



UNIVERSITAT DE
BARCELONA

Gene regulatory networks in pancreatic islets and insulinoma

Richard Norris

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



UNIVERSITAT DE
BARCELONA

Programa de Doctorat en Biomedicina

Facultat de Biologia

Gene regulatory networks in pancreatic islets and insulinoma

Richard Norris

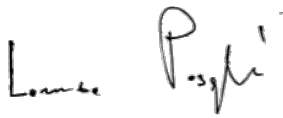
Thesis submitted for the degree of Doctor of Philosophy
to the Doctoral Program in Biomedicine.

Supervised by Lorenzo Pasquali MD PhD

Universitat de Barcelona

May 2021

Thesis developed under the supervision of Dr. Lorenzo Pasquali in the Endocrine
Regulatory Genomics Lab at the Germans Trias i Pujol Research Institute (IGTP) and
Universitat Pompeu Fabra, and advised by Professor Josep Francesc Abril Ferrando at the
Universitat de Barcelona



Lorenzo Pasquali
Director



Josep Francesc Abril Ferrando
Tutor

Digitally signed by JOSE
FRANCISCO ABRIL FERRANDO
- DNI 40932452L
DN: c=ES, sn=ABRIL
FERRANDO, givenName=JOSE
FRANCISCO,
serialNumber=IDCES-4093245
2L, cn=JOSE FRANCISCO
ABRIL FERRANDO - DNI
40932452L
Date: 2021.05.25 20:02:03
+02'00'



Richard Norris
Candidate

Acknowledgements

Many people who have completed a PhD will tell you that it's really tough, and they would be right, but the people you have around you can make it easier. Since long before I began working on my doctorate my wife has been by my side, encouraging me, making sacrifices and generally being amazing. It's been a long journey, but I couldn't have done it without her.

I always wanted to have the experience of living outside the UK, and it was not easy to find the right opportunity, but I feel very lucky that I found the Endocrine Regulatory Genomics lab. Barcelona has been an amazing place to live and do research. I want to thank my supervisor Lorenzo for giving me the opportunity, for challenging me, for pushing me, and for generally being a great guy to have a beer with. I also want to thank the rest of the team for their contributions to the project, especially Helena for her training and support in all things wet lab, Mireia for making me a better bioinformatician, Marc for the WGS analysis, and everyone, including the newest member of the team, Bea, for being great people to work with and hang out with. I'm proud of what we've achieved in the four years that I've been in the lab. I'd also like to thank my tutor at UB Professor Josep Francesc Abril Ferrando for his guidance and enthusiasm.

Outside of our lab there are many people who I worked alongside at Institute Recerca Germans Trias i Pujol, who made contributions large and small, to my doctoral work. I'd like to thank all the groups that we worked with including those of David Sanchez, Mireia Jorda, Tanya Vivouri, Lauro Sumoy, Marcus Buschbeck, Sergio Alonso, and many others. In addition I'm so grateful to all the support staff at IGTP who made PhD life so much easier, including lab manager David Izquierdo, and sys admins Iñaki and Lloyd.

Many thanks to the labs who provided the samples, the San Raffaele Scientific Institute in Milan, Italy, Instituto Universitario del Hospital Italiano de Buenos Aires, Argentina and Germans Trias i Pujol Hospital, Badalona, Spain, and to CRG and BGI for providing

excellent sequencing services. I also have to thank those who funded the project, La Marató, Plan Nacional, ciberdem and Sociedad Española de Diabetes.

Last but not least, a big thank you to all my family and friends for all their support and encouragement.

Abstract

Insulinomas are extremely rare pancreatic neuroendocrine (PNET) tumours that develop from the β -cells of the pancreas. β -cells are the only cells capable of producing the hormone insulin, and play a major role in glucose homeostasis but have extremely low proliferative ability under normal physiological conditions. Insulinomas feature dysregulation of insulin secretion and aberrant proliferation. Insulinoma development has been associated with genetic and epigenetic alterations leading to loss of cell fate.

We collected a large set of insulinoma and human islet samples, and utilised published data from other PNETs. ChIP-seq of H3K27ac, ATAC-seq and RNA-seq were performed to infer the regulatory landscape of insulinomas and control samples. In silico pipelines and analyses methods were developed in order to characterise the β -cell physiological and aberrant regulatory networks, across different layers of gene regulation.

By comparing unaffected pancreatic islets and insulinomas I have identified widespread changes in enhancer activity and gene expression. I have described the gene-regulatory landscape of insulinomas, identified putative drivers and mechanisms of tumour development, and built the first insulinoma-specific gene regulatory networks.

Contents

List of Figures	viii
List of Tables	x
List of abbreviations / acronyms	xi
1 Introduction	1
1.1 Pancreatic islets, β -cells and insulin secretion.	2
1.1.1 β -cells, the most abundant cell type in the pancreatic islets.	2
1.1.2 Development and identity of β -cells (and other islet cells).	4
1.1.3 Insulin production and secretion	7
1.1.4 β -cell proliferation	9
1.2 Insulinoma - a functional pancreatic neuroendocrine tumour.	12
1.2.1 Epidemiology, diagnosis and treatment.	12
1.2.2 Mutations associated with insulinoma	16
1.3 Gene regulation	17
1.3.1 Deoxyribonucleic acid (DNA) and chromatin.	18
1.3.2 Genes defined	20
1.3.3 The human genome and the origin of complexity.	20
1.3.4 Transcription regulation by cis-regulatory elements.	21
1.3.5 Epigenetic modification of chromatin at CREs.	24
1.3.6 Gene regulatory alterations in cancer.	26
1.4 Regulatory genomics in the era of high-throughput sequencing	27
1.5 Motivation	28
2 Hypothesis and objectives	29
2.1 Hypothesis	30
2.2 Objectives	30

3	Materials and methods	31
3.1	Ethics.	32
3.2	Wet lab	32
3.2.1	Samples	32
3.2.2	ChIP-seq.	35
3.2.3	ATAC-seq.	37
3.2.4	RNA-seq.	38
3.2.5	<i>YY1</i> T372R genotyping.	38
3.3	Computational analysis	39
3.3.1	NGS analysis pipelines.	39
3.3.2	Count matrices	39
3.3.3	Transcript quantification	40
3.3.4	Regulatory element conservation.	40
3.3.5	Regulatory element coverage and sharing.	40
3.3.6	Correlation clustering.	41
3.3.7	Differential analysis	42
3.3.8	Pre-ranked gene set enrichment analysis.	44
3.3.9	Chromatin accessibility within CREs	45
3.3.10	Super-enhancers	45
3.3.11	Comparison to chromHMM and H3K27me3 data	45
3.3.12	Selection of CRE to TSS distance cutoff	46
3.3.13	Overlap-permutation tests.	46
3.3.14	Networks	47
3.3.15	Non-functional PNETs.	47
4	Genome-wide profiling of gene expression and CREs in pancreatic islets and insulinoma.	48
4.1	Transcriptome profiling of insulinomas.	49
4.2	Optimisation of chromatin immunoprecipitation	50

4.3	Insulinoma and control ChIPs showed strong signal enrichment at key β -cell factors.	53
4.4	QC, genome alignment and peak calling for ChIP-seq data.	54
4.5	A comprehensive profile of active CREs.	55
4.6	Characteristics of H3K27ac enriched regions.	57
4.7	Evaluation of accessible chromatin within CREs.	60
4.8	Summary and concluding remarks.	61
5	The gene-regulatory landscape of insulinoma.	64
5.1	Genotyping known recurrent mutations.	65
5.2	Insulinoma development is driven by changes in gene expression.	66
5.2.1	Data correlation and hierarchical clustering	67
5.2.2	Differential analysis of gene expression.	68
5.2.3	Upregulated genes are enriched for chromatin modifiers.	70
5.3	Profiling CREs enriched for H3K27ac genome-wide can distinguish insulinoma from human islets and non-functional PNETs.	72
5.4	Insulinoma-enriched CREs.	75
5.5	Motif analysis for putative insulinoma CREs	78
5.6	Summary and concluding remarks	79
6	Insulinoma-specific gene-regulatory networks.	81
6.1	Insulinoma-specific CREs are linked to changes in gene expression.	82
6.2	Putative enhancer-promoter links hint at mechanisms of insulinoma development.	83
6.3	Insulinoma chromatin architecture develops via de-repression and activation of CREs.	85
6.4	Gene-regulatory networks.	86

7 Discussion	91
8 Conclusions	97
9 Supplementary methods	98
9.1 CHIPmentation	98
9.2 Scripts	106
References	107

List of Figures

1	The composition of pancreatic islets in the human pancreas.	3
2	Development of islet cell types from Endocrine progenitors.	5
3	Functional interactions of key transcription factors involved in the development of cells of the endocrine pancreas in humans.	6
4	Glucose-stimulated insulin secretion from pancreatic β -cells.	9
5	Model of β -cell proliferation in murine cells.	10
6	Locations of neuroendocrine tumours in the human body.	13
7	Incidence and clinical features of functional pancreatic neuroendocrine tumours.	14
8	Classification and grading system for insulinoma.	15
9	Schematic representation of the structure of DNA and a nucleosome.	19
10	Enhancer-promoter interactions	23
11	Schematic representation of histone modifications associated with repressed, poised and active enhancer states.	25
12	Density plot of proliferation index (Ki67) values for insulinoma samples.	34
13	Schematic of the ChIPmentation protocol.	36
14	Sonication results.	51
15	The relationship between immunoprecipitate SDS concentration and ChIP-seq library concentration.	52
16	Enrichment qPCR results.	53
17	Fastqc analysis results for insulinoma ChIP-seq of H3K27ac.	54
18	Coverage plots for H3K27ac enriched regions in insulinoma and human islets.	56
19	Distribution of H3K27ac enriched regions.	58
20	Conservation of insulinoma CREs.	59
21	Identifying TF binding sites within CREs.	61
22	Genotyping of the T372R mutation in <i>YY1</i>	66
23	Insulinoma development is characterised by significant changes in gene expression.	67

24	Differential analysis of transcript abundance.	69
25	Gene-set enrichment analysis identifies up-regulated gene-sets in insulinoma.	71
26	Genes in up-regulated pathways show higher than average expression.	72
27	Cluster analysis of H3K27ac ChIP-seq data.	73
28	Principal component analysis of H3K27ac data	74
29	Differential analysis of H3K27ac ChIP-seq data	75
30	Conservation of insulinoma-specific CREs.	76
31	Distribution of super-enhancers across human islets and PNETs.	77
32	Motif enrichment in differentially active CREs.	79
33	Linking enhancer activity to gene expression.	82
34	Results of overlap permutation tests.	84
35	Overlap of CREs with H3K27me3 enriched regions and Chromm HMM analysis from unaffected human islets.	86
36	Stringdb interaction network 1.	87
37	Stringdb interaction network 2.	90

List of Tables

1	Patient data for human islet samples.	33
2	Primer sequences for positive and negative control qPCRs.	37
3	Quality control for Insulinoma RNA samples.	49
4	RNA-seq read and quality data for insulinoma samples.	50
5	Genome alignment and peak calling statistics for human islet and insulinoma samples.	55
6	Summary of experiments for all insulinoma samples in the study cohort . . .	62
7	Summary of data from motif analysis.	79

List of abbreviations / acronyms

5-IT	5-iodo-tubericidin
AJCC	American Joint Committee on Cancer-specific
ARX	Aristlass-related homonox gene-level
ATAC-seq	Assay for transposase-accessible chromatin with sequencing
ATP	Adenosine triphosphate
BAM	Binary alignment map
BRD4	Bromodomain-containing protein 4
CPS	Consensus peak set
CRE	Cis regulatory element
CDK1	Cyclin-dependent kinase 1
CDKN1C	Cyclin Dependent Kinase Inhibitor 1C
ChIP-seq	Chromatin immunoprecipitation with high-throughput sequencing
DE	Differential expression
DNA	Deoxybrinucleic acid
DP	Double positive
DYRK1A	Dual specificity tyrosine-phosphorylation-regulated kinase 1A
EBF1	Early B cell factor 1
EGFR	Epidermal growth factor receptor
ENETS	European Neuroendocrine Tumour Society
EPI	Enhancer-promoter interaction
ERG	ETS-related gene
eRNA	enhancer RNA
ES	Enrichment score
ETS	E26 transormation-specific
FGSEA	Fast GSEA
FDR	False discovery rate

GAGE	Generally applicable GSEA
GLP-1	Glucagon-like peptide-1
GLUT1	Glucose transporter 1
GLUT2	Glucose transporter 2
GREAT	Genomic Regions Enrichment of Annotations Tool
GRN	Gene regulatory network
GSEA	Gene set enrichment analysis
GSIS	Glucose-stimulated insulin secretion
GSK3	Glucose synthase kinase 3
GTPase	Guanine triphosphatase
H3K4me1	Histone 3 Lysine 4 monomethylation
H3K4me3	Histone 3 Lysine 4 trimethylation
H3K27ac	Histone 3 Lysine 27 acetylation
H3K27me2	Histone 3 Lysine 27 trimethylation
HAT	Histone acetyltransferase
HDM	Histone demethylase
HGF	Hepatocyte growth factor
HI	Human islets
hPSC	human pluripotent stem cells
iPSC	induced pluripotent stem cells
ICGC	International Cancer Genome Consortium
IGF	Insulin growth factor
INSR	Insulin receptor
IPF1	Insulin promoter factor 1
IRS	Insulin receptor substrates
ISL1	ISL LIM Homeobox 1
KAT	Lysine acetyltransferase
mTOR	mammalian target of rapamycin

MAFA	MAF BZIP Transcription Factor A
MEF2C	Myocyte enhancer factor 2C
MEN1	Multiple endocrine neoplasia type 1
mESC	mouse embryonic stem cell
MPC	Multipotent pancreatic progenitors
NB	Negative binomial
NDR	Nucleosome depleted region
NET	Neuroendocrine tumour
NF1	Neurofibromatosis type 1
NFAT	Nuclear factor activated in T cells
NFR	Nucleosome free region
NGN3	Neurogenin 3
NKX6.1	NK6 Transcription Factor Related, Locus 1
NEUROD1	Neurogenic differentiation factor D1
PAX4	Paired box gene 4
PCR	Polymerase chain reaction
PDX1	Pancreatic and duodenal homeobox
PI3K	Phosphatidylinositol-3 kinase
PDK-1	Phosphoinositide-dependent kinase-1
PKC	Protein kinase C
QL	Quasi likelihood
PNET	Pancreatic neuroendocrine tumour
RiN	RNA integrity number
RNA	Ribonucleic acid
mRNA	messenger RNA
RNA polII	RNA polymerase II
RTK	Receptor tyrosine kinase
SAM	Sequence alignment map

SE	Super-enhancer
SMARC	SWI/SNF related, matrix associated, actin dependent, regulator of chromatin
SV	Structural variant
T2D	Type 2 diabetes
TF	Transcription factor
TNM	Tumour nodes, metastases
VEGF	Vascular endothelial growth factor
WHO	World Health Organisation
YY1	Yin Yang 1

1 Introduction

1.1 Pancreatic islets, β -cells and insulin secretion.

The pancreas could be described as two (or even four) organs in one, as its different sections and cell types have different developmental origins and functions. The bulk of the mature pancreas in humans is an exocrine gland comprising acinar and ductal cells, that produce digestive enzymes and sodium bicarbonate respectively, responsible for digestion and absorption of foodstuffs. The endocrine gland is distributed throughout the pancreas and is comprised of five different types of secretory islet cells. The principal function of islets is the maintenance of glucose homeostasis through the production and release of insulin and glucagon. Pancreatic secretory functions are tightly regulated by a variety of mechanisms, and aberrant activation or inactivation of regulatory pathways has significant consequences in terms of human health and disease.

1.1.1 β -cells, the most abundant cell type in the pancreatic islets.

Pancreatic islets, first described by Paul Langerhans in 1869 are clusters of approximately 1000 cells scattered throughout the pancreas. There is significant variation in cellular composition and architecture of pancreatic islets between species. But human islet samples are difficult to obtain, so much of our knowledge regarding human islet structure and function is derived from experiments using rodent islet samples. Human islets consist of insulin-producing β -cells ($\approx 70\%$), glucagon-producing α -cells ($\approx 20\%$), with the remainder ($\approx 10\%$) made up of somatostatin-producing δ -cells, ghrelin-producing ϵ -cells and pancreatic polypeptide-producing PP cells [1]. Murine islets, the most well studied, feature a β -cell core surrounded by the other islet cell types, and a slightly higher $\beta : \alpha$ cells ratio [2]. Human islets have been described as having a more scattered organisation of endocrine cells [3], and this scattering may extend to the positioning of islets within the pancreas [4]. Reports of variation in insulin secretion in patients after partial pancreatectomy also suggests functional differences in islets from different parts of the pancreas (e.g. head vs tail) [1].

β -cells are the only cells capable of producing and secreting insulin, and loss of β -cell mass results in the dysregulation of glucose homeostasis which leads to diabetes mellitus. The discovery of insulin and its ability to lower blood glucose earned Frederick Banting and Charles Best a Nobel Prize in 1923, and has since been the subject of extensive research due to the loss of β -cell mass/function associated with diabetes. But while the loss of pancreatic β -cell mass leads to disease, there appears to be little or no effect on physiology from near complete loss of α -cell mass [5]. The major function of glucagon is as a counter-regulatory response to insulin, and the prevention of hypoglycaemia, and elevated glucagon secretion has been observed in type 2 diabetes (T2D) in addition to reduced insulin levels. However, despite recent efforts, a lot less is known about α -cells compared to β -cells.

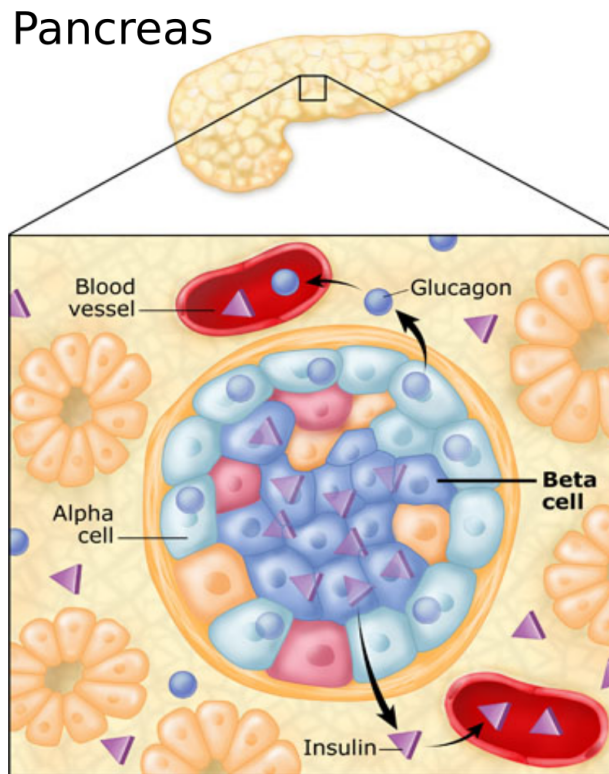


Figure 1: The composition of pancreatic islets in the human pancreas.

In simple terms a β -cell is a cell that produces insulin and secretes it in response to physiological glucose levels. Insulin is stored in secretory granules as a complex with zinc and released in response to high glucose concentrations [6] and stimulation by neurotransmitters

[7]. Several other factors also contribute to the regulation of insulin secretion, including inhibition by somatostatin and ghrelin. Furthermore, studies have suggested significant heterogeneity in β -cell identity and function within an islet [8]. Mapping of islet functional architecture revealed the spatial arrangement of β -cells within islets, including hub cells with pacemaker properties, enabling specific cell-cell communication patterns and insulin release via 'rhythmic activity'.

1.1.2 Development and identity of β -cells (and other islet cells).

β -cells develop from a subset of precursor cells that have adopted an endocrine fate and themselves represent a subset of cells of the posterior foregut with a pancreatic identity [9]. It is possible to build a putative map of β -cell development using expression patterns and knockout phenotypes of associated genes (fig.2). Although more is known about β -cell development in mice, evidence of the involvement of the various factors in this lineage map is derived from experiments in human multipotent pancreatic progenitors (MPCs) and pluripotent stem cells (hPSCs). The most important regulator of endocrine specification is Neurogenin 3 (NGN3) [10], expression of which activates several transcription factors (TFs) including PAX4, ARX, NKX2.2, NKX6.1, ISL1, NEUROD and INSM1, which are responsible for the differentiation of endocrine precursors into mono-hormonal islet cells [11]. *NGN3* expression is transient and disappears in mature endocrine cells, but the timing and level of expression is important in determining the fate of endocrine progenitors [12]. One feature of endocrine cell lineages that stands out is the number of factors implicated in α and β cell development compared to the other endocrine cell types. This may be reflective of specific biological needs, although it may also be due to the fact that the majority of studies have focused on α and β cells and therefore much more is known about them.

One of the earliest TFs to be expressed in pancreatic progenitors is *SOX9*, which plays a major role in regulating the expression of TFs, such as *HNF6* and *FOXA2*, in the transcriptional network of pancreatic progenitors. Cells of both endocrine and exocrine pancreatic

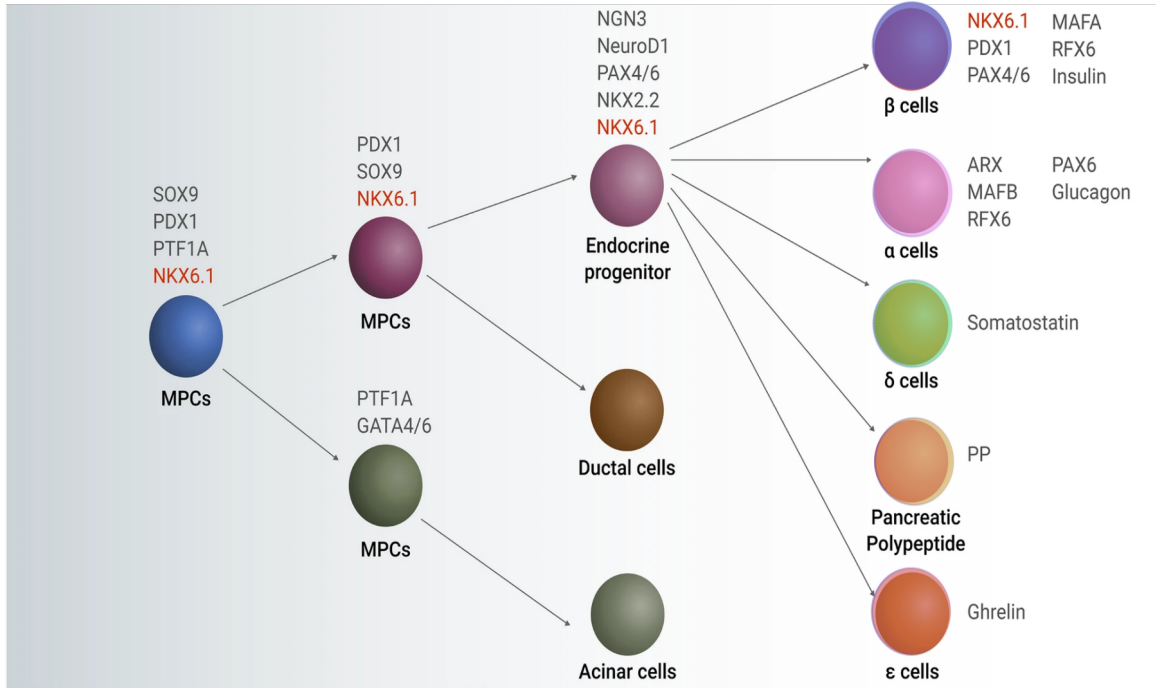


Figure 2: Development of islet cell types from Endocrine progenitors. A lineage diagram of pancreatic endocrine cell development from $Ngn3^+$ progenitor cells [12].

lineages are derived from SOX9-positive progenitors [13]. SOX9 activates *NGN3* and along with FOXA2 and TCF2 controls the activation of genes involved in endocrine cell differentiation [14]. During pancreatic development in mice *Foxa1* and *Foxa2* regulate pancreatic and duodenal homeobox 1 (*Pdx1*) (also known as insulin promoter factor 1 (IPF1)), which is expressed in all pancreatic precursor cells, by binding to a distal enhancer [15]. *PDX1* expression levels vary across pancreas developmental stages, but it is upregulated in the later stages of β -cell development. *PDX1* is a master regulator of β -cell fate, activating genes that specify β -cell identity and repressing genes that promote α -cell identity [16]. For example, overexpression of *Pdx1* in a murine cell line resulted in significant upregulation of insulin (*Ins1* and *Ins2*) [17].

NKX6.1 is expressed at early stages of pancreatic development and is ever-present in the β -cell lineage, playing a critical role in β -cell development [18]. During the MPC stage, expression of *NKX6.1* and *SOX9* is crucial, and these TFs interact to promote *NGN3* ex-

pression and the generation of endocrine progenitors . An antagonistic mechanism exists between NKX6.1 and PTF1A, with NKX6.1 promoting endocrine lineages and PTF1A a main regulator of acinar gene transcription. PTF1A represses *SOX9* and *NGN3* but is directly regulated by NKX6.1, enabling the switch to endocrine cell development [12]. This switch occurs during a specific window, after which progenitors are committed to an acinar or ductal fate, suggesting that this antagonism between TFs dictates the relative numbers of newly differentiated endocrine vs exocrine cells in the pancreas. Poly-hormonal and mono-hormonal endocrine cells appear at different stages in humans and mice [19], but *Nkx6.1*-knockout studies have shown that it is important in the transition from poly-hormonal cells to mature β -cells [18, 20].

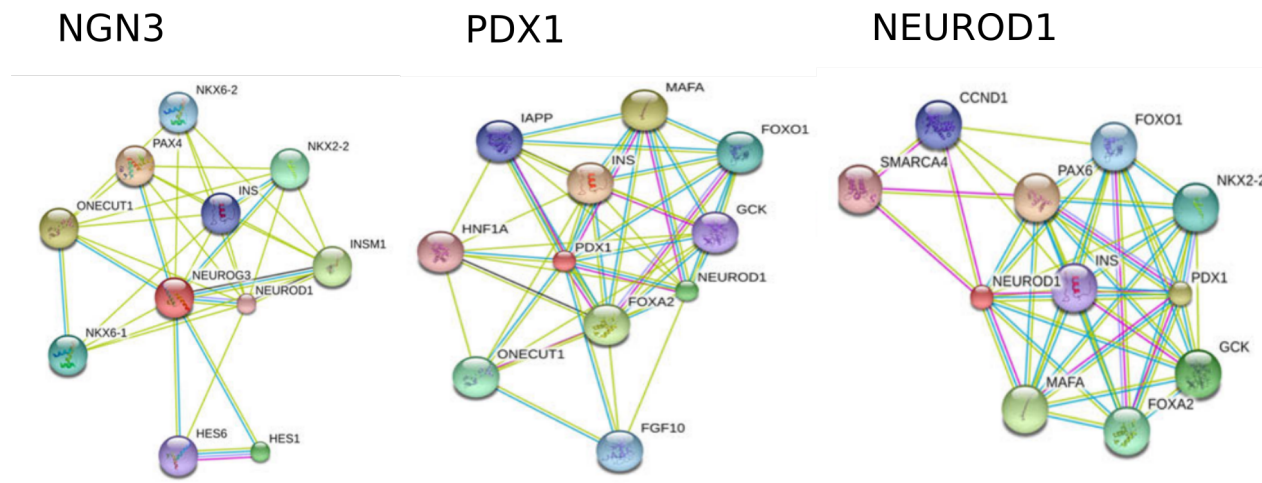


Figure 3: Functional interactions of key transcription factors involved in the development of cells of the endocrine pancreas in humans. Thicker lines represent stronger associations (minimum confidence value 0.4/1). Protein interaction maps created using STRINGdb and featured in Al-Khawaga et al 2018 [11].

Neurogenic differentiation factor D1 (*NEUROD1*) is another crucial TF for development of the endocrine pancreas and β -cell maturation. *NEUROD1* is activated by *NGN3* and is a transactivator of the insulin gene. It interacts with *INSM1* and *FOXA2* and co-occupies β -cell-specific regulatory elements. Disruption of *NEUROD1* binding sites at these genomic regions has been shown to result in β -cell dysfunction in humans [21]. Another *Ngn3* target, *Pax4* (paired box gene 4), is expressed strongly in β -cell precursors but is limited in α -

cells [22]. Its counterpart *Arx* (aristaless-related homeobox gene) drives specification of glucagon-producing α -cells. Precursor cells split into separate populations according to their expression of *Pax4* or *Arx*. Selective inhibition of *ARX* is sufficient to promote the conversion of α -cells into β -cells [23], but β -cells are the first and most predominant islet cell type to appear in human fetal development [24].

The specification of cell identity in cells of the endocrine pancreas occurs via sequential activation and repression of specific genes. Transcription factors with key roles in maintaining the mature β -cell phenotype include MAFA and RFX6. *Mafa* knockout mice have reduced β : α cell ratio, reduced expression of *Ins1* and *Ins2*, and upregulation of genes that are normally repressed in mature β -cells [25]. Gene set enrichment analysis in *Mafa*^{-/-} mice identified several β -cell signalling pathways including protein-binding, ion-binding, and receptor and mitogen-activated protein kinase signalling indicating widespread functional importance of *Mafa* in β -cells [26]. RFX6 is a downstream effector of NGN3 which promotes insulin production and secretion via upregulation of *INS* gene expression and calcium channels [27]. *FOXA2*, *MAFB* and *NKX2.2* are expressed in β -cells and in other islet cell types, while *NKX6.1* is specific to β -cells and *PDX1* is expressed primarily in β -cells but also at low levels in δ -cells. Down-regulation of MAFA, *NKX6.1* and FOXO1 contributes to de-differentiation in type 2 diabetic (T2D) patients [28], highlighting their importance in the maintenance of β -cell identity. ChIP-seq analysis in human islets revealed that these TFs form a remarkably interconnected network, with overlapping DNA-binding patterns suggesting auto- and cross-regulatory interactions. Chromatin conformation analysis also identified clusters of enhancers involved in combinatorial transcription factor binding [29].

1.1.3 Insulin production and secretion

Several of the factors described above have a role in regulating insulin biosynthesis, thus controlling the most important aspects of the identity of a β -cell: the production and secretion of insulin in a tightly regulated manner. In mammals, control of insulin gene expression,

including metabolic regulation, is centered around a highly conserved promoter/enhancer region located within 350bp immediately upstream of the transcription start site. Within this region are a series of regulatory elements, including E, A and C boxes, that are major determinants of insulin gene expression [30]. Pancreatic/duodenal homeobox-1 (PDX1) binds to the A boxes and functionally interacts with proteins of the basic helix-loop-helix family (including NeuroD1) that bind to E boxes [31], whilst MAFA binds to C boxes. PAX6 has also been shown to regulate insulin gene expression both directly (by binding to the insulin gene promoter) and indirectly via interactions with TFs including MAFA [32]. Together these interactions between transcription factors and the proximal cis-regulatory region make up the principal regulatory machinery of the insulin gene under normal conditions [33].

Healthy β -cells act as glucose sensors, matching insulin secretion to the circulating glucose concentration, and glucose regulates all steps of insulin gene expression, including transcription, preRNA splicing and mRNA stability. Glucose-responsive transcription control elements include A3, E1 and C1 as well as a more distal regulatory element that binds a glucose-sensitive complex [34]. Glucose infusion results in an increase in Akt activation, which increases *PDX1* expression [35]. Glucose also promotes the binding of PDX1 to the insulin promoter, in turn promoting recruitment of transcriptional co-activators including histone acetyltransferases [36].

Insulin secretion from β -cells is achieved via changes in electrical activity, characterised by membrane depolarisation and bursts of action potentials [37]. Glucose-stimulated insulin secretion (GSIS) begins with glucose entering β -cells via glucose transporter 1 and 2 (GLUT1 and GLUT2). Glucose is broken down by glycolysis to pyruvate, which then enters the mitochondria and is used to generate adenosine triphosphate (ATP) via the TCA cycle. This ATP is then transported to the cytoplasm where it acts to close ATP-sensitive potassium channels, thus depolarising the cell membrane, which in turn induces the opening of voltage-gated Ca^{2+} channels. Bursts of Ca^{2+} influx into β -cells push insulin vesicles to the cell membrane and insulin is released by exocytosis [38].

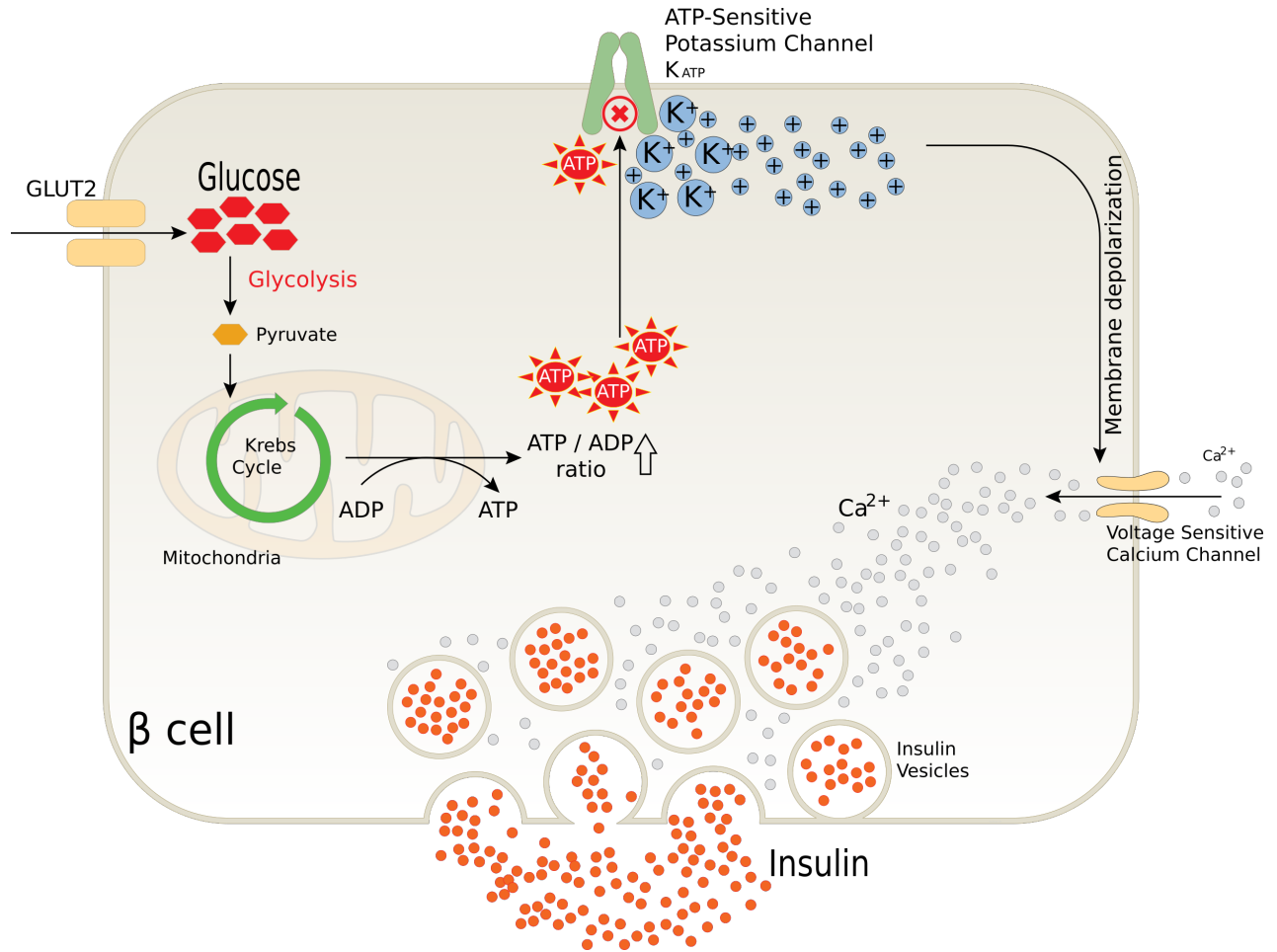


Figure 4: Glucose-stimulated insulin secretion from pancreatic β -cells.

1.1.4 β -cell proliferation

While early $PDX1^+$ pancreatic progenitor cells are proliferative, replication rates of differentiated β -cells decline rapidly in early childhood. With this comes a marked change in cell identity; as the secretory capacity (limited in immature cells) of β -cells is enhanced, their ability to replicate declines rapidly [39]. Mature human β -cells replicate at low rates ($\approx 2-4\%$ /day) and this replicative ability is limited to the first few years of life. By contrast cell proliferation in several other tissues is often 10-fold higher. Furthermore there is insufficient regeneration of β -cells to offset the loss of β -cell mass in diabetes and a major area of research is focused on finding ways to reverse these effects. Little is known about how

β -cell proliferation is regulated during the perinatal period in humans, so here again, findings from studies in murine β -cells provide most of the insights. There is a large and expanding intracellular signalling map describing how β -cell proliferation occurs in rodents (fig.5). Multiple mitogenic signalling pathways such as IrsPi3kAkt, Gsk3, mTor, ChREBP/cMyc, Ras/Raf/Erk, and Nfats integrate signals from growth factors and nutrients, including insulin/insulin growth factor.

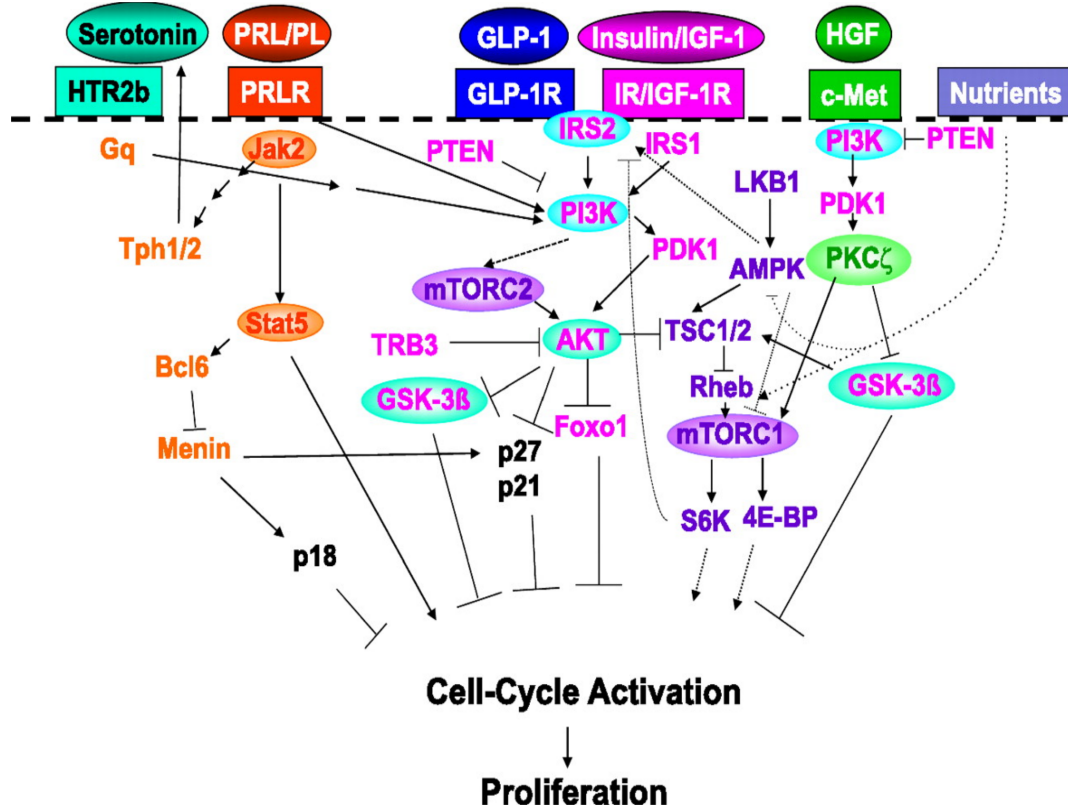


Figure 5: Model of β -cell proliferation in murine cells. Pathways and key factors involved in β -cell proliferation in mouse cell lines [40]. β -cell proliferation in mice features signalling by growth factors and hormones including insulin/insulin growth factor (IGF), hepatocyte growth factor (HGF) and glucagon-like peptide-1 (GLP-1). Subsequent phosphorylation of insulin receptor substrates (IRS) activates the phosphatidylinositol-3 kinase (PI3K)/Akt pathway. PI3K activation results in the production of phosphatidylinositol trisphosphate, which directly binds to phosphoinositide-dependent kinase-1 (PIDK-1). PIDK-1 then binds to and activates protein kinase C γ (PKC γ). (PKC γ) inactivates glucose synthase kinase 3 β (GSK3 β) and activates mTORC1. mTORC1 signalling controls cell growth and metabolism via further downstream effectors [40].

Mitogenic signals regulate the expression of downstream cell cycle regulators, pushing quies-

cent β -cells to re-enter the cell cycle. Circulating factors derived from other organs including intestinal peptides and adipose tissue-derived adipokines can also regulate β -cell proliferation during puberty, pregnancy and obesity. Curiously, rodent β -cells can display relatively large proliferative responses ($\approx 10 - 15\%/day$), which is never seen in humans, even in fetal and neonatal stages [40]. This enforced quiescence in adult human β -cells may have several molecular causes, but recent evidence points towards epigenetic factors including DNA methylation and chromatin modifying enzymes [41, 42]. Significant research efforts have been directed towards identifying drugs or growth factors able to induce beta cell replication. Many putative pathways, that may act as drug targets, and classes of drugs have been discovered. DIRK1A inhibitors provide the highest proliferation rates (up to 5%) in human β -cells and also represent the most widely replicated findings.

DYRK1A (Dual specificity tyrosine-phosphorylation-regulated kinase 1A) is a dual specificity kinase, possessing both serine/threonine and tyrosine kinase activities. Several different chemical entities, including harmine (a beta-carboline)[43] and 5-IT (5-iodo-tubericidin) [44], have been shown to increase the rate of β -cell proliferation via inhibition of DYRK1A. Dyrk1a is a negative regulator of the NFAT pathway, which is of crucial importance to murine β -cell proliferation [45]. The nuclear factor activated in T cells (NFAT) family of TFs bind to and activate cell cycle activating genes, including *CDK1* (cyclin-dependent kinase 1), and repress cell cycle inhibitor genes such as *CDKN1C*, thus activating cell cycle progression [43]. DYRK1A phosphorylates nuclear NFATs, preventing entry to the nucleus and thus terminating their mitogenic signal. DYRK1A inhibitor action prevents NFAT phosphorylation, allowing continued stimulation of cell cycle activation [46]. Harmine treatment also increased markers of β -cell differentiation including PDX1, NKX6.1 and MAFA [43].

In insulinomas, β -cells proliferate whilst maintaining some of the functional features of normal β -cells, including the capacity to synthesize and secrete insulin. But they lack the capacity to process, store and limit insulin secretion in response to the physiological glucose range such that proliferation leads to fasting hypoglycemia. Tumour cells in insulinomas therefore present an intriguing possibility to study the molecular mechanisms governing cell

proliferation and cell identity.

1.2 Insulinoma - a functional pancreatic neuroendocrine tumour.

Neuroendocrine tumours (NETs) are neoplasms that arise from neuroendocrine cells, which are characterised by the release of hormones or neuropeptides into the circulating blood in response to a neural stimulus. NETs can develop in several locations in the human body, the most common occurring in the gastrointestinal tract, lungs and appendix, and around 20% develop from the cells of pancreatic islets. Pancreatic neuroendocrine tumours (PNETs) are divided into two classes, functional (characterised by the abnormal secretion of hormones) and non-functional, with the later representing the majority of cases. Functional PNETs are named according to the hormone produced and secreted by the cells from which they derive, including insulinoma, glucagonoma and somatostatinoma.

Insulinomas are the most common type of functional PNET ($\approx 50\%$) and develop from β -cells. As described above, β -cells replicate readily in fetal and neonatal stages of development but their proliferative ability declines rapidly after these stages. The proliferation of β -cells in insulinoma therefore presents an intriguing phenomenon, and by investigating the aberrant molecular mechanisms driving insulinoma development we hope to gain insights into the factors affecting β -cell proliferation, and a broader understanding of β -cell identity.

1.2.1 Epidemiology, diagnosis and treatment.

PNETs, previously known as islet cell tumours, comprise less than 2% of all pancreatic neoplasms, with an incidence of ≈ 1 case per 100,000 individuals. Their incidence has increased significantly in recent years, although improvements in imaging technology has also improved detection. Functional PNETs typically present with clinical syndromes related to the hypersecretion of hormones (fig.7), which in the case of insulinoma is hyperinsulinaemic hypoglycemia. Cells in non-functional PNETs still produce and secrete the hormone associ-

Anatomical Distribution of Neuroendocrine Tumors

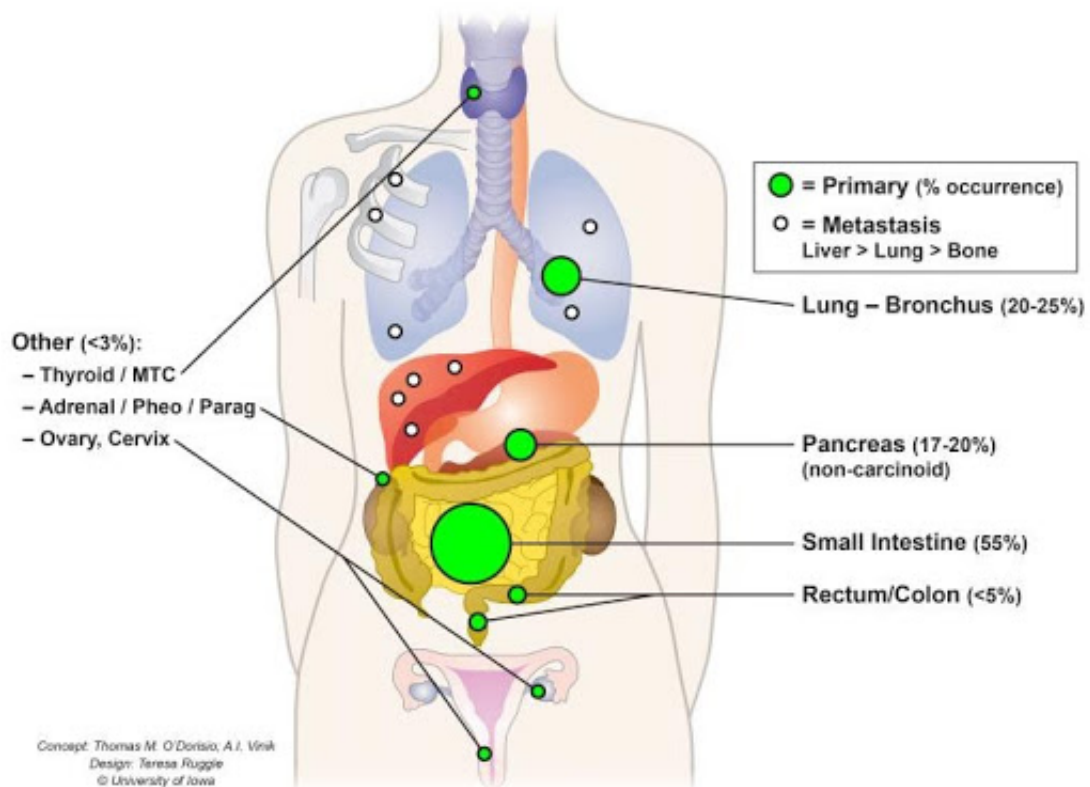


Figure 6: Locations of neuroendocrine tumours in the human body.

ated with the cell type from which they originate, but the cause of morbidity and mortality is associated with expansion of the tumour mass rather than hormone-related symptoms. Insulinomas are extremely rare (1 to 4 cases per million of population) and present as single, small (usually around 2cm in diameter) lesions. More than 90% of insulinomas are benign and yet if left untreated the uncontrolled release of insulin can cause severe symptoms associated with hypoglycaemia including convulsions and coma, and can even be fatal.

Benign NETs tend to develop slowly, and the average time between symptoms appearing and diagnosis of insulinoma is about 3 years. Insulinomas can develop at any age, but the peak incidence is between 30 and 60 years of age, and are more prevalent in females. They can be difficult to diagnose as the milder symptoms (including confusion and weakness) can have several causes. When insulinomas are suspected, blood sugar is monitored and

Tumor	Incidence	Location	Secreted hormone	Malignant (%)	Clinical features	Biochemical diagnosis
Insulinoma	40–55%	Pancreas (>99%)	Insulin	<10	Hypoglycemic syndromes (Whipple's triad)	Insulin ≥ 5 mIU/l ^a Glucose 40 mg/dl C-peptide 0.6 ng/ml Proinsulin ≥ 20 pmol/l (or >25% of immunoreactive insulin)
Gastrinoma	25–50%	Duodenum (70%) Pancreas (25%) Others (5%)	Gastrin	60–90	Zollinger–Ellison syndrome (abdominal pain, gastroesophageal reflux, diarrhea, duodenal ulcers, PUD/GERD)	Serum gastrin level ≥ 10 times normal range + gastric pH < 2
Glucagonoma	Rare	Pancreas (100%)	Glucagon	50–80	Rash, glucose intolerance, necrolytic migratory erythema, weight loss	Glucagon > 500 pg/ml
Somatostatinoma	Rare	Pancreas (55%) Duodenum-jejunum (45%)	Somatostatin	>70	Diabetes mellitus, cholelithiasis, diarrhea	Somatostatin-fasting serum level
VIPoma (Verner–Morrison)	Rare	Pancreas (90%) Other (10%)	Vasoactive intestinal peptide	40–70	WHDA	VIP fasting serum level
ACTHoma	Rare	Pancreas (4–16% all ectopic Cushing's)	Adreno Cortico Tropic Hormone (ACTH)	95	Cushing's syndrome	
pNET-causing carcinoid syndrome	Rare	Pancreas (100%)	Serotonin, tachyins	60–90	Carcinoid syndrome	Urinary 5-HIAA in a 24-h urine collection

GERD, gastroesophageal reflux disease; 5-HIAA, 5-hydroxyindoleacetic acid; PUD, peptic ulcer disease; VIP, vasoactive intestinal peptide; WHDA, watery diarrhea, hypokalemia, achlorhydria.

^aShould be performed at the time of hypoglycemia during prolonged fasting (up to 72 h).

Figure 7: Incidence and clinical features of functional pancreatic neuroendocrine tumours [47].

imaging tests are used to check the size and precise location of the tumour. The gold standard test for diagnosing insulinoma is a 72-hour fast during which levels of insulin, plasma glucose, C peptide, proinsulin and beta-hydroxybutyrate are monitored in order to establish whether hypoglycemia induction is due to hyperinsulinemia. Histopathology tests can also support insulinoma diagnosis. Immunohistochemistry for insulinoma includes staining for insulin, chromogranin A (a protein released from neuroendocrine cells) and Ki-67 (a marker of proliferative cells) [48]. There are several classification and grading systems, including World Health Organisation (WHO), European Neuroendocrine Tumour Society (ENETS) and American Joint Committee on Cancer (AJCC) for PNETs, providing essential prognostic values. Figure 8 outlines a TNM (tumour, nodes, metastases) staging system for

insulinoma.

TNM staging

- TX - Tumor cannot be assessed
- T1 - Tumor limited to the pancreas, less than 2 cm
- T2 - Tumor limited to the pancreas, 2 to 4 cm
- T3 - Tumor limited to the pancreas, greater than 4 cm; or tumor invading the duodenum or common bile duct
- T4 - Tumor invasion of adjacent organs (e.g., stomach, spleen, colon, adrenal gland), or the walls of large vessels (celiac axis or the superior mesenteric artery)
- NX - Regional lymph nodes cannot be assessed
- N0 - No regional lymph node involvement
- N1 - Regional lymph node involvement
- M0 - No distant metastasis
- M1 - Distant metastasis
- M1a - Metastasis confined to the liver
- M1b - Metastasis in at least one extrahepatic site (e.g., lung, ovary, nonregional lymph node, peritoneum, bone)
- M1c - Both hepatic and extrahepatic metastases

Figure 8: Classification and grading system for insulinoma.

Single, sporadic insulinoma is curable by surgical intervention involving either enucleation or laparoscopic resection. Prior to surgery treatment strategies include dietary modification, MedicAlert bracelets, glucagon pens, somatostatin analogs, and steroids. In cases where surgery is not an option drugs like diazoxide and everolimus can be used. Diazoxide decreases insulin release and enhances glycogenolysis, and eliminates symptoms in approximately 60% of patients [49]. Everolimus, an mTOR (mammalian target of rapamycin) inhibitor with high

affinity to the intracellular receptor FKBP12, is often used in cases of metastatic PNETs [50]. Inhibition of mTOR by the Everolimus-FKBP12 complex prevents downstream signalling required for cell cycle progression [51]. Treatment strategies have become more aggressive in recent years, enabled by advances in surgical intervention techniques, and based on the assumption that patients benefit from reducing the tumour burden. Aggressive approaches involve removing as much (primary or metastatic) tumour as possible. The overall 10 year survival rate for insulinoma patients, following successful surgical removal of the tumour, is 88%. However, this drops to less than 30% for malignant cases.

1.2.2 Mutations associated with insulinoma

The majority of PNETs are sporadic, with no known cause, but a small percentage arise as the result of inherited syndromes including Multiple Endocrine Neoplasia type 1 (MEN1), Von Hippel-Lindau syndrome and Neurofibromatosis type 1 (NF1). Early approaches to finding the molecular alterations responsible for tumour phenotypes focused on mutations in protein-coding genes, and the catalogue of cancer-associated mutations in these regions is largely complete for major cancer types. But due to their rarity insulinomas were not prioritised by large consortia such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). Several smaller studies have investigated exonic mutations in insulinoma and identified germline and somatic mutations associated with the insulinoma phenotype.

Mutations in *MEN1*, which codes for Menin, a tumour suppressor with roles in cell growth and apoptosis, as well as control of transcriptional machinery and gene expression, account for 4 - 6% of insulinomas [52] [53]. Wang et al also identified several exonic mutations in insulinoma including recurrent mutational signals in epigenetic modifiers functionally related to MEN1 [42], including *KDM6A*, *EZH2*, *PCGF5*, *KMT2C* and *CREBBP*. The authors of this study suggested that multiple specific "epigenetic roads" may lead to insulinoma. Menin is a ubiquitously expressed scaffold protein with no intrinsic enzymatic activity. But

it does contain a binding site through which it interacts with various chromatin modifiers and transcription factors, affecting the expression of genes involved in cell proliferation [54]. Recurrent mutations affecting another ubiquitously expressed protein, YY1 (YinYang 1), including T372R, are reported to be present in up to 30% of sporadic insulinomas [55] [56]. YY1 is a transcription factor with roles in glucose metabolism and epigenetic regulation. The T372R mutation affects the third zinc-finger domain, part of the repression domain of YY1, altering its DNA-binding specificity [57].

So there is evidence that aberrant transcriptional regulation may be a key driver of insulinoma development, but no study has yet investigated the role of cis-regulatory elements (CREs) in insulinoma development. Elucidating insulinoma-specific gene-regulatory networks could provide important insights into the pathways driving cell proliferation and dys-regulated insulin secretion in insulinoma.

1.3 Gene regulation

The regulation of gene expression is a complex process involving many different factors, including genomic regulatory elements, epigenetic modifications and transcription factors. Control and regulation of these factors enables precise control of gene expression and, in turn, specific processes to occur within each cell type, thus maintaining cell identity. Gene-regulatory networks (GRNs) are composed of two main components: nodes and edges. The network nodes are the factors involved (genes and their regulators), while edges are the physical/regulatory relationships between the nodes. Disruption of GRNs has been associated with a wide range of human diseases including cancer. Much remains to be discovered about gene regulatory mechanisms, but the knowledge that enables us to investigate it is the result of decades of scientific discoveries.

1.3.1 Deoxyribonucleic acid (DNA) and chromatin.

DNA is composed of nucleotides: a deoxyribose sugar, a phosphate group and one of four nitrogenous bases (adenine, thymine, cytosine and guanine or A, T, C and G). The famous double helix structure of DNA showed the sugar and phosphate groups form the backbone of the molecule while the bases are paired up (A with T and C with G) via hydrogen bonds to connect the two strands. A human nucleus is $\approx 10\mu\text{m}$ in diameter but contains more than 2 meters of DNA. Fitting all that DNA into such a small space is achieved via a 'beads-on-a-string' structure incorporating sections of about 150bp wrapped around an octamer of proteins called histones to form a nucleosome [58]. Each nucleosome features a central tetramer of two H3 and two H4 histones flanked by two H2A/H2B heterodimers [59] fig(9). Another histone protein (H1) binds to the linker DNA between nucleosomes which allows further folding of this DNA-protein complex to form what is known as chromatin fiber.

So nucleosomes are essentially the core unit of chromatin, but chromatin is not a static structure, and features large variation in accessibility, which persists across the human genome. Chromatin accessibility refers to the degree to which nuclear macromolecules are able to physically interact with chromatinised DNA and is determined by nucleosome positioning and occupancy [61]. The positioning of nucleosomes is defined with respect to genomic DNA sequence incorporated by each nucleosome, and occupancy refers to the fraction of cells in a population in which a given region of DNA is occupied by a histone octamer [62]. Nucleosome-free regions (NFR) (or nucleosome-depleted regions (NDR)) of chromatin allow proteins (including transcription and replication machineries) to bind to the DNA. Thus while nucleosomes enable the packaging of DNA in the nucleus, this variation in chromatin accessibility also provides a layer of control over the initiation of transcription [63].

The dynamic and flexible nature of eukaryotic chromatin enables it to respond to environmental, developmental and metabolic cues. The organisation of DNA into chromatin generates a 'default' state of inaccessibility, preventing, for example, DNA binding proteins from finding their binding sites. The cell overcomes this state by employing enzymes that are

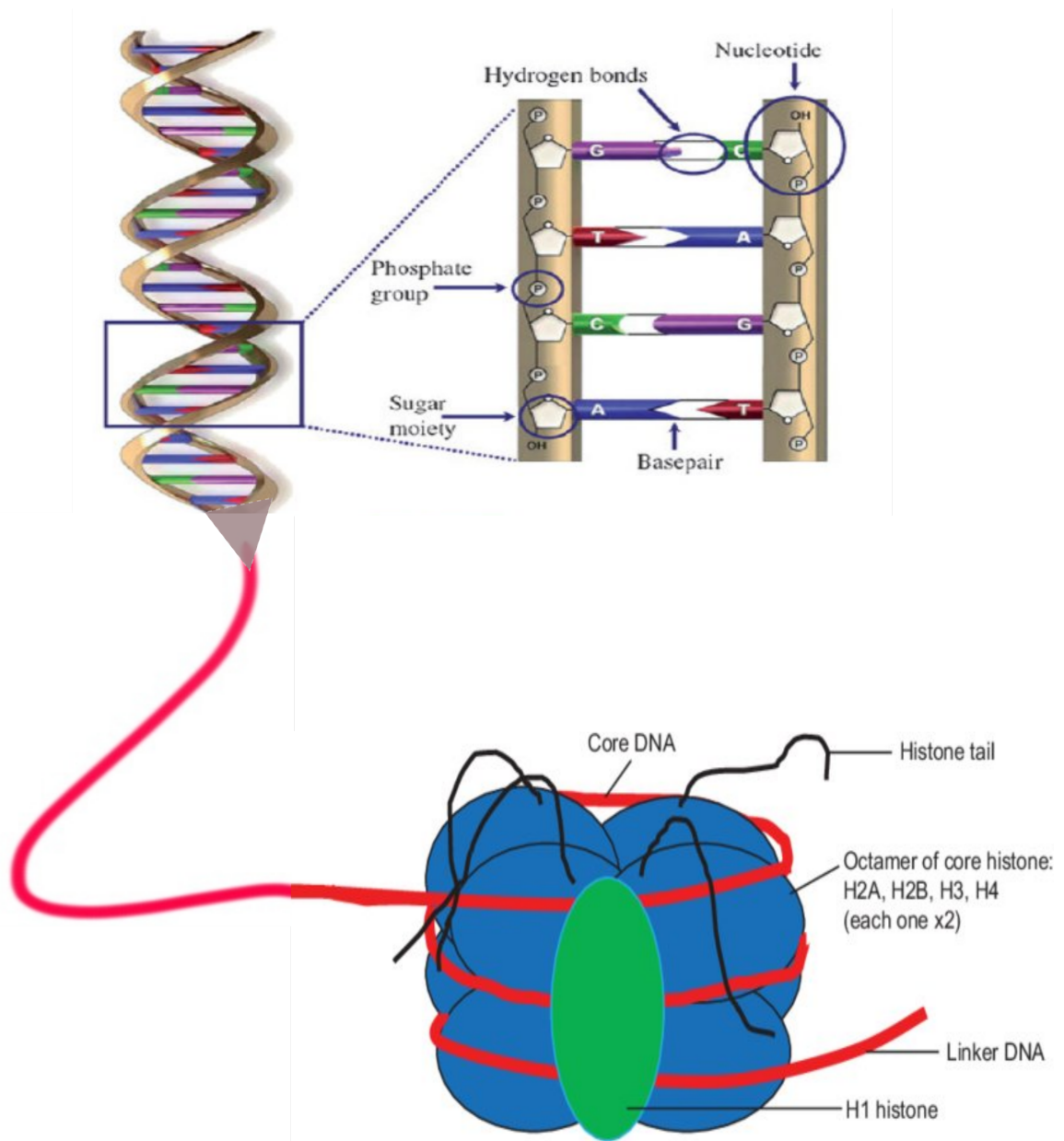


Figure 9: Schematic representation of the structure of DNA and a nucleosome [60].

able to remodel nucleosomes. Nucleosome remodelling enzymes may partially disassemble nucleosomes or incorporate histone variants and thereby change the interactions between DNA and histones. The precise mechanism depends on the combination of remodelling enzymes and co-factors [64].

1.3.2 Genes defined

In order to investigate gene expression and how it is regulated it is important to consider how a gene is defined. The earliest definition of a gene as 'the basic unit of heredity' is useful as a starting point, but through decades of research this picture has become increasingly complicated. The core principals that genotype determines phenotype and that DNA within genes contains the code to produce functional molecules [60] remain and form the basis of current descriptions. When information flows from DNA to protein, the DNA is first transcribed into a form of ribonucleic acid (RNA) called messenger RNA (mRNA) via an intermediate precursor (pre-mRNA). The simplest view is that one DNA sequence codes for one protein, but each gene may code for several different mRNA molecules (or transcripts) to produce multiple versions of a protein. Furthermore, many gene transcripts are non-coding RNA molecules which do not serve as the basis for protein production. Non-coding RNAs have a variety of functions within a cell, including regulation of gene expression.

A gene contains not just coding regions (exons) but also non-coding regions (introns). The sequence corresponding to the introns is removed from pre-mRNA by splicing to produce a mature mRNA. During this process alternative splicing events may occur, whereby one or more exons are also removed. This is a simplified version of events, and the mechanisms governing alternative splicing are not fully understood, but the fact remains that the DNA in each gene can be used to make several different functional products. The crucial component here is function, and regulatory mechanisms ensure that specific proteins and non-coding RNA molecules are generated from the same DNA blueprint. Cell type-specific utilisation of genetic information, including timing, dosage and transcript usage, forms the basis of cell function and identity.

1.3.3 The human genome and the origin of complexity.

In 2003, 50 years after the paper describing the structure of DNA, the human genome project was declared complete. The project, an international collaboration, determined the

order (sequence) of all the bases in the human genome and the locations of genes for major sections of each chromosome. The genome of an organism is the complete DNA sequence in one haploid cell. In humans this amounts to more than 3 billion base pairs and almost every cell in the human body contains the same genome sequence. Given the complexity of humans compared to other species, initial estimates of the number of protein-coding genes in the human genome ran to 6 figures. However, following the completion of the human genome project only around 20,500 genes, covering approximately 2% of the genome, were identified. To put this into context, the mouse genome has around 30,000 genes and *Caenorhabditis elegans*, a nematode worm about 1 mm in length with no circulatory or respiratory systems and only a primitive neural system, has around 18,000.

So it was clear that organismal complexity was not merely a function of the number of genes an organism has, and that much remained to be discovered regarding the molecular mechanisms involved [65]. The genome project gave us the genetic code and the ability to predict genes from sequence, but in order to understand how this information is used to generate complexity, and give each cell type its functional identity, we would need to understand the 'regulatory code'. Subsequently the ENCODE project was established with the aim of cataloguing all functional elements in the human genome (and select model organisms), including the 'non-coding' regions outside of genes [66]. This resource would enable developments in our understanding of the regulation of gene expression and the genetic basis of disease.

1.3.4 Transcription regulation by cis-regulatory elements.

Generation of the multitude of cell types required by a complex organism is facilitated by specific gene expression programs in each cell type. Gene expression is initiated along a promoter, a sequence of DNA located immediately upstream (up to 2kb) of the coding region of each gene. Promoters contain binding sites for RNA polymerase II (RNA polII), the enzyme responsible for the transcription of all genes (with the exception of those coding

for ribosomal or transfer RNA) in eukaryotes, and transcription factors (TFs). Once bound to DNA, TFs interact with RNA polII and other factors and in doing so initiate and regulate transcription of the DNA [67]. But this mechanism alone is only sufficient for low levels of transcription relative to those which may be required by the cell. Many cell-type-specific processes require significantly higher levels of transcription of specific genes. This is enabled by enhancers, another type of CRE located distal to the transcription start site (TSS) of genes.

Enhancers are short (200 - 1500 bp) DNA segments that can exert their effect over distances of hundreds of thousands of base pairs. Like promoters they contain binding sites for transcription factors, co-factors and RNA pol II. They are able to integrate multiple signals and regulatory determinants, enabling precise, cell-type-specific and state-specific control of spatiotemporal gene expression. Hundreds of thousands of putative enhancers have been mapped in various human cell lines by the ENCODE project, an order of magnitude more than the number of protein-coding genes, and it has been estimated that tens of thousands of enhancers are active in any given cell type [66]. Enhancers may be located upstream, downstream or in introns of target genes or unrelated genes, and connect to promoter elements via long-range physical interactions [68], often bypassing more proximally located genes [69] (fig.10). In fact, up to 50% of enhancers skip over the most proximal gene and regulate more distal gene(s) [70].

Several mechanisms for the establishment of enhancer-promoter interactions (EPIs) have been proposed, including chromatin looping (involving structural protein complexes cohesin and CTCF, and mediator) and tracking of molecular motors along chromatin [71]. Regardless of the precise mechanism, it is clear that contact between enhancers and promoters, and associated factors, is critical for precise control of gene transcription and ultimately cell-type specific gene expression. Some EPIs are prevented by insulators, DNA sequence elements that are able to protect genes from inappropriate signals from the surrounding environment, either by blocking the action of an enhancer or by affecting local chromatin organisation [72]. Aside from genomic location, the line between promoter and enhancer in terms of

gene-regulatory properties is somewhat blurred. In fact it has been observed that enhancers and promoters share remarkably similar chromatin and sequence architecture [73]. There are many examples of promoters with enhancer activity [74], and transcription can be initiated directly from enhancers, generating enhancer RNAs [75].

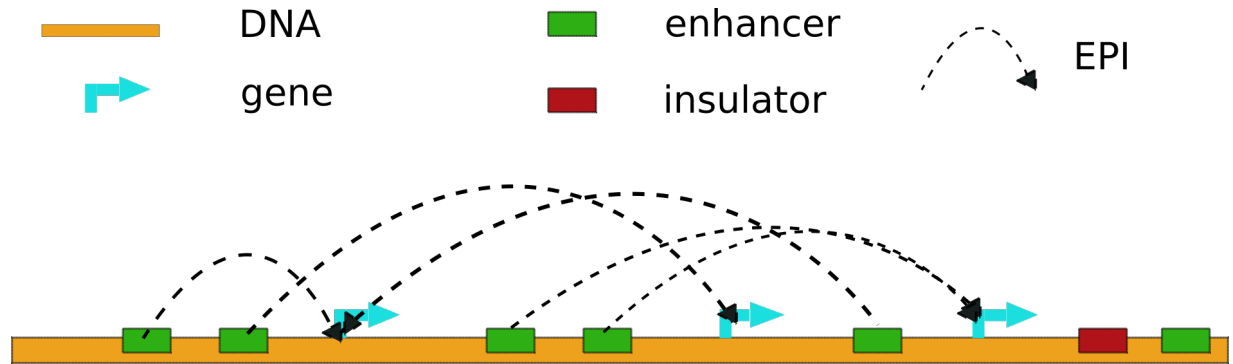


Figure 10: Enhancer-promoter interactions. Illustration of gene-regulatory interactions between distal cis-regulatory elements and genes. Assumes interaction with promoter regions immediately upstream of each gene. Insulator regions may prevent some EPIs.

Clusters of enhancers called 'super-enhancers' have also been functionally linked to cell identity [76]. Super-enhancers are described as groups of putative enhancers in close genomic proximity (within 12.5kb of each other), densely occupied by Mediator (a multi-subunit complex that mediates gene expression) and bound by master TFs Oct4, Sox2 and Nanog [77]. The median size of super-enhancers (as described in mouse embryonic stem cells (mESC)) is an order of magnitude larger than that of regular enhancers [76]. Several factors are enriched in both regular and super-enhancers, including RNA polII, enhancer RNA (eRNA), the histone acetyltransferases p300 and CBP, cohesin, histone modifications H3K27ac and H3K4me1, and increased chromatin accessibility. In mESCs super-enhancers are associated (by proximity) with genes encoding factors involved in pluripotency. Subsequent studies in various cell types, including pro-B cells and helper T cells, identified super-enhancers on the basis of enrichment for H3K27ac and revealed that they are enriched for sequence motifs corresponding to cell type-specific master transcription factors relative to normal enhancers [78]. Recent studies have also revealed super-enhancers that control tissue-specific gene expression in human pancreatic islets [29, 79, 80]. Super-enhancers are also enriched at genes

with known oncogenic function such as the MYC locus [81].

1.3.5 Epigenetic modification of chromatin at CREs.

The core histone proteins in nucleosomes acquire post-translational chemical modifications at various residues of their N-terminal tails. Active CREs are located in regions of open chromatin, featuring nucleosome depletion and accessibility to transcription factors and transcriptional co-activators. These regions are enriched for histone variants H3.3 and H2A.Z [82]. Incorporation of these histone variants creates domains of nucleosome hypermobility, facilitating the binding of TFs and TF-dependent recruitment of chromatin remodelling complexes [83]. But the most prominent marker of CRE activity is the epigenetic modification of adjacent nucleosomes, including the (post-translational) addition of methyl and acetyl groups to histone tails. Histones are modified by chromatin modifying enzymes which may be activating (e.g. p300) or repressive (e.g. PRC2 (polycomb repressive complex 2)). Modification of lysine residues of histones 3 and 4 make up the enhancer signature, as identified by the ENCODE project. Enhancers can be classified as active, poised or repressed depending on the combination of histone tail modifications present (fig. 11).

Tri-methylation of lysine 27 on histone 3 (H3K27me3) is a repressive histone mark associated with PRC2 and linked to chromatin compaction and gene silencing. Acetylation of the same lysine residue (H3K27ac) is a key feature of nucleosomes flanking both active enhancers and promoters. There is also an intermediate stage in which enhancers are poised for activity, and are enriched for both activating and repressive histone marks. Nucleosomes at poised enhancers still feature the repressive H3K27me3 mark and lack H3K27ac, but they gain monomethylation of histone 3 lysine 4 (H3K4me1). H3K4me3 is a marker of active promoters (in the developmental context) and when this mark is found in combination with H3K27me3, associated genomic regions were found to be in a poised state. However, markers typically enriched at enhancers such as H3K4me1 have also been observed at promoters [85], so here again the line between promoter and enhancer is not so distinct, and the distribution of

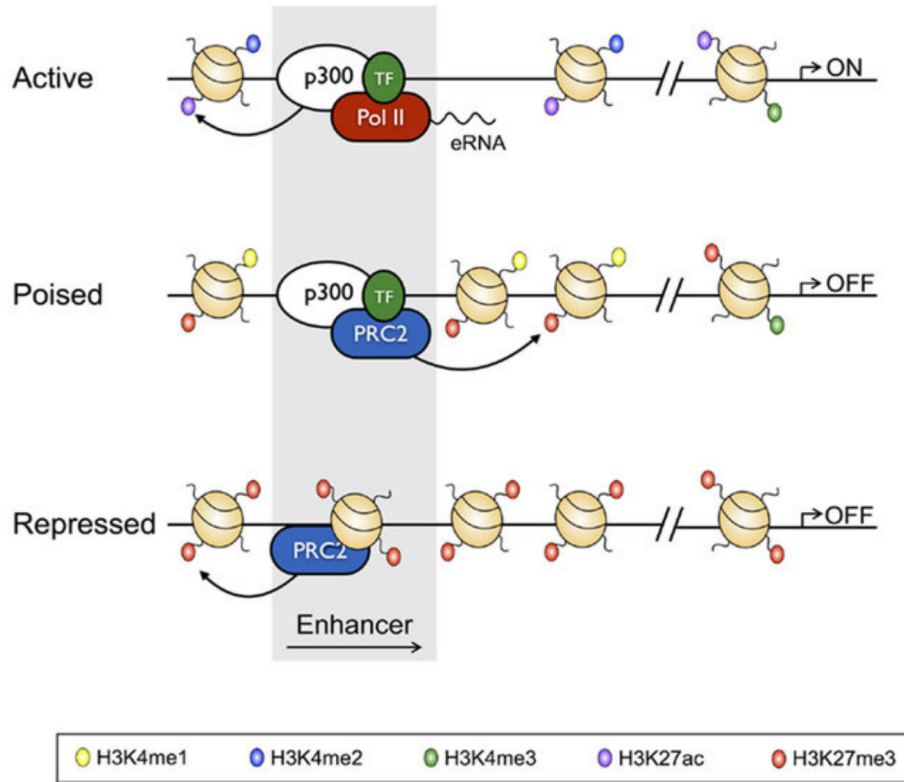


Figure 11: Schematic representation of histone modifications associated with repressed, poised and active enhancer states [84].

histone marks may be context dependent. What is clear though is the distinction between poised and active CREs, with the loss of repressive marks and gain of H3K27ac marking the transition [73].

Deposition of H3K27ac, specifically at enhancer regions, is mediated by the SWI/SNF multiprotein complex, one of four major families of chromatin remodelling complexes containing an ATPase subunit as the main catalytic subunit. ChIP-seq for the core SWI/SNF subunits Smarcc1 and Smarca4 in a murine cell line demonstrated that over 95% of distal CREs showed enrichment. Deletion of the SWI/SNF DNA-binding subunit *Smarca1* resulted in significant reduction in SWI/SNF at enhancers (relative to promoters) and regions showing the greatest loss of SWI/SNF binding also showed the strongest loss of H3K27ac. Gain-of-function of *Smarca1* in a human cell line resulted in a global increase in H3K27ac levels, as well as a 'subtle increase' in H3K4me1. In addition, levels of p300 (an H3K27 acetyltrans-

ferase) and enhancer-associated factors BRD4 and Mediator were also increased. These findings demonstrate that SWI/SNF is essential for enhancer activation [86].

1.3.6 Gene regulatory alterations in cancer.

It has long been known that changes in gene expression levels drive tumourigenesis. Studies in pancreatic cancer cell lines showed differential expression of genes involved in a variety of cellular functions including control of transcription, cell cycle regulation, cell adhesion and cell signalling [87]. Differences in gene expression levels have also been shown to account for differences in cancer subtypes, such as those observed in breast cancer [88].

More recent studies have shown that in cancer, cell-type-specific enhancers can become dys-regulated or 'hijacked', resulting in the activation of genes that promote tumourigenesis [89]. Activation of enhancers associated with oncogenes or multipotentiality factors may drive tumour formation [90]. Several studies have described aberrant enhancer activity, resulting from both mutations and epigenetic changes in cancer cells [91, 92, 93]. Specific 'enhancer signatures' were also recently described in a cohort of non-functional PNETs (those that do not produce excess hormone), including variation in the level of activity at specific genomic regions that indicated the likelihood of relapse following surgery. In addition to differential enhancer activation, the 'unexpected yet consistent' observation of structural variants (SVs) resulting in juxtaposition of specific oncogenes to putative cis-regulatory elements (including enhancers) was observed in a cohort of medulloblastoma samples. Subsequent analysis of H3K27ac in the regions surrounding the SVs predicted the presence of multiple enhancers, with enrichment consistent with the presence of super-enhancers [89].

There is also significant diversity in terms of tumour initiating GRNs. A pan-cancer screen of accessible regions of the genome in 23 cancer types identified > 500,000 novel distal regulatory elements. Cluster analysis showed that almost half of the putative CREs identified were present in a single cluster or small group of clusters. These cluster-specific regions were enriched for binding sites of transcription factors associated with cancer and tissue identity

[94]. With such heterogeneity, and the challenges associated with assigning function to distal regulatory elements, the task of establishing components of GRNs driving aberrant gene expression in human disease phenotypes is not trivial.

1.4 Regulatory genomics in the era of high-throughput sequencing

The first draft of the human genome took 10 years to complete, at a cost of nearly \$3 billion, using DNA sequencing technology developed by Fred Sanger. Since then the technology for molecular biology experiments has developed rapidly, akin to the development of microchips, to the point where an entire human genome can be sequenced in a matter of days and at a fraction of the cost. This enables us to perform genome-wide assays across large sample cohorts, increasing the statistical power of the conclusions.

Until recently, performing assays using samples with relatively few cells (< 1 million) was technically very challenging. This may be due to loss of input material at various stages of a protocol, including DNA preparation and enzymatic reactions. But new techniques, including developments of existing methodology, have made experiments with limited cell numbers more routine. A good example of this is chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq), which allows the mapping of histone modifications and chromatin-associated proteins on a genome-wide scale. ChIP-seq methodology has developed rapidly from its initial use, becoming quicker and more high-throughput, as well as enabling the use of more challenging samples. A major improvement to the protocol came with the development of ChIPmentation [95] in which sequencing-compatible adaptors are attached to bead-bound chromatin using a Tn5 transposase. This reduces the number of steps in the protocol by eliminating the need for DNA purification prior to library amplification, and in doing so reduces time and sample losses.

With this increase in the generation of large data sets comes additional challenges in processing and analysing large volumes of data, requiring increasingly complex computational methods. Raw sequence must be subject to quality control checks to highlight any prob-

lems that might affect downstream analysis. Once aligned to a genome and processed to a binary output, the data may be utilised in a myriad of ways to try to tease out meaningful results. Care is required when comparing data sets generated in different labs, and with variations in protocol. There are now a multitude of algorithms for performing tasks such as normalisation and differential analysis, and within each alternate statistical methods may be used. Furthermore, designing computational protocols for integrating data from various assays, such as epigenomic screens, transcript abundance and chromatin accessibility adds additional challenges.

1.5 Motivation

Despite developments in experimental and computational biology, deciphering the gene-regulatory networks involved in tumour development remains a major challenge. In addition both cancer and diabetes remain widespread health problems, so increasing our knowledge of the molecular mechanisms involved could be of great benefit to the scientific and medical community. This study has the potential to uncover novel gene-regulatory mechanisms involved in β -cell proliferation and insulin secretion, which could be useful in the diagnosis and treatment of both insulinoma and other pathologies.

This work may be viewed in two parts; the first largely descriptive part focused on identifying differentially expressed genes and differentially active CREs as putative markers of insulinoma development. The second, integrating multiple data sets to search for evidence of links between groups of genes and CREs, and to build putative gene-regulatory networks that could characterise functional β -cell tumours.

2 Hypothesis and objectives

2.1 Hypothesis

Changes to the gene regulatory architecture including the epigenetic landscape, chromatin structure and polymorphisms in cis-regulatory elements may result in aberrant proliferation and the loss of cell identity, and contribute to the development of functional pancreatic neuroendocrine tumours.

2.2 Objectives

To create the first gene regulatory maps of a functional pancreatic neuroendocrine tumour (insulinoma) by profiling active gene regulatory elements and gene expression. Using these maps we aim to identify the key gene regulatory pathways underlying the expansion of the β -cell mass and the maintenance of β -cell identity.

Profile the landscape of active CREs in pancreatic islets and insulinoma:

- Complete ChIP-seq assays to profile the deposition of the active histone mark H3K27ac genome wide for all insulinoma samples.
- Complete RNA-seq to detect changes in gene expression levels for all insulinoma samples.

Identify insulinoma-specific gene regulatory elements and pathways by applying bioinformatic techniques to epigenomic and transcriptomic data.

- Perform differential analysis to identify genomic regions significantly enriched in insulinoma compared to pancreatic islets.
- Investigate the regulatory potential of the insulinoma-specific genomic regions by integrating gene expression data, identifying super-enhancers and comparing putative insulinoma enhancers to other pancreatic tumour enhancers.
- Build insulinoma-specific gene regulatory networks.

3 Materials and methods

3.1 Ethics.

Human islets were isolated from brain-dead organ donors in accordance with national laws and institutional ethical requirements at the Istituto Scientifico San Raffaele, Milan, Italy. Insulinoma samples were obtained upon removal of the tumour without interfering with the clinical patient management. All experiments were performed according to protocols approved by the institutional research committees of the Istituto Scientifico Ospedale San Raffaele and the Institute for Health Science Research Germans Trias i Pujol and the study was conducted in accordance with the Declaration of Helsinki. All patients or their parents gave informed consent. All samples and data were handled protecting patients' privacy.

3.2 Wet lab

3.2.1 Samples

We elected to use human islets as controls rather than a β -cell line such as EndoC or FACS purified β -cells. EndoC cells are essentially β -cells that have been induced to proliferate in a cell culture. As the aim of this work is to discover in vivo drivers of β -cell proliferation, using EndoC cells might be inappropriate as they are already biased towards proliferating. FACS purified β -cells have been used in previous studies to investigate insulinoma, but the purification process induces significant cell stress, which may in turn lead to gene-regulatory changes that diverge from those of a normal beta cell.

Sample	Gender	Age	BMI	% Purity	Cause of Death	Experiment
HI6*	Female	23	22.5	46	Cardiac arrest	RNA-seq
HI7*	Male	31	27.8	66	Cerebral bleeding	RNA-seq
HI8*	Male	77	24.5	59	Cerebral bleeding	RNA-seq
HI9*	Female	64	29.4	47	Cerebral bleeding	RNA-seq
HI10*	Female	58	21.3	67	Cerebral bleeding	RNA-seq
HI25**	Male	59	24.2	93.5	ischemic haemorrhage	RNA-seq
HI32**	Male	38	22.9	93.8	trauma	RNA-seq
HI.19	Male	34	23.1	80	Cerebral bleeding	ChIP-seq
HI.22	Male	52	25.1	85	Trauma	ChIP-seq
HI.32	Male	62	23.1	90	Cerebral bleeding	ChIP-seq
HI.37	Female	53	21.8	85	Cerebral bleeding	ChIP-seq
HI.40	Female	62	29.3	90	Cerebral bleeding	ChIP-seq
HI.D2***	Female	47	28	85	Stroke	ChIP-seq

Table 1: Patient data for human islet samples. Previously reported data: * [96] ** [97] *** [80].

So, although human islets are a heterogeneous group of cells, at least 70% of those cells are normal β -cells, lacking significant proliferative activity, and thus they represent the closest approximation to the cells from which insulinomas would develop. Healthy human islet samples were obtained from cadaveric donors (with no premortem diagnosis of diabetes) from San Raffaele hospital in Milan, Italy (table 1). Additional ChIP-seq (HI.D2) and RNA-seq data from healthy human islet samples was obtained from previously published reports [96, 97, 80].

Insulinoma samples were obtained from hospitals in Spain, Italy and Argentina. A 20–30 cm^3 segment of pancreas was resected from each insulinoma patient during surgery, from which a tumour mass of 4 – 6 cm^3 was isolated. Part of each resected tumour was frozen and part was fixed with formaldehyde to preserve the DNA-protein contacts. Following histological

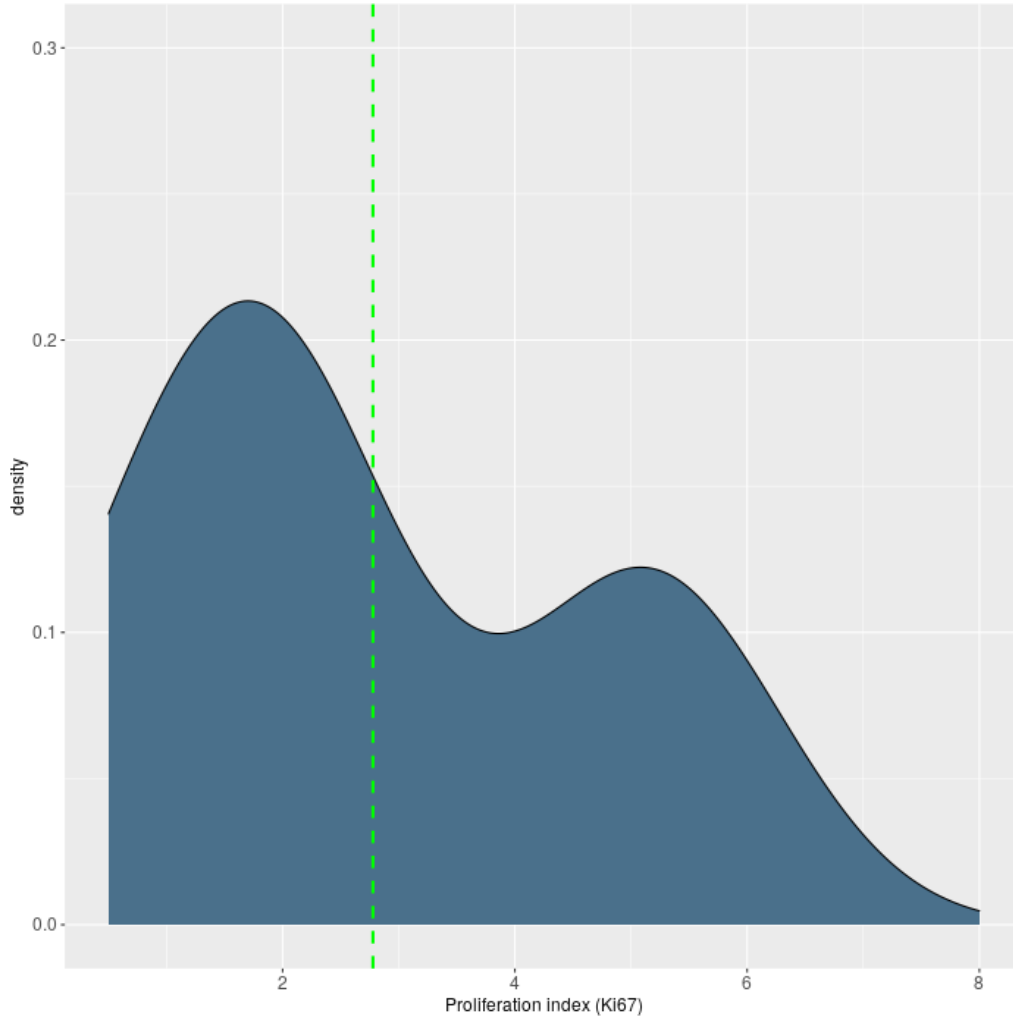


Figure 12: Density plot of proliferation index (Ki67) values for insulinoma samples. The dashed green line indicates the mean average Ki67 value for all insulinoma samples in our cohort.

evaluation, 75% of the samples were characterised as grade 1 (NET G1 or WHO 1) and the remainder as grade 2 (NET G2 or WHO 2). These grades relate to the speed at which the tumour is growing, so in this case the majority are growing slowly. This can further be represented in terms of the proliferation index, or mitotic index, which is a measure of the percentage of cells that stain positive for Ki67, a marker of cell proliferation. Slow growing tumours have a Ki67 value of less than 2%, whilst aggressive tumours may have values up to 50%. The majority of insulinoma samples in our cohort had Ki67 value of 1 – 2% and the

remainder had values of 5–6%, with a mean average Ki67 value of 3.8% (fig.12). Insulinoma phenotype was determined via positive immunoreactivity tests, with tumour cells positive for insulin and negative for glucagon, confirming the β -cell origin.

Data for non-functional PNETs, including ChIP-seq and RNA-seq was obtained from db-GAP and uploaded by the authors of 'Enhancer signatures stratify and predict outcomes of non-functional pancreatic neuroendocrine tumours' [98]. Samples are characterised as 'PDX+' (β -cell-like), 'ARX+' (α -cell-like) and 'DP' (double positive) based on a combination of immunohistochemistry and enhancer signatures. ChIP-seq data includes genome-wide H3K27ac enrichment profiles for 4 PDX+, 9 ARX+ and 7 DP samples. RNA-seq data includes transcriptome profiles for 3 PDX+, 5 ARX+ and 2 DP samples.

3.2.2 ChIP-seq.

ChIP-seq is a well established and popular technique for identifying the binding sites of DNA-binding proteins and the locations of histone modifications, and has been widely utilised by the ENCODE consortium [66]. It is extremely versatile, enabling profiling of multiple factors and modifications in multiple cell types, but there are also a multitude of options for the various buffers and conditions that may be used at different stages of the technique. A level of optimisation is thus required to adapt the technique to the specific cells and antibodies under investigation. Furthermore, challenges still remained to obtain good quality data from ChIP-seq with low cell numbers. We decided to utilise a version of the ChIPmentation technique [95] as Schmidl et al had demonstrated results with as few as 500,000 cells to be of similar quality to those from assays using several million cells. The ChIPmentation protocol was a significant improvement on previous ChIP experiments using low cell numbers as it reduced the number of steps involved and the overall cost per assay, enabling faster and cheaper results. The protocol utilises a Tn5 transposase which enables DNA fragmentation and adaptor ligation in the same reaction, applied directly to bead bound chromatin (fig.13).

ChIP-seq assays were performed with an anti-histone H3 acetylK27 antibody (abcam ab4729)

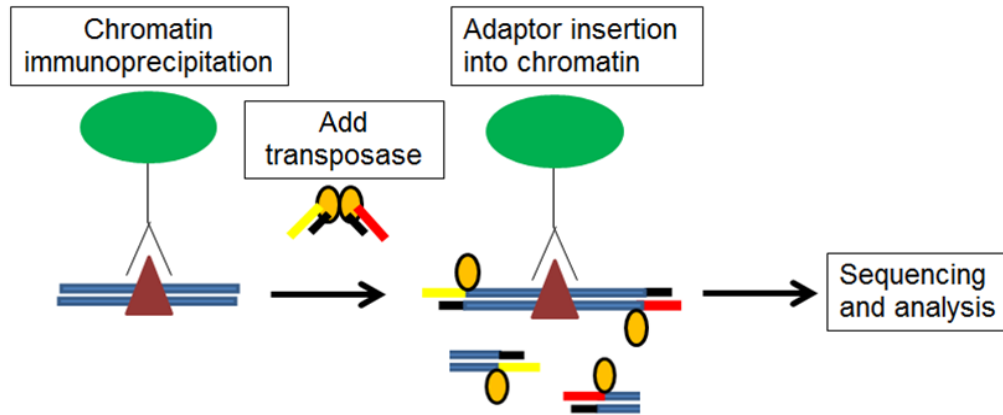


Figure 13: Schematic of the ChIPmentation protocol, adapted from Schmidl et al 2015 [95].

using 12 insulinoma and 4 human islet samples. These samples typically have low cell numbers and the total number of cells in each tumour mass is unknown. The ChIP-seq protocol was based on the ChIPmentation technique [95] and can be found in full in the supplementary section. Briefly, fixed and frozen samples were sonicated to achieve an average fragment size of around 200bp. $35\mu\text{g}$ chromatin was immunoprecipitated in a 0.4% SDS IP buffer with $1.5\mu\text{l}$ anti-H3K27ac and $50\mu\text{l}$ 10% BSA. Following incubation, IPs were hybridised to protein A+G beads and washed with low-salt, high-salt and LiCl wash buffers. IPs were then incubated with $1\mu\text{l}$ Tagment DNA enzyme for 10 minutes, followed by washes with RIPA and Tris-EDTA buffers. ChIP libraries were eluted in a 1% SDS, 0.1M NaHCO_3 buffer. qPCR with SYBRgreen polymerase mix was used to determine the number of cycles to amplify each library. Libraries were amplified using Illumina primers and Nextera Taq mix and enrichment of ChIP libraries was determined by qPCR using custom primers (table 2). Sequencing of ChIP libraries was performed using a single-end protocol with minimum 40 million, 50bp reads. An additional 2 human islet ChIP-seq data sets from previous studies were utilised for analysis.

Name	Oligonucleotide sequence
NEUROD_ChIP-F	CTTGGCTTCTTCTCCTTGGC
NEUROD_ChIP-R	GGAAGTGGGAGAGGACGATC
MAFB_ChIP-F	GAATGAGCGCCGGGACAA
MAFB_ChIP-R	CGCCCTCCCCAACATACAAA
ISL1_ChIP-F	GTGTGCCCTCCAAAGCTCTA
ISL1_ChIP-R	GTTGTCACTCGTTCTCTTTTGCA
NANOG_prom-F	TCCCATTCCCTGTTGAACCAT
NANOG_prom-R	TCCCGTCTACCAGTCTCACC
AMILASE-F	GGCTTCTGGACCACTTGTTT
AMILASE-R	ACAGGTATAAATGCGAACCCC
PDX1_R8B-F	CATGAAAGCGGGTTAATCGT
PDX1_R8B-R	GGCCCCTCACTCTTCTTACC

Table 2: Primer sequences for positive and negative control qPCRs.

3.2.3 ATAC-seq.

The assay for transposase accessible chromatin followed by high-throughput sequencing (ATAC-seq) was used to profile regions of accessible chromatin in insulinoma and human islet samples. Preservation of the native chromatin structure and the original nucleosome distribution patterns are key to successful ATAC-seq experiments and therefore use of freshly isolated tissue rather than fixed or frozen material represents the most efficient approach. For this reason ATAC-seq assays were performed immediately following surgical resection at the San Raffaele Scientific Institute in Milan, Italy.

A fraction of the fresh insulinoma (corresponding to approximately 50,000 cells) was used for each sample. The protocol used is based on that originally developed by Buenrostro et al [99]. Cells were lysed in 300 μ l cold lysis buffer and centrifuged for 15 min at 500*xg*. Nuclei pellets were then resuspended in 25 μ l transposase reaction mix containing 2 μ l of Tn5 transposase, 12.5 μ l TD buffer (Nextera DNA Library Prep Kit, 15028212, Illumina) and 10.5 μ l DEPC

treated water per reaction, and incubated at 37°C for 1 hour. Following incubation, $5\mu\text{l}$ clean up buffer ((900 mM NaCl, 300 mM EDTA)), $2\mu\text{l}$ 5% SDS and $2\mu\text{l}$ Proteinase K was added and the mix incubated for 30 min at 40°C . Tagmented DNA was isolated using a 2x SPRI bead cleanup (Agencourt AMPure XP5 ml, A63880, Beckman Coulter). Isolated DNA samples were stored at -20°C before library amplification. Tagmented DNA samples were shipped to the Endocrine Regulatory Genomics lab where we amplified the libraries. An initial amplification of 5 PCR cycles was performed (see ChIP protocol for library amplification conditions), after which a qPCR assay was used to determine how many extra cycles were required. The number of extra cycles used was $x + 1$ where x was determined by rounding the Ct value of the qPCR reaction up to the nearest whole number. Enrichment of ATAC libraries was determined by qPCR using custom primers [99].

3.2.4 RNA-seq.

RNA was extracted from frozen tumour samples using the Qiagen AllPrep DNA/RNA extraction kit. RNA concentration was measured using a nanodrop spectrophotometer. RNA quality was established using gel electrophoresis (to check for degradation) and RNA integrity number (RiN) tests, which compares several features of RNA integrity, including 28S:18S ratio, to produce an overall score [100]. RNA libraries were prepared using a ribosomal RNA depletion method and libraries were sequenced using a 150bp paired-end protocol. After sequencing the raw reads were filtered to remove adaptor sequences, contamination and low quality reads

3.2.5 YY1 T372R genotyping.

20 insulinoma samples were genotyped for the heterozygous C/T mutation T372R as described in [56] using the following primers:

F: 5'-CACCCAGGGCAGGAATG-3'

R: 5'-CCTGTCTCCGGTATGGA-3'.

3.3 Computational analysis

3.3.1 NGS analysis pipelines.

An initial cleanup of raw sequencing data was performed at the sequencing facility. This included filtering to remove adaptor sequences, contamination and low quality reads. A pipeline was created to perform all steps from quality control of raw data to peak calling. The pipeline consisted of an R script with functions calling software using the Bash shell via system commands. Each function included if statements allowing the customisation of the pipeline for different NGS data sets including ATAC-seq and various ChIP-seq assays. QC for datasets was performed using fastqc to identify any problems with the data before further processing. Raw reads were aligned to hg38 using Bowtie2 [101] using standard settings. Aligned reads were filtered for duplicates and repetitive sequences before conversion of the sequence alignment map (SAM) to a binary alignment map (BAM) using Samtools [102].

ChIP-seq and ATAC-seq peaks were called using MACS2 [103]. For ChIP-seq of H3K27ac the 'broad' flag was used, with q-value cutoff 0.05 and broad cutoff q-value = 0.1. Although H3K27ac covers relatively narrow domains compared to other histone marks, in the context of peak calling with MACS2 it is a broad peak, as the standard ('narrow') setting is designed to detect peaks reflective of TF binding sites. The broad setting links nearby highly enriched regions (max length 4x fragment length).

3.3.2 Count matrices

ChIP-seq data for 12 insulinoma, 6 human islet and 20 PNET samples was analysed in order to identify differentially active (enriched for H3K27ac) regulatory elements. Initially a consensus peak set (cps) was created using the cat command in Bash and the merge function from bedtools [104]. This cps was imported into R [105] and then modified to remove X, Y and non-canonical chromosomes. Aligned reads were then assigned to each consensus region for each sample using featureCounts [106] from the Rsubread package to create a count

matrix.

3.3.3 Transcript quantification

Transcript quantification for 11 insulinoma and 7 human islet samples was performed using Salmon [107], which includes a 'pseudo alignment' algorithm for mapping reads to the genome. Salmon combines quasi-mapping with a two-phase inference procedure to achieve accurate and 'wicked-fast' expression estimates. It takes a transcriptome index and unaligned reads in fastq format and performs quantification directly, without generating intermediate alignment files. Its algorithm also takes into account sequence fragment biases such as GC bias and fragment length distribution, learning auxiliary models that describe the relevant distributions. Read count data (Salmon output) was imported into R using tximport [108] and merged with transcript information using Ensembl biomaRt.

3.3.4 Regulatory element conservation.

All H3K27ac enriched regions were converted to a standard size (3 or 5kb) by extending 1.5 or 2.5kb in each direction from the midpoint of the region. Each region was then divided into 50bp bins for which a conservation score, derived from hg38.phastCons100way.bw (phastCons scores for multiple alignments of 46 vertebrate genomes to the human genome) was assigned. The mean score for each bin across all regions was then calculated. Conservation scores were calculated for insulinoma CREs and compared to a set of random regions of similar size and genomic distribution (blacklist regions including repetitive sequences excluded).

3.3.5 Regulatory element coverage and sharing.

CREs were split into proximal and distal groups (proximal defined as within 2kb upstream and 200bp downstream of the nearest transcription start site). A consensus peak set was then created, combining the coordinates of every peak (corresponding to a proximal or distal

regulatory element) in ChIP-seq data from the 12 insulinomas. Bedtools intersect was used to find every peak in each individual dataset that overlapped with the consensus set. For each combination of a set number of samples (from 1 to 12) the sum (Mb) of all H3K27ac peaks represented was calculated in R. For each CRE in the consensus set a sharing index score (equivalent to the number of individual samples from each tissue type where the CRE is enriched for H3K27ac) was also calculated.

3.3.6 Correlation clustering.

To investigate the inter- and intra-group variability between insulinoma and unaffected human islets for ChIP-seq and RNA-seq data we utilised correlation tests, specifically the Spearman correlation. Correlation coefficients describe the directionality and strength of the relationship between two variables. The more similar the profiles (e.g. gene expression) between two samples, the higher the correlation. This allows a global overview of the data and to identify any outliers. Spearman's correlation is a rank-based algorithm that can handle non-normal distribution of data and capture various relationships between the data [72]. Thus it is less restrictive than the Pearson correlation, which assumes a normal distribution of data and a linear relationship between variables. As our initial aim is to get a general overview of the data, using a less restrictive methods made sense.

Correlation coefficients were used as input for an unsupervised clustering algorithm to generate a graphical representation of the relationship between tumour and control samples for counts of transcript abundance and enrichment for H3K27ac. Unsupervised learning algorithms aim to infer the inner structure of data sets and then group or cluster them based on similarities between them, where similarity refers to the distance between data points. Hierarchical clustering is one of the most commonly used types of unsupervised learning, and builds a multi-level hierarchy of clusters based on patterns or groups in the data. Each observation is initially treated as a separate cluster, and the most similar clusters are merged iteratively. Unsupervised hierarchical clustering of correlation values was performed, and

results annotated, using the ComplexHeatmap package [109] in R.

3.3.7 Differential analysis

There are several different algorithms designed to perform differential analysis of count data, and there is little consensus about the most appropriate pipeline or protocol for this task [110]. Selection of appropriate methods depends on the type of data (experimental origin) being analysed, as well as prior knowledge or assumptions about its distribution. Many popular differential expression analysis methods use some kind of parametric test where certain assumptions are made about the distribution of the data. Such approaches can make the process much faster (compared to non-parametric tests) by providing a form of likelihood or functional distribution to the parameters. They also have greater statistical power, meaning they are more likely to find true significant results. A certain level of heterogeneity in both normal and aberrant cell types, as well as slight variations between experiments, is expected. EdgeR [111] calculates relative changes in gene expression rather than estimating absolute expression levels, meaning that any technical factor not related to the experimental setup, such as gene length, should have no effect on differential expression analysis.

The expression profile of each sample is the set of genewise counts and the expected size of each count is a combination of the library size and the relative transcript abundance for that specific gene and sample. The number of reads mapped to gene g in sample i may be denoted as Y_{gi} . Variation in the magnitude of Y_{gi} between RNA samples is mainly due to biological causes. Variation in Y_{gi} due to technical differences can be captured using a Poisson model [112], with $< 0.5\%$ of genes deviating from this model. Poisson distribution is a discrete probability distribution expressing the probability of x number of events occurring in time t , if the mean and variance are constant. EdgeR uses the negative binomial (NB) distribution, which has one more parameter than the Poisson distribution, that adjusts the variance independently from the mean using the following equation, where M_i is the total

number of reads, ϕg is the dispersion and Pgi is the relative abundance of gene g in sample i of experimental group j .

$$Y_{gi} \approx NB(MiPgi, \phi g)$$

For RNA-seq analysis we utilised quasi likelihood (QL) dispersion estimation, an extension of the NB model which accounts for uncertainty in estimates of the NB dispersion for each gene, allowing for overdispersion in the data and increasing error control. In this case, the NB dispersion describes the overall biological variability across all genes and the QL dispersion controls for any gene-specific variability above or below the overall level. Gene-wise dispersion estimates are maximised using an empirical Bayes approach, whereby QL estimates of dispersion are squeezed towards a mean-dependent trend, reducing uncertainty in the estimates and improving testing power.

This is a common analysis for gene expression, and several of the tools developed for this type of analysis were optimised for gene expression data, but the noise and variability in ChIP-seq often present additional challenges. In recent years many new methods have been developed or modified for use with ChIP-seq data, utilising different algorithms, varying in applicability and all have advantages and disadvantages. For consistency we used edgeR for differential analysis of ChIP-seq data, but to account for the uncertainty in variability we incorporated the RUVr method from the RUVseq [113] package. RUVr enables control of 'unwanted variation' by calculating residuals, from a first-pass GLM regression of raw read counts, which can be used in the normalization of raw count data. Importantly, RUVr assumes that the true biological effects are much larger than the unwanted variation.

For both RNA-seq and ChIP-seq differential expression (DE) was assessed for each transcript or CRE using an exact test analogous to Fisher's exact test, but adapted for overdispersed data [114]. For RNA-seq the main analyses are focused on differential analysis between insulinoma and unaffected human islets, but an additional DE analysis was performed between non-functional PNETs and human islets in order to investigate genes related to the dysregulation of insulin secretion that distinguishes the two tumour types. Two individual analyses

(insulinoma vs human islets & insulinoma vs non-functional PNETs) were also performed for ChIP-seq data sets, producing \log_2 fold change and adjusted p-values for all regions in the cps for each comparison. The two sets of results were then combined in a data frame and results grouped using FDR (< 0.05) and $\log_2\text{FC}$ ($\geq [1.5]$) cutoffs for significance.

3.3.8 Pre-ranked gene set enrichment analysis.

Pre-ranked GSEA produces an enrichment score for each gene set which is reflective of the frequency with which individual genes in the set occur at the top or bottom of the ranked data set. Transcript data was sorted according to $\log\text{FC}$ from the differential analysis from highest to lowest and the transcript with the highest $\log\text{FC}$ for each gene was selected for the analysis. Transcripts described as 'nonsense-mediated decay' by Ensembl were removed. Pre-ranked GSEA was performed using pathway gene sets from Reactome and two different algorithms.

Analysis was first performed using the multi-level method from the fast gene set enrichment analysis (fgsea) R package which enables accurate estimation of arbitrarily low p-values for individual gene sets based on an adaptive multi-level split Monte Carlo scheme [115]. To account for variation in experimental design, including the number of data sets in test and control groups, we also applied a generally applicable GSEA (GAGE) [116]. GAGE calculates an enrichment score (ES) that reflects the degree to which a gene set is over-represented at the extremes (top or bottom) of the ranked list of genes. The significance of the ES is calculated using an empirical phenotype-based permutation test. A null distribution is calculated for each ES and an empirical p-value for the observed ES is calculated relative to this null distribution. Only gene sets with adjusted p-values < 0.05 obtained using both algorithms were considered significant.

3.3.9 Chromatin accessibility within CREs

Peaks from ATAC-seq assays in insulinoma samples were combined with ATAC-seq peaks from unaffected human islets and a pan-cancer cohort from Corces et al [94] to produce a consensus peak set incorporating more than 400,000 regions of accessible chromatin. In addition to this the nfr algorithm from HOMER was used to identify putative nucleosome free regions from insulinoma ChIP-seq peaks. These two data sets were combined and overlapped with CREs of interest in order to assign region(s) of accessible chromatin to each H3K27ac-enriched CRE.

3.3.10 Super-enhancers

Super-enhancers were called for all H3K27ac data (insulinoma, islets and non-functional PNETs) using the ROSE algorithm [76, 81]. As with regular enhancers, comparisons of SEs from different tissues were performed by creating a consensus peak set, but in this case a minimum overlap of 0.2 was used to merge peaks. Bedtools was used to find overlaps between SEs from individual samples and the cps. Overlap data was loaded into R in order to identify tissue-specific SEs and the extent to which SEs are shared between insulinoma and non-functional PNETs.

3.3.11 Comparison to chromHMM and H3K27me3 data

CRE regions derived from ChIP-seq data were compared to regions from chromHMM analysis and H3K27me3 ChIP-seq data sets from unaffected human islets using the findOverlaps function from the GenomicRanges package in R. For H3K27me3 data overlap permutation tests were then used to test if the overlaps were significant.

3.3.12 Selection of CRE to TSS distance cutoff

As described, enhancers may act over large genomic distances. But in order to search for putative insulinoma-specific enhancer-promoter interactions it is necessary to define a cutoff (distance in bp between enhancer and promoter). Using a small cutoff risks eliminating a significant number of functional interactions from the analysis, while a large cutoff risks incorporating many false positives. A genome-wide map of regulatory interactions in the human genome (using cells lines from ENCODE and analysing several histone marks and transcription factors) observed a median distance of 120kb [117]. We further analysed interaction distance data from this study and found that over 99% of interactions occurred within 200kb. Another study using human induced pluripotent stem cells (iPSCs) and iPSC-derived cardiomyocytes (CMs) observed a median EPI distance of 170kb and 164kb respectively [118]. These studies suggest 200kb to be a suitable range to investigate functional EPIs, whilst limiting spurious results. Our approach was thus to extend the coordinates of each TSS by 200kb upstream and downstream to produce a window to explore potential functional relationships between differentially active enhancers and up-regulated genes in insulinoma.

3.3.13 Overlap-permutation tests.

Overlap permutation tests were performed using the `overlapPermTest` function from the `regionR` package [119]. Tests were performed for each group of CREs (identified from the plot of differentially enriched CREs (fig.29)) vs TSS coordinates (± 200 kb) corresponding to the leading edge of each up-regulated gene-set from the `gsea` analysis. For each comparison, random region sets consisting of regions of a similar size to the CREs were generated 1000 times and the number of overlaps with the TSS windows calculated. The profile of overlap scores created was then used to calculate the likelihood that the number of overlaps of the CRE group was significant. Tests were performed iteratively for every gene set (from `GSEA` analysis) against each CRE group. The p-value for each test was recorded and then corrected for multiple testing using the Bonferroni approach. Results with adjusted p-values < 0.05

were considered significant for overlap.

3.3.14 Networks

Network analysis was performed using Stringdb. Gene lists were produced by selecting genes from the leading edge of pathways identified using gene set enrichment analysis. Only genes above a \log_2FC cutoff and overlapping specific CREs (as described in results) were selected. A minimum confidence value of 0.4 (scale from 0 to 1) was used along with Markov clustering.

3.3.15 Non-functional PNETs.

H3K27ac ChIP-seq data for a collection of non-functional PNET samples was obtained from Cejas et al [98] following an application for permission via dbGAP.

4 **Genome-wide profiling of gene expression and CREs in pancreatic islets and insulinoma.**

4.1 Transcriptome profiling of insulinomas.

Cellular function is dictated by transcriptional profiles and, as described above, the development and functional identity of normal β -cells is the result of the action of several transcription factors which act to regulate the expression of various genes. Components of multi-protein complexes involved in activation and repression of gene expression are associated with insulinoma [42]. Furthermore the two most frequently mutated genes associated with insulinoma (*MEN1* and *YY1*) have critical roles in transcription regulation. Profiling the transcriptome of insulinomas is therefore an important step toward identifying the mechanisms driving the cell proliferation and insulin secretion observed in these tumours. Whole RNA extracted from 11 frozen insulinoma samples was of sufficient quality to build libraries for sequencing. An RNA integrity score of ≥ 7 is normally the minimum standard, but we proceeded with sequencing of NET17 (RiN = 6.9). NET16 was incorrectly stored in formalin following surgical resection, hence it's low RiN score, so an RNA-seq protocol designed for formalin-fixed samples was applied to this sample.

Sample ID	RiN	Concentration (ng/ μ l)
NET10	9.1	1190
NET11	8.1	383
NET14	9	1155
NET16	3.1	63
NET17	6.9	196
NET20	7.5	90
NET25	8.4	860
NET26	7.2	740
NET29	9.2	605
NET30	7.5	1715
NET38	8.9	267

Table 3: Quality control for Insulinoma RNA samples. RiN = RNA integrity number; concentration measured by nanodrop spectrophotometer.

Paired-end transcriptome sequencing was completed for all 11 insulinoma samples. More than 150 million reads (75 million in each direction) were generated for each sample (following cleanup of raw data) with an average Q20 score of 98%. Fastqc analysis of the raw data

Sample name	Clean reads	Q20(%)	GC(%)
NET10	150,310,928	97.68%	53.69%
NET11	181,913,670	97.41%	50.33%
NET14	198,368,952	97.32%	52.11%
NET16	194,295,974	98.38%	60.88%
NET17	216,028,968	98.04%	51.75%
NET20	214,629,428	98.31%	55.43%
NET25	205,432,668	97.41%	52.93%
NET26	213,410,396	97.45%	53.24%
NET29	196,272,612	98.20%	55.20%
NET30	217,511,018	97.51%	53.30%
NET38	160,346,214	97.93%	52.50%

Table 4: RNA-seq read and quality data for insulinoma samples. 'Clean reads' corresponds to the number of reads for each sample following data cleanup (described in methods. The Q20 score is the percentage of reads for which the probability of an incorrect base call is 1 in 100).

revealed that the majority of reads for each sample reached Q40 (probability of incorrect base call 1 in 10,000). This data enabled comparisons of insulinoma transcript profiles with existing data sets from unaffected human islets and thus the potential to identify genes and pathways that are differentially regulated in insulinoma.

4.2 Optimisation of chromatin immunoprecipitation

To enable investigation of molecular mechanisms driving gene expression changes in insulinoma I also profiled CRE activity in insulinomas and healthy human islets. Acetylation of lysine 27 on histone 3 (H3K27ac) is a prominent marker of activity for both proximal and distal CREs. Genomic regions enriched for this mark were identified using chromatin immunoprecipitation (with an anti-H3K27ac antibody) followed by high-throughput sequencing. For this assay, all stages from cell lysis and sonication to library preparation and quality control were completed in the Endocrine Regulatory Genomics lab. Optimisation of the protocol was performed using samples from a colorectal cancer cell line before proceeding with (more precious) insulinoma and human islet samples.

The chromatin immunoprecipitation stage of the experiments was very similar to other protocols. As with all such protocols the overall goal is to map the regions of DNA that bind to an antibody with the maximum signal-to-noise ratio and genome coverage [120]. There are several steps for which sub-optimal conditions may compromise the success of the assay. Firstly, to achieve maximum resolution it is important to fragment the chromatin to an average size of 150 - 200bp (approximate size of DNA wrapped around a nucleosome) whilst preserving the interactions between modified histones and DNA. The challenge here lies in the fact that the efficiency of sonication varies from sample to sample, particularly in solid tissue. I used a standard protocol for the initial sonication cycles across all samples (see methods) followed by additional cycles as required. Figure 14 shows an example of the fragmentation achieved, with the majority of fragments in the desired range.

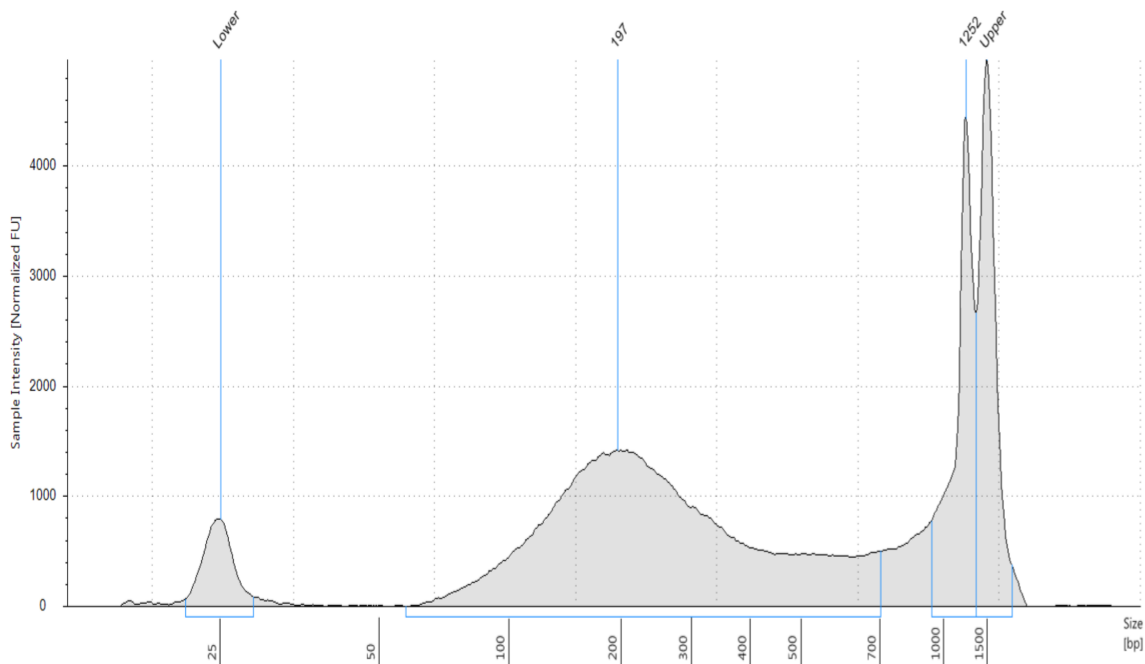


Figure 14: Sonication results. Tape-station profile of sonicated chromatin. The main peak around 200bp accounts for the majority ($\approx 80\%$) of the chromatin in the sample. A larger peak between 2 and 5kb is relatively undigested material.

Optimisation of the immunoprecipitate reaction involved a series of ChIP experiments using various SDS concentrations. I observed that the yield of immunoprecipitated DNA was inversely correlated with the concentration of SDS. However, the library concentration was

not the only factor to be affected by SDS concentration. The final step in a ChIP protocol before sequencing is to check the enrichment of the amplified library. Regions expected to be enriched (gene regulatory regions known to be active in β -cells - positive controls) should have a significantly higher signal compared to regions expected to have low or zero enrichment (negative controls) - evaluated using a quantitative assay such as RT qPCR. Higher SDS concentrations in the IP reduced the enrichment levels of positive controls, most likely by limiting the efficacy of the antibody. Optimal results were thus achieved using an SDS concentration that allowed the best compromise between yield and enrichment.

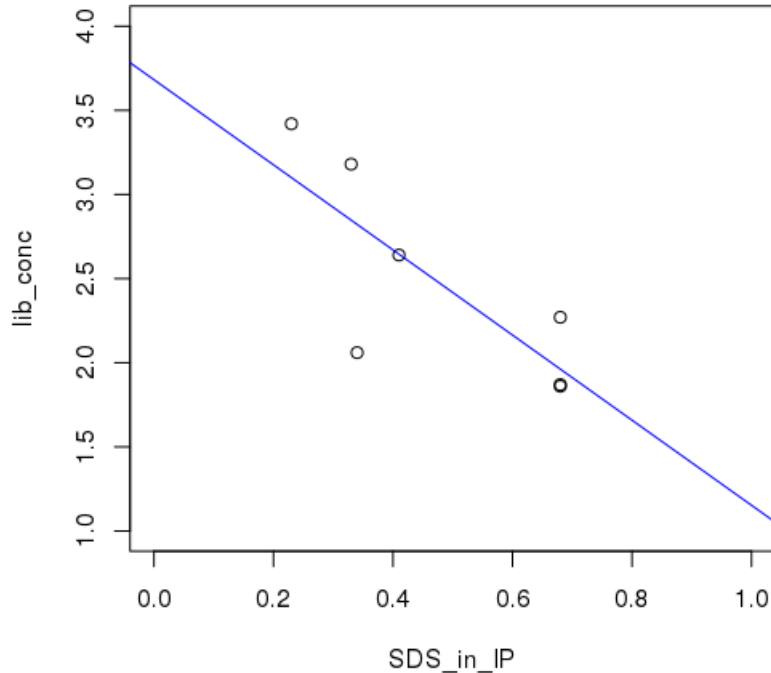


Figure 15: The relationship between immunoprecipitate SDS concentration and ChIP-seq library concentration. SDS concentration of the IP plotted against library concentration for the first few ChIPmentation experiments of the study. A line of best fit shows a an approximately linear relationship.

4.3 Insulinoma and control ChIPs showed strong signal enrichment at key β -cell factors.

Signal-to-noise ratio of the ChIP libraries was evaluated by quantitative real time PCR prior to proceeding with library sequencing. qPCR enables the quantification of initial starting material by measuring the point (PCR cycle number) at which a reaction reaches a fluorescent intensity above background levels. The enrichment of signal was profiled at regulatory regions of genes important for the proper functioning of β -cells.

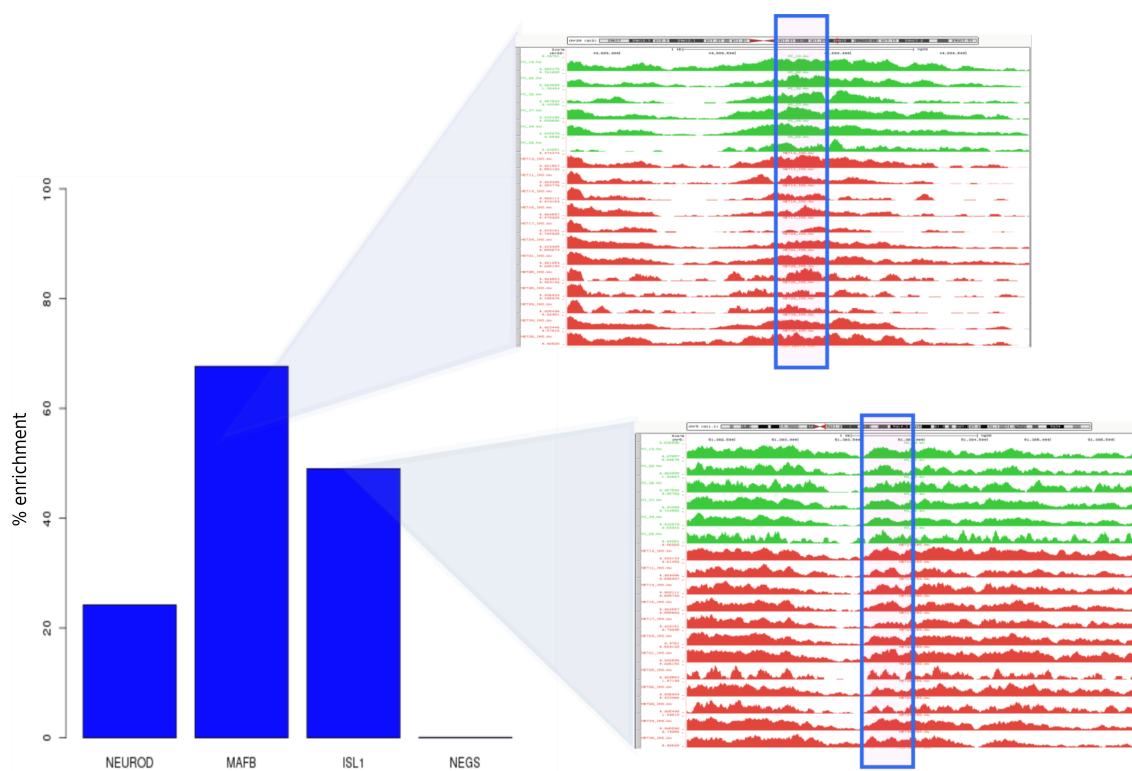


Figure 16: Enrichment qPCR results. Bar plot shows the enrichment of immunoprecipitated material (concentration derived from qPCR Ct values) for positive controls (regions proximal to the genes NEUROD, MAFB and ISL1) compared to negative controls (NEGS, corresponding to regions of chromatin that are inactive in islets) - see primer info (table 2). UCSC web browser images for *MAFB* (top) and *ISL1* (bottom) loci; green and red big-wig tracks correspond to healthy human islet and insulinoma data respectively. Blue boxes highlight the region covered by the qPCR primers.

NEUROD is expressed in all islet cell types but crucially in β -cells it acts as a transactivator of the insulin gene [121]. *MAFB* is required for insulin and glucagon transcription and is required for the maintenance of key β -cell markers *PDX1* and *SLC2A2* [122, 32]. *ISL1* (Insulin Gene Enhancer Protein) binds to the enhancer region of the insulin gene and is an important factor in its transcription [123]. All insulinoma and control ChIP libraries showed strong enrichment (> 10 fold) for these factors over negative controls (fig. 16). The high signal-to-noise ratio and its consistency across samples was confirmed through visual analysis of ChIP profiles via the UCSC genome browser and the subsequent computational analysis.

4.4 QC, genome alignment and peak calling for ChIP-seq data.

Initial analysis of sequence data from ChIP-seq experiments using Fastqc showed the data to be of consistently high quality across all data sets. An example of per base and per sequence quality metrics for insulinoma is given in figure 17.

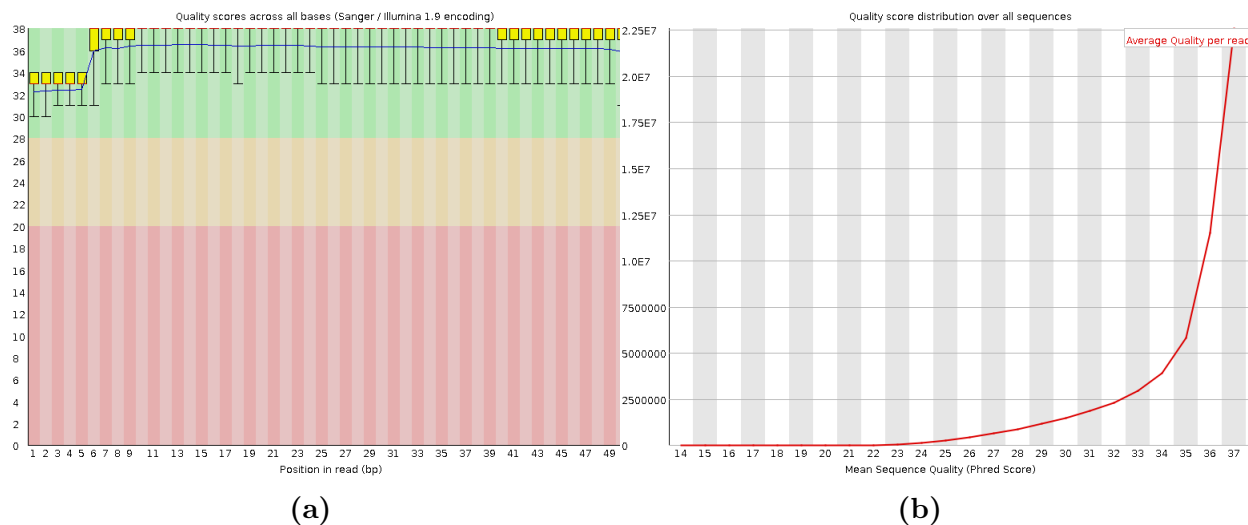


Figure 17: Fastqc analysis results for insulinoma ChIP-seq of H3K27ac. Per base (A) and per sequence (B) quality results from fastqc analysis of ChIP-seq data for an insulinoma sample (NET30).

Insulinoma ChIP-seq data aligned to genome hg38 with a 99% total average alignment rate

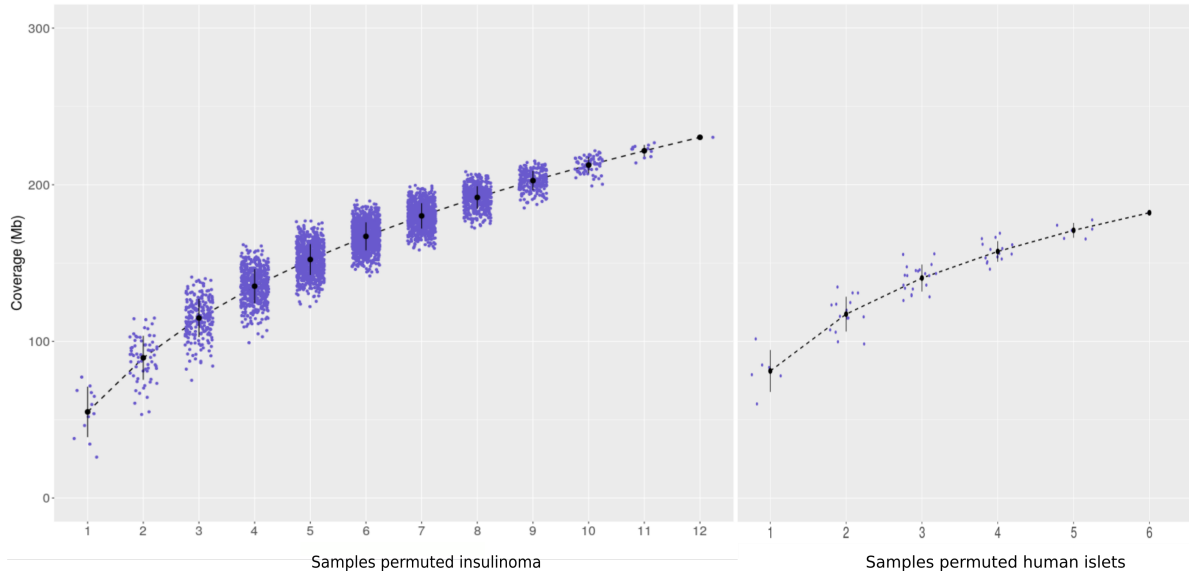
and 80% single alignment rate (table 5). The number of peaks called per sample ranged from 36,000 to 73,000 with a mean average of 54,700. These statistics suggest the successful implementation of experimental protocols and the generation of high quality data.

Sample name	# Raw reads	Genome alignment rate		# Peaks called
		Single (%)	All (%)	
HI_19	105751270	83.5	98.5	75261
HI_22	48182652	83.5	99.1	65290
HI_32	23812683	77.9	93.1	60953
HI_37	85131100	84	99	64152
HI_40	78309325	84	99.2	62420
HI_D2	30967085	81.9	97.3	80598
NET_10	59802823	81.6	98.9	52324
NET_11	75691688	81.4	97.9	72950
NET_14	53599712	81.3	99.2	59213
NET_16	46796968	80.1	99	42875
NET_17	63105141	80.7	98.7	61870
NET_20	63259531	78.6	98.6	51623
NET_21	57372840	80	99.1	56605
NET_25	54135363	79.4	98.9	35933
NET_26	30908538	81.1	98.6	54039
NET_29	44014639	81.6	98.8	49270
NET_30	56360741	79.1	99	56577
NET_38	57015729	76.4	98.7	63074

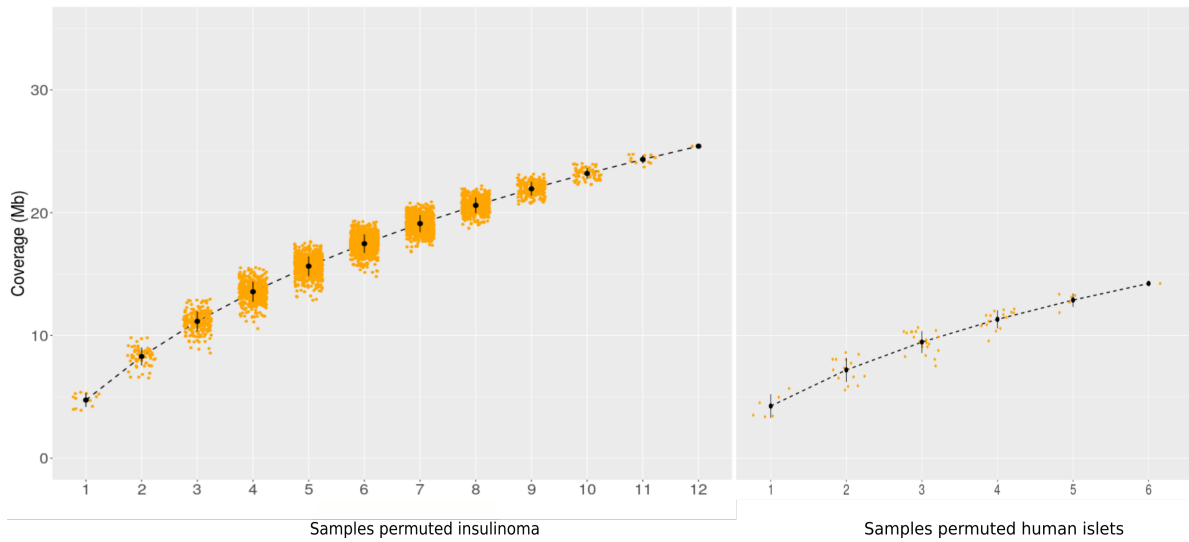
Table 5: Genome alignment and peak calling statistics for human islet and insulinoma samples. HI = human islets, NET - insulinoma.

4.5 A comprehensive profile of active CREs.

Having high quality data is an important first step, but it’s usefulness is limited unless it represents a comprehensive map of the factor in question. This means a set of regions enriched for H3K27ac across the sample cohorts that is representative of the vast majority of active CREs in insulinoma and control samples is required. This is also important in order to give confidence in the statistical analysis and any conclusions drawn from downstream analysis. With a data set incorporating data from x samples, the question is, how many



(a) Distal CREs



(b) Proximal CREs

Figure 18: Coverage plots for H3K27ac enriched regions in insulinoma and human islets. LHS plots = insulinoma (x12); RHS plots = healthy human islets (x6). Dots represent the total coverage (Mb) of peaks of enrichment for H3K27ac for each combination of the number of samples permuted. Black dots represent the mean average coverage and vertical black lines the standard deviation from the mean. Dashed lines connect the means at each level. Proximal defined as within 2kb upstream and 200bp downstream of the nearest TSS.

more regions would be added to the set if we were to profile $x + 1$ samples? Starting with one

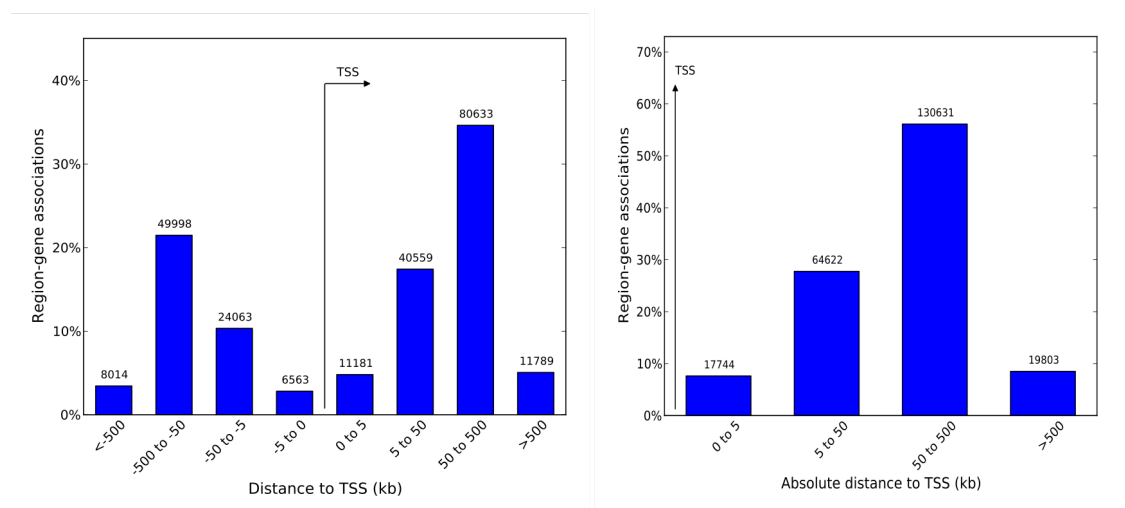
sample and adding one at a time, plotting the coverage (Mb) of CREs for every combination of samples, the difference in average coverage between x and $x + 1$ will be reduced with each iteration until the addition of a new sample ($x + 1$) will not result in an increment of the genomic coverage of the CRE data.

As the insulinoma H3K27ac coverage curve reaches a plateau we are getting close to saturation in terms of approximating a comprehensive set of CREs, and adding extra data would add only a minimal number of extra regions to the set. Figure 18 shows the coverage (Mb) of genomic regions enriched for H3K27ac in our sample cohorts. In each case (insulinoma and unaffected human islets; distal and proximal CREs) coverage approaches a peak, indicating that the data is sufficient to approximate saturation in terms of generating a comprehensive set of active CREs. Importantly, this suggests that we are not excluding a large number of enriched regions from downstream analysis. Calculating the difference between mean coverage for 11 and 12 samples indicates that by adding the 12th sample we gain less than 4% extra enriched sites. Extrapolating forward, by adding a 13th sample we would gain approximately 3.5%, compared to the 28% difference between 2 and 3 samples. The difference in coverage between proximal and distal CREs indicates, in line with the previous analysis, that distal elements are approximately 10 times more abundant. The range of coverage values for each permutation also suggests, as we would expect, more diversity in distal CREs in comparison to proximal regions.

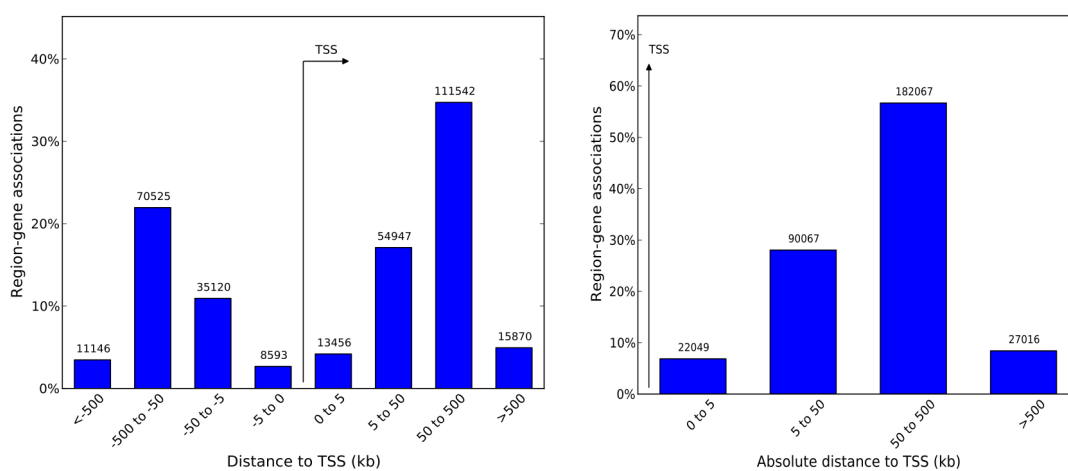
4.6 Characteristics of H3K27ac enriched regions.

Consensus peak sets (cps) generated from ChIP-seq data for healthy human islets and insulinoma contained 131,119 and 178,675 regions respectively. Part of this difference in the number of active regions could be explained by the number of samples in each cohort (12 insulinoma vs 6 human islet). However, from the previous analysis we can say that the number of samples is not a limiting factor in profiling comprehensive data sets of H3K27ac enrichment. This would suggest therefore, that insulinoma development is characterised by

an overall increase in CRE activity.



(a) Human islets



(b) Insulinoma

Figure 19: Distribution of H3K27ac enriched regions. Proportion of H3K27 enriched regions from consensus peak sets of a) human islets and b) insulinoma located within a set distance of annotated transcription start sites as calculated by GREAT analysis.

As mentioned, this data includes both enhancers and promoters, and by analysing specific characteristics of the regions we can begin to describe the landscape of CREs in insulinoma and compare it to that of unaffected human islets. Promoters are defined as proximal CREs, which means they are located within 5kb (usually 2kb) of the transcription start site of the gene they control. Conversely, regions with enhancer function may be located either

proximal or distal to the TSS(s) that they regulate. Figure 19 shows the distribution of regions in each cps relative to transcription start sites annotated to GRCh38. Here we can see that for both islets and insulinomas the majority of regions are distal to TSSs, with less than 10% located within 5kb, which fits with observations from the coverage analysis. As we can expect a proportion of the proximally located regions to feature enhancer activity we can approximate that between 5 and 10% of regions within each cps are promoters.

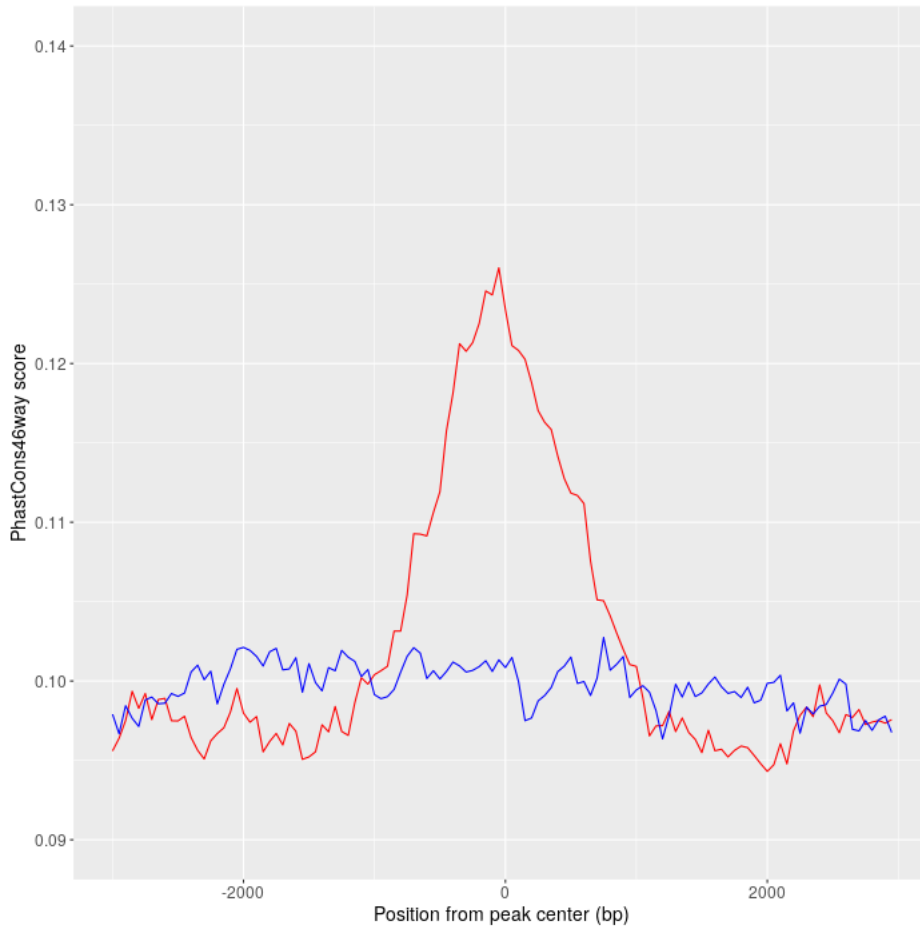


Figure 20: Conservation of insulinoma CREs. Mean phastCons conservation scores for 50bp binned regions across insulinoma CREs (red line) and random genomic regions (blue line).

Another defining feature of gene-regulatory elements is the level of evolutionary conservation. Whole genome sequencing projects focusing on a wide range of animals, from vertebrates to nematodes, have revealed that non-coding sequence is conserved at a much greater level than

was previously expected. In particular, extremely conserved CREs have been found to cluster around genes involved in regulating development [124]. In vertebrates, CREs are often more conserved than protein-coding sequences. Evolutionary conservation of H3K27ac enriched regions is therefore a strong indication of functional importance. Comparing phastCons conservation scores for H3K27ac enriched regions from insulinoma samples to a random control set of genomic regions we observe high mean conservation scores for the insulinoma data (fig. 20). Mean conservation scores are highest (above background) at the centre of each region, but high levels of conservation are recorded for regions encompassing almost 1kb up and downstream of the centre, encompassing a region size that is characteristic of most CREs. These results provide further evidence that the regions identified from ChIP-seq data of H3K27ac enrichment are functionally important in gene regulation.

4.7 Evaluation of accessible chromatin within CREs.

A key part of the identity of any cell type is the action of specific transcription factors. Under normal conditions TFs are the molecular orchestrators, activated by signalling pathways and in turn activating transcriptional programs. Changes to CRE activity in a tumour cell may alter or initiate de novo activity of TFs, which may in turn affect the identity of the cell type from which the tumour develops. CRE regions are defined by the enrichment of flanking histones for specific epigenetic marks, but the CRE itself is located in the accessible regions of chromatin between enriched nucleosomes. Therefore, rather than searching the entirety of regions represented by ChIP-seq peaks for TF binding sites, a more targeted approach is to use accessible regions (NFRs) overlapping ChIP-seq peaks. This concept is illustrated in figure 21.

Quality control of ATAC-seq libraries showed suboptimal enrichment ($< 5x$) for positive controls (regions expected to be accessible in cells with a β -cell identity) and fragment distribution, for most insulinoma samples assayed. Ensuring efficient chromatin fragmentation (whilst preserving native chromatin structure) and Tn5 transposase distribution is more

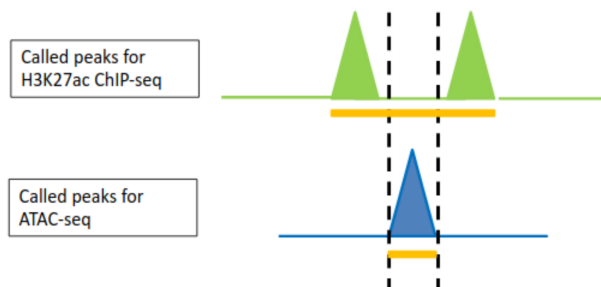


Figure 21: Identifying TF binding sites within CREs. Schematic representation of the identification of TF binding sites by overlapping ChIP-seq and ATAC-seq data.

challenging in solid tissues than in cell lines, and these factors may account for the limited success seen here. Finally, I sequenced five ATAC-seq libraries and four human islet control libraries. Narrow peak calling using MACS2 produced an average of $\approx 100,000$ peaks for islet ATAC-seq data. In contrast 3 of the insulinoma ATAC-seq data sets produced fewer than 10,000 peaks and only one produced more than 30,000. We thus decided to combine this data with other ATAC-seq profiles, to locate possible TF binding sites within the active CREs already identified. ATAC-seq data from multiple sources was used to direct the search for TF binding DNA motifs overlapping peaks from insulinoma ChIP-seq data. I incorporated data from genome-wide chromatin accessibility profiling of more than 400 tumour samples [94]. The combined data set, generated using data from insulinomas, human islets and various cancer types, as well as the results of nfr analysis of insulinoma ChIP-seq reads, included more than 500,000 accessible regions. This was sufficient to assign one or more accessible regions to the vast majority of active CREs from our insulinoma ChIP-seq data sets, enabling a comprehensive search for TFs involved in insulinoma development.

4.8 Summary and concluding remarks.

I have generated the first genome-wide profiles of H3K27ac enrichment in a cohort of insulinoma samples. This is a big step towards identifying insulinoma-specific gene-regulatory elements. Moreover such data can be used to identify TF binding sites within active CREs by

integrating chromatin accessibility data. One potential limitation of studying a rare tumour, in this case with a sample cohort of less than 20, is heterogeneity. But by analysing the coverage of H3K27ac enriched regions across our cohort we have demonstrated that we are able to capture most of the variability due to tumour heterogeneity for this histone mark. For the majority of insulinomas I also generated corresponding RNA-seq data enabling the profiling of the insulinoma transcriptome which could be used to infer gene regulatory networks.

Sample ID	YY1 T372R	RNA	H3K27ac	ATAC
NET_10	WT			
NET_11	WT			
NET_14	WT			
NET_16	T372R			
NET_17	WT			
NET_20	WT			
NET_21	WT			
NET_23	WT			
NET_25	WT			
NET_26	WT			
NET_29	WT			
NET_30	WT			
NET_33	WT			
NET_34	WT			
NET_35	WT			
NET_36	WT			
NET_38	T372R			

Table 6: Summary of experiments for all insulinoma samples in the study cohort. Green cells indicate experiments completed successfully, blank cells indicate no experiment completed, WT = wild type, T372R = samples heterozygous for that mutation.

I have demonstrated that the data is of sufficient quality to use for downstream analy-

ses. This data enables comparisons of enhancer activity and gene expression in insulinoma and unaffected human islets and thus identification and characterisation of changes in the gene-regulatory architecture that lead to insulinoma development. Table 6 summarises the experiments conducted for each insulinoma sample. Where possible RNA-seq analysis was performed for each sample for which ChIP-seq was successfully completed. YY1 T372R mutation analysis was performed for all insulinoma samples, regardless of the success of other assays.

5 The gene-regulatory landscape of insulinoma.

Bioinformatics may be defined in several ways, but essentially it involves developing and applying computational methods to analyse data from biological experiments to produce an interpretable output. It requires using robust statistical methods, but there is no one-size-fits-all approach. The methodology to conduct experiments in molecular biology has developed at an incredible pace, generating vast amounts of data, so the design and use of in silico methods for handling this data is of ever increasing importance. Using the data described in the previous chapter I take on this challenge in order to describe the gene-regulatory landscape of insulinomas. First though, it is necessary to genotype each insulinoma for known recurrent mutations, in order to enable accurate interpretations of any novel findings.

5.1 Genotyping known recurrent mutations.

Despite excluding all subjects with a family history of MEN1, I nevertheless detected one insulinoma in our cohort (NET14_ INS) carrying a coding mutation in the *MEN1* gene. I also genotyped all insulinoma samples in our cohort for the T372R mutation in *YY1*. Sanger sequencing of the region surrounding this mutation identified heterozygous mutations in 2 (NET16_ INS & NET38_ INS) out of the 20 insulinomas (10%) in our cohort (fig.22). This represents a much lower mutation rate for T372R than the 30% reported in other studies [56]. In addition the minor allele was at a low level in both samples, suggesting mosaicism of the T372R genotype within the tumour samples. Unfortunately the low number of cases detected limits the possibility of performing solid statistical analysis when comparing tumours carrying this mutation against *YY1* wild type tumours. Nevertheless, we can speculate on the potential effect of this mutation on downstream analysis as the findings from RNA-seq data (fig.23) and cluster analysis of CRE activity (fig.27b), described below, suggest it results in widespread changes in CRE activity and subsequently gene expression.

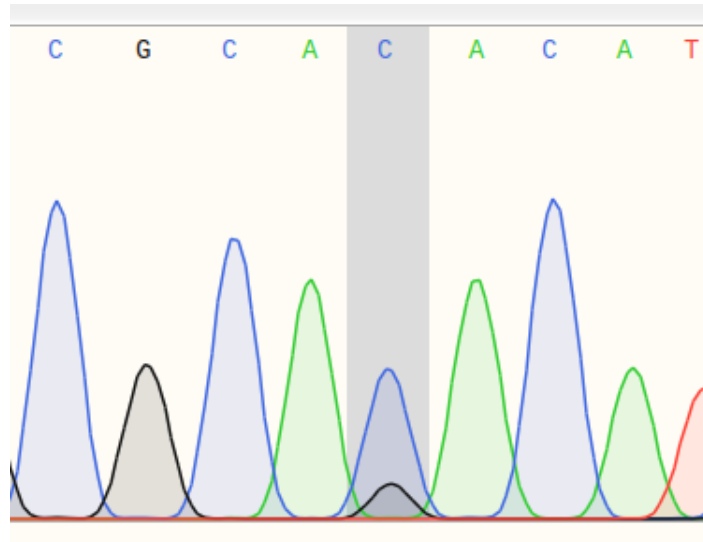


Figure 22: Genotyping of the T372R mutation in *YY1*. Chromatogram of Sanger sequencing for NET38_ INS, one of the insulinomas heterozygous for the c.C1115G/p.T372R mutation in the *YY1* gene.

5.2 Insulinoma development is driven by changes in gene expression.

The identity of a cell is dictated by the action of signalling pathways, transcription factors and gene-regulatory elements. The observable output of these factors is the transcriptome, the amount of RNA corresponding to each possible gene transcript in the cell, in other words, the level of expression of each gene. As described, this output is frequently modified in disease states, and therefore identifying gene expression that is significantly different in an abnormal cell or tissue compared to the normal state is extremely important. The starting point for delineating the relationship between data sets from high-throughput sequencing experiments is a matrix of count data where rows represent transcripts or genomic regions and columns represent samples. Various algorithms may then be applied, the majority of which enable modelling of the data in order to estimate variance.

5.2.1 Data correlation and hierarchical clustering

Using unsupervised hierarchical clustering of pairwise correlations of transcript abundance in insulinomas ($n = 11$) and unaffected human islets ($n = 7$) we found a clear distinction between tumour samples and healthy islets indicating that tumour development involves widespread changes in gene expression (fig.23). Insulinoma samples heterozygous for the T372R mutation in YY1 (NET16_ INS & NET38_ INS) cluster together, and separately from the rest of the insulinomas, suggesting that this mutation has a significant effect on gene expression. NET10_ INS also features a significantly different gene expression profile, and appears similar to NET16_ INS & NET38_ INS. However, NET10 does not carry a mutation in YY1 and we are yet to identify a mutation that would account for the gene expression profile of this sample.

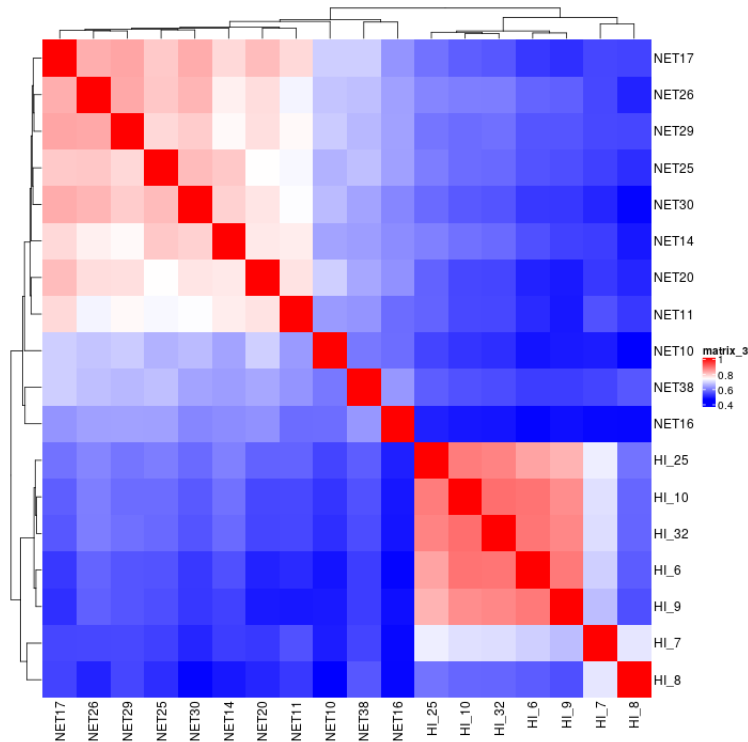


Figure 23: Insulinoma development is characterised by significant changes in gene expression. Unsupervised hierarchical clustering of pairwise Spearman correlations of normalised transcript abundance counts for insulinoma (NET) and unaffected human islets (HI).

There is also some heterogeneity within the unaffected human islet samples, but crucially all human islet control data clusters in a distinct group separate from the insulinoma data, enabling a direct comparison of normal and pathogenic gene expression profiles.

5.2.2 Differential analysis of gene expression.

Given that insulinomas can be distinguished from unaffected human islets on the basis of gene expression profiles, we sought to identify genes that are differentially expressed in insulinoma. Count data from RNA-seq for a total of 50513 transcripts was analysed across 11 insulinoma and 7 human islet samples. 5980 transcripts were significantly ($\log_2FC > [1.5]$ & $FDR < 0.05$) up-regulated (insulinoma vs human islets) and 5168 were significantly down-regulated (fig.24). Focusing on the insulin gene, expression of *INS* in our cohort was 10 fold higher in insulinoma vs human islets. So in addition to dysregulation of insulin secretion, the expression of *INS* is significantly higher in insulinomas. Interestingly, expression of *INS* was $\approx 1000x$ higher in insulinoma than a small cohort of PDX+ (derived from β -cells) non-functional PNETs, suggesting a significant loss of insulin production in non-functional PNETs.

Two *PAX6* protein-coding transcripts were also 10 fold more abundant, which may explain, at least in part, the upregulation of insulin, given the role of PAX6 in insulin gene expression. Interestingly, PDX1 was significantly ($FDR < 0.01$) downregulated (log fold change -2.3). Although PDX1 plays a significant role in establishing β -cell identity, studies have shown that islets with reduced PDX1 expression tend to mount a larger Ca^{2+} response to glucose, possibly mediated by an increase in expression of Ca^{2+} channel subunits [35]. Taken together, these results point towards a mechanism involving significant changes in gene expression that is responsible for the dysregulation of insulin production and secretion observed in insulinoma. Other key TFs that are involved in establishing β -cell identity, including NKX6.1 and MAFA are expressed equally in insulinoma and unaffected human islets.

The majority of genes encoding chromatin modifiers, in which recurrent variants described

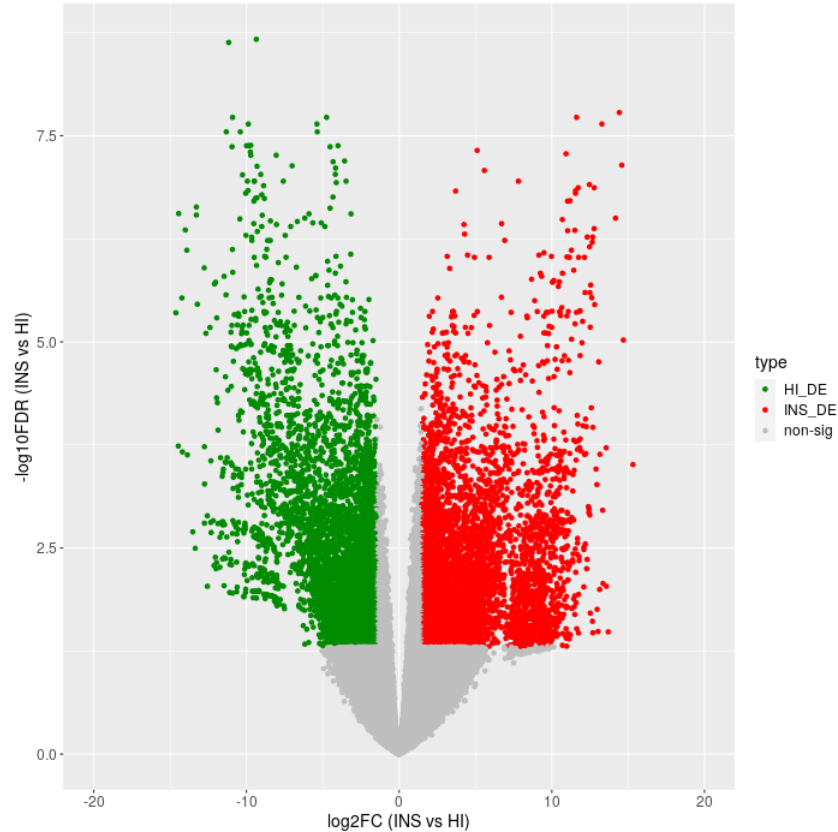


Figure 24: Differential analysis of transcript abundance. Fold change (\log_2) vs FDR adjusted p-values ($-\log_{10}$) for differential analysis of transcript abundance (insulinoma vs human islets). Green and red dots indicate significantly ($\log_2FC > [1.5]$ & $FDR < 0.05$) downregulated or upregulated transcripts in insulinoma respectively. Grey dots indicate non-significant results ($FDR \geq 0.05$).

by Wang et al [42] are located, are stably expressed (insulinoma vs human islets) in our cohort. The exception to this is the lysine methyl transferase *KMT2C*, which is significantly upregulated in insulinoma ($\log_2FC + 2.8$, $FDR 0.003$). *KMT2C* (along with *KMT2D*) methylates lysine 4 of histone 3, increasing genome-wide deposition of H3K4me1. In doing so it facilitates the activation of enhancers, partly by promoting the recruitment of the coactivator and H3K27 acetyltransferase p300 [125, 126]. Wang et al observed a recurrent copy number gain variant in *KMT2C* in insulinoma samples (compared to healthy β -cells), and as copy number gain is associated with increased gene expression (as observed in a pan-cancer study [127]) this is in line with what we would expect if *KMT2C* is overexpressed.

5.2.3 Upregulated genes are enriched for chromatin modifiers.

Given that gene expression changes are associated with the insulinoma phenotype, I next sought to find specific gene sets or pathways that are overrepresented in the data. Expression changes in individual genes may be the result of a multitude of different factors, and may have only a small effect on the identity of a cell. However, finding gene sets that form part of specific pathways, enriched for genes with expression changes in the same direction, would be very significant, and enable us to tease out the changes most likely to have a large effect. GSEA, a widely used method for analysis of gene expression data, determines whether differences between biological states are significantly associated with a priori defined gene sets. However, classical GSEA methods are suboptimal for data sets of different sample sizes and experimental designs. Initial analysis with a standard GSEA approach produced inconsistent results, making interpretation challenging.

A more recent development of GSEA enables the use of pre-ranked gene expression data. Pre-ranked GSEA produces an enrichment score for each gene set which is reflective of the frequency with which individual genes in the set occur at the top or bottom of the ranked data set. Pre-ranked gene-set enrichment analysis of normalised counts of transcript abundance identified 17 significantly ($FDR < 0.05$) up-regulated Reactome pathways (insulinoma vs unaffected human islets), the most significant of which are displayed in fig.25. Interestingly, one of the most significantly up-regulated gene sets was 'chromatin modifying enzymes', indicating that chromatin modification is a key part of insulinoma development. More specifically, histone demethylases (HDMs) and histone acetyl transferases (HATs) were up-regulated, suggesting that a mechanism involving the derepression of CREs in human islets is an important factor in insulinoma development. This supports the hypothesis that up-regulation of enhancer activity is a key driver of insulinoma development and an important factor in insulinoma-specific gene-regulatory networks.

Genes involved in neuronal cell development and signalling by Rho GTPases and receptor tyrosine kinases (RTK) are also upregulated. One of the most significantly up-regulated

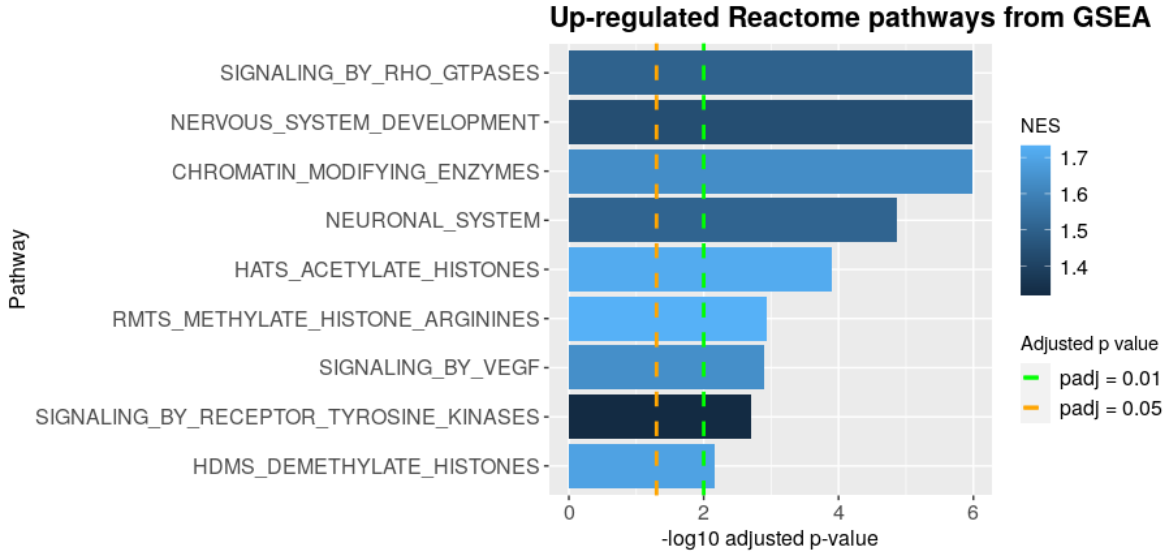


Figure 25: Gene-set enrichment analysis identifies up-regulated gene-sets in insulinoma. The most significantly up-regulated (insulinoma vs unaffected human islets) gene-sets (adjusted p – value < 0.01) identified by pre-ranked GSEA. Lighter blue bars represent higher normalised enrichment scores; vertical dashed lines mark adjusted p -value thresholds.

genes ($\log_2FC = 8.7, FDR = 0.0005$) in the Rho GTPase gene set is *CDC42*. *CDC42* is involved in the regulation of multiple cellular functions and is known to be an important mediator of cell proliferation and insulin granule mobilisation [128]. The leading edge of the RTK gene-set includes the insulin receptor gene (*INSR*) and epidermal growth factor receptor (*EGFR*). *EGFR* expression is required for embryonic β -cell maturation, islet migration and maintenance of β -cell mass and proliferation [129, 130].

I further demonstrated the up-regulation of genes in pathways identified by GSEA analysis by comparing the fold change of transcripts within each upregulated gene set to all other transcripts (control set) from differential analysis of RNA-seq data. The average \log_2FC for the control set is, as expected, very close to zero, whilst the average for each GSEA gene is ≈ 5 , confirming the upregulation of genes in these pathways (fig. 26). Curiously, no gene sets were significantly ($padj < 0.05$) downregulated in insulinoma, which suggests that activation, rather than repression, of gene-regulatory networks is the most prominent driver of insulinoma development.

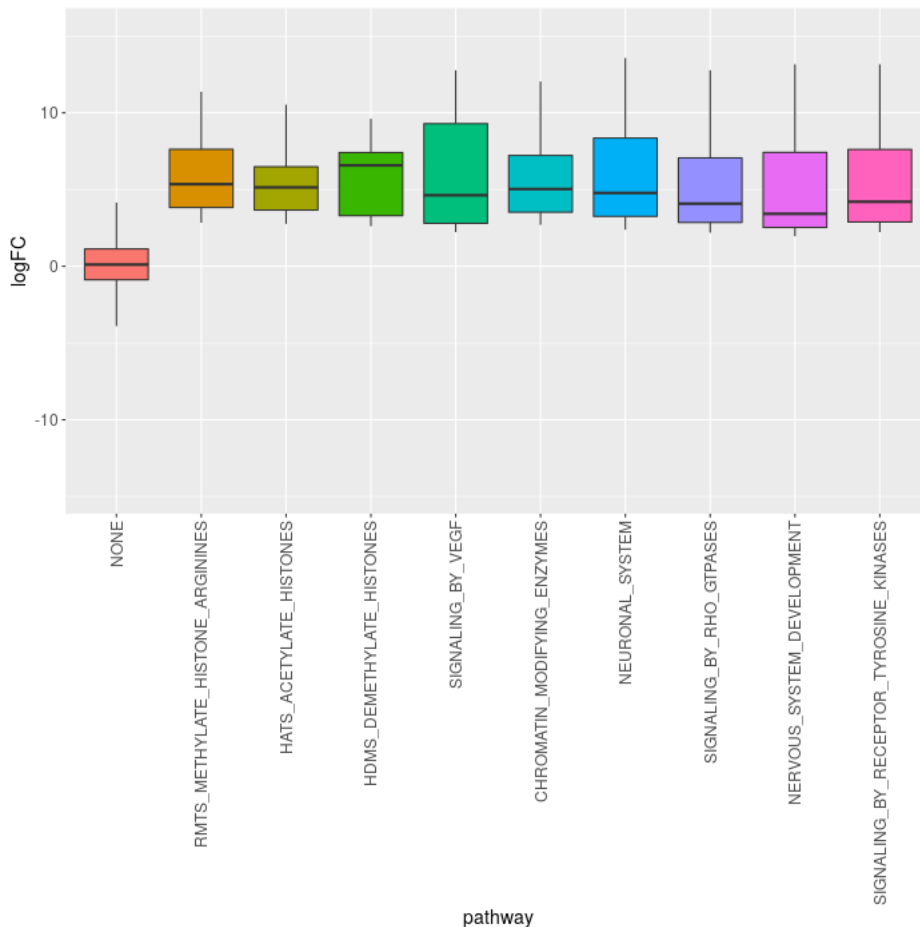


Figure 26: Genes in up-regulated pathways show higher than average expression. Distribution of \log_2FC (insulinoma vs unaffected human islets) of normalised RNA-seq counts for transcripts from up-regulated gene-sets identified by gene-set enrichment analysis and a control group of all remaining transcripts from RNA-seq analysis.

5.3 Profiling CREs enriched for H3K27ac genome-wide can distinguish insulinoma from human islets and non-functional PNETs.

If the development of insulinoma is driven by significant changes in gene expression and chromatin modification by HATs, we would expect to observe changes in the profile of H3k27ac enrichment in insulinoma samples compared to unaffected human islets. Correlation clustering of ChIP-seq data (fig.27) demonstrates a clear distinction between insulinomas and human islets suggesting that changes to enhancer activity are a key part of the development

of these tumours. I also integrated H3K27ac ChIP-seq data for non-functional PNETs from published research. Here too we observe a clear distinction from insulinoma, suggesting that alteration of gene regulatory mechanisms is a key marker of both cell proliferation and dysregulation of insulin secretion that is characteristic of insulinomas.

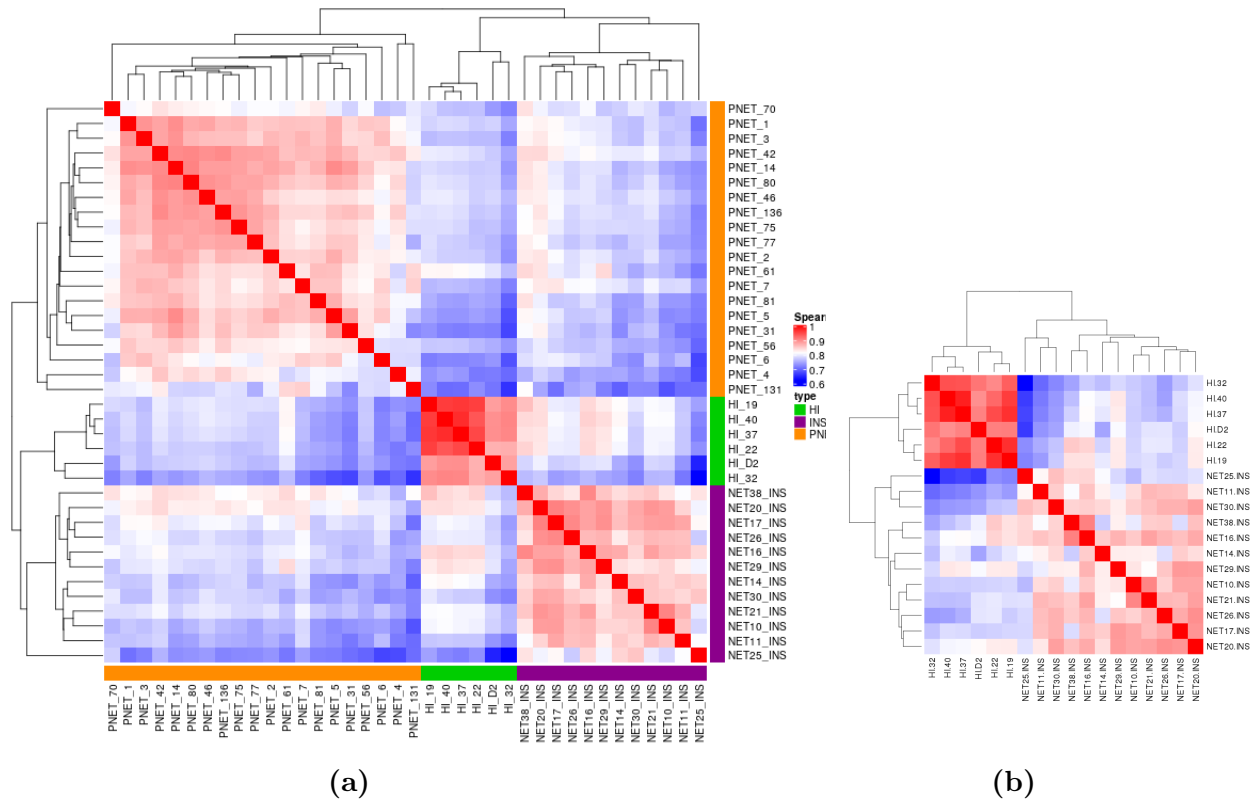


Figure 27: Cluster analysis of H3K27ac ChIP-seq data. Unsupervised hierarchical clustering of pairwise Spearman correlations of normalised ChIP-seq read counts. **a)** Insulinoma (NET, purple), unaffected human islets (HI, green) and non-functional PNETs (PNET, orange). **b)** Insulinoma (NET) and unaffected human islets (HI).

Performing correlation cluster analysis using large sample sets (such as the non-functional PNETs) can mask some relationships between samples. For example, differences between samples in one cohort may appear less significant when compared to samples from different cells and tissue types than if they were compared to cells from the same origin. For this reason I performed a separate correlation cluster analysis comparing just insulinoma and human islet samples (fig.27b). This analysis highlights a high degree of similarity between insulinomas heterozygous for YY1 T372R. Once again we do not have enough samples carrying this

mutation to enable statistically robust analysis, but we can postulate that this mutation is effecting CRE activity genome-wide in these tumours, possibly via changes in H3K27ac deposition resulting from aberrant DNA binding of YY1.

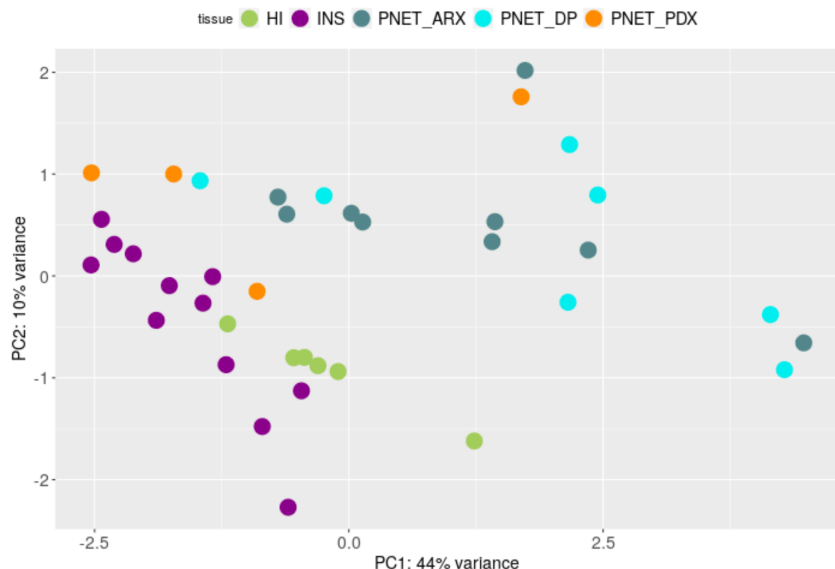


Figure 28: Principal component analysis of H3K27ac data. PCA plot of insulinoma (x_{12}) (purple dots), unaffected human islets (x_6) (green dots), PDX+ PNETs (x_4) (orange dots, ARX+ PNETs (x_9) (grey dots), double positive (DP) PNETs (x_7) (blue dots).

In the clustering analysis I treated all non-functional PNETs as one cohort, but Cejas et al described differences in the enhancer profiles of these samples and classified them by the enrichment of H3K27ac activity around key genes encoding cell-type-specific TFs [98]. Thus 'alpha-like' and 'beta-like' non-functional PNETs featured signal enrichment around *ARX* and *PDX1* respectively, whilst a third group (double positive (DP)) showed signal enrichment around both genes. PCA analysis (fig.28) revealed that overall the enhancer landscape of PDX+ PNETs is more similar to that of insulinoma than ARX+ or DP PNETs. So whilst we identify distinct gene regulatory mechanisms driving insulinoma we also observe changes to the gene-regulatory architecture that are common to both functional and non-functional PNETs derived from the same cell type. Furthermore, there is some overlap between insulinomas and unaffected human islets in terms of CRE activity, and we could

potentially split the insulinomas into two or more groups on the basis of this factor. Moving forward though, for the purposes of statistical power in deciphering insulinoma-specific gene-regulatory networks, we will consider all insulinomas as a single group.

5.4 Insulinoma-enriched CREs.

I next sought to determine the differences in CRE activity between insulinoma and unaffected human islets in order to identify CRE activity involved in insulinoma development.

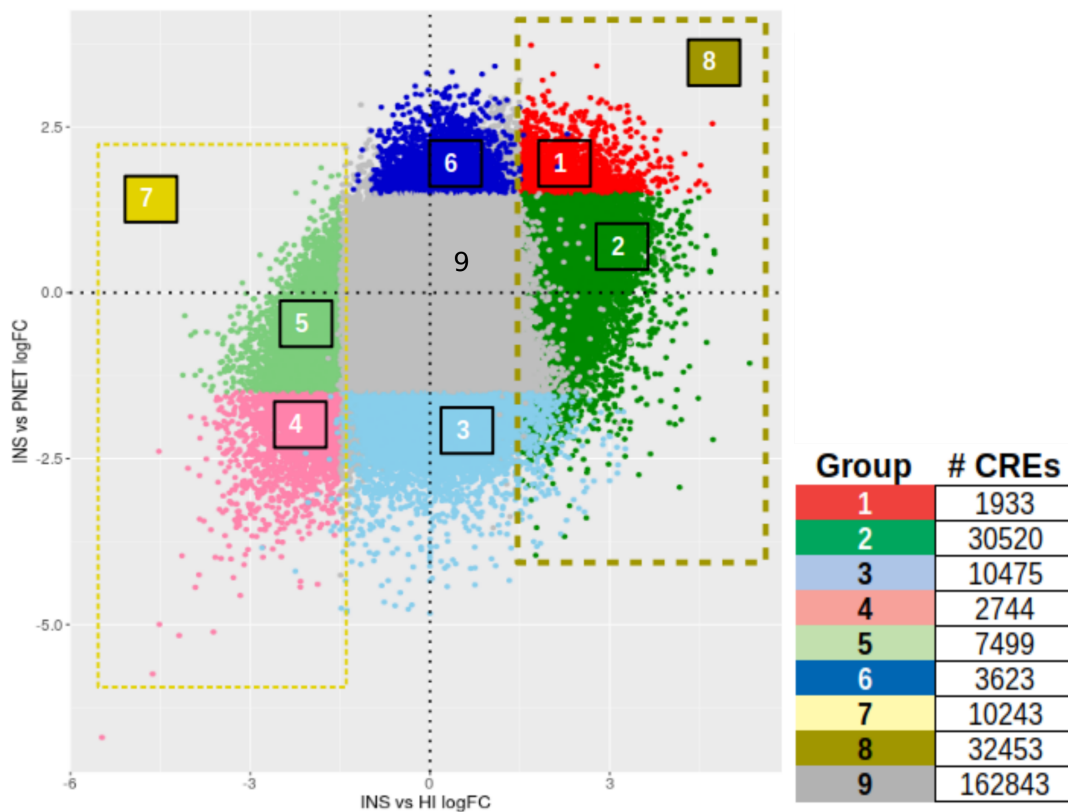


Figure 29: Differential analysis of H3K27ac ChIP-seq data. Fold changes (\log_2) of H3K27ac enrichment at CREs in insulinoma vs unaffected human islets (x axis) and insulinoma vs non-functional PNETs (y axis). Coloured dots represent CRE with $FDR < 0.05$ and $\log_2 FC \geq [1.5]$. Grey dots represent regions that fall outside of the FDR and fold change cutoffs and are therefore considered equally active in the 3 tissues.

Differential analysis of ChIP-seq data identified groups of CREs that are differentially enriched for the H3K27ac mark in insulinoma compared to unaffected human islets and non-

functional PNETs (fig.29). I compared $> 250,000$ H3K27ac enriched regions comprising a consensus peak set generated using data from all samples. 32,453 regions display insulinoma-specific activity when compared to unaffected islets (group 8), and a small subset of almost 2000 CREs (group 1) are more active in insulinoma compared to both unaffected islets and non-functional PNETs. These CREs are phylogenetically conserved (fig 30) and 97% are distal to the nearest TSS. Interestingly, there are three times more CRE regions that are up-regulated in insulinoma (compared to unaffected islets) than down-regulated.



Figure 30: Conservation of insulinoma-specific CREs. Mean phastCons conservation scores for 50bp binned regions across insulinoma CREs (red line) from group 8 of figure 29 and a set of random genomic regions (blue line).

Often the most significant changes (such as exonic mutations) are associated with loss-of-function of proteins or pathways associated with tumour suppression, but here we see clear evidence of a gain-of-function mechanism leading to tumour development. Regions falling outside of the fold change and adjusted p-value cutoffs are considered equally active in the

three tissues and will be referred to as 'stable'. The groups of CREs identified from this analysis (1-9 in fig.29) formed the basis of downstream analysis aimed at deciphering gene-regulatory networks driving the development of functional β -cell tumours.

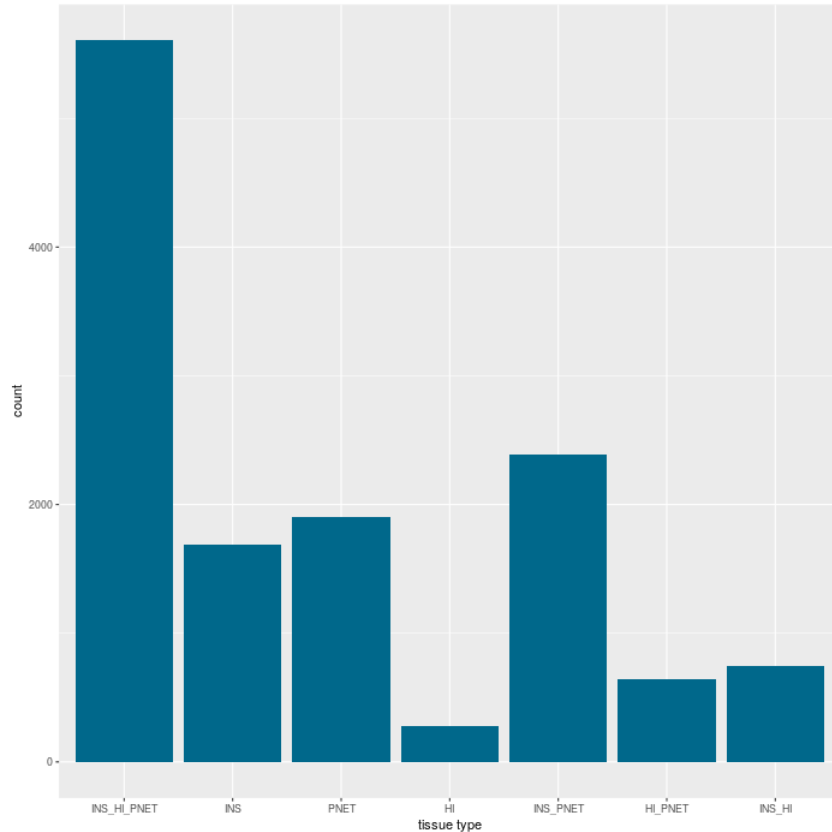


Figure 31: Distribution of super-enhancers across human islets and PNETs. Blue bars indicate the number of super-enhancers called in each tissue. X axis labels indicate the tissue profiled (INS = insulinoma, HI = unaffected human islets, PNET = non-functional PNETs).

Since the vast majority of changes in H3K27ac enrichment were found distal to TSSs (i.e. not promoter regions) I also profiled super-enhancers in unaffected human islets, insulinomas and non-functional PNETs and generated a set of 'insulinoma-specific' super-enhancers (SEs). Over 5500 SEs were called in the 12 insulinoma samples, half of which were specific to insulinomas compared to unaffected human islets and almost 2000 were specific compared to islets and non-functional PNETs. Furthermore, the number of SEs specific to either tumour type is more than six times higher than SEs specific to unaffected islets. This can

be accounted for in part by the fact that there are more samples in both tumour cohorts, but this data also suggests that de novo establishment of SEs maybe a key marker of PNET development. As SEs are characterised by very high enrichment for H3K27ac deposition, this result further highlights the importance of CRE activation in insulinoma development. De novo SE formation could account for some of the most significant changes in gene expression observed in insulinomas compared to unaffected islets.

5.5 Motif analysis for putative insulinoma CREs

Motif analysis was performed on all regions of accessible chromatin overlapping distal (outside of loci parameters defined for proximal regions) CREs from group 8 of figure 29. For each significantly enriched motif I searched for associated transcription factors expressed in our insulinoma cohort with $\log_2 FC > 0$ compared to unaffected human islets. Three TFs with the most significantly enriched motifs are shown in figure 32. Early B-Cell Factor 1 (EBF1) is required for B cell lineage commitment, repressing factors that promote alternative cell fates [131]. It has also been shown to be enriched in regions of hypomethylated DNA in breast cancer. Pathway analysis of hypomethylated regions correlated to nearby EBF1 motifs showed enrichment for regulation of cell proliferation and apoptosis [132]. Myocyte enhancer factor 2C (MEF2C) is one of several forms of MEF2 TF (and the predominant form in the mammalian cerebral cortex) and induces a mixed neuronal/myogenic phenotype in P19 precursor cells [133]. Constitutive activation of MEF2C rescued P19 cells dominant negative for the mitogen-activated protein kinase p38 α from apoptosis, suggesting that MEF2C prevents cell death during neuronal differentiation. ETS-related gene (ERG) is one of 28 E-26 transformation-specific (ETS) genes in the human genome and was first described in human colorectal carcinoma cells [134]. ERG has a key role in embryonic development via regulation of the WNT/ β -catenin signalling pathway. It is also overexpressed in a high proportion of prostate carcinomas and is one of the most consistently overexpressed oncogenes in malignant prostate cancer. High levels of ERG are associated with loss of cell polarity

and changes in cell adhesion [135].

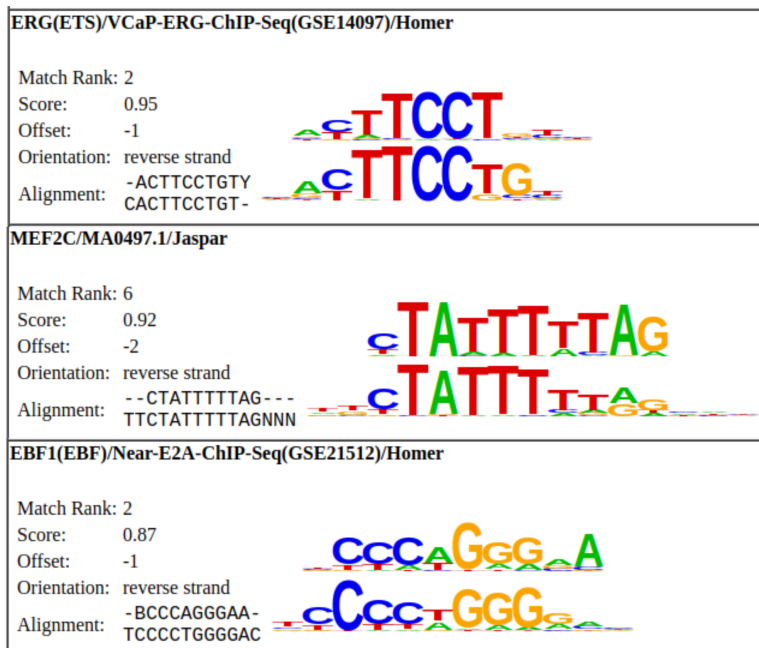


Figure 32: Motif enrichment in differentially active CREs. Information on TFs and associated consensus binding sites obtained from motif analysis using HOMER.

Transcription factor	TF match score	p-value	RNA-seq $\log_2 FC$ INS vs HI
ERG	0.95	$1e^{-67}$	+5.0
MEF2C	0.92	$1e^{-27}$	+4.5
EBF1	0.87	$1e^{-26}$	+5.0

Table 7: Summary of data from motif analysis.

5.6 Summary and concluding remarks

I have shown that the insulinoma samples in our cohort feature genome-wide enhancer activity and gene expression profiles that are significantly different from unaffected human islets. More than 30,000 CREs and 5000 transcripts (including chromatin modifying enzymes) are significantly upregulated (insulinoma vs human islets). I have identified over 2000 de novo super-enhancers in insulinoma, adding further evidence of the importance of enhancer activity in insulinoma development. I have also identified DNA binding motifs for TFs (known

to regulate cell identity and cell proliferation) that are enriched in insulinoma-specific CREs and for which corresponding transcripts are upregulated in insulinoma (vs unaffected human islets). Further investigation of the functional relationship between upregulated CREs and genes may enable insights into the structure of insulinoma-driving gene regulatory networks.

6 Insulinoma-specific gene-regulatory networks.

6.1 Insulinoma-specific CREs are linked to changes in gene expression.

Having established that the development of insulinomas is coupled with broad changes in gene expression and CRE activity, the next challenge is to find functional links between activated CREs and up-regulated genes. To establish whether, on a broad scale, insulinoma-specific CREs have a functional impact on gene expression, I overlapped the genomic coordinates of enhancers in each group of figure 29, with transcript coordinates from insulinoma and human islet RNA-seq data (extended by 200kb (see methods)).

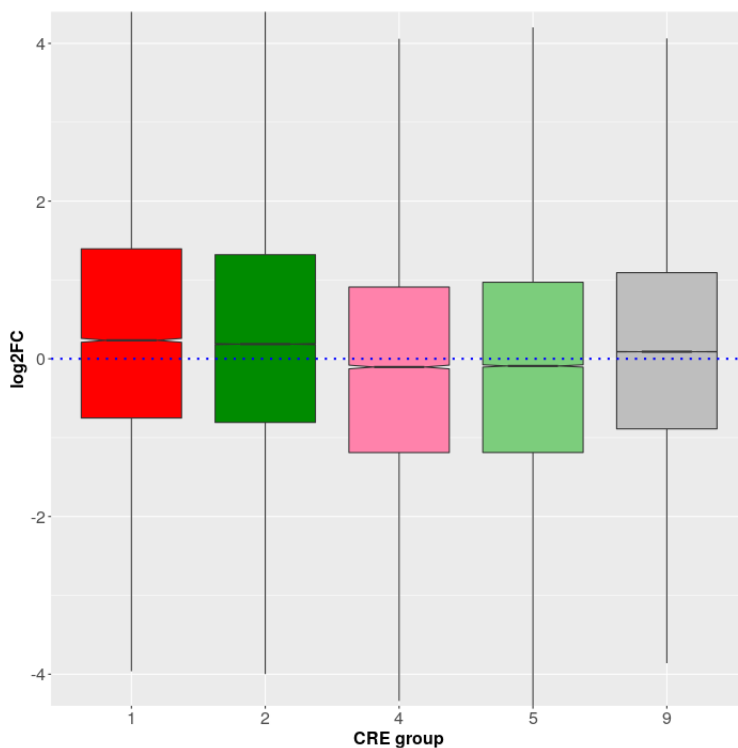


Figure 33: Linking enhancer activity to gene expression. Box plots showing the range of \log_2FC (from differential analysis of RNA-seq data: insulinoma vs human islets) for transcripts with TSS within 200kb of enhancers in CRE groups (CRE classification is as in figure 29 (dark (red and green), light (red and green) and grey boxes represent transcripts associated with CREs that are up-regulated, down-regulated and stable respectively). Box plot limits show upper and lower quartiles.

The average \log_2FC of transcript abundance (insulinoma vs human islet) is significantly

higher for genes associated with differentially active enhancers in insulinoma than for stable enhancers (fig.33). Concordantly, enhancers that are more active in islets are associated with genes with reduced expression levels in insulinoma compared to unaffected islets. Statistical tests confirmed that the increase in average \log_2FC (INS vs HI) of genes associated with up-regulated enhancers compared to those associated with stable enhancers is significant ($p - value < 2.2 \times 10^{-16}$).

6.2 Putative enhancer-promoter links hint at mechanisms of insulinoma development.

Linking enhancer activity to changes in gene expression is a major challenge in modern molecular biology given the large variation in location of distal regulatory elements relative to the promoter(s) they interact with. Chromatin conformation experiments can be used to identify such contacts genome-wide, but often have limited resolution. 3C and 4C approaches can provide higher resolution but the results are limited to the specific genomic loci analysed. In silico approaches to predicting EPIs rely on the availability of multiple epigenetic and transcriptomic datasets to build models of chromatin architecture surrounding CREs and describe how these models are related to gene expression. Without these data sets, distal CREs can be assigned to genes simply by proximity, using a window around transcription start sites, but this approach has significant limitations. The number of positive (and false positive/negative) interactions identified will vary greatly depending on the size of the window used and the gene-specific regulatory landscape. In fact one study found that less than 50% of distal CREs regulate the nearest gene [136].

Initial analysis of the potential function of insulinoma-specific CREs utilised the Genomic Regions Enrichment of Annotations Tool (GREAT), which uses a binomial test to assign ontology terms to CREs within a user defined window. GREAT analysis resulted in significant hits for RTK signalling, regulation of apoptosis and regulation of ion transport when using CREs in group 8 of figure 29. This is a useful starting point but relies on evidence from

studies of CRE-gene relationships from a large number of studies in a wide range of tissues. A more robust and informative approach could take advantage of the association between large groups of CREs and upregulated gene sets identified in the tissue(s) of interest, rather than approaching the problem from the perspective of individual EPIs.

gene set	1	4	2	5	8	7	Super_enhancers
<u>RMTS METHYLATE HISTONE ARGININES</u>	0.6204	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
<u>HATS ACETYLATE HISTONES</u>	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
<u>HDMS DEMETHYLATE HISTONES</u>	1.0000	1.0000	1.0000	0.7463	1.0000	1.0000	1.0000
<u>CHROMATIN MODIFYING ENZYMES</u>	0.2787	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
<u>SIGNALING BY VEGF</u>	0.2697	1.0000	0.0090	1.0000	0.0090	1.0000	0.0090
<u>NEURONAL SYSTEM</u>	0.0090	1.0000	0.0090	0.3956	0.0090	0.5574	0.0090
<u>SIGNALING BY RHO GTPASES</u>	0.0450	1.0000	0.0090	1.0000	0.0090	1.0000	0.0539
<u>NERVOUS SYSTEM DEVELOPMENT</u>	0.0090	1.0000	0.0090	0.8452	0.0090	1.0000	0.0090
<u>SIGNALING BY RECEPTOR TYROSINE KINASE</u>	0.0090	1.0000	0.0090	1.0000	0.0090	1.0000	0.0180

Figure 34: Results of overlap permutation tests. Adjusted p-values for overlap permutation tests between TSS +/- 200kb for each gene set enriched for genes up-regulated in insulinoma (rows) and groups of CREs identified by differential analysis, plus insulinoma-specific super-enhancers (columns).

By performing overlap permutation tests between CREs and regions surrounding up-regulated genes from the leading edge of enriched gene sets (fig.25), I found evidence pointing to a potential mechanism in insulinoma development. The question I was asking when performing these tests is 'do these two sets of regions overlap more than would be expected by chance?'. In this case the regions corresponded to different groups of CREs (the genomic coordinates of H3K27ac enriched regions) based on the classifications made in figure 29 and genes (TSS +/- 200kb) from the leading edge of each enriched gene set identified by GSEA (fig.25). An overlap higher than expected by chance suggests enhancer activation has a functional effect on gene expression. Permutation tests showed that enhancers activated during insulinoma development are enriched in regions surrounding up-regulated genes involved in nervous system development and function, signalling by receptor tyrosine kinases, Rho GTPases and VEGF (fig.34). Surprisingly though, there was no significant association between activated enhancers and upregulated chromatin modifying enzymes. This does not rule out a functional effect of insulinoma-specific enhancers on genes encoding chromatin modifiers, but it does suggest that an alternative mechanism, such as a mutation or copy number alteration,

is more likely to be the primary cause of the observed upregulation of these genes. This observation is consistent with findings from Wang et al, 2017 [42], who reported an accumulation of mutations in chromatin modifiers in an insulinoma cohort, particularly copy number gains such as that described in *KMT2C*. Super-enhancers follow the same pattern as regular enhancers, being significantly associated with the aforementioned signalling pathways and nervous system genes (with the possible exception of Rho GTPase signalling, although this result is very close to significance) and no association with chromatin modifiers.

6.3 Insulinoma chromatin architecture develops via de-repression and activation of CREs.

Upregulated genes (insulinoma vs unaffected human islets) within the chromatin modifier gene set include components of the SWI/SNF and MOZ/MORF complexes. The SWI/SNF complex (BRG1/SMARCA4 and associated factors in humans) has a key role in activation of enhancers via H3K27ac and is also required for p300 activity at enhancers [86]. Similarly the MOZ/MORF complex (KAT6A and KAT6B) along with bromodomain-containing co-activators have H3K27 acetyltransferase activity. In addition to the genes that increase H3K27ac, our data suggest a role for histone demethylases, such as KDM6B. This observation suggests a mechanism by which CREs that are normally inactive in human islets are de-repressed by removal of H3K27me3 and subsequently activated by acetylation at H3K27. To further investigate this mechanism I overlapped CREs (both up- and down-regulated in insulinoma) in groups from figure 29, plus CREs upregulated in non-functional PNETs vs unaffected human islets, with data from ChromHMM analysis of unaffected human islets. This analysis showed that more than half of the insulinoma-specific CREs are activated de novo. 10–15% of differentially active CREs in insulinoma and non-functional PNETs overlap polycomb-repressed CREs in human islets (fig.35a). To confirm this finding I then overlapped CREs from the same groups with regions enriched for the repressive H3K27me3 mark in healthy human islets. I found significant overlap between CREs activated in insulinoma

and those repressed in human islets, whilst stable and downregulated CREs did not overlap significantly (fig.35b). Taken together these analyses showed activation of CREs in regions that are actively repressed in human islets, highlighting a potentially relevant mechanism in insulinoma development.

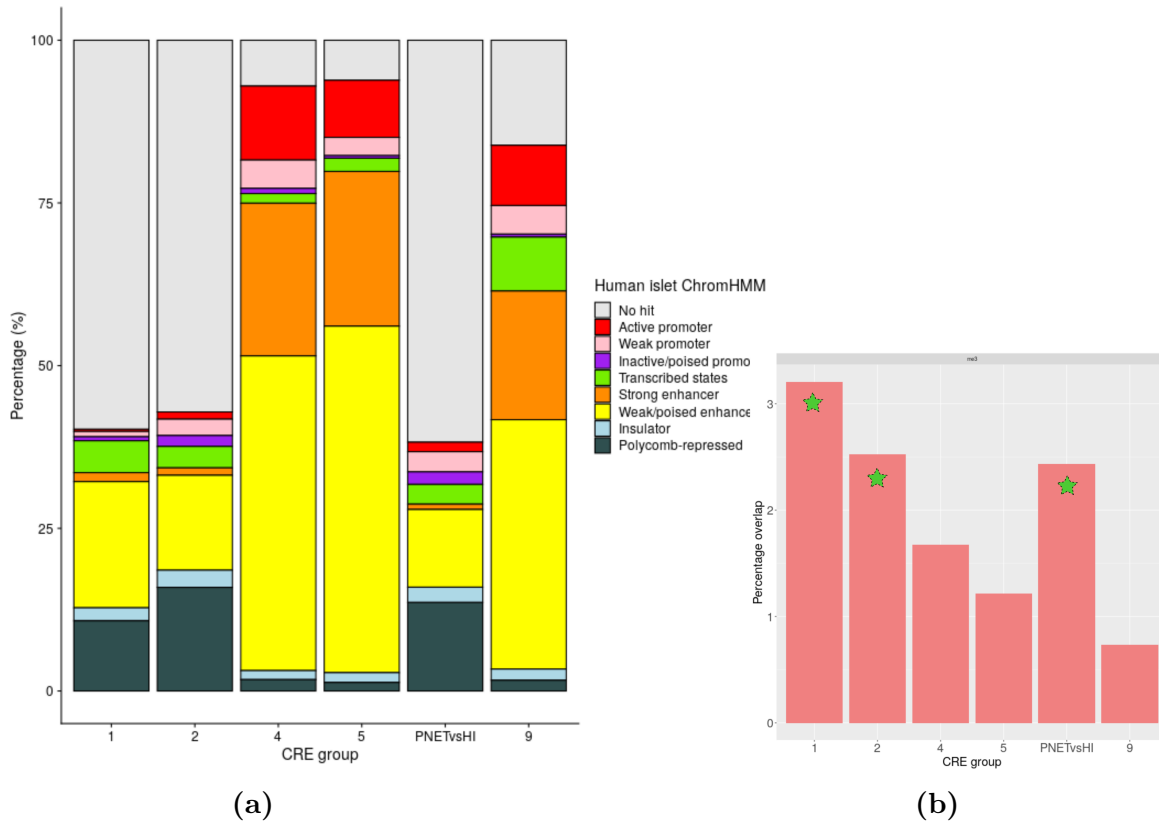


Figure 35: Overlap of CREs with H3K27me3 enriched regions and ChromHMM analysis from unaffected human islets. A Overlap of CRE regions from fig.29 (plus CREs upregulated in non-functional PNETs vs unaffected human islets) with regions classified by chromatin state (see key) as calculated by chromHMM analysis of a set of unaffected human islet samples. **B** Red bars show the percentage overlap between groups of CREs from fig.29 and regions enriched for H3K27me3 in unaffected human islets. Green stars indicate groups with a statistically significant overlap.

6.4 Gene-regulatory networks.

Combining the results of the previous analyses, I built putative insulinoma gene-regulatory networks. The first (fig. 36) includes significantly upregulated genes from the chromatin

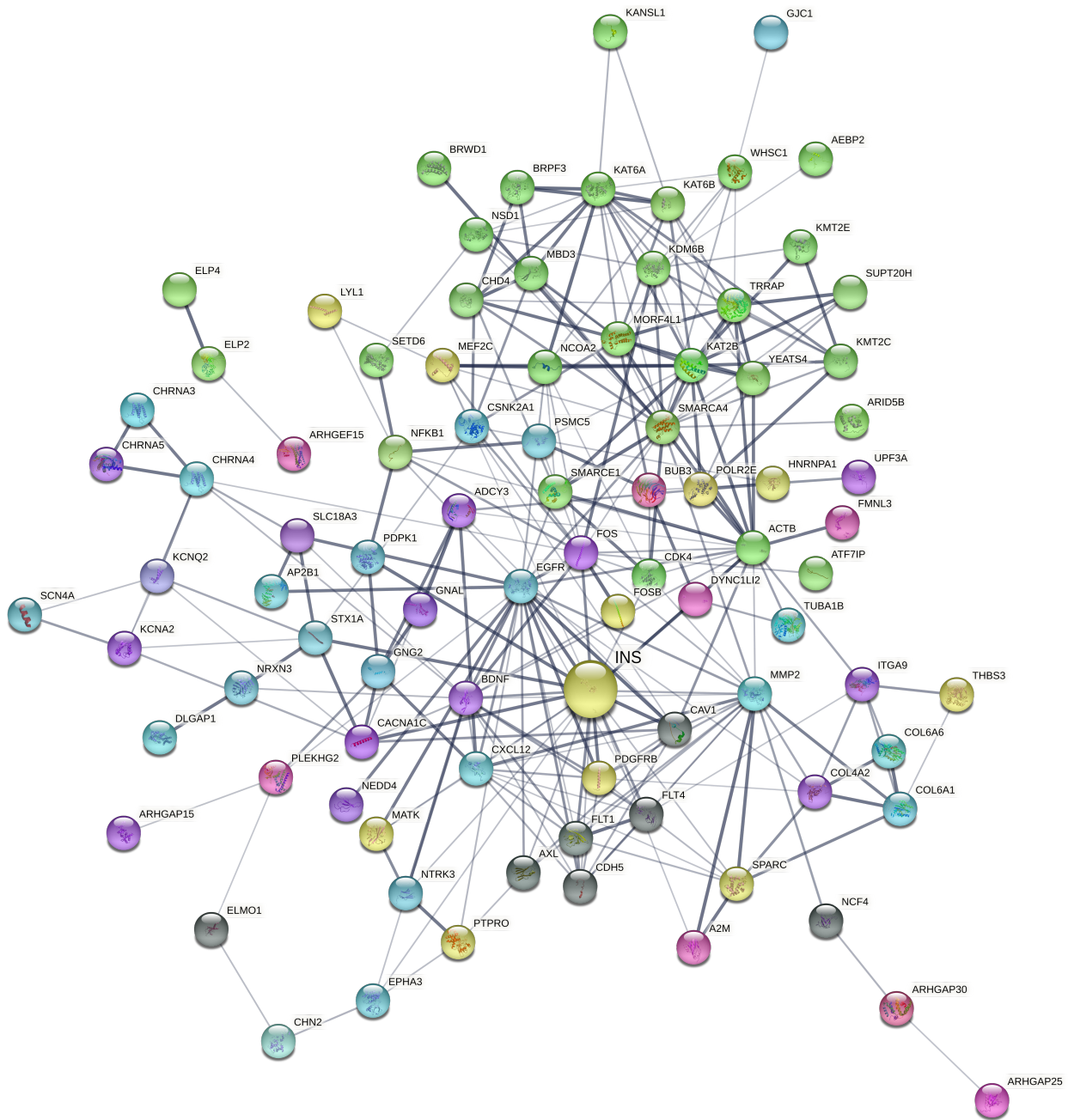


Figure 36: Stringdb interaction network 1. Interaction network featuring chromatin modifier genes with significantly ($FDR < 0.05$) upregulated (insulinoma vs unaffected human islets) protein-coding transcripts, plus genes from up-regulated gene sets with derepressed CREs within 200kb. Colours represent individual gene sets: Chromatin modifiers (green), Rho-GTPases (pink), VEGF (dark grey), RTK (yellow) and nervous system (blue) gene sets. Genes associated with super-specific CREs are coloured purple. Lines in-between gene names represent evidence of molecular interaction with line thickness corresponding to the confidence level assigned to each interaction. The marker for the INS gene is larger only to highlight its position within the network.

modifier gene set and leading edge genes from the other upregulated gene sets (RTK, Rho GTPase and VEGF signalling, and nervous system development) with associated (by proximity) derepressed CREs from group 8 (differentially active in insulinoma vs human islets or non-functional PNETs). I also included the insulin gene in the network. Whilst we would not expect repressed CREs to be associated with *INS* in unaffected human islets, several genes in the network feature strong interactions with *INS*. This suggests that an increase in insulin production may itself be partly responsible for expansion of the β -cell mass in insulinoma. This phenomenon was reported in murine β -cells [137] and in other tumour cells [138]. The group of CREs used includes some regions that are differentially active compared to both unaffected human islets and non-functional PNETs, but the majority are equally active in both functional and non-functional PNETs. This allowed us to build a much larger network than that which would have been possible using only CREs in group 1 of fig.29, the 'super-specific' CREs in insulinoma. However, by highlighting proteins in the network that are linked only with these super-specific' CREs we are able to describe a network with different elements, some of which are common to both types of PNET and some of which are specific to insulinoma.

Another interesting feature of this network is MEF2C (myocyte enhancer factor 2), a transcription factor that was a significant hit from motif analysis of group 8 CREs. MEF2C is activated by lysine acetyltransferase 2B (KAT2B), an upregulated chromatin modifier in the network. MEF2 TFs are regulators of neuronal survival [139] and MEF2C is a p38-binding protein. Phosphorylation of MEF2 by p38 has been linked to the protection of neurons from apoptosis [133]. One study investigating epigenetic programs in mouse cortical neurons found that MEF2C binds enhancer regulatory elements close to target genes involved in neuronal plasticity and calcium signalling [140]. If these functions are conserved in human cells then it is very likely that MEF2C is a key orchestrator of insulinoma development.

The majority of derepressed CREs linked (by proximity) to genes in this network are differentially active in both insulinoma and non-functional PNETs compared to unaffected human islets. However, there were interesting examples of genes with proximity to 'super-

specific' derepressed CREs (enriched for H3K27ac compared to both unaffected human islets and non-functional PNETs) including genes linked to cell proliferation and insulin secretion. *NEDD4*, an E3 ubiquitin-protein ligase, overexpression of which has been associated with the growth of breast tumours [74] and *FOS*, a component of the AP-1 TF, could play major roles in the proliferative ability of beta cells in insulinoma. Also linked to super-specific derepressed CREs was *CACNA1C* which encodes the voltage-gated Ca²⁺ channel Cav1.2. Cav1.2 knockout mice featured impaired insulin secretion [141]. Overexpression of this gene in insulinoma could explain (at least in part) the dysregulation of insulin secretion seen in insulinomas.

While this network encapsulates the key findings from this study, it represents a limited number of CREs that are 'super-specific' (differentially active in insulinoma vs both human islets and non-functional PNETs) to insulinoma. I have shown that the upregulation of chromatin modifiers is a key part of insulinoma development, and therefore it seems important to create a network incorporating all genes linked to these super-specific CREs (fig. 37). This network may include false positive components (genes that aren't regulated by super-specific CREs) but all proteins displayed are coded for by up-regulated genes (insulinoma vs human islet). CDC42 which, as described above, is an important mediator of cell proliferation and insulin granule mobilisation, appears as a major node. Several other proteins of the Rho GTPase family appear to interact with CDC42, but recent evidence suggests that CDC42, by itself is an important modulator of insulin expression [142]. The most differentially expressed gene (insulinoma vs human islets $\log_2FC = 13$) is *COL4A1* which encodes a subunit of type IV collagen. There are 15 super-specific, differentially active CREs within 200kb of this gene, strongly suggesting that insulinoma-specific increases in enhancer activity are responsible, at least in part, for this increase in its expression. Type IV collagen is abundant in the peripheral matrix of human islets, affecting the stiffness of the extra-cellular matrix. The 'islet-matrix relationship' has been shown to be a key determinant of β -cell function and survival in vitro [143].

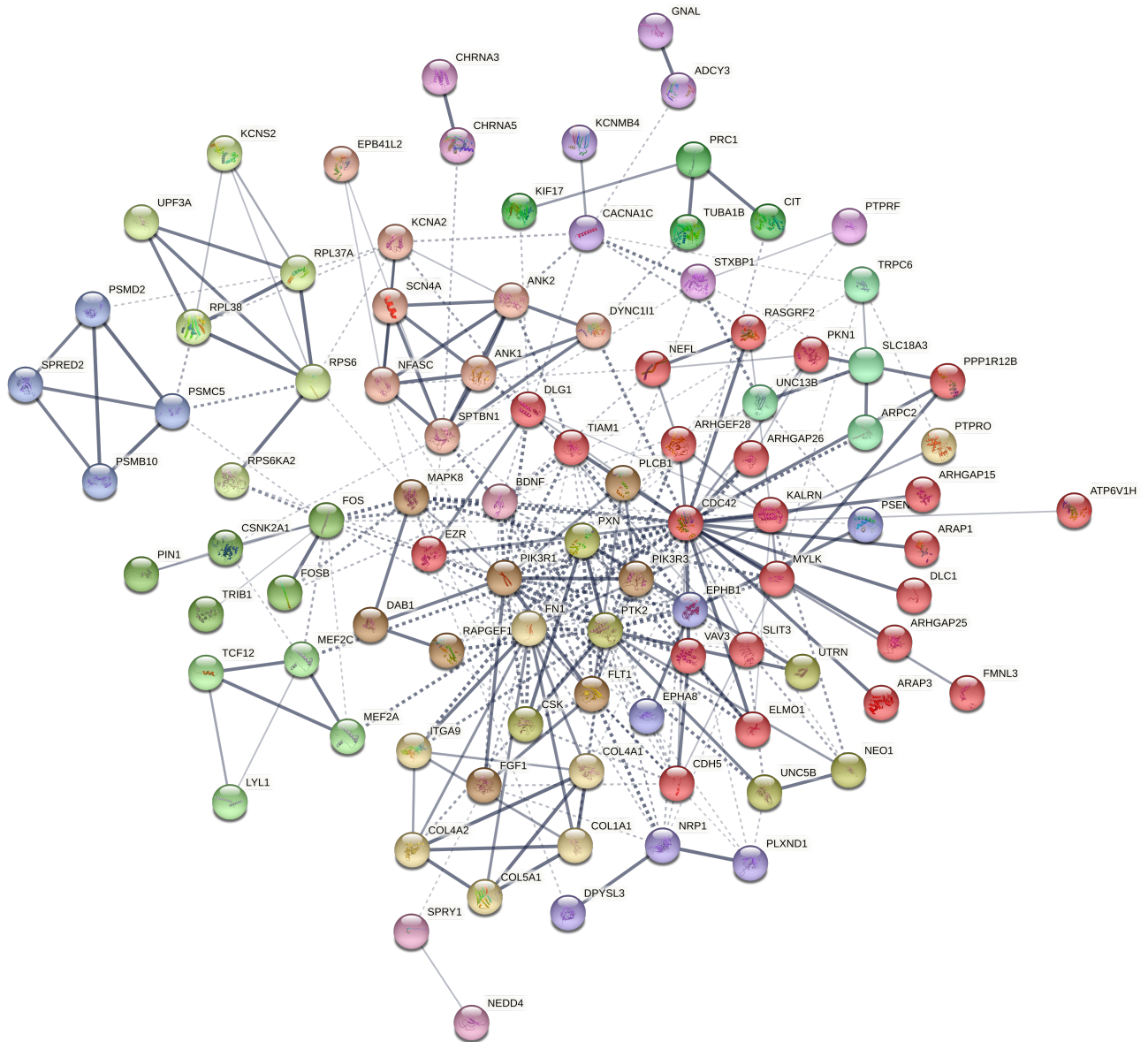


Figure 37: Stringdb interaction network 2. Interaction network featuring genes from the leading edge of all pathways identified by GSEA analysis with significantly ($FDR < 0.05$) upregulated (insulinoma vs unaffected human islets) protein-coding transcripts, with insulinoma super-specific CREs within 200kb. Colours are derived from Markov clustering. Lines in-between gene names represent evidence of molecular interaction with line thickness corresponding to the confidence level assigned to each interaction.

7 Discussion

Insulinoma is a very rare disease, and as such, opportunities for performing studies on large sample datasets have been limited compared to common diseases such as breast cancer or prostate cancer. Rare diseases account for a relatively small fraction of the healthcare burden, and therefore fewer resources are directed at investigating the mechanisms that cause their development. However, given the role that β -cells play in glucose homeostasis, the prevalence of diabetes, and the low proliferative ability of β -cells under normal physiological conditions, investigation of the mechanisms driving insulinoma development is a relevant area of research. Although several studies have investigated exonic mutations and gene expression in insulinoma development (including YY1 and MEN1), no previous study has attempted to map cis-regulatory elements in insulinoma samples, or to build insulinoma-specific gene regulatory networks. As such this study represents a significant step forward in insulinoma research.

Perhaps the most comprehensive prior study of the molecular mechanisms driving insulinoma was conducted by Wang et al [42], in which whole exome sequencing was used to identify variation, including SNVs, CNVs and indels in a cohort of 26 human insulinomas. Recurrent exonic variants identified in this cohort were significantly enriched in genes encoding epigenetic regulators. However, the authors did not extend the study to profile the enrichment of histone marks genome-wide. These findings influenced the direction of our study, as we focused efforts on profiling changes in the activity of cis-regulatory elements, by completing genome-wide profiles of H3K27ac enrichment in our insulinoma cohort. Another key finding from Wang et al was the enrichment of upregulated genes (insulinoma vs normal β -cells) for the repressive mark H3K27me3. Our findings shed further light on this mechanism, showing that this de-repression extends to associated enhancers, confirming the importance of the differential regulation of chromatin modifiers to the insulinoma phenotype.

Investigating rare diseases inevitably involves working with a limited number of samples. However, collaborations with labs in several countries have enabled the collection of a cohort of 20 insulinoma samples. The relatively small size of the tumours presented challenges in generating high quality genome-wide epigenomic and transcriptomic profiles. Here we have

developed a robust protocol for chromatin immunoprecipitation with small cell numbers, utilising the tn5 transposase enzyme (tagmentation), that could be used for future studies. I have demonstrated the high quality of the data generated, including genome-wide profiles of the histone mark H3K27ac and transcriptome profiles for insulinoma samples. I have also shown that this data is sufficient to approximate a comprehensive set of cis-regulatory elements in insulinoma and to perform robust statistical analysis. Initial analyses focused on describing the landscape of enhancer activity and gene expression in insulinoma and to compare and contrast it with that of a cohort of unaffected human islets.

The landscape of H3K27ac enriched chromatin in our insulinoma cohort consists of almost 178,000 regions, nearly 40,000 more than the number seen in unaffected human islets. Furthermore, differential analysis of ChIP-seq data showed 3 times as many gained active regions compared to those lost (32,453 vs 10,243). The key finding from this analysis therefore is that insulinoma development is characterised by an overall increase in chromatin activation. The heterogeneity of the control samples prevents us from drawing firm conclusions regarding down-regulated CREs or genes, as although β -cells are the most abundant cell type in human islets, it is possible that any increase in enrichment (of H3K27ac) relative to insulinomas could originate from other islet cell types. But this point should not detract from the significance of the results obtained and the conclusions drawn. We observe that CREs upregulated in insulinoma map to phylogenetically conserved sequences, a further indication of their potential function as gene regulatory elements. More than 90% of H3K27ac enriched regions in insulinoma are located distal to the nearest TSS, and the largest group of such regions is located more than 50kb from a TSS. Overlap of ChIP-seq data for H3K27ac with similar data for the repressive histone mark H3K27me3 and ChromHMM data enabled further characterisation of the landscape of insulinoma-specific cis-regulatory element activity. This analysis showed that activation of CREs that are actively repressed under normal physiological conditions is a major factor in insulinoma development. It also provided further evidence that the insulinoma gene regulatory landscape includes a large number of de novo (compared to unaffected human islets) regulatory elements.

Given that the function of enhancers is to positively regulate the expression of associated genes, we would expect to observe a corresponding increase in transcript abundance, and indeed there are more upregulated transcripts than downregulated. Differential analysis of transcript abundance identified putative factors involved in the dysregulation of insulin production and secretion in insulinoma. Gene set enrichment analysis showed that genes that code for chromatin modifiers are some of the most upregulated in insulinoma compared to unaffected human islets. This finding is further evidence that changes in the activity of gene regulatory elements are a major part of the molecular mechanisms responsible for the insulinoma phenotype. In addition we identified signalling pathways enriched for genes upregulated in insulinoma (vs human islets). I believe that the approach taken in this analysis (pre-ranked GSEA using only significant hits from two algorithms) was very robust statistically and gave a clear evaluation of the pathways and mechanisms most associated with the insulinoma phenotype.

Integration of ChIP-seq and RNA-seq data formed the basis of *in silico* investigations aimed at elucidating more precisely the aberrant pathways and networks involved in insulinoma development. In the absence of data from experiments that profile chromatin contacts, assigning function to distal CREs in terms of the genes they regulate is extremely challenging. We can assign limits to their activity based on established knowledge and evidence in the literature, but a complete understanding of enhancer action is thus far elusive. However, to begin to understand the relationship between CRE activity and gene expression, and to build putative gene regulatory networks, it's not necessary to evaluate every single element, but simply to look for relationships at a more general level. To do this some boundaries must be drawn, limits within which the majority of functional relationships operate. As described above, the key limit (distance from the nearest TSS) was arrived at by careful evaluation of the existing literature.

As described, up-regulated genes encoding chromatin modifiers are not significantly associated with activated enhancers within the genomic window investigated. Perhaps chromatin modifiers are regulated by fewer enhancers or by enhancers located more distally than the

200kb window around each TSS. However it also seems likely that another mechanism is involved in the activation of genes within this gene set. One working hypothesis is that gain-of-function mutations, including gain of CNV could be associated with the upregulation of chromatin modifier gene(s) in insulinoma. This hypothesis is consistent with the observation by Wang et al [42] that insulinomas are enriched for copy number gains in this class of genes. Taken together, data from this study and Wang et al suggest that a mutational mechanism affecting chromatin modifier genes is a strong candidate for driving the insulinoma phenotype. Finally, activation of polycomb-repressed regions in unaffected human islets would likely be facilitated by the upregulation of histone demethylases and acetyltransferases observed from the RNA-seq data. This therefore potentially represents a key feature of enhancer-driven development of insulinomas.

Incorporating insights from all analyses in this study I built the first insulinoma gene regulatory networks. The first is a snapshot of the most significant findings presented here and shows putative links between chromatin modifiers, transcription factors, signalling pathways and insulin. Whilst I believe that this network encapsulates some very important findings, it represents only a handful of genes that are associated with 'super-specific' (upregulated in insulinoma vs human islets and non-functional PNETs) enhancers, meaning that several of the tumour-driving factors in the network could be common to both insulinoma and non-functional β -cell tumours. The second network presented addresses this by including only proteins encoded by genes associated with super-specific enhancers. As mentioned, due to the challenges of assigning distal CREs to genes, the confidence with which we assign individual proteins to this network has some limitations. However, the strict inclusion criteria ($\log_2FC \geq 2$ transcript abundance insulinoma vs human islet) ensures that false positives are limited.

Parallel studies using whole-genome sequence data have shown that all insulinoma and human islet samples in this study show high purity levels (75-85%) and as β -cells are by far the largest cell type in human islets we can be confident that these results are not significantly limited by heterogeneity of the tissues studied. However, future analysis using single-cell and

chromatin contact technologies would provide the opportunity to validate these findings and provide insights into the cell sub-population that might comprise pancreatic islets affected by insulinoma tumours. Further studies including gain and loss of function assays in human β -cell lines could provide experimental confirmation for the role of candidate genes identified in the studies described in this thesis. Insulinoma is not just a rare tumour, it is one that develops from cells with extremely low proliferative ability (under normal physiological conditions). The proliferation of β -cells observed in insulinomas is essentially the opposite phenotype to that which characterises the very common disease, diabetes. So, in this case, focusing research efforts on a rare disease could also provide benefits in terms of advances in treatment approaches for common disease.

8 Conclusions

The results presented here represent important insights into the gene regulatory networks involved in the development of insulinomas. Through successful implementation of ChIP-seq and RNA-seq in a large cohort of tumour samples and unaffected pancreatic islets I have generated genome-wide profiles of gene expression and the activity of cis-regulatory elements in insulinomas. I have shown that insulinomas display widespread changes in gene expression and CRE activity compared to unaffected human islets and non-functional PNETs. I have also integrated these data sets to identify components of gene-regulatory networks as candidate drivers of cell proliferation and insulin dysregulation in insulinomas.

I have uncovered potentially key epigenetic mechanisms driving insulinoma development. Our results show that insulinoma development is associated with the upregulation of key chromatin modifiers, resulting in widespread activation of CREs, including a significant number of regions that are normally repressed in human islets. Activated CREs promote expression of genes involved in nervous system development, and signalling pathways involving Rho GTPases, receptor tyrosine kinases and VEGF. Overexpression of genes resulting from elevated CRE activity specific to insulinoma (inactive or stable in non-functional PNETs) may explain the dysregulation of insulin and subsequent hyperinsulinemic hypoglycemia seen in insulinoma patients.

9 Supplementary methods

9.1 ChIPmentation

Sonication:

keep sample on ice

- Prepare lysis buffer - aliquot 1ml and add 1ul PIC just before use. Use 600ul (can adjust depending on size of sample).
- Incubate for 5 mins with lysis buffer
- Break up tissue with douncer (after approximately 5 mins and 15 mins of incubation) - need to grind the tissue to try to break up the lumps complete lysis and homogenisation are crucial for efficient sonication. Total incubation = 20-25 mins.
- Each cycle of sonication = 5 x 1 min (40 sec ON & 20 sec OFF) Initial sonication for all samples = 4 cycles

NOTES: make sure water is cold add ice to cool water after each cycle but make sure there is no ice floating in the water afterwards

- After 4 cycles of sonication transfer 40ul (or 5 10%) of chromatin to another tube for QC and put the rest in freezer @ -80 or on ice in fridge (for up to 48 hours)
- Extract DNA from sample (as described below)
- Send DNA for Tapestation analysis

Successful sonication = around 80% fragments in the 100 500bp range

- Decide if further sonication is required usually one or two more cycles

Prepare IP

- Measure concentration of DNA from sonicated chromatin and (if required) pipette appropriate amount of chromatin to a new tube
- Spin sonicated chromatin at full speed for 15 mins @ 4C
- Transfer supernatant to another tube
- Add dilution buffer (and lysis buffer if required) to give a volume of 1ml and an SDS concentration of 0.4% (for H3K27ac antibody).
- Add 50ul of 10% BSA (100mg/ml = 10g in 100ml)
- Add 1.5ul antibody (depends on antibody) and incubate at 4C (cold room) O/N with rotation

Beads:

- Aliquot 20ul Protein A + G bead slurry (= 10ul beads + 10 solution)
- Wash with PBS: Add 1ml PBS, mix, place on magnet, wait for beads to move towards magnet, remove supernatant, repeat.
- Resuspend in 20ul PBS
- Add beads to chromatin/antibody sample and incubate at 4C (cold room) for 2 hours (max) with rotation

Washes:

NOTE : it is recommended that buffers are ice cold : pipette mixing should be gentle

- Place sample on magnetic rack (on ice), wait for beads to move towards magnet, remove supernatant.

NOTE: need to keep chromatin cold make a trough for magnet stand in the ice

- Remove sample from magnetic rack, add 1ml LOW SALT Immune Complex Wash Buffer, pipette up and down to mix and incubate for 4 mins at 4C (cold room) with rotation. Place sample on magnetic rack and remove the supernatant.
- Repeat for HIGH SALT Immune Complex Wash buffer and LiCl Immune Complex Wash buffer

ChIPmentation:

All washes use magnetic rack

- Wash beads twice with 10mM Tris-HCl (pH 8.0) (brief no incubation)

During second wash, before placing sample on magnetic rack, transfer sample to another tube this will decrease tagmentation of unspecific chromatin fragments
- Resuspend beads in 30ul tagmentation reaction buffer and add 1ul Tagment DNA Enzyme (from the Nextera DNA Sample Prep Kit Illumina)
- Incubate for 10 mins @ 37C with gentle agitation. During incubation (twice roughly every 2-3 mins) briefly remove sample from thermo block and flick mix to resuspend beads to maintain contact between beads and reaction mix.
- Wash twice with RIPA buffer (from version 2 of ChIPmentation protocol)
- Wash twice with Tris-EDTA transfer sample to another tube during second wash (as above)

Elution:

- Place the sample at room temperature and add 150ul Elution buffer to the beads.
- Pipette to mix and incubate with rotation for 15 mins at room temperature

- Place on magnetic rack, allow beads to settle to magnet and transfer the supernatant to another Eppendorf
- Re-elute by adding another 150ul Elution buffer to the beads, pipette to mix, separate beads using magnet.
- Combine both elutions in the same tube

DNA extraction:

First step for checking sonication add cold TE to make 300ul and spin at full speed @ 4C for 5 mins, transfer the supernatant to a new tube.

- Add 0.75ul (20mg/ml) OR 1.5ul (10mg/ml) RNase A, flick tube to mix and incubate at 37C for 30 mins with gentle agitation
- Add 4.5ul Proteinase K and 12ul 5M NaCl, flick tube to mix and incubate at 65C for 5 hrs or O/N
- Add 150ul TE to make 450ul
- Add 450ul Phenol Chloroform to a new tube and then add the sample, vortex for 30 secs to mix then spin at full speed for 5 mins @ room temperature
- Transfer the top phase to a new tube
- Add 450ul Chloroform, vortex for 30 secs to mix then spin at full speed for 5 mins @ room temperature
- Transfer the top phase to a new tube
- Add 1/10 volume 3M Na Acetate (pH 5.2), 2.5 x volume of ice-cold 100% EtOH (calculate ethanol volume after addition of Na Acetate) and 1ul glycogen (volume depends on concentration). Invert sample several times to mix and incubate at -20C for 4hrs (over night for ChIP DNA)

- Spin at full speed for 20 mins @ 4C and remove the supernatant
- Wash with 1ml ice-cold 70% EtOH, spin at full speed for 5 mins @ 4C and remove the supernatant
- Allow sample to air dry (to remove the excess EtOH)
- Re-suspend in appropriate volume of dH2O (10ul to check sonication, 22ul for CHIP)
- Incubate @ 37C for 10 mins

qPCR (to determine number of cycles for library amplification):

Prepare 2.5M index primer dilutions and make a mix as follows -

1 l 10x Sybr Green (final concentration in reaction mix = 1x) 0.6 l primer (index Ad1) 2.5M (0.15M in reaction mix) 0.6 l primer (chosen index) 2.5M (0.15M in reaction mix) 2 l library 5 l Taq mix (Next) 0.8 l dH2O

Light cycler program:

activation: 72C 5 mins

denaturation: 98C 30 s

then 24 cycles of: 98C 10 s

63C 30 s

72C 30 s

final elongation: 72C 1 min

- Determine optimum number of cycles for library amplification based on Cq value

Notes on Sybr green dilution:

stock @ 10,000x dilute 1:10 1,000x dilute 1:100 10x

Library amplification:

Prepare 25M index primer dilutions and make a mix as follows -

1.5 l primer (index Ad1) 25M (0.75M in reaction mix) 1.5 l primer (chosen index) 25M (0.75M in reaction mix) 20l library 25 l Taq mix (Next) 2 l dH₂O

PCR program:

activation: 72C 5 mins

denaturation: 98C 30 s

then n cycles (where n = C_q rounded up to the nearest whole cycle) of: 98C 10 s

63C 30 s

72C 30 s

final elongation: 72C 1 min

Bead cleanup - SPRI based size selection (0.7x)

For a 50ul reaction...

- Add 35 ul AmpPure beads to the sample
- Pipette up and down at least 10 times and incubate at RT for 1 minute
- Separate beads using a magnet and transfer the supernatant to a new tube
- Add 93.5 ul ((1.8 0.7) x 85) beads to the sample
- Pipette up and down at least 10 times and incubate at RT for 1 minute

- Separate beads using a magnet and discard the supernatant
- With beads still on the magnet add 180ul 85% Ethanol to the tube, incubate for 30 secs and remove
- To elute add at least 20ul dH₂O, pipette up and down to mix and incubate at RT for 1 minute
- Transfer the elute to a new tube Check concentration and enrichment of ChIP library

Buffers:

Lysis buffer: 2% Triton X-100 (10 mL 10% Triton X-100) 1% SDS (5 mL 10% SDS) 100 mM NaCl (1 mL 5M SDS) 10 mM Tris-HCl pH 8.0 (500ul 1M Tris-HCl) 1 mM EDTA (100ul 0.5M EDTA pH 8.0) 1x protease inhibitor cocktail (add just before adding the buffer to the pellet)

Dilution Buffer: 50 mM Hepes pH8.0 (2.5 mL 1M Hepes-KOH pH 8.0) 140 mM NaCl (1.4 mL 5M NaCl) 1 mM EDTA (100ul 0.5 M EDTA pH 8.0) 0.75% Triton X-100 (3.75mL 10% Triton X-100) 0.1% Na-deoxycholate (50mg Na-deoxycholate) 1x protease inhibitor cocktail (add just before adding the buffer to the pellet)

Low Salt Immune Complex Wash Buffer: 1% Triton X-100 (5 mL 10% Triton X-100) 150 mM NaCl (1.5 mL 5M NaCl) 20 mM Tris-HCl, pH 8.0 (1mL 1M Tris-HCl pH 8.0) 0.1% SDS (500ul 10% SDS) 2mM EDTA (200ul 0.5M EDTA pH 8.0)

High Salt Immune Complex Wash Buffer: 500 mM NaCl (5mL 5M NaCl) 1% Triton X-100 (5mL 10% Triton X-100) 20 mM Tris-HCl, pH 8.0 (1mL 1M Tris-HCl pH 8.0) 0.1% SDS (500ul 10% SDS) 2mM EDTA (200ul 0.5 M EDTA pH 8.0)

LiCl Immune Complex Wash Buffer: 0.25 M LiCl (2.5 mL 5M LiCl) 1% deoxycholate sodium (0.5g deoxycholate sodium) 10 mM Tris-HCl, pH 8.0 (500ul 1M Tris-HCl pH 8.0) 1% NP40 (5mL 10% NP40) 1 mM EDTA (100ul 0.5 M EDTA pH 8.0)

1xTE: 10 mM Tris-HCl, pH 8.0 (500ul 1M Tris-HCl pH 8.0) 1 mM EDTA (100ul 0.5 M EDTA pH 8.0)

RIPA buffer: 10 mM Tris HCl, pH 8.0 1 mM EDTA, pH 8.0 140 mM NaCl 1% Triton x100 0.1% SDS 0.1% Sodium Deoxycholate 1x protease inhibitor cocktail

Elution buffer: (5mL) Add water before (4mL) 1% SDS (500ul 10% SDS) 0.1M NaHCO₃ (500ul 1M NaHCO₃) Prepare at room temperature

9.2 Scripts

Scripts developed in the R programming language for the integrative analysis of ChIP-seq, ATAC-seq and RNA-seq data sets can be found at

<https://github.com/rnorris1260/hello-world>

References

- [1] Gabriela Da Silva Xavier. The Cells of the Islets of Langerhans. *J. Clin. Med.*, 7(3):54, 2018.
- [2] Donald J. Steiner, Abraham Kim, Kevin Miller, and Manami Hara. Pancreatic islet plasticity: Interspecies comparison of islet architecture and composition. *Islets*, 2(3):135–145, 2010.
- [3] Over Cabrera, Dora M. Berman, Norma S. Kenyon, et al. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proc. Natl. Acad. Sci. U. S. A.*, 103(7):2334–2339, 2006.
- [4] Constantin Ionescu-Tirgoviste, Paul A. Gagniuc, Elvira Gubceac, et al. A 3D map of the islet routes throughout the healthy human pancreas. *Sci. Rep.*, 5:1–14, 2015.
- [5] Fabrizio Thorel, Nicolas Damond, Simona Chera, et al. Normal glucagon signaling and β -cell function after near-total α -cell ablation in adult mice. *Diabetes*, 60(11):2872–2882, 2011.
- [6] P. Rorsman and E. Renström. Insulin granule dynamics in pancreatic beta cells. *Diabetologia*, 46(8):1029–1045, 2003.
- [7] Nigel Irwin and Peter R. Flatt. Enteroendocrine hormone mimetics for the treatment of obesity and diabetes. *Curr. Opin. Pharmacol.*, 13(6):989–995, 2013.
- [8] Natalie R. Johnston, Ryan K. Mitchell, Elizabeth Haythorne, et al. Beta Cell Hubs Dictate Pancreatic Islet Responses to Glucose. *Cell Metab.*, 24(3):389–401, 2016.
- [9] L. Charles Murtaugh. Pancreas and beta-cell development: From the actual to the possible. *Development*, 134(3):427–438, 2007.

- [10] Stuart B. Smith, Rosa Gasa, Hirotaka Watada, et al. Neurogenin3 and hepatic nuclear factor 1 cooperate in activating pancreatic expression of Pax4. *J. Biol. Chem.*, 278(40):38254–38259, 2003.
- [11] Sara Al-Khawaga, Bushra Memon, Alexandra E. Butler, et al. Pathways governing development of stem cell-derived pancreatic β cells: lessons from embryogenesis. *Biol. Rev.*, 93(1):364–389, 2018.
- [12] Idil I. Aigha and Essam M. Abdelalim. NKX6.1 transcription factor: a crucial regulator of pancreatic β cell development, identity, and proliferation. *Stem Cell Res. Ther.*, 11(1):1–14, 2020.
- [13] Haruhiko Akiyama, Jung Eun Kim, Kazuhisa Nakashima, et al. Osteochondroprogenitor cells are derived from Sox9 expressing precursors. *Proc. Natl. Acad. Sci. U. S. A.*, 102(41):14665–14670, 2005.
- [14] F. C. Lynn, S. B. Smith, M. E. Wilson, et al. Sox9 coordinates a transcriptional network in pancreatic progenitor cells. *Proc. Natl. Acad. Sci. U. S. A.*, 104(25):10500–10505, 2007.
- [15] Blair K. Gage, Travis D. Webber, and Timothy J. Kieffer. Initial cell seeding density influences pancreatic endocrine development during in vitro differentiation of human embryonic stem cells. *PLoS One*, 8(12):1–13, 2013.
- [16] Manjula Alejandro López-Juárez, Sara Morales-Lázaro, Roberto Sánchez-Sánchez, Andrew J. Morris Sunkara, Hilda Lomelí, Iván Velasco, and Diana Escalante-Alcalde. Pdx1 maintains β -cell identity and function by repressing an α -cell program. *Cell Metab.*, 100(2):130–134, 2012.
- [17] Atsushi Kubo, Robert Stull, Mitsuaki Takeuchi, et al. Pdx1 and Ngn3 overexpression enhances pancreatic differentiation of mouse ES cell-derived endoderm population. *PLoS One*, 6(9), 2011.

- [18] N. Sander, L. Sussel, J. Conners, et al. Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of β -cell formation in the pancreas. *Development*, 127(24):5533–5540, 2000.
- [19] Rachel E. Jennings, Andrew A. Berry, Rebecca Kirkwood-Wilson, et al. Development of the human pancreas from foregut to endocrine commitment. *Diabetes*, 62(10):3514–3522, 2013.
- [20] Shelley B. Nelson, Ashleigh E. Schaffer, and Maike Sander. The transcription factors Nkx6.1 and Nkx6.2 possess equivalent activities in promoting beta-cell fate specification in Pdx1+ pancreatic progenitor cells. *Development*, 134(13):2491–2500, 2007.
- [21] Sara S. Roscioni, Adriana Migliorini, Moritz Gegg, and Heiko Lickert. Impact of islet architecture on β -cell heterogeneity, plasticity and function. *Nat. Rev. Endocrinol.*, 12(12):695–709, 2016.
- [22] Junfeng Wang, Lynda Elghazi, Susan E. Parker, et al. The concerted activities of Pax4 and Nkx2.2 are essential to initiate pancreatic β -cell differentiation. *Dev. Biol.*, 266(1):178–189, 2004.
- [23] Monica Courtney, Elisabet Gjernes, Noémie Druelle, et al. The Inactivation of Arx in Pancreatic α -Cells Triggers Their Neogenesis and Conversion into Functional β -Like Cells. *PLoS Genet.*, 9(10):1–18, 2013.
- [24] Rachel E. Jennings, Andrew A. Berry, James P. Strutt, David T. Gerrard, and Neil A. Hanley. Human pancreas development. *Dev.*, 142(18):3126–3137, 2015.
- [25] Wataru Nishimura, Satoru Takahashi, and Kazuki Yasuda. MafA is critical for maintenance of the mature beta cell phenotype in mice. *Diabetologia*, 58(3):566–574, 2015.
- [26] Yan Hang, Tsunehiko Yamamoto, Richard K.P. Benninger, et al. The MafA transcription factor becomes essential to islet β -cells soon after birth. *Diabetes*, 63(6):1994–2005, 2014.

- [27] Vikash Chandra, Olivier Albagli-Curiel, Benoit Hastoy, et al. RFX6 Regulates Insulin Secretion by Modulating Ca²⁺ Homeostasis in Human β Cells. *Cell Rep.*, 9(6):2206–2218, 2014.
- [28] Ji-Ann Lee Kelsey C. Martin Mhatre V. Ho. Pancreatic β -Cell Dedifferentiation As Mechanism Of Diabetic β -Cell Failure. *Bone*, 23(1):1–7, 2012.
- [29] Lorenzo Pasquali, Kyle J. Gaulton, Santiago A. Rodríguez-Seguí, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.*, 46(2):136–143, 2014.
- [30] M. German, S. Ashcroft, K. Docherty, et al. The insulin gene promoter: A simplified nomenclature. *Diabetes*, 44(8):1002–1004, 1995.
- [31] B. Peers, J. Leonard, S. Sharma, G. Teitelman, and M. R. Montminy. Insulin expression in pancreatic islet cells relies on cooperative interactions between the helix loop helix factor E47 and the homeobox factor STF-1. *Mol. Endocrinol.*, 8(12):1798–1806, 1994.
- [32] M. Maral Dias, Vera Junn, Eunsung Mouradian. Preferential reduction of β cells derived from Pax6- MafB- pathway in MafB deficient mice. *Bone*, 23(1):1–7, 2008.
- [33] Vincent Poitout, Derek Hagman, Roland Stein, et al. Regulation of the insulin gene by glucose and fatty acids. *J. Nutr.*, 136(4):873–876, 2006.
- [34] Maike Sander, Steven C. Griffen, Juemin Huang, and Michael S. German. A novel glucose-responsive element in the human insulin gene functions uniquely in primary cultured islets. *Proc. Natl. Acad. Sci. U. S. A.*, 95(20):11572–11577, 1998.
- [35] A. Fornoni, A. Pileggi, R. D. Molano, et al. Inhibition of c-jun N terminal kinase (JNK) improves functional beta cell mass in human islets and leads to AKT and glycogen synthase kinase-3 (GSK-3) phosphorylation. *Diabetologia*, 51(2):298–308, 2008.
- [36] Tessy Iype, Joshua Francis, James C. Garmey, et al. Mechanism of insulin gene regulation by the pancreatic transcription factor Pdx-1: Application of pre-mRNA analysis

- and chromatin immunoprecipitation to assess formation of functional transcriptional complexes. *J. Biol. Chem.*, 280(17):16798–16807, 2005.
- [37] Carina Ammala, Olof Larsson, Per-olof Berggren, et al. Inositol trisphosphate-dependent periodic activation of a Ca^{2+} -activated K^{+} conductance in glucose-stimulated pancreatic β -cells. 353(October):849–852, 1991.
- [38] Piero Marchetti, Marco Bugliani, Vincenzo De Tata, Mara Suleiman, and Lorella Marselli. Pancreatic Beta Cell Identity in Humans and the Role of Type 2 Diabetes. *Front. Cell Dev. Biol.*, 5(May):55, 2017.
- [39] Susan Bonner-Weir, Cristina Aguayo-Mazzucato, and Gordon C. Weir. Dynamic development of the pancreas from birth to adulthood. *Ups. J. Med. Sci.*, 121(2):155–158, 2016.
- [40] Rohit N. Kulkarni, Ernesto Bernal Mizrachi, Adolfo Garcia Ocana, and Andrew F. Stewart. Human β -cell proliferation and intracellular signaling: Driving in the dark without a road map. *Diabetes*, 61(9):2205–2213, 2012.
- [41] Dana Avrahami, Changhong Li, Jia Zhang, et al. Aging-Dependent Demethylation of Regulatory Elements Correlates with Chromatin State and Improved β Cell Function. 22(4):619–632, 2016.
- [42] Huan Wang, Aaron Bender, Peng Wang, et al. Insights into beta cell regeneration for diabetes via integration of molecular landscapes in human insulinomas. *Nat. Commun.*, 8(1):767, 2017.
- [43] Peng Wang, Juan-Carlos Alvarez-Perez, Dan P Felsenfeld, et al. Induction of human pancreatic beta cell replication by inhibitors of dual specificity tyrosine regulated kinase Corresponding. *Nat Med*, 21(4):383–388, 2015.
- [44] Ercument Dirice, Deepika Walpita, Amedeo Vetere, et al. Inhibition of DYRK1A stimulates human β -cell proliferation. *Diabetes*, 65(6):1660–1671, 2016.

- [45] Jeremy J. Heit, Åsa A. Apelqvist, Xueying Gu, et al. Calcineurin/NFAT signalling regulates pancreatic β -cell growth and function. *Nature*, 443(7109):345–349, 2006.
- [46] Esra Karakose, Courtney Ackeifi, Peng Wang, and Andrew F. Stewart. Advances in drug discovery for human beta cell regeneration. *Diabetologia*, 61(8):1693–1699, 2018.
- [47] Giampaolo Perri, Laura R. Prakash, and Matthew H.G. Katz. Pancreatic neuroendocrine tumors. *Curr. Opin. Gastroenterol.*, 35(5):468–477, 2019.
- [48] Dermot O’Toole, Reza Kianmanesh, and Martyn Caplin. ENETS 2016 consensus guidelines for the management of patients with digestive neuroendocrine tumors: An update. *Neuroendocrinology*, 103(2):117–118, 2016.
- [49] G. V. Gill, O. Rauf, and I. A. MacFarlane. Diazoxide treatment for insulinoma: A national UK survey. *Postgrad. Med. J.*, 73(864):640–641, 1997.
- [50] Ruben Mujica-Mota, Jo Varley-Campbell, Irina Tikhonova, et al. Everolimus, lutetium-177 DOTATATE and sunitinib for advanced, unresectable or metastatic neuroendocrine tumours with disease progression: A systematic review and cost-effectiveness analysis. *Health Technol. Assess. (Rockv)*., 22(49):1–325, 2018.
- [51] Melanie E. Royce and Diaa Osman. Everolimus in the treatment of metastatic breast cancer. *Breast Cancer Basic Clin. Res.*, 9:73–79, 2015.
- [52] S. Fontanière, J. Tost, A. Wierinckx, et al. Gene expression profiling in insulinomas of Men1 beta-cell mutant mice reveals early genetic and epigenetic events involved in pancreatic beta-cell tumorigenesis. *Endocr. Relat. Cancer*, 13(4):1223–1236, 2006.
- [53] Aldo Scarpa, David K. Chang, Katia Nones, et al. Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature*, 2017.
- [54] Satyajit K. Karnik, Christina M. Hughes, Xueying Gu, et al. Menin regulates pancreatic islet growth by promoting histone methylation and expression of genes encoding p27Kip1 and p18INK4c. *Proc. Natl. Acad. Sci. U. S. A.*, 102(41):14659–14664, 2005.

- [55] M Kyle Cromer, Murim Choi, Carol Nelson-Williams, et al. Neomorphic effects of recurrent somatic mutations in Yin Yang 1 in insulin-producing adenomas. *Proc. Natl. Acad. Sci. U. S. A.*, 112(13):4062–4067, 2015.
- [56] Yanan Cao, Zhibo Gao, Lin Li, et al. Whole exome sequencing of insulinoma reveals recurrent T372R mutations in YY1. *Nat. Commun.*, 4:2810, 2013.
- [57] Hristo B. Houbaviy, Anny Usheva, Thomas Shenk, and Stephen K. Burley. Cocystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24):13577–13582, 1996.
- [58] Ada L. Olins and Donald E. Olins. Spheroid chromatin units (v bodies). *Science (80-.)*, 183(4122):330–332, 1974.
- [59] J. L. Workman and R. E. Kingston. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.*, 67:545–579, 1998.
- [60] Victor Chukwudi Osamor, Shalom N. Chinedu, Dominic E. Azuh, Emeka Joshua Iweala, and Olubanke Olujoke Ogunlana. The interplay of post-translational modification and gene therapy. *Drug Des. Devel. Ther.*, 10(February):861–871, 2016.
- [61] Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, 2009.
- [62] Kevin Struhl and Eran Segal. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, 20(3):267–273, 2013.
- [63] Bing Li, Michael Carey, and Jerry L. Workman. The Role of Chromatin during Transcription. *Cell*, 128(4):707–719, 2007.
- [64] Peter B. Becker and Jerry L. Workman. Nucleosome remodeling and epigenetics. *Cold Spring Harb. Perspect. Biol.*, 5(9), 2013.
- [65] Fernanda Moraes and Andréa Góes. A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. *Biochem. Mol. Biol. Educ.*, 44(3):215–

223, 2016.

- [66] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [67] I. M. Adcock, K. Ito, and G. Caramori. Transcription Factors: Overview. *Encycl. Respir. Med. Four-Volume Set*, pages 243–251, 2006.
- [68] Jennifer L. Plank and Ann Dean. Enhancer function: Mechanistic and genome-wide insights come together. *Mol. Cell*, 55(1):5–14, 2014.
- [69] Amartya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012.
- [70] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014.
- [71] Eileen E.M. Furlong and Michael Levine. Developmental enhancers and chromosome topology. *Science (80-.)*, 361(6409):1341–1345, 2018.
- [72] Adam G. West, Miklos Gaszner, and Gary Felsenfeld. Insulators: Many functions, many mechanisms. *Genes Dev.*, 16(3):271–288, 2002.
- [73] M. P. Creighton, A. W. Cheng, G. G. Welstead, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.*, 107(50):21931–21936, 2010.
- [74] Lingfeng Wan, Tao Liu, Zhipeng Hong, et al. NEDD4 expression is associated with breast cancer progression and is predictive of a poor prognosis. *Breast Cancer Res.*, 21(1):1–16, 2019.
- [75] Francesca de Santa, Iros Barozzi, Flore Mietton, et al. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol.*, 8(5), 2010.

- [76] Warren A Whyte, David A Orlando, Denes Hnisz, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, 2013.
- [77] Sebastian Pott and Jason D. Lieb. What are super-enhancers? *Nat. Genet.*, 47(1):8–12, 2015.
- [78] Denes Hnisz, Brian J. Abraham, Tong Ihn Lee, et al. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934, 2013.
- [79] Kyle J. Gaulton, Takao Nammou, Lorenzo Pasquali, et al. A map of open chromatin in human pancreatic islets. *Nat. Genet.*, 42(3):255–259, 2010.
- [80] Stephen C J Parker, Michael L Stitzel, D Leland Taylor, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, 110(44):17921–17926, 2013.
- [81] Jakob Lovén, Heather A. Hoke, Charles Y. Lin, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334, 2013.
- [82] Aaron D. Goldberg, Laura A. Banaszynski, Kyung Min Noh, et al. Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions. *Cell*, 140(5):678–691, 2010.
- [83] Zhaoyu Li, Paul Gadue, Kaifu Chen, et al. Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell*, 151(7):1608–1616, 2012.
- [84] Michelle L.T. Nguyen, Sarah A. Jones, Julia E. Prier, and Brendan E. Russ. Transcriptional enhancers in the regulation of T cell differentiation. *Front. Immunol.*, 6(SEP), 2015.
- [85] Artem Barski, Suresh Cuddapah, Kairong Cui, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

- [86] Burak H. Alver, Kimberly H. Kim, Ping Lu, et al. The SWI/SNF chromatin remodeling complex is required for maintenance of lineage specific enhancers. *Nat. Commun.*, 8:1–10, 2017.
- [87] Edoardo Missiaglia, Ekaterini Blaveri, Benoit Terris, et al. Analysis of gene expression in cancer cell lines identifies candidate markers for pancreatic tumorigenesis and metastasis. *Int. J. Cancer*, 112(1):100–112, 2004.
- [88] Hector L. Franco, Anusha Nagari, Venkat S. Malladi, et al. Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. *Genome Res.*, 28(2):159–170, 2018.
- [89] Paul A. Northcott, Catherine Lee, Thomas Zichner, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, 511(7510):428–434, 2014.
- [90] Ken J Kron, Swneke D Bailey, and Mathieu Lupien. Enhancer alterations in cancer : a source for a cell identity crisis. pages 1–12, 2014.
- [91] Danielle Welter, Jacqueline MacArthur, Joannella Morales, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, 42(D1):1001–1006, 2014.
- [92] Inderpreet Sur and Jussi Taipale. The role of enhancers in cancer, 2016.
- [93] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.
- [94] M Ryan Corces, Jeffrey M Granja, Shadi Shams, et al. The chromatin accessibility landscape of primary human cancers. 1898, 2018.
- [95] Christian Schmidl, André F Rendeiro, Nathan C Sheffield, and Christoph Bock. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods*, 12(10):963–5, 2015.

- [96] Ignasi Morán, Ildem Akerman, Martijn Van De Bunt, et al. Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab.*, 16(4):435–448, 2012.
- [97] Sergio Gonzalez-Duque, Marie Eliane Azoury, Maikel L. Colli, et al. Conventional and Neo-antigenic Peptides Presented by β Cells Are Targeted by Circulating Naïve CD8+ T Cells in Type 1 Diabetic and Healthy Donors. *Cell Metab.*, 28(6):946–960.e6, 2018.
- [98] Paloma Cejas, Yotam Drier, Koen M.A. Dreijerink, et al. Enhancer signatures stratify and predict outcomes of non-functional pancreatic neuroendocrine tumors. *Nat. Med.*, 25(8):1264–1269, 2019.
- [99] Howard Chang Jason Buenrostro, Beijing Wu and William Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Physiol. Behav.*, 176(1):139–148, 2016.
- [100] Andreas Schroeder, Odilo Mueller, Susanne Stocker, et al. The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.*, 7:1–14, 2006.
- [101] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, 2012.
- [102] Heng Li, Bob Handsaker, Alec Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [103] Yong Zhang, Tao Liu, Clifford A. Meyer, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9), 2008.
- [104] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.
- [105] R RCoreTeam. *R: a Language and Environment for Statistical Computing.*, volume 2. 2008.

- [106] Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.
- [107] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419, 2017.
- [108] Charlotte Sonesson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research*, 4:1–23, 2016.
- [109] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.
- [110] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One*, 12(12):1–18, 2017.
- [111] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.
- [112] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, 2008.
- [113] Davide Risso, John Ngai, Terence P. Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32(9):896–902, 2014.
- [114] Mark D. Robinson and Gordon K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.

- [115] Zdravko I. Botev and Dirk P. Kroese. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodol. Comput. Appl. Probab.*, 10(4):471–505, 2008.
- [116] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, 2005.
- [117] Nastaran Heidari, Douglas H Phanstiel, Chao He, et al. Title: Genome-wide map of regulatory interactions in the human genome Running Title: Global map of human regulatory interactions. *Cold Spring Harb. Lab. Press*, pages 1905–1917, 2015.
- [118] Lindsey E. Montefiori, Debora R. Sobreira, Noboru J. Sakabe, et al. A promoter interaction map for cardiovascular disease genetics. *Elife*, 7:1–35, 2018.
- [119] Bernat Gel, Anna Díez-Villanueva, Eduard Serra, et al. RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, 32(2):289–291, 2016.
- [120] Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9):1813–1831, 2012.
- [121] Chunyan Gu, Gretchen H. Stein, Ning Pan, et al. Pancreatic β Cells Require NeuroD to Achieve and Maintain Functional Maturity. *Cell Metab.*, 11(4):298–310, 2010.
- [122] Isabella Artner, Bruno Bianchi, Jeffrey C. Raum, et al. MafB is required for islet β cell maturation. *Proc. Natl. Acad. Sci. U. S. A.*, 104(10):3853–3858, 2007.
- [123] Yunshin Jung, Ruyi Zhou, Toshiki Kato, et al. Isl1 β Overexpression With Key b Cell Transcription Factors Enhances Glucose-Responsive Hepatic Insulin Production and Secretion. *Endocrinology*, 159(2):869–882, 2018.

- [124] Tanya Vavouri and Ben Lehner. Conserved noncoding elements and the evolution of animal body plans. *BioEssays*, 31(7):727–735, 2009.
- [125] D. Hu, X. Gao, M. A. Morgan, et al. The MLL3/MLL4 Branches of the COMPASS Family Function as Major Histone H3K4 Monomethylases at Enhancers. *Mol. Cell. Biol.*, 33(23):4745–4754, 2013.
- [126] Ji Eun Lee, Chaochen Wang, Shiliyang Xu, et al. H3K4 mono- And di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *Elife*, 2013(2):1–25, 2013.
- [127] Xin Shao, Ning Lv, Jie Liao, et al. Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Med. Genet.*, 20(1):1–14, 2019.
- [128] Qi Yuan Huang, Xing Ning Lai, Xian Ling Qian, et al. Cdc42: A novel regulator of insulin secretion and diabetes-associated diseases. *Int. J. Mol. Sci.*, 20(1), 2019.
- [129] Päivi J. Miettinen, Mari Anne Huotari, Tarja Koivisto, et al. Impaired migration and delayed differentiation of pancreatic islet cells in mice lacking EGF-receptors. *Development*, 127(12):2617–2627, 2000.
- [130] Päivi J. Miettinen, Jarkko Ustinov, Päivi Ormio, et al. Downregulation of EGF receptor signaling in pancreatic islets causes diabetes due to impaired postnatal β -cell growth. *Diabetes*, 55(12):3299–3308, 2006.
- [131] Robert Nechanitzky, Duygu Akbas, Stefanie Scherer, et al. Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. *Nat. Immunol.*, 14(8):867–875, 2013.
- [132] Nora Fernandez-Jimenez, Athena Sklias, Szilvia Ecsedi, et al. Lowly methylated region analysis identifies EBF1 as a potential epigenetic modifier in breast cancer. *Epigenetics*, 12(11):964–972, 2017.

- [133] Shu Ichi Okamoto, Dimitri Krainc, Katerina Sherman, and Stuart A. Lipton. Anti-apoptotic role of the p38 mitogen-activated protein kinase-myocyte enhancer factor 2 transcription factor pathway during neuronal differentiation. *Proc. Natl. Acad. Sci. U. S. A.*, 97(13):7561–7566, 2000.
- [134] E. S.P. Reddy, V. N. Rao, and T. S. Papas. The *erg* gene: A human gene related to the *ets* oncogene. *Proc. Natl. Acad. Sci. U. S. A.*, 84(17):6131–6135, 1987.
- [135] P. Adamo and M. R. Ladomery. The oncogene *ERG*: A key factor in prostate cancer. *Oncogene*, 35(4):403–414, 2016.
- [136] Amartya Sanyal and Job Dekker Bryan Lajoie, Gaurav Jain. The long-range interaction landscape of gene promoters. *Physiol. Behav.*, 176(10):139–148, 2017.
- [137] Jennifer L. Beith, Emilyn U. Alejandro, and James D. Johnson. Insulin stimulates primary β -cell proliferation via Raf-1 kinase. *Endocrinology*, 149(5):2251–2260, 2008.
- [138] Chi Cheng Lu, Pei Yi Chu, Shih Min Hsia, et al. Insulin induction instigates cell proliferation and metastasis in human colorectal cancer cells. *Int. J. Oncol.*, 50(2):736–744, 2017.
- [139] Timothy A. McKinsey, Chun Li Zhang, and Eric N. Olson. MEF2: A calcium-dependent regulator of cell division, differentiation and death. *Trends Biochem. Sci.*, 27(1):40–47, 2002.
- [140] Qi Ma and Francesca Telese. Genome-wide epigenetic analysis of MEF2A and MEF2C transcription factors in mouse cortical neurons. *Commun. Integr. Biol.*, 8(6):1–5, 2015.
- [141] Verena Schulla, Erik Renström, Robert Feil, et al. Impaired insulin secretion and glucose tolerance in β cell-selective Cav1.2 Ca²⁺ channel null mice. *EMBO J.*, 22(15):3844–3854, 2003.
- [142] Xiang Qin He, Ning Wang, Juan Juan Zhao, et al. Specific deletion of CDC42 in pancreatic β cells attenuates glucose-induced insulin expression and secretion in mice.

Mol. Cell. Endocrinol., 518, 2020.

- [143] L. Alberto Llacua, Marijke M. Faas, and Paul de Vos. Extracellular matrix molecules and their potential contribution to the function of transplanted pancreatic islets. *Diabetologia*, 61(6):1261–1272, 2018.