

Imperial College London
Department of Medicine

**Genetic and epigenetic regulation of gene
expression in pancreatic islets**

Nikolina Nakić

Thesis submitted to Imperial College London for the degree of
Doctor of Philosophy

February 29, 2016

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

T2D is a complex disease with evidence of a strong genetic basis. Studies of the human pancreatic islets have provided valuable insight into the islet regulatory landscape and identified enrichment of T2D-associated variants in islet enhancers. The relationship between *cis*-regulatory variation and changes in gene expression, however, remains unclear. This question is challenging to address in human, as it calls for a systematic analysis of *cis*-regulatory variation and the impact it has on gene expression in the native genomic context. While a mouse model can be used for this purpose, the regulatory landscape of the mouse pancreatic islets has been poorly characterized. This thesis addresses the questions of genetic and epigenetic regulation of gene expression in pancreatic islets in two parts. First, a genome-wide map of several histone modifications and transcription factor binding sites is created for the mouse pancreatic islet. This enabled identification of the mouse islet regulatory elements and characterization of enhancer clustering, transcription factor occupancy and conservation. A systematic comparison of enhancer clustering between mouse and human islets identified species-common and species-specific subsets of the islet regulatory program, each associated to a distinct biological function of the pancreatic islet. Second, a hybrid mouse was used as a model where naturally occurring genetic variation drives changes in gene expression. High-density mapping of allelic regulatory activity, chromatin accessibility and transcription provided insight into the properties of both genes subject to *cis*-regulatory variation and the regulatory elements driving the change in expression. As a result, tissue-specific genes associated to clustered enhancers were shown to be most influenced by *cis*-regulatory variation. Additionally, enhancer clustering emerged as the dominant property of regulatory elements associated to changes in gene expression. Overall, this thesis advanced our knowledge of the mouse islet regulatory landscape and provided novel insight into the properties of functional *cis*-regulatory variation.

Acknowledgements

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

Thanks to all the members of the Marie Curie “Biology of Liver and Pancreatic Development and Disease” (BOLD) Initial Training Network for supporting me through the early stages of my studentship and enabling me to interact with an amazingly talented group of scientists.

Thanks to all current and past members of the Ferrer lab for their company on this journey, especially to Vane and Miguel Angel who helped making my work possible. Thanks to Joris, Myriam and Xavi for introducing me to the world of mice; to Ildem for the tea breaks and molecular biology wisdom; to Lorenzo, Ignasi, Loris and Joan for the bioinformatic discussions and Natalia and Mar for the moral support. My profound gratitude goes to Claire, Irene and Meritxell for all the useful comments and proofreading the thesis. To them and all the other member of the lab, in Barcelona and in London - it has been an honour to be able to work with such talented scientist and wonderful people.

My deepest thanks goes to my parents and sisters for all the words of encouragement over the years.

Finally, thanks to my partner Hrvoje, for the endless patience and unconditional love and support. Thank you.

Declaration of Originality

This dissertation, in any form, has not been submitted for any other degree, diploma or qualification. It does not contain more than 100,000 words, excluding the tables, figures, bibliography and appendices.

I hereby declare that all work described in this dissertation has been carried on by myself and that all information derived from the work of others has been appropriately credited and referenced.

I would like to acknowledge the following contributions to my work.

Vanessa Grau Martinez contributed to the project by taking care of the animal colonies and assisting in islet isolation procedures. Miguel Angel Maestro performed cannulations and further assisted with mouse islet isolation procedures. Ana Zekavati generated sequencing libraries for the majority of sequenced ChIP samples.

Contents

Copyright Declaration	1
Abstract	2
Acknowledgements	3
Declaration of Originality	4
Contents	5
List of Figures	10
List of Tables	12
List of Abbreviations	13
1 General Introduction	16
1.1 Chromatin structure and transcription regulation	16
1.1.1 Chromatin modifications	16
1.1.1.1 Histone variants	17
1.1.1.2 Histone post-translational modifications	17
1.1.1.3 DNA modifications	18
1.1.2 Regulation of RNA Polymerase II mediated transcription	20
1.1.3 <i>Cis</i> -regulatory elements	23
1.1.3.1 Core promoter	23
1.1.3.2 Silencers	26
1.1.3.3 Enhancers	27
1.1.4 3D organisation of <i>cis</i> -regulation	33
1.1.4.1 Spatial interactions between enhancers and promoters	34
1.1.4.2 Target promoter selection	36
1.1.4.3 Predicting target genes	36
1.1.4.4 Topologically associating domains	39

1.2	Genetic variation and phenotype variability	40
1.2.1	Major types of sequence variation	40
1.2.2	Genotype to phenotype	42
1.2.2.1	Genome-wide association studies	42
1.2.2.2	Genetic variation of gene expression	43
1.3	Pancreatic islets of Langerhans	47
1.3.1	Pancreatic islet cell types	48
1.3.2	Transcription factors and pancreatic development	50
1.3.2.1	The primary transition	50
1.3.2.2	The secondary transition	51
1.3.2.3	Postnatal period	52
1.3.3	Diabetes Mellitus	52
1.3.3.1	Type 1 diabetes	53
1.3.3.2	Type 2 diabetes	54
1.3.3.3	Gestational diabetes	55
1.3.3.4	Monogenic forms of diabetes	56
1.4	Mouse as an experimental model	57
1.4.1	The mouse model in genetic research	57
1.4.2	The mouse model in diabetes research	60
1.4.2.1	Comparison of mouse and human pancreatic islets	60
1.4.2.2	The mouse model and interpretation of diabetes GWAS	62
1.5	Thesis aims	64
2	General Methods	65
2.1	Animal work	65
2.1.1	Mouse maintenance	65
2.1.2	Islet isolation	65
2.2	Molecular biology methods	67
2.2.1	RNA isolation	67
2.2.2	ChIP	67
2.2.2.1	Islet fixation	68
2.2.2.2	Chromatin preparation	68
2.2.2.3	Chromatin pre-clearing	68

2.2.2.4	Immunoprecipitation	69
2.2.2.5	DNA isolation	69
2.2.3	ATAC	70
2.2.3.1	Nuclei preparation	70
2.2.3.2	Transposition reaction	70
2.2.3.3	Library preparation	71
3	Regulatory Architecture of Mouse Pancreatic Islets	72
3.1	Introduction	72
3.1.1	The pancreatic islet regulome	72
3.1.1.1	The human islet regulome	72
3.1.1.2	The mouse islet regulome	73
3.1.2	Evolutionary conservation of the <i>cis</i> -regulatory landscape	74
3.1.3	Summary	75
3.2	Methods	76
3.2.1	ChIP-Seq	76
3.2.2	ATAC-Seq	76
3.2.3	Processing of raw sequenced reads	76
3.2.4	Peak calling	77
3.2.5	Defining <i>cis</i> -regulatory elements	78
3.2.6	Defining highly occupied active enhancers	78
3.2.7	Defining enhancer clusters	78
3.2.8	Transcription factor motif analysis	79
3.2.9	Gene ontology analysis	80
3.2.10	Conservation analysis	80
3.2.11	Quantification of gene expression	81
3.3	Results	82
3.3.1	Defining <i>cis</i> -regulatory elements of mouse pancreatic islets	82
3.3.2	Mapping genome-wide binding of five islet transcription factors	84
3.3.3	<i>De novo</i> annotation of mouse islet regulatory motif sequences	88
3.3.4	Categorizing active enhancers	91
3.3.4.1	Defining highly occupied active enhancers	91
3.3.4.2	Defining clustered active enhancers	93

3.3.5	Conservation of <i>cis</i> -regulation by enhancer clusters in mouse and human	95
3.3.5.1	Categorizing active enhancer clusters according to syntenic and sequence conservation	95
3.3.5.2	Enhancer cluster conservation and biology of pancreatic islets	99
3.4	Discussion	103
3.4.1	Mouse pancreatic islet regulation resource	103
3.4.2	Patterns of mouse-human islet regulatory conservation	104
3.4.3	Species-specific subset of islet regulatory architecture	107
4	Dissecting <i>cis</i>-regulatory Variation in Pancreatic Islets of a Hybrid Mouse Model	110
4.1	Introduction	110
4.1.1	<i>Cis</i> -regulatory variation	110
4.1.2	Allele-specific analysis in a hybrid mouse model	111
4.1.3	Summary	113
4.2	Methods	115
4.2.1	Allele-specific analysis pipeline	115
4.2.1.1	Allele-specific analysis in individual samples	115
4.2.1.2	Determining consistent and statistically significant allelic im- balance across samples	119
4.2.2	Control set of regions without allelic imbalance	120
4.2.3	Quantification of transcription factor motif disruption	120
4.2.4	Tissue specificity	121
4.2.4.1	Genes	121
4.2.4.2	Active enhancers	121
4.2.5	GLM analysis of properties associated to imbalanced genes	122
4.2.5.1	Gene and promoter properties	122
4.2.5.2	Enhancer properties	122
4.3	Results	124
4.3.1	Mapping allele-specific expression and activity of regulatory elements in mouse pancreatic islets	124
4.3.1.1	Allele-specific analysis pipeline	124
4.3.2	Properties of genes subject to <i>cis</i> -regulatory variation in pancreatic islets	132
4.3.3	Effects of genetic variation on enhancer chromatin properties	136

4.3.4	Relationship between allelic expression and chromatin state of associated regulatory regions	139
4.3.5	Properties of imbalanced enhancers associated to imbalanced genes . .	142
4.3.6	Allelic expression in T2D GWAS loci	146
4.4	Discussion	147
4.4.1	Allelic expression in pancreatic islets	147
4.4.2	Tissue-specificity of genes susceptible to <i>cis</i> -regulatory variation	148
4.4.3	Allelic chromatin states in pancreatic islets	150
4.4.4	Enhancer variation and changes in gene expression	151
4.4.5	Properties of allelic enhancers associated with allelic expression	152
5	Final Remarks	154
	Bibliography	201
	Appendix: Copyright Permissions	202

List of Figures

1.1	Distribution of epigenetic modifications across gene regulatory regions.	19
1.2	Overview of RNAPII transcription initiation.	22
1.3	Cis-regulatory elements.	23
1.4	Models of transcription factor recruitment to enhancers.	29
1.5	Illustration of enhancer-promoter looping.	35
1.6	Major classes of genetic variants.	41
1.7	Allele-specific expression and transcription factor binding.	46
1.8	Anatomy of the human pancreas.	48
1.9	Paracrine and autocrine regulation of islet cells.	50
1.10	Transcription factors and beta cell development.	53
1.11	Origin of classical inbred mouse strains.	59
3.1	Enrichment and consistency of ChIP-Seq and ATAC-Seq signal.	83
3.2	Utilizing chromatin properties for identifying regulatory elements in mouse pancreatic islets.	84
3.3	Binding of five islet transcription factors across the regulatory regions of the mouse pancreatic islets.	86
3.4	Chromatin and transcription map of the Nkx6-1 locus.	87
3.5	Categorizing active enhancers according to occupancy by the five assayed tran- scription factors.	92
3.6	Categorizing active enhancers according to clustering.	93
3.7	Gene ontology analysis of clustered and orphan mouse active enhancers.	94
3.8	Categories of mouse enhancer clusters based on conservation between mouse and human.	96
3.9	Properties of conservation-based enhancer cluster categories.	98
3.10	Expression levels of genes associated to conservation-based enhancer cluster categories.	98
3.11	Gene ontology analysis on genes associated to conservation-based categories of clustered enhancers.	101

3.12	Enhancer and transcription maps of the <i>Ghr</i> gene locus in mouse and human. . .	102
4.1	Schematic overview of the experimental design.	125
4.2	Schematic overview of the allele-specific analysis pipeline for determining al- lelic imbalance consistent across samples.	128
4.3	Estimates of allelic ratios across different stages of the allele-specific analysis pipeline.	129
4.4	Allelic imbalance of transcripts across individual and combined samples. . . .	131
4.5	Illustration of the allele-specific events mapped across a genomic locus. . . .	132
4.6	Allelic imbalance of gene expression and generation of the control gene set. . .	134
4.7	Properties of imbalanced genes and their promoters.	135
4.8	Allelic expression and properties of gene-associated regions.	137
4.9	Genome-wide allelic imbalance of H3K27ac and ATAC signal.	138
4.10	Enrichment of imbalanced H3K27ac and ATAC signal in the vicinity of imbal- anced and control genes.	140
4.11	Fraction of genes associated with imbalanced enhancers and promoters.	141
4.12	Properties of H3K27ac-imbalanced enhancers associated with allelic expression.	144
4.13	Clustering and conservation of H3K27ac-imbalanced enhancers associated to allelic expression.	145

List of Tables

1.1	An overview of major core promoter elements.	24
1.2	Forms of maturity onset diabetes of the young (MODY).	56
2.1	List of primers used for amplification of ATAC libraries.	71
3.1	Binding of TFs in regulatory elements of TF genes.	87
3.2	Motifs enriched in mouse pancreatic islet active enhancers.	89
4.1	Allelically expressed genes associated with T2D in humans.	146

List of Abbreviations

3C - Chromosome Conformation Capture
5hmC - 5-hydroxymethylcytosine
5mC - 5-Methylcytosin
ADP - Adenosine Diphosphate
ASA - Allele Specific Analysis
ATAC - Assay for Transposase-Accessible Chromatin
ATAC-Seq - ATAC Sequencing
ATP - Adenosine Triphosphate
BAM - Binary sequence Alignment/Map format data
bp - base pair(s)
BSA - Bovine Serum Albumin
CAGE - Capped Analysis of Gene Expression
cast - Mus Musculus Castaneus pseudogenome
CBP - CREB-Binding Protein
ChIA-PET - Chromatin Interaction Analysis by Paired-End Tag sequencing
ChIP - Chromatin ImmunoPrecipitation
ChIP-Seq - ChIP Sequencing
CIGAR - Compact Idiosyncratic Gapped Alignment Report
CORE - Clusters of Open Regulatory Elements
CpG - 5'—C—phosphate—G—3'
CRISPR - Clustered Regularly Interspaced Short Palindromic Repeat
CRM - Cis-Regulatory Module
CTCF - CCCTC-binding Factor
DHS - DNase I hypersensitive site
DNA - Deoxyribonucleic Acid
DRE - DNA Replication-related Element
EB - Elution Buffer
EDTA - Ethylene Diamine Triacetic Acid
eQTL - expression Quantitative Trait Loci

eRNA - enhancer RNA
ES - Embryonic Stem (cells)
EtOH - Ethanol
FAIRE - Formaldehyde-Assisted Isolation of Regulatory Elements
FBS - Fetal Bovine Serum
FGF - Fibroblast Growth Factor
GLM - Generalised Linear Models
GDM - Gestational diabetes
GRO-Seq - Global Run-On and Sequencing
GTEx - Genotype-Tissue Expression
GTF - General Transcription Factor
GWAS - Genome-wide Association Study
HAT - Histone Acetyl Transferase
HBSS - Hanks' Balanced Salt Solution
HEPES - N-2-hydroxyethylpiperazine-N'-ethanesulphonic acid
hPSC - human Pluripotent Stem Cells
IDDM - Insulin Dependent Diabetes Mellitus
IEQ - Islet Equivalent number
INDEL - Insertion or the Deletion of bases in the DNA
kb - kilobase pair(s)
KO - Knockout
LD - Linkage Disequilibrium
MAF - Minor Allele Frequency
mm9 - Mus Musculus reference genome, NCBI build 37
mmol/L - millimoles per litre
MODY - Maturity Onset Diabetes of the Young
mRNA - messenger Ribonucleic Acid
MSigDB - Molecular Signatures Database
NGF - Nerve Growth Factor
NGS - Next Generation Sequencing
NIDDM - Non-insulin-dependent *Diabetes Mellitus*
PBS - Phosphate-Buffered Saline

PCR - Polymerase Chain Reaction
PE - Pair End (read)
PEAT - Paired-End Analysis of Transcript start sites
PIC - Pre-Initiation Complex
PTM - Post-Translational Modification
qPCR - real-time quantitative PCR
RIN - Ribonucleic acid Integrity Number
RNA - Ribonucleic Acid
RNA-Seq - RNA Sequencing
RNAPI - Ribonucleic Acid Polymerase I
RNAPII - Ribonucleic Acid Polymerase II
RNAPIII - Ribonucleic Acid Polymerase II
RNaseA - Ribonuclease A
RPKM - Reads Per Kilobase of transcript per Million mapped reads
RPMI - Roswell Park Memorial Institute medium
rRNA - ribosomal Ribonucleic Acid
RT - Room Temperature
SAM - Sequence Alignment/Map format data
SE - Single End (read)
SINE - Short Interspersed Element
SNP - Single Nucleotide Polymorphism
SNV - Single Nucleotide Variant
T1D - Type 1 Diabetes
T2D - Type 2 Diabetes
TAD - Topologically Associating Domains
TALENs - Transcription Activator-Like Effector Nucleases
TCC - Tethered Conformation Capture
TF - Transcription Factor
TFBS - Transcription Factor Binding Site
tRNA - transfer Ribonucleic Acid
TSS - Transcription Start Site
ZNFs - Zinc Finger Nucleases

Chapter 1

General Introduction

1.1 Chromatin structure and transcription regulation

Genomic DNA of the eukaryotic cell is organised into a compact protein-DNA structure called chromatin (Kornberg, 1975). The structure of chromatin is condensed in a hierarchical manner. The basic unit is the nucleosome - a 146bp long stretch of DNA that forms a superhelix around a protein octamer made of four core histones. The nucleosomes, separated by approximately 200 bp intervals along a DNA molecule, form a nucleosomal array or a *bead-on-a-string* 10nm fibre (Luger et al., 1997). The fibres are coordinated by a linker histone (H1 or H5) into a hypothetical 30nm chromatin fibre (Fussner et al., 2011; Robinson et al., 2006). Further folding creates a highly organised structure that is accurately replicated in daughter cells upon cell division (reviewed in Groth et al., 2007). However, chromatin structure is far from static - its architecture is dynamic, highly regulated, and has a close relationship with the transcriptional activity of the genomic region (Berger, 2007; Lorch et al., 1987). Transcriptionally active regions of non-dividing chromatin with more relaxed nucleosomes are historically referred to as *euchromatin*. The transcriptionally inactive chromatin can be roughly divided into two functional compartments – *constitutive* and *facultative heterochromatin*. Constitutive heterochromatin contains invariably transcriptionally silenced regions and includes centromeres, pericentric DNA, telomeres and silenced mobile elements. Facultative heterochromatin, on the other hand, is conditionally silenced and includes genes subject to tissue-specific regulation, inactive X chromosomes, imprinted genes and perinucleolar DNA. (Passarge, 1979; Saksouk et al., 2015).

1.1.1 Chromatin modifications

Chromatin structure and transcriptional activity must be responsive to developmental or external stimuli, yet the defined patterns should be maintained through cell divisions. This complex

behaviour is regulated not only by the DNA sequence, but also through another layer of control that includes histone modifications and variants, DNA methylation and hydroxymethylation and some classes of small RNAs (Bernstein et al., 2007; Margueron and Reinberg, 2010). Some of these modifications may be transmissible through mitosis or meiosis, in which case they are termed *epigenetic modifications* (Bird, 2007; Holliday and Pugh, 1975).

1.1.1.1 Histone variants

The classic nucleosome histone octamer is made of two copies of four core histones - H2A, H2B, H3 and H4 (Kornberg and Lorch, 1999; Luger et al., 1997). The histone genes are present in multiple copies in most organisms and their sequences and expression patterns are deeply conserved from yeast to human (Marino-Ramirez et al., 2006; Marzluff et al., 2002; Thatcher and Gorovsky, 1994). However, the canonical histones are not universally present in nucleosomes and can be replaced with alternative variants that differ in structure and/or time of expression (Marzluff et al., 2002). The core histone H3, for example, has one alternative (H3.3) that differs in only 4 amino acids and is associated with activation of transcription (Schwartz and Ahmad, 2005). In contrast, H2A has four alternatives (H2AZ, H2ABbd, H2A.X, macroH2A) associated with transcriptional activation/repression, transcriptional activation, repression and X chromosome inactivation, respectively (Freeman et al., 1980; Thatcher and Gorovsky, 1994). Presence of alternative histones impacts nucleosome stability and represents an important regulatory mechanism of transcriptional activity.

1.1.1.2 Histone post-translational modifications

The globular regions of the nucleosome core histones are tightly packed, but their amino-terminal tails are extended and exposed to histone modifying enzymes (Allfrey et al., 1964; Luger et al., 1997). These enzymes introduce a variety of post-translational modifications (PTMs), including methylation of arginine, methylation, acetylation, ubiquitination, ADP-ribosylation, and sumoylation of lysine and phosphorylation of serine and threonine (reviewed in Bannister and Kouzarides, 2011). The PTMs are introduced by enzymes called *writers*, removed by *erasers* and recognized by *readers*, with every modification having a dedicated set of such protein complexes (Minguez et al., 2012; Seet et al., 2006).

Histone modifications are proposed to affect chromosome structure and function via two main

mechanisms. The first mechanism includes the direct effect these modifications have on the inter-nucleosomal interactions. Histone acetylation, for example, disrupts electrostatic interactions between histones and DNA, effectively leading to a more open chromatin state (Görisch et al., 2005). Acetylation occurs on numerous histone tail lysines (H3K9, H3K14, H3K18, H4K5, H4K8 and H4K12) and was proposed to promote accessibility of chromatin to DNA-binding proteins and transcriptional complexes (Grunstein, 1997). The second mechanism of PTM activity includes recruitment of chromatin remodelers, protein complexes that are capable of mobilizing and ejecting nucleosomes and thus changing accessibility to nucleosomal DNA (Saha et al., 2006). Mobilization of histones as well as proper replication of their PTMs is controlled and promoted by *chromatin chaperons*, a diverse family of proteins that can act independently or cooperatively with chromatin remodelers (Falbo and Shen, 2010).

One histone can be modified by more than one type of PTM or multiply modified by the same PTM at different residues. In the year 2000, the *histone code* hypothesis was introduced, where the combination of PTMs on one or multiple histone tails specifies a unique downstream function (Strahl and Allis, 2000). It has since been documented that PTMs have dynamic and context-dependent effects, with particular combinations of PTMs occurring at different regulatory regions (Figure 1.1). Patterns of chromatin modifications across *cis*-regulatory elements will be discussed in more detail in Section 1.1.3.

1.1.1.3 DNA modifications

The genomic DNA itself can also acquire chemical modifications. DNA methylation is the covalent addition of a methyl group to the 5-position of the cytosine ring (5mC) (Bird, 1992). It was the first epigenetic modification to be described (Avery et al., 1944; Holliday and Pugh, 1975) and the best characterized one to date. In differentiated mammalian cells, DNA methylation occurs predominantly in the context of individual CpG dinucleotides, the majority of which are methylated (Ehrlich et al., 1982). In contrast, *CpG islands* (CpG clusters approximately 1,000 bp long), are thought to be typically unmethylated (Bird, 1992; Bird et al., 1985). CpG islands tend to be associated with promoters (Gardiner-Garden and Frommer, 1987) and can be subject to methylation, as observed in imprinting, X chromosome inactivation or tissue-specific silencing (Ferguson-Smith et al., 1993; Riggs, 1975; Schilling and Rehli, 2007; Sharp et al., 2011). It has long been established that DNA methylation of gene promoters is closely linked to silencing of gene transcription (Deaton and Bird, 2011; Jones, 1985). In contrast,

methylation of the gene body was demonstrated to actively increase gene expression levels (Yang et al., 2014).

Methylated cytosine can undergo enzymatic oxidation that produces 5-hydroxymethylcytosine (5hmC) - a residue that was recognized not only as an intermediary product of a putative active DNA demethylation pathway, but an epigenetic mark in its own right. In contrast to 5mC, 5hmC of regulatory elements is positively correlated with gene expression (Ficz et al., 2011; Tahiliani et al., 2009; Wu and Zhang, 2011).

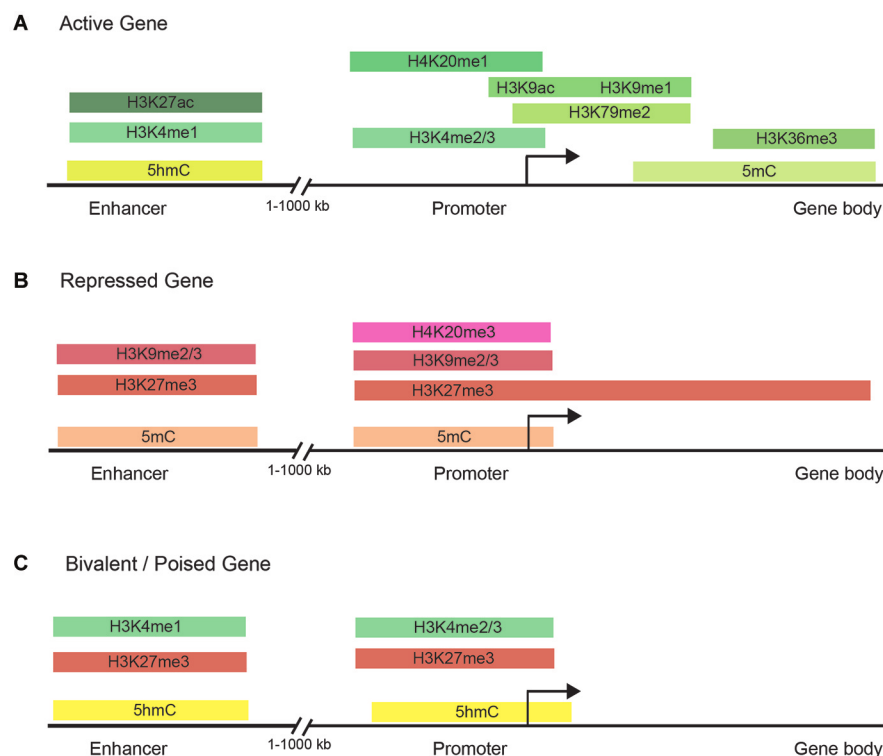


Figure 1.1 Distribution of epigenetic modifications across gene regulatory regions. Specific combinations of DNA and histone modifications are associated with a) active, b) repressed or c) bivalent/poised state of proximal and distal regulatory elements, as well as expression of the associated gene. Figure reproduced from Layman and Zuo (2015).

1.1.2 Regulation of RNA Polymerase II mediated transcription

Transcription in eukaryotes is achieved by three different types of RNA polymerase - RNAPI, RNAPII and RNAPIII (Carter and Drouin, 2009). These polymerases are responsible for transcription of different classes of RNAs. RNAPI transcribes ribosomal RNAs (rRNAs), RNAPII transcribes RNAs that will become messenger RNAs (mRNAs) as well as other regulatory RNAs, and RNAPIII transcribes small RNAs such as transfer RNAs (tRNAs) (Cramer et al., 2008).

RNAPII mediated transcription is the major transcription event of the eukaryotic cell and has been extensively studied since it was first described (Kornberg, 1999; Lopata et al., 1986; Weiss and Gladstone, 1959). This highly complex event consists of a series of several precisely timed steps, each one regulated by a large number of specialized sequence-specific DNA binding proteins (transcription factors) (Davidson et al., 1983; Shandilya and Roberts, 2012).

RNAPII recruitment takes place on the core promoter - a regulatory element that positions the transcriptional machinery and directs unidirectional transcription (Danino et al., 2015; Verrijzer et al., 1995). Typically, binding of *activator* transcription factors (TFs) distal from the core promoter leads to the recruitment of *adaptor* complexes (Green, 2005). Adaptors help activators to recruit Mediator (Leroy et al., 2006) and facilitate sequential binding of the subunits of the basic pre-initiation complex (PIC); five general transcription factors (GTFs) TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and RNAPII itself (Roeder, 1996; Thomas and Chiang, 2006). The basic PIC is sufficient for accurate start-site selection and initiation of basal transcription *in vitro* (Das et al., 1995; Hampsey, 1998). However, PIC components *in vivo* are assembled around the Mediator complex, which is necessary for responding to transcriptional activators and achieving active levels of transcription (Allen and Taatjes, 2015; Holstege et al., 1998).

Transcription initiation starts with ATP hydrolysis powered structural remodeling of PIC and directing the DNA template strand to the active site of RNAPII for DNA-templated RNA chain synthesis (Treutlein et al., 2012; Wang et al., 1992).

Transcription initiation is a discrete stage of transcription (Dvir, 2002; Pal et al., 2001) during which the transcription machinery often goes through *abortive transcription* - synthesis of multiple short abortive RNA products (Dvir, 2002; Holstege et al., 1997). Once RNAPII is phosphorylated and dissociated from Mediator, the escape is complete. TFIID, TFIIE, TFIIH and the Mediator complex remain on the core promoter (now making the *scaffolding complex*),

enabling rapid re-initiation of the subsequent cycle of transcription. (Dahmus, 1996; Søgaaard and Svejstrup, 2007; Wong et al., 2014).

Promoter-proximal pausing occurs at the end of the promoter escape, while the transcription complex is still not fully prepared for elongation. Once released, the transcription machinery associates with a number of *elongation factors*, many of which have apparently redundant functions (Jonkers and Lis, 2015; Pokholok et al., 2002). Through the elongation process, factors involved in capping, splicing, and cleavage/polyadenylation are recruited and act in parallel with transcription (Bentley, 2014; Kornblihtt et al., 2004; McCracken et al., 1997).

Transcription termination commences after the successful completion of mRNA synthesis and dissociation of RNAPII from the DNA template. Once termination is completed, RNAPII is recycled and used in new rounds of transcription, though not much is known about the mechanisms of recycling (Cho et al., 1999; Svejstrup, 2004; Zawel et al., 1995).

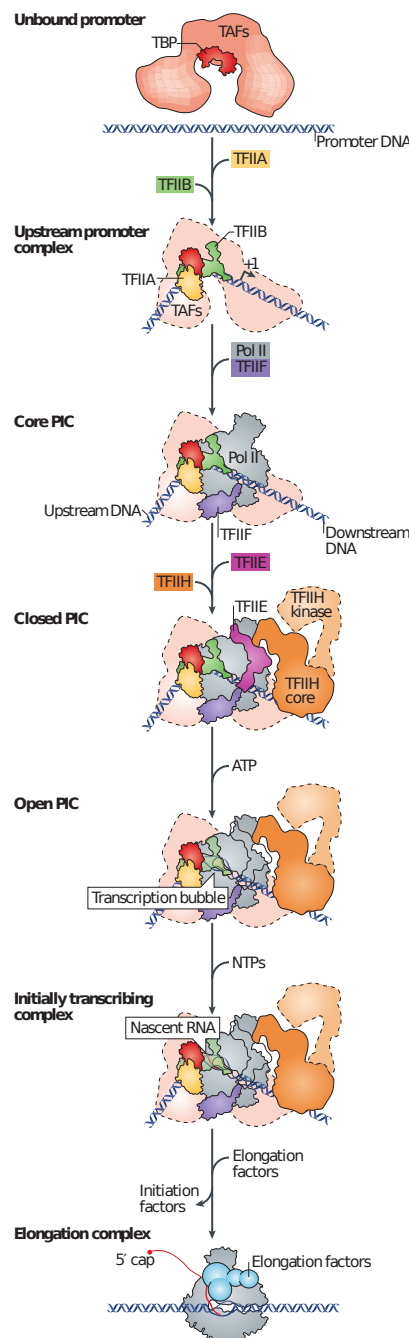


Figure 1.2 Overview of RNAPII transcription initiation. The *unbound promoter* recruits TFIID via its TATA box-binding protein (TBP) subunit and TBP-associated factors (TAFs). Addition of TFIIB and TFIIA creates the *upstream promoter complex*, which together with RNAPII (Pol II) and TFIIF makes the core pre-initiation complex (PIC). Subsequent binding of TFIIE and TFIIH complete the *closed PIC* formation. Transcription is initiated with ATP-powered structural remodelling, *opening of PIC* and creation of the transcription bubble. The nascent RNA goes through several rounds of abortive transcription. Once prepared for elongation, the transcription machinery associates with a number of elongation factors, forming the *elongation complex*. Factors involved in capping, splicing, and cleavage/polyadenylation may be recruited to act in parallel with transcription. Figure adapted from Sainsbury et al. (2015).

1.1.3 *Cis*-regulatory elements

Stretches of DNA known as *cis-regulatory elements* control transcription from close proximity or great distance. *Cis*-regulatory elements usually harbour binding sites for *trans*-regulatory proteins and carry a particular combination of chromatin modifications. Several types of *cis*-regulatory elements have been defined so far, including core promoters, proximal promoters, enhancers, silencers and insulators (Figure 1.3) (Hardison and Taylor, 2012; Mathelier et al., 2015; Symmons and Spitz, 2013).

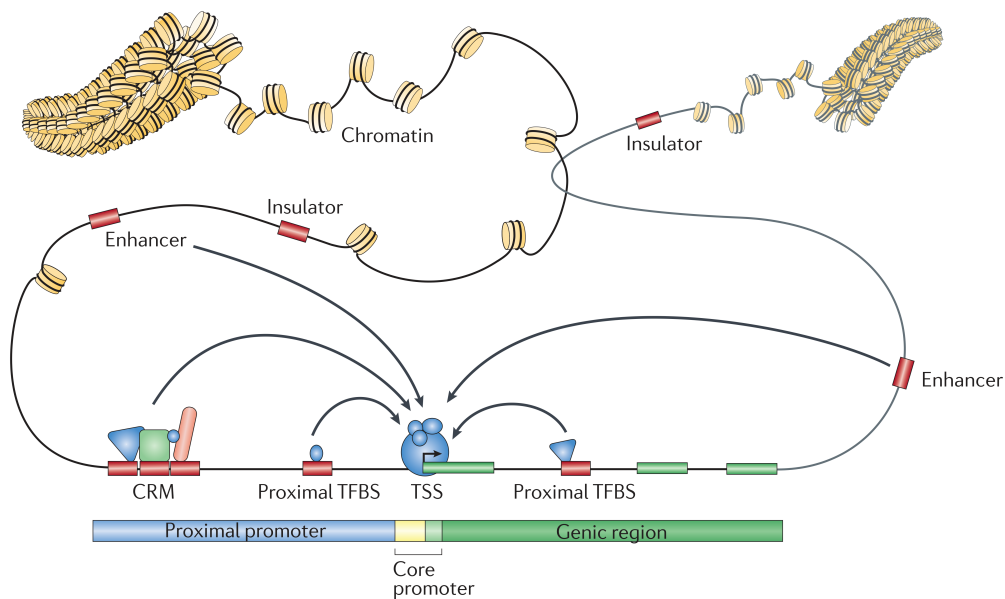


Figure 1.3 *Cis*-regulatory elements. A schematic representation of the transcription regulatory input. *Cis*-regulatory elements can be proximal or distal to the transcription start site (TSS). They are embedded into open chromatin and contain sequence-specific transcription factor binding sites (TFBSs). TFBSs can form clusters, often called *cis*-regulatory modules (CRMs). The transition between the compact and open chromatin may be marked by insulators. The regulatory input of enhancers and proximal promoters converges into the target gene core promoter that directs transcription. Figure adapted from Lenhard et al. (2012).

1.1.3.1 Core promoter

The core promoter is the central component of the regulation of gene transcription (Butler and Kadonaga, 2002; Sandelin et al., 2007). It is often defined as the minimum sequence necessary for accurate initiation of gene transcription (Smale and Kadonaga, 2003; Struhl, 1987). Its basic function is to properly position the RNAPII and the components of the PIC and direct the synthesis of mRNA from the DNA template (Kadonaga, 2004; Verrijzer et al., 1995; Weil

et al., 1979). Typically, the core promoter overlaps the gene transcription start site (TSS) and then extends for approximately 35bp upstream or downstream from it (Smale and Kadonaga, 2003). A gene, however, is not restricted to a single TSS. Instead, one core promoter can span several TSSs in a 50-100 bp region (Carninci et al., 2006; Haberle et al., 2014). Interestingly, some promoters show differences in preferential use of a TSS in different tissues, indicating a regulated TSS selection (Kawaji et al., 2006).

The sequence underlying the core promoter region harbours a series of motifs that bind PIC factors (overview in Danino et al., 2015). The TATA box was the first core promoter element to be identified (Gannon et al., 1989) and was once considered to be present at all core promoters. Today, it is estimated to be present in 10–20% of metazoan core promoters (Carninci et al., 2006; Kim et al., 2005). In fact, no motif is universal for all core promoters. Different combinations of these motifs are associated with different modes of transcription initiation and represent an important layer of gene regulation (Butler and Kadonaga, 2002; Lenhard et al., 2012; Sandelin et al., 2007). An overview of the major core promoter elements, their DNA sequence, ability to bind particular proteins and co-occurrence in the same core promoter is adapted from Sandelin et al. (2007) and summarized in Table 1.1.

Table 1.1 An overview of major core promoter elements. An overview of major core promoter elements. Abbreviations; TATA box (TATA), initiator (INR), downstream promoter element (DPE), TFIIB recognition element (BRE), 5'-C-phosphate-G-3' (CpG).

Element	Sequence	Position relative to TSS	Binding factors	Frequently co-occurring motifs	Gene type
TATA	TATA	-25	TFIID	INR	Tissue-specific
INR	YYA+1NWYY	+30	TFIID	TATA, DPE	Housekeeping
DPE	DSWYVY	-20	TFIID	INR	Housekeeping
BRE	SSRCGCC / RTDKKKK	-35,-20	TFIIB	TATA	Tissue-specific
CpG	CpG	0.5 – 2Kb spanning TSS	/	/	/

Immediately upstream of the core promoter lies a region termed the *proximal promoter* (Allen et al., 1992; Maston et al., 2006). It is typically only a few hundred bases away and contains multiple binding sites for activators (Sandelin et al., 2007; Soravia et al., 1995). Interestingly, regions further upstream (-1,000 to -500 bp) tend to contain elements for negatively regulators of transcription (Cooper et al., 2006).

Recent advances in NGS technology gave rise to a series of genome-wide assays for identification of TSSs, for example; capped analysis of gene expression (CAGE Shiraki et al., 2003); global run-on and sequencing (Gro-Seq Core et al., 2008) and paired-end analysis of transcript start sites (PEAT Ni et al., 2010). In combination with open chromatin assays, such as; formaldehyde-assisted isolation of regulatory elements (FAIRE Giresi et al., 2007); Deoxyribonuclease I (DNaseI Workman and Kingston, 1992) and chromatin immunoprecipitation (ChIP Furey, 2012), these assays enabled the identification of large number of core promoter and in-depth dissection of their characteristics. Comparative analysis of these characteristic across metazoans revealed classes of promoters that differ according to enrichment of motifs, their size and directionality, nucleosome arrangement and histone signatures (Carninci et al., 2006; Sandelin et al., 2007). Nonetheless, there is no firm consensus on the number of classes of promoters. The prevalent classification separates promoters to *sharp* and *broad* promoters, based on the preferential use of one or several TSSs for transcription initiation (Carninci et al., 2006; Sandelin et al., 2007; Schug et al., 2005). Nucleosome configuration was shown to be one of the most prominent differences between *sharp* and *broad* promoters with nucleosomes in *sharp* promoters flanking the TSS in a less ordered fashion (Rach et al., 2011). However, other classifications of promoters exist, such as the tripartition (Types I - III) suggested by Lenhard et al. (2012).

Type I (“adult”) promoters are described as *sharp* and often have a combination of a TATA box and the Inr element (Ponjavic et al., 2006). These promoters generally have a low CpG content (Yamashita et al., 2005) and a specific pattern of H3K4me3 which is only present downstream of the TSS (Ernst and Kellis, 2010). The TSSs within these promoters show relaxed evolutionary constraint and are generally less conserved than the *broad* promoter TSSs (Forrest et al., 2014). The TATA-containing subset of these promoters, however, were shown to have more conserved sequence than the *broad* promoter classes (Carninci et al., 2006; Taylor et al., 2006). Type I promoters are associated with tissue-specific genes expressed in fully differentiated adult tissues. These genes tend to be regulated predominantly with enhancers and show less diversity across promoter states (Andersson et al., 2014; Forrest et al., 2014; Schug et al., 2005).

Type II (“ubiquitous”) promoters, a subcategory of *broad* promoters, harbour either a DNA replication-related element (DRE) or an array of novel, less defined motifs (Matsukage et al., 2008). Usually only one short CpG island is present, overlapping the strongly conserved TSS

(Forrest et al., 2014). The H3K4me3 follows the short CpG island and overlaps only the 5' end of the gene (Deaton and Bird, 2011; Lenhard et al., 2012). The genes controlled by type II promoters are typically active in all cell types, have few associated enhancers and receive the majority of their regulatory input from the promoter (Andersson et al., 2014; Forrest et al., 2014).

Type III (“developmental”) promoters are a subcategory of the *broad* promoters, but are somewhat less “broad” than type II promoters. They have a higher number of TSS and typically contain several large CpG islands that extend into the gene body (Akalin et al., 2009). This class of promoters carries an Inr element alone or in combination with DPE. H3K4me3, follows the CpG islands and extends to gene body and a strong H3K4me3 signal indicates an actively transcribed gene. Moderate H3K4me3 in combination with H3K27me3 (so called “bivalent” state), however, indicates a poised or Polycomb silenced gene. Absence of any of the two histone modifications on the flanking nucleosomes indicates an inactive gene (Akalin et al., 2009; Bernstein et al., 2006; Lenhard et al., 2012; Schwartz et al., 2010). Type III promoters are associated with developmental genes. These genes have a high number of associated enhancers contributing to their regulation at the same level as the diverse state promoters (Akalin et al., 2009; Andersson et al., 2014; Forrest et al., 2014).

A striking number of promoters, 10-22%, are coupled with an independent promoter located less than 1,000 bp on the opposite strand (Trinklein et al., 2004; Uesaka et al., 2014). These “bidirectional” promoters are often rich in CpG islands and conserved between human and mouse (Lavia et al., 1987). Genes regulated by the bidirectional pairs are often (though not exclusively) co-expressed (Trinklein et al., 2004).

1.1.3.2 Silencers

On the opposite side of the regulatory spectrum lie silencers. Silencers are defined as sequence elements that bind negative transcription factors and convey a repressive effect on the promoter activity in an orientation- and position-independent fashion (Brand et al., 1985; Donnelly et al., 1999; Rusche et al., 2003; Shei and Broach, 1995). Elements with described properties are considered to be *classical silencers* and make the largest group of silencing elements described in the genome (Ogbourne and Antalis, 1998). An example of a *classical* silencer is the AT-rich, Oct1-binding region found 128-480 bp upstream of the transcription initiation site of the human

thyrotropin-beta gene (Kim et al., 1996).

The non-classical, *position-dependent silencers* (NREs) generally act passively by physically inhibiting the activation of a positive transcription signal. They can act through blocking the binding of a neighbouring activator (Harris et al., 2005), directly competing for the activator binding site (Chin et al., 1995), establishing a repressive chromatin structure (Bi et al., 1999; Srinivasan and Atchison, 2004) or inhibiting PIC assembly (Chen and Widom, 2005).

1.1.3.3 Enhancers

Enhancers are a major category of regulatory elements, usually defined as *cis*-acting DNA sequences that can increase the transcription of target genes (Banerji et al., 1983, 1981; Khoury and Gruss, 1983; Ong and Corces, 2011). They act independently of their orientation, though exceptions to this rule have been described (Swamynathan and Piatigorsky, 2002; Wei and Brennan, 2000). Enhancers are found in intergenic regions (Bulger and Groudine, 2010), introns (Kowalczyk et al., 2012), and even exons of genes (Birnbaum et al., 2012; Ritter et al., 2012), with approximately 51% of them located within the gene body (Heintzman et al., 2007). As such, they can be situated upstream, downstream or within the target gene at distances that range from a few kilobases to over a megabase away from the target gene promoter (Lettice, 2003; Sanyal et al., 2012). Consistent with that, enhancers do not necessarily regulate the closest gene and can “skip” several genes before reaching their target promoter (Sanyal et al., 2012; Waszak et al., 2015).

1.1.3.3.1 Models of enhancer action

Enhancers are usually several hundred base pairs long and contain clusters of short, degenerate nucleotide sequence patterns - *motifs* - recognized by multiple transcription factors. Transcription factors themselves are modular in nature and have regions with specialized functions; a DNA-binding domain that ensures binding to a specific DNA sequence, a multimerization domain that allows protein-protein interactions, and one or more effector domains that can activate or repress the rate of transcription (Pabo and Sauer, 1992; Tjian and Maniatis, 1994). Binding of TFs to the enhancer can be *additive*, where the concentration of the bound TFs is proportional to the transcriptional activation (Giorgetti et al., 2010), or *cooperative*, where the transcription is “switched on” once TF concentration reaches a certain threshold (Lebrecht

et al., 2005). Cooperative binding is often associated to developmental enhancers and could be important for buffering variation in the concentrations of individual TFs (Karczewski et al., 2011; Spivakov et al., 2012).

The relative order, orientation and spacing of TF motifs within an enhancer is often referred to as *motif grammar* and can be critical for ensuring proper regulatory activity of the enhancer, even though different grammar can have the same regulatory output (Reményi et al., 2004; Swanson et al., 2010; Zinzen et al., 2009). Three models of TF binding to enhancers have been described so far (Figure 1.4) (reviewed in Spitz and Furlong, 2012). In the *enhanceosome* model, the recruitment of TFs is highly cooperative, the motif grammar and protein-protein-DNA interaction is fixed and the integrated activity of all TFs is required for achieving the regulatory output (Merika and Thanos, 2001; Thanos and Maniatis, 1995). This leads to rapid activation of transcription, which was shown to be important for immune related genes (Senger et al., 2004). In the *billboard* model, the motif grammar is flexible, the TF binding cooperative and/or additive and the regulatory output can be achieved with only a subset of bound TFs (Kulkarni and Arnosti, 2003; Swanson et al., 2010). As demonstrated *in vivo* Drosophila embryos, an enhancer can serve to “display” both active and repressed signals, and the transcriptional machinery can actively interpret those signals (Kulkarni and Arnosti, 2003). The *TF collective* describes a model where flexible motif grammar ultimately achieves binding of the same TFs, some of which are recruited via protein-protein interactions (Junion et al., 2012).

1.1.3.3.2 Pioneering transcription factors

Enhancers have to be primed for binding of TFs by a select subset of TFs capable of accessing their designated binding sites, irrespective of nucleosomes that may be occluding these sequences - *pioneering* TF. This phenomenon was first described in yeast (Almer et al., 1986) but was since recognized as a general mechanism of establishing transcriptional competence in development and cell reprogramming (Cirillo and Zaret, 1999; Zaret and Carroll, 2011). The propensity of Oct4, Sox2, and Klf4 to bind and displace nucleosomes was recently attributed to the ability of their DNA-binding domains to recognize a part of their canonical motif that is displayed on the nucleosome surface (Soufi et al., 2015). Other TFs, such as FoxA1, have a histone-like structure and are proposed to open chromatin via modifying or replacing the nu-

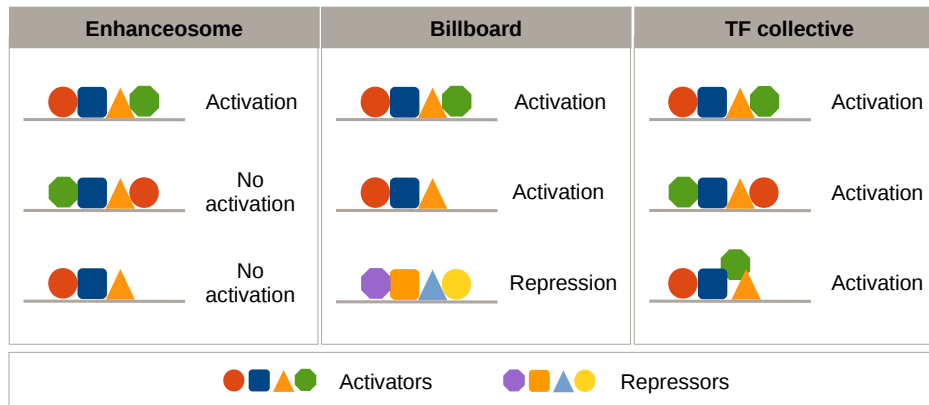


Figure 1.4 Models of transcription factor recruitment to enhancers. There are currently three accepted models of TF-mediated enhancer activation. a) The enhanceosome model describes a highly cooperative TF recruitment to the enhancer. b) The billboard model describes a more flexible recruitment of TFs, with few constraints on their relative positioning, or the type of TF recruited. c) In the TF collective model, activation of enhancers can be achieved by different combinations of TFs, some of which might bind indirectly to the DNA. Information presented in this figure was adapted from Spitz and Furlong (2012).

cleosome (Caravaca et al., 2013; Cirillo et al., 2002). Nucleosome displacement is followed by demethylation of DNA and, in some cases, deposition of an active histone modification (Xu et al., 2007).

Enhancers bound only by pioneering TFs are usually not capable of activation of transcription (Xu et al., 2009). Two scenarios can lead to enhancer activation. Following priming, the pioneering TFs may remain bound to be appropriately used in different subsequent lineages, as demonstrated in the hematopoietic cells (Mercer et al., 2011). Alternatively, they can be replaced by other TFs, as was described for members of the Sox family in the neural lineage (Bergsland et al., 2011). These other TFs are developmentally regulated and expressed in a specific spatio-temporal window, resulting in time and tissue specific activity of the enhancer (Zinzen et al., 2009). Binding of developmentally regulated factors to primed enhancers is followed by the recruitment of co-activators (Mediator, p300/CREB-binding protein (CBP)) that mediate contacts between enhancers and general transcription factors at the promoters and/or function as chromatin remodelers (Brown et al., 2000; Kuras et al., 2003). Apart from the chromatin remodelers, cooperative TF binding (Adams and Workman, 1995; Workman and Kingston, 1992), chromatin structure remodeling (RSC) complex (Cairns et al., 1996; Lia et al., 2006; Lorch et al., 2001) and actively transcribing RNAPII (Kireeva et al., 2002; Kulaeva et al., 2010) were all shown to mediate histone displacement.

1.1.3.3 Chromatin properties of enhancers

Priming and activation of enhancers is associated with nucleosome mobility and changes in histone modifications, even so the precise timing of acquiring unique chromatin signatures has been under investigated. H2A.Z and H3.3, alternative and less stable histone variants that mark open chromatin, were found to be enriched at both active and poised enhancer and promoter regions of multiple cell-types (Chen et al., 2013; Creyghton et al., 2010b). H2A.Z was demonstrated to promote binding of both activators and repressors as well as chromatin modifiers at regulatory elements during embryonic stem cell differentiation, indicating a mutually dependent relationship between the nucleosomes and TFs (Hu et al., 2013).

The same might hold true for histone modifications. H3K4me1, H3K9me1, and H3K27me1 were shown to be present at the enhancers of genes poised for later activation in the hematopoietic lineage (Cui et al., 2009). Concordantly, mono- and di-methylation of H3K4 was shown to coincide with the binding of pioneering TFs FoxA1 and PU.1 (Heinz et al., 2010; Sérandour et al., 2011). However, H3K4me1 and H3K4me2 were also shown to mark nucleosomes that are not yet, but about to be bound by pioneering factor FoxA1, hence pre-determining a set of lineage-specific transcriptional enhancers (Lupien et al., 2008). Presence of H3K4me1 on enhancers prior to their activation was documented in multiple developmental model systems, from the fruit fly and zebrafish to human (Bogdanović et al., 2012; Bonn et al., 2012; Rada-Iglesias et al., 2010). H3K4me1 is maintained on enhancers during activation and even after they lose their activating properties (Bogdanović et al., 2012).

Many of the co-activators recruited by pioneering TFs have histone acetyltransferase activity (Brown et al., 2000; Brownell et al., 1996). A major product of the p300/CBP co-activator is acetylation of H3K27 (Jin et al., 2011; Tie et al., 2009). H3K27ac is now established as a mark characteristic for active enhancers and promoters, acquired almost exclusively in the context of pre-existing H3K4me1 regions (Bonn et al., 2012; Cui et al., 2009). This residue is recognized by bromodomain proteins, ATP-dependent remodelers, TFIID components, and factors regulating transcriptional pause release, all of which might contribute to the transcription process (Yun et al., 2011). H3K27ac is particularly highly correlated with enhancer activity and the expression of the closest proximal gene (Rada-Iglesias et al., 2010). As such, it is perhaps the best studied histone modification, although it is not the only acetylation mark that is found at enhancers. H3K18, for example, was found to be present in over 20% of enhancers where it is

associated with gene activation (Jin et al., 2011; Wang et al., 2008c).

A particular subset of enhancers, *poised* enhancers, was found marked by H3K4me1, H3K27me3 and bound by the Polycomb complex PRC2, while lacking H3K27ac (Rada-Iglesias et al., 2010; Zentner et al., 2011). These enhancers are bound by CBP/p300, but the lack of H3K27ac distinguishes them from active enhancers (Creyghton et al., 2010a). Poised enhancers are located in the vicinity of key early developmental genes and, surprisingly, were found to already be in contact with their target promoters (Sanyal et al., 2012). A subset of enhancers marked by H3K4me1 was found to lack both H3K27ac and H3K27me3 (Creyghton et al., 2010a; Rada-Iglesias et al., 2010; Zentner et al., 2011) and may represent a separate class of inactive enhancers.

Tri-methylation of H3K4, a modification primarily associated with RNAPII occupied promoters, may also be present on active enhancers and was reported to be proportional to enhancer activity (Pekowska et al., 2011). However, most enhancers do not acquire H3K4me3 and a high ratio of H3K4me1 to H3K4me3 is considered a feature that distinguishes enhancers from promoters (Heintzman et al., 2007; Koch and Andrau, 2011). Some less mentioned methylation residues, H3K9me1 and H4K20me1, were also reported on activated enhancers (Cui et al., 2009).

Methylation of DNA is also associated with enhancers and was shown to be inversely correlated with active enhancer histone marks, TF occupancy and binding of co-activator proteins (Thurman et al., 2012). Enhancer 5mC is dynamic and can be introduced upon binding of repressors or removed upon loss of activators (Stadler et al., 2012). In contrast, hydroxymethylation at enhancers coincides with H3K4me1 and H3K27ac and activation of nearby genes (Ficz et al., 2011; Pastor et al., 2011; Wu and Zhang, 2011). While it can be found in promoters, 5hmC is more abundant in poised and active enhancers (Yu et al., 2012).

Upon recognizing specific chromatin signatures of enhancers, profiling of open chromatin, chromatin modifications and TF binding sites has been extensively used for their prediction across different tissues and cell lines (Heintzman et al., 2009; Kim et al., 2010; Visel et al., 2009). In the latest such endeavour, The NIH Roadmap Epigenomics Mapping Consortium identified over 399,000 putative enhancer elements, covering 8.1% of the non-coding genome across 147 different tissue or cell types, demonstrating the power of this approach (Kundaje et al., 2015).

1.1.3.3.4 Functional organisation of enhancers

Activity of many single enhancers has been tested and confirmed by *in vivo* and *in vitro* experimental approaches. This implies that these enhancers are able to act independently of the chromosomal context. However, in endogenous loci, enhancers are embedded in a complex regulatory environment with multiple enhancers and other *cis*-regulatory sequences present. In fact, metazoan genes have been shown to be typically regulated by more than one enhancer (Levine and Tjian, 2003).

Developmental genes, in particular, are known to have loci that can be fragmented into elements that each drive a different portion of total gene expression (Visel et al., 2007). This property, often termed *enhancer modularity* enables enhancers to regulate transcription in a strict spatio-temporal window, independently of each other and of the basal promoter, creating complex patterns of expression (Frankel, 2012). An example of modular regulation is APOE, a human gene expressed in seven different tissues, in each tissue driven by an exclusive set of *cis*-regulatory elements (Simonet et al., 1991; Zheng et al., 2004).

Enhancers can also have mutually overlapping activity, whether partially or completely. In that scenario, *shadow* enhancers drive the same pattern of transcription and bind the same TFs as the *primary* enhancer (which is more proximal to the target promoter) (Barolo, 2012). Shadow enhancers are proposed to provide robustness to genetic and environmental perturbations and are generally associated with developmental genes (Barolo, 2012; Khoueiry et al., 2016). They are, therefore, not necessarily redundant, but might be advantageous in conditions not tested in experimental settings. A classic example of such regulation is the *shavenbaby* (*svb*) gene in *Drosophila*. Whereas three enhancer elements are sufficient for recapitulating *svb* expression *in vivo*, presence of two more distal regulatory elements is necessary for achieving a robust transcriptional output, impervious to temperature fluctuation (Frankel et al., 2010). A very recent study concluded shadow enhancers to be much more pervasive than originally thought, with 64% of examined *Drosophila* developmental loci harbouring two or more partially redundant enhancers (Khoueiry et al., 2016).

In the linear genomic space, enhancers can be located in close proximity of one another. This phenomenon was recognized in enhancer clusters (Pasquali et al., 2014), open regulatory elements (COREs) (Gaulton et al., 2010), genomic regulatory blocks (Akalin et al., 2009), enhancer stretches (Parker et al., 2013) and super-enhancers (Lovén et al., 2013; Whyte et al.,

2013). These terms are considered related, but were each defined with their own set of criteria. super-enhancers, for example, were originally defined as groups of enhancers less than 12.5kb apart and with an unusually strong enrichment of the Mediator complex (Whyte et al., 2013). In contrast, stretched enhancers were described as H3K27Ac-rich chromatin regions over 3kb in length (Parker et al., 2013). The number of elements described by these studies differs, suggesting that the terms they are describing are not equivalent (Pott and Lieb, 2015). However, all of them were found to be associated with tissue-specific transcription. Moreover, several studies found enrichment of disease-associated variants at these loci (Hnisz et al., 2013; Parker et al., 2013; Pasquali et al., 2014), further emphasising the importance of this novel concept in transcription regulation. There is very little explanation to why these enhancers cluster or what is the relationship between individual enhancers within the unit. Recently, a functional dissection of one super-enhancer in an *in vitro* model of human erythropoiesis suggested that, despite the indistinguishable chromatin features and TF occupancy between the individual enhancers, there is a functional hierarchy within the super-enhancer (Huang et al., 2016). Further studies are needed to see if this is a general characteristic of clustered enhancers and/or super-enhancers.

1.1.4 3D organisation of *cis*-regulation

Interactions between linearly separated *cis*-regulatory elements must be viewed in the context of the three-dimensional organisation of the genome. Formation of DNA loops in the *Escherichia coli* lac operator (Oehler et al., 1990) led to the idea that the same concept can be extrapolated to the much more complex architecture of the eukaryotic chromatin. A hypothesis has been proposed where the *cis*-regulatory elements are in physical contact via formation of chromatin loops (Bulger and Groudine, 1999; Kleinjan and van Heyningen, 2005). This hypothesis has since been tested with two independent approaches - fluorescence *in situ* hybridization (FISH Ronshaugen and Levine, 2004) and chromosome conformation capture based methods (Dekker et al., 2002; Kleinjan and van Heyningen, 2005). Although FISH experiments do not always support the physical interactions found in cross-linking experiments (Williamson et al., 2012), this model of long-range regulation in *cis* is now widely accepted.

1.1.4.1 Spatial interactions between enhancers and promoters

Enhancer-promoter interactions are considered central to the current view of transcriptional regulation. It is unclear what the mechanism is that enables this interaction and whether the mechanism is universal for all types of promoters/enhancers or whether there are multiple mechanisms. Two models of promoter-enhancer communication are currently under consideration. The *looping* model implies a direct contact between the two chromosomal regions by looping out the intermediate DNA sequence (Rippe et al., 1995). The *tracking* model, describes a gradual progress of enhancer-bound proteins along the chromatin towards the promoter, potentially “scanning” the chromatin on the way, and ultimately forming a loop between the enhancer and the target promoter (Blackwood, 1998; Louie et al., 2003). This model is consistent with the long-distance and orientation-independent properties of enhancer activity and insulators, though it is unlikely to be feasible in the case of larger enhancer-promoter distances.

A number of proteins have been found to mediate the enhancer-promoter loop formation, including GATA1 (Vakoc et al., 2005), EKLF (Drissen et al., 2004), Brg1 (Bayer et al., 2003), Mediator (Kagey et al., 2010), CTCF (Mishiro et al., 2009), SATB1 (Cai et al., 2006) and cohesin (Haering and Jessberger, 2012; Kagey et al., 2010). Mediator, in particular, seems to be important for loop formation. This large protein complex is a general target of TF activation domains and multiple TFs may bind to the same or different Mediator subunits simultaneously (Allen and Taatjes, 2015). A study in murine embryonic stem cells demonstrated that Mediator interacts with cohesin, a protein complex which can form rings that connect two DNA regions (Kagey et al., 2010).

The purpose of looping is generally considered to be increasing the concentration of relevant factors at the promoter (Figure 1.5). Enhancers were demonstrated to recruit both RNAPII and GTFs and thus could serve as centres for PIC assembly and delivery to the promoter (Koch et al., 2011). Furthermore, enhancer-promoter looping and gene activation was shown to correlate with RNAPII transcription of enhancer DNA itself (De Santa et al., 2010; Kim et al., 2010; Wang et al., 2011a). This transcription is bi-directional and generates low-level transcripts of various length, stability and polyadenylation status, termed *eRNAs* (enhancer RNAs) (Sambasivarao, 2014). Induction of eRNAs was suggested to be a robust indicator of functional enhancer-promoter looping and a better predictor of enhancer activity than any enhancer-bound transcription activator or chromatin mark (Wang et al., 2011a; Zhu et al., 2013). eRNAs interact

directly with the Mediator complex (Lai et al., 2013) and were proposed to serve as a scaffold for protein-mediated looping (Hsieh et al., 2014). There are indications that eRNAs can act through different mechanisms. For example, eRNAs production occurs prior to appearance of H3K4me1 in some cases (Kaikkonen et al., 2013) and post-transcription of the target genes in others (Arner et al., 2015). Moreover, some studies reported eRNA production to depend on the target gene transcription (Kim et al., 2010), while others did not find that type of dependence (Yoo et al., 2012). In addition to initiation, eRNAs were shown to affect elongation by binding to the NELF (negative elongation factor) and destabilizing its association with paused RNAPII (Schaukowitch et al., 2014).

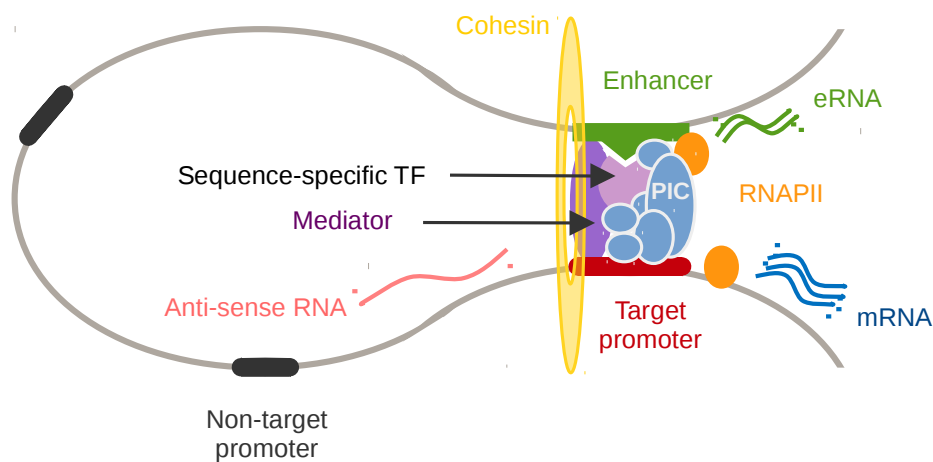


Figure 1.5 Illustration of enhancer-promoter looping. A simplified representation of elements involved in enhancer-promoter looping. Sequence-specific TFs bind to enhancers and enable recruitment of the Mediator complex. This complex facilitates the assembly of PIC and mediates the formation of an enhancer-promoter loop. The loop, further stabilized by cohesin, increases the concentration of PIC proteins at the core promoter of the target gene. Upon initiation, the RNAPII machinery is present at both enhancer and promoters, generating eRNAs and mRNAs, respectively. Anti-sense lncRNAs are also often generated.

An independent line of research discovered that a subset of long non-coding RNAs (*lncRNAs*) possesses enhancer-like properties. These transcripts were called ncRNA-activating (*ncRNA-a*) and they also act through a Mediator recruiting mechanism (Ørom et al., 2010; Wang et al., 2011b). Other lncRNAs acting through different mechanisms have also been implicated in transcription activation, for example LUNAR1 (Trimarchi et al., 2014), HOTTIP (Wang et al., 2011b), NeST (Gomez et al., 2013). Considering their similarities, it is quite possible that some of the transcripts annotated as ncRNA-as can also be defined as eRNAs (Vučićević et al., 2015).

The enhancer-promoter tissue specific interactions are enriched within topologically associating domains (TADs Dixon et al., 2012), which will be discussed in more detail in Section 1.1.4.4.

1.1.4.2 Target promoter selection

It was established very early on that immediate proximity of an enhancer to a gene does not indicate functional relation between those two elements (Choi and Engel, 1988). In fact, the percentage of enhancers that are not associated to the neighbouring gene is currently estimated to be 21-25% (Kvon et al., 2014; Mifsud et al., 2015), including 21% of intragenic enhancers that do not regulate their host genes (Kvon et al., 2014). The hypothesis of *promoter specificity* stipulates that some inherent properties of promoters and enhancers might favor certain combinations of the two (Merli et al., 1996). This inherent property was demonstrated to lie in the underlying sequence elements. For example, Butler and Kadonaga discovered a set of *Drosophila* enhancers specific for promoters that contain either DPE or TATA box elements (Butler and Kadonaga, 2001). A more recent study in *Drosophila* used a genome-wide screening of enhancer activity (STARR-seq) (Arnold et al., 2013) and clearly demonstrated motif-based enhancer-promoter specificity (Zabidi et al., 2014). The authors established that thousands of enhancers exhibit a marked specificity either to promoters associated to developmentally regulated genes which are enriched in TATA, Inr, MTE, and DPE motifs or promoters characteristic for housekeeping genes enriched for the polypyrimidine initiator (TCT) motif. This could indicate that housekeeping and developmental genes use different enhancer-promoter recognition mechanisms or reflect independent evolution of the two types of enhancers (Lorberbaum and Barolo, 2015).

1.1.4.3 Predicting target genes

Given the large range of distances across which enhancers can act and the propensity to skip the closest neighbouring gene, pinpointing the exact gene regulated by an enhancer has proven to be challenging. Further hindering this already complex task, genes are often regulated by more than one enhancer and one enhancer can regulate more than one gene (Link et al., 2013). There are three general approaches for linking distal regulatory elements to target genes: predictions based on physical interaction, predictions based on association of enhancer sequence/activity

to gene/promoter activity, and experimental approaches.

1.1.4.3.1 Chromatin interaction maps

Over the past decade, a series of methods was developed for studying the physical interaction of different genomic regions. These *chromosome conformation capture* (3C) based methods rely on formaldehyde crosslinking of interacting chromatin fragments, thus capturing pairs of loci in sufficient proximity to be crosslinked (approximately 10–100 nm) (Dekker et al., 2002). Several versions of this approach have been developed, focusing on individual loci (3C, 4C) or genome-wide testing (5C, Hi-C, and TCC) (Dekker et al., 2002; Sanyal et al., 2012; Wit and Laat, 2012). Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is a ChIP-based method which recruits antibodies to a specific protein in order to enrich for protein-bound interacting fragments (Rusk, 2009). Capture Hi-C (CHi-C) is a recently developed approach that uses sequence capture technology to enrich a Hi-C library for annotated promoters (Jäger et al., 2015). When interpreting results from these assays, it is important to have in mind that they represent a physical interaction, not a functional one. It remains to be seen how many enhancers that are in contact with a promoter contribute to regulation of the gene transcription.

1.1.4.3.2 Association of enhancer sequence to gene activity

Expression quantitative trait loci (*eQTL*) is one of the most commonly used methods for determining the effects of genetic variants on gene expression levels (Westra and Franke, 2014). When those variants are located within a distal regulatory element, this method can infer a functional connection between the regulatory element and the affected gene. This approach looks for association between a genetic polymorphism and variation of gene expression across many different individuals (Cookson et al., 2009; Montgomery and Dermitzakis, 2011). A disadvantage of this approach is that not all enhancers have a genetic variant necessary to determine these associations. Another caveat is that associations between genomic loci and gene expression predicted by eQTL represent a mixture of direct and indirect regulation which are often not easy to distinguish (Li et al., 2013; Westra and Franke, 2014). eQTL approach is primarily used for studies of genetic variation in gene expression and will be further discussed in Section 1.2.2.2.1.

1.1.4.3.3 Association of enhancer activity to gene activity

Enhancer targets can also be predicted by matching changes in enhancer activity to changes in gene expression across tissues and/or developmental stages. PresSTIGE, for example, is an algorithm that predicts enhancer-gene linkages using H3K4me1 or H3K27ac in combination with RNA-seq data (Corradin et al., 2014). Any enhancer associated signal could be used instead of H3K4me1/H3K27ac, such as TF binding or chromatin accessibility. For example, a study by Sheffield et al. used DNase I hypersensitive sites (DHSs) as markers of regulatory elements and interrogated correlations between distal DNase I hypersensitive sites (DHSs) and gene expression across 72 cell types, predicting genes targeted by 530,000 regulatory regions (Sheffield et al., 2013). Another large-scale study by ENCODE used DHSs to identify 578,905 putative couples of distal regulatory elements and promoters across 125 cells and tissue types. These pairs were overrepresented in regions identified by 3C-based methods, confirming the validity of both types of approaches (Thurman et al., 2012). Although this approach is based on statistical associations and limited by the number of samples available for a study, the abundance of genetic and epigenetic profiles for various cell lines and tissues released by big consortia often makes it time and cost efficient.

1.1.4.3.4 Experimental approaches

Regardless of the prediction method, experimental validation of enhancer target selection is necessary for confirmation of such interactions *in vivo* or *in vitro*. For the target gene to be revealed, enhancer activity must be modified within its natural chromosomal context. A series of genomic-editing tools is now available for disrupting the activity of an enhancer. This can be done through deletion or introduction of a deleterious mutation with zinc finger nucleases (ZNFs) (Durai, 2005), transcription activator-like effector nucleases (TALENs) (Joung and Sander, 2013) or clustered regularly interspaced short palindromic repeat (CRISPR/Cas9 system) (Sander and Joung, 2014). The most recent of these methods, CRISPR/Cas9, is perhaps the most commonly used one today. Apart from disruption of sequence, enhancer activity can be modified through introduction of an arteficial chromatin modification. This can be achieved by targeting the region with a genomic-editing enzyme fused with a chromatin-modifying domain (Hilton et al., 2015). Although powerful, genomic-editing tools are labor intensive and, at this point, only suitable for validation of limited number of loci at a time.

1.1.4.4 Topologically associating domains

Genes often require protection from the unintended influence of adjacent positive or negative transcription signals. Studies in *Drosophila* identified *insulators* - orientation-independent sequence elements involved in establishing or maintaining discrete genomic boundaries (Cai et al., 1996; Udvardy et al., 1985). In vertebrates, insulators are associated with the binding of CTCF - an essential, highly conserved and ubiquitously expressed zinc-finger protein (Bell et al., 1999; Moon et al., 2005). CTCF molecules were shown to form complexes with itself and other target proteins in distal loci, resulting in formation of discrete genomic domains (Splinter et al., 2006; Wallace and Felsenfeld, 2007; Yusufzai et al., 2004). These *topologically associating domains* (TADs), as described in *Drosophila*, mouse and human, were found to be very stable across species and cell types (Dixon et al., 2012; Nora et al., 2012).

Positioning the enhancer and promoter in separate functional expression domains is proposed to be the basis of *enhancer-blocking insulators* activity (Handoko et al., 2011). In line with that, loss of TAD boundaries leads to partial fusion of the adjacent TADs, formation of ectopic contacts and long-range transcriptional miss-regulation (Lupiáñez et al., 2015; Nora et al., 2012). Apart from CTCF and cohesin, TAD boundaries were found enriched for housekeeping genes, tRNAs and short interspersed elements (SINEs), indicating that high levels of transcription might be involved in the creation of topological boundaries (Dixon et al., 2012). TADs consist of variable number of smaller 3D structures. While TAD boundaries are remarkably stable, sub-TAD structures were found to be dynamic and cell-type specific (Dixon et al., 2015; Phillips-Cremins et al., 2013).

Taken together, these studies shed light on long-range genome organisation and provide insight into the 3D context for enhancer-promoter interactions. The mechanisms behind establishing TADs are still not well understood. Further characterization of this level of genome organisation is necessary for understanding the functional impact that regulatory elements have on gene expression.

1.2 Genetic variation and phenotype variability

Genetic variation drives evolutionary adaptation and creates diversity. Evolutionary biologists King and Wilson first suggested that small regulatory changes in the genome may be the driving force behind speciation (King and Wilson, 1975). Genetic variation results from genetic drift and selection and is abundant within the human species. The latest large-scale whole genome sequencing project by the 1000 Genomes Project Consortium analyzed 2,504 genomes from 26 populations characterized 84.7 million variants, 3.6 million short insertions/deletions, and 60,000 structural variants (Auton et al., 2015; Sudmant et al., 2015). Genetic variability between populations and individuals reflects on their disease susceptibility, disease progress and response to treatment, highlighting the importance of understanding the relationship between genotype and phenotype.

1.2.1 Major types of sequence variation

There are five major types of sequence variation (Figure 1.6). Single nucleotide variants (SNVs) are single-nucleotide substitutions of one base for another between members of one species. A subset of SNVs that occurs in at least 1% of the general population are termed single nucleotide polymorphisms (SNPs) (Antonarakis, 2010). SNPs are the most common form of sequence variation in the human genome, estimated to make 90 % of all human sequence variants (Ke et al., 2008). SNPs may have two or more alleles, meaning that there are two or more occurring base-pair possibilities within a population. Every SNP can be assigned a minor allele frequency (MAF) - the frequency at which the least common allele occurs in a given population. SNPs with MAF >5% are considered *common* and the rest of them *rare*. SNP allele frequencies vary between human populations, so an allele common in one population may be rare in another (Nielsen et al., 2012; Sachidanandam et al., 2001; Wakeley et al., 2001).

Structural variants (SVs) can be broadly defined as changes in the DNA sequences that are not single base-pair substitutions (Antonarakis, 2010; Mills et al., 2006; Weber et al., 2002). These encompass INDELs (insertions and deletions), block substitutions, inversions of DNA sequences and copy number differences (Khaja et al., 2006; Kidd et al., 2008). SVs have been less explored than SNPs, as their discovery was long impeded by technical limitations. Large (>100 kb), intermediate (500 bp-100 kb) and fine-scale (1-500 bp) structural variants

Single nucleotide variant	ATTGGCCTTAACCCCGATTATCAGGAT ATTGGCCTTAACCTCCGATTATCAGGAT	Structural variants
Insertion–deletion variant	ATTGGCCTTAACCCGATCCGATTATCAGGAT ATTGGCCTTAACCC---CCGATTATCAGGAT	
Block substitution	ATTGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTAACAGTGATTATCAGGAT	
Inversion variant	ATTGGCCTTAACCCCGATTATCAGGAT ATTGGCCTTCGGGGGTTATTATCAGGAT	
Copy number variant	ATTGGCCTTAGGCCTTAACCCCGATTATCAGGAT ATTGGCCTTA-----ACCTCCGATTATCAGGAT	

Figure 1.6 Major classes of genetic variants. Five classes of major genetic variants can be described. Single nucleotide variant (SNV) is an alteration of a single nucleotide. Insertion/deletion variants (INDELs) occur when one or more base pairs are present in some genomes but absent in others. Block substitutions occur when a string of adjacent nucleotides varies between two genomes. Inversions are cases in which the order of the base pairs is reversed in a defined section of a chromosome. Copy number variants describe repetitions of identical or nearly identical sequences in some chromosomes, but not others. The figure is reproduced from (Frazer et al., 2009).

have now been described in the human genome (Khaja et al., 2006). The first high-resolution sequence map of human structural variation demonstrated that they were more abundant in the genome than originally presumed, with 40% of them assigned to copy number variations (CNVs) (Kidd et al., 2008). SVs involve both intergenic and intragenic sequences and are created predominantly by non-allelic homologous recombination (Kidd et al., 2008).

Association of alleles at different loci is not random, constituting a phenomenon called *linkage disequilibrium* (LD) (Lewontin and Kojima, 1960; Slatkin, 2008). In other words, individuals who carry a particular allele at one site often predictably carry specific alleles at other nearby variant sites (Altshuler et al., 2005). LD arises from shared ancestry of contemporary chromosomes and is influenced by natural selection, rate of recombination, rate of mutation, genetic drift and genetic linkage (Altshuler et al., 2005). A set of variants that tend to be inherited together are called a *haplotype*. Patterns of LD across the genome revealed *haplotype blocks* that vary in size from a few kb to more than 100 kb (Slatkin, 2008; Wall and Pritchard, 2003). Characterizing the patterns of LD is of great importance for studies of both genetic association and demographic history.

1.2.2 Genotype to phenotype

The connection between a genotype and phenotype was established as early as 1911 (Johannsen, 2014). Although many genetic variants are neutral, others determine phenotypic traits, from appearance and behaviour to predisposition to certain diseases. *Monogenic traits* formed the basis of “classic” genetics. They are strongly influenced by variation within a single gene and follow Mendelian laws of inheritance (Peltonen, 2006; Weatherall, 2000). *Complex traits* are much more common and are believed to result from variation of multiple genes and their interaction with environmental factors (Relton and Smith, 2010; Teperino et al., 2013). These two categories are not always mutually exclusive as complex diseases predominantly influenced by one gene have been reported (Stefansson et al., 2007; Weatherall, 2000). Although the polygenic nature of complex traits is generally accepted, the relative contribution of common variants and rare variants to complex traits is still under debate (Frazer et al., 2009; Kilpinen and Dermitzakis, 2012).

Genetic dissection of complex traits has proven challenging. Over the years, several approaches have been employed for this purpose, including linkage analysis and allele-sharing methods (based on studying relatives in a pedigree), studies of experimental crosses (of mouse and rat) and association studies that compare genotypes of affected and unaffected individuals (Lander and Schork, 1994). The development of high-throughput genotyping technologies allowed genome-wide characterization of large number of individual genomes and led to the rise of *genome-wide association studies* (GWAS).

1.2.2.1 Genome-wide association studies

Genome-wide association studies (GWAS) examine a large number of genetic variants in different individuals, looking for association with a particular trait. This approach was designed in 1996 (Risch and Merikangas, 1996), yet only became possible upon the release of the International HapMap resource, availability of dense genotyping chips and existence of large and well characterized sample cohorts (Altshuler et al., 2005; WTCCC, 2007). A typical GWAS has a *case-control design* that compares genotypes of two groups looking for SNPs that occur more frequently in individuals with a particular phenotype (case) compared to individuals without it (control) (Frazer et al., 2009). Given that the number of possible SNP variants in the human genome surpassed 149 million (as of June 2015, Sherry et al. (2001)), the GWA studies exploit

the haplotype structure of the genome and generally use customizable arrays to genotype only a subset of the variants identified as *tag* SNPs.

Since these first landmark GWAS (Ozaki et al., 2002) to August 2014, 1569 GWAS publications reported 11,115 hits for 866 diseases (Welter et al., 2014), with the numbers rising since. Although GWA studies were successful in robustly identifying common trait associated variants, those variants managed to capture only a small fraction of the heritability estimated for these complex traits (Gibson, 2010). The classical example of *missing* heritability is height, which has a heritability of 80-90%, but with common SNPs identified by GWAS explaining only 5% of it (Visscher, 2008; Weedon and Frayling, 2008). This spoke against the premise GWA studies were based on, the *common disease - common variant* hypothesis, which postulated that common genetic variants with small to modest effect sizes underlie susceptibility to most common traits (Reich and Lander, 2001). Several explanations were offered for this phenomenon. The first was that the effect sizes of individual SNPs are so small that they do not reach genome-wide significance in GWAS. This *hidden* heritability view was supported by Yang et al. who demonstrated that considering all SNPs simultaneously in one model can account for 45 percent of the variance in height (Yang et al., 2010). The second explanation was that causal variants (with presumably much lower MAF) are not in sufficient LD with SNPs on the commercial arrays that contain common *tag* SNPs (Cirulli and Goldstein, 2010; Eyre et al., 2012; Trynka et al., 2012a). Finally, current methodologies might not account for genotypic, environmental and epigenetic interactions (Gibson, 2012). Some of the mentioned scenarios may be resolved with larger sample sizes, more narrowly defined phenotypes, improved modeling and denser, whole genome sequencing-based genotyping.

Considering that the genotyped SNPs are “tagging” haplotypes on which the directly functional variants reside, fine-scale mapping is usually necessary for identifying the strongest candidate causal variants in the risk loci (Gaulton et al., 2015). The following task, pinpointing the actual causal variant(s) and elucidating the mechanism by which it influences the expression of its target gene(s), represents a major challenge of the post-GWAS era.

1.2.2.2 Genetic variation of gene expression

One approach to inferring a connection between a SNP and activity of a protein is to associate the SNP with changes in an intermediary phenotype - expression levels of the coding gene.

Several genome-wide methods address this question, including eQTL studies and allele-specific analysis.

1.2.2.2.1 eQTL

As described in Section 1.1.4.3.2, the eQTL approach looks for association between a genetic polymorphism and variation of gene expression across many different individuals (Cookson et al., 2009; Montgomery and Dermitzakis, 2011). GWAS SNPs are more likely to predict gene expression levels than MAF-matched SNPs, an observation valid across a range of complex traits (Nicolae et al., 2010). Considering the spatio-temporal specificity of many regulatory elements, integration of GWAS and eQTL analysis is expected to be informative if the eQTL was performed in the disease-relevant tissue. The Genotype-Tissue Expression (GTEx) pilot analysis encompassed nine different tissues and observed enrichment of eQTLs in SNPs associated with traits relevant for the underlying tissue. Whole blood eQTLs, for example, showed enrichment for autoimmune disease associated SNPs, but not for bipolar disorder or Type 2 diabetes (GTExConsortium, 2015). Several studies demonstrated that a limited number of eQTLs are shared across multiple tissues (20-50%) (Dimas et al., 2009; Emilsson et al., 2008). Furthermore, even if they are shared, the fold difference in expression can vary substantially between tissues (Nica et al., 2011). The GTEx pilot study, however, reported 50% of all detected eQTLs to be common to all analyzed tissues, suggesting that eQTL sharing across tissues might be more frequent than previously described. Overall, these findings demonstrate the power of the eQTL approach in prioritizing candidate GWAS SNPs, as well as the utility of an multi-tissue eQTL catalogue in pinpointing the disease-relevant tissue.

One of the surprising discoveries by GWAS was that approximately 88% of GWAS tag SNPs lie in intergenic or intronic regions (The ENCODE Project Consortium et al., 2012; Welter et al., 2014). Assuming the same is true for the linked causal SNPs, one can infer that in many cases these SNPs influence *cis*-regulatory elements. GWAS-derived SNPs falling into regulatory elements are thus often considered candidate causal variants. Large consortia have generated extensive catalogues of regulatory elements in various cell types and tissues (Romanoski et al., 2015; The ENCODE Project Consortium et al., 2012), supplemented by various studies performed in additional cell types (for example Ernst et al., 2011; Gaulton et al., 2010; Reilly et al., 2015). Numerous studies found GWAS SNPs to be enriched within non-coding

functional elements (Dimas et al., 2009; Myers et al., 2007; Stranger et al., 2007; The ENCODE Project Consortium et al., 2012). eQTLs shared across several tissues were shown to be more restricted to the TSS and have larger effect sizes while tissue-specific eQTLs were distal, more abundant and exhibited weaker effects (Dimas et al., 2009). In addition, a recent eQTL study on both SNPs and INDELs found that, for both types of variants, tissue-specific associations are more likely to act in *trans* (Huang et al., 2015). Together, these findings highlight the importance of enhancers and tissue-specific gene regulation in complex traits.

1.2.2.2.2 Allele-specific analysis

An alternative approach to eQTL is allele specific analysis of the transcriptome. This method estimates the abundance of transcripts coming from each allele within an individual and looks for *allelic imbalance* - difference in transcriptional activity between the alleles (Pastinen, 2010). The main advantage of this type of analysis is the “internal control” - the fact that the impacted and the control allele are located in the same cell and exposed to the same environmental and *trans* effects. For that reason, presence of allelic imbalance implies the existence of a regulatory variant driving the changes in transcription in *cis* (Figure 1.7). The main limitation, compared to GWA and QTL studies, is that it can not include homozygous individuals in the analysis and only genes/regions that have a polymorphism themselves are “informative” of their allelic status (Pastinen, 2010; Rozowsky et al., 2011).

The rise of single-base resolution NGS technologies enabled genome-wide assessments not only of allele-specific gene expression (RNA-Seq Pickrell et al., 2010; Rozowsky et al., 2011), but also DNA methylation (bisulfite sequencing Fang et al., 2012), DNA-protein interactions and histone modifications (ChIP-Seq Rozowsky et al., 2011; Waszak et al., 2015), chromatin states (DNaseI-Seq, FAIRE-Seq, ATAC-Seq Buenrostro et al., 2013; Degner et al., 2012) and more recently, chromatin conformation (3C-based methods Dixon et al., 2015; Visser et al., 2015).

Reports on the prevalence of allele-specific activity vary greatly between studies. Early reports of allele-specific expression suggested that a great majority of genes is expected to have allele-specific expression differences in at least some individuals (Buckland, 2004). Other studies reported 6.5% to 24% of informative genes to be imbalanced, per individual (Dixon et al., 2015; Lappalainen et al., 2013; Reddy et al., 2012; Rozowsky et al., 2011). Similarly, reported

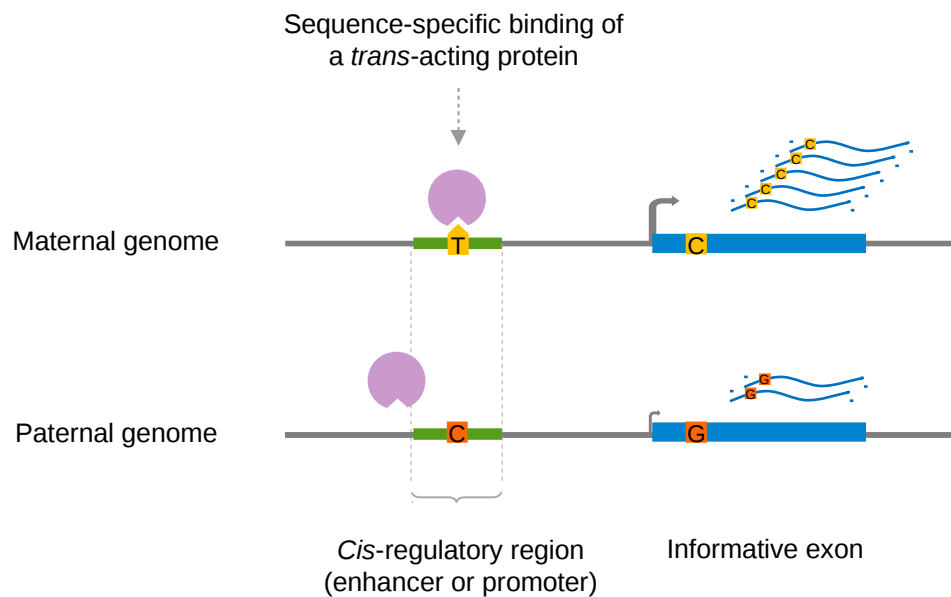


Figure 1.7 Allele-specific expression and transcription factor binding. Allele-specific expression is an imbalance in the amount of transcript generated by the two alleles in the genome. For allele-specific expression of a gene to be detectable, the gene must harbour a genetic variant that will enable assigning transcripts to their genome of origin. This *informative* variant could also be the cause of allelic imbalance, through modifying transcript stability or splicing efficiency. Alternatively, allele-specific expression can occur as a result of imbalanced activity in a *cis*-regulatory element, enhancer or promoter. A *trans*-acting protein has equal opportunity to bind to each of the regulatory alleles, yet binding in one of the alleles may be impacted by a motif-disrupting variation, causing allelic imbalance.

allelic imbalance in TF binding ranged from 2% to 25% (Kasowski et al., 2010; McDaniel et al., 2010; Rozowsky et al., 2011) 6 to 30% for histone modifications (Kilpinen et al., 2013). Despite the variation in their estimations, these studies suggest that allelic effects on activity of regulatory elements and gene expression are common across the genome.

While QTL and allele-specific analysis studies allow a direct evaluation of the impact *cis*-regulatory polymorphisms have on both gene expression and chromatin states, they do not explain the mechanism behind it. Genetic variation in sequences targeted by sequence-specific proteins presumably leads to disruption of TF binding, leading to disruption of regulatory activity. However, several studies found that the majority of allele-specific TF binding is not triggered by TF-motif disruption (Kasowski et al., 2010; Kilpinen et al., 2013; Lappalainen et al., 2013; Reddy et al., 2012). Furthermore, the majority of motif-disrupting SNPs do not show significant allelic effects (Kilpinen et al., 2013). Even though this phenomenon can be partially contributed to collaborative TF binding (Karczewski et al., 2011; Kasowski et al.,

2010; Stefflova et al., 2013), it was observed that there is a strong correlation between allele-specific events across several hundreds to few thousands of base pairs, suggesting that activity of regulatory elements might be coordinated across a larger genomic space (Kilpinen et al., 2013; McVicker et al., 2013). Following this, Waszak et al. (2015) recently described *variable chromatin modules* (VCMs) which harbour promoters and/or enhancers with coordinated chromatin states. Grubert et al. (2015) concurrently demonstrated that this genetically coordinated variation occurs through physical contacts between the coordinated regulatory elements and takes place within TADs. These studies emphasise the complexity of the *cis*-regulatory mechanisms and highlight the importance of dissecting the properties of regulatory elements within their native chromosomal context.

1.3 Pancreatic islets of Langerhans

The pancreas is a glandular organ present in all vertebrate species in functionally similar forms. Anatomically, the human pancreas is divided into four parts, all situated in a particular orientation within the abdominal cavity and referred to as the *head*, *neck*, *body* and *tail* (Figure 1.8). Functionally, the pancreas is separated into endocrine tissue dedicated to production of several important hormones and exocrine tissue that secretes digestive enzymes (Ellis, 2007; Henry Gray, 1918).

The exocrine pancreas makes up about 99% of the human pancreas by weight. It secretes the *pancreatic fluid* which is a mixture of enzymes that digest fats (pancreatic lipase) proteins (trypsinogen and chymotrypsinogen) and to some extent sugars (pancreatic amylase). It also secretes phospholipase A2, lysophospholipase, and cholesterol esterase. These enzymes are produced and secreted by *acinar* cells organised into small clusters known as *acini*. Acini are surrounded with small *ducts* that connect and ultimately converge to the large pancreatic duct (Figure 1.8). The ductal cells secrete bicarbonate ions which help neutralize the acidic content arriving to the duodenum from the stomach (Ellis, 2007). The part of the tissue dedicated to endocrine function is made of cell clusters called *pancreatic islets* or *islets of Langerhans*.

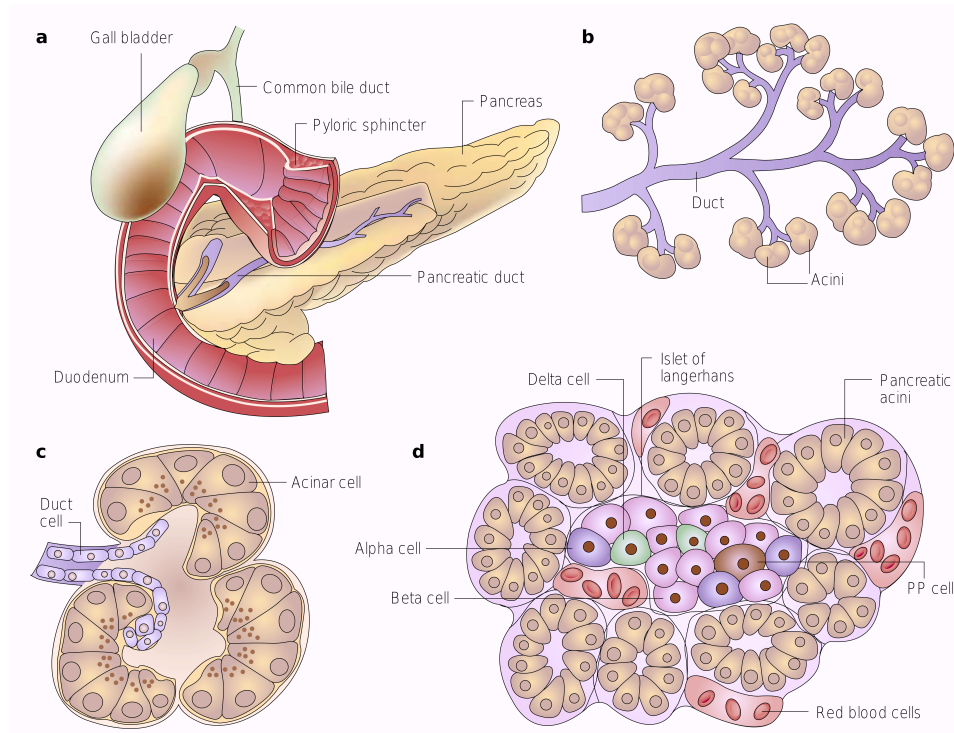


Figure 1.8 Anatomy of the human pancreas. The anatomy of a) the whole pancreas, b) the exocrine pancreas, c) a single acinus, d) a pancreatic islet embedded in exocrine tissue. Figure adapted from Bardeesy and DePinho (2002).

1.3.1 Pancreatic islet cell types

Islets are composed of five cell types, each one specialized for production of different hormones.

Alpha cells produce *glucagon*, a catabolic hormone that targets mainly liver and some kidney cells. It mobilizes glucose, fatty acids and amino acids into the blood which elevates blood glucose levels (Rutter, 2009; Suckale, 2008).

Beta cells are the most abundant cell type in the pancreatic islet and are dedicated to production of *insulin*, *C-peptide* and *amylin* (IAPP) (reviewed in Rutter, 2001).

Insulin is the major hormone produced by the pancreas and has a central role in glucose homeostasis. Insulin targets skeletal muscles and fat tissue (where it promotes absorption of glucose), liver (where it inhibits production of glucose) and hypothalamus (where it influences food intake) (Kulkarni, 2004; Rutter, 2001). Insulin receptors are found in almost all types of cells and trigger a PIP3-dependent signalling cascade when activated (Kido et al., 2001).

Insulin secretion is biphasic and is regulated by a variety of stimulatory and inhibitory glucose

metabolism-related factors (Komatsu et al., 2013; Newgard et al., 2002). Glucose is the primary modulator of insulin secretion and changes in blood glucose levels can be directly “sensed” by beta cells (MacDonald et al., 2005; Schuit et al., 2001). The glucose-sensing mechanism is essential for maintenance of glucose homeostasis and is enabled by several properties of the beta cell. First, a high capacity transporter ensures a high rate of glucose influx, proportional to blood glucose concentrations up to 10 mmol/L. Following, glucokinase mediates glycolysis in a rate proportional to the influx of glucose. The intracellular ATP:ADP ratio is increased via the citrate cycle, which causes suppression of ATP-sensitive potassium channels. This leads to membrane depolarization and an influx of extracellular calcium through voltage dependent calcium channels (VDCCs) which finally triggers exocytosis of insulin-containing secretory granules (MacDonald et al., 2005; Rutter, 2001).

Insulin secretion is the immediate response beta cells have to rise in blood glucose levels. The short-term response is mainly translational and besides insulin includes boosting production of insulin granule component proteins (Hou et al., 2009; Kulkarni et al., 2011; Leibiger et al., 1998). Long-term adaptations of beta cells to the environment might involve changes in transcriptional regulation of genes involved in control of ER load, granule counting, and cell cooperation (Ishikawa et al., 2015; Suckale, 2008).

Delta cells produce *stomatostatin* (SST), a hormone with inhibitory properties secreted by several tissues in the body. SST secretion is also stimulated by increased extracellular glucose, although the threshold concentration is higher than the one for insulin (Hauge-Evans et al., 2009; Rutter, 2009; Woods et al., 2006).

Gamma cells (or PP cells) are responsible for production of the *pancreatic polypeptide* (PP). Secretion of PP is stimulated by food intake and acts to reduce it, presumably via hypothalamus (Batterham et al., 2003; Koska et al., 2004).

Epsilon cells producing *ghrelin*, also known as the “hunger hormone”. As its name implies, ghrelin stimulates appetite by acting on hypothalamus and also by preparing the gastrointestinal system for food intake (Inui et al., 2004; Klok et al., 2007).

Taken together, the cell of the pancreatic islet form a nutrient sensing network and have a major role in the homeostasis of energy. The hormone secretion is regulated via paracrine and autocrine mechanisms and requires cooperation with the nervous and digestive system (Aronoff et al., 2004; Caicedo, 2013; Kiba, 2004).

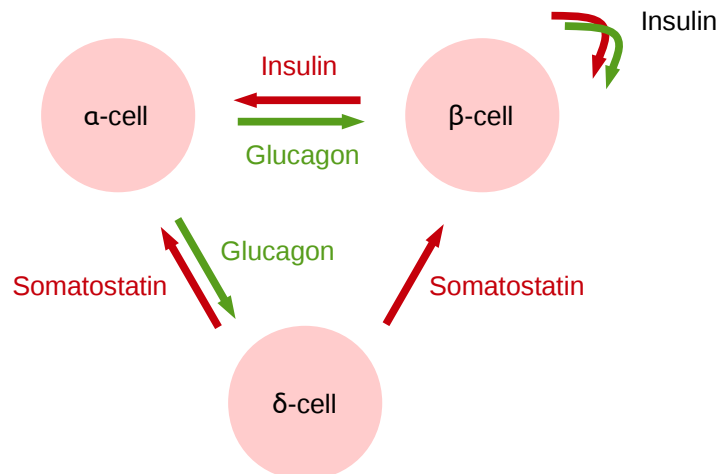


Figure 1.9 Paracrine and autocrine regulation of islet cells. Apart for their system-wide roles, pancreatic islet hormones also serve as local signalling molecules. Here is a schematic representation of putative paracrine and autocrine pathways between three major types of pancreatic islet cell types. The arrows represent the direction of the signalling pathway.

1.3.2 Transcription factors and pancreatic development

The pancreas develops from the endoderm, the innermost of the three germ layers formed during gastrulation. It originates from the gut tube, where the endodermal epithelium thickens and forms buds. These buds grow, branch and eventually form the liver, the gallbladder, and the pancreas (Gittes, 2009; Spooner et al., 1970).

The development of the pancreas can be described in a series of steps, each one characterized by a particular set of regulators (Figure 1.10). As studies of pancreas development are limited in human, much of the knowledge we have today comes from studies done in mice. Mouse pancreatic organogenesis, differentiation and maturation can be described in three major stages: primary transition (embryonic day 9.5 [E9.5] - E12.5), secondary transition (E12.5 - birth) and the postnatal period (from E16.5 to weaning, coinciding with the onset of adolescence in human).

1.3.2.1 The primary transition

Pancreatic development from the endoderm is triggered by signalling from neighbouring mesoderm-derived tissues around embryonic day 9.5 in mouse (E9.5) which corresponds to the 4th week of human gestation. The dorsal and ventral pancreatic buds emerge from the subregion of the

foregut with repressed Sonic hedgehog, followed by expression of Pdx1 and Ptf1a that together define the primordial pancreatic endoderm (Burlison et al., 2008; Chiang and Melton, 2003; Hebrok et al., 1998). Few endocrine cells are generated at this time, with the majority of cells formed being alpha cells (Herrera et al., 1991). At E11.5, the gut tube begins to rotate, which leads to proximity and fusion of dorsal and ventral buds later on at E12.5.

1.3.2.2 The secondary transition

At E12.5, the pancreatic epithelium begins to evaginate and branch into the neighbouring mesenchyme, creating a tubular structure organised into tip and trunk regions. The trunk contains progenitors that will give rise to endocrine and ductal cells, while the tip gives rise to acinar cells (Seymour, 2014; Villasenor et al., 2010). Further dedication to endocrine or exocrine lineage is decided by notch signalling (Apelqvist et al., 1999). Absence of notch signalling permits expression of Ngn3, presumably via Hnf1a (Jacquemin et al., 2000), which is both necessary and sufficient to give rise to unipotent endocrine precursors (Apelqvist et al., 1999; Schwitzgebel et al., 2000). Ngn3 then directly activates Pax4, NeuroD1 and Nkx2.2 (Schwitzgebel et al., 2000). Ngn3 negative cells that express HES-1 and p48 adopt the exocrine fate (Apelqvist et al., 1999).

Following the initiation of the endocrine program, additional transcription factors are necessary to differentiate Ngn3 expressing cells into one of the five islet cell types, such as Arx (for alpha cells) and Pax4 (for beta/delta cell fate) (Collombat, 2005; Collombat et al., 2003). Beta cell differentiation is activated by a coordinated action of Pax4 and Nkx2.2 (Wang et al., 2004). Downstream of Nkx2.2 lies Nkx6.1 whose expression during differentiation becomes restricted to beta cells only (Sander et al., 2000). Similarly, cells expressing both Pax4 and Pax6 become beta, gamma and delta cells, whereas cells expressing only Pax6 assume alpha cell identity (St-Onge et al., 1997).

Yet another array of TFs is involved in late differentiation and/or maintenance of islet cell subtype identity. Isl1 and Pdx1, for example, are active in several developmental stages, but also vital for production of endocrine hormones and the maintenance of endocrine cell mass (Ahlgren et al., 1997; Du et al., 2009; Serup et al., 1995). Pax6, originally present in all endocrine cells, may play a role in maintaining a hormone expressing phenotype of the alpha cell (Ashery-Padan et al., 2004). Mouse MafB was proposed to have a role in differentiation of

both alpha and beta cells, with preference of expression in alpha cells upon maturation. MafA becomes active in late differentiation of beta cells upon deactivation of MafB (Kataoka et al., 2004; Stalvey et al., 2006).

1.3.2.3 Postnatal period

Upon birth, the mouse endocrine pancreas undergoes remodeling through a process that involves apoptosis and beta cell replication (Bonner-Weir, 2001). In the adult pancreas, under normal conditions, the beta cell turnover is slow and apoptosis is balanced with replication of existing cells (Dor et al., 2004; Georgia and Bhushan, 2004). However, beta cell mass can be increased to compensate for increased metabolic demands that occur with pregnancy and obesity (Linnemann et al., 2014; Rieck and Kaestner, 2010). New beta cells are thought to derive predominantly by proliferation of preexisting beta cells (Dor et al., 2004; Georgia and Bhushan, 2004), although other studies have also suggested neogenesis from various cell types (Hao et al., 2006; Rosenberg, 1999; Xiao et al., 2013). Pdx1 emerged as a crucial regulator of beta cell maintenance and proliferation (Johnson et al., 2003; Kulkarni et al., 2004), although a number of TFs are implicated, such as CREB, NFAT, FoxM1, Pax4 and PPARα (Bernardo et al., 2008). Glucose, insulin and prolactin are just some of the factors that stimulate beta cell proliferation (Heit et al., 2006).

1.3.3 Diabetes Mellitus

Diabetes Mellitus is a group of metabolic diseases characterized by chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both. Long-term effects of diabetes include damage, dysfunction and failure of various organs. In 2014, 9% of adults 18 years and older were diagnosed with diabetes (World Health Organization, 2014b). In 2012 diabetes was reported as the direct cause of 1.5 million deaths. More than 80% of diabetes deaths occur in low- and middle-income countries (World Health Organization, 2014a). Without a doubt, diabetes is a disease of worldwide epidemic nature.

There are two main types of diabetes, termed Type 1 and Type 2.

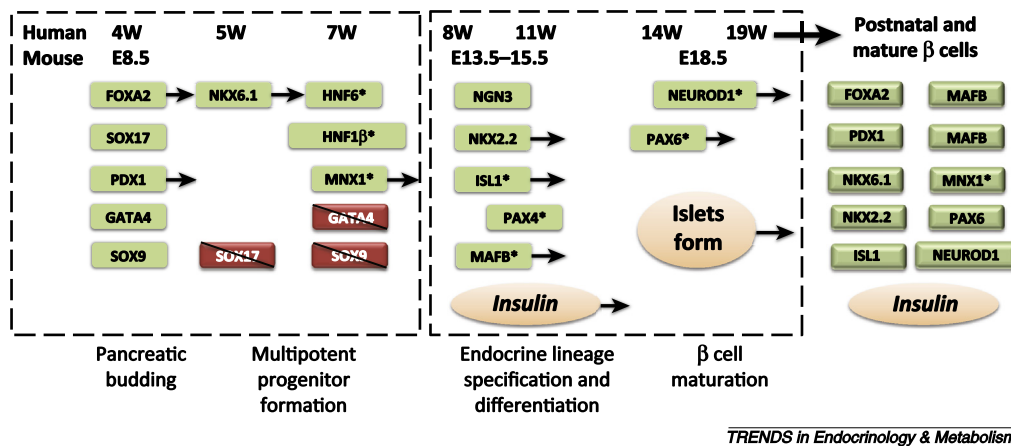


Figure 1.10 Transcription factors and beta cell development. Expression of transcription factors during development of pancreatic beta (β) cells in mouse and human. The time-line shows relative time-points in mouse (embryonic days, E) and human (weeks gestation, W). Transcription factors are listed under each time-point where their expression is first observed, or lost. Transcription factors with arrows extending denote expression that persists into postnatal and/or mature beta cells. (*) Expression of the transcription factor was derived from whole pancreas mRNA, thus cell type-specificity was not determined. The figure is reproduced from Conrad et al. (2014).

1.3.3.1 Type 1 diabetes

Type 1 diabetes (T1D, insulin-dependent *Diabetes Mellitus*) is a consequence of organ-specific autoimmune destruction of insulin-producing beta cells. Loss of blood glucose control can lead to both acute (ketoacidosis, severe hypoglycemia) and secondary complications (heart disease, blindness and kidney failure) (Gubitosi-Klug, 2014; Maahs and Rewers, 2006). The onset of T1D was once thought to typically occur in childhood, yet more recent numbers report a large number of adult onset T1D cases (Bluestone et al., 2010; Harjutsalo et al., 2008). T1D results from a combination of genetic predisposition, environmental factors, and stochastic events (Lorenzen et al., 1994; Pociot and McDermott, 2002; Reddy et al., 2011). Human leukocyte antigen (HLA) encodes cell surface proteins that interact with immune cells was estimated to contribute up to 40% of T1D risk (Erlich et al., 2008; Risch, 1987). The latest T1D GWA studies identified 44 risk loci with a strong enrichment of SNPs in enhancers of immunologically relevant tissues (Onengut-Gumuscu et al., 2015). Several immunological cell types were implicated in T1D pathogenesis, but the primary mediators of beta cell destruction were identified to be T lymphocytes (Roep, 2003; Wallet and Tisch, 2006).

1.3.3.2 Type 2 diabetes

Type 2 diabetes (T2D), also known as non-insulin-dependent *Diabetes Mellitus* (NIDDM) is an array of diseases characterized by chronic insulin resistance and loss of beta cell mass and function (DeFronzo, 1992; Stumvoll et al., 2005). Over 90% of all diabetes cases are T2D. In the early stage of Type 2, insulin resistance is manifested as an inappropriate response to insulin by cells in the body which leads to increase of blood glucose, impairment of metabolism and eventually causing increased intracellular fat content in several tissues. As a response to high blood glucose, beta cells increase the production of insulin, leading to high blood insulin level. In some individuals, the beta cell mass adapts and compensates for the increased metabolic load. In others, they are not able to keep up with the rising demand for insulin. In those individuals, this leads to a reduction in insulin secretion which marks the onset of mild and progression to severe T2D (DeFronzo, 1992).

Beta cell failure has been proposed to result from increased beta cell apoptosis (Butler et al., 2003; Rhodes, 2005), possibly induced by a combination of chronic hyperglycemia (Donath et al., 1999), hyperlipidemia (Alonso et al., 2014), and inflammatory cytokines (Donath et al., 2009) that disrupt beta cell growth and survival signalling pathways. T2D is a multiorgan disease, with defects in muscle tissue (Minokoshi et al., 2003), liver (Michael et al., 2000), fat and the central nervous system (Sandoval et al., 2009) contributing to disease pathophysiology. It is also a complex disease with a genetic and lifestyle component. Obesity is a major factor in T2D pathophysiology, followed by lack of physical activity, poor diet, stress, and urbanization (Cheema et al., 2014; Salas-Salvadó et al., 2011). Food high in sugar, saturated fats and trans fatty acids is also associated with increased risk of T2D (Imamura et al., 2015; Riserus et al., 2009).

The strong genetic basis of T2D is supported by multiple lines of evidence. The concordance of T2D is 20-30% in dizygotic twins and reaches 70% in monozygotic ones (Kaprio et al., 1992). The lifetime risk of developing the disease is approximately 40% in offspring of one Type 2 diabetic parent and approaching 70% if both parents have diabetes (Groop et al., 1996; Lyssenko et al., 2005). In fact, properties reflecting the pancreatic function (insulin sensitivity and insulin secretion) were shown to be heritable in non-diabetic individuals, as well (Elbein et al., 1999; Vaag et al., 1995). A number of T2D relevant loci emerged through linkage or candidate gene approach studies, such as *CAPN10* (related to glucose transport in muscle),

HNF4a, *PTPN1* (negative regulator of insulin signalling), *CASQ1*, *AMP1*, *TCFL7* (related to Wnt regulation of proglucagon), *PPARGgamma* (target of insulin sensitizers), *KCNJ11*, *SUR1*, *ENPP1* and *WFS1* (Ridderstråle and Groop, 2009; Thomsen and Gloyn, 2014).

Throughout the GWAS era, T2D diabetes was the centre of numerous studies across populations (Cho et al., 2012; Cornelis et al., 2009; Dupuis et al., 2010; Hara et al., 2014; Imamura et al., 2012; Kooner et al., 2011; Locke et al., 2015; Long et al., 2012; Mahajan et al., 2014; Morris et al., 2012; Palmer et al., 2012; Parra et al., 2011; Saxena et al., 2012, 2013; Scott et al., 2006; Shu et al., 2010; Sladek et al., 2007; Steinthorsdottir et al., 2014; Thompson, Ian M, Lucia et al., 2006; Voight et al., 2010; Zeggini et al., 2007). GWA studies confirmed many of previously identified loci and highlighted new ones, such as fat mass and obesity associated *FTO* (Scott et al., 2007). To date, these studies combined have identified T2D-associated variants mapping to more than 80 loci. Broad functional characterization of Type 2 diabetic human islets suggested that many of these variants have a role in regulation of glucose levels, insulin secretion and/or action, demonstrating that they impact primarily beta cell function rather than insulin sensitivity (Grarup et al., 2010; Rosengren et al., 2012). Understanding the biological mechanisms behind T2D has been impeded by difficulties in defining the precise causal variant and effector transcripts. This is largely due to the fact that the vast majority of association signals map to non-coding regions, whose architecture we are only starting to unravel. This warrants the need for a detailed characterization of the *cis*-regulatory landscape and the transcriptional programme of pancreatic beta cells.

1.3.3.3 Gestational diabetes

Gestational diabetes (GDM) is diagnosed when the blood glucose of a pregnant women falls in the upper 5-10% of the population distribution. The hyperglycemia originates from an autoimmune beta cell dysfunction (a minority of the cases), a genetic predisposition to impaired insulin secretion or a beta cell dysfunction associated with chronic insulin resistance (Metzger et al., 2007). The women that suffered through GDM are in high risk from developing postpartum T1D or T2D (Bertoni et al., 2010; Xiang et al., 1999). These consequences extend to the offspring, as well. Children and young adults that were exposed to GDM *in utero* have a higher prevalence of obesity, diabetes, and impaired glucose tolerance (Dabelea, 2007; Malcolm et al., 2006; Pettitt et al., 1985).

1.3.3.4 Monogenic forms of diabetes

1.3.3.4.1 Maturity onset diabetes of the young

Maturity onset diabetes of the young (MODY) is a group of monogenic disorders characterized by autosomal, dominantly inherited, non-insulin dependent form of diabetes. MODY makes 1% of all diabetes, though this might be an underestimation as it is frequently misdiagnosed as T1D or T2D. It typically manifests before the age of 25 and in some rare cases as late as the age of 50. The described forms of MODY are all due to insults in insulin production or release which occur as a result of mutations in beta cell-relevant genes (Table 1.2). About half of patients classified as MODY have no mutation in the known MODY gene (Clearinghouse, 2014).

Table 1.2 Forms of maturity onset diabetes of the young (MODY).

Form	Mutated gene	References
MODY1	HNF4 α	Yamagata et al., 1996
MODY2	GCK	Froguel et al., 1993, 1992
MODY3	HNF1 α	Frayling et al., 1997, Ellard, 2000; Vaxillaire et al., 1997
MODY4	PDX1	Stoffers et al., 1997
MODY5	HNF1 β	Lindner et al., 1999; Horikawa et al., 1997
MODY6	NEUROD1	Kristinsson et al., 2001; Gonsorčíkova et al., 2008
MODY7	KLF11	Bonnefond et al., 2011
MODY8	CEL	Johansson et al., 2011
MODY9	PAX4	Sujitjorn et al., 2015; Jo et al., 2011; Plengvidhya et al., 2007
MODY10	INS	Edghill et al., 2008
MODY11	BLK	Bonnefond et al., 2013; Borowiec et al., 2009

1.3.3.4.2 Neonatal diabetes

Neonatal diabetes is a form of diabetes that occurs in the first 6 months of life. It is very rare and occurs in 100,000 to 500,000 live births. It can be *permanent* (caused by mutations in *KCNJ11* (~30%), *ABCC8* (~19%), *INS* (~20%), *GCK* (~4%), and *PDX1* (<1%)) (Nicolino et al., 2010;

Njølstad et al., 2003; Polak and Cavé, 2007; Sagen et al., 2004) or *transient* (brought on by mutations in *ABCC8* or *KCNJ11* gene or disruption of imprinting in *ZAC/HYMAI* locus) (Ioannou et al., 2011; Shim et al., 2015; Temple and Shield, 2002).

1.4 Mouse as an experimental model

A major part of our knowledge about fundamental properties of living cells, organisms and disease has come from studies conducted in models. A wide range of models include *in vivo* (in a whole, living organism), *in vitro* (in a controlled environment outside of a living organism) and *in silico* models (performed on computer or via computer simulation). When it comes to *in vivo* studies, invertebrate organisms, including the fruit fly and the nematodes are excellent models for examining the signalling pathway or protein interactions. Vertebrates such as the mouse, zebrafish, frog, chicken and in some cases primates, provide insight into human development and disease (Bier and McGinnis, 2008).

The mouse, a small mammal of the order *Rodentia*, is the most commonly used mammalian research model. Several reasons propelled this organism to the top of the biomedical research. As a mammalian species, mice share the developmental overview, cell types and tissue organisation with humans. Mice-derived material is easily accessible across developmental points and can serve as a source for primary culture cells. Mice are small, have a short reproductive cycle (20 day gestation), a large number of offspring (3-12 pups) and a life span of two to three years. They are fairly easy to handle, nourish and maintain in a controlled environment. Furthermore, mice succumb to a range of Mendelian and polygenic and complex diseases that impact humans, including diabetes.

1.4.1 The mouse model in genetic research

The mouse model is particularly relevant for the field of genetics and complex diseases. Mouse genetics started over a century ago and has since accumulated a wealth of resources and tools for dissecting the basis of human development and disease (Eppig et al., 2015).

The mouse genome is well characterized, with the first genome draft released in 2002, shortly after the human genome (Waterston et al., 2002). Comparison of the two genomes revealed that the mouse one is somewhat smaller (2.5Gb vs 2.9Gb), but surprisingly well conserved.

Approximately 80% of the mouse coding genes have a direct 1:1 orthologue in human. Furthermore, more than four out of five conserved human/mouse elements appear to reside in non-coding DNA. The conservation was also evident in the linear organisation of the DNA, with 217 conserved syntenic blocks identified (mean length 23.2 Mb) (Waterston et al., 2002).

Traditional mouse genetics was based on inbred strains which are derived by generations of inbreeding and are homozygous at all loci. The first inbred strain of mice was generated by Abbie Lathrop in 1910 (Steensma et al., 2010). Today, hundreds of inbred strains are available, accompanied by catalogues of exhaustive phenotypic information. Existence of these strains, combined with the syntenic conservation of the mouse and human genomes, gave weight to extensive QTL mapping in mouse (McPeck, 2000).

Several types of inbred-derived strains can be used for this purpose (Peters et al., 2007). *Coisogenic strains*, for example, come from inbred strains that develop spontaneous mutations. A classic example of coisogenic strains are *ob/ob* and *db/db* mice that linked the leptin (*ob*) and leptin receptor genes (*db*) genes with obesity and diabetes before the exact mutations underlying the phenotype were discovered (Chen et al., 1996; Coleman, 1978, 1982). Other inbred-derived strains are generated through complex breeding pattern and include *recombinant inbred* (have a fixed combination of the original parental genomes), *congenic strains* (control for differences in phenotype due to strain background), *consomic strains*, *genome tagged mice*, and *advanced intercross strains* (Flint and Eskin, 2012; Peters et al., 2007).

Most commonly used laboratory strains of mice are *classical inbred strains* - hybrids generated by breeding between several subspecies of the *Mus musculus* species (Figure 1.11), predominantly of *Mus musculus domesticus* origin (92%) (Wade et al., 2002; Yang et al., 2007). C57BL/6J (the “black” mouse) is the mouse reference genome strain and one of the more commonly used laboratory strains (Dejager et al., 2009; Yoshiki and Moriwaki, 2006). *Wild-derived* strains were established from progenitors of either *Mus musculus musculus*, *Mus musculus castaneus* or *Mus spretus*, which were trapped in the wild. These strains are smaller than classical laboratory strains, and have retained features of their behaviour from the wild (Dejager et al., 2009; Yoshiki and Moriwaki, 2006). More importantly, these strains contain more genetic diversity (Frazer et al., 2007). Whole genome sequences of 17 inbred strains were reported in 2011, including four wild-derived inbred strains; CAST/EiJ, PWK/PhJ, WSB/EiJ and SPRET/EiJ (Keane et al., 2011), creating a catalogue of great use for studies focused on exploring the relationship between genotype and phenotype.

The genetic diversity combined with the reduced genome complexity makes mice a useful surrogate for GWA studies. Compared to the studies done in human, mouse GWA studies require a much smaller number of samples and were found to describe a greater percentage of the phenotypic variance for a particular trait (Burke et al., 2012). However, GWA studies are based on the assumption of independence of individuals. This is not the case with mice strains whose population structure reflects their breeding history and must be addressed with appropriate statistical methods (Flint and Eskin, 2012; McPeck, 2000; Wade et al., 2002).

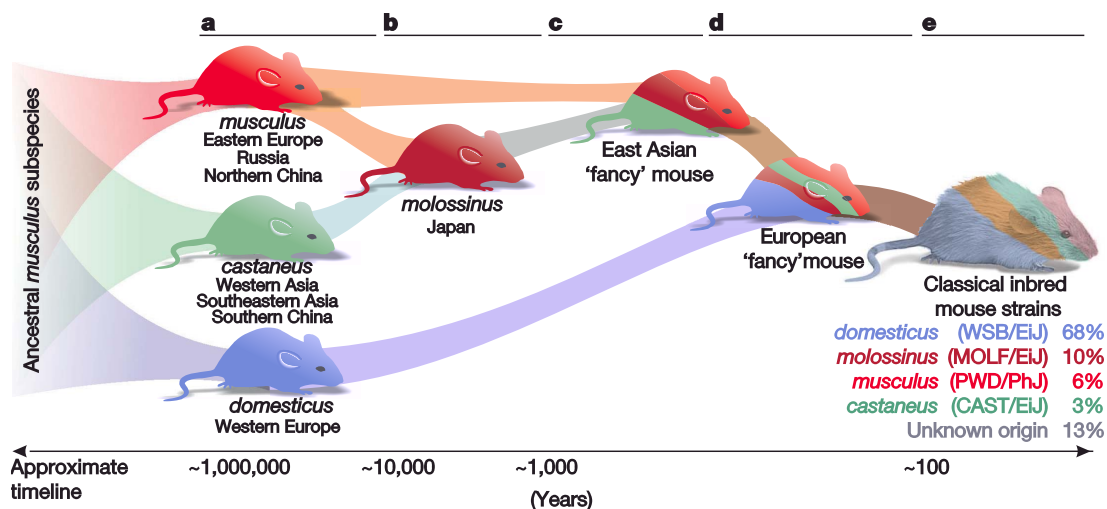


Figure 1.11 Origin of classical inbred mouse strains. Classical inbred strains were generated from several evolutionary distinct subspecies of *Mus musculus*. a) *Mus musculus musculus*, *castaneus* and *domesticus* diverged from a common ancestor approximately one million years ago. b) *Mus musculus molossinus* is a more recent hybrid between *musculus* and *castaneus*. c) East Asian “fancy mouse” was bred in the 18th century from various Asian *Mus musculus* subspecies. d) European “fancy” mice derived from the Asian “fancy” mouse and local *Mus musculus domesticus*. e) Modern classical inbred strains are a result of inbreeding programs, with various relative contributions of the *Mus musculus* subspecies to their haplotypes. The figure is reproduced from Frazer et al. (2007).

An array of techniques available for manipulation of genetic material opened a way for mice to be used for functional testing of loci identified in human GWAS. Genetically modified animals can carry knockout alleles, overexpressed alleles, dominant-negative alleles and point mutations (Doyle et al., 2013). For example, one such study focused on a common SNP associated with cancers of the prostate, large bowel, and thyroid. Sur et al. (2012) generated a mouse in which an entire conserved enhancer element was conditionally deleted to confirm a direct functional effect of the enhancer on tumorigenesis. In another example, a recent study from Multhaup et al. took an interesting approach to identifying 30 candidate genes involved in obe-

sity and T2D through epigenetic mechanisms. In short, they examined the DNA methylation changes in adipocytes induced by high-fat diet in both mouse and human and then overlapped the conserved implicated regions with GWAS loci, linking obesity-derived phenotypes with human T2D (Multhaup et al., 2015). While the utility of the mouse model is clear, there are several caveats to its use. Firstly, studies using genetically modified mice are restricted to the phenotypic spectrum of 129 and C57BL/6 background, as embryonic stem cells for homologous recombination are only available for those strains (van der Weyden et al., 2011). Secondly, and more importantly, species-specific aspects of tissue function between mice and human do exist. Therefore, interpretation of results observed in the mouse model and extrapolation to human must be made with these differences in mind.

1.4.2 The mouse model in diabetes research

Through the years, the mouse has been a major source of knowledge about pancreas biology. Mouse models have been extensively used in dissecting the pancreatic development (reviewed in Benitez et al., 2012; Jørgensen et al., 2007), transcription regulation (Bramswig and Kaestner, 2014), diabetes (Atkinson and Leiter, 1999; King, 2012) and other pathological processes (Bardeesy and DePinho, 2002; Herreros-Villanueva, 2012).

1.4.2.1 Comparison of mouse and human pancreatic islets

Core developmental processes and the function of mature pancreatic islets - the stimulus coupled secretion of endocrine hormones - are well conserved between the mouse and human (King, 2012). However, the islets of these two species differ in a number of aspects, including cellular composition and organisation, vasculature, innervation, glucose sensing, insulin secretion and the proliferative capacity of the beta cell.

1.4.2.1.1 Cellular composition and organisation

The differences in organisation of hormone secreting cells between mouse and human islets have long been acknowledged (Brissova et al., 2005; Cabrera et al., 2006; Eizirik et al., 1994). Mouse islets are composed of 70-80% beta cells, 10-20% alpha cells, and <10% delta cells as opposed to 50-60% beta cells, 20-40% alpha cells, and <10% delta cells that make the human

islet. Mouse islets are organised in a *core-to-mantle* fashion where beta cells are clustered in the islet centre, surrounded by a *mantle* of the remaining cell types. Human islets, on the other hand, lack this structure and have beta cells dispersed between other cell types. However, human islets also tend to vary in size and smaller islets were recently described to have cellular composition and structure similar to the one found in mouse (Bonner-Weir et al., 2015; Bosco et al., 2010; Farhat et al., 2013).

1.4.2.1.2 Vasculature and innervation

Both islet vasculature and innervation are considered to play a vital role in beta cell survival and function. Early reports tend to emphasise differences between these aspects in the two species. Mouse, unlike human, was reported to have an increased vascular and nerve density in islets compared to the exocrine pancreas (Dai et al., 2013; Rodriguez-Diaz et al., 2011). Furthermore, sympathetic neurovascular contacts were originally found only in human islets (Rodriguez-Diaz et al., 2011), even though this is disputed by a later study (Chiu et al., 2012). Similarly, reported differences between vascular patterns (Bosco et al., 2010) were contradicted by findings of others (Bonner-Weir et al., 2015) that report blood vessels reaching beta cells in both mouse and human, not just human.

1.4.2.1.3 Glucose sensing and insulin secretion

The kinetics of glucose-stimulated insulin secretion was shown to be similar in mouse and human islets (Henquin et al., 2006). The threshold and magnitude of insulin secretion, however, is much higher in mouse than in human islets (Dai et al., 2012; Henquin et al., 2006; Rorsman and Braun, 2013). The differences go down to transcriptional and protein level. The principal glucose sensing protein in mouse is proposed to be GLUT2. In human, however, GLUT2 is not highly expressed and it is unclear if that role belongs to GLUT2 or GLUT1 (De Vos et al., 1995; Ferrer et al., 1995; McCulloch et al., 2011; Schuit, 1997). Finally, mice have two genes encoding for insulin (*Ins1* and *Ins2*), while humans have one (Bell et al., 1980; Wentworth et al., 1986).

1.4.2.1.4 Beta cell proliferative capacity

As loss of beta cells is a hallmark of both T1D and T2D, understanding the mechanisms behind beta cell proliferation is one of the major aims of diabetes research. Proliferation of beta cells in both mouse and human was shown to be a function of age, with the peak of proliferation at neonatal period (Chen et al., 2011; Halter and Chang, 2003; Tschen et al., 2009). However, while in humans the maximum rate of proliferation is 2%, in mice it reaches 20% (Finegood et al., 1995; Meier et al., 2008). Additionally, islets of aged mice retain detectable proliferative capacity and will proliferate if exposed to the right cue (Stolovich-Rain et al., 2012). A wide array of mitogenic agents and small molecules that induce proliferation of beta cells in mouse, fail to do so in human (reviewed in Wang et al., 2015), calling for further research on the subject. However, an important thing to have in mind when interpreting results or designing new studies is that most research is performed on mice aged 8-12 weeks (young adult) whereas the human islet samples generally come from individuals over 40 years of age (Ihm et al., 2006; Wang et al., 2015). Together, these considerations create a need to understand the pathways that define species-specific competence for beta cell proliferation, as they could signal targets to induce human beta cell growth.

1.4.2.2 The mouse model and interpretation of diabetes GWAS

Mice may be a useful model for following up on GWAS loci found to be associated to pancreas-related diseases. Although numerous previously unknown susceptibility loci emerged from T2D association studies, only a handful were followed up by functional studies in the mouse model (reviewed in Cox and Church, 2011).

Thus far, the most extensive research has been performed on the TCF7L2 loci, a gene encoding for a component of TCF, a factor involved in the control of cell growth downstream of Wnt signalling with context dependent activator/repressor function (Jin and Liu, 2008). TCF7L2 loci harbours SNPs with the one of the strongest and consistent associations with T2D ever reported (Dupuis et al., 2010; Grant et al., 2006; Kooner et al., 2011; Mahajan et al., 2014; Morris et al., 2012; Zeggini et al., 2007). The risk allele is associated with impaired insulin secretion, insulin processing and decreased sensitivity to incretin in T2D patients (Lyssenko et al., 2007; Villareal et al., 2010). *In vitro* studies linked TCF7L2 silencing to impaired function and apoptosis of beta cells (da Silva Xavier et al., 2009; Shu et al., 2008). Mouse models have been extensively

used to further elucidate the role of TCF7L2 *in vivo*. Tcf7l2 is not considered to have a role in pancreatic development, as whole body knockout mice (Tcf7l2 ^{-/-}) have apparently normal pancreas upon birth (Boj et al., 2012; Korinek et al., 1998). A pancreas-specific Tcf7l2 ^{-/-} demonstrated a glucose intolerance phenotype that was exacerbated upon exposure to high-fat diet (da Silva Xavier et al., 2012). A beta cell-specific knock out of Tcf7l2, however, casted doubt in beta cells being the tissue of interest when it comes to glucose homeostasis (Boj et al., 2012), a discrepancy that could be the result of a heterogeneous role of Tcf7l2 across tissues and developmental time points. Savic et al. opted for a transgenic approach and designed a mouse containing human TCF7L2 regulatory sequences in a copy-number allelic series. With this model, they demonstrated a direct, dose-dependent role effect of altered Tcf7l2 expression on glucose metabolism (Savic et al., 2011). They also found that the introduced enhancers are active in intestine and neonatal islets, but not in adult islets, supporting the heterogeneous role of Tcf7l2 in glucose homeostasis (Savic et al., 2011).

The example of Tcf7l2 demonstrates the versatility of the mouse model in diabetes research. This model facilitates assessing different phenotypes associated with diabetes, such as obesity, hyperglycaemia, hyperinsulinaemia and insulin resistance. Whole body or tissue-specific KO can reveal a developmental role of the investigated transcript or its involvement in a particular signalling pathway. At the same time, the KO models enable the identification of potential pleiotropic effects of the knockout mutations. The mouse model enables the study of interaction between the genetic sequence and the environment, such as high-fat diet. In conclusion, despite the differences addressed in Section 1.4.2.1, the mouse model has a major and meaningful place in diabetes research.

1.5 Thesis aims

This thesis addresses the questions of genetic and epigenetic regulation of gene expression in pancreatic islets in two chapters.

The specific aims of the first chapter are:

1. To map the chromatin signatures and target sites of representative beta cell transcription factors in the mouse pancreatic islets
2. Characterize the mouse islet *cis*-regulatory elements and their underlying sequences
3. Identify common and species-specific aspects of the mouse and human islet regulatory architecture

The specific aims of the second chapter are:

1. To construct a pipeline for allele-specific analysis of RNA-Seq, ChIP-Seq and ATAC-Seq focused on allelic events driven by *cis*-regulatory variation
2. Create a genome-wide map of allelic expression and chromatin states in pancreatic islets
3. Characterize genes susceptible to *cis*-regulatory variation
4. Define properties of *cis*-regulatory variation associated with allelic expression

Chapter 2

General Methods

2.1 Animal work

2.1.1 Mouse maintenance

C57BL/6J mice originated from Charles Rivers (#000664) and the CAST/Ei from MRC Harwell (#E1510). Local mouse colonies were established and housed two to seven animals per cage, in a specific pathogen free facility (SPF) with 12 hour (h) light/dark cycle and free access to standard mouse diet (Global 14% Protein Rodent Maintenance Diet, #2014 Teklad). F1 mice generated for the purposes of research in Chapter 4 (Figure 4.1) were randomly exposed to either control (20% protein, #D10123103, Research Diets Inc.) or low-protein diet (8% protein, #D10123102, Research Diets Inc.) *in utero* and during lactation period.

C57BL/6J, commonly referred to as the black mouse, is a classical inbred mouse strain. In contrast, CAST/Ei is a strain derived by inbreeding wild *Mus musculus castaneus* mice (discussed in Section 1.4.1). CAST/Ei is characterised by aggressive behaviour and stress response to changes in the environment, such as light, noise levels, diet or manipulation. In addition, CAST/Ei is a poor breeder with tendencies of infanticide. In order to generate the number of mice sufficient for performing the assays described in Chapter 4, crosses were performed through three years, ultimately yielding 180 male F1 mice.

2.1.2 Islet isolation

The mouse pancreatic islet isolation was performed following the protocol developed by Maestro et al. (2001), with minor modifications. For the purposes of the reference mouse regulome (Chapter 3) the islets were isolated from 12-14 weeks old male C57BL/6J mice, in groups of 3-8 animals. For the purposes of the experiments performed in Chapter 4, the islets were isolated from 16 weeks old male mice, in groups of 3-8 animals of the same cross.

The mice were anesthetized by inhalation of 2% isoflourane (Forane isofluranum, Abbott Laboratories Ltd) and sacrificed by decapitation. 1X Cold Hanks' balanced salt solution (HBSS) was prepared from 10x HBSS (Invitrogen, #14060-040) and 10mM HEPES (Sigma, #H7523-50G), with the pH adjusted to 7.4 with 1 molar (M) NaHCO₃. Following, 1 milligram per millilitre (mg/mL) of Collagenase P (Roche, #11-213-865 001) was prepared in 1 x HBSS and left to solubilise on ice. The mouse pancreas was cannulated via the main pancreatic duct and perfused with 2.5mL of cold Collagenase P solution, removed from the mouse and digested for 10 minutes (min) at 37 °C, with continuous agitation. The samples were placed on ice and washed with 15mL of ice cold HBSS-BSA (1x HBSS, 10mM HEPES, 0.5% BSA, pH7.4 adjusted with 1M NaHCO₃) and resuspended by a 20 gauge (G) needle (Beckton Dickinson, #305175). The cell suspension was washed 4x with ice cold HBSS-BSA, each time sedimented for 3min. The samples were resuspended in 15mL of ice cold HBSS-BSA and centrifugated at 650xg for 15min at 4 °C. The histopaque mix was prepared by mixing Histopaque 1077 (#11191 Sigma) and Histopaque 1119 (#11191 Sigma) in the ratio of 7:3. The islets were purified by a discontinuous gradient of 7mL of the histopaque mix as the bottom phase and 7mL of HBSS-BSA as the top phase. The gradients were separated with centrifugation at 1000xg for 20min at room temperature, with deceleration breaks removed. The islet-containing interphase was collected, washed in ice cold HBSS-BSA and pelleted by centrifugation at 1000xg for 4min at 4 °C. The islets were hand picked under a stereoscopic zoom microscope (SMZ1500 Nikon) and cultivated in nonadhesive plates in Roswell Park Memorial Institute medium (RPMI 1604, #be12-702F, Lonza) supplemented with 10% Fetal Bovine Serum (FBS #14-801, Cambrex), 2mM L-glutamine and penicillin and streptomycin to final concentrations of 0.5U/mL and 0.5µg/mL, respectively (#17-745E, Lonza). Approximately 100 islet equivalent number (IEQ) were plated in 5mL of medium. On the third day, the islets were collected, pooled and used fresh for ATAC or stored for RNA isolation or ChIP-Seq, as described in 2.2.

All experiments were conducted in the August Pi i Sunyer Biomedical Research Institute (IDIBAPS) in Barcelona, Spain, according to the guidelines for animal care and handling approved by the Ethics Committee on Animal and Human Research of the University of Barcelona and Generalitat de Catalunya (DARP: 8624).

2.2 Molecular biology methods

2.2.1 RNA isolation

Isolation of RNA was performed from pools of approximately 300 IEQ, derived from minimum three mice, that were collected and snap frozen upon cultivation (described in Section 2.1.2). The isolation was done with TRIzol (Life Technologies, #15596), following the guidelines recommended by the manufacturer. In short, the islet pellet was resuspended in 1mL of the TRIzol reagent, vortexed and incubated on ice for 5min. The cell suspension was passed through a 30G needle (Beckton Dickinson, #304000) until no trace of islet tissue was visible. 200 microlitre (μ l) of chloroform (Sigma-Aldrich, #288306) was added, vortexed and incubated for 3min on ice. The phases were separated by centrifugation at 13400xg for 15min at 4 °C, the water phase was collected and mixed with an equivalent volume of 2-propanol. After 10min incubation on ice, the samples were centrifuged at 13400xg for 15min at 4 °C, the supernatant discarded and the pellet rinsed with 1mL of ice cold 75% EtOH. The washed pellets were air dried and dissolved in 10 μ g of nuclease-free H₂O. The concentration was determined using Qubit (RNA BR Assay Kit, #Q10210) and the quality was checked with the Eukaryote Total RNA Nano assay on the Agilent Bioanalyzer 2100 Expert (Agilent Technologies). Samples with RINs >8 were considered suitable for sequencing.

2.2.2 ChIP

The chromatin immunoprecipitation protocol described here was based on the protocol used by Maestro et al. (2001) and optimized for use with lower starting amount of islets. Optimisation included prolonging the thawing of frozen islets from 10min to 30min and passing the lysate through a needle to make sure chromatin is fully solubilised, modifying sonication conditions and introducing tRNA to the immunoprecipitation reaction as a carrier nucleic acid. The number of islets and the amount of antibody used per reaction depends on the protein targeted for immunoprecipitation and are listed in Section 3.2.1.

2.2.2.1 Islet fixation

After cultivation (as described in Section 2.1.2), islets from a minimum of three mice were hand picked and pooled into aliquots approximately 300 islet equivalents (IEQ). The islets were washed 3x with 1mL of cold phosphate-buffered saline (PBS) (centrifugation done at 100xg for 1min at 4 °C), and fixed at room temperature for 10min in a total volume of 1mL of PBS with 1% (w/v) formaldehyde (37% HCHO–10% methanol stock solution; Merck). The reaction was quenched by adding glycine to the final concentration of 0.125M and incubating for 5min. The islets were then pelleted by centrifugation at 1000xg for 2min in 4 °C and washed 3x with cold PBS containing 1x proteinase inhibitor (PI) (cOmplete EDTA-free Protease Inhibitor Cocktail Tablets, Roche #4693132001). Supernatant-free pellets of fixed islets were snap frozen in liquid nitrogen and stored in -80 °C.

2.2.2.2 Chromatin preparation

A total amount of 300-1200 IEQ was re-suspended in 300 μ l of the ChIP Lysis Buffer (2% Triton X-100, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH 8.0, 1mM EDTA, 1x freshly added protease inhibitor cocktail) and lysed on ice for 30min with gentle shaking. The lysed islets were passed through a 30G needle until there was no visible pieces of islets left. In order to obtain chromatin of fragment sizes 150-400bp long, the lysate was sonicated in ice cold water (Bioruptor UCD-200, Diagenode), for a total of 15min, with 1 minute cycles of 40sec ON + 20sec OFF and strength setting “HIGH”. The debris were cleared by centrifuging at 13400xg for 5min in 4 °C and the supernatant containing sonicated chromatin was collected.

2.2.2.3 Chromatin pre-clearing

The sonicated chromatin was diluted with 4 volumes of the ChIP Dilution Buffer (50mM HEPES (pH8.0), 140mM NaCl, 1mM EDTA (pH8.0), 0.75% Triton X-100, 0.1% Na-deoxycholate, 1x freshly added protease inhibitor cocktail). The sepharose beads were prepared by mixing equal volumes of sepharose A and sepharose G beads (Protein A Sepharose CL-4B and Protein G Sepharose 4 Fast Flow from GE) and washing 3x in 1mL of cold ChIP Working Buffer (prepared by mixing the Lysis Buffer and 5x Dilution Buffer in the ratio 1:4). The beads were pelleted by centrifugation at 400xg for 2min in 4 °C and the final bead slurry was prepared

by mixing beads and the Working Buffer in the ratio 1:1. The diluted sonicated chromatin was pre-cleared with 10 μ l of the bead slurry and incubating for 1h at 4 °C with gentle mixing.

2.2.2.4 Immunoprecipitation

An aliquot of 10% of the pre-cleared chromatin was taken and saved as *Input*. The rest of the pre-cleared chromatin was supplemented with bovine serum albumin (BSA) to a final concentration of 0.5% and *Escherichia coli* tRNA to a final concentration of 0.5 μ g/ μ l. The chromatin was incubated with the primary antibody overnight at 4 °C and with the secondary antibody for 1h at 4 °C, if the secondary antibody was required. 10 μ l of washed beads was added per reaction and incubated for 2h at 4 °C. The beads were washed three times, each wash being for 5min with 1mL of the following ice cold buffer: Low Salt Immune Complex Wash Buffer (1% Triton X-100, 150mM NaCl, 20mM Tris-HCl (pH8.0), 0.1% SDS, 2mM EDTA (pH8.0)); High Salt Immune Complex Wash Buffer (500mM NaCl, 1% Triton X-100, 20mM Tris-HCl (pH8.0), 0.1% SDS, 2mM EDTA (pH8.0)); LiCl Immune Complex Wash Buffer (0.25M LiCl, 1% deoxycholate sodium, 10mM Tris-HCl (pH8.0), 1% IGEPAL CA-630 (Sigma-Aldrich), 1mM EDTA (pH8.0)). This was followed by 3x washes with cold TE buffer (10mM Tris-HCl (pH8.0), 1mM EDTA (pH8.0)). Elution from the beads was done with 2 x 150 μ l of Elution Buffer (1% SDS, 0.1M NaHCO₃), incubating with shaking for 2 x 15min at room temperature.

2.2.2.5 DNA isolation

The precipitated chromatin and *Input* were treated in parallel from this point on. The samples were treated with 1.5 μ g of RNaseA (Qiagen, #19101) for 1h at 65 °C, followed by reverse-crosslinking with 90 μ g of Proteinase K (Roche, #03 115 852 001) at 65 °C overnight. The DNA was isolated using the Phenol Chloroform protocol. Briefly, two extractions were done, one with 450 μ l of phenol chloroform and one with 450 μ l of chloroform, each time vortexing vigorously, centrifuging for 5min at 13400xg at room temperature and saving the top aqueous phase. The DNA was precipitated by adding 1/10 volume of Na Acetate, 2.5 volume of ice-cold 100% EtOH and 5 μ g of glycogen (Invitrogen, #10814-010), inverting the tube several times and incubating at -20 °C overnight. The samples were then centrifuged 13400xg for 20min at 4 °C and washed with 1mL of ice-cold 70% EtOH. The pellet was dried using a speedvac

concentrator (Savant SVC-100H, Thermo Scientific) for 10min at room temperature. The DNA was re-suspended in 15 μ l of nuclease-free H₂O (#10977 Gibco). The DNA concentration was checked with Qubit (dsDNA HS Assay Kit #Q32854) and the fragment size with the DNA 1000 assay on the Agilent Bioanalyzer 2100 Expert (Agilent Technologies).

2.2.3 ATAC

ATAC-Seq is a relatively novel method for measuring chromatin accessibility across the genome. It is based on the action of the Tn5 transposase that preferentially cuts nucleosome-free DNA and simultaneously ligates adaptor sequences (Buenrostro et al., 2013). The original ATAC protocol was optimised for use in cell lines with a quick nuclei isolation step from a single-cell suspension. The pancreatic islets, however, are a compact structure of approximately 1000 islets cells, making isolation of clean nuclei more challenging. For that reason, the nuclei preparation step described in the original protocol was modified by increasing the volume of the ATAC lysis buffer, prolonging the lysis time, releasing the nuclei from the cells by syringe re-suspension and introducing an additional nuclei washing step. The details of the protocol are described below.

2.2.3.1 Nuclei preparation

50 IEQ were picked from a plate of cultivated islets (isolated and cultivated as described in Section 2.1.2), pooled from a minimum of three different mice. The islets were washed in 500 μ l of cold PBS, centrifuged at 100xg for 1min at 4 °C and the supernatant was removed. The islets were lysed in 300 μ l of cold ATAC lysis buffer (10mM Tris-HCl (pH7.4), 10mM NaCl, 3mM MgCl₂ and 0.1% IGEPAL CA-630) for 20min on ice, with gentle shaking. The nuclei were recovered by passing the islets 3x through a 27G syringe and centrifugation at 500xg for 10min at 4 °C. The nuclei were then washed with 100 μ l of cold ATAC lysis buffer at 500xg for 10min at 4 °C.

2.2.3.2 Transposition reaction

The transposase reaction mix was prepared at room temperature by mixing 25 μ l 2xTD Buffer and 2.5 μ l transposase enzyme from the Nextera DNA Sample Preparation Kit (Illumina Inc,

#FC-121-1030) and 22.5 μ l nuclease-free water. Supernatant free nuclei were re-suspended in the transposase reaction mix and incubated for 30min at 37 °C, with gentle shaking. After 30min, the reaction mix was put on ice and the tagged DNA was purified using the Qiagen MinElute Reaction Cleanup Kit (Qiagen, #28204), following the kit instructions. The DNA was finally eluted in 2 x 10 μ l of elution buffer (EB).

2.2.3.3 Library preparation

The ATAC libraries were amplified on the 7900HT Fast Real-Time PCR System, using 1 x NEBnext PCR master mix (NEB, #M0541S) and 1.25 μ M of Primers 1 and 2 (primers designed by Buenrostro et al. (2013) and described in Table 2.1), under the following PCR conditions: 72 °C for 5 min; 98 °C for 30s; 12 cycles of 98 °C for 10s, 63 °C for 30s and 72 °C for 1min. The libraries were purified with Qiagen MinElute PCR Purification Kit (Qiagen, #28004), following kit instructions and eluting in 2 x 10 μ l of EB. The libraries were quantified with Qubit (dsDNA HS Assay Kit #Q32854) and the fragment size was checked with the DNA 1000 assay on the Agilent Bioanalyzer 2100 Expert (Agilent Technologies).

Table 2.1 List of primers used for amplification of ATAC libraries.

	Primer name and index	Primer sequence
Primer 1	Ad1_noIndex	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Primer 2	Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
	Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
	Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
	Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
	Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
	Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
	Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
	Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
	Ad2.9_GCTACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
	Ad2.10_CGAGGCTG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
	Ad2.11_AAGAGGCA	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
	Ad2.12_GTAGAGGA	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT

Chapter 3

Regulatory Architecture of Mouse Pancreatic Islets

3.1 Introduction

3.1.1 The pancreatic islet regulome

Understanding how gene transcription is regulated in pancreatic islets is of central importance for diabetes mellitus. Pancreatic islets are clusters of hormone-producing cells that play a central role in glucose homeostasis. Impaired function of the insulin producing beta cells, the major islet cell type, is one of the hallmarks of type 2 diabetes (T2D) (discussed in Section 1.3.3.2). Understanding the mechanisms underlying islet-cell function is thus a major focus of diabetes research (Ashcroft and Rorsman, 2012; Kahn, 2001). Regulation of gene transcription is a highly complex process directed by non-coding elements and mediated by sequence-specific transcription factors (discussed in Section 1.1.3). Tissue specificity of this process calls for characterization of the regulatory elements and transcription factor target sequences in every tissue of interest (The ENCODE Project Consortium et al., 2012).

3.1.1.1 The human islet regulome

Availability of human pancreatic islet tissue for research is generally scarce, given that the islets can be harvested only from deceased donors and that the islet isolation procedure is technically demanding (Kulkarni and Stewart, 2014). For this reason, the characterization of the human pancreatic islets transcriptome and regulome started relatively recently.

The first comprehensive study of regulatory elements in the human pancreatic islets was performed by Gaulton et al. (2010) who defined ~80,000 open chromatin sites and 3,348 islet-selective clusters of open chromatin sites (COREs) using FAIRE. These chromatin maps were

later refined by Pasquali et al. (2014) and Bhandare et al. (2010) who generated ChIP-Seq data of several histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K79me2), two chromatin defining proteins (H2A.Z and CTCF) and five transcription factors essential for mature beta cell function (FoxA2, MafB, Nkx2.2, Nkx6.1 and Pdx1). Pasquali et al. (2014) also integrated the histone modifications and TF binding with FAIRE and characterized five different classes of accessible chromatin in pancreatic islets - promoters (C1), poised/inactive enhancers (C2), active enhancers (C3), CTCF-bound sites (C4) and other accessible chromatin lacking active histone modifications (C5) (Pasquali et al., 2014). The authors found that the driving force behind the pancreatic islet cell identity is the combined binding of TFs within enhancer clusters. Furthermore, both Pasquali et al. (2014) and Parker et al. (2013) found that enhancer clusters (or stretched enhancers) were enriched in T2D and fasting glycemia associated variants, providing a direct link between islet *cis*-regulatory networks and T2D-related traits. Together, these findings significantly advanced the understanding of the regulatory landscape of the human beta cells and pancreatic islets.

3.1.1.2 The mouse islet regulome

Considering the widespread use of the mouse model in pancreas research, efforts were made early on to annotate mouse islet regulatory regions. Before whole-genome approaches were available, findings of regulatory elements of individual genes were often based on analysis of sequence conservation of regulatory elements. Comparison of mouse and human gene-proximal and distal sequences led to successful characterization of conserved enhancer regions associated to *Pdx1* (Marshak et al., 2000), *Pax4* (Brink et al., 2001) and *NeuroD* genes (Huang et al., 2000). Hoffman et al. (2010) provided the first characterization of genome-wide binding maps of TF binding in mouse pancreatic islets, where they defined the binding sites of Pdx1 and FoxA2. Khoo et al. (2012) later redefined the Pdx1 sites. The first characterization of regulatory elements in murine pancreatic islets resulted in a list of 22,223 enhancers, 1,799 of which were identified as islet-specific (Tennant et al., 2013). This study was based on H3K4me1 and four TFs (Pdx1, NeuroD1, MafA and FoxA2), but was limited by the inability of four TFs to capture all open chromatin sites. Furthermore, lack of H3K27ac, the histone mark that was found to correlate with enhancer activity, prevented distinguishing active from inactive enhancers. A comprehensive genome-wide identification and characterization of mouse islet regulatory elements is, thus, still pending.

3.1.2 Evolutionary conservation of the *cis*-regulatory landscape

Conservation of sequence across species is considered to be indicative of functional importance and is often used for identification of regulatory regions in the non-coding genome (Pennacchio et al., 2006, 2007; Visel et al., 2008; Woolfe et al., 2005). *Cis*-regulatory regions harbour short stretches of DNA targeted by transcription factors and whose sequence is considered to be the major determinant of TF binding (Wilson et al., 2008). Conservation of TF binding in mammals, however, seems to be more complex and goes beyond simple sequence conservation of the TF binding motif. Individual TF binding events are rapidly lost and gained across species, and the conservation of a binding event in the vicinity of a particular gene is not necessarily coupled with sequence constraint (Blow et al., 2010; Hare et al., 2008; Schmidt et al., 2010). Combinatorial binding of key TFs tends to be more conserved than individual binding events, as shown by human to mouse comparisons, as well as studies in different strains of mice (Cheng et al., 2014; Göke et al., 2011; Stefflova et al., 2013). Furthermore, comparison of mouse and human regulatory networks found the overall architecture of TF networks to be overwhelmingly conserved (Cheng et al., 2014; Stergachis et al., 2014). Together, these studies describe the general principles of conservation of TF binding, where flexibility of *cis*-regulatory regulation is enabled by conservation of higher-level organisation of TF networks.

Enhancers harbouring the TF target sequences vary in their level of conservation. Enhancers associated to developmental genes, for example, have long been known for their particularly conserved sequences (Bejerano et al., 2004; Woolfe et al., 2005). However, looking at the regulatory elements of terminally differentiated liver tissue, a recent large scale analysis of regulatory elements across 20 mammalian species found that enhancers generally evolve quite rapidly (Villar et al., 2015). The authors also found that recently evolved enhancers seem to be equally important as the conserved ones and are overrepresented in the vicinity of positively selected genes. Differences in the non-coding genome were long thought to underlie species-specific phenotypes (King and Wilson, 1975), a hypothesis now supported by numerous studies (Cotney et al., 2013; McLean et al., 2011; Shibata et al., 2012). A genome-wide study of enhancer activity performed in several *Drosophila* species recently demonstrated that, even between closely related species, differences in enhancer activity can translate into significant differences in transcriptional phenotypes *in-vivo* (Arnold et al., 2014). Such differences between mouse and human pancreatic islets were recently described in a study that identified a

subset of beta cell genes differentially expressed between the two species (Benner et al., 2014). This poses the question whether a species-specific subset of pancreatic islet regulatory architecture can be identified between human and mouse and opens the possibility of revealing the mechanism behind the differences in biology of mouse and human islets.

3.1.3 Summary

Despite the widespread use of the mouse as the model for pancreas development and diabetes pathophysiology, the regulatory landscape of the mouse pancreatic islets has been poorly characterized. While investigating the non-coding part of the genome, sequence conservation has often been utilized as a predictor of functional regulatory elements. Recent studies, however, have shown major rewiring of individual regulatory elements between genomes of different species. It is clear that extensive functional annotation of genomes in disease-relevant tissues of model organisms is crucial for proper design and interpretation of functional studies performed in those models. At the same time, establishing the particularities of species-specific aspects of regulatory architecture provides insight into the mechanisms behind the biology of differences in tissue function.

Thus, the specific aims of this chapter are:

1. To map the chromatin signatures and target sites of representative beta cell transcription factors in the mouse pancreatic islets
2. Characterize the mouse islet *cis*-regulatory elements and their underlying sequences
3. Identify common and species-specific aspects of the mouse and human islet regulatory architecture

3.2 Methods

3.2.1 ChIP-Seq

The chromatin immunoprecipitations for histone modifications were performed as described in General Methods, Section 2.2.2, using 150-300IEQ per pull-down, with islets coming from minimum three animals. Each reaction was performed in biological duplicate, using the following antibodies: H3K4me3 (Upstate, 05-745; 1 μ g/reaction), H3K27ac (ab4729, Abcam; 1 μ g/reaction), H3K4me1 (ab-8895, Abcam; 1 μ g/reaction). The ChIPs performed against transcription factors in general require a much higher starting amount of chromatin, due to lower number of epitopes available for a pull down. The optimization of the ChIP protocol described in General Methods enabled only 1,000-1,500IEQ to be used per reaction. Each ChIP was performed once and the following antibodies were used: Nkx2.2 (HPA003468, Sigma; 1 μ g/reaction), Nkx6.1 (HPA036774, Sigma; 1 μ g/reaction), FoxA2 (sc-6554, Santa Cruz Biotechnology; 0.8 μ g/reaction).

Each ChIP yielded 1.5 - 3ng of DNA, 25% of which was used for quality control (as described in General Methods, Section 2.2.2.4), and 1-2ng was used for constructing the sequencing libraries using Nugen Ovation Kit (Nugen, Part No. 0330). The libraries were sequenced using 50 single-end (SE) reads to the depth of 30-60 million (M) reads per library.

3.2.2 ATAC-Seq

ATAC libraries were prepared in duplicate, using the ATAC-Seq protocol optimized for pancreatic islets, as described in General Methods, Section 2.2.3. The sequencing of the libraries was performed using 100PE reads to the depth of 100-150M fragments per library.

3.2.3 Processing of raw sequenced reads

The ChIP-Seq (both in-house and data published by Tennant et al. (2013) and Khoo et al. (2012) and ATAC-Seq reads were processed in a similar manner, the differences accounting for the fact that the ATAC-Seq was performed using PE reads, and ChIP-Seq using SE reads. *PrinSeq* was used for checking the quality of the raw reads (adapters removed by the sequencing facility) and

filtering out the low quality reads (Schmieder and Edwards, 2011). The reads were aligned to the *mm9* genome (NCBI build 37) using *bowtie2* with default settings (Langmead and Salzberg, 2012) and accounting for SE or PE reads. Uniquely aligned reads were selected and used for identifying read-enriched regions (peak calling).

3.2.4 Peak calling

Peak calling programs use different algorithms to determine regions in the genome where the enrichment of ChIP-Seq reads are indicative of binding of the protein of interest. A similar logic can be applied to ATAC-Seq, where the enrichment of reads would indicate regions of open chromatin, bound by unknown proteins. The choice of the peak calling algorithm depends on the properties of the peaks one expects to see, as some algorithms are more suitable for detecting wide peaks (for example, SICER Zang et al., 2009) and others for narrow, well defined peaks (for example, MACS Zhang et al., 2008).

For H3K27ac, H3Kme1 and H3K4me3 peaks were called with SICER, using the following parameters:

```
SICER.sh output_folder chip_file.bed input_file.bed . mm9 1000 200 150 0.74  
200 0.01
```

For all the TFs, the peaks were called with MACS, calling:

```
macs2 callpeak -t chip_file.bam -c input_file.bam -f BAM -g 2150570000 -n  
outfile_name -B -q 0.01
```

For ATAC-Seq, the peaks were called using MACS with the following parameters:

```
macs2 callpeak -t atac_file.bed -f BED -g 2150570000 -n outfile_name --  
nomodel --extsize 200
```

Input (a sample of sonicated pre-cleared chromatin, described in General Methods), was sequenced to the depth of 40M 50SE reads and was used as a control for peak calling of all histone modifications and TFs, but not ATAC.

Consistent peaks were defined as the intersect of the peaks called on the two replicates for the histone modifications, or the intersect of the ChIP and ATAC peaks for transcription factors.

The intersection was performed with the help of *bedtools* (Quinlan and Hall, 2010), a suite of tools that allows easy manipulation of genomic coordinates, using default settings. Consistent peaks overlapping blacklisted regions were filtered out from the final list (Bailey et al., 2013).

3.2.5 Defining *cis*-regulatory elements

The regulatory elements were defined by overlapping consistent peaks for the three histone modifications available and consistent ATAC peaks (schematic representation of overlaps (Figure 3.2 in Section 3.2). Using the ATAC peaks as the base, three general types of regulatory elements were defined; active promoters (AP) enriched in H3K4me3 and H3K27ac with absence of H3K4me1, active enhancers (AE) enriched in H3K27ac and H3K4me1 with absence of strong H3K4me4 and inactive enhancers (IE) enriched in H3K4me1 but absence of H3K27ac or H3K4me4. Weak H3K4me3 signals were defined as signal weaker than the 25th percentile of the H3K4me3 signal located in RefSeq promoters. The overlap between the genomic coordinates of the histone marks and ATAC regions was performed with *bedtools*, using default settings (Quinlan and Hall, 2010).

3.2.6 Defining highly occupied active enhancers

Enhancers highly occupied by the five assayed TFs (FoxA2, Nkx2-2, Nkx6-1, Pdx1, NeuroD1) were defined following an approach used by Whyte et al. (2013) and Hnisz et al. (2013) for defining super-enhancers. Enhancers were ordered according to the sum of MACS-derived ChIP signal scores for all five TFs. The ordered sum signal scores formed a curve (bottom panel of Figure 3.5a in Section 3.3.4.1). Enhancers falling beyond the point where the slope of the tangent to the values of ranked enhancers reached 1 were classified as *highly occupied*. The enhancers occupied by TFs, but not classifying as highly occupied were considered *moderately occupied*. The rest of the enhancers were *not occupied* by neither of the assayed TFs.

3.2.7 Defining enhancer clusters

The enhancer clusters were defined using an approach adapted from Pasquali et al. (2014). A cluster is defined as a group of minimum three enhancers separated by a maximum of 24110bp,

which represents the median of the inter-enhancer distance thresholds defined for each chromosome in Pasquali et al. (2014). To enable a direct comparison of mouse and human enhancer clusters, the human clusters were here redefined, using the same minimum distance for all chromosomes.

3.2.8 Transcription factor motif analysis

A *de novo* motif analysis of the islet regulome elements was performed with Homer (Heinz et al., 2010), a tool that looks for enrichment of sequence motifs against a custom made set of background genomic sequences normalized for length and GC-content. The computational demands of the motif search increase greatly with increasing the target size of the motif (Heinz et al., 2010). One way of maximising the efficiency of the analysis is to provide the algorithm with a narrow range of motif lengths. As the majority of annotated motifs are 6-20bp long (Mathelier et al., 2014), motif discovery was performed in three rounds, looking for motifs 6-10bp, 12-16bp and 18-20bp long. For example;

```
findMotifsGenome.pl regulatory_elements.bed mm9 regulatory_elements_6-10 -  
size 250 -homer2 -p 6 -len 6,8,10
```

The motifs were later converted to *meme* format, and compared to each other using TomTom (Gupta et al., 2007), a tool from the MEME tool suite (<http://meme.nbcr.net/meme>), using the parameters below. If redundant, motif with the lowest *p*-value would be chosen. Finally, a set of non-redundant motifs with *p*-values > 1e-50 was obtained.

```
tomtom -thresh 0.01 -text -verbosity 1 homerMotifs.meme homerMotifs.meme >  
homerMotifs.compared.tomtom
```

In parallel with *de novo* motif discovery, Homer performs motif annotation to Jaspar database (Mathelier et al., 2014). To ensure exhaustive motif annotation, additional comparison to the Jaspar database was performed with TomTom. First, genomic locations of the *de novo* motifs were mapped with FIMO, another tool from the MEME tool suite. Second, for each motif, the sequences underlying the motif locations were extracted and used for construction of the motif count matrix. The new count matrices were then converted to *meme* format and compared with the Jaspar database, using TomTom and the same parameters as above.

The comparison of mouse and human motifs was done on the final list of curated *de novo* mouse motifs and human pancreatic islet active enhancer motifs taken from Pasquali et al. (2014). The motifs were compared using TomTom, with a significance threshold of 0.05

3.2.9 Gene ontology analysis

Two closest genes were assigned to each enhancer using GREAT, a tool for functional interpretation of genomic regions (McLean et al., 2010). The genes were required to be less than 250kb away from the enhancer, and were further filtered for residing in the same TAD as the enhancer. The mouse ES cell topological domains were taken from Dixon et al. (2012) and extended for 40kb in each direction (40kb being the size of bins used for the TAD definition). Molecular Signatures Database (MSigDB) categories were used as representative gene ontology terms (Subramanian et al., 2005).

3.2.10 Conservation analysis

Mouse enhancer coordinates were translated from *mm9* to *hg19* with bnMapper using default settings (Denas et al., 2015). A chain file that allows 1-to-many mapping was used to ensure exhaustive mapping of mouse sequences to human genome.

```
bnMapper.py mouse_regulatory_element.bed mm9ToHg19.over.chain -f BED12 -o  
mouse_regulatory_element.hg19.bed
```

Human pancreatic regulatory regions were taken from Pasquali et al. (2014). A mouse active enhancer would be classified as sequence conserved if the coordinates were mappable to the human genome and overlapped a “C3” element in human (an element with a chromatin signature of an active enhancer). A mouse cluster was characterized as syntenically conserved if its enhancers were associated to a gene that was also associated to clustered enhancers in human. Genes were associated to enhancers by proximity, using GREAT and applying the same criteria as for the gene ontology analysis.

3.2.11 Quantification of gene expression

Gene expression in mouse pancreatic islets was quantified by RNA-Seq of the samples processed as described in General Methods, Section 2.2.1 and Chapter 4, Section 4.2.1.1.1. The aligned reads falling into Ensembl catalogued transcripts were counted with HTSeq (Anders et al., 2014) and Reads Per Kilobase of transcript per Million mapped reads (RPKM) values calculated with the RSeQC tool (Wang et al., 2012). Gene expression levels for human pancreatic islets were taken from Morán et al. (2012). After compiling a list of genes present in both species and expressed > 0.1 RPKM in at least one of them, the percentile values of gene expression were calculated separately for both mouse and human genes.

3.3 Results

3.3.1 Defining *cis*-regulatory elements of mouse pancreatic islets

Different types of *cis*-regulatory elements have been shown to be marked by particular combinations of chromatin modifications (Barski et al., 2007). Therefore, in order to define these elements in mouse pancreatic islets, ChIP-Seq was performed on three relevant histone modifications: H3K4me3 (found in active promoters), H3K27ac (found in active promoters and enhancers) and H3K4me1 (characteristic for all enhancers) (Creyghton et al., 2010a; Heintzman et al., 2007) (Figure 3.1). While histone modifications generally tend to cover relatively broad genomic regions (Barski et al., 2007), the functional part of regulatory elements is hypothesized to be much more narrow. Nucleosome displacement at regulatory elements leaves the chromatin open, which marks the more narrow, protein-binding sites (Thurman et al., 2012). Thus, in order to refine the regulatory regions of the mouse islets, an additional assay was used to identify regions of open chromatin - assay for transposase-accessible chromatin using sequencing (ATAC-Seq). This method, described by Buenrostro et al. (2013), maps open chromatin, TF and nucleosome occupancy (discussed in Section 2.2.3).

All assays were performed in biological duplicates. The number of signal-enriched regions of H3K4me3 (28,215) and H3K4me1 (95,398) was similar to the number of regions obtained with the same antibody in human pancreatic islets by Pasquali et al. (2014) (23,541 for H3K4me3 and 78,165 for H3K4me1, Figure 3.1). The H3K27ac analysis, however, yielded more regions in mouse (58,848) than in human (28,592, Figure 3.1). The increased number of H3K27ac-enriched regions in mouse islets could reflect differences in the approach to defining enriched regions (different settings of the peak-calling algorithm, for example). However, given that H3K27ac marks both active promoters and active enhancers, it is plausible that the increase of the number of H3K27ac-enriched regions reflects a higher quality of the ChIP assay and capture of a larger number of regulatory regions.

Although the correlation of signal intensities was high for H3K4me3 and ATAC (0.889 and 0.779, respectively), it was fairly modest between replicates of H3K27ac and H3K4me1 (0.360 and 0.527, respectively, Figure 3.1). These two assays yield broad regions with a dispersed signal, unlike H3K4me1 and ATAC that produce narrow, peak-like enriched regions. Discrepancies in the signal intensities skewed towards one of the axis indicate that the assay was of

higher quality in one of the replicates. The high overlap, however, indicates that the quality of the assay in both replicates was sufficiently good for accurate calling of enriched regions.

The final set of regulatory elements was obtained by intersecting consistent peaks of appropriate histone modifications with the consistent ATAC peaks (Figure 3.2a). This approach resulted in a compendium of 78,921 regulatory elements, carrying chromatin signatures of active promoters (AP, n=20,303), active enhancers (AE, n=30,259) and inactive enhancers (IE, n=28,359) (Figure 3.2b).

Histone modification, ATAC and TF sequencing data, as well as the regulatory regions, were visualised in the UCSC browser and can be accessed through "User; Nikolina" and "Session name; NNakic_PancIsletRegulome_mm9".

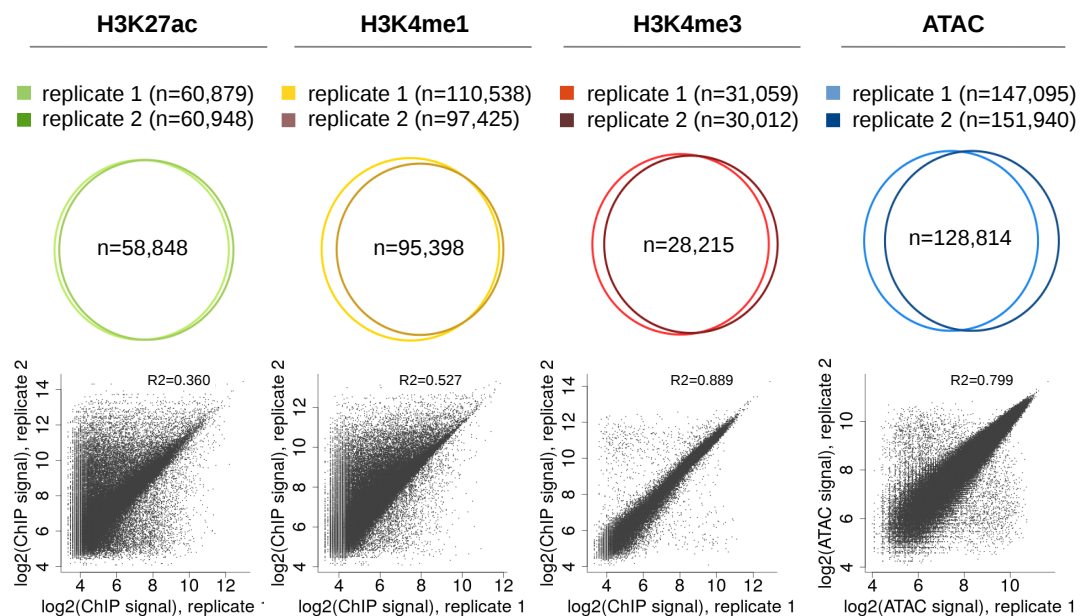


Figure 3.1 Enrichment and consistency of ChIP-Seq and ATAC-Seq signal. The top panel shows the overlap between regions found to be enriched for histone modification or ATAC signal in individual replicates. The bottom panel shows correlation of signal intensities (Sicer peak scores) between the two biological replicates in the consistently enriched regions. The correlation is expressed as goodness of fit (R² value).

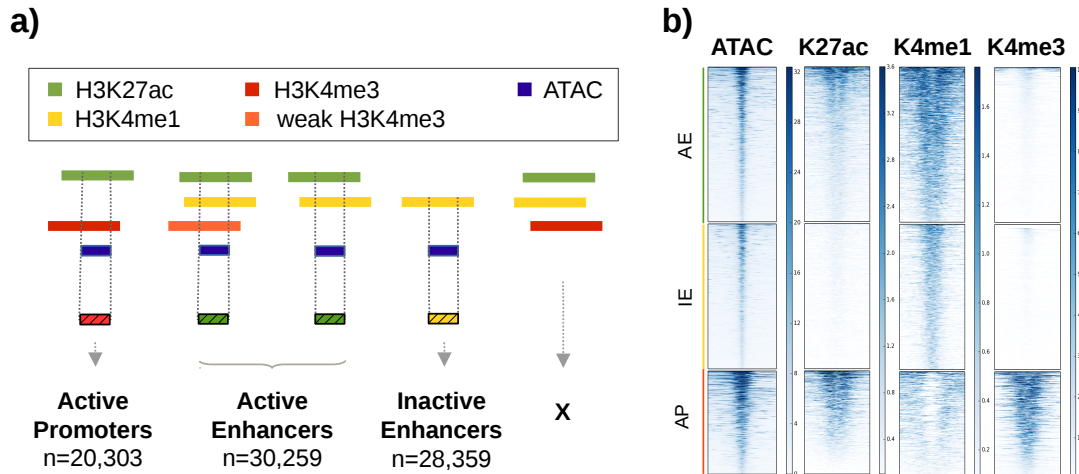


Figure 3.2 Utilizing chromatin properties for identifying regulatory elements in mouse pancreatic islets. a) A schematic representation of the intersection-based approach to identifying the regulatory elements in mouse pancreatic islets. The empty bars represent the consistent peaks of ChIP-Seq/ATAC-Seq. The striped bars represent the final regulatory elements obtained by intersecting the respective chromatin-marked regions. The weak H3K4me3 peaks were defined as peaks with intensity scores below that of the 25th percentile of the intensity scores of peaks located in RefSeq promoters. b) The chromatin signatures of the active promoters (AP), active enhancers (AE) and inactive enhancers (IE) identified by the approach described in a).

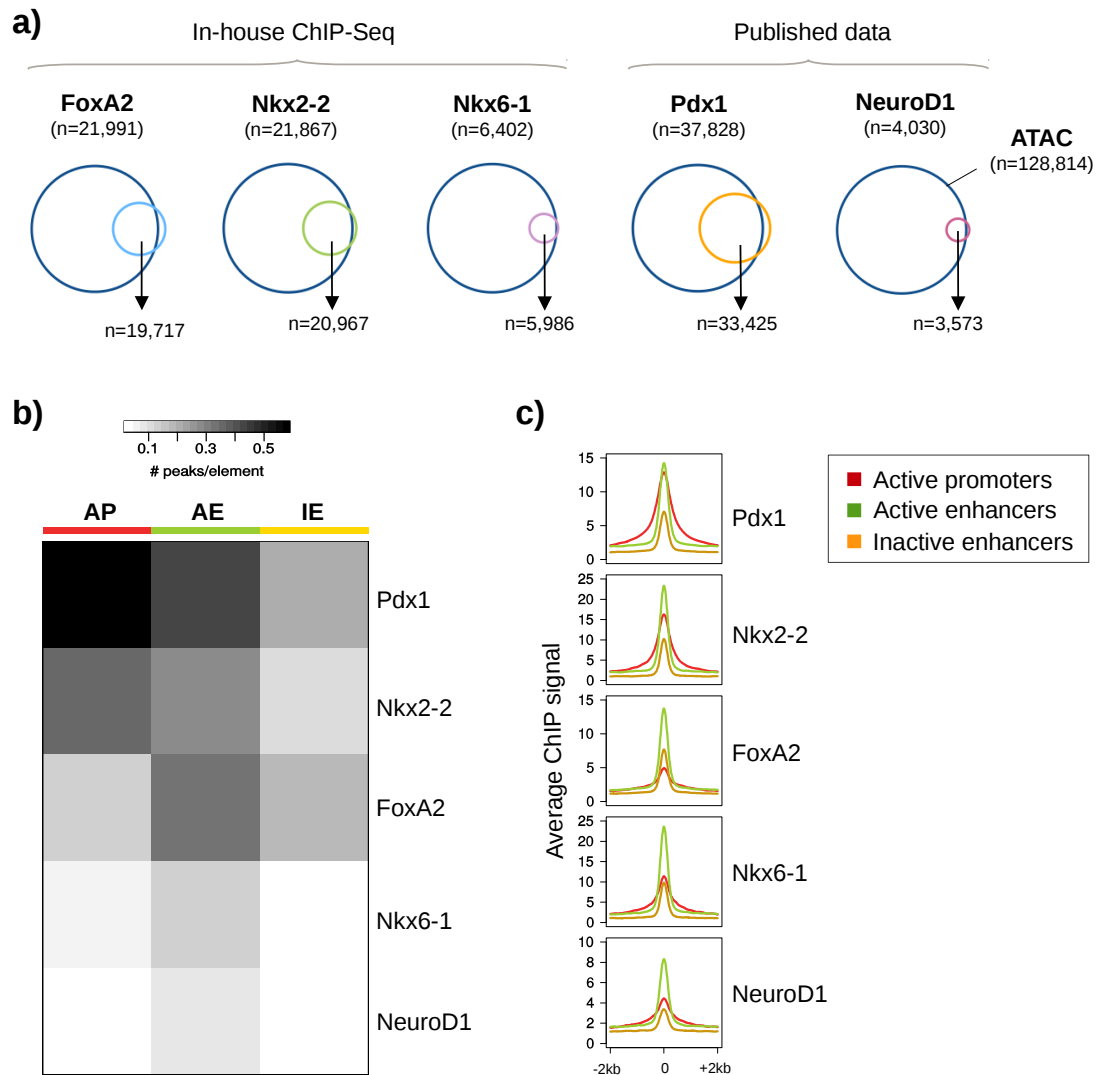
3.3.2 Mapping genome-wide binding of five islet transcription factors

To define binding patterns of lineage-selective TFs in mouse islets, five representative transcription factors were studied: FoxA2, Nkx2-2, Nkx6-1 (in house ChIP-Seq) and Pdx1 and NeuroD1 as published by Tennant et al. (2013) and Khoo et al. (2012), respectively. These factors play an essential role in beta cell differentiation or maturation and are all expressed in the adult insulin-producing beta cell (discussed in Section 1.3.2 and summarised in Figure 1.10). High quality ChIP-grade antibodies are available for all five TF. Furthermore, binding of four out of five of these factors was previously mapped in the human pancreatic islets (Pasquali et al., 2014), allowing for a direct comparison of their binding profiles.

To enrich for TF bindings sites that are likely to be important for gene regulation, final sets of TF-bound sites were obtained by overlapping the TF peaks with open chromatin regions (ATAC peaks). This yielded sets of 3,573 - 33,425 enriched sites per transcription factor (Figure 3.3a), reflecting differences in number of genomic regions occupied by the proteins *in vivo*, although some differences likely stem from variability in the quality between assays. The distribution of TF binding events and signal intensities is not uniform across different regulatory

elements (Figure 3.3b,c). Pdx1 and Nkx2-2, for example, have a higher average number of peaks located in active promoters while the rest tend to be bound in active enhancers (Figure 3.3b). However, the strongest overall TF signal intensity, apart from Pdx1, was observed in active enhancers, further reinforcing the central role of lineage-selective TFs in active enhancers (Figure 3.3c).

Earlier studies of the human islet regulome revealed auto- and cross-regulatory interactions between the interrogated TFs (Pasquali et al., 2014). The same pattern now emerged in the mouse islets, with TFs binding to the regulatory regions in the vicinity of their own gene and genes of the other assayed TFs (Table 3.1). Nkx6-1 is the only TF that did not bind to its own promoter, although it did bind to enhancers associated to its own gene, as all other TFs (Figure 3.4). Together, these findings define general features of lineage-selective transcription factors binding in mouse islets.



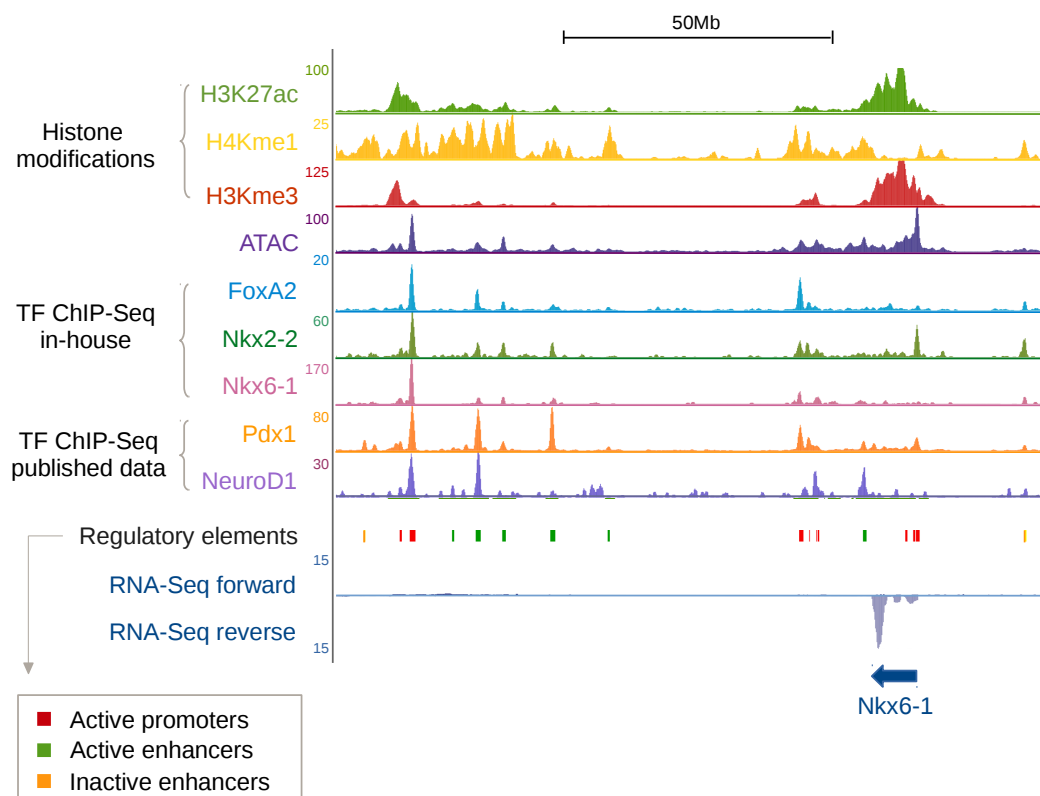


Figure 3.4 Chromatin and transcription map of the Nkx6-1 locus. Each track represents read density for an assay, calculated across 10bp windows and normalized to mappable genome size. A single replicate of histone modification ChIP-Seq, ATAC-Seq and RNA-Seq is shown here as representative. The “Regulatory elements” track displays the genomic locations of active promoters, active enhancers and inactive enhancers.

Table 3.1 Binding of TFs in regulatory elements of TF genes. Presence (+) or absence (-) of a TF peak at the active promoter (AP) or active enhancer (AE) less than 100kb away from the gene TSS.

		Binding site									
		Pdx1		Nkx2-2		FoxA2		Nkx6-1		NeuroD1	
		AP	AE	AP	AE	AP	AE	AP	AE	AP	AE
TF binding	Pdx1	+	+	+	+	+	+	+	+	+	+
	Nkx2-2	+	+	+	+	+	+	+	+	+	+
	FoxA2	+	+	+	+	+	+	-	+	+	+
	Nkx6-1	-	+	+	+	+	+	-	+	+	+
	NeuroD1	+	+	+	+	-	+	-	+	+	+

3.3.3 *De novo* annotation of mouse islet regulatory motif sequences

To expand the analysis and identify additional putative transcriptional regulators binding to the mouse islet active enhancers, a motif analysis was performed on their underlying sequences. This revealed enrichment of 25 non-redundant motifs, including motifs of an array of known beta cell regulators, such as FoxA1, Nkx6-1, Hnf1a/b and MafA (Table 3.2). Furthermore, many of those motifs were novel and could not be matched to known Jaspar motifs under stringent settings.

To assess the conservation of the regulatory code between mouse and human, a comparison of islet enhancer-enriched motifs was performed between the 25 mouse motifs identified here and the motifs found enriched in human active enhancers by Pasquali et al. (2014). The majority of sequences found enriched in the mouse islet matched a human islet-enriched motif, including known and novel motifs (Table 3.2). As the comparison between mouse and human motifs was done with a more relaxed similarity threshold, some novel mouse motifs were assigned to known human motifs. The CTCF motif was found enriched in mouse active enhancers but not in human C3 elements. This reflects the differences in definition of regulatory elements, as Pasquali et al. eliminated CTCF-binding regions from their active enhancer elements. Overall, motif similarities between mouse and human active enhancers reveal a general conservation of islet enhancer transcriptional regulators.

Table 3.2 Motifs enriched in mouse pancreatic islet active enhancers. *De novo* motif analysis on active enhancer sequences was performed with HOMER, in comparison with length and GC-content matched background sequences. Curated motifs with a p -value $\leq 1e - 50$ were considered significantly enriched. The motifs were assigned to TFs using HOMER and TomTom. If the two algorithms assigned the motif to different TFs, both matches are shown. Novel motifs did not match any motif catalogued in the JASPAR database under stringent settings (significance values threshold less than 0.01). Comparison of mouse and human motifs was performed with less stringent settings (significance values threshold less than 0.05).

Assigned TF (HOMER or TomTom)	HOMER de-novo motif enrichment (p-value)	Motif enriched in mouse islet active enhancers	Matched motif enriched in human enhancers	Matched motif similarity (q-value)
FoxA1	1e-1906		FOX	3.103e-03
Nkx6.1	1e-1065		PDX1-NKX6.1	2.382e-03
Atoh1/MyoD1	1e-757		NEUROD	1.812e-04
Rfx5	1e-647		RFX.2	1.010e-05
AP1/MAF	1e-463		AP1	2.164e-08
NovelMotif-1	1e-460		FOXA.3	2.151e-03
NovelMotif-2	1e-336		bHLH.1	1.485e-03
NovelMotif-3	1e-68		IEF.15	0.0638
NovelMotif-5	1e-208		FOXA.1	0.0534
NovelMotif-7	1e-152		/	/
NovelMotif-8	1e-613		HNF1	0.105
NovelMotif-9	1e-413		MAF-CREB1	0.0285
NovelMotif-10	1e-254		IEF.5	8.147e-03

Assigned TF (HOMER or TomTom)	HOMER <i>de-novo</i> motif enrichment (p-value)	Motif enriched in mouse islet active enhancers	Matched motif enriched in human enhancers	Matched motif similarity (q-value)
FKH1/2	1e-357		HNF1	5.996e-10
HNFA/B	1e-340		FOXA	0.172
NovelMotif-11	1e-272		bHLH.1	3.250e-04
MEF2A/MEF2C	1e-153		MEF2.2	7.038e-06
HNF4A/HNF4G	1e-98		NR	0.104
HOXC9/Dfd	1e-230		IEF.7	1.390e-04
Irf1	1e-76		PRDM-IRF	3.151e-04
NovelMotif-12	1e-200		FOXA.1	0.0832
RORA_2	1e-114		/	/
NovelMotif-16	1e-51		/	/
CTCF	1e-284		/	/
TLX1/NFIC	1e-130		/	/

3.3.4 Categorizing active enhancers

3.3.4.1 Defining highly occupied active enhancers

Level of occupancy by TFs varies greatly across binding sites. Studies in *Drosophila* indicated that the low-occupied sites lack functionality (Bradley et al., 2010; Fisher et al., 2012; MacArthur et al., 2009), while strong combinatorial binding predicts expression of target genes (Dogan et al., 2015). In order to identify a subset of active enhancers highly occupied by the five assayed TFs, an approach described by Whyte et al. was here adapted for use with multiple transcription factors (Whyte et al., 2013). MACS derived peak intensity values were used as proxy of TF occupancy and peak intensity scores of all TFs were combined across enhancers as an indicator of total TF occupancy (Figure 3.5a, bottom panel). A geometrical point determined the intensity signal cutoff above which enhancers were considered highly occupied by TFs (Figure 3.5a, bottom panel). Three categories of active enhancers emerged; enhancers not occupied by assayed TFs, enhancers moderately occupied by assayed TFs and enhancers highly occupied by assayed TFs. Even though highly occupied enhancers made only 4.5% of total number of enhancers, they harboured 33.5% of TF signal, indicating highly imbalanced distribution of TF binding across the regulome. The same enhancers have shown a high degree of combinatorial TF binding, with an average of 4 heterotypic TFs bound (Figure 3.5b). H3K27ac and ATAC signal intensities generally followed the cumulative TF intensities (Figure 3.5a, top panel). The three occupancy-based categories of enhancers showed a dramatic difference of enrichment for all histone modifications and ATAC signal, indicating major differences in enhancer activity between the three categories (Figure 3.5c). Overall, these results suggest great heterogeneity between active enhancers and demonstrate the relevance of TFs used for this analysis.

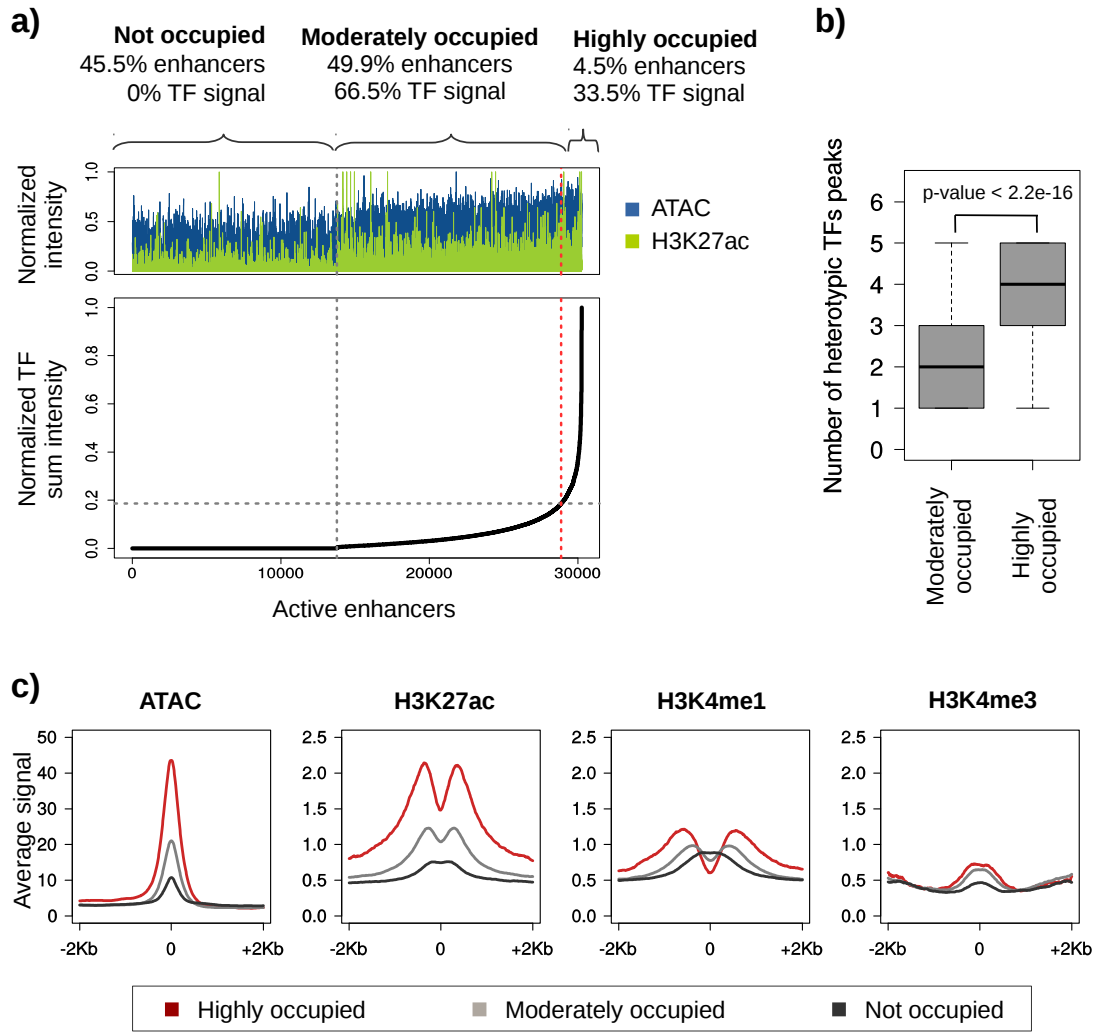


Figure 3.5 Categorizing active enhancers according to occupancy by the five assayed transcription factors. Categorizing active enhancers according to occupancy by the five assayed TFs. a) Active enhancers are sorted on the x-axis according to the cumulative TF intensity scores. Geometrical properties of the intensity score (bottom panel, black line) separate three categories: enhancers not occupied, moderately occupied and highly occupied by the assayed TFs. The top panel shows the normalized intensities of H3K27ac and ATAC signal for the same enhancers. b) The number of heterotypic binding events in moderately occupied vs. highly occupied enhancers. The p -value is calculated with the Wilcoxon rank sum test. c) Chromatin signatures of TF occupancy-based categories of enhancers. Distribution of the average ATAC and histone modification signal is shown relative to the centre of the enhancer.

3.3.4.2 Defining clustered active enhancers

Enhancers that form clusters were found to be of particular importance for islet cell identity (Parker et al., 2013; Pasquali et al., 2014). In order to define enhancer clusters in mouse pancreatic islets, the approach by Pasquali et al. was adapted and enhancer clusters were defined as groups of minimum three enhancers, each of which is separated by maximum 24,110bp from the nearest enhancer in the cluster. This distance represents the median of distances defined separately for each chromosome in Pasquali et al. (2014).

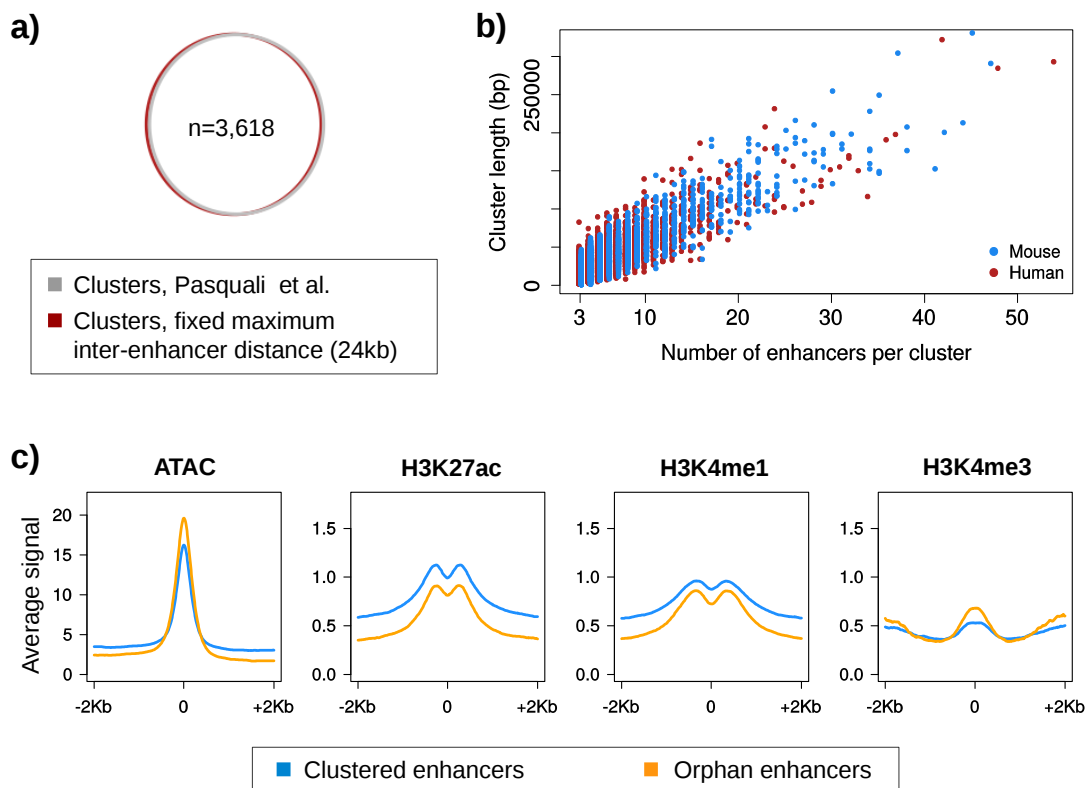


Figure 3.6 Categorizing active enhancers according to clustering. a) Overlap between human pancreatic islet enhancer clusters originally identified by Pasquali et al. (n=3,677) and number of clusters identified in this study (n=3,705). b) Total length and enhancer content of enhancer clusters in mouse and human. c) Chromatin signatures of clustered and orphan mouse active enhancers.

In total 3,252 enhancer clusters were identified in mouse islets, harbouring 19,539 enhancers (64.6% of total number of active enhancers). To enable a fair comparison between the mouse and human enhancer clusters, the human clusters were here redefined, starting from the list of individual enhancers from Pasquali et al. (2014) and following the same approach used for the

mouse. Here, 3,705 clusters were identified in the human pancreatic islets, capturing 98.4% of the clusters identified in the original study (Figure 3.6a). This overlap made it possible to compare the findings described in Pasquali et al. (2014) with the results found in mouse. The size and number of enhancers that formed clusters were similar between mouse and human (Figure 3.6b). Clustered enhancers showed overall higher levels of H3K27ac, suggesting stronger activity, even if ATAC-Seq did not indicate that chromatin was more accessible in comparison with orphan enhancers (Figure 3.6c). Furthermore, following the finding that clustered enhancers drive islet-specific transcription (Parker et al., 2013; Pasquali et al., 2014), gene ontology analysis demonstrated that, unlike orphan enhancers, clustered enhancers associate with genes involved in the core programme of pancreatic beta cells (Figure 3.7). Together, these results demonstrate the differences in chromatin properties and type of proximal genes between clustered and orphan enhancers.

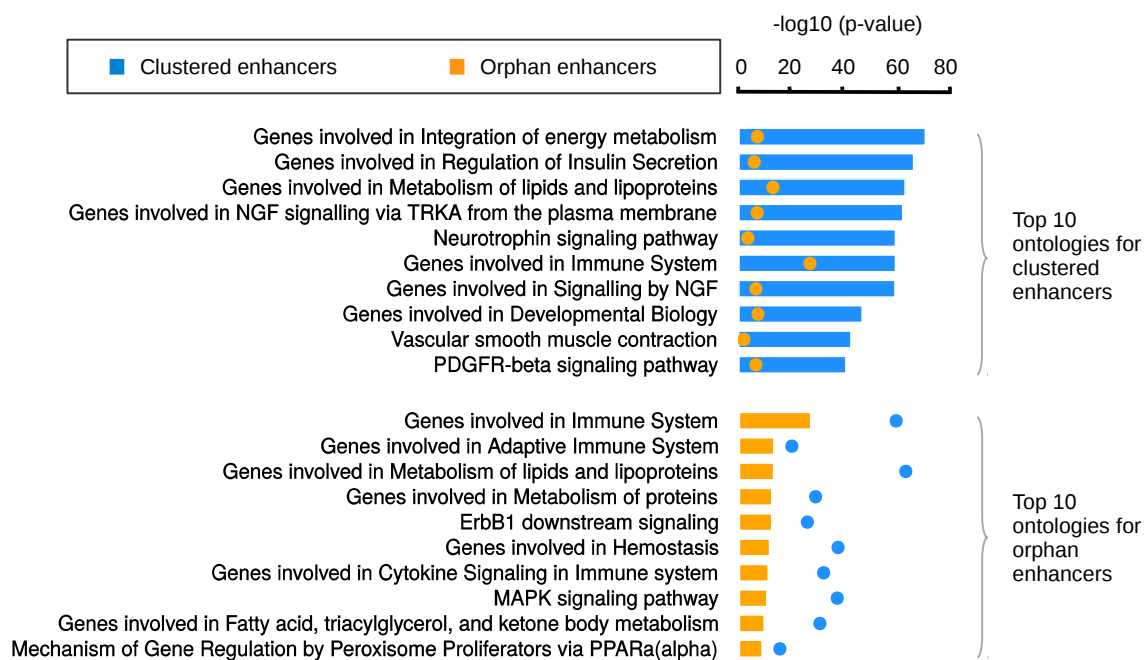


Figure 3.7 Gene ontology analysis of clustered and orphan mouse active enhancers. Molecular signatures of genes associated with clustered and orphan enhancers. Ten most enriched MSigDB ontologies are shown, ranked according to the enrichment binomial p -value. The colour code applies to both bars and the points.

3.3.5 Conservation of *cis*-regulation by enhancer clusters in mouse and human

3.3.5.1 Categorizing active enhancer clusters according to syntenic and sequence conservation

Enhancer clusters were shown to associate with genes essential for beta cell identity in both mouse and human pancreatic islets, suggesting conservation of the regulatory landscape between the two species. To further characterize these commonalities and explore species-specific aspects of islet gene regulation, clustered enhancers were compared between the two species. Taking mouse active enhancers as the starting point of the analysis, both syntenic conservation (association of a cluster to the same gene in both mouse and human) and sequence conservation (alignment of mouse sequence to human) were considered. As a result, five mutually exclusive categories of regulatory conservation were described (Figure 3.8a);

1. **C-C-S** (Cluster - Cluster with Sequence conservation) - a gene is associated with a cluster in both mouse and human, at least one of the enhancers is sequence conserved.
2. **C-C** (Cluster - Cluster without sequence conservation) - a gene is associated with a cluster in both mouse and human, none of the enhancers are sequence conserved.
3. **C-O-S** (Cluster - Orphan with Sequence conservation) - a gene is associated with a cluster in mouse and with an orphan enhancer in human, at least one of the enhancers is sequence conserved.
4. **C-O** (Cluster - Orphan without sequence conservation) - a gene is associated with a cluster in mouse and with an orphan enhancer in human, none of the enhancers are sequence conserved.
5. **C-A** (Cluster - Absence of any regulatory element) - a gene is associated with a cluster in mouse, but is not associated to a regulatory element in human.

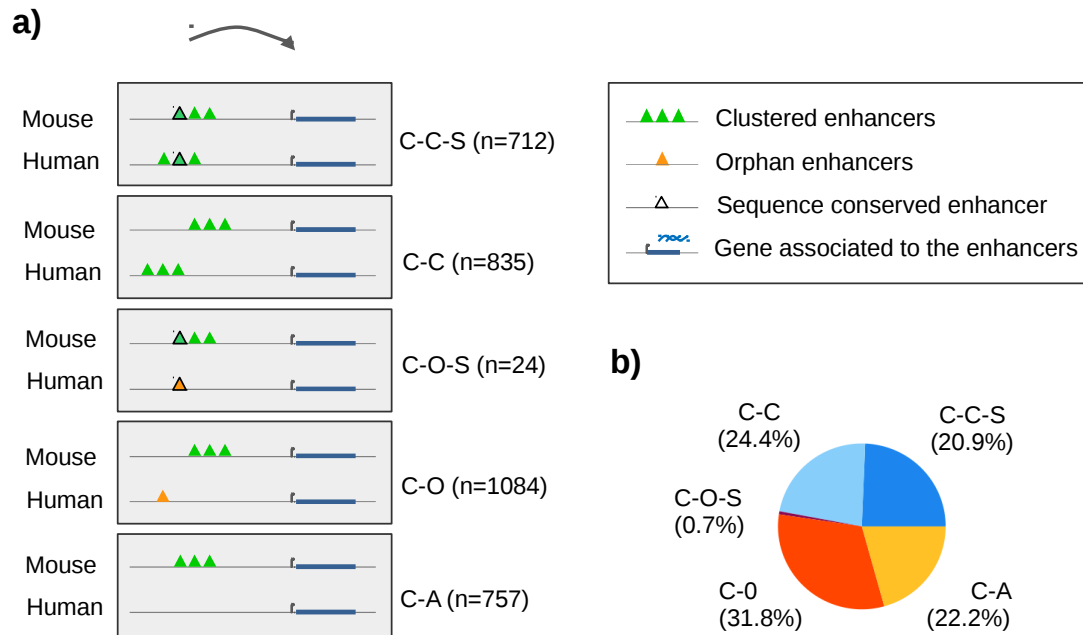


Figure 3.8 Categories of mouse enhancer clusters based on conservation between mouse and human. a) Schematic representation of the cluster categories. The enhancers were associated to the closest expressed gene located within the same topologically associating domain (TAD) and maximum 250kb away from the gene TSS. The numbers represent the number of clusters falling into each category. b) Fractions of clusters falling into each category. Abbreviations: C-C-S (cluster is associated to the same gene in both mouse and human, harbouring a sequence conserved enhancer), C-C (cluster is associated to the same gene in both mouse and human, not harbouring a sequence conserved enhancer), C-O-S (mouse cluster is associated to orphan-associated gene in human, harbouring a sequence conserved enhancer), C-O (mouse cluster is associated to orphan-associated gene in human, not harbouring a sequence conserved enhancer), C-A (mouse cluster is associated to a gene that is not associated to an active enhancer in human).

The majority of the clustered enhancers showed evolutionary flexibility - a total of 32.5% of the clusters in mouse had orphan enhancers associated to the same genes in human (Figure 3.8b). These enhancers overwhelmingly lacked sequence conservation, with only 0.7% of enhancers in total falling into C-O-S category. As a result, the latter category was excluded from further analysis.

Compared to the C-O-S category, sequence conservation was much more abundant in clusters that were associated to the same genes in both species. Almost half of these clusters contained sequence conserved enhancers (C-C-S), making 20.9% of all clusters (Figure 3.8b). This sequence conservation goes beyond the mouse-human comparison and extends across the vertebrate species (Figure 3.9b). Analysis of chromatin signatures revealed that the same category,

C-C-S, distinguishes itself from all other categories by stronger enrichment of both ATAC signal and all chromatin modifications (Figure 3.9c). This difference was particularly noticeable for H3K27ac, the mark that reflects enhancer activity level. Additionally, over 10% of C-C-S enhancers belonged to the category of enhancers highly occupied by TFs (as defined in Section 3.3.4.1) (Figure 3.9c). Together, these results suggest that C-C-S clusters play an important role in pancreatic islets.

Remarkably, absence of an active enhancer in human (C-A) occurred for 22.2% of the mouse clusters (Figure 3.8b). To assess the functional consequences of these events, expression levels of genes associated to the categorized enhancer clusters were compared between mouse and human. The results show that the absence of a cluster in human is indeed correlated with a significant loss of expression levels in human (Figure 3.10). The difference in expression is subtle in the case of C-O events, but quite prominent in the event of complete absence of a regulatory element in human (C-A).

Taken together, these results characterize the commonalities and differences between the regulatory landscapes of enhancer clusters in mouse and human pancreatic islets. While the overall landscape showed great flexibility between the species, clustered enhancers with both syntenic and sequence conservation emerged as highly active and often highly occupied by islet TFs. Furthermore, the results identify a species-specific subset of enhancer cluster regulatory events and demonstrate that it is often associated with changes of gene expression between species.

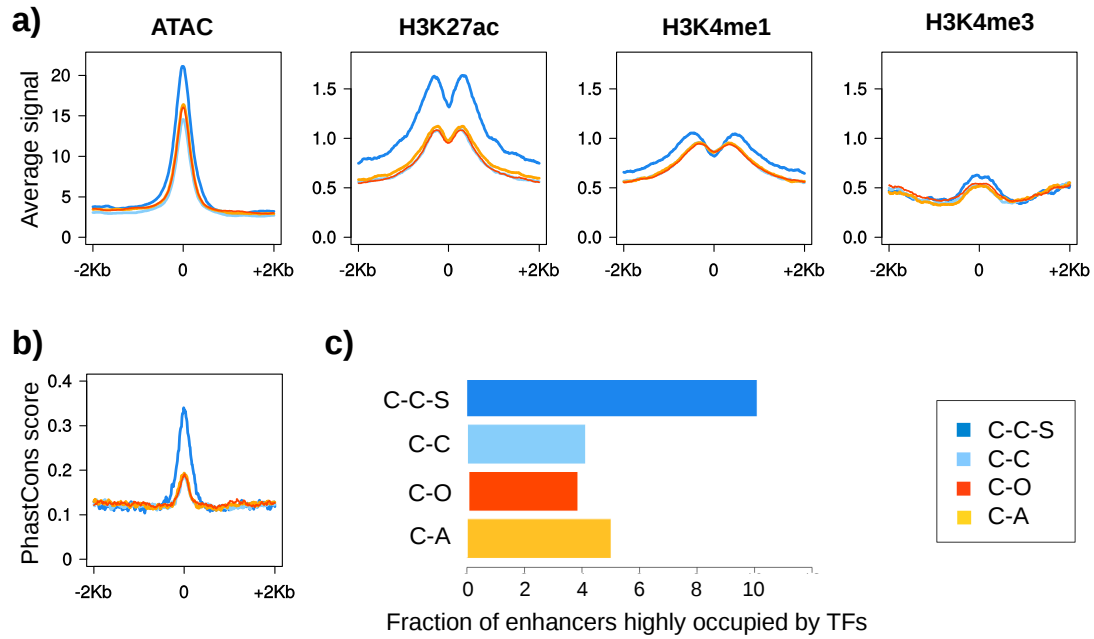


Figure 3.9 Properties of conservation-based enhancer cluster categories. a) Chromatin signatures, b) conservation (PhastCons scores based on 30 vertebrate genomes) and c) TF occupancy of clustered enhancer categories. Abbreviations are explained in the main text. The colour legend (bottom right) applies to all sub-figures.

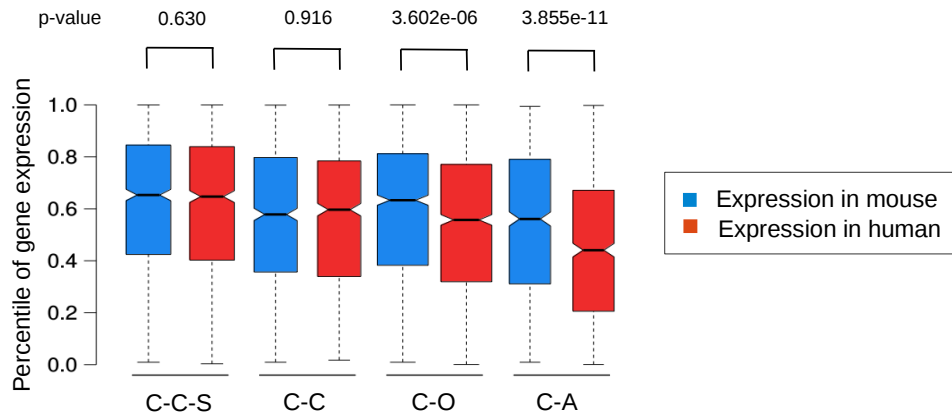


Figure 3.10 Expression levels of genes associated to conservation-based enhancer cluster categories. The significance of gene expression differences was assessed with the Wilcoxon rank sum test. Abbreviations are explained in the main text.

3.3.5.2 Enhancer cluster conservation and biology of pancreatic islets

To correlate the changes in the regulatory landscape of enhancer clusters to biology of pancreatic islets, gene ontology analysis was performed on the cluster-associated genes. Remarkably, the analysis revealed that the four categories of clustered enhancers regulate biologically distinct categories of genes. The C-C-S clustered enhancers were shown to control the core programme of pancreatic beta cells (Figure 3.11). Genes regulating beta cell development, insulin signalling and genes associated with ChREBP, a glucose-regulated transcription factor, were all present in the top 10 enriched ontologies for clustered enhancers strongly conserved between mouse and human. These results are consistent with previous reports of high sequence conservation of enhancers associated to genes with tissue-specific functions (Villar et al., 2015; Visel et al., 2009).

The C-C category of clusters was surprisingly enriched for several immune-related ontologies (Figure 3.11). This enrichment was partially driven by genes involved in transforming growth factor-beta/Smad3 signalling, a pathway essential for both T-cells and pancreatic islets (Delisle et al., 2013; El-Gohary et al., 2008). The largest category, C-O, showed enrichment for an array of ontologies, from immune-related to trans-membrane and small-molecule transport. For example, genes involved in Wnt-signalling pathway were associated with this category of clusters. This is another pathway well characterized in immune cells, but also essential for regulating beta cell mass in mouse islets (Rulifson et al., 2007; Staal et al., 2008).

C-A, the category of enhancers that has shown the most dramatic change in expression between mouse and human, revealed the most interesting set of gene ontologies. These genes, that are apparently deprived of association to active regulatory elements in human islets, are highly enriched in ontologies related to growth factors and cell survival signalling pathways (Figure 3.11). Some of these pathways were already implicated in beta cell survival and proliferation (Hart et al., 2000; Navarro-Tableros et al., 2004).

An example of a gene associated to clustered enhancers in mouse but not human is *Ghr*. As previously reported by Benner et al. (2014), this gene is highly expressed in mouse, but lowly expressed in human islets. *Ghr* encodes for a receptor of the the growth hormone (GH), a factor involved in beta cell proliferation (Sun and Bartke, 2015; Wu et al., 2011), a property that greatly differs between mouse and human islets (discussed in Section 1.4.2.1.4). The *Ghr* locus is syntenically conserved between the two species, with sequence conservation of 23 out

of 26 of the mouse enhancers in the human locus (Figure 3.12). The chromatin properties of those regions, however, differ between species. While mouse *Ghr* harbours a large cluster of active enhancers, the corresponding regions in the human genome exhibits very low levels of H3K27ac. This particular scenario demonstrates that conservation of sequence does not necessarily lead to a conservation of function, opening the question of the mechanism underlying the differences in enhancer activity.

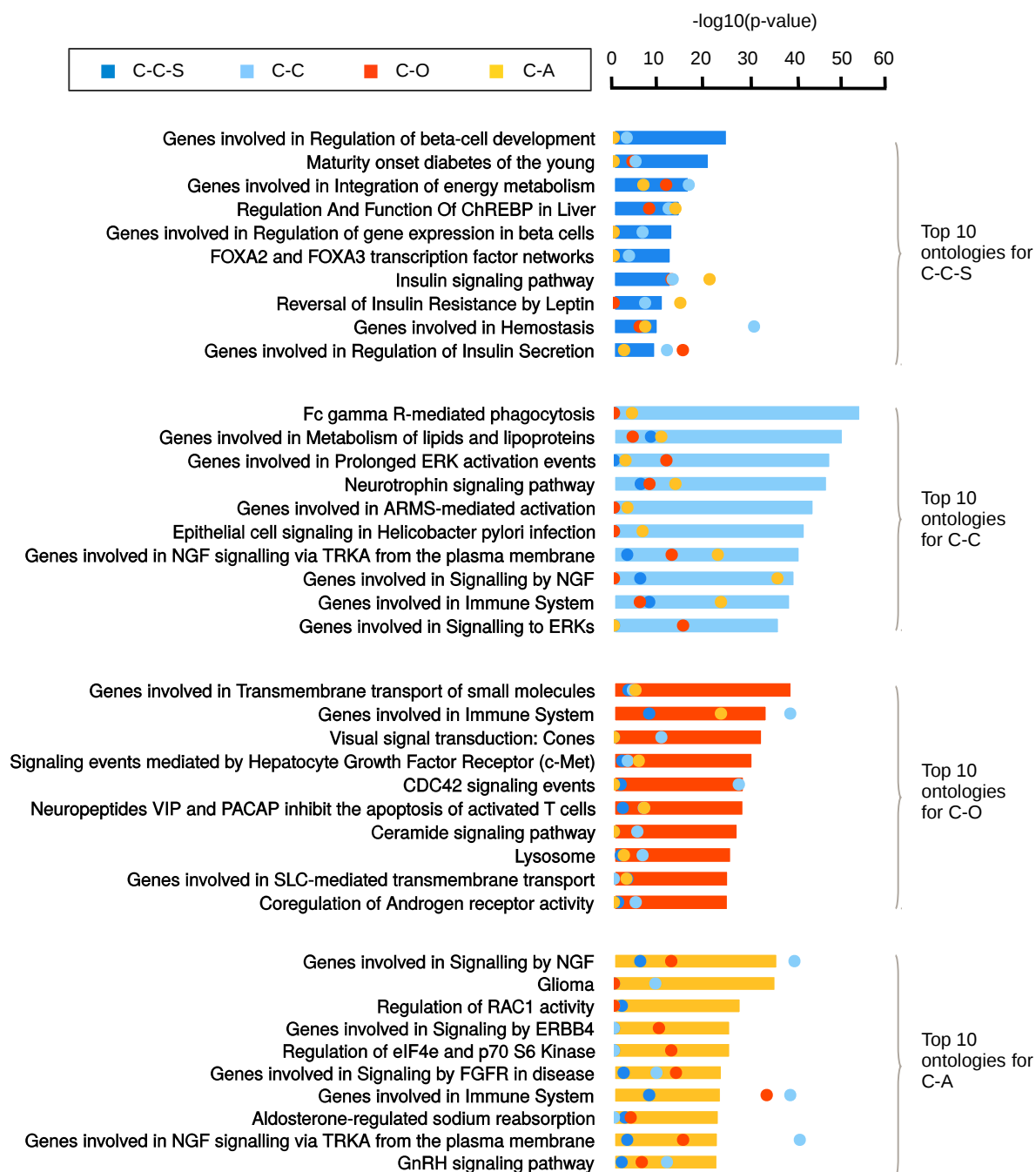


Figure 3.11 Gene ontology analysis for genes associated to conservation-based categories of clustered enhancers. Molecular signatures of genes associated with all four categories of clustered enhancers. Ten most enriched MSigDB ontologies are shown for each category, ranked according to the enrichment binomial p -value. The colour code applies to both bars and the points. Abbreviations are explained in the main text.

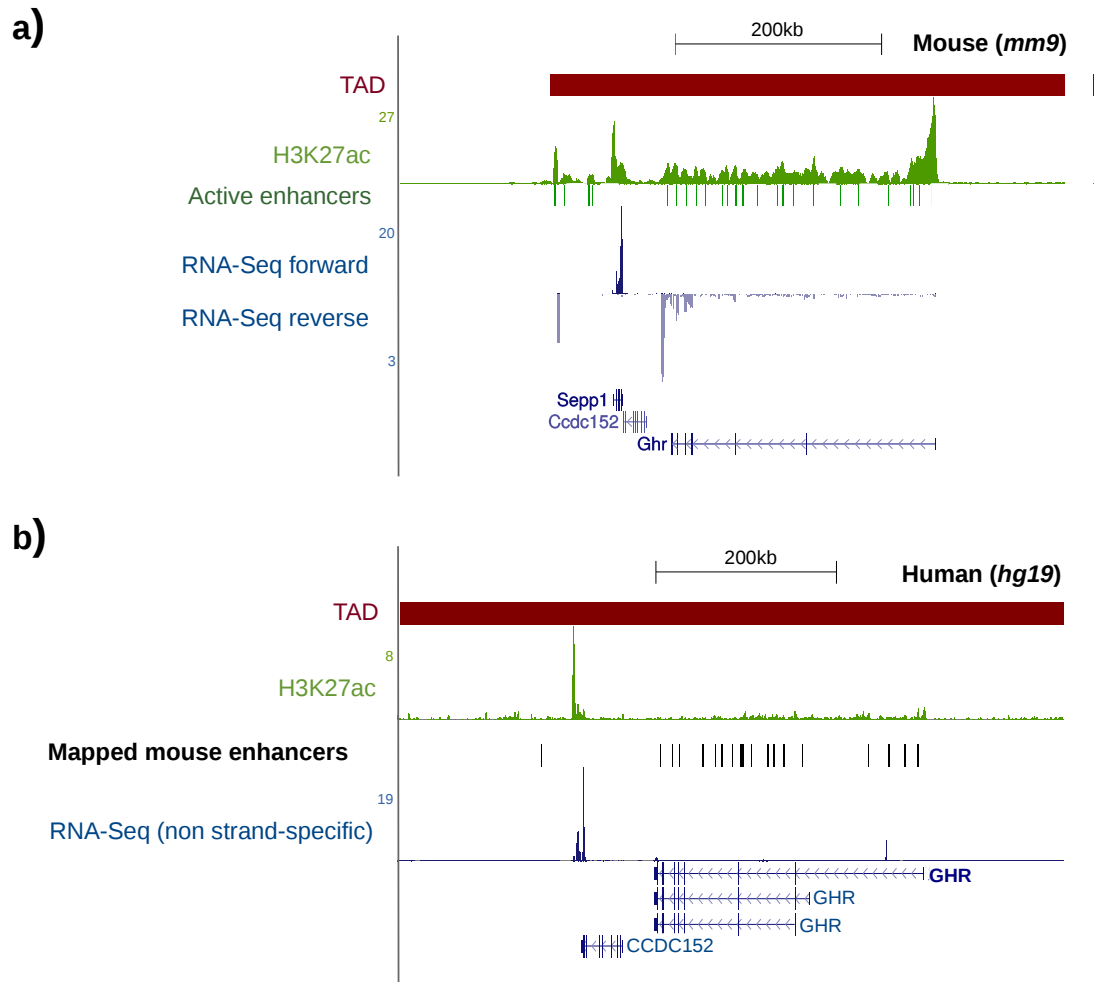


Figure 3.12 Enhancer and transcription maps of the *Ghr* gene locus in mouse and human. Genomic locus of the *Ghr* gene is syntenically conserved between mouse and human. a) Mouse *Ghr* gene is well expressed and the gene locus harbours a large number of active clustered enhancers, as reflected in levels of H3K27ac. b) Human *GHR* gene is poorly expressed. Genomic regions corresponding to the mouse active enhancers exhibit low levels of H3K27ac and are not characterized as active enhancers in human islets.

3.4 Discussion

3.4.1 Mouse pancreatic islet regulation resource

In this study, I present the most comprehensive characterization of mouse islet regulatory regions to date. Combination of H3K4me3, H3K27ac and H3K4me1 histone marks with open chromatin regions enabled me to create a reference catalogue of 78,921 islet regulatory elements and classify them as active promoters, active enhancers and inactive enhancers (Figure 3.1 and Figure 3.2 in Section 3.3.1). Additionally, I annotated genome-wide binding of five representative islet transcription factors, two of which (Nk2-2, Nkx6-1) were never previously assayed in mouse islets, to my knowledge (Figure 3.3 in Section 3.3.2). I further focused my analysis on active enhancers, elements that are thought to control time- and cell-specific patterns of gene expression.

Identification of active enhancers allowed me to examine the sequences underlying these regulatory elements and identify 25 non-redundant enriched sequence motifs (Table 3.2 in Section 3.3.3). These sequences reflected current knowledge of islet transcription and were highly enriched for target sequences of known islet regulators, such as FoxA1, Nkx6-1, Hnf1a/b and MafA. Motif degeneration, together with similarity of motifs targeted by different members of the same family of proteins, however, sometimes leads to imprecise matching of *de novo* motifs to motifs in annotated databases. Enriched motif matched to MyoD1, for example, is most likely targeted by NeuroD1 - a transcription factor that belongs to the same bHLH protein family, yet unlike MyoD1, plays an essential role in beta cell development and function (Gasa et al., 2004). Similarly, a motif was assigned to Rfx5, a factor with a known role in lymphocyte function and no known associations to pancreatic islets (Steimle et al., 1995). However, other members of the Rfx family, Rfx3 and Rfx6, have a well documented role in islet formation and production of hormones (Smith et al., 2010; Soyer et al., 2010). Overall, this analysis provided insight into the sequences directing gene regulation in mouse pancreatic islets. The compilation of transcription factor target sequences described in this chapter may be used in mouse model studies of genetic variation in pancreatic islets.

Not all active enhancers have the same properties. In this chapter, active enhancers were further categorized considering their organisation in the linear genome and occupancy by transcription factors. A number of studies demonstrated that the number and strength of TF binding is highly

predictive of enhancer activity (Chen et al., 2008; Dogan et al., 2015; He et al., 2011; Vanhille et al., 2015; Xu et al., 2012). Therefore, to identify highly active enhancers in mouse pancreatic islets, the combinatorial signal of the five islet TFs was used to categorize active enhancers as highly occupied, moderately occupied or not occupied by TFs (Figure 3.5 in Section 3.3.4.1). This approach of identification of highly occupied enhancers was adapted from Whyte et al. (2013) who based their categorization on the levels of occupancy by the Mediator complex. Similarly to the super-enhancers identified by Whyte et al. (2013), the highly occupied enhancers were shown to be an exclusive category, making only 4.5% of all active enhancers. Furthermore, these elements were characterized by strong activity and chromatin accessibility, indicating an important role in pancreatic islets. This analysis, however, was limited to the five TFs analyzed in this study, potentially missing additional islet-relevant enhancers.

Several approaches for defining groups of adjacent regulatory elements have been described (Gaulton et al., 2010; Parker et al., 2013; Pasquali et al., 2014; Whyte et al., 2013). Clustered enhancers in this chapter were defined following the approach of Pasquali et al. (2014), while the rest of the active enhancers were considered orphans. Clustered enhancers make the major group of islet active enhancers (64.6%). In line with the findings described in Pasquali et al. (2014), they are highly active and associate with genes involved in core processes of the pancreatic beta cell (Figure 3.7 in Section 3.3.4.2). These results open an interesting question of the identity of orphan active enhancers. While they carry histone signatures of active enhancers, they apparently do not associate with genes involved in tissue-specific processes. Although it is possible that some of them might be miss-annotated inactive enhancers, alternative explanations should be explored. A more detailed comparison of their properties with the clustered enhancers, such as as a comparative motif analysis or assessment of the functional impact they have on gene expression, might shed light on the results presented here.

Overall, this study produced a compendium of information on mouse islet regulatory elements and sequences, which will be of great use in study of pancreatic islet physiology and disease.

3.4.2 Patterns of mouse-human islet regulatory conservation

Integrative analysis of chromatin states revealed 30,259 open chromatin regions that carried chromatin marks characteristic for active enhancers (Figure 3.2 in Section 3.3.1). This number

is very similar to the number of putative active enhancers described in human pancreatic islets (30,816) by Pasquali et al. (2014). Considering that both genomes hold a similar number of genes, 80% of which are estimated to be orthologous (Waterston et al., 2002), this would suggest that a similar average number of enhancers controls expression of each gene in both mouse and human islets.

Analysis of mouse islet transcription factor regulatory networks (Figure 3.3 in Section 3.3.2) echoes the properties already described in human islets. In both species, the binding of lineage-specific transcription factors was shown to be confined to open chromatin sites and unequally distributed across the types of regulatory elements. In line with the binding patterns observed in humans, Pdx1 and Nkx2.2, were found to bind more frequently to promoters. The binding signal of all analyzed TFs apart from Pdx1, however, is strongest in active enhancers. This property is particularly highlighted in Nkx6.1 and NeuroD1, the two TFs with an expression profile most specific for mature islet beta cells (Itkin-Ansari et al., 2005; Taylor et al., 2013). This finding is consistent with the currently accepted view of enhancers as drivers of tissue-specific gene expression (Heintzman et al., 2009; Visel et al., 2009). Mirroring the binding patterns in humans, the five analyzed TFs form a self- and inter-regulatory network by binding to regulatory regions associated with their own other TF genes (Figure 3.1 in Section 3.3.2). Together, these results reveal an overall conservation of transcription factor binding patterns between mouse and human islets.

Analysis of the enhancer sequences discovered similarities between motifs found enriched in mouse and human, indicating a conservation of the regulatory language between the two species (Table 3.2 in Section 3.3.3). The similarities extended beyond the sequences of known transcription factors. For example, a novel mouse motif (NovelMotif10) was best matched with a novel motif identified in human (IEF.5). This strong motif could thus potentially bind a yet unknown, but evolutionary conserved, regulator of pancreatic islet function. Identification of the exact protein binding to this motif would require further investigation. Comparing the NovelMotif10 with databases of known TF motifs using a more relaxed similarity cut-off could provide information about the protein family the target TF belongs to. This can be followed up on by investigating which members of that family are expressed in pancreatic islets. In addition, complexes bound to the NovelMotif10 sequence can be isolated and further dissected by mass spectrometry, identifying candidate target proteins (as described in Lee et al. (2015)). Once candidate proteins are selected, the target protein can be identified by an mobility super-shift

assay (Pares-Matos, 2005) using antibodies for all candidate proteins and establishing which one leads to a super-shift.

The number and overall properties of larger regulatory structures - clustered enhancers, was also very similar between the two species. Almost half of the clusters were associated with the same gene in mouse and human, reflecting syntenic and functional conservation (Figure 3.8 in Section 3.3.5). Recent work from Yang et al. demonstrated that sequence conservation is not necessary for conservation of function in a syntenic environment (Yang et al., 2015). Supporting that notion, more than half of syntenically conserved clusters in mouse islet did not harbour sequence conserved elements. Interestingly, approximately 4% of enhancers in those clusters were shown to be highly occupied with islet transcription factors, despite lacking sequence conservation. This suggests that the motif grammar of those enhancers diverged in a way that still allowed binding of the same transcription factors. As discussed in Section 1.1.3.3.1, differences in enhancer sequence that would impact the motif grammar (order, orientation and spacing of TF motifs) may still result in the same regulatory output. In addition, although binding of one TF is conserved, co-binding factors (and their target sequences) may differ between mouse and human (Yang et al., 2015). The traditional approach of using sequence conservation for identification of orthologous regulatory elements would not have identified these regions as relevant. However, lack of sequence conservation does not seem to impair the functionality of these elements, as there was no difference in expression of genes associated to these enhancers between mouse and human.

It is conservation of sequence that was shown to underlie the enhancers associated with genes crucial for islet and beta-cell development and function (Figure 3.9 in Section 3.3.5). While strong sequence conservation of developmental enhancers has long been documented (Bejerano et al., 2004; Woolfe et al., 2005), high conservation of mature tissue-specific enhancers was described more recently (Villar et al., 2015). In line with that, even though relaxed criteria for sequence conservation were used for this analysis, enhancers composing these clusters are not only conserved between mouse and human, but across vertebrates. Compared to other cluster categories, this particular subset of clustered enhancers displayed chromatin properties indicating very high activity and occupation with assayed islet transcription factors, further reinforcing their central role in beta cell function.

Overall, the results presented here demonstrate that the regulatory mechanisms of mature mouse and human pancreatic islets share the general architecture and transcription factor net-

works. Moreover, the regulatory landscape of central tissue-specific genes is conserved in both function and sequence, affirming mouse as a valid model in studies of pancreatic islet gene regulation.

3.4.3 Species-specific subset of islet regulatory architecture

Comparative analysis of regulatory elements suggests that the divergence between mouse and human species resulted in reorganisation of the majority of the pancreatic islets active enhancer regions. This is to be expected, as changes in regulatory regions were found to be the major determinant of intra- and inter-species differences (Varki and Altheide, 2005; Wray, 2007). In fact, Villar et al. (2015) discovered that recently evolved enhancers are often associated with positively selected genes and may point to regulatory elements behind species-specific adaptation. In this study, I focused on enhancer clusters and characterized a species-specific subset of the mouse pancreatic islet regulome.

The mouse-specific regulatory regions are associated with genes that are apparently not related to beta cell function (Figure 3.11 in Section 3.3.5). Interestingly, the gene ontology revealed enrichment of growth-factor signalling and survival genes. Nerve growth factor (NGF) signalling, the top ontology enriched for genes associated to mouse-specific enhancers, was identified as essential in beta cell survival and apoptosis evasion in both rat islets and human insulinoma cell lines (Navarro-Tableros et al., 2004; Pierucci et al., 2001). Fibroblast growth factor (FGF) signalling in mature mouse islets was implicated in T2D-like characteristics, including reduced number of beta cells, and suggested that FGF might be a key factor for the stimulation of beta cell expansion (Hart et al., 2000). Differences in beta cell proliferative capacity between human and rodent islets have long been acknowledged (Chandrasekera and Pippin, 2013; Eizirik et al., 1994). Results described in this chapter suggests that these species-specific differences in phenotype can at least partially be explained by changes in *cis*-regulation.

Beta cell destruction and incapacity to raise a proliferative challenge, such as gestation or obesity, are considered hallmarks of diabetes. Great efforts have thus been directed into research that would discover ways of preserving or expanding beta cell mass (Ashcroft and Rorsman, 2012; Donath and Halban, 2004). Due to difficulties in obtaining human pancreatic tissue and a lack of appropriate in vitro human model, a large part of this research is performed in the mouse model. Differences in mouse and human islet biology, however, often make that research not

translatable to human (Chandrasekera and Pippin, 2013; Chatzigeorgiou et al., 2009; Vetere et al., 2014). Growth hormone (GH), for example, is an important factor involved in both insulin synthesis and beta cell proliferation (Sun and Bartke, 2015; Wu et al., 2011). Growth hormone receptor (*Ghr*) forms a part of the JAK-STAT pathway and is implicated in beta cell proliferation during gestation in mice (Lindberg et al., 2005; Nielsen et al., 2001). However, beta cell mass increase during human gestation is not as drastic as in mice, nor is it mediated through beta cell proliferation (Butler et al., 2010). Furthermore, growth factor stimulation successfully induces *in vitro* proliferation of murine beta cells while the same treatment fails to elicit a response in human cells (Parnaud et al., 2008). *Ghr* was recently found to be differentially expressed between mouse and human islets (Benner et al., 2014). In this study, I reveal species-specific activity of the *Ghr cis*-regulatory landscape, suggesting that the differences in phenotype might arise from differences in the *cis*-regulatory landscape (Figure 3.12 in Section 3.3.5).

Sequence conservation of a large number of clustered enhancers associated to the *Ghr* gene raises the question of the mechanism behind the cluster loss of function. As this study did not seek to identify ultra-conserved elements (with 100% sequence identity), it is possible that the variation in sequence present between mouse and human was sufficient to disrupt crucial TF binding sites leading to failure of enhancer activation. A further possibility is that the enhancer key for activation of the whole enhancer cluster was not conserved between the two species. Out of 26 clustered enhancers associated with *Ghr* in mouse, 23 were successfully mapped to human, making this scenario possible (Figure 3.12 in Section 3.3.5). The loss of enhancer activity could be a consequence of appearance of a novel transcriptional repressor. An example of this scenario was described by Ochi et al. (2012), who reported that a difference in activity of *pax4* and *pax8*, two paralogs with well conserved non-coding regulatory elements, was a consequence of a new silencer motif present in the proximal promoter of *pax8*. Further research will be necessary to answer these questions, including a more detailed comparison of sequences underlying the enhancers inactive in human and investigating whether the human *GHR* locus is subject to *trans*-silencing processes.

The results described in this chapter contribute to our understanding of the differences in biology between mouse and human islets. Further in-depth analysis of chromatin and underlying sequence changes between mouse and human could provide insight into the mechanism behind species-specific enhancers activity and gene regulation. Moreover, the same approach has

the potential of revealing regulatory mechanisms driving beta cell proliferation, which would greatly contribute to the research efforts made by the diabetes community.

Chapter 4

Dissecting *cis*-regulatory Variation in Pancreatic Islets of a Hybrid Mouse Model

4.1 Introduction

Regulation of gene expression includes a wide range of mechanisms that define and maintain cell types and shape higher level phenotypes in health and disease. These mechanisms are subject to genetic variation, much of which is harboured in the non-coding regulatory sequences of the genome. Understanding the relationship between genetic variation and gene expression is challenging, particularly for complex traits which are believed to result from the interaction of multiple genetic events and the environment. T2D is a complex disease with evidence of a strong genetic basis (discussed in Section 1.3.3.2). To date, GWA studies successfully identified statistical association between T2D and approximately 80 loci across the genome. The transition between identification of these genetic associations to elucidating the molecular mechanisms behind them has been very slow. The majority of T2D-associated variants were associated with beta-cell function and map to non-coding regions (discussed in Section 1.3.3.2), urging for better understanding of the *cis*-regulatory networks of the pancreatic islet beta cells.

4.1.1 *Cis*-regulatory variation

Genetic variation within a *cis*-regulatory element has a potential of disrupting the interaction between a transcription factor (TF) and the target sequence. Several studies have demonstrated genetically variable binding of RNA polymerase II and tissue-specific transcription factors, as well as their correlation with gene expression (Birney et al., 2010; Kasowski et al., 2010; Kilpinen et al., 2013; Musunuru et al., 2010; Reddy et al., 2012). McVicker et al. (2013) and Kilpinen et al. (2013) further identified a subset of disrupted TF sites correlating to changes in

histone modifications and chromatin accessibility. Together, these studies propose a chain of events where changes in TF binding modify the activity of regulatory elements, which in turn impact gene expression.

Open chromatin sites, particularly enhancers, tend to be enriched for disease-associated variants (Ernst et al., 2011; Farh et al., 2014; Maurano et al., 2012). Several recent studies have reported that enhancers displaying allele-specific activity generally tend to be located in the vicinity of allelically imbalanced genes (Dixon et al., 2015; Leung et al., 2015) and are in more frequent contact with the target promoter (Dixon et al., 2015). However, it is not clear if changes in activity of regulatory elements always lead to changes in gene expression. Kasowski et al. (2013), for example, reported enhancers to be the most variable regulatory element between individuals, but observed that only a small subset of variable enhancer activity correlates with changes in gene expression. Taken together, these studies raise questions about what determines the functional impact of enhancers on gene expression and which genes tend to be susceptible to variation in regulatory elements.

To address these questions, it is necessary to conduct a systematic assessment of the effect genetic variation has on both enhancer activity and associated gene expression. In the last few years several high-throughput reporter assays have been designed for this purpose (Arnold et al., 2014; Melnikov et al., 2012; Patwardhan et al., 2012; White et al., 2013). Although these studies provided valuable insight into the principles of regulatory variation, they are limited to assessing the activity of regulatory elements outside of their genomic context. This calls for an approach that allows high-throughput assessment of *cis*-regulatory effects *in vivo*.

4.1.2 Allele-specific analysis in a hybrid mouse model

In this study, a mouse hybrid is used as a model of *in vivo* mutagenesis. The genetic diversity in the hybrid mouse is ensured by breeding two evolutionary distinct strains of mice - C57BL/6J, the most commonly used laboratory strain, and CAST/Ei, a wild derived strain (use of mice in genetic research discussed in Section 1.4.1). The naturally occurring genetic variation between these two strains serves two purposes. A SNP or an INDEL may disrupt a functionally important sequence and drive differences in activity of a regulatory element and/or gene expression between the two genomes. At the same time, a genetic variant can act as a genetic marker and allow assignment of transcripts or chromatin marks to one genome or the other. In combination

with transcriptome and chromatin assays, this model enables an assessment of the differences in gene expression and chromatin states driven by genetic variation.

Differential activity between alleles of genes or regulatory regions is measured by allele-specific analysis (ASA) (discussed in Section 1.2.2.2.2). Single-base resolution of high-throughput sequencing technology makes allele-specific analysis a powerful tool for genome-wide detection of genetic effects on gene regulation. The advantages of ASA compared to the more commonly used quantitative trait loci (QTL) approaches lie in elimination of environmental and *trans* effects on gene expression within an individual, given that both maternal and paternal genome are exposed to the same *trans*-acting factors and environmental conditions (Pastinen, 2010). While allele-specific analysis can be, and has been performed in human tissues (Dixon et al., 2015; Kasowski et al., 2013; Kilpinen et al., 2013; Leung et al., 2015; McVicker et al., 2013), there are a number of advantages to using a hybrid mouse model.

Allele-specific analysis is restricted to informative genes and regulatory regions - regions that carry at least one heterozygous SNP. Compared to many species, humans have a relatively low level of intraspecies sequence diversity (Chen and Li, 2001). An average human genome differs from the standard reference genome in approximately 4.1 million to 5.0 million sites (Auton et al., 2015). Furthermore, single nucleotide polymorphisms (SNPs) are inherited in linkage disequilibrium (LD) blocks, which often hinders identification of the variant causing the allele-specific activity. In contrast, the genomes of C57BL/6J and CAST/Ei differ in 17.6 million SNPs and 2.5 million INDELs (Keane et al., 2011). In addition, these two strains are inbred, ensuring extensive linkage disequilibrium and facilitating integration of allele-specific expression with allele-specific activity of regulatory regions.

The majority of studies using allele-specific analysis performed to date aimed to estimate allelic imbalance within individual samples. However, allelic imbalance detected in an individual sample may be a result of artefacts related to inaccurate or incomplete genotyping or library preparation and sequencing (DeVeale et al., 2012). In addition, allele-specific analysis reveals imbalance driven not only by *cis* variation, but also monoallelic activity that is parent-of-origin (Reik and Walter, 2001), X-inactivation related (Marks et al., 2015) or stochastic (Reinius and Sandberg, 2015). This study was designed to enrich for allelic imbalance resulting from *cis*-acting variants. To reduce technical and stochastic noise, an allele-specific analysis pipeline was developed to detect allelic imbalance consistent across multiple samples. Furthermore, to remove parent-of-origin effects, samples were derived from hybrid offspring from both recip-

rocal crosses of pure-strain mice.

Finally, a general advantage of using a mouse model is facilitated accessibility to the tissue of interest. Pancreatic islets are central to diabetes pathogenesis, but are often hard to obtain from human in sufficient number or quality. As established in Chapter 3, the *cis*-regulatory architecture of mouse and human islets is generally well conserved, making mouse islets a valuable source of knowledge on *cis*-regulation in pancreatic islets in general.

Although allele-specific analysis of next generation sequencing data was first implemented almost a decade ago (Babak et al., 2008; Serre et al., 2008; Wang et al., 2008b), there is no well established pipeline for such analysis. Numerous approaches have recently been published, applicable to RNA-Seq and/or ChIP-Seq data derived from human or animal models (Andergassen et al., 2015; Harvey et al., 2015; León-Novelo et al., 2014; Romanel et al., 2015; van de Geijn et al., 2015). Approaches developed for use on human samples generally aim to establish allelic imbalance on a per-sample basis, while approaches developed for use in animal models often aim to detect parent-of-origin effects. In order to focus on allelic events driven by *cis*-regulatory variation, a new approach to allele-specific analysis was taken in this study, aiming to remove biases of technical origin, as well as stochastic and parent-of-origin effects.

Ultimately, a comprehensive survey of allelic activity of regulatory elements, chromatin accessibility and gene expression conducted in this study provides insights into gene regulation of pancreatic islets that may be relevant to human genetic variation and disease.

4.1.3 Summary

Genetic variation in the non-coding genome is thought to explain the majority of both inter-species and a large part of intraspecies differences in phenotypes, including an array of heritable disease-related traits. Identifying the principles of *cis*-regulatory variation would therefore provide insight into how regulatory elements control gene expression, as well as how noncoding variants alter disease risk. It is unclear, however, if all *cis*-regulatory variants have a functional impact and what are the properties of the variants that do lead to changes in gene expression. To answer these questions, a genome-wide analysis of *cis*-regulatory variation must be performed in its native genomic context. While this type of analysis is challenging to perform in human, a hybrid mouse model provides the naturally occurring genetic variation which can be used for dissecting the properties of functional *cis*-variation that acts in the tissue of interest.

The specific aims of this chapter thus are;

1. To construct a pipeline for allele-specific analysis of RNA-Seq, ChIP-Seq and ATAC-Seq focused on allelic events driven by *cis*-regulatory variation
2. Create a genome-wide map of allelic expression and chromatin states in pancreatic islets
3. Characterize genes susceptible to *cis*-regulatory variation
4. Define properties of *cis*-regulatory variation associated with allelic expression

4.2 Methods

4.2.1 Allele-specific analysis pipeline

A customized analysis pipeline was developed for assessing allelic imbalance in the transcriptome, open chromatin and H3K27ac (see Section 4.3.1.1, Figure 4.2). The pipeline is able to process RNA-Seq, ChIP-Seq and ATAC-Seq reads. The choice of the read alignment algorithm depended on the assay and read type. The pipeline consists of two major parts, aiming to;

- 1) Determine the allelic ratios of regions in individual samples
- 2) Establish consistent and statistically significant allelic imbalance in regions of interest across multiple samples

4.2.1.1 Allele-specific analysis in individual samples

This part of the analysis aims to estimate allelic imbalance across regions in an individual sample, while accounting for possible artefacts originating from alignment biases, clonality and mis-genotyping.

4.2.1.1.1 Unbiased alignment

In order to avoid alignments biases, the NGS reads were aligned independently to the genomes of both C57BL/6J (*mm9*, the reference mouse genome) and CAST/Ei. The compiled CAST/Ei genome is not yet available and had to be constructed from data available.

- 1) Constructing the CAST genome

CAST/Ei pseudogenome was compiled using *vcftools* (Danecek et al., 2011). The *mm9* genome was used as the backbone, while CAST/Ei SNPs and INDELs were filtered for a Phred score >20 (released by the Sanger Institute, Keane et al., 2011). In total, 17,673,726 SNPs and 2,534,008 INDELs were incorporated, calling:

```
java -Xmx16000m -jar vcf2diploid.jar -id CAST -chr mm9.fa -vcf  
cast_SNPs_filtered.vcf cast_INDELs_filtered.vcf
```

Incorporation of INDELs resulted in a shift in coordinates between the new CAST/Ei pseudogenome, *cast*, and the reference *mm9*. Coordinates were tracked in a “chain file” created by *vcftools*. This file was used in the analysis downstream together with *liftOver* (Hinrichs et al., 2006) to translate regions, SNPs or INDELs from *cast* to *mm9* genome coordinates.

2) Alignments

Two independent alignments were performed for each NGS data set, one to the *mm9* genome, and the other to the *cast* pseudogenome. The RNA-Seq was aligned in a non-strand-specific manner using *Tophat2* (Kim et al., 2013), and the chromatin data by *bowtie2* (Langmead and Salzberg, 2012), accounting for single (SE) or paired-end (PE) reads.

RNA-Seq;

```
tophat -p 8 -N 4 --read-gap-length 3 --read-edit-dist 4 -g 1 --no-coverage-search --segment-length 22 --library-type fr-firststrand --b2-N 1 genome reads_1.fastq reads_2.fastq
```

ChIP-Seq, SE reads;

```
bowtie2 -p 8 -x genome reads.fastq -S outfile.sam
```

ATAC-Seq, PE reads;

```
bowtie2 -p 8 -x genome -X2000 -1 reads_1.fastq -2 reads_2.fastq -S outfile.sam
```

3) Compiling a non-redundant read set

Reads from alignments to the two genomes were compared and combined as a non-redundant set of reads, using an in-house Perl script. As each read was aligned to both *mm9* and *cast* genomes, alignment scores were compared between the two alignments. The aligned read with a higher alignment score (reflecting less mismatches) was selected for the final non-redundant set of reads. If the alignment score was identical between the two alignments, one of the aligned reads was picked randomly.

4.2.1.1.2 Read counts

Numbers of reads originating from C57BL/6J or CAST/Ei genomes were derived from counting the number of reads covering a heterozygous SNP that carry the nucleotide found in *mm9* or *cast* genome. The reads were first piled up over SNP locations separately for the two genomes using *samtools* (Li et al., 2009). Only reads with a Phred score greater than 20 at the base covering the SNP were taken into account.

```
samtools mpileup -B -Q20 -f genome -l snp_list.bed aligned_reads.bam >
outfile.pileup
```

samtools output format was converted into read counts and the counts coming from the two genomes were combined using in-house Perl scripts.

4.2.1.1.3 Filtering

In order to ensure the reliability of information provided by each SNP and remove biases introduced by library preparation or alignment, extensive filtering was done. This was implemented using a series of in-house Perl and Bash scripts:

1) Clonality

To remove biases caused by preferential PCR amplification of reads derived from one of the two genomes, two criteria were used for identifying whether the SNP will be kept for further analysis;

- A minimum number of 20 non-clonal reads covering the SNP. Clonal reads were defined as reads aligned to identical coordinates and harbouring identical sequences.
- Preservation of the allelic ratio after the removal of clonal reads. Ratio of the number of *mm9*-derived reads to the total number of reads covering a SNP (allelic ratio to reference) was calculated before and after removal of clonal reads. The difference between the allelic ratios (*delta*) was calculated across all SNPs within a sample. The SNPs with a *delta* value below the 25th percentile of the sample distribution was considered unchanged.

2) Read independence

One read can span multiple SNPs. Therefore, to ensure that one read provides information about the allelic imbalance of the region only once, SNPs were filtered for inter-SNP

distance. In the case multiple SNPs were separated by a distance smaller than the read length (90 or 100bp) only one SNP was kept for downstream analysis - the SNP with the highest total read coverage.

3) Presence of non-genotyped variance

Incomplete/inaccurate genotyping or sequencing errors lead to presence of read nucleotides that match neither *mm9* nor *cast* genomes. SNPs covered with more than 10% of such read nucleotides were removed from the analysis.

4) Allelic ratio in pure strains

In the case of a true unbiased read alignment, the allelic ratio of reads derived from pure-strain C57BL/6J and CAST/Ei, processed in the allele-specific way, are expected to be 1 and 0, respectively. In the case of preferential alignment of reads to one of the genomes, one of the ratios would be distorted. Therefore, for the purpose of removing SNPs located in regions with biased read alignment, pure-strain derived datasets were processed with the allele-specific pipeline (as described up to this point) and allelic ratios of the SNPs were assessed on a per-sample basis. SNPs with allelic ratios smaller than 0.98 (in C57BL/6J) or greater than 0.2 (in CAST/Ei datasets) in more than two datasets were removed from the analysis.

4.2.1.1.4 Allelic ratios of informative regions

Final allelic ratios were expressed at the level of regions - genes (for the transcriptome) or regulatory elements (for the chromatin data). For each sample, the read counts covering filtered SNPs falling into the region of interest were summed. The allelic ratio (AR_{region}) was then expressed as the number of reads covering all informative SNPs in exons of a gene/regulatory element derived from *mm9* (r_s^{mm9}) over the number of reads derived from both *mm9* and *cast* ($r_s^{mm9} + r_s^{cast}$).

$$AR_{region} = \frac{\sum_{s=1}^{N_{SNP}} r_s^{mm9}}{\sum_{s=1}^{N_{SNP}} r_s^{mm9} + \sum_{s=1}^{N_{SNP}} r_s^{cast}}$$

Regions containing SNPs with inconsistent allelic ratios that deviated from 0.5 in opposite directions were removed from the analysis. The statistical significance of the deviation from 0.5 was tested with a two-tailed binomial test and reported as *p*-values.

4.2.1.2 Determining consistent and statistically significant allelic imbalance across samples

In order to detect *cis*-regulatory driven allelic imbalance in the hybrid mouse islets, the allelic information was combined across all hybrid samples. For regions that were informative in 3 or more samples, information from individual samples was combined in the following manner, using a series of in-house *R* scripts;

- The representative allelic ratio was calculated as the median of allelic ratios in individual samples
- Imbalance of combined allelic ratio was tested with a weighted Z-method (Whitlock, 2005).

$$Z_W = \frac{\sum_{i=1}^n Z_i w_i d_i}{\sqrt{\sum_{i=1}^n w_i^2}}$$

Each individual sample i was tested for allelic imbalance with a binomial test. p -values obtained that way were transformed back to Z-statistic, Z_i . To take into account differences in number of reads in each sample, that is, statistical power of the test, each Z-statistic is assigned a weight, w_i . The median number of informative reads across a region in a sample was used as a weight. The median allelic ratio of the region was used as a direction, d_i (1 for allelic ratios greater than 0.5 and -1 for allelic ratios smaller than 0.5). n denotes the number of individual hybrid samples in the region. p -value for the weighted Z-statistic, Z_W , was based on normal distribution where parameters (mean and standard deviation) were permutation based. The regions were grouped to 5 quantile bins according to read coverage and 1,000 permutations were performed by shuffling region names within a bin. p -values were corrected for multiple-hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995), implemented as *p.adjust* function in *R*. Regions with adjusted p -values smaller than 0.05 were considered to show statistically significant allelic imbalance.

Finally, to ensure parent-of-origin effects are not represented, regions with standard deviation of the allelic ratio between samples greater than 0.15 were removed from the downstream analysis.

4.2.2 Control set of regions without allelic imbalance

For the purposes of comparing properties of regions showing consistent and statistically significant allelic imbalance across samples with regions where such imbalance is absent, a control set of regions was constructed. The control regions were matched to the significantly imbalanced on total number of regions, number of informative reads covering the regions and standard deviation of allelic ratios across samples. This was achieved in two steps, using an in-house *R* script:

- 1) The significantly imbalanced regions (previously filtered for standard deviation (SD) of allelic ratios smaller than 0.15) were grouped to 10 quantile bins according to read coverage.
- 2) Regions with adjusted *p*-values greater than 0.05, filtered for SD of allelic ratios smaller than 0.15, and grouped to bins determined by read coverage of significantly imbalanced regions (in step 1). One region was randomly selected from each bin until the number of regions reached the number of significantly imbalanced regions or until there were no more regions left in either of the bins.

4.2.3 Quantification of transcription factor motif disruption

In order to assess potential differences between transcription factor (TF) binding strength between the sequences present in *mm9* and *cast* genome, a customized version of *sTRAP* algorithm was used (Manke et al., 2010; Thomas-Chollier et al., 2011). In short, this algorithm looks into pairs of genomic regions (sequences of the same region from two genomes), scans it for known TF motifs (Jaspar database, Mathelier et al., 2014), predicts the binding affinities of scanned TFs for both sequences and normalizes them to the affinity distribution of random background sequences (Manke et al., 2008). The log-ratio of the binding affinities *p*-values is used to rank all TFs according to their change in binding affinity for a particular region. Here, the algorithm was customized in the following manner;

- 1) A comprehensive and non-redundant set of *de-novo* motifs enriched in regulatory elements of the mouse pancreatic islet was added to the motif models already present in the Jaspar database. The *de-novo* motif discovery process is described in the Methods section of Chapter 3.

- 2) A custom set of background sequences was used for calculating the background motif affinity distribution. To account for differences in GC content, two sets of 5,000 background sequences (200-1,000bp long) were created using HOMER - one for active enhancers and another for active promoters. The affinity distributions were calculated using *sTRAP* function *fit-gev.ed.R*.

The customized *sTRAP* algorithm was run separately on enhancer and promoter sequences, using corresponding background affinity distributions. Affinity changes with log-ratio greater than 1.2 and corresponding *p*-value smaller than 0.01 were considered significant.

4.2.4 Tissue specificity

4.2.4.1 Genes

To assess tissue specificity of genes transcribed in mouse pancreatic islets, their expression levels were compared to expression levels in 19 non-pancreatic adult and fetal tissues (Shen et al., 2012). Tissue specificity was also calculated for pancreatic beta cells, using expression values obtained from FACS purified mouse beta cells published by Benner et al. (2014). The RNA-Seq derived expression levels were obtained as FPKM values, and the mean FPKM value was used when replicates for a tissue were available. Gene expression levels across all genes were quantile normalized and the tissue specificity was expressed as a Z-score of the normalized values. The Z-score was calculated for both pancreatic islets (pure strain C57BL/6J) and FACS purified beta cells.

4.2.4.2 Active enhancers

In order to establish the level of tissue specific activity of mouse pancreatic enhancers, enrichment of H3K27ac across those regions was examined in 23 non-pancreatic mouse tissues (Shen et al., 2012). The data from each tissue was obtained as ChIP-Seq-derived sets of MACS peaks. The overlap between the peaks and islet enhancers (n_i) was determined with *intersectBed* (Quinlan and Hall, 2010), asking for minimum 1bp overlap. Finally, for each enhancer (i), tissue specificity was expressed as a score TS_i ranging from 0 (no overlap in any of the tissues) to 1 (overlap found in all tissues):

$$TS_i = 1 - \frac{n_i}{23}$$

where n represents the number of tissues with an overlapping H3K27ac peak.

4.2.5 GLM analysis of properties associated to imbalanced genes

4.2.5.1 Gene and promoter properties

Several gene and promoter-associated properties were tested for association with allelic imbalance of gene expression; islet and beta cell tissue specificity scores (TS_i^{pi} and TS_i^b , respectively); promoter coverage by H3K4me3 ($K4_i$); average GC content (GC_i^p) and average PhastCons score (PC_i^p). Average RNA read coverage of informative SNPs across samples was introduced as an internal control of gene expression (RNA_i^g). For each gene (i), these properties were used as independent variables in a generalised linear model (GLM) while the imbalance of the gene was the response variable (y_i). Logit link function was used as response variable was binary, either balanced (coded as 0) or imbalanced gene (coded as 1). Full specification was the following

$$\text{logit}(y_i) = \beta_0 + \beta_1 TS_i^{pi} + \beta_2 TS_i^b + \beta_3 K4_i + \beta_4 GC_i^p + \beta_5 PC_i^p + \beta_6 RNA_i^g$$

where β_0 is the intercept and β_{1-6} represent the independent effect sizes of each property on the log odds of allelic expression. This GLM model was implemented in R through function *glm*.

4.2.5.2 Enhancer properties

Several enhancer properties were tested for association with allelic expression of genes located within 60kb from the enhancer; enhancer location (L_i , intergenic versus intronic or exonic); clustering (C_i , orphan versus clustered); TF occupancy (TF_i , not occupied versus highly occupied or moderately occupied); the strength of H3K27ac and ATAC signal ($K27_i^e$ and $ATAC_i^e$); tissue specificity (TS_i^e), GC content (GC_i^e) and sequence conservation (PC_i^e). To ensure that

the results are not biased by variation in expression of the associated gene, coverage of the associated gene was also included in the analysis (RNA_i^{eg}). The described properties were characterized for every H3K27ac-imbalanced enhancer (i) and used as independent variables associated with the presence of an allelically imbalanced gene in the region, which represented a binary predicted variable in the generalised linear model (y_i).

$$logit(y_i) = \beta_0 + \beta_1 L_i + \beta_2 C_i + \beta_3 T F_i + \beta_4 K 27_i^e + \beta_5 AT AC_i^e + \beta_6 T S_i^e + \beta_7 GC_i^e + \beta_8 PC_i^e + \beta_9 RNA_i^{eg}$$

where β_0 is the intercept and β_{1-9} represent the independent effect sizes of each property on the log odds of allelic expression. This GLM model was implemented in R through function *glm*.

4.3 Results

4.3.1 Mapping allele-specific expression and activity of regulatory elements in mouse pancreatic islets

In order to profile allele-specific events in adult pancreatic islets, the tissue was isolated from 16 week old male F1 progeny of both reciprocal crosses and pure strain controls (Figure 4.1). Genomes of C67BL/6J (C57) and CAST/Ei (CAST) differ approximately every 140bp, providing a base for high-density mapping of allele-specific events in hybrid mice. Several assays were performed in multiple replicates, each replicate containing a pool of pancreatic islets derived from a minimum of three animals, following the protocols described in Sections 2.2 and 3.2.1. The islet transcriptome was quantified by RNA-Seq (Wilhelm et al., 2008), chromatin accessibility with ATAC-Seq (Buenrostro et al., 2013) while H3K27ac ChIP-Seq signal served as a proxy for the level of activity of regulatory elements (Zhu et al., 2013). Finally, analysis of motif disruption was performed in order to determine the allele-specific events potentially driven by changes in binding affinity for canonical TF motifs. Together, this model provides a comprehensive insight in the sequence-driven changes of allele-specific transcription and regulatory elements in pancreatic islets.

Histone modification, ATAC and RNA sequencing data were visualised in the UCSC browser and can be accessed through "User; Nikolina" and "Session name; NNakic_HybridExperimental_mm9".

4.3.1.1 Allele-specific analysis pipeline

The approach to allele-specific analysis (ASA) designed for this study aims to remove biases of technical origin, as well as stochastic and parent-of-origin effects, thus enriching for allelic events driven by *cis*-regulatory variation. The ASA pipeline presented here is applicable to both transcriptome and chromatin-derived NGS data sets. It can be separated into two analysis modules (Figure 4.2).

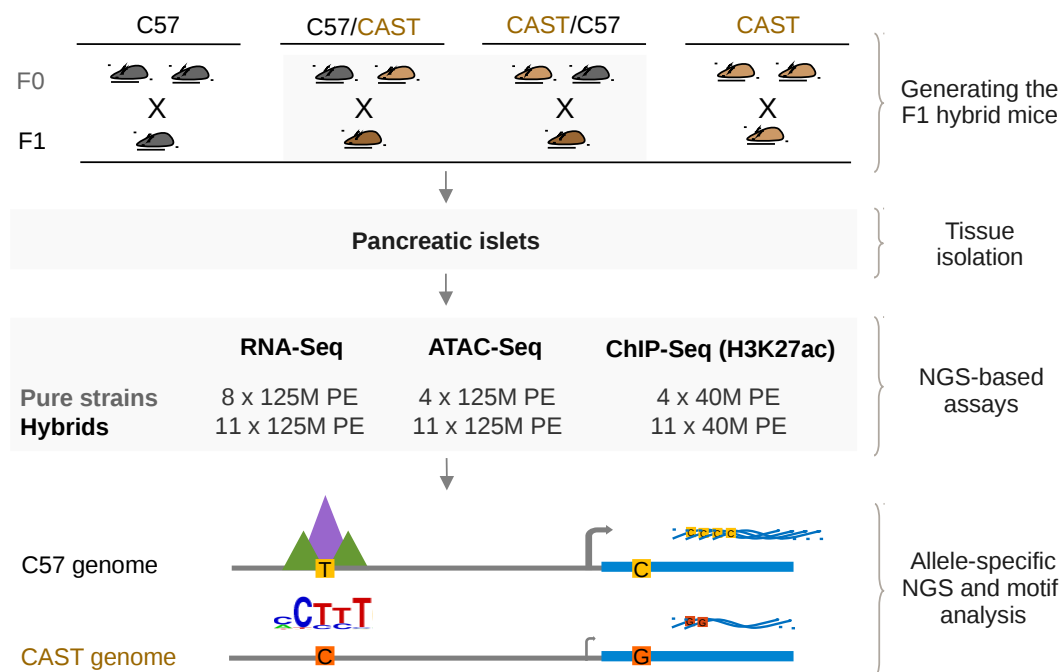


Figure 4.1 Schematic overview of the experimental design. The F1 hybrid offspring of two evolutionary distinct strains of mice, C57BL/6J (C57) and CAST/Ei (CAST), was utilized as an *in vivo* mutagenesis model for assessing allele-specific activity driven by *cis*-regulatory variation. Pancreatic islets were isolated from male hybrid and pure-strain mice at the age of 16 weeks and assayed for gene expression (RNA), enhancer activity (H3K27ac) and chromatin accessibility (ATAC). Pancreatic islets pooled from a minimum of three animals were used for each replicate of every assay. RNA-Seq was performed on 8 pure-strain and 11 hybrid-strain samples and sequenced to approximately 125 million paired-end reads (PE). ATAC-Seq was performed on 4 pure-strain and 11 hybrid-strain samples and sequenced to approximately 125 million PE reads. ChIP-Seq was performed on 4 pure-strain and 11 hybrid-strain samples and sequenced to approximately 40 million single-end (SE) reads. High density of genetic variants (17.6 million SNPs and 2.5 million INDELs) and high-depth sequencing of the assays enabled comprehensive mapping of allele-specific events throughout the pancreatic transcriptome and regulatory elements.

4.3.1.1.1 Removing artefacts of technical origin

The first module establishes the degree of allelic bias across regions of interest in an individual sample. In this module, published approaches for addressing known sources of technical artefacts are adapted for use in the hybrid mouse model and extended with novel solutions.

Reference bias effect is a consequence of the alignment bias that occurs when the reference genome is used as the backbone for alignment of reads derived from a heterozygous organism. SNPs or INDELs present in the reads originating from the non-reference allele hamper alignment of those reads to the reference genome. This results in fewer non-reference reads being aligned, creating an arteficial allelic imbalance in the direction of the reference genome. To avoid the reference bias effect, the *cast* genome was reconstructed by incorporating CAST/Ei SNPs and INDELs into the reference genome. The hybrid reads were then aligned independently to each genome (Rozowsky et al., 2011). Allelic ratios across SNPs derived from the unbiased alignment resulted in a distribution of allelic ratios that centres on 0.5 (Figure 4.3, Pre-filtering SNPs). However, the distribution of ratios varied across samples, particularly for RNA-Seq, which indicated the need for filtering out SNPs impacted by other sources of technical artefacts.

Preferential alignment of reads derived from one pure strain to the genome of the other strain indicates an alignment artefact. This can be caused by unannotated variants not captured in the genome assembly and it is expected to happen more frequently for reads produced from the CAST genome, considering that the reference genome is better annotated. To remove such artefacts, reads derived from pure-strain mice were aligned to both genomes. SNPs with allelic ratios that showed major deviation from expected (1 in the case of aligning reads from C57, 0 when aligning reads from CAST) marked loci with alignment biases and were filtered out from the datasets.

Further steps were taken to ensure that the allelic ratio was not impacted by poor read diversity, while maintaining the maximum number of reads. Preparation of the NGS libraries generally includes PCR amplification of fragments prior to sequencing. Preferential amplification of a fragment, which can be sequence-driven or stochastic, may cause arteficial allelic imbalance in a locus. To avoid artefacts caused by poor read diversity, a minimum of 20 non-clonal reads (reads identical by mapping coordinates and sequence) was required for each SNP. The high depth of sequencing resulted in high read coverage of SNPs, allowing this threshold, higher

than the one used by Kilpinen et al. (2013), to be set while still retaining a representative number of informative SNPs. As this still permitted for distortions of allelic ratios caused by clonality of one of those reads, the allelic ratio was calculated before and after removal of all clonal reads. If the two allelic ratios showed major differences, the SNP was removed from analysis. A SNP that was kept after all the filtering steps was considered informative of its allelic ratio. The filtering process resulted in a more uniform distribution of allelic ratios across samples, which was most evident in the case of RNA-Seq, which had the least uniform starting distribution of ratios (Figure 4.3, Post-filtering SNPs).

Considering the high density of SNPs between the genomes, one region of interest often harboured more than one informative SNP. To obtain a singular allelic ratio for a region and increase the statistical power, counts derived from each SNP in a region were merged (Figure 4.3, Regions). The ratios obtained in this manner and the corresponding p -values were used as input for the second module of the ASA pipeline (Figure 4.2).

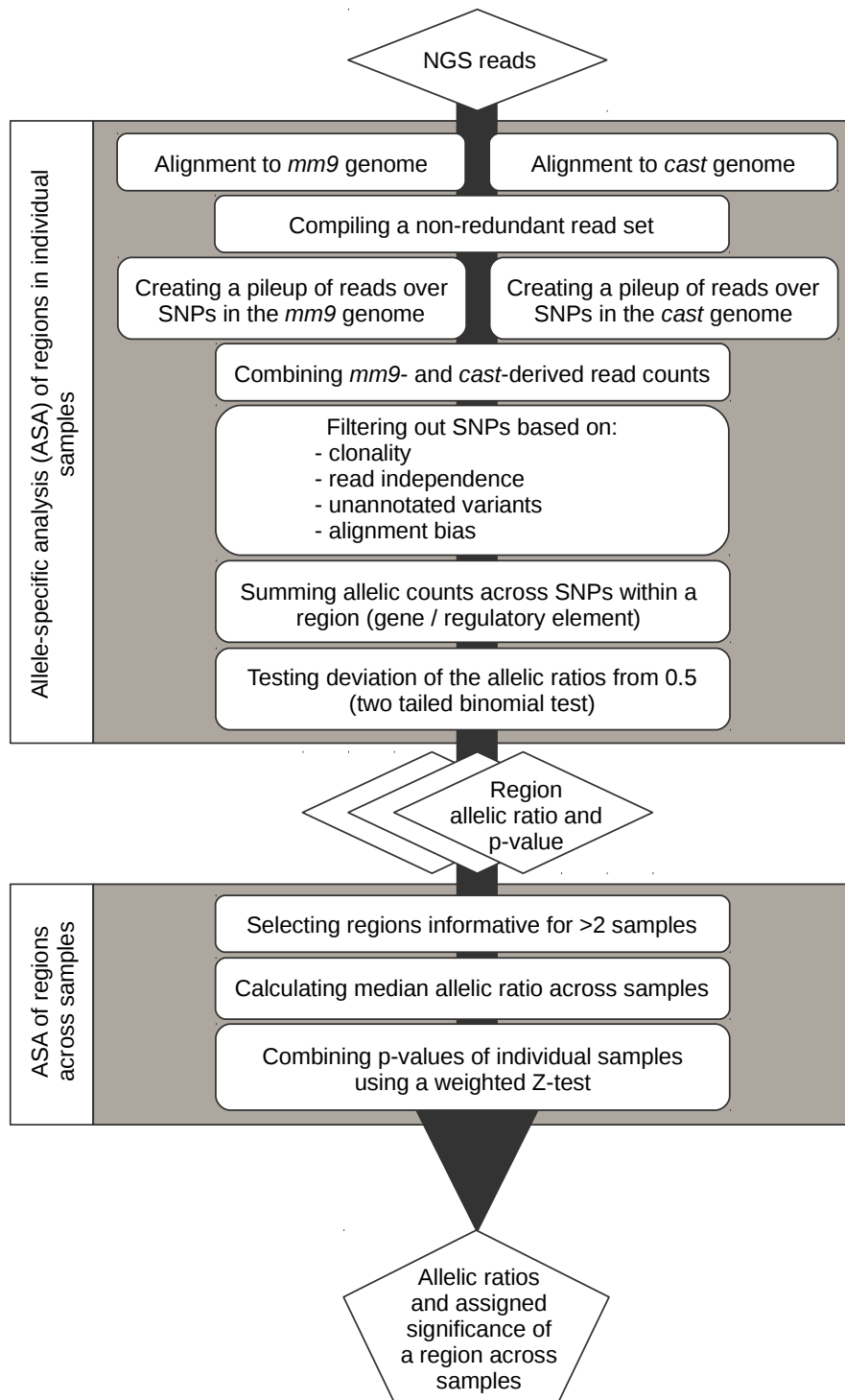


Figure 4.2 Schematic overview of the allele-specific analysis pipeline for determining allelic imbalance consistent across samples. Allele-specific analysis (ASA) pipeline developed for this study aims to detect allele-specific events driven by *cis*-regulatory variation. The first module receives NGS reads derived from a single sample and returns allelic ratios and associated *p*-values of all informative regions. The second module combines information derived from individual samples and assesses allelic ratios and the significance of allelic imbalance consistent across multiple samples.

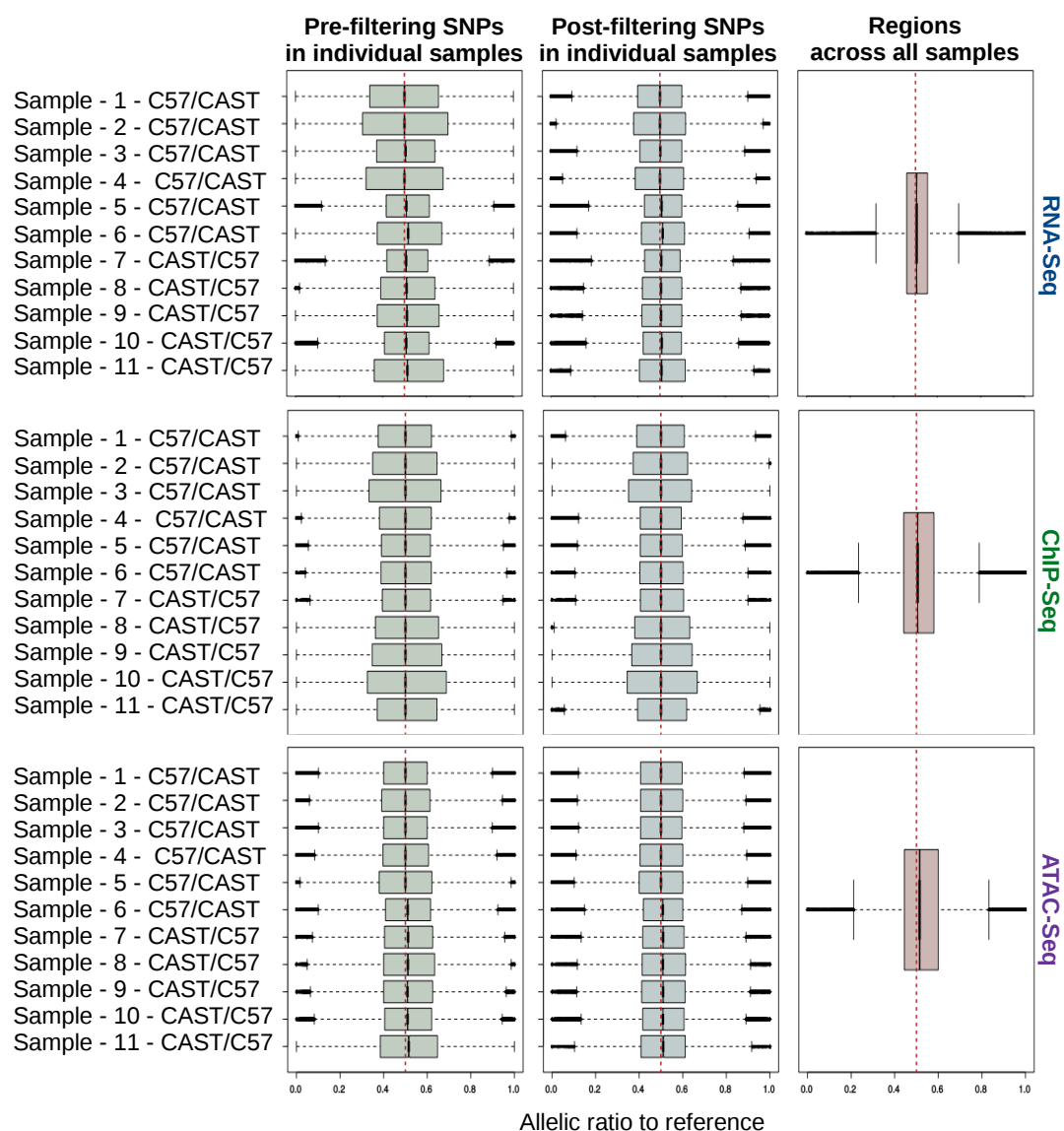


Figure 4.3 Estimates of allelic ratios across different stages of the allele-specific analysis pipeline. Allelic ratios estimated in individual samples across SNPs (Pre-filtering SNPs) were uniformed by the filtering process that accounted for artefacts of technical origin (Post-filtering SNPs). Ratios originating from individual SNP were then combined across regions, resulting in a unique allelic ratio for every gene and regulatory region in the mouse genome (Regions).

4.3.1.1.2 Removing stochastic and parent-of-origin effects

The second module of the ASA pipeline aimed to remove stochastic allelic imbalance across samples and allelic imbalance that is parent-of-origin (Figure 4.2). In order to measure allelic imbalance that is significant and consistent across samples, the p -values derived from individual samples were combined with a weighted Z-test (Whitlock, 2005). This method had been used for meta-analysis in evolutionary biology and GWAS (Willer et al., 2010), but was here adapted for assessment of consistent allelic imbalance, as described in 4.2.1. The resulting allelic ratios and assigned significance reflected consistence across samples while giving more weight to samples with more informative reads. ASA of RNA-Seq found 22.8% of informative genes to be consistently and significantly imbalanced (Figure 4.4a). This percentage is notably lower from the average 65% found significantly imbalanced in individual samples, indicating that the majority of allelic imbalance detected in individual samples comes from stochastic biological events or sample-related artefacts.

Allelic imbalance that is parent-of-origin goes in opposite directions between samples coming from opposite reciprocal crosses. As samples derived from all crosses were assessed together, parent-of-origin imbalance resulted in a final non-significant Z-test derived value. Efficiency of this approach was demonstrated by checking the allelic status of known imprinted genes. Although approximately 80% of informative imprinted genes were significantly imbalanced in individual samples, none of them tested significant in the combined ASA, demonstrating an effective removal of parent-of-origin effects (Figure 4.4b).

Allele-specific analysis of RNA-Seq, H3K27ac ChIP-Seq and ATAC-Seq enabled genome wide mapping of allele-specific events across the genomic loci of the mouse genome. The results of the analysis are illustrated in the example of the *Adh1* locus (Figure 4.5). Allele specific analysis revealed strong allelic bias of *Adh1* expression towards the reference genome (Figure 4.5a), consistent with the differences in expression levels detected between the pure-strains (Figure 4.5b), where the gene was strongly expressed in C57 but not expressed in CAST mice. Furthermore, several enhancers and promoters surrounding *Adh1* display allelic imbalance of H3K27ac and ATAC signal in the same direction as *Adh1* expression, suggesting a functional link between them (Figure 4.5a). Systematic assessment of allelic events in genes and regulatory regions across the genome will provide insight into the principles of *cis* regulation in pancreatic islets.

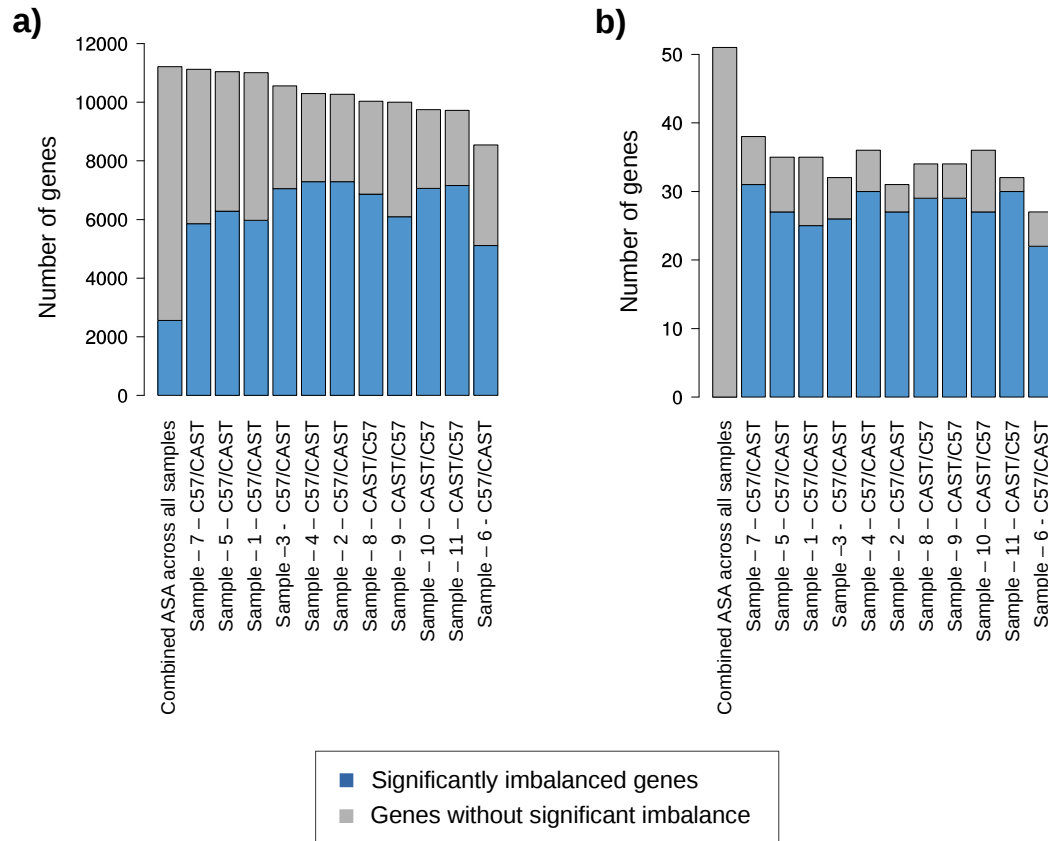


Figure 4.4 Allelic imbalance of transcripts across individual and combined samples. Number of genes found informative and significantly imbalanced in individual samples and combined samples, considering a) all informative genes or b) informative genes known imprinted genes. For individual samples, significance of deviation of allelic ratios from 0.5 was estimated by a two-tailed binomial test. The combined allele specific analysis (ASA) was based on weighted Z-score across minimum three samples and the significance of the Z-score was calculated against a permutation-based background distribution. Both binomial and Z-score-derived *p*-values were adjusted for multiple-hypothesis testing using the Benjamini-Hochberg method and reported significant if less than 0.05.

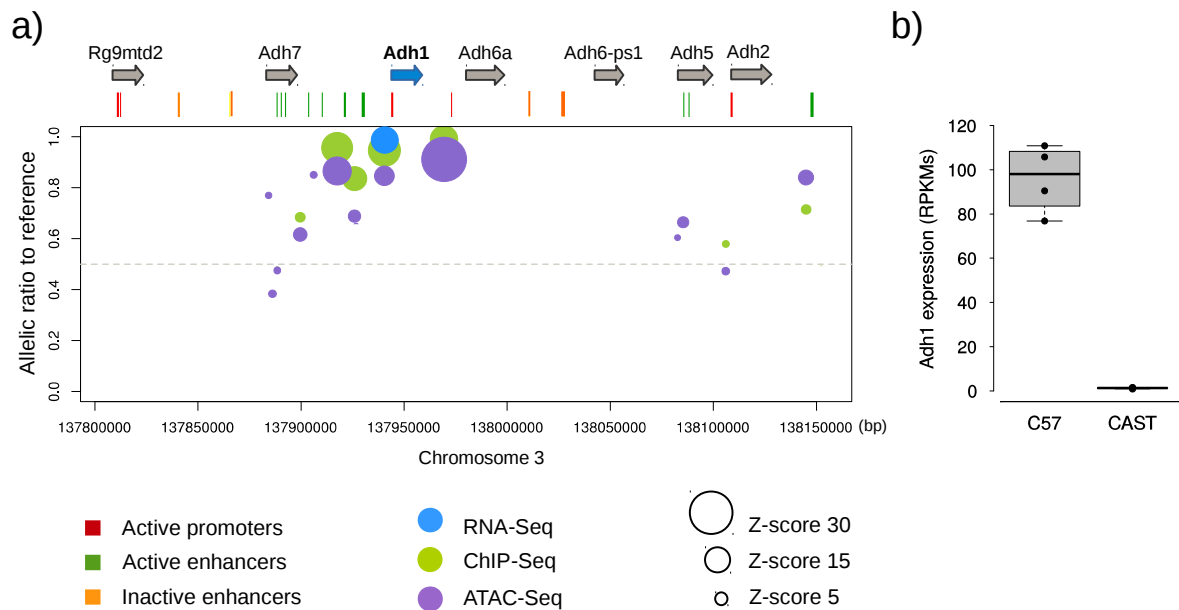


Figure 4.5 Illustration of the allele-specific events mapped across a genomic locus. a) Allele-specific activity in the *Adh1* locus. The *Adh1* gene shows strong allelic bias towards the reference genome. The gene is surrounded by a large number of enhancers and promoters, some of which display allelic imbalance of H3K27ac and ATAC signal in the same direction. Systematic assessment of such events across the genome will provide insight into the principles of *cis*-regulatory regulation in pancreatic islets. Coordinates of RefSeq-annotated genes in the region are displayed in the top track. Coordinates of regulatory elements defined in Chapter 3 are displayed in the second track, colour coded according to the legend. The main plot represents allelic imbalance of RNA, H3K27ac and ATAC-Seq signal, colour coded according to the legend. Only events with a Z-score smaller than 3 are represented and the size of the circle is proportional to the Z-score. b) Expression levels of *Adh1* in the pure-strain mice.

4.3.2 Properties of genes subject to *cis*-regulatory variation in pancreatic islets

To understand what kind of genes tend to be impacted by *cis*-regulatory variation, genes derived from allele-specific analysis of RNA-Seq data were further characterized. The ASA pipeline captured the allelic status of 32.2% of Ensembl annotated genes, out of which 22.8% were significantly imbalanced (Figure 4.6a). A great majority of imbalanced genes were moderately imbalanced, with ratios ranging from 0.1 to 0.9, indicating that *cis*-regulatory variants rarely cause complete inactivation of one of the alleles (Figure 4.6b). Future work (for example, confirmation of allelic ratios in hybrid mice by comparison with gene expression levels in pure

parental strains) may establish if the allelic ratios of extremely imbalanced genes are artefactual or if these genes represent a group of biological interest.

Allele-specific analysis is powered by the number of informative reads, resulting in a bias for highly expressed genes. Furthermore, some genes displayed a higher variation of allelic ratios across individual samples than others (Figure 4.6c). To enable a fair comparison between imbalanced genes and genes not displaying significant imbalance, a control gene set was created. This control set contained genes that are not significantly imbalanced, but match the imbalanced genes by coverage and dispersion of allelic ratios across replicates.

To understand what kind of genes are subject to variation in *cis*, several gene and promoter-associated properties were tested for association with allelic imbalance. Islet and beta cell tissue specificity scores were calculated by comparing the level of expression of genes in islets and FACS purified beta cells to 19 non-pancreatic adult and fetal tissues. Promoter properties included coverage by H3K4me3, a histone mark that is associated with TF binding at the TSS of genes (Cheng et al., 2012), average GC content, a property that differs between tissue-specific and ubiquitously expressed genes (Roider et al., 2009; Schug et al., 2005; Yamashita et al., 2005) and average PhastCons score, reflecting sequence conservation across 30 vertebrate species (Siepel et al., 2005). These properties were used as independent variables in a generalised linear model (GLM) where the imbalance of the gene was a binary response variable, as described in Section 4.2.5.1. RNA read coverage of informative SNPs was introduced as an internal control of gene expression. The results indicate that tissue specificity and promoter conservation are significantly associated to imbalanced genes, independently of each other (Figure 4.7a). While higher tissue specificity was positively associated with imbalance, the association of promoter conservation with imbalanced genes was negative. A direct comparison of those properties between imbalanced genes and the control gene set further confirmed that imbalanced genes do have higher tissue specificity scores but lower conservation of promoters (Figure 4.7b,c). Interestingly, while increased conservation of promoters was accompanied with a lower rate of canonical motif disruption (Figure 4.7e), there was no such trend when it came tissue specificity (Figure 4.7d). Overall, this suggested that tissue specific genes are more impacted by *cis*-regulatory variation. Furthermore, this impact can not be attributed to the frequency of motif disruption occurring in their promoters.

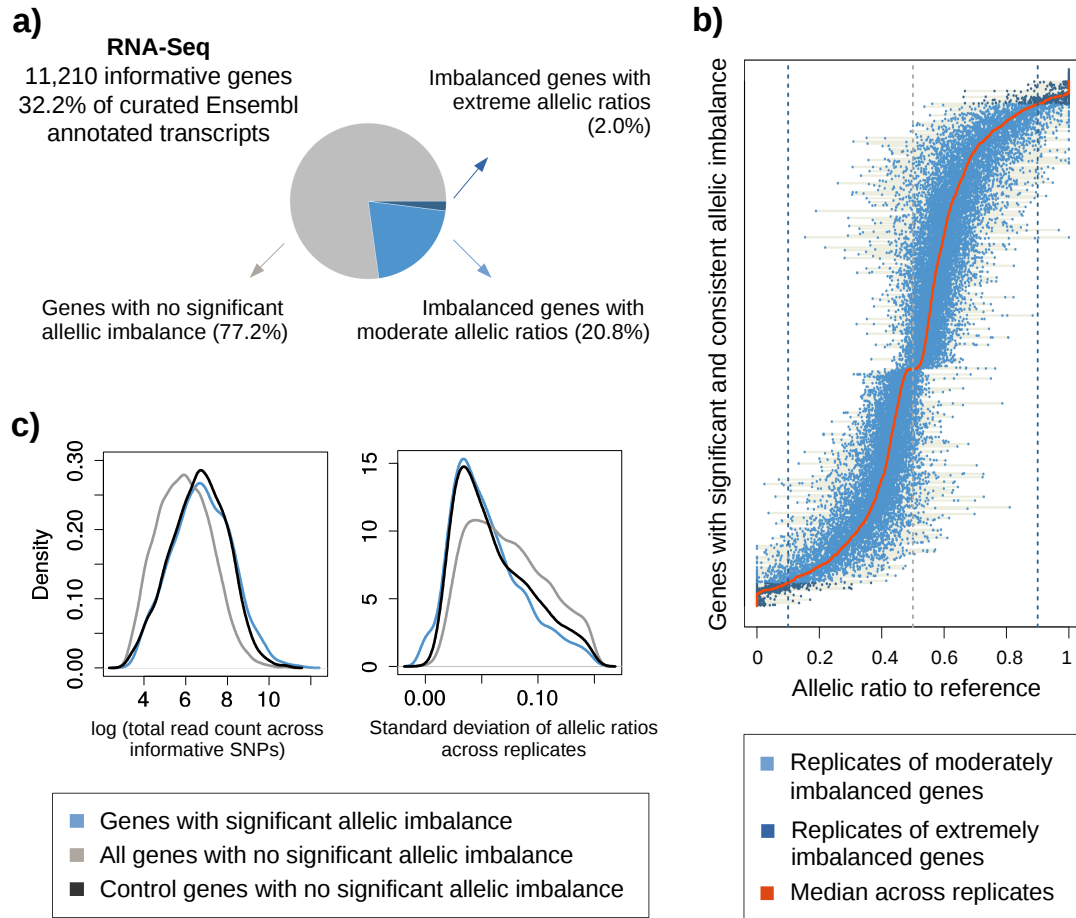


Figure 4.6 Allelic imbalance of gene expression and generation of the control gene set. a) Allele-specific analysis (ASA) of gene expression. ASA-derived p -values were adjusted for multiple-hypothesis testing using the Benjamini-Hochberg method and considered significant if smaller than 0.05. Allelic ratios were considered extreme if smaller than 0.1 or larger than 0.9, or moderate if between 0.1 and 0.9. b) Distribution of individual sample-derived allelic ratios for genes that were found significantly imbalanced across all samples. c) Distribution of coverage across informative SNPs and allelic ratios across samples for the three categories of genes (significantly imbalanced genes, all imbalanced genes and the matched control gene set). The control genes are not significantly imbalanced, but match imbalanced genes by coverage over informative SNPs and standard deviation of allelic ratios across replicates.

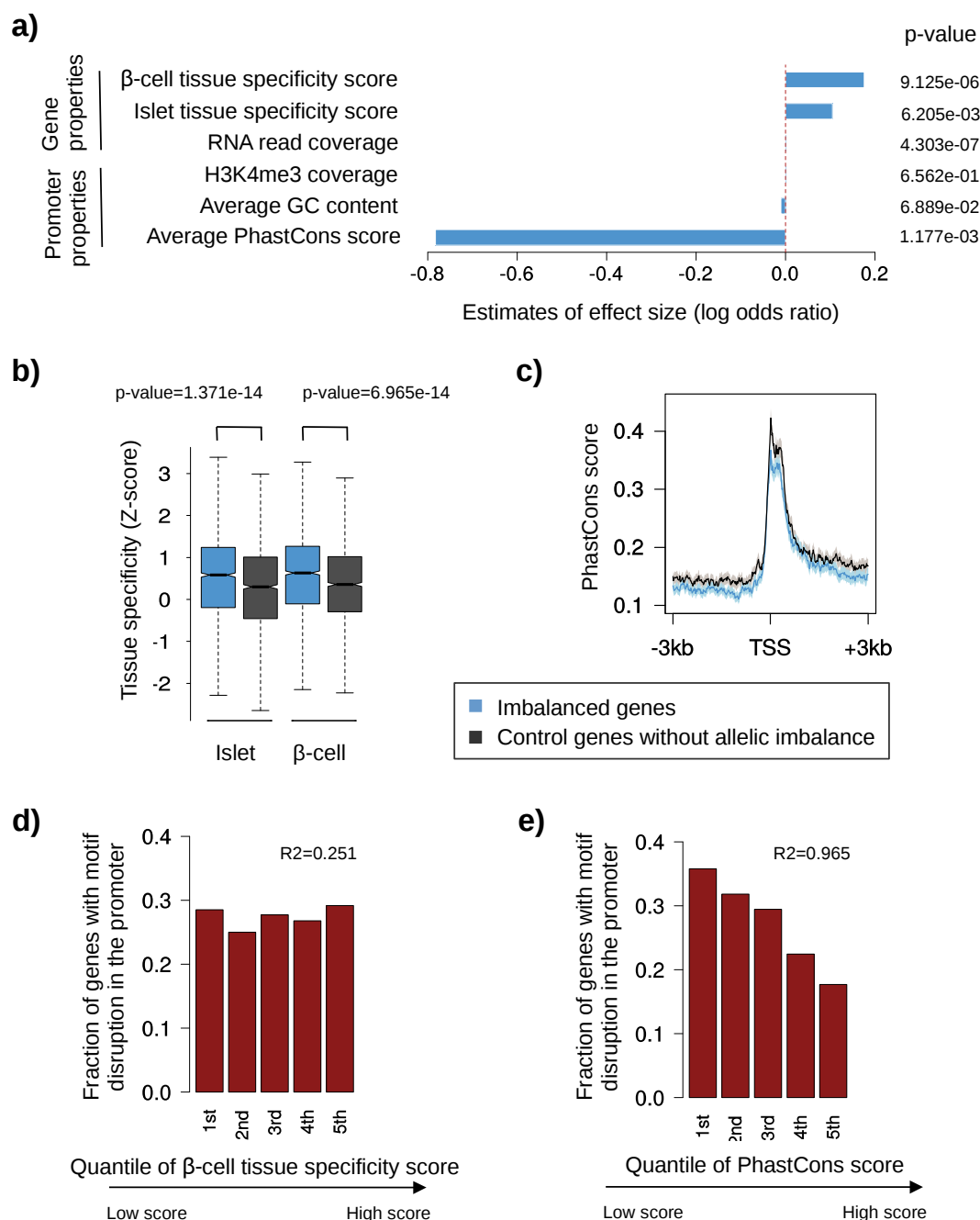


Figure 4.7 Properties of imbalanced genes and their promoters. a) Association of gene and promoter properties with allelic expression was estimated with a generalised linear model where the properties were independent variables and allelic imbalance the binary response variable. b) Comparison of tissue specificity scores between the imbalanced and the control gene set. Significance was calculated using the Wilcoxon rank-sum test. c) Distribution of sequence conservation expressed as PhastCons scores (based on 30 vertebrate genomes) around the TSS of imbalanced and control genes. The full line represents the median value and the shaded areas mark the 95% confidence intervals. The figure legend refers to both b) and c). Fraction of genes with a motif disruption in a promoter, calculated for genes falling into quantile-based groups according to d) gene beta cell tissue specificity or e) sequence conservation of the promoter. The R² values in d) and e) represent the goodness of fit of the linear regression between the fraction of genes with a motif disruption in a promoter and the median PhastCons scores of promoters in each quantile.

To investigate if imbalanced genes are associated with a particular type of regulatory architecture, the density of genes and regulatory elements was inspected within the same TAD, in the region $\pm 250\text{kb}$ from the TSS of a gene. The most noticeable difference was the increased number of enhancers in regions around imbalanced genes (Figure 4.8a). This increase was not driven by the number of genes in the area, as density of genes and active promoters was slightly lower in the same regions (Figure 4.8b,c). These results suggest that imbalanced genes might be associated to clustered enhancers. To confirm this hypothesis, the distribution of ASA-derived p -values was compared between genes associated to clustered enhancers, orphan enhancers or not associated to enhancers. These three categories of genes, however, differ in levels of expression (Figure 4.8d). Therefore, to ensure that a comparison of allelic imbalance between them is not biased by coverage, a subset of genes with matching distribution of reads covering informative SNPs was selected from each category (Figure 4.8d). Distribution of the p -values for coverage-controlled cluster associated genes deviates from the one of orphan enhancers and even more from the one of genes not associated to enhancers (Figure 4.8e). These results indicate that genes associated to clustered enhancers tend to be impacted by *cis*-regulatory variation.

Overall, the results described here indicate that genes whose expression is influenced by *cis*-regulatory variation tend to be tissue-specific and associated to clustered enhancers.

4.3.3 Effects of genetic variation on enhancer chromatin properties

In order to assess if a genetic variant in a regulatory region can drive changes in gene expression, one must first assess if the variant is associated with a change of the regulatory region itself. Therefore, to investigate the changes in activity and chromatin accessibility driven by genetic variation, allelic imbalance of H3K27ac and ATAC signal was estimated across pancreatic regulatory regions defined in Chapter 3. As shown in (Figure 4.9a), 30,539 regulatory regions were informative of their H3K27ac allelic status, capturing 60.4% of active enhancers and promoters (regulatory elements marked by H3K27ac). ATAC was informative for 48,002 regions, which represents 60.8% of all islet regulatory regions (Figure 4.9b). For both assays, most of the imbalanced regions showed moderate allelic ratios (0.1-0.9). Overall, allele-specific analysis of the chromatin properties captured the majority of regulatory regions defined in mouse pancreatic islets.

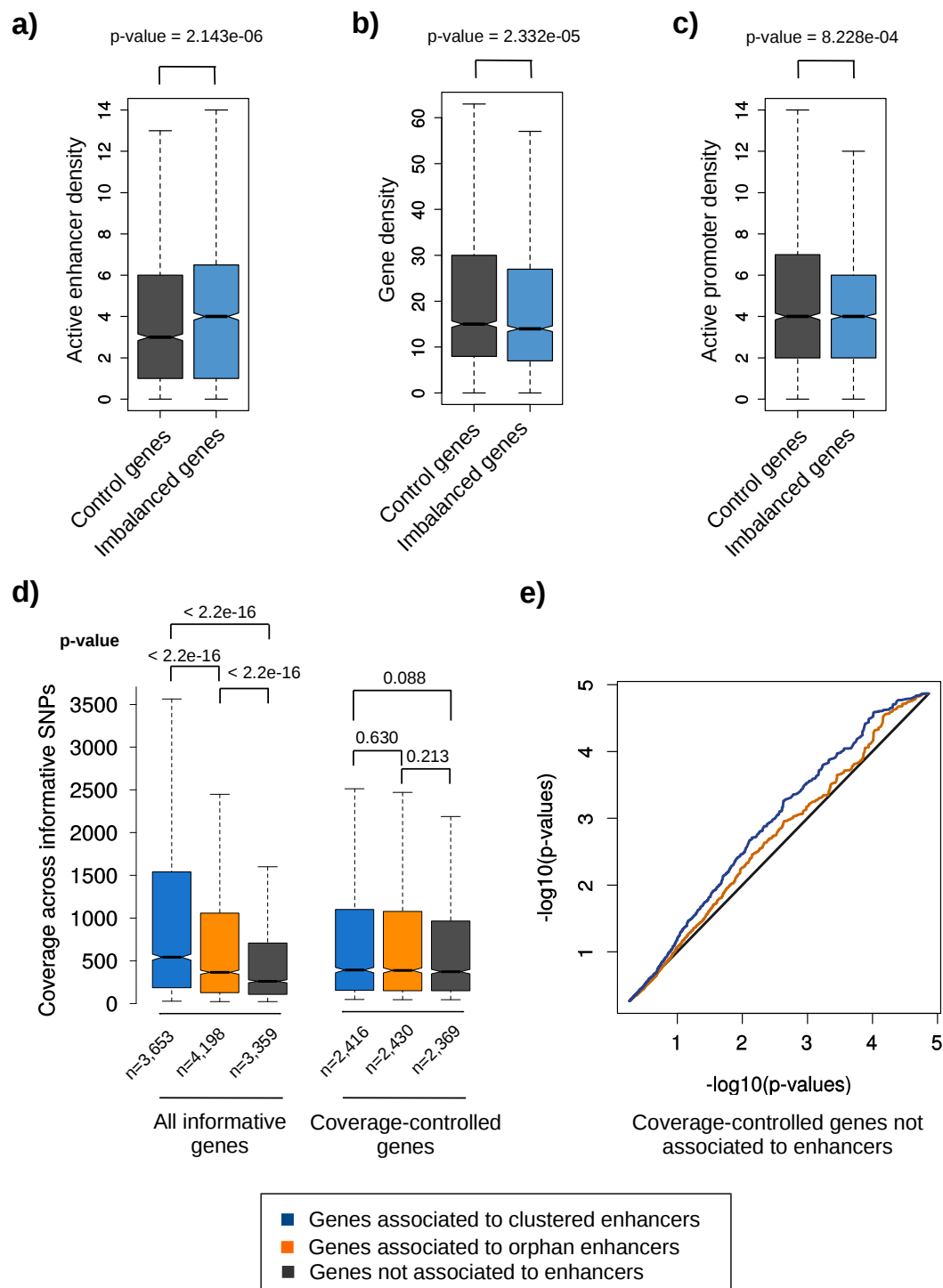


Figure 4.8 Allelic expression and properties of gene-associated regions. Number of a) active enhancers b) TSS of annotated genes and c) active promoters was counted in a space of 250kb upstream or downstream from the gene TSS, or less if interjected with a TAD border. Active promoters and active enhancers were defined in Chapter 3. d) Coverage-controlled genes are a subset of all informative genes, matched for the distribution of the number or reads across informative SNPs. p -values were calculated with the Wilcoxon rank-sum test. e) A two-way qq-plot was created using the distribution of the ASA-derived p -values of the genes not associated to enhancers as a reference (on the x-axis) and then plotting the distributions of all three gene categories on the y-axis.

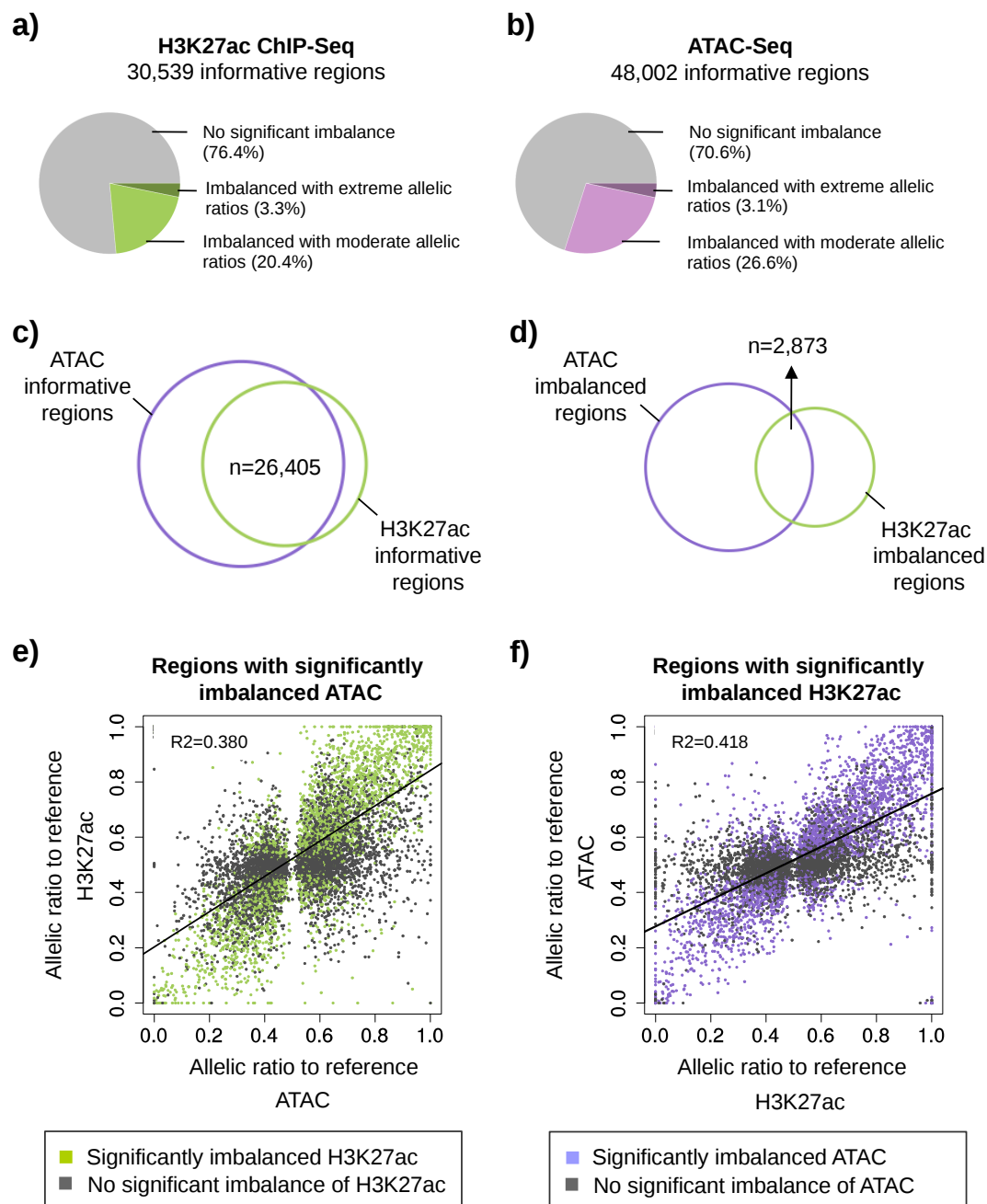


Figure 4.9 Genome-wide allelic imbalance of H3K27ac and ATAC signal. Allele specific analysis was performed on a) H3K27ac ChIP-Seq and b) ATAC-Seq in the same manner as described for RNA-Seq, across regulatory element defined in Chapter 3. A subset of those regions was found c) informative for both assays or d) significantly imbalanced for both assays. e) Allelic ratios of ATAC and H3K27ac across regions with significantly imbalanced ATAC. f) Allelic ratios of H3K27ac and ATAC across regions with significantly imbalanced H3K27ac. The lines in e) and f) represent the linear fit between the plotted values. The R^2 values reflect the goodness of fit. The R^2 value for allelic ratios significantly imbalanced for both H3K27ac and ATAC is 0.642 (green/purple points).

To explore the relationship between enhancer activity and chromatin accessibility, regions that were informative for both H3K27ac and ATAC allelic status were further analyzed. Even though 86.6% of H3K27ac-informative regions were also informative for their ATAC allelic status (Figure 4.9c), the overlap of significantly imbalanced regions was modest (Figure 4.9d). This indicated that not all allelically active H3K27ac regions are underlined with significant changes in open chromatin. Furthermore, allelic ratios of H3K27ac across imbalanced ATAC regions (Figure 4.9e) demonstrated greater dispersion than the distribution of ATAC across imbalanced H3K27ac regions (Figure 4.9f). This leads to the conclusion that changes in chromatin accessibility are not necessarily accompanied with significant changes in enhancer activity. Additionally, changes in H3K27ac are more often, although not always, coupled with coordinated changes in chromatin accessibility.

Together, these results demonstrate the impact of genetic variation on the regulatory landscape of pancreatic islets. Furthermore, they show that changes in enhancer activity and chromatin accessibility are not always coupled.

4.3.4 Relationship between allelic expression and chromatin state of associated regulatory regions

To determine if there is a connection between the allelic status of the regulatory regions and allelic imbalance of gene expression, enrichment of the enhancer allelic signal was determined in a function of distance from the target gene TSS (Figure 4.9e). More specifically, the occurrence of an imbalanced enhancer within a particular distance window from TSS was estimated around imbalanced and control genes. The enrichment of imbalanced enhancers was calculated over a background created by permutations of p -values across all enhancers, thus controlling for number of enhancers associated with each gene. A separate background was created in the same manner for the promoters. The analysis revealed a significant enrichment of H3K27ac-imbalanced enhancers within 20kb from the TSS of imbalanced genes (Figure 4.10a). Furthermore, these enhancers were depleted up to 60kb from the control gene TSSs. In parallel, ATAC-imbalanced enhancers showed a more modest correlation with the allelic status of the gene, with the only enhancer signal being enrichment within 20kb of imbalanced genes (Figure 4.10b). Promoters followed the same trend, with H3K27ac-imbalance of promoters of imbalanced and control genes being significantly over and underrepresented, respectively.

Overall, these results suggest that allelic imbalance of H3K27ac is better associated to the allelic status of the gene than ATAC signal, with an effect observable up to 60kb from the target gene TSS.

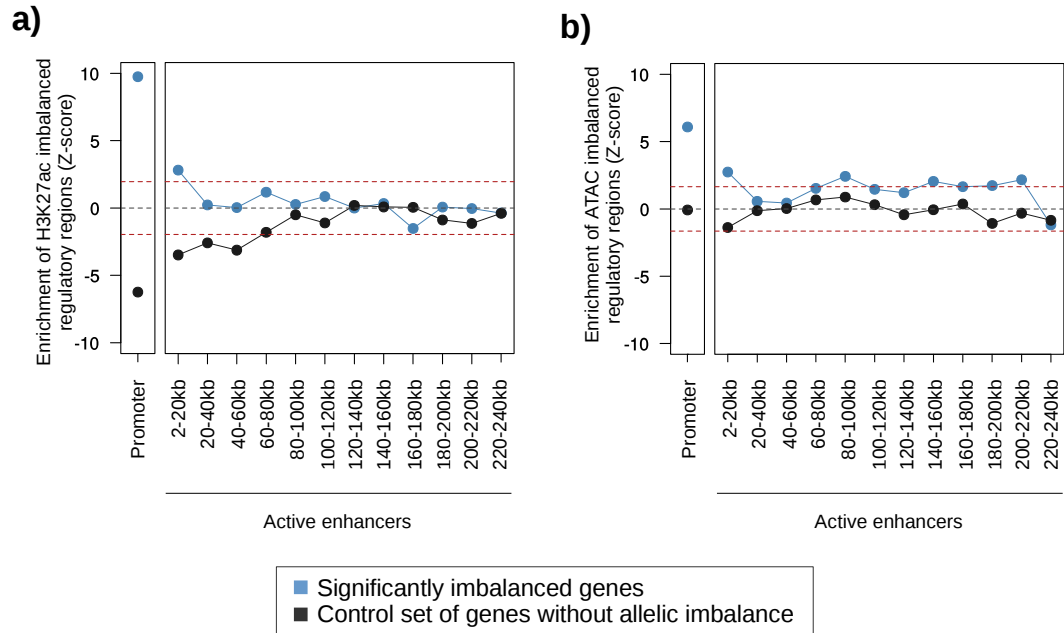


Figure 4.10 Enrichment of imbalanced H3K27ac and ATAC signal in the vicinity of imbalanced and control genes. Enrichment of imbalanced promoters and enhancers in the vicinity of imbalanced and control genes. The imbalance of a) H3K27ac and b) ATAC signal in promoters and enhancers was estimated against a background distribution obtained by 1,000 permutations of allelic activity derived p -values across all enhancers. Fractions of genes associated to an imbalanced enhancer/promoter were calculated for the empirical and permuted data sets. Deviation of the empirical fraction from the permuted distribution was tested using a one-sample Z-test. The Z-test derived p -value was considered significant if less than 0.05. Enrichment was calculated separately for every bin. The significant Z-statistic threshold is represented by the red line.

To further dissect the relationship between allelic expression and enhancer activity, enhancers were assigned to imbalanced and control genes, based on proximity (distance to TSS less than 60kb). The results indicate that, even though there is an enrichment of H3K27ac-imbalanced enhancers in the vicinity of allelically expressed genes, not all changes in gene expression are associated with changes in enhancer activity (Figure 4.11a). In fact, only 29% of allelically expressed genes had an H3K27ac-imbalanced enhancer within 60kb from their TSS. A subset of enhancers that act on genes from a distance larger than 60kb is missed by this approach, which is expected to account for some of the remaining imbalanced genes. Additionally, some im-

balanced enhancers might not have been captured due to low coverage, or were removed in the filtering process. However, this result may also suggest that other mechanisms are responsible for allelic expression of a fraction of these genes, such as changes in promoter activity or allele-specific splicing or polyadenylation (summarised in Figure 1.7 in Section 1.2.2.2.2).

To investigate the relationship between allelic states of genes, enhancers and promoters, genes with an H3K27ac-informative promoter were categorized according to allelic activity of the gene and associated enhancers. The fraction of genes without an imbalanced enhancer in their vicinity was significantly higher for imbalanced genes than control, suggesting that allelic promoter activity does underlie some cases of allelic expression. However, H3K27ac imbalance was more frequent for promoters of genes that also have an associated H3K27ac-imbalanced enhancer. This result suggests that allelic promoter activity is more frequently a result of allelic enhancer activity, rather than an independent event.

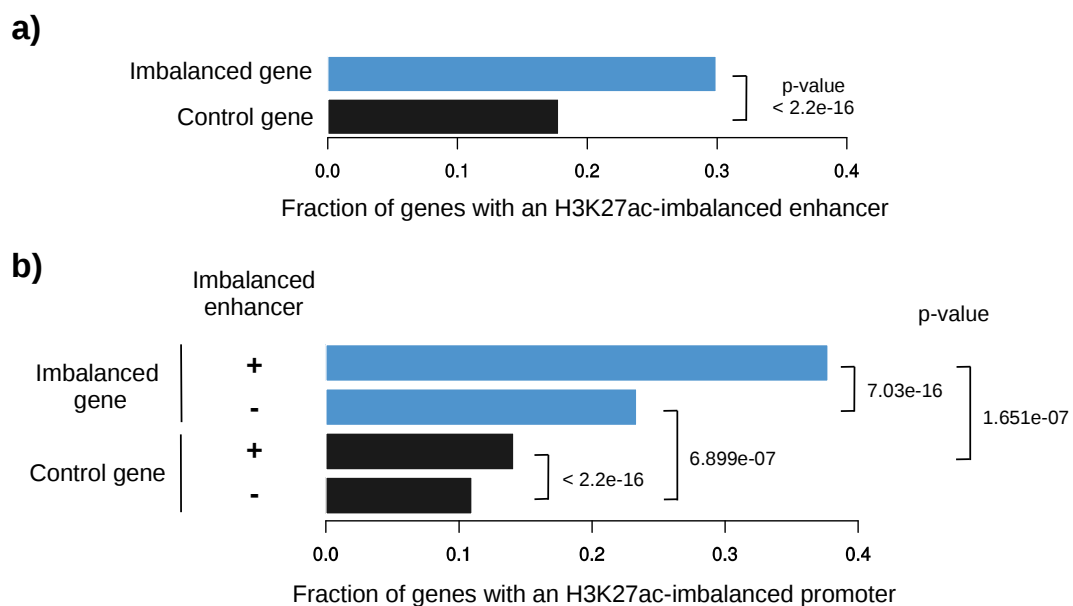


Figure 4.11 Fraction of genes associated with imbalanced enhancers and promoters. a) Fraction of all imbalanced or control genes that have an H3K27ac-imbalanced enhancer within 60kb from their TSS. The 60kb window was selected based on the previous analysis that showed a relationship between allelic imbalance of gene expression and allelic activity of enhancers within that distance. b) Fraction of all imbalanced or control genes, with or without an H3K27ac-imbalanced enhancer within 60kb from their TSS, that have an H3K27ac-imbalanced promoter. Significance of the difference between fractions was calculated with Fisher's exact test.

Interestingly, approximately 17% of control non-imbalanced genes also had an H3K27ac-imbalanced enhancer in their vicinity. Although many of these enhancers might regulate another gene in the region, this result indicates that some allelically active enhancers are not associated with allelic expression.

Overall, these results show enrichment of allelically active enhancers in the vicinity of allelically expressed genes, which is often associated with allelic activity of the gene promoter. Furthermore, the results show that some genes associated to an allelically active enhancers do not show significant allelic expression.

4.3.5 Properties of imbalanced enhancers associated to imbalanced genes

The presence of allelically active enhancers in the vicinity of non-imbalanced genes suggests that not all enhancers are capable of changes in gene expression. Following the hypothesis that not all enhancers are equal, a number of enhancer properties was examined in order to assess what kind of imbalanced enhancers are associated to allelic gene expression. For this purpose, an enhancer-centric approach was taken, where genes were assigned to enhancers if their TSS was within the same TAD and closer than 60kb.

Regarding their location, enhancers can be categorized as intronic (the majority), intergenic or exonic (the least frequent category). Regarding their organisation, enhancers can be found forming enhancer clusters (as described in Chapter 3) or can remain orphan. The number and strength of TF binding can also vary between enhancers. This property was represented by the TF occupancy categories defined in Chapter 3 and based on the combined ChIP-Seq signal of five islet TFs. Enhancer strength and chromatin accessibility were represented by the strength of H3K27ac and ATAC signal. Tissue specificity was estimated in comparison to the H3K27ac signal derived from 23 non-pancreatic tissues. Enhancer sequence properties were also examined, including average GC content and sequence conservation. The described enhancer properties were characterized for every H3K27ac-imbalanced enhancer and used as independent variables associated with allelic status of the gene, which represented the binary predicted variable of a generalised linear model (described in Section 4.2.5.2). To ensure that the results are not biased by variation in expression of the associated gene, coverage of the associated gene was also included in the analysis. Together, these categories provided an extensive overview of the properties of enhancers of interest.

Several properties of imbalanced enhancers showed significant independent association to the allelic status of the gene (Figure 4.12). Exonic enhancers, for example, showed the strongest positive association with allelic imbalance. These enhancers are a rare type of regulatory elements and make only 1.4% of all enhancers defined in pancreatic islet. Intergenic enhancers showed stronger association to allelic expression than intronic, in line with Keane et al. (2011). Clustered enhancers, who are a major enhancer category, showed strong association to imbalanced genes in comparison with orphan enhancers. These results are consistent with the finding that cluster-associated genes tend to be more imbalanced, as presented earlier (Figure 4.8d). In addition, presence of an ATAC signal imbalanced in the same direction as H3K27ac was positively associated with the imbalance of the gene. The same was true for the strength of H3K27ac and ATAC signal (Figure 4.12). Interestingly, higher enhancer GC content also showed significant, though small, association with gene imbalance. Sequence conservation of enhancers showed strong negative association with gene imbalance and was the only property negatively associated with gene imbalance. Overall, these results demonstrate that certain types of imbalanced enhancers have a stronger association with imbalanced genes. Clustering of enhancers has emerged as a property with a strong positive association with coordinated allelic expression of nearby genes.

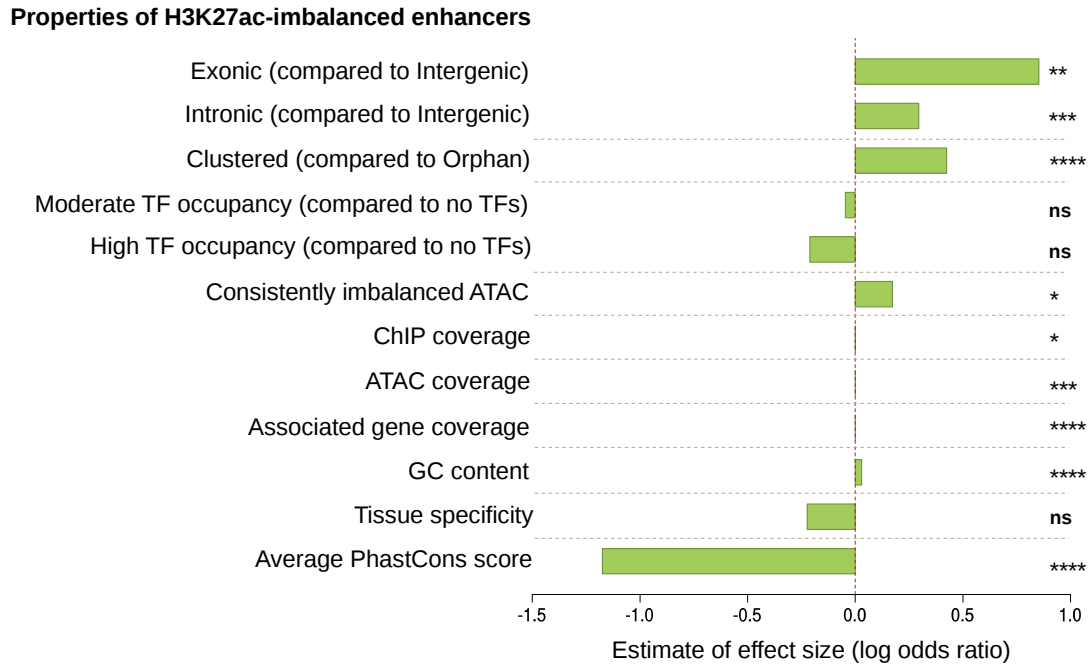


Figure 4.12 Properties of H3K27ac-imbalanced enhancers associated with allelic expression. Properties of enhancers were used as independent variables in a generalised linear model in which the imbalance of the gene in the same direction as enhancer represented the binary predicted variable. The p -values are represented as following; p -value > 0.05 as ns, p -value < 0.05 as *, p -value < 0.01 as **, p -value < 0.001 as *** and p -value < 0.0001 as ****.

To confirm that the differences established between clustered and orphan enhancers are not co-founded by differences in sequence conservation, the enrichment of imbalanced genes was examined in their vicinity, over a range of PhasCons scores (Figure 4.13a). Remarkably, imbalanced clustered enhancers were consistently enriched for occurrence of local genes imbalanced in the same direction, but not for control genes. This enrichment was significant across PhastCons values, even though the fold enrichment was falling as the sequence conservation was rising. In contrast, there was an overwhelming lack of enrichment of imbalanced genes in the vicinity of significantly imbalanced orphan enhancers (Figure 4.13a). Furthermore, apart from the most conserved category, there was no differences in the rate of motif disruption between clustered and orphan enhancers (Figure 4.13b). Together, these results indicate that clustered enhancers have a stronger *cis*-regulatory impact on gene expression. Moreover, this difference between clustered and orphan enhancers is independent of other properties, such as sequence conservation, and is not driven by rates of canonical motif disruption.

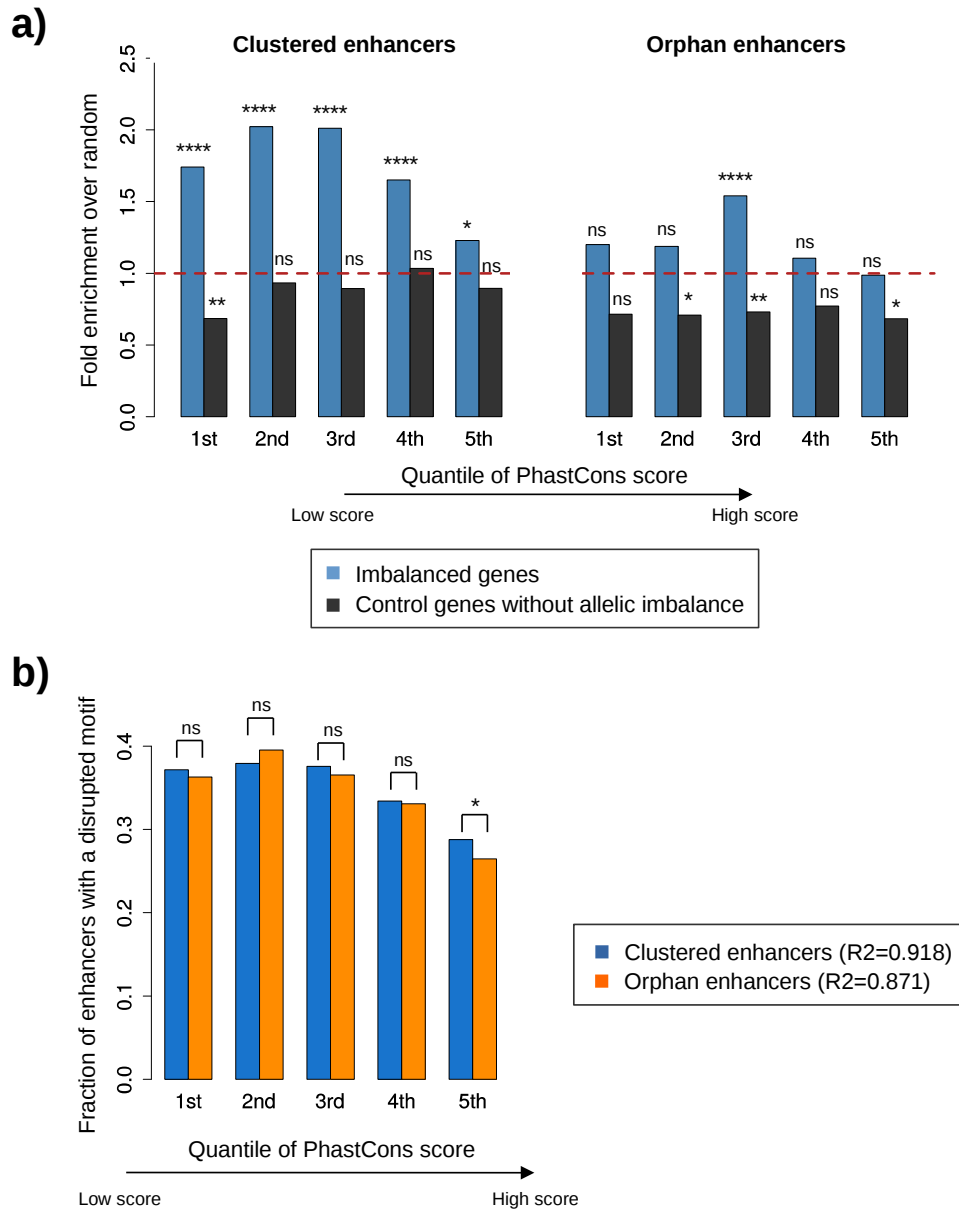


Figure 4.13 Clustering and conservation of H3K27ac-imbalanced enhancers associated to allelic expression. a) Clustered and orphan enhancers were separated in quantile-based categories according to their average PhastCons score (based on 30 vertebrate genomes). Fold enrichment of imbalanced/control genes occurring within the same TAD and less than 60kb from an imbalanced enhancer was calculated separately for each category of enhancers, over a background distribution obtained by permuting allelic expression p -values across genes. Significance was calculated using a one-sample Z-test and considered significant if less than 0.05. b) The fraction of imbalanced enhancers harbouring a disruption of a canonical motif, estimated separately for every category of enhancers presented. The p -value was calculated with a Wilcoxon rank-sum test. The p -values are represented as following; p -value > 0.05 as ns, p -value < 0.05 as *, p -value < 0.01 as **, p -value < 0.001 as *** and p -value < 0.0001 as ****. The R^2 values represent the goodness of fit of the linear regression between the fraction of clustered/orphan enhancers with a motif disruption and the median PhastCons scores of clustered/orphan enhancers in each quantile.

4.3.6 Allelic expression in T2D GWAS loci

In order to explore if the results presented in this study are directly informative of T2D biology, allelic imbalance was assessed in mouse orthologues of human genes with T2D-associated variants in their loci. A list of 84 putative T2D-associated genes reported by Cho et al. (2012), Kooner et al. (2011), Ma et al. (2013), Mahajan et al. (2014), Morris et al. (2012), Voight et al. (2010) and Steinthorsdottir et al. (2014) was assembled. Out of those, 56 genes had an mouse orthologue informative for their allelic status, 19 of which were significantly allelically imbalanced (Table 4.1). The fraction of T2D-associated allelically imbalanced genes was not significantly higher than the average percentage of imbalanced genes (33.9% versus 22.8% respectively, Fisher's exact p -value=0.0551). This result, however, does not necessarily reflect an accurate representation of allelic expression in T2D-associated genes as it is possible that allelic expression of genes differs between species. Thus, future work might include assessing allelic expression of genes in human pancreatic islets.

Table 4.1 Allelically expressed genes associated with T2D in humans. Allelically expressed genes with T2D-associated SNPs reported in the loci of their human orthologues. SNP coverage is the average number of reads aligning over the informative SNPs across replicates.

Gene name	SNP coverage	Allelic ratio	Adjusted p -value
<i>Pdx1</i>	17033	0.43	1.364e-5
<i>Slc30a8</i>	61854	0.56	1.364e-5
<i>Ccnd2</i>	11825	0.61	1.367e-5
<i>Ap3s2</i>	3610	0.42	1.281e-4
<i>Kcnk16</i>	5248	0.55	1.501e-4
<i>Pepd</i>	1493	0.41	1.802e-4
<i>Fitm2</i>	5077	0.43	2.100e-4
<i>Wfs1</i>	2821	0.43	4.107e-4
<i>Pparg</i>	170	0.79	6.064e-4
<i>Tspan8</i>	992	0.59	9.277e-4
<i>Hnf1B</i>	376	0.63	1.814e-3
<i>Psmc6</i>	649	0.41	4.381e-3
<i>Maea</i>	2067	0.44	4.814e-3
<i>PAM</i>	2158	0.61	5.062e-3
<i>Klhdc5</i>	665	0.61	7.725e-3
<i>Tcf19</i>	196	0.35	1.256e-3
<i>Irs1</i>	534	0.41	2.040e-3
<i>Ankrd55</i>	52.5	0.02	2.579e-3
<i>Ssr1</i>	17780	0.47	4.103e-3

4.4 Discussion

4.4.1 Allelic expression in pancreatic islets

In this study, I utilize a hybrid mouse rich in naturally occurring genetic variation as a model for studying the relationship between gene expression, enhancer activity and chromatin accessibility in pancreatic islets. By using mice derived from both reciprocal crosses and assessing allelic states consistent across all samples, I filtered out stochastic and parent-of-origin effects and enriched for *cis*-acting variation. When applied to high-depth RNA-Seq, H3K27ac ChIP-Seq and ATAC-Seq data, this approach yielded 11,210 genes informative for their allelic expression, 30,539 regulatory regions informative for their allelic activity and 48,002 regions informative for allelic chromatin accessibility (Figure 4.6 in Section 4.3.2 and Figure 4.9 in Section 4.3.3). Together, these results provide a high-resolution genome-wide map of regions impacted by genetic variation in mouse pancreatic islet.

Several studies have explored allelic expression in hybrid mice using transcriptome sequencing. The majority of them, however, used allele specific analysis to investigate parent-of-origin effects (DeVeale et al., 2012; Gregg et al., 2010a,b; Wang et al., 2008b). The results derived from these studies have shown very little overlap, even if performed in the same tissue and developmental stage (DeVeale et al., 2012). These inconsistencies highlight the importance of controlling for technical artefacts and stochastic events that can potentially inflate the findings of allelic imbalance.

Studies that have used the hybrid mouse model for exploration of *cis*-regulation have been less numerous, but equally inconsistent in their reports. In this chapter, I report that 22.8% of informative mouse pancreatic islet genes vary in expression as a result of *cis*-regulatory variation (Figure 4.6 in Section 4.3.2). In comparison, a recent study by Crowley et al. (2015), reported 80% of mouse genes having *cis*-regulatory variation. This number, however, was estimated across three different strains of mice and four different tissues, capturing much more variation than the analysis performed in this chapter. Another study, performed across six different tissues, estimated 43% of informative genes to be significantly imbalanced (Keane et al., 2011). Goncalves et al. (2012), used the same hybrid model described in this chapter and had a similar number of informative genes. The authors attributed approximately 31% of differences in gene expression in liver to variation in *cis*, which is more in line with the

number reported here. However, apart from the differences in the statistical methods used, Goncalves et al. used fewer biological replicates than were analyzed in this chapter, which would impact their signal to noise ratio. Overall, it is clear that the number of genes reported to be allelically imbalanced depends heavily on the study design, analysis method and significance thresholds.

Analysis of genes with T2D-associated variants in their loci did not show a significant over-representation of allelic imbalance in their mouse orthologues (Section 4.3.6). This analysis, however, has several limitations. Firstly, genes allelically expressed in mouse pancreatic islets may not necessarily be allelically expressed in human islets. Secondly, the genes reported by GWAS are generally genes that are closest to the GWAS reporter SNP. The reporter SNPs tagging the loci are not necessarily the causal variants. Furthermore, the gene closest to the SNP of interest is not necessarily the gene that is driving the GWAS signal. Future work thus may include assessing allelic expression in human islets and relating the results to all genes in GWAS loci.

4.4.2 Tissue-specificity of genes susceptible to *cis*-regulatory variation

While numerous studies reported allele specific expression, it is still not clear if there is a particular subset of genes that tends to vary in expression due to genetic variation in *cis*. The results presented in this chapter demonstrate that allelically expressed genes tend to be more tissue specific than non-imbalanced genes. The first report of tissue-specific characteristics of allelically expressed genes came from a small microarray-based study in hybrid mice (Campbell et al., 2008). Next, Keane et al. (2011) found allelically imbalanced genes to be related to the function of the tissue where the imbalance is detected, in line with tissue-specific gene expression. Soon after, a study in human lymphoblastoid cell lines found allelically imbalanced genes to be expressed in fewer tissues than the genes not impacted by *cis* variation (Reddy et al., 2012). While these studies lead to the same conclusion, they were based on a relatively low number of informative genes and did not control for differences in coverage or signal variation between the imbalanced and non-imbalanced genes. A more comprehensive analysis presented in this chapter confirms the tissue specific nature of genes that are subject to *cis*-regulatory variation.

Evolutionary genetics provides a possible explanation to why tissue specific genes tend to be

more variable in expression. Several studies reported that tissue-specific genes have a less constrained expression patterns both within species as well as between species (Blekhman et al., 2008; Khaitovich, 2005; Yang et al., 2005; Zheng-Bradley et al., 2010). The proposed hypothesis is that variation of tissue-specific genes results in fewer, potentially harmful, pleiotropic effects. Other studies, however, have reported higher constriction of tissue-specific expression, leaving this hypothesis open to debate (Liao and Zhang, 2006; Yang et al., 2005). Evolution of coding sequences was found to be positively correlated with tissue specificity of genes (Duret and Mouchiroud, 2000; Kryuchkova-Mostacci and Robinson-Rechavi, 2015; Winter et al., 2004). This could have led to tissue specific genes being over-represented in the subset of genes informative of their allelic expression. However, the analysis of tissue specificity was conducted within the informative gene-set, alleviating the potential bias caused by this phenomena.

Interestingly, the tissue specificity of allelically imbalanced genes was shown to be independent from promoter properties such as GC content or sequence conservation (Figure 4.7 in Section 4.3.2). While promoter sequence conservation itself did show a significant association to allelic gene expression, this association could be attributed to a decreased rate of motif disruption occurring in highly conserved promoters. Tissue specificity, by contrast, was not correlated with changes in promoter motif disruption rates, suggesting that allelic expression of tissue specific genes is caused by alternative mechanisms, such as distal regulatory elements or changes in transcript stability or splicing efficiency.

In line with their expression pattern and promoter properties, allelically imbalanced genes were shown to be embedded in genomic regions with increased enhancer density, despite somewhat smaller gene density (Figure 4.8 in Section 4.3.3). Together, these results suggest that it is variation in distal regulatory regions that is driving changes in gene expression. Enhancers are considered to provide the key regulatory input responsible for tissue-specific gene expression (Heintzman et al., 2009; Visel et al., 2009). Additionally, numerous studies have shown enrichment of *cis*-regulatory variation in enhancer regions (Akhtar-Zaidi et al., 2012; del Rosario et al., 2015; Ernst et al., 2011; Gutierrez-Arcelus et al., 2015; Kasowski et al., 2010; The ENCODE Project Consortium et al., 2012; Trynka et al., 2012b). These findings are thus consistent with a model where a disruption of enhancer activity leads to changes in associated tissue-specific genes. Remarkably, clustered enhancers demonstrated an even more extreme distribution of allele-imbalance *p*-values than orphan enhancers. This suggests that some types

of enhancers might be more associated with allelic expression, a question that is addressed later in this chapter.

4.4.3 Allelic chromatin states in pancreatic islets

The next step in understanding the connection between regulatory elements and gene expression was investigating allele specific activity of the regulatory elements themselves. Thus, I have described genome-wide allelic differences of enhancer activity and chromatin accessibility in mouse pancreatic islet (Figure 4.9 in Section 4.3.3). The analysis captured approximately 60% of regulatory regions defined in mouse islets, enabling a genome wide comparative analysis of the enhancer activity and chromatin accessibility.

Approximately 45% of regions informative for their ATAC allelic state were not informative for H3K27ac (Figure 4.9 in Section 4.3.3). This is not surprising, considering that ATAC also marks inactive enhancers which are characterized by absence of H3K27ac (Creyghton et al., 2010a; Zentner et al., 2011). Although inactive enhancers were not in the focus of this study, they represent an interesting category of regulatory elements. A study by Ostuni et al. (2013) demonstrated that inactive enhancers in differentiated tissue may activate upon exposure of the tissue to external stimuli. Pancreatic islets are a tissue with a strong transcriptional response to changes in environment, such as changes in glucose concentration (Schuit et al., 2002; Webb et al., 2000). Failure to open chromatin due to genetic variation could thus underlie an aberrant response to stimuli, leading to a phenotype related to glucose homeostasis.

Regions with significant imbalance of both H3K27ac and ATAC signal generally showed a good correlation of their allelic ratios (Figure 4.9 in Section 4.3.3). This is consistent with the findings of del Rosario et al. (2015), Dixon et al. (2015) and McVicker et al. (2013) that suggested that a single genetic variant may simultaneously affect histone modifications and chromatin accessibility of a region. Interestingly, some regions were imbalanced exclusively for H3K27ac or ATAC, demonstrating that the signal captured by the two assays can be independent. While these cases may be a result of lack of statistical power for testing allelic imbalance of one of the assays, it is possible that they reflect the true nature of regulatory events. Allelic changes of H3K27ac that are not coupled with a coordinated change in ATAC signal indicate that some changes in enhancer activity are not dependent on nucleosome or TF occupancy. Several studies proposed a model in which variants in one region generate allelic

changes of histone modifications across a larger number of regulatory elements (Kilpinen et al., 2013; McVicker et al., 2013; Waszak et al., 2015). It is possible that the suggested model acts through mechanisms that do not involve changes of the chromatin structure, such as activation of proteins already present in the region rather than recruitment of new ones.

4.4.4 Enhancer variation and changes in gene expression

Upon further investigation of the relationship between allelic activity, chromatin accessibility and gene expression, I found allelically expressed genes to be enriched in allele-specific changes of H3K27ac and ATAC signal in local regulatory elements (Figure 4.10 in Section 4.3.4). Accordingly, allelically imbalanced regulatory elements were underrepresented in the vicinity of control non-imbalanced genes. This result is consistent with the recent findings of the Roadmap Epigenomics Program (Dixon et al., 2015; Leung et al., 2015), and suggests that allelic expression might be driven by allele-specific changes in enhancer activity.

The enrichment of both allelically imbalanced H3K27ac and ATAC signal was stronger at promoters rather than enhancers (Figure 4.10 in Section 4.3.4). This could indicate that the allelic imbalance of promoters themselves is often driving allelic expression. However, imbalanced promoters are overrepresented in genes that have an imbalanced enhancer in vicinity, suggesting an alternative mechanism for most of these cases (Figure 4.11 in Section 4.3.4). Considering the currently accepted model of gene regulation, where the regulatory input of distal regulatory elements converges into the core promoter, allelic activity of the enhancer may drive the changes in the promoter activity and, consequently, gene expression levels. Additionally, some instances of allelic expression are expected to result from events other than disruption of the promoter activity, such as allele-specific changes in splicing or transcript stability (Li et al., 2012; Nembaware et al., 2008, 2004; Wang et al., 2008a).

Although there is an enrichment of allelically active enhancers in the vicinity of allelically expressed genes, H3K27ac-imbalanced enhancers can also be found in the vicinity of non-imbalanced genes (Figure 4.11 in Section 4.3.4). This finding might be partially attributed to mis-assignment of enhancers to genes, which is to be expected when assigning enhancers to genes based on proximity only. Alternatively, not all enhancers might have the capacity to drive changes in expression. Understanding the mechanisms behind functional *cis*-variation is one of the major goals of functional genomics today and was further explored in this chapter.

4.4.5 Properties of allelic enhancers associated with allelic expression

In order to understand what characterizes enhancers that drive allelic expression, I took an enhancer-centric approach and compared properties of imbalanced enhancers associated to imbalanced genes versus control genes (Section 4.3.5). This analysis revealed several properties distinguishing active enhancers that potentially drive allelic expression from the enhancers that do not.

Enhancer sequence conservation was found to be in strong negative association with allelic expression (Figure 4.12 in Section 4.3.5). According to the neutral theory of molecular evolution, majority of sequence changes between species arise through mutations with poor functional impact (Ohta, 1992). Mutations in functional elements, by contrast, impact fitness and are eliminated by purifying selection. Sequence conservation, thus, implies function and is often used for identification of biologically relevant non-coding regions (Berman et al., 2004; Goode et al., 2003; Pennacchio et al., 2007; Visel et al., 2008; Woolfe et al., 2005). Results presented in this chapter, however, indicate that changes of activity of highly conserved enhancers often do not lead to changes in gene expression. Although not intuitive, the idea that some conserved enhancers do not dictate gene expression has been presented before, by Ahituv et al. (2007). In their study, the authors deleted four ultraconserved enhancer elements, active in transgenic assays and located in the vicinity of developmental genes, yet found no detectable changes in the mouse phenotype. There are several possible explanations for these results. Expression of genes associated to these particular enhancers might be too robust to be impacted by a change in activity of one enhancer or small number of enhancers. For example, Jeong et al. (2006) reported that expression of sonic hedgehog (*Shh*) in some tissues is disrupted only after a deletion of two enhancers, both of which are conserved across most vertebrates. While redundancy implies weaker evolutionary constraint, such regulators might appear redundant in a basal state, yet lose the ability to complement each other in different environmental conditions (Frankel et al., 2010; Perry et al., 2010). Finally, it is possible that this result is driven by mis-assignment of target genes. In this chapter, genes were assigned to enhancers if located within the same TAD and less than 60kb away. As many enhancers act across distances much larger than 60kb, it is possible that the real target genes are not accounted for. However, to my knowledge, there is no evidence to support the premise that highly conserved enhancers tend to be located further away from their target genes than less conserved enhancers. Nevertheless,

implementing more reliable approaches of enhancer-gene coupling, such as 3C-based methods, would help to confirm these findings.

The enhancer-centric approach once more exposed enhancer clustering as a property strongly associated with allelic imbalance. Essentially, this result suggests that, compared to an orphan enhancer, a change in activity of an clustered enhancer is more likely to have a functional impact on gene expression. Clustered enhancers were initially defined by Pasquali et al. (2014) and are conceptually similar to stretched enhancers (Parker et al., 2013) and super-enhancers (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). The same studies found these enhancers to be associated to expression of tissue-specific genes and enriched for disease-associated genetic variations. In fact, Whyte et al. (2013) suggested that perturbations of TF binding in super-enhancers might have a stronger impact on the expression of the associated genes than binding changes in typical enhancers. While the results presented in this chapter support a central role of clustered enhancers in cell identity and disease, they do not explain what differentiates them from orphan enhancers. It is possible that these two types of regulatory elements have a fundamentally different role in gene regulation. While clustered enhancers tend to be associated with tissue specific genes, orphans might be associated to genes that are under a stronger regulatory directive from promoters and receive modest input from associated enhancers. These features are consistent with the characteristics of ubiquitously expressed genes which are associated to fewer enhancers and primarily controlled by highly conserved promoters (Andersson et al., 2014; Forrest et al., 2014). Furthermore, the stronger functional impact of clustered enhancers might come from their capacity to propagate their regulatory signal from one enhancer to many, a novel mechanism recently proposed by Waszak et al. (2015) and Grubert et al. (2015). To understand the functional differences between these two types of regulatory elements, a more detailed analysis of their properties is needed. Future work might include an analysis of their motif composition, capacity to produce eRNA, recruitment of Mediator and frequency of physical contact with target genes.

Chapter 5

Final Remarks

The overall aim of the work presented here was to dissect the genetic and epigenetic complexities of gene expression in pancreatic islets. This was carried out in the mouse model by exhaustive characterization of the islet *cis*-regulatory landscape and relaying changes in *cis*-regulatory activity to changes in gene expression.

In Chapter 3, I created genome-wide maps of chromatin accessibility, H3K4me3, H3K4me1 and H3K27ac as well as binding of five transcription factors, two of which were not profiled before in mouse islets. I used the described datasets to defined three general classes of *cis*-regulatory elements - active promoters, active enhancers and inactive enhancers. I further focused on active enhancers and compiled a list of 25 non-redundant transcription factor binding motifs enriched in these regulatory regions, some of which were previously uncharacterized. Moreover, I defined enhancers that are highly, moderately or not occupied by the five assayed TFs. In addition, I sub-classified active enhancers to clustered and orphan, according to their organisation in the linear chromosome. Furthermore, I utilized the clustered enhancers to identify species-common and species-specific subsets of active enhancers. I found that these subsets are associated with distinct categories of genes and reflect on the level of gene expression between mouse and human.

The datasets presented in Chapter 3 provide a valuable resource for future studies of the mouse islet epigenome. Furthermore, my characterization of regulatory elements provides the basis necessary for studies of *cis*-regulatory variation in mouse islets, such as the one conducted in Chapter 4. This compendium of information may be expanded in the future. For example, addition of repressive histone modifications, such as H3K9me3 or the Polycomb protein-associated repressive H3K27me3 mark, would enable identifying poised enhancers with bivalent histone marks (Rada-Iglesias et al., 2010; Zentner et al., 2011). Furthermore, although H3K27ac is considered to be a good predictor of enhancer activity, assessing Mediator occupancy or production of eRNA would give a more accurate readout of the enhancer activity levels (Wang et al., 2011a; Whyte et al., 2013; Zhu et al., 2013).

While the analysis of gene expression performed in this chapter focused on transcription of annotated genes, a portion of transcripts, largely non-coding, remains to be described (Kashi et al., 2016). A systematic *de novo* annotation of lncRNAs in human islets exposed a large number of transcripts, some of which exhibited dynamic expression, dependent on developmental stage or glucose stimulation (Morán et al., 2012). So far, two studies have described novel non-coding transcripts in mouse islets and beta cells (Benner et al., 2014; Ku et al., 2012). Neither of them, however, used an independent estimation of promoter activity. The catalogue of active promoters I described in Chapter 3 presents an opportunity for an integrative analysis of the mouse islet promoters and transcriptome, which would result in a more comprehensive and accurate annotation of novel lncRNA transcripts.

Identification of a species-specific subset of the mouse active enhancer clusters provides a novel perspective on the differences in mouse and human islet biology. Mouse-specific activity of clustered enhancers was associated to genes that are expressed more in mouse than in human and tend to be involved in cell proliferation signalling. As discussed in 1.4.2.1, beta cell proliferative capacity is one of the major differences between mouse and human islets. The results described in Chapter 3 suggest that some of these differences are driven by changes in *cis*-regulatory landscape. These findings provide a framework for future work necessary to understand if the changes in enhancer activity are driven by divergence of sequence or active silencing of the corresponding sequences in human. This will require an in-depth comparison of mouse-human sequences and the extent of motif disruption introduced by sequence divergence. In addition, this calls for a search for repressor-targeted motifs as well as analysis of repressive marks in the corresponding loci of human pancreatic islets. The inability of human beta cells to replicate the proliferative response of murine beta cells to mitogens is one of the major obstacles in the search for T2D therapies (Chandrasekera and Pippin, 2013; Chatzigeorgiou et al., 2009; Vetere et al., 2014). Therefore, these results are of great interest for the diabetes research community.

I employed the knowledge acquired in Chapter 3 to investigate the relationship between *cis*-regulatory variation and gene expression in Chapter 4. First, I developed an approach that focuses on allele-specific imbalance resulting from *cis*-regulatory variation and estimated its significance across replicates while accounting for consistency of the imbalance. Next, I employed the allele-specific analysis pipeline to map allelic imbalance of transcription, enhancer activity and chromatin accessibility. This enabled me to investigate the relationship between allele

specific expression and chromatin states. I determined that genes subjected to *cis*-regulatory variation tend to be tissue-specific and associated to clustered enhancers. Moreover, the results described in Chapter 4 suggest that clustered enhancers have a stronger functional impact on gene expression. This is consistent with the findings of Pasquali et al. (2014) and Parker et al. (2013) who reported an enrichment of T2D-associated variants in clustered enhancers. In addition, I find that the differences of clustered and orphan enhancers are independent of their genomic location, sequence conservation or rate of motif disruption.

I assigned enhancers to putative target genes based on proximity, which is a frequently used approach of target gene assignment, albeit an imprecise one (McLean et al., 2010). Therefore, this analysis does not provide evidence of a direct functional impact that individual enhancers have on a gene. Future studies would greatly benefit from information on chromatin interactions between the functional elements of the mouse islet regulome, provided by a 3C-based method. Chromatin maps would enable including more distal regulatory elements in the analysis, while focusing on enhancer-target gene couples that have a physical connection, thus improving the signal to noise ratio. The same maps would serve to better understand the nature of the disparity between clustered and orphan enhancers by providing information about the frequency of contacts with their target genes and each other. In parallel, a comparative analysis of the sequences underlying clustered and orphan enhancers would clarify if these characteristics are driven by differences in motifs and/or motif combinations. In addition, in order to directly confirm functional differences of these two classes of active enhancers, future studies might employ CRISPR technology for disrupting the activity of a select group of clustered and orphan enhancers, matched for other properties such as sequence conservation, TF occupancy or genomic location.

Gene expression and chromatin states described here represent a snapshot of the pancreatic islet transcriptome and regulome in mice of a particular age and assessed under certain conditions. Gene regulation, however, is a dynamic process that responds to queues from the environment. It is quite possible that the effect of some *cis*-regulatory variants becomes detectable only under particular conditions. These questions are essential for the field of complex diseases and particularly difficult to answer in humans. The mouse model can be subjected to controlled changes in the environment at any age and monitored throughout the mouse lifespan. Such studies would greatly benefit from the characterization of *cis*-regulatory elements provided here, and are likely to contribute new insights into the mechanisms behind complex diseases. This once

more highlights the importance of the mouse model in diabetes research and the utility of the results presented in this thesis.

Allelic activity mapped in mouse pancreatic islets may not be directly translated to allelic activity in human pancreatic islets (discussed in Section 4.4.1). The basic mechanisms behind *cis*-regulatory variation, however, are likely to be conserved across species. The hybrid mouse provided an opportunity to dissect these basic principles in a healthy viable model with abundance of naturally occurring genetic variation. The results presented in Chapter 4 of this study suggest that genetic variation harboured by clustered enhancers and enhancers that are not extremely conserved often have functional consequences for gene expression. These findings suggest a new set of criteria to take into consideration when searching for the causal SNPs within the GWAS loci. It would, therefore, be of great interest to investigate which genes in T2D-associated loci are allelically expressed and which T2D-associated SNPs fall into clustered enhancers and/or enhancers that are not extremely conserved.

Although extensive analysis of motif disruption was outside of the scope of this thesis, the hybrid mouse model has a great potential for exploring the properties of sequence motifs that, when disrupted, have a strong functional impact on gene expression. Considering the overall conservation of transcriptional networks between the mouse and human islets, as demonstrated in Chapter 3, such work would likely be translatable to human and would thus greatly contribute to the efforts of decoding the mechanisms behind T2D associated variants.

Bibliography

- Adams, C. C. and Workman, J. L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Molecular and Cellular Biology*, 15(3):1405–1421.
- Ahituv, N., Zhu, Y., Visel, A., et al. (2007). Deletion of ultraconserved elements yields viable mice. *PLoS Biology*, 5(9):1906–1911.
- Ahlgren, U., Pfaff, S. L., Jessell, T. M., Edlund, T., and Edlund, H. (1997). Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells. *Nature*, 385(6613):257–260.
- Akalın, A., Fredman, D., Arner, E., et al. (2009). Transcriptional features of genomic regulatory blocks. *Genome Biology*, 10(4):R38.
- Akhtar-Zaidi, B., Cowper-Sallari, R., Corradin, O., et al. (2012). Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. *Science*, 336(6082):736–739.
- Allen, B. L. and Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nature Reviews. Molecular Cell Biology*, 16(3):155–166.
- Allen, J. M., Forbush, K. A., and Perlmutter, R. M. (1992). Functional dissection of the Ick proximal promoter. *Molecular and Cellular Biology*, 12(6):2758–2768.
- Allfrey, V., Faulkner, R., and Mirsky, A. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Biochemistry*, 315(1938):786–794.
- Almer, A., Rudolph, H., Hinnen, A., and Hörz, W. (1986). Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *The EMBO Journal*, 5(10):2689–2696.
- Alonso, R., Sharma, B., and C., L. (2014). Lipotoxicity in the Pancreatic Beta Cell: Not Just Survival and Function, but Proliferation as Well? *Current Diabetes Reports*, 14(6):1199–1216.
- Altshuler, D., Brooks, L. D., Chakravarti, A., et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.
- Andergassen, D., Dotter, C. P., Kulinski, T. M., et al. (2015). Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Research*, 43(21):e146.
- Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.
- Andersson, R., Gebhard, C., Miguel-Escalada, I., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.
- Antonarakis, S. E. (2010). Human Genome Sequence and Variation. *Human Genomics*, 5(1):73–74.
- Apelqvist, A., Li, H., Sommer, L., et al. (1999). Notch signalling controls pancreatic cell differentiation. *Nature*, 400(6747):877–881.

- Arner, E., Daub, C. O., Vitting-Seerup, K., et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014.
- Arnold, C. D., Gerlach, D., Spies, D., et al. (2014). Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genetics*, 46(7):685–692.
- Arnold, C. D., Gerlach, D., Stelzer, C., et al. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077.
- Aronoff, S. L., Berkowitz, K., Shreiner, B., and Want, L. (2004). Glucose Metabolism and Regulation: Beyond Insulin and Glucagon. *Diabetes Spectrum*, 17(3):183–190.
- Ashcroft, F. M. and Rorsman, P. (2012). Diabetes mellitus and the beta cell: The last ten years. *Cell*, 148(6):1160–1171.
- Ashery-Padan, R., Zhou, X., Marquardt, T., et al. (2004). Conditional inactivation of Pax6 in the pancreas causes early onset of diabetes. *Developmental Biology*, 269(2):479–488.
- Atkinson, M. and Leiter, E. H. (1999). The NOD mouse model of type 1 diabetes: As good as it gets? *Nature Medicine*, 5(6):601–604.
- Auton, A., Abecasis, G. R., Altshuler, D. M., et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.
- Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medicine*, 79(2):137–158.
- Babak, T., Deveale, B., Armour, C., et al. (2008). Global survey of genomic imprinting by transcriptome sequencing. *Current Biology*, 18(22):1735–1741.
- Bailey, T., Krajewski, P., Ladunga, I., et al. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Computational Biology*, 9(11):e1003326.
- Banerji, J., Olson, L., and Schaffner, W. (1983). A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes. *Cell*, 33(3):729–740.
- Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308.
- Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395.
- Bardeesy, N. and DePinho, R. A. (2002). Pancreatic cancer biology and genetics. *Nature Reviews. Cancer*, 2(12):897–909.
- Barolo, S. (2012). Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 34(2):135–141.
- Barski, A., Cuddapah, S., Cui, K., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.

- Batterham, R. L., Le Roux, C. W., Cohen, M. A., et al. (2003). Pancreatic Polypeptide Reduces Appetite and Food Intake in Humans. *The Journal of Clinical Endocrinology & Metabolism*, 88(8):3989–3992.
- Bayer, T. A., Schäfer, S., Simons, A., et al. (2003). BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):14187–14192.
- Bejerano, G., Pheasant, M., Makunin, I., et al. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.
- Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell*, 98(3):387–396.
- Bell, G. I., Pictet, R. L., Rutter, W. J., et al. (1980). Sequence of the human insulin gene. *Nature*, 284(5751):26–32.
- Benitez, C. M., Goodyer, W. R., and Kim, S. K. (2012). Deconstructing pancreas developmental biology. *Cold Spring Harbor Perspectives in Biology*, 4(6):1–17.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300.
- Benner, C., van der Meulen, T., Cacères, E., et al. (2014). The transcriptional landscape of mouse beta cells compared to human beta cells reveals notable species differences in long non-coding RNA and protein-coding gene expression. *BMC Genomics*, 15(1):620.
- Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics*, 15(3):163–175.
- Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature*, 447(7143):407–412.
- Bergsland, M., Ramsköld, D., Zaouter, C., et al. (2011). Sequentially acting Sox transcription factors in neural lineage development. *Genes & Development*, 25(23):2453–2464.
- Berman, B. P., Pfeiffer, B. D., Laverty, T. R., et al. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biology*, 5(9):R61.
- Bernardo, A. S., Hay, C. W., and Docherty, K. (2008). Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic β cell. *Molecular and Cellular Endocrinology*, 294(1-2):1–9.
- Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The Mammalian Epigenome. *Cell*, 128(4):669–681.
- Bernstein, B. E., Mikkelsen, T. S., Xie, X., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326.
- Bertoni, A. G., Burke, G. L., Owusu, J. A., et al. (2010). Inflammation and the Incidence of Type 2 Diabetes. *Diabetes Care*, 33(4):804–810.
- Bhandare, R., Schug, J., Le Lay, J., et al. (2010). Genome-wide analysis of histone modifications in human pancreatic islets. *Genome Research*, 20(4):428–433.

- Bi, X., Braunstein, M., Shei, G. J., and Broach, J. R. (1999). The yeast HML I silencer defines a heterochromatin domain boundary by directional establishment of silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 96(21):11934–11939.
- Bier, E. and McGinnis, W. (2008). Model organisms in the study of development and disease. In Erikson, J. and Wynshaw-Boris, A., editors, *Molecular Basis of Inborn Errors of Development*, pages 25–48. Oxford University Press, Epstein.
- Bird, A. (1992). The essentials of DNA methylation. *Cell*, 70(1):5–8.
- Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396–398.
- Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99.
- Birnbaum, R. Y., Clowney, E. J., Agamy, O., et al. (2012). Coding exons function as tissue-specific enhancers of nearby genes. *Genome Research*, 22(6):1059–1068.
- Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E., and Iyer, V. R. (2010). Allele-specific and heritable chromatin signatures in humans. *Human Molecular Genetics*, 19(R2):R204–209.
- Blackwood, E. M. (1998). Going the Distance: A Current View of Enhancer Action. *Science*, 281(5373):60–63.
- Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K., and Gilad, Y. (2008). Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genetics*, 4(11):e1000271.
- Blow, M. J., McCulley, D. J., Li, Z., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nature Genetics*, 42(9):818–822.
- Bluestone, J. A., Herold, K., and Eisenbarth, G. (2010). Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature*, 464(7293):1293–1300.
- Bogdanović, O., Fernandez-Miñán, A., Tena, J. J., et al. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Research*, 22(10):2043–2053.
- Boj, S. F., van Es, J. H., Huch, M., et al. (2012). Diabetes Risk Gene and Wnt Effector Tcf7l2/TCF4 Controls Hepatic Response to Perinatal and Adult Metabolic Demand. *Cell*, 151(7):1595–1607.
- Bonn, S., Zinzen, R. P., Girardot, C., et al. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156.
- Bonnefond, A., Lomberg, G., Buttar, N., et al. (2011). Disruption of a novel kruppel-like transcription factor p300-regulated pathway for insulin biosynthesis revealed by studies of the c.-331 INS mutation found in neonatal diabetes mellitus. *Journal of Biological Chemistry*, 286(32):28414–28424.
- Bonnefond, A., Philippe, J., Durand, E., et al. (2012). Whole-exome sequencing and high throughput genotyping identified KCNJ11 as the thirteenth MODY gene. *PloS One*, 7(6):e37423.

- Bonnefond, A., Yengo, L., Philippe, J., et al. (2013). Reassessment of the putative role of BLK-p.A71T loss-of-function mutation in MODY and type 2 diabetes. *Diabetologia*, 56(3):492–496.
- Bonner-Weir, S. (2001). Beta Cell Turnover; Its Assessment and Implications. *Diabetes*, 50(February):S20–24.
- Bonner-Weir, S., Sullivan, B. A., and Weir, G. C. (2015). Human Islet Morphology Revisited: Human and Rodent Islets Are Not so Different After All. *Journal of Histochemistry & Cytochemistry*, 63(8):604–612.
- Borowiec, M., Liew, C. W., Thompson, R., et al. (2009). Mutations at the BLK locus linked to maturity onset diabetes of the young and beta-cell dysfunction. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14460–14465.
- Bosco, D., Armanet, M., Morel, P., et al. (2010). Unique Arrangement of alpha- and beta-cells in Human Islets of Langerhans. *Diabetes*, 59(5):1202–1210.
- Bowman, P., Flanagan, S. E., Edghill, E. L., et al. (2012). Heterozygous ABCC8 mutations are a cause of MODY. *Diabetologia*, 55(1):123–127.
- Bradley, R. K., Li, X. Y., Trapnell, C., et al. (2010). Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species. *PLoS Biology*, 8(3):e1000343.
- Bramswig, N. C. and Kaestner, K. H. (2014). Transcriptional and epigenetic regulation in human islets. *Diabetologia*, 57(3):451–454.
- Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., and Nasmyth, K. (1985). Characterization of a "silencer" in Yeast: A DNA Sequence with Properties Opposite to Those of a Transcriptional Enhancer. *Cell*, 41(1):41–48.
- Brink, C., Chowdhury, K., and Gruss, P. (2001). Pax4 regulatory elements mediate beta cell specific expression in the pancreas. *Mechanisms of Development*, 100(1):37–43.
- Brissova, M., Fowler, M. J., Nicholson, W. E., et al. (2005). Assessment of human pancreatic islet architecture and composition by laser scanning confocal microscopy. *Journal of Histochemistry & Cytochemistry*, 53(9):1087–1097.
- Brown, C. E., Lechner, T., Howe, L., and Workman, J. L. (2000). The many HATs of transcription coactivators. *Trends in Biochemical Sciences*, 25(1):15–19.
- Brownell, J. E., Zhou, J., Ranalli, T., et al. (1996). Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, 84(6):843–851.
- Buckland, P. R. (2004). Allele-specific gene expression differences in humans. *Human Molecular Genetics*, 13(R2):255–260.
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218.
- Bulger, M. and Groudine, M. (1999). Looping versus linking: Toward a model for long-distance gene activation. *Genes & Development*, 13(19):2465–2477.

- Bulger, M. and Groudine, M. (2010). Enhancers: The abundance and function of regulatory sequences beyond promoters. *Developmental Biology*, 339(2):250–257.
- Burke, D. T., Kozloff, K. M., Chen, S., et al. (2012). Dissection of complex adult traits in a mouse synthetic population. *Genome Research*, 22(8):1549–1557.
- Burlison, J. S., Long, Q., Fujitani, Y., Wright, C. V. E., and Magnuson, M. A. (2008). Pdx-1 and Ptf1a concurrently determine fate specification of pancreatic multipotent progenitor cells. *Developmental Biology*, 316(1):74–86.
- Butler, A. E., Galasso, R., Rizza, R. A., et al. (2010). Adaptive changes in pancreatic beta cell fractional area and beta cell turnover in human pregnancy. *Diabetologia*, 53(10):2167–2176.
- Butler, A. E., Janson, J., Bonner-Weir, S., et al. (2003). Beta Cell Deficit and Increased Beta Cell Apoptosis in Humans With Type 2 Diabetes. *Diabetes*, 52(1):102–110.
- Butler, J. and Kadonaga, J. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, 16(20):2583–2592.
- Butler, J. E. F. and Kadonaga, J. T. (2001). Enhancer - promoter specificity mediated by DPE or TATA core promoter motifs service. *Genes & Development*, 15(1):2515–2519.
- Cabrera, O., Berman, D. M., Kenyon, N. S., et al. (2006). The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2334–2339.
- Cai, H. N., Arnosti, D. N., and Levine, M. (1996). Long-range repression in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 93(18):9309–9314.
- Cai, S., Lee, C. C., and Kohwi-Shigematsu, T. (2006). SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nature Genetics*, 38(1):1278–1288.
- Caicedo, A. (2013). Paracrine and autocrine interactions in the human islet: more than meets the eye. *Seminars in Cell & Developmental Biology*, 18(9):1199–1216.
- Cairns, B. R., Lorch, Y., Li, Y., et al. (1996). RSC, an essential, abundant chromatin-remodeling complex. *Cell*, 87(7):1249–1260.
- Campbell, C. D., Kirby, A., Nemesh, J., Daly, M. J., and Hirschhorn, J. N. (2008). A survey of allelic imbalance in F1 mice. *Genome Research*, 18(4):555–563.
- Caravaca, J. M., Donahue, G., Becker, J. S., et al. (2013). Bookmarking by specific and non-specific binding of FoxA1 pioneer factor to mitotic chromosomes. *Genes & Development*, 27(3):251–260.
- Carninci, P., Sandelin, A., Lenhard, B., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635.
- Carter, R. and Drouin, G. (2009). Structural differentiation of the three eukaryotic RNA polymerases. *Genomics*, 94(6):388–396.
- Chandrasekera, P. C. and Pippin, J. J. (2013). Of Rodents and Men: Species-Specific Glucose Regulation and Type 2 Diabetes Research. *Altex*, 31(2):157–176.

- Chatzigeorgiou, A., Halapas, A., Kalafatakis, K., and Kamper, E. (2009). The use of animal models in the study of diabetes mellitus. *In Vivo*, 23(2):245–258.
- Cheema, A., Adeloye, D., Sidhu, S., Sridhar, D., and Chan, K. Y. (2014). Urbanization and prevalence of type 2 diabetes in Southern Asia: A systematic analysis. *Journal of Global Health*, 4(1):102–110.
- Chen, F. C. and Li, W. H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics*, 68(2):444–456.
- Chen, H., Charlat, O., Tartaglia, L. A., et al. (1996). Evidence That the Diabetes Gene Encodes the Leptin Receptor: Identification of a Mutation in the Leptin Receptor Gene in db/db Mice. *Cell*, 84(3):491–495.
- Chen, H., Gu, X., Liu, Y., et al. (2011). PDGF signalling controls age-dependent proliferation in pancreatic β -cells. *Nature*, 478(7369):349–355.
- Chen, L. and Widom, J. (2005). Mechanism of Transcriptional Silencing in Yeast. *Cell*, 120(1):37–48.
- Chen, P., Zhao, J., Wang, Y., et al. (2013). H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes & Development*, 27(19):2109–2124.
- Chen, X., Xu, H., Yuan, P., et al. (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, 133(6):1106–1117.
- Cheng, C., Alexander, R., Min, R., et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, 22(1):1658–1667.
- Cheng, Y., Ma, Z., Kim, B.-H., et al. (2014). Principles of regulatory information conservation between mouse and human. *Nature*, 515(7527):371–375.
- Chiang, M. K. and Melton, D. A. (2003). Single-cell transcript analysis of pancreas development. *Developmental Cell*, 4(3):383–393.
- Chin, K., Oda, N., Shen, K., and Noguchi, C. T. (1995). Regulation of transcription of the human erythropoietin receptor gene by proteins binding to GATA-1 and Spi motifs. *Nucleic Acids Research*, 23(15):3041–3049.
- Chiu, Y. C., Hua, T. E., Fu, Y. Y., Pasricha, P. J., and Tang, S. C. (2012). 3-D imaging and illustration of the perfusive mouse islet sympathetic innervation and its remodelling in injury. *Diabetologia*, 55(12):3252–3261.
- Cho, H., Kim, T.-k., Mancebo, H., et al. (1999). A protein phosphatase functions to recycle RNA polymerase II A protein phosphatase functions to recycle RNA polymerase II. *Genes & Development*, 13(12):1540–1552.
- Cho, Y. S., Chen, C.-H., Hu, C., et al. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature Genetics*, 44(1):67–72.

- Choi, O. R. and Engel, J. D. (1988). Developmental regulation of beta-globin gene switching. *Cell*, 55(1):17–26.
- Cirillo, L. A., Lin, F. R., Cuesta, I., et al. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular Cell*, 9(2):279–289.
- Cirillo, L. A. and Zaret, K. S. (1999). An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Molecular Cell*, 4(6):961–969.
- Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425.
- Clearinghouse, N. D. I. (2014). Monogenic Forms of Diabetes : Neonatal Diabetes Mellitus and Maturity-onset Diabetes of the Young. Technical report, U.S. Department of Health and Human Services.
- Coleman, D. L. (1978). Obese and diabetes: two mutant genes causing diabetes-obesity syndromes in mice. *Diabetologia*, 14(3):141–148.
- Coleman, D. L. (1982). Diabetes-Obesity Syndromes in Mice. *Diabetes*, 31(1):1–6.
- Collombat, P. (2005). The simultaneous loss of Arx and Pax4 genes promotes a somatostatin-producing cell fate specification at the expense of the alpha and beta cell lineages in the mouse endocrine pancreas. *Development*, 132(13):2969–2980.
- Collombat, P., Mansouri, A., Hecksher-Sørensen, J., et al. (2003). Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes & Development*, 17(1):2591–2603.
- Conrad, E., Stein, R., and Hunter, C. S. (2014). Revealing transcription factors during human pancreatic β cell development. *Trends in Endocrinology and Metabolism*, 25(8):1–8.
- Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194.
- Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., and Myers, R. M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Research*, 16(1):1–10.
- Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909):1845–1848.
- Cornelis, M. C., Qi, L., Zhang, C., et al. (2009). Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. *Annals of Internal Medicine*, 150(8):541–550.
- Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., et al. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Research*, 24(1):1–13.
- Cotney, J., Leng, J., Yin, J., et al. (2013). The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, 154(1):185–196.
- Cox, R. D. and Church, C. D. (2011). Mouse models and the interpretation of human GWAS in type 2 diabetes and obesity. *Disease Models & Mechanisms*, 4(2):155–164.

- Cramer, P., Armache, K.-J., Baumli, S., et al. (2008). Structure of Eukaryotic RNA Polymerases. *Annual Review of Biophysics*, 37(1):337–352.
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., et al. (2010a). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.
- Creyghton, M. P., Markoulaki, S., Levine, S. S., et al. (2010b). The histone variant H2AZ is enriched at Polycomb group target genes in ES cells and is necessary for proper execution of developmental programs. *Cell*, 135(4):649–661.
- Crowley, J. J., Zhabotynsky, V., Sun, W., et al. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics*, 47(4):353–360.
- Cui, K., Zang, C., Roh, T.-Y., et al. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, 4(1):80–93.
- da Silva Xavier, G., Loder, M. K., McDonald, A., et al. (2009). TCF7L2 Regulates Late Events in Insulin Secretion From Pancreatic Islet Beta-Cells. *Diabetes*, 58(4):894–905.
- da Silva Xavier, G., Mondragon, A., Sun, G., et al. (2012). Abnormal glucose tolerance and insulin secretion in pancreas-specific Tcf7l2-null mice. *Diabetologia*, 55(10):2667–2676.
- Dabelea, D. (2007). The Predisposition to Obesity and Diabetes in Offspring of Diabetic Mothers. *Diabetes Care*, 30(2):S169–S174.
- Dahmus, M. E. (1996). Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *Journal of Biological Chemistry*, 271(32):19009–19012.
- Dai, C., Brissova, M., Hang, Y., et al. (2012). Islet-enriched gene expression and glucose-induced insulin secretion in human and mouse islets. *Diabetologia*, 55(3):707–718.
- Dai, C., Brissova, M., Reinert, R. B., et al. (2013). Pancreatic islet vasculature adapts to insulin resistance through dilation and not angiogenesis. *Diabetes*, 62(12):4144–4153.
- Danecek, P., Auton, A., Abecasis, G., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.
- Danino, Y. M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The core promoter: at the heart of gene expression. *Biochimica et Biophysica Acta*, 1849(8):1116–1131.
- Das, G., Hinkley, C. S., and Herr, W. (1995). Basal promoter elements as a selective determinant of transcriptional activator function. *Nature*, 374:657–660.
- Davidson, E. H., Jacobs, H. T., and Britten, R. J. (1983). Eukaryotic gene expression: Very short repeats and coordinate induction of genes. *Nature*, 301(10):468–470.
- De Santa, F., Barozzi, I., Mietton, F., et al. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biology*, 8(5):e1000384.
- De Vos, A., Heimberg, H., Quartier, E., et al. (1995). Human and rat beta cells differ in glucose transporter but not in glucokinase gene expression. *Journal of Clinical Investigation*, 96(5):2489–2495.

- Deaton, A. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010–1022.
- DeFronzo, R. A. (1992). Pathogenesis of Type 2 (non-insulin dependent) diabetes mellitus: a balanced overview. *Diabetologia*, 35(4):389–397.
- Degner, J. F., Pai, A. A., Pique-Regi, R., et al. (2012). DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394.
- Dejager, L., Libert, C., and Montagutelli, X. (2009). Thirty years of *Mus spretus*: a promising future. *Trends in Genetics*, 25(5):234–241.
- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation Job. *Science*, 295(5558):1306–1311.
- del Rosario, R. C.-H., Poschmann, J., Rouam, S. L., et al. (2015). Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nature Methods*, 12(5):458–470.
- Delisle, J.-S., Giroux, M., Boucher, G., et al. (2013). The TGF- β -Smad3 pathway inhibits CD28-dependent cell growth and proliferation of CD4 T cells. *Genes and Immunity*, 14(2):115–26.
- Denas, O., Sandstrom, R., Cheng, Y., et al. (2015). Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics*, 16(1):1–9.
- DeVeale, B., van der Kooy, D., and Babak, T. (2012). Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genetics*, 8(3):e1002600.
- Dimas, A. S., Deutsch, S., Stranger, B. E., et al. (2009). Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science*, 325(September):1246–1250.
- Dixon, J. R., Jung, I., Selvaraj, S., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336.
- Dixon, J. R., Selvaraj, S., Yue, F., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.
- Dogan, N., Wu, W., Morrissey, C. S., et al. (2015). Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics & Chromatin*, 8(1):16.
- Donath, M. Y., Böni-Schnetzler, M., Ellingsgaard, H., and Ehse, J. A. (2009). Islet inflammation impairs the pancreatic beta-cell in type 2 diabetes. *Physiology*, 24(12):325–331.
- Donath, M. Y., Gross, D. J., Cerasi, E., and Kaiser, N. (1999). Hyperglycemia-induced beta-cell apoptosis in pancreatic islets of *Psammomys obesus* during development of diabetes. *Diabetes*, 48(4):738–744.
- Donath, M. Y. and Halban, P. A. (2004). Decreased beta-cell mass in diabetes: Significance, mechanisms and therapeutic implications. *Diabetologia*, 47(3):581–589.
- Donnelly, S. R., Hawkins, T. E., and Moss, S. E. (1999). A conserved nuclear element with a role in mammalian gene regulation. *Human Molecular Genetics*, 8(9):1723–8.

- Dor, Y., Brown, J., Martinez, O. I., and Melton, D. A. (2004). Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. *Nature*, 429(6987):41–46.
- Doyle, A., McGarry, M. P., Lee, N. A., and Lee, J. J. (2013). The Construction of Transgenic and Gene Knockout/Knockin Mouse Models of Human Disease. *Transgenic Research*, 21(2):327–349.
- Drissen, R., Palstra, R.-J., Gillemans, N., et al. (2004). The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes & Development*, 18(20):2485–2490.
- Du, A., Hunter, C. S., Murray, J., et al. (2009). Islet-1 is Required for the Maturation, Proliferation, and Survival of the Endocrine Pancreas. *Diabetes*, 58(9):2059–2069.
- Dupuis, J., Langenberg, C., Prokopenko, I., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, 42(2):105–116.
- Durai, S. (2005). Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Research*, 33(18):5978–5990.
- Duret, L. and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, 17(1):68–74.
- Dvir, A. (2002). Promoter escape by RNA polymerase II. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1577(2):208–223.
- Edghill, E. L., Flanagan, S. E., Patch, A.-M., et al. (2008). Mutations in the INS Gene Are a Common Cause of Neonatal Diabetes but a Rare Cause of Diabetes Diagnosed in Childhood or Adulthood. *Diabetes*, 57(4):1034–1042.
- Ehrlich, M., A-Gama-Sosa, M., Huang, L.-H., et al. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8):1498–1500.
- Eizirik, D. L., Pipeleers, D. G., Ling, Z., et al. (1994). Major species differences between humans and rodents in the susceptibility to pancreatic beta-cell injury. *Proceedings of the National Academy of Sciences of the United States of America*, 91(20):9253–9256.
- El-Gohary, Y., Tulachan, S., Guo, P., et al. (2008). Smad signaling pathways regulate pancreatic endocrine development. *Developmental Biology*, 141(4):520–529.
- Elbein, S. C., Hasstedt, S. J., Wegner, K., and Kahn, S. E. (1999). Heritability of pancreatic beta-cell function among nondiabetic members of Caucasian familial type 2 diabetic kindreds. *The Journal of Clinical Endocrinology & Metabolism*, 84(4):1398–1403.
- Ellard, S. (2000). Hepatocyte Nuclear Factor 1 Alpha (HNF-1 α) Mutations in Maturity-Onset Diabetes of the Young. *Human Mutation*, 16(5):377–385.
- Ellis, H. (2007). Anatomy of the pancreas. In *Pancreas - Pathological Practice and Research*, volume 25, pages 72–73.
- Emilsson, V., Thorleifsson, G., Zhang, B., et al. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428.

- Eppig, J. T., Blake, J. A., Bult, C. J., et al. (2015). The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. *Nucleic Acids Research*, 43(1):D726–D736.
- Erlich, H., Valdes, A. M., Noble, J., et al. (2008). HLA DR-DQ Haplotypes and Genotypes and Type 1 Diabetes Risk. *Diabetes*, 57(4):1084–1092.
- Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–838.
- Ernst, J., Kheradpour, P., Mikkelsen, T. S., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.
- Eyre, S., Bowes, J., Diogo, D., et al. (2012). High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics*, 44(12):1336–1340.
- Falbo, K. B. and Shen, X. (2010). The Tango of Histone Marks and Chaperones at Replication Fork. *Molecular Cell*, 37(5):595–596.
- Fang, F., Hodges, E., Molaro, A., et al. (2012). Genomic landscape of human allele-specific DNA methylation. *Proceedings of the National Academy of Sciences*, 109(19):7332–7337.
- Farh, K. K.-h., Marson, A., Zhu, J., et al. (2014). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343.
- Farhat, B., Almelkar, A., Ramachandran, K., et al. (2013). Small human islets comprised of more Beta-cells with higher insulin content than large islets. *Islets*, 5(2):87–94.
- Ferguson-Smith, A. C., Sasaki, H., Cattanach, B. M., and Surani, M. A. (1993). Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature*, 362(6422):751–755.
- Ferrer, J., Benito, C., and Gomis, R. (1995). Pancreatic islet GLUT2 glucose transporter mRNA and protein expression in humans with and without NIDDM. *Diabetes*, 44(12):1369–1374.
- Ficz, G., Branco, M. R., Seisenberger, S., et al. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473(7347):398–402.
- Finegood, D. T., Scaglia, L., and Bonner-Weir, S. (1995). Dynamics of Beta-cell Mass in the Growing Rat Pancreas: Estimation With a Simple Mathematical Model. *Diabetes*, 44(3):249–256.
- Fisher, W. W., Li, J. J., Hammonds, A. S., et al. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proceedings of the National Academy of Sciences*, 109(52):21330–21335.
- Flint, J. and Eskin, E. (2012). Genome-wide association studies in mice. *Nature Reviews Genetics*, 13(11):807–817.
- Forrest, A. R. R., Kawaji, H., Rehli, M., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.
- Frankel, N. (2012). Multiple layers of complexity in cis-regulatory regions of developmental genes. *Developmental Dynamics*, 241(12):1857–1866.

- Frankel, N., Davis, G. K., Vargas, D., et al. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305):490–493.
- Frayling, T. M., Bulman, M. P., Ellard, S., et al. (1997). Mutations in the Hepatocyte Nuclear Factor-1alpha Gene Are a Common Cause of Maturity-Onset Diabetes of the Young in the U.K. *Diabetes*, 46(4):720–772.
- Frazer, K. A., Eskin, E., Kang, H. M., et al. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448(7157):1050–1053.
- Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251.
- Freeman, W. H., Francisco, S., West, M. H. P., and Bonner, W. M. (1980). Histone 2A, a Heteromorphous Family of Eight Protein Species. *Biochemistry*, 19(1979):3238–3245.
- Froguel, P., Vaxillaire, M., Sun, F., et al. (1992). Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature*, 356(6365):162–164.
- Froguel, P., Zouali, H., Vionnet, N., et al. (1993). Familial hyperglycemia due to mutations in glucokinase. *The New England Journal of Medicine*, 328(10):697–702.
- Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 13(12):840–852.
- Fussner, E., Ching, R. W., and Bazett-Jones, D. P. (2011). Living without 30nm chromatin fibers. *Trends in Biochemical Sciences*, 36(1):1–6.
- Gannon, F., O’Hare, K., Perrin, F., et al. (1989). Organisation and sequenced at the 5’ end of a cloned complete ovalbumin gene. *Nature*, 278(5703):428–434.
- Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282.
- Gasa, R., Mrejen, C., Leachman, N., et al. (2004). Proendocrine genes coordinate the pancreatic islet differentiation program in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13245–13250.
- Gaulton, K. J., Ferreira, T., Lee, Y., et al. (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature Genetics*, 47(12):1415–1425.
- Gaulton, K. J., Nammo, T., Pasquali, L., et al. (2010). A map of open chromatin in human pancreatic islets. *Nature Genetics*, 42(3):255–259.
- Georgia, S. and Bhushan, A. (2004). Beta Cell Replication Is the Primary Mechanism for Maintaining Postnatal Beta Cell Mass. *Journal of Clinical Investigation*, 114(7):963–968.
- Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature Genetics*, 42(7):558–560.
- Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145.
- Giorgetti, L., Siggers, T., Tiana, G., et al. (2010). Noncooperative Interactions between

- Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. *Molecular Cell*, 37(3):418–428.
- Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6):877–85.
- Gittes, G. K. (2009). Developmental biology of the pancreas: a comprehensive review. *Developmental Biology*, 326(1):4–35.
- Göke, J., Jung, M., Behrens, S., et al. (2011). Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Computational Biology*, 7(12):e1002304.
- Gomez, J. A., Wapinski, O. L., Yang, Y. W., et al. (2013). NeST, a long noncoding RNA, controls microbial susceptibility and epigenetic activation of the *Ifng* locus. *Cell*, 152(4):743–754.
- Goncalves, A., Leigh-Brown, S., Thybert, D., et al. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research*, 22(12):2376–2384.
- Gonsorcikova, L., Pruhova, S., Cinek, O., et al. (2008). Autosomal inheritance of diabetes in two families characterized by obesity and a novel H241Q mutation in *NEUROD1*. *Pediatric Diabetes*, 9(4 Pt 2):367–72.
- Goode, D. K., Snell, P., and Elgar, G. (2003). Comparative analysis of vertebrate *Shh* genes identifies novel conserved non-coding sequence. *Mammalian Genome*, 14(3):192–201.
- Görisch, S. M., Wachsmuth, M., Tóth, K. F., Lichter, P., and Rippe, K. (2005). Histone acetylation increases chromatin accessibility. *Journal of Cell Science*, 118(Pt 24):5825–5834.
- Grant, S. F. A., Thorleifsson, G., Reynisdottir, I., et al. (2006). Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nature Genetics*, 38(3):320–323.
- Grarup, N., Sparsø, T., and Hansen, T. (2010). Physiologic characterization of type 2 diabetes-related loci. *Current Diabetes Reports*, 10(6):485–497.
- Green, M. R. (2005). Eukaryotic transcription activation: right on target. *Molecular Cell*, 18(4):399–402.
- Gregg, C., Zhang, J., Butler, J. E., Haig, D., and Dulac, C. (2010a). Sex-Specific Parent-of-Origin Allelic Expression in the Mouse Brain. *Science*, 682(5992):682–685.
- Gregg, C., Zhang, J., Weissbourd, B., et al. (2010b). High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain. *Science*, 329(5992):643–648.
- Groop, L., Forsblom, C., Lehtovirta, M., et al. (1996). Metabolic consequences of a family history of NIDDM (the Botnia study): Evidence for sex-specific parental effects. *Diabetes*, 45(11):1585–1593.
- Groth, A., Rocha, W., Verreault, A., and Almouzni, G. (2007). Chromatin Challenges during DNA Replication and Repair. *Cell*, 128(4):721–733.
- Grubert, F., Zaugg, J., Kasowski, M., et al. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065.

- Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. *Nature*, 389(6649):349–352.
- GTExConsortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.
- Gubitosi-Klug, R. A. (2014). The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study at 30 Years: Summary and Future Directions. *Diabetes Care*, 37(1):44–49.
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*, 8(2):R24.
- Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., et al. (2015). Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. *PLoS Genetics*, 11(1):1–25.
- Haberle, V., Li, N., Hadzhiev, Y., et al. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, 507(7492):381–385.
- Haering, C. H. and Jessberger, R. (2012). Cohesin in determining chromosome architecture. *Experimental Cell Research*, 318(12):1386–1393.
- Halter, J. B. and Chang, A. M. (2003). Aging and insulin secretion. *Handbook of the Biology of Aging*, 284(1):E7–E12.
- Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews*, 62(2):465–503.
- Handoko, L., Xu, H., Li, G., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, 43(7):630–638.
- Hao, E., Tyrberg, B., Itkin-Ansari, P., et al. (2006). Beta-cell differentiation from nonendocrine epithelial cells of the adult human pancreas. *Nature Medicine*, 12(3):310–316.
- Hara, K., Fujita, H., Johnson, T. A., et al. (2014). Genome-wide association study identifies three novel loci for type 2 diabetes. *Human Molecular Genetics*, 23(1):239–246.
- Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7):469–483.
- Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., and Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genetics*, 4(6):e1000106.
- Harjutsalo, V., Sjöberg, L., and Tuomilehto, J. (2008). Time trends in the incidence of type 1 diabetes in Finnish children: a cohort study. *Lancet*, 371(9626):1777–1782.
- Harris, M. B., Mostecky, J., and Rothman, P. B. (2005). Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *Journal of Biological Chemistry*, 280(13):13114–13121.
- Hart, A. W., Baeza, N., Apelqvist, A., and Edlund, H. (2000). Attenuation of FGF signalling in mouse beta-cells leads to diabetes. *Nature*, 408(6814):864–868.

- Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., et al. (2015). QuASAR: Quantitative allele-specific analysis of reads. *Bioinformatics*, 31(8):1235–1242.
- Hauge-Evans, A. C., King, A. J., Carmignac, D., et al. (2009). Somatostatin secreted by islet δ -cells fulfills multiple roles as a paracrine regulator of islet function. *Diabetes*, 58(2):403–411.
- He, A., Kong, S. W., Ma, Q., and Pu, W. T. (2011). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14):5632–5637.
- Hebrok, M., Kim, S. K., and Melton, D. A. (1998). Notochord repression of endodermal sonic hedgehog permits pancreas development. *Genes & Development*, 12(11):1705–1713.
- Heintzman, N. D., Hon, G. C., Hawkins, R. D., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.
- Heintzman, N. D., Stuart, R. K., Hon, G., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318.
- Heinz, S., Benner, C., Spann, N., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589.
- Heit, J. J., Karnik, S. K., and Kim, S. K. (2006). Intrinsic regulators of pancreatic beta-cell proliferation. *Annual Review of Cell and Developmental Biology*, 22(1):311–338.
- Henquin, J.-C., Dufrane, D., and Nenquin, M. (2006). Nutrient Control of Insulin Secretion in Isolated Normal Human Islets. *Diabetes*, 55(12):3470–3477.
- Henry Gray, F. (1918). *Anatomy of the human body*. Lea & Febiger, Philadelphia and New York, 20 edition.
- Herrera, P. L., Huarte, J., Sanvito, F., et al. (1991). Embryogenesis of the murine endocrine pancreas; early expression of pancreatic polypeptide gene. *Development*, 113(4):1257–1265.
- Herreros-Villanueva, M. (2012). Mouse models of pancreatic cancer. *World Journal of Gastroenterology*, 18(12):1286–1294.
- Hilton, I. B., D’Ippolito, A. M., Vockley, C. M., et al. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33(5):510–7.
- Hinrichs, A., Karolchik, D., Baertsch, R., et al. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(Database issue):D590–D598.
- Hnisz, D., Abraham, B. J., Lee, T. I., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–47.
- Hoffman, B. G., Robertson, G., Zavaglia, B., et al. (2010). Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Research*, 20(8):1037–1051.
- Holliday, R. and Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232.

- Holstege, F., Fiedler, U., and Timmers, H. (1997). Three transitions in the RNA polymerase II transcription complex during initiation. *The EMBO Journal*, 16(24):7468–80.
- Holstege, F. C., Jennings, E. G., Wyrick, J. J., et al. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–728.
- Horikawa, Y., Iwasaki, N., Hara, M., et al. (1997). Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY. *Nature Genetics*, 17(4):384–385.
- Hou, J. C., Min, L., and Pessin, J. E. (2009). Insulin Granule Biogenesis, Trafficking and Exocytosis. *Vitamins & Hormones*, 80(1):473–506.
- Hsieh, C.-L., Fei, T., Chen, Y., et al. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences*, 111(20):7319–7324.
- Hu, G., Cui, K., Northrup, D., et al. (2013). H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, 12(2):180–92.
- Huang, H. P., Liu, M., El-Hodiri, H. M., et al. (2000). Regulation of the pancreatic islet-specific gene BETA2 (neuroD) by neurogenin 3. *Molecular and Cellular Biology*, 20(9):3292–3307.
- Huang, J., Chen, J., Esparza, J., et al. (2015). eQTL mapping identifies insertion- and deletion-specific eQTLs in multiple tissues. *Nature Communications*, 8(6):6821.
- Huang, J., Liu, X., Li, D., et al. (2016). Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Developmental Cell*, 36(1):9–23.
- Ihm, S.-H., Matsumoto, I., Sawada, T., et al. (2006). Effect of Donor Age on Function of Isolated Human Islets. *Diabetes*, 55(5):1361–1368.
- Imamura, F., O'Connor, L., Ye, Z., et al. (2015). Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction. *British Journal of Sports Medicine*, 50(8):496–504.
- Imamura, M., Maeda, S., Yamauchi, T., et al. (2012). A single-nucleotide polymorphism in ANK1 is associated with susceptibility to type 2 diabetes in Japanese populations. *Human Molecular Genetics*, 21(13):3042–3049.
- Inui, A., Akihiro, A., Bowers, C. Y., et al. (2004). Ghrelin, appetite, and gastric motility: the emerging role of the stomach as an endocrine organ. *The FASEB Journal*, 18(3):439–456.
- Ioannou, Y. S., Ellard, S., Hattersley, A., and Skordis, N. (2011). KCNJ11 activating mutations cause both transient and permanent neonatal diabetes mellitus in Cypriot patients. *Pediatric Diabetes*, 12(2):133–7.
- Ishikawa, K., Tsunekawa, S., Ikeniwa, M., et al. (2015). Long-Term Pancreatic Beta Cell Exposure to High Levels of Glucose but Not Palmitate Induces DNA Methylation within the Insulin Gene Promoter and Represses Transcriptional Activity. *Plos One*, 10(2):e0115350.
- Itkin-Ansari, P., Marcora, E., Geron, I., et al. (2005). NeuroD1 in the endocrine pancreas:

- Localization and dual function as an activator and repressor. *Developmental Dynamics*, 233(3):946–953.
- Jacquemin, P., Durviaux, S. M., Jensen, J., et al. (2000). Transcription factor hepatocyte nuclear factor 6 regulates pancreatic endocrine cell differentiation and controls expression of the proendocrine gene *ngn3*. *Molecular and Cellular Biology*, 20(12):4445–4454.
- Jäger, R., Migliorini, G., Henrion, M., et al. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, 6(2):6178.
- Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M., and Epstein, D. J. (2006). A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development*, 133(4):761–772.
- Jin, Q., Yu, L.-R., Wang, L., et al. (2011). Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *The EMBO Journal*, 30(2):249–262.
- Jin, T. and Liu, L. (2008). The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus. *Molecular Endocrinology*, 22(11):2383–2392.
- Jo, W., Endo, M., Ishizu, K., Nakamura, A., and Tajima, T. (2011). A novel PAX4 mutation in a Japanese patient with maturity-onset diabetes of the young. *The Tohoku Journal of Experimental Medicine*, 223(2):113–118.
- Johannsen, W. (2014). The genotype conception of heredity. *The American Naturalist*, 30(355):536–553.
- Johansson, B. B., Torsvik, J., Bjorkhaug, L., et al. (2011). Diabetes and Pancreatic Exocrine Dysfunction Due to Mutations in the Carboxyl Ester Lipase Gene-Maturity Onset Diabetes of the Young (CEL-MODY). *Journal of Biological Chemistry*, 286(40):34593–34605.
- Johnson, J. D., Ahmed, N. T., Luciani, D. S., et al. (2003). Increased islet apoptosis in Pdx1 +/- mice. *The Journal of Clinical Investigation*, 111(8):1147–1160.
- Jones, P. A. (1985). Altering gene expression with 5-azacytidine. *Cell*, 40(3):485–486.
- Jonkers, I. and Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, 16(3):167–177.
- Jørgensen, M. C., Ahnfelt-Rønne, J., Hald, J., et al. (2007). An illustrated review of early pancreas development in the mouse. *Endocrine Reviews*, 28(6):685–705.
- Joung, J. K. and Sander, J. D. (2013). TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews. Molecular Cell Biology*, 14(1):49–55.
- Junion, G., Spivakov, M., Girardot, C., et al. (2012). A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell*, 148(3):473–486.
- Kadonaga, J. T. (2004). Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell*, 116(2):247–257.
- Kagey, M. H., Newman, J. J., Bilodeau, S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435.

- Kahn, S. E. (2001). Clinical review 135: The importance of beta-cell failure in the development and progression of type 2 diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 86(9):4047–4058.
- Kaikkonen, M. U., Spann, N. J., Heinz, S., et al. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell*, 51(3):310–325.
- Kaprio, J., Tuomilehto, J., Koskenvuo, M., et al. (1992). Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia*, 35(11):1060–1067.
- Karczewski, K. J., Tatonetti, N. P., Landt, S. G., et al. (2011). Cooperative transcription factor associations discovered using regulatory variation. *Proceedings of the National Academy of Sciences*, 108(32):13353–13358.
- Kashi, K., Henderson, L., Bonetti, A., and Carninci, P. (2016). Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochimica et Biophysica Acta journal*, 1859(1):3–15.
- Kasowski, M., Grubert, F., Heffelfinger, C., et al. (2010). Variation in transcription factor binding among humans. *Science*, 328(5975):232–235.
- Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., et al. (2013). Extensive Variation in Chromatin States Across Humans. *Science*, 342(6159):750–752.
- Kataoka, K., Shioda, S., Ando, K., et al. (2004). Differentially expressed Maf family transcription factors, c-Maf and MafA, activate glucagon and insulin gene expression in pancreatic islet alpha- and beta-cells. *Journal of Molecular Endocrinology*, 32(1):9–20.
- Kawaji, H., Frith, M. C., Katayama, S., et al. (2006). Dynamic usage of transcription start sites within core promoters. *Genome Biology*, 7(12):R118.
- Ke, X., Taylor, M. S., and Cardon, L. R. (2008). Singleton SNPs in the human genome and implications for genome-wide association studies. *European Journal of Human Genetics*, 16(4):506–515.
- Keane, T. M., Goodstadt, L., Danecek, P., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.
- Khaitovich, P. (2005). Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. *Science*, 309(5742):1850–1854.
- Khaja, R., Zhang, J., MacDonald, J. R., et al. (2006). Genome assembly comparison identifies structural variants in the human genome. *Nature Genetics*, 38(12):1413–1418.
- Khoo, C., Yang, J., Weinrott, S. A., et al. (2012). Research resource: the Pdx1 cistrome of pancreatic islets. *Molecular Endocrinology*, 26(3):521–533.
- Khoueiry, P., Garfield, D. A., Ciglar, L., et al. (2016). Shadow Enhancers Are Pervasive Features of Article Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Current Biology*, 26(1):38–51.
- Khoury, G. and Gruss, P. (1983). Enhancer elements. *Cell*, 33(2):313–314.

- Kiba, T. (2004). Relationships between the autonomic nervous system and the pancreas including regulation of regeneration and apoptosis: recent developments. *Pancreas*, 29(2):e51–e58.
- Kidd, J. M., Cooper, G. M., Donahue, W. F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.
- Kido, Y., Nakae, J., and Accili, D. (2001). Clinical review 125: The insulin receptor and its cellular targets. *The Journal of Clinical Endocrinology & Metabolism*, 86(3):972–979.
- Kilpinen, H. and Dermitzakis, E. T. (2012). Genetic and epigenetic contribution to complex traits. *Human Molecular Genetics*, 21(R1):R24–8.
- Kilpinen, H., Waszak, S. M., Gschwind, A. R., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, 342(6159):744–747.
- Kim, D., Pertea, G., Trapnell, C., et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36.
- Kim, M. K., Lesoon-Wood, L. A., Weintraub, B. D., and Chung, J. H. (1996). A soluble transcription factor, Oct-1, is also found in the insoluble nuclear matrix and possesses silencing activity in its alanine-rich domain. *Molecular and Cellular Biology*, 16(8):4366–4377.
- Kim, T. H., Barrera, L. O., Zheng, M., et al. (2005). A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880.
- Kim, T.-K., Hemberg, M., Gray, J. M., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.
- King, A. J. F. (2012). The use of animal models in diabetes research. *British Journal of Pharmacology*, 166(3):877–894.
- King, M.-C. and Wilson, A. (1975). Evolution at Two Levels in Humans and Chimpanzees. *Science*, 188(4184):107–116.
- Kireeva, M. L., Walter, W., Tchernajenko, V., et al. (2002). Nucleosome Remodeling Induced by RNA Polymerase II. *Molecular Cell*, 9(3):541–552.
- Kleinjan, D. A. and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics*, 76(1):8–32.
- Klok, M. D., Jakobsdottir, S., and Drent, M. L. (2007). The role of leptin and ghrelin in the regulation of food intake and body weight in humans: a review. *Obesity Reviews*, 8(1):21–34.
- Koch, F. and Andrau, J.-C. (2011). Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specificity. *Transcription*, 2(6):263–268.
- Koch, F., Fenouil, R., Gut, M., et al. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Structural & Molecular Biology*, 18(8):956–963.
- Komatsu, M., Takei, M., Ishii, H., and Sato, Y. (2013). Glucose-stimulated insulin secretion: A newer perspective. *Journal of Diabetes Investigation*, 4(6):511–516.

- Kooner, J. S., Saleheen, D., Sim, X., et al. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics*, 43(10):984–989.
- Korinek, V., Barker, N., Moerer, P., et al. (1998). Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nature Genetics*, 19(4):379–383.
- Kornberg, R. D. (1975). Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139):868–871.
- Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends in Biochemical Sciences*, 24(12):46–49.
- Kornberg, R. D. and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–294.
- Kornblihtt, A. R., Mata, M. D. E. L. A., Fededa, J. P., and Muñoz, M. J. (2004). Multiple links between transcription and splicing. *RNA*, 10(10):1489–1498.
- Koska, J., DelParigi, A., De Courten, B., Weyer, C., and Tataranni, P. A. (2004). Pancreatic polypeptide is involved in the regulation of body weight in Pima Indian male subjects. *Diabetes*, 53(12):3091–3096.
- Kowalczyk, M. S., Hughes, J. R., Garrick, D., et al. (2012). Intragenic Enhancers Act as Alternative Promoters. *Molecular Cell*, 45(4):447–458.
- Kristinsson, S. Y., Thorolfsson, E. T., Talseth, B., et al. (2001). MODY in Iceland is associated with mutations in HNF-1 α and a novel mutation in NeuroD1. *Diabetologia*, 44(11):2098–2103.
- Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2015). Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. *PloS One*, 10(6):e0131673.
- Ku, G. M., Kim, H., Vaughn, I. W., et al. (2012). Research resource: RNA-Seq reveals unique features of the pancreatic β -cell transcriptome. *Molecular Endocrinology*, 26(10):1783–92.
- Kulaeva, O. I., Hsieh, F.-K., and Studitsky, V. M. (2010). RNA polymerase complexes cooperate to relieve the nucleosomal barrier and evict histones. *Proceedings Of The National Academy Of Sciences (Of The United States Of America)*, 107(25):11325–11330.
- Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, 130(26):6569–6575.
- Kulkarni, R. N. (2004). The islet β -cell. *The International Journal of Biochemistry & Cell Biology*, 36(3):365–371.
- Kulkarni, R. N., Jhala, U. S., Winnay, J. N., et al. (2004). PDX-1 haploinsufficiency limits the compensatory islet hyperplasia that occurs in response to insulin resistance. *Journal of Clinical Investigation*, 114(6):828–836.
- Kulkarni, R. N. and Stewart, A. F. (2014). Summary of the Keystone Islet Workshop (April 2014): The Increasing Demand for Human Islet Availability in Diabetes Research. *Diabetes*, 63(12):3979–3981.

- Kulkarni, S. D., Muralidharan, B., Panda, A. C., et al. (2011). Glucose-stimulated translation regulation of insulin by the 5' UTR-binding proteins. *Journal of Biological Chemistry*, 286(16):14146–14156.
- Kundaje, A., Meuleman, W., Ernst, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.
- Kuras, L., Borggreffe, T., and Kornberg, R. D. (2003). Association of the Mediator complex with enhancers of active genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):13887–13891.
- Kvon, E. Z., Kazmar, T., Stampfel, G., et al. (2014). Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature*, 512(7512):91–95.
- Lai, F., Orom, U. A., Cesaroni, M., et al. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature*, 494(7438):497–501.
- Lander, E. and Schork, N. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.
- Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.
- Lappalainen, T., Sammeth, M., Friedländer, M. R., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.
- Lavia, P., Macleod, D., and Bird, A. (1987). Coincident start sites for divergent transcripts at a randomly selected CpG-rich island of mouse. *The EMBO Journal*, 6(9):2773–2779.
- Layman, W. S. and Zuo, J. (2015). Epigenetic regulation in the inner ear and its potential roles in development, protection, and regeneration. *Frontiers in Cellular Neuroscience*, 8(446):1–11.
- Lebrecht, D., Foehr, M., Smith, E., et al. (2005). Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13176–13181.
- Lee, D., Gorkin, D. U., Baker, M., et al. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, 47(8):955–61.
- Leibiger, B., Moede, T., Schwarz, T., et al. (1998). Short-term regulation of insulin gene transcription by glucose. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8):9307–9312.
- Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233–245.
- León-Novelo, L. G., McIntyre, L. M., Fear, J. M., and Graze, R. M. (2014). A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics*, 15(920):1–11.
- Leroy, C., Cormier, L., and Kuras, L. (2006). Independent recruitment of mediator and SAGA by the activator Met4. *Molecular and Cellular Biology*, 26(8):3149–3163.
- Lettice, L. A. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14):1725–1735.

- Leung, D., Jung, I., Rajagopal, N., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539):350–354.
- Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147–151.
- Lewontin, R. C. and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472.
- Li, G., Bahn, J. H., Lee, J. H., et al. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, 40(13):1–13.
- Li, H., Handsaker, B., Wysoker, A., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Li, Q., Seo, J.-H., Stranger, B., et al. (2013). Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. *Cell*, 152(3):633–641.
- Lia, G., Praly, E., Ferreira, H., et al. (2006). Direct observation of DNA distortion by the RSC complex. *Molecular Cell*, 21(3):417–425.
- Liao, B. Y. and Zhang, J. (2006). Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Molecular Biology and Evolution*, 23(6):1119–1128.
- Lindberg, K., Rønn, S. G., Tornehave, D., et al. (2005). Regulation of pancreatic β -cell mass and proliferation by SOCS-3. *Journal of Molecular Endocrinology*, 35(2):231–243.
- Lindner, T. H., Njølstad, P. R., Horikawa, Y., et al. (1999). A novel syndrome of diabetes mellitus, renal dysfunction and genital malformation associated with a partial deletion of the pseudo-POU domain of hepatocyte nuclear factor-1 β . *Human Molecular Genetics*, 8(11):2001–2008.
- Link, N., Kurtz, P., Neal, M. O., Garcia-Hughes, G., and Abrams, J. M. (2013). A p53 enhancer region regulates target genes through chromatin conformations in cis and in trans. *Genes & Development*, 27(22):2433–2438.
- Linnemann, A. K., Baan, M., and Davis, D. B. (2014). Pancreatic Beta-Cell Proliferation in Obesity. *Advances in Nutrition*, 5(3):278–288.
- Locke, J. M., Hysenaj, G., Wood, A. R., Weedon, M. N., and Harries, L. W. (2015). Targeted allelic expression profiling in human islets identifies cis-regulatory effects for multiple variants identified by type 2 diabetes genome-wide association studies. *Diabetes*, 64(4):1484–91.
- Long, J., Edwards, T., Signorello, L. B., et al. (2012). Evaluation of genome-wide association study-identified type 2 diabetes loci in African Americans. *American Journal of Epidemiology*, 176(11):995–1001.
- Lopata, M. A., Cleveland, D. W., and Sollner-Webb, B. (1986). RNA polymerase specificity of mRNA production and enhancer action. *Proceedings of the National Academy of Sciences of the United States of America*, 83(18):6677–6681.
- Lorberbaum, D. S. and Barolo, S. (2015). Enhancers: holding out for the right promoter. *Current Biology*, 25(7):R290–R293.

- Lorch, Y., LaPointe, J. W., and Kornberg, R. D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, 49(2):203–210.
- Lorch, Y., Zhang, M., and Kornberg, R. D. (2001). RSC unravels the nucleosome. *Molecular Cell*, 7(1):89–95.
- Lorenzen, T., Pociot, E., and Hougaard (1994). Long-term risk of IDDM in first-degree relatives of patients with IDDM. *Diabetologia*, 37(3):321–327.
- Louie, M. C., Yang, H. Q., Ma, A.-H., et al. (2003). Androgen-induced recruitment of RNA polymerase II to a nuclear receptor-p160 coactivator complex. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2226–2230.
- Lovén, J., Hoke, H., Lin, C., et al. (2013). Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell*, 153(2):320–334.
- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260.
- Lupiáñez, D. G., Kraft, K., Heinrich, V., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012–1025.
- Lupien, M., Eeckhoute, J., Meyer, C. A., et al. (2008). FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell*, 132(6):958–970.
- Lyssenko, V., Almgren, P., Anevski, D., et al. (2005). Predictors of and Longitudinal Changes in Insulin Sensitivity and Secretion Preceding Onset of Type 2 Diabetes. *Diabetes*, 54(1):166–174.
- Lyssenko, V., Lupi, R., Marchetti, P., et al. (2007). Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *The Journal of Clinical Investigation*, 117(8):2155–2163.
- Ma, R. C. W., Hu, C., Tam, C. H., et al. (2013). Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4. *Diabetologia*, 56(6):1291–1305.
- Maahs, D. M. and Rewers, M. (2006). Editorial: Mortality and renal disease in type 1 diabetes mellitus—progress made, more to be done. *The Journal of Clinical Endocrinology and Metabolism*, 91(10):3757–3759.
- MacArthur, S., Li, X.-Y., Li, J., et al. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology*, 10(7):R80.
- MacDonald, P. E., Joseph, J. W., and Rorsman, P. (2005). Glucose-sensing mechanisms in pancreatic beta cells. *Philosophical Transactions of the Royal Society*, 360(1464):2211–2225.
- Maestro, M. A., Boj, S. F., Paniagua, A., et al. (2001). Hepatic Nuclear Factor 1- α Directs Nucleosomal Hyperacetylation to Its Tissue-Specific Transcriptional Targets. *Molecular and Cellular Biology*, 21(9):3234–3243.

- Mahajan, A., Go, M. J., Zhang, W., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–44.
- Malcolm, J. C., Lawson, M. L., Gaboury, I., Lough, G., and Keely, E. (2006). Glucose tolerance of offspring of mother with gestational diabetes mellitus in a low-risk population. *Diabetic medicine : a journal of the British Diabetic Association*, 23(5):565–70.
- Manke, T., Heinig, M., and Vingron, M. (2010). Quantifying the effect of sequence variation on regulatory interactions. *Human Mutation*, 31(4):477–483.
- Manke, T., Roider, H. G., and Vingron, M. (2008). Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Computational Biology*, 4(3):e1000039.
- Margueron, R. and Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, 11(4):285–296.
- Marino-Ramirez, L., Jordan, I. K., and Landsman, D. (2006). Multiple independent evolutionary solutions to core histone gene regulation. *Genome Biology*, 7(12):1–17.
- Marks, H., Kerstens, H. H. D., Barakat, T. S., et al. (2015). Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome Biology*, 16(149):1–20.
- Marshak, S., Benshushan, E., Shoshkes, M., et al. (2000). Functional conservation of regulatory elements in the pdx-1 gene: PDX-1 and hepatocyte nuclear factor 3beta transcription factors mediate beta-cell-specific expression. *Molecular and Cellular Biology*, 20(20):7583–7590.
- Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J., and Maltais, L. J. (2002). The Human and Mouse Replication-Dependent Histone Genes. *Genomics*, 80(5):487–498.
- Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59.
- Mathelier, A., Shi, W., and Wasserman, W. W. (2015). Identification of altered cis -regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76.
- Mathelier, A., Zhao, X., Zhang, A. W., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(D1):D142—D147.
- Matsukage, A., Hirose, F., Yoo, M. A., and Yamaguchi, M. (2008). The DRE/DREF transcriptional regulatory system: a master key for cell proliferation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1779(2):81–89.
- Maurano, M. T., Humbert, R., Rynes, E., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195.
- McCracken, S., Fong, N., Yankulov, K., et al. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, 385(6614):357–361.
- McCulloch, L. J., van de Bunt, M., Braun, M., et al. (2011). GLUT2 (SLC2A2) is not the principal glucose transporter in human pancreatic beta cells: Implications for understanding genetic association signals at this locus. *Molecular Genetics and Metabolism*, 104(4):648–653.

- McDaniell, R., Lee, B.-K., Song, L., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, 328(5975):235–239.
- McLean, C. Y., Bristor, D., Hiller, M., et al. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501.
- McLean1, C. Y., Reno, P. L., Pollen, A. A., et al. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, 471(5):216–219.
- McPeck, M. S. (2000). From mouse to human: fine mapping of quantitative trait loci in a model organism. *Proceedings of the National Academy of Sciences of the United States of America*, 97(23):12389–90.
- McVicker, G., Geijn, B. V. D., Degner, J. F. J., et al. (2013). Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*, 342(6159):747–749.
- Meier, J. J., Butler, A. E., Saisho, Y., et al. (2008). Beta-cell replication is the primary mechanism subserving the postnatal expansion of beta-cell mass in humans. *Diabetes*, 57(6):1584–1594.
- Melnikov, A., Murugan, A., Zhang, X., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277.
- Mercer, E. M., Lin, Y. C., Benner, C., et al. (2011). Multilineage Priming of Enhancer Repertoires Precedes Commitment to the B and Myeloid Cell Lineages in Hematopoietic Progenitors. *Immunity*, 35(3):413–425.
- Merika, M. and Thanos, D. (2001). Enhanceosomes. *Current Opinion in Genetics & Development*, 11(1):205–208.
- Merli, C., Bergstrom, D. E., Cygan, J. A., and Blackman, R. K. (1996). Promoter specificity mediates the independent regulation of neighboring genes. *Genes & Development*, 10(1):1260–1270.
- Metzger, B. E., Buchanan, T. A., Coustan, D. R., et al. (2007). Summary and Recommendations of the Fifth International Workshop-Conference on Gestational Diabetes Mellitus. *Diabetes Care*, 30(Supplement 2):S251–S260.
- Michael, M. D., Kulkarni, R. N., Postic, C., et al. (2000). Loss of insulin signaling in hepatocytes leads to severe insulin resistance and progressive hepatic dysfunction. *Molecular Cell*, 6(1):87–97.
- Mifsud, B., Tavares-Cadete, F., Young, A. N., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606.
- Mills, R., Luttig, C., and Larkins, C. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, 16:1182–1190.
- Minguez, P., Parca, L., Diella, F., et al. (2012). Deciphering a global network of functionally associated post-translational modifications. *Molecular Systems Biology*, 8(599):1–16.
- Minokoshi, Y., Kahn, C. R., and Kahn, B. B. (2003). Tissue-specific Ablation of the GLUT4 Glucose Transporter or the Insulin Receptor Challenges Assumptions about Insulin Action and Glucose Homeostasis. *Journal of Biological Chemistry*, 278(36):33609–33612.

- Mishiro, T., Ishihara, K., Hino, S., et al. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *The EMBO Journal*, 28(9):1234–1245.
- Montgomery, S. B. and Dermitzakis, E. T. (2011). From expression QTLs to personalized transcriptomics. *Nature Reviews Genetics*, 12(4):277–282.
- Moon, H., Filippova, G., Loukinov, D., et al. (2005). CTCF is conserved from *Drosophila* to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Reports*, 6(2):165–170.
- Morán, I., Akerman, I., Van De Bunt, M., et al. (2012). Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metabolism*, 16(4):435–448.
- Morris, A. P., Voight, B. F., Teslovich, T. M., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990.
- Multhaup, M., Seldin, M., Jaffe, A., et al. (2015). Mouse-Human Experimental Epigenetic Analysis Unmasks Dietary Targets and Genetic Liability for Diabetic Phenotypes. *Cell Metabolism*, 21(1):138–149.
- Musunuru, K., Strong, A., Frank-Kamenetsky, M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719.
- Myers, A. J., Gibbs, J. R., Webster, J. A., et al. (2007). A survey of genetic human cortical gene expression. *Nature Genetics*, 39(12):1494–1499.
- Navarro-Tableros, V., Sanchez-Soto, M. C., Garcia, S., and Hiriart, M. (2004). Autocrine regulation of single pancreatic beta-cell survival. *Diabetes*, 53(8):2018–2023.
- Nembaware, V., Lupindo, B., Schouest, K., et al. (2008). Genome-wide survey of allele-specific splicing in humans. *BMC Genomics*, 9(265):1–15.
- Nembaware, V., Wolfe, K. H., Bettoni, F., Kelso, J., and Seoighe, C. (2004). Allele-specific transcript isoforms in human. *FEBS Letters*, 577(1-2):233–238.
- Newgard, C. B., Lu, D., Jensen, M. V., et al. (2002). Stimulus/secretion coupling factors in glucose-stimulated insulin secretion: insights gained from a multidisciplinary approach. *Diabetes*, 51 Suppl 3(7):S389–S393.
- Ni, T., Corcoran, D. L., Rach, E. A., et al. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods*, 7(7):521–527.
- Nica, A. C., Parts, L., Glass, D., et al. (2011). The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genetics*, 7(2):e1002003.
- Nicolae, D. L., Gamazon, E., Zhang, W., et al. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4):e1000888.
- Nicolino, M., Claiborn, K. C., Senée, V., et al. (2010). A novel hypomorphic PDX1 mutation responsible for permanent neonatal diabetes with subclinical exocrine deficiency. *Diabetes*, 59(3):733–40.
- Nielsen, J. H., Galsgaard, E. D., Møldrup, A., et al. (2001). Regulation of beta-cell mass by hormones and growth factors. *Diabetes*, 50(S1):S25–S29.

- Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS One*, 7(7):e37558.
- Njølstad, P. R., Sagen, J. V., Bjørkhaug, L., et al. (2003). Permanent neonatal diabetes caused by glucokinase deficiency: inborn error of the glucose-insulin signaling pathway. *Diabetes*, 52(11):2854–2860.
- Nora, E. P., Lajoie, B. R., Schulz, E. G., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385.
- Ochi, H., Tamai, T., Nagano, H., et al. (2012). Evolution of a tissue-specific silencer underlies divergence in the expression of pax2 and pax8 paralogues. *Nature Communications*, 3(5):848.
- Oehler, S., Eismann, E. R., Krämer, H., and Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO Journal*, 9(4):973–979.
- Ogbourne, S. and Antalis, T. M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, 331(Pt1):1–14.
- Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, 23(1):263–286.
- Onengut-Gumuscu, S., Chen, W.-M., Burren, O., et al. (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics*, 47(4):381–386.
- Ong, C.-T. and Corces, V. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet*, 12(4):283–293.
- Ørom, U. A., Derrien, T., Beringer, M., et al. (2010). Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell*, 143(1):46–58.
- Ostuni, R., Piccolo, V., Barozzi, I., et al. (2013). Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1-2):157–71.
- Ozaki, K., Ohnishi, Y., Iida, A., et al. (2002). Functional SNPs in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650–4.
- Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, 61(1):1053–1095.
- Pal, M., McKean, D., and Luse, D. S. (2001). Promoter clearance by RNA polymerase II is an extended, multistep process strongly affected by sequence. *Molecular and Cellular Biology*, 21(17):5815–5825.
- Palmer, N. D., McDonough, C. W., Hicks, P. J., et al. (2012). A Genome-Wide Association Search for Type 2 Diabetes Genes in African Americans. *PLoS One*, 7(1):e29202.
- Pares-Matos, E. I. (2005). Electrophoretic Mobility-Shift and Super-Shift Assays for Studies and Characterization of Protein-DNA Complexes. In *Methods in Molecular Biology: Gene Regulation*, volume 977, pages 159–167.

- Parker, S. C. J., Stitzel, M. L., Taylor, D. L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17921–17926.
- Parnaud, G., Bosco, D., Berney, T., et al. (2008). Proliferation of sorted human and rat beta cells. *Diabetologia*, 51(1):91–100.
- Parra, E. J., Below, J. E., Krithika, S., et al. (2011). Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. *Diabetologia*, 54(8):2038–2046.
- Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, 46(2):136–143.
- Passarge, E. (1979). Emil Heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. *American journal of human genetics*, 31(2):106–115.
- Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Publishing Group*, 11(8):533–538.
- Pastor, W. A., Pape, U. J., Huang, Y., et al. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, 473(7347):394–397.
- Patwardhan, R. P., Hiatt, J. B., Witten, D. M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, 30(3):265–270.
- Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., et al. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO Journal*, 30(20):4198–4210.
- Peltonen, L. (2006). Lessons from studying monogenic disease for common disease. *Human Molecular Genetics*, 15(1):R67–R74.
- Pennacchio, L. A., Ahituv, N., Moses, A. M., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.
- Pennacchio, L. A., Loots, G. G., Nobrega, M. A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. *Genome Research*, 17(2):201–211.
- Perry, M. W., Boettiger, A. N., Bothma, J. P., and Levine, M. (2010). Shadow enhancers foster robustness of drosophila gastrulation. *Current Biology*, 20(17):1562–1567.
- Peters, L. L., Robledo, R. F., Bult, C. J., et al. (2007). The mouse as a model for human biology: a resource guide for complex trait analysis. *Nature Reviews Genetics*, 8(1):58–69.
- Pettitt, D. J., Bennett, P. H., Knowler, W. C., Baird, H. R., and Aleck, K. A. (1985). Gestational Diabetes Mellitus and Impaired Glucose Tolerance During Pregnancy: Long-Term Effects on Obesity and Glucose Tolerance in the Offspring. *Diabetes*, 34(June):119–122.
- Phillips-Cremins, J., Sauria, M., Sanyal, A., et al. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, 153(6):1281–1295.
- Pickrell, J. K., Marioni, J. C., Pai, A. A., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772.

- Pierucci, D., Cicconi, S., Bonini, P., et al. (2001). NGF-withdrawal induces apoptosis in pancreatic beta cells in vitro. *Diabetologia*, 44(10):1281–1295.
- Plengvidhya, N., Kooptiwut, S., Songtawee, N., et al. (2007). Brief report: PAX4 mutations in Thais with maturity onset diabetes of the young. *Journal of Clinical Endocrinology and Metabolism*, 92(7):2821–2826.
- Pociot, F. and McDermott, M. F. (2002). Genetics of type 1 diabetes mellitus. *Genes and Immunity*, 3(5):235–249.
- Pokholok, D. K., Hannett, N. M., and Young, R. A. (2002). Exchange of RNA polymerase II initiation and elongation factors during gene expression in vivo. *Molecular Cell*, 9(4):799–809.
- Polak, M. and Cavé, H. (2007). Neonatal diabetes mellitus: a disease linked to multiple mechanisms. *Orphanet Journal of Rare Diseases*, 2(2):1–12.
- Ponjavic, J., Lenhard, B., Kai, C., et al. (2006). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biology*, 7(8):R78.
- Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nature Publishing Group*, 47(1):8–12.
- Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.
- Rach, E. A., Winter, D. R., Benjamin, A. M., et al. (2011). Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. *PLoS Genetics*, 7(1):e1001274.
- Rada-Iglesias, A., Bajpai, R., Swigut, T., et al. (2010). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283.
- Reddy, M. V. P. L., Wang, H., Liu, S., et al. (2011). Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population. *Genes and Immunity*, 12(3):208–212.
- Reddy, T. E., Gertz, J., Pauli, F., et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research*, 22(5):860–869.
- Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510.
- Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32.
- Reilly, S. K., Yin, J., Ayoub, A. E., et al. (2015). Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science*, 347(6226):1155–1159.
- Reinius, B. and Sandberg, R. (2015). Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nature Reviews Genetics*, 16(11):653–664.
- Relton, C. L. and Smith, G. D. (2010). Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment. *PLoS Medicine*, 7(10):e1000356.

- Reményi, A., Schöler, H. R., and Wilmanns, M. (2004). Combinatorial control of gene expression. *Nature Structural & Molecular Biology*, 11(9):812–815.
- Rhodes, C. J. (2005). Type 2 Diabetes - a Matter of Beta-Cell Life and Death? *Science*, 307(1):380–384.
- Ridderstråle, M. and Groop, L. (2009). Genetic dissection of type 2 diabetes. *Molecular and Cellular Endocrinology*, 297(1-2):10–17.
- Rieck, S. and Kaestner, K. H. (2010). Expansion of β -cell mass in response to pregnancy. *Trends Endocrinol Metab.*, 21(3):151–158.
- Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetics*, 14(1):9–25.
- Rippe, K., von Hippel, P., and Langowski, J. (1995). Action at a distance: DNA-looping and initiation of transcription. *Trends in Biochemical Sciences*, 20(12):500–506.
- Risch, N. (1987). Assessing the role of HLA-linked and unlinked determinants of disease. *The American Journal of Human Genetics*, 40(1):1–14.
- Risch, N. and Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science*, 273(9):1516–1517.
- Riserus, U., Willet, W., and Hu, F. (2009). Dietary fats and prevention of type 2 diabetes. *Progress in Lipid Research*, 48(1):44–51.
- Ritter, D. I., Dong, Z., Guo, S., and Chuang, J. H. (2012). Transcriptional enhancers in protein-coding exons of vertebrate developmental genes. *PloS One*, 7(5):e35202.
- Robinson, P. J. J., Fairall, L., Huynh, V. A. T., and Rhodes, D. (2006). EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proceedings of the National Academy of Sciences of the United States of America*, 103(17):6506–6511.
- Rodriguez-Diaz, R., Abdulreda, M. H., Formoso, A. L., et al. (2011). Innervation patterns of autonomic axons in the human endocrine pancreas. *Cell Metabolism*, 14(1):45–54.
- Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, 21(9):327–335.
- Roep, B. O. (2003). The role of T-cells in the pathogenesis of Type 1 diabetes: from cause to cure. *Diabetologia*, 46:305–321.
- Roider, H. G., Lenhard, B., Kanhere, A., Haas, S. A., and Vingron, M. (2009). CpG-depleted promoters harbor tissue-specific transcription factor binding signals - Implications for motif overrepresentation analyses. *Nucleic Acids Research*, 37(19):6305–6315.
- Romanel, A., Lago, S., Prandi, D., Sboner, A., and Demichelis, F. (2015). ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Medical Genomics*, 8(9):1–12.
- Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L., and Almouzni, G. (2015). Epigenomics: Roadmap for regulation. *Nature*, 518(7539):314–316.
- Ronshaugen, M. and Levine, M. (2004). Visualization of trans-homolog enhancer-promoter in-

- teractions at the Abd-B hox locus in the *Drosophila* embryo. *Developmental Cell*, 7(6):925–932.
- Rorsman, P. and Braun, M. (2013). Regulation of Insulin Secretion in Human Pancreatic Islets. *Annual Review of Physiology*, 75(1):155–179.
- Rosenberg, L. (1999). In vivo cell transformation: neogenesis of beta cells from pancreatic ductal cells. *Cell Transplantation*, 25(9):1417–1424.
- Rosengren, A. H., Braun, M., Mahdi, T., et al. (2012). Reduced insulin exocytosis in human pancreatic beta cells with gene variants linked to type 2 diabetes. *Diabetes*, 61(7):1726–1733.
- Rozowsky, J., Abyzov, A., Wang, J., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7(522):1–15.
- Rulifson, I. C., Karnik, S. K., Heiser, P. W., et al. (2007). Wnt signaling regulates pancreatic beta cell proliferation. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6247–6252.
- Rusche, L. N., Kirchmaier, A. L., and Rine, J. (2003). The establishment, inheritance, and function of silenced chromatin in *Saccharomyces Cerevisiae*. *Annual Review of Biochemistry*, 72(1):481–516.
- Rusk, N. (2009). When ChIA PETs meet Hi-C. *Nature Methods*, 6(12):863.
- Rutter, G. A. (2001). Nutrient-secretion coupling in the pancreatic islet beta cell: recent advances. *Molecular Aspects of Medicine*, 22(6):247–284.
- Rutter, G. A. (2009). Regulating glucagon secretion: Somatostatin in the spotlight. *Diabetes*, 58(2):299–301.
- Sachidanandam, R., Weissman, D., Schmidt, S., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933.
- Sagen, J. V., Raeder, H., Hathout, E., et al. (2004). Permanent neonatal diabetes due to mutations in KCNJ11 encoding Kir6.2: patient characteristics and initial response to sulfonylurea therapy. *Diabetes*, 53(10):2713–2718.
- Saha, A., Wittmeyer, J., and Cairns, B. R. (2006). Chromatin remodelling: the industrial revolution of DNA around histones. *Nature Reviews Molecular Cell Biology*, 7(6):437–447.
- Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, 16(3):129–143.
- Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. *Epigenetics & Chromatin*, 8(3):1–17.
- Salas-Salvadó, J., Martínez-González, M., Bulló, M., and Ros, E. (2011). The role of diet in the prevention of type 2 diabetes. *Nutrition, Metabolism and Cardiovascular Diseases*, 21(S2):B32–B48.
- Sambasivarao, S. V. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends in Biochemical Sciences*, 39(4):1199–1216.

- Sandelin, A., Carninci, P., Lenhard, B., et al. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, 8(6):424–436.
- Sander, J. D. and Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, 32(4):347–355.
- Sander, M., Sussel, L., Connors, J., et al. (2000). Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta-cell formation in the pancreas. *Development*, 127(24):5533–5540.
- Sandoval, D. A., Obici, S., and Seeley, R. J. (2009). Targeting the CNS to treat type 2 diabetes. *Nature Reviews. Drug Discovery*, 8(5):386–398.
- Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113.
- Savic, D., Ye, H., Aneas, I., et al. (2011). Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Research*, 21(9):1417–25.
- Saxena, R., Elbers, C., Guo, Y., et al. (2012). Large-Scale Gene-Centric Meta-Analysis across 39 Studies Identifies Type 2 Diabetes Loci. *The American Journal of Human Genetics*, 90(3):410–425.
- Saxena, R., Saleheen, D., Been, L. F., et al. (2013). Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. *Diabetes*, 62(5):1746–55.
- Schaukowitch, K., Joo, J.-Y., Liu, X., et al. (2014). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell*, 56(1):29–42.
- Schilling, E. and Rehli, M. (2007). Global, comparative analysis of tissue-specific promoter CpG methylation. *Genomics*, 90(3):314–323.
- Schmidt, D., Wilson, M. D., Ballester, B., et al. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, 1036(5):1036–1041.
- Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.
- Schug, J., Schuller, W.-P., Kappen, C., et al. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biology*, 6(4):R33.
- Schuit, F., Flamez, D., Vos, A. D., and Pipeleers, D. (2002). Glucose-Regulated Gene Expression Maintaining the. *Diabetes*, 51(S3):326–332.
- Schuit, F. C. (1997). Is GLUT2 required for glucose sensing? *Diabetologia*, 40(1):104–111.
- Schuit, F. C., Huypens, P., Heimberg, H., and Pipeleers, D. G. (2001). Glucose sensing in pancreatic beta-cells: A model for the study of other glucose-regulated cells in gut, pancreas, and hypothalamus. *Diabetes*, 50(1):1–11.
- Schwartz, B. E. and Ahmad, K. (2005). Transcriptional activation triggers deposition and removal of the histone variant H3.3. *Genes & Development*, 19(7):804–14.
- Schwartz, Y. B., Kahn, T. G., Stenberg, P., et al. (2010). Alternative epigenetic chromatin states of polycomb target genes. *PLoS Genetics*, 6(1):e1000805.

- Schwitzgebel, V. M., Scheel, D. W., Connors, J. R., et al. (2000). Expression of neurogenin3 reveals an islet cell precursor population in the pancreas. *Development*, 127(16):3533–3542.
- Scott, L. J., Bonnycastle, L. L., Willer, C. J., et al. (2006). Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a finnish sample. *Diabetes*, 55(9):2649–2653.
- Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., et al. (2007). A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science*, 316(6):1341–1345.
- Seet, B. T., Dikic, I., Zhou, M.-M., and Pawson, T. (2006). Reading protein modifications with interaction domains. *Nature Reviews Molecular Cell Biology*, 7(7):473–483.
- Senger, K., Armstrong, G. W., Rowell, W. J., et al. (2004). Immunity Regulatory DNAs Share Common Organizational Features in Drosophila. *Molecular Cell*, 13(1):19–32.
- Sérandour, A. A., Avner, S., Percevault, F., et al. (2011). Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Research*, 21(4):555–565.
- Serre, D., Gurd, S., Ge, B., et al. (2008). Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genetics*, 4(2):e1000006.
- Serup, P., Petersen, H. V., Pedersen, E. E., et al. (1995). The homeodomain protein IPF-1/STF-1 is expressed in a subset of islet cells and promotes rat insulin 1 gene expression dependent on an intact E1 helix-loop-helix factor binding site. *The Biochemical Journal*, 310(Pt3):997–1003.
- Seymour, P. A. (2014). Sox9: A master regulator of the pancreatic program. *Review of Diabetic Studies*, 11(1):51–83.
- Shandilya, J. and Roberts, S. G. (2012). The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(5):391–400.
- Sharp, A., Stathaki, E., E, M., et al. (2011). DNA methylation profiles of human active and inactive X chromosomes. *Genome Research*, 21(10):1592–1600.
- Sheffield, N. C., Thurman, R. E., Song, L., et al. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding , and long-range interactions. *Genome Research*, 23(5):777–788.
- Shei, G. J. and Broach, J. R. (1995). Yeast silencers can act as orientation-dependent gene inactivation centers that respond to environmental signals. *Molecular and Cellular Biology*, 15(7):3496–506.
- Shen, Y., Yue, F., McCleary, D. F., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120.
- Sherry, S. T., Ward, M. H., Kholodov, M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–11.
- Shibata, Y., Sheffield, N. C., Fedrigo, O., et al. (2012). Extensive evolutionary changes in

- regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genetics*, 8(6):e1002789.
- Shim, Y. J., Kim, J. E., Hwang, S.-K., et al. (2015). Identification of Candidate Gene Variants in Korean MODY Families by Whole-Exome Sequencing. *Hormone Research in Paediatrics*, 83(4):242–251.
- Shiraki, T., Kondo, S., Katayama, S., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15776–15781.
- Shu, L., Sauter, N. S., Schulthess, F. T., and Matveyenko, A. V. (2008). Transcription Factor 7-Like 2 Regulates B-cell Survival and Function in Human Pancreatic Islets. *Diabetes*, 57(3):645–653.
- Shu, X. O., Long, J., Cai, Q., et al. (2010). Identification of new genetic risk variants for Type 2 Diabetes. *PLoS Genetics*, 6(9):e1001127.
- Siepel, A., Bejerano, G., Pedersen, J. S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050.
- Simonet, W. S., Bucay, N., Pitas, R. E., Lauer, S. J., and Taylor, J. M. (1991). Multiple tissue-specific elements control the apolipoprotein E/C-I gene locus in transgenic mice. *The Journal of Biological Chemistry*, 266(14):8651–8654.
- Sladek, R., Rocheleau, G., Rung, J., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885.
- Slatkin, M. (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485.
- Smale, S. T. and Kadonaga, J. T. (2003). The RNA Polymerase II Core Promoter. *Annual Review of Biochemistry*, 72(1):449–479.
- Smith, S. B., Qu, H.-Q., Taleb, N., et al. (2010). Rfx6 Directs Islet Formation and Insulin Production in Mice and Humans. *Nature*, 463(7282):775–780.
- Søgaard, T. M. M. and Svejstrup, J. Q. (2007). Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *Journal of Biological Chemistry*, 282(19):14113–14120.
- Soravia, E., Grebe, A., De Luca, P., et al. (1995). A conserved TATA-less proximal promoter drives basal transcription from the urokinase-type plasminogen activator receptor gene. *Blood*, 86(2):624–635.
- Soufi, A., Garcia, M. F., Jaroszewicz, A., et al. (2015). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell*, 161(3):555–568.
- Soyer, J., Flasse, L., Raffelsberger, W., et al. (2010). Rfx6 is an Ngn3-dependent winged helix transcription factor required for pancreatic islet cell development. *Development*, 137(2):203–212.
- Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626.

- Spivakov, M., Akhtar, J., Kheradpour, P., et al. (2012). Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biology*, 13(9):R49.
- Splinter, E., Heath, H., Kooren, J., et al. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development*, 20(17):2349–2354.
- Spooner, B. S., Walther, B. T., and Rutter, W. J. (1970). The development of the dorsal and ventral mammalian pancreas in vivo and in vitro. *Journal of Cell Biology*, 47(1):235–246.
- Srinivasan, L. and Atchison, M. L. (2004). YY1 DNA binding and PcG recruitment requires CtBP. *Genes & development*, 18(21):2596–601.
- St-Onge, L., Sosa-Pineda, B., Chowdhury, K., Mansouri, A., and Gruss, P. (1997). Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. *Nature*, 387(6631):406–409.
- Staal, F. J. T., Luis, T. C., and Tiemessen, M. M. (2008). WNT signalling in the immune system: WNT is spreading its wings. *Nature Reviews. Immunology*, 8(8):581–93.
- Stadler, M. B., Murr, R., Burger, L., et al. (2012). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 484(7395):550–550.
- Stalvey, M. S., Muller, C., Schatz, D. A., et al. (2006). An Activator of the Glucagon Gene Expressed in Developing Islet alpha- and beta-Cells. *Diabetes*, 55(2):15–17.
- Steensma, D. P., Kyle, R. A., and Shampo, M. A. (2010). Abbie Lathrop, the "Mouse Woman of Granby": Rodent Fancier and Accidental Genetics Pioneer. *Mayo Clinic Proceedings*, 85(11):e83.
- Stefansson, H., Rye, D. B., Hicks, A., et al. (2007). A Genetic Risk Factor for Periodic Limb Movements in Sleep. *The New England Journal of Medicine*, 357(7):639–647.
- Stefflova, K., Thybert, D., Wilson, M. D., et al. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–40.
- Steimle, V., Durand, B., Barras, E., et al. (1995). A novel DNA-binding regulatory factor is mutated in primary MHC class II deficiency (bare lymphocyte syndrome). *Genes and Development*, 9(9):1021–1032.
- Steinthorsdottir, V., Thorleifsson, G., Sulem, P., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature Genetics*, 46(3):294–298.
- Stergachis, A. B., Neph, S., Sandstrom, R., et al. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–370.
- Stoffers, D. A., Ferrer, J., Clarke, W. L., and Habener, J. F. (1997). Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nature Genetics*, 17(2):138–139.
- Stolovich-Rain, M., Hija, a., Grimsby, J., Glaser, B., and Dor, Y. (2012). Pancreatic Beta Cells in Very Old Mice Retain Capacity for Compensatory Proliferation. *Journal of Biological Chemistry*, 287(33):27407–27414.
- Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45.

- Stranger, B. E., Nica, A. C., Forrest, M. S., et al. (2007). Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224.
- Struhl, K. (1987). Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast. *Cell*, 49(3):295–297.
- Stumvoll, M., Goldstein, B. J., and Van Haeften, T. W. (2005). Type 2 diabetes: Principles of pathogenesis and therapy. *Lancet*, 365(9467):1333–1346.
- Subramanian, A., Tamayo, P., Mootha, V. K., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.
- Suckale, J. (2008). Pancreas islets in metabolic signaling - focus on the beta-cell. *Frontiers in Bioscience*, 1(13):7156–7171.
- Sudmant, P. H., Rausch, T., Gardner, E. J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.
- Sujitjorn, J., Kooptiwut, S., Chongjaroen, N., et al. (2016). Aberrant mRNA splicing of paired box 4 (PAX4) IVS7-1GtoA mutation causing maturity-onset diabetes of the young, type 9. *Acta Diabetologica*, 53(2):205–216.
- Sun, L. Y. and Bartke, A. (2015). Tissue-specific GHR knockout mice: Metabolic phenotypes. *Frontiers in Endocrinology*, 5(243):1–4.
- Sur, I. K., Hallikas, O., Vähärautio, A., et al. (2012). Mice Lacking a Myc Enhancer That Includes Human SNP rs6983267 Are Resistant to Intestinal Tumors. *Science*, 338(6112):1360–1363.
- Svejstrup, J. Q. (2004). The RNA polymerase II transcription cycle: Cycling through chromatin. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1677(1-3):64–73.
- Swamynathan, S. K. and Piatigorsky, J. (2002). Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse Shsp/alphaB-crystallin and Mkbp/HspB2 genes. *Journal of Biological Chemistry*, 277(51):49700–49706.
- Swanson, C. I., Evans, N. C., and Barolo, S. (2010). Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Developmental Cell*, 18(3):359–370.
- Symmons, O. and Spitz, F. (2013). From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620):20120358.
- Tahiliani, M., Koh, K. P., Shen, Y., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929):930–935.
- Taylor, B., Liu, F. F., and Sander, M. (2013). Nkx6.1 Is Essential for Maintaining the Functional State of Pancreatic Beta Cells. *Cell Reports*, 4(6):1262–1275.
- Taylor, M. S., Kai, C., Kawai, J., et al. (2006). Heterotachy in mammalian promoter evolution. *PLoS Genetics*, 2(4):e30.

- Temple, I. K. and Shield, J. P. H. (2002). Transient neonatal diabetes, a disorder of imprinting. *Journal of Medical Genetics*, 39(12):872–875.
- Tennant, B. R., Robertson, A. G., Kramer, M., et al. (2013). Identification and analysis of murine pancreatic islet enhancers. *Diabetologia*, 56(3):542–552.
- Teperino, R., Lempradl, A., and Pospisilik, J. A. (2013). Bridging epigenomics and complex disease: The basics. *Cellular and Molecular Life Sciences*, 70(9):1609–1621.
- Thanos, D. and Maniatis, T. (1995). Virus induction of human IFN β gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–1100.
- Thatcher, T. H. and Gorovsky, M. A. (1994). Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Research*, 22(2):174–179.
- The ENCODE Project Consortium, Dunham, I., Kundaje, A., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.
- Thomas, M. C. and Chiang, C.-M. (2006). The General Transcription Machinery and General Cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, 41(3):105–178.
- Thomas-Chollier, M., Hufton, A., Heinig, M., et al. (2011). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols*, 6(12):1860–1869.
- Thompson, Ian M, Lucia, M. S., Parnes, H. L., Minasian, L. M., et al. (2006). TCF7L2 Polymorphisms and Progression to Diabetes in the Diabetes Prevention Program. *The New England Journal of Medicine*, 355(3):2239–2246.
- Thomsen, S. K. and Gloyn, A. L. (2014). The pancreatic beta cell: recent insights from human genetics. *Trends in Endocrinology & Metabolism*, 25(8):425–434.
- Thurman, R. E., Rynes, E., Humbert, R., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.
- Tie, F., Banerjee, R., Stratton, C. A., et al. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development*, 136(18):3131–3141.
- Tjian, R. and Maniatis, T. (1994). Transcriptional activation: A complex puzzle with few easy pieces. *Cell*, 77(1):5–8.
- Treutlein, B., Muschielok, A., Andrecka, J., et al. (2012). Dynamic Architecture of a Minimal RNA Polymerase II Open Promoter Complex. *Molecular Cell*, 46(2):136–146.
- Trimarchi, T., Bilal, E., Ntziachristos, P., et al. (2014). Genome-wide Mapping and Characterization of Notch-Regulated Long Noncoding RNAs in Acute Leukemia. *Cell*, 158(3):593–606.
- Trinklein, N. T., Force Aldred, S., Hartman, S. J., et al. (2004). An abundance of bidirectional promoters in the human genome. *Genome Research*, 14(1):62–66.
- Trynka, G., Hunt, K. A., Bockett, N. A., et al. (2012a). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*, 43(12):1193–1201.

- Trynka, G., Sandor, C., Han, B., et al. (2012b). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.
- Tschen, S.-I., Dhawan, S., Gurlo, T., and Bhushan, A. (2009). The capacity of beta-cell regeneration in mice. *Diabetes*, 58(6):1312–1320.
- Udvardy, A., Maine, E., and Schedl, P. (1985). The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *Journal of Molecular Biology*, 185(2):341–58.
- Uesaka, M., Nishimura, O., Go, Y., et al. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC genomics*, 15(35):1–14.
- Vaag, A., Henriksen, J. E., Madsbad, S., Holm, N., and Beck-Nielsen, H. (1995). Insulin secretion, insulin action, and hepatic glucose production in identical twins discordant for non-insulin-dependent diabetes mellitus. *The Journal of Clinical Investigation*, 95(2):690–698.
- Vakoc, C. R., Letting, D. L., Gheldof, N., et al. (2005). Proximity among Distant Regulatory Elements at the β -Globin Locus Requires GATA-1 and FOG-1. *Molecular Cell*, 17(3):453–462.
- van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12(11):1061–1063.
- van der Weyden, L., White, J. K., Adams, D. J., and Logan, D. W. (2011). The mouse genetics toolkit: revealing function and mechanism. *Genome Biology*, 12(224):1–11.
- Vanhille, L., Griffon, A., Maqbool, M. A., et al. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, 15(6):6905.
- Varki, A. and Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Research*, 15(12):1746–1758.
- Vaxillaire, M., Rouard, M., Yamagata, K., et al. (1997). Identification of nine novel mutations in the hepatocyte nuclear factor 1 alpha gene associated with maturity-onset diabetes of the young (MODY3). *Human Molecular Genetics*, 6(4):583–586.
- Verrijzer, C. P., Chen, J. L., Yokomori, K., and Tjian, R. (1995). Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. *Cell*, 81(7):1115–1125.
- Vetere, A., Choudhary, A., Burns, S. M., and Wagner, B. K. (2014). Targeting the pancreatic β -cell to treat diabetes. *Nature Reviews. Drug Discovery*, 13(4):278–289.
- Villar, D., Berthelot, C., Flicek, P., et al. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566.
- Villareal, D. T., Robertson, H., Bell, G. I., et al. (2010). TCF7L2 variant rs7903146 affects the risk of type 2 diabetes by modulating incretin action. *Diabetes*, 59(2):479–485.
- Villasenor, A., Chong, D. C., Henkemeyer, M., and Cleaver, O. (2010). Epithelial dynamics of pancreatic branching morphogenesis. *Development*, 137(24):4295–4305.

- Visel, A., Blow, M. J., Li, Z., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.
- Visel, A., Bristow, J., and Pennacchio, L. A. (2007). Enhancer identification through comparative genomics. *Seminars in Cell & Developmental Biology*, 18(1):140–152.
- Visel, A., Prabhakar, S., Akiyama, J. A., et al. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genetics*, 40(2):158–160.
- Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics*, 40(5):489–490.
- Visser, M., Palstra, R.-J., and Kayser, M. (2015). Allele-specific transcriptional regulation of IRF4 in melanocytes is mediated by chromatin looping of the intronic rs12203592 enhancer to the IRF4 promoter. *Human Molecular Genetics*, 24(9):2649–2661.
- Voight, B. F., Scott, L. J., Steinthorsdottir, V., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 42(7):579–589.
- Vučičević, D., Corradin, O., Ntini, E., Scacheri, P. C., and Ørom, U. A. (2015). Long ncRNA expression associates with tissue-specific enhancers. *Cell Cycle*, 14(2):253–260.
- Wade, C. M., Kulbokas, E. J., Kirby, A. W., et al. (2002). The mosaic structure of variation in the laboratory mouse genome. *Nature*, 420(6915):574–578.
- Wakeley, J., Nielsen, R., Liu-Cordero, S. N., and Ardlie, K. (2001). The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *American Journal of Human Genetics*, 69(6):1332–1347.
- Wall, J. D. and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587–597.
- Wallace, J. A. and Felsenfeld, G. (2007). We gather together: insulators and genome organization. *Current Opinion in Genetics and Development*, 17(5):400–407.
- Wallet, M. A. and Tisch, R. (2006). Type 1 diabetes, inflammation and dendritic cells. *Drug Discovery Today: Disease Mechanisms*, 3(3):373–379.
- Wang, D., Garcia-Bassets, I., Benner, C., et al. (2011a). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351):390–394.
- Wang, E. T., Sandberg, R., Luo, S., et al. (2008a). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.
- Wang, J., Elghazi, L., Parker, S. E., et al. (2004). The concerted activities of Pax4 and Nkx2.2 are essential to initiate pancreatic β -cell differentiation. *Developmental Biology*, 266(1):178–189.
- Wang, K. C., Yang, Y. W., Liu, B., et al. (2011b). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341):120–124.
- Wang, L., Wang, S., and Li, W. (2012). RSeQC: Quality Control of RNA-seq experiments. *Nucleic Acids Research*, 40(10):e72–e72.
- Wang, P., Fiaschi-Taesch, N. M., Vasavada, R. C., et al. (2015). Diabetes mellitus—advances and challenges in human β -cell proliferation. *Nature Reviews Endocrinology*, 11(4):201–212.

- Wang, W., Carey, M., and Gralla, J. D. (1992). Polymerase II Promoter Activation: Closed Complex Formation and ATP-Driven Start Site Opening. *Science*, 255(6):6–9.
- Wang, X., Sun, Q., McGrath, S. D., et al. (2008b). Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One*, 3(12):e3839.
- Wang, Z., Zang, C., Rosenfeld, J. A., et al. (2008c). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903.
- Waszak, S. M., Delaneau, O., Gschwind, A. R., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, 162(5):1–12.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62.
- Weatherall, D. J. (2000). Single gene disorders or complex traits: lessons from the thalassemias and other monogenic diseases. *British Medical Journal (Clinical research ed.)*, 324(5):1135–1138.
- Webb, G. C., Akbar, M. S., Zhao, C., and Steiner, D. F. (2000). Expression profiling of pancreatic beta cells: glucose regulation of secretory and metabolic pathway genes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11):5773–5778.
- Weber, J. L., David, D., Heil, J., et al. (2002). Human diallelic insertion/deletion polymorphisms. *The American Journal of Human Genetics*, 71(4):854–862.
- Weedon, M. N. and Frayling, T. M. (2008). Reaching new heights: insights into the genetics of human stature. *Trends in Genetics*, 24(12):595–603.
- Wei, W. and Brennan, M. D. (2000). Polarity of transcriptional enhancement revealed by an insulator element. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26):14518–14523.
- Weil, P. A., Luse, D. S., Segall, J., and Roeder, R. G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. *Cell*, 18(2):469–484.
- Weiss, S. B. and Gladstone, L. (1959). A Mammalian System for the Incorporation of Cytidine Triphosphate Into Ribonucleic Acid. *Journal of the American Chemical Society*, 81(15):4118–4119.
- Welter, D., MacArthur, J., Morales, J., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006.
- Wentworth, B. M., Schaefer, I. M., Villa-Komaroff, L., and Chirgwin, J. M. (1986). Characterization of the two nonallelic genes encoding mouse preproinsulin. *Journal of Molecular Evolution*, 23(4):305–312.
- Westra, H.-J. and Franke, L. (2014). From genome to function by studying eQTLs. *Biochimica et Biophysica Acta*, 1842(10):1896–1902.
- White, M. A., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29):11952–11957.

- Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, 18(5):1368–1373.
- Whyte, W. A., Orlando, D. A., Hnisz, D., et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319.
- Wilhelm, B. T., Marguerat, S., Watt, S., et al. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243.
- Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191.
- Williamson, I., Eskeland, R., Lettice, L. A., et al. (2012). Anterior-posterior differences in HoxD chromatin topology in limb development. *Development*, 139(17):3157–3167.
- Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., et al. (2008). Species-Specific Transcription in Mice Carrying Human Chromosome 21. *Science*, 322(5900):434–438.
- Winter, E. E., Goodstadt, L., and Ponting, C. P. (2004). Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research*, 14(1):54–61.
- Wit, E. D. and Laat, W. D. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26(1):11–24.
- Wong, K. H., Jin, Y., and Struhl, K. (2014). TFIIF Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Molecular Cell*, 54(4):601–612.
- Woods, S. C., Lutz, T. A., Geary, N., and Langhans, W. (2006). Pancreatic signals controlling food intake; insulin, glucagon and amylin. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1471):1219–1235.
- Woolfe, A., Goodson, M., Goode, D. K., et al. (2005). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology*, 3(1):e7.
- Workman, J. L. and Kingston, R. E. (1992). Nucleosome Core Displacement in Vitro via a Metastable Transcription Factor- Nucleosome Complex. *Science*, 258(5089):1780–1784.
- World Health Organization (2014a). Global status report on noncommunicable diseases 2014. Technical report.
- World Health Organization (2014b). World Health statistics 2014. Technical report.
- Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Genetics*, 8(3):206–216.
- WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.
- Wu, H. and Zhang, Y. (2011). Tet1 and 5-hydroxymethylation. *Cell Cycle*, 10(15):2428–2436.
- Wu, Y., Liu, C., Sun, H., et al. (2011). Growth hormone receptor regulates beta cell hyperplasia and glucose-stimulated insulin secretion in obese mice. *The Journal of Clinical Investigation*, 121(6):2422–2426.
- Xiang, A. H., Peters, R. K., Trigo, E., et al. (1999). Multiple metabolic defects during late pregnancy in women at high risk for type 2 diabetes. *Diabetes*, 48(4):848–854.

- Xiao, X., Chen, Z., Shiota, C., et al. (2013). No evidence for beta cell neogenesis in murine adult pancreas. *Journal of Clinical Investigation*, 123(5):2207–2217.
- Xu, J., Pope, S. D., Jazirehi, A. R., et al. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12377–12382.
- Xu, J., Shao, Z., Glass, K., et al. (2012). Combinatorial Assembly of Developmental Stage-Specific Enhancers Controls Gene Expression Programs during Human Erythropoiesis. *Developmental Cell*, 23(4):796–811.
- Xu, J., Watts, J. A., Pope, S. D., et al. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. *Genes & Development*, 23(24):2824–2838.
- Yamagata, K., Furuta, H., Oda, N., et al. (1996). Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). *Nature*, 384(6608):458–460.
- Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. (2005). Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene*, 350(2):129–136.
- Yang, H., Bell, T. A., Churchill, G. A., and Pardo-Manuel de Villena, F. (2007). On the sub-specific origin of the laboratory mouse. *Nature Genetics*, 39(9):1100–1107.
- Yang, J., Benyamin, B., McEvoy, B. P., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.
- Yang, J., Su, A. I., and Li, W. H. (2005). Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Molecular Biology and Evolution*, 22(10):2113–2118.
- Yang, S., Oksenberg, N., Takayama, S., et al. (2015). Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics*, 16(882):1–13.
- Yang, X., Han, H., De Carvalho, D. D., et al. (2014). Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell*, 26(4):577–590.
- Yoo, E. J., Cooke, N. E., and Liebhaber, S. A. (2012). An RNA-Independent Linkage of Non-coding Transcription to Long-Range Enhancer Function. *Molecular and Cellular Biology*, 32(10):2020–2029.
- Yoshiki, A. and Moriwaki, K. (2006). Mouse phenome research: implications of genetic background. *ILAR journal*, 47(2):94–102.
- Yu, M., Hon, G., Szulwach, K., et al. (2012). Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell*, 149(6):1368–1380.
- Yun, M., Wu, J., Workman, J. L., and Li, B. (2011). Readers of histone modifications. *Cell Research*, 21(4):564–578.
- Yusufzai, T. M., Tagami, H., Nakatani, Y., and Felsenfeld, G. (2004). CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species. *Molecular Cell*, 13(2):291–298.

- Zabidi, M. A., Arnold, C. D., Schernhuber, K., et al. (2014). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 1(7540):556–559.
- Zang, C., Schones, D. E., Zeng, C., et al. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25(15):1952–1958.
- Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, 25(21):2227–2241.
- Zawel, L., Kumar, K. P., and Reinberg, D. (1995). Recycling of the general transcription factors during RNA polymerase II transcription. *Genes & Development*, 9(12):1479–1490.
- Zeggini, E., Weedon, M. N., Lindgren, C. M., et al. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341.
- Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Research*, 21(8):1273–1283.
- Zhang, Y., Liu, T., Meyer, C. A., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137.
- Zheng, P., Pennacchio, L. A., Le Goff, W., Rubin, E. M., and Smith, J. D. (2004). Identification of a novel enhancer of brain expression near the apoE gene cluster by comparative genomics. *Biochimica et Biophysica Acta*, 1676(1):41–50.
- Zheng-Bradley, X., Rung, J., Parkinson, H., and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11(12):R124.
- Zhu, Y., Sun, L., Chen, Z., et al. (2013). Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Research*, 41(22):10032–10043.
- Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70.

Appendix: Copyright Permissions

RightsLink Printable License

<https://s100.copyright.com/App/PrintableLicense...>

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Jan 12, 2016

This is a License Agreement between Nikolina Nakic ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3786610989430
License date	Jan 12, 2016
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Metazoan promoters: emerging characteristics and insights into transcriptional regulation
Licensed content author	Boris Lenhard, Albin Sandelin and Piero Carninci
Licensed content date	Apr 1, 2012
Volume number	13
Issue number	4
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	2
High-res required	no
Figures	Figure 1., Figure 2.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Genetic and epigenetic regulation of gene expression in pancreatic islets.
Expected completion date	Feb 2016
Estimated size (number of pages)	250
Customer Tax ID	GB649926678
Total	0.00 USD

TERMS AND CONDITIONS

Jun 04, 2016

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

License Number	3882051129167
License date	Jun 04, 2016
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	Structural basis of transcription initiation by RNA polymerase II
Licensed Content Author	Sarah Sainsbury, Carrie Bernecky, Patrick Cramer
Licensed Content Date	Feb 18, 2015
Licensed Content Volume Number	16
Licensed Content Issue Number	3
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	1
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Genetic and epigenetic regulation of gene expression in pancreatic islets.
Expected completion date	Feb 2016
Estimated size (number of pages)	250
Requestor Location	Nikolina Nakic Imperial College, Dpt. Medicine Hammersmith Campus, ICTEM, 5th floor Du Cane Rd Lodnon, W12 0NN United Kingdom Attn: Nikolina Nakic
Customer VAT ID	GB649926678
Billing Type	Invoice
Billing Address	Nikolina Nakic Imperial College, Dpt. Medicine Hammersmith Campus, ICTEM, 5th floor Du Cane Rd Lodnon, United Kingdom W12 0NN Attn: Nikolina Nakic
Total	0.00 GBP
Terms and Conditions	

**NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS**

Jan 12, 2016

This is a License Agreement between Nikolina Nakic ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3786621205161
License date	Jan 12, 2016
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Human genetic variation and its contribution to complex traits
Licensed content author	Kelly A. Frazer, Sarah S. Murray, Nicholas J. Schork and Eric J. Topol
Licensed content date	Apr 1, 2009
Volume number	10
Issue number	4
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 1.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Genetic and epigenetic regulation of gene expression in pancreatic islets.
Expected completion date	Feb 2016
Estimated size (number of pages)	250
Customer Tax ID	GB649926678
Total	0.00 USD
Terms and Conditions	

**NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS**

Jan 12, 2016

This is a License Agreement between Nikolina Nakic ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3786631167072
License date	Jan 12, 2016
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Cancer
Licensed content title	Pancreatic cancer biology and genetics
Licensed content author	Nabeel Bardeesy and Ronald A. DePinho
Licensed content date	Dec 1, 2002
Volume number	2
Issue number	12
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 1.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Genetic and epigenetic regulation of gene expression in pancreatic islets.
Expected completion date	Feb 2016
Estimated size (number of pages)	250
Customer Tax ID	GB649926678
Total	0.00 USD
Terms and Conditions	

**ELSEVIER LICENSE
TERMS AND CONDITIONS**

Jan 12, 2016

This is a License Agreement between Nikolina Nakic ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard, Langford Lane Kidlington, Oxford, OX5 1GB, UK
Registered Company Number	1982084
Customer name	Nikolina Nakic
Customer address	Imperial College, Dpt. Medicine Lodnon, W12 0NN
License number	3786640373265
License date	Jan 12, 2016
Licensed content publisher	Elsevier
Licensed content publication	Trends in Endocrinology & Metabolism
Licensed content title	Revealing transcription factors during human pancreatic β cell development
Licensed content author	Elizabeth Conrad, Roland Stein, Chad S. Hunter
Licensed content date	August 2014
Licensed content volume number	25
Licensed content issue number	8
Number of pages	8
Start Page	407
End Page	414
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No

**NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS**

Jan 12, 2016

This is a License Agreement between Nikolina Nakic ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3786640745834
License date	Jan 12, 2016
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature
Licensed content title	A sequence-based variation map of 8.27 million SNPs in inbred mouse strains
Licensed content author	Kelly A. Frazer, Eleazar Eskin, Hyun Min Kang, Molly A. Bogue, David A. Hinds et al.
Licensed content date	Aug 30, 2007
Volume number	448
Issue number	7157
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 1.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Genetic and epigenetic regulation of gene expression in pancreatic islets.
Expected completion date	Feb 2016
Estimated size (number of pages)	250
Customer Tax ID	GB649926678
Total	0.00 USD