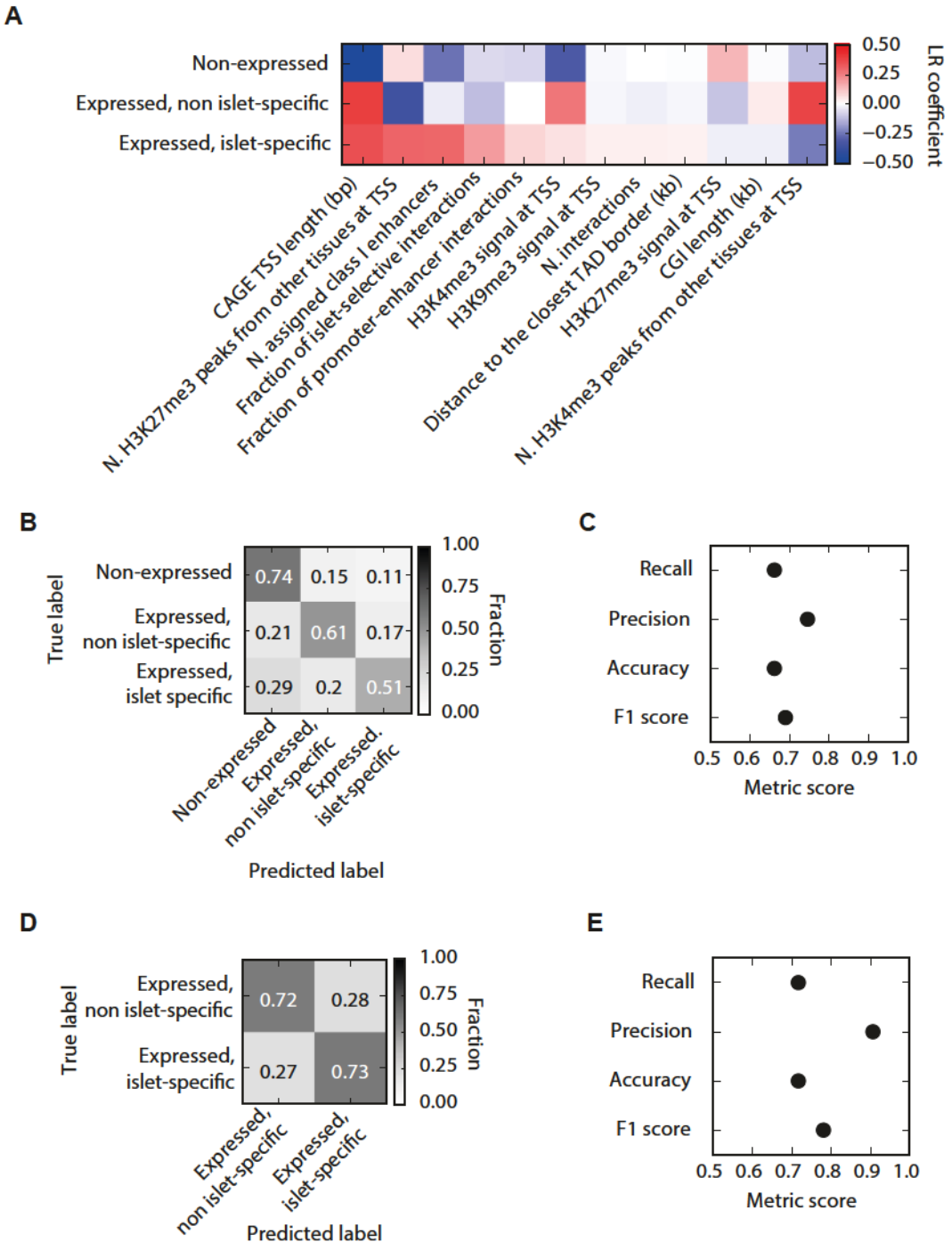


Fig. 61. Logistic Regression analysis to determine features associated to gene classes. (A). Heatmap showing Logistic Regression coefficients per epigenomic feature and gene class. (B and D) Confusion matrix assessing the proportion of labels / gene classes correctly predicted, interrogating all 3 classes (C) or differentiating between expressed genes (D). (E and C). Values for 4 metric scores were used to assess a logistic regression model to predict all gene classes (C) or only differentiate between expressed genes (E). Those metric scores were: recall, precision, accuracy and f1 score.

8.25. PAD classification based on enhancer content

PADs were classified based on their content of assigned active enhancers (see section 8.1, 8.21). Among all active enhancers defined in the islet regulome, PAD classification was based on class I enhancers. Class I enhancers were markedly enriched in epigenomic features associated to active enhancers (H3K27ac and MED1 signal) and were most enriched at tissue-selective interactions.

Thus, PADs were classified as:

- Enhancer-less. PADs without any assigned enhancer.
- Enhancer-poor. PADs with one or more assigned class I-III enhancers, but less no more than 2 assigned class I enhancers.
- Enhancer-rich. PADs with 3 or more assigned class I enhancers.

8.26. Enrichment of genomic variants at *cis*-regulatory regions

My colleague Irene Miguel-Escalada compiled an updated list of genomic variants associated to type-2 diabetes (T2D) and fasting glycemia (FG) variation. I used this list to interrogate whether enhancers contained in different PAD classes were particularly enriched in disease associated variants using Variant Set Enrichmet (VSE) (Yang et al., 2011). VSE accounts for linkage disequilibrium (LD) between the interrogated genomic variants so that the biases due to LD structures are minimalised. As a negative control, I used a set of breast cancer associated variants (obtained from the VSE GitHub repository), as I do not expect any association with cis-regulatory elements active in pancreatic islets.

8.27. Overlap between enhancer-rich PADs

Enhancer-rich PADs were merged into a single genomic interval using mergeBed (Bedtools). The number of PAD or TAD compartments was determined by overlap using intersectBed (BedTools).

8.28. STARR-seq

Selection of candidate regions

I selected genomic regions of approximately 650 bp for testing enhancer activity by STARR-seq (Arnold et al., 2013). This pilot collection of 69 genomic regions included 45 active

enhancers, 4 primed enhancers, and 5 open chromatin regions without enhancer marks, all previously identified in human pancreatic islets based on epigenomic marks (Pasquali et al., 2014). Moreover, I included 15 regions that did not show evidences of chromatin accessibility in human pancreatic islets, neither in the different human cell line characterised as part of the ENCODE project (Dunham et al., 2012). It was reasonable to hypothesise that these 15 compacted chromatin region would not show enhancer activity therefore were used as negative control regions to determine background signal.

Library preparation

This collection of 69 genomic DNA loci was amplified by PCR using the high-fidelity DNA polymerase Q5 (NEB) for few (12) cycles to avoid PCR amplification mistakes. After checking the PCR products in an agarose gel, PCR products were pooled and purified using the PCR purification kit (QIAGEN). This was used as starting material to create a STARR-seq library as described in Arnold et al., 2013 with the exception that I used custom adaptors for sequencing. These custom adaptors included a 3bp Unique Molecular Identifier (UMIs) to facilitate the quantification and avoid bias due to PCR amplification artefacts. Moreover, these custom adaptors also included a 6bp library specific index (Note 1). The library index did not have utility in the current experiment, but in the future it could be used to differentiate between libraries containing risk or non-risk alleles for genomic variants as the sequencing reads would probably not cover the cloned fragment completely.

Adaptor 1: 5' - ACACTCTTTCCCTACACGACGCTCTTCCGATC*TNNNXXXXXX*T - 3'
Adaptor 2: 5' - [Phos]XXXXXXNNNAGATCGGAAGAGCACACGTCT - 3'

Note 1: Custom STARR-seq adaptors. The position of UMI is indicated with Ns and it is formed by 3 random nucleotides. Position custom library index is indicated with Xs and formed by 6bp nucleotides following Illumina sequencing index list. Adaptor 2 contains a 5' phosphatase modification ([Phos]) and adaptor 1 phosphorothioate bonds (*).

Briefly, purified PCR products were dA-tailed to allow adaptor ligation, which was performed in a molarity ration 10:1 in favour to the annealed custom adaptors using the T4 Ligase (NEB). Homologous regions to the STARR-seq vector were incorporated by PCR in order to enable vector ligation through isothermal *in vitro* recombination (Gibson et al., 2009) using In-fusion HD kit (Clonatech). The resultant plasmid library was electroporated and amplified using Endura ElectroCompetent Cells (Cambridge bioscience) that were

cultivated in 10 LB +ampicillin plates. Bacteria was collected and purified using PureYield Plasmid Midiprep System (Promega). The presence of the 69 elements in the final library was confirmed by PCR.

Cell transfection

4 million Min6 cells (Ishihara et al., 1993) were plated in 48 well plates, 120.000 cells per well. Each well was transfected with 600 ng STARR-seq library using lipofectamine 2000 (Thermo Fisher scientific). At 48h after transfection cell were tripsinised, collected and stored at -80C until used. ¾ of the pooled cells were used for RNA extraction and the remaining ¼ for plasmid DNA extraction.

RNA-extraction and preparation for sequencing

Total RNA was extracted using RNAeasy Kit (QIAGEN), and mRNA enrichment was performed using Dynabeads Oligo (dT)25 (Life technologies) following manufacturer's instructions. Contaminant DNA was degraded using Turbo DNAse (Life technologies). After quantifying the RNA concentration by Qubit (Thermo Fisher scientific), the sample was used as template to generate cDNA by RT-PCR using the SuperScript III kit (Thermo Fisher scientific) and the reporter-RNA specific primer indicated in Arnold et al., 2013. Resultant cDNA was used as a template in a nested PCR in which the first set of primers amplifies the cloned fragments and then adds Illumina indexes for multiplexing.

Plasmid purification and preparation for sequencing

Plasmid DNA was purified using the Mini-prep purification kit (QIAgen). The resultant material was amplified and indexed for NGS as described in Arnold et al., 2013.

Sequencing

Each sample was spiked (1%) in a lane of an Illumina Hiseq 2500 run in high output mode by the Imperial BRC Genomics Facility (Imperial College London). The number of 100bp single-end read obtained per sample is indicated in Table 12.

Table 12: Number of reads obtained per sample.

Sample	Number of reads
STARR cDNA replicate 1	3.5 M
STARR cDNA replicate 2	2.83 M
STARR input replicate 1	1.95 M
STARR input replicate 2	2.13 M

- **Computational analysis**

Low quality and low complexity reads were filtered out using Prinseq (Schmieder and Edwards, 2011). Raw reads were mapped using Bowtie2 (Langmead and Salzberg, 2012) exclusively to the interrogated regions. Raw counts were computed per loci and UMI. Per each locus, I computed the median signal among UMIs. The two biological replicates showed high correlations (r2> 0.9, Pearson's linear correlation) ensuring reproducibility (Fig. 62).

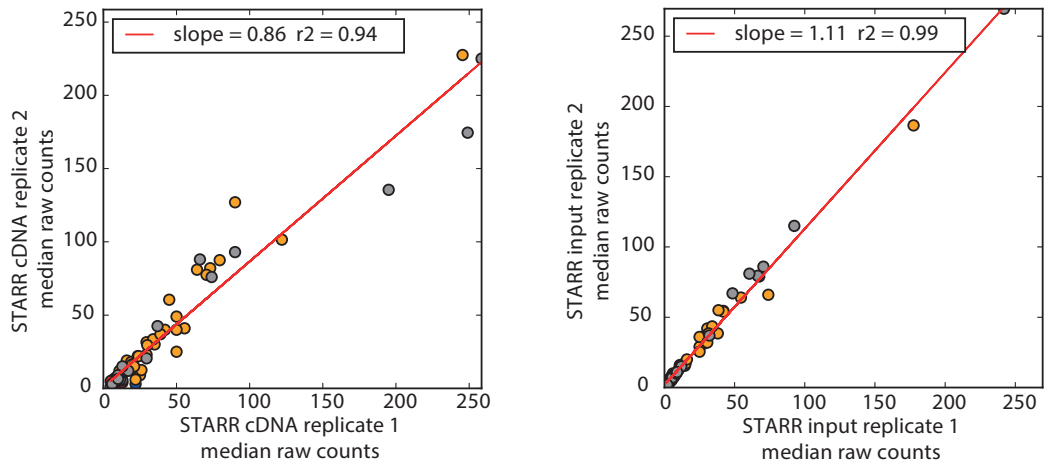


Fig. 62: High correlation between STARR-seq biological replicates. Both STARR-seq cDNA and input showed a high correlation between biological replicates tested by Pearson's correlation (r2). No significant differences were observed between the tested genomic classes; active enhancers (orange), primed enhancers (blue), open chromatin regions (red) or negative control regions (grey).

To correct for possible library representation bias among the tested fragments RNA STARR-seq raw read counts were divided by their counterpart in input STARR-seq library (Arnold et al., 2013).

Enhancer activity was assessed as fold change (F.C.), in a log2 scale, over the mean background signal determined using all negative control regions. A fragment was considered to present enhancer activity if the signal was greater than 2 standard deviations from the mean background signal in both biological replicates (Fig. 42).

8.29. Luciferase assays

- **Cloning**

Forward primers used to amplify genomic loci for the STARR seq library were adapted for TOPO-cloning (Thermo Fisher scientific) by adding the sequence "CACC" to the 5' edge. PCR amplification products were purified and cloned into the pENTR/D-TOPO vectors following manufacturer's instructions. Later, interrogated loci were cloned into the destination vector by applying the Gateway method (Thermo Fisher scientific).

- **Transfection**

Each vector was tested in at least two biological replicates, each biological replicate was formed by 3 technical replicates. Each test was performed in a well from a 48-wells plate containing 120.000 cells. 300 ng of the tested plasmid were co-transfected with 2 ng of pCMV-Renilla using Lipofectamine-2000 and following manufacturer's instructions.

After 48h of the transfection, cells were washed twice with PBS and treated with the Dual-Luciferase kit (Promega) following manufacturer's instructions.

- **Read-out**

The plate was read with a *GloMax® 96 Microplate* Luminometer with Dual Injectors (Promega).

- **Enhancer-activity**

After normalising for the loading control (pCMV-Renilla) enhancer activity was computed as fold change over the mean background signal determined using 4 negative control regions.

Bibliography

Adams, C.C., and Workman, J.L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol. Cell. Biol. *15*, 1405–1421.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene annotation system. Database *2016*, baw093.

Akerman, I., Tu, Z., Beucher, A., Rolando, D.M.Y., Sauty-Colace, C., Benazra, M., Nakic, N., Yang, J., Wang, H., Pasquali, L., et al. (2017). Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory Networks. Cell Metab. *25*, 400–411.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat. Rev. Mol. Cell Biol. *16*, 155–166.

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

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

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. Science (80-.). *340*, 1234167–1234167.

Ardlie, K.G., Deluca, D.S., Segre, A. V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.). *348*, 648–660.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science *1074*.

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

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell *143*, 470–484.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. Cell *98*, 387–396.

Benazra, M., Lecomte, M.-J., Colace, C., Müller, A., Machado, C., Pechberty, S., Bricout-Neveu, E., Grenier-Godard, M., Solimena, M., Scharfmann, R., et al. (2015). A human beta cell line with drug inducible excision of immortalizing transgenes.

Bewick, V., Cheek, L., and Ball, J. (2005). Statistics review 14: Logistic regression. Crit. Care *9*, 112.

Bhandare, R., Schug, J., Le Lay, J., Fox, A., Smirnova, O., Liu, C., Naji, A., and Kaestner, K.H. (2010). Genome-wide analysis of histone modifications in human pancreatic islets. Genome Res. *20*, 428–433.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG Island Mapping by Epigenome Prediction. PLoS Comput. Biol. *3*, e110.

Bonev Boyan, and Cavalli Giacomo (2016). Organization and function of the 3D genome. Nat. Rev. Genet. *17*, 661–678.

Branco, M.R., Branco, T., Ramirez, F., and Pombo, A. (2008). Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. Chromosom. Res. *16*, 413–426.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 1–8.

Burke, B., and Stewart, C.L. (2013). The nuclear lamins: flexibility in function. Nat Rev Mol Cell Biol *14*, 13–24.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.-M., Osborne, C., et al. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. *17*, 127.

Calo, E., and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell *49*, 825–837.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635.

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

Chen, T., and Dent, S.Y.R. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat. Rev. Genet. *15*, 93–106.

Chen, J., Zhang, Z., Li, L., Chen, B.-C., Revyakin, A., Hajj, B., Legant, W., Dahan, M., Lionnet, T., Betzig, E., et al. (2014). Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. Cell *156*, 1274–1285.

Chi, K.R. (2016). The dark side of the human genome. Nature *538*, 275–277.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol. Cell *9*, 279–289.

Cooper, G.M., and Hausman, R.E. (2007). The Cell: A Molecular Approach 2nd Edition.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D., and Lander, E.S. (2002). Detection of regulatory variation in mouse genes. Nat. Genet. *32*, 432–437.

Cremer, M., Grasser, F., Lanctôt, C., Müller, S., Neusser, M., Zinner, R., Solovei, I., and Cremer, T. (2008). Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. Methods Mol. Biol. *463*, 205–239.

Crow, D. (2016). “Junk DNA” tells mice—and snakes—how to grow a backbone. Science (80-.).

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.-Y., Cui, K., and Zhao, K. (2008). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24–32.

Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R., and Hughes, J.R. (2015). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat. Methods *13*, 74–80.

Davies, J.O.J., Oudelaar, A.M., Higgs, D.R., and Hughes, J.R. (2017). How best to identify chromosomal interactions : a comparison of approaches. Nat. Publ. Gr. *14*.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Dekker, J. (2002). Capturing Chromosome Conformation. Science (80-.). *295*, 1306–1311.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization

of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A.Y., Dixon, J., Maliskova, L., Guan, K., Shen, Y., and Ren, B. (2016). A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res. *26*, 397–405.

Le Dily, F., Baù, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H.G., Ballare, C., Filion, G., et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. *28*, 2151–2162.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dowen, J.M.M., Fan, Z.P.P., Hnisz, D., Ren, G., Abraham, B.J.J., Zhang, L.N.N., Weintraub, A.S.S., Schuijers, J., Lee, T.I.I., Zhao, K., et al. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell *159*, 374–387.

Drissen, R. (2004). The active spatial organization of the -globin locus requires the transcription factor EKLF. Genes Dev. *18*, 2485–2490.

Dukler, N., Gulko, B., Huang, Y., and Siepel, A. (2017). Is a super-enhancer greater than the sum of its parts ? *49*, 2–7.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ediger, B.N., Du, A., Liu, J., Hunter, C.S., Walp, E.R., Schug, J., Kaestner, K.H., Stein, R., Stoffers, D.A., and May, C.L. (2014). Islet-1 Is Essential for Pancreatic -Cell Function. Diabetes *63*, 4206–4217.

Engel, N., West, A.G., Felsenfeld, G., and Bartolomei, M.S. (2004). Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. Nat. Genet. *36*, 883–888.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Fadista, J., Vikman, P., Laakso, E.O., Mollet, I.G., Esguerra, J.L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R.B., et al. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc. Natl. Acad. Sci. *111*, 13924–13929.

Fantom Consortium, T. (2005). The Transcriptional Landscape of the Mammalian Genome. Science (80-.). *309*, 1559–1563.

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet *15*, 7–21.

Flannick, J., and Florez, J.C. (2016). Type 2 diabetes: genetic data sharing to advance complex disease research. Nat. Rev. Genet. *17*, 535–549.

Forrest, A.R.R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol *11*, 1–14.

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S.W., Cairns, J., Collier, A.J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C.S., Fraser, P.J., et al. (2017). Global reorganisation of cis - regulatory units upon lineage commitment of human embryonic stem cells. Elife *6*.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. Cell *166*, 1–11.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science (80-.). *354*, 769–773.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature *462*, 58–64.

Gall, J.G. (2003). Timeline: The centennial of the Cajal body. Nat. Rev. Mol. Cell Biol. *4*, 975–980.

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

Gaulton, K.J., Ferreira, T., Lee, Y., Raimondo, A., Mägi, R., Reschen, M.E., Mahajan, A., Locke, A., William Rayner, N., Robertson, N., et al. (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat. Genet. *2*.

Gaur, U., Li, K., Mei, S., and Liu, G. (2013). Research progress in allele-specific expression and its regulatory mechanisms. J. Appl. Genet. *54*, 271–283.

Gause, M., Schaaf, C.A., and Dorsett, D. (2008). Cohesin and CTCF: cooperating to control chromosome conformation? BioEssays *30*, 715–718.

Geyer, P.K., Vitalini, M.W., and Wallrath, L.L. (2011). Nuclear organization: Taking a position on gene expression. Curr. Opin. Cell Biol. *23*, 354–359.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Gómez-Marín, C., Tena, J.J., Acemel, R.D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., et al. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc. Natl. Acad. Sci. U. S. A. *112*, 7542–7547.

Gonzalez-sandoval, A., and Gasser, S.M. (2016). On TADs and LADs : Spatial Control Over Gene Expression. Trends Genet. *xx*, 1–11.

Graf, T., and Enver, T. (2009). Forcing cells to change lineages. Nature *462*, 587–594.

Gruenbaum, Y., and Foisner, R. (2015). Lamins: Nuclear Intermediate Filament Proteins with Fundamental Functions in Nuclear Mechanics and Genome Regulation. Annu. Rev. Biochem. *84*, 131–164.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015).

CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

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

Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2016). CTCF and Cohesin Regulate Chromatin Loop Stability with Distinct Dynamics. 1–10.

Hatta, M., and Cirillo, L.A. (2007). Chromatin opening and stable perturbation of core histone:DNA contacts by FoxO1. J. Biol. Chem. *282*, 35583–35593.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat. Genet. 1–12.

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

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

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

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

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat. Biotechnol. *33*, 510–517.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature *463*, 474–484.

Højfeldt, J.W., Agger, K., and Helin, K. (2013). Histone lysine demethylases as targets for anticancer therapy. Nat. Rev. Drug Discov. *12*, 917–930.

Iborra, F.J., Pombo, A., Mcmanus, J., Jackson, D.A., and Cook, P.R. (1996a). The Topology of Transcription by Immobilized Polymerases. Exp. Cell Res. *229*, 167–173.

Iborra, F.J., Pombo, A., Jackson, D.A., and Cook, P.R. (1996b). Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. J. Cell Sci. *109 (Pt 6*, 1427–1436.

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

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2016). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P., and Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nat. Methods.

Ishihara, H., Asano, T., Tsukuda, K., Katagiri, H., Inukai, K., Anai, M., Kikuchi, M., Yazaki, Y., Miyazaki, J.-I., and 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*, 1139–1145.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369–1384.e19.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D.,

Pegoraro, G., Lee, T.I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262–275.

Jiang, F.-X., and Morah, G. (2011). Pancreatic Stem Cells: Unresolved Business. In Stem Cells in Clinic and Research, (InTech), p.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C.O., et al. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res. *21*, 1150–1159.

Karmodiya, K., Krebs, A.R., Oulad-Abdelghani, M., Kimura, H., and Tora, L. (2012). H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. BMC Genomics *13*, 424.

Kellum, R., and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. Cell *64*, 941–950.

Kheradpour, P., Ernst, J., Melnikov, a., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.

Kim, T., and Shiekhattar, R. (2015). Review Architectural and Functional Commonalities between Enhancers and Promoters. Cell *162*, 948–959.

Kim, K.-D., Tanizawa, H., Iwasaki, O., and Noma, K. (2016). Transcription factors mediate condensin recruitment and global chromosomal organization in fission yeast. Nat. Genet. 1–12.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K. a, I, D., Green, R.D., Zhang, M.Q., Lobanenkov, V. V, and Ren, B. (2008). Analysis of the vertebrate insulator protein CTCF binding sites in the human genome. Cell *128*, 1231–1245.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G.J., Ariyurek, Y., den Dunnen, J.T., and Lamond, A.I. (2010). High-Resolution Whole-Genome Sequencing Reveals That Specific Chromatin Domains from Most Human Chromosomes Associate with Nucleoli. Mol. Biol. Cell *21*, 3735–3748.

Korkmaz, G., Lopes, R., Ugalde, A.P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. *34*, 192–198.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., et al. (2012). Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature *493*, 231–235.

Krijger, P.H.L., and de Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biol. *17*, 771–782.

Kundaje, A. (2013). A comprehensive collection of signal artifact blacklist regions in the human genome. … Site/Anshulkundaje/Projects/Blacklists (Last Accessed 30 ….

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P.,

Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Kvon, E.Z. (2015). Using transgenic reporter assays to functionally characterize enhancers in animals. Genomics *106*, 185–192.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Lam, M.T.Y., Li, W., Rosenfeld, M.G., and Glass, C.K. (2014). Enhancer RNAs and regulated transcriptional programs. Trends Biochem. Sci. *39*, 170–182.

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

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

Leisch, F. (2006). A toolbox for -centroids cluster analysis. Comput. Stat. Data Anal. *51*, 526–544.

Lemieux, K., and Gaudreau, L. (2004). Targeting of Swi/Snf to the yeast GAL1 UASG requires the Mediator, TAFIIs, and RNA polymerase II. EMBO J. *23*, 4040–4050.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233–245.

Lieberman-Aiden, E., and Berkum, N. van (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science (80-.). *326*, 289–293.

Lopes, R., Korkmaz, G., and Agami, R. (2016). Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers. Nat Rev Mol Cell Biol *17*, 597–604.

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

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. *32*, 225–237.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol. *1*, REVIEWS001.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-Guided Human Genome Engineering via Cas9. Science (80-.). *339*, 823–826.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat. Methods *13*, 366–370.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

Mattout, A., Cabianca, D.S., and Gasser, S.M. (2015). Chromatin states and nuclear organization in development--a view from the nuclear lamina. Genome Biol. *16*, 174.

Mazza, D., Abernathy, A., Golob, N., Morisaki, T., and McNally, J.G. (2012). A benchmark for chromatin binding measurements in live cells. Nucleic Acids Res. *40*, e119–e119.

McClellan, J., and King, M.C. (2010). Genetic heterogeneity in human disease. Cell *141*, 210–217.

Melé, M., and Rinn, J.L. (2016). "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. Mol. Cell *62*, 657–664.

Merkenschlager, M., and Nora, E.P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Annu. Rev. Genomics Hum. Genet. *17*, annurev-genom-083115-022339.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P. a, et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Morán, I., Akerman, I., Van De Bunt, M., Xie, R., Benazra, M., Nammo, T., Arnes, L., Nakić, N., García-Hurtado, J., Rodríguez-Seguí, S., et al. (2012). Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. Cell Metab. *16*, 435–448.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: Efficient and sensitive analysis of protein-directed genome architecture.

Nagano, T., Lubling, Y., Yaffe, E., Wingett, S.W., Dean, W., Tanay, A., and Fraser, P. (2015). Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. Nat. Protoc. *10*, 1986–2003.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science (80-.). *347*, 1017–1021.

Nasmyth, K., and Haering, C.H. (2009). Cohesin: Its Roles and Mechanisms. Annu. Rev. Genet. *43*, 525–558.

Németh, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Péterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Längst, G. (2010). Initial Genomics of the Human Nucleolus. PLoS Genet. *6*, e1000889.

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J.A. (2012). Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. Cell *150*, 1274–1286.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72–77.

Nishizaki, S.S., and Boyle, A.P. (2016). Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends Genet. *xx*, 1–12.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Oliphant, T.E. (2007). SciPy: Open source scientific tools for Python. Comput. Sci. Eng. *9*, 10–20.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Pabo, C., and Sauer, R.T. (1992). TRANSCRIPTION FACTORS : Structural Families and Principles of DNA Recognition.

Padeken, J., and Heun, P. (2014). Nucleolus and nuclear periphery: Velcro for heterochromatin. Curr. Opin. Cell Biol. *28*, 54–60.

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

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

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., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. *46*, 136–143.

Pederson, T. (2011). The Nucleolus. Cold Spring Harb. Perspect. Biol. *3*, a000638–a000638.

Pedregosa, F., and Varoquaux, G. (2011). Scikit-learn: Machine learning in Python.

Phillips-Cremins, J., and Corces, V. (2013). Chromatin Insulators: Linking Genome Organization to Cellular Function. Mol. Cell *50*, 461–474.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T.T., Hookway, T. a., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Van der Ploeg, L.H.T., Konings, A., Oort, M., Roos, D., Bernini, L., and Flavell, R.A. (1980). γ-β-Thalassaemia studies showing that deletion of the γ- and δ-genes influences β-globin gene expression in man. Nature *283*, 637–642.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Pradeepa, M.M., Grimes, G.R., Kumar, Y., Olley, G., Taylor, G.C. a, Schneider, R., and Bickmore, W. a (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. Nat. Genet.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell *152*, 1173–1183.

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

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

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., Syed, T., Emons, B.J.M., Gifford, D.K., and Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174.

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Rao, S.S.P.S.P., Huntley, M.H.H., Durand, N.C.C., Stamenova, E.K.K., Bochkov, I.D.D., Robinson, J.T.T., Sanborn, A.L.L., Machol, I., Omer, A.D.D., Lander, E.S.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1–16.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell *129*, 1311–1323.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.

Romer, A.I., and Sussel, L. (2015). Pancreatic islet cell development and regeneration. Curr. Opin. Endocrinol. Diabetes Obes. *22*, 255–264.

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc. Natl. Acad. Sci. *105*, 8309–8314.

Sahlén, P., Abdullayev, I., Ramsköld, D., Matskova, L., Rilakovic, N., Lötstedt, B., Albert, T.J., Lundeberg, J., and Sandberg, R. (2015). Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol. *16*, 156.

Saint-André, V., Federation, A.J., Lin, C.Y., Abraham, B.J., Reddy, J., Lee, T.I., Bradner, J.E., and Young, R.A. (2016). Models of human core transcriptional regulatory circuitries. Genome Res. *26*, 385–396.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci.

112, E6456–E6465.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. *103*, 1412–1417.

Schalch, T., Duda, S., Sargent, D.F., and Richmond, T.J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. Nature *436*, 138–141.

Scharfmann, R., Xiao, X., Heimberg, H., Mallet, J., and Ravassard, P. (2008). Beta Cells within Single Human Islets Originate from Multiple Progenitors. PLoS One *3*, e3559.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep *17*, 2042–2059.

Schneider, R., and Grosschedl, R. (2007). Dynamics and interplay of nuclear architecture, genome organization, and gene expression. Genes Dev. *21*, 3027–3043.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. *25*, 582–597.

Schwartz, Y.B., and Pirrotta, V. (2013). A new world of Polycombs: unexpected partnerships and emerging functions. Nat. Rev. Genet. *14*, 853–864.

Shen, S.Q., Myers, C.A., Hughes, A.E.O., Byrne, L.C., Flannery, J.G., and Corbo, J.C. (2016). Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res. *26*, 238–255.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shin, H.Y., Willi, M., Yoo, K.H., Zeng, X., Wang, C., Metser, G., and Hennighausen, L. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat. Genet.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. *100*, 15776–15781.

Shlyueva, D., Stampfel, G., and Stark, A. (2014a). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Shlyueva, D., Stelzer, C., Gerlach, D., Yáñez-Cuna, J.O., Rath, M., Boryń, Ł.M., Arnold, C.D., and Stark, A. (2014b). Hormone-Responsive Enhancer-Activity Maps Reveal Predictive Motifs, Indirect Repression, and Targeting of Closed Chromatin. Mol. Cell 1–13.

Shogren-Knaak, M. (2006). Histone H4-K16 Acetylation Controls Chromatin Structure and Protein Interactions. Science (80-.). *311*, 844–847.

Simon, J. a, and Kingston, R.E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nat. Rev. Mol. Cell Biol. *10*, 697–708.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat. Genet. *38*, 1348–1354.

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555–568.

Splinter, E. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. *20*, 2349–2354.

Splinter, E., and de Laat, W. (2011). The complex transcription regulatory landscape of our genome: control in three dimensions. EMBO J. *30*, 4345–4355.

Stitzel, M.L., Sethupathy, P., Pearson, D.S., Chines, P.S., Song, L., Erdos, M.R., Welch, R., Parker, S.C.J., Boyle, A.P., Scott, L.J., et al. (2010). Global Epigenomic Analysis of Primary Human Pancreatic Islets Provides Insights into Type 2 Diabetes Susceptibility Loci. Cell Metab. *12*, 443–455.

Sutherland, H., and Bickmore, W.A. (2009). Transcription factories: gene expression in unions? Nat. Rev. Genet. *10*, 457–466.

Tallapragada, D.S.P., Bhaskar, S., and Chandak, G.R. (2015). New insights from monogenic diabetes for "common" type 2 diabetes. Front. Genet. *6*.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1–17.

Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., Schaffner, S.F., et al. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell *165*, 1519–1529.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Mol. Cell *10*, 1453–1465.

Towbin, B.D., Meister, P., Pike, B.L., and Gasser, S.M. (2010). Repetitive Transgenes in C. elegans Accumulate Heterochromatic Marks and Are Sequestered at the Nuclear Envelope in a Copy-Number- and Lamin-Dependent Manner. Cold Spring Harb. Symp. Quant. Biol. *75*, 555–565.

Tronche, F., and Yaniv, M. (1992). HNF1, a homeoprotein member of the hepatic transcription regulatory network. BioEssays *14*, 579–587.

Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S., et al. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell.

Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T.M., Fernandez, N., Ballester, B., Andrau, J.C., and Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat. Commun. *6*, 6905.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Varala, K., Li, Y., Marshall-Colón, A., Para, A., and Coruzzi, G.M. (2015). "Hit-and-Run" leaves its mark: Catalyst transcription factors and chromatin modification. BioEssays *37*, 851–856.

Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., and Reddy, T.E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res. *25*, 1206–1214.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. *22*, 1680–1688.

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., Lango Allen, H., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat. Genet. *46*, 61–64.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at

Key Cell Identity Genes. Cell *153*, 307–319.
Wilcox, G. (2005). Insulin and insulin resistance. Clin. Biochem. Rev. *26*, 19–39.
Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. (2007). DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. *36*, D88–D92.
Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. F1000Research.
Wit, E. De, and Laat, W. De (2012). A decade of 3C technologies-insights into nuclear organization. 11–24.
de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Mol. Cell *60*, 676–684.
Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G.C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell *66*, 285–299.e5.
Xu, Z., Wei, G., Chepelev, I., Zhao, K., and Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. Nat. Struct. Mol. Biol. *18*, 372–378.
Yan, H. (2002). Allelic Variation in Human Gene Expression. Science (80-.). *297*, 1143–1143.
Yang, W., de las Fuentes, L., Dávila-Román, V.G., and Charles Gu, C. (2011). Variable set enrichment analysis in genome-wide association studies. Eur. J. Hum. Genet. *19*, 893–900.
Ye, Z., Chen, Z., Sunkel, B., Frietze, S., Huang, T.H.-M., Wang, Q., and Jin, V.X. (2016). Genome-wide analysis reveals positional-nucleosome-oriented binding pattern of pioneer factor FOXA1. Nucleic Acids Res. gkw659.
Zabidi, M. a., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2014). Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation. Nature *1*.
Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: Establishing competence for gene expression. Genes Dev. *25*, 2227–2241.
Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.
Zhang, X., Choi, P.S., Francis, J.M., Imielinski, M., Watanabe, H., Cherniack, A.D., and Meyerson, M. (2015). Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nat. Genet. *48*, 176–182.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.
Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M.Q., et al. (2016). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. *44*, D203–D208.
Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. *38*, 1341–1347.
Zhu, H., Wang, G., and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. Nat. Rev. Genet. *17*, 551–565.
Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. U. S. A. *111*, 996–1001.

Appendix A

Copyright permissions

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057120492570
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Metazoan promoters: emerging characteristics and insights into transcriptional regulation
Licensed Content Author	Boris Lenhard, Albin Sandelin and Piero Carninci
Licensed Content Date	Apr 1, 2012
Licensed Content Volume	13
Licensed Content Issue	4
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Table 2: Promoter types Promoter.
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa Du Cane Road

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057130312159
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	Regulation of disease-associated gene expression in the 3D genome
Licensed Content Author	Peter Hugo Lodewijk Krijger, Wouter de Laat
Licensed Content Date	Nov 9, 2016
Licensed Content Volume	17
Licensed Content Issue	12
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
High-res required	no
Figures	Figure 1 | Mechanism of gene expression regulation by enhancers. Figure 3 | Erroneous regulatory wiring between enhancers and target genes causing disease.
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057130602261
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	Regulation of disease-associated gene expression in the 3D genome
Licensed Content Author	Peter Hugo Lodewijk Krijger, Wouter de Laat
Licensed Content Date	Nov 9, 2016
Licensed Content Volume	17
Licensed Content Issue	12
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Box 2 | Roadmap: from disease-associated genetic variants to molecular mechanisms causing disease
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa Du Cane Road

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057110948442
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	The selection and function of cell type-specific enhancers
Licensed Content Author	Sven Heinz, Casey E. Romanoski, Christopher Benner, Christopher K. Glass
Licensed Content Date	Feb 4, 2015
Licensed Content Volume	16
Licensed Content Issue	3
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 3 | Cell type-specific enhancer selection and activation
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa Du Cane Road

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057110120430
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Drug Discovery
Licensed Content Title	Histone lysine demethylases as targets for anticancer therapy
Licensed Content Author	Jonas W. Højfeldt, Karl Agger, Kristian Helin
Licensed Content Date	Nov 15, 2013
Licensed Content Volume	12
Licensed Content Issue	12
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1: Readers, writers and erasers.
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa Du Cane Road

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Apr 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4097100851665
License date	Apr 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Methods
Licensed Content Title	How best to identify chromosomal interactions: a comparison of approaches
Licensed Content Author	James O J Davies, A Marieke Oudelaar, Douglas R Higgs, Jim R Hughes
Licensed Content Date	Jan 31, 2017
Licensed Content Volume	14
Licensed Content Issue	2
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1 | Common principles in 3C-based techniques
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057120014016
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Structural and Molecular Biology
Licensed Content Title	Structure and mechanism of the RNA polymerase II transcription machinery
Licensed Content Author	Steven Hahn
Licensed Content Date	Apr 27, 2004
Licensed Content Volume	11
Licensed Content Issue	5
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1 The pathway of transcription initiation and reinitiation for RNA Pol II
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Feb 27, 2017

This Agreement between Joan Ponsa ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4057120953983
License date	Feb 27, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data
Licensed Content Author	Job Dekker, Marc A. Marti-Renom, Leonid A. Mirny
Licensed Content Date	May 9, 2013
Licensed Content Volume	14
Licensed Content Issue	6
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
High-res required	no
Figures	Box 2: Genome compartments.
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Three-dimensional chromatin organisation in human pancreatic islets
Expected completion date	May 2017
Estimated size (number of pages)	100
Requestor Location	Joan Ponsa Du Cane Road Londin, W12 0NN United Kingdom Attn: Joan Ponsa
Billing Type	Invoice
Billing Address	Joan Ponsa Du Cane Road

Londin, United Kingdom W12 0NN
Attn: Joan Ponsa

Total 0.00 GBP

Terms and Conditions

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer*, the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

   Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

   Note: For translation from the *British Journal of Cancer,* the following credit line applies.
   Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.
Special Terms:
v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B

Publications

This section contains the candidate's published papers, whose results are not part of the thesis.

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., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. 46, 136–143.

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

Zhao, L., Oliver, E., Maratou, K., Atanur, S.S., Dubois, O.D., Cotroneo, E., Chen, C.-N., Wang, L., Arce, C., Chabosseau, P.L., et al. (2015). The zinc transporter ZIP12 regulates the pulmonary vascular response to chronic hypoxia. Nature 524, 356–360.