FUNCTIONAL ANALYSIS OF HUMAN ENHANCERS USING THE ZEBRAFISH EMBRYO

by

IRENE MIGUEL ESCALADA

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Department of Medical and Molecular Genetics
School of Clinical and Experimental Medicine
College of Medical and Dental Sciences
The University of Birmingham

June 2014

Abstract

In the post-genomic era the availability of genome-wide datasets has revealed an unexpected complexity of transcriptional regulation. In this context, where most enhancer predictions are based on computational analyses, functional validations are lacking. This thesis investigated the utility of the transgenic zebrafish embryo as an *in vivo* vertebrate model to study the function of candidate human enhancers, and detect subtle changes in enhancer function caused by disease-associated variants. Our functional validations indicated that despite the evolutionary distance between human and fish, 60% of the conserved enhancers predicted by a combination of chromatin signatures, TF binding events and bidirectional transcription, lead to reporter expression that recapitulates the patterns of either zebrafish or human genes. To improve the reliability of zebrafish transgenesis, a targeted integration system mediated by PhiC31 integrase was validated for enhancer testing. I demonstrated that this method overcomes position effect variation commonly found in transposon-based assays. However, enhancer-driven expression could not be detected when I attempted to quantitate *TCF7L2*-associated enhancer variants, indicating the need for further studies to understand the limitations of the zebrafish model. Taken together, my results provide strong support for zebrafish as a valuable *in vivo* model to study the function of mammalian transcriptional regulatory elements.

Acknowledgements

First and foremost I would like to thank my supervisor Professor Ferenc Müller, for his constant support, enthusiasm and help throughout these years, for offering me exciting projects and for trusting me with new collaborations. I also want to thank my co-supervisor, Professor Paul Gissen for his support; and Marie Curie Actions for funding my PhD.

I would like to acknowledge our collaborators Lorenzo Pasquali (IDIBAPS, Spain) and Jorge Ferrer (Imperial College, UK), with whom we have actively worked on all the pancreas-related projects for the last 3 years. Thanks to Robin Andersson and Albin Sandelin (The Bioinformatic Centre, Copenhagen University, Denmark) and the FANTOM5 Consortium for letting us be part of the enhancer-ome project. Thanks to all for sharing their extensive bioinformatic datasets and analyses, and for their enthusiasm for the zebrafish experiments.

Thanks to all the members of the Marie Curie "Biology of Liver and Pancreatic Development and Disease" (BOLD) Initial Training Network for fruitful discussions and wonderful meetings. Thanks to all current and past members of the Müller lab, especially Yavor Hadzhiev and Nan Li, the best post-docs I could have asked for; thanks for being willing to help at any time (literally); Andreas Zaucker, for introducing me to the zebrafish world; Jenny Roberts, Harmeet Gill, Aliesha Griffins and Silvia Parajes for their invaluable help and for being excellent colleagues, and to all of them for being even better friends after hours.

Finally, thanks to my family and friends for their daily words of encouragement. Last but certainly not least, thanks to my partner Nuno, for his constant support and love. Thanks for being my touchstone.

Without all of you this PhD would not have been possible.

Table of Contents

List of figures

List of tables

List of abbreviations

BAC	Bacterial Artificial Chromosome
BSA	Albumin from bovine serum
bp	base pair(s)
BCIP	5-Bromo 4-chloro 3 indolyl phosphate
CAGE	Capped Analysis of Gene Expression
CNE	Conserved Non-coding Element
CFP	Cyan Fluorescent Protein
ChIP	Chromatin Immunoprecipitation
ChIP-Seq	ChIP Sequencing
CNS	Central Nervous System
CPE	Core Promoter Element
CRE	*Cis*-regulatory element
CTCF	CCCTC-binding factor
DHS	DNase I Hypersensitive Sites
dNTPs	deoxynucleotide triphosphates
EPTS-LMPCR	Extension Primer Tag Selection LMPCR
eRNA	Enhancer RNAs
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FISH	Fluorescent In Situ Hybridization
GWAS	Genome-Wide Association Studies
hESC	Human Embryonic Stem Cells
GFP	Green Fluorescent Protein
GTF	General Transcription Factor
HAT	Histone Acetyl Transferase
NBT	Nitro blue tetrazolium
hpf	hours post-fertilisation
kb	kilobase pairs
LB	Luria Bertani
LD	Linkage Disequilibrium
LMPCR	Linker-Mediated PCR
mESC	Mouse Embryonic Stem Cells

NGS	Next Generation Sequencing
NRE	Negative Regulatory Elements
PCR	Polymerase Chain Reaction
PFA	Paraformaldehyde
PGC	Primordial Germ Cell
PIC	Pre Initiation Complex
PTU	N-Phenylthiourea
RNA-Seq	RNA Sequencing
RFP	Red fluorescent protein
RT	Room Temperature
SNP	Single Nucleotide Polymorphism
SSR	Site-specific Recombinases
T2D	Type 2 Diabetes
TAF	TBP-associated Factors
TBE	Tris-borate-EDTA
TBP	TATA-box binding protein
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site
WISH	Whole-mount in situ hybridization
WT	Wild Type
YFP	Yellow Fluorescent Protein
YSL	Yolk Syncytial layer

Chapter One: GENERAL INTRODUCTION

1.1 Transcriptional regulation mediated by RNA Polymerase II: overview

The faithful execution of all biological processes in a cell including proliferation, differentiation and homeostasis, relies on the proper temporal and spatial control of gene expression. In eukaryotes protein-coding genes are transcribed by RNA Polymerase II (aka PolII, Weil 1979), whose activity is regulated through the integrated action of DNA binding proteins, including the general transcriptional machinery, DNA-specific transcription factors, activators and co-activators; and *cis*-regulatory elements, including promoters, enhancers, silencers and insulators, in response to developmental and environmental signals (Levine et al., 2014).

Transcription mediated by PolII is divided into three stages: initiation, when the initiation complex is recruited to the promoter and mRNA synthesis begins, elongation, when the mRNA transcript is extended and termination, when both the transcript and PolII are released. It is traditionally believed that the most stringent regulation of gene expression occurs during transcriptional initiation, although there is a growing body of evidence suggesting that elongation is a limiting step as well (Nechaev and Adelman, 2011).

Transcription is initiated when the highly conserved General or basal transcription factors (GTFs) that include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH bind sequentially to the core promoter to assemble the Pre Initiation Complex (PIC), which will bring RNA PolII to the Transcriptional Start Site (TSS) of the target gene (Orphanides et al., 1996; Roeder, 1996). In the traditional multi-step model of PIC assembly TFIID, a protein complex that comprises the TATA-box binding protein (TBP) and TBP-associated factors (TAFs), binds through its different subunits to core promoter elements. Subsequently TFIIB is recruited, followed by TFIIF-PolII complex and the binding of TFIIE and TFIIH, completing the formation of the PIC. Following PIC assembly TFIIH, which has ATPase, kinase and helicase activities, will unwind the double helix of DNA and position PolII polymerase to start transcription (Goodrich and Tjian, 1994; Douziech et al., 2000). This early-elongating PolII will clear the core promoter facilitated by the phosphorylation of Serine-5 residue on its C-terminal domain by a kinase subunit of TFIIH (Komarnitsky et al., 2000; Hirose and Ohkuma, 2007). When the nascent mRNA transcript reaches 25-50 nucleotides, most of the GTFs are released, the elongation complex is established and PolII pauses (Rougvie and Lis, 1988; Marshall and Price, 1992; Liu et al., 2013). Phosphorylation of the Ser2 residue of PolII by P-TEFb kinase favours processive transcription and is required to recruit factors that are important for transcription elongation, termination and mRNA processing (Marshall and Price, 1995; Marshall et al., 1996; Ni et al., 2004; Hirose and Ohkuma, 2007). After this, the Polymerase is released and ready to reinitiate another transcription cycle.

This textbook view of transcription, which portrays a universal and common mechanism of transcriptional initiation where PIC and the core promoter are static elements, has been challenged at different levels over the past years: i) even though GTFs were identified in TATA-box containing genes, they are actually a minority and core promoter elements are diverse (Cooper et al., 2006; Juven-Gershon et al., 2006; Kadonaga, 2012); ii) there are alternative functional TBP-related factors in metazoa that can recognize TATA-box element (Dantonel et al., 1999); iii) apart from TBP some other TAFs, TFIIB and variant paralogs can generate alternative promoter recognition complexes (Verrijzer et al., 1995; Burke and Kadonaga, 1996; Burke and Kadonaga, 1997; Lagrange et al., 1998); iv) TFIID is not ubiquitous, in fact there are tissue-specific subunits of TAFs in tissues like testis and ovary (Freiman et al., 2001; Pointud et al., 2003; Falender et al., 2005); v) during muscle differentiation TFIID complex is rapidly degraded and replaced by a novel TRF3/TAF3 complex (Deato and Tjian, 2007). The existence of alternative initiation complexes that can assemble at distinct sets of promoters during different developmental times suggests an unexpected complexity of transcriptional initiation (Goodrich and Tjian, 2010; Muller et al., 2010; Muller and Tora, 2014). Several aspects of such diversity will be discussed below.

1.2 *Cis*-Regulatory Elements

There are two main classes of *cis*-regulatory elements controlling genes transcribed by PolII: proximal elements, including the core promoter and proximal promoter and distal regulatory regions, including enhancers, silencers and insulators **(Figure 1.1)**. All these *cis*-regulatory elements contain binding sites for *trans*-regulatory proteins (general transcription factors, activators or co-activators (Maston et al., 2006).

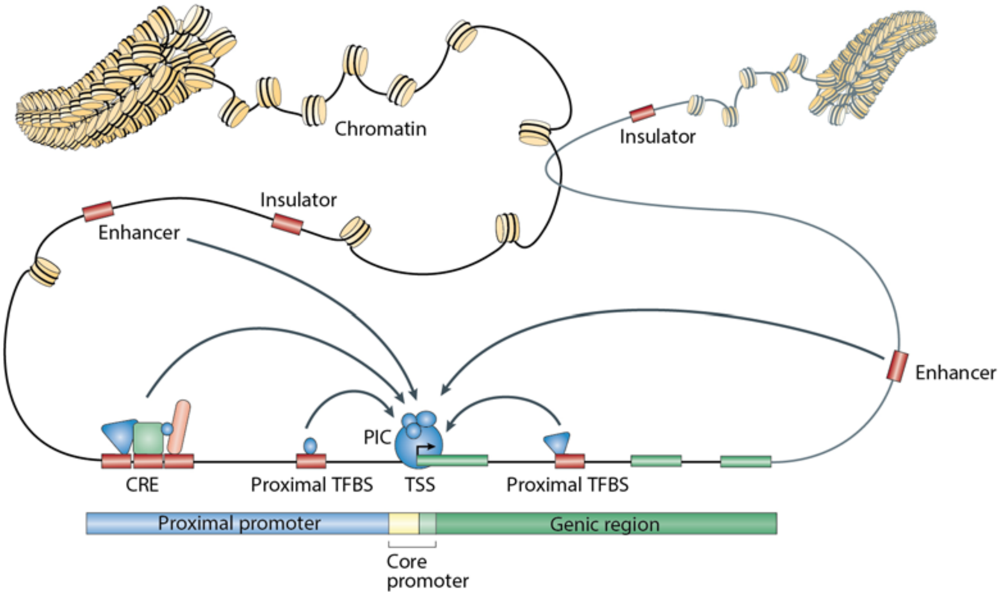


Figure 1.1 Overview of transcriptional regulation.
Schematic representation of proximal and distal *cis*-regulatory elements (CRE) and the integration of regulatory input in the core promoter region, encompassing the TSS of a gene (black arrow). Transcription factors bind to CREs by transcription factor binding sites (TFBS). Reproduced from (Lenhard et al., 2012).

1.2.1 Core promoters

The core promoter is a region in the immediate vicinity of the TSS of a gene, which contains multiple motifs, or core promoter elements (CPEs), for docking the PIC (Sandelin et al., 2007).

However, the core-promoter does not only serve this role, it integrates the total regulatory input from transcription factors (TF) binding to proximal and distal elements such as enhancer or repressors, and translates it into a refined and regulated rate of transcriptional initiation (Lenhard et al., 2012). Core promoter elements are diverse: there is no element present in all promoters and most likely more CPEs remain yet to be discovered (**Figure 1.2**, reviewed in (Kadonaga, 2012)).

- The TATA-box was the first element to be discovered and is the best characterized position-dependent core promoter motif. It is located about 30 nt upstream of the TSS and it is predominantly bound by the TBP (TATA-box Binding Protein) subunit of the TFIID complex (Lifton et al., 1978). Both the TATA box and TBP are conserved from archaebacteria to humans. However, TATA-containing promoters are a minority compared to TATA-less ones (Juven-Gershon et al., 2006). This element can also be recognised in metazoans by other members of the TBP family including TRF1, specific to *Drosophila*, TRF3, described in vertebrates and TBP2, described in zebrafish (Crowley et al., 1993; Hansen et al., 1997; Dantonel et al., 1999; Persengiev et al., 2003; Bartfai et al., 2004).

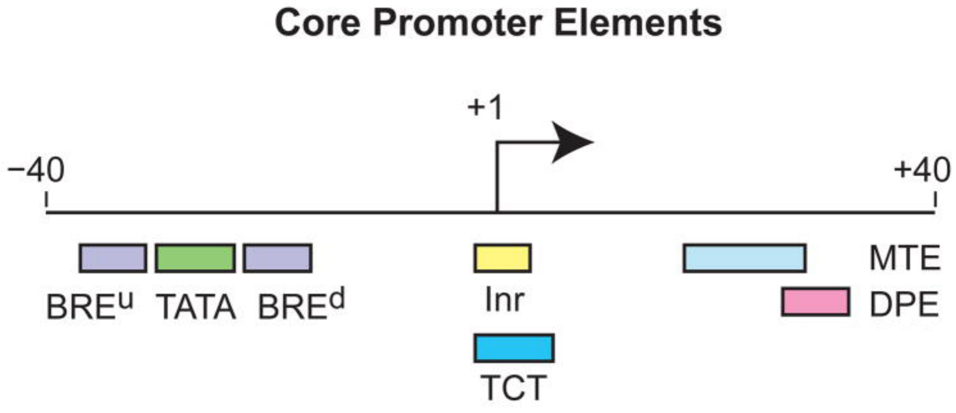


Figure 1.2 Relative positioning of CPEs with roles in PolII-mediated transcription.
All elements depicted in the image are approximately drawn to scale and their position relative to the dominant TSS of the gene (black arrow, +1). Importantly, promoters only exhibit a subset of these CPEs and there is no universal CPE present in all promoters. Adapted from (Kadonaga, 2012)

- The Initiator motif is a pyrimidine-rich 17 bp element that encompasses the TSS, thus located at +1, and can be found both in TATA-containing or TATA-less promoters (Smale and Baltimore, 1989). It can be recognized by the TAFI and TAFII subunits of the TFIID complex (Chalkley and Verrijzer, 1999).
- The DPE (or Downstream Promoter Element) is a 7 bp motif highly conserved from Drosophila to humans located 30 nt downstream of TSS, mostly in TATA-less promoters. It is bound by TBP-associated factor TAF6 and TAF9 subunits of TFIID, but not by TBP (Burke and Kadonaga, 1996; Burke and Kadonaga, 1997).
- The MTE (or Motif Ten Element) is located between +18 and +27 nt relative to the initiator, immediately downstream of the DPE. It can promote transcription independently of the TATA box and DPE but requires the initiator motif for its function (Lim et al., 2004).
- BREs (or TFIIB Recognition Elements) are 7 bp elements located immediately upstream (BREu) or downstream (BREd) of the TATA element and act in conjunction with it. BRE motifs are recognised by TFIIB and it is believed that binding stabilizes the TFIID-DNA complex (Lagrange et al., 1998).
- The TCT or polypyrimidine initiator is a recently described novel CPE that resembles the initiator motif and is specifically required for the transcription of ribosomal genes in Drosophila and mammals. It spans from -2 to +6, encompassing the TSS (Parry et al., 2010).

1.2.1.1 Classification of core promoters

The development of high-throughput technologies that permits mapping of the 5' end of mRNAs with single base pair resolution, such as CAGE (Shiraki et al., 2003), has shown that metazoan promoters are heterogeneous and exhibit a differential architecture. Based on the TSS distribution, core promoters in mammals can be divided into two classes (Carninci et al., 2006). A minority of promoters is considered “sharp”, as they have a single TSS. Sharp promoters often contain a TATA box and are associated with tissue-specific expression. On the contrary, “broad” promoters typically contain a cluster of TSS and can start transcription in a range of 100 bp, generating mRNAs of different lengths but with the same coding content. These promoters are TATA-less, enriched in CpG islands, and usually correlate with ubiquitous or developmentally regulated expressed genes (Sandelin et al., 2007). These observations have recently been confirmed by the FANTOM5 Consortium, where 573 human and 128 mouse primary cell samples were analysed by single molecule CAGE (FANTOM Consortium, 2014). Additionally, due to the high depth coverage of the sequencing data, they were also able to determine the preferred TSS in broad promoters (FANTOM Consortium, 2014). A global description of TSS usage during the development of the zebrafish embryo has unveiled that zebrafish also contains "sharp" and "broad" promoters and interestingly, the usage of broad promoters increased after zygotic gene activation (Nepal et al., 2013).

This dichotomy of promoters has been challenged since it has been shown in mammals that some "sharp" promoters overlap CpG islands and some "broad" promoters are not enriched in CpG nucleotides (Ponjavic et al., 2006). It has been proposed therefore to use a tripartition to distinguish mammalian promoter types (Rach et al., 2011; Lenhard et al., 2012). Type I promoters are TATA-enriched, have a low CpG content, and are associated with tissue-specific expression of genes. Type II promoters usually contain a short CpG island overlapping with the TSS and correlate with ubiquitously expressed genes. Type III promoters contain CpG islands that extend onto the gene body and are characteristic of developmentally regulated genes (Lenhard et al., 2012).

1.2.2 Proximal promoter

The proximal promoter is the region extending upstream of the core promoter (generally up to 200 and 250 nt upstream of the TSS) and contains multiple binding sites for activators (Butler and Kadonaga, 2002). Proximal promoters might also contain tethering elements to recruit specific distal enhancers through homotypic binding of transcription factors (Calhoun et al., 2002).

Functional dissection of the proximal promoter area has revealed that the region from -300 to -50 relative to the TSS contributes positively to core promoter activity, while there is often a negative regulatory region from -1000 to -500 (Cooper et al., 2006).

1.2.3 Silencers

Silencers are *cis*-regulatory elements that act by silencing or repressing the transcription of a gene. They were originally identified in yeast, and defined as elements that could switch off heterologous promoters in an orientation-independent manner and with relative independence to the position of the target promoter (Brand et al., 1985). These classical silencer elements are located upstream of the TSS and are bound by repressors that can interfere with PIC assembly (Maston et al., 2006). Examples of classical silencers include the distal element of the *zen* gene in *Drosophila.* It is located between -1.4 Kb and -1.2 Kb of the TSS and can actively repress transcription of the gene in the ventral region of the embryo (Doyle et al., 1989) through binding of the DSP1 repressor, which recruits TBP, displacing TFIIA and impeding PIC formation (Kirov et al., 1996). There is a second class of silencers that are also known as Negative Regulatory Elements (NRE). These silencers are position dependent and can passively repress transcriptional regulation by preventing the binding of TFs to their regulatory motifs, or by interfering with signals controlling splicing, polyadenylation or elongation (Ogbourne and Antalis, 1998).

1.2.4 Insulators

Insulators are boundary elements that limit the action of regulatory modules to certain genomic domains, protecting genes from neighbouring effects (Maston et al., 2006). They are usually short pieces of DNA between 0,5-3Kb in size that act in an orientation-independent but position dependent manner, reviewed in (Heger and Wiehe, 2014). They have two main properties: they can directionally block enhancer-promoter interactions, i.e. only if the insulator is located between them (Geyer and Corces, 1992); or they can prevent or limit the spread of heterochromatin, acting as a barrier (Sun and Elgin, 1999). There are several DNA sequences that can function as insulators in the *Drosophila* genome. The first insulators to be identified were scs and scs´ flanking the proximal and distal ends of the *hsp70* locus (Udvardy et al., 1985), which are bound by Zw5 and BEAF proteins respectively (Hart et al., 1997; Gaszner et al., 1999). Another well-known enhancer-blocking insulator is *gypsy*, a retrotransposable element responsible for multiple mutations in *Drosophila*, whose interaction with su(Hw) protein is essential for its function (Parkhurst et al., 1988; Geyer and Corces, 1992; Kuhn et al., 2003). However, in vertebrates all enhancer-blocking insulators identified so far contain binding sites for CCCTC-binding factor (CTCF) (Heger and Wiehe, 2014). CTCF is an 11-zinc finger DNA binding protein, which is highly conserved in vertebrates (Filippova et al., 1996). Computational prediction of CTCF binding sites has revealed that they are spread throughout the genome and particularly enriched in conserved non-coding regions (Xie et al., 2007). Furthermore, CTCF-binding sites showing enhancer-blocking properties *in vivo* are evolutionarily conserved and syntenic between human and chicken (Martin et al., 2011). These sites act as barriers that define regulatory domains and flank transcription factor encoding genes, whose disruption could potentially affect development and lead to disease. Consistent with this idea, CTCF was also found enriched in the boundaries of the "topological domains" defined by genome-wide maps of tridimensional interactions (see **section 1.3.4**, (Dixon et al., 2012)).

1.3 Enhancers

Enhancers were originally described thirty years ago as regulatory elements acting in *cis*, capable of enhancing transcription in an orientation and distance independent manner (Banerji et al., 1981). Enhancers can regulate gene expression with high spatio-temporal precision and are essential for vertebrate development. They serve as platforms for the binding of multiple TFs through 6-12 bp binding sites (Istrail and Davidson, 2005). When appropriate occupancy of TFs occurs, transcriptional co-factors and chromatin remodelers are recruited, which then facilitates enhancer-promoter interaction, and ultimately activates transcription (Visel et al., 2009b). Enhancers are commonly found in intergenic regions or in the introns or exons of the genes they regulate or of neighbouring genes (Abbasi et al., 2007; Birnbaum et al., 2012; Ritter et al., 2012; Sanyal et al., 2012). Enhancers can also be clustered forming "super-enhancers" with crucial roles in regulating cell identity (Hnisz et al., 2013; Loven et al., 2013; Whyte et al., 2013) or spread at high frequencies along Genomic Regulatory Blocks (GRB), i.e. syntenic chromosomal segments conserved among vertebrates that contain their developmental target genes and unrelated "bystander genes" (Kikuta et al., 2007; Akalin et al., 2009). Enhancers can be very far away from their target gene: an extreme example is the limb enhancer of *Shh*, which resides 1 Mb away in the fifth intron of *Lmbr1* (Lettice et al., 2003), highlighting the difficulty in predicting the distance of regulatory interactions and the challenges in assigning a target gene to an annotated enhancer.

Approximately 80,000 putative enhancers have been identified in the human genome using multiple cell lines and a combination of genome-wide methods including DNase I hypersensitivity, TF binding and chromatin marks (Bernstein et al., 2012; Thurman et al., 2012). The importance of annotating enhancers has been underscored by the fact that enhancers have been shown to contribute to disease and that SNPs linked to human disorders by GWAS are enriched in non-coding functional elements (Bernstein et al., 2012).

1.3.1 Modularity and redundancy of enhancers

An important characteristic of enhancers is their modular nature: complex patterns of gene expression across tissues result from the combined activity of multiple independent elements, each contributing to a sub-pattern of the complete activity (Visel et al., 2007). The apolipoprotein E (*APOE*) gene serves as a good example to explain this property. Human *APOE* gene is located in a gene cluster on chromosome 19 spanning 45 Kb. It encodes for a structural protein that is a major component of mammalian lipoproteins, whose role is the re-distribution of cholesterol between the liver and peripheral tissues (Simonet et al., 1991). Tissue specificity is controlled by multiple independent *cis*-regulatory elements: a proximal region is responsible for expression in the kidney (Simonet et al., 1991), two enhancers located a few kb downstream the TSS drive expression in the liver (Simonet et al., 1993; Allan et al., 1995), one enhancer in an intergenic region is active in the skin (Simonet et al., 1991), two distal enhancers specify expression in macrophages and adipocytes (Shih et al., 2000), two distal enhancers control the production of apoE in astrocytes (Grehan et al., 2001) and a highly conserved enhancer was demonstrated to act in the brain (Zheng et al., 2004).

Certain genes can also be regulated by several "redundant enhancers" driving very similar or overlapping expression patterns. This is the case of enhancers regulating *Shh* expression in mouse: while expression in the midbrain and diencephalon areas is controlled by unique elements, redundant enhancers regulate *Shh* expression in hindbrain, ventral spinal cord and telencephalic regions (Jeong et al., 2006). Similarly, a study mapping the occupancy of several TFs involved in dorso-ventral patterning in *Drosophila* embryos revealed two enhancers regulating the *vnd* locus, which were several kb away and directed very similar spatio-temporal expression patterns (Zeitlinger et al., 2007). This observation was extended to approximately half of the genes involved in dorso-ventral patterning in *Drosophila*, where a distant secondary enhancer directing overlapping activities with the primary enhancer was termed "shadow enhancer", in contrast to the primary enhancer that was close to the TSS of the target gene (Hong et al., 2008). It has been suggested that these apparently "redundant" enhancers confer phenotypic robustness against environmental and genetic fluctuations, thus although they might have a minimal contribution under laboratory conditions, they are maintained over long periods of time (Frankel et al., 2010).

Enhancers can also function in synergy. Prominent examples are the LT and M enhancers of *Dll*, which are involved in leg patterning in *Drosophila* (Estella et al., 2008). Neither element on its own can activate *Dll* expression pattern in the leg discs, but combined in a single construct, they recapitulate *Dll* activity, suggesting a synergistic mode of action (Estella et al., 2008). Similarly, the proximal and distal early stripe elements (PESE and DESE) in *Drosophila* on their own direct an incomplete activity but when combined in a composite construct they can fully recapitulate the pattern of their target gene *slp1* (Prazak et al., 2010).

1.3.2 Transcription factors and enhancer function

Transcription factors are sequence specific DNA-binding proteins that bind to CREs by recognising small degenerate DNA sequences that will ultimately bring the transcriptional apparatus to the TSS of the gene leading to gene-specific transcriptional activation, as reviewed in (Levine, 2010; Todeschini et al., 2014). There are around 1,500 TFs in the human genome grouped into families that share similar properties (Vaquerizas et al., 2009). TFs have a modular structure typically composed of a DNA binding domain linked to one or more activating or repressing modules (Triezenberg, 1995; Ptashne and Gann, 1997) that can establish contacts with chromatin remodelers, histone modifiers or the basal transcriptional apparatus (Lee and Young, 2000).

TFs bound to enhancers mediate the recruitment of a wide range of co-activator complexes (Kadonaga, 2004; Taatjes et al., 2004). Among these co-regulators there are complexes that serve as a "bridge" between the basal transcriptional machinery and the activators and facilitate the assembly of the PIC, such as the evolutionarily conserved Mediator (Flanagan et al., 1991; Malik and Roeder, 2010). Other co-regulators include chromatin remodeling complexes that mediate nucleosome eviction, such as the ATP-dependent SWI-SNI complex (Schwabish and Struhl, 2007); and enzymes that catalyze covalent histone modifications, such as p300/CBP or SAGA histone-acetyl transferases (**Figure 1.3**).

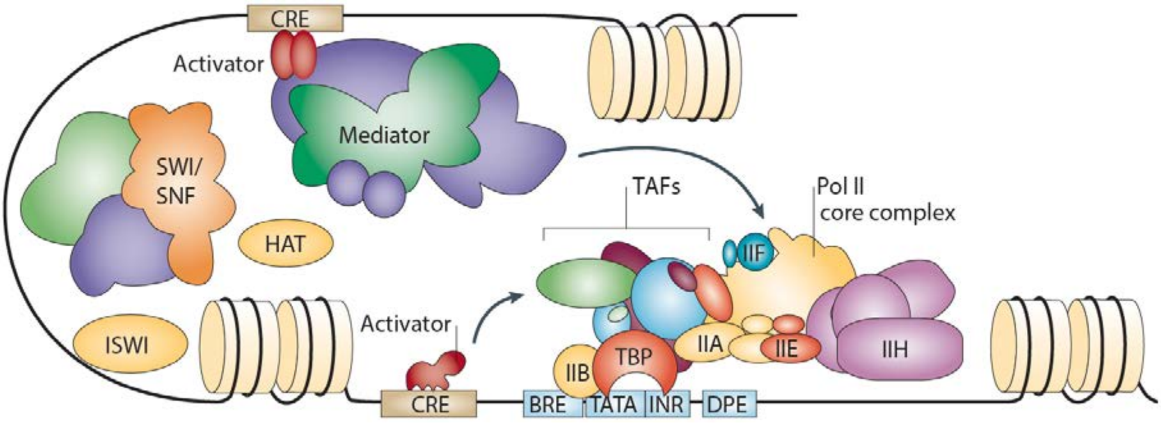


Figure 1.3 Transcriptional initiation mediated by co-factors and PIC
Activators assembled at distal enhancers (CRE) can recruit co-activator complexes, including Mediator, chromatin remodelling complexes (SWI/SNF or ISWI) and histone acetyl transferases (HAT). Together they create an accessible chromatin environment for the assembly and binding of PIC and PolII to the core promoter elements. Adapted from (Taatjes et al., 2004).

Several studies using FISH (Fluorescent In Situ Hybridization) techniques and electron microscopy have suggested that active PolII and nascent mRNA might be concentrated at nuclear foci termed "transcription factories", where many units of PolII are pre-assembled and anchored, as reviewed in (Sutherland and Bickmore, 2009). This model suggests it is the loci that move to the factories, rather than the other way around (Iborra et al., 1996), which is consistent with the idea that active distal foci physically co-localize during transcription (Osborne et al., 2004).

1.3.2.1 Binding of sequence-specific TFs to enhancers

Enhancers need to integrate information coming from signalling cascades and TFs present in the cellular context and respond accordingly. TFs bind to enhancers by recognising small degenerate DNA sequences that usually cluster within them, creating a combinatorial code that results in tissue-specific expression patterns (Dynan, 1989), but how is this binding regulated?

Several studies have demonstrated that TFs bind to hundreds of TFBS in a certain cell at a certain time (Walter et al., 1994; Farnham, 2009; Fisher et al., 2012). It has been proposed that the large size of eukaryote genomes coupled with the degenerate nature of binding sites leads to widespread non-functional unspecific binding at several hundred binding sites at a time (Wunderlich and Mirny, 2009). Functional specificity is then achieved by clustering several binding sites, for example within an enhancer. We refer to "additive binding" when the transcriptional response of the enhancer is directly correlated with the concentration of the TFs, as demonstrated for genes regulated by NF-κB (Giorgetti et al., 2010). When increasing concentrations of TFs translates into an on/off binary response we refer to "cooperative binding", which is facilitated by protein-protein interactions between TFs bound adjacently or by another TF already bound to the DNA (Adams and Workman, 1995; Senger et al., 2004). This type of binding is common during development, leading to sharp expression patterns, such as with Bicoid target gene regulation in *Drosophila* (Lebrecht et al., 2005).

1.3.2.2 Enhancer architecture: models of TF recruitment

The orientation, order, and spacing of TFBS are often referred to as the enhancer's "grammar" or enhancer architecture **(Figure 1.4)**. Depending on the enhancer´s grammar three types of TF recruitment can be distinguished.

The "enhanceosome" model of action assumes that only the perfect formation of a TF complex following a rigid grammar would activate transcription of the target gene, generating a binary on/off type of response (Merika and Thanos, 2001). It entails the formation of a very stable nucleoprotein complex (enhanceosome), where specific binding of multiple TFs to the enhancer region through protein-protein interactions, triggers a transcriptional response. One of the best examples is the case of the interferon β (*INF-β*) gene (Thanos and Maniatis, 1995). *INF-β* is normally silenced, but following a viral infection it can be induced at very high levels. This induction is caused by an enhancer located between -110 and -45 bp relative to the TSS of the gene, which contains, among others, 4 binding sites for the high mobility group protein HMG I(Y). Upon binding of HMG I(Y) to the DNA, and through multiple protein-protein interactions requiring a very strict spacing of binding sites, the "enhanceosome" is assembled (Thanos and Maniatis, 1992; Thanos and Maniatis, 1995; Merika and Thanos, 2001).

The "billboard" or "TF display" model implies an inherent flexibility of the enhancer where the TFs are more flexibly disposed and the bound TFs do not function as a single unit (Arnosti and Kulkarni, 2005). This model was exemplified by an experiment performed *in vivo* in *Drosophila* embryos, in which compact constructs containing different combinations of activators and repressors linked to a reporter gene were built. It was shown that the basal transcriptional

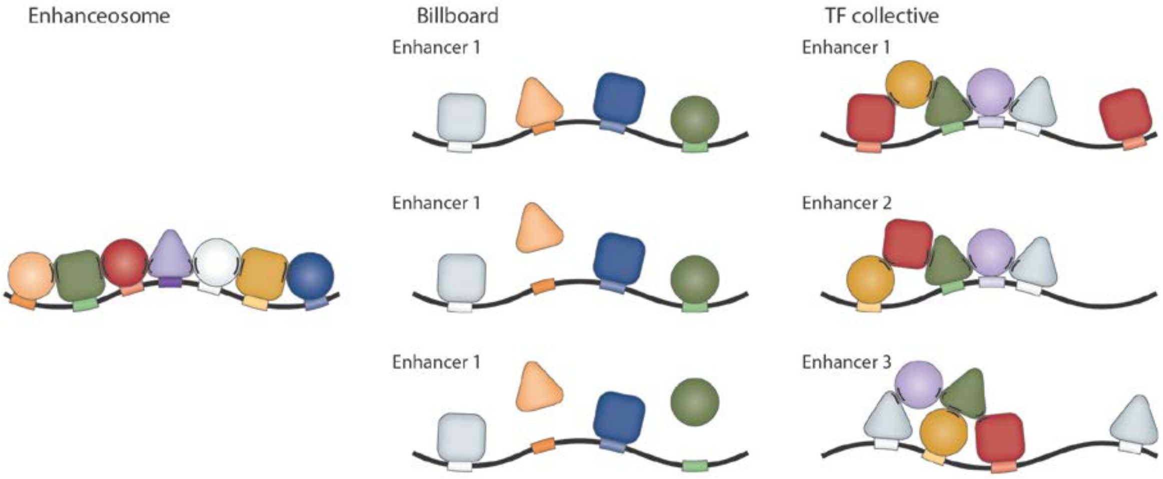


Figure 1.4 Current models of enhancer action.
In the enhanceosome model the DNA acts as a scaffold for the binding of several transcription factors that operate as a single unit. Disruption of a single binding site renders the enhancer inactive. In the billboard model the architecture is more flexible; the enhancer contains multiple binding sites that can interact independently with their targets and activate gene expression. The TF collective mode of action implies that subset of TFs can act in different enhancers, occupying each in a different manner. Adapted from (Spitz and Furlong, 2012).

machinery "samples" activating and repressing binding sites within the enhancer, which would ultimately dictate the transcriptional output (Kulkarni and Arnosti, 2003). This model allows for a more flexible order of occupancy of TFBS, where TF can bind cooperatively or additively (Spitz and Furlong, 2012).

Contrary to these models, the "TF collective" model, described for a set of 5 cardiac TFs in *Drosophila* (Junion et al., 2012), proposes that certain enhancers function by recruiting several transcription factors as a collective unit in the absence of grammar, to the extent that not all the TFBS need to be present. This model proposes that some TFs bind with high-affinity to their sites while the others are recruited through protein-protein interactions or through co-factors such as CBP/p300 (Junion et al., 2012).

1.3.2.3 Pioneer factors during development

Chromatin environment also determines the efficiency of binding of TFs to their cognate TFBS. While most TFs cannot access their binding sites in compacted chromatin, "pioneer" factors are a special class that can access TFBS at developmental enhancers by actively mediating chromatin decompaction or passively acting as a landmark to recruit additional TFs (Zaret and Carroll, 2011). Members of the Forkhead Box (Fox) and GATA families of TFs are considered "pioneer" factors since they are thought to scan the chromatin fibre and bind to regulatory elements before there is active gene expression or lineage commitment (Bossard and Zaret, 1998; Zaret, 1999).

The albumin gene enhancer *Alb1* in mouse liver has been extensively studied in this regard. In non-liver tissues where the albumin gene is silent, the enhancer is not occupied by TFs and the chromatin is highly compact with nucleosomes randomly positioned over the enhancer (McPherson et al., 1993). However, in liver precursor cells, binding of FOXA1 and GATA4 mediates chromatin decompaction and creates a new region of DNase hypersensitivity (McPherson et al., 1993; Cirillo et al., 2002). This renders the enhancer active and competent for further binding, which is essential to trigger liver developmental program (Lee et al., 2005). FOX factors have been shown to be more efficient than GATA factors in binding condensed chromatin (Cirillo et al., 2002). It is thought that the structural similarity between their "winged-helix" DNA binding domain and the globular domain of linker histone H1 might explain this ability, since it allows them to bind simultaneously to the minor and major groove of DNA while still being able to recruit TFs on the other side (Clark et al., 1993; Zaret et al., 2010).

1.3.3 Enhancer-promoter specificity

Enhancers can be very far away from their target promoters, as reviewed in (Krivega and Dean, 2012), and scattered in large syntenic chromosomal regions containing "bystander genes" that they are not regulating (Kikuta et al., 2007), however, enhancers appear to selectively interact with core promoters.

Early observations of lack of enhancer-promoter specificity emerged from studies in *Drosophila*, where enhancers could not interact with heterologous promoters of non-cognate genes (Li and Noll, 1994; Merli et al., 1996). In the case of neighbouring *gsb* and *gsbn* genes that are divergently transcribed, enhancers located in a common upstream region could not activate non-target genes because of an incompatible interaction (Li and Noll, 1994) that was unrelated to the presence of insulator elements (Merli et al., 1996). It was soon demonstrated that enhancer-promoter-specificity was greatly determined by the presence of different CPEs (Ogbourne and Antalis, 1998; Butler and Kadonaga, 2001). Ohtsuki and colleagues used transgenesis assays in *Drosophila* to test the effect of TATA-box CPE in heterologous promoters and its capacity to interact with various developmental enhancers. They showed that AE1 and IAB5 enhancers preferentially interact with a TATA-containing promoter compared with a TATA-less one, while the enhancer NEE does not discriminate(Ohtsuki et al., 1998). Furthermore, Butler and Kadonaga, used transgenic lines containing DPE or TATA promoters in the same chromosomal location to demonstrate *in vivo* that transcriptional enhancers are specific for promoters and that the specificity depends, at least partly, on CPE composition (Butler and Kadonaga, 2001). Similarly, Caudal TF was shown to activate transcription in Hox genes with a higher preference for DPE-containing CREs, relative to TATA-box (Juven-Gershon et al., 2008).

In zebrafish, a high-throughput analysis of 202 enhancer-promoter combinations coupled with automated imaging software that could annotate tissue-specific reporter activity showed that promoters interact with enhancers with variable degrees of efficiency (Gehrig et al., 2009). This analysis identified a subset of heterologous promoters (*krt4*, *hsp70* and *eng2b*) with a broad ability to interact with enhancers, which provides a useful resource for transgenesis studies in zebrafish and underlines the importance of enhancer-promoter specificity.

1.3.4 Enhancer-promoter interactions

Packaging DNA inside the nucleus imposes tremendous organisational challenges. Furthermore, given the evidence supporting long-range promoter-enhancer interactions, it is expected that tridimensional nuclear architecture is crucial for the regulation of gene expression in the cell. The mechanism by which enhancers establish a selective interaction with their target promoters is still under debate. The tracking or scanning theory suggests that activators bound to enhancers can move along the chromatin fibre until they find their target promoter (Blackwood and Kadonaga, 1998). The looping theory postulates a direct interaction between enhancers and promoters through the looping of the DNA strand (Rippe et al., 1995). Specific protein-protein interactions between the activators that bind the enhancer and the transcription factors binding the promoter ensure the activation of the correct gene. These interactions have been shown experimentally using chromosome conformation capture technologies (3C) and its derivatives 4C (circularized 3C) and 5C (carbon-copy 3C) (reviewed in (de Wit and de Laat, 2012), **Figure 1.5**).

3C-derived technologies use fixation with formaldehyde, so that all potentially distant physical contacts between DNA and proteins are cross-linked. While 3C Technology quantifies the frequency of interactions between 2 loci (Dekker et al., 2002), Chromosome conformation Capture-on-Chip or 4C technology is known as the "one versus all" strategy, since this method screens the genome for all the sequences that interact with a selected genomic site (Simonis et al., 2006).

Application of this methodology has shown how the β-globin locus interacts with a completely different set of genomic sites in a tissue where it is active (foetal liver) versus a tissue where it is inactive (brain), and how interactions are preferentially established with transcriptionally active loci (Simonis et al., 2006). On the contrary, when examining a housekeeping gene in the same tissues, the contacts are very similar, suggesting a differential nuclear architecture between active and inactive chromatin (de Laat and Duboule, 2013).

An interesting question arising from these studies is how local and long-range interactions are established and whether transcription itself could help to re-shape the genome. It appears that in mammals, the DNA contacts established by the β-globin locus are not changed after drug-induced transcription inhibition (Palstra et al., 2008), suggesting that transcription is unnecessary for the formation of enhancer-promoter loops. However, loops are necessary for transcription, as demonstrated by experiments with engineered enhancer-promoter loops in the β-globin locus (Deng et al., 2012).

5C Technology or Chromosome Conformation Capture Carbon Copy is a 3C-based method that allows to study interactions occurring between multiple loci, that is why it is also known as the "many to many "strategy (de Wit and de Laat, 2012). It requires the creation of a 3C library, to which oligonucleotides are ligated, and multiplex amplification; followed by analysis using microarrays or quantitative PCR (Dostie et al., 2006). The main limitation of this technique is the high number of oligos that would be needed to evaluate the entire genome (Dostie and Bickmore, 2012). Hi-C technology overcomes this limitation, since it measures three-dimensional interactions of whole genomes ("all to all" strategy), at a resolution of 1 Mb in mammalian systems (Lieberman-Aiden et al., 2009)

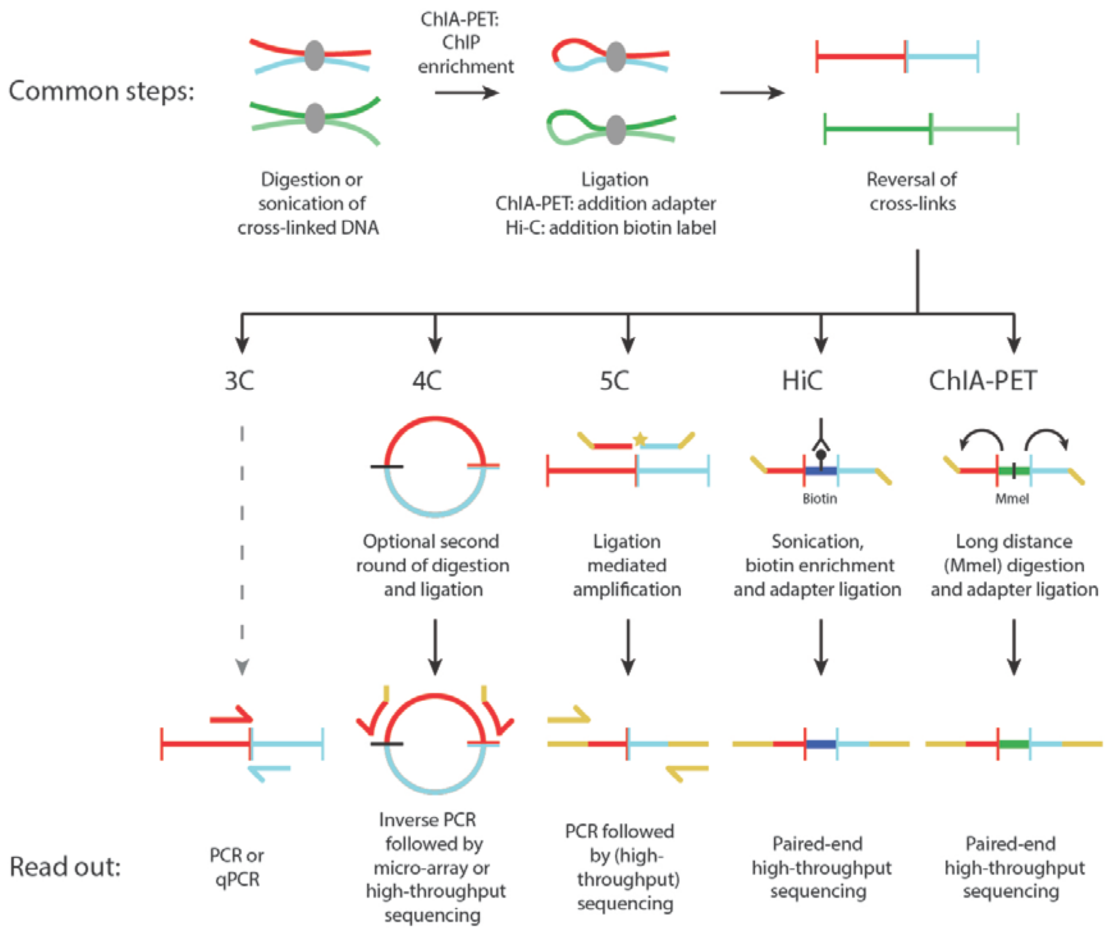


Figure 1.5 Overview of 3C-derived methodologies.
Top panel indicates processes that are common for all technologies (crosslinking of chromatin, digestion, ligation and reverse crosslinking); while the vertical panels show the specific steps for each technique. Reproduced from (Noordermeer and Duboule, 2013).

The analysis of the human and mouse genomes using Hi-C showed that it is pervasively segmented in megabase-size portions named "topological domains" (Dixon et al., 2012). Such organization seems to be a property of the genomes, since it is very well conserved between human and mouse, and stable between cell types, although the regions within each domain can be dynamic, potentially representing cell-type specific regulatory events. High levels of transcription might be involved in the creation of topological boundaries, since they are enriched in CTCF insulator, housekeeping genes and TSS, while enhancer or chromatin marks (H3K4me1 or H3K9me3) were depleted. (Dixon et al., 2012). A study focusing on X-chromosome inactivation in *Drosophila* embryos also showed a similar partitioning of the genome into discrete units at the sub-megabase scale, which they termed "Topologically Associated Domains" or TADs, the disruption of which results in long-range misregulation (Nora et al., 2012).

ChIA-PET (or chromatin interaction analysis with paired-end tag sequencing) has been introduced as a technique that combines chromatin immunoprecipitation (ChIP) with 3C analysis, enabling the unbiased study of genome-wide long-range chromatin interactions bound to a certain protein or histone modification mark (Fullwood et al., 2009; Li et al., 2010a). The main limitation is that ChIA-PET can only analyse contacts bound by the selected factor.

Overall, the chromosome capture technologies indicate that a physical contact is not equivalent to a functional interaction. It will be interesting to couple 3C-based technologies with high-resolution single-cell live imaging, to measure cell-to-cell variations and investigate whether there is a connection between function and structure.

1.4 Genome-wide strategies for enhancer prediction

Contrary to promoters, which can be located by sequencing the 5' end of its mRNA, the lack of distinguishing features in enhancers makes them inherently difficult to identify. Traditionally, non-coding regulatory regions were identified using arduous serial deletion assays (promoter bashing), whereby the function of the resulting fragments was tested in reporter gene assays (Arnosti et al., 1996; Muller et al., 1999; Heckman et al., 2003). In the post-genomic era, genome-wide computational and experimental strategies for enhancer prediction have used various analyses of the DNA sequence and chromatin structure, as reviewed in (Wang et al., 2013). International Research Consortia including the FANTOM and ENCODE Projects set out to assign a function to all the elements in the mouse and human genomes, respectively. In parallel, the modENCODE project's aim was to annotate the genomes of two model organisms: *Drosophila melanogaster* and *Caenorhabditis elegans*. The release of the ENCODE data, with a pilot phase covering 1% of the genome (Birney et al., 2007) and a final report consisting of more than 40 papers in 2012, has unloaded an impressive amount of data on distinct functional elements and shed light on strategies for CRE prediction.

1.4.1 Comparative genomics as a tool for the identification of non-coding functional elements

The comparison of sequences across different species with the aim of identifying conserved elements that could be potential regulatory regions has been termed phylogenetic footprinting (Zhang and Gerstein, 2003). Cross-species sequence comparison methods rely on the principle that functionally relevant sequences are under purifying selection, i.e. conserved across long evolutionary periods, whereas non-functional regions evolve neutrally and eventually diverge beyond recognition (Ahituv et al., 2004). Interestingly non-coding conserved elements tend to cluster around genes coding for transcription factors involved in embryonic developmental processes shared by vertebrates, suggesting that they are CREs controlling gene expression during development (Sandelin et al., 2004; Woolfe et al., 2005)

Selecting the species to be used in comparative genomics usually represents a compromise. Comparing closely related species often obscures the identification of functional elements, due to the high degree of similarities between the two genomes. On the other hand, comparing distantly related species (e.g. human with a non-primate mammal or with a non-mammal vertebrate) might impair the discovery of lineage-specific elements; or elements might not be readily identifiable because they will have diverged too much (Boffelli et al., 2004; Nobrega and Pennacchio, 2004). In order to overcome these limitations and balance evolutionary distances, multiple species comparisons have been used to identify functional non-coding regions (Dubchak et al., 2000; Frazer et al., 2001; Gottgens et al., 2002; Nobrega et al., 2003; Santagati et al., 2003; Dermitzakis et al., 2004).

Identification of human enhancers has been performed based solely on human-mouse sequence comparisons (Hardison et al., 1997; Wasserman et al., 2000; Dermitzakis et al., 2002; Patwardhan et al., 2012). Even though alignment of the human and mouse genomes, which diverged around 75 million years ago, revealed a similarity of almost 40% at the nucleotide level (Schwartz et al., 2003), their divergence is sufficient to identify functional elements.

Distant species comparisons such as human-fish comparisons have also proven to be effective tools. Given the extensive divergence time between fish and human (around 450 million years ago, (Kumar and Hedges, 1998)) and inherent biological differences, it is possible that conserved sequences will represent CREs that are essential for common developmental processes and would have the most dramatic biological effect if disrupted. Nobel laureate Sydney Brenner proposed in 1993 that the Tetraodontoid fish *Fugu rubripes* (pufferfish), with a compact genome of 400 Mb, which is 7.5 times smaller than the human genome, would represent an ideal model for deciphering the human genome, as it has all the specialized functions of higher vertebrates (Brenner et al., 1993). Not only have pufferfish-human comparisons allowed the identification of around 1000 novel human genes (Aparicio et al., 2002) but also a pioneering study carried out years before the human genome sequence was available showed how these comparisons could be used to unveil non-coding functional elements (Marshall et al., 1994). Multiple studies have efficiently identified functional human enhancers using human-fish comparisons as the only filter (Miles et al., 1998; Bagheri-Fam et al., 2001; Nobrega et al., 2003; Woolfe et al., 2005; Allende et al., 2006; Pennacchio et al., 2006). Based on some of these reports, an arbitrary threshold was established for the identification of human-fish conserved elements, requiring 70% of conservation over 100 bp (Ahituv et al., 2004).

Another approach to the identification of functional elements is to increase the stringency of conservation alignments between two closely related species (Bejerano et al., 2004; Visel et al., 2008; McBride et al., 2011). Bejerano et al identified ultraconserved elements (100% sequence identity for at least 200bp) between human, mouse and rat, and found 481 conserved regions located in exons of genes involved in RNA processing, or in intergenic regions and introns of genes implicated in transcriptional regulation and embryonic development. Interestingly, some of these segments were also conserved in fish (Bejerano et al., 2004). This and other studies validate the use of both ancient human-fish conservation and human-mouse-rat ultraconservation parameters for the identification of functional non-coding elements that behave as enhancers *in vivo*.

While these studies assume that a high level of sequence conservation is indicative of functionality there is evidence that some non-conserved fragments can maintain their regulatory function and direct transgene specific expression in several species (Fisher et al., 2006a; Hare et al., 2008; McGaughey et al., 2009; Swanson et al., 2010; Chatterjee et al., 2011). In zebrafish Fisher and colleagues tested discrete regulatory sequences of the human RET locus, a gene encoding a receptor tyrosine kinase, and their data suggested that functional conservation in fish was possible without sequence similarity (Fisher et al., 2006a). There are also studies showing that TFBSs in *cis*-regulatory elements may be rapidly gained or lost during evolution. Schmidt et al showed that the majority of TF binding events in adult livers of 5 vertebrates are species-specific, while ultraconserved events are very rare (Schmidt et al., 2010). Recent reports have also proposed that the level of conservation may vary depending on the tissue type. Blow and colleagues demonstrated that heart enhancers are three times less well conserved than other tissue-specific enhancers in vertebrate evolution, such as forebrain enhancers (Blow et al., 2010). This suggests that the use of sequence conservation for enhancer identification should be complemented with other strategies.

1.4.2 Open chromatin sites as indicators of CRE regions

Packaging of histones into DNA normally prevents the binding of non-histone proteins (Luger et al., 1997). For stable binding the displacement or disruption of nucleosomes is necessary. Thus, open chromatin sites or nucleosome-depleted regions are good indicators of the presence of concentrated TFs and chromatin remodelling complexes, and therefore of regulatory elements (Workman, 2006).

Disruption of chromatin structure can be experimentally identified by hypersensitivity to DNase I digestion (Galas and Schmitz, 1978; Wu, 1980). DNase I footprinting is a protection assay where cleavage of double-stranded DNA is inhibited by the specific binding of a ligand. Protected fragments will be indicated by a "gap" or footprint on a gel where the digestion products are resolved (Fox, 1997; Hampshire et al., 2007). Dnase I hypersensitive assays coupled to next-generation sequencing have been used to predict CREs in cell lines (Mito et al., 2007; Li et al., 2011; Song et al., 2011). Identification of DNase I hypersensitive Sites (DHS) in 6 human cell lines encompassing 1% of the human genome revealed that 22% of DHSs are shared among all cell types and mostly overlap with promoters or insulator elements, whereas cell type-specific DHSs correlate with known enhancer elements (Xi et al., 2007). The second phase of the ENCODE Project has mapped 2.89 million unique DNase I hypersensitive sites in 125 cell types by DNase I-

Seq, most of which are distal to TSS, with an average of 205,109 sites per cell type (Bernstein et al., 2012; Thurman et al., 2012). Although DNase I analyses have so far only been applied to cell lines, the data can be used to identify functional developmental enhancers in embryos, as demonstrated by the identification of novel *Pax6* enhancers in mouse that lacked evolutionary conservation (Kleinjan et al., 2001; McBride et al., 2011).

A more recent technique called FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) has also been successfully used to map open chromatin sites (Giresi et al., 2007). FAIRE detects regions depleted of nucleosomes that strongly correlate with DNase I hypersensitive sites, allowing an effective identification of regulatory regions (Gaulton et al., 2010; Song et al., 2011). Nevertheless, the open chromatin sites detected by both techniques seem to differ slightly. While DNase I enzyme captures chromatin structures that are close to TSS, FAIRE tends to detect distal CREs that DNase I cannot cut (Song et al., 2011).

1.4.3 Combinatorial binding of transcription factors define enhancers

Because combinatorial binding of TFs is required for enhancer function, effort has been devoted to the computational and experimental prediction and mapping of TF binding sites. Computationally, motif clustering of TFBS has been used to predict the location of CREs, as reviewed in (Aerts, 2012). Experimentally, chromatin immunoprecipitation (ChIP) has been the primary method used to identify protein-DNA interactions, as reviewed in (Furey, 2012). It has been demonstrated that combinatorial binding predicts the location of enhancers (Hallikas et al., 2006; Zinzen et al., 2009; He et al., 2011; Stefflova et al., 2013). Zinzen and colleagues profiled the genome-wide occupancy of 5 TFs involved in *Drosophila* mesoderm differentiation at multiple developmental times using ChIP-Chip, and used the observed profiles to computationally predict spatio-temporal patterns, which were very accurately matched *in vivo* in transgenic embryos (Zinzen et al., 2009). However, unlike approaches such as DNase I hypersensitivity, ChIP requires prior knowledge about the factors to be tested, and each factor needs to be evaluated separately. Furthermore, ChIP-grade antibodies are not always available, hindering the systematic identification of enhancers. Currently, therefore, information derived from TFBS is coupled with other approaches such as sequence conservation, binding of general transcriptional co-factors or/ and histone modification marks, in order to predict more accurately transcriptional enhancers.

1.4.4 Binding of general transcriptional co-factors: CBP/p300 mark enhancers

Binding of general transcriptional co-factors has also been shown to help locate developmental enhancers (Visel et al., 2009a; Blow et al., 2010; Rada-Iglesias et al., 2011; Xu et al., 2011; May et al., 2012; Rada-Iglesias et al., 2012). CBP and p300 were originally identified as proteins binding cAMP-responsive transcription factor CREB (Chrivia et al., 1993) or the adenovirus E1A oncoprotein (Eckner et al., 1994), respectively. Both proteins exhibit adaptor properties and have high homology at the sequence level, which is why they are considered functional homologs and usually referred to as CBP/p300 (Eckner et al., 1996). CBP and p300 also contains intrinsic HAT activity (Bannister and Kouzarides, 1996; Ogryzko et al., 1996) mostly for H3K27 residue (Tie et al., 2009).

Heintzman and colleagues demonstrated that distal p300 binding sites displayed many enhancer-like features, such as overlap with DHSs and contained evolutionarily conserved motifs (Heintzman et al., 2007). Subsequently, it was demonstrated *in vivo* that mapping of p300 binding sites by ChIP could be used to accurately predict active enhancers in numerous cell lines, in mouse embryonic forebrain, midbrain and limb tissues (Visel et al., 2009a), in flies (Negre et al., 2011), and human embryonic stem cells (hESC) (Rada-Iglesias et al., 2011; Rada-Iglesias et al., 2012; Rada-Iglesias et al., 2013) among others. Nevertheless, there are subsets of enhancers that lack p300 binding (Heintzman et al., 2007; Heintzman et al., 2009; He et al., 2011), so mapping of p300 is usually complemented with other strategies such as the profiling of tissue-specific TFs or histone modification marks.

1.4.5 Histone modification marks as predictors of enhancers

Histones serve a role beyond packaging DNA (Li et al., 2007). Initial evidence for the existence of a complex structure between histones and the double helix of DNA came from electron microscopy studies, X-ray crystallographic observations, and nucleosome digestions in the 1970s, as reviewed in (Kornberg, 1977). The nucleosome core contains two copies of H2A, H2B, H3 and H4 histone proteins assembled into an octamer with 145-147 bp of DNA wrapped around it. These core particles are further stabilised by linker histone H1 and linker DNA (Luger et al., 1997). Nucleosome arrays are folded like "beads on a string" into a 30nm diameter fibre (Finch and Klug, 1976; Widom and Klug, 1985; Graziano et al., 1994). This chromatin fibre is further condensed in the form of metaphase chromosomes, where DNA is compacted 10,000 fold (Belmont et al., 1987; Woodcock and Ghosh, 2010). Nucleosomes are highly dynamic, both in terms of positioning and in terms of biochemical modifications. The NH2 terminus of the histones (or histone tails) protrude from the centre of the nucleosome core and are amenable to a wide-array of post-translational modifications such as phosphorylation, methylation, acetylation, ubiquitilation etc., as reviewed in (Kouzarides, 2007).

Recently, some modifications have been associated with different states of transcription, or have even been used to predict the location of *cis*-regulatory modules such as promoters or enhancers. Whether the presence of these marks is a cause or a consequence of enhancer function or TF occupancy remains unknown (Spitz and Furlong, 2012). The first indication that enhancers could be marked by histone modification marks came from the pilot ENCODE project and Bing Ren's lab. Ren and colleagues succeeded in identifying chromatin features that could distinguish promoters from enhancers (Heintzman et al., 2007). They showed that active promoters are enriched in trimethylation of histone H3 lysine 4 (H3K4me3), lack mono-methylation of H3 lysine 4 (H3K4me1) and are marked by a nucleosome-free region near the TSS. In contrast, enhancers bound by the co-activator p300 are enriched in H3K4me1, lack H3K4me3, and usually lie on DNase I hypersensitivity sites and in evolutionary conserved regions. Furthermore, enhancers show cell-type specific histone modifications that correlate strongly with gene expression, in contrast to the invariant pattern of chromatin state in promoters (Heintzman et al., 2009). These initial observations in cell lines were quickly confirmed in several genome-wide studies using different embryo model systems such as *Drosophila* (Negre et al., 2011), zebrafish (Aday et al., 2011) and mouse (Visel et al., 2009a). In *Drosophila* H3K9ac, H3K27ac and H3K4me1 were also found in promoters and H3K4me1 plus CBP-p300 was a common signature of enhancers (Negre et al., 2011). In zebrafish it was also demonstrated that the 5' end of expressed genes were enriched in H3K4me3 (Wardle et al., 2006) and H3K4me1 (Aday et al., 2011) and that this association correlated well with the level of gene expression. It was also shown that H3K4me3 and H3K27ac modification marks were present in developmental genes, at the onset of zygotic transcription irrespective of their transcriptional activation (Vastenhouw et al., 2010), and that enhancers were enriched in H3K4me1 (Aday et al., 2011).

Furthermore, it was shown that developmental enhancers in mouse and human ESC could exist in two different states: poised and active (**Figure 1.6**), which may be distinguished by the presence of H3K27ac mark (Creyghton et al., 2010; Rada-Iglesias et al., 2011). In both states, enhancer regions are marked by H3K4me1, bound by the co-activator p300, display low nucleosomal density and are depleted of H3K4me3 modification. In hESCs an association was found between “active enhancers” and H3K27ac enrichment, absence of H3K27me3 and a high level of transcription; while "poised enhancers” are enriched in H3K27me3, lack H3K27ac and are poorly expressed. Upon hESC differentiation poised enhancers acquire H3K27ac and can recruit RNA

PolII. More importantly, when tested in zebrafish, these elements can drive cell and stage specific expression, and function as developmental enhancers (Rada-Iglesias et al., 2011). This classification could explain why some previous studies had found H3K4me1 associated with non-active enhancers (Heintzman et al., 2009; Rada-Iglesias et al., 2011). In mESC however, poised enhancers do not display any modification in the K27 residue of H3 (Creyghton et al., 2010) and the repressive mark H3K9me3 may also be distinctive of this category (Zentner et al., 2011).

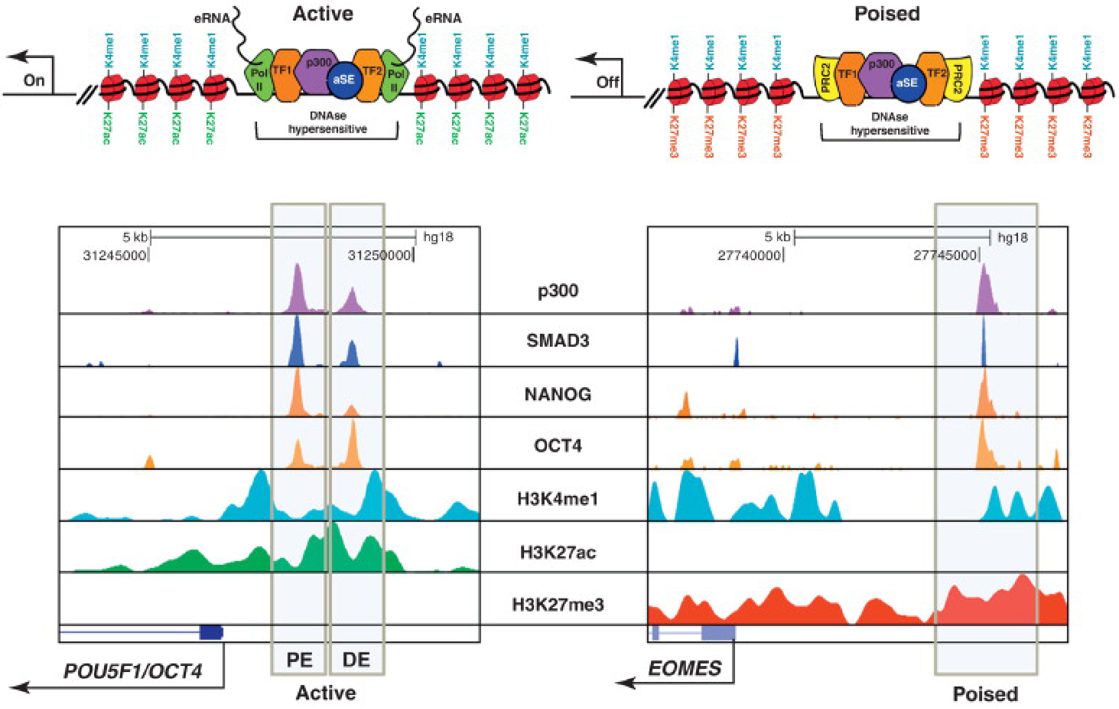


Figure 1.6 Chromatin signatures at active and poised developmental enhancers in hESCs.
Both enhancers lie on DNase I hypersensitive sites, are bound by co-activator p300 and several transcription factors (TF1 and TF2), and are marked by H3K4me1. Active enhancers are also bound by PolII, which produces eRNA bidirectionally. Poised enhancers recruit Polycomb Repressive Complex 2 (PRC2) and are marked by H3K27me3 repressive mark. Bottom panels are Genome Browser screenshots of POU5F1/OCT4 distal and proximal active enhancers (DE and PE respectively), and EOMES poised enhancer, and their enrichment in p300, SMAD3, NANOG and OCT4 binding as well as enrichment in several histone modifications marks (H3K4me1, H3K27ac, H3K27me3). Reproduced from (Calo and Wysocka, 2013).

There are additional histone modification marks that have been associated with enhancers. In mESC, active enhancers were shown to produce RNA transcripts, so phosphorylation of Ser2-5 of PolII and H3K36me3, which are associated with transcript elongation (Kouzarides, 2007; Kolasinska-Zwierz et al., 2009), can be considered predictive marks (Zentner et al., 2011). In addition to that, H4K16ac modification has been shown to overlap with a subset of active enhancers in mESC that are not bound by p300 or enriched in H3K27ac mark, suggesting that it may be a novel mark for p300-independent regulatory regions (Taylor et al., 2013).

Recently a computational method has been developed to predict enhancers on a genome-wide scale based on histone modification marks (Rajagopal et al., 2013). By integrating information from 24 histone modification profiles, DNase I hypersensitive sites and p300 binding sites in mESC and lung fibroblasts, the new algorithm can measure the relative importance of each modification mark in defining an enhancer and demonstrated that H3K4me1, H3K4me2 and H3K4me3 are the minimal marks required for robust enhancer prediction across cell lines and replicates, leading to a more accurate prediction of enhancers in different tissues.

1.4.6 H2AZ histone variant correlates with nucleosome-free regions

H2AZ is a highly conserved histone variant that diverges from the canonical H2A histone in terms of length and sequence of the C-terminal domain (Kamakaka and Biggins, 2005). In yeast, H2AZ prevents the spread of heterochromatin and is preferentially enriched in promoter regions, specifically on the nucleosomes that are displaced during transcriptional initiation (Raisner et al., 2005; Zhang et al., 2005). A similar pattern has been found in humans, where H2AZ is found upstream and downstream of the TSS and also in nucleosome-free regions of enhancers and insulators (Barski et al., 2007). It has been suggested that nucleosomes containing the H2AZ histone variant are less stable and more easily displaced (Placek et al., 2005; Zhang et al., 2005). Consistent with this idea, the H3.3/H2AZ histone combination is preferentially located at nucleosome-free regions and correlates well with *cis*-regulatory elements including promoters, silencers and enhancers (Jin et al., 2009). Thus, H2AZ has been successfully used as a marker for open chromatin in order to identify enhancers (Pham et al., 2012).

1.4.7 Transcription from enhancers is pervasive and might aid in enhancer detection

Even though most recent approaches to identify regulatory regions have been based on the analysis of chromatin marks, complementary strategies have been implemented. Although initially surprising, transcription from enhancers is now well documented across cell types and seems to be an additional feature of this type of CREs (Kim et al., 2010; Melgar et al., 2011; Wang et al., 2011a; Djebali et al., 2012). The transcripts produced by enhancers have been named eRNA (enhancer RNAs) and unlike protein-coding transcripts they are not marked by H3K4me3 and H3K36me3 along their length (Mikkelsen et al., 2007; De Santa et al., 2010). The exploitation of bidirectional transcription from enhancers as a prediction tool will be discussed in more depth in **Chapter 5**.

Overall, in terms of predictive approaches, the integration of multiple layers of genomic information, such as sequence conservation, enrichment of key epigenetic marks, general co-factor and transcription factors binding events, and possibly bidirectional transcription, will potentially yield the most comprehensive method for identifying active developmental enhancers.

1.5 *Cis*-regulation and disease

Over the past decades, thousands of mutations causing monogenic disorders have been identified in coding regions. By contrast, the number of disease-causing mutations in distal CREs and other non-coding CREs has been minimal, mainly because of the difficulties in identifying regulatory elements (Visel et al., 2009b). Nevertheless, variations in any component of the transcriptional machinery including transcription factors, co-factors, chromatin regulators or in regulatory regions and non-coding RNAs are often associated with human disorders and can increase susceptibility to multifactorial diseases such as cancer, autoimmune diseases, diabetes, schizophrenia or obesity among others, as reviewed in (Lee and Young, 2013).

Diseases can be caused by the removal or repositioning of distal regulatory elements leading to altered gene expression (Kleinjan and Lettice, 2008). Classic examples include some cases of thalassemias, where deletions or chromosome rearrangements reposition enhancers needed for normal β-globin gene expression (Kioussis et al., 1983; Semenza et al., 1984). In these cases the coding region of the gene was not disrupted but its transcriptional regulation was. Since their discovery, the term "position effect" is used to refer to changes in the level of gene expression when the location of the regulatory regions is changed in the chromosome, usually by translocation (Kleinjan and Lettice, 2008).

There are also examples of human diseases caused by single-nucleotide changes in *cis*-regulatory regions. Preaxial polydactyly is among the most commonly observed human limb malformations and includes a broad range of digit abnormalities in hands and feet. Studies in transgenic mice have shown that Preaxial polydactyly is caused by mutations in the highly conserved limb-specific enhancer ZRS of Sonic Hedgehog (*Shh*), which lies 1 Mb away from the target gene within the intron of the neighbouring gene *Lmbr1* (Lettice et al., 2002; Lettice et al., 2003). These single-point mutations cause ectopic *Shh* expression and subsequently supernumerary digits and were also found in four families affected of PDD (Lettice et al., 2003; Lettice et al., 2008). Because there are several single point changes found throughout the length of ZRS, it is unlikely that PDD arises from the disruption of a single TFBS (Kleinjan and Lettice, 2008).

Not only mutations in enhancers, but also mutations in proteins regulating enhancer function can lead to disease, as reviewed in (Smith and Shilatifard, 2014). For example, mutations in CBP/p300

HAT can lead to Rubinstein-Taybe syndrome (Roelfsema et al., 2005). Missense and non-sense mutations in MLL complexes, histone methyl transferases responsible for catalysing H3K4me1 in enhancers (Hu et al., 2013), have been associated with Kabuki syndrome (Ng et al., 2010). Also, mutations in pioneer transcription factors can lead to anemia or cancer (Zheng and Blobel, 2010).

Common variants located in CREs can increase the susceptibility to certain multifactorial diseases. Hirschsprung disease (HSCR), or congenital aganglionosis with megacolon, is a multigenic disease in which RET proto-oncogen is the main gene implicated (Emison et al., 2005). A highly conserved region in the first intron of RET was found to be an enhancer by *in vitro* and *in vivo* transgenesis assays (Grice et al., 2005). This enhancer contains 3 SNPs that are in linkage disequilibrium (LD) with HSCR, one of which can explain a 10 to 20-fold greater susceptibility to the disease (Emison et al., 2005). Fuelled by the power of Genome-Wide Association Studies (GWAS) a growing number of single nucleotide polymorphisms (SNPs) located in non-coding regions have been linked to human diseases (Bernstein et al., 2012; Maurano et al., 2012), such as T2D (Dupuis et al., 2010; Stitzel et al., 2010). Interestingly, so far validation of disease-associated enhancers has not been quantified *in vivo*, and *in vitro* analyses have revealed very subtle effects (Gaulton et al., 2010; Stitzel et al., 2010; Rada-Iglesias et al., 2013).

1.6 Zebrafish as a model organism

Zebrafish (*Danio rerio*) is a tropical, freshwater bony fish of the Cyprinidae family. There are around 44 species of Danio distributed throughout East and South-East Asia, with the highest diversity in North-East India, Bangladesh and Myanmar, reviewed in (Spence et al., 2008). They are distinguished by their small size (usually < 120 mm) in adulthood and their distinctive pattern of alternative dark and white horizontal stripes.

Within the class of bony fishes or Teleostei there are three other fish model organisms worth mentioning. Medaka (*Oryzias latipes*), which has also been used as a model for developmental genetics and to functionally validate regulatory elements (Wittbrodt et al., 2002) and two pufferfish species, the Japanese pufferfish *Takifugu rubripes* (also known as *Fugu rubripes*) and the green spotted pufferfish or *Tetraodon nigroviridis.* The genome sequence of these three species is publicly available (Aparicio et al., 2002; Jaillon et al., 2004; Kasahara et al., 2007). The pufferfish species own the most compact genome known of all vertebrates, while maintaining the specialized functions of higher vertebrates. Therefore, pufferfish was proposed as a valuable model in the identification of conserved *cis*-regulatory elements through comparative genomic studies (Brenner et al., 1993; Aparicio et al., 2002). However, *Fugu rubripes* cannot be bred or kept in laboratory conditions, and the breeding and biology of *Tetraodon nigroviridis* is mostly unknown, hindering their use as a model system beyond comparative genomics (Muller et al., 2002). The inability to carry out functional work with pufferfish has further contributed to widen the use of zebrafish, and Medaka to a lesser extent, as a developmental model among the scientific community.

1.6.1 General characteristics

Multiple features have resulted in zebrafish becoming a widely used vertebrate model system. Adult zebrafish can breed all year round in laboratory conditions and are easy and relatively cheap to maintain in the laboratory. A single female can spawn clutches of several hundreds of eggs and generation time is short, which makes it a high-throughput model when compared to rodents. Zebrafish present the additional advantage of embryo transparency, which allows for observation of internal organs and imaging using fluorescent marker proteins. Moreover zebrafish fertilization is external, which eases manipulation of the embryos. They exhibit a rapid development, with primary organ systems present between 36 hpf and 72 hpf, and all embryonic stages have been described in detail and can be easily identified under the dissecting microscope (Kimmel et al., 1995).

1.6.2 Resources and tools available to the zebrafish community

Zebrafish is increasingly being used to model human diseases, which requires a high quality sequenced genome. The zebrafish reference genome sequence was recently published (Howe et al., 2013b). However, the initial shotgun draft has been available and accessible to users since 2002. There are 26,206 genes annotated in the zebrafish genome, 69% of which have an orthologous in human. Reciprocally 71.4% of human genes have an orthologous zebrafish gene (Howe et al., 2013b). Compared to humans and mice, the higher number of genes displayed by zebrafish is most likely a result of an ancient whole-genome duplication process specific to teleosts (Meyer and Schartl, 1999).

In addition to the availability of the reference genome sequence, all mutant and transgenic lines available, morpholinos, antibodies, anatomical structures, phenotypes, publications and expression data are systematically curated and incorporated onto ZFIN website (Zebrafish Model Organism Database, http://zfin.org), which is a highly valuable resource for the zebrafish research community (Howe et al., 2013a).

Zebrafish has become one of the most prominent vertebrate model systems used to study development, disease and regulatory mechanisms. The first studies using *Danio rerio* as a model organism for genetics began in 1981 (Streisinger et al., 1981). Since then, a large number of tools for forward and reverse genetics have been developed. Forward genetic approaches seek to identify genes and pathways involved in different developmental processes through the screening of populations of individuals displaying phenotypes induced by random mutations in their genome. Classical genetic screens have been successfully carried out using chemical mutagens such as N-ethyl-N-nitrosourea (Driever et al., 1996) or retroviral insertions (Amsterdam et al., 1999). All these approaches have identified thousands of mutations in genes that affect embryonic development. The small size of zebrafish larvae, fitting inside wells of a standard 96-well plate), and the availability of large number of embryos are essential features for the automation and high-throughput scale needed for these screens.

Reverse genetics seek to elucidate the function of known genes by knock-out and knock-down approaches. Methods for reverse genetics include morpholino oligonucleotides and engineered nucleases. Morpholinos are short antisense oligonucleotides (18-25 bp) that can effectively bind a target mRNA and function through an RNase-H independent mechanism as efficient translational inhibitors *in vivo* (Nasevicius and Ekker, 2000). When injected into zebrafish one-cell stage embryos, morpholinos mediate knock-down of a target gene, generating phenotypes that resemble loss-of-function mutants (Nasevicius and Ekker, 2000). However, they have some limitations: they are transient in nature, only are completely penetrant during the first 48 hpf and there are common off-target effects caused by non-specific binding (Dahm and Geisler, 2006). An additional reverse genomic tool is TILLINGS (Targeting Induced Local Lesions IN Genomes), which relies on the screen of mutagenized populations, generally produced by ENU, and subsequent analysis by sequencing of induced mutations (Wienholds et al., 2003). This approach has recently been able to identify mutations in 3,188 zebrafish genes that are orthologous to some of the 5,494 human genes currently associated with human diseases (Kettleborough et al., 2013).

More recently a wide array of engineered nucleases including ZFN (Zinc Finger Nucleases), TALENs (Transcription Activator-Like Effector Nucleases) and CRISPR-Cas (Clustered Regulatory Interspaced Short Palindromic Repeats) have emerged as more refined tools for genome-editing in zebrafish (Maeder et al., 2008; Sander et al., 2011; Hwang et al., 2013). These tools contain a specific DNA binding domain fused to a non-specific DNA cleavage domain that induces double-strand breaks (DBS), stimulating NHEJ (Non-Homologous End Joining) DNA repair mechanisms in the cell and promoting introduction of deletions, insertions or substitutions at target loci, as reviewed in (Gaj et al., 2013). ZFNickases (Zinc Finger Nickases) have been introduced as powerful alternatives to avoid unwanted indels caused by error-prone NHEJ repair of DBS (Kim et al., 2012; Ramirez et al., 2012). They create single-strand breaks or nicks and trigger DNA repair by the homologous directed repair pathway, reducing the mutagenesis rate of classical ZFNs. The application of these tools to the field of transcriptional regulation could be revolutionary, as it will allow dissection of the structure of CREs by disrupting potentially relevant motifs with high precision. Along these lines, a recent study has proven that a cancer-associated enhancer lost its function following deletion of a 7-bp motif using TALEN nucleases (Webster et al., 2014). As a result of the genetic lesion, the *MET* gene enhancer could no longer interact with its cognate promoter and gene regulation was altered, demonstrating the importance of a single TFBS.

1.7 Transgenesis in zebrafish for the study of transcriptional regulation

Transgenesis is the most effective method for assess the potential of putative regulatory elements *in vivo*. In these assays constructs containing the region to be tested are linked to a reporter gene, the activity of which can be accurately measured when introduced into the cells of the animal model. If the element to be tested is an enhancer, a construct is produced so that the enhancer is placed upstream of a weak promoter and a reporter gene, typically fluorescent proteins, as reviewed in (Narlikar and Ovcharenko, 2009). Transgenesis techniques have been described in several animal models such as mouse, *Xenopus* and zebrafish (Stuart et al., 1990; Khokha and Loots, 2005; Fisher et al., 2006b; Pennacchio et al., 2006). These methods require the injection of the vector into fertilized eggs. The vector is randomly integrated in the host genome and tissue specific activity is evaluated by assessing the *in vivo* expression pattern of the injected embryos.

Zebrafish has proven to be unique in its capacity for transgenesis screens due to the low cost of transgenesis and the availability of high throughput approaches. The production of around two hundred eggs per female per week makes it an ideal model for the injection and screening of large number of transient transgenic embryos. In addition to this, the short generation time (3 months) permits generation of stable transgenic lines. Establishment of transgenic lines is more time-consuming and lower-throughput. Although transient transgenics allow for high-throughput screens, one main limitation is the high mosaicism between embryos. Mosaicism complicates the identification and analysis of transgene activities in small tissue domains (Pashos et al., 2008). Nevertheless, transient transgenic embryos can be used, where each injected embryo represents a different integration site. Thus, position effects can be statistically eliminated by observing common patterns in a large number of embryos (Muller et al., 1997; Muller et al., 1999; Dickmeis et al., 2004; Gehrig et al., 2009). In transgenesis assays it is also common to find that reporter constructs drive expression in unexpected cells or tissues, referred to as "ectopic expression" (Summerbell et al., 2000; Adachi et al., 2003; Chao et al., 2010) or at a stage that does not match temporally that of the tested gene, known as "heterochronic expression" (Lin et al., 2010), hindering the analysis of the CRE of interest.

1.7.1 Zebrafish tests of human *cis*-regulatory elements

Zebrafish has been successfully used to test fish and human candidate regulatory regions using transient and stable transgenic assays (de la Calle-Mustienes et al., 2005; Woolfe et al., 2005; Abbasi et al., 2007; Navratilova et al., 2009; Li et al., 2010b; Narlikar et al., 2010; Ritter et al., 2010; Rada-Iglesias et al., 2011; Gorkin et al., 2012; Ritter et al., 2012; Bhatia et al., 2013; Ravi et al., 2013). Most of these studies tested conserved non-coding regulatory regions among evolutionary distant species, such as human and zebrafish or *Fugu*, and confirmed that CNEs are good indicators of putative enhancers that function *in vivo*. Although some studies have addressed in detail how well enhancer function is conserved across species, particularly between human and zebrafish (Navratilova et al., 2009; Ritter et al., 2010), the test of enhancers predicted in a tissue-specific fashion by novel genome-wide methods has not been explored in depth.

1.7.2 Transgenesis methods in zebrafish

Several protocols exist for the creation of stable transgenic lines in zebrafish, but two basic approaches have been employed to analyse CREs by transgenesis: naked DNA microinjections and transposon-mediated transgenesis.

The first experiments to create transgenic zebrafish embryos used linearized plasmid DNA containing *cis*-regulatory elements (Stuart et al., 1988; Stuart et al., 1990; Westerfield et al., 1992). The injected linear plasmid is first replicated during cleavage stages before being mostly degraded during gastrulation. The vast majority of replicated DNA is not stably integrated but lives in an extra-chromosomal state, which results in a highly mosaic individual. The mosaicism is most likely caused both by late integration into the genomic DNA (prompted by the rapid 15-minutes cleavages), uneven distribution of the exogenous DNA in the embryo, and/or tissue-specific activation of the transgene (Stuart et al., 1988; Stuart et al., 1990; Westerfield et al., 1992). These and other experiments proved that mosaic analysis of transient transgenic zebrafish provides a rapid method for the dissection of the activity of *cis*-regulatory elements, either by injecting promoters alone or by co-injecting enhancer and promoter fragments (Muller et al., 1997; Woolfe et al., 2005; Allende et al., 2006). Only about 5% of the mosaic injected fish in these experiments carry germline integration of the transgene (Stuart et al., 1988; Stuart et al., 1990; Amsterdam et al., 1999) Nevertheless, germline transmission rates can be improved with the use of transposons.

1.7.3 Use of transposons for random integration of transgenes

Transposons or DNA-based transposable elements are mobile DNA segments. In their simplest form they consist of a genetic sequence flanked by short terminal repeats and the encoded transposase enzyme, which triggers the replicative spread through the genome by recognising target DNA sequences and mediating integration by cutting, exchanging and fusing DNA strands (Plasterk, 1993). Autonomous transposons carry the transposase gene whilst non-autonomous transposons depend on an external enzyme for their mobilization in the genome (Cui et al., 2002).

Transposon-mediated transgenesis represents an advantage over naked DNA microinjection, as reviewed in (Kawakami, 2005), particularly with regards to germline transmission rates, which increase from around 5% to 12.5%, or even 50% (Kawakami et al., 2000; Kawakami et al., 2004).

1.7.3.1 Sleeping Beauty

Sleeping Beauty (SB) is a synthetic transposon generated from a consensus sequence of the salmonoid Tc1-like elements. The system consists of a synthetic gene encoding the transposase and a cloned Tc1-like element containing the inverted repeats needed for transposition (Ivics et al., 1997). It is very active in higher vertebrates (including fish, mouse and human) and preferentially integrates one copy in sequences flanking TA dinucleotides (Ivics et al., 1997). In zebrafish, SB is a valuable tool for transgenesis: it enhances the expression rates of transgenes from 5% to 31% over standard plasmid injection and exhibits a germline transmission efficiency of around 10% of multiple single-copies of transgenes (Davidson et al., 2003). However, this two-component system has a limited cargo carrying capacity (up to 7Kb for efficient transposition rates in human cells) and lowered transposition activity at high SB doses (Geurts et al., 2003; Balciunas et al., 2006).

1.7.3.2 Tol2 transposon

Tol2 belongs to the hAT family of transposons and was isolated from the tyrosinase gene locus of the Medaka fish *Oryzias latipes* (Koga et al., 1996). Koga and collaborators identified a 4.7 Kb insertion on the 5th exon of i^4 allele displaying "cut and paste" transposon features: it was present in multiple copies in the host genome and was flanked by short inverted terminal repeats (Koga et al., 1996). In order to determine whether the Tol2 element was autonomous and active in zebrafish, a construct harbouring the Tol2 element was injected into zebrafish eggs. Analysis by PCR of the injected embryos showed that Tol2 was excised from the plasmid leaving small deletions and insertions at the excision sites, which is characteristic of transposons (Kawakami et al., 1998). In addition, when mRNA transcribed *in vitro* was injected in embryos, Tol2 produced a functional transposase capable of catalysing transposition in the zebrafish germline (Kawakami et al., 2000). These results indicated that Tol2 was an autonomous element and, more importantly, that it was active in zebrafish.

The minimal sequence required for transposition was identified from the Tol2 element, indicating that 200 bp from the left end and 150 bp from the right end were sufficient for Tol2-mediated transposition (Urasaki et al., 2006). Subsequently, constructs containing miniTol2 versions were designed (Balciunas et al., 2006; Urasaki et al., 2006). These vectors offered the additional advantage of being able to carry efficiently DNA elements larger than 10 Kb between the left and right Tol2 ends (Balciunas et al., 2006; Urasaki et al., 2006), and even Bacterial Artificial Chromosome (BAC) transgenes (Suster et al., 2009).

Several features such as high rates of germline transmission (around 50%), single copy transgene insertions, an average of 5-6 insertions, clean integrations in the genome, and the passage of transgenes through generations make Tol2 an ideal system for creating transgenic zebrafish (Kawakami et al., 2000; Kawakami et al., 2004).

1.7.4 Zebrafish as a model for studying *cis*-regulation of vertebrate pancreas development and function

1.7.4.1 Zebrafish shares a common pancreas structure and function with mammals

In this thesis, pancreas specific human enhancers will be tested by transgenesis in zebrafish. Zebrafish has become a popular model for studying development and the fact that the zebrafish pancreas shares a common basic structure and cellular composition with mammals, implies that the lessons we can learn from zebrafish will be broadly applicable (Kinkel and Prince, 2009).

Pancreas is a vital organ controlling glucose homeostasis, food digestion and nutrition, and consists of both endocrine and exocrine compartments. In adult mammals the endocrine pancreas represents 1-2% of the total mass of the organ and is embedded within the exocrine pancreas, as reviewed in (Slack, 1995).The exocrine component is a lobulated gland consisting of acinar and ductal cells. Acinar cells are arranged into spherical structures called acini that produce and secrete digestive enzymes (zymogens) from their apical surface including proteases, amylases, lipases and nucleases. Ductal cells form a network that allows the transport of zymogens into the intestine (Wang et al., 2011b). The endocrine pancreas consists of islets that contain four different cell types secreting hormones into the bloodstream: insulin producing β-cells, glucagon producing α-cells, somatostatin producing δ-cells, and pancreatic polypeptide producing PP-cells. The pancreatic islet is organized in a mantle-core structure, with the most abundant β-cells in a central core surrounded by non β-cells at the periphery (Montague, 1983). It has been shown in most species that during development and early post-natal period there is a unique fifth type of endocrine cells: ghrelin producing ε-cells (Heller, 2010). In humans, the islets of Langerhans are organized as several mantle-core subunits. The adult endocrine pancreas contains approximately one million islets of Langerhans embedded in exocrine tissue (Li et al., 2009). This description of the human pancreas is broadly applicable to reptiles, amphibia, birds and some fishes.

1.7.4.2 Zebrafish pancreas development

At early larval stages, the zebrafish pancreas consists of exocrine tissue surrounding a single principal islet. The islet is composed of a core of β-cells, enclosed by an outer layer of δ-, α- and ε-cells. Whether PP-cells form part of the zebrafish embryonic islet remains unclear, with some authors claiming that PP cells can only be detected in the adult pancreas (Argenton et al., 1999).

The endocrine and exocrine compartments originate from two contiguous areas in the gut that bud at two different times during development **(Figure 1.7)**: the anteroventral bud will give rise to the endocrine pancreas, while the posterodorsal bud will form the exocrine pancreas and a supply of endocrine cells (Field et al., 2003).

The first pancreatic bud forms from two sets of endodermal cells that originate at both sides of the gastrula and merge at the midline by the 12-somite stage. These cells express *pdx1*, a pancreatic and duodenal homeobox gene, which is an early pancreatic and gut marker (Biemar et al., 2001). These cells bud off dorsally from the gut before 24 hpf. The posterodorsal bud gives rise to endocrine tissue expressing the main pancreatic hormones. Due to the rotation of the gut between 24-48 hpf, the dorsal bud is dislocated to the right side of the gut.

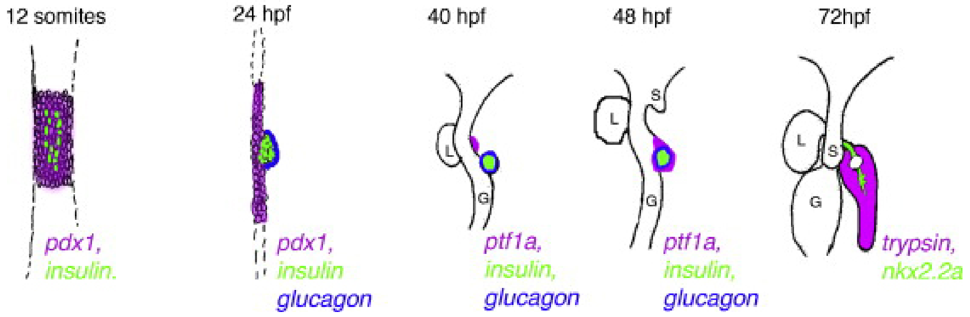


Figure 1.7 Schematic diagram of the development of zebrafish pancreas.
Relevant genes in pancreas morphogenesis are labelled in green, blue and magenta. L=liver, G=gut, S=swim bladder. All images are dorsal view with anterior at the top. Adapted from (Tiso et al., 2009)

By 48 hpf the first islet can be detected and it is organized like a mammalian one with a core of β-cells surrounded by α- and δ-cells (Tiso et al., 2009). Ghrelin cells are not detected in the zebrafish pancreas, but have been found in catfish (Heller, 2010). The anteroventral bud emerges at 40 hpf from the ventral portion of the gut, anterior to the first islet. It can be detected by the expression of the exocrine marker *ptf1a* (Lin et al., 2004). This second bud will grow towards the posterior end of the gut, enveloping the islet. By 5 dpf the pancreas lies on the right side of the gut and consists of a head, where the principal islet and *nkx2.2* -expressing extra-pancreatic ducts are, and a tail, which is posterior to the head and consists of exocrine pancreas (Parsons et al., 2009). The exocrine pancreas of a 5 dpf larvae contains approximately 260 cells (Jiang et al., 2008).

While the embryonic and larval zebrafish pancreas consists of a single principal islet, the pancreas in the adult zebrafish is organized as a large principal islet located at the head of the pancreas and several secondary islets within the tail of the pancreas (Tiso et al., 2009). The cells that form the secondary islets do not appear during the first 6 days of development, but are established during the first three weeks of larval life and are embedded in exocrine tissue and scattered along the gut (Parsons et al., 2009). Their formation can be explained by two alternative hypothetical models. The first implies the replication and migration along the gut tube of existing β-cells from the primary islet. The second model proposes the differentiation of nkx2.2a-expressing ductal precursor cells. A paper published by (Wang et al., 2011b) confirmed this second hypothesis. There are two differentiated cell populations in the anteroventral bud: a *ptf1a* expressing domain and a Notch-responsive domain. The Notch responsive pluripotent cells align along the ducts and differentiate during development to form ductal and centroacinar cells and the insulin producing cells that will eventually form the secondary islets.

1.8 Aims of this thesis

In the post-genomic era, international efforts have been devoted to design strategies that can identify CREs on a genome-wide scale. However, the vast majority of the putative enhancers have not been experimentally characterized *in vivo* and their contribution to disease is unclear. Therefore there is a demand for an animal model that can be used in validation assays with the capacity to be upscaled. Zebrafish has been used to study transcriptional regulation before but its relevance as an evolutionarily distant model for the functional test of mammalian *cis*-regulatory elements remains poorly understood. Therefore, I aimed to investigate the utility of the transgenic zebrafish embryo to study the function of candidate human enhancers identified by several prediction methods. I also aimed to develop a robust transgenesis method that would permit the detection of subtle quantitative changes in enhancer function caused by disease-associated SNPs.

Chapter Two: MATERIAL AND METHODS

2.1 Materials

All the chemicals used were molecular biology quality grade. Some commonly used chemicals, consumables and equipment are not listed here.

2.1.1 Antibiotics

Ampicillin sodium salt	A9518	Sigma-Aldrich, UK
Gentamicin sulfate	BPE918-1	Fisher Bioreagents, UK
Kanamycin sulfate from *Streptomyces kanamyceticus*	K1377	Sigma-Aldrich, UK

2.1.2 Bacterial strains

Alpha-Select Silver Efficiency chemically competent cells	BIO-85026	Bioline Ltd., UK
Alpha-Select Bronze Efficiency chemically competent cells	BIO-85025	Bioline Ltd., UK
Alpha-Select Silver Efficiency electrocompetent cells	BIO-85028	Bioline Ltd., UK
Stellar™ Competent Cells	636763	Clontech, UK

2.1.3 Chemical Reagents

Agarose	BIO-41025	Bioline Ltd., UK
Albumin from bovine serum, further purified fraction V, 99% pure (BSA)	A3059	Sigma-Aldrich, UK
5-Bromo 4-chloro 3 indolyl phosphate (BCIP)	11383221001	Roche Diagnostics, UK
Chloroform	C2432	Sigma-Aldrich, UK
CTP alpha-^{32}P	BLU008H001MC	Perkin Elmer, UK
Denhardt's Solution (50X)	750018	Invitrogen, UK
Deoxyribonucleic acid from calf thymus	D4522-1MG	Sigma-Aldrich, UK
dNTPs 10mM	R0192	Thermo Scientific, UK
Dynabeads® M-280 Streptavidin	11205D	Invitrogen, UK
Ethyl-3-aminobenzoate methanesulfonate 98% (MESAB/E-222/Tricaine)	E10521	Sigma-Aldrich, UK
Formamide (deionized)	AM9342	Ambion, UK
Gel Loading Dye, Blue 6X	B7021S	NEB, UK
Glycerol	G6279	Sigma-Aldrich, UK
Heparin sodium salt	H3393	Sigma-Aldrich, UK
Human genomic DNA	G3041	Promega, UK
LB (Luria Bertani) Agar	L2897	Sigma-Aldrich, UK

LB (Luria Bertani)Broth	L3022	Sigma-Aldrich, UK
New Born Calf Serum	N4637	Sigma-Aldrich, UK
N-Phenylthiourea Grade I (PTU)	P7629	Sigma-Aldrich, UK
Nitro blue tetrazolium (NBT)	11383213001	Roche Diagnostics, UK
Paraformaldehyde (PFA)	P6148	Sigma-Aldrich, UK
Phenol:Chloroform 5:1, pH 4.7	P1944	Sigma-Aldrich, UK
Phenol:Chloroform:Isoamyl Alcohol 25:24:1 Saturated with 10 mM Tris, pH 8.0, 1 mM EDTA	P3803	Sigma-Aldrich, UK
Phenol solution Equilibrated with 10 mM Tris HCl, pH 8.0, 1 mM EDTA	P4557	Sigma-Aldrich, UK
Phenol red solution	P0290	Sigma-Aldrich, UK
Ready-To-Go DNA Labelling Beads (-dCTP)	27924001	GE Life Sciences, UK
SSC buffer substance	85639	Sigma-Aldrich, UK
Sheep anti-digoxigenin-AP Fab fragments	1093274	Roche Diagnostics, UK
Sheep serum	013000121	Interchim, US
tRNA from wheat germ type V, lyophilized powder	R7876	Sigma-Aldrich, UK
Tween-20	P1379	Sigma-Aldrich, UK
UltraPure™ 1M Tris-HCI, pH 8.0	15568-025	Invitrogen, UK
1 kb DNA Ladder	N3232S	NEB, UK
100 bp DNA Ladder	N3231S	NEB, UK

2.1.4 Consumables

Borosilicate glass capillaries (OD 1mm, ID 0.78mm)	BF100-7810	Warner Instruments, USA
Electroporation cuvettes 1 mm	EP-01	Geneflow, UK
GeneScreen Plus transfer membrane	NEF987001PK	Perkin Elmer, US
Illustra MicroSpin G25 Columns	27532501	GE Life Sciences, UK
Illustra ProbeQuant G-50 Micro Columns	28903408	GE Life Sciences, UK
MicroAmp Optical 96-well Reaction Plates	N8010560	Applied Biosystems, UK
Microcon-30kDa Centrifugal Filter Unit with Ultracel-30 membrane	MRCF0R030	Millipore, UK

2.1.5 Enzymes

Cloned *Pfu* DNA Polymerase	600353	Agilent Technologies, UK
Gateway BP Clonase™ II Enzyme Mix	11789-100	Invitrogen, UK
Gateway LR Clonase™ II Enzyme Mix	11791-020	Invitrogen, UK
GoTaq® Hot Start Polymerase	M5006	Promega, UK
Herculase Hotstart DNA Polymerase	600310	Agilent Technologies, UK
In-Fusion® HD Cloning Kit	639643	Clontech, UK
PCR Extender System	2200500	5Prime, UK
Protease from *Streptomyces griseus* Type XIV (Pronase)	P5147	Sigma-Aldrich, UK
Proteinase K from *Tritirachium album*	P4850	Sigma-Aldrich, UK
Ribonuclease A from bovine pancreas (RNase)	R4642	Sigma-Aldrich, UK
Restriction Enzymes	various	NEB, UK
T4 DNA ligase	M1801	Promega, UK

2.1.6 Kits

DAB Peroxidase Substrate Kit	SK4100	Vector Laboratories, UK
DIG RNA Labelling Kit SP6/T7	11175025910	Roche Diagnostics, UK
NucleoSpin Gel and PCR Clean-up Kit	NZ74060950	Macherey-Nagel, UK
NucleoSpin Plasmid Kit	NZ74058850	Macherey-Nagel, UK
NucleoBond Xtra Midi Kit	NZ74041010	Macherey-Nagel, UK
mMESSAGE mMACHINE® Kit	AM1340	Ambion, UK

2.1.7 Oligonucleotides

All oligonucleotides described in this thesis were designed with Primer 3 software version 4.0 (Untergasser et al., 2012) and were purchased from Sigma-Aldrich, UK.

2.1.8 Buffers and solutions

-2X BW Buffer: 2M NaCl, 10mM Tris-HCl (pH 7.5), 1 mM EDTA

- 4% Paraformaldehyde (PFA): PFA was dissolved in sterile PBS at 60° with constant stirring.

- E3 Medium: 0.5 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM $MgSO_4$.

- Mammalian lysis buffer: 10 mM Tris-HCl (pH 8.0), 0.1 M EDTA (pH 8.0), 0.5% (w/v) SDS, 20 µg/ml of DNase-free Ribonuclease A from bovine pancreas.

-PBST: PBS, 0.1% Tween-20.

-Southern Blot hybridization buffer: 6× SSC, 0.5% SDS, 5× Denhardt's solution, 10 mg/ml of denatured calf thymus genomic DNA

-WISH Blocking Solution: 1x PBST, 2% sheep serum (vol/vol), 2 mg/ml BSA

-WISH Hybridization Buffer (+):50% deionized formamide, 5X SSC, 0.1% Tween-20, 50 µg/ml of heparin bile salts, 500 µg/ml of extracted RNase-free tRNA adjusted to pH 6.0 by adding citric acid (460 µl of 1 M citric acid solution per 50 ml of hybridization buffer).

WISH Hybridization Buffer (-):50% deionized formamide, 5X SSC, 0.1% Tween-20

-WISH Staining Buffer: 1 M Tris-HCl, adjusted to pH9.5, 50 mM $MgCl_2$, 100 mM NaCl, 0.1% Tween-20.

-WISH Staining Solution: Dilute 225 µl of 50 mg/ml NBT and 175µl of 50 mg/ml BCIP with 50 ml of WISH Staining Buffer (Light sensitive).

-WISH Stop solution: PBS pH 5.5, 1 mM EDTA, 0.1% Tween-20.

-Antibody staining blocking buffer: 10% NBCS in PBST.

2.1.9 Zebrafish strains

Wild-type embryos used are derived from AB and AB* strains. For pancreas co-localization experiments the following stable transgenic lines were used: Tg (ins-mCherry)^{jh2} (Pisharath et al., 2007).

2.1.10 Equipment

Digital microINJECTOR System	MINJ-1	Tritech Research, US
Flaming Brown micropipette puller	P-97	Sutter Instruments, USA
Heraeus Function line incubator	B6	Thermo Scientific, UK
Leica TCS LSI zoom confocal microscope	LSI6000	Leica, UK
MicroPulser Electroporation Apparatus	165-2000	BIO-RAD, UK
NanoDrop ND-1000 Spectophotometer	ND-1000	NanoDrop Technologies
Scan^R High-Content Screening Station for Life Science	scan^r	Olympus, Germany
Sensoquest Thermocycler	613-3908	GeneFlow, UK
Stereoscopic Zoom microscope	SMZ1500	Nikon, UK
Storage Phosphor Screen	28956474	GE Life Sciences, UK
Stratalinker UV crosslinker	400075	Stratagene, US
Stuart Scientific tube rotator	SB3	VWR, UK
Typhoon 9200 Variable Mode Imager	375-635	GE Life Sciences, UK
Zoom stereomicroscope	SMZ645	Nikon, UK
6-Tube Magnetic Separation Rack	S1506S	New England Biolabs, UK

2.2 Molecular Biology Methods

2.2.1 Phenol-chloroform extraction of nucleic acids

Phenol-chloroform extraction is a widely used molecular biology method designed for the purification of nucleic acids (Chomczynski and Sacchi, 1987). One volume of Tris-HCl buffer saturated-phenol was added to the aqueous solution containing nucleic acid and proteins, forming two immiscible layers: an upper aqueous phase and a heavy phenolic phase. The two phases were mixed and separated by centrifugation at 10,000 rpm for 10 minutes at room temperature (RT). The upper aqueous phase containing nucleic acids was transferred to a fresh tube. One volume of chloroform was added, mixed briefly and centrifuged at 10,000 rpm for 10 minutes at RT. The aqueous phase was transferred to a fresh tube for precipitation.

2.2.2 Precipitation of nucleic acids

DNA was precipitated by adding 0.2 volumes of 3M Na-acetate and 2 volumes of 100% ethanol. RNA was precipitated by adding one volume of isopropanol and 1/10 volumes of 10M ammonium acetate. Solutions were incubated at least for 1 hour at -20°C and centrifuged at 4°C for at least 30 minutes. The precipitate was washed with 70-80% ethanol and pelleted at 10,000 rpm for 5-10 minutes. The pellet was air-dried and resuspended in nuclease-free water. DNA was stored at -20°C.

2.2.3 Isolation of plasmid DNA

Isolation of plasmid DNA was carried out using commercially available kits from Macherey Nagel, UK. These kits are based on the alkaline lysis extraction protocol described by (Birnboim and Doly, 1979). Bacteria from liquid cultures are harvested by centrifugation and plasmid DNA liberated by alkaline-SDS lysis. The lysate is neutralised and plasmid DNA containing solution is applied to silica or anion-exchange membrane column. Genomic DNA, proteins and cell debris are removed by centrifugation. Other contaminants, like salts and metabolites, are removed by washing with ethanolic buffers. Purified plasmid DNA is eluted under low ionic strength conditions with alkaline TE buffer or nuclease-free water.

2.2.3.1 Miniprep of plasmid DNA

Bacterial plasmids were purified using NucleoSpin Plasmid Kit (Macherey-Nagel, UK) following manufacturer's instructions, which are described briefly below. Firstly, 5 ml LB cultures supplemented with the appropriate antibiotic were inoculated with a single colony picked from streaked agar plates and were incubated with constant shaking (200-250 rpm) overnight at 37°C. Cells were pelleted in a standard benchtop microcentrifuge for 30 seconds at 11,000 g. After discarding the supernatant, 250 µl of A1 Buffer supplemented with RNase A was added to resuspend the cell pellet. 250 µl of A2 Buffer was then added to lyse the cells and the tube was inverted 6-8 times carefully to avoid shearing of bacterial genomic DNA. After 5 minutes of incubation of the lysate, 300 µl of A3 Buffer were added to neutralize the reaction. The lysate was clarified by centrifugation at 11,000 g for 5 minutes. The supernatant was loaded onto a NucleoSpin Plasmid QuickPure Column and was centrifuged for 1 min at 11,000 g so that the DNA could bind to the silica membrane of the column. 450 µl of A4 Buffer supplemented with 100% ethanol were added to wash the column and after 1 minute of centrifugation at 11,000 g the flow-through was carefully discarded. An additional centrifugation step at 11,000 g for 2 minutes was performed to ensure that the silica membrane was dry. The bacterial plasmid was eluted by adding 30 µl of nuclease-free water to the column, incubating for 1 min and centrifuging for 1 min at 11,000 g. The flow-through was collected in a fresh tube.

2.2.3.2 Midiprep of plasmid DNA

High-quality bacterial plasmid DNA was prepared using NucleoBond Xtra Midi (Macherey-Nagel, UK) following manufacturer's instructions, which are described briefly below. A 200 ml LB culture containing the appropriate selective antibiotic was grown overnight at 37°C and 200 rpm. Cells were harvested by centrifugation at 6,000 g for 10 min at 4°C. After discarding the supernatant, the cell pellet was resuspended with 8 ml of Resuspension Buffer supplemented with RNase A. Then 8 ml of Lysis Buffer were added to the suspension, which was inverted 5 times carefully and incubated for 5 minutes at RT. The reaction was neutralised by adding 8 ml of Neutralization Buffer. The tube was inverted 10 times and the lysate was loaded into a NucleoBond Xtra Column filter previously equilibrated with 12 ml of Equilibration Buffer. After allowing the column to empty by gravity flow, 5 ml of Equilibration Buffer was added to the column to wash out the remaining lysate from the filter. The column filter was discarded and the column was washed with 8 ml of Wash Buffer. Plasmid DNA was eluted with 5 ml of Elution Buffer and collected in a fresh 50 ml centrifuge tube. To precipitate the DNA, 3.5 ml of isopropanol at RT were added to the tube. After vortexing, the precipitate was centrifuged at 10,000 g for 30 minutes at 4°C. The supernatant was removed and the pellet washed with 2 ml of 70% room-temperature ethanol by centrifuging at 15,000 g for 5 min. The DNA pellet was dried at RT and reconstituted with 100-300 µl of nuclease-free water.

2.2.3.3 Isolation of high molecular weight genomic DNA

Genomic DNA was isolated according to the standard protocol described by (Sambrook and Russell, 2006). Snap-frozen zebrafish embryos were thawed on ice and homogenised using a clean pestle. One ml of Mammalian Cell Lysis Buffer (recipe in **section 2.1.8 above**) was added to the samples, and the solution was incubated for 1 hour at 37°C. Proteinase K was added to the lysate to a final concentration of 100 µg/ml and the mixture was incubated from 3h to overnight in a water bath at 50°C at 50-60 rpm. After cooling the solution at RT, phenol extraction was carried out twice. Phases were mixed by rotating the tubes in an end-over-end rotator at 2 rpm for 1 hour and separated by centrifugation at low speed to avoid breaking the DNA molecule (5,000 g for 20 minutes at RT). A wide-bore pipette was used to transfer the viscous aqueous phase to a fresh centrifuge tube and a phenol-chloroform-isoamylalcohol (25:24:1) extraction was performed.

DNA was precipitated by adding 0.2 volumes of 10 M ammonium acetate and 2 volumes of 100% ethanol at RT. The sample was rotated for 15 minutes in an end-over-end rotor at 2 rpm. The precipitated DNA was removed in one piece with a sterile crook and transferred to a solution of 70% ethanol. DNA was washed again in 70% ethanol and was dried at RT on a crook. When the precipitated DNA turned transparent, it was resuspended in 300 µl of Tris- buffer (pH 8.0) and was dissolved by gently rotating the solution overnight.

2.2.4 Quantification of nucleic acids

Nucleic acids yield and purity were assessed by UV spectrophotometry using a NanoDrop ND 1000 Spectrophotometer (Thermo Scientific, UK). The concentration (ng/µl) is based on the 260λ absorbance. The purity of DNA and RNA was assessed by evaluating the 260/280λ ratio of the sample's absorbance. A ratio of 1.8 was accepted as pure for DNA and a ratio of 2.0 was accepted as pure for RNA. The 260/230λ absorbance ratio was used as a second measurement of the nucleic acids purity (absence of co-purified contaminants). A ratio in the range of 1.8 – 2-2 was considered as pure.

2.2.5 Polymerase Chain Reaction (PCR)

Amplification of DNA sequences was performed by PCR using a Sensoquest Thermocycler (GeneFlow, UK). For cloning purposes, the proofreading PCR Extender System (Primer5, UK) for high fidelity amplifications was used. This system combines a blend of highly thermostable DNA Polymerases with high processivity that enable amplification of fragments up to 40 kb with minimal error rate. When high-fidelity was not needed, ie, when testing colonies after ligation reactions, a Taq Polymerase without proofreading capability was used (GoTaq® Hot Start Polymerase, Promega, UK). Reactions were assembled following manufacturer's instructions. Annealing temperatures and extension times were adjusted depending on primer and template conditions. Typically, a standard PCR cycle consists on a hot-start step of 95°C for 5 minutes, followed by 35-40 cycles of denaturing, annealing and extension, and a final extension step at 72 °C for 7 minutes.

2.2.6 Agarose Gel electrophoresis

The size and specificity of PCR products, or digested DNA was assessed in agarose gels by electrophoresis. Gels were prepared by dissolving agarose in Tris-borate-EDTA (TBE) buffer to a final concentration ranging from 0.8 to 1.5% depending on the fragment size. To allow visualization of DNA under UV light 10ng/ml of ethidium bromide was added to the gels. The ladders commonly used were either 100 bp DNA Ladder or 1 Kb DNA Ladder from NEB UK, depending on the size of the fragments to be analyzed. Gels were run in TBE 1X electrophoresis buffer at 90 V for a minimum of 40 minutes or until proper separation of the bands was achieved.

2.2.7 Purification of DNA from agarose gels

PCR products were purified to remove remaining primers and other PCR components for downstream procedures. Also DNA products were excised and extracted from agarose gels when unspecific bands were present. Digested plasmid DNA was purified in order to ensure a higher quality for further procedures (usually ligation or recombination reactions) using the NucleoSpin Gel and PCR Clean-up Kit from Macherey-Nagel, UK.

Gel purification was carried out by loading DNA samples into an agarose gel and the fragment of interest was excised using a clean scalpel under UV light. UV exposure was minimized to avoid damaging the DNA. The gel slice was purified following manufacturer’s instructions. For each 100 mg of agarose 200 µl of NT Buffer was added. The sample was incubated at 50°C for 10 minutes. In the case of PCR clean-up, the volume of the PCR reaction was adjusted to 100 µl and 200 µl of Buffer NT were added. The sample was loaded into a NucleoSpin Column attached to a collection tube and centrifuged at 11,000 g for 1 minute. The flow-through was discarded and 700 µl of NT3

Buffer supplemented with ethanol were added to wash the silica membrane of the column. This was centrifuged for 1 minute at 11,000 g and after removal of the flow-through the column was centrifuged again for 2 minutes at maximum speed to remove residual ethanol. DNA was eluted by adding 15-50 µl of nuclease-free water into the column and centrifuging for 1 minute at 11,000 g. The eluate was collected into a fresh tube.

2.2.8 Molecular cloning

Molecular cloning refers to the construction of plasmids made by directional ligation of DNA fragments flanked by complementary cohesive ends created by endonuclease restriction digest. Inserts coming from PCR products were amplified using primers containing restriction sites.

2.2.8.1 Restriction digest of DNA

Around 10 µg of vector and insert DNA were digested using appropriate Restriction Endonuclease enzymes from New England Biolabs, UK. Reactions were set up in a total volume of either 50 or 20 µl, by adding suitable amounts of DNA, 2 to 10-fold excess of restriction enzyme (2-10 units per µg of DNA), appropriate NEBuffer (10X) and BSA to a final concentration of 100 ng/µl. Unless stated by the manufacturer, the reaction was incubated at 37°C for 2-3h.

2.2.8.2 Ligation of DNA fragments

T4 DNA ligase was chosen to catalyze the ligation of DNA fragments containing cohesive ends generated by restriction digest. Ligation reactions were assembled using a 1:3 vector:insert molar ratio, 1 µL of T4 DNA ligase and 2 µL of 10X Ligase Buffer in a total volume of 20 µL. The reaction was incubated at RT overnight and then 10 µl of this reaction was used to transform an aliquot of 50µl of DH5α cells as described in **section 2.3.1 below**.

2.2.9 Gateway Cloning

Expression vectors containing a human putative enhancer linked to a zebrafish promoter and fluorescent reporter gene were cloned using the Multisite Gateway Cloning System by Invitrogen, UK. The Gateway strategy is based on phage lambda site-specific recombination i*n vitro* between PCR products and vectors containing attB and attP recombination sites (Roure et al., 2007).

Expression Vectors are created through two recombination steps. The BP reaction is a recombination reaction between an attB-site flanked PCR product (either a human enhancer or zebrafish promoter PCR product) and a donor vector that generates a “shuttle vector” or Entry Vector. These entry vectors are recombined with a variety of Destination Vectors to create a final Expression **(Figure 2.1)**.

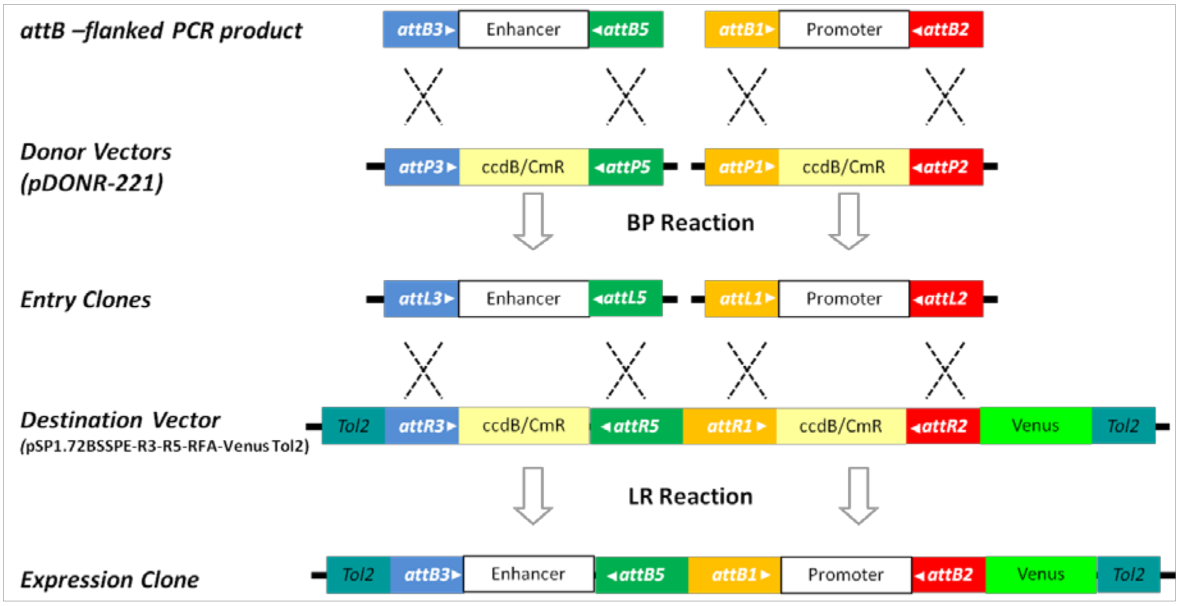


Figure 2.1. Schematic representacion of Gateway Cloning Technology: from attB-flanked PCR products to the generation of an Expression Clone.

2.2.9.1 Production of attB-flanked PCR products

Human enhancers and zebrafish promoters were amplified using the proofreading PCR Extender System for high fidelity amplifications. Human genomic DNA containing disease-linked SNPs or mutations for amplification was provided by our collaborators.

Two rounds of PCR were carried out to produce a final product flanked by attB recombination sites (29bp). In the first round the 12bp most 3'-prime to the gene-specific sequence was added by the adaptor primers. In the second round, the whole 29 bp attB tail was reconstructed by primers containing the complete attB site (**Table 2.1**): attB3 and attB5 sites were used to flank enhancers while attB1 and attB2 sites were flanking promoters.

Table 2.1 Primers used to amplify attB-flanked products for Gateway Cloning

Oligo Name	Sequence (5' to 3')
attB1-primer	ggggacaagtttgtacaAAAAAGCAGGCT
attB2-primer	ggggaccactttgtacaAGAAAGCTGGGT
attB3-primer	ggggacaagtttgtataATAAAGTAGGCT
attB5 –primer	ggggaccactttgtataCAAAAGTTGGGT

After assembling a high-fidelity PCR reaction (**section 2.2.5** above) and assessing the size and specificity of the PCR product by electrophoresis, the PCR product was cleaned-up or gel extracted as described in **section 2.2.7 above**. Purification of the PCR product was carried out to remove attB primers and any attB primer-dimers, which could clone efficiently into the Entry Vector.

2.2.9.2 Creation of Entry Clones via BP Reaction

By the BP reaction attB-flanked PCR products were transferred as a unit into a Donor Vector (pDONR221) to create Entry Clones. Two separate BP reactions were needed to create promoter and enhancer containing Entry Clones.

The BP Reaction was assembled at room temperature as follows:

	Negative Control	Sample
Component	**Tube 1**	**Tube 2**
attB PCR product (use 50 fmol*)	----	1-10 µl
pDONR221 Vector (150 ng/ml)	2 µl	2 µl
BP Clonase Enzyme Mix	2 µl	2 µl
TE buffer	10 µl	To 10 µl

The BP reaction was incubated at 25°C for 1 hour. To stop the reaction 2 µl of Proteinase K were added and incubated for 15 minutes at 37°C. Five µl of BP Reaction were transformed into an aliquot of 50 µl DH5α chemically competent cells, as described in **section 2.3.1**. The next day, single colonies were picked from LB Agar plates and colony PCR using M13 forward and reverse primers (flanking recombination sites) were used for testing positive Entry Clones. Positive Entry Clones were isolated by minipreps (see **section 2.2.3.1 above**) and sent for sequencing (see **section 2.2.11 below**).

2.2.9.3 Creating Expression Vectors via LR Reaction

From the Entry Clones, the non-coding sequences were shuttled by LR recombination to the Gateway Destination Vector pSP1.72BSSPE-R3-R5-RFA-Venus Tol2 (Roure et al., 2007). This vector contains two Tol2 arms which are recognized by Tol2 transposase, enabling single copy integration of the construct into the genome by co-injection with Tol2 mRNA. LR Reactions were assembled at room temperature as follows:

Component	**Negative Control**	**Sample**
Enhancer containing Entry Clones	----	100 femtomoles
Promoter containing Entry Clone		100 femtomoles
Destination Vector	100 femtomoles	100 femtomoles
LR Clonase Enzyme Mix	2 µl	2 µl
TE	Up to 10 µl	Up to 10 µl

The LR reaction was incubated overnight at RT. To stop the reaction 2 µl of Proteinase K was added and incubated for 15 minutes at 37°C. For multisite LR cloning, 10 µl of the LR Reaction was transformed into 40 µl electrocompetent *E. coli* DH5α cells by electroporation, as described in **section 2.3.2**. To screen for positive colonies attB3 forward primer and the Venus reverse primers (5'-AACTCCAGCAGGACCATGT-3') were used. Positive Expression Clones were isolated using NucleoBond Xtra Midi Kit, as per manufacturer's instructions. All final Expression vectors were sent for sequencing.

2.2.10 In-fusion cloning

In-Fusion® HD Cloning Kit system allows directional cloning of one or many DNA fragments into a vector, using only one restriction site. Clontech´s In-fusion enzyme can efficiently fuse DNA fragments that share a 15 bp homology region at both ends. This technology was used when only one restriction site was available within a vector, and directional cloning of the insert was required. Manufacturer´s instructions were followed.

2.2.10.1 Design of primers for PCR

Primers were designed so that the 5'end would contain a 15 bp fragment overlapping with the end of the linearized vector where it was going to be cloned, and the 3´end would contain 18 to 25 bp specific to the target gene. The 3´end region of the oligonucleotide was designed using Primer3 software. For the final primer design an online tool by Clontech was used: http://bioinfo.clontech.com/infusion/**.**

2.2.10.2 Preparation of linearized vector and insert

The vector was linearized using one or two restriction enzymes as described and it was gel extracted using Macherey Nagel Gel and PCR Clean-up Kit. The PCR fragment was amplified using the proofreading hot-start Polymerase provided by the Kit: CloneAmp™ HiFi PCR Premix.

The PCR reaction was assembled as follows:

Component	**Volume**
CloneAmp™ HiFi PCR Premix	0.2 µL (1 unit)
Template DNA	30-100 ng
dNTPs (10 mM)	0.4 µl
Primers containing 15 bp tail (10 µM)	0.5 µl/each
Nuclease-free water	up to 20 µL

The specificity and size of the amplified fragment was checked by agarose gel electrophoresis and then the PCR fragment was purified by gel extraction.

2.2.10.3 In-fusion cloning reaction

Following manufacturer´s recommendations between 50-200 ng of both insert and vector were used to obtain good cloning efficiencies. For inserts smaller than 0.5 Kb, less than 50 ng were used and for vectors larger than 10 Kb more than 200 ng were used. The cloning reaction was assembled as follows:

Component	**Negative Control**	**Sample**
Purified PCR fragment (insert)	----	10-200 ng
Linearized vector	50-200 ng	50-200 ng
5X Fusion HD Enzyme Premix	2 µl	2 µl
Nuclease-free water	Up to 10 µl	Up to 10 µl

This reaction was incubated at 50°C for 15 minutes and then placed on ice. Then 2.5 µl of the reaction was used for transformation.

2.2.10.4 Transformation using Stellar™ Competent Cells

An aliquot of 50 µl of Stellar™ Competent Cells was used to transform 2.5 µl of the cloning reaction. After thawing the cells, the mixture was incubated on ice for 30 minutes. Then, the cells were heat-shocked for 45 seconds in a 42°C water bath and placed on ice for 3 minutes. Cells were diluted in 500 µl of pre-warmed SOC Medium and then recovered for one hour at 37°C in a shaker at ~250 rpm. 1/10th, 1/25th and the remaining transformation volume were used to streak

LB Agar plates containing the appropriate selective antibiotic. Colony PCR was performed the next day to screen for positive colonies.

2.2.11 Sequencing

To ensure no mutations were introduced during PCR reactions or after cloning, all plasmids produced and critical PCR products were sequenced by an external service (Beckman Coulter Genomics, UK). For each sequencing reaction, 1.5µg of each plasmid DNA or 250 ng of a purified PCR product together with 10 µl of appropriate primers (5 mM) were barcoded and sent for Quick Lane Express Sequencing to Beckman Coulter Genomics (Takeley, UK).

2.2.12 Site-directed mutagenesis

Site-directed mutagenesis by PCR was used in order to generate a PROX1 putative human enhancer carrying the non-common variant of rs3242786 SNP (G>A). This point mutation was generated following Higuchi*'s* method (Higuchi et al., 1988) using PROX1 element Entry Clone as a template. Higuchi*'s* method is based on two PCR rounds: in the first round two overlapping DNA fragments carrying the desired mutation are produced (**Figure 2.2**) using primers specified in **Table 2.2**. Mutations are introduced via primer mismatch. Then, products from the first round serve as a template for the second round and produce a duplex fragment that can be extended by a standard DNA polymerase after denaturation and renaturation.

Table 2.2 Primers used to create a single site mutation in PROX1 element

Primer name	Sequence (5' to 3')	PCR Round
PROX1-pF1	GCAAAAATGAACTTGAGAAATCC	First round: PCR product A
PROX1-G>A-pR	TGATTACAAAGA**T**GATAATTTATGACTGACATAC	
PROX1-G>A-pF	TCATAAATTATC**A**TCTTTGTAATCATTAAGGATC	First round: PCR product B
PROX1-pR1	CATTCCCTTTAATATCCCATGC	
PROX1-pF2	GCAAAAATGAACTTGAGAAATCC	Second round
PROX1-pR2	CATTCCCTTTAATATCCCATGC	

For the first round of PCR, the following conditions were used:

95°C	95°C	60°C	72°C	95°C	65°C	72°C	72°C
2 min	20s	20s	30s	30s	15s	30s	5 min
	Repeat 10 times			**Repeat 20 times**			

Products A and B from the first round were purified from an agarose gel as described above. For the second round of PCR 7.5 µl of the purified A and B products were mixed and serve as template after denaturalization for 95 ºC for 10 minutes. The following conditions were used:

95°C	95°C	50°C	72°C	95°C	65°C	72°C	72°C
2 min	30s	30s	50s	30s	10s	30s	5 min
	Repeat 10 times			**Repeat 20 times**			

The final PCR product bearing the mutation was sequenced and was used to create a Gateway Expression Vector following the protocol detailed in **section 2.2.9.3 above**.

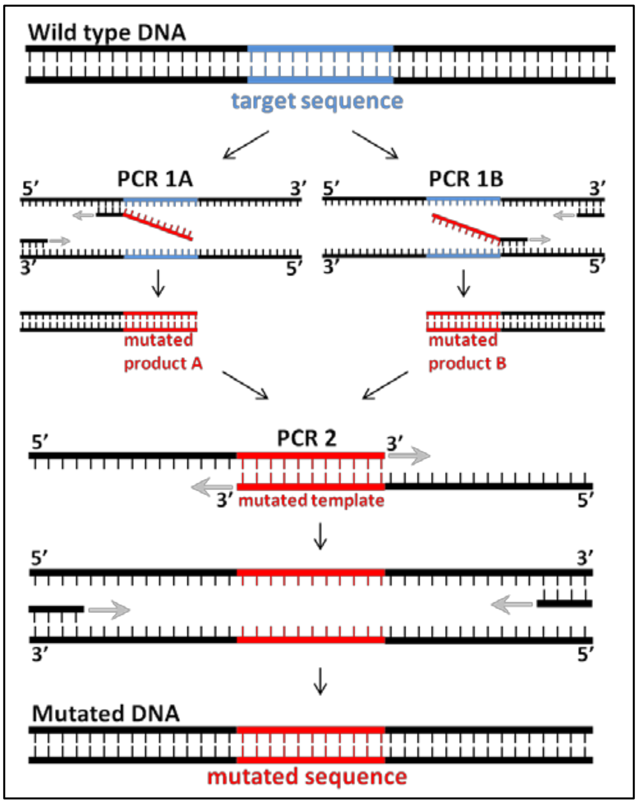
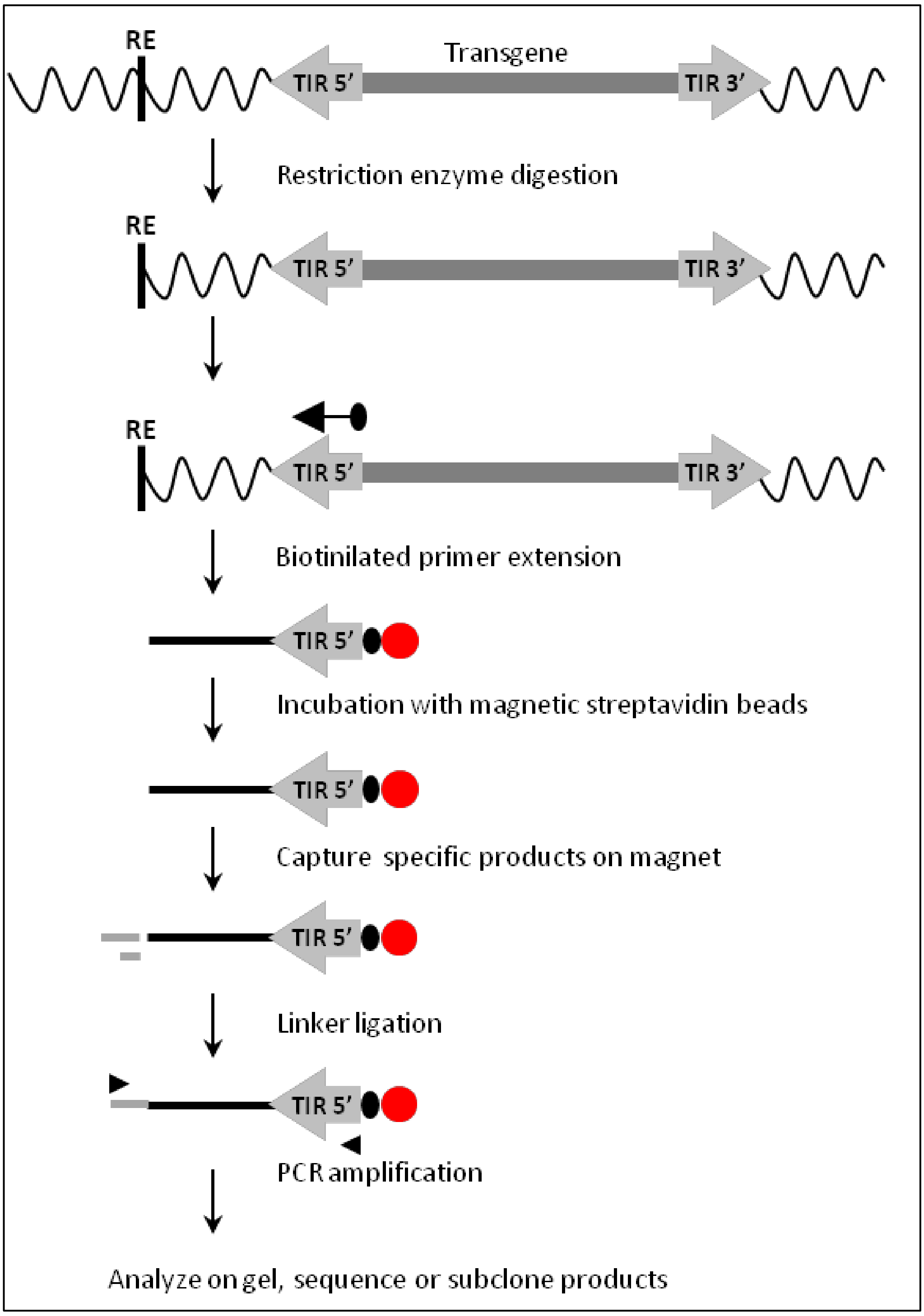


Figure 2.2. Schematic representation of Higuchi´s Method used for site-directed mutagenesis. Adapted from (Hadzhiev, 2007)

2.2.13 Extension primer tag selection linker-mediated PCR (EPTS-LMPCR)

In order to map the genomic insertion site of a transgene in several zebrafish transgenic lines generated in the lab, I used a modified linker-mediated PCR protocol adapted from (Yergeau et al., 2007). In standard linker-mediated PCR (LMPCR) protocols, genomic DNA is digested by restriction enzymes and linkers are ligated to the DNA fragments. PCR is then carried out using primers specific to the adaptors and to the terminal ends of the transposon. A variation of LMPCR is Extension Primer Tag Selection (EPTS)-LMPCR, where the fragmented genomic DNA regions containing the transgene are purified from the rest of genomic DNA fragments using a biotinylated primer specific to the transposon arms (**Figure 2.3**). Briefly, genomic DNA is digested with a restriction enzyme that does not cut in the transposon terminal end. A transposon-specific biotinylated primer is added and DNA is extended using a DNA polymerase that generated biotinylated products, which can be specifically purified by streptavidin beads. Genomic DNA fragments that do not contain the transposon are therefore removed. Through several steps, excess adaptors are washed away and the purified biotinylated product is used as a template for one or two rounds of PCR that used transposon and linker specific primers. The resultant product(s) are sequenced or subcloned in order to identify the genomic DNA sequences flanking the transposon integration site.

The following protocol was followed: two µg of high-quality genomic DNA from the zebrafish transgenic lines of interest were digested at 37°C overnight with either AluI (for mapping the 3' end of Tol2 transgene) or DpnII (to map the 5'end). 1/10th of 3M sodium acetate and 3 volumes of 100% ethanol were added and the samples were incubated on dry ice for 30 minutes. Samples were centrifuged at 16,000g for 10 minutes at RT and all liquid was removed and DNA pellet was air-dried.



RE
Transgene
TIR 5'
TIR 3'
Restriction enzyme digestion
RE
TIR 5'
TIR 3'
RE
TIR 5'
TIR 3'
Biotinilated primer extension
TIR 5'
Incubation with magnetic streptavidin beads
TIR 5'
Capture specific products on magnet
TIR 5'
Linker ligation
TIR 5'
PCR amplification
Analyze on gel, sequence or subclone products

Figure 2.3. Overview of EPTS-LMPCR protocol.
After digesting genomic DNA with restriction enzymes (RE), a biotinylated primer (black circle) annealing the Tol2 inverse repeats (TIR) is used to extend DNA fragments containing the transgene insertion. These products are purified using streptavidin beads (red circle), adaptors are linked (grey lines), and amplification using oligos specific to TIRs and linkers takes place (black triangles). Resultant PCR products are either sequenced or subcloned to identify the genomic sequence flanking the transposon integration site. Figure adapted from (Yergeau et al., 2007).

The primer extension reaction was set up on ice as follows:

Component	**Volume**
10X Pfu DNA Polymerase reaction buffer	5 µL
dNTPs (10 mM)	1 µL
Biotinylated Tol2 primer (0.125 µM)	2 µL (250 µM)
Cloned Pfu DNA Polymerase	1 µL (2.5 units)
Nuclease-free water	41 µL

DNA pellet was resuspended in 50 µl of primer extension mix and carefully mixed. Each DNA and primer extension mix was transferred to a thin-walled 200 µl PCR tube and it was incubated in a thermocycler using the following parameters: 98°C for 3 minutes, 68°C for 30 minutes and 72°C for 30 minutes. 450 µl of nuclease-free water was added to each tube and the extended products were purified using a Millipore Microcon-30 spin column. Briefly, samples were spun for 12 minutes at 12,000 g at RT to remove excess of primers and enzymes. In the elution step, the column insert was inverted and inserted in a fresh tube. A small concentrated volume ranging from 5 to 20 µl was visible after spinning for 3 minutes at 1,000 g at RT. Forty µl of nuclease-free water were added to each sample and they were kept at RT while streptavidin beads were processed.

For each reaction 200 µg of beads were used. The amount of beads required to process all reactions was transferred to a fresh 1.5 ml eppendorf tube and it was placed on a magnetic holder for 1 minute. All liquid was removed with a pipette tip and the tubes were removed from the magnet holder. An equal volume of 2X BW buffer was added, mixed and placed in the magnet. Beads were washed this way for a total of 3 times. After the third was, beads were resuspended at a concentration of 200 µg per 40 µl of 2X BW buffer.

Forty µl of streptavidin beads (Dynabeads® M-280) were mixed with 40 µl of the primer extension product and the mixture was rotated in an end-to-end rotator for at least 30 minutes at RT, followed by 30 minutes of shaking at approximately 200 rpm.

The beads were washed twice with 100 µl of nuclease-free water using the magnetic holder in order to remove genomic DNA fragments not containing the transgene, and therefore, were not bound by streptavidin beads. After the second wash, beads were resuspended in 10 µl of ligation mix, which was prepared as follows:

Component	**Volume**
10X T4 DNA ligase buffer	1 µL
50 pmol/l of annealed linker cassette	2 µL
T4 DNA ligase	2 µL (6 units)
Nuclease-free water	5 µL (final volume of 10 µL)

Ligation reaction was incubated overnight at 16°C. The following morning beads were washed twice with 100 µL of nuclease-free water using magnetic holder as described. After the second wash, magnetic beads were resuspended in 10 µL of nuclease-free water and the first round of PCR was set up as follows:

Component	**Volume**
Resuspended magnetic beads	1 µL
dNTPs (10 mM)	1 µL
OCI primer	2 µL (500 mM)
Tol2 5N1 primer	2 µL (500 mM)
10X Herculase Hot Start Polymerase buffer	5 µL
Herculase Hot Start Polymerase	0.5 µL (2.5 units)
Nuclease-free water	38.5 µL

The following thermocyler program was used to amplify the product:

Cycle number	**Denature**	**Anneal**	**Extend**
1	92°C for 2 min	---	---
2-11	94°C for 30 sec	50°C for 30 sec	72°C for 2 min
12-31	94°C for 30 sec	50°C for 30 sec	72°C for 2 min plus 10 sec per cycle
32			72°C for 5 min

After amplification 15 µL were run on a 1.25% agarose gel in 1X TAE. If products were visible, all bands from each sample were purified, as described above. If no products were visible or present at very low concentrations, a nested PCR was carried out using the following components:

Component	Volume
Primary PCR product	1 µL
dNTPs (10 mM)	1 µL
OCII primer	2 µL (500 mM)
Tol2 5N2 primer	2 µL (500 mM)
10X Herculase Hot Start Polymerase buffer	5 µL
Herculase Hot Start Polymerase	0.5 µL (2.5 units)
Nuclease-free water	38.5 µL

Nested PCR was carried out using the following thermocyler program:

Cycle number	Denature	Anneal	Extend
1	92°C for 2 min	---	---
2-11	94°C for 30 sec	48°C for 30 sec	72°C for 2 min
12-31	94°C for 30 sec	48°C for 30 sec	72°C for 2 min plus 10 sec per cycle
32			72°C for 5 min

The oligonucleotides used in this protocol are listed in **Table 2.3**.

Table 2.3 Primers used in EPTS-LMPCR protocol.

Oligo name	Sequence (5´-3´)	Modification	Description
BIO-Tol2-3'end	AAACTGGGCATCAGCGCAATTCAAT	5´ biotinylation	Used for primer extension reaction of Tol2 3'end
BIO-Tol2-5'end	ATTCCATGGATATCAAGCTTAAACA	5´ biotinylation	Used for primer extension reaction of Tol2 5'end
NZ117	GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG	None	Linker primer
NZ118-P	CCTAACTGCTGTGCCACTGAATTCAGATCTCCC	5´phosphorylation	Linker primer phosphorylated on the 5'end to ligate extended fragments
Tol2-N1-3'end	CGCAATTCAATTGGTTTGG	None	Primer specific to Tol2 5'end (used in primary PCR)
Tol2-N3-5'end	aAGCTTAAACAAGAATCTC	None	Primer specific to Tol2 3'end (used in primary PCR)
OCI	GACCCGGGAGATCTGAATTC	None	Primer specific to linker sequence (used in primary PCR)
Tol2-N2-3'end-nested	GCAAGGGAAAATAGAATGAAG	None	Primer specific to Tol2 5'end (used in nested PCR)
Tol2-N4-5'end-nested	TCTTTCTTGCTTTTACTTTTA	None	Primer specific to Tol2 3'end (used in nested PCR)
OCII-nested	AGTGGCACAGCAGTTAGG	None	Primer specific to linker sequence (used in nested PCR)

2.2.14 Capped mRNA *in vitro* transcription

Tol2, CFP and PhiC31 integrase mRNA for microinjections were *in vitro* transcribed from pCS2+-Tol2, pCS2+-CFP and pCS2+PhiC31 plasmids respectively using mMESSAGE mMACHINE® Kit from Ambion, UK. Linearized plasmid DNA was used as a template for transcription. Digestion of pCS2+ plasmids was carried out with an appropriate digestion enzyme cutting downstream of the polyA of the insert to be transcribed (NotI). After linearization the plasmid DNA was purified using phenol-chloroform extraction, as described in **section 2.2.1 above**), and precipitated with 3M Na acetate and ethanol and resuspended in nuclease-free water. The capped transcription reaction was assembled at room temperature as follows:

Component	**Volume**
2X NTP/CAP	10 µl
10X Reaction Buffer	2 µl
Purified linear plasmid DNA	100ng-1 µg
Enzyme Mix	2 µL
Nuclease-free water	up to 20 µL

The reaction was mixed thoroughly and incubated for 2-4 hours at 37°C. To remove plasmid DNA template, 1 µl of Turbo DNase was added and it was incubated for 15 minutes at 37°C. To remove the enzymes as well as unincorporated nucleotides, transcribed RNA was further purified by phenol-chloroform extraction, as described. First, 115 µl of nuclease-free water and 15 µl of Ammonium Acetate Stop Solution were added to the reaction and were thoroughly mixed. RNA was extracted with 1 volume of 5:1 phenol: chloroform equilibrated with citric acid (pH= 3.8). RNA was precipitated by adding isopropanol and RNA was eluted in nuclease-free water. RNA correct size and purity was checked by agarose electrophoresis and spectophotometry.

2.3 General microbiological methods

2.3.1 Transformation using chemically competent cells

Retransformation of existing plasmids was performed using an aliquot of 30 µl Alpha-Select Bronze Efficiency DH5α chemically competent cells. Transformation of ligation reactions was performed using an aliquot of 50µl of Alpha -Select Silver Efficiency DH5α chemically competent cells or 100µl of Stellar™ Competent Cells following manufacturer's instructions.

Between 10 and 50 ng of DNA or 2.5-10 µl of ligation reactions were incubated on ice for 30 minutes, heat-shocked in a 42°C water bath for 45 seconds and placed on ice for 3 minutes. Cells were recovered with 1 ml of LB Broth 1X and incubated at 37°C for at least 1 hour in a shaker at approximately 300 rpm. 100-500 µl of cell transformation mixture was plated by spreading on fresh LB Agar plates containing the appropriate selective antibiotic. Plates were incubated overnight at 37°C.

2.3.2 Transformation of electrocompetent cells

For multisite LR cloning DH5α electrocompetent cells were used. Ten µl of the LR Reaction were transformed into an aliquot of 40 µl cells. The mixture was incubated on ice for 15 minutes in a 0.1 cm cuvette and was electroporated using a voltage of 1.8kV in a BIO-RAD Micropulser. Cells were recovered with 1 ml of LB Broth 1X and incubated at 37°C for at least 1 hour in a shaker at ~200 rpm. All the cells were plated by spreading onto fresh LB Agar plates the appropriate selective antibiotic. Plates were incubated overnight at 37°C.

2.4 Fish husbandry and embryo methods

2.4.1 Zebrafish husbandry

Zebrafish were kept in the BMSU according to Home Office Regulations. Seven to fifteen pairs of adult zebrafish were maintained in 3.5 litre polycarbonate tanks in a ZebTEC recirculating housing system (Tecniplast, UK) on a regular light-dark cycle, with 14 hours of light and 10 hours of darkness. Water temperature was kept at 26°C. Adults were fed three times a day combining ZM Medium Premium Granular dry food and hatched brine shrimp cysts (ZMSystems, UK).

2.4.2 Production of zebrafish embryos and eggs

Crosses of adult zebrafish were carried out in 1 litre breeding cages containing an inlay with a bottom mesh, which allowed for embryo collection and prevented the adult fish from eating their eggs. In the evening, one female and one male were placed in a crossing cage separated by a divider and left overnight. Fertilized eggs were collected by filtering the water in each crossing cage with a net.

2.4.3 Raising zebrafish larvae

Zebrafish embryos were kept in 90 cm Petri dishes in E3 Medium supplemented with gentamicin (0,01%) in an incubator at 28.5 ºC for 5 days after fertilization (Westerfield, 1993). Medium was changed every day and dead embryos were removed to avoid bacterial infections. PTU was added to E3 Medium (0, 03%) in embryos older than 24 hpf, when removal of pigment was necessary to allow visualisation.

2.4.4 Dechorionation

Embryos were dechorionated before 48 hpf by adding 1 ml of Pronase stock solution (10 mg/ml) to 10 ml of E3 medium. After swirling the plate, the enzymatic reaction was carefully controlled under the microscope. When 2/3 of the embryos were out of the chorion, they were washed at least three times with fish water to remove all traces of the enzyme and placed in fresh plates with E3 Medium supplemented with gentamicin. Alternatively, embryos older than 24 hpf were dechorionated manually using fine and sharp forceps.

2.4.5 Microinjections

Fertilized embryos were microinjected using an analogue pressure-controlled microinjector from Tritech Research, US. Needles for microinjection were pulled using a Flaming Brown needle puller. Microinjection solutions contained 15 ng/µl of high-quality plasmid DNA, 30 ng/µl of CFP mRNA (when appropriate), 15 ng/µl of Tol2 mRNA or 15 ng/µl of PhiC31 integrase mRNA (unless otherwise indicated), filtered phenol red solution (0,2%) and nuclease-free water up to 10 µl.

Embryos within 10 minutes of fertilization were collected in Petri dishes and water was removed, so that just a fine layer of water was covering the eggs. Under the control of the stereomicroscope eggs were injected through the chorion with approximately 1 nl of microinjection solution. Successful microinjections were visible due to phenol red solution. Immediately after injections, embryos were transferred to a dish containing E3 Medium and gentamicin and were incubated at 28,5 ºC for a maximum of 5 days. Embryos expressing high levels of CFP mRNA at 24 hpf were considered correctly injected and further analyzed.

2.4.6 Generation of stable transgenic lines

Injected embryos were returned to the fish facility and raised as a stock to screen for germline transmission of the transgenes. Approximately 100 embryos injected with each construct were raised to create stable transgenic lines. Larvae and young adults were maintained in breeding cages for the first 3 weeks and fed at least three times a day with fry diets (ZM-000, ZM-100 and ZM-200; ZMSystems, UK), in combination with live food (paramecia, ZMSystems, UK). After three weeks, fish were moved to regular 3.5 tanks in a regular recirculating housing system (ZebTEC, Tecniplast).

2.5 Whole-mount in situ hybridization (WISH)

Whole-mount in situ hybridization (WISH) is a widely used technique to describe the pattern of expression of developmental genes in preserved organisms (Herrmann, 1991; Schulte-Merker et al., 1992). It allows the detection of mRNA in whole embryos by using an antisense RNA probe, which is synthesized *in vitro* and labelled with digoxygenin-UTP. After permeabilization of the embryos and hybridization of the probe, the transcript of interest is visualized by an anti-digoxygenin antibody conjugated to alkaline phosphatase, which catalyzes a chromogenic reaction. WISH was carried out following the protocol described by (Thisse and Thisse, 2008).

2.5.1 WISH RNA probe synthesis

Antisense RNA probes were labelled with digoxigenin-UTP by *in vitro* transcription with SP6 or T7 polymerases using a DIG RNA labelling kit from Roche Diagnostics. As a template, constructs that contained SP6 or T7 promoters downstream the DNA to be transcribed were used. Vectors were linearized with a suitable enzyme and the RNA polymerases were used to produce a run-off transcript. Labelling reaction was assembled as follows:

Component	**Volume**
10X dNTP labelling mixture	2µl
10X transcription buffer	2µl
RNase inhibitor	1µl
SP6/T7 RNA Polymerases	2µl
Linearized purified DNA template	1 µg

These components were mixed and incubated at 37 °C for 2 hours. DNA template was removed by incubating the sample with 2 µl of RNase-free DNase I for 15 min at 37 °C. This reaction was stopped by adding 2 µl of 0.2 M EDTA (pH 8.0). The DIG-labelled probe was further purified by filtering it through resin based Illustra MicroSpin G25 Columns. Probes were quantified and their size and quality were checked by agarose electrophoresis.

2.5.2 Fixation of embryos

Dechorionated embryos of the appropriate developmental stages were fixed in 4% paraformaldehyde diluted in PBS overnight at 4°C. The next day, embryos were dehydrated by replacing PBS with methanol. Washes of 75%, 50% and 25% PBS-Methanol were performed for 5 minutes each. Embryos were washed twice in 100% methanol and stored at -20°C.

2.5.3 Permeabilization of embryos

Embryos were rehydrated in successive dilutions of methanol in PBST (75%, 50%, 25% methanol-PBST). Washes were performed at RT for 5 minutes each with gentle agitation (40 rpm). Embryos were washed 4 times in PBST, 5 minutes per wash. Embryos were permeabilized by digestion with proteinase K (10 µg/ml) at RT for 10 minutes if they were 24 hpf and for 30 minutes if they were older than 24 hpf.

2.5.4 Hybridization of embryos

Digestion was stopped by incubating the embryos for 20 minutes in 4% PFA. Residual fixative was removed by washing 4 times in PBST. Embryos were pre-hybridized in 700 µl of hybridization buffer for 2-5 h in a 70°C water bath. Embryos were hybridized overnight at 70°C, in 200 µl of hybridization mix containing 50-100 ng of DIG-labelled RNA probe (for recipe see section 2.1.8 above).

2.5.5 Washes and incubation with anti-DIG antibody

The next day, hybridization buffer was replaced with 2X SSC through a series of 10 minutes washes at 70°C with hybridization buffer (-) diluted in 2X SSC with gentle agitation (for recipe see **section 2.1.8 above**). Embryos were washed twice in 0.2X SSC at 70°C for 30 minutes. Next, 0.2 X SSC was replaced with PBST through a series of 10 minutes washes at RT with 0.2X SSC diluted in PBST. In order to prevent non-specific binding of the antibody, embryos were incubated in blocking buffer for 3-4 hours at RT with gentle agitation. Anti-DIG antibody was diluted at 1:4000 with blocking buffer and incubated overnight at 4°C on a horizontal orbital shaker.

2.5.6 Washes and staining

The antibody solution was discarded and embryos washed with gentle agitation in PBST 4 times, for 30 minutes each wash. Embryos were washed twice with staining buffer for 5 minutes. In order to stain the embryos, they were transferred to a 24-well plate and staining buffer was replaced with 1 ml of fresh staining solution. The colorimetric reaction was monitored closely under the microscope and embryos were kept in the dark between checks.

When the staining intensity and pattern desired was reached, the reaction was stopped by adding stop solution for 3 minutes. Embryos were washed 3 times in PBST and then fixed in PFA 4% for 30 minutes at 4°C and dehydrated through a series of washes with methanol diluted in PBST.

2.5.7 Embryo mounting

Dehydrated embryos were mounted in 100% glycerol. Methanol was replaced with glycerol through a series of washes with glycerol diluted in water. Washes were performed at RT for 10 minutes with gentle agitation. For imaging, embryos were placed in a small glycerol drop on glass slide, covered with a glass coverslip and oriented accordingly.

2.6 Antibody staining

Embryos were fixed, dehydrated and permeabilized as described above. After proteinase K digestion, embryos were washed 4 times in PBST for 30 minutes with gentle agitation and then embryos were blocked for at least 1 hour in blocking buffer (for recipe see **section 2.1.8 above**). The primary antibody was added and incubated overnight at 4°C with gentle agitation. For vasa staining 1:500 dilution of anti-vasa antibody was used, a kind gift of Holger Knaut (Knaut et al., 2000). The following day, the primary antibody was removed and embryos were washed 4 times in PBST for 30 minutes with gentle agitation. Embryos were blocked for at least 1 hour in blocking buffer, and then incubated at 4°C with the secondary antibody overnight. The embryos were washed 4 times in PBST for 30 minutes. For secondary antibodies conjugated with HRP, detection was carried out with DAB Staining Kit, following manufacturer's instructions. Embryos were briefly incubated in DAB staining solution in the dark. The reaction was carefully monitored under the brightfield microscope and was stopped by washing the embryos twice with PBST. For long-term storage, embryos were then fixed in PFA 4% and dehydrated in methanol as described above.

2.7 Southern blotting

Southern blot was carried out with invaluable technical help of Elizabeth Marsh (University of Birmingham, UK) following the protocol described in (Wilson et al., 2007). Briefly, 5 µg of high-quality genomic DNA from clutches of GFP-positive and GFP-negative 5 dpf embryos from an F3 outcross of F3 Tg(Xla.crygc:attP-GFP)*uobL6* and Tg(Xla.crygc:attP-GFP)*uobL12* lines were digested with an excess of the appropriate enzymes at 37°C overnight. To control for copy number integration, the plasmid used to generate these lines: pTol2/Xla.crygc:attP-GFP (pDB896) vector (Roberts et al., 2014) was spiked in WT genomic DNA and digested in parallel to the samples. Digested DNA was separated on 0.8% agarose gel at 45 V overnight and visualized quickly the morning after. The gel was then washed twice for 20 minutes in 0.25M HCl, and twice for 30 minutes in 0.4 M of NaOH with gentle shaking. DNA was transferred to a GeneScreen™ nylon membrane (PerkinElmer, MA) by capillarity overnight in 0.4 NaOH solution. The membrane was washed twice in 2X SSC and DNA was crosslinked to the nylon membrane using a Stratalinker® UV crosslinker (Stratagene, US). The membrane was blocked with 5 ml of hybridization buffer (recipe in **section 2.1.8 above**) containing 10 mg/ml of denatured calf thymus genomic DNA for a minimum of 60 minutes in a rotating oven at 42°C.

The probe containing attP-GFP and Tol2 3’arm was generated by digesting pTol2/Xla.crygc:attP-GFP (pDB896) vector with ApaI and BamHI enzymes and purified as described in **section 2.2.7**. Fifty ng of the probe was labelled with [α-32P]-CTP (Perkin Elmer, UK) using Ready-To-Go DNA Labelling Beads (GE Life Sciences, UK) and purified using Illustra ProbeQuant G-50 Micro Columns (GE Life Sciences, UK). The labelled probe was denatured and added to the hybridization buffer. The membrane was incubated with the radiolabelled probe at 42°C overnight. The following morning, it was washed at RT for 15 minutes in 2X SSC/0.1% SDS and then with 0.1X SSC/0.1% to remove excess background. Finally, it was wrapped in plastic film and exposed to autoradiograph film for at least 16 hours. To analyze the copy number of the transgenic recipient lines, the nylon membrane was blocked by light for 30 minutes and transferred to a Storage Phosphor Screen (Molecular Dynamics, UK). It was scanned on a Typhoon 9200 Variable Mode Imager (Amersham Biosciences, UK) with Typhoon Scanner Control software (v5.0) and analyzed using ImageQuant 5.1.

2.8 Fluorescence imaging methods

Injected embryos were individually screened for transgenic expression using a Nikon SMZ1500 epifluorescence microscope. Agarose coated plates were used for orienting anaesthetised embryos. Relevant expression patterns were documented using NIS Elements imaging software.

For high resolution imaging of mosaic or stable transgenic expression Olympus ScanR high content screening microscope was used. A brass template adapted for 96-well plate was employed to orient and image 24-72 hpf embryos (Peravali et al., 2011). 96-well plates were coated by 60 µl of 1.5% liquid agarose. The level was made even by adding 20 µl of 100% ethanol. The brass template was placed and agarose was allowed to set for around 30 minutes. After cooling down, template was removed and ethanol was washed away by rinsing the plate several times with tap water. Individual anesthetized embryos in 0.03% MESAB were then pipetted into each well with a cut yellow tip in a volume of 100 µl of medium. They were oriented with a bent 19 G hypodermic needle attached to a plastic Pasteur pipette under the stereomicroscope.

Stacks containing 100 slices of 4 µm thickness were taken in both fluorescence and brightfield channels. For fluorescence channels, out of focus slices were manually removed and maximum projections were made using an extended depth of field plug-in for ImageJ software (Forster et al., 2004). Brightfield images were processed using Photoshop CS6 (Adobe). For high resolution imaging of whole embryos six overlapping stacks were taken. Each stack was processed separately and afterwards reconstructed using a "Grid and Stitch" plug-in for Image J software (Preibisch et al., 2009). Contrast and brightness was adjusted linearly afterwards using Photoshop CS6 (Adobe).

2.9 Statistical analysis

Statistical analyses were carried out using GraphPad Instat Software version 3.05 for Windows (GraphPad Software, US). Chi-square test or Fisher's exact test where used to compare rates of expression patterns between injection groups. P values smaller than 0.05 were considered to be statistically significant.

Chapter Three: FUNCTION OF ISLET-SPECIFIC HUMAN CANDIDATE ENHANCERS IN ZEBRAFISH

Foreword:

The results presented in this chapter have been partially published in (Pasquali et al., 2014).

This project is the result of a collaboration with Lorenzo Pasquali and Jorge Ferrer, affiliated to IDIBAPS (Spain) and Imperial College London (UK), and was funded by Marie Curie BOLD ITN.

The computational analysis was carried out by Lorenzo Pasquali.

3.1 INTRODUCTION

3.1.1 Type-two diabetes as a disease model

Type 2 diabetes (T2D) is a complex metabolic disorder that accounts for approximately 90% of all cases of diabetes and affects hundreds of millions of people worldwide (Stitzel et al., 2010). It is characterized by hyperglycaemia (elevated blood sugar) and insulin resistance. Insulin is the key hormone for regulation of blood glucose levels. It is secreted by β-cells of the pancreatic islets, which take glucose from blood and metabolize it to produce energy (**Figure 3.1**, (Stumvoll et al., 2005)). Normally, the pancreatic β-cells can adapt to changes in insulin by up-regulating the levels of secretion. However, when the insulin production by β-cells is unable to meet the metabolic demand of peripheral tissues diabetes occurs (Oliver-Krasinski and Stoffers, 2008). Type-one diabetes is characterized by the autoimmune attack of beta cells by auto-reactive T cells, resulting in deficient insulin production and a severely reduced β-cell mass (Stankov et al., 2013). Type-two diabetes on the other hand is characterized by insulin resistance, which implies a reduced ability

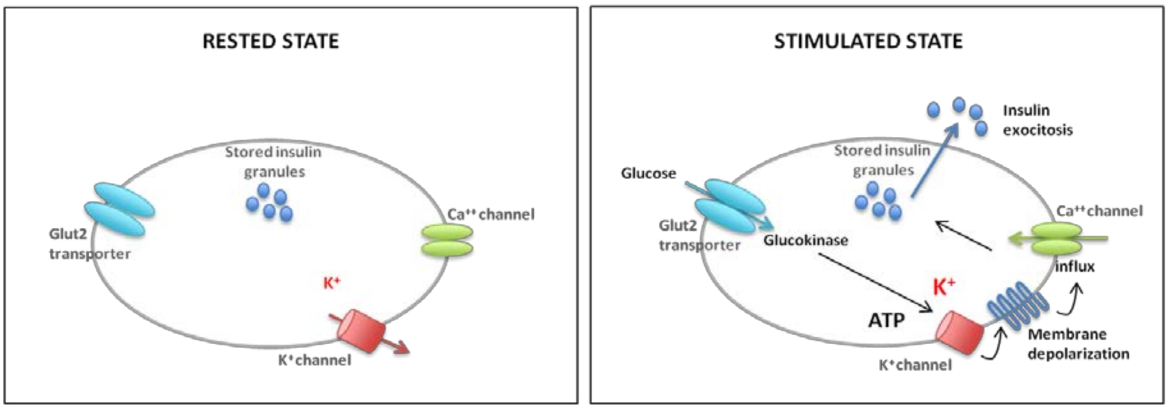


Figure 3.1 Regulation of insulin secretion in a pancreatic beta cell.
Beta cells sense concentration of glucose and regulate insulin secretion. In the resting state, the ATP-sensitive K_{ATP} channel is open, the beta-cell membrane is hyperpolarized and the voltage-gated calcium channel is closed. Upon ingestion of food, the serum glucose levels are elevated and glucose enters the beta cell through a non-insulin dependent GLUT2 transporter and becomes phosphorylated by the glucokinase enzyme. Further glucose metabolism in the mitochondria produces ATP, which is involved in membrane depolarisation and closure of the potassium channel. The change in ADP/ATP ratio induces the closure of the K+ channel. This depolarizes the beta-cell membrane, opening the Calcium channel, and triggering the exocytosis of insulin. Image adapted from (Sperling, 2005).

of the beta cells to respond to insulin, and which leads to disruption of function and to a modest beta cell mass loss (Mastracci and Sussel, 2012).

3.1.2 Type 2 diabetes as a disease model to unravel *cis*-regulatory networks

Unravelling the mechanisms that govern tissue-specific *cis*-regulation is essential to understand development and disease. In the case of endocrine pancreas and beta cell function, the motivation is underlined by the fact that epidemic diabetes is becoming a global health challenge. The prevalence of diabetes mellitus has more than doubled over the last 30 years, accounting in 2010 for more than 285 million affected people worldwide, 90% of whom suffer from type-two diabetes (Chen et al., 2012). Type-two diabetes (T2D) is a complex multifactorial metabolic disorder that results from the interplay between genetic, environmental and behavioural factors, such as smoking, obesity, sleeping disorders and depression (Chen et al., 2012). Around 60 T2D susceptibility loci have been identified so far (Morris et al., 2012), however it is calculated that this susceptibility can only explain between 5 and 10% of the risk (Bramswig and Kaestner, 2014).

3.1.3 Transcriptional regulation in the pancreatic islet

There are several key transcription factors that control both pancreas development and β-cell specification. Some of them, including PDX1, PTF1A, PAX6, and NKX2.2 are also conserved in zebrafish (Biemar et al., 2001). PDX1 or Pancreas Duodenum Homeobox-1 is a marker of pancreatic progenitors essential for pancreas development and β-cell differentiation and it is conserved from human to zebrafish (Milewski et al., 1998). *Pdx1* regulates the expression of several key genes in the beta-cell such as *insulin* and *glucagon* by binding directly to their promoters (Leonard et al., 1993; Petersen et al., 1994). Deletion of *Pdx1* in mouse causes pancreatic agenesis (Jonsson et al., 1994) and heterozygous or homozygous mutations in *PDX1* gene can lead to permanent neonatal diabetes (Stoffers et al., 1997b; Schwitzgebel et al., 2003; Thomas et al., 2009) or monogenic diabetes (MODY, maturity-onset diabetes of the young, (Stoffers et al., 1997a). *Pdx1* is initially expressed in the pancreatic progenitors and later restricted to the insulin producing β-cells and more scarcely in somatostatin-producing δ cells (Ohlsson et al., 1993). Precisely, it is the combined expression of *Pdx1* and *Ptf1a* (Pancreas-specific Transcription Factor) that specifies pancreatic commitment from multipotent progenitors cells, which can give rise to endocrine, exocrine or ductal lineages (Kawaguchi et al., 2002; Burlison et al., 2008). Other important transcription factors involved in pancreas development are *FoxA* gene family. It has been shown that Foxa1 and Foxa2 can regulate *Pdx1* expression by specifically binding to conserved enhancer sequences located around 6 kb upstream of *Pdx1* TSS. Moreover, deletion of *Foxa1* and *Foxa2* in mice causes pancreatic hypoplasia and loss of *Pdx1* expression. (Gao et al., 2008).

Nk homeodomain factors Nkx2.2 and Nkx6.1 are also essential for pancreas development and β-cell differentiation. Both TFs are expressed in the pancreatic buds but Nkx6.1 becomes restricted to the β-cells in adults, similarly to Pdx1, while Nkx2.2 is expressed in α, β and PP cells of the mature pancreas (Sussel et al., 1998). In mice lacking Nkx2.2 TF β-cells fail to differentiate, they develop severe hyperglycaemia and die shortly after birth (Sussel et al., 1998). Interestingly, the phenotypes displayed by Nkx6.1/Nkx2.2 double mutant mice is identical to the Nkx2.2 mouse, implying that Nkx6.1 is downstream of Nkx2.2 in β-cell differentiation pathway (Sander et al., 2000).

3.1.4 Chromatin profiling in the pancreatic islet: how much do we know

Despite the fast development in the field of epigenetics and the global prevalence of diabetes, the pancreas epigenome has not been systematically assessed. The first attempt to map genome wide regulatory regions in purified human islets used a technique called FAIRE-Seq, that is, Formaldehyde-Assisted Isolation of Regulatory Elements, coupled with NGS (Gaulton et al., 2010). The authors found that open chromatin sites defined by FAIRE were not isolated, but clustered in the genome in what they termed COREs (Clusters of Regulatory Elements). These clusters are in average 25 kb in size and are usually linked to single islet-specific genes such as *PDX1* or *NKX6.1* (Gaulton et al., 2010). The first study that profiled chromatin in human pancreatic islets mapped several histone modification marks by ChIP-Seq (H3K4me1-3 and H3K27me3). They found that insulin and glucagon promoters were sparsely enriched in H3K4 methylation and also made global predictions of which described T2D-associated polymorphisms were overlapping with H3K4me1 enriched regions and therefore could be functional (Bhandare et al., 2010). However, this study lacked functional validations. Following a similar strategy, around 18,000 putative promoters and 34,000 distal elements were found when profiling chromatin in pancreatic islets by a combination of DNase I-Seq and ChIP-Seq on CTCF and H3K4me3, H3K4me1 and H3K79me2 (Stitzel et al., 2010). Upon validation of some of the putative enhancers in a murine islet cell line they found that 4 out of the 12 enhancers that worked *in vitro* harboured T2D-associated SNPs (Stitzel et al., 2010). Other studies have focused on master regulators of the pancreas. Khoo and colleagues investigated PDX1 occupancy and targets both in human and mouse islets to find that binding sites are not generally conserved; whereas in mouse islets PDX1 is enriched in promoter regions, in humans it is preferentially bound to intragenic regions (Khoo et al., 2012). Nevertheless, what most of these studies have in common is the lack of functional validation of predicted CREs in the context of a complex organism.

3.1.5 Preliminary data leading to the project

In an effort to elucidate the linkage between transcriptional regulation and epigenetics, our collaborators decided to profile islet-specific transcription factor binding sites and several key regulatory histone modification marks, which together with expression data would help to understand pancreas *cis*-regulation in humans.

Genome-wide integrated maps of key pancreatic transcription factors, including PDX1, NKX6.1, NKX2.2, FOXA2 and MAFB and active chromatin marks including H3K4me1, H3K4me3 and H327ac, were generated by Lorenzo Pasquali (IDIBAPS, Barcelona) using purified human islets as starting material. Bioinformatic analysis revealed ~95,000 “open” chromatin regions in human islets marked either by FAIRE-Seq or by H2A.Z enrichment. Among these sites, discrete subclasses of open chromatin regions can be distinguished that correlate with distinct CRE categories (**Figure 3.2**). The cluster termed “C1” contained ~14,000 regions that showed strong H3K4me3-enrichment, a typical promoter chromatin signature. ”C2“ regions (~30,000) resembled inactive or “poised” enhancers (unimodal H3K4me1-enrichment lacking H3K27ac (Creyghton et al., 2010; Rada-Iglesias et al., 2011), and “C3” regions (~31,000) had a typical active transcriptional enhancer signature (strong bimodal enrichment in H3K4me1 and H3K27ac but not H3K4me3. Among remaining open chromatin sites, ~8,900 (C4) were enriched in CCCTC-binding factor, while ~9,600 (C5) lacked active histone modifications.

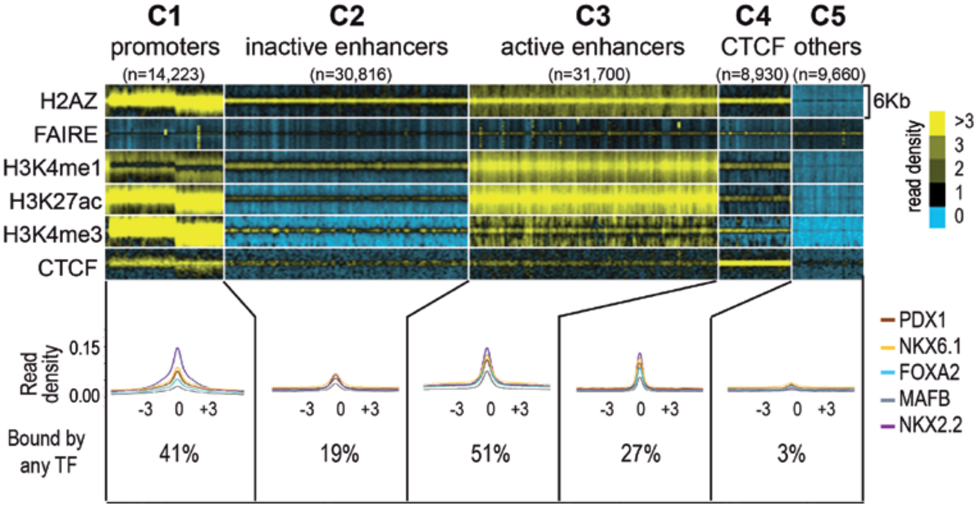


Figure 3.2 Clustering of~95,000 open chromatin regions based on active histone modifications revealed 5 distinct epigenetic classes.

Clusters of open chromatin sites (C1-C5) defined by H2AZ or FAIRE (top). Transcription factor distribution in each cluster is depicted (bottom). Figure adapted from (Pasquali et al., 2014).

3.1.6 Aims:

These initial observations led to the hypothesis that regions within C3 cluster could act as developmental islet enhancers, since they were open chromatin sites enriched in H3K27ac and H3K4me1. Harnessing the advantages of zebrafish as an effective transgenesis model, we proposed the use of zebrafish as a tool to test pancreatic-specific CREs. Our main motivations included the exploitation of easy transgenesis screens, the conservation of the expression and function of pancreas-specific transcriptional network between mammals and zebrafish and a shared anatomical structure of the pancreas.

In order to test the function of putative CREs, several candidate regions conserved at the sequence level between human and fish were selected for transgenesis assays. In addition, regions within C2 and C5 clusters were used as control regions to evaluate the accuracy of the enhancer selection process. Within this global aim, several specific objectives were set:

- To test the potential of these regions to direct tissue-specific activity in the transient transgenic zebrafish embryo.
- To verify autonomous independent activity of the human candidate enhancers.
- To verify whether the transient patterns were valid in stable transgenic lines and to evaluate how faithful are transient transgenic assays to fully recapitulate the expression pattern of a human CRE.

3.2 METHODS

Putative human elements were cloned into 2 different contexts: linked to *hsp70* zebrafish promoter using Multisite Gateway Cloning, and linked to *gata2* promoter using T4 ligation-mediated cloning.

3.2.1 Multisite-Gateway Cloning of Entry Clones and Expression Vectors

Expression vectors containing a human putative enhancer linked to *hsp70* zebrafish promoter and reporter gene (Venus fluorescent protein) were cloned using the Multisite Gateway Cloning System by Invitrogen, following manufacturer's instructions. Selected human putative enhancers were amplified from human genomic DNA provided by our collaborators using the primers listed in **Table 3.1**. In order to control zebrafish promoter functionality, a non-conserved region from *Fugu rubripes* showing no regulatory activity (fr2(assembly 2004) chrUn:54,537,362-54,537,937) was used (Sanges et al., 2006). As additional negative controls, two genomic regions from C2 and C5 clusters were also cloned (**Table 3.1**). Final expression vectors were generated through LR recombination reactions between Entry Clones and pSP1.72BSSPE-R3-R5-hsp70-Venus Destination Vector, as described (Roure et al., 2007). These vectors contained Tol2 transposase arms to facilitate single-copy genome integration of the construct into the host genome by co-injection with Tol2 transposase mRNA (**Figure 3.3**).

3.2.2 Molecular cloning of putative human enhancers in a different construct

In order to test the autonomous capacity of candidate enhancers to drive tissue-specific expression, the pDB896-hsp70-mCherry vector available in the lab, a kind gift from Darius Balciunas (Balciunas et al., 2006), was modified to contain a multicloning site where the human element of interest, zebrafish *gata2* promoter and a reporter protein (YFP) could be inserted using T4 ligation-mediated cloning. The three elements were flanked by restriction sites, which allows for an easy exchange of both reporter genes and *cis*-regulatory elements. The resulting vector was named pDB896-gata2-YFP. Zebrafish promoters can be cloned using AgeI and XhoI sites, and the reporter proteins using XhoI and SnaBI sites. The putative elements were cloned using the primers described in **Table 3.1** adding EcoRV and SpeI tails to the forward and reverse primers respectively.

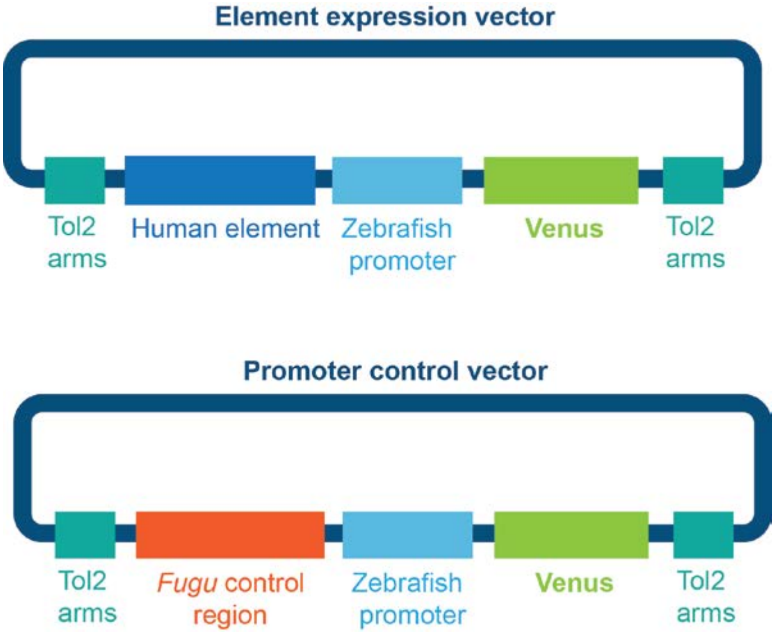


Figure 3.3 Schematic representation of the vectors cloned for each human element.

Table 3.1 Primers used to amplify human putative enhancers and zebrafish promoters from genomic DNA

CRE	Genomic coordinates (Hg18 or Zv9)	Forward primer (5’-3’)	Reverse primer (5’-3’)	Product size (bp)
C3-1	chr4:85339334-85339883	TGCAGTCACATGCACAAAG	AGAAACTAGGGCTGTGTTTA	550
C3-3	chr5:51786984-51788169	TTAAGGTCCCTCTGCCATGT	AACTCTTCCCAAGCCTCATT	1186
C3-4	chr1:212242123-212243697	AATTTTCTTCCTCCGCTTTC	CATTCCCTTTAATATCCCATGC	1575
C3-5	chr19:6066434-6067353	GAAAAGCGCTCCAGAAATTG	AGTTCCCTTTGCACTTGTT	920
C3-6	chr10:114736989-114737824	CCAAGGCTTGAAAATGGATG	AGAGCTTTTTCTAGGCCTCC	836
C2-11	chr1:244375378-244375874	AGGCATCTGAGCTTCACTGG	AGTCAGACAGACCTGGAATA	491
C5-14	chr4:139712081-139712568	ACGCATATGGTCGGATATGA	AAGGCCTGTAGAGAAAGAAT	488
hsp70 zebrafish promoter	chr3:26911324-26911472	TTGATTGGTCGAACATGCTGG	CAGTCCGCTCGCTGTCTCGCT	149
gata2 zebrafish promoter	chr11:3,922,100-3,923,130	ATTCATTAATAGAATAGAGGCATT	CTCAAGTGTCCGCGCTTA	1031

3.3 RESULTS

3.3.1 Selection of candidate islet-specific human enhancers based on epigenetic marks and sequence conservation analysis

Based on the epigenetic clusters uncovered by our collaborators, we decided to test whether human putative enhancers, predicted by the presence of a strong bimodal enrichment in H3K4me1 and H3K27ac in the human islet, could function as developmental enhancers in the zebrafish pancreas. We selected 5 enhancer candidates from cluster C3 based on human-zebrafish conservation (>70% sequence identity over 100 nt, hg18 vs DanRer7). Tested fragments were around 1 kb in size around the central point of FAIRE or H2AZ. Element boundaries were determined manually by examining epigenetic enrichment profiles. In parallel, we also checked as a proof of principle two additional regions that belonged to C2 cluster, resembling inactive enhancers and C5 cluster lacking histone modification marks (**Table 3.1**).

3.3.2 Functional testing of putative islet-specific developmental enhancers using the zebrafish embryo

Zebrafish transient transgenesis assays have been successfully applied in the past to functionally uncover long-range acting *cis*-regulatory elements (Muller et al., 1997; Muller et al., 1999; Dickmeis et al., 2004; Woolfe et al., 2005). To investigate the function of human putative enhancers *in vivo*, selected candidates were linked to zebrafish *hsp70* promoter and YFP reporter gene and were injected in zebrafish embryos as described (Gehrig et al., 2009). Reporter gene expression pattern was assessed during the first 5 days of development and all tissues of expression from at least three experimental replicates were annotated (**Table 3.2**). Representative patterns of expression were documented at 72 hpf, as the zebrafish pancreas has developed completely at this stage (Tiso et al., 2009). As a control a *Fugu* region with no enhancer activity linked to *hsp70* minimal promoter was used (Sanges et al., 2006).

Table 3.2 Frequency of transgene expression in 3 dpf injected zebrafish embryos.

Tested CRE	Nearby gene upstream (kb from TSS)	Nearby gene downstream (kb from TSS)	Domains of expression at 72 hpf (%)						No of injected embryos	No of replicates
			Lens	Pancreatic islet	Neurons	Hindbrain	Floor plate	Pronephric duct		
C3-1	*AGPAT9* (663)	*NKX6-1* (299)	**49.9**	0.0	**18.6**	0.0	0.0	0.0	746	4
C3-3	*ISL1* (1,108)	*PELO* (296)	**94.9**	**38.3**	0.0	0.0	0.0	0.0	352	3
C3-4	*PROX1* (14)	*SMYD2* (277)	0.0	**28.7**	0.0	0.0	0.0	0.0	230	3
C3-5	*LOC100128568* (88)	*RFX2* (43)	0.0	**32.9**	0.0	**94.7**	**86.8**	**44.7**	228	3
C3-6	*TCF7L2* (27)	*HABP2* (573)	**41.3**	0.0	**46.6**	0.0	0.0	0.0	496	4
C2-11	*KIF26B* (990)	*SMYD3* (271)	0.0	0.0	0.0	0.0	0.0	0.0	466	4
C5-14	*LINC00499* (481)	*CCRN4L* (224)	0.0	0.0	0.0	0.0	0.0	0.0	407	4
Fugu	*n/a*	*n/a*	0.0	0.0	0.0	0.0	0.0	0.0	217	3

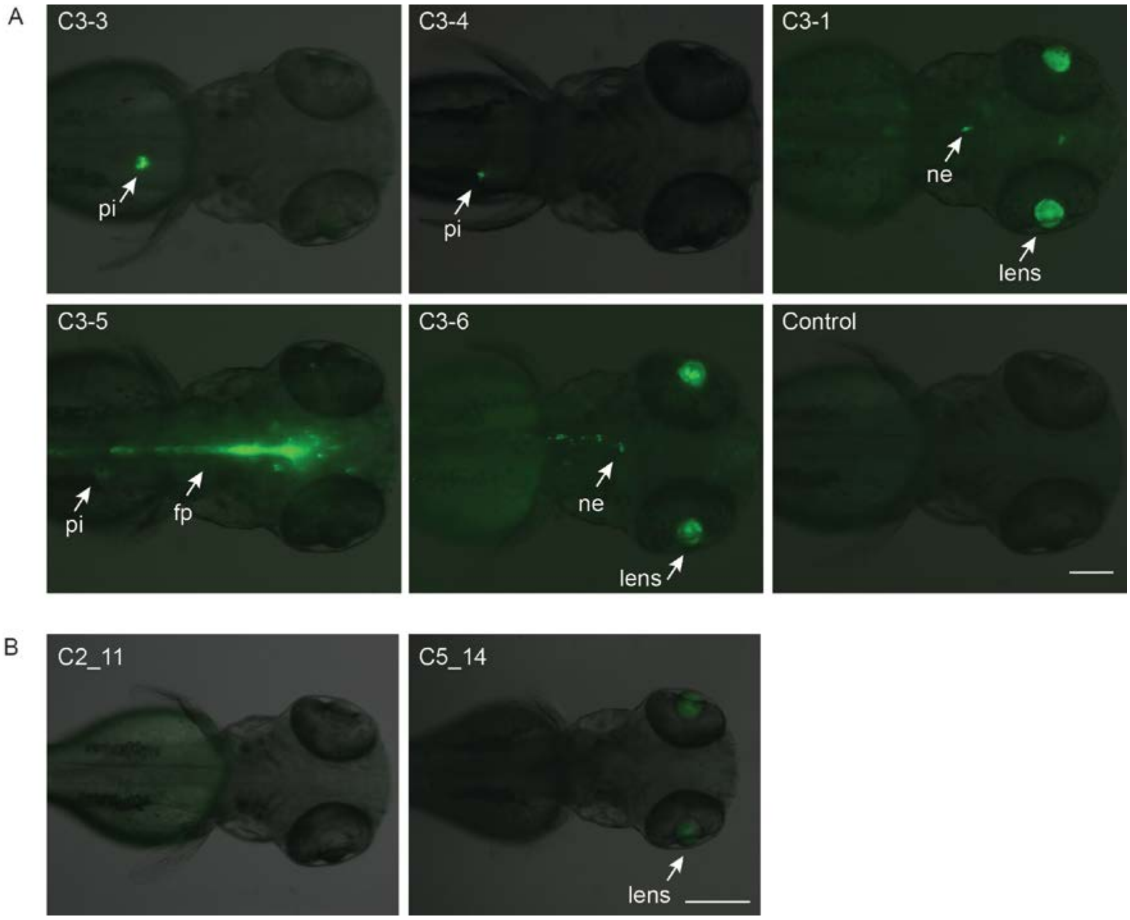


Figure 3.4 C3 human active enhancer regions drove tissue-specific transient expression patterns in zebrafish.

A. Merged images of YFP and brightfield channels for zebrafish embryos injected with C3 candidate enhancers or a control construct containing the zebrafish *hsp70* promoter linked to a region lacking enhancer function. **B**. Representative embryos injected with C2 and C5 regions. YFP expression is observed in the pancreatic islet (pi), neurons (ne), lens (lens), and floor plate (fp). Expression in the lens is ectopic activity from the *hsp70* promoter (Blechinger et al., 2002). All embryos are 72 hpf, oriented dorsally anterior to the right. Scale bar indicates 100 µm.

The enhancer-less control and C2 and C5 sequences showed no YFP activity. In contrast, 5 conserved elements tested from C3 cluster behaved as enhancers, as they were able to direct tissue-specific expression; and three of them, C3-3, C3-4 and C3-5, were active in the pancreatic islet (**Figure 3.4, Table 3.2**).

3.3.2.1 C3-3 candidate enhancer

C3-3 element contained a 1,185 bp fragment located in an intergenic region that is 1.1 Mb downstream of *ISL1* and 300 kb upstream of *PELO*. It was in an open chromatin site, bound by PDX1, NKX6.1, FOXA2 and NKX2.2 TFs and showed a significant enrichment for H3K27ac and H3K4me1 epigenetic marks (**Figure 3.5A**). In transient transgenic assays C3-3 element linked to zebrafish *hsp70* promoter could drive islet-specific expression in 38.4% of the injected embryos in all replicates (total number of injected embryos n=352). In contrast, none of the embryos injected with the control region could reproduce the pancreas expression pattern (n=217). Lens activity was shown by 94.9% (n=352) of the expressing embryos. Lens ectopic expression was caused by the endogenous promoter itself and not by the human element linked to it (Blechinger et al., 2002). Islet-specific activity in transient transgenic assays was verified by injecting the putative enhancer-containing constructs into Tg(ins-mCherry)^{jh2} transgenic line (a kind gift from Elke Ober), which labels insulin-producing beta cells in red (**Figure 3.5B**). Co-localization experiments verified that C3-3-driven activity was specific to the endocrine pancreas.

Given that enhancers can be very far away from their regulatory target, it is challenging to identify target genes (Krivega and Dean, 2012). Both in mammals and zebrafish, *ISL1* gene is essential for the development of endocrine cells (Ahlgren et al., 1997; Wilfinger et al., 2013), while *PELO* gene is unrelated to pancreas or endodermal development (Shamsadin et al., 2000). Interestingly, chromosome capture conformation experiments carried out elsewhere demonstrated that there

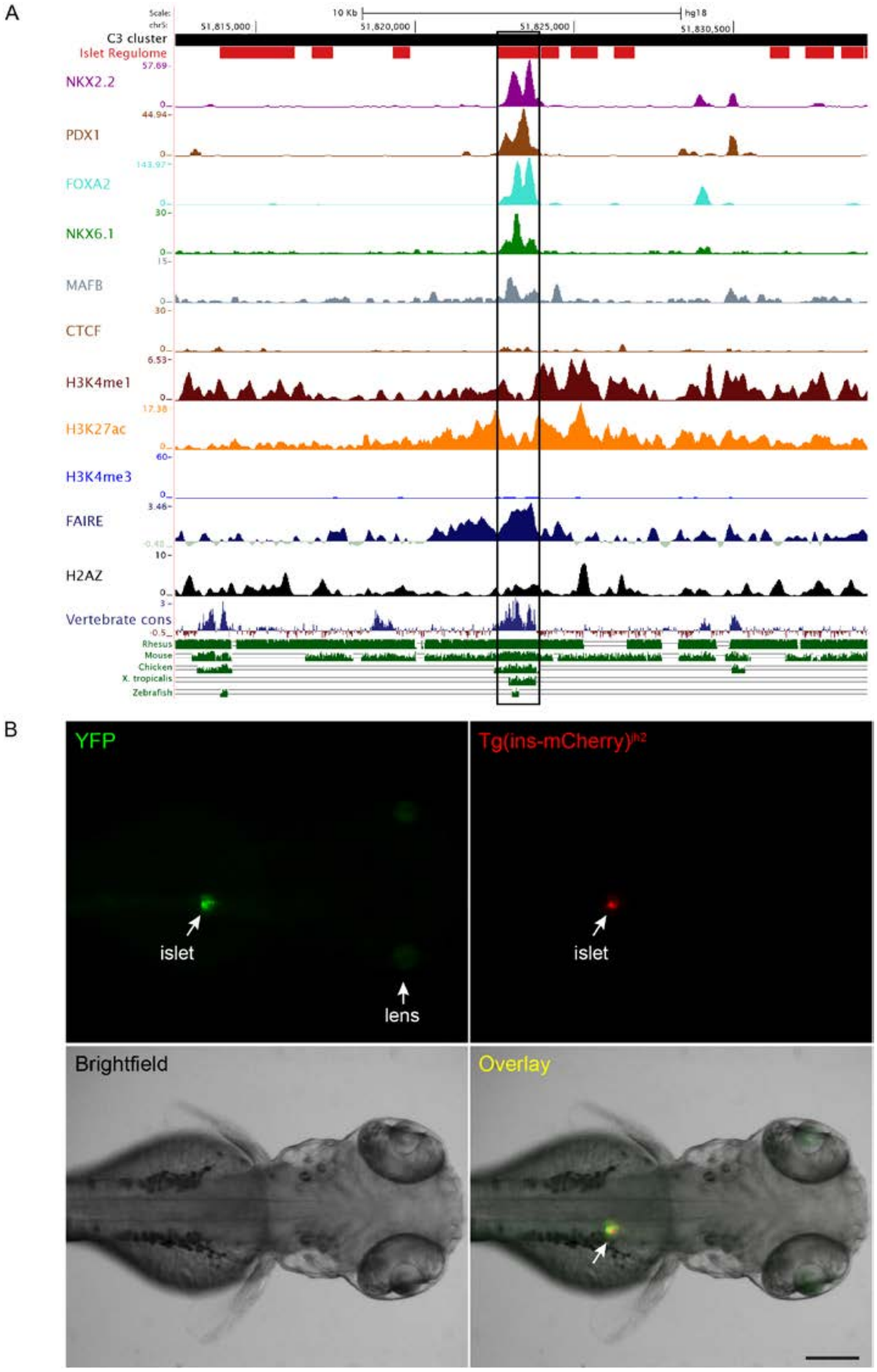


Figure 3.5 Intergenic enhancer located 1 Mb downstream of *ISL1* can direct islet-specific expression in the zebrafish embryo.

A. UCSC genome browser screenshot of C3-3 putative enhancer location and epigenomic landscape. Note that C3-3 (black rectangle) is in an open chromatin region 1.1 Mb downstream of *ISL1* and 300 kb upstream of *PELO*. It is bound by NKX2.2, PDX1, NKX6.1 and FOXA2 TFs and shows a bimodal enrichment of H3K4me1 and H3K27ac, which are considered typical enhancer marks. **B.** C3-3 can drive robust expression in the pancreatic islet as demonstrated by co-localization assays in Tg(ins-mCherry)^{jh2} line. Embryos are 72 hpf oriented dorsally anterior to the right. Scale bar indicates 100 μm.

was a direct interaction between C3-3 tested enhancer and *ISL1* promoter, which is located > 1 Mb away (Pasquali et al., 2014), supporting the pancreas-specific activity.

3.3.2.2 C3-4 candidate enhancer

CRE3-4 is located on the second intron of PROX1 (1,574 bp) and could also drive pancreas-specific reporter expression in 28.7% (n=230) of the injected embryos. This region was also shown to be bound by pancreas-specific TFs, including NKX6.1, FOXA2, MAFB, NKX2.2.

3.3.2.3 C3-1 and C3-6 candidate enhancers

In contrast to the previous elements, which were active in the endocrine pancreas, neither a region of 835 bp in the second intron of *TCF7L2* gene (C3-6) nor a 549 bp fragment upstream of the *NKX6.1* gene (C3-1) showed islet-specific activity in four experimental replicates (**Table 3.2**). Both elements behaved as weak and broad neuronal enhancers in the zebrafish embryo (**Figure 3.4**).

3.3.2.4 C3-5 candidate enhancer

A very interesting example among the elements tested is C3-5, a 919 bp region from the second intron of the *RFX2* gene. When C3-5 was cloned upstream of the zebrafish *hsp70* promoter, it could drive hindbrain, floor plate and pronephric duct expression at 3 dpf, recapitulating almost completely the endogenous pattern of *rfx2* zebrafish gene (**Figure 3.6**) consistently among the experimental replicates.

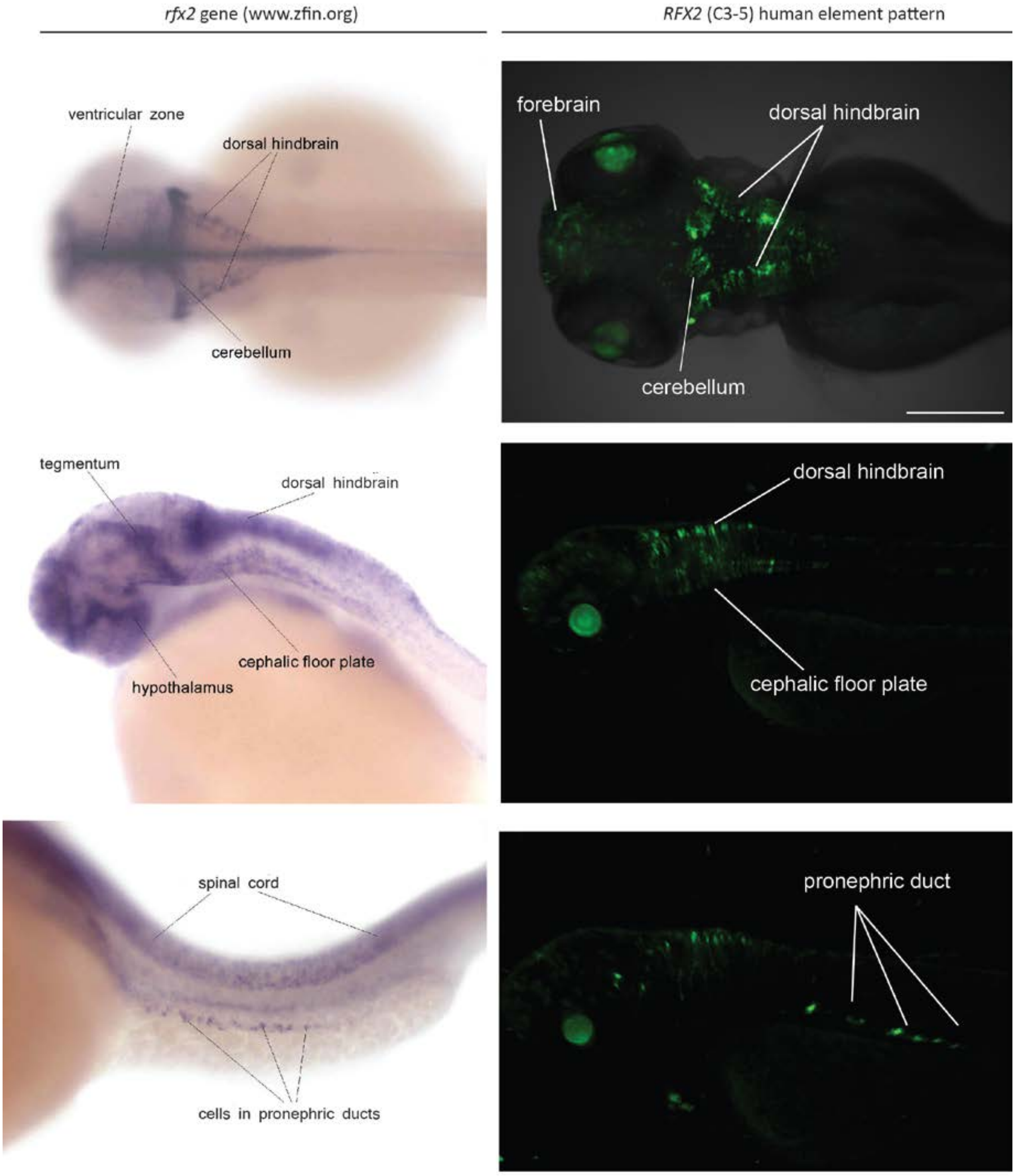


Figure 3.6 C3-5 candidate human enhancer recapitulates activity domains of zebrafish *rfx2* gene.

Zebrafish mosaic embryos injected with C3-5 human enhancer linked to *hsp70* promoter and YFP reporter gene protein (right panels) recapitulated broadly zebrafish *rfx2* endogenous expression patterns as demonstrated by *rfx2* WISH (left panels, taken from www.zfin.org). Expression domains include forebrain, dorsal hindbrain, cerebellum, cephalic floor plate and pronephric ducts. Lens expression is driven by *hsp70* promoter. Scale bar indicates 100 µm.

3.3.2.5 Validation of control regions

In order to test the hypothesis that only regions showing enhancer-associated epigenetic marks would be able to drive tissue-specific activity, two regions from C2 and C5 clusters (**Figure 3.2**), bound by PDX1 but devoid of active epigenetic marks were tested by transient transgenesis. These regions could not direct tissue-specific expression on zebrafish embryos during the first days of development (**Table 3.2, Figure 3.4**), suggesting that epigenetic marks are indeed predictive of enhancer function**.**

3.3.3 Enhancer-promoter interactions *in vivo*.

In order to test whether C3 candidate enhancers were autonomously capable of driving expression in the endocrine pancreas, C3 candidates were tested with a second independent minimal promoter and compared their activity to *hsp70* promoter. Candidates were cloned into a pD896 vector containing mini Tol2 arms (Balciunas et al., 2006), YFP reporter gene and *gata2* promoter, which had been successfully used to validate enhancers in zebrafish (Bessa et al., 2009; Navratilova et al., 2009; Royo et al., 2012).

All human enhancers from cluster C3 could activate *gata2* promoter with similar frequencies and patterns except for C3-5. Interestingly, when C3-5 (located on the second intron of *RFX2* gene) was linked to *gata2* zebrafish promoter, it was able to activate reporter expression in the endocrine pancreas as an additional expression domain (**Figure 3.7**) in 32.9% of injected embryos (n=228, **Table 3.2**). This domain had not been detected in embryos injected with *hsp70*-containing construct (n= 278, **Figure 3.6)** and underlined the importance of selecting an appropriate promoter for validations as well as the specificity enhancer-promoter interactions occurring *in vitro*. These differences were also prevalent in the transgenic lines established from both constructs, which will be discussed below.

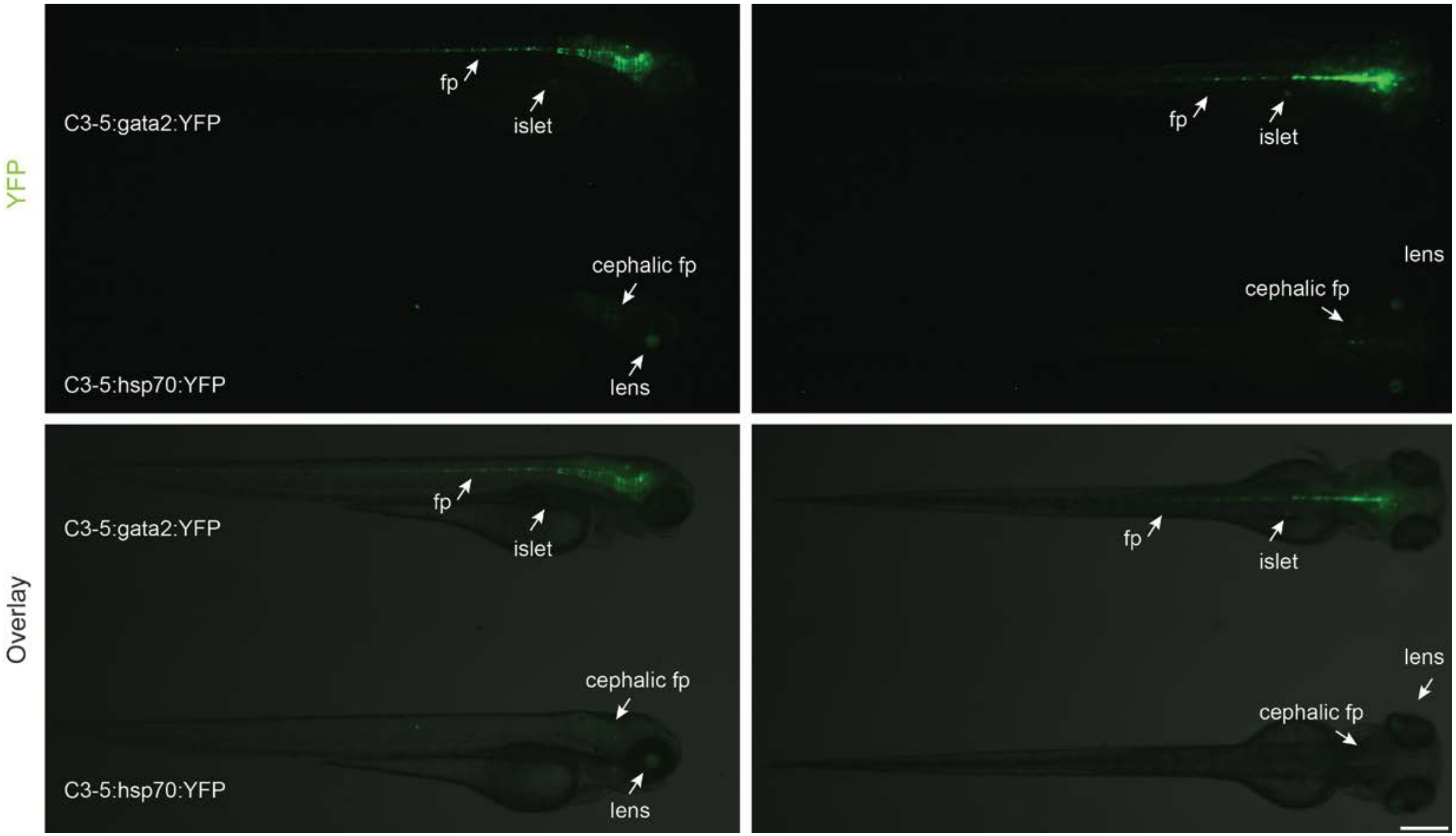


Figure 3.7 Specificity of enhancer promoter interactions *in vivo* using C3-5 human enhancer.

Representative images of 72 hpf embryos injected with constructs containing C3-5 human enhancer linked to zebrafish *gata2* promoter (C3-5:gata2:YFP, top) or C3-5 enhancer linked to *hsp70* promoter (C3-5:hsp70:YFP, bottom). Note how *gata2* promoter is overall more sensitive to regulatory input from the enhancer and can drive an additional domain of expression in the pancreatic islet. Scale bar indicates 250 µm.

3.3.4 Verification of enhancer activity detected in mosaic transient transgenics by the establishment of stable transgenic lines

In order to verify the pancreas-specific activity shown by transient transgenic embryos I decided to establish transgenic lines for all the enhancers that were active in the pancreatic islet (C3-3, C3-4 and C3-5). Injected embryos showing reporter gene expression were raised to sexual maturity and outcrossed with wild type fish. The offspring of these outcrosses was screened during the first 3 dpf (**Table 3.3**). The expression patterns of the transgenic lines were documented using Olympus ScanR microscope and representative images are shown in **Figure 3.8** and **Figure 3.9**.

I identified three adult founders containing C3-3 CRE linked to *hsp70* promoter whose offspring was positive for reporter gene expression (**Table 3.3**). Two of them displayed specific expression in the pancreatic islet and the third one showed an additional domain of expression, the spinal cord, which had not been seen in transient transgenic assays but is a sub-expression domain of the *isl1* gene (**Figure 3.5**). Expression in the lens was considered as ectopic activity from the *hsp70* promoter (Blechinger et al., 2002). One positive founder containing C3-4 candidate enhancer gave rise to positive progeny expressing in the pancreatic islet and in a specific group of midbrain neurons, reproducing the specific expression in the islet cells seen in transient embryos. Because different activities were displayed by C3-5 human enhancer when linked to different zebrafish promoters, I decided to establish transgenic lines using both constructs. Two lines established using C3-5 linked to *hsp70* promoter (C3-5 L1 and L2) showed expression in forebrain, hindbrain, spinal cord and pronephric duct, reproducing endogenous pattern of *rfx2* and the activity seen in transient transgenic assays (**Figure 3.9**). In contrast, the 3 lines established with a construct containing C3-5 linked to *gata2* promoter (C3-5 L3, L4 and L5) were active in all of the aforementioned *rfx2* endogenous domains plus the pancreatic islet (**Figure 3.8**). It is interesting to note how transgenic lines containing C3-5 and *hsp70* minimal promoter failed to activate the pancreas domain, resembling mosaic injected embryos. Slight differences between C3-5 lines might be attributed to position effects caused by Tol2-mediated transgenesis. Nevertheless, it needs to be noted that the additional domains, namely olfactory bulb and ventricular zone, are also *rfx2* endogenous domains.

Islet-specific expression of the transgenic lines was verified by out-crossing the adult transgenic founders with individuals from the Tg (ins-mCherry)^{jh2} line (**Figure 3.10**), confirming activity in the zebrafish endocrine pancreas. Taken together these results suggest that the transgenic lines validate the tissue-specific patterns seen in the transient transgenic embryos.

Table 3.3 Analysis of the offspring of C3 human enhancer-derived transgenic lines

Transgenic line name	No. of screened founders	No. of positive founders	Germline transmission rate (%)	Line ID	Domains of expression	No. of embryos analysed	Transmission rate per pattern (%)
Tg(hs-PELO-0.1dr-hsp70-pr:YFP)	8	2	25.0	C3-3-L1	**Pancreatic islet**, lens	107	57.9
				C3-3-L2	**Pancreatic islet**, lens	173	75.7
				C3-3-L3	**Pancreatic islet**, floor plate, lens		24.3
Tg(1.5hs-PROX1-0.1dr-hsp70-pr:YFP)	10	1	10.0	C3-4-L1	**Pancreatic islet**, midbrain neurons	63	34.9
Tg(hs-RFX2-0.1dr-hsp70-pr:YFP)	4	2	50.0	C3-5-L1	Forebrain, dorsal hindbrain, cephalic floor plate, hindbrain, pronephric duct, lens	252	45.2
				C3-5-L2	Forebrain, dorsal hindbrain, cephalic floor plate, hindbrain, pronephric duct, lens	190	42.1
Tg(hs-RFX2-1.0dr-gata2-pr:YFP)	7	3	42.9	C3-5-L3	Olfactory bulb, hindbrain, floor plate, pronephric duct, **islet**	378	3.9
				C3-5-L4	Forebrain, hindbrain, floor plate, pronephric duct, **islet**	51	23.5
				C3-5-L5	Forebrain, hindbrain, floor plate, pronephros, **islet**	153	47.1

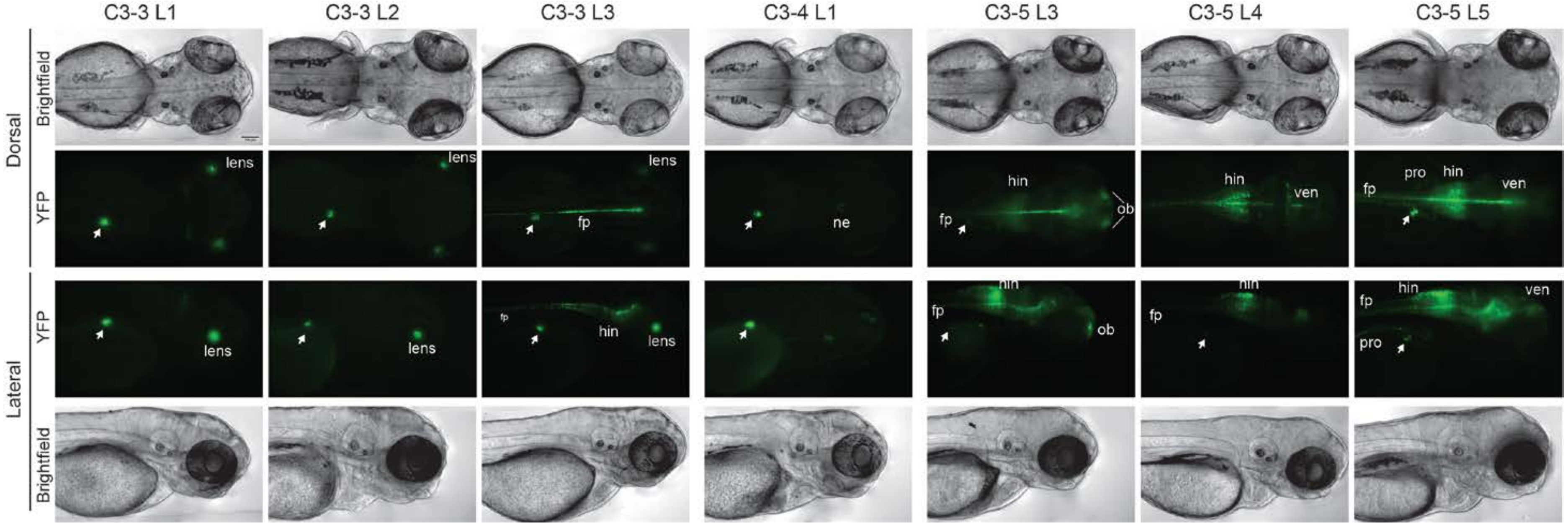


Figure 3.8 Stable transgenic lines derived from predicted enhancers recapitulate transient patterns.
Three stable transgenic lines derived from C3-3 element (C3-3 L1-3), one line derived from C3-4 (C3-4 L1) and three lines derived from C3-5 (C3-5 L1-3) confirm pancreas specific activity (arrow) displayed by mosaic transgenic embryos. Additional domains of expression include lens, fp=floor plate, pro=pronephros, hin=hindbrain, ne=neurons, ob=olfactory bulb, ven=ventricular zone All embryos are 72 hpf oriented anterior to the right. One dorsal and one lateral brightfield and YFP images are shown.

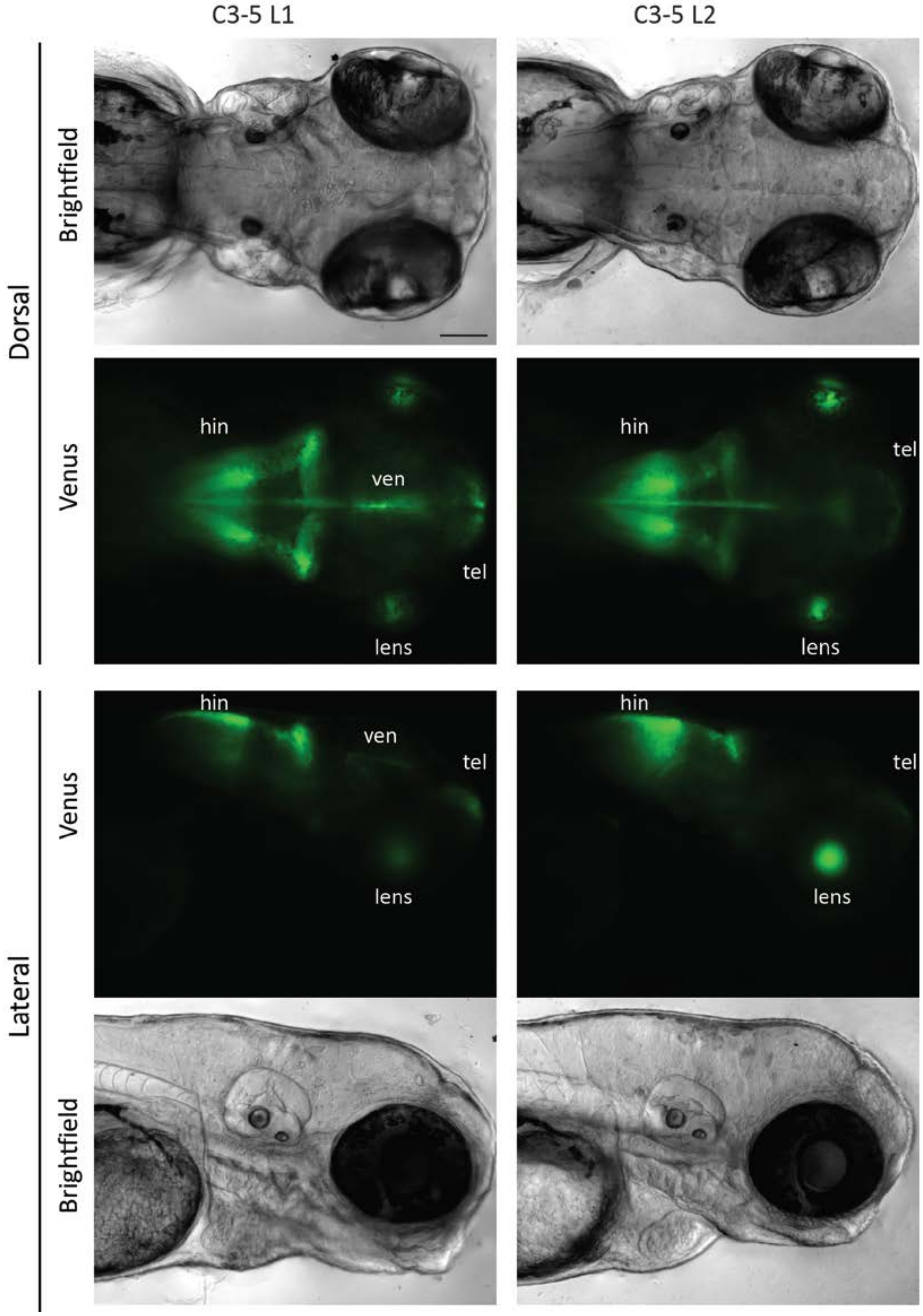


Figure 3.9 Transgenic lines from C3-5 element linked to *hsp70* promoter fail to activate the islet. Two transgenic lines established with a construct containing C3-5 human enhancer linked to *hsp70* promoter displayed expression domains that broadly overlap *rfx2* endogenous pattern, similarly to our observations in transient transgenic assays (**Figure 3.6**). YFP domains include hin=hindbrain, lens, tel=telencephalon and ven=ventricular zone. Embryos are 72 hpf oriented anterior to the right. Scale bar is 100 µm.

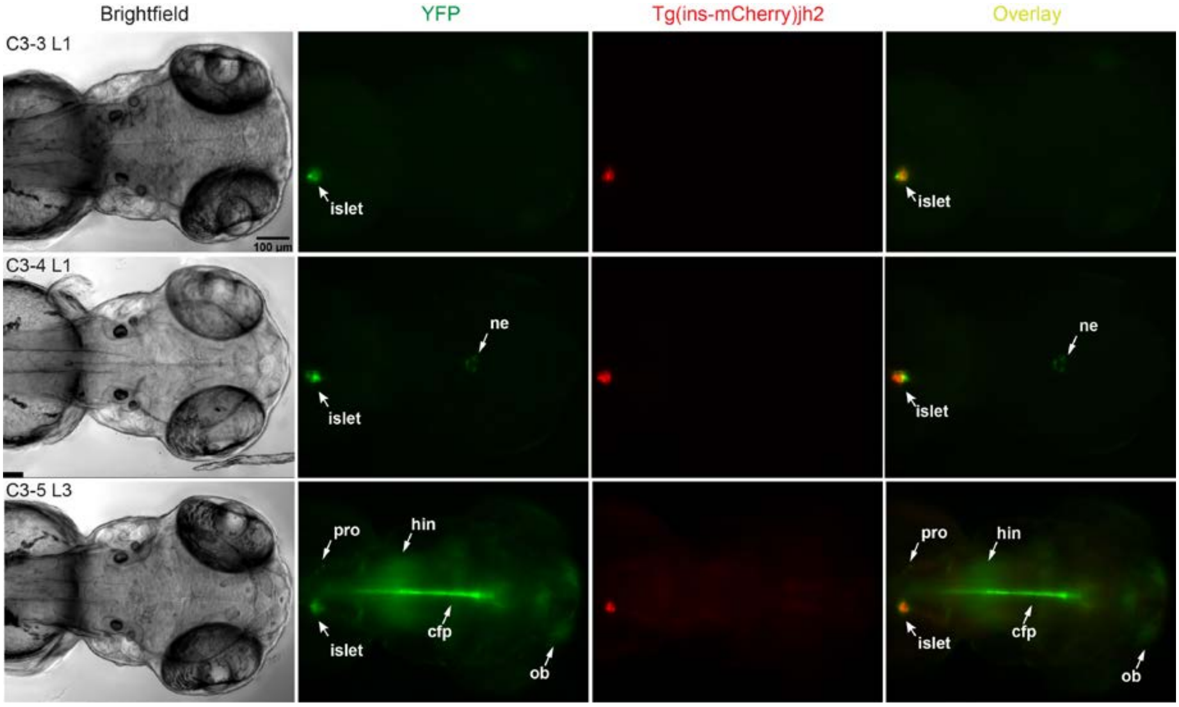


Figure 3.10 C3-3,4,5 derived stable transgenic lines show pancreas-specific activity when outcrossed with Tg(ins-mCherry)^{jh2} line.

One representative founder is shown per CRE tested. Note the overlap between insulin producing beta cells in red and enhancer driven activity in green. Additional domains of expression include cfp=cephalic floor plate, hin=hindbrain, ne=neurons, ob=olfactory bulb and pro=pronephros. All embryos are 72 hpf oriented anterior to the left.

3.4 DISCUSSION

In this chapter, I studied the function of several human putative enhancers selected through a combination of transcription factor binding events and the presence of key epigenetic marks to address the utility of the zebrafish as a validation model for human enhancers. I demonstrated that all 5 human predicted human enhancers can drive tissue-specific reporter activity in zebrafish embryo. From these elements, 3 are active in the zebrafish embryonic islet (C3-3, C3-4 and C3-5), while the remaining 2 were broad neuronal enhancers (C3-1 and C3-6). Additionally, two control regions devoid of epigenetic marks failed to drive any tissue-specific expression. These results were verified not only in transient transgenesis assays, where a high percentage of injected mosaic embryos were displaying very reproducible and specific patterns beyond background, but also in stable transgenic lines. Taking into account the negligible background level of our enhancer-less construct we can conclude that the specific expression patterns are attributable to the human elements tested. This argues for the specificity of the enhancer selection process and leaves 3 out of 5 enhancers behaving as predicted by their epigenetic marks.

A challenge in enhancer tests is the unambiguous identification of the target gene. In the case of the three tested human enhancers active in the pancreatic islet, the expression patterns reproduced broadly the endogenous activity of the predicted target gene. As a general rule, we associated the enhancer to the closest coding gene. C3-4 which was located on an intronic region of *PROX1* proved to be a very specific pancreatic enhancer in zebrafish, and indeed the endocrine pancreas is included among the expression domains of *PROX1*, a master regulator during embryogenesis (Sosa-Pineda et al., 2000; Burke and Oliver, 2002; Wang et al., 2005; Pistocchi et al., 2008). C3-5 located in the second intron of *RFX2* reproduced most of the endogenous domains of activity of zebrafish *rfx2,* including ventral floor plate, dorsal hindbrain and pronephric duct. Interestingly, only when C3-5 was paired with *gata2* endogenous promoter an additional domain in the pancreatic islet was active. This result, which was consistent, both in transient and stable transgenic assays, argues for *in vivo* evidence of enhancer-promoter interactions and underlines the importance of promoter choice in enhancer validation assays. In the case of C3-3, an element located in an intergenic region 1Mb away from *ISL1* and 400 Mb away from *PELO*, it was demonstrated a very strong and reproducible expression in the endocrine pancreas, which is an active domain of *ISL1* gene in human and zebrafish. Additionally, 4C experiments carried out elsewhere demonstrated that the tested enhancer was directly interacting with *ISL1* promoter (Pasquali et al., 2014).

Even though a higher number of tested elements would be necessary to establish a statistically significant correlation, in our hands 60% of the predicted human islet enhancers showing sequence conservation with zebrafish were interpreted accurately by the zebrafish embryo. This figure is similar to the rates shown in the literature of conserved human elements directing tissue-specific expression in fish: 58% out of 24 candidate enhancers (Ritter et al., 2012), 34.5% out of 113 sequences tested (Royo et al., 2011), 47% out of 34 enhancers tested (Punnamoottil et al., 2010) and 77% out of 31 tested candidates (Ritter et al., 2010), among others.

It is difficult to predict why some of these human enhancers could not drive specific enhancer activity in the pancreas, despite showing similar levels of sequence conservation, strength and enrichment of histone modification marks and transcription factor occupancy. Statistical analysis on any of these parameters would require a higher number of elements to be tested *in vivo*, however, imposing the sequence conservation filter drastically reduces the number of human candidates that can be validated. Nevertheless, it needs to be noted that the neuronal expression pattern displayed by some candidates is not completely unexpected. The neuroendocrine nature of islet cells has been recognized: endocrine cells are electrically excitable and share the expression of a large battery of genes with neurons (Atouf et al., 1997). In fact, the expression profile of pancreatic endocrine cells is more similar to ectoderm-derived neuronal tissues than to tissues such as liver or lung, which share a common endodermal origin (van Arensbergen et al., 2010).

Recently, several studies have highlighted the benefits of applying the knowledge derived from genome-wide analyses of epigenetic marks to the vertebrate embryo (Rada-Iglesias et al., 2011; Bernstein et al., 2012). Nevertheless, this is the first study where an isolated tissue from human donors was interrogated and predicted pancreas enhancers were tested using an embryonic model. Most publications have based their validations on cell lines, which cannot reproduce the complexity of an organism (Gaulton et al., 2010; Stitzel et al., 2010). A recent study based on the profiling of the chromatin and transcriptome in human islets, has also identified chromatin states that could overlap with enhancers, promoters and insulators (Parker et al., 2013). Based on this chromatin segmentation, 2 predicted enhancers were tested *in vivo*, demonstrating tissue-specific activity in the pancreas primordium of mouse embryos (Parker et al., 2013). Interestingly, the predicted pancreatic enhancers appear clustered contiguously along regions larger than 3 kb, what was termed “stretch enhancers”, similarly to the clustered C3 elements described by (Pasquali et al., 2014).

Our set of validations, together with previous studies testing human pancreas-specific enhancers in zebrafish (Ragvin et al., 2010) argue that despite over 460 million years of evolutionary divergence between human and teleosts (Meyer and Schartl, 1999), the zebrafish can interpret functionally certain conserved pancreatic enhancers maintaining the tissue-specificity; demonstrating conserved regulatory codes acting in vertebrate endocrine pancreas during development.

Chapter Four: ANALYSIS OF NON-CODING SEQUENCE VARIANTS DOWNSTREAM OF *PTF1A* RELATED TO PANCREAS AGENESIS

FOREWORD

Within the BOLD Network, we initiated a collaboration with Andrew Hattersley and Michael Weedon, from the University of Exeter Medical School (UK), with the common aim of testing disease-associated variants in human enhancers. This led to the design of functional experiments in zebrafish where I tested the function of a candidate enhancer associated with pancreas agenesis.

4.1 INTRODUCTION

4.1.1 Genetic basis of pancreatic agenesis

Pancreatic agenesis is a rare condition resulting from a developmental abnormality of the pancreas that causes permanent neonatal diabetes mellitus and exocrine insufficiency (Winter et al., 1986). There are 2 forms of neonatal diabetes mellitus: transient and permanent. Transient neonatal diabetes mellitus (TND) is resolved by 18 months of age and predisposes to adult T2D (Greeley et al., 2011). The main cause for TND is abnormalities chromosome 6q24 including overexpression of paternally imprinted genes, methylation defects or uniparental chromosome disomy (Temple et al., 1996; Gardner et al., 2000; Temple et al., 2000; Mackay et al., 2008). The permanent form of neonatal diabetes mellitus (PND) is mainly caused by activating mutations of the K_{ATP} channel (*KCNJ11* gene) or heterozygous mutations in the insulin gene (*INS*) but can also be caused by pancreas agenesis or hypoplasia (Gloyn et al., 2004). There are less than 50 examples of permanent neonatal diabetes mellitus caused by pancreatic agenesis described in the literature and the genetic basis is mostly unknown; only in a few cases mutations in the coding regions of pancreas-specific transcription factors, such as *PDX1*, *PTF1A* or *GATA6* have been found (Stoffers et al., 1997b; Schwitzgebel et al., 2003; Sellick et al., 2004; Thomas et al., 2009; Al-Shammari et al., 2011; Allen et al., 2012; De Franco et al., 2013).

Stoffers and colleagues showed a case of pancreatic agenesis caused by a homozygous single point deletion within codon 63 of the *PDX1* gene. This mutation caused a frame shift in the C-terminal end of the PDX1 protein, which leads to the translation of 59 novel codons before termination, altering the transactivation domain of PDX1 that is essential for DNA-binding (Stoffers et al., 1997b). Another case of neonatal diabetes caused by a homozygous mutation of the same gene was also described recently by (Thomas et al., 2009). In the case reported by (Schwitzgebel et al., 2003) a compound heterozygous mutation in two highly conserved sites of helix 1 and 2 of *PDX1* homeodomain caused a decreased half-life form of PDX1 protein, which subsequently led to pancreas agenesis and neonatal diabetes mellitus. Several other studies also identified a couple of mutations (705insG and C886T) and a homozygous deletion (c.437-460del) in *PTF1A* transcription factor gene in patients presenting permanent neonatal diabetes mellitus accompanied by pancreatic and cerebellar agenesis (Sellick et al., 2004; Al-Shammari et al., 2011). More recently, Allen and colleagues showed that *GATA6* was a common cause of syndromic pancreatic agenesis (Allen et al., 2012), which was later confirmed in a larger cohort of 795 individuals (De Franco et al., 2013).

However, in around 30% of described cases of PND caused by pancreatic agenesis, the genetic cause remains to be elucidated (Greeley et al., 2011). Interestingly, most papers that could not find a genetic cause excluded the presence of mutations in the coding sequence of several TFs involved in pancreas development or differentiation, such as *PDX1, PTF1A, SOX9, SOX17, HNF6 or HLXB9* (Chen et al., 2008; Balasubramanian et al., 2010; Salina et al., 2010). Indeed, in all of the PND case reports described above the causal mutations were affecting the coding gene sequence or intron-exon boundaries. However, there is now ample evidence describing human disorders caused by non-coding sequence variations (Kleinjan and van Heyningen, 2005; Epstein, 2009; Lee and Young, 2013). Suspecting that non-coding variants could be responsible for several cases of pancreatic agenesis, we established a collaboration with Michael Weedon and Andrew Hattersley, partners of the BOLD Network, in order to carry out some functional analysis of a potential *cis*-regulatory element.

4.1.2 Preliminary data leading to the project

Weedon and colleagues had found 6 subjects with isolated non-syndromic pancreatic agenesis coming from 3 unrelated consanguineous families. Initially they performed whole-genome sequencing and linkage analysis to look for causative homozygous mutations. Homozygosity mapping in the families revealed an inherited 25-100 Mb region in chromosome 10, which included *PTF1A* candidate gene. However, coding mutations in *PTF1A* and 24 other genes from this region were excluded by Sanger Sequencing. Whole-exome sequencing did not reveal either a potential causative homozygous coding mutation. After filtering common variants, they found 7 variants overlapping with putative regulatory elements and particularly 1 variant that was shared by 2 consanguineous and unrelated individuals, overlapped with the shared region on chr10 previously identified by homozygosity mapping. This conserved region comprised 485 bp, was located 25 Kb downstream of the *PTF1A* gene and was conserved between human and zebra finch, but not with zebrafish.

4.1.3 PTF1A transcription factor has a conserved role in specification of pancreatic fate

PTF1A (previously called PTF1-p48) encodes for a pancreas-specific basic helix-loop-helix DNA-binding subunit of the Pancreas Transcription Factor-1 PTF1 complex (Krapp et al., 1996). In adults it is restricted to the acinar compartment of the exocrine pancreas (Rose et al., 2001) but during development *Ptf1a* is essential for both endocrine and exocrine pancreas formation: lineage-tracing experiments indicate that *Ptf1a* is expressed in the progenitors of acinar, endocrine and exocrine cells, contributing to acquisition of lineage identity (Kawaguchi et al., 2002). Consistent with this role in cell commitment, in *Ptf1a*-/- embryos the exocrine pancreas is absent and cells with endocrine functions relocate to the spleen (Krapp et al., 1998). Additionally *Ptf1a* contributes to neurogenesis: *Ptf1a* mRNA can be detected in E9.5 in neural tube and myelencephalon, and is further expressed in the central nervous system including medulla oblongata, cerebellum and the spinal cord until E12.5 (Obata et al., 2001).

In zebrafish *ptf1a* is an early marker of the anteroventral bud that will give rise to the exocrine pancreas (Zecchin et al., 2004). Its mRNA is first detected at 12-somite stage, when it is expressed in the developing hindbrain (Zecchin et al., 2004) and by 32 hpf it is already expressed in the ventrolateral endoderm sharing the domain of Pdx1 positive cells (Lin et al., 2004). The neural expression pattern peaks at 2 dpf in the hindbrain, retinal anlagen and the rhombic lip and is undetectable by 72 hpf, mimicking the transient neural pattern observed in mice (Obata et al., 2001). At 48 hpf *ptf1a* is additionally expressed in the exocrine bud surrounding insulin positive cells, where it will be restricted to from this moment on in development, reproducing the distribution of exocrine tissue (Biemar et al., 2001). While *Ptf1a* is essential for the formation of the endocrine pancreas in mouse (Krapp 1998), morpholino knock-down experiments in zebrafish cause the loss of exocrine pancreas but do not affect the formation or the spatial organization of the endocrine islet (Lin et al., 2004).

4.1.4 Aims:

Given the conserved evolutionary role of PTF1A in human and zebrafish pancreas development and suspecting that the inherited region downstream of *PTF1A* in patients with pancreas agenesis could be a long-range acting *cis*-regulatory element, I set out to use the zebrafish embryo to determine whether this human putative regulatory region located 25 kb downstream of *PTF1A* gene held regulatory potential. Also, I wanted to investigate whether the genomic variants found in the consanguineous probands suffering from pancreatic agenesis could affect the regulatory activity of *PTF1A* putative enhancer by transient transgenesis assays.

4.2 METHODS

Genomic DNA from patients containing recessive mutations in homozygosis in the putative enhancer element was provided by our collaborators at the University of Exeter: alongside with the wild-type gDNA, one sample containing a homozygous g.23508363A>G mutation and a third sample containing a homozygous g.23508305A>G mutation were used for all experiments.

A short 485pb region from chromosome 10 encompassing the variants from the three patients (**Table 4.1)** was cloned into a Gateway multisite destination vector provided by Darius Balciunas that contained miniTol2 arms (Balciunas et al., 2006), *hsp70* zebrafish promoter and mCherry reporter gene (pDB896_GWB3B5_HSP70-PRM_mCherry).

A large region encompassing a vertebrate conservation block of 1445 bp was also amplified from the patient gDNA using oligos described in **Table 4.1**. Due to technical problems previously encountered when trying to clone enhancer fragments larger than 1 Kb using Gateway Technology, two additional Tol2-based constructs containing zebrafish *hsp70* or *gata2* promoters linked to mCherry were also designed, where the enhancer and its variants could be introduced by cloning. For this purpose, mCherry reporter gene was purified from pCS2+-mCherry and cloned into pDB896_HSP70-PRM_YFP construct (**Chapter 3**) using XhoI and SnaBI sites. The resulting vector was named pDB896_HSP70-PRM_mCherry. Zebrafish *gata2* promoter was amplified from zebrafish genomic DNA and cloned into this vector using AgeI and XhoI sites. The resulting vector was named pDB896_GATA2-PRM_mCherry. In order to control zebrafish promoter functionality, a non-conserved region from *Fugu rubripes* showing no regulatory activity (fr2(assembly 2004) chrUn:54,537,362-54,537,937) was used (Sanges et al., 2006), as discussed in **Chapter 3**. A total of 9 vectors containing variations of the putative human enhancer and minimal promoters and two additional control-containing constructs were sequenced and subsequently used for functional validations (**Table 4.1**).

Table 4.1 Constructs generated to test *PTF1A* enhancer variants in zebrafish transgenesis assays

Construct	Putative enhancer size (bp)	Putative enhancer coordinates (hg19)	Forward primer (5' to 3')	Reverse primer (5' to 3')	Tails used for cloning	gDNA sample used	Zebrafish Promoter
pDB896_Hs.PTF1A-short-WT:Dr.hsp70:mCherry	485bp	chr10:23,508,102-23,508,586	ATCACCCCCTGGATGATTCT	GGTGCATGCAACATAGAAAG	attB3/attB5	WT	*hsp70*
pDB896_Hs.PTF1A-short-305:Dr.hsp70:mCherry						23508305A>G	
pDB896_Hs.PTF1A-large-363:Dr.hsp70:mCherry						23508363A>G	
pDB896_Hs.PTF1A-large-WT-:Dr.hsp70:mCherry	1445bp	chr10:23,507,374-23,508,818	GCCCCAGGTTTTAATTTATCA	CAGCCTCCTCTGCTTCTTTA	EcoRV/SpeI	WT	*hsp70*
pDB896_Hs.PTF1A-large-305:Dr.hsp70:mCherry					HindIII/EcoRI	23508305A>G	
pDB896_Hs.PTF1A-large-363:Dr.hsp70:mCherry					HindIII/EcoRI	23508363A>G	
pDB896-PTF1A-large-WT:Dr.gata2:mCherry	1445bp	chr10:23,507,374-23,508,818	GCCCCAGGTTTTAATTTATCA	CAGCCTCCTCTGCTTCTTTA	EcoRV/SpeI	WT	*gata2*
pDB896-PTF1A-large-305:Dr.gata2:mCherry					HindIII/EcoRI	23508305A>G	
pDB896-PTF1A-large-363:Dr.gata2:mCherry					HindIII/EcoRI	23508363A>G	
pDB896_Fugu-control:Dr.hsp70:mCherry	n/a	chrUn:54,537,362-54,537,937	GTGTGTCATCCTCATCCACG	CCATGATGGTGCTCTGCC	attB3/attB5	*Fugu* gDNA	*hsp70*
pDB896-Fugu-control:Dr.gata2:mCherry	n/a				EcoRV/SpeI		*gata2*

4.3 RESULTS

4.3.1 Transient transgenesis assays to test *PTF1A* putative human enhancer

In order to test the regulatory potential of the non-conserved element located 25 Kb downstream of *PTF1A,* I initially verified the pattern of *ptf1a* gene in zebrafish and whether it was also expressed in the pancreas. The endogenous pattern of *ptf1a* in zebrafish is very similar to mouse (Obata et al., 2001), being expressed temporally in the central nervous system, in the pancreas primordium and later in the acini of the exocrine pancreas (Lin et al., 2004; Zecchin et al., 2004). The Tg(ptf1a:EGFP)^{jh1}, a transgenic line created by replacing the coding sequence of *ptf1a* gene by eGFP in a genomic BAC (kindly provided by Elke Ober) reproduced *ptf1a* expression pattern and provided an appropriate anatomical marker and potential co-localization tool (Godinho et al., 2005).

In order to determine whether *PTF1A* putative enhancer could function as a regulatory element in zebrafish and whether the variants found in patients could affect the regulatory potential of this element, the 485bp conserved region from chromosome 10 and its mutations (**Figure 4.1A**) were tested in Tol2-based transient transgenesis assays in zebrafish embryos. For functional tests, two promoters that had previously worked in combination with human enhancers were chosen: *hsp70* and *gata2* promoter*.* Zebrafish *hsp70* promoter was chosen because it drove minimal background and had a general capacity of interacting with enhancers (Gehrig et al., 2009); zebrafish *gata2* promoter was selected because it was highly sensitive in recapitulating expression in the pancreas driven by human enhancers (see **Chapter 3**). As a control region, a non-conserved region from *Fugu rubripes* showing no regulatory activity was used (Sanges et al., 2006). These elements were placed upstream of a red fluorescent protein (mCherry) that would potentially allow for co-localization studies using Tg(ptf1a:EGFP)^{jh1} transgenic line.

Initially three vectors containing the WT sequence and two mutated variants of the inherited 485 bp region from human chromosome 10 linked to zebrafish *hsp70* minimal promoter were tested in transient transgenic assays; but no tissue-specific activity was detected during the first 5 dpf. The putative pathogenic variants did not show any reporter expression either, being comparable to the WT sequence or to the control construct (**Table 4.2**).

Given the lack of activity shown by the 485 bp candidate enhancer I decided to amplify a region of 1,445 bp encompassing the vertebrate conservation block between human and zebra finch (**Figure 4.1B**), with the expectation that a larger fragment could represent the full functional element required for autonomous enhancer activity, and thus could be easily interpreted by our evolutionarily distant model despite the lack of sequence conservation. This large fragment containing variants was cloned upstream of the *hsp70* minimal promoter and the more sensitive *gata2* promoter and injected in zebrafish embryos (**Table 4.1**). However, no regulatory activity driven by the candidate enhancer could be detected when we compared embryos injected with the enhancer-containing constructs and promoter only controls (**Table 4.2**). Taken together, the above results showed that neither variant nor the WT candidate *PTF1A* enhancer could direct detectable activity in the transient transgenic zebrafish embryo when linked to two zebrafish promoters, suggesting lack of functionality in zebrafish.

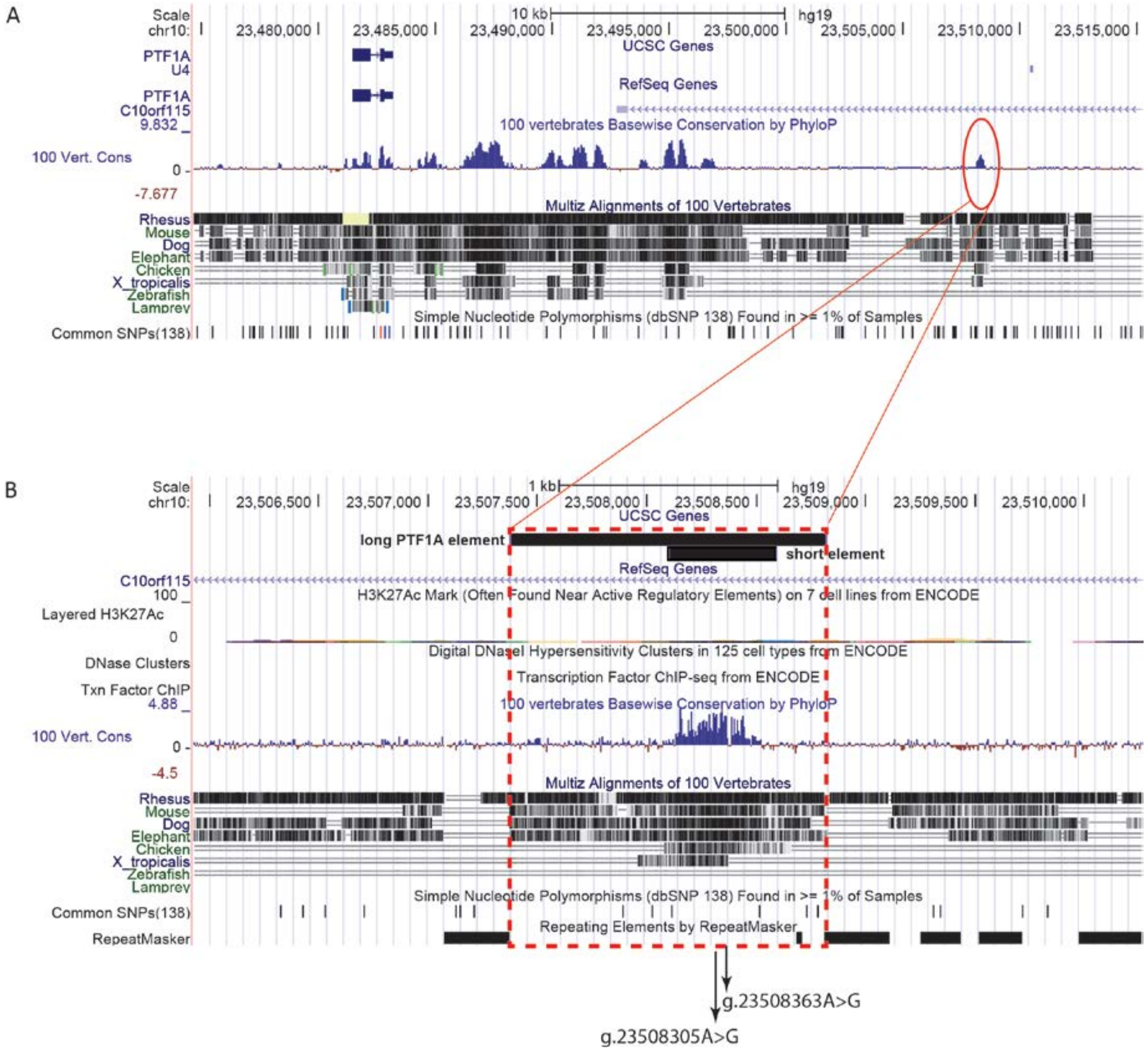


Figure 4.1 USCS screenshot showing the location of putative PTF1A regulatory region.

A. The region of 485 bp identified by whole exome sequencing is located 25 Kb downstream of the *PTF1A* coding gene (red ellipse). **B.** Putative enhancer regions used for zebrafish transgenesis assays are depicted by black rectangles. The large *PTF1A* putative element encompassing 1,445 bp and the vertebrate conservation block is labelled with a dashed red rectangle.

Table 4.2 Summary of transient transgenesis assays performed using *PTF1A* putative human enhancers

Construct	Variant tested	Putative enhancer size	Number of injected embryos	Number of expressing embryos (%)	Number of replicates
pDB896_Hs.PTF1A-short-WT:Dr.hsp70:mCherry	WT	485 bp	310	0 (0)	2
pDB896_Hs.PTF1A-short-305:Dr.hsp70:mCherry	23508305A>G		419	0 (0)	2
pDB896_Hs.PTF1A-large-363:Dr.hsp70:mCherry	23508363A>G		183	0 (0)	2
pDB896_Hs.PTF1A-large-WT-:Dr.hsp70:mCherry	WT	1445 bp	253	0 (0)	2
pDB896_Hs.PTF1A-large-305:Dr.hsp70:mCherry	23508305A>G		212	0 (0)	2
pDB896_Hs.PTF1A-large-363:Dr.hsp70:mCherry	23508363A>G		182	0 (0)	2
pDB896-PTF1A-large-WT:Dr.gata2:mCherry	WT	1445 bp	184	0 (0)	2
pDB896-PTF1A-large-305:Dr.gata2:mCherry	23508305A>G		228	0 (0)	2
pDB896-PTF1A-large-363:Dr.gata2:mCherry	23508363A>G		230	0 (0)	2
pDB896_Fugu-control:Dr.hsp70:mCherry	Fugu gDNA	n/a	217	0 (0)	2
pDB896-Fugu-control:Dr.gata2:mCherry			205	0 (0)	2

4.4 DISCUSSION

In this chapter, I aimed to test the regulatory potential of a region downstream of *PTF1A* containing variants inherited in unrelated patients suffering from pancreatic agenesis by zebrafish transient transgenesis assays. However, none of the nine constructs made to test the candidate enhancer with various combinations of length, linked to two different minimal promoters could show enhancer effect in zebrafish. Moreover, neither the wild type sequence nor the element containing potentially pathogenic variants could drive any tissue-specific activity in transient transgenic assays.

The role of *cis*-regulatory variants in disease is still not very well understood and finding the functional causing variant is still a challenge. We and others have proposed to use the zebrafish model as a tool to validate the function of human enhancers (Ishibashi et al., 2013), and the variants within. Because sequencing of all these constructs was carried out before the reporter expression assays took place, we can exclude the possibility that mutations introduced during cloning are responsible for the lack of activity. In the same line, both the Gateway Expression Vectors and the classical screening vectors used to validate *PTF1A* putative enhancers had been previously used by us to test the activity of human enhancers (as discussed in **Chapters 3** and **5**), with reliable detection of function.

From these results we can propose that either lack of sufficient degree of sequence conservation, lack of enhancer-promoter interaction, or lack of enhancer function *in vivo,* could be the reason for the lack of reporter activity in zebrafish. 3C analysis showed that the candidate enhancer is interacting with *PTF1A* promoter in human pancreatic progenitor cells (Weedon et al., 2014). To test in zebrafish whether the specificity of the enhancer-promoter interaction is the reason for the lack of function, the human putative enhancer would have to be cloned upstream of the endogenous *ptf1a* zebrafish promoter. Nevertheless, the promoters used for the assays: *hsp70* and *gata2*, have been shown to have a general ability to interact with enhancers (Gehrig et al., 2009) and had been active in combination with several pancreas-specific human enhancers before (as discussed in **Chapter 3**). Moreover, zebrafish *gata2* promoter has been routinely used by us and others in similar enhancer tests and has proven very sensitive to unravel the full regulatory potential of human CREs (Bessa et al., 2009; Navratilova et al., 2009; Ragvin et al., 2010; Royo et al., 2012).

Despite evidence demonstrating that zebrafish can recapitulate regulatory function in the lack of sequence conservation (Fisher et al., 2006a; McGaughey et al., 2009), our experience with testing a limited set is that non-conserved elements are less likely to work than conserved elements. *PTF1A* putative enhancer shares no sequence similarity to the zebrafish genome (0% sequence alignment), therefore we could attribute the lack of activity to the lack of conservation.

One piece of evidence that suggests that the negative results from zebrafish assays might not be necessarily due to the large evolutionary distance between human and zebrafish, is provided by additional functional assays carried out with human cells lines and mouse embryos. In parallel to the zebrafish assays, Jorge Ferrer´s lab tested the functionality of *PTF1A* element using luciferase assays in progenitor and adult pancreas cell lines. Their results, which have been published in (Weedon et al., 2014), argue that the 485 bp element is active at low levels in pancreatic progenitor cells (not in adult exocrine cell lines) and that the 23508363A>G variant abolishes enhancer activity through the disruption of a FOXA2 binding site, while the 23508305A>G variant disrupts the binding of a non-identified protein as demonstrated by EMSA (Electrophoretic Mobility Shift Assay). Furthermore, the putative *PTF1A* enhancer region is enriched in typical enhancer marks such as H3K27ac or H3K4me1 in hESCs derived from pancreatic progenitors, although at very low levels (Weedon et al., 2014).

In zebrafish, pancreatic progenitors giving rise to the anteroventral and dorsolateral buds appear at the 12-somite stage, express *pdx1* and are located at both sides of the midline (Argenton et al., 1999; Biemar et al., 2001), while *ptf1a* expressing cells appear first at 32 hpf (Lin et al., 2004). We could therefore argue that our Tol2-based transgenesis system was not sensitive enough to detect such small cell numbers or even the stage at which embryos were analysed might not have been suitable.

Interestingly, when Weedon's collaborators tested *PTF1A* putative enhancer *in vivo* by transgenesis assays in mice using a construct where *PTF1A* enhancer was linked to a minimal viral promoter and GFP, E9.5-E.10-5 embryos did not show enhancer-driven reporter gene expression; not even when they tried to amplify the signal using immunofluorescence techniques (Inês Cebola, personal communication).

Conflicting results between *in vitro* and *in vivo* data in two species could be attributed to a suboptimal genomic environment of the isolated enhancer. The fact that certain isolated enhancer sequences can function in transgenesis assays does not mean that all are able to function independently outside of their regulatory context; and therefore, might require interactions with other *cis*-regulatory elements to regulate transcription (Frankel, 2012). Such physical interactions have been described during digit development in mouse, where several enhancer elements spread in a gene desert forming a "regulatory archipelago" contribute quantitatively or qualitatively to full transcriptional response of *HoxD* genes (Montavon et al., 2011).

Epigenetic silencing of the transgene might also be a potential cause for lack of *in vivo* function. The lab of Mary Goll has reported transgenerational epigenetic silencing of integrated transgenes mediated by DNA methylation, where transgene reactivation was possible but occurred in a cell-specific manner (Goll et al., 2009). We could hypothesize that instead of transgenerational inactivation, our transgene is being silenced through cell differentiation, which could point to an inherent property of certain enhancers. Testing the enhancer in its endogenous regulatory context would require the development of more elaborate loss of function assays mediated by genome editing tools such as CRISPR (Shen et al., 2014).

Attributing the lack of *PTF1A* activity in zebrafish to lack of enhancer function is thus partially refuted by the work of Weedon and colleagues. On one hand the region is enriched in typical enhancer marks such as H3K27ac or H3K4me1 in hESCs derived from pancreatic progenitors, although at very low levels. The putative CRE acts as a developmental enhancer in cell culture assays but does not function *in vivo* either in mouse or in zebrafish transgenesis experiments. Taken together, conflicting functional results suggest that this enhancer could be very stage-specific and possibly have a very weak transcriptional effect. We can conclude that the lack of activity of this region could be explained by either lack of enhancer conservation at the sequence level, by lack of enhancer-promoter interaction, by the lack of a sensitive transgenesis system able to detect subtle activity in early development affecting a few number of cells; or because it is not an autonomous enhancer and it is only able to function on transient cell culture assays in a context dependent fashion.

Chapter Five: VALIDATION OF BIDIRECTIONALLY TRANSCRIBED HUMAN ENHANCERS

FOREWORD

The results presented in this chapter have been partially published in (Andersson et al., 2014).

This project is the result of a collaboration between our lab and Robin Andersson and Albin Sandelin from the University of Copenhagen (Denmark), members of the FANTOM5 International Consortium. They provided all the computational data presented in this chapter, which formed the basis of *in vivo* functional analysis of predicted human enhancers.

5.1 INTRODUCTION

5.1.1 Limitations of current approaches for enhancer prediction

Genome-wide strategies for enhancer prediction have been based on the use of multiple layers of information that include comparative genomics, searching for open chromatin sites and/or profiling of tissue-specific TFs, general co-factors and histone modification marks (Visel et al., 2009a; Zinzen et al., 2009; Rada-Iglesias et al., 2011; Bernstein et al., 2012; Rada-Iglesias et al., 2012). Nevertheless, there are limitations to all of these approaches. Conservation is not always indicative of function and there are subsets of functional enhancers that are not bound by general co-factors. Furthermore, ChIP-Seq, which is frequently used to map protein-DNA interactions, has technical limitations (Furey, 2012). Restrictions include the lack of appropriate antibodies (particularly in zebrafish) and the large numbers of cells required per experiment (in the order of tens of millions); factors that represent a challenge when dealing with small cell-numbered organisms or cells isolated from a tissue. Therefore, there is an opportunity for the implementation of a complementary strategy to predict enhancers.

5.1.2 Enhancers produce transcripts

Multiple lines of evidence have recently indicated that transcription from enhancers is a general property (Kim et al., 2010; Melgar et al., 2011; Wang et al., 2011a; Djebali et al., 2012). Initial studies demonstrated that mouse neuronal enhancers marked by H3K4me1 and CBP co-activator binding can recruit PolII and transcribe bidirectionally a novel class of RNA, which was then termed enhancer RNA (eRNA, (Kim et al., 2010)). Subsequent RNA-Seq analysis excluded the possibility that PolII was in fact associated to the promoter and had been crosslinked to the enhancer during ChIP experiments, and it clearly revealed bidirectional transcription towards H3K4me1 enriched nucleosomes (Kim et al., 2010). This reported phenomenon was then used for modelling a strategy to predict enhancer-associated bidirectional expression of short transcripts in a single cell line (Melgar et al., 2011). This study showed that there is widespread bidirectional transcription not associated with promoters and co-occurring in open chromatin regions enriched in H3K27ac and H3K18ac (potentially marking active enhancers), while lacking H3K27me3 and H3K9me3 repressive marks. These observations allowed the authors to use the bidirectional signature of transcription as a genome-wide enhancer predictive tool (Melgar et al., 2011). Global run-on sequence analysis (GRO-Seq, which maps the position, amount and orientation of transcriptionally engaged PolII genome-wide, (Core et al., 2008) also confirmed that active enhancers in human prostate adenocarcinoma cells produce non-coding eRNAs (Wang et al., 2011a).

The fact that both gene-rich and gene-poor regions are pervasively transcribed was already hinted by the first phase of the ENCODE Project, which analysed 1% of the human genome (Birney et al., 2007). The second phase of the Project interrogated the human transcriptome genome-wide and showed that in one of their cell lines there was a significant overlap of CAGE tags with ENCODE-predicted enhancers (Djebali et al., 2012). These tags extended a few kb and could be seen in both polyadenylated and non-polyadeylated RNA fractions. The enhancer regions showing transcription also displayed a different chromatin signature than non-transcribed enhancers, consisting of a high enrichment of H3K4 methylation, H3K27 acetylation as well as high levels of PolII recruitment (Djebali et al., 2012). What these studies left unanswered is whether bidirectional transcription could be used as a novel enhancer predictive tool and whether predicted enhancers would be functional in the context of a vertebrate organism.

5.1.3 Preliminary data leading to the project

FANTOM (Functional ANnoTation Of the Mammalian genome) is an international consortium established by Dr. Hayashizaki at RIKEN (Japan) that aims to assign functional annotations to mammalian genomes. FANTOM5 Project has aimed to generate a map of human promoters and enhancers through the generation of a TSS atlas that includes 432 human primary cell types and 135 human tissue samples (FANTOM Consortium, 2014) using single molecule CAGE (Kanamori-Katayama et al., 2011).

Andersson and colleagues found that there was bidirectional transcription of short RNAs at enhancer locations and they used this signature to systematically predict active enhancers and characterize their usage across the FANTOM5 tissue atlas, which covers the majority of human cell types and tissues. Enhancer locations were defined by the co-occurrence of P300, H3K27ac and H3K4me1 marks. When this data was overlaid with CAGE tags, they could identify a genome-wide bidirectional transcription pattern where tags would extends away from the enhancer centre, flanked by H3K27ac and H3K4me1 ChIP signals (**Figure 5.1**).

These initial observations proved to be valid with well-known enhancers from VISTA and were confirmed as a general feature across the FANTOM5 CAGE atlas, proceeding to identify more than 43,000 enhancer candidates in 808 human CAGE libraries. The RNA produced by enhancers is approximately 350 bp in length, capped, unspliced, contains termination sequences and is rapidly targeted by the exosome. Most of the transcripts are nuclear and non-polyadenylated. In contrast, mRNA produced from promoters are in average 1246 bp in size, are biased towards the sense direction and contain splice sites.

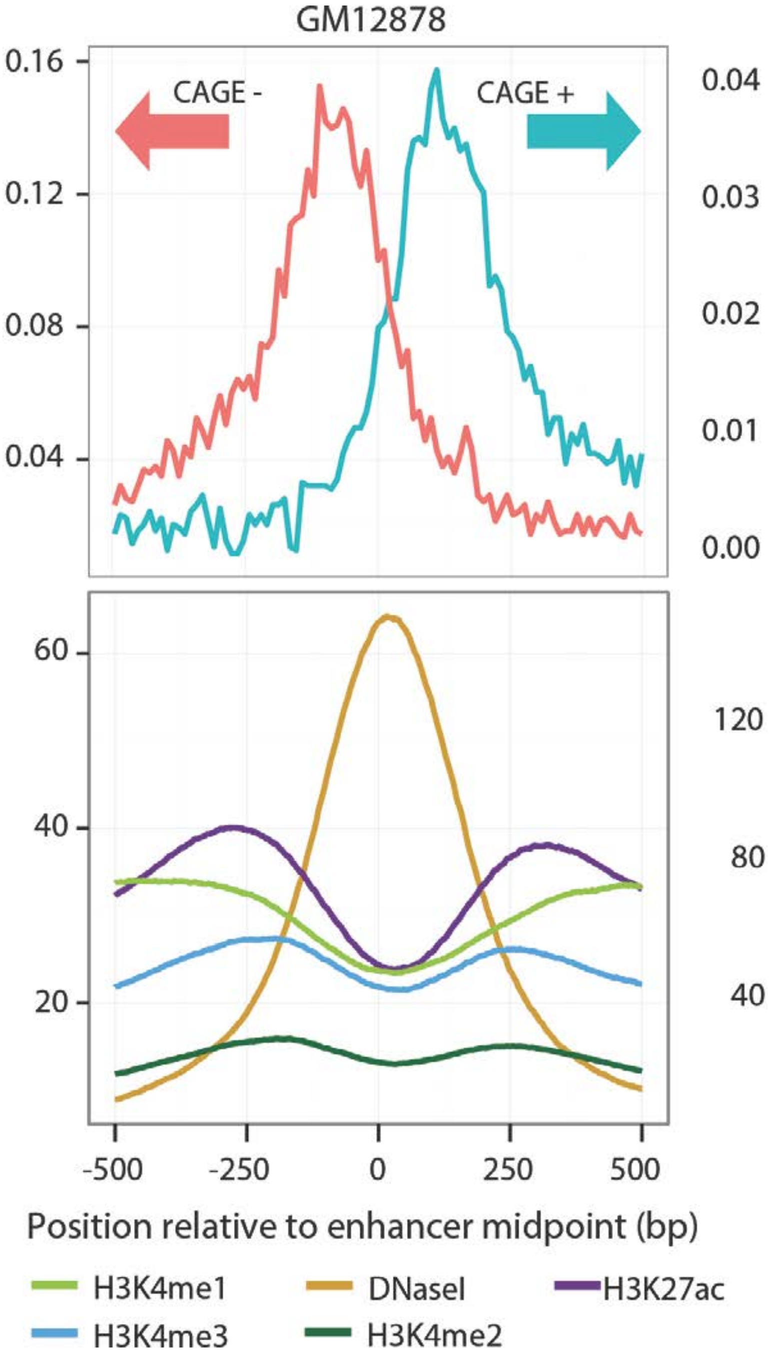


Figure 5.1 Bidirectional transcription is a signature of active enhancers across the human body. Overlay of enhancer sites identified by the presence of H3K4me1 and H3K27ac (bottom), with CAGE tags showing a bimodal distribution relative to the P300 centre point (top). This figure was produced by Robin Andersson and has been adapted from (Andersson et al., 2014).

5.1.4 Aims:

To address whether candidate enhancers predicted by the presence of bidirectional transcription were able to drive tissue-specific expression in the complexity of the whole organism, I decided to carry out transient and stable transgenic assays using the zebrafish embryo. Within this global aim, the following objectives were also set:

- To investigate whether bidirectionally transcribed sequences, predicted as enhancers, are active in zebrafish.
- Determine whether enhancers recapitulate human or zebrafish endogenous expression patterns.
- Evaluate if ubiquitously transcribed candidate enhancers show activity in the zebrafish embryo.

5.2 RESULTS

5.2.1 Selection of bidirectionally transcribed human putative enhancers based on sequence conservation.

To ask whether human enhancers could function in a living organism, we selected bidirectionally transcribed enhancers that were only expressed in a subset of tissues or cells from the FANTOM5 CAGE Atlas. Selection was based on human-zebrafish conservation (>70% sequence identity over 100 nt, hg19 vs DanRer7), in order to take into account the large evolutionary separation between the two species. We prioritized candidates that had zebrafish orthologous target genes with known endogenous patterns (www.zfin.org), so as to be able to compare potential reporter activity to both human and zebrafish patterns. Enhancer candidates that were active in human tissues with no zebrafish correspondence, such as lungs or trachea, were excluded. Epigenetic marks including histone modifications or DHS sites were not taken into consideration during the selection process, so as to rely only on the prediction power of the transcriptional mark. Three control regions were also chosen randomly from the human genome with the following constraints: low sequence conservation with zebrafish and no other enhancer-selective feature, that is, no DNase hypersensitivity, no H3K4me1 or H3K27ac signals and CAGE signal only at noise levels. The enhancer candidates selected for zebrafish validations are listed in **Table 5.1**.

5.2.2 Transient transgenesis assays of bidirectionally transcribed human enhancers

To test if candidate enhancers could drive tissue-specific expression in the complexity of the whole organism, I tested evolutionarily conserved CAGE-defined human enhancers by Tol2-mediated transgenesis in zebrafish embryos. I cloned ~1kb regions surrounding the *cis*-regulatory element centre point into a pDB896 vector containing *gata2* zebrafish promoter and YFP reporter gene (**Table 5.2**). Zebrafish *gata2* promoter was used because previous enhancer screens carried out in the lab showed that it is more sensitive than core promoters such as *hsp70* (see **Chapter 3**).

Table 5.1 Tissue specificity of selected human enhancers and control regions based on CAGE expression atlas and zebrafish counterparts

***Cis*-regulatory element ID**	**Nearest human coding gene(s)**	**Human tissue expression (FANTOM 5 Atlas)**	**Zebrafish orthologous gene expression (www.zfin.org)**
CRE1	*TMEM161b* and *MEF2C*	Fetal brain, neurons, optic nerve	Branchial arches, myotome boundaries, pectoral fin muscles, posterior hypaxial muscles
CRE2	*POU3F2*	Brain specific	Central nervous system
CRE3	*SOX7*	Endothelial cells, salivary gland, pineal gland and aorta	Vasculature, endoderm, ysl, rhombomeres
CRE4	*PAX6* and *RCN1*	Retina, heart and brain	Retina, pancreas and brain
CRE5	*DLX1* and *DLX2*	Brain	Brain
CRE6	*HNF1B*	Liver	Liver
CRE7	*TBR1* and *PSMD14*	Brain specific	Olfactory bulb and telencephalon
CRE8	*PTPRN2*	Brain specific	Central nervous system
Control region 1	*C11orf74*	No expression	Not spatially restricted
Control region 2	*DSCAM*	No expression	Central nervous system and eye
Control region 3	*PBX1*	No expression	Central nervous system and posterior branchial arches

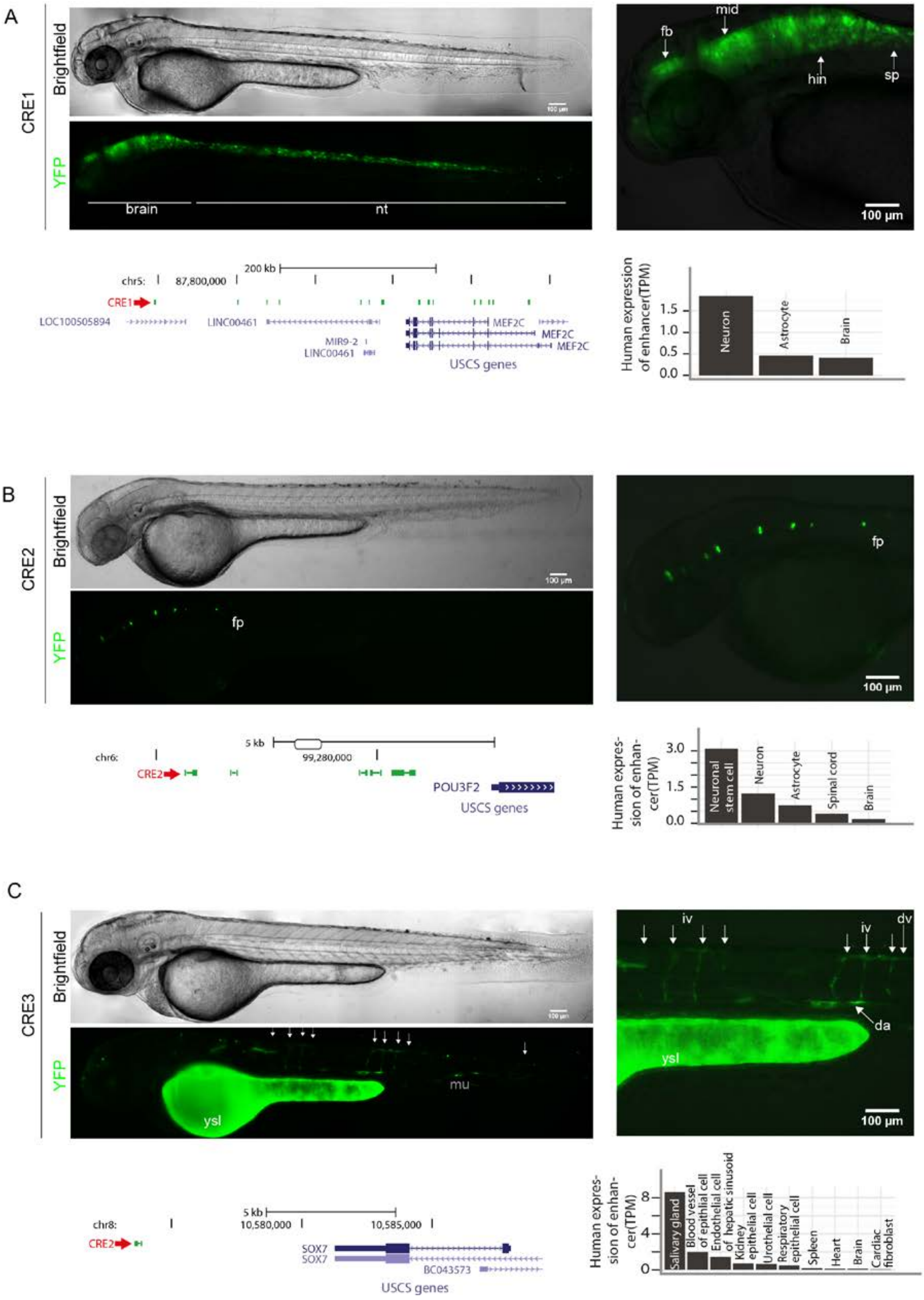
Table 5.2. Selected human *Cis*-Regulatory Elements (CRE) and control regions used in zebrafish transient reporter assays.

Cis-regulatory element ID	Nearest human gene(s)	Coordinates (Hg19 or Zv7)	Size (bp)	Forward primer (5' to 3')	Reverse primer (5' to 3')	5' Tail	3' Tail
CRE5	*DLX1* and *DLX2*	chr2:172958266-172959477	1211	CCAGACCCATCCTCCTATCTTGA	GCGGTAGAGACAAAAGAAGAGCC	HindIII	EcoRI
CRE4	*PAX6* and *RCN1*	chr11:31840987-31842070	1083	AGAAAAGAGGTTTCTTTCCCGCT	GGGAGCTTTGGCTGAGAAGTTT	EcoRV	SpeI
CRE3	*SOX7*	chr8:10573085-10574291	1206	CTTTGCTCTCATGCTGCTTGTCT	AAGATGACACTGAAAAGGGGGAG	EcoRV	SpeI
CRE2	*POU3F2*	chr6:99275060-99276226	1166	GTTTTCCCCTCACTCTTCTGAGC	CTAGTACTTCGGTCTGGGGTGCT	EcoRV	SpeI
CRE1	*TMEM161b* and *MEF2C*	chr5:87692532-87693408	877	CCCAAGGGAAGTCACGTAAA	TGAGCCTTGGGTTTTTGTTT	EcoRV	SpeI
Control region 1	*C11orf74*	chr11:38087624-38088486	863	CATGGCAATCACCACTTCTG	CCGACTGGGATGATTGATCT	HindIII	EcoRI
Control region 2	*DSCAM*	chr21:41847729-41848678	950	GCCTGGGCAACAGAGTAAGA	ATTTTTGAGCCCTTCCCATC	EcoRV	SpeI
Control region 3	*PBX1*	chr1:163991571-163992470	900	CAGGCCTCAACCCATTTCTA	AACAGGTGGCACTCCTATGG	EcoRV	SpeI
Zebrafish *gata2* promoter	n/a	chr11:3922100-3923130	1031	ATTCATTAATAGAATAGAGGCATT	CTCAAGTGTCCGCGCTTA	AgeI	XhoI

CRE6-8 overlapped with highly repetitive genomic regions and could not be amplified, therefore, only 5 conserved elements (CRE1-5) were used for transgenesis assays. Additionally, three control regions were also cloned and analysed to check the specificity of the enhancer selection process. Constructs were injected at one-cell stage and reporter expression patterns were analysed at long-pec stage (48 hpf) by fluorescent microscopy (**Table 5.3**). The levels of expression are depicted as a percentage between the number of expressing embryos versus the total number of injected embryos. In order to control for overall background activity from the construct (i.e., promoter, backbone) an empty pDB896 vector containing the zebrafish *gata2* promoter linked to YFP reporter gene but lacking an enhancer sequence was injected and analysed in parallel. Any tissue-specific enrichment shown by enhancer-containing vectors over the activity shown by the empty control vector (ectopic expression or background) was considered enhancer-specific.

I observed tissue-specific enhancer activity with 3 out of 5 human CREs, which corresponded to the human enhancer tissue expression (**Table 5.1**). CRE1, which is located in an intronic region of the *TMEM161b* gene and is 450 Kb downstream of the *MEF2C* gene, was robustly expressed in the central nervous system (CNS) including the forebrain, midbrain, hindbrain and spinal cord regions, reproducing the activity found with the human enhancer in fetal brain and neurons (**Figure 5.2A**). Around 82% (n=202) of the expressing embryos showed this neural-specific expression pattern, with minimal background in the yolk syncytial layer (ysl, **Table 5.3**).

CRE2, which is 7 Kb upstream of *POU3F2* was weakly expressed in the floor plate of the zebrafish embryo, that is, the ventral section of the neural tube. The neuronal pattern matched the tissue-specificity found in the human CAGE expression atlas, as it was active in human neuronal stem cells (**Figure 5.2B**). The resolution of this experiment did not allow us to address whether these are the same sets of neurons.



(See next page for figure legend)

Figure 5.2. Validations of in vivo activity of CAGE-defined human enhancers CRE1-3 in zebrafish embryos at long-pec stage.

Representative YFP and brightfield images of embryos injected with human enhancer-containing constructs are shown. UCSC browser left sub-panels depict the human genomic landscape (including USCS gene track) around the validated enhancer (red arrow). Right sub-panels represent CAGE expression in human tissues or cell types for the enhancer, measured in Transcripts Per Million (TPM). Note the correspondence between zebrafish and human enhancer expression patterns. **A.** CRE1, ~230kb upstream of the *MEFC2* gene, drives highly robust expression in the brain (brain) and neural tube (nt). Right panel shows zoom-in overlay image showing expression in the forebrain (fb), midbrain (mid), hindbrain (hin) and spinal cord (sp). **B.** CRE2, 5kb upstream of the *POU3F2* gene, is active in the floor plate (fp). Right panel is a zoom-in overlay image. **C.** CRE3, 10kb upstream of the *SOX7* gene TSS, shows specific expression in the vasculature, including intersegmental vessels (iv), dorsal vein (dv) and dorsal aorta (da). Detail is shown on the right panel. Muscle (mu) and yolk syncytial layer (ysl) activities are background expression coming from the *gata2* promoter-containing reporter construct (**Table 5.3**). All embryos are 48 hpf oriented laterally with the head to the left.

Table 5.3 Quantitation of transient expression displayed by long-pec zebrafish embryos injected with selected human *Cis*-Regulatory Elements (CRE) and control regions.

CRE ID	No. of injected embryos	Ectopic expression (background, %)					Enhancer driven specific expression (%)			Total % of expressing embryos	Unspecific expression (%)	Specific expression (%)	Non-expressing embryos (%)
		Muscle	YSL	Heart	Ubiquitous	Combination of tissues	CNS	Floor plate	Vasculature				
CRE5	218	**14.7**	**4.6**	0.0	**18.3**	**9.2**	0.0	0.0	0.0	46.8	46.8	0.0	53.2
CRE4	234	**24.8**	**6.8**	**0.8**	**7.3**	**5. 6**	0.0	0.0	0.0	45.3	45.3	0.0	54.7
CRE3	162	0.0	**17.5**	0.0	0.0	**25.0**	0.0	0.0	**12.5**	55.0	42.5	12.5	45.0
CRE2	160	**10.0**	**6.2**	0.0	0.0	**5.6**	0.0	**7.5**	0.0	29.4	21.9	7.5	70.6
CRE1	202	0.0	**0.5**	0.0	0.0	0.0	**81.2**	0.0	0.0	81.7	0.5	81.2	18.3
Control 1	198	**19.7**	**14.1**	0.0	**4.5**	**3.5**	0.0	0.0	0.0	41.9	41.9	0.0	58.1
Control 2	187	**17.7**	**9.6**	0.0	**5.5**	**7.5**	0.0	0.0	0.0	40.1	40.1	0.0	59.9
Control 3	205	**25.4**	**6.8**	**0.5**	**5.4**	**4.9**	0.0	0.0	0.0	42.9	42.9	0.0	57.1
Promoter control	156	**23.7**	**8.9**	**2.5**	**8.3**	**13.5**	0.0	0.0	0.0	57.0	57.0	0.0	42.9

***Table legend:** Transgene-driven reporter expression at 48 hpf is depicted as a ratio between the number of embryos showing tissue-specific activity versus the total number of embryos injected with the enhancer-containing construct or “enhancer-less” *gata2* promoter-containing control vector. Merged numbers from 3 independent injection experiments are shown. Ectopic expression (background) includes all expression domains displayed by the control vector. Enhancer-specific expression includes tissue-specific enrichment beyond the activity shown by the control vector.

CRE3, which is 8Kb downstream of the 3´ end of the *SOX7* gene, acted as a weak enhancer in the zebrafish vasculature, with expression in the intersegmental vessels, dorsal vein and dorsal aorta. This expression pattern partially resembled the human activity, since CRE3 showed expression in salivary gland and blood vessels in human tissues (**Figure 5.2C**). Notably, moderate to high levels of background in tissues such as ysl, muscle, heart (or a combination of them) were found in embryos injected with CRE2 and CRE3 (**Table 5.3**). Thus, we considered CRE2 and CRE3 to be weak enhancers, since there was tissue enrichment over the activity shown by the empty control vectors (7.5% (n=160) and 12.5% (n=162), respectively) but also ectopic background. This is depicted graphically in (**Figure 5.3**).

CRE4 and CRE5 could not direct tissue-specific expression. Similarly, none of three control fragments without CAGE signal activated the *gata2* promoter (**Table 5.3**). Taken together the above results showed that 3 candidate enhancers reproduced the activity displayed in the CAGE human tissue atlas, suggesting that bidirectional transcription could be used as a predictive enhancer tool.

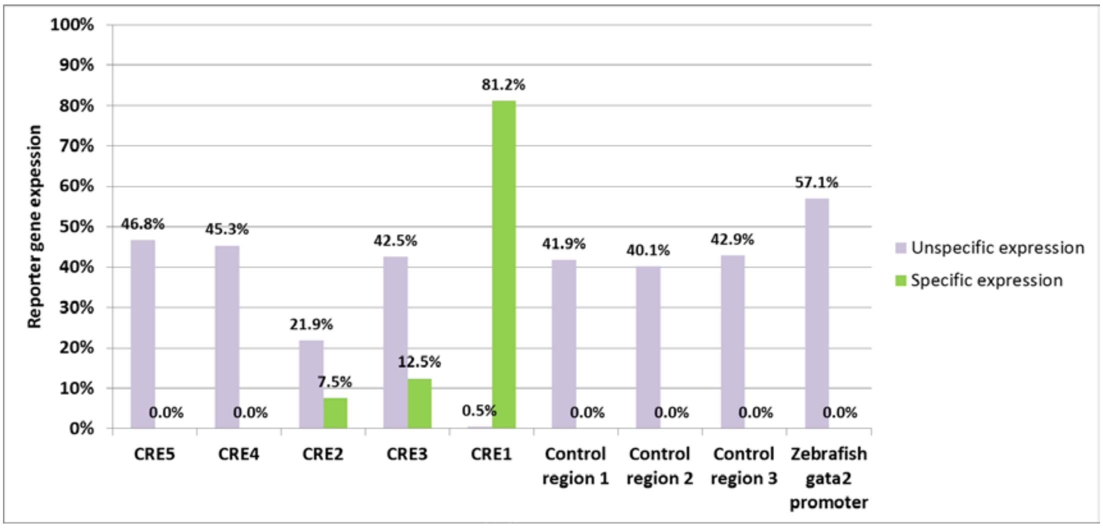


Figure 5.3 Tissue specificity of enhancer activity in mosaic transgenic zebrafish embryos
Percentage of zebrafish injected embryos at 48 hpf showing tissue specific expression (driven by human CRE1-3, green bars) and unspecific expression coming from the enhancer-less *gata2* promoter containing vector (lilac bars). For detailed expression patterns refer to **Table 5.3.**

It has been reported that SINE B2 sequences in mice could actively repress *cis*-regulatory activity by harbouring CTCF binding sites that act as insulators (Schmidt et al., 2012). Although a genome-wide enrichment of CTCF sites was not found in human SINE sequences, taking into account that CTCF binding events are more conserved than tissue-specific TFs, we wanted to test whether repetitive sequences were responsible for the weak activity of two of our enhancers. Thus, we decided to scan the candidate CREs and remove any SINE sequences, if found. Among CRE1-5, only CRE3 contained 168 bp SINE repetitive region on its 5´end (**Figure 5.4**) and was bound by CTCF in several ENCODE lines. Thus, the CRE3 element excluding the repetitive fragment was re-amplified and its activity was re-evaluated in Tol2-based transient zebrafish assays. I found that in three experimental replicates where both the original element and the shortened CRE were injected in parallel under the same conditions, 6.6% (n=389) and 8.2% (n=365) of the total injected embryos respectively were active in the vasculature. Thus, reporter assays with the shortened 911 bp fragment did not show any statistically significant change in terms of either expression pattern or frequencies (p= 0.4876), suggesting that the low activity of this enhancer was not directly related to the presence of the SINE fragment and that the SINE element was not required either for autonomous activity of the enhancer.

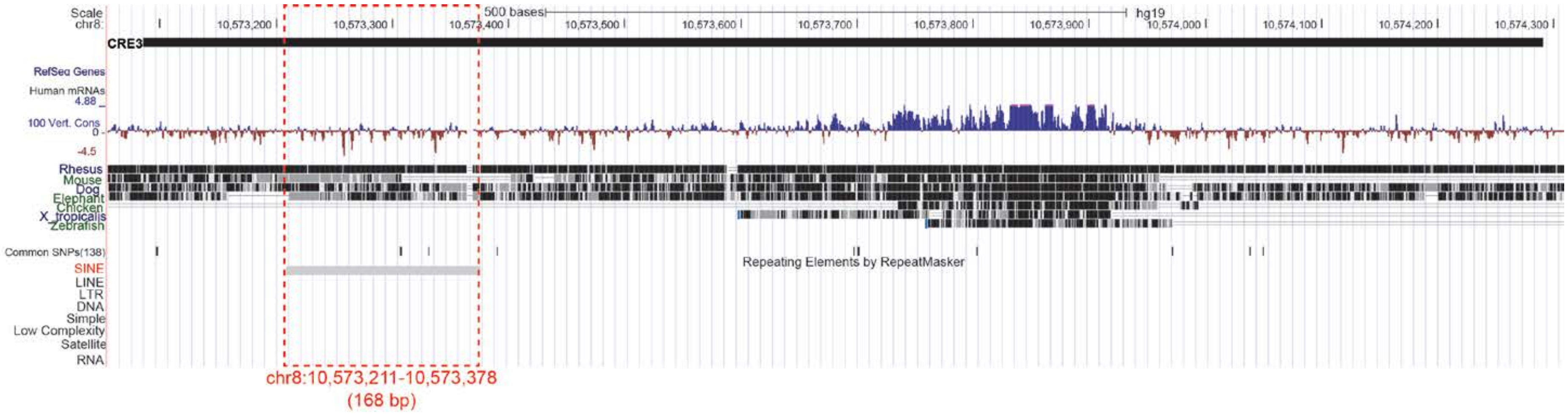


Figure 5.4 UCSC screenshot depicting genomic location of CRE3 putative enhancer and SINE region
CRE3 (depicted as a black horizontal bar) contained a SINE region located on the 5´end (labelled by a dashed red rectangle). A new fragment of 911 bp that excludes this repetitive region was tested (chr8:10,573,380-10,574,291).

5.2.3 Validation of transient transgenesis results by establishing stable transgenic lines

In order to verify whether the patterns seen in transient transgenic embryos were confirmed in stable lines, I decided to grow up embryos injected with CRE1-3-containing constructs. Adults were bred to WT fish and analysis of their offspring led to the identification of 5 different transgenic lines containing CRE1 candidate enhancer (**Table 5.4**). All transgenic lines displayed a similar pattern to that seen in transient transgenic embryos (**Figure 5.5**). No positive founders containing CRE2 candidate enhancer could be identified, despite screening a comparable number of fish. The 2 founders with CRE3 displayed expression patterns that were not similar to the enhancer activity seen in transient transgenic embryos or that could be related to the zebrafish orthologous gene expression pattern, suggesting enhancer traps events (**Figure 5.6**). Taken together, stable transgenic lines allowed us to verify the activity of CRE1. More transgenic lines would be needed to conclude on the activity of CRE2 and CRE3 candidate enhancers.

Table 5.4 Analysis of the offspring of CRE1-3 containing founders

Transgenic line name	No. of screened founders	No. of positive founders	Germline transmission rate (%)	Line ID	No. of embryos analysed	Transmission rate per pattern (%)	Expression pattern
Tg(hs-TMEM161-1.0dr-gata2-pr:YFP)	8	5	62.5	CRE1-L1	57	26.3	Brain, neural tube
				CRE1-L2	106	3.8	Brain, neural tube
				CRE1-L3	275	8.7	Brain, neural tube
				CRE1-L10	140	8.6	Brain, neural tube
				CRE1-L30	68	17.6	Brain, neural tube
Tg(hs-POU3F2-0.1dr-hsp70-pr:YFP)	22	0	0.0	n/a	n/a	n/a	n/a
Tg(hs-SOX7-0.1dr-hsp70-pr:YFP)	20	1	5.0	CRE3-L1	120	10.8	Lens, tectum, neural tube, jaws
				CRE3-L2		5.8	Lens, neurons

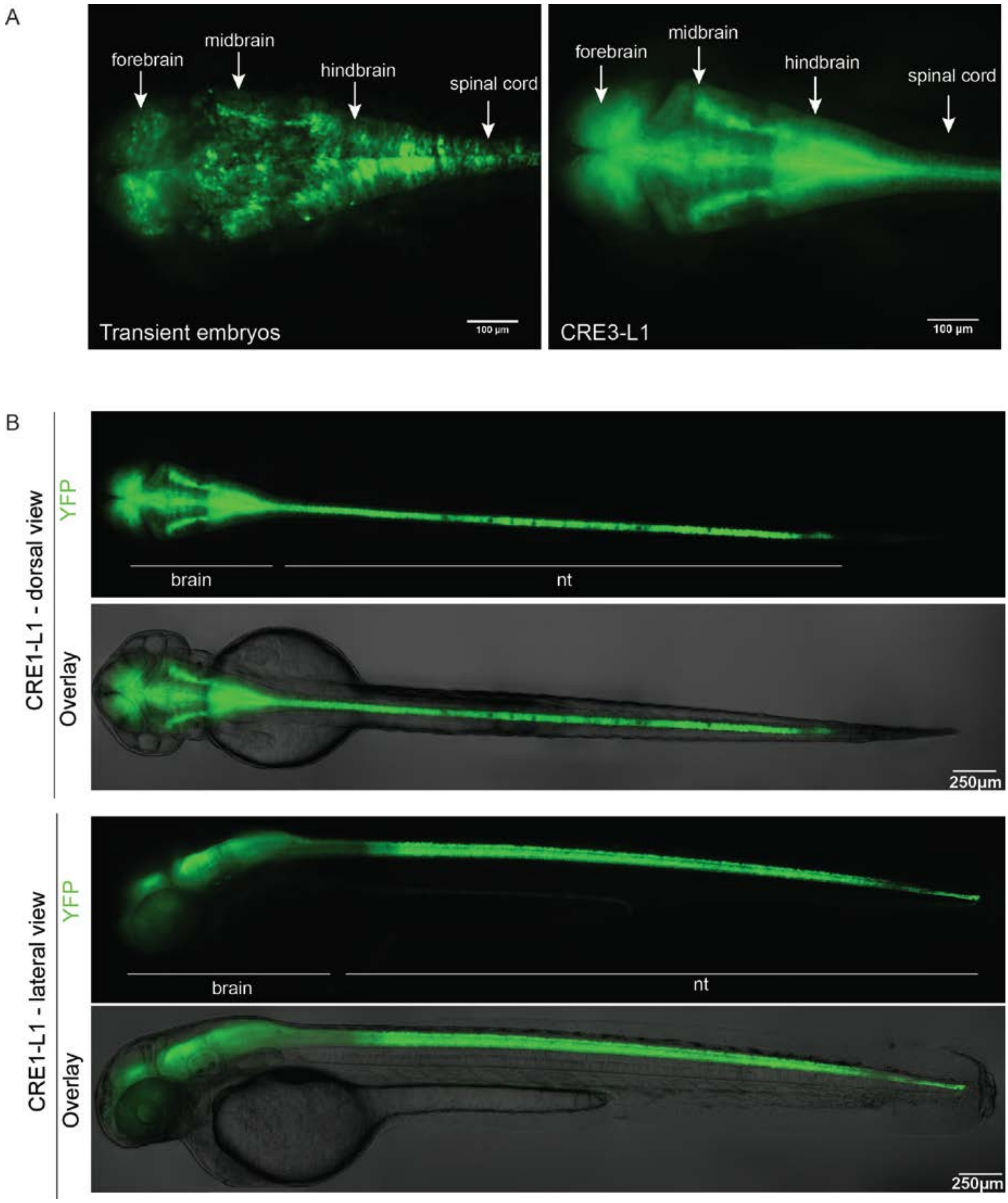


Figure 5.5 Stable lines containing CRE1 recapitulated faithfully the expression pattern seen in transient transgenic zebrafish embryos

A. Embryos injected transiently with CRE1 (left panel) displayed a very similar expression pattern to the established transgenic lines (right panel), recapitulating human tissue activity. **B.** Full expression pattern driven by CRE1-containing transgenic lines. All embryos are 48 hpf anterior to the left.

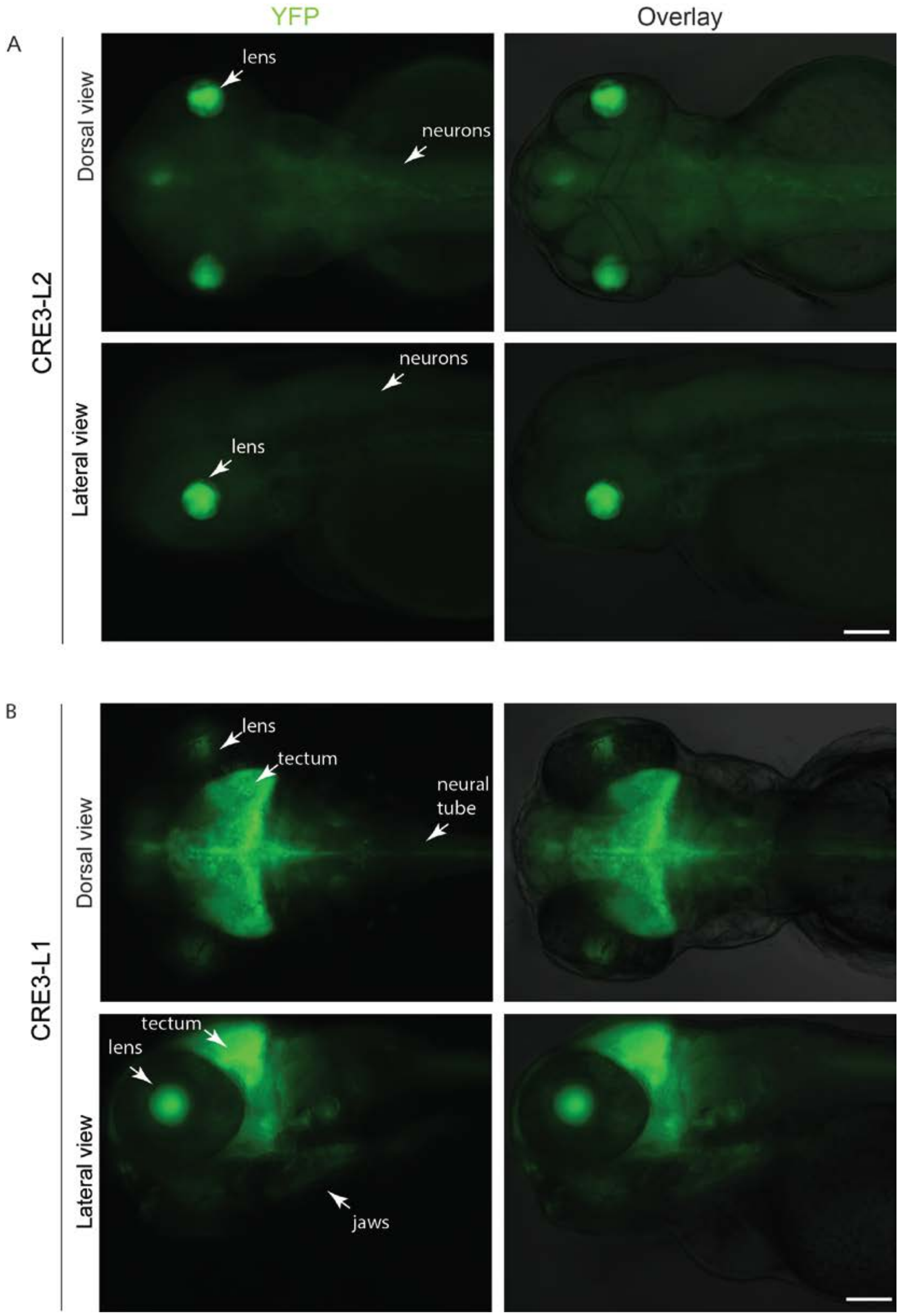


Figure 5.6 Stable lines containing CRE3 showed expression patterns not relevant to the human or zebrafish activities.

Two transgenic lines displayed different expression patterns to one another that are not related to the human or zebrafish expected activities, suggesting position effects. Embryos are 42-48 hpf oriented anterior to the left. Scale bar indicates 100 µm.

5.2.4 Test of *in vivo* function of ubiquitously expressed candidate enhancers

Aside from enhancer candidates that were expressed in specific human tissues or cell types, Andersson and colleagues identified sequences that were transcriptionally active in the majority of tissues and cell lines encompassing the CAGE-expression atlas. We asked whether these elements would be able to drive any expression in the zebrafish embryo. In a proof of principle experiment, we selected 3 candidates that were bidirectionally transcribed, conserved between human and fish and also had enhancer-associated histone modification marks, overlapped DHs and were bound by transcription factors (**Table 5.5**). I cloned the three candidates upstream of the zebrafish *gata2* promoter linked to YFP, and co-injected the constructs with Tol2 mRNA in zebrafish one-cell stage embryos (**Table 5.6**). Analysis of the expression pattern during the first 5 days of development showed that these elements were not active or tissue-specific, as they were not able to control any tissue-specific activity beyond promoter background (**Table 5.7**).

Table 5.5 Tissue specificity of selected ubi-enhancers based on CAGE expression atlas and zebrafish counterparts

CRE ID	Nearest human coding gene(s)	Human tissue expression (FANTOM 5 Atlas)	Zebrafish orthologous gene expression (www.zfin.org)
ubi1	*GTF2A*	Ubiquitous	not spatially restricted
ubi2	*GPR84*	Ubiquitous	not spatially restricted
ubi3	*ZBTB16*	Ubiquitous	central nervous system, branchial arches, pectoral fin, pronephric ducts

Table 5.6 Selected candidate enhancers ubiquitously expressed in human CAGE atlas.

Cis-regulatory element ID	Coordinates (Hg19)	Size (bp)	Forward primer (5' to 3')	Reverse primer (5' to 3')	5' Tail	3' Tail
ubi1	chr14:81685343-81686243	901	ACCAAAACCAAGTCCTCTGC	CAGTTCCTTCCGAATGGGTA	HindIII	EcoRI
ubi2	chr12:54752583-54753476	854	TGCCTTCCTTCTCTCCTCAA	TCTCGGGATGTGTGTGTGTT	HindIII	EcoRI
ubi3	chr11:114033182-114034045	864	AGATGGCTTCCCCTCATCTT	GTGAATCAGCAGCAGGGTTT	HindIII	EcoRI

Table 5.7 Quantitation of transient expression displayed by high-pec zebrafish embryos injected with ubiquitously expressed candidate enhancers.

CRE ID	No. of injected embryos	Ectopic expression (background, %)					Enhancer-driven expression (%)	Total % of expressing embryos	Unspecific expression (%)	Specific expression (%)	Non-expressing embryos (%)
		Muscle	YSL	Heart	Ubiquitous	Combination of tissues					
ubi1	253	14.2	5.5	2.8	7.9	9.9	**0.0**	40.3	40.3	0.0	59.7
ubi2	257	16.7	4.3	1.9	9.7	8.6	**0.0**	41.2	41.2	0.0	58.8
ubi3	273	17.6	6.6	2.9	11.0	11.7	**0.0**	49.8	49.8	0.0	50.2
Promoter control	156	23.7	9.0	2.6	8.3	13.5	**0.0**	57.1	57.1	0.0	42.9

5.3 DISCUSSION

In this study I tested whether predicted bidirectionally transcribed candidate enhancers were active in the context of a vertebrate embryo, based on the observations made by the FANTOM5 Consortium. Out of 5 evolutionary conserved human enhancers, I observed tissue-specific activity driven by 3 CREs in zebrafish transient transgenesis assays, whereas none of the three control fragments without CAGE signal could function as enhancers, suggesting that bidirectional transcription is predictive of active enhancers. In contrast, none of the ubiquitously expressed enhancers displayed any activity in the zebrafish transgenic embryo, suggesting lack of functionality in zebrafish.

Interestingly, the tissue-specific pattern displayed by the zebrafish embryos injected with the 3 functional CREs corresponded to the human tissues where the putative enhancers were shown to be active *in vitro* and not to the zebrafish endogenous patterns (**Table 5.1**). However, the relatively small set of tested enhancers hindered a meaningful statistical analysis of this correlation.

Recently, the prediction of enhancers has been based on the analysis of histone modification marks, nucleosome-depleted regions, the presence of transcription factor binding sites or a combination of these approaches (Rada-Iglesias et al., 2011; Djebali et al., 2012; Rada-Iglesias et al., 2012). However, although several studies had reported that enhancers could effectively bind to PolII and produce non-coding transcripts (Kim et al., 2010; Melgar et al., 2011), bidirectional transcription had never been used as a signature to predict active enhancers throughout the human body and to our knowledge this is the first study where bidirectionally transcribed enhancers has been functionally tested *in vivo*.

We decided to use transient transgenesis as these type of analyses have proven successful in the past to validate CREs (Navratilova et al., 2009; Narlikar et al., 2010), among others, arguing that reproducible expression patterns seen in hundreds of injected embryos can indicate enhancer function. The establishment of stable transgenic lines on the other hand has been used here to validate the expression patterns seen in transient transgenic assays. In this study, however, I could distinguish between two categories among the 3 functional human CREs. Transient transgenic assays demonstrated that CRE1 was a strong enhancer in the central nervous system of the zebrafish embryo and the expression pattern was validated in 5 independent transgenic lines. These lines showed an almost identical pattern to each other and to the injected embryos. In opposition to that, CRE2 and CRE3 could drive tissue-specific expression, in the floor plate and in the vasculature respectively, but also displayed ectopic expression domains in yolk syncytial layer, muscle, and other tissues that were present in embryos injected with the enhancer-less *gata2* promoter only. Consistent with this, we were unable to confirm these patterns in stable transgenic lines, obtaining offspring that exhibited activity in tissues that were irrelevant to activities expected from the human and zebrafish potential target genes. These results suggested position effects, which are caused by the random integration in the host genome of the transgene flanked by the transposase arms (Wilson et al., 1990; Rossant et al., 2011). Both position effects and partial recapitulation of the expression pattern displayed in transient transgenesis hindered the unambiguous identification of a CRE regulatory potential. In order to conclude better, a higher number of founders would have to be established. Nevertheless, a comparable number of adults were screened for all 3 elements, which suggests that an unknown inherent property of these weak enhancers might be responsible for the lack of reproducible gene expression.

In order to investigate whether the presence of repetitive SINE sequences were responsible for the weak activity of CRE, a “SINE-less” CRE3 was cloned and tested. Nevertheless, the percentage of tissue-specific activity obtained with the “SINE-less” fragment was not significantly different from the whole region, suggesting that this region was not related to the low activity of the human enhancer. As an alternative, lack of enhancer-promoter specificity could also be responsible for weak activity displayed by CRE2 and CRE3, which underlines the often underestimated importance of promoter selection in validation assays.

Overall, although we have demonstrated that human enhancers recapitulate expression patterns relevant to the human activity in 3 tissue-specific CREs, higher-throughput analysis would be desirable to conclude statistically on the validation rates in zebrafish. Using sequence conservation as a filter significantly reduced the number of enhancer candidates suitable for zebrafish assays. As I will discuss throughout the Thesis, in our experience non-conserved human enhancers are not functional in zebrafish, so we believe that this filter, although restrictive is backed up by our results.

Chapter Six: DETECTION OF QUANTITATIVE DIFFERENCES CAUSED BY ENHANCER VARIANTS

FOREWORD

For my studies on developing a method to detect quantitative differences between enhancer variants we established a collaboration with Lorenzo Pasquali and Jorge Ferrer, affiliated to IDIBAPS (Spain) and Imperial College London (UK), and members of the Marie Curie BOLD Network.

The computational analysis of epigenetic marks and selection of appropriate T2D-associated SNPs was carried out by Lorenzo Pasquali.

6.1 INTRODUCTION

6.1.1 Genome-wide association studies as a source of non-coding variability

Several publications have recently confirmed that functional *cis*-regulatory elements, particularly enhancers marked by DNase I hypersensitivity and histone modification marks, are enriched in disease-associated SNPs (Hindorff et al., 2009; Bernstein et al., 2012; Maurano et al., 2012; Lee and Young, 2013; Trynka et al., 2013).

In the last ten years, the capacity to identify common variants has increased enormously thanks to Genome Wide Association Studies (GWAS), which nurtured from the sequencing of the Human Genome (Human Genome Sequencing Consortium, 2004) and from HapMap and 1000 Genomes Projects (HapMap Consortium, 2005; Frazer et al., 2007; Altshuler et al., 2010; Abecasis et al., 2012). GWAS allow for millions of independent SNP association tests with a trait of interest without considering candidate genes, aiming to identify multiple variants that contribute with a small effect to the phenotype (Prokopenko et al., 2008). In contrast, linkage analysis and candidate gene approaches find rare loci that strongly influence a disease (Billings and Florez, 2010).

It is estimated that there are around 10 million common SNPs in the human genome, around one SNP every 300 nt, where the frequency of allelic variants is above 1% (HapMap Consortium, 2003). Their analysis is facilitated by the presence of block-like structures of variants that are in linkage disequilibrium (LD). Two allelic variants are in LD when their genotypes are correlated and one SNP allele can predict the genotype of a closely positioned second variant (Mohlke and Scott, 2012). The particular combination of allelic variants on a chromosome region that is co-inherited is termed haplotype (HapMap Consortium, 2005). LD is commonly measured by 2 parameters: D´and r^2. D´ is a measure of the frequency of the haplotype and ranges from 0 (no LD) to 1 (LD is 100%). New haplotypes can be generated by the occurrence of new mutations or by chromosomal recombination events (HapMap Consortium, 2003), thus, the rate of recombination between the 2 allelic variants will directly influence the decrease of D´ value. The statistic r^2 measures the correlation coefficient between 2 SNPs, and ranges from 0 to 1, being 1 when the association between the two is not disrupted by recombination (HapMap Consortium, 2005). Although GWAS are a source of disease-associated SNPs, they usually point to associations with other correlated variants with which they are in LD but are not powerful enough to identify the variant that is causing the phenotype or disease (Cooper et al., 2010; Schaub et al., 2012).

6.1.2 Common variants associated to disease lie on functional regulatory regions

Several reports have confirmed that single base pair nucleotide variation in *cis*-regulatory regions can have phenotypic effects and increase the susceptibility to human disorders, mostly through the disruption or creation of TFBS (Pomerantz et al., 2009; Musunuru et al., 2010; Harismendy et al., 2011; Smemo et al., 2014). For example, the risk allele of a SNP present in the enhancer of Interferon Regulatory Factor 6 disrupts an AF2-α binding site and confers a higher risk to Van der Woude syndrome (Rahimov et al., 2008). A similar study has demonstrated that the SNP rs6983267, which has been strongly associated to colorectal cancer, lies on an intergenic region that behaves as an enhancer *in vitro*. Interestingly, the risk variant displays differential enhancer activity, and an increased capacity for TCF7L2 TF binding; uncovering a possible molecular mechanism for the increased risk to cancer pathogenesis (Pomerantz et al., 2009). Another example of a disease-associated SNP with clinical relevance is rs12740374, located on a chromosomal region linked to myocardial infarction in humans (Musunuru et al., 2010). The infrequent allelic variant of rs12740374 creates a novel binding site for the liver-specific TF CBPEA, influencing the expression of *SORT1* in the liver. This in turn modulates hepatic secretion of low-density lipoproteins and contributes to the altered phenotype (Musunuru et al., 2010).

Although most of these publications have not used *in vivo* models, a study has recently utilized zebrafish transgenesis to demonstrate differential activity of a SNP linked to Restless Leg Syndrome (Spieler et al., 2014). The authors showed that the risk allele, located in a highly conserved non-coding region within *MEIS1* locus, could reduce significantly the enhancer activity in the brain and spinal cord (Spieler et al., 2014), demonstrating the utility of zebrafish in this type of assays.

6.1.3 Type two diabetes associated polymorphisms identified by GWAS

T2D is a complex disease caused by both genetic and environmental factors. The genetic basis of T2D has been widely studied through linkage analysis and association methods, reviewed in (Pal and McCarthy, 2013). GWAS have made progress towards understanding the inherited base of T2D by detecting more than 60 disease-associated DNA variants in European and Asian populations (Morris et al., 2012). Interestingly, out of the 18 most strongly T2D-associated SNPs, only 3 are mis-sense variants, while the rest are non-coding (Prokopenko 2008).

One of the most interesting non-coding variants lies in *TCF7L2*, where the SNP rs7903146 has the strongest link effect to T2D demonstrated so far, in most, but not all ethnicities (Grant et al., 2006; Cauchi et al., 2007; Guo et al., 2007; Helgason et al., 2007; Dupuis et al., 2010; Morris et al., 2012). *TCF7L2* encodes a high-mobility group TF involved in Wnt signalling pathway. Gaulton and colleagues demonstrated that the risk allele of the SNP s7903146 (C>T) showed greater enhancer activity *in vitro* than the common allelic variant (Gaulton et al., 2010). Nevertheless, only a subtle effect could be detected, and the question remained of whether it could lead to an alteration in the spatio-temporal control of gene expression *in vivo*.

6.1.4 Aims

In this project, I asked whether quantitative differences in expression driven by candidate enhancer variants associated with T2D could be detected using zebrafish. Given that most studies to date had used luciferase assay in cell lines to show functionality (Gaulton et al., 2010; Stitzel et al., 2010), we argued that differential enhancer activity had to be verified in the context of a vertebrate organism that shared a common pancreas-specific transcriptional regulatory network and anatomical structure with humans, such as zebrafish. The following objectives were set:

- To test *in vivo* the functionality candidate enhancers associated with T2D using Tol2-based zebrafish transgenesis assays.
- To develop a proof of principle method able to quantitate *in vivo* allelic differences from enhancer variants associated with T2D.

6.2 METHODS

Selected human putative enhancers with T2D-associated SNPs were amplified from human genomic DNA provided by our collaborators using the primers listed in **Table 6.1**. Due to the unavailability of human genomic DNA carrying the infrequent allelic variant of T2D-associated SNP rs3242786 on *PROX1* intronic region, a point mutation was generated following Higuchi method (see **Chapter 2** for primers and a detailed protocol). In order to control zebrafish promoter functionality, a non-conserved region from *Fugu rubripes* showing no regulatory activity (fr2(assembly 2004) chrUn:54,537,362-54,537,937) was used (Sanges et al., 2006). Final expression vectors were generated using pSP1.72BSSPE-R3-R5-R1-R2-Venus Destination Vector (Roure et al., 2007). These vectors contained Tol2 transposase arms to facilitate single-copy genome integration of the construct into the host genome by co-injection with Tol2 transposase mRNA.

Table 6.1 Primers used to amplify human putative enhancers and zebrafish promoters from genomic DNA

CRE	Genomic coordinates (Hg18 or Zv9)	Forward primer (5'-3')	Reverse primer (5'-3')	Product size (bp)
PROX1	chr1:212242977-212243697	GCAAAAATGAACTTGAGAAATCC	CATTCCCTTTAATATCCCATGC	721
TCF7L2	chr10:114748248-114748506	AATTCATGGGCTTTCTCTGC	GTGAAGTGCCCAAGCTTCTC	239
DGKB-TMEM195	chr7:15030251-15031281	AGTCTAATACCTCTCAGTGGATA	TGGTTGATTGACAGAATTCATT	1031
ADCY5	chr3:124547851-124549052	GATTCAGCCAGGGGCAGCCTT	GAAGCAACACCAGCCGCTTTG	1202
Zebrafish *prox1* promoter	chr17:33071557-33071979	TCCGCACAGAGAACGTATTG	TGAGCTTCTTCGCGATAGTG	423
Zebrafish *tcf7l2* promoter	chr12:32790190-32790830	TCAGCCTCTTCTGTTTTGAGCAG	TTTAAGTTTAGGGACTCGCAGTGG	641
Zebrafish *hsp70* promoter	chr3:26911324-26911472	TTGATTGGTCGAACATGCTGG	CAGTCCGCTCGCTGTCTCGCT	149

6.3 RESULTS

6.3.1 Selection of putative enhancers associated with T2D

To ask whether T2D-associated variants were active in zebrafish, we initially selected four T2D-associated SNPs uncovered by a global meta-analysis (**Table 6.2,** (Dupuis et al., 2010). The SNPs were located in or near *ADCY5*, *PROX1*, *TCF7L2* and *DGKB-TMEM195* loci and were significantly associated with elevated fasting glucose in non-diabetic individuals, which increased the risk to T2D (Dupuis et al., 2010). Next, candidate enhancer sequences harbouring these T2D-associated SNPs were identified by Lorenzo Pasquali. Candidate enhancers for *in vivo* assays were defined by the presence of open chromatin sites and/or enhancer-associated marks, including H3K4me1 or H3K27ac (Pasquali et al., 2014), and were approximately 1 kb in length (**Table 6.1**).

Table 6.2 List of T2D SNPs selected for *in vivo* assays.

T2D-associated SNP	T2D-associated locus	Location	Allelic variants	GWAS Reference
rs4282786	*PROX1*	Intronic	G>A	(Dupuis et al., 2010)
rs7903146	*TCF7L2*	Intronic	C>T	(Cauchi et al., 2007)
rs2191349	*DGKB-TMEM195*	Intergenic	G>T	(Dupuis et al., 2010)
rs11708067	*ADCY5*	Intronic	A>G	(Dupuis et al., 2010)

6.3.2 Identification of zebrafish endogenous promoter using a CAGE dataset

For reporter assays, we decided to use zebrafish endogenous promoters, as well as the heterologous *hsp70* promoter. In order to identify the endogenous promoters of *prox1* and *tcf7l2* genes in zebrafish, a CAGE dataset was utilized, which predicts promoters at a single nucleotide resolution in several embryonic developmental stages (Nepal et al., 2013).

The transcription start site (TSS) of genes was determined by CAGE tag cluster distribution (depicted as blue peaks in **Figure 6.1**), where the peak represents the 5' end position of aligned CAGE reads and its height is indicative of the enrichment of the aligned segments.

A region of 423 bp in length, surrounding the TSS of *prox1* gene was identified as the minimal promoter and was amplified for cloning (**Figure 6.1A**) Approximately 50 bp downstream and 400 bp upstream of the dominant CAGE-predicted TSS were used.

Zebrafish *tcf7l2* promoter was identified as a broad promoter due to the presence of two enrichment peaks (**Figure 6.1B**). The CAGE peak overlapping with the reference TSS was chosen as the main promoter as it was consistently supported by the highest number of tags at the zygotic stages of development. A region of 641 bp around the main TSS was used for cloning (**Table 6.1**).

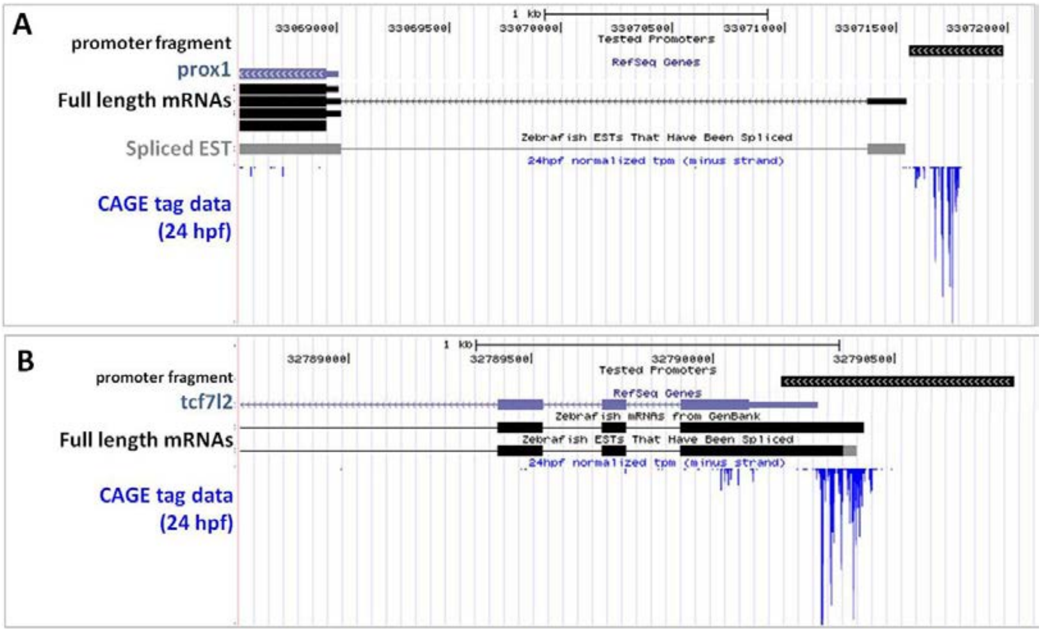


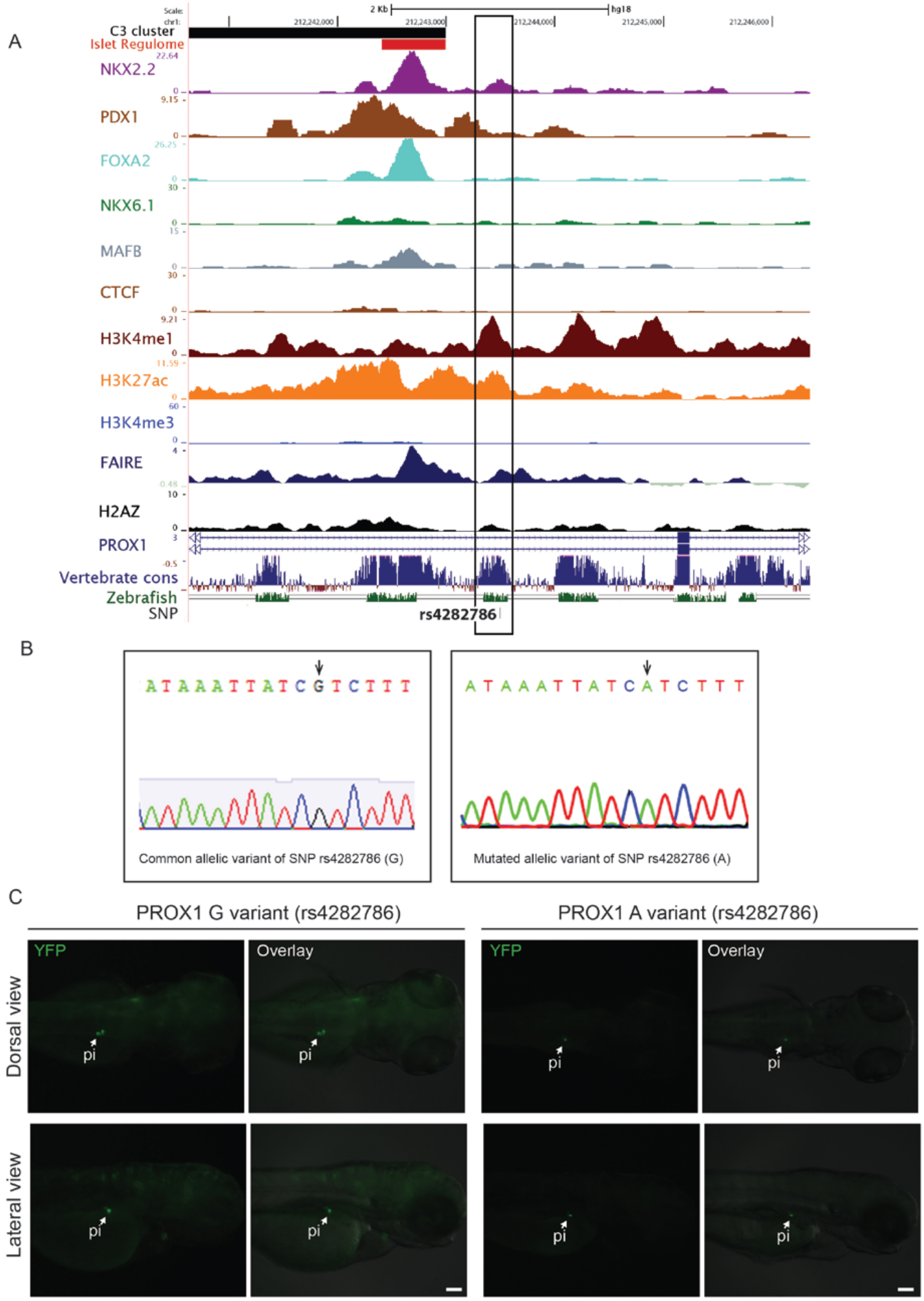
Figure 6.1 Identification of zebrafish *prox1* and *tcf7l2* endogenous promoters.
UCSC Genome Browser screenshot displaying custom tracks of the fragments cloned, RefSeq genes, full length mRNAs, spliced EST and CAGE tag data for 24 hpf and 33 hpf developmental stages. The solid black box with arrowheads, which point the direction of transcription, in the "promoter fragment" track represents the region cloned for reporter assays. **A.** Identification of *prox1* minimal promoter (423 bp around the main CAGE TSS). **B.** Identification of *tcf7l2* minimal promoter (641 bp) around the main CAGE TSS.

6.3.3 *In vivo* testing of T2D-associated candidate enhancers by Tol2-based transient transgenic assays in zebrafish

To determine whether predicted regulatory elements could function as enhancers, and whether disease-associated SNPs could result in differential enhancer activities in a vertebrate embryo model, Tol2-based transient transgenesis assays were carried out. For each T2D-associated locus three constructs were produced by linking the human enhancer variants to a zebrafish promoter and a reporter gene. In order to control zebrafish promoter functionality, a non-conserved region from *Fugu rubripes* showing no regulatory activity was used (as described in **Chapter 3**, see also **Figure 3.3**). These three expression vectors were injected in parallel into WT zebrafish fertilized eggs and reporter expression was assessed at protruding mouth stage (72 hpf) using fluorescence microscopy. All tissues of expression from three experimental replicates were annotated following the criteria set in previous chapters and representative patterns of expression were documented. The time point chosen was 72 hpf because the zebrafish pancreas has developed completely at this stage (Tiso et al., 2009).

6.3.3.1 Validation of *PROX1* putative human enhancer containing SNP rs4282786

GWAS revealed that SNP rs340874, which was located 2kb upstream of *PROX1* TSS, was significantly associated to T2D (Dupuis et al., 2010). However, rs340874 was neither conserved in zebrafish nor enriched in chromatin marks typically associated with enhancers. In contrast, intronic SNP rs4282786 (G>A) was in LD with T2D-associated rs340874, and was located in an evolutionary conserved region of *PROX1* gene. Additionally, it was enriched in H3K4me1 and H3K27ac histone modification marks, suggesting that it could be a developmental enhancer (**Figure 6.2A**). Thus, we selected a 721 bp region harbouring SNP rs4282786 for functional analysis in zebrafish. Due to the unavailability of human genomic DNA carrying the infrequent variant in homozygosis, I used site directed mutagenesis by PCR to generate the infrequent A allele carried by the candidate enhancer associated to T2D (**Figure 6.2B**).



(See next page for figure legend)

Figure 6.2 Chromatin landscape and activity of putative human regulatory element *PROX1*.
A. UCSC Genome Browser screenshot displaying custom tracks of FAIRE-Seq data, histone modifications and DNA sequence conservation. *PROX1* putative element is depicted by a black solid line box. Screenshot is courtesy of Lorenzo Pasquali. **B.** Sequencing chromatograms of rs4282786 SNP within *PROX1* putative element. Left panel shows the wild type sequence containing the common variant (G). Right panel shows the site-directed introduced mutation to mimic the infrequent allelic variant of the SNP (A). **C**. Transient expression pattern driven by *PROX1* candidate enhancer. Representative dorsal and lateral view images of embryos injected with G (top) and A (bottom) variants of SNP rs4282786 are displayed. All embryos are 72 hpf oriented with anterior to the right. Scale bar indicates 100 µm.

In order to test the functionality of the conserved *PROX1* putative regulatory element in zebrafish, this region was linked to the zebrafish *prox1* endogenous minimal promoter (**Figure 6.1**) and Venus reporter gene and domains of expression were annotated at 3 dpf (**Table 6.3**, **Figure 6.3**). Constructs containing *PROX1* putative enhancer variants could drive reporter gene expression in the pancreatic islet, yolk syncytial layer (ysl), muscle, heart and neurons (**Figure 6.2C**, **Figure 6.3**). However, no statistically significant differences could be detected between the A and G alleles in terms of frequency of expression domains (**Table 6.3**). Interestingly, none of the embryos injected with the *Fugu* control region linked to *prox1* zebrafish promoter led to detectable expression in the pancreas (n=246), suggesting that the human *PROX1* candidate enhancer was responsible for this activity. Constructs containing *PROX1* candidate also showed a significant enrichment in tissues such as heart and neurons, when compared to the *Fugu* control, which suggests that these domains are attributable to the candidate enhancer (**Table 6.3**).

In order to find out whether *PROX1* was autonomously capable of driving expression in the pancreas, and to determine whether heart and neuronal expression patterns were an artefact from the construct used, both enhancer variants and the *Fugu* control region were cloned into a different context: upstream of a heterologous minimal promoter (*hsp70*) in a pDB896 vector containing short Tol2 arms. Three experimental replicates confirmed that both G and A allelic variants could drive pancreas specific expression in 2.7% (n=298) and 1.85% (n=334) of the injected embryos respectively, whereas reporter gene expression was not detectable in this domain with the control vector (n=215). These results suggested that *PROX1* human enhancer, and not the zebrafish prox1 promoter, was responsible for the activity in the endocrine pancreas. Ectopic lens expression driven by *hsp70* zebrafish promoter (Blechinger et al., 2002) was detected in embryos injected with *PROX1* putative enhancer variants, although surprisingly it was not shown by the Fugu region linked to the promoter only. On average, the frequency of heart expression was reduced 12% in embryos injected with *PROX1* variants (**Figure 6.3B**), suggesting that either a cryptic vector sequence from the Gateway vector, the Tol2 arms or a specific interaction between *PROX1* candidate enhancer and *prox1* promoter were responsible for the enhanced expression in this domain.

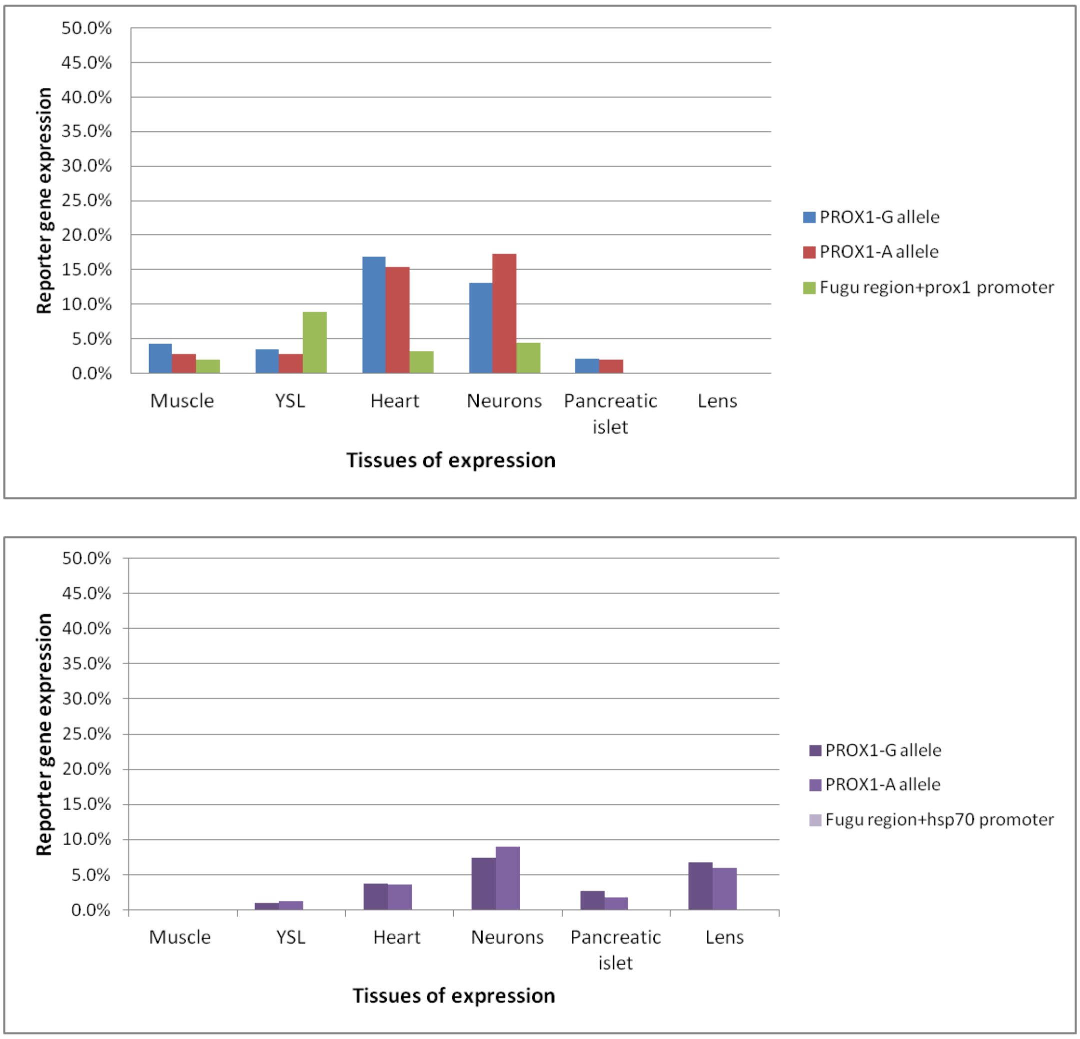


Figure 6.3 Rates of tissue expression driven by *PROX1* human element constructs linked to zebrafish *prox1* and *hsp70* zebrafish promoters.

Percentage of 72 hpf zebrafish injected embryos showing reporter gene expression driven by *PROX1* candidate enhancer variants and the *Fugu* control region linked to prox1 promoter (top) and *hsp70* promoter (bottom). Reporter expression could be detected in muscle, yolk syncytial layer (ysl), heart, neurons and endocrine islet when *PROX1* candidate enhancer was linked to prox1 promoter. When *PROX1* is linked to *hsp70* promoter, the candidate enhancer is autonomously capable of driving reporter gene expression in the ysl, heart, neurons and pancreatic. Expression in the lens is ectopic activity from *hsp70* core promoter (Blechinger et al., 2002). For detailed expression frequencies refer to **Table 6.3.**

Table 6.3 Quantitation of reporter gene expression driven by *PROX1* putative enhancer in protruding-mouth embryos.

Tested CRE	Zebrafish promoter used	No. of injected embryos	Reporter gene expression (%)						Percentage of expressing embryos	No. of replicates
			Muscle	YSL	Heart	Neurons	Pancreatic islet	Lens		
PROX1-G allele	*prox1*	284	4.2	3.5	16.9	13.0	2.1	0.0	35.9	3
PROX1-A allele	*prox1*	352	2.8	2.8	15.3	17.3	2.0	0.0	50.9	3
Fugu control region	*prox1*	246	2.0	8.9	3.3	4.5	0.0	0.0	44.7	3
PROX1-G allele	*hsp70*	298	0.0	1.0	3.7	7.4	2.7	6.7	19.8	3
PROX1-A allele	*hsp70*	334	0.0	1.2	3.6	9.0	1.8	6.0	28.7	3
Fugu control region	*hsp70*	215	0.0	0.0	0.0	0.0	0.0	0.0	0.9	3

Table legend: Transgene-driven reporter expression at 72 hpf is depicted as a ratio between the number of embryos showing tissue-specific activity versus the total number of embryos injected with the enhancer-containing construct or *Fugu* region and promoter-containing control vector. Merged numbers from 3 independent injection experiments are shown. YSL= yolk syncytial layer.

6.3.3.2 Verification of pancreatic islet specific activity of *PROX1* candidate enhancer by co-localization analysis with Tg(ins-mCherry)^{jh2} line

In order to confirm the pancreatic islet-specific activity attributed to *PROX1* candidate enhancer as well as to attempt to quantitate differences between *PROX1* allelic variants, I carried out an additional reporter analysis thereby I compared *PROX1* driven Venus reporter activity to embryos from the Tg(ins:mCherry)^{jh2} line, which accurately label insulin-producing β-cells in red (Pisharath et al., 2007). To this end, vectors containing *PROX1* allelic variants linked to a *prox1* endogenous promoter (namely pTol2/Hs.PROX1(G)-Dr.prox1:Venus and pTol2/Hs.PROX1(A)-Dr.prox1:Venus) were injected into heterozygous embryos of the Tg(ins:mCherry)^{jh2} line, and co-localization of the two fluorescence reporters was evaluated at 3 dpf (**Table 6.4**, **Figure 6.4**).

Table 6.4 Summary of co-localization events with *PROX1* set of constructs

Construct	No. of injected embryos	No. of expressing embryos	No. of embryos with co-localization	Co-localization frequency (%)
pTol2/Hs.PROX1(G)-Dr.prox1:Venus	230	82	15	18,3
pTol2/Hs.PROX1(A)-Dr.prox1:Venus	196	68	10	14,7
pTol2/Fugu:Dr.prox1:Venus	205	72	0	0,0

Co-localization between *PROX1* enhancer-driven reporter expression and insulin-driven reporter in the endocrine pancreas was found in 18.3% of embryos (n=230) injected with the common rs4282786 allele (G) and in 14.7% embryos (n=196) injected with the infrequent variant (A). None of the embryos injected with the *Fugu* control region linked to *prox1* promoter (pTol2/Fugu:Dr.prox1:Venus vector) showed co-localization events (**Table 6.4**). However, there was no statistically significant difference in the frequency of co-localization events between the two alleles (P=0.6617), suggesting that rs4282786 SNP was not essential for the regulation of the pancreas-specific activity.

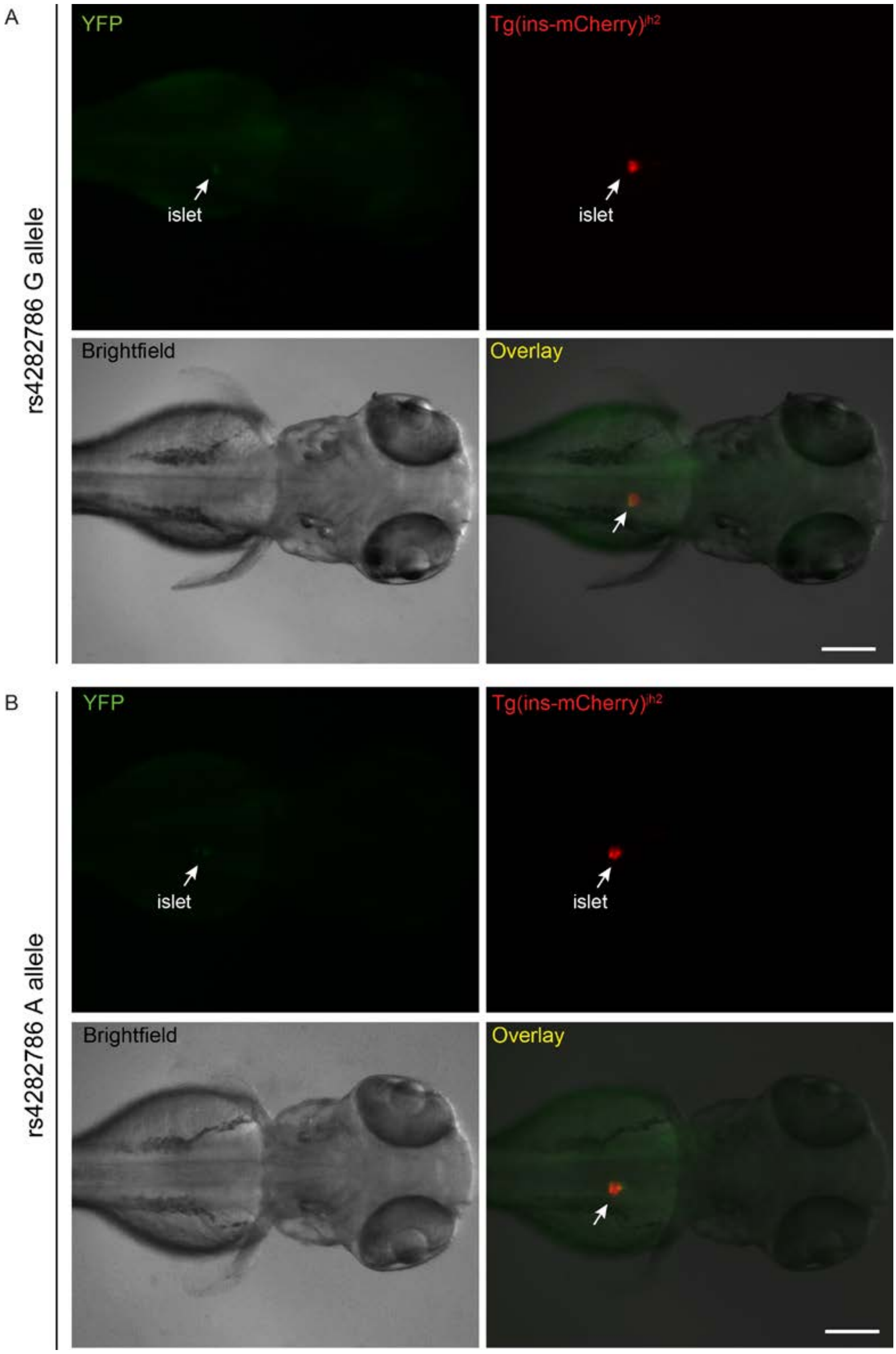


Figure 6.4 Co-localization events between *PROX1* candidate enhancer (green) and pancreatic islet specific transgenic line (red).

Reporter expression (YFP), insulin transgenic expression (mCherry), brightfield image and an overlay of channels are shown for representative embryos injected with *PROX1*-containing variants linked to *prox1* zebrafish promoter. All zebrafish embryos are 72 hpf oriented with anterior to the right. Scale bar indicates 250 µm.

6.3.3.3 Validation of *TCF7L2* putative human enhancer containing SNP rs7903146

In order to test whether the 239 bp open chromatin region evaluated by Gaulton and colleagues was active in zebrafish, and whether the allelic variants of SNP rs7903146 (C>T) could lead to a differential enhancer activity, *TCF7L2* candidate enhancer was cloned upstream of the zebrafish *tcf7l2* endogenous minimal promoter (**Figure 6.1**) and Venus reporter gene, and its activity analysed in transgenesis assays. This region was not conserved at a sequence level with zebrafish and was enriched in H3K4me1, H3K27ac and H2AZ histone variant (**Figure 6.5A**).

Fluorescent microscopy analysis of reporter expression during the first five days of embryo development revealed that *TCF7L2* enhancer variants could not drive any tissue specific activity that differed from the *Fugu* control vector in three experimental batches of independently injected embryos (n=878). Because the lack of activity could have been caused by the limited sensitivity of the stereo-fluorescence microscope used, we decided to increase the transgene detection sensitivity. Thus, we carried out three additional experiments where 72 hpf injected embryos were plated and imaged by the ScanR high-throughput imaging station (Liebel et al., 2003). Twenty images from each embryo group were overlaid using the Adobe Photoshop CS6 software to create an overview of mosaic expression patterns (**Figure 6.5B**). A diversity of ectopic expression was found (mainly in the yolk syncytial layer) but neither pancreas specific expression nor diencephalic expression (which is relevant to the expression of the zebrafish *tcf7l2* gene) was detected, suggesting that there was no specific reporter activity driven by *TCF7L2* human putative enhancer (**Figure 6.5B**).

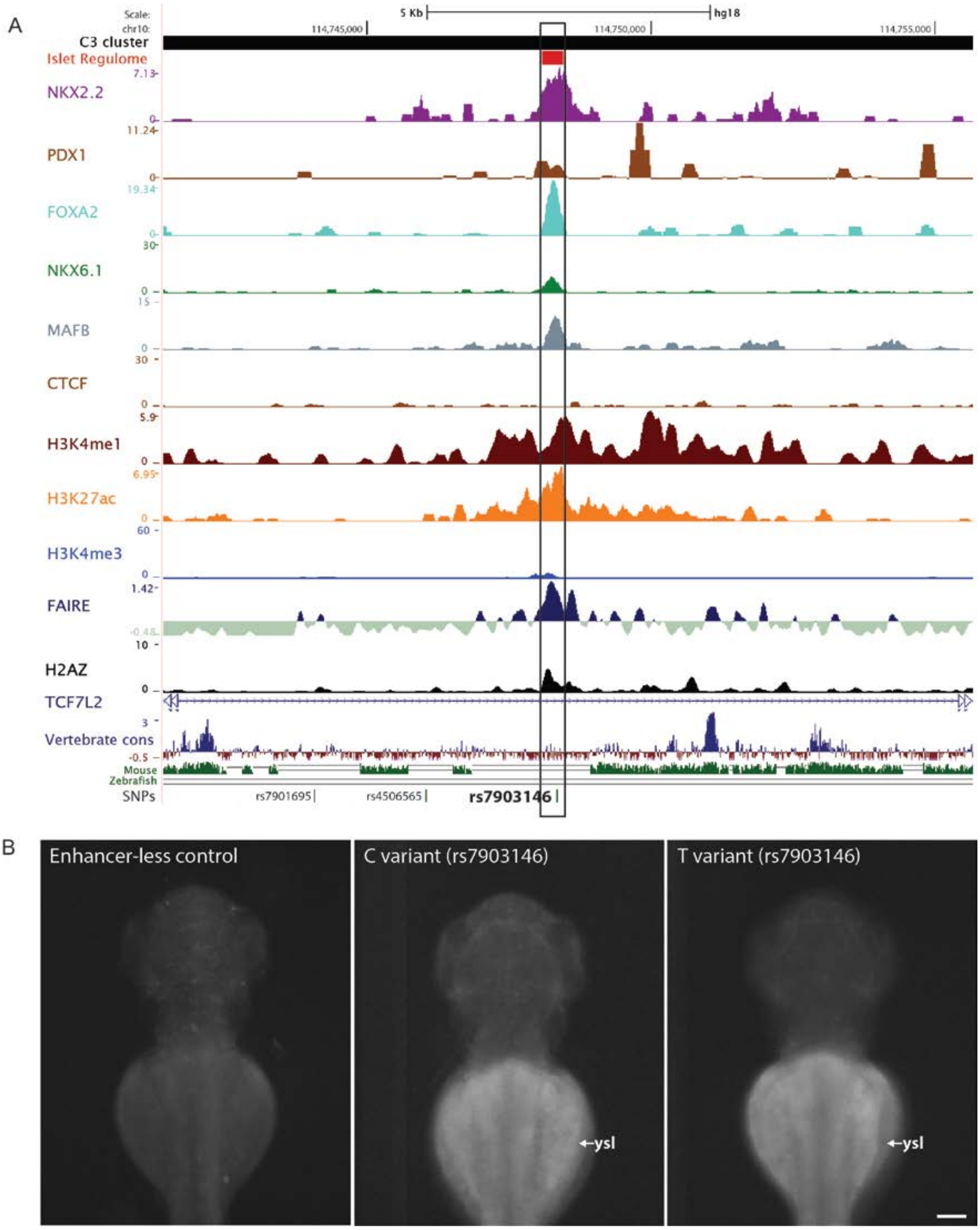


Figure 6.5 Chromatin landscape and activity of putative human regulatory element *TCF7L2*.
A. UCSC Genome Browser screenshot, courtesy of Lorenzo Pasquali, displaying custom tracks of FAIRE-Seq data, histone modifications, RNA Seq and DNA sequence conservation. The solid line box contains the cloned element. **B.** Composite overviews of *in vivo* Venus expression data from 20 individual zebrafish embryos. All embryos are 72 hpf oriented dorsally. Yolk syncytial layer (ysl) ectopic expression was detected in embryos injected with *TCF7L2* human element. Scale bar indicates 100 µm.

6.3.3.4 Validation of *DGKB-TMEM195* putative human enhancer and SNP rs2191349

SNP rs2191349 (G>T), which is located on a gene desert between *DGKB* and *TMEM195*, was also significantly associated to T2D by GWAS (Dupuis et al., 2010). *DGKB* encodes a catalytic domain of diacylglycerol kinase and *TMEM195* encodes a transmembrane protein expressed in liver. A region of 1,031 bp surrounding the lead SNP rs2191349 that overlapped with an open chromatin site and was enriched in H3K4me1 and H3K27ac marks, was used for *in vivo* tests in zebrafish (**Figure 6.6A**).

In order to test the functionality of this SNP, both allelic variants were cloned upstream of the zebrafish *hsp70* minimal promoter and injected in zebrafish WT embryos. Pancreas specific activity was not detected in embryos injected with any of the constructs (total number of injected embryos n=979, three experimental repeats). Only lens activity was detected in 71.7% (n=211) of embryos injected with the common allele variant and by 63.8% (n=232) of embryos injected with the T2D-associated variant (**Figure 6.6B**). Zebrafish *hsp70* minimal promoter can drive autonomously expression in the lens (Blechinger et al., 2002), suggesting that the lens reporter activity seen in embryos injected with constructs containing the enhancer variants was due to the *hsp70* promoter itself and that the candidate enhancer could not drive any tissue-specific activity in injected embryos.

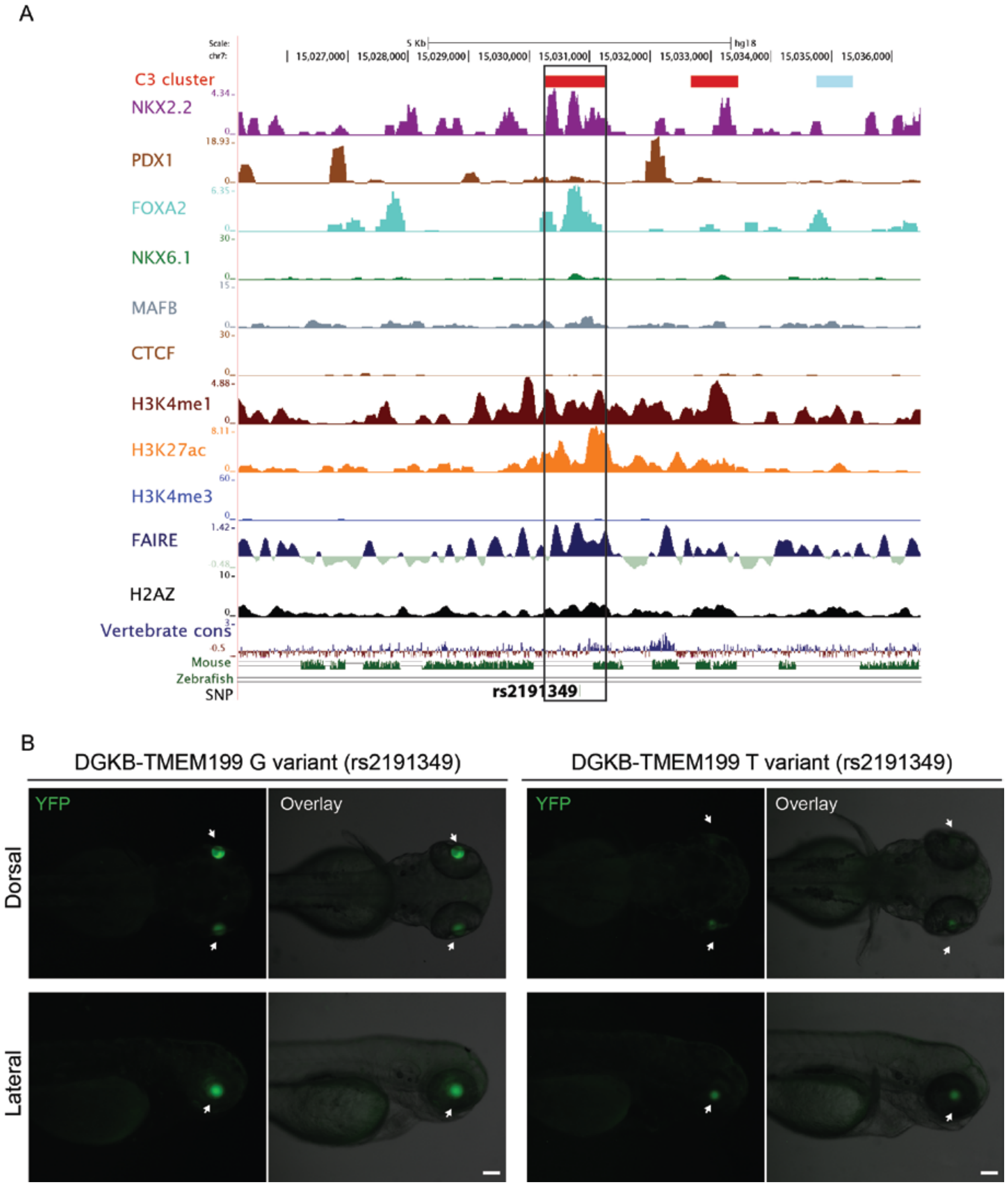


Figure 6.6 Chromatin landscape and activity of putative human regulatory element *DGKB-TMEM195*.

A. UCSC Genome Browser screenshot (courtesy of Lorenzo Pasquali) displaying custom tracks of FAIRE-Seq data, histone modifications, RNA Seq and DNA sequence conservation. The solid line box contains the cloned element. **B.** Transient expression pattern driven by *DGKB-DMEM195* candidate enhancer. Representative dorsal and lateral view images of 72 hpf embryos injected with G (top) and T (bottom) variants of SNP rs2191349 are shown. Scale bar indicates 100 µm.

6.3.3.5 Validation of *ADCY5* putative human enhancer and SNP rs11708067

T2D-associated SNP rs11708067 (A>G) is located on an intronic region of *ADCY5*, which encodes Adenylate Cyclase-5 (Dupuis et al., 2010). This enzyme catalyses the generation of cAMP, a signalling molecule in beta cells of the pancreatic islet. An open chromatin site enriched in H3K4me1 and H3K27ac marks of 1,200 bp in length, was identified as a putative regulatory enhancer and used in reporter assays (**Figure 6.7A**). This region was cloned upstream of the *hsp70* minimal promoter and tested, as described before. No tissue-specific expression pattern could be detected with either the common or the disease-associated variant across three experimental replicates (total number of injected embryos n=379 and n=408, respectively, **Figure 6.7B**), suggesting that human *ADCY5* putative element was not able to direct tissue specific activity in zebrafish.

6.3.4 Establishing Tol2-based stable transgenic lines in order to detect differences

To ask whether enhancer-specific reporter function could have been missed by the potentially low resolution and sensitivity of transient transgenic assays, around 200 embryos were injected with each construct to generate stable transgenic lines. The injected embryos were grown up in the animal facility to study transmission of transgenes to the next generations. We reasoned that by avoiding the mosaicism typical of transient transgenic embryos, the analysis of the stable transgenic lines could potentially reveal more faithfully the expression pattern driven by the candidate enhancers. Adult founders were outcrossed with WT fish and the F1 progeny were screened for potential enhancer-driven Venus expression. For each founder at least 100 or more offspring embryos were analysed before deemed to be negative for transgene activity.

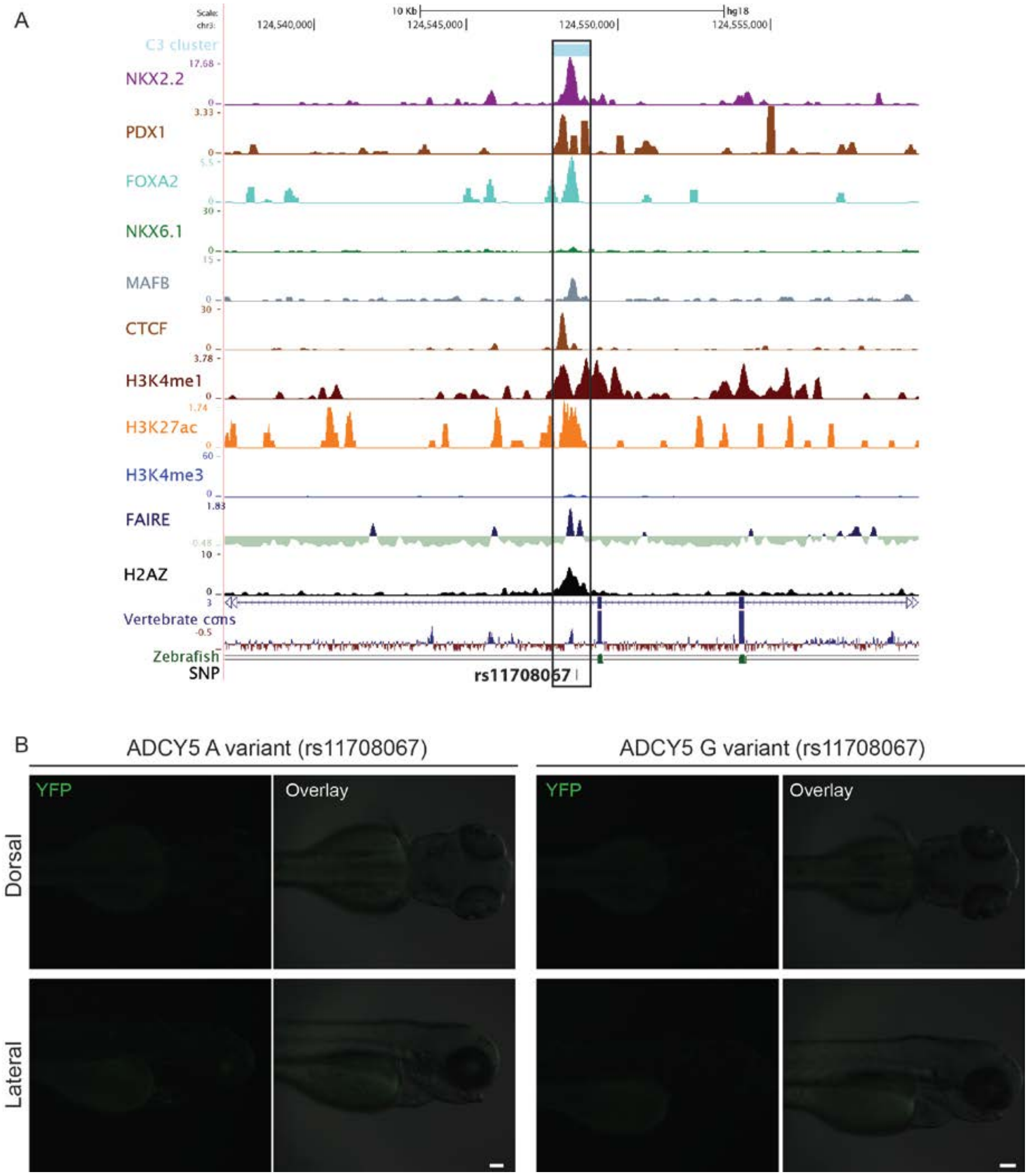


Figure 6.7 Chromatin landscape and activity of putative human regulatory element *ADCY5*.
A. UCSC Genome Browser screenshot displaying custom tracks of FAIRE-Seq data, histone modifications, RNA Seq and DNA sequence conservation. The solid line box contains the cloned candidate enhancer. This screenshot is courtesy of Lorenzo Pasquali. **B.** Transient expression pattern driven by ADCY5 human candidate enhancer. Representative embryo injected with the common variant of SNP rs11708067 (left) and with the infrequent variant (right). Representative dorsal and lateral view images of 72 hpf embryos are shown. Scale bar indicates 100 µm.

6.3.4.1 Analysis of *PROX1* candidate enhancer-containing stable transgenic lines

Four stable transgenic lines with *PROX1* rs4282786 common G allele (named Tg(Hs.PROX1(G):-0.4Dr.prox1:Venus G1-G7) and 3 transgenic lines with the disease-associated A allele (named Tg(Hs.PROX1 (A):-0.4Dr.prox1:Venus A1-A12) were identified by reporter gene expression (**Table 6.5**). Furthermore, 3 lines were generated with the *Fugu* control region linked to the *prox1* promoter (named Tg(Fr.ek:-0.4Dr.prox1:Venus) L1-5), which served as a control for the background activity expected without the enhancer (**Figure 6.8, Figure 6.9, Figure 6.10**). In order to dissect which domains of reporter activity were driven by the candidate human enhancer and distinguish them from position effects, we identified all the expression domains in enhancer-containing transgenic lines and evaluated their reproducibility in comparison to the promoter only control transgenic lines, and to publicly available WISH data of the zebrafish orthologous gene patterns (**Table 6.6**). From 17 patterns identified among all transgenic lines, trigeminal and lateral line ganglia (#4 and #5) and spinal cord neurons (#6) expression domains were shared between the enhancer-less control lines and the *PROX1*-containing lines, and therefore could be attributed to *prox1* promoter function. The pancreatic islet (#13), which is an expression domain of *prox1* gene, was only present in one of the lines containing the human enhancer, but not in the promoter-only lines. Although the presence of a single line being active in the pancreas might not allow us to draw a significant correlation with the candidate enhancer, together with the transient transgenic data, it suggested that it could be driven by the *PROX1* putative enhancer. Reporter expression in midbrain neurons (#15) and heart (#16) domains were only present in *PROX1* containing lines, which suggested that they were driven by the human enhancer. This is supported by their presence in transient transgenesis. Nevertheless, there were other ectopic expression domains that could not be attributed to the enhancer unambiguously. This variation of patterns is commonly seen in Tol2-based transgenesis, a system prone to position effects due to the random integration of the transgene into the host genome.

Table 6.5 Summary of transgenic lines established using *PROX1* human element and T2D-associated SNP rs4282786

Transgenic Line	No of screened founders	No of positive founders	Germline transmission rate (%)	Line ID	Domains of expression	No. of embryos analysed	Transmission rate per pattern (%)
Tg(Hs.PROX1 (G):-0.4Dr.prox1:Venus)	20	4	20.0	G1	Pan neuronal	120	12.5
				G2	**Pancreatic islet**	154	11.0
				G3	Forebrain, Midbrain, Hindbrain neurons	70	30.0
				G7	Hindbrain neurons	98	39.8
Tg(Hs.PROX1 (A):-0.4Dr.prox1:Venus)	14	3	21.4	A1	Forebrain,	90	7.8
				A4	Forebrain, tectum, midbrain-hindbrain boundary, pericardium		8.9
				A5	Tectum, hindbrain	150	13.3
				A12	Tectum, tegmentum, hindbrain, pericardium	70	14.3
Tg(Fr.ek:-0.4Dr.prox1:Venus)	15	3	20.0	L1	Central nervous system	240	23.3
				L3	Cranial ganglia	50	10.0
				L5	Retina, optic tectum	82	19.5

Table 6.6 Analysis of the domains of Venus reporter expression and their distribution among *PROX1*-containing Tol2 transgenic lines.

Domains of reporter gene expression		Zebrafish *prox1* expression pattern	Tol2-based stable transgenic lines										
			Tg(Fr.ek:-0.4Dr.prox1:Venus)			Tg(Hs.PROX1 (G):-0.4Dr.prox1:Venus)				Tg(Hs.PROX1 (A):-0.4Dr.prox1:Venus)			
			L1	L3	L5	G1	G2	G3	G7	A1	A4	A5	A12
	Habenula	🟪											
	Liver	🟪											
1	Retina	🟪			▨								
2	Diencephalon	🟪											
3	Dorsal hindbrain	🟪						🟦	🟦			🟩	🟩
4	Trigeminal ganglia	🟪		▨									
5	Lateral line ganglia	🟪		▨						🟩	🟩	🟩	
6	Spinal cord neurons	🟪	▨		▨				🟦				
7	Telencephalon		▨							🟩	🟩		
8	Ventral midbrain		▨										
9	Tegmentum		▨										🟩
10	Optic tectum				▨								🟩
11	Lens					🟦							
12	Yolk syncytial layer					🟦							
13	Pancreatic islet	🟪					🟦						
14	Forebrain neurons							🟦					
15	Midbrain neurons							🟦	🟦			🟩	
16	Heart							🟦			🟩		🟩
17	Midbrain-Hindbrain boundary										🟩		

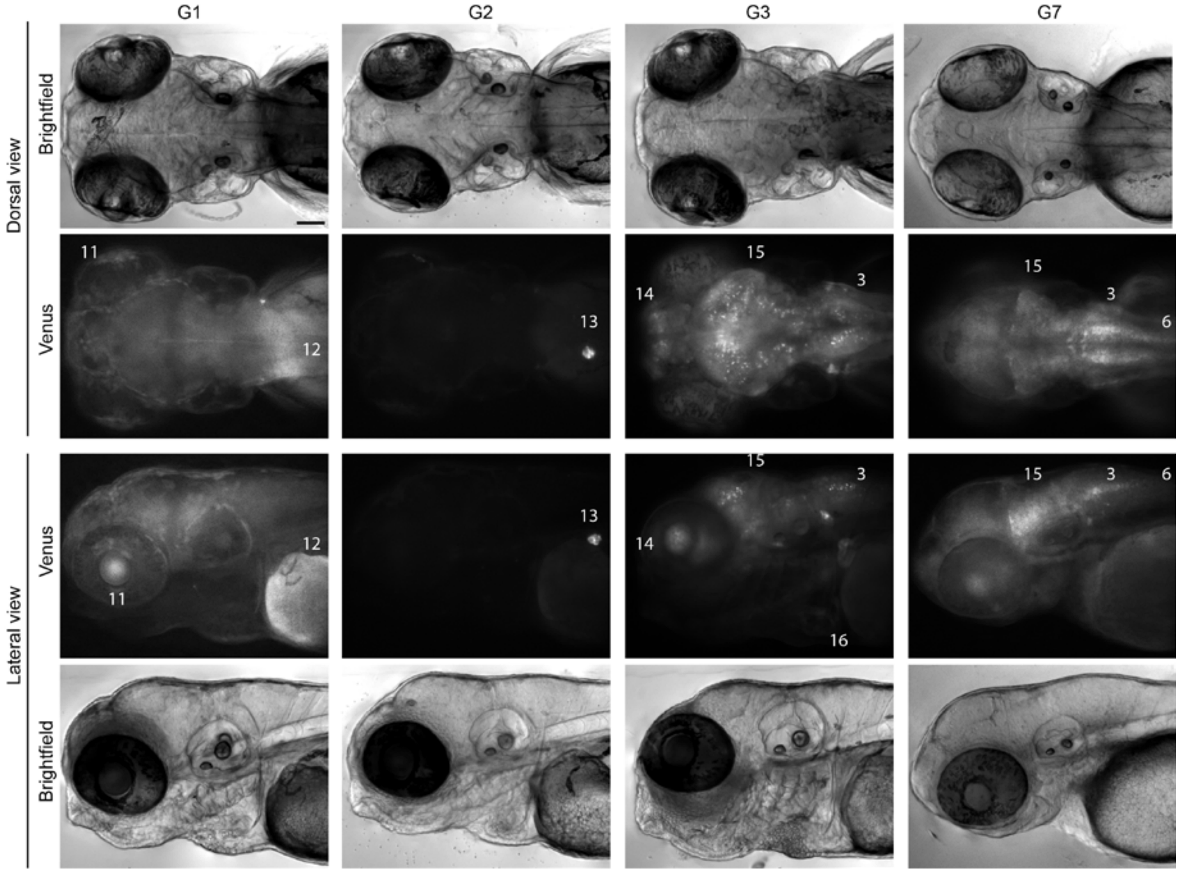


Figure 6.8 Tol2-based transgenic lines established with *PROX1* containing rs4282786 G allele
Expression patterns shown by Tol2 transgenic lines injected with Hs.PROX1(G):Dr.prox1:Venus construct. Tol2 patterns are numbered from 1 to 17 as described in **Table 6.6**. Dorsal and lateral views of YFP and brightfield channels are shown. Larvae are protruding mouth stage (approximately 3 dpf) with anterior to the left. Scale bar indicates 100 µm.

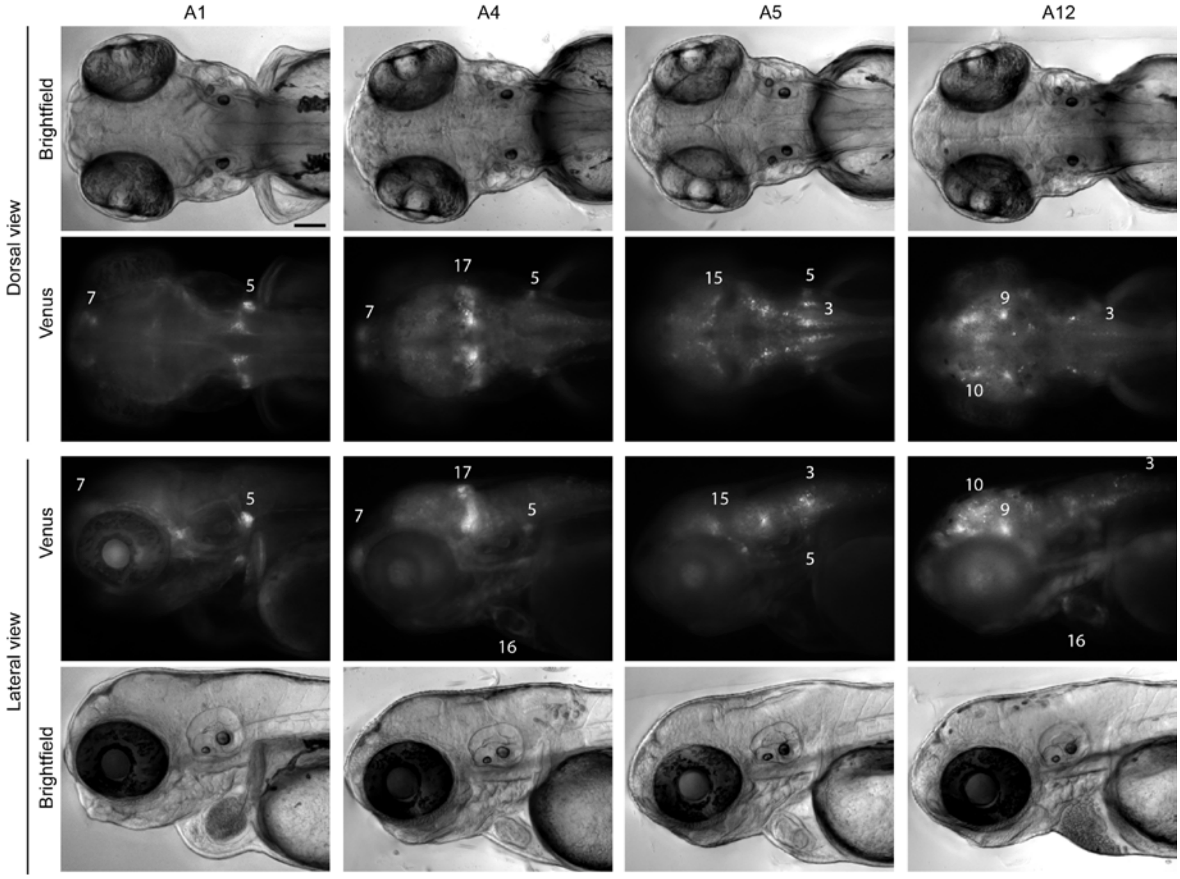


Figure 6.9 Tol2-based transgenic lines established with *PROX1*-containing rs4282786A allele
Expression patterns shown by Tol2 transgenic lines injected with Hs.PROX1(A):Dr.prox1:Venus construct. Tol2 patterns are numbered from 1 to 17 as described in **Table 6.6**. Dorsal and lateral views of YFP and brightfield channels are shown. Larvae are protruding mouth stage (approximately 3 dpf) with anterior to the left. Scale bar indicates 100 µm.

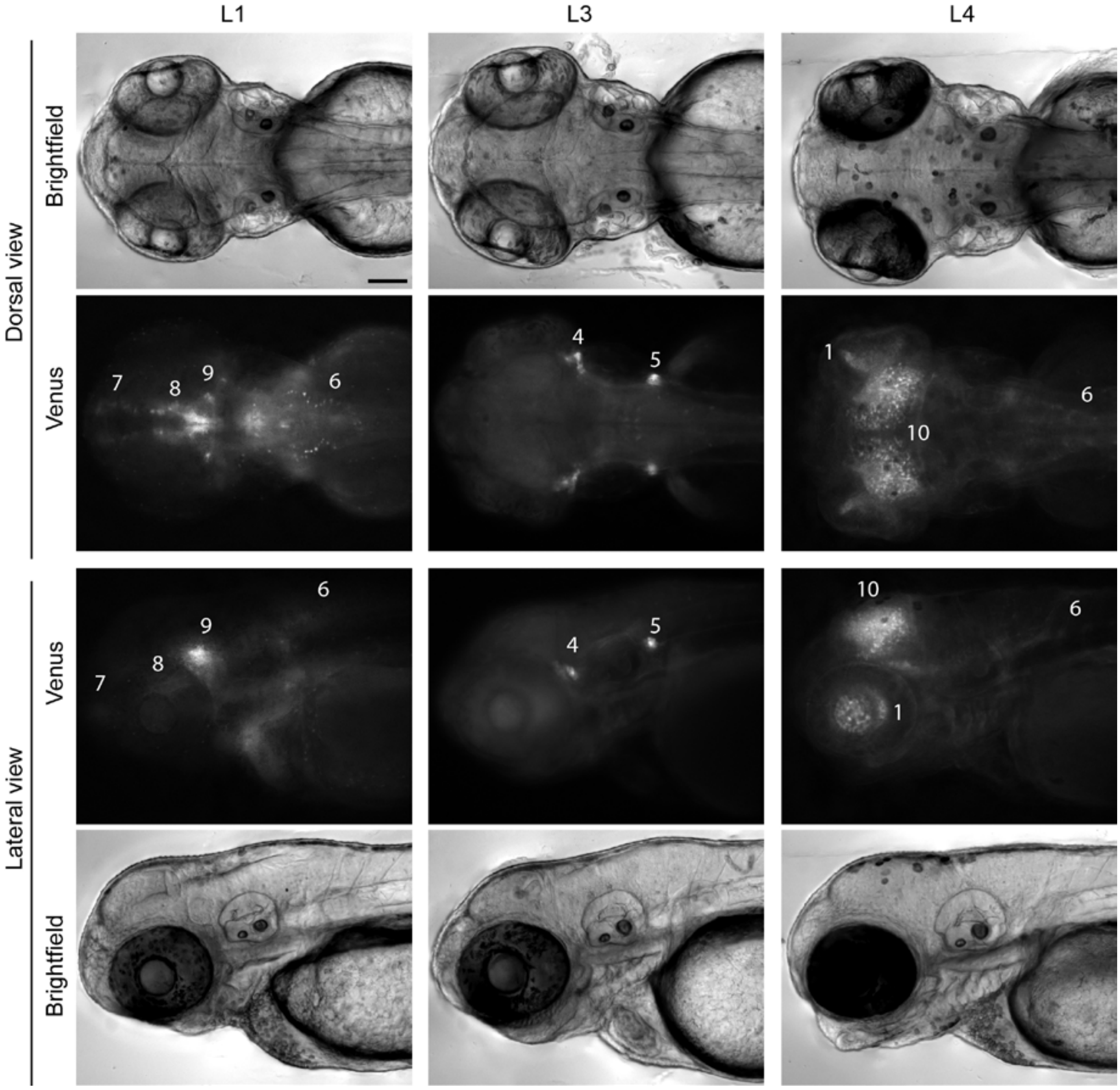


Figure 6.10 Tol2-based transgenic lines established with zebrafish *prox1* endogenous promoter Expression patterns shown by Tol2 transgenic lines injected with pTol2/Fr.ek:Dr.prox1:Venus construct. Tol2 patterns are numbered from 1 to 17 as described in **Table 6.6**. Dorsal and lateral views of YFP and brightfield channels are shown. Larvae are protruding mouth stage (approximately 3 dpf) with anterior to the left. Scale bar indicates 100 µm.

6.3.4.2 *Analysis of TCF7L2 candidate enhancer-containing transgenic lines*

In order to analyse the expression patterns attributable to *TCF7L2* human putative enhancer we undertook the same approach described for *PROX1* human enhancer candidate. One line carrying the *tcf7l2* zebrafish promote only (named Tg(Fr.ek:-0.6Dr.tcf7l2:Venus) L1) , and 6 lines containing *TCF7L2* human enhancer variants (named Tg(Hs.TCF7L2(C):-0.6Dr.tcf7l2:Venus) C1-5 and Tg(Hs.TCF7L2(T):-0.6Dr.tcf7l2:Venus) L1) were identified (**Table 6.7**, **Figure 6.11, Figure 6.12**). There was a lack of overlap in reporter expression patterns among the transgenic lines established with the enhancer-containing constructs, as well as between these lines and the zebrafish *tcf7l2* gene endogenous pattern, suggesting strong position effect variability, and the lack of identifiable tissue-specificity attributed to *TC7FL2* candidate enhancer (**Table 6.8**). Because a reduced number of cells specify endocrine pancreas fate, the presence of islet activity in C4 transgenic line (**Table 6.7**) suggested enhancer-driven activity. However, lack of reproducibility in other enhancer-containing lines prevented us from making accurate conclusions on this.

Table 6.7 Summary of transgenic lines established using *TCF7L2* human element and T2D-associated SNP rs7903146.

Transgenic Line	No. of screened founders	No. of positive founders	Germline transmission rate (%)	Line ID	Domains of expression	N of embryos analysed	Transmission rate per pattern (%)
Tg(Hs.TCF7L2(C):-0.6Dr.tcf7l2:Venus)	10	2	20.0	C1	Diencephalon	97	10.3
				C2	Diencephalon + hindbrain		7.2
				C3	Neural tube		15.5
				C4	**Pancreatic islet**, trigeminal ganglia	155	43.9
				C5	Diencephalon		13.5
Tg(Hs.TCF7L2(T):-0.6Dr.tcf7l2:Venus)	15	1	6.7	T1	Midbrain-Hindbrain boundary, pericardium, cranial ganglia	96	8.3
Tg(Fr.ek:-0.6Dr.tcf7l2:Venus)	15	1	6.7	L1	Hindbrain, jaw	100	10.0

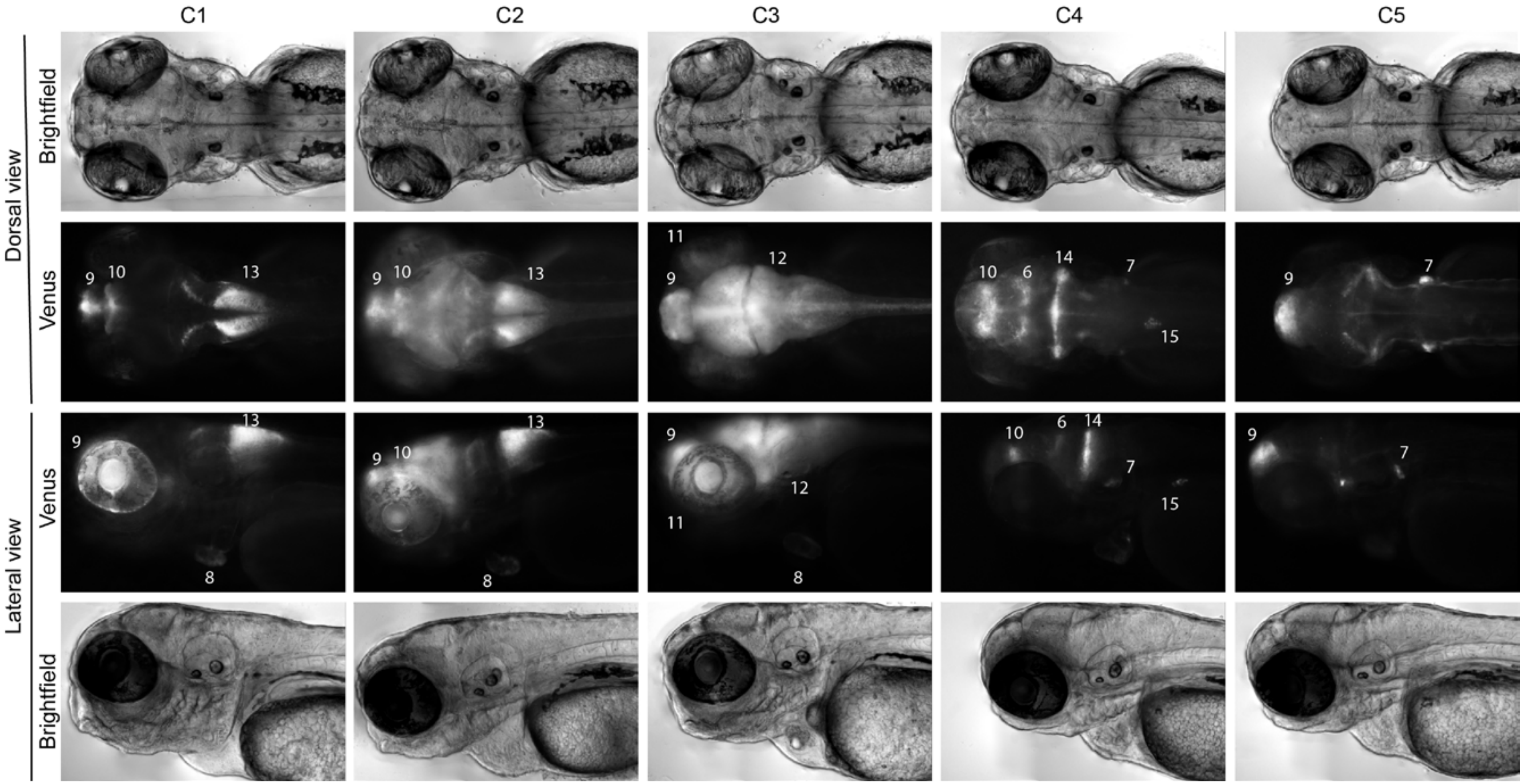


Figure 6.11 Tol2-based transgenic lines established with *TCF7L2* common C allele

Expression patterns shown by Tol2 transgenic lines injected with pTol2/Hs.TCF7L2(C):Dr.tcf7l2:Venus construct. Expression patterns are numbered from 1 to 15 as described in **Table 6.8**.

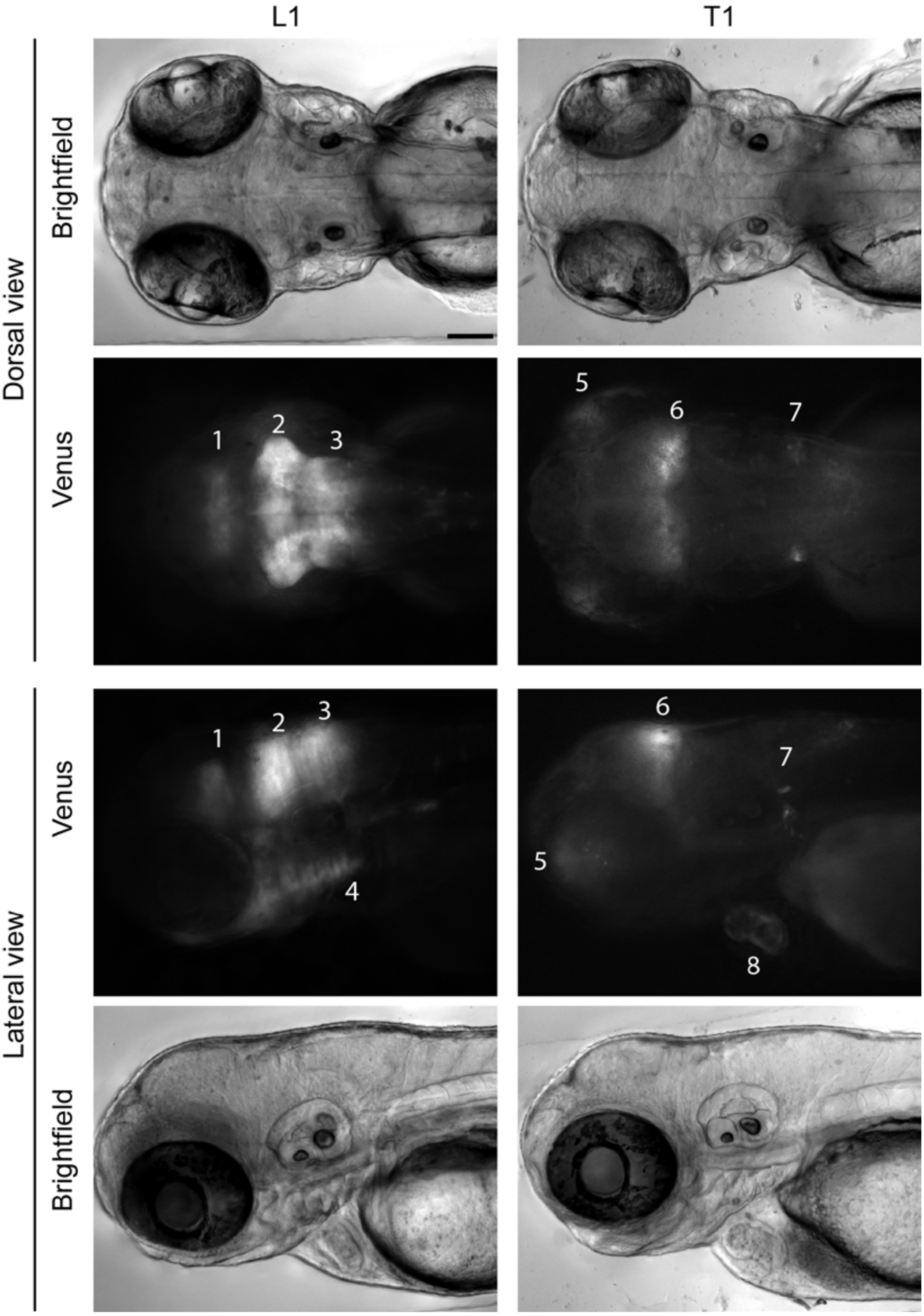


Figure 6.12 Tol2-based transgenic lines established with *TCF7L2* infrequent T allele and zebrafish *tcf7l2* promoter only.

Expression patterns shown by Tol2 transgenic lines injected with pTol2/Hs.TCF7L2(C):Dr.tcf7l2:Venus construct. Expression patterns are numbered from 1 to 15 as described in **Table 6.8**). Dorsal and lateral views of YFP and brightfield channels are shown. Larvae are protruding mouth stage (approximately 3 dpf) with anterior to the left. Scale bar indicates 100 µm.

Table 6.8 Analysis of the domains of Venus reporter expression and their distribution among *TCF7L2* enhancer candidate containing Tol2 transgenic lines compared to endogenous patterns.

Domains of expression		Zebrafish *tcf7l2* expression pattern	Tol2-based stable transgenic lines						
			Tg(Fr.ek:-0.6Dr.tcf7l2:Venus)	**Tg(Hs.TCF7L2(C):-0.6Dr.tcf7l2:Venus)**					**Tg(Hs.TCF7L2(T):-0.6Dr.tcf7l2:Venus)**
			L1	**C1**	**C2**	**C3**	**C4**	**C5**	**T1**
1	Dorsal diencephalon	🟪	▨						
2	Midbrain		▨						
3	Rostral hindbrain		▨						
4	Brachial arches		▨						
5	Ventral telencephalon								🟩
6	Tectum						🟦		🟩
7	Cranial ganglia						🟦	🟦	🟩
8	Heart			🟦	🟦	🟦			🟩
9	Telencephalon			🟦	🟦	🟦		🟦	
10	Optic stalks			🟦	🟦		🟦		
11	Retina					🟦			
12	Central nervous system					🟦			
13	Dorsal hindbrain			🟦	🟦				
14	Midbrain-Hindbrain boundary						🟦		
15	Pancreatic islet						🟦		

6.3.5 Analysis of *DGKB-TMEM195* and *ADCY5* candidate enhancer containing-transgenic lines

Despite screening an equivalent number of adult founder fish injected with constructs containing *DGKB-TMEM* and *ADCY5* enhancer variants, I could not identify any founder whose offspring was positive for reporter gene expression (**Table 6.9**). This result combined with the lack of activity in transient transgenic experiments argued for these candidate enhancers not functioning as transcriptional enhancers in zebrafish.

Table 6.9 Summary of transgenic lines established using *DGKB-TMEM195* and *ADCY5* putative enhancers.

Line	No. of screened founders	No. of Venus positive founders	Germline transmission rate (%)
Tg(Hs.DGKB-TMEM195 (G):Dr.hsp70:Venus)	25	0	0.0
Tg(Hs.DGKB-TMEM195 (G):Dr.hsp70:Venus)	21	0	0.0
Tg(Hs.ADCY5 (A):Dr.hsp70:Venus)	22	0	0.0
Tg(Hs.ADCY5 (G):Dr.hsp70:Venus)	16	0	0.0
Tg(Fr.ek:Dr.hsp70:Venus)	20	0	0.0

6.4 DISCUSSION

In this chapter I aimed to establish a method to detect *in vivo* quantitative differences between T2D-associated allelic variants using Tol2-based transient and stable transgenesis assays in zebrafish embryos. Functional test of 4 candidate human enhancers with T2D-associated SNPs (Dupuis et al., 2010) revealed that only a conserved region in *PROX1* human gene could drive reporter gene expression in the developing endocrine pancreas consistently in transient and stable transgenesis experiments. However, no differential expression patterns caused by the allelic variants could be detected with our approaches so far. The lack of detectable reporter gene expression with *DGKB-TMEM195* and *ADCY5* candidate enhancers argued that neither the putative CRE nor the allelic variants were functional in the zebrafish embryo or that these regions lacked enhancer function. Additionally, despite previous evidence suggesting that TCF7L2 candidate enhancer was active in pancreas-specific cell lines (Gaulton et al., 2010), the lack of detectable reporter gene activity in transient transgenesis experiments and the lack of reproducibility in stable transgenic lines prevented us to conclude accurately on *TCF7L2* function.

6.4.1 Utilizing Tol2-based transient transgenesis assays for allele quantification

Transient transgenesis assays in zebrafish allow the generation of hundreds of injected embryos, with a mosaic reporter gene expression, where typically each embryo displays reporter gene activity in a few cells (Pashos et al., 2008). Nevertheless, because each injected embryo represents a different transgene integration event, the position effects might be statistically eliminated by observing reproducible patterns in large numbers of embryos (Muller et al., 1997; Muller et al., 1999; Dickmeis et al., 2004; Gehrig et al., 2009; Rada-Iglesias and Wysocka, 2011; Taher et al., 2011). Mosaicism could however hinder the analysis of the activity of certain CREs in small tissues, such as the endocrine pancreas, which has approximately 20 β-cells at 3 dpf (Parsons et al., 2009). Our transient reporter assays with *PROX1* candidate enhancer revealed that on average 2% of the injected embryos were showing a pancreatic-specific signal. It could be argued that pancreas expression was not specific, and that it was a result of the random cell targeting of an otherwise non-tissue specific element. However, when *PROX1* candidate enhancer was linked to a minimal *hsp70* promoter, the pancreas activity was still reproducible, suggesting that *PROX1* could autonomously drive expression in the pancreatic islet.

Interestingly, *PROX1* reporter gene expression was not restricted to this domain. The rest of the expressing embryos showed reporter activity in tissues such as muscle, neurons and heart. From these domains, only expression in the neurons and in the heart was reproducible when two different zebrafish promoters were tested, suggesting that the activity in neurons and heart was attributable to the *PROX1* enhancer. Consistent with this, zebrafish *prox1* gene expression is detected in the pancreas, liver, otic and lens placodes, lateral line and central nervous system (Pistocchi et al., 2008), and has proven to be essential for liver organogenesis, neuronal development and functionality of the lateral line (Pistocchi et al., 2009).

Overall, the analysis of transient transgenic data did not reveal statistically significant differences between the candidate enhancer variants. And even when the co-localization approach increased our accuracy in detecting pancreatic specific expression, the mosaicism of the embryos could have hindered the detection of expression differences, suggesting that an alternative approach was required.

6.4.2 Establishing Tol2-based stable transgenic lines to quantitate allelic variants

The establishment of stable transgenic lines can be advantageous in order to detect weak enhancer function by avoiding the mosaicism of injected embryos and could thus be used to validate expression patterns detected by transient transgenic assays. Nevertheless, we observed a high variability of expression patterns among *PROX1* and *TCF7L2* candidate enhancer containing transgenic lines. The variability of expression patterns caused by Tol2-based transgenesis is larger than the subtle changes expected by the SNPs, which hindered our attempt to quantify subtle differences caused by disease-associated SNPs.

A BAC enhancer-trap study established that regulatory variation within *Tcf7l2* was leading to over-expression of the gene, which could be the underlying cause of T2D. However, this work also failed to detect allele-specific differences *in vivo* (Savic et al., 2011). Conflicting observations between *in vivo* and *in vitro* models may also be explained because the interaction of the region of interest outside the pancreas cannot be reproduced, as it is when using an animal model. In fact, when Savic and colleagues tested the same 239 bp region in a variety of cell lines, they detected enhancer activity in beta-pancreatic, as well as in neuron, bone and myoblast cell lines (Savic et al., 2013). Certainly, an experimental design where allelic specific differences can be measured will be difficult to achieve using transgenesis assays due to position effect, copy number variation and mosaicism. This is why the preferred approach in most studies relies on the use of cell lines.

6.4.3 Alternative approaches for quantitation of variant enhancer activity

Since differences in tissue specificity between allelic variants could not be detected by manual inspection of mosaic transgenic embryos, it could be argued that the use of an high-throughput automated measurement of the number of cells and intensity of fluorescence in embryos may provide a more accurate and unbiased spatial detection of the reporter activity (Gehrig et al., 2009). Nevertheless, the published pipeline was developed for the analysis of 36-42 hpf embryos and re-designing the existing software required computational and engineering support that was not available during this project.

An alternative approach would be to use BAC transgenesis to improve synergistic enhancer function detection (Balciunas et al., 2006; Suster et al., 2009). The use of BACs is more advantageous than the use of smaller vectors because their large size can include regulatory elements that are scattered in long distances (Yang et al., 2006). Besides, it permits to control the genomic context. Although these results have not been shown in this chapter, I created a modified human BAC containing *PROX1* human locus and followed an enhancer trap strategy to replace the 5´end of the second intron, where the candidate enhancer was located, with *hsp70* minimal promoter and eGFP reporter gene. However, the injection of the modified BAC of *PROX1* human gene did not drive any specific tissue-specific expression in the zebrafish embryos. This could be explained by the BAC itself, by the enhancer trap approach or by weakness of the pattern. Due to the lack of activity, we decided not to proceed with this time-consuming method and decided to investigate alternative approaches for allelic quantification.

6.4.4 Site-directed transgenesis as an alternative to Tol2-based transgenesis

Random transgene integration and position effects are a major drawback of Tol2-based transgenesis. As I have shown in this chapter, it often leads to the creation of a large number of transgenic lines per construct, among which there is substantial variation of expression patterns that hinders the analysis of function. A promising alternative to aid the screening of CREs by reducing the position effect variation would be the development of a transgenesis system mediated by site-specific recombinases. The validation of a targeted integration system for enhancer testing will be discussed in the next chapter.

Chapter Seven: DEVELOPMENT OF PHIC31 TARGETED INTEGRATION SYSTEM FOR SNP QUANTITATION

FOREWORD

The results presented in this chapter have been partially published in (Roberts et al., 2014).

The site-specific integration system mediated by PhiC31 integrase in zebrafish was designed by a collaboration between the labs of Ferenc Müller and Darius Balciunas (Temple University, US). Jennifer Roberts (former PhD student from the lab) initiated the development of this system as her PhD project (Roberts, 2013) until November 2012. I decided to extend this preliminary work with the final aim of using this system to quantitate SNP variants. I also carried out a series of experiments aiming to validate the targeting system for enhancer testing and to prove that the system was indeed capable of eliminating position effect variation.

- All the plasmids used for EL161 enhancer testing were made by Jennifer Roberts, as well as the docking transgenic lines Tg(Xla.crygc:attP-GFP)uobL6 and uobL12*.*
- The plasmid pCS2+PhiC31o-nos1-3'UTR was a kind gift from Shannon Fisher
- The plasmid pTol2/Xla.crygc:attP-GFP (pDB896) was made by Darius Balciunas

Information about the construction of these and other plasmids is presented in (Roberts et al., 2014).

7.1 INTRODUCTION

7.1.1 Site-specific recombination systems used in transgenesis

Functional analysis in zebrafish usually involves the injection of constructs containing predicted *cis*-regulatory elements placed upstream of a minimal promoter linked to a fluorescent reporter gene. The expression of fluorescent reporter gene in the injected embryos is then used as readout for CRE activity. In most cases, Tol2 transposase-based transgenesis is used to facilitate the integration of the transgene into the host genome (Kawakami, 2005; Fisher et al., 2006b). However, one of the main limitations of Tol2-based transgenesis is position effect variation caused by the random integration of the transgene into the genome. The term "position effect" describes alterations of expression pattern and subsequent variation among individuals injected with the same transgene, due to the influence of the genomic context where the construct was integrated (Jaenisch et al., 1981; Wilson et al., 1990; Rossant et al., 2011). As a consequence of position effects, stable lines generated from the same construct usually display different expression patterns that hinder the assessment of the function of the tested CRE, besides forcing the generation of many stable transgenic lines to conclude on the correct expression pattern of an element.

Position effect variation can be reduced by the use of site-specific recombinases (SSR). They function by recognizing target DNA sequences, bringing them together in a synapse and exchanging DNA strands after cleavage (Stark et al., 1992). There are two main families of SSR with different mechanisms of action. Cre from *Bacteriophage* P1 and Flp recombinase from *S. cerevisiae* belong to the λ family of integrases and can recombine DNA without accessory co-factors (Bischof and Basler, 2008). They recognize loxP and FRT target sites respectively that consist of two identical 13 bp palindromic repeats separated by an asymmetric 8 bp core (Hoess et al., 1982; McLeod et al., 1986). Depending on the directionality of these target sites, the recombination process can result in excision, inversion or integration of DNA fragments. A main limitation of Cre/lox and Flp/FRT systems is the bidirectional and reversible nature of the recombination reaction, which makes integration particularly unstable (Sauer and Henderson, 1990). In contrast, PhiC31 integrase from *Streptomyces* phage Φ31, a member of the resolvase family of SRR, can catalyse directional or irreversible recombination of heterotypic attP and attB sites autonomously (Thorpe and Smith, 1998). The resultant attL and attR hybrid sites cannot be used as a substrate by the integrase, which is advantageous when integration is the only goal (Thorpe et al., 2000). The natural function of PhiC31 integrase is to facilitate the integration of the phage genome into the bacterial host genome by binding to two recognition sites, attP and attB, cleaving four DNA strands in a synaptic complex, rotating 180° one pair of half sites and ligating DNA strands (Smith et al., 2004). PhiC31 integrase has been shown to be active and efficient in mediating site-specific integration in several animal systems that include human cells (Groth et al., 2000), mice (Belteki et al., 2003), flies (Olivares et al., 2002; Groth et al., 2004) and frogs (Allen and Weeks, 2005).

7.1.2 PhiC31 integrase-based systems in zebrafish

In zebrafish the first reports of PhiC31 integrase being active in somatic cells and the germline used a cassette-excision based strategy (Lister, 2010; Lu et al., 2011). This system´s design allowed them to monitor the rate of PhiC31-mediated intramolecular recombination by flanking a GFP fluorescent reporter gene with attP and attB sites, which would be excised if recombination happened (Lister, 2010). They also proved that the mouse-codon optimized version of the PhiC31 integrase was twice as efficient in zebrafish than the native one (Raymond and Soriano, 2007).

PhiC31 also allows for recombination-based cassette exchange, as demonstrated by (Hu et al., 2011). This strategy, which is also functional in *Drosophila* (Bateman et al., 2006), was employed to exchange a reporter gene on targeted transgenes that had been incorporated into the zebrafish germline through Tol2-based experiments. They also tested PhiC31 integrase fused with 3´UTR of a primordial germ cell-specific *nanos1* gene and concluded that it was significantly increasing the rate of recombination events in the germline (Hu et al., 2011).

The great potential of PhiC31 integrase-based technology is the creation of attB/attP containing docking sites in the host genome, where site-directed recombination is possible, using either cassette exchange or integration approaches. This method would circumvent the need to create multiple transgenic lines and reduce significantly the position effects variation derived from the random integration of the transgenes in the genome. More importantly, a robust targeted integration system would aid in the exploitation of the transgenic zebrafish model for the functional analysis of SNP variants in CREs identified in large numbers by genome wide disease association studies (Maurano et al., 2012), particularly on those cases where the SNP was expected to cause only subtle, quantitative changes in gene expression (Gaulton et al., 2010; Rada-Iglesias et al., 2013), that would not be detected with a conventional technology prone to position effects, such as the commonly used Tol2-based transgenesis.

7.1.3 Aims:

In order to be able to quantitate expression differences caused by enhancer variants, we decided to develop a PhiC31-based method for transgene integration in zebrafish. We hypothesized that PhiC31 integrase-mediated transgenesis could reduce the variability to allow for a sensitive system that could detect subtle variations in expression patterns, such as those caused by disease-linked SNP variants. Within this global aim we set the following specific objectives:

- To identify a single-copy recipient transgenic line that allows for robust enhancer tests.
- To examine PhiC31-mediated recombination frequency and test the toxicity of the two PhiC31 integrase mRNA variants described in the literature.

- To use PhiC31 mediated transgenesis to quantitate subtle differences caused by disease-associated SNP variants within enhancers.

7.2 Preliminary data leading to the project: Development of PhiC31-based transgenesis system in the zebrafish embryo

In collaboration with Darius Balciunas (Temple University, US), Ferenc Müller and Jennifer Roberts designed a site-specific integration system mediated by PhiC31 integrase in zebrafish. This system consists of two components: a "recipient transgenic line" containing a docking attP site in the genome and a targeting plasmid containing a donor attB site. To facilitate the screening of site-specific integrations, a phenotypic selection tool based on fluorescent reporter colour change in the lens was designed (**Figure 7.1A**).

Recipient transgenic lines Tg(Xla.crygc:attP-GFP) contained a lens-specific promoter (*gamma-crystallin* from *Xenopus laevis* (Davidson et al., 2003), and attP site located upstream of GFP. The crygc:attP-GFP cassette drives green fluorescent expression in the lens. When the recipient transgenic line is injected with a targeting vector (pJET-attB-mCherry, (Roberts et al., 2014)) containing attB site upstream of a red fluorescent reporter gene (mCherry) and integrase mRNA, PhiC31 mediated integration is expected to produce crygc:attR-mCherry recombinant site, which can be scored by green to red colour change in the lens (**Figure 7.1B**).

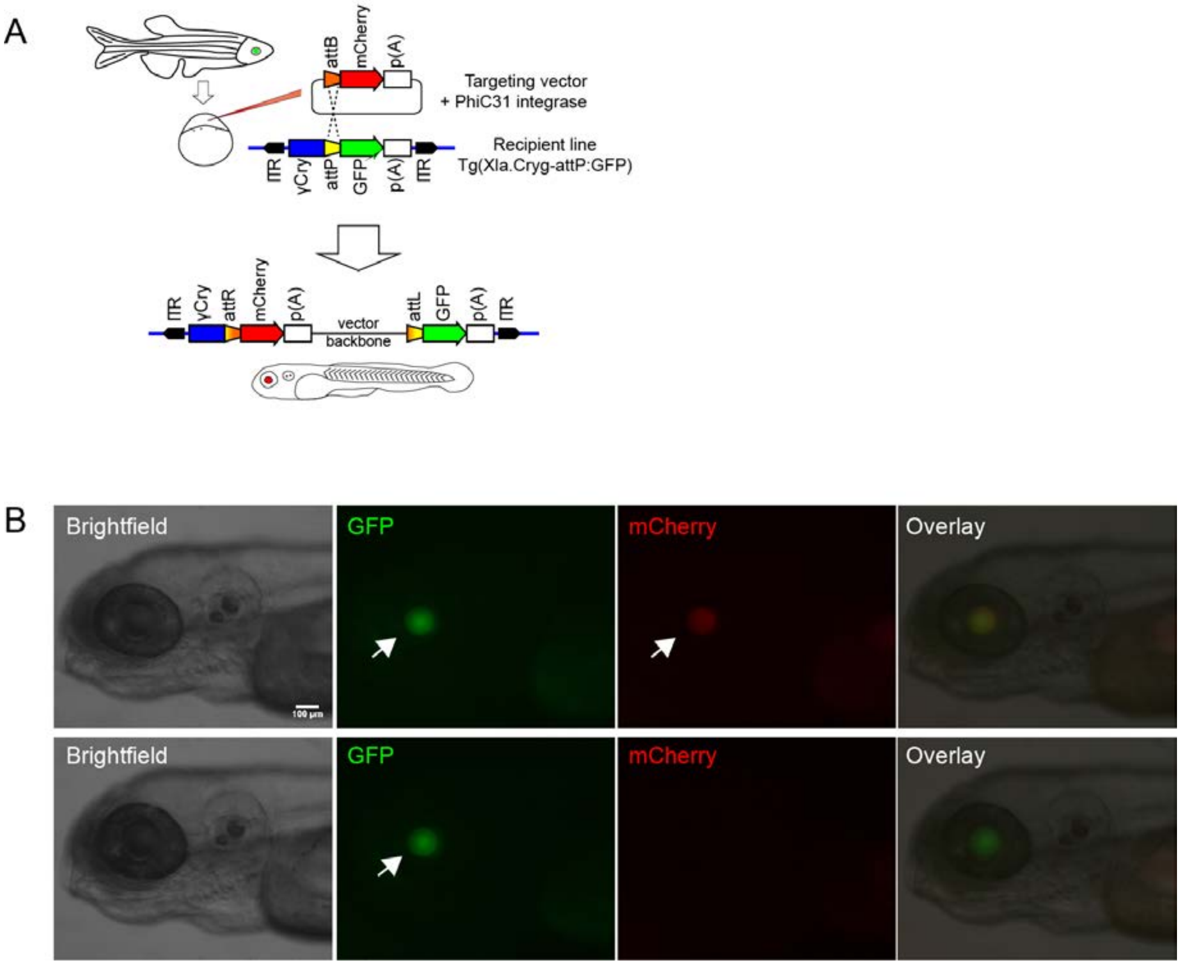


Figure 7.1 *In vivo* detection system for targeted integration of reporter constructs in zebrafish.
A. Schematic of PhiC31 targeted integration system. Transgenic embryo containing attP docking site shows green lens activity due to crystalline promoter driving GFP reporter gene. ITR labels Tol2 recognition sequences. Injection of a circular plasmid with attB and red reporter (targeting vector) leads to eye colour change upon PhiC31 targeted integration **B.** Detection of eye colour change upon PhiC31 mediated integration in the transgenic recipient Tg(Xla.crygc:attP-GFP)uobL6 line. Top row shows a 5 dpf transgenic larva where PhiC31 legitimate recombination of attB-mCherry cassette into the attP-GFP docking site has taken place. Bottom row shows an embryo where recombination did not occur. Side views of larvae with anterior to the left are shown. Scale bar indicates 100 µm.

7.3 RESULTS

7.3.1 Analysis of the zygosity of zebrafish recipient transgenic lines

In order to study changes in gene expression due to enhancer SNPs by targeted transgenesis, the transgenic recipient lines must carry a single and stable landing site, that is, only one copy of the Xla.crygc:attP-GFP transgene in the genome. Tol2-based Tg(Xla.crygc:attP-GFP)uobL6 and Tg(Xla.crygc:attP-GFP)uobL12 recipient transgenic lines (*aka* uobL6 and uobL12 for short) had been established by Jennifer Roberts by injecting WT embryos with pTol2/Xla.crygc:attP-GFP (pDB896) vector and growing up lens-GFP positive embryos, as described in (Roberts et al., 2014).

To establish that the recipient line contained only a single copy of the attP site, I studied the germline transmission rates of multiple individuals from different generations of the recipient lines available at that time in the lab (uobL6F2 and F3 and uobL12 F2) in order to assess whether they approached Mendelian rates. Transgenic adults were bred to wild-type fish (1:1 crosses) and the offspring was analysed for green lens expression at 5 dpf. On average, transmission rates were close to 50% (**Table 7.1**) for individuals from uobL6 and uobL12, suggesting that there was only one integration site in the genome. I also identified two adult females from Tg(Xla.crygc:attP-GFP)uobL6 F2 that produced 100% of GFP positive embryos, indicating that they were homozygous and had resulted from an incross of uobL6F1 individuals.

Table 7.1. Germline transmission rate of existing transgenic recipient lines

Transgenic line ID	Generation	Number of individuals outcrossed	Total number of embryos analysed at 5 dpf	Germline transmission, average (%)
Tg(pTol2/*Xla.*Cryg-attP:GFP)*uobL6*	F3	8	648	51.4
	F2	2	269	100.0
Tg(pTol2/*Xla.*Cryg-attP:GFP)*uobL12*	F2	15	2764	49.8

7.3.2 Molecular characterization of location, integrity and copy number of integration events in transgenic recipient lines

As a first step towards the potential generation of a library of attP-containing recipient transgenic lines, I characterized molecularly the transgene integration events of the uncharacterized lines existing in the lab. In order to map the genomic locus harbouring the Xla.crygc:attP-GFP transgene in uobL6 and uobL12 recipient lines, extension primer tag selection linker-mediated PCR (EPTS-LMPCR) was carried out on GFP+ clutches of outcrossed embryos from uobL12 F2, uobL6 F2 and F3 transgenic recipient lines. EPTS-LMPCR is a variation of LMPCR where genomic DNA fragments containing the transgene(s) are purified using a biotinylated primer that anneals to the Tol2 terminal arms, therefore only allowing the mapping of transposon-mediated integration events. High-quality genomic DNA was extracted from 5 dpf embryo clutches and EPTS-LMPCR was carried out following a protocol described for *Xenopus* (Yergeau et al., 2007).

By EPTS-LMPCR I could map in uobL6F2 and F3 lines a single Tol2-mediated transgene insertion in an intergenic region of chromosome 18 (**Figure 7.2**), 40 Kb away from the nearest coding gene. This integration site was further verified by PCR using primers specific to the zebrafish genome (not shown). However, backbone sequence from pTol2/Xla.crygc:attP-GFP (pDB896) were found flanking the transposon integration sites in the uobL12 F2 line, suggesting that there had been an illegitimate insertion of Tol2, where not only the Tol2-flanked cassette but also plasmid backbone had been integrated. Inverse PCR carried out by Jorune Balciunene and Darius Balciunas could map the integration site of uobL12 to an intergenic region of chromosome 1 (Roberts et al., 2014).

In order to verify the transgene copy number in both lines, I performed Southern blotting of 5 dpf lens-GFP+ and GFP- clutches of embryos coming from uobL6 F2 homozygous and uobL12 F2 heterozygous outcrosses (**Figure 7.2B,C**). We reasoned that if uobL6 homozygous recipient line contained one integration site in homozygosis, 100% of its offspring would be harbouring the docking Xla.crygc:attP-GFP cassette and therefore could be used for efficient SNP testing. Two sets of enzymes had to be used to be able to confirm copy numbers: SacI with BamHI, which cut once in the construct backbone and once upstream attP site, respectively and BamHI and ApaI, cutting twice in the transgene (**Figure 7.2D**). Southern blot analyses confirmed that uobL12 carried a Tol2-mediated multimer integration of the transgene with an approximate copy number of 11.6 (**Figure 7.2B, C, D**). Surprisingly, the Southern blot and Phosphoimager quantification also indicated that uobL6 homozygous individuals carried two independent copies of the Xla.crygc:attP-GFP transgene, one of them potentially being an insertion of the whole (pTol2/Xla.crygc:attP-GFP (pDB896) vector plasmid (**Figure 7.2B**).

Because uobL6 F1 had been incrossed to produce F2, we hypothesized that there might have been a mixed population of transgenic fish in F1, where some fish were carrying the legitimate Tol2-mediated chr18 insertion identified by LMPCR, and some others were carrying the latter plus a full plasmid copy, which was located on a different chromosome and could therefore segregate. In order to test this hypothesis, regular PCRs were carried out with oligos present in the backbone of the pTol2/Xla.crygc:attP-GFP (pDB896) vector using genomic DNA from different uobL6 F2 and F3 individuals (**Figure 7.3**). PCR results indicated that the extra plasmid insertion was only present in F2 homozygous female 3 (uobL6F2.3) and therefore could segregate. These PCR results together with Southern blot carried out on uobL6 heterozygous fish by collaborator Darius Balciunas, showed that uobL6 F3 individuals did not carry the extra copy, and were therefore used for further experiments and to establish the next generation of uobL6.

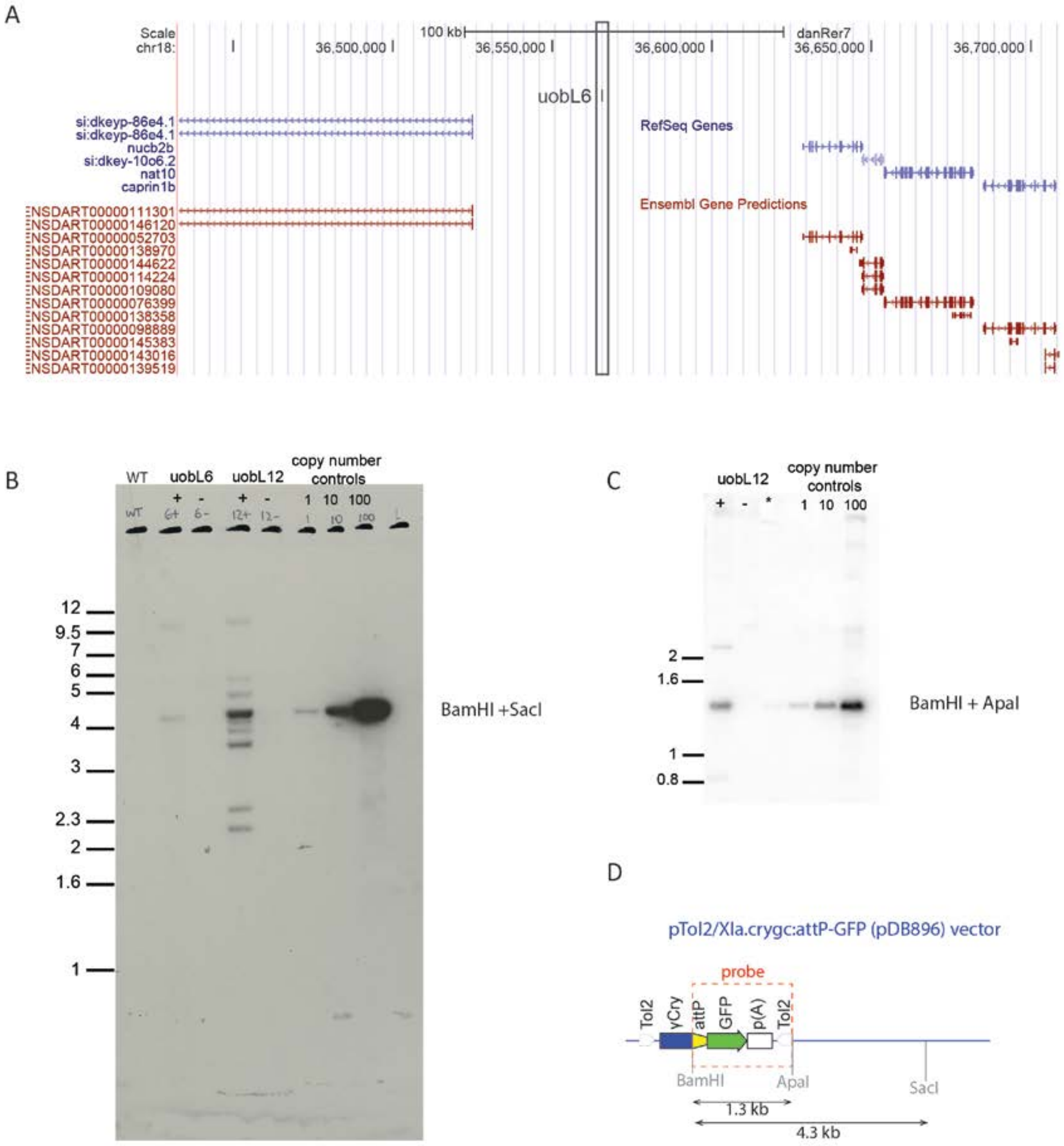


Figure 7.2 Molecular analysis of uobL6 and uobL12 transgenic recipient lines by EPTS-LMPCR and Southern blot.

A. USCS screenshots displaying the genomic context of transgene integration in uobL6 recipient line (grey vertical rectangle). The insertion site is 40 kb downstream of *si:dkeyp-86e4.1* and 65 kb upstream of the *nucb2b* gene. **B.** Southern blot analysis of uobL6 and uobL12 recipient transgenic lines. Genomic DNA from batches of GFP-positive and GFP-negative embryos was digested with *BamHI* and *ApaI,* cutting once in the transgene and once in the backbone. Southern hybridization was performed with a radioactively labelled attP:GFP:Tol2 1.3 Kb probe (depicted in **D**). **C.** Copy number analysis of uobL12 was performed by digesting genomic DNA with BamHI and ApaI, which cuts in the backbone of pTol2/Xla.crygc:attP-GFP (pDB896) vector producing an expected band of 1.3 Kb. Phosphoimager analysis suggests uobL12 carries 11.6 copies. **D.** Schematic of pTol2/Xla.crygc:attP-GFP (pDB896) vector including restriction sites used for Southern blot analysis and distances between them. Probe sequence is depicted by a dashed red rectangle.

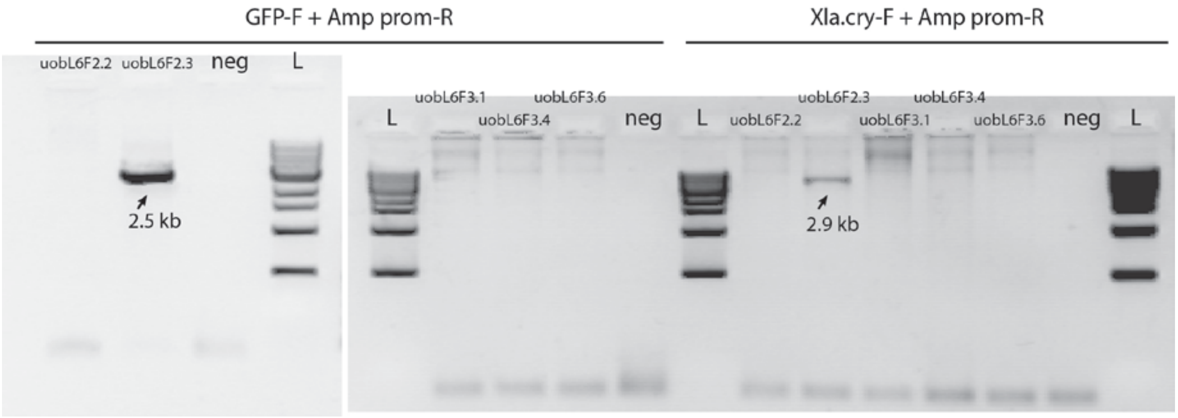


Figure 7.3 Diagnostic PCR to evaluate the presence of plasmid backbone in the uobL6 F2 homozygous fish

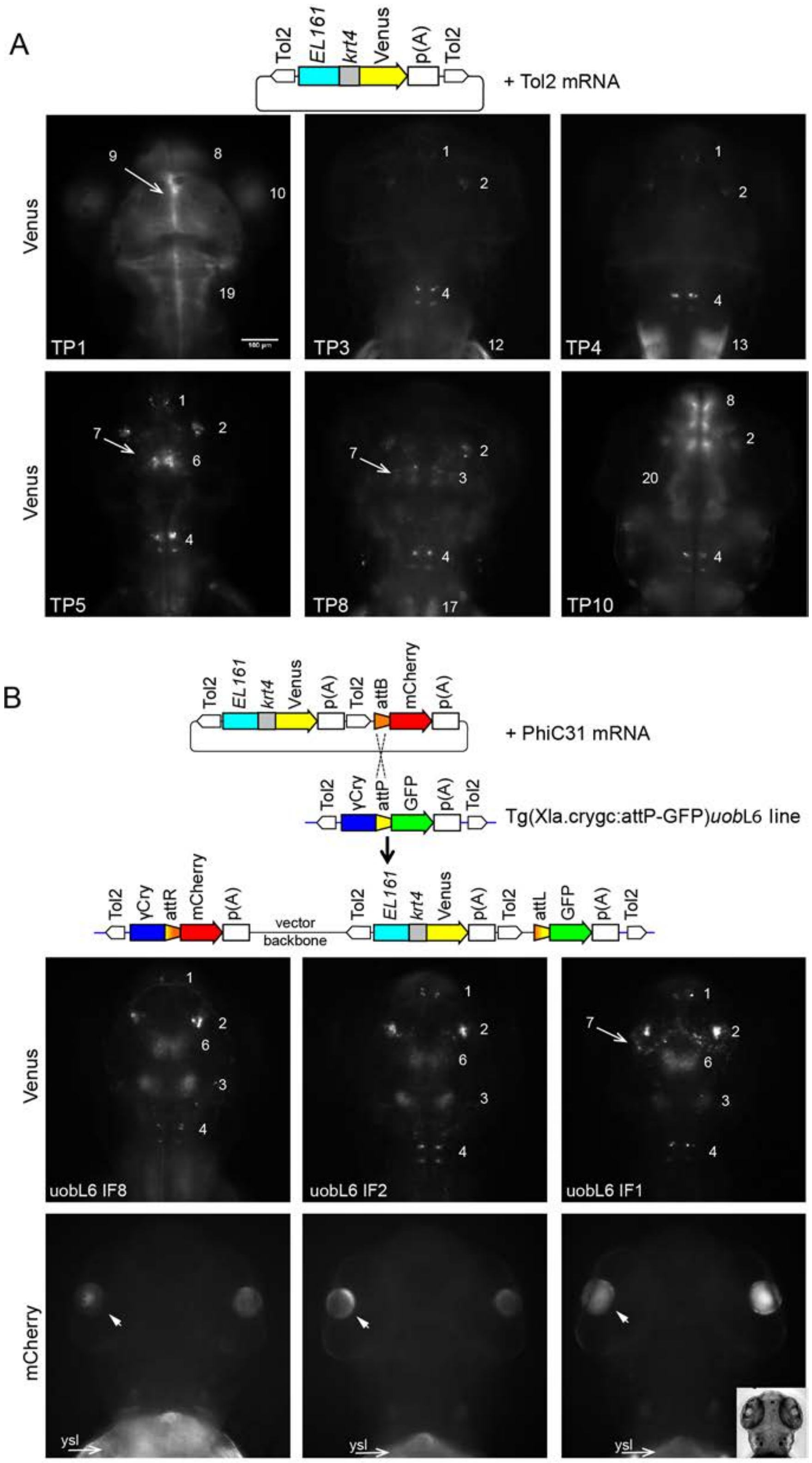
Two sets of primers annealing in the Xla.crygc:attP-GFP transgene (GFP and *X. laevis gamma-crystallin* promoter) and in pTol2/Xla.crygc:attP-GFP (pDB896) backbone construct (Ampicillin promoter reverse) were used to test whether the two homozygous individuals from uobL6 (uobL6F2.2 and uobL6F2.3), and heterozygous individuals from uobL6F3 (F3.1, F3.4, F3.6), carried a plasmid integration non-mediated by Tol2 that could explain the double bands found at the Southern blot (**Figure 5.3**). Black arrow points at the expected band produced by a successful amplification of pTol2/Xla.crygc:attP-GFP (pDB896) backbone. L indicates 100 bp DNA Ladder from New England Biolabs.

7.3.3 Analysis of position effect variability in Tol2 and PhiC31-mediated transgenesis

A common obstacle in transgenesis assays aiming to characterize enhancer function is position effect variability derived from the random integration of the transgene into the host genome, as discussed in **Chapter 6**. In order to test whether site-directed transgenesis mediated by PhiC31 would help to reduce the position effect variability of expression patterns, we decided to evaluate enhancer activity driven by a highly conserved zebrafish element after integration in the genome by Tol2 or PhiC31-based transgenesis. The selected candidate enhancer, named EL161, showed more than 80% sequence identity with humans and was identified by collaborator Remo Sanges (Stazione Zoologica Anton Dohrn, Italy) as described in (Roberts et al., 2014). It is located on the last intron of *esrrga* zebrafish gene (Bardet et al., 2004; Thisse et al., 2004) and was postulated to regulate *esrrga* expression. In a preliminary study carried out in the lab, EL161 was cloned upstream of the *krt4* gene minimal promoter, which shows minimal transcriptional activity (Gehrig et al., 2009), and tested in a Tol2-based transient assay in zebrafish. EL161 was shown to be active in the brain, however the expression was highly variable due to position effects (Yavor Hadzhiev, unpublished data).

The same Tol2-based cassette containing EL161 element linked to *krt4* promoter and Venus (Tol2/EL161-krt4:Venus) was cloned and tested by both Tol2 and PhiC31 transgenesis. The resulting vectors, namely pTol2/EL161-krt4:Venus and pattB-mCherry,Tol2/EL161-krt4:Venus, were injected in WT or uobL6/L12 embryos, respectively, as described in (Roberts et al., 2014). Tol2-mediated transgenesis is expected to integrate only the Tol2/EL161-krt4:Venus cassette into the genome (**Figure 7.4A top**), while PhiC31-mediated transgenesis integrates the whole vector, including Tol2/EL161-krt4:Venus and attB:mCherry cassettes (**Figure 7.4B top**). When injected embryos reached adulthood, founders were screened for positive offspring. Among 20 adults injected with Tol2 construct, 4 positive founders were identified (TF1-4, **Table 7.2**), where ten different enhancer-driven complex expression patterns were recovered (numbered TP1-10, **Table 7.3, Figure 7.5**). These results suggest multiple integration sites and that Tol2/EL161-krt4:Venus is highly susceptible to position effects (**Figure 7.4A**). Screening the offspring of all 29 surviving adult founders from *uobL6* and *uobL12* injected with PhiC31 integrase mRNA resulted in the identification of 4 positive founders (**Table 7.2**). Their offspring showed eye colour change during larval development together with EL161-driven YFP expression pattern, indicating site-specific integration events (**Figure 7.4B**, **Figure 7.6**). Notably, the 3 founders with targeted integration into the *uobL6* line showed remarkable similarity in their expression patterns, in contrast to Tol2 mediated transgene integrations, which were characterized by widely varying patterns (**Figure 7.4**).

Figure 7.4 Variability of position effects is sharply reduced among PhiC31-mediated transgenic lines when compared to Tol2 transgenesis
A. Schematic of pTol2/EL161-krt4:Venus construct co-injected with Tol2 transposase mRNA. Below examples of expression patterns of transgenic F1 larvae (3 dpf). TP indicates Tol2 mediated expression patterns. Venus expression domains are labelled as in **Figure 7.7B**. **B.** Schematic drawing on top indicates the recombination of the targeting plasmid pattB-mCherry,Tol2/EL161-krt4:Venus into the attP docking site of transgenic recipient line Tg(Xla.crygc:attP-GFP)*uobL6*. Larvae from three different founders (IF1, IF2, IF8) targeted with PhiC31 integrase are shown. Venus expression domains are labelled as in **Figure 7.7B**. Lens activity driven by mCherry (arrows in bottom row) demonstrates PhiC31-mediated integration. Insert in bottom right shows bright field representation of the head region imaged. Arrows in red channel indicate auto-fluorescence of the yolk syncytial layer. Dorsal views of larvae head are shown. Scale bar is 100µm.



(See previous page for figure legend)

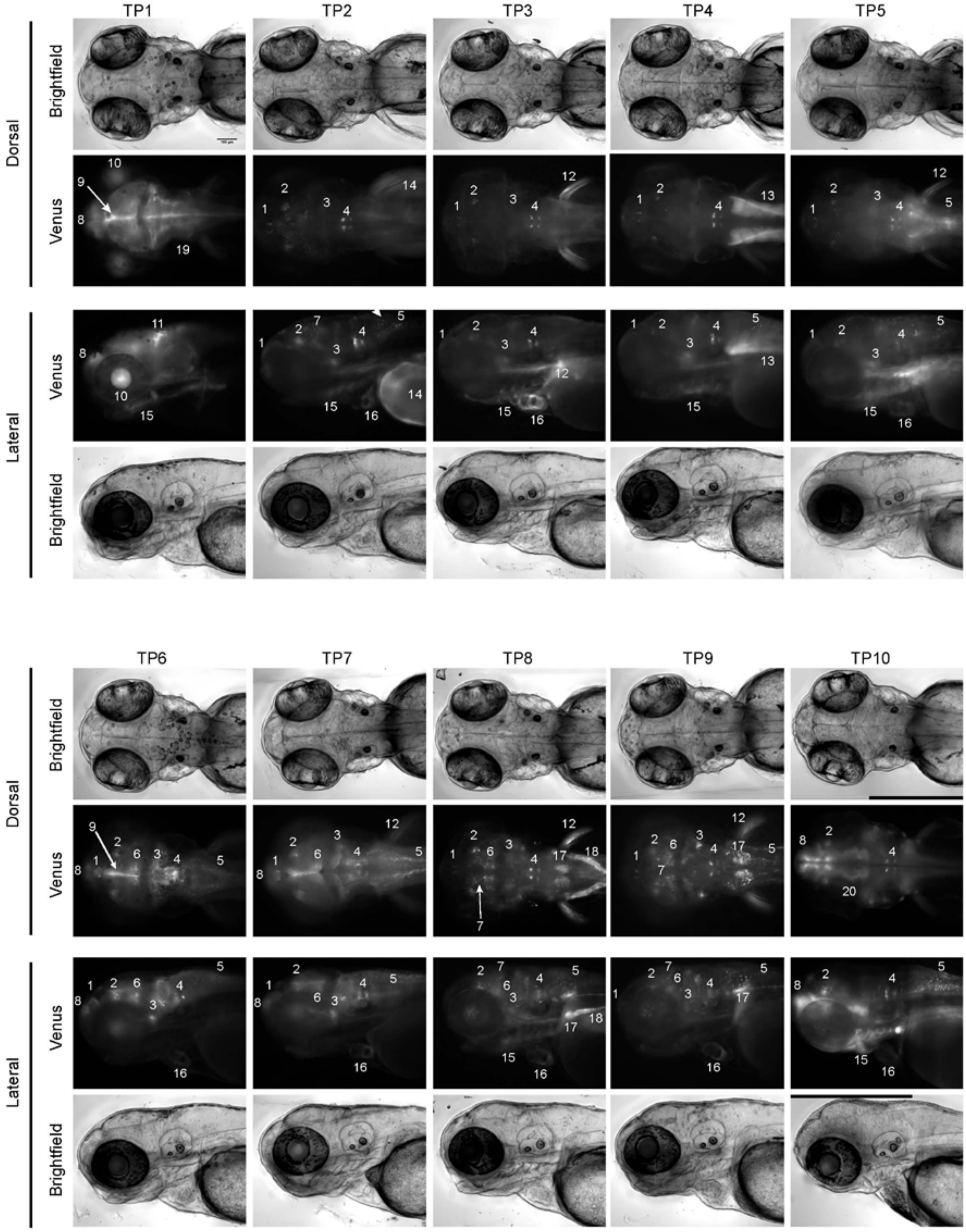


Figure 7.5 Tol2 transgenic founders show variable position effects on EL161 driven reporter activity.

All of the different expression patterns shown by Tol2 transgenic lines injected with pTol2/EL161-krt4: Venus construct. Tol2 patterns (TP) are numbered from 1 to 10. Dorsal and lateral views of YFP and brightfield channels are shown and annotated domains correspond to the expression domains described in **Figure 7.7**. Larvae are protruding mouth stage (approximately 3 dpf) with anterior to the left. Scale bar indicates 100 μm.

Table 7.2 Survival and transmission rates of PhiC31 and Tol2 founders

Strain injected	Method of transgenesis	Targeting vector	Number of screened individuals	Transmission rates % (n)
AB*WT	Tol2 transposase	pTol2/EL161-krt4:Venus	20	20 (4)
Tg(pTol2/Xla.Cryg-attP:GFP)uobL6	PhiC31 integrase	pattB-mCherry,Tol2/EL161-krt4:Venus	25	12.0 (3)
Tg(pTol2/Xla.Cryg-attP:GFP)uobL12	PhiC31 integrase	pattB-mCherry,Tol2/EL161-krt4:Venus	4	25.0 (1)

Table 7.3 Analysis of the offspring of Tol2 founders

Founder ID	Total number of embryos analysed	Transmission rates (%)	Tol2 patterns	Transmission rate per pattern (%)
Tol2 founder 1 (TF1)	246	15.0	TP1	2.8
			TP2	2.0
			TP3	3.3
			TP4	5.7
			TP5	1.2
Tol2 founder 2 (TF2)	109	5.5	TP6	4.0
			TP7	6.7
Tol2 founder 3 (TF3)	236	10.7	TP8	4.7
			TP9	0.8
Tol2 founder 4 (TF4)	150	12.6	TP10	12.6

Table 7.4 Analysis of the offspring of integrase founders

Founder ID	Total number of F1 embryos analysed	Percentage of green lens embryos (n)	Percentage of red lens embryos with YFP pattern (n)	Percentage of embryos with YFP pattern only (n)	Percentage of embryos with green and red lenses(n)	Percentage of negative embryos (n)
uobL6- Founder1 (IF1)	228	35.2 (81)	11.7 (27)	0.0 (0)	0.0 (0)	52.6 (120)
uobL6- Founder 2 (IF2)	116	36.2 (42)	6.0 (7)	0.0 (0)	0.0 (0)	57.8 (67)
uobL6- Founder8 (IF8)	372	43.0 (160)	11.0 (41)	0.0 (0)	0.0 (0)	51.1 (190)
uobL12- Founder1 (IF1)	100	22.0 (22)	15.0 (15)	0.0 (0)	21.0 (21)	42.0 (42)

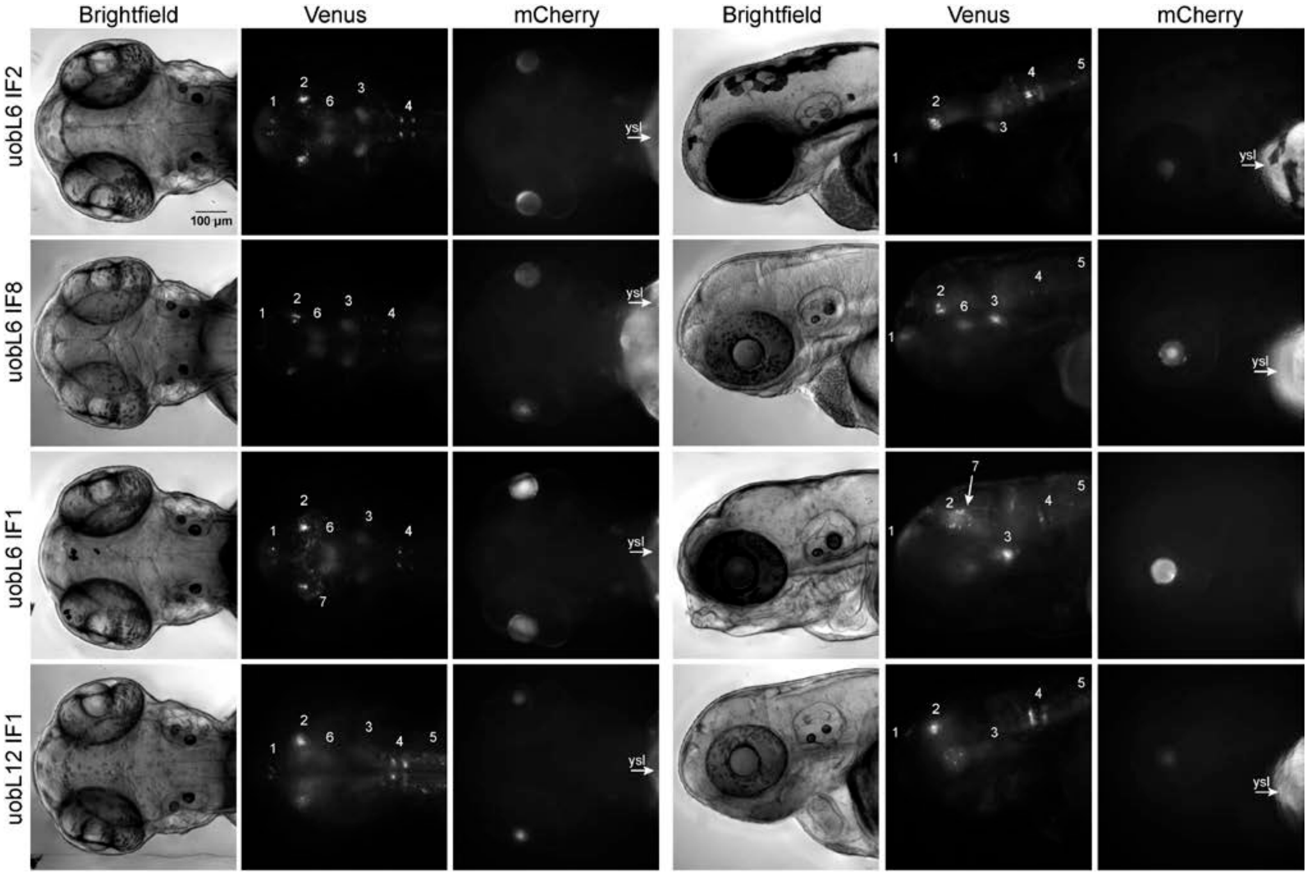


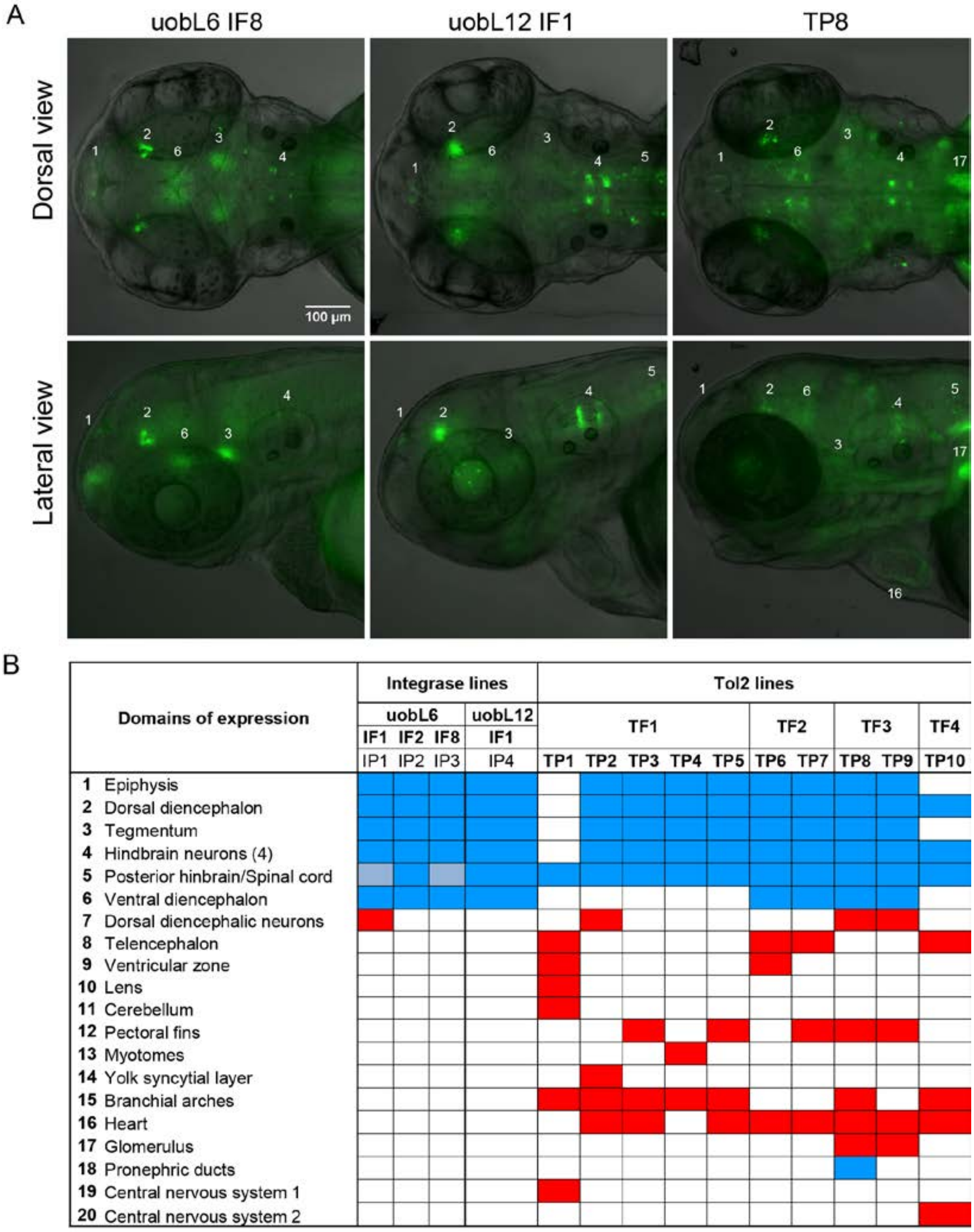
Figure 7.6 PhiC31-mediated transgenic founders show highly reproducible expression patterns.

Neuronal expression patterns obtained in integrase injected founders uobL6 IF1, uobL6 IF2, uobL6 IF8 and uobL12 IF1. Red lens indicates legitimate PhiC31-mediated integration into the Xla.crygc:attP-GFP cassette from either Tg(Xla.crygc:attP-GFP)uobL6 or Tg(Xla.crygc:attP-GFP)uobL12. Auto-fluorescent yolk syncytial layer (ysl) is marked by arrows in mCherry images. YFP channel shows EL161-driven expression patterns are listed in **Figure 7.7**. Larvae are at protruding mouth stage (3 dpf) with anterior to the left. Scale bar indicates 100 µm.

7.3.4 Evaluation of EL161 enhancer function using PhiC31-mediated transgenesis

To evaluate the specific reporter activity driven by EL161 element and to dissect enhancer driven activity from position effects, we identified distinct expression domains and analysed the frequency of their occurrence in targeted and random transgene integration loci. A total of 20 domains of activity labelled 1-20 were identified (**Figure 7.7**). PhiC31 integrase-mediated activity was present in 7 distinguishable expression domains in the neural tube (**Figure 7.4B**, **Figure 7.6**), six out of which were shared between the 3 integrase lines with targeting events in the same uobL6 docking site. Only one additional domain (Domain 7) was registered in one of these lines (uobL6 IF1) indicating mild variation (**Figure 7.7B**) that was not due to mutations of the integrated transgene. In contrast, Tol2 lines showed a total of 21 expression domains with a variation between 7-12 domains per pattern, which included a variety of tissues such as somatic muscle, pectoral fins and heart besides neural tube activity (**Figure 7.7**, **Figure 7.5)**. This result indicated that integrase-mediated targeting of a single locus (in uobL6) led to reduced variability of expression patterns induced by a neural enhancer among transgenic lines.

The expression domains 1-6 were shared among uobL6 and uobL12 integrase founders and were present in 8 out of 10 Tol2-mediated patterns, which together suggested autonomous enhancer activity in these neural domains. This result demonstrated that the shared expression patterns were unlikely to be due to position effect at the *uobL6* locus and were an autonomous property of the targeting transgene (**Figure 7.7**). Notably, these 6 shared expression domains including epiphysis, diencephalic and hindbrain nuclei and tegmentum, overlapped with expression domains of the *esrrga* gene (**Figure 7.8**, www.zfin.org, (Bardet et al., 2004) suggesting that EL161 may contribute to the activity of *esrrga* in these neural subdomains. Low level of variability was still observed in the integrase transgenic lines, at least in part explained by variation in focal planes of imaging as well as by potential differences in developmental stage of the individuals imaged. For stage-dependent variation of transgene activity see **Figure 7.9**. Nevertheless, the variability of expression patterns among PhiC31-mediated targeted integrants was significantly lower than those found among Tol2-mediated transgene integrants (Likelihood Ratio = 15.0, DF = 1, p = 0.0001). Taken together, the low variability of transgene activity observed in PhiC31 targeted integration events demonstrated superior reproducibility and robustness of enhancer driven expression patterns as compared to that obtained by transposase mediated integration. These experiments demonstrated that PhiC31-mediated system could limit if not completely overcome position effect variability, and therefore represented a refined alternative for SNP quantitation.



Domains of expression	Integrase lines				Tol2 lines									
	uobL6			uobL12	TF1					TF2		TF3		TF4
	IF1	IF2	IF8	IF1										
	IP1	IP2	IP3	IP4	TP1	TP2	TP3	TP4	TP5	TP6	TP7	TP8	TP9	TP10
1 Epiphysis	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	
2 Dorsal diencephalon	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
3 Tegmentum	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	
4 Hindbrain neurons (4)	🟦	🟦	🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
5 Posterior hinbrain/Spinal cord	▨	🟦	▨	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
6 Ventral diencephalon	🟦	🟦	🟦	🟦						🟦	🟦	🟦	🟦	
7 Dorsal diencephalic neurons	🟥					🟥						🟥	🟥	
8 Telencephalon					🟥					🟥	🟥			🟥
9 Ventricular zone					🟥					🟥				
10 Lens					🟥									
11 Cerebellum					🟥									
12 Pectoral fins							🟥		🟥		🟥	🟥	🟥	
13 Myotomes								🟥						
14 Yolk syncytial layer						🟥								
15 Branchial arches					🟥	🟥	🟥	🟥	🟥			🟥		🟥
16 Heart						🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥
17 Glomerulus												🟥	🟥	
18 Pronephric ducts												🟦		
19 Central nervous system 1					🟥									
20 Central nervous system 2														🟥

Figure 7.7 Reproducible Venus expression patterns upon targeted integration of EL161 enhancer construct demonstrate *cis*-regulatory function in the brain

A. Brain-specific enhancer effect of a transgene inserted in different integration sites. Domains of Venus activity are specified in panel B. Dorsal (top) and lateral (bottom) views onto the head of 3 dpf F1 transgenic larvae. Scale bar is 100 µm. **B**. Domains of Venus reporter expression and their distribution among targeted (PhiC31 integrase) and randomly (Tol2 transposase) transgenic lines. Abbreviations: IF- integrase injected founders, IP- integrase mediated expression patterns, TF-Tol2 injected founders and TP-Tol2 mediated expression patterns. Blue depicts expression domains overlapping with *esrrga* activity, whereas red colour indicates additional domains. Light blue indicates weak expression.

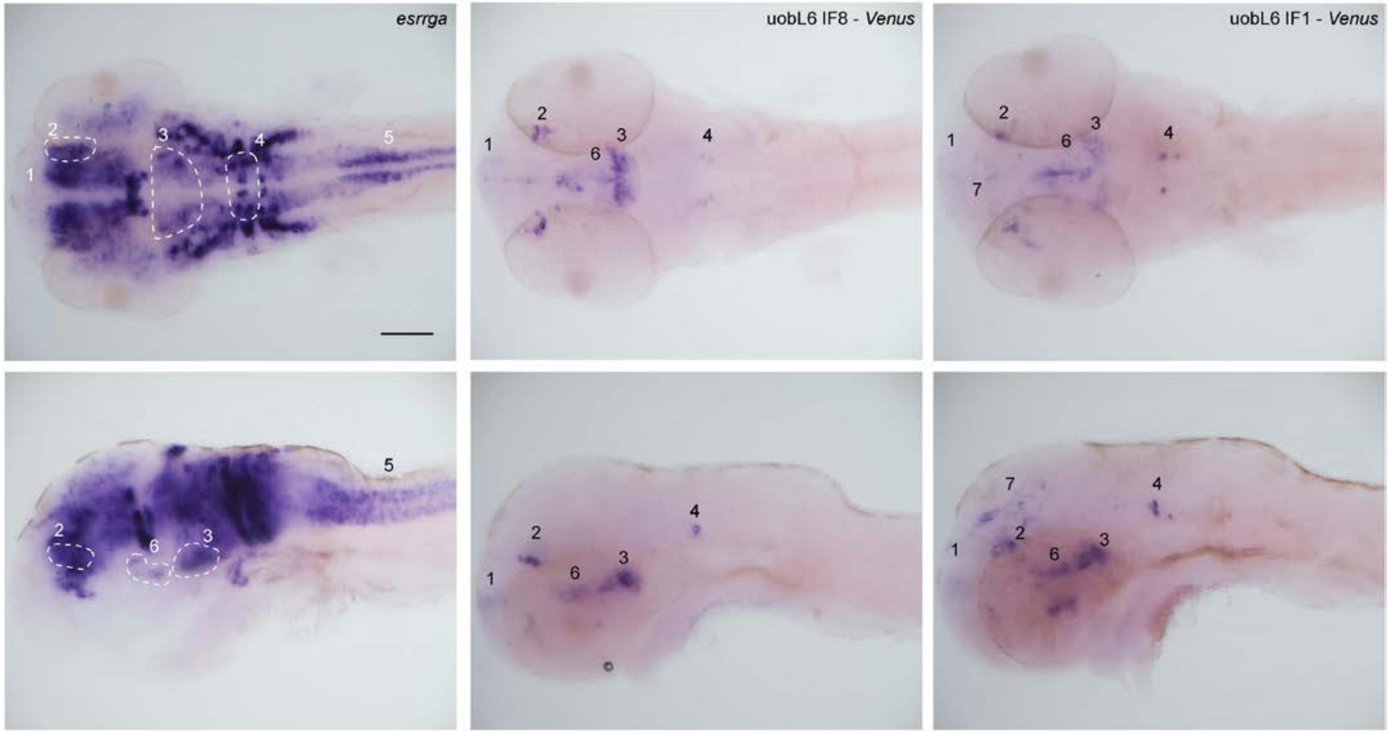


Figure 7.8 Tissue-specific activity driven by EL161 element overlaps with endogenous expression domains of endogenous esrrga gene.
60 hpf embryos probed with *esrrga* (left) share 5 expression domains with integrant founders from uobL6 (right) probed with Venus. Dorsal and lateral images are shown. Expression domains are labelled as listed in **Figure 7.7**. Note that *esrrga* domain 1 (epiphysis) is weakly expressed at this stage, but shows a strong expression earlier in development. Lens from bottom left embryo was removed to allow visualization of inner brain domains. Scale bar indicates 100 µm.

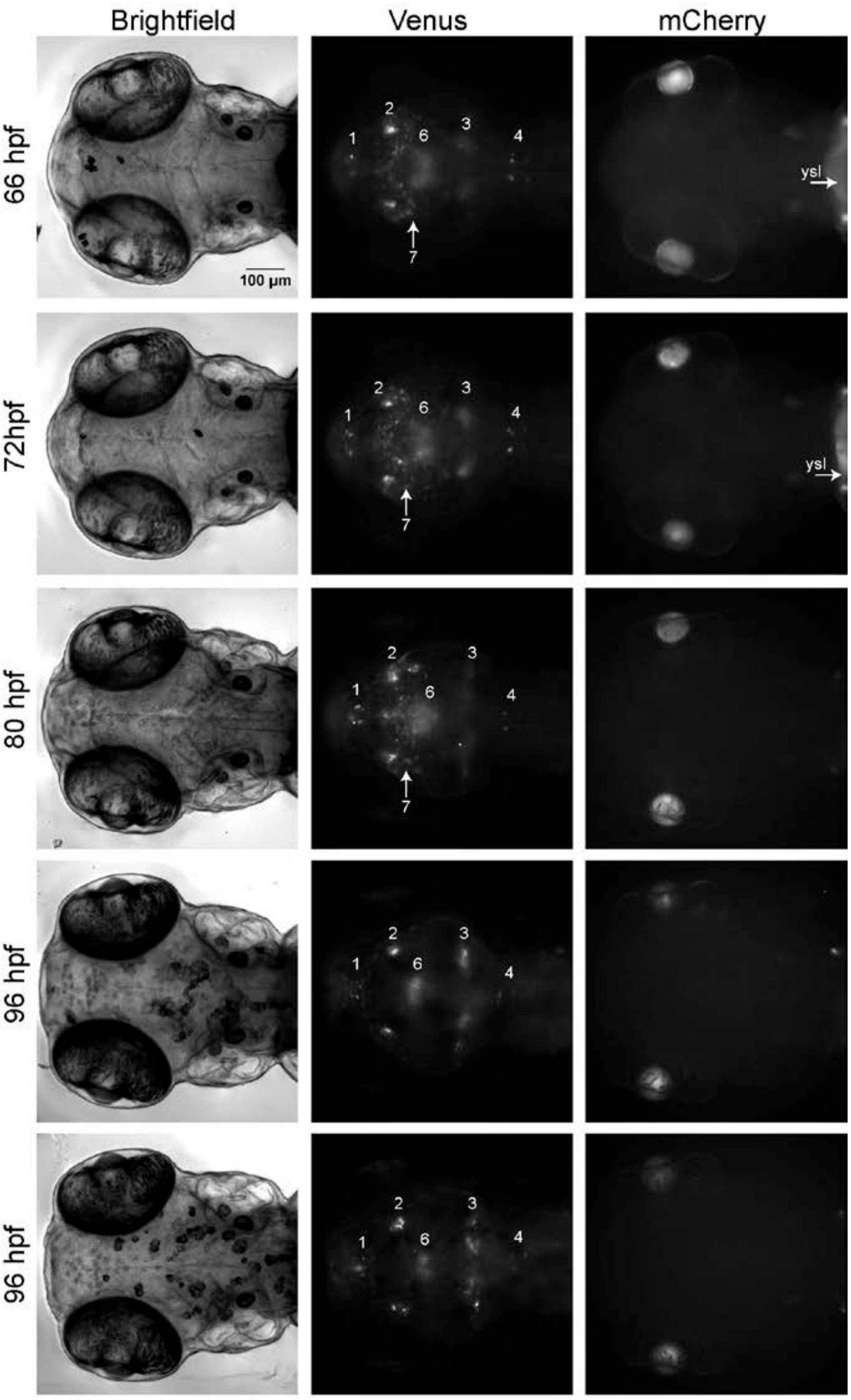


Figure 7.9 EL161 putative enhancer from essrga locus drives a complex and dynamic neuronal pattern.

Complexity and dynamic pattern of pattB-mCherry,Tol2/EL161-krt4:Venus construct integrated in uobL6 IF1 are shown from 60 hpf to 96 hpf. Red lens indicates PhiC31-mediated legitimate integration. YFP channel shows the dynamic expression pattern driven by the injected construct. Domains of expression are listed in **Figure 7.7**. All larvae are oriented anterior to the left. Scale bar indicates 100 µm.

7.3.5 Analysis of the effect of PhiC31 integrase mRNA on embryo and PGC survival

It was recently demonstrated that PhiC31-nos1-3'UTR to PhiC31 mRNA lead to improved embryo survival and germline transmission of recombination events in zebrafish (Hu et al., 2011). We reasoned that the usage of an integrase variant that led to higher survival and similar germline transmission rates could be advantageous for our SNP experiments, where large numbers of transgenic lines would be needed in order to quantify predictably subtle differences in expression pattern and/or intensity. Therefore, I injected uobL6 F3 embryos (the recipient line carrying one single integration site) with both the WT and PhiC31-nos1-3'UTR version of PhiC31 integrase mRNA and analysed the survival rates over 72 hpf (**Table 7.5**).

In accordance with previous findings (Hu et al., 2011), we observed that PhiC31 mRNA *in vitro* transcribed from pCS2+ vector was highly toxic to zebrafish embryos using as little as 15 pg of integrase (**Table 7.5**). We also consistently observed improved survival upon PhiC31-nos1-3'UTR injection when compared to PhiC31 wild type version, while the red lens conversion efficiency remained similar (8.52% n=171 embryos using PhiC31-nos1-3'UTR versus 5.74%, n=100 using the regular PhiC31 mRNA).

Table 7.5: PhiC31 integrase mRNA toxicity rates and embryo survival analysis

Injection solution	Total number of injected embryos	Survival rate at 72 hpf (average % ± st dev)	Number of replicates
15 ng/µL pCS2+PhiC31 integrase RNA	465	25.8 ± 0.15	3
30 ng/µL pCS2+PhiC31 integrase RNA	422	13.8 ± 0.05	3
15 ng/µL pCS2+PhiC31-nos1-3'UTR integrase RNA	385	85.8 ± 0.03	3
30 ng/µL pCS2+PhiC31-nos1-3'UTR integrase RNA	395	74.8 ± 0.03	3

Although the mechanism by which PhiC31 exerts its actions is not known, we thought that if PhiC31-nos1-3'UTR mRNA was differentially more stable in the primordial germ cells (PGCs, (Koprunner et al., 2001)), its toxicity could potentially deplete them and produce a high level of sterility among adults, even though previous studies with PhiC31-nos1-3’UTR had not reported it (Hu et al., 2011). In order to test the viability of PGCs, uobL6 F3 embryos were injected with PhiC31 and PhiC31o-nos1-3’UTR mRNA (15 and 30 ng/µl) and PGCs were stained using anti-vasa antibody (kind gift from Holger Knaut). Vasa protein is restricted to the germ line (Knaut et al., 2000) and therefore it is a suitable marker for the assessment of PGC number and viability. Our results showed that injection of PhiC31-nos1-3’UTR mRNA did not substantially reduce the viability of primordial germ cells (**Table 7.6**, **Figure 7.10**), and therefore there was no indication against using PhiC31o-nos1-3’UTR for further experiments.

Table 7.6 Primordial germ cells survival rates upon PhiC31 integrase injections.

Injection solution	Total number of embryos analysed	Number of PGCs at 24 hpf (average ± st dev)
15 ng/µL pCS2+PhiC31 integrase RNA	21	19.4 ± 4.19
30 ng/µL pCS2+PhiC31 integrase RNA	17	18.5 ± 4.77
15 ng/µL pCS2+PhiC31-nos1-3'UTR integrase RNA	25	19.2 ± 5.55
30 ng/µL pCS2+PhiC31-nos1-3'UTR integrase RNA	12	19.6 ± 5.13
Uninjected group	30	20.3 ± 5.22

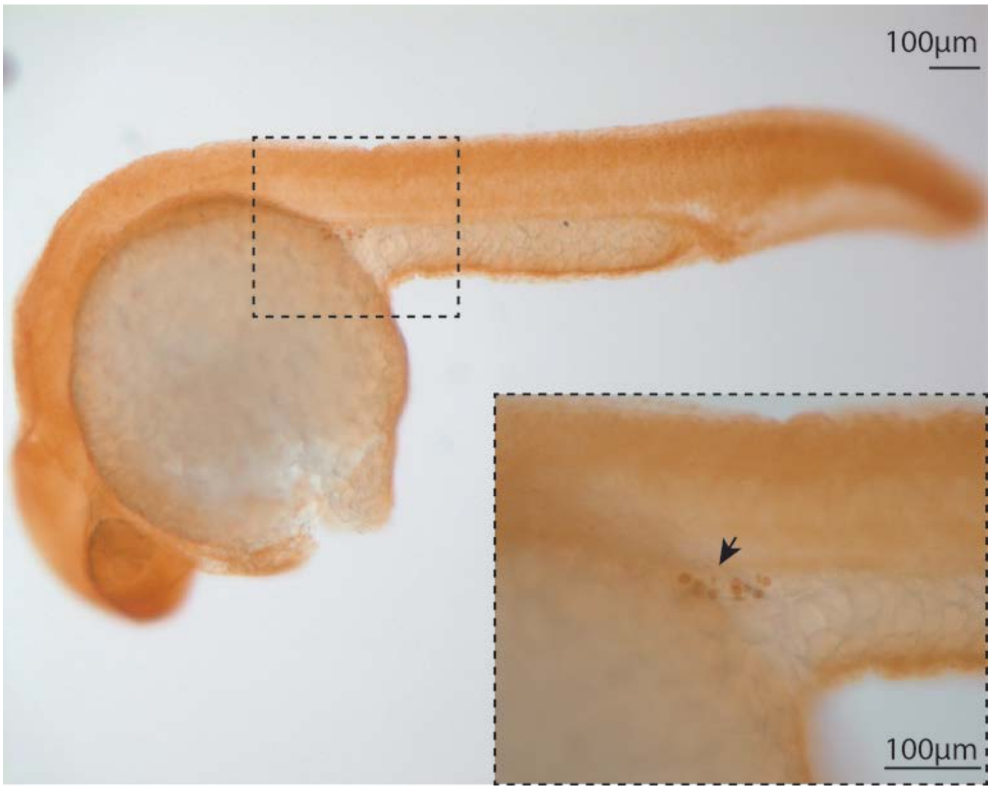


Figure 7.10 Analysis of PGC survival in embryos injected with PhiC31 mRNA
Lateral view of 24 hpf embryo (anterior to the left) stained with anti-vasa antibody (kind gift from Holger Knaut). Magnified view of the genital ridge area (bottom right insert) showing stained PGCs (arrow).

7.3.6 Design of an attB-containing vector to harbour enhancers and measure allelic differences

In order to facilitate the test of multiple enhancers and promoters, a vector containing a multicloning site upstream attB-mCherry targeting cassette in pJET-attB-mCherry plasmid (Roberts et al., 2014) was designed, where *cis*-regulatory elements of interest could be easily exchanged using cloning (**Figure 7.11A**).

A multicloning site was placed upstream attB targeting site and as a reporter gene YFP was chosen, so that there was no spectrum overlap with mCherry reporter gene. This design allows for the insertion of promoters using AgeI/XhoI sites and the cloning of enhancers using EcoRI, EcoRV, SpeI or HindIII sites. For enhancers, in-fusion cloning (see **Chapter 2**) is the preferred option, since only one restriction site from the vector is needed. The expectation is that when this construct containing an enhancer:promoter:YFP cassette upstream attB-mCherry transgene is co-injected with nanos-Phic31 mRNA in uobL6 transgenic recipient line and PhiC31-mediated recombination happens, transgenic fish will exhibit red lens in approximately 8-9% of the cases (see above) together with enhancer pattern driven YFP expression **(Figure 7.11B)**. Previous experience testing this system with wild-type version of PhiC31 mRNA and the zebrafish EL161 enhancer indicated that random integration of the targeting vector in the genome was extremely unlikely to result in red lens co-occurring with enhancer-driven YFP pattern.

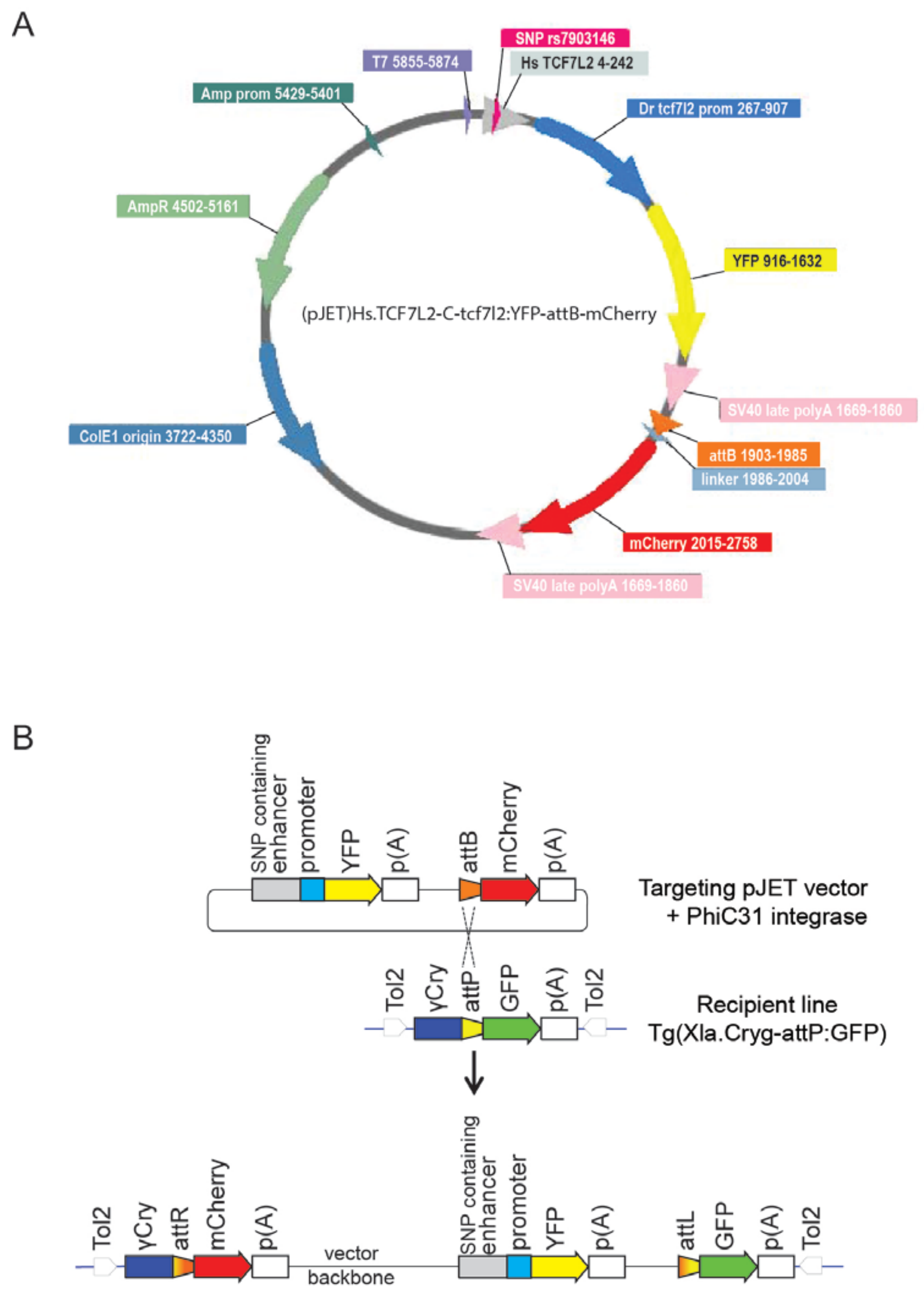


Figure 7.11 Overview of PhiC31 transgenesis system for the detection of SNP variants.
A. Map of the construct designed to test enhancer containing SNP variants using PhiC31-mediated transgenesis system. A vector containing attB-mCherry targeting site, was modified to contain a putative enhancer linked to a promoter and YFP reporter gene. In this example TCF7L2 putative enhancer (Gaulton et al., 2010) containing the C common variant of SNP rs7903146 was cloned upstream endogenous zebrafish *tcf7l2* promoter. **B.** Schematic drawing of the targeting plasmid pJETattB-mCherry,TCF7L2:tfc7l2:YFP co-injected with PhiC31 mRNA into the docking site of Tg(Xla.crygc:attP-GFP) transgenic recipient line and the resultant genomic recombination occurring in injected embryos.

7.3.7 Selection of SNP-containing enhancers to be tested by PhiC31-mediated transgenesis

To test whether PhiC31-mediated transgenesis system could be used for the quantitative analysis of SNP variants, human enhancers containing disease-associated SNPs were cloned into our customised attB-containing construct and the system efficiency was evaluated by establishing transgenic lines. In order to test the system, four sets of human enhancers associated with disease were selected (**Table 7.7**). These putative enhancers contained SNPs linked to type-two diabetes (T2D) and for all of them there is *in vitro* evidence that the SNP variant was causing a significant difference in the enhancer activity by luciferase assays in pancreas cell lines (Gaulton et al., 2010; Pasquali et al., 2014).

Table 7.7. **Type-2-diabetes associated putative enhancers selected for SNP quantitation assays**

Human *Cis*-regulatory element	**Genomic coordinates (Hg18)**	**Disease-associated SNP**	**Allelic variants**	**GWAS Reference**	**Zebrafish promoter used in assays**
TCF7L2	chr10:114748258-114748496	rs7903146	C>T	(Cauchi et al., 2007)	*tcf7l2*
ZFAND3	chr6:37883254-37883831	rs58692659	C>T	(Cho et al., 2012)	*hsp70*
C2CD4A	chr15:60178441-60179347	rs7163757	C>T	(Schaub et al., 2012)	*hsp70*
PROX1	chr1:212242977-212243697	rs4282786	G>A	(Stitzel et al., 2010)	*prox1*

The selection of an appropriate promoter is important for testing enhancer function, as it has been discussed throughout this thesis (**Chapters 3, 4** and **6**). In the case of putative enhancers located in introns (*PROX1* and *TCF7L2*), we decided to take the minimal promoter of the endogenous gene harbouring the enhancer candidates. In the case of putative enhancers located in gene deserts (12 kb upstream the closest coding gene in the case of ZFAND3 enhancer and 27kb downstream the closest coding gene for *C2CD4A* enhancer), we decided to use a minimal promoter *hsp70*, which had been shown to give a minimal to non-existing background in previous analyses (Gehrig et al., 2009; Pasquali et al., 2014).

7.3.8 Transient analyses of attB containing enhancer constructs injected in uobL6

For each tested putative human enhancer tested *in vivo* three constructs were made: two vectors contained the enhancer carrying either the common or the non-common allelic variant of the T2D-associated SNP, linked to the zebrafish promoter and YFP fluorescent protein. Moreover, an enhancer-less (promoter only) control was used to assess the regulatory activity originating from the zebrafish gene promoter (**Table 7.8).**

These constructs were co-injected with 15 ng/µl PhiC31o-nos1-3’UTR mRNA in embryos resulting from outcrosses between uobL6 F3 individuals and AB* WT zebrafish adults. YFP-driven expression pattern and GFP to mCherry lens colour switch was monitored and at 5 dpf, all transgenic fish were grown up (**Table 7.9**). The lack of enhancer driven-YFP expression was initially attributed to the lack of expression shown by this element when tested in Tol2-mediated transgenesis (as discussed in the previous chapter) and due to the PGCs-specific targeting driven by PhiC31o-nos1-3’UTR.

Table 7.8 Primers used to clone human putative enhancers and zebrafish promoters into attB containing vector

***Cis* regulatory element**	**Forward primer (5’-3’)**	**Reverse primer(5’-3’)**	**Element size (Hg18 or Zv9)**
TCF7L2	TTTTGATATCAATTCATGGGCTTTCTCTGC	AAAAACTAGTGTGAAGTGCCCAAGCTTCTC	239 bp
ZFAND3	CATGGAATTCACTAGTTCATGTTTCCCCCGTATGT	AGGCGCCAAAACTAGTCCTGCCCCAAGTTGCACAG	578 bp
C2CD4A	CATGGAATTCACTAGTACATCCCTTACCCTTACTGGA	AGGCGCCAAAACTAGTGGCAATGCGGGCTCTTTTT	907 bp
PROX1	CATGGAATTCACTAGTGCAAAAATGAACTTGAGAAATCC	AGGCGCCAAAACTAGCATTCCCTTTAATATCCCATGC	721 bp
hsp70 zebrafish promoter	AAAAAACCGGTTTGATTGGTCGAACATGCTGG	AAAAACTCGAGCAGTCCGCTCGCTGTCTCGCT	149 bp
prox1 zebrafish promoter	AAAAACCGGTTCCGCACAGAGAACGTATTG	AAAAGTCGACTGAGCTTCTTCGCGATAGTG	423 bp
tcf7l2 zebrafish promoter	CCCCACCGGTTCAGCCTCTTCTGTTTTGAGCAG	TTTTCTCGAGTTTAAGTTTAGGGACTCGCAGTGG	641 bp

7.3.9 Analysis of T2D-associated *TCF7L2* enhancer using PhiC31-mediated transgenesis

Due to time constraints only *TCF7L2* enhancer was tested and included in this thesis. *TCF7L2* putative enhancer was one of the most interesting examples, since regulatory variation within SNP rs7903146 has the strongest link effect to T2D demonstrated so far (Cauchi et al 2007). In order to determine whether we could effectively use PhiC31-based transgenesis in zebrafish to quantitate small differences in either intensity or pattern coming from *TCF7L2*, we decided to analyse the offspring of sexually mature injected founders outcrossed with WT fish. We found at least three transgenic founders per injected construct (**Table 7.9**) giving mosaic offspring with red lens during development (**Table 7.10**), indicating that PhiC31-mediated integration had occurred and that PhiC31o-nos1-3'UTR mRNA variant could efficiently mediate site-specific integration. However, the progeny did not show any enhancer-driven YFP tissue-specific expression (**Figure 7.12**). Suspecting YFP translational problems, we decided to fix 96 hpf embryos showing red-lens conversion and performed WISH using a DIG-labelled YFP antisense probe (**Figure 7.13**), but none of the embryos showed any enhancer driven expression, whilst sibling negative embryos probed with *esrrga* transcription factor showed a clear neuronal pattern that matched the expected endogenous gene activity. These negative results could be attributed to an extreme weakness of the YFP signal caused by the presence of a single copy in the genome. Alternative explanations include that the docking site in uobL6 might not provide the right genomic environment for this enhancer or that zebrafish cells cannot interpret functionally this non-conserved human element; or that this element does not hold regulatory potential.

Table 7.9 Germline transmission rates of PhiC31 founders

Line	Number of lens-GFP + embryos grown up (5 dpf)	Number of screened individuals	Number of red lens positive founders	Germline transmission rates % (n)
Tg(γcry:ATTR-mCherry,Hs.TCF7L2-C-tcf7l2:YFP-ATTR-GFP)	164	54	5	9.3 (5)
Tg(γcry:ATTR-mCherry,Hs.TCF7L2-T-tcf7l2:YFP-ATTR-GFP)	179	22	3	13.6 (3)
Tg(γcrya:ATTR-mCherry,tcf7l2:YFP-ATTR-GFP)	134	40	3	7.5 (3)

Table 7.10 Analysis of the offspring of integrase founders

Line	Founder ID	Total number of F1 embryos analysed	Percentage of red lens embryos %,(n)	Percentage of embryos with enhancer YFP pattern %,(n)
Tg(γcry:ATTR-mCherry,Hs.TCF7L2-C-tcf7l2:YFP-ATTR-GFP)	C9	69	13.0 (9)	0.0 (0)
	C11	108	1.8 (2)	0.0 (0)
	C12	21	4.8 (1)	0.0 (0)
	C18	126	15.9 (20)	0.0 (0)
	C19	58	29.3 (17)	0.0 (0)
Tg(γcry:ATTR-mCherry,Hs.TCF7L2-T-tcf7l2:YFP-ATTR-GFP)	T1	120	5.8 (7)	0.0 (0)
	T3	348	10.9 (38)	0.0 (0)
	T28	126	20.6 (26)	0.0 (0)
Tg(γcrya:ATTR-mCherry,tcf7l2:YFP-ATTR-GFP)	ek1	225	13.3 (30)	0.0 (0)
	ek13	98	16.3 (16)	0.0 (0)
	ek31	234	23.1 (54)	0.0 (0))

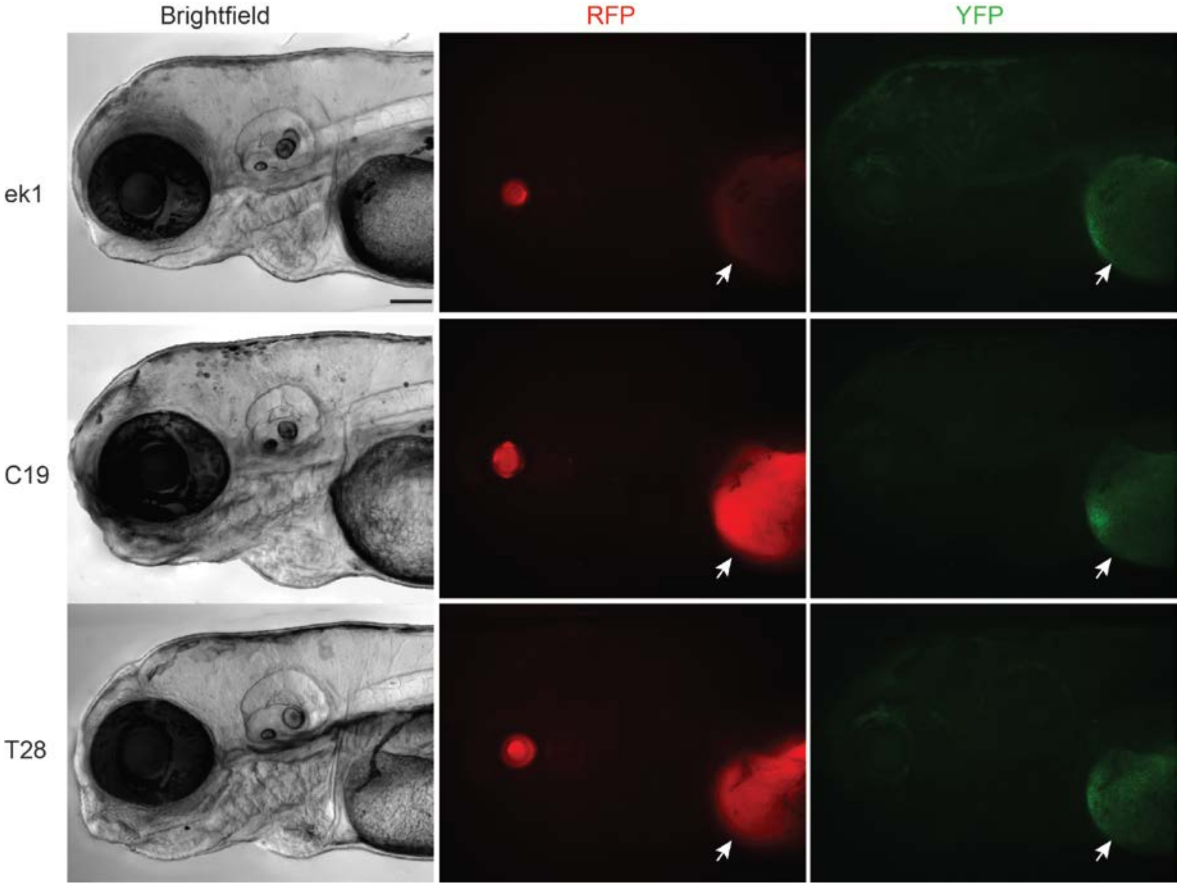


Figure 7.12 PhiC31-mediated integration detected by conversion of reporter expression in *TCF7L2*-containing transgenic lines

96 hpf representative embryos from *TCF7L2* containing transgenic lines. Lens activity driven by mCherry demonstrates PhiC31-mediated integration. Transgenic founders are named as in **Table 7.10.** Arrows in green and red channels indicate auto-fluorescence of the yolk syncytial layer. Lateral views of larvae head to the left are shown. Scale bar represents 100µm.

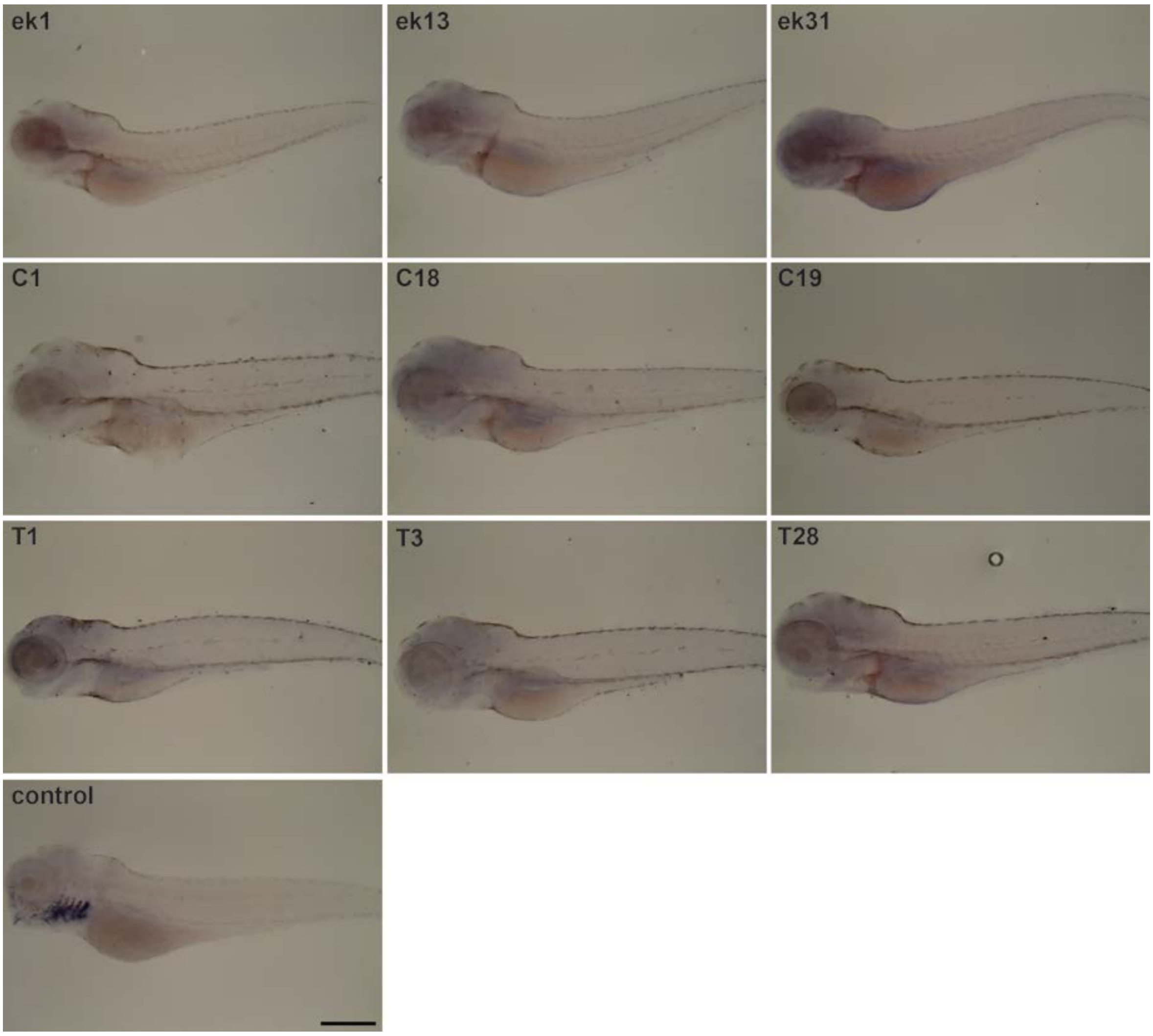


Figure 7.13 Absence of *TCF7L2* enhancer-driven pattern demonstrated by WISH
Representative embryos from each transgenic line probed with DIG-labelled YFP antisense probe. None of the embryos showed any tissue-specific activity despite PhiC31-mediated integration evidenced by red lens conversion. Technical control embryos of Tg(ARE:hsp70:YFP) transgenic line probed with YFP showed the expected jaw-specific expression pattern (bottom left panel). *TCF7L2* transgenic founders are named as in **Table 7.10**. All embryos are 96 hpf, oriented laterally, anterior to the left. Scale bar represents 100µm.

7.3.10 Establishment of new transgenic recipient lines

Due to the finding that uobL12 contained a Tol2-mediated multimer integration, which contradicted previous literature of Tol2 transposon only mediating a single insertion site into the genome (Kawakami et al., 2000), I decided to establish new transgenic recipient lines, widening the choice of landing sites for enhancer testing.

Wild-type embryos were co-injected with 15 ng/µl of pTol2/Xla.crygc:attP-GFP (pDB896) and Tol2 mRNA and positive embryos were grown up to sexual maturity. Then, adults were outcrossed with wild type fish and the offspring was screened for green lens. Out of 10 individuals screened 2 positive founders were identified. EPTS-LMPCR results from one of the lines suggest that there is a single integration site on an intergenic region on chromosome 17 (coordinates for Zv9 are chr17:38,691,215). It is 215 kb downstream *mbip* and 50 kb upstream *sptb* coding genes and could potentially be used for further enhancer assays.

7.3.11 Further perspectives

At the time this thesis was prepared, founder fish injected with *ZFAND3*, *PROX1* and *C2CD4A* enhancers associated with T2D had reached sexual maturity. Outcrosses will be carried out, as described in this chapter for *TCF7L2* and the co-occurrence of YFP enhancer expression and red lens will be used as indicators of PhiC31 mediated integration into the germline. Expectations are that if these human elements can be interpreted by the zebrafish embryo, SNP differences will be quantified by fluoresce imaging, WISH and quantitative PCR of the reporter gene.

7.4 DISCUSSION

In this chapter we report the design of a site-directed integration system mediated by PhiC31 integrase in zebrafish, where docking transgenic lines containing attP sites in known genomic locations can be targeted by attB containing donor vectors that include CREs of interest. By using site-specific transgenesis, we expect to reduce position effect variation and therefore create a more robust system for the detection of subtle differences coming from SNP variants in CREs.

There are examples in the literature of SSR functioning in zebrafish (Boniface et al., 2009) but PhiC31 is particularly efficient because of its directional recombination between heterotypic attB and attP target sites (Thorpe et al., 2000). Using this system reproducible expression pattern driven by EL161 predicted transcriptional enhancer was found, where the observed expression domains appeared to match broadly the endogenous activity of *esrrga*, the predicted target gene. Given that enhancers can act at distances as far as 1 Mb, reviewed by (Krivega and Dean, 2012), unambiguous identification of target genes remains challenging. Nevertheless, several *esrrga* expression domains with reproducible activity in independent genomic integration sites could be identified. Tol2 transgenics also showed several of the enhancer-specific patterns but were mostly coupled to a variety of additional ectopic patterns.

In concordance with previous reports, PhiC31 integrase mRNA can be highly toxic in zebrafish embryos (Hu et al., 2011). We have shown that the use of PhiC31 fused with 3´UTR of *nanos1* can significantly reduce the toxicity after injection (**Table 7.5**), without affecting the recombination rate measured by red lens colour change or the viability of the PGCs. We have also demonstrated germline transmission of site-specific integration of around 15% in adult founders, which is higher than our previous work with codon-optimized PhiC31 integrase (Roberts et al., 2014).

For *TCF7L2* enhancer test we chose *uobL6*, a well characterised landing line that had proven to be an excellent docking site for the analysis of *esrrga* zebrafish enhancer. An additional advantage of a site-directed transgenesis system based on the creation of attP landing sites is that it allows for evaluation of the genomic landscape the transgene was integrated into. However, careful molecular analysis of the docking lines is required to ensure that a single attP containing transgene is harboured. Our transgenic docking lines were created by Tol2 transgenesis and despite having been described to integrate a single copy of Tol2 arms-flanked transgenes (Kawakami et al., 2000), Tol2 independent integration events of full plasmid or fragments are not unheard of and are actually expected to happen in low frequency via NHEJ.

Similar complications during the selection of recipient transgenic lines have been reported by Mosimann and colleagues, who designed a similar site-specific approach in zebrafish based on the creation of several transgenic lines harbouring attP landing sites targeted by attB containing donor Gateway constructs (Mosimann et al., 2013). They characterized three functional Tol2 transgenic lines carrying a single attP site that allows for attB flanked transgene integration. However, the authors encountered difficulties during the characterization process, since one of the lines was defective and another one showed ectopic transgene expression, attributed to enhancer trapping of the attP containing Tol2 transgene (Mosimann et al., 2013).

Nevertheless, such systems provide the basis for creating a library of transgenic docking sites where optimal genomic environments for CREs can be identified and where the effect of genomic locations on different enhancers or enhancer-promoter combinations could be tested. Such efforts have started in *Drosophila*, where around 100 attP transgenic lines have been characterized (Groth et al., 2004; Bateman et al., 2006; Bischof and Basler, 2008; Venken and Bellen, 2012) and attempts to measure position effects in various loci have led to the conclusion that an attP docking site that permits optimal transgene expression in one tissue might not be as effective in another, and that the perfect attP site might not even exist in any vertebrate genome (Markstein et al., 2008).

Despite describing a readily available system consisting of fully characterized transgenic lines carrying single functional attP sites, custom-made attB containing donor vectors that allow easy cloning of enhancers and promoters and highly efficient PhiC31 mRNA, we have not been able to prove *TCF7L2* human enhancer function using this system or have been able to quantitate differences coming from SNP variants within this CRE. We can possibly attribute the negative enhancer results to a suboptimal genomic location of the landing site or to the fact that this non-conserved enhancer does not actually harbour regulatory potential in zebrafish. As discussed in the previous chapter, we have obtained inconclusive results regarding *TCF7L2* enhancer specificity with Tol2-based transgenesis. Although mosaic transient transgenic experiments showed no reporter activity, the fact that islet-specific expression was found among the expression patterns displayed by stable transgenic lines, suggested that this activity was enhancer driven. The fact that SNP rs7903146 is the strongest link to T2D so far also underlines the relevance of this experiment.

To identify optimal sites for CREs, more docking lines need to be created and molecularly characterized, which is both laborious and time consuming. A short term improvement in the current recombination efficiency would require the creation of homozygous attP adult fish, where 100% of the offspring would be bearing the docking site and therefore would double the number of recombinants per injected clutch and reduce labour time by skipping the selection of positive transgenic embryos after injections. Future experiments should also include testing a known human enhancer that is functional in zebrafish and that includes a disease associated SNP. Breeding of the fish already injected with human enhancers associated with T2D (**Table 7.7**) should shed light on the feasibility of this system. Additionally, the application of recent genome editing tools such as TALENs or CRISPRs opens the future to engineered attP site-containing transgenic recipient lines, which would circumvent drawbacks derived from the random integration mediated by commonly used transposons and provide a higher flexibility in choosing the genomic landscape of landing sites.

Chapter Eight: GENERAL DISCUSSION AND PRESPECTIVES

8.1 Using zebrafish transgenics to validate human enhancers

The identification of cis-regulatory elements remains a challenging task due to the lack of distinguishing signatures at the sequence level and our poor understanding of CRE structure and genomic distribution. Recently, the combination of approaches based on cross-species sequence conservation, genome-wide maps of epigenetic marks, binding data of transcriptional co-factors and open chromatin sites, have proven successful in predicting regulatory regions (Visel et al., 2009a; Rada-Iglesias et al., 2011; Djebali et al., 2012). However, computational predictions have led to the identification of large numbers of candidate CREs with unknown function. Thus, the main objective of this thesis was to evaluate the utility of the transgenic zebrafish embryo for detecting the functionality of human regulatory regions predicted by various genome-wide strategies

I studied the function of 5 evolutionarily conserved candidate enhancers defined in human isolated islets by the presence of pancreas-specific TF binding events and active histone modification marks, including H3K4me1 and H3K27ac (**Chapter 3**). Rapid transient transgenesis assays demonstrated that all of the tested candidates were active in the zebrafish embryo, whereas control regions devoid of epigenetic marks failed to drive any tissue-specific expression, arguing for the specificity of the enhancer selection process. Three of the tested CREs showed reproducible reporter expression in the endocrine pancreas, both in transient and stable lines, and two of them in neuronal tissues, broadly reproducing the expression pattern of the predicted zebrafish candidate genes. Overall, these results suggest that there are conserved regulatory codes acting in vertebrate pancreas development, indicating that zebrafish would be a valuable tool to study pancreas *cis*-regulation in humans.

Interestingly, these experiments also showed that enhancers behave differently when coupled to different promoters *in vivo.* Functional test of CRE3-5, which is located in the second intron of *RFX2*, revealed that the *gata2* promoter was more sensitive to regulatory input from the candidate enhancer than *hsp70* and could activate an additional expression pattern in the pancreatic islet of zebrafish; confirming the findings of other studies that have tested putative enhancers with several promoters (Allende et al., 2006; Bessa et al., 2009; Gehrig et al., 2009; Navratilova et al., 2009). In fact, there is a wide variety of promoters used for enhancer tests in the literature, ranging from heterologous minimal promoters, such as EB1 (Li et al., 2010b; Ritter et al., 2010; Ritter et al., 2012) or *hsp70* (Rada-Iglesias et al., 2011), to large gene promoters such as cardiac myosin (Shin et al., 2005) or *gata2* promoters (Bessa et al., 2009; Navratilova et al., 2009; Royo et al., 2011; Royo et al., 2012; Bhatia et al., 2013). Our results and others mentioned here highlight the importance of enhancer-promoter interactions *in vivo* and argue that the promoter choice will influence our capacity to detect enhancer function.

In **Chapter 5** we tested whether bidirectional transcription, detected at enhancer locations by the FANTOM5 Consortium, could be used as a novel enhancer predictive tool, and whether predicted enhancers would be functional in the context of an organism. Transient transgenesis assays demonstrated that 3 out of 5 bidirectionally transcribed enhancers were active in zebrafish, recapitulating the tissue-specificity shown by the human sequences *in vitro.* These results suggested that bidirectional transcription is predictive of active enhancers and that the zebrafish could interpret 60% of the enhancers showing sequence homology. The limited set of enhancers tested prevented us from establishing statistical correlations between sequence conservation and conserved expression. However, there are a few studies in the literature that have analysed this in detail by testing the same candidate enhancer in more than one model and orthologous enhancers in the same species (Navratilova et al., 2009; Ritter et al., 2010). To distinguish between cis- and trans-regulatory changes Ritter and colleagues analysed human CREs tested in zebrafish (HZ), human CREs tested in mouse (HM) and orthologous conserved zebrafish CREs tested in zebrafish (ZZ). Their results indicated that cis-regulatory changes, which are caused when function was discordant between HZ and ZZ tests, are two times more frequent than trans-regulatory changes, which happen when function was disparate between HZ and HM experiments (Ritter et al., 2010).

We also used the zebrafish embryo to test whether an inherited region downstream of *PTF1A* in patients with pancreas agenesis could act as a long-range CRE, and whether the potentially pathogenic variants found in the probands could affect the regulatory activity of the candidate enhancer (**Chapter 4**). Transient transgenesis assays were unable to detect any tissue-specific activity in the zebrafish embryo, suggesting that either lack of sequence homology, enhancer-promoter interaction, or enhancer function *in vivo,* could be the reason for the lack of reporter activity in zebrafish. Similarly negative results were obtained when we tested T2D-associated enhancers that were not conserved at a sequence level in zebrafish but were enriched in key histone modification marks and bound by pancreas-specific TFs (**Chapter 6**), suggesting that prediction tools require further refinement and that further studies are needed to understand the limitations of zebrafish to test human enhancers.

8.2 Correlation between enhancer structure and predicted function

There are several characteristics that could aid in predicting enhancer activity, such as the degree of sequence conservation, their length, the strength of the predictors used in the computational analyses or the presence of TFBS.

The length of the elements tested in this thesis was on average 1 kb, in accordance with similar studies in the literature (Shin et al., 2005; Abbasi et al., 2007; Ritter et al., 2010; Rada-Iglesias et al., 2011; Ritter et al., 2012). We argued that longer sequences could represent the full functional element required for autonomous activity. In agreement with this, recent studies using bimodal enrichment of H3K27ac and H3K4me1 histone tails as a predictive tool indicated that enhancers are in average 3 nucleosomes in length, that is, approximately 600 bp (**Figure 5.1**, (Bernstein et al., 2012). However, despite the candidate enhancers tested had comparable lengths, not all of them were able to modulate reporter gene expression, suggesting that the chosen length may not reflect the true size of the enhancer.

The strength of predictors used, such as enrichment in histone modification marks and bidirectional transcription, were comparable within groups, as demonstrated by the robust computational analysis performed by our collaborators. It could be argued that TFBS composition could correlate with enhancer function; however, it was out of the scope of this thesis to perform an in-depth analysis of TFBS in the tested candidate enhancers. Due to the lack of appropriate zebrafish antibodies there is limited experimental ChIP-Seq data on the specific binding of TFs of interest. In addition to that, certain computational predictions of TFs have been shown to be poorly correlated with *in vivo* measurements (Kaplan et al., 2011), and discerning whether predicted binding is functional is not trivial, suggesting that this approach would be unreliable.

Interestingly, sequence conservation seems to be strongly correlated with enhancer function in our dataset (**Table 8.1**). In this study, the activity of 27 predicted enhancers and enhancer variants was evaluated, along with 5 control regions. Among the candidate enhancers, 15 were conserved at the sequence level (using as criteria 70% of conservation over 100 bp, **Chapters 3, 5, 6**) and 12 were not conserved (**Chapter 4**, **Chapter 6**). Our transient transgenic analyses showed reproducible reporter expression in 10 out of the 15 the conserved enhancers, whereas none of the 12 non-conserved enhancer variants could drive tissue-specific activity (**Table 8.1**). These results suggest that sequence conservation has a strong influence on functionality, in keeping

Table 8.1 Overview of human candidate enhancers tested in this thesis

Human enhancer ID	**Closest coding gene(s)**	**Conserved with zebrafish**	**Reproducible enhancer-driven pattern**		**Enriched in H3K4me1 and/or H3K27ac**	**Bidirectionally transcribed**
			Transient assays	**Transgenic lines**		
C3-1	*AGPAT9* and *NKX6-1*	**Yes**	**Yes**	-	**Yes**	-
C3-3	*ISL1* and *PELO*	**Yes**	**Yes**	3/3	**Yes**	-
C3-4	*PROX1* and *SMYD2*	**Yes**	**Yes**	1/1	**Yes**	-
C3-5	*LOC100128568* and *RFX2*	**Yes**	**Yes**	5/5	**Yes**	-
C3-6	*TCF7L2* and *HABP2*	**Yes**	**Yes**	-	**Yes**	-
C2-11	*KIF26B* and *SMYD3*	No	No	-	No	-
C5-14	*LINC00499* and *CCRN4L*	No	No	-	No	-
PROX1- A	*PROX1*	**Yes**	**Yes**	0/4	**Yes**	-
PROX1-G	*PROX1*	**Yes**	**Yes**	0/4	**Yes**	-
TCF7L2- C	*TCF7L2*	No	No	0/5	**Yes**	-
TCF7L2-T	*TCF7L2*	No	No	0/1	**Yes**	-
DGKB-TMEM195-G	*DGKB* and *TMEM195*	No	No	-	**Yes**	-
DGKB-TMEM195-T	*DGKB* and *TMEM195*	No	No	-	**Yes**	-
ADCY5-A	*ADCY5*	No	No	-	**Yes**	-
ADCY5-G	*ADCY5*	No	No	-	**Yes**	-
PTF1A-short-WT	*PTF1A*	No	No	-	**Yes**	-
PTF1A-short-363G	*PTF1A*	No	No	-	**Yes**	-
PTF1A-short-305G	*PTF1A*	No	No	-	**Yes**	-
PTF1A-long-WT	*PTF1A*	No	No	-	**Yes**	-
PTF1A-long-363G	*PTF1A*	No	No	-	**Yes**	-
PTF1A-long-305G	*PTF1A*	No	No	-	**Yes**	-
CRE1	*TMEM161b* and *MEF2C*	**Yes**	**Yes**	5/5	n/a	**Yes**
CRE2	*POU3F2*	**Yes**	**Yes**	-	n/a	**Yes**
CRE3	*SOX7*	**Yes**	**Yes**	0/2	n/a	**Yes**
CRE4	*PAX6* and RCN1	**Yes**	No	-	n/a	**Yes**
CRE5	*DLX1* and DLX2	**Yes**	No	-	n/a	**Yes**
ubi1	*GTF2A*	**Yes**	No	-	**Yes**	**Yes**
ubi2	*GPR84*	**Yes**	No	-	**Yes**	**Yes**
ubi3	*ZBTB16*	**Yes**	No	-	**Yes**	**Yes**
Control region 1	*C11orf74*	No	No	-	No	No
Control region 2	*DSCAM*	No	No	-	No	No
Control region 3	*PBX1*	No	No	-	No	No

with studies that have argued that sequence conservation is relevant to enhancer function (de la Calle-Mustienes et al., 2005; Shin et al., 2005; Woolfe et al., 2005; Abbasi et al., 2007; Navratilova et al., 2009; Li et al., 2010b; Punnamoottil et al., 2010; Ritter et al., 2010; Chatterjee et al., 2011; Ritter et al., 2012). Nevertheless, our observations contrast with other publications where non-conserved human CREs maintained functionality in zebrafish (Fisher et al., 2006a; Rada-Iglesias et al., 2011); and with reports that found a comparable level of sequence conservation between enhancers that showed specific activity and those that did not (Li et al., 2010b). Along these lines, Ritter and colleagues concluded that a higher degree of conservation is not a better indicator of conserved function, whereas TFBS distribution is (Ritter et al., 2010).

8.3 Using zebrafish transgenesis to quantitate enhancer variants

Several studies have demonstrated that common SNP variants can lead to phenotypical consequences and increased susceptibility to human disorders (Lettice et al., 2003; Rahimov et al., 2008; Pomerantz et al., 2009; Musunuru et al., 2010; Harismendy et al., 2011; Smemo et al., 2014). Nevertheless, most studies have revealed subtle expression differences between enhancer variants *in vitro* (Gaulton et al., 2010; Stitzel et al., 2010; Rada-Iglesias et al., 2013), and it remained largely unexplored whether these variants would elicit detectable changes in the spatial or temporal activity of enhancers *in vivo*. Therefore, a second aim of this thesis was to investigate if zebrafish transgenesis could be used to detect change in enhancer function caused by sequence variation. Thus, we attempted to quantitate T2D-associated enhancer allelic variants *in vivo* using Tol2 transgenesis (**Chapter 6**). Our results showed that two non-conserved candidate enhancers were not functional in zebrafish. Furthermore, it was not possible to determine unambiguously the function of *TCF7L2* enhancer variants, due to the lack of activity in transient transgenic assays and lack of reproducibility in transgenic lines (**Table 8.1**). Along these lines, the variability of position effects in the transgenic lines containing *PROX1* enhancer hindered quantitation of enhancer variants using transposon-based transgenesis (**Chapter 6**). These results contrast with a recent study that has used Tol2 transgenesis to demonstrate that a risk SNP can alter enhancer function substantially (Spieler et al., 2014), suggesting that certain enhancer might be more prone to position effects and their activity may only be detected in a system that controls the genomic environment of the integration site.

In order to improve the reliability of zebrafish transgenesis, I contributed significantly to the development of a more refined site-directed transgenesis system mediated by PhiC31 integrase. This technology was validated for enhancer testing (**Chapter 7**) and proved to reduce position effect variation commonly found in conventional, transposon-based transgenesis experiments. Nevertheless, when the system was used to quantify *TCF7L2*-associated enhancer variants, enhancer-driven expression pattern could not be detected. These negative results may be explained by lack of sequence homology, by a suboptimal genomic or epigenomic environment of the integration site, because the tested element does not represent the full autonomous enhancer, or because it is not transcriptionally active in zebrafish. Genomic influence cannot be completely avoided using PhiC31-based transgenesis, since the recipient transgenic lines are created by randomly integrating the attP containing transgene into the genome. However, the creation of libraries of recipient lines represents the first step towards a more robust transgenesis system for enhancer tests. These results also indicate the need for further studies to understand the limitations of zebrafish model in functional analysis of human enhancers.

Overall, by utilizing zebrafish in novel projects of enhancer prediction and aiding in the development of a site-directed transgenesis method that can reduce position effect variability, the results presented in this thesis contribute to the establishment of the zebrafish embryo as a valuable model to test the function of predicted human enhancers.

Chapter Nine: REFERENCES

Abbasi, A. A., Paparidis, Z., Malik, S., Goode, D. K., Callaway, H., Elgar, G. and Grzeschik, K. H. (2007) 'Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers', *PLoS One* 2(4): e366.

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. and McVean, G. A. (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature* 491(7422): 56-65.

Adachi, Y., Hauck, B., Clements, J., Kawauchi, H., Kurusu, M., Totani, Y., Kang, Y. Y., Eggert, T., Walldorf, U., Furukubo-Tokunaga, K. et al. (2003) 'Conserved cis-regulatory modules mediate complex neural expression patterns of the eyeless gene in the Drosophila brain', *Mech Dev* 120(10): 1113-26.

Adams, C. C. and Workman, J. L. (1995) 'Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative', *Mol Cell Biol* 15(3): 1405-21.

Aday, A. W., Zhu, L. J., Lakshmanan, A., Wang, J. and Lawson, N. D. (2011) 'Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites', *Dev Biol* 357(2): 450-62.

Aerts, S. (2012) 'Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets', *Curr Top Dev Biol* 98: 121-45.

Ahituv, N., Rubin, E. M. and Nobrega, M. A. (2004) 'Exploiting human--fish genome comparisons for deciphering gene regulation', *Hum Mol Genet* 13 Spec No 2: R261-6.

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): 257-60.

Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J. C., Suzuki, H., Daub, C. O., Hayashizaki, Y. and Lenhard, B. (2009) 'Transcriptional features of genomic regulatory blocks', *Genome Biol* 10(4): R38.

Al-Shammari, M., Al-Husain, M., Al-Kharfy, T. and Alkuraya, F. S. (2011) 'A novel PTF1A mutation in a patient with severe pancreatic and cerebellar involvement', *Clin Genet* 80(2): 196-8.

Allan, C. M., Walker, D. and Taylor, J. M. (1995) 'Evolutionary duplication of a hepatic control region in the human apolipoprotein E gene locus. Identification of a second region that confers high level and liver-specific expression of the human apolipoprotein E gene in transgenic mice', *J Biol Chem* 270(44): 26278-81.

Allen, B. G. and Weeks, D. L. (2005) 'Transgenic Xenopus laevis embryos can be generated using phiC31 integrase', *Nat Methods* 2(12): 975-9.

Allen, H. L., Flanagan, S. E., Shaw-Smith, C., De Franco, E., Akerman, I., Caswell, R., Ferrer, J., Hattersley, A. T. and Ellard, S. (2012) 'GATA6 haploinsufficiency causes pancreatic agenesis in humans', *Nat Genet* 44(1): 20-2.

Allende, M. L., Manzanares, M., Tena, J. J., Feijoo, C. G. and Gomez-Skarmeta, J. L. (2006) 'Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos', *Methods* 39(3): 212-9.

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Bonnen, P. E., de Bakker, P. I., Deloukas, P., Gabriel, S. B. et al. (2010) 'Integrating common and rare genetic variation in diverse human populations', *Nature* 467(7311): 52-8.

Amsterdam, A., Burgess, S., Golling, G., Chen, W., Sun, Z., Townsend, K., Farrington, S., Haldi, M. and Hopkins, N. (1999) 'A large-scale insertional mutagenesis screen in zebrafish', *Genes Dev* 13(20): 2713-24.

Andersson, R. Gebhard, C. Miguel-Escalada, I. Hoof, I. Bornholdt, J. Boyd, M. Chen, Y. Zhao, X. Schmidl, C. Suzuki, T. et al. (2014) 'An atlas of active enhancers across human cell types and tissues', *Nature* 507(7493): 455-61.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. et al. (2002) 'Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes', *Science* 297(5585): 1301-10.

Argenton, F., Zecchin, E. and Bortolussi, M. (1999) 'Early appearance of pancreatic hormone-expressing cells in the zebrafish embryo', *Mech Dev* 87(1-2): 217-21.

Arnosti, D. N., Barolo, S., Levine, M. and Small, S. (1996) 'The eve stripe 2 enhancer employs multiple modes of transcriptional synergy', *Development* 122(1): 205-14.

Arnosti, D. N. and Kulkarni, M. M. (2005) 'Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?', *J Cell Biochem* 94(5): 890-8.

Atouf, F., Czernichow, P. and Scharfmann, R. (1997) 'Expression of neuronal traits in pancreatic beta cells. Implication of neuron-restrictive silencing factor/repressor element silencing transcription factor, a neuron-restrictive silencer', *J Biol Chem* 272(3): 1929-34.

Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G. and Pfeifer, D. (2001) 'Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions', *Genomics* 78(1-2): 73-82.

Balasubramanian, M., Shield, J. P., Acerini, C. L., Walker, J., Ellard, S., Marchand, M., Polak, M., Vaxillaire, M., Crolla, J. A., Bunyan, D. J. et al. (2010) 'Pancreatic hypoplasia presenting with neonatal diabetes mellitus in association with congenital heart defect and developmental delay', *Am J Med Genet A* 152A(2): 340-6.

Balciunas, D., Wangensteen, K. J., Wilber, A., Bell, J., Geurts, A., Sivasubbu, S., Wang, X., Hackett, P. B., Largaespada, D. A., McIvor, R. S. et al. (2006) 'Harnessing a high cargo-capacity transposon for genetic applications in vertebrates', *PLoS Genet* 2(11): e169.

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. (1996) 'The CBP co-activator is a histone acetyltransferase', *Nature* 384(6610): 641-3.

Bardet, P. L., Obrecht-Pflumio, S., Thisse, C., Laudet, V., Thisse, B. and Vanacker, J. M. (2004) 'Cloning and developmental expression of five estrogen-receptor related genes in the zebrafish', *Dev Genes Evol* 214(5): 240-9.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) 'High-resolution profiling of histone methylations in the human genome', *Cell* 129(4): 823-37.

Bartfai, R., Balduf, C., Hilton, T., Rathmann, Y., Hadzhiev, Y., Tora, L., Orban, L. and Muller, F. (2004) 'TBP2, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish', *Curr Biol* 14(7): 593-8.

Bateman, J. R., Lee, A. M. and Wu, C. T. (2006) 'Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange', *Genetics* 173(2): 769-77.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. and Haussler, D. (2004) 'Ultraconserved elements in the human genome', *Science* 304(5675): 1321-5.

Belmont, A. S., Sedat, J. W. and Agard, D. A. (1987) 'A three-dimensional approach to mitotic chromosome structure: evidence for a complex hierarchical organization', *J Cell Biol* 105(1): 77-92.

Belteki, G., Gertsenstein, M., Ow, D. W. and Nagy, A. (2003) 'Site-specific cassette exchange and germline transmission with mouse ES cells expressing phiC31 integrase', *Nat Biotechnol* 21(3): 321-4.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. and Snyder, M. (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature* 489(7414): 57-74.

Bessa, J., Tena, J. J., de la Calle-Mustienes, E., Fernandez-Minan, A., Naranjo, S., Fernandez, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F. et al. (2009) 'Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish', *Dev Dyn* 238(9): 2409-17.

Bhandare, R., Schug, J., Le Lay, J., Fox, A., Smirnova, O., Liu, C., Naji, A. and Kaestner, K. H. (2010) 'Genome-wide analysis of histone modifications in human pancreatic islets', *Genome Res* 20(4): 428-33.

Bhatia, S., Bengani, H., Fish, M., Brown, A., Divizia, M. T., de Marco, R., Damante, G., Grainger, R., van Heyningen, V. and Kleinjan, D. A. (2013) 'Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia', *Am J Hum Genet* 93(6): 1126-34.

Biemar, F., Argenton, F., Schmidtke, R., Epperlein, S., Peers, B. and Driever, W. (2001) 'Pancreas development in zebrafish: early dispersed appearance of endocrine hormone expressing cells and their convergence to form the definitive islet', *Dev Biol* 230(2): 189-203.

Billings, L. K. and Florez, J. C. (2010) 'The genetics of type 2 diabetes: what have we learned from GWAS?', *Ann N Y Acad Sci* 1212: 59-77.

Birnbaum, R. Y., Clowney, E. J., Agamy, O., Kim, M. J., Zhao, J., Yamanaka, T., Pappalardo, Z., Clarke, S. L., Wenger, A. M., Nguyen, L. et al. (2012) 'Coding exons function as tissue-specific enhancers of nearby genes', *Genome Res* 22(6): 1059-68.

Birnboim, H. C. and Doly, J. (1979) 'A rapid alkaline extraction procedure for screening recombinant plasmid DNA', *Nucleic Acids Res* 7(6): 1513-23.

Birney, E. Stamatoyannopoulos, J. A. Dutta, A. Guigo, R. Gingeras, T. R. Margulies, E. H. Weng, Z. Snyder, M. Dermitzakis, E. T. Thurman, R. E. et al. (2007) 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature* 447(7146): 799-816.

Bischof, J. and Basler, K. (2008) 'Recombinases and their use in gene activation, gene inactivation, and transgenesis', *Methods Mol Biol* 420: 175-95.

Blackwood, E. M. and Kadonaga, J. T. (1998) 'Going the distance: a current view of enhancer action', *Science* 281(5373): 60-3.

Blechinger, S. R., Evans, T. G., Tang, P. T., Kuwada, J. Y., Warren, J. T., Jr. and Krone, P. H. (2002) 'The heat-inducible zebrafish hsp70 gene is expressed during normal lens development under non-stress conditions', *Mech Dev* 112(1-2): 213-5.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. (2010) 'ChIP-Seq identification of weakly conserved heart enhancers', *Nat Genet* 42(9): 806-10.

Boffelli, D., Nobrega, M. A. and Rubin, E. M. (2004) 'Comparative genomics at the vertebrate extremes', *Nat Rev Genet* 5(6): 456-65.

Boniface, E. J., Lu, J., Victoroff, T., Zhu, M. and Chen, W. (2009) 'FlEx-based transgenic reporter lines for visualization of Cre and Flp activity in live zebrafish', *Genesis* 47(7): 484-91.

Bossard, P. and Zaret, K. S. (1998) 'GATA transcription factors as potentiators of gut endoderm differentiation', *Development* 125(24): 4909-17.

Bramswig, N. C. and Kaestner, K. H. (2014) 'Transcriptional and epigenetic regulation in human islets', *Diabetologia* 57(3): 451-4.

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): 41-8.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S. (1993) 'Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome', *Nature* 366(6452): 265-8.

Burke, T. W. and Kadonaga, J. T. (1996) 'Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters', *Genes Dev* 10(6): 711-24.

Burke, T. W. and Kadonaga, J. T. (1997) 'The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila', *Genes Dev* 11(22): 3020-31.

Burke, Z. and Oliver, G. (2002) 'Prox1 is an early specific marker for the developing liver and pancreas in the mammalian foregut endoderm', *Mech Dev* 118(1-2): 147-55.

Burlison, J. S., Long, Q., Fujitani, Y., Wright, C. V. and Magnuson, M. A. (2008) 'Pdx-1 and Ptf1a concurrently determine fate specification of pancreatic multipotent progenitor cells', *Dev Biol* 316(1): 74-86.

Butler, J. E. and Kadonaga, J. T. (2001) 'Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs', *Genes Dev* 15(19): 2515-9.

Butler, J. E. and Kadonaga, J. T. (2002) 'The RNA polymerase II core promoter: a key component in the regulation of gene expression', *Genes Dev* 16(20): 2583-92.

Calhoun, V. C., Stathopoulos, A. and Levine, M. (2002) 'Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex', *Proc Natl Acad Sci U S A* 99(14): 9243-7.

Calo, E. and Wysocka, J. (2013) 'Modification of enhancer chromatin: what, how, and why?', *Mol Cell* 49(5): 825-37.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C. et al. (2006) 'Genome-wide analysis of mammalian promoter architecture and evolution', *Nat Genet* 38(6): 626-35.

Cauchi, S., El Achhab, Y., Choquet, H., Dina, C., Krempler, F., Weitgasser, R., Nejjari, C., Patsch, W., Chikri, M., Meyre, D. et al. (2007) 'TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis', *J Mol Med (Berl)* 85(7): 777-82.

Chalkley, G. E. and Verrijzer, C. P. (1999) 'DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator', *EMBO J* 18(17): 4835-45.

Chao, C. H., Wang, H. D. and Yuh, C. H. (2010) 'Complexity of cis-regulatory organization of six3a during forebrain and eye development in zebrafish', *BMC Dev Biol* 10: 35.

Chatterjee, S., Bourque, G. and Lufkin, T. (2011) 'Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes', *BMC Dev Biol* 11: 63.

Chen, L., Magliano, D. J. and Zimmet, P. Z. (2012) 'The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives', *Nat Rev Endocrinol* 8(4): 228-36.

Chen, R., Hussain, K., Al-Ali, M., Dattani, M. T., Hindmarsh, P., Jones, P. M. and Marsh, P. (2008) 'Neonatal and late-onset diabetes mellitus caused by failure of pancreatic development: report of 4 more cases and a review of the literature', *Pediatrics* 121(6): e1541-7.

Cho, Y. S., Chen, C. H., Hu, C., Long, J., Ong, R. T., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T. et al. (2012) 'Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians', *Nat Genet* 44(1): 67-72.

Chomczynski, P. and Sacchi, N. (1987) 'Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction', *Anal Biochem* 162(1): 156-9.

Chrivia, J. C., Kwok, R. P., Lamb, N., Hagiwara, M., Montminy, M. R. and Goodman, R. H. (1993) 'Phosphorylated CREB binds specifically to the nuclear protein CBP', *Nature* 365(6449): 855-9.

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M. and Zaret, K. S. (2002) 'Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4', *Mol Cell* 9(2): 279-89.

Clark, K. L., Halay, E. D., Lai, E. and Burley, S. K. (1993) 'Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5', *Nature* 364(6436): 412-20.

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J. and Nickerson, D. A. (2010) 'Single-nucleotide evolutionary constraint scores highlight disease-causing mutations', *Nat Methods* 7(4): 250-1.

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 Res* 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): 1845-8.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A. et al. (2010) 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proc Natl Acad Sci U S A* 107(50): 21931-6.

Crowley, T. E., Hoey, T., Liu, J. K., Jan, Y. N., Jan, L. Y. and Tjian, R. (1993) 'A new factor related to TATA-binding protein has highly restricted expression patterns in Drosophila', *Nature* 361(6412): 557-61.

Cui, Z., Geurts, A. M., Liu, G., Kaufman, C. D. and Hackett, P. B. (2002) 'Structure-function analysis of the inverted terminal repeats of the sleeping beauty transposon', *J Mol Biol* 318(5): 1221-35.

Dahm, R. and Geisler, R. (2006) 'Learning from small fry: the zebrafish as a genetic model organism for aquaculture fish species', *Mar Biotechnol (NY)* 8(4): 329-45.

Dantonel, J. C., Wurtz, J. M., Poch, O., Moras, D. and Tora, L. (1999) 'The TBP-like factor: an alternative transcription factor in metazoa?', *Trends Biochem Sci* 24(9): 335-9.

Davidson, A. E., Balciunas, D., Mohn, D., Shaffer, J., Hermanson, S., Sivasubbu, S., Cliff, M. P., Hackett, P. B. and Ekker, S. C. (2003) 'Efficient gene delivery and gene expression in zebrafish using the Sleeping Beauty transposon', *Dev Biol* 263(2): 191-202.

De Franco, E., Shaw-Smith, C., Flanagan, S. E., Shepherd, M. H., Hattersley, A. T. and Ellard, S. (2013) 'GATA6 mutations cause a broad phenotypic spectrum of diabetes from pancreatic agenesis to adult-onset diabetes without exocrine insufficiency', *Diabetes* 62(3): 993-7.

de la Calle-Mustienes, E., Feijoo, C. G., Manzanares, M., Tena, J. J., Rodriguez-Seguel, E., Letizia, A., Allende, M. L. and Gomez-Skarmeta, J. L. (2005) 'A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts', *Genome Res* 15(8): 1061-72.

de Laat, W. and Duboule, D. (2013) 'Topology of mammalian developmental enhancers and their regulatory landscapes', *Nature* 502(7472): 499-506.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C. L. and Natoli, G. (2010) 'A large fraction of extragenic RNA pol II transcription sites overlap enhancers', *PLoS Biol* 8(5): e1000384.

de Wit, E. and de Laat, W. (2012) 'A decade of 3C technologies: insights into nuclear organization', *Genes Dev* 26(1): 11-24.

Deato, M. D. and Tjian, R. (2007) 'Switching of the core transcription machinery during myogenesis', *Genes Dev* 21(17): 2137-49.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) 'Capturing chromosome conformation', *Science* 295(5558): 1306-11.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., Dean, A. and Blobel, G. A. (2012) 'Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor', *Cell* 149(6): 1233-44.

Dermitzakis, E. T., Kirkness, E., Schwarz, S., Birney, E., Reymond, A. and Antonarakis, S. E. (2004) 'Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment', *Genome Res* 14(5): 852-9.

Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B. J., Flegel, V., Bucher, P., Jongeneel, C. V. et al. (2002) 'Numerous potentially functional but non-genic conserved sequences on human chromosome 21', *Nature* 420(6915): 578-82.

Dickmeis, T., Plessy, C., Rastegar, S., Aanstad, P., Herwig, R., Chalmel, F., Fischer, N. and Strahle, U. (2004) 'Expression profiling and comparative genomics identify a conserved regulatory region controlling midline expression in the zebrafish embryo', *Genome Res* 14(2): 228-38.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature* 485(7398): 376-80.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. et al. (2012) 'Landscape of transcription in human cells', *Nature* 489(7414): 101-8.

Dostie, J. and Bickmore, W. A. (2012) 'Chromosome organization in the nucleus - charting new territory across the Hi-Cs', *Curr Opin Genet Dev* 22(2): 125-31.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C. et al. (2006) 'Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements', *Genome Res* 16(10): 1299-309.

Douziech, M., Coin, F., Chipoulet, J. M., Arai, Y., Ohkuma, Y., Egly, J. M. and Coulombe, B. (2000) 'Mechanism of promoter melting by the xeroderma pigmentosum complementation group B helicase of transcription factor IIH revealed by protein-DNA photo-cross-linking', *Mol Cell Biol* 20(21): 8168-77.

Doyle, H. J., Kraut, R. and Levine, M. (1989) 'Spatial regulation of zerknullt: a dorsal-ventral patterning gene in Drosophila', *Genes Dev* 3(10): 1518-33.

Driever, W., Solnica-Krezel, L., Schier, A. F., Neuhauss, S. C., Malicki, J., Stemple, D. L., Stainier, D. Y., Zwartkruis, F., Abdelilah, S., Rangini, Z. et al. (1996) 'A genetic screen for mutations affecting embryogenesis in zebrafish', *Development* 123: 37-46.

Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M. and Frazer, K. A. (2000) 'Active conservation of noncoding sequences revealed by three-way species comparisons', *Genome Res* 10(9): 1304-6.

Dupuis, J. Langenberg, C. Prokopenko, I. Saxena, R. Soranzo, N. Jackson, A. U. Wheeler, E. Glazer, N. L. Bouatia-Naji, N. Gloyn, A. L. et al. (2010) 'New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk', *Nat Genet* 42(2): 105-16.

Dynan, W. S. (1989) 'Modularity in promoters and enhancers', *Cell* 58(1): 1-4.

Eckner, R., Ewen, M. E., Newsome, D., Gerdes, M., DeCaprio, J. A., Lawrence, J. B. and Livingston, D. M. (1994) 'Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor', *Genes Dev* 8(8): 869-84.

Eckner, R., Yao, T. P., Oldread, E. and Livingston, D. M. (1996) 'Interaction and functional collaboration of p300/CBP and bHLH proteins in muscle and B-cell differentiation', *Genes Dev* 10(19): 2478-90.

Emison, E. S., McCallion, A. S., Kashuk, C. S., Bush, R. T., Grice, E., Lin, S., Portnoy, M. E., Cutler, D. J., Green, E. D. and Chakravarti, A. (2005) 'A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk', *Nature* 434(7035): 857-63.

Epstein, D. J. (2009) 'Cis-regulatory mutations in human disease', *Brief Funct Genomic Proteomic* 8(4): 310-6.

Estella, C., McKay, D. J. and Mann, R. S. (2008) 'Molecular integration of wingless, decapentaplegic, and autoregulatory inputs into Distalless during Drosophila leg development', *Dev Cell* 14(1): 86-96.

Falender, A. E., Freiman, R. N., Geles, K. G., Lo, K. C., Hwang, K., Lamb, D. J., Morris, P. L., Tjian, R. and Richards, J. S. (2005) 'Maintenance of spermatogenesis requires TAF4b, a gonad-specific subunit of TFIID', *Genes Dev* 19(7): 794-803.

FANTOM Consortium (2014) 'A promoter-level mammalian expression atlas', *Nature* 507(7493): 462-70.

Farnham, P. J. (2009) 'Insights from genomic profiling of transcription factors', *Nat Rev Genet* 10(9): 605-16.

Field, H. A., Dong, P. D., Beis, D. and Stainier, D. Y. (2003) 'Formation of the digestive system in zebrafish. II. Pancreas morphogenesis', *Dev Biol* 261(1): 197-208.

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J. and Lobanenkov, V. V. (1996) 'An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes', *Mol Cell Biol* 16(6): 2802-13.

Finch, J. T. and Klug, A. (1976) 'Solenoidal model for superstructure in chromatin', *Proc Natl Acad Sci U S A* 73(6): 1897-901.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. and McCallion, A. S. (2006a) 'Conservation of RET regulatory function from human to zebrafish without sequence similarity', *Science* 312(5771): 276-9.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L., Urasaki, A., Kawakami, K. and McCallion, A. S. (2006b) 'Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish', *Nat Protoc* 1(3): 1297-305.

Fisher, W. W., Li, J. J., Hammonds, A. S., Brown, J. B., Pfeiffer, B. D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. A., Eisen, M. B. et al. (2012) 'DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila', *Proc Natl Acad Sci U S A* 109(52): 21330-5.

Flanagan, P. M., Kelleher, R. J., 3rd, Sayre, M. H., Tschochner, H. and Kornberg, R. D. (1991) 'A mediator required for activation of RNA polymerase II transcription in vitro', *Nature* 350(6317): 436-8.

Forster, B., Van De Ville, D., Berent, J., Sage, D. and Unser, M. (2004) 'Complex wavelets for extended depth-of-field: a new method for the fusion of multichannel microscopy images', *Microsc Res Tech* 65(1-2): 33-42.

Fox, K. R. (1997) 'DNase I footprinting', *Methods Mol Biol* 90: 1-22.

Frankel, N. (2012) 'Multiple layers of complexity in cis-regulatory regions of developmental genes', *Dev Dyn* 241(12): 1857-66.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F. and Stern, D. L. (2010) 'Phenotypic robustness conferred by apparently redundant transcriptional enhancers', *Nature* 466(7305): 490-3.

Frazer, K. A. Ballinger, D. G. Cox, D. R. Hinds, D. A. Stuve, L. L. Gibbs, R. A. Belmont, J. W. Boudreau, A. Hardenbol, P. Leal, S. M. et al. (2007) 'A second generation human haplotype map of over 3.1 million SNPs', *Nature* 449(7164): 851-61.

Frazer, K. A., Sheehan, J. B., Stokowski, R. P., Chen, X., Hosseini, R., Cheng, J. F., Fodor, S. P., Cox, D. R. and Patil, N. (2001) 'Evolutionarily conserved sequences on human chromosome 21', *Genome Res* 11(10): 1651-9.

Freiman, R. N., Albright, S. R., Zheng, S., Sha, W. C., Hammer, R. E. and Tjian, R. (2001) 'Requirement of tissue-selective TBP-associated factor TAFII105 in ovarian development', *Science* 293(5537): 2084-7.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H. et al. (2009) 'An oestrogen-receptor-alpha-bound human chromatin interactome', *Nature* 462(7269): 58-64.

Furey, T. S. (2012) 'ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions', *Nat Rev Genet* 13(12): 840-52.

Gaj, T., Gersbach, C. A. and Barbas, C. F., 3rd (2013) 'ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering', *Trends Biotechnol* 31(7): 397-405.

Galas, D. J. and Schmitz, A. (1978) 'DNAse footprinting: a simple method for the detection of protein-DNA binding specificity', *Nucleic Acids Res* 5(9): 3157-70.

Gao, N., LeLay, J., Vatamaniuk, M. Z., Rieck, S., Friedman, J. R. and Kaestner, K. H. (2008) 'Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development', *Genes Dev* 22(24): 3435-48.

Gardner, R. J., Mackay, D. J., Mungall, A. J., Polychronakos, C., Siebert, R., Shield, J. P., Temple, I. K. and Robinson, D. O. (2000) 'An imprinted locus associated with transient neonatal diabetes mellitus', *Hum Mol Genet* 9(4): 589-96.

Gaszner, M., Vazquez, J. and Schedl, P. (1999) 'The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction', *Genes Dev* 13(16): 2098-107.

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., Bosco, D. et al. (2010) 'A map of open chromatin in human pancreatic islets', *Nat Genet* 42(3): 255-9.

Gehrig, J., Reischl, M., Kalmar, E., Ferg, M., Hadzhiev, Y., Zaucker, A., Song, C., Schindler, S., Liebel, U. and Muller, F. (2009) 'Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos', *Nat Methods* 6(12): 911-6.

Geurts, A. M., Yang, Y., Clark, K. J., Liu, G., Cui, Z., Dupuy, A. J., Bell, J. B., Largaespada, D. A. and Hackett, P. B. (2003) 'Gene transfer into genomes of human cells by the sleeping beauty transposon system', *Mol Ther* 8(1): 108-17.

Geyer, P. K. and Corces, V. G. (1992) 'DNA position-specific repression of transcription by a Drosophila zinc finger protein', *Genes Dev* 6(10): 1865-73.

Giorgetti, L., Siggers, T., Tiana, G., Caprara, G., Notarbartolo, S., Corona, T., Pasparakis, M., Milani, P., Bulyk, M. L. and Natoli, G. (2010) 'Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs', *Mol Cell* 37(3): 418-28.

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 Res* 17(6): 877-85.

Gloyn, A. L., Pearson, E. R., Antcliff, J. F., Proks, P., Bruining, G. J., Slingerland, A. S., Howard, N., Srinivasan, S., Silva, J. M., Molnes, J. et al. (2004) 'Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes', *N Engl J Med* 350(18): 1838-49.

Godinho, L., Mumm, J. S., Williams, P. R., Schroeter, E. H., Koerber, A., Park, S. W., Leach, S. D. and Wong, R. O. (2005) 'Targeting of amacrine cell neurites to appropriate synaptic laminae in the developing zebrafish retina', *Development* 132(22): 5069-79.

Goll, M. G., Anderson, R., Stainier, D. Y., Spradling, A. C. and Halpern, M. E. (2009) 'Transcriptional silencing and reactivation in transgenic zebrafish', *Genetics* 182(3): 747-55.

Goodrich, J. A. and Tjian, R. (1994) 'Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II', *Cell* 77(1): 145-56.

Goodrich, J. A. and Tjian, R. (2010) 'Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation', *Nat Rev Genet* 11(8): 549-58.

Gorkin, D. U., Lee, D., Reed, X., Fletez-Brant, C., Bessling, S. L., Loftus, S. K., Beer, M. A., Pavan, W. J. and McCallion, A. S. (2012) 'Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes', *Genome Res* 22(11): 2290-301.

Gottgens, B., Barton, L. M., Chapman, M. A., Sinclair, A. M., Knudsen, B., Grafham, D., Gilbert, J. G., Rogers, J., Bentley, D. R. and Green, A. R. (2002) 'Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci', *Genome Res* 12(5): 749-59.

Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A. et al. (2006) 'Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes', *Nat Genet* 38(3): 320-3.

Graziano, V., Gerchman, S. E., Schneider, D. K. and Ramakrishnan, V. (1994) 'Histone H1 is located in the interior of the chromatin 30-nm filament', *Nature* 368(6469): 351-4.

Greeley, S. A., Naylor, R. N., Philipson, L. H. and Bell, G. I. (2011) 'Neonatal diabetes: an expanding list of genes allows for improved diagnosis and treatment', *Curr Diab Rep* 11(6): 519-32.

Grehan, S., Tse, E. and Taylor, J. M. (2001) 'Two distal downstream enhancers direct expression of the human apolipoprotein E gene to astrocytes in the brain', *J Neurosci* 21(3): 812-22.

Grice, E. A., Rochelle, E. S., Green, E. D., Chakravarti, A. and McCallion, A. S. (2005) 'Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer', *Hum Mol Genet* 14(24): 3837-45.

Groth, A. C., Fish, M., Nusse, R. and Calos, M. P. (2004) 'Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31', *Genetics* 166(4): 1775-82.

Groth, A. C., Olivares, E. C., Thyagarajan, B. and Calos, M. P. (2000) 'A phage integrase directs efficient site-specific integration in human cells', *Proc Natl Acad Sci U S A* 97(11): 5995-6000.

Guo, T., Hanson, R. L., Traurig, M., Muller, Y. L., Ma, L., Mack, J., Kobes, S., Knowler, W. C., Bogardus, C. and Baier, L. J. (2007) 'TCF7L2 is not a major susceptibility gene for type 2 diabetes in Pima Indians: analysis of 3,501 individuals', *Diabetes* 56(12): 3082-8.

Hadzhiev, Yavor (2007) Phylogenomic and Functional Analyses of Enhancer Evolution of Sonic Hedgehog Paralogs, vol. PhD Thesis: Karlsruhe Institute of Technology.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E. and Taipale, J. (2006) 'Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity', *Cell* 124(1): 47-59.

Hampshire, A. J., Rusling, D. A., Broughton-Head, V. J. and Fox, K. R. (2007) 'Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands', *Methods* 42(2): 128-40.

Hansen, S. K., Takada, S., Jacobson, R. H., Lis, J. T. and Tjian, R. (1997) 'Transcription properties of a cell type-specific TATA-binding protein, TRF', *Cell* 91(1): 71-83.

HapMap Consortium (2003) 'The International HapMap Project', *Nature* 426(6968): 789-96.

HapMap Consortium (2005) 'A haplotype map of the human genome', *Nature* 437(7063): 1299-320.

Hardison, R. C., Oeltjen, J. and Miller, W. (1997) 'Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome', *Genome Res* 7(10): 959-66.

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 Genet* 4(6): e1000106.

Harismendy, O., Notani, D., Song, X., Rahim, N. G., Tanasa, B., Heintzman, N., Ren, B., Fu, X. D., Topol, E. J., Rosenfeld, M. G. et al. (2011) '9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response', *Nature* 470(7333): 264-8.

Hart, C. M., Zhao, K. and Laemmli, U. K. (1997) 'The scs' boundary element: characterization of boundary element-associated factors', *Mol Cell Biol* 17(2): 999-1009.

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', *Proc Natl Acad Sci U S A* 108(14): 5632-7.

Heckman, C. A., Cao, T., Somsouk, L., Duan, H., Mehew, J. W., Zhang, C. Y. and Boxer, L. M. (2003) 'Critical elements of the immunoglobulin heavy chain gene enhancers for deregulated expression of bcl-2', *Cancer Res* 63(20): 6666-73.

Heger, P. and Wiehe, T. (2014) 'New tools in the box: An evolutionary synopsis of chromatin insulators', *Trends Genet*.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W. et al. (2009) 'Histone modifications at human enhancers reflect global cell-type-specific gene expression', *Nature* 459(7243): 108-12.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A. et al. (2007) 'Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome', *Nat Genet* 39(3): 311-8.

Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I. et al. (2007) 'Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution', *Nat Genet* 39(2): 218-25.

Heller, R. S. (2010) 'The comparative anatomy of islets', *Adv Exp Med Biol* 654: 21-37.

Herrmann, B. G. (1991) 'Expression pattern of the Brachyury gene in whole-mount TWis/TWis mutant embryos', *Development* 113(3): 913-7.

Higuchi, R., Krummel, B. and Saiki, R. K. (1988) 'A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions', *Nucleic Acids Res* 16(15): 7351-67.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proc Natl Acad Sci U S A* 106(23): 9362-7.

Hirose, Y. and Ohkuma, Y. (2007) 'Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression', *J Biochem* 141(5): 601-8.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., Hoke, H. A. and Young, R. A. (2013) 'Super-enhancers in the control of cell identity and disease', *Cell* 155(4): 934-47.

Hoess, R. H., Ziese, M. and Sternberg, N. (1982) 'P1 site-specific recombination: nucleotide sequence of the recombining sites', *Proc Natl Acad Sci U S A* 79(11): 3398-402.

Hong, J. W., Hendrix, D. A. and Levine, M. S. (2008) 'Shadow enhancers as a source of evolutionary novelty', *Science* 321(5894): 1314.

Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K., Knight, J., Mani, P., Martin, R., Moxon, S. A. et al. (2013a) 'ZFIN, the Zebrafish Model Organism Database:

increased support for mutants and transgenics', *Nucleic Acids Res* 41(Database issue): D854-60.

Howe, K. Clark, M. D. Torroja, C. F. Torrance, J. Berthelot, C. Muffato, M. Collins, J. E. Humphray, S. McLaren, K. Matthews, L. et al. (2013b) 'The zebrafish reference genome sequence and its relationship to the human genome', *Nature* 496(7446): 498-503.

Hu, D., Gao, X., Morgan, M. A., Herz, H. M., Smith, E. R. and Shilatifard, A. (2013) 'The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers', *Mol Cell Biol* 33(23): 4745-54.

Hu, G., Goll, M. G. and Fisher, S. (2011) 'PhiC31 integrase mediates efficient cassette exchange in the zebrafish germline', *Dev Dyn* 240(9): 2101-7.

Human Genome Sequencing Consortium (2004) 'Finishing the euchromatic sequence of the human genome', *Nature* 431(7011): 931-45.

Hwang, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., Peterson, R. T., Yeh, J. R. and Joung, J. K. (2013) 'Efficient genome editing in zebrafish using a CRISPR-Cas system', *Nat Biotechnol* 31(3): 227-9.

Iborra, F. J., Pombo, A., McManus, J., Jackson, D. A. and Cook, P. R. (1996) 'The topology of transcription by immobilized polymerases', *Exp Cell Res* 229(2): 167-73.

Ishibashi, M., Mechaly, A. S., Becker, T. S. and Rinkwitz, S. (2013) 'Using zebrafish transgenesis to test human genomic sequences for specific enhancer activity', *Methods* 62(3): 216-25.

Istrail, S. and Davidson, E. H. (2005) 'Logic functions of the genomic cis-regulatory code', *Proc Natl Acad Sci U S A* 102(14): 4954-9.

Ivics, Z., Hackett, P. B., Plasterk, R. H. and Izsvak, Z. (1997) 'Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells', *Cell* 91(4): 501-10.

Jaenisch, R., Jahner, D., Nobis, P., Simon, I., Lohler, J., Harbers, K. and Grotkopp, D. (1981) 'Chromosomal position and activation of retroviral genomes inserted into the germ line of mice', *Cell* 24(2): 519-29.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. et al. (2004) 'Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype', *Nature* 431(7011): 946-57.

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): 761-72.

Jiang, Z., Song, J., Qi, F., Xiao, A., An, X., Liu, N. A., Zhu, Z., Zhang, B. and Lin, S. (2008) 'Exdpf is a key regulator of exocrine pancreas development controlled by retinoic acid and ptf1a in zebrafish', *PLoS Biol* 6(11): e293.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K. and Felsenfeld, G. (2009) 'H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions', *Nat Genet* 41(8): 941-5.

Jonsson, J., Carlsson, L., Edlund, T. and Edlund, H. (1994) 'Insulin-promoter-factor 1 is required for pancreas development in mice', *Nature* 371(6498): 606-9.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E. and Furlong, E. E. (2012) 'A transcription factor collective defines cardiac cell fate and reflects lineage history', *Cell* 148(3): 473-86.

Juven-Gershon, T., Hsu, J. Y. and Kadonaga, J. T. (2006) 'Perspectives on the RNA polymerase II core promoter', *Biochem Soc Trans* 34(Pt 6): 1047-50.

Juven-Gershon, T., Hsu, J. Y. and Kadonaga, J. T. (2008) 'Caudal, a key developmental regulator, is a DPE-specific transcriptional factor', *Genes Dev* 22(20): 2823-30.

Kadonaga, J. T. (2004) 'Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors', *Cell* 116(2): 247-57.

Kadonaga, J. T. (2012) 'Perspectives on the RNA polymerase II core promoter', *Wiley Interdiscip Rev Dev Biol* 1(1): 40-51.

Kamakaka, R. T. and Biggins, S. (2005) 'Histone variants: deviants?', *Genes Dev* 19(3): 295-310.

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C. O. et al. (2011) 'Unamplified cap analysis of gene expression on a single-molecule sequencer', *Genome Res* 21(7): 1150-9.

Kaplan, T., Li, X. Y., Sabo, P. J., Thomas, S., Stamatoyannopoulos, J. A., Biggin, M. D. and Eisen, M. B. (2011) 'Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development', *PLoS Genet* 7(2): e1001290.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y. et al. (2007) 'The medaka draft genome and insights into vertebrate genome evolution', *Nature* 447(7145): 714-9.

Kawaguchi, Y., Cooper, B., Gannon, M., Ray, M., MacDonald, R. J. and Wright, C. V. (2002) 'The role of the transcriptional regulator Ptf1a in converting intestinal to pancreatic progenitors', *Nat Genet* 32(1): 128-34.

Kawakami, K. (2005) 'Transposon tools and methods in zebrafish', *Dev Dyn* 234(2): 244-54.

Kawakami, K., Koga, A., Hori, H. and Shima, A. (1998) 'Excision of the tol2 transposable element of the medaka fish, Oryzias latipes, in zebrafish, Danio rerio', *Gene* 225(1-2): 17-22.

Kawakami, K., Shima, A. and Kawakami, N. (2000) 'Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage', *Proc Natl Acad Sci U S A* 97(21): 11403-8.

Kawakami, K., Takeda, H., Kawakami, N., Kobayashi, M., Matsuda, N. and Mishina, M. (2004) 'A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish', *Dev Cell* 7(1): 133-44.

Kettleborough, R. N., Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., de Bruijn, E., van Eeden, F., Sealy, I., White, R. J., Herd, C., Nijman, I. J. et al. (2013) 'A systematic genome-wide analysis of zebrafish protein-coding gene function', *Nature* 496(7446): 494-7.

Khokha, M. K. and Loots, G. G. (2005) 'Strategies for characterising cis-regulatory elements in Xenopus', *Brief Funct Genomic Proteomic* 4(1): 58-68.

Khoo, C., Yang, J., Weinrott, S. A., Kaestner, K. H., Naji, A., Schug, J. and Stoffers, D. A. (2012) 'Research resource: the pdx1 cistrome of pancreatic islets', *Mol Endocrinol* 26(3): 521-33.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engstrom, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K. et al. (2007) 'Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates', *Genome Res* 17(5): 545-55.

Kim, E., Kim, S., Kim, D. H., Choi, B. S., Choi, I. Y. and Kim, J. S. (2012) 'Precision genome engineering with programmable DNA-nicking enzymes', *Genome Res* 22(7): 1327-33.

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S. et al. (2010) 'Widespread transcription at neuronal activity-regulated enhancers', *Nature* 465(7295): 182-7.

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. and Schilling, T. F. (1995) 'Stages of embryonic development of the zebrafish', *Dev Dyn* 203(3): 253-310.

Kinkel, M. D. and Prince, V. E. (2009) 'On the diabetic menu: zebrafish as a model for pancreas development and function', *Bioessays* 31(2): 139-52.

Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. and Grosveld, F. G. (1983) 'Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia', *Nature* 306(5944): 662-6.

Kirov, N. C., Lieberman, P. M. and Rushlow, C. (1996) 'The transcriptional corepressor DSP1 inhibits activated transcription by disrupting TFIIA-TBP complex formation', *EMBO J* 15(24): 7079-87.

Kleinjan, D. A. and Lettice, L. A. (2008) 'Long-range gene control and genetic disease', *Adv Genet* 61: 339-88.

Kleinjan, D. A., Seawright, A., Schedl, A., Quinlan, R. A., Danes, S. and van Heyningen, V. (2001) 'Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6', *Hum Mol Genet* 10(19): 2049-59.

Kleinjan, D. A. and van Heyningen, V. (2005) 'Long-range control of gene expression: emerging mechanisms and disruption in disease', *Am J Hum Genet* 76(1): 8-32.

Knaut, H., Pelegri, F., Bohmann, K., Schwarz, H. and Nusslein-Volhard, C. (2000) 'Zebrafish vasa RNA but not its protein is a component of the germ plasm and segregates asymmetrically before germline specification', *J Cell Biol* 149(4): 875-88.

Koga, A., Suzuki, M., Inagaki, H., Bessho, Y. and Hori, H. (1996) 'Transposable element in fish', *Nature* 383(6595): 30.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S. and Ahringer, J. (2009) 'Differential chromatin marking of introns and expressed exons by H3K36me3', *Nat Genet* 41(3): 376-81.

Komarnitsky, P., Cho, E. J. and Buratowski, S. (2000) 'Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription', *Genes Dev* 14(19): 2452-60.

Koprunner, M., Thisse, C., Thisse, B. and Raz, E. (2001) 'A zebrafish nanos-related gene is essential for the development of primordial germ cells', *Genes Dev* 15(21): 2877-85.

Kornberg, R. D. (1977) 'Structure of chromatin', *Annu Rev Biochem* 46: 931-54.

Kouzarides, T. (2007) 'Chromatin modifications and their function', *Cell* 128(4): 693-705.

Krapp, A., Knofler, M., Frutiger, S., Hughes, G. J., Hagenbuchle, O. and Wellauer, P. K. (1996) 'The p48 DNA-binding subunit of transcription factor PTF1 is a new exocrine pancreas-specific basic helix-loop-helix protein', *EMBO J* 15(16): 4317-29.

Krapp, A., Knofler, M., Ledermann, B., Burki, K., Berney, C., Zoerkler, N., Hagenbuchle, O. and Wellauer, P. K. (1998) 'The bHLH protein PTF1-p48 is essential for the formation of the exocrine and the correct spatial organization of the endocrine pancreas', *Genes Dev* 12(23): 3752-63.

Krivega, I. and Dean, A. (2012) 'Enhancer and promoter interactions-long distance calls', *Curr Opin Genet Dev* 22(2): 79-85.

Kuhn, E. J., Viering, M. M., Rhodes, K. M. and Geyer, P. K. (2003) 'A test of insulator interactions in Drosophila', *EMBO J* 22(10): 2463-71.

Kulkarni, M. M. and Arnosti, D. N. (2003) 'Information display by transcriptional enhancers', *Development* 130(26): 6569-75.

Kumar, S. and Hedges, S. B. (1998) 'A molecular timescale for vertebrate evolution', *Nature* 392(6679): 917-20.

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. and Ebright, R. H. (1998) 'New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB', *Genes Dev* 12(1): 34-44.

Lebrecht, D., Foehr, M., Smith, E., Lopes, F. J., Vanario-Alonso, C. E., Reinitz, J., Burz, D. S. and Hanes, S. D. (2005) 'Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila', *Proc Natl Acad Sci U S A* 102(37): 13176-81.

Lee, C. S., Friedman, J. R., Fulmer, J. T. and Kaestner, K. H. (2005) 'The initiation of liver development is dependent on Foxa transcription factors', *Nature* 435(7044): 944-7.

Lee, T. I. and Young, R. A. (2000) 'Transcription of eukaryotic protein-coding genes', *Annu Rev Genet* 34: 77-137.

Lee, T. I. and Young, R. A. (2013) 'Transcriptional regulation and its misregulation in disease', *Cell* 152(6): 1237-51.

Lenhard, B., Sandelin, A. and Carninci, P. (2012) 'Metazoan promoters: emerging characteristics and insights into transcriptional regulation', *Nat Rev Genet* 13(4): 233-45.

Leonard, J., Peers, B., Johnson, T., Ferreri, K., Lee, S. and Montminy, M. R. (1993) 'Characterization of somatostatin transactivating factor-1, a novel homeobox factor that stimulates somatostatin expression in pancreatic islet cells', *Mol Endocrinol* 7(10): 1275-83.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E. and de Graaff, E. (2003) 'A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly', *Hum Mol Genet* 12(14): 1725-35.

Lettice, L. A., Hill, A. E., Devenney, P. S. and Hill, R. E. (2008) 'Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly', *Hum Mol Genet* 17(7): 978-85.

Lettice, L. A., Horikoshi, T., Heaney, S. J., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., Joosse, M., Akarsu, N., Oostra, B. A., Endo, N. et al. (2002) 'Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly', *Proc Natl Acad Sci U S A* 99(11): 7548-53.

Levine, M. (2010) 'Transcriptional enhancers in animal development and evolution', *Curr Biol* 20(17): R754-63.

Levine, M., Cattoglio, C. and Tjian, R. (2014) 'Looping back to leap forward: transcription enters a new era', *Cell* 157(1): 13-25.

Li, B., Carey, M. and Workman, J. L. (2007) 'The role of chromatin during transcription', *Cell* 128(4): 707-19.

Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., Ariyaratne, P. N., Mohamed, Y. B., Ooi, H. S., Tennakoon, C. et al. (2010a) 'ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing', *Genome Biol* 11(2): R22.

Li, Q., Ritter, D., Yang, N., Dong, Z., Li, H., Chuang, J. H. and Guo, S. (2010b) 'A systematic approach to identify functional motifs within vertebrate developmental enhancers', *Dev Biol* 337(2): 484-95.

Li, X. and Noll, M. (1994) 'Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo', *EMBO J* 13(2): 400-6.

Li, X. Y., Thomas, S., Sabo, P. J., Eisen, M. B., Stamatoyannopoulos, J. A. and Biggin, M. D. (2011) 'The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding', *Genome Biol* 12(4): R34.

Li, Z., Wen, C., Peng, J., Korzh, V. and Gong, Z. (2009) 'Generation of living color transgenic zebrafish to trace somatostatin-expressing cells and endocrine pancreas organization', *Differentiation* 77(2): 128-34.

Liebel, U., Starkuviene, V., Erfle, H., Simpson, J. C., Poustka, A., Wiemann, S. and Pepperkok, R. (2003) 'A microscope-based screening platform for large-scale functional protein analysis in intact cells', *FEBS Lett* 554(3): 394-8.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O. et al. (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science* 326(5950): 289-93.

Lifton, R. P., Goldberg, M. L., Karp, R. W. and Hogness, D. S. (1978) 'The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications', *Cold Spring Harb Symp Quant Biol* 42 Pt 2: 1047-51.

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U. and Kadonaga, J. T. (2004) 'The MTE, a new core promoter element for transcription by RNA polymerase II', *Genes Dev* 18(13): 1606-17.

Lin, J. W., Biankin, A. V., Horb, M. E., Ghosh, B., Prasad, N. B., Yee, N. S., Pack, M. A. and Leach, S. D. (2004) 'Differential requirement for ptf1a in endocrine and exocrine lineages of developing zebrafish pancreas', *Dev Biol* 274(2): 491-503.

Lin, Q., Lin, L. and Zhou, J. (2010) 'Chromatin insulator and the promoter targeting sequence modulate the timing of long-range enhancer-promoter interactions in the Drosophila embryo', *Dev Biol* 339(2): 329-37.

Lister, J. A. (2010) 'Transgene excision in zebrafish using the phiC31 integrase', *Genesis* 48(2): 137-43.

Liu, X., Bushnell, D. A. and Kornberg, R. D. (2013) 'RNA polymerase II transcription: structure and mechanism', *Biochim Biophys Acta* 1829(1): 2-8.

Loven, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I. and Young, R. A. (2013) 'Selective inhibition of tumor oncogenes by disruption of super-enhancers', *Cell* 153(2): 320-34.

Lu, J., Maddison, L. A. and Chen, W. (2011) 'PhiC31 integrase induces efficient site-specific excision in zebrafish', *Transgenic Res* 20(1): 183-9.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. and Richmond, T. J. (1997) 'Crystal structure of the nucleosome core particle at 2.8 A resolution', *Nature* 389(6648): 251-60.

Mackay, D. J., Callaway, J. L., Marks, S. M., White, H. E., Acerini, C. L., Boonen, S. E., Dayanikli, P., Firth, H. V., Goodship, J. A., Haemers, A. P. et al. (2008) 'Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57', *Nat Genet* 40(8): 949-51.

Maeder, M. L., Thibodeau-Beganny, S., Osiak, A., Wright, D. A., Anthony, R. M., Eichtinger, M., Jiang, T., Foley, J. E., Winfrey, R. J., Townsend, J. A. et al. (2008) 'Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification', *Mol Cell* 31(2): 294-301.

Malik, S. and Roeder, R. G. (2010) 'The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation', *Nat Rev Genet* 11(11): 761-72.

Markstein, M., Pitsouli, C., Villalta, C., Celniker, S. E. and Perrimon, N. (2008) 'Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes', *Nat Genet* 40(4): 476-83.

Marshall, H., Studer, M., Popperl, H., Aparicio, S., Kuroiwa, A., Brenner, S. and Krumlauf, R. (1994) 'A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1', *Nature* 370(6490): 567-71.

Marshall, N. F., Peng, J., Xie, Z. and Price, D. H. (1996) 'Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase', *J Biol Chem* 271(43): 27176-83.

Marshall, N. F. and Price, D. H. (1992) 'Control of formation of two distinct classes of RNA polymerase II elongation complexes', *Mol Cell Biol* 12(5): 2078-90.

Marshall, N. F. and Price, D. H. (1995) 'Purification of P-TEFb, a transcription factor required for the transition into productive elongation', *J Biol Chem* 270(21): 12335-8.

Martin, D., Pantoja, C., Fernandez Minan, A., Valdes-Quezada, C., Molto, E., Matesanz, F., Bogdanovic, O., de la Calle-Mustienes, E., Dominguez, O., Taher, L. et al. (2011) 'Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes', *Nat Struct Mol Biol* 18(6): 708-14.

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

Mastracci, T. L. and Sussel, L. (2012) 'The endocrine pancreas: insights into development, differentiation, and diabetes', *Wiley Interdiscip Rev Dev Biol* 1(5): 609-28.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J. et al. (2012) 'Systematic localization of common disease-associated variation in regulatory DNA', *Science* 337(6099): 1190-5.

May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C. et al. (2012) 'Large-scale discovery of enhancers from human heart tissue', *Nat Genet* 44(1): 89-93.

McBride, D. J., Buckle, A., van Heyningen, V. and Kleinjan, D. A. (2011) 'DNaseI hypersensitivity and ultraconservation reveal novel, interdependent long-range enhancers at the complex Pax6 cis-regulatory region', *PLoS One* 6(12): e28616.

McGaughey, D. M., Stine, Z. E., Huynh, J. L., Vinton, R. M. and McCallion, A. S. (2009) 'Asymmetrical distribution of non-conserved regulatory sequences at PHOX2B is reflected at the ENCODE loci and illuminates a possible genome-wide trend', *BMC Genomics* 10: 8.

McLeod, M., Craft, S. and Broach, J. R. (1986) 'Identification of the crossover site during FLP-mediated recombination in the Saccharomyces cerevisiae plasmid 2 microns circle', *Mol Cell Biol* 6(10): 3357-67.

McPherson, C. E., Shim, E. Y., Friedman, D. S. and Zaret, K. S. (1993) 'An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array', *Cell* 75(2): 387-98.

Melgar, M. F., Collins, F. S. and Sethupathy, P. (2011) 'Discovery of active enhancers through bidirectional expression of short transcripts', *Genome Biol* 12(11): R113.

Merika, M. and Thanos, D. (2001) 'Enhanceosomes', *Curr Opin Genet Dev* 11(2): 205-8.

Merli, C., Bergstrom, D. E., Cygan, J. A. and Blackman, R. K. (1996) 'Promoter specificity mediates the independent regulation of neighboring genes', *Genes Dev* 10(10): 1260-70.

Meyer, A. and Schartl, M. (1999) 'Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions', *Curr Opin Cell Biol* 11(6): 699-704.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P. et al. (2007) 'Genome-wide maps of chromatin state in pluripotent and lineage-committed cells', *Nature* 448(7153): 553-60.

Miles, C., Elgar, G., Coles, E., Kleinjan, D. J., van Heyningen, V. and Hastie, N. (1998) 'Complete sequencing of the Fugu WAGR region from WT1 to PAX6: dramatic compaction and conservation of synteny with human chromosome 11p13', *Proc Natl Acad Sci U S A* 95(22): 13068-72.

Milewski, W. M., Duguay, S. J., Chan, S. J. and Steiner, D. F. (1998) 'Conservation of PDX-1 structure, function, and expression in zebrafish', *Endocrinology* 139(3): 1440-9.

Mito, Y., Henikoff, J. G. and Henikoff, S. (2007) 'Histone replacement marks the boundaries of cis-regulatory domains', *Science* 315(5817): 1408-11.

Mohlke, K. L. and Scott, L. J. (2012) 'What will diabetes genomes tell us?', *Curr Diab Rep* 12(6): 643-50.

Montague, W. (1983) 'Diabetes and the Endocrine Pancreas: A biochemical Approach', *Arch Dis Child* 58(11): 942.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F. and Duboule, D. (2011) 'A regulatory archipelago controls Hox genes transcription in digits', *Cell* 147(5): 1132-45.

Morris, A. P. Voight, B. F. Teslovich, T. M. Ferreira, T. Segre, A. V. Steinthorsdottir, V. Strawbridge, R. J. Khan, H. Grallert, H. Mahajan, A. et al. (2012) 'Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes', *Nat Genet* 44(9): 981-90.

Mosimann, Christian, Puller, Ann-Christin, Lawson, Katy L., Tschopp, Patrick, Amsterdam, Adam and Zon, Leonard I. (2013) 'Site-directed zebrafish transgenesis into single landing sites with the phiC31 integrase system', *Developmental Dynamics* 242(8): 949-963.

Muller, F., Blader, P. and Strahle, U. (2002) 'Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements', *Bioessays* 24(6): 564-72.

Muller, F., Chang, B., Albert, S., Fischer, N., Tora, L. and Strahle, U. (1999) 'Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord', *Development* 126(10): 2103-16.

Muller, F. and Tora, L. (2014) 'Chromatin and DNA sequences in defining promoters for transcription initiation', *Biochim Biophys Acta* 1839(3): 118-28.

Muller, F., Williams, D. W., Kobolak, J., Gauvry, L., Goldspink, G., Orban, L. and Maclean, N. (1997) 'Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos', *Mol Reprod Dev* 47(4): 404-12.

Muller, F., Zaucker, A. and Tora, L. (2010) 'Developmental regulation of transcription initiation: more than just changing the actors', *Curr Opin Genet Dev* 20(5): 533-40.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M. et al. (2010) 'From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus', *Nature* 466(7307): 714-9.

Narlikar, L. and Ovcharenko, I. (2009) 'Identifying regulatory elements in eukaryotic genomes', *Brief Funct Genomic Proteomic* 8(4): 215-30.

Narlikar, L., Sakabe, N. J., Blanski, A. A., Arimura, F. E., Westlund, J. M., Nobrega, M. A. and Ovcharenko, I. (2010) 'Genome-wide discovery of human heart enhancers', *Genome Res* 20(3): 381-92.

Nasevicius, A. and Ekker, S. C. (2000) 'Effective targeted gene 'knockdown' in zebrafish', *Nat Genet* 26(2): 216-20.

Navratilova, P., Fredman, D., Hawkins, T. A., Turner, K., Lenhard, B. and Becker, T. S. (2009) 'Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes', *Dev Biol* 327(2): 526-40.

Nechaev, S. and Adelman, K. (2011) 'Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation', *Biochim Biophys Acta* 1809(1): 34-45.

Negre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R. et al. (2011) 'A cis-regulatory map of the Drosophila genome', *Nature* 471(7339): 527-31.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A. M., Sheng, Y., Abdelhamid, R. F., Anand, S. et al. (2013) 'Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis', *Genome Res* 23(11): 1938-50.

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C. et al. (2010) 'Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome', *Nat Genet* 42(9): 790-3.

Ni, Z., Schwartz, B. E., Werner, J., Suarez, J. R. and Lis, J. T. (2004) 'Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes', *Mol Cell* 13(1): 55-65.

Nobrega, M. A., Ovcharenko, I., Afzal, V. and Rubin, E. M. (2003) 'Scanning human gene deserts for long-range enhancers', *Science* 302(5644): 413.

Nobrega, M. A. and Pennacchio, L. A. (2004) 'Comparative genomic analysis as a tool for biological discovery', *J Physiol* 554(Pt 1): 31-9.

Noordermeer, D. and Duboule, D. (2013) 'Chromatin looping and organization at developmentally regulated gene loci', *Wiley Interdiscip Rev Dev Biol* 2(5): 615-30.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J. et al. (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature* 485(7398): 381-5.

Obata, J., Yano, M., Mimura, H., Goto, T., Nakayama, R., Mibu, Y., Oka, C. and Kawaichi, M. (2001) 'p48 subunit of mouse PTF1 binds to RBP-Jkappa/CBF-1, the intracellular mediator of Notch signalling, and is expressed in the neural tube of early stage embryos', *Genes Cells* 6(4): 345-60.

Ogbourne, S. and Antalis, T. M. (1998) 'Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes', *Biochem J* 331 (Pt 1): 1-14.

Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. and Nakatani, Y. (1996) 'The transcriptional coactivators p300 and CBP are histone acetyltransferases', *Cell* 87(5): 953-9.

Ohlsson, H., Karlsson, K. and Edlund, T. (1993) 'IPF1, a homeodomain-containing transactivator of the insulin gene', *EMBO J* 12(11): 4251-9.

Ohtsuki, S., Levine, M. and Cai, H. N. (1998) 'Different core promoters possess distinct regulatory activities in the Drosophila embryo', *Genes Dev* 12(4): 547-56.

Olivares, E. C., Hollis, R. P., Chalberg, T. W., Meuse, L., Kay, M. A. and Calos, M. P. (2002) 'Site-specific genomic integration produces therapeutic Factor IX levels in mice', *Nat Biotechnol* 20(11): 1124-8.

Oliver-Krasinski, J. M. and Stoffers, D. A. (2008) 'On the origin of the beta cell', *Genes Dev* 22(15): 1998-2021.

Orphanides, G., Lagrange, T. and Reinberg, D. (1996) 'The general transcription factors of RNA polymerase II', *Genes Dev* 10(21): 2657-83.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S., Reik, W. et al. (2004) 'Active genes dynamically colocalize to shared sites of ongoing transcription', *Nat Genet* 36(10): 1065-71.

Pal, A. and McCarthy, M. I. (2013) 'The genetics of type 2 diabetes and its clinical relevance', *Clin Genet* 83(4): 297-306.

Palstra, R. J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B. and de Laat, W. (2008) 'Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription', *PLoS One* 3(2): e1661.

Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L. et al. (2013) 'Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants', *Proc Natl Acad Sci U S A* 110(44): 17921-6.

Parkhurst, S. M., Harrison, D. A., Remington, M. P., Spana, C., Kelley, R. L., Coyne, R. S. and Corces, V. G. (1988) 'The Drosophila su(Hw) gene, which controls the phenotypic effect of the gypsy transposable element, encodes a putative DNA-binding protein', *Genes Dev* 2(10): 1205-15.

Parry, T. J., Theisen, J. W., Hsu, J. Y., Wang, Y. L., Corcoran, D. L., Eustice, M., Ohler, U. and Kadonaga, J. T. (2010) 'The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery', *Genes Dev* 24(18): 2013-8.

Parsons, M. J., Pisharath, H., Yusuff, S., Moore, J. C., Siekmann, A. F., Lawson, N. and Leach, S. D. (2009) 'Notch-responsive cells initiate the secondary transition in larval zebrafish pancreas', *Mech Dev* 126(10): 898-912.

Pashos, E. E., Kague, E. and Fisher, S. (2008) 'Evaluation of cis-regulatory function in zebrafish', *Brief Funct Genomic Proteomic* 7(6): 465-73.

Pasquali, L., Gaulton, K. J., Rodriguez-Segui, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J. J., Moran, I., Gomez-Marin, C., van de Bunt, M. et al. (2014) 'Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants', *Nat Genet* 46(2): 136-43.

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S. I., Cooper, G. M. et al. (2012) 'Massively parallel functional dissection of mammalian enhancers in vivo', *Nat Biotechnol* 30(3): 265-70.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D. et al. (2006) 'In vivo enhancer analysis of human conserved non-coding sequences', *Nature* 444(7118): 499-502.

Peravali, R., Gehrig, J., Giselbrecht, S., Lutjohann, D. S., Hadzhiev, Y., Muller, F. and Liebel, U. (2011) 'Automated feature detection and imaging for high-resolution screening of zebrafish embryos', *Biotechniques* 50(5): 319-24.

Persengiev, S. P., Zhu, X., Dixit, B. L., Maston, G. A., Kittler, E. L. and Green, M. R. (2003) 'TRF3, a TATA-box-binding protein-related factor, is vertebrate-specific and widely expressed', *Proc Natl Acad Sci U S A* 100(25): 14887-91.

Petersen, H. V., Serup, P., Leonard, J., Michelsen, B. K. and Madsen, O. D. (1994) 'Transcriptional regulation of the human insulin gene is dependent on the homeodomain protein STF1/IPF1 acting through the CT boxes', *Proc Natl Acad Sci U S A* 91(22): 10465-9.

Pham, T. H., Benner, C., Lichtinger, M., Schwarzfischer, L., Hu, Y., Andreesen, R., Chen, W. and Rehli, M. (2012) 'Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states', *Blood* 119(24): e161-71.

Pisharath, H., Rhee, J. M., Swanson, M. A., Leach, S. D. and Parsons, M. J. (2007) 'Targeted ablation of beta cells in the embryonic zebrafish pancreas using E. coli nitroreductase', *Mech Dev* 124(3): 218-29.

Pistocchi, A., Feijoo, C. G., Cabrera, P., Villablanca, E. J., Allende, M. L. and Cotelli, F. (2009) 'The zebrafish prospero homolog prox1 is required for mechanosensory hair cell differentiation and functionality in the lateral line', *BMC Dev Biol* 9: 58.

Pistocchi, A., Gaudenzi, G., Carra, S., Bresciani, E., Del Giacco, L. and Cotelli, F. (2008) 'Crucial role of zebrafish prox1 in hypothalamic catecholaminergic neurons development', *BMC Dev Biol* 8: 27.

Placek, B. J., Harrison, L. N., Villers, B. M. and Gloss, L. M. (2005) 'The H2A.Z/H2B dimer is unstable compared to the dimer containing the major H2A isoform', *Protein Sci* 14(2): 514-22.

Plasterk, R. H. (1993) 'Molecular mechanisms of transposition and its control', *Cell* 74(5): 781-6.

Pointud, J. C., Mengus, G., Brancorsini, S., Monaco, L., Parvinen, M., Sassone-Corsi, P. and Davidson, I. (2003) 'The intracellular localisation of TAF7L, a paralogue of transcription factor TFIID subunit TAF7, is developmentally regulated during male germ-cell differentiation', *J Cell Sci* 116(Pt 9): 1847-58.

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., Beckwith, C. A., Chan, J. A., Hills, A., Davis, M. et al. (2009) 'The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer', *Nat Genet* 41(8): 882-4.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Sandelin, A. (2006) 'Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters', *Genome Biol* 7(8): R78.

Prazak, L., Fujioka, M. and Gergen, J. P. (2010) 'Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors', *Dev Biol* 344(2): 1048-59.

Preibisch, S., Saalfeld, S. and Tomancak, P. (2009) 'Globally optimal stitching of tiled 3D microscopic image acquisitions', *Bioinformatics* 25(11): 1463-5.

Prokopenko, I., McCarthy, M. I. and Lindgren, C. M. (2008) 'Type 2 diabetes: new genes, new understanding', *Trends Genet* 24(12): 613-21.

Ptashne, M. and Gann, A. (1997) 'Transcriptional activation by recruitment', *Nature* 386(6625): 569-77.

Punnamoottil, B., Herrmann, C., Pascual-Anaya, J., D'Aniello, S., Garcia-Fernandez, J., Akalin, A., Becker, T. S. and Rinkwitz, S. (2010) 'Cis-regulatory characterization of sequence conservation surrounding the Hox4 genes', *Dev Biol* 340(2): 269-82.

Rach, E. A., Winter, D. R., Benjamin, A. M., Corcoran, D. L., Ni, T., Zhu, J. and Ohler, U. (2011) 'Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level', *PLoS Genet* 7(1): e1001274.

Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S. A., Swigut, T. and Wysocka, J. (2012) 'Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest', *Cell Stem Cell* 11(5): 633-48.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A. and Wysocka, J. (2011) 'A unique chromatin signature uncovers early developmental enhancers in humans', *Nature* 470(7333): 279-83.

Rada-Iglesias, A., Prescott, S. L. and Wysocka, J. (2013) 'Human genetic variation within neural crest enhancers: molecular and phenotypic implications', *Philos Trans R Soc Lond B Biol Sci* 368(1620): 20120360.

Rada-Iglesias, A. and Wysocka, J. (2011) 'Epigenomics of human embryonic stem cells and induced pluripotent stem cells: insights into pluripotency and implications for disease', *Genome Med* 3(6): 36.

Ragvin, A., Moro, E., Fredman, D., Navratilova, P., Drivenes, O., Engstrom, P. G., Alonso, M. E., de la Calle Mustienes, E., Gomez Skarmeta, J. L., Tavares, M. J. et al. (2010) 'Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3', *Proc Natl Acad Sci U S A* 107(2): 775-80.

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., Domann, F. E., Govil, M., Christensen, K., Bille, C. et al. (2008) 'Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip', *Nat Genet* 40(11): 1341-7.

Raisner, R. M., Hartley, P. D., Meneghini, M. D., Bao, M. Z., Liu, C. L., Schreiber, S. L., Rando, O. J. and Madhani, H. D. (2005) 'Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin', *Cell* 123(2): 233-48.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M. and Ren, B. (2013) 'RFECS: a random-forest based algorithm for enhancer identification from chromatin state', *PLoS Comput Biol* 9(3): e1002968.

Ramirez, C. L., Certo, M. T., Mussolino, C., Goodwin, M. J., Cradick, T. J., McCaffrey, A. P., Cathomen, T., Scharenberg, A. M. and Joung, J. K. (2012) 'Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects', *Nucleic Acids Res* 40(12): 5560-8.

Ravi, V., Bhatia, S., Gautier, P., Loosli, F., Tay, B. H., Tay, A., Murdoch, E., Coutinho, P., van Heyningen, V., Brenner, S. et al. (2013) 'Sequencing of Pax6 loci from the elephant shark reveals a family of Pax6 genes in vertebrate genomes, forged by ancient duplications and divergences', *PLoS Genet* 9(1): e1003177.

Raymond, C. S. and Soriano, P. (2007) 'High-efficiency FLP and PhiC31 site-specific recombination in mammalian cells', *PLoS One* 2(1): e162.

Rippe, K., von Hippel, P. H. and Langowski, J. (1995) 'Action at a distance: DNA-looping and initiation of transcription', *Trends Biochem Sci* 20(12): 500-6.

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.

Ritter, D. I., Li, Q., Kostka, D., Pollard, K. S., Guo, S. and Chuang, J. H. (2010) 'The importance of being cis: evolution of orthologous fish and mammalian enhancer activity', *Mol Biol Evol* 27(10): 2322-32.

Roberts, J. A., Miguel-Escalada, I., Slovik, K. J., Walsh, K. T., Hadzhiev, Y., Sanges, R., Stupka, E., Marsh, E. K., Balciuniene, J., Balciunas, D. et al. (2014) 'Targeted transgene integration overcomes variability of position effects in zebrafish', *Development* 141(3): 715-24.

Roberts, Jennifer Anne (2013) Development of a PhiC31 system for functional characterisation of cis-regulatory elements in reporter transgenic zebrafish *School of Clinical and Experimental Medicine*, vol. Ph.D.: University of Birmingham.

Roeder, R. G. (1996) 'The role of general initiation factors in transcription by RNA polymerase II', *Trends Biochem Sci* 21(9): 327-35.

Roelfsema, J. H., White, S. J., Ariyurek, Y., Bartholdi, D., Niedrist, D., Papadia, F., Bacino, C. A., den Dunnen, J. T., van Ommen, G. J., Breuning, M. H. et al. (2005) 'Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease', *Am J Hum Genet* 76(4): 572-80.

Rose, S. D., Swift, G. H., Peyton, M. J., Hammer, R. E. and MacDonald, R. J. (2001) 'The role of PTF1-P48 in pancreatic acinar gene expression', *J Biol Chem* 276(47): 44018-26.

Rossant, J., Nutter, L. M. and Gertsenstein, M. (2011) 'Engineering the embryo', *Proc Natl Acad Sci U S A* 108(19): 7659-60.

Rougvie, A. E. and Lis, J. T. (1988) 'The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged', *Cell* 54(6): 795-804.

Roure, A., Rothbacher, U., Robin, F., Kalmar, E., Ferone, G., Lamy, C., Missero, C., Mueller, F. and Lemaire, P. (2007) 'A multicassette Gateway vector set for high throughput and comparative analyses in ciona and vertebrate embryos', *PLoS One* 2(9): e916.

Royo, J. L., Bessa, J., Hidalgo, C., Fernandez-Minan, A., Tena, J. J., Roncero, Y., Gomez-Skarmeta, J. L. and Casares, F. (2012) 'Identification and analysis of conserved cis-regulatory regions of the MEIS1 gene', *PLoS One* 7(3): e33617.

Royo, J. L., Hidalgo, C., Roncero, Y., Seda, M. A., Akalin, A., Lenhard, B., Casares, F. and Gomez-Skarmeta, J. L. (2011) 'Dissecting the transcriptional regulatory properties of human chromosome 16 highly conserved non-coding regions', *PLoS One* 6(9): e24824.

Salina, A., Pasquali, L., Aloi, C., Lugani, F., d'Annunzio, G. and Lorini, R. (2010) 'Neonatal diabetes caused by pancreatic agenesia: which other genes should be used for diagnosis?', *Diabetes Care* 33(8): e112.

Sambrook, J. and Russell, D. W. (2006) 'Isolation of High-molecular-weight DNA from Mammalian Cells Using Proteinase K and Phenol', *CSH Protoc* 2006(1).

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P. G., Klos, J. M., Wasserman, W. W., Ericson, J. and Lenhard, B. (2004) 'Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes', *BMC Genomics* 5(1): 99.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D. A. (2007) 'Mammalian RNA polymerase II core promoters: insights from genome-wide studies', *Nat Rev Genet* 8(6): 424-36.

Sander, J. D., Cade, L., Khayter, C., Reyon, D., Peterson, R. T., Joung, J. K. and Yeh, J. R. (2011) 'Targeted gene disruption in somatic zebrafish cells using engineered TALENs', *Nat Biotechnol* 29(8): 697-8.

Sander, M., Sussel, L., Conners, J., Scheel, D., Kalamaras, J., Dela Cruz, F., Schwitzgebel, V., Hayes-Jordan, A. and German, M. (2000) 'Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta-cell formation in the pancreas', *Development* 127(24): 5533-40.

Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F. and Stupka, E. (2006) 'Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage', *Genome Biol* 7(7): R56.

Santagati, F., Abe, K., Schmidt, V., Schmitt-John, T., Suzuki, M., Yamamura, K. and Imai, K. (2003) 'Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny', *Genetics* 165(1): 235-42.

Sanyal, A., Lajoie, B. R., Jain, G. and Dekker, J. (2012) 'The long-range interaction landscape of gene promoters', *Nature* 489(7414): 109-13.

Sauer, B. and Henderson, N. (1990) 'Targeted insertion of exogenous DNA into the eukaryotic genome by the Cre recombinase', *New Biol* 2(5): 441-9.

Savic, D., Park, S. Y., Bailey, K. A., Bell, G. I. and Nobrega, M. A. (2013) 'In vitro scan for enhancers at the TCF7L2 locus', *Diabetologia* 56(1): 121-5.

Savic, D., Ye, H., Aneas, I., Park, S. Y., Bell, G. I. and Nobrega, M. A. (2011) 'Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism', *Genome Res*.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. and Snyder, M. (2012) 'Linking disease associations with regulatory information in the human genome', *Genome Res* 22(9): 1748-59.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P. and Odom, D. T. (2012) 'Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages', *Cell* 148(1-2): 335-48.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S. et al. (2010) 'Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding', *Science* 328(5981): 1036-40.

Schulte-Merker, S., Ho, R. K., Herrmann, B. G. and Nusslein-Volhard, C. (1992) 'The protein product of the zebrafish homologue of the mouse T gene is expressed in nuclei of the germ ring and the notochord of the early embryo', *Development* 116(4): 1021-32.

Schwabish, M. A. and Struhl, K. (2007) 'The Swi/Snf complex is important for histone eviction during transcriptional activation and RNA polymerase II elongation in vivo', *Mol Cell Biol* 27(20): 6987-95.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003) 'Human-mouse alignments with BLASTZ', *Genome Res* 13(1): 103-7.

Schwitzgebel, V. M., Mamin, A., Brun, T., Ritz-Laser, B., Zaiko, M., Maret, A., Jornayvaz, F. R., Theintz, G. E., Michielin, O., Melloul, D. et al. (2003) 'Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1', *J Clin Endocrinol Metab* 88(9): 4398-406.

Sellick, G. S., Barker, K. T., Stolte-Dijkstra, I., Fleischmann, C., Coleman, R. J., Garrett, C., Gloyn, A. L., Edghill, E. L., Hattersley, A. T., Wellauer, P. K. et al. (2004) 'Mutations in PTF1A cause pancreatic and cerebellar agenesis', *Nat Genet* 36(12): 1301-5.

Semenza, G. L., Delgrosso, K., Poncz, M., Malladi, P., Schwartz, E. and Surrey, S. (1984) 'The silent carrier allele: beta thalassemia without a mutation in the beta-globin gene or its immediate flanking regions', *Cell* 39(1): 123-8.

Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M. and Levine, M. (2004) 'Immunity regulatory DNAs share common organizational features in Drosophila', *Mol Cell* 13(1): 19-32.

Shamsadin, R., Adham, I. M., von Beust, G. and Engel, W. (2000) 'Molecular cloning, expression and chromosome location of the human pelota gene PELO', *Cytogenet Cell Genet* 90(1-2): 75-8.

Shen, B., Zhang, W., Zhang, J., Zhou, J., Wang, J., Chen, L., Wang, L., Hodgkins, A., Iyer, V., Huang, X. et al. (2014) 'Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects', *Nat Methods* 11(4): 399-402.

Shih, S. J., Allan, C., Grehan, S., Tse, E., Moran, C. and Taylor, J. M. (2000) 'Duplicated downstream enhancers control expression of the human apolipoprotein E gene in macrophages and adipose tissue', *J Biol Chem* 275(41): 31567-72.

Shin, J. T., Priest, J. R., Ovcharenko, I., Ronco, A., Moore, R. K., Burns, C. G. and MacRae, C. A. (2005) 'Human-zebrafish non-coding conserved elements act in vivo to regulate transcription', *Nucleic Acids Res* 33(17): 5437-45.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T. et al. (2003) 'Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage', *Proc Natl Acad Sci U S A* 100(26): 15776-81.

Simonet, W. S., Bucay, N., Lauer, S. J. and Taylor, J. M. (1993) 'A far-downstream hepatocyte-specific control region directs expression of the linked human apolipoprotein E and C-I genes in transgenic mice', *J Biol Chem* 268(11): 8221-9.

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', *J Biol Chem* 266(14): 8651-4.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) 'Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)', *Nat Genet* 38(11): 1348-54.

Slack, J. M. (1995) 'Developmental biology of the pancreas', *Development* 121(6): 1569-80.

Smale, S. T. and Baltimore, D. (1989) 'The "initiator" as a transcription control element', *Cell* 57(1): 103-13.

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gomez-Marin, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F. et al. (2014) 'Obesity-associated variants within FTO form long-range functional connections with IRX3', *Nature* 507(7492): 371-5.

Smith, E. and Shilatifard, A. (2014) 'Enhancer biology and enhanceropathies', *Nat Struct Mol Biol* 21(3): 210-9.

Smith, M. C., Till, R., Brady, K., Soultanas, P., Thorpe, H. and Smith, M. C. (2004) 'Synapsis and DNA cleavage in phiC31 integrase-mediated site-specific recombination', *Nucleic Acids Res* 32(8): 2607-17.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., Sheffield, N. C., Graf, S., Huss, M., Keefe, D. et al. (2011) 'Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity', *Genome Res* 21(10): 1757-67.

Sosa-Pineda, B., Wigle, J. T. and Oliver, G. (2000) 'Hepatocyte migration during liver development requires Prox1', *Nat Genet* 25(3): 254-5.

Spence, R., Gerlach, G., Lawrence, C. and Smith, C. (2008) 'The behaviour and ecology of the zebrafish, Danio rerio', *Biol Rev Camb Philos Soc* 83(1): 13-34.

Sperling, M. A. (2005) 'Neonatal diabetes mellitus: from understudy to center stage', *Curr Opin Pediatr* 17(4): 512-8.

Spieler, D., Kaffe, M., Knauf, F., Bessa, J., Tena, J. J., Giesert, F., Schormair, B., Tilch, E., Lee, H., Horsch, M. et al. (2014) 'Restless Legs Syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon', *Genome Res* 24(4): 592-603.

Spitz, F. and Furlong, E. E. (2012) 'Transcription factors: from enhancer binding to developmental control', *Nat Rev Genet* 13(9): 613-26.

Stankov, K., Benc, D. and Draskovic, D. (2013) 'Genetic and epigenetic factors in etiology of diabetes mellitus type 1', *Pediatrics* 132(6): 1112-22.

Stark, W. M., Boocock, M. R. and Sherratt, D. J. (1992) 'Catalysis by site-specific recombinases', *Trends Genet* 8(12): 432-9.

Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D. J., Talianidis, I., Marioni, J. C. et al. (2013) 'Cooperativity and rapid evolution of cobound transcription factors in closely related mammals', *Cell* 154(3): 530-40.

Stitzel, M. L., Sethupathy, P., Pearson, D. S., Chines, P. S., Song, L., Erdos, M. R., Welch, R., Parker, S. C., Boyle, A. P., Scott, L. J. et al. (2010) 'Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci', *Cell Metab* 12(5): 443-55.

Stoffers, D. A., Ferrer, J., Clarke, W. L. and Habener, J. F. (1997a) 'Early-onset type-II diabetes mellitus (MODY4) linked to IPF1', *Nat Genet* 17(2): 138-9.

Stoffers, D. A., Zinkin, N. T., Stanojevic, V., Clarke, W. L. and Habener, J. F. (1997b) 'Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence', *Nat Genet* 15(1): 106-10.

Streisinger, G., Walker, C., Dower, N., Knauber, D. and Singer, F. (1981) 'Production of clones of homozygous diploid zebra fish (Brachydanio rerio)', *Nature* 291(5813): 293-6.

Stuart, G. W., McMurray, J. V. and Westerfield, M. (1988) 'Replication, integration and stable germ-line transmission of foreign sequences injected into early zebrafish embryos', *Development* 103(2): 403-12.

Stuart, G. W., Vielkind, J. R., McMurray, J. V. and Westerfield, M. (1990) 'Stable lines of transgenic zebrafish exhibit reproducible patterns of transgene expression', *Development* 109(3): 577-84.

Stumvoll, M., Goldstein, B. J. and van Haeften, T. W. (2005) 'Type 2 diabetes: principles of pathogenesis and therapy', *Lancet* 365(9467): 1333-46.

Summerbell, D., Ashby, P. R., Coutelle, O., Cox, D., Yee, S. and Rigby, P. W. (2000) 'The expression of Myf5 in the developing mouse embryo is controlled by discrete and dispersed enhancers specific for particular populations of skeletal muscle precursors', *Development* 127(17): 3745-57.

Sun, F. L. and Elgin, S. C. (1999) 'Putting boundaries on silence', *Cell* 99(5): 459-62.

Sussel, L., Kalamaras, J., Hartigan-O'Connor, D. J., Meneses, J. J., Pedersen, R. A., Rubenstein, J. L. and German, M. S. (1998) 'Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due to arrested differentiation of pancreatic beta cells', *Development* 125(12): 2213-21.

Suster, M. L., Sumiyama, K. and Kawakami, K. (2009) 'Transposon-mediated BAC transgenesis in zebrafish and mice', *BMC Genomics* 10: 477.

Sutherland, H. and Bickmore, W. A. (2009) 'Transcription factories: gene expression in unions?', *Nat Rev Genet* 10(7): 457-66.

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', *Dev Cell* 18(3): 359-70.

Taatjes, D. J., Marr, M. T. and Tjian, R. (2004) 'Regulatory diversity among metazoan co-activator complexes', *Nat Rev Mol Cell Biol* 5(5): 403-10.

Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S. and Ovcharenko, I. (2011) 'Genome-wide identification of conserved regulatory function in diverged sequences', *Genome Res* 21(7): 1139-49.

Taylor, G. C., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M. M. and Bickmore, W. A. (2013) 'H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction', *Genome Res* 23(12): 2053-65.

Temple, I. K., Gardner, R. J., Mackay, D. J., Barber, J. C., Robinson, D. O. and Shield, J. P. (2000) 'Transient neonatal diabetes: widening the understanding of the etiopathogenesis of diabetes', *Diabetes* 49(8): 1359-66.

Temple, I. K., Gardner, R. J., Robinson, D. O., Kibirige, M. S., Ferguson, A. W., Baum, J. D., Barber, J. C., James, R. S. and Shield, J. P. (1996) 'Further evidence for an imprinted gene for neonatal diabetes localised to chromosome 6q22-q23', *Hum Mol Genet* 5(8): 1117-21.

Thanos, D. and Maniatis, T. (1992) 'The high mobility group protein HMG I(Y) is required for NF-kappa B-dependent virus induction of the human IFN-beta gene', *Cell* 71(5): 777-89.

Thanos, D. and Maniatis, T. (1995) 'Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome', *Cell* 83(7): 1091-100.

Thisse, B., Heyer, V., Lux, A., Alunni, V., Degrave, A., Seiliez, I., Kirchner, J., Parkhill, J. P. and Thisse, C. (2004) 'Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening', *Methods Cell Biol* 77: 505-19.

Thisse, C. and Thisse, B. (2008) 'High-resolution in situ hybridization to whole-mount zebrafish embryos', *Nat Protoc* 3(1): 59-69.

Thomas, I. H., Saini, N. K., Adhikari, A., Lee, J. M., Kasa-Vubu, J. Z., Vazquez, D. M., Menon, R. K., Chen, M. and Fajans, S. S. (2009) 'Neonatal diabetes mellitus with pancreatic agenesis in an infant with homozygous IPF-1 Pro63fsX60 mutation', *Pediatr Diabetes* 10(7): 492-6.

Thorpe, H. M. and Smith, M. C. (1998) 'In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family', *Proc Natl Acad Sci U S A* 95(10): 5505-10.

Thorpe, H. M., Wilson, S. E. and Smith, M. C. (2000) 'Control of directionality in the site-specific recombination system of the Streptomyces phage phiC31', *Mol Microbiol* 38(2): 232-41.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B. et al. (2012) 'The accessible chromatin landscape of the human genome', *Nature* 489(7414): 75-82.

Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M. O., Scacheri, P. C. and Harte, P. J. (2009) 'CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing', *Development* 136(18): 3131-41.

Tiso, N., Moro, E. and Argenton, F. (2009) 'Zebrafish pancreas development', *Mol Cell Endocrinol* 312(1-2): 24-30.

Todeschini, A. L., Georges, A. and Veitia, R. A. (2014) 'Transcription factors: specific DNA binding and specific gene regulation', *Trends Genet*.

Triezenberg, S. J. (1995) 'Structure and function of transcriptional activation domains', *Curr Opin Genet Dev* 5(2): 190-6.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S. and Raychaudhuri, S. (2013) 'Chromatin marks identify critical cell types for fine mapping complex trait variants', *Nat Genet* 45(2): 124-30.

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', *J Mol Biol* 185(2): 341-58.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. and Rozen, S. G. (2012) 'Primer3--new capabilities and interfaces', *Nucleic Acids Res* 40(15): e115.

Urasaki, A., Morvan, G. and Kawakami, K. (2006) 'Functional dissection of the Tol2 transposable element identified the minimal cis-sequence and a highly repetitive sequence in the subterminal region essential for transposition', *Genetics* 174(2): 639-49.

van Arensbergen, J., Garcia-Hurtado, J., Moran, I., Maestro, M. A., Xu, X., Van de Casteele, M., Skoudy, A. L., Palassini, M., Heimberg, H. and Ferrer, J. (2010) 'Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program', *Genome Res* 20(6): 722-32.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009) 'A census of human transcription factors: function, expression and evolution', *Nat Rev Genet* 10(4): 252-63.

Vastenhouw, N. L., Zhang, Y., Woods, I. G., Imam, F., Regev, A., Liu, X. S., Rinn, J. and Schier, A. F. (2010) 'Chromatin signature of embryonic pluripotency is established during genome activation', *Nature* 464(7290): 922-6.

Venken, K. J. and Bellen, H. J. (2012) 'Genome-wide manipulations of Drosophila melanogaster with transposons, Flp recombinase, and PhiC31 integrase', *Methods Mol Biol* 859: 203-28.

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): 1115-25.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. (2009a) 'ChIP-seq accurately predicts tissue-specific activity of enhancers', *Nature* 457(7231): 854-8.

Visel, A., Bristow, J. and Pennacchio, L. A. (2007) 'Enhancer identification through comparative genomics', *Semin Cell Dev Biol* 18(1): 140-52.

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M. and Pennacchio, L. A. (2008) 'Ultraconservation identifies a small subset of extremely constrained developmental enhancers', *Nat Genet* 40(2): 158-60.

Visel, A., Rubin, E. M. and Pennacchio, L. A. (2009b) 'Genomic views of distant-acting enhancers', *Nature* 461(7261): 199-205.

Walter, J., Dever, C. A. and Biggin, M. D. (1994) 'Two homeo domain proteins bind with similar specificity to a wide range of DNA sites in Drosophila embryos', *Genes Dev* 8(14): 1678-92.

Wang, C., Zhang, M. Q. and Zhang, Z. (2013) 'Computational identification of active enhancers in model organisms', *Genomics Proteomics Bioinformatics* 11(3): 142-50.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A. et al. (2011a) 'Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA', *Nature* 474(7351): 390-4.

Wang, J., Kilic, G., Aydin, M., Burke, Z., Oliver, G. and Sosa-Pineda, B. (2005) 'Prox1 activity controls pancreas morphogenesis and participates in the production of "secondary transition" pancreatic endocrine cells', *Dev Biol* 286(1): 182-94.

Wang, Y., Rovira, M., Yusuff, S. and Parsons, M. J. (2011b) 'Genetic inducible fate mapping in larval zebrafish reveals origins of adult insulin-producing beta-cells', *Development* 138(4): 609-17.

Wardle, F. C., Odom, D. T., Bell, G. W., Yuan, B., Danford, T. W., Wiellette, E. L., Herbolsheimer, E., Sive, H. L., Young, R. A. and Smith, J. C. (2006) 'Zebrafish promoter microarrays identify actively transcribed embryonic genes', *Genome Biol* 7(8): R71.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. and Lawrence, C. E. (2000) 'Human-mouse genome comparisons to locate regulatory sites', *Nat Genet* 26(2): 225-8.

Webster, D. E., Barajas, B., Bussat, R. T., Yan, K. J., Neela, P. H., Flockhart, R. J., Kovalski, J., Zehnder, A. and Khavari, P. A. (2014) 'Enhancer-targeted genome editing selectively blocks innate resistance to oncokinase inhibition', *Genome Res* 24(5): 751-60.

Weedon, M. N., Cebola, I., Patch, A. M., Flanagan, S. E., De Franco, E., Caswell, R., Rodriguez-Segui, S. A., Shaw-Smith, C., Cho, C. H., Lango Allen, H. et al. (2014) 'Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis', *Nat Genet* 46(1): 61-4.

Westerfield, M., Wegner, J., Jegalian, B. G., DeRobertis, E. M. and Puschel, A. W. (1992) 'Specific activation of mammalian Hox promoters in mosaic transgenic zebrafish', *Genes Dev* 6(4): 591-8.

Westerfield, Monte (1993) *The zebrafish book : a guide for the laboratory use of zebrafish (Brachydanio rerio)*, Eugene, OR: M. Westerfield.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I. and Young, R. A. (2013) 'Master transcription factors and mediator establish super-enhancers at key cell identity genes', *Cell* 153(2): 307-19.

Widom, J. and Klug, A. (1985) 'Structure of the 300A chromatin filament: X-ray diffraction from oriented samples', *Cell* 43(1): 207-13.

Wienholds, E., van Eeden, F., Kosters, M., Mudde, J., Plasterk, R. H. and Cuppen, E. (2003) 'Efficient target-selected mutagenesis in zebrafish', *Genome Res* 13(12): 2700-7.

Wilfinger, A., Arkhipova, V. and Meyer, D. (2013) 'Cell type and tissue specific function of islet genes in zebrafish pancreas development', *Dev Biol* 378(1): 25-37.

Wilson, C., Bellen, H. J. and Gehring, W. J. (1990) 'Position effects on eukaryotic gene expression', *Annu Rev Cell Biol* 6: 679-714.

Wilson, R., Ryan, G. B., Knight, G. L., Laimins, L. A. and Roberts, S. (2007) 'The full-length E1E4 protein of human papillomavirus type 18 modulates differentiation-dependent viral DNA amplification and late gene expression', *Virology* 362(2): 453-60.

Winter, W. E., Maclaren, N. K., Riley, W. J., Toskes, P. P., Andres, J. and Rosenbloom, A. L. (1986) 'Congenital pancreatic hypoplasia: a syndrome of exocrine and endocrine pancreatic insufficiency', *J Pediatr* 109(3): 465-8.

Wittbrodt, J., Shima, A. and Schartl, M. (2002) 'Medaka--a model organism from the far East', *Nat Rev Genet* 3(1): 53-64.

Woodcock, C. L. and Ghosh, R. P. (2010) 'Chromatin higher-order structure and dynamics', *Cold Spring Harb Perspect Biol* 2(5): a000596.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K. et al. (2005) 'Highly conserved non-coding sequences are associated with vertebrate development', *PLoS Biol* 3(1): e7.

Workman, J. L. (2006) 'Nucleosome displacement in transcription', *Genes Dev* 20(15): 2009-17.

Wu, C. (1980) 'The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I', *Nature* 286(5776): 854-60.

Wunderlich, Z. and Mirny, L. A. (2009) 'Different gene regulation strategies revealed by analysis of binding motifs', *Trends Genet* 25(10): 434-40.

Xi, H., Shulha, H. P., Lin, J. M., Vales, T. R., Fu, Y., Bodine, D. M., McKay, R. D., Chenoweth, J. G., Tesar, P. J., Furey, T. S. et al. (2007) 'Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome', *PLoS Genet* 3(8): e136.

Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M. and Lander, E. S. (2007) 'Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites', *Proc Natl Acad Sci U S A* 104(17): 7145-50.

Xu, C. R., Cole, P. A., Meyers, D. J., Kormish, J., Dent, S. and Zaret, K. S. (2011) 'Chromatin "prepattern" and histone modifiers in a fate choice for liver and pancreas', *Science* 332(6032): 963-6.

Yang, Z., Jiang, H., Chaichanasakul, T., Gong, S., Yang, X. W., Heintz, N. and Lin, S. (2006) 'Modified bacterial artificial chromosomes for zebrafish transgenesis', *Methods* 39(3): 183-8.

Yergeau, D. A., Kuliyev, E. and Mead, P. E. (2007) 'Injection-mediated transposon transgenesis in Xenopus tropicalis and the identification of integration sites by modified extension primer tag selection (EPTS) linker-mediated PCR', *Nat Protoc* 2(11): 2975-86.

Zaret, K. (1999) 'Developmental competence of the gut endoderm: genetic potentiation by GATA and HNF3/fork head proteins', *Dev Biol* 209(1): 1-10.

Zaret, K. S., Caravaca, J. M., Tulin, A. and Sekiya, T. (2010) 'Nuclear mobility and mitotic chromosome binding: similarities between pioneer transcription factor FoxA and linker histone H1', *Cold Spring Harb Symp Quant Biol* 75: 219-26.

Zaret, K. S. and Carroll, J. S. (2011) 'Pioneer transcription factors: establishing competence for gene expression', *Genes Dev* 25(21): 2227-41.

Zecchin, E., Mavropoulos, A., Devos, N., Filippi, A., Tiso, N., Meyer, D., Peers, B., Bortolussi, M. and Argenton, F. (2004) 'Evolutionary conserved role of ptf1a in the specification of exocrine pancreatic fates', *Dev Biol* 268(1): 174-84.

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A. and Levine, M. (2007) 'Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo', *Genes Dev* 21(4): 385-90.

Zentner, G. E., Tesar, P. J. and Scacheri, P. C. (2011) 'Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions', *Genome Res* 21(8): 1273-83.

Zhang, H., Roberts, D. N. and Cairns, B. R. (2005) 'Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss', *Cell* 123(2): 219-31.

Zhang, Z. and Gerstein, M. (2003) 'Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements', *J Biol* 2(2): 11.

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', *Biochim Biophys Acta* 1676(1): 41-50.

Zheng, R. and Blobel, G. A. (2010) 'GATA Transcription Factors and Cancer', *Genes Cancer* 1(12): 1178-88.

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. and Furlong, E. E. (2009) 'Combinatorial binding predicts spatio-temporal cis-regulatory activity', *Nature* 462(7269): 65-70.

Chapter Ten: LIST OF PUBLICATIONS

Robin Andersson*, Claudia Gebhard*, **Irene Miguel-Escalada**, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A. Maxwell Burroughs, J. Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J. Mungall, Terrence F. Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O. Daub,Peter Heutink, David A. Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R.R. Forrest, Piero Carninci, Michael Rehli, Albin Sandelin. An atlas of active enhancers across human cell types and tissues. ***Nature*** 2014 Mar 27;507(7493):455-61. doi: 10.1038/nature12787

Lorenzo Pasquali*, Kyle J. Gaulton*, Santiago A. Rodríguez-Seguí*, Loris Mularoni, **Irene Miguel-Escalada**, İldem Akerman, Juan J. Tena, Carlos Gómez-Marín, Martijn van de Bunt, Joan Ponsa-Cobas, Ignasi Morán, Natalia Castro, Takao Nammo, Inês Cebola, Javier García-Hurtado, Miguel Angel Maestro, François Pattou, Lorenzo Piemonti, Thierry Berney, Anna Gloyn, Philippe Ravassard, José Luis Gómez Skarmeta, Ferenc Müller, Mark I. McCarthy, Jorge Ferrer. Pancreatic islet epigenomics reveals enhancer clusters that are enriched in Type 2 diabetes risk variants. ***Nature Genetics*** 2014 Feb;46(2):136-43. doi: 10.1038/ng.2870

Jennifer Anne Roberts*, **Irene Miguel-Escalada***, Katherine Joan Slovik, Kathleen Theodora Walsh, Yavor Hadzhiev, Remo Sanges, Elia Stupka, Elizabeth Kate Marsh, Jorune Balciuniene, Darius Balciunas, Ferenc Müller. Targeted transgene integration overcomes variability of position effects in zebrafish. ***Development*** 2014 Feb;141(3):715-24. doi: 10.1242/dev.100347

***** These authors contributed equally.