CHAPTER 1 INTRODUCTION

1.1. Overview of the pancreas

The pancreas is a glandular organ originating from two separate primordia, the dorsal and ventral buds that arise from either side of the distal foregut endoderm. The organ is made up of a variety of distinct cell types and has a mix of both exocrine and endocrine functions.

The exocrine gland serves as a digestive organ with its acinar cells performing gastrointestinal functions by secreting digestive enzymes and a duct system that allows these digestive enzymes to drain into the intestine.

The endocrine gland is arranged as cell clusters known as islets of Langerhans which functions in regulating blood glucose homeostasis and other hormone secretions. Each cluster comprises of multiple distinct cell types, each secreting unique hormones into the body's circulation (α-cells, glucagon to increase blood glucose; β-cells, insulin to decrease blood glucose; δ-cells, somatostatin which regulates α-cells and β-cells; ε-cells, ghrelin which stimulates hunger and functions as a neuropeptide in the central nervous system; and γ or pancreatic polypeptide (Rashid et al.)-cells, pancreatic polypeptide which regulates pancreatic secretions, hepatic glycogen levels and gastrointestinal secretions). (Figure 1, adapted from (Shih et al., 2013)).

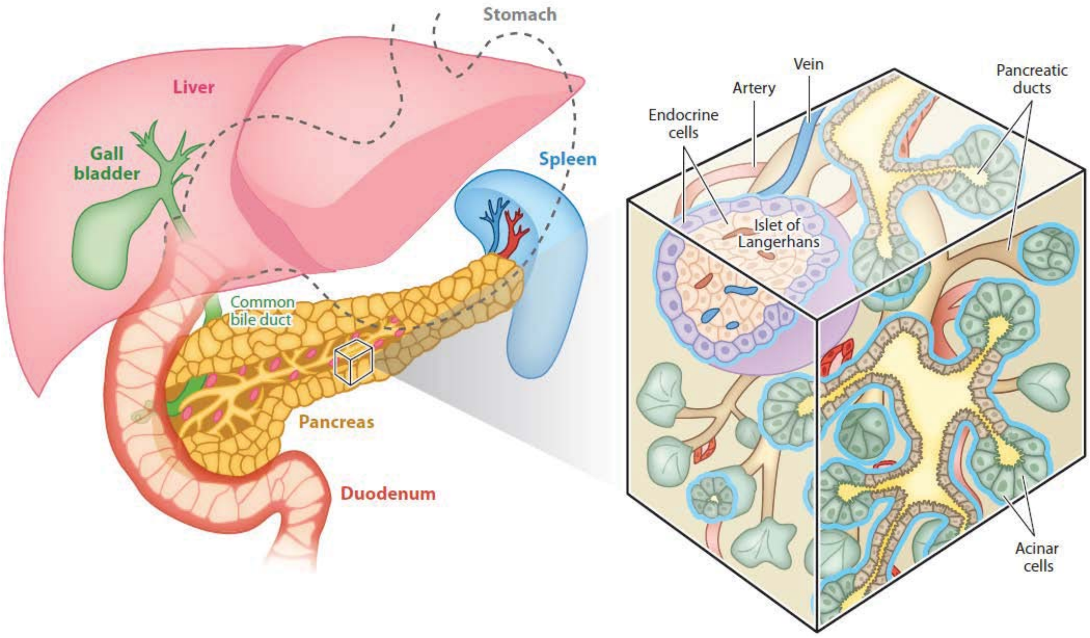
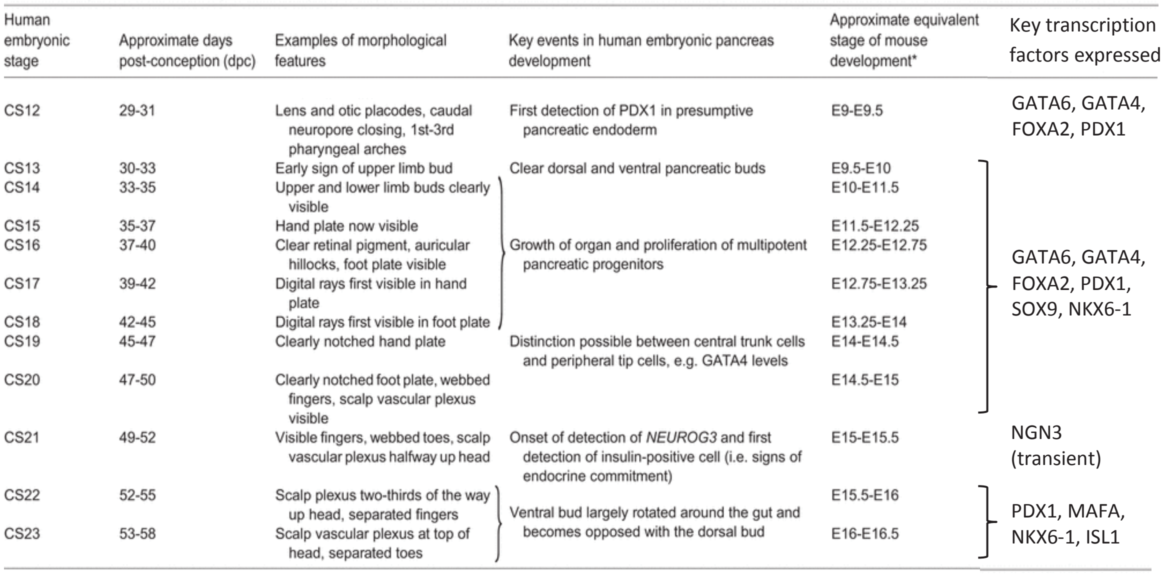


Figure 1. **The pancreas as a mixed exocrine and endocrine organ.** The mature pancreas lies behind the stomach and is adjacent to the duodenum. Acinar cells form the exocrine pancreas. The endocrine pancreas consists of small cell clusters, called islets of Langerhans, containing five endocrine cell types. Adapted from (Shih et al., 2013).

1.1.1. Development of the human pancreas

Human embryogenesis spans from fertilisation to approximately 8 weeks post-conception. After which, the embryo is referred to as a foetus. During the development of the embryo, specification of the three germ layers: ectoderm, mesoderm and endoderm occur, from which all adult tissues are formed. A recent publication by O'Rahilly and Müller has proposed a staging classification covering embryonic development (O'Rahilly and Müller, 2010). Based on a morphological scheme and staged by extension of time i.e. days post-conception (dpc), human embryonic development was divided into 23 different Carnegie Stages (CS). The key developmental stages of the pancreas during human embryogenesis, along with the approximate equivalent stage of mouse development, are mapped onto the 23 different CS stages (CS12 to CS23) (Table 1).



Human embryonic stage	Approximate days post-conception (dpc)	Examples of morphological features	Key events in human embryonic pancreas development	Approximate equivalent stage of mouse development*	Key transcription factors expressed
CS12	29-31	Lens and otic placodes, caudal neuropore closing, 1st-3rd pharyngeal arches	First detection of PDX1 in presumptive pancreatic endoderm	E9-E9.5	GATA6, GATA4, FOXA2, PDX1
CS13	30-33	Early sign of upper limb bud	Clear dorsal and ventral pancreatic buds	E9.5-E10	GATA6, GATA4, FOXA2, PDX1, SOX9, NKX6-1
CS14	33-35	Upper and lower limb buds clearly visible	Growth of organ and proliferation of multipotent pancreatic progenitors	E10-E11.5	
CS15	35-37	Hand plate now visible		E11.5-E12.25	
CS16	37-40	Clear retinal pigment, auricular hillocks, foot plate visible		E12.25-E12.75	
CS17	39-42	Digital rays first visible in hand plate		E12.75-E13.25	
CS18	42-45	Digital rays first visible in foot plate		E13.25-E14	
CS19	45-47	Clearly notched hand plate	Distinction possible between central trunk cells and peripheral tip cells, e.g. GATA4 levels	E14-E14.5	
CS20	47-50	Clearly notched foot plate, webbed fingers, scalp vascular plexus visible		E14.5-E15	
CS21	49-52	Visible fingers, webbed toes, scalp vascular plexus halfway up head	Onset of detection of *NEUROG3* and first detection of insulin-positive cell (i.e. signs of endocrine commitment)	E15-E15.5	NGN3 (transient)
CS22	52-55	Scalp plexus two-thirds of the way up head, separated fingers	Ventral bud largely rotated around the gut and becomes opposed with the dorsal bud	E15.5-E16	PDX1, MAFA, NKX6-1, ISL1
CS23	53-58	Scalp vascular plexus at top of head, separated toes		E16-E16.5	

Table 1. Stages of human pancreas development and their respective Carnegie Stages (CS). CS stages are shown together with their estimates of corresponding days post-conception (dpc). Key events in human embryonic pancreas development along with the approximate equivalent stage of mouse development are mapped to the CS stages. Table edited and adapted from O'Rahilly and Müller (2010).

Pancreas organogenesis is a highly complex and orchestrated process, comprising of coordinated signalling events that occur in a step-wise manner, as well as transcriptional networks that result in a cascade of transcription factors driving pancreatic specification (Figure 2 adapted from (Jennings et al., 2015)).

Pancreas induction occurs at CS9, where the definitive endoderm (DE) maintains communication with the visceral endoderm of the yolk sac (Jennings et al., 2013), and ventral and dorsal thickenings of the epithelial cells in the distal foregut occurs (Piper et al., 2002, Piper et al., 2004).

At CS10, endodermal folding gives rise to the foregut and hindgut, thus restricting the opening of the yolk sac to the intervening midgut (Jennings et al., 2013). The anterior end of this opening, known as the anterior intestinal portal (AIP), constitutes the foregut-midgut boundary and is the site of pancreatic specification. In other species, early specification of the pancreas within the gut endoderm occurs in the absence of sonic hedgehog (Shh) signalling (Apelqvist et al., 1997, Kim and Melton, 1998, Hebrok et al., 2000). In chick embryos, Activin secreted from the notochord and the close proximity of the resulting dorsal foregut endoderm to it causes exclusion of Shh expression, allowing for the expression of key transcription factor pancreatic and duodenal homeobox factor 1 (*Pdx1*) (Kim and Melton, 1998, Hebrok et al., 1998). In humans, patterning was similarly observed where *PDX1* was first detected at CS12, even though SHH could still be detected at CS10, which suggests a slightly later timing for the exclusion of SHH in humans (Jennings et al., 2013). The dorsal foregut endoderm subsequently develops into the dorsal pancreatic bud. One difference worth noting that was observed between human embryos and mouse or chick embryos is that in humans, there has been no detection of early pancreatic endocrine differentiation (Villasenor et al., 2008, Jennings et al., 2013), whereas early pro-endocrine patterning has been observed in mouse and chick (Lammert et al., 2001, Bonal and Herrera, 2008). This observation could possibly be explained by the lack of proximity of the paired dorsal aortae to the early pancreatic endoderm, thus reducing the opportunity for early pro-endocrine patterning.

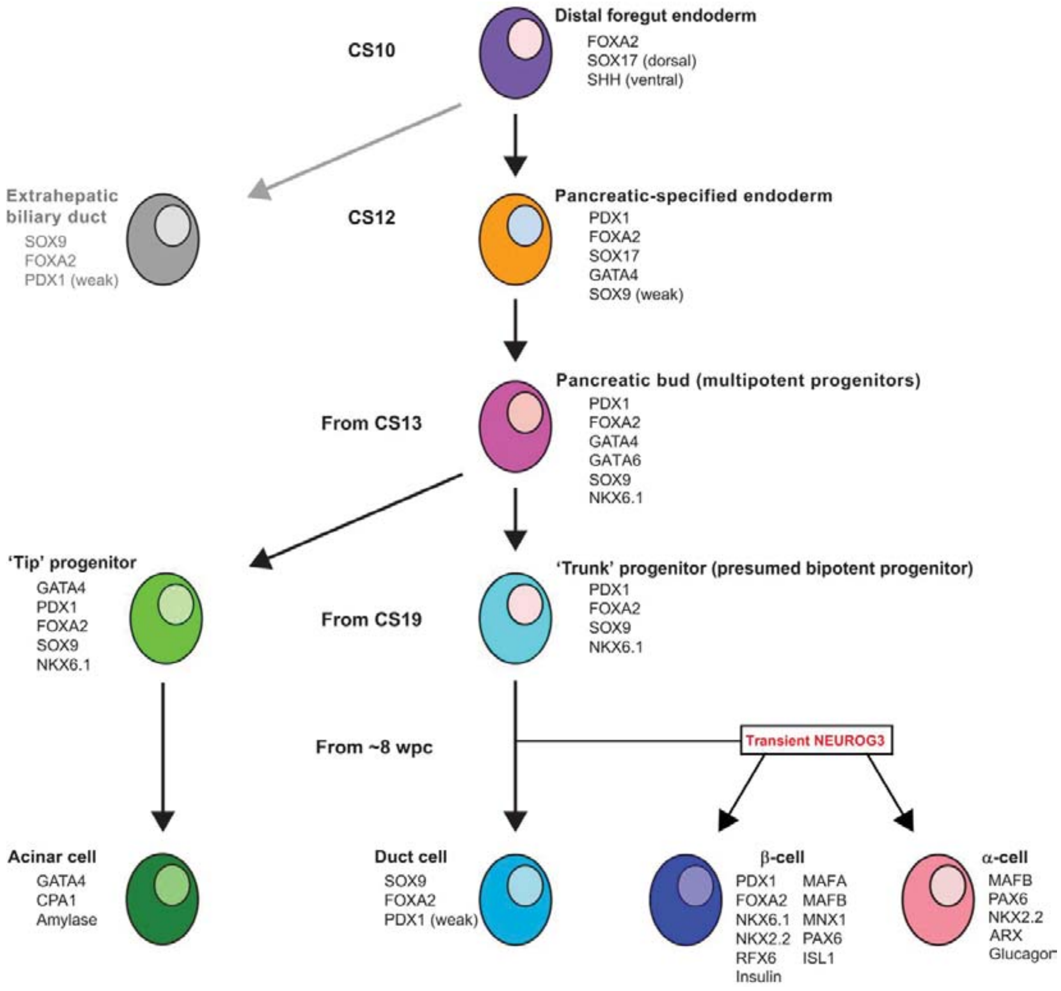


Figure 2. Developmental stages of human pancreas development and its respective transcription factor network. Illustration of transcription factors and key markers that identify the various key stages of early pancreas specification in the dorsal pancreas and commitment to subsequent lineages.

At CS13, the dorsal and ventral pancreatic buds are clearly defined and are marked by the transcription factors SRY (sex-determining region Y)-box 9 (*SOX9*), *PDX1*, GATA binding protein 4 (*GATA4*) and *GATA6* (Piper et al., 2004, Jennings et al., 2013), all of which play a pivotal role in promoting human pancreatic growth (Stoffers et al., 1997, Piper et al., 2002, Allen et al., 2012, Shaw-Smith et al., 2014). The human pancreas continues its expansion of proliferative progenitor cells for the remainder of the embryonic period. Another notable difference between the human and mouse pancreatic development is that transcription factor Nirenberg and Kim homeobox factor (NKX) 2.2 (*Nkx2-2*) is detected in these progenitor cells of the mouse but not human (Jennings et al., 2013).

At CS19, divergence into “tip” or “trunk” progenitor cells is marked by the expression levels of *GATA4* (Jennings et al., 2013). “Trunk” cells, which commit to central duct-like structures, express lower levels of *GATA4* as compared to the more peripheral clustered pro-acinar “tip” cells, even though both progenitor cell types express several common pancreatic markers such as *PDX1*, *SOX9* and *NKX6-1* (Figure 2). A similar separation of acinar cells was also observed in the mouse (Esni et al., 2004, Solar et al., 2009, Schaffer et al., 2010), although *Sox9* is lost more promptly in these peripheral tip cells in mouse (Schaffer et al., 2010) than in humans, where the loss of *SOX9* is delayed to between 10 and 14 weeks post-conception (wpc) (Jennings et al., 2013).

The commencement of endocrine specification is marked by the transient expression of transcription factor neurogenin 3 (*NGN3*) (Figure 2). In the mouse, *Ngn3* is transiently expressed to enable progenitor cells within the central duct-like structures to commit into the endocrine lineage (Gradwohl et al., 2000, Schwitzgebel et al., 2000, Gu et al., 2002). In humans, *NGN3* expression is detected at CS21 (8 wpc) around the end of the first trimester of human pregnancy upon the formation of foetal β-cells, which function as true endocrine cells by this time (Piper et al., 2004, Lyttle et al., 2008). The peak expression level of *NGN3* occurs at 10 to 14 wpc and declines at and after 18 wpc, and diminishes in human foetuses after 35 wpc (Salisbury et al., 2014). By contrast, *SOX9* is absent in cells robustly expressing *NGN3* and continues to be absent in subsequent endocrine cells, but is present in pancreatic duct cells (Jennings et al., 2013). By 10 wpc, β-cells are well vascularised and at 12-13 wpc, islets containing α-cells, β-cells, δ-cells and γ-cells are apparent (Piper et al., 2004, Jennings et al., 2013).

1.1.2. Diabetes Mellitus as a pancreatic disease

Diabetes Mellitus ('diabetes') represents a family of metabolic disease caused primarily by dysfunction in the pancreas. Diabetes is a growing health problem worldwide. The global prevalence of diabetes has nearly doubled since 1980, rising from 4.7% to 8.5% in the adult population and causing 1.5 million deaths in 2012 (World Health Organisation, 2016). It is estimated that 366 million people were diagnosed with diabetes in 2011; by 2030 this would have risen to 552 million (International Diabetes Federation, 2011). The need to understand human pancreas development is, therefore, critical because of its relevance to the different types of diabetes and therapies for this disease. There are multiple forms of diabetes, such as polygenic forms Type 1 diabetes (T1D), Type 2 diabetes (T2D), and monogenic forms maturity onset diabetes of the young (MODY), and neonatal diabetes mellitus (NDM). Polygenic forms of diabetes i.e. T1D and T2D make up an overwhelming majority (98%) of diabetic cases and its risk is related to multiple genes. Rare, monogenic forms of diabetes such as MODY and NDM result from mutations in a single gene and account for about 1 to 5% of all cases of diabetes in young people. Gene mutations in monogenic diabetes can either be inherited or spontaneous (*de novo*).

T1D, or juvenile-onset diabetes, accounts for approximately 5-10% of diabetic patients and is caused by the chronic autoimmune destruction of insulin-secreting β-cells, usually leading to complete insulin deficiency or hyperglycaemia (Anderson and Bluestone, 2005). Hyperglycaemia occurs when blood glucose levels are high, and this can lead to serious health conditions such as ketoacidosis, kidney failure, heart disease, stroke, and blindness. Despite being able to affect children and adult with normal weight, the childhood onset is most prevalent. Although the main effector mechanism of T1D is clearly an autoimmune reaction, T1D is also suggested to result from genetic susceptibility and/or environmental triggers (reviewed in (Atkinson and Eisenbarth, 2001, Van Belle et al., 2011)). T1D can be fatal if not treated with exogenous insulin to compensate for the lack of insulin production by the body.

T2D, or adult-onset diabetes, on the other hand, accounts for approximately 90% of diabetic patients and is usually associated with obesity or older age. T2D is characterised by insulin resistance, where insulin-sensitive target tissues such as the pancreatic β-cells, liver, muscle or adipose cells do not respond adequately to normal levels of insulin produced by intact β-cells. Consequently, this leads to disruption of the pancreatic β-cell function and decreased β-cell mass. Although T2D is most prevalent in adults, there has been increasing evidence of this form of diabetes affecting younger individuals (Fajans et al., 2001). T2D is a polygenic disease influenced by many environmental and behavioural risk factors. Thus, it has not only been challenging to identify the underlying genetic causes for this disease, but also to devise universal therapeutic strategies. No cure has yet been found for the disease. Several forms of treatment that currently exist, especially for obese patients; include lifestyle modifications, treatment of obesity, oral hypoglycaemic drugs, and insulin sensitizers such as metformin that reduces insulin resistance.

MODY was first recognised by Tattersall (Tattersall, 1974, Tattersall and Fajans, 1975) and is characterised by autosomal dominant inheritance, onset typically before 25 years of age and β-cell dysfunction leading to hyperglycaemia. The prevalence of MODY is higher than NDM, accounting for about 1% of people with diabetes in Europe (Frayling et al., 2001, Ledermann, 1995). Many MODY genes are pancreatic development transcription factors with the exception of glucokinase (*GCK*), acinar cell carboxyl ester lipase (*CEL*) and insulin (*INS*). Common MODY genes include MODY1 (Hepatic Nuclear Factor 4 Alpha; *HNF4A*-MODY), MODY2 (*GCK*-MODY), MODY3 (Hepatic Nuclear Factor 1 Alpha; *HNF1A*-MODY) which account for 70% of MODY cases, and MODY5 (Hepatic Nuclear Factor 1 Beta; *HNF1B*-MODY).

NDM is a rare, genetically heterogeneous monogenic form of diabetes occurring in approximately 1 in 200,000 live births (Stanik et al., 2007, Kanakatti Shankar et al., 2013). It presents in the days and weeks after birth, and almost always before 6 months of age (Iafusco et al., 2002, Edghill et al., 2006). NDM may be transient or permanent. In transient NDM (TNDM), the condition remits during infancy but may reappear later in life whereas in permanent NDM (PNDM), hyperglycaemia persists during life. Approximately 50% of NDM patients have heterozygous activating mutations in the potassium inwardly-rectifying channel, subfamily J, member 11 (*KCNJ11*) and ATP-binding cassette, sub-family C, member 8 (*ABCC8*) genes encoding the adenosine triphosphate-sensitive potassium channel subunits (De Franco et al., 2015). Failure of the potassium channels to close appropriately in response to rising glucose, thus inhibiting the release of insulin from β-cells, leads to diabetes. Sulfonylurea therapy permits insulin secretion through closure of the channel (Pearson et al., 2006, Rafiq et al., 2008).

Other rare gene mutations leading to monogenic NDM include eukaryotic translation initiation factor 2 alpha kinase 3 (*EIF2AK3*), forkhead box P3 (*FOXP3*), *GATA6*, GLIS family zinc finger 3 (*GLIS3*), neuronal differentiation 1 (*NEUROD1*), *NGN3*, *PDX1*, pancreas specific transcription factor 1a (*PTF1A*), regulatory factor X6 (*RFX6*) and methylation defects at chromosome 6q24.

Studies of rare monogenic diseases provide an invaluable opportunity to learn about underlying molecular mechanisms, thereby contributing significantly to our understanding of the molecular genetic basis of common, complex diseases (Antonarakis and Beckmann, 2006).

1.1.3. Pancreatic agenesis

Congenital pancreatic agenesis is an extremely rare cause of NDM with a prevalence of less than 1/1 000 000 and around 50 cases being reported in the literature so far. It is caused by an impaired formation of the pancreas during embryonic development. Morphologically, the pancreas can either be totally absent or extremely reduced in size (pancreatic hypoplasia).

Clinically, pancreatic agenesis is defined as insulin dependent neonatal diabetes diagnosed before 6 months of age and pancreatic exocrine insufficiency requiring enzyme replacement therapy. Patients affected by pancreatic agenesis usually present with intrauterine growth retardation (IUGR) as a result of reduced insulin secretion *in utero* and are diagnosed with hyperglycaemia in the first days of life. Patients with pancreatic agenesis usually require insulin treatment.

Diagnosis of pancreatic agenesis can be made by imaging (MRI or ultrasound) showing reduction or absence of pancreatic tissue, measurement of fecal elastase which is often undetectable in patients with pancreatic agenesis as a result of exocrine dysfunction, or clinically by the presence of insulin-dependent neonatal diabetes and exocrine insufficiency requiring enzyme replacement therapy.

A genetic diagnosis is also possible for over 80% of patients with pancreatic agenesis and transcription factor *GATA6* has recently been identified to be a major cause of pancreatic agenesis (Lango Allen et al., 2012). In these patients, pancreatic agenesis is commonly associated with other extrapancreatic malformations such as cardiac malformation, neurocognitive defects, hypothyroidism, gut abnormalities and gallbladder agenesis/biliary atresia (De Franco et al., 2013). These defects affect organs of endodermal origin, suggesting a defect in early embryonic differentiation.

1.2. Human pluripotent stem cells as an *in vitro* system to model the development of the human pancreas

Stem cells are cells with unique properties such as the capacity to self-renew indefinitely and the ability to differentiate into many diverse cell types. Being pluripotent, these cells are able to differentiate into all derivatives of the three primary germ layers, namely endoderm, mesoderm and ectoderm (Evans and Kaufman, 1981). This is in contrast to adult stem cells which are multipotent and more restrictive in their differentiation to various cell types (Suda et al., 1987, Zwaka and Thomson, 2005). With advances in pluripotent stem cell (PSC) technology, a large number of stem cells can now be expanded and maintained *in vitro* whilst retaining their unique properties (Suda et al., 1987, Solter, 2006). This allows for studies that were once difficult using primary tissues or biopsies to progress. In the clinical setting, human PSCs (hPSCs) bring vast potential in providing opportunities for treating and curing diseases.

hPSCs can be broadly categorised into two categories; embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs). ESCs were first derived from the inner cell mass (ICM) of the mouse embryo at the early post-implantation blastocyst stage by Evans and Kaufman, and Martin in 1981 (Evans and Kaufman, 1981, Martin, 1981). The ICM in the blastocyst is a transient pluripotent pool of cells that rapidly differentiates during gastrulation into the primary germ layers. They can be maintained indefinitely *in vitro* in their pluripotent state in the presence of cytokine leukaemia inhibitory factor (LIF) on a layer of mitotically inactivated mouse embryonic fibroblasts (MEF) feeder cells (Smith et al., 1988, Williams et al., 1988).

There was a significant lag before the first human ESC (hESC) line was successfully isolated from human blastocysts in 1998 (Thomson et al., 1998). This was largely due to the fact that human embryos were much more difficult to obtain than mouse embryos and the ethical dilemmas that accompanied it. Prior to this, the first primate ESC line from the blastocyst of a rhesus monkey was also isolated and successfully derived (Thomson et al., 1995). The derivation of hESCs paved the way to an accelerated expansion on stem cell research.

Approximately a decade later, pioneering studies describing human iPSCs (hiPSCs) emerged (Takahashi and Yamanaka, 2006, Takahashi et al., 2007, Gurdon and Melton, 2008). These studies showed that by introducing a cocktail of four specific transcription factors (SOX2, KLF4, c-MYC and OCT4) *ex-vivo*, differentiated fibroblasts could be converted to a pluripotent state resembling ESCs derived from the blastocyst ICM. Like ESCs, iPSCs also had the ability to form teratomas in mice (Wernig et al., 2007). iPSC technology has the advantage over ESCs in that it was able to circumvent the ethical issues associated with human embryos. Furthermore, since iPSCs can be derived from patients' cells, it brings with it the potential application of excluding immunosuppression treatments that are required in conventional cell replacement or transplantation therapies to prevent tissue rejection when cells are transferred between genetically different individuals. iPSC is therefore a robust and ethical way of re-programming differentiated cells to a pluripotent state. Similarly to ESCs, the iPSCs can then be directed, by growth factors important and specific for development and differentiation, to form functional differentiated cells of a variety of lineages. It has been suggested that hiPSCs and hESCs are functionally equivalent since they both utilise similar signalling pathways to maintain pluripotency and drive differentiation, and the mechanisms controlling the early cell fate decision of these pluripotent stem cells are similar (Vallier et al., 2009a).

Since their discovery, both hESCs and hiPSCs have proven to be powerful tools in biomedical research, overcoming ethical limitations in human embryonic development studies where access to human embryos is scarce. hPSCs have the potential to be used in disease models for studying the molecular basis of diseases, including genetically inherited human diseases (Yusa et al., 2011). It brings tremendous potential not only in disease modelling, but also in regenerative medicine, cell replacement therapy, drug testing and targeted gene-repair strategies, such as homologous recombination to repair genetic defects (Figure 3). Thus, they serve as ideal model systems for human developmental scientific studies. This dissertation focuses on using hPSCs to model the development of the human pancreas by specifying the cells down the pancreatic lineage.

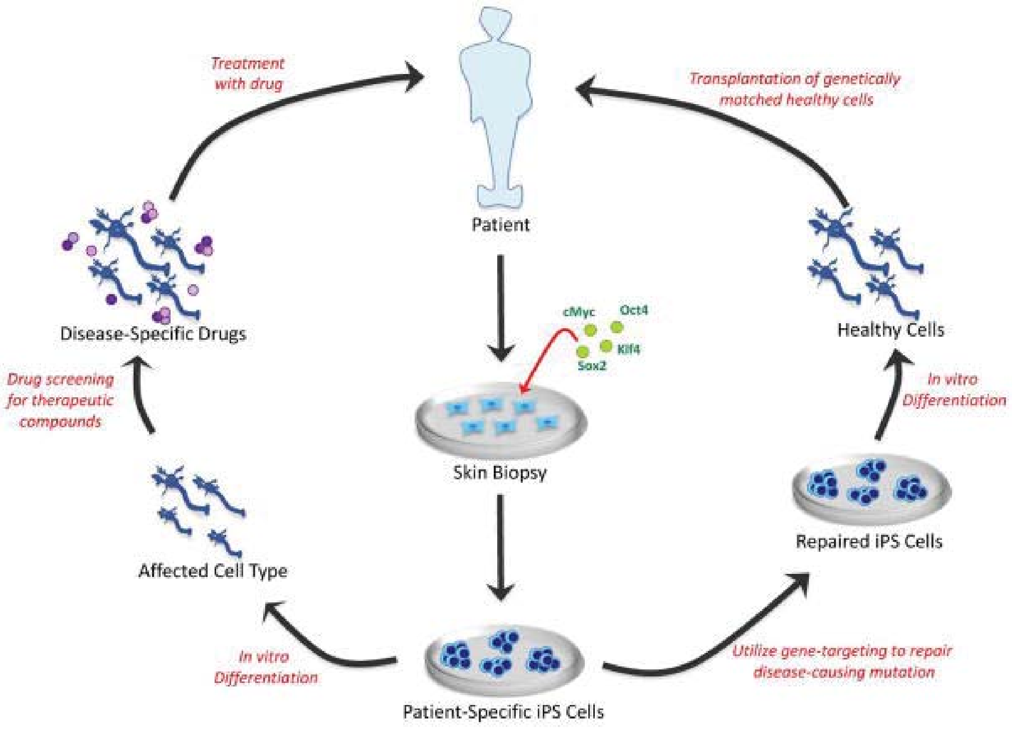

Treatment with drug
Patient
Transplantation of genetically matched healthy cells
Disease-Specific Drugs
cMyc
Oct4
Klf4
Sox2
Healthy Cells
Drug screening for therapeutic compounds
In vitro Differentiation
Skin Biopsy
Repaired iPS Cells
Affected Cell Type
In vitro Differentiation
Utilize gene-targeting to repair disease-causing mutation
Patient-Specific iPS Cells

Figure 3. Applications of hiPSCs. hiPSCs have the potential to be used to model and treat human disease, in this case, via drug testing or targeted gene-repair strategies. Patient-specific hiPSCs are derived from co-transfection of pluripotency-inducing transcription factors in cells isolated from a skin biopsy. The hiPSCs are differentiated *in vitro* into the affected cell type where they could be used in a drug screen to test for suitable drugs for treatment of the disease. Alternatively, the disease-causing mutation is corrected and repaired hiPSCs are differentiated *in vitro* into healthy cells of the affected cell type, and the genetically matched cells are subsequently transplanted into the patient. Adapted from (Robinton and Daley, 2012).

1.2.1 Pancreatic specification protocols to date

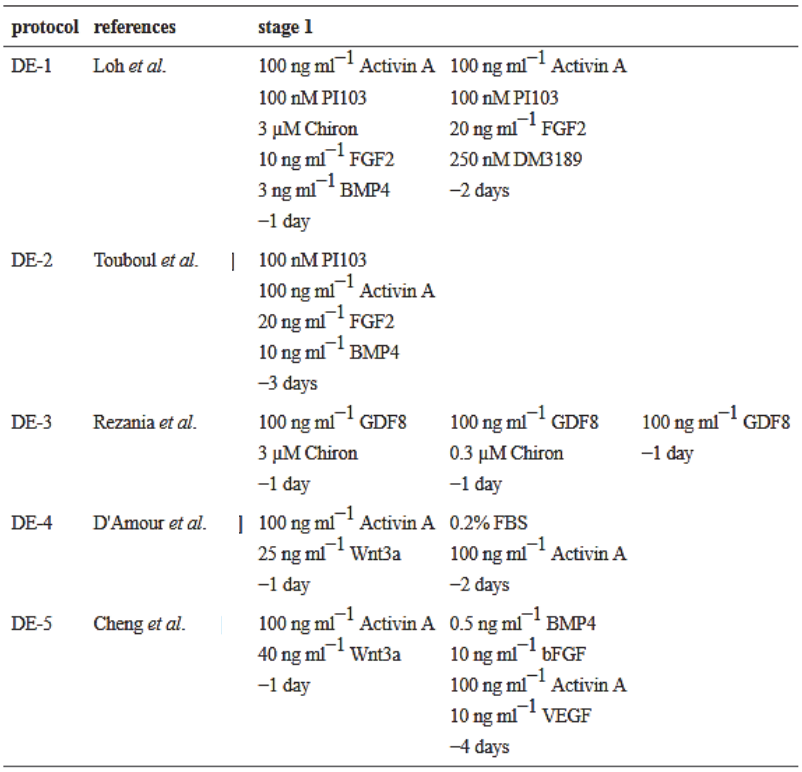
Pancreatic differentiation was first published by Soria *et al.* in a landmark study reporting the successful differentiation of ESCs into insulin-producing cells (Soria et al., 2000). In this study, mouse ESCs constructed to contain a neomycin resistance gene under the control of the human insulin promoter were able to correct hyperglycaemia a week upon implantation into streptozotocin-induced diabetic mice. A subsequent study described the specification of mouse ESCs into the definitive endoderm (DE) in the presence of activin A and absence of fetal bovine serum, establishing the first stepping stone for directed differentiation to many organ systems (Kubo et al., 2004), namely the digestive and respiratory tracts, thyroid, liver, and pancreas. The high levels of activin A mimics the action of Nodal, a ligand for transforming growth factor-β (TGF-β) superfamily, which activates downstream signalling pathways and transcriptional networks that regulates the formation of the DE (Lowe et al., 2001, Champeris Tsaniras and Jones, 2010). Later, hESCs were also efficiently differentiated into DE using elevated concentration of activin A (D'Amour et al., 2005).

Shortly after, a comprehensive stepwise pancreatic specification protocol was introduced, describing the differentiation of hESCs to endocrine cells capable of synthesising pancreatic hormones such as insulin, glucagon, somatostatin, pancreatic polypeptide and ghrelin with the use of specific growth factors and chemical compounds (D'Amour et al., 2006). With this *in vitro* differentiation protocol, the cells mimic *in vivo* pancreas organogenesis by being directed through stages resembling DE, gut-tube endoderm, pancreatic endoderm and endocrine precursor, thus recapitulating the major stages of normal pancreatic endocrine development. Each stage is recognised by the expression of specific markers. One striking difference between *in vitro* differentiation protocols and *in vivo* pancreas organogenesis is the duration, which are 2-3 weeks and 12-13 weeks respectively. The cells produced from this protocol have an insulin content almost mimicking that of adult islets, and released C-peptide in response to various secretagogues, but only minimally to glucose. The presence of immature polyhormonal cells e.g. insulin and glucagon, or insulin and somatostatin double-positive cells, however, suggested a lack of precision in the endocrine specification, as mature β-cells solely secrete insulin.

The publication of an improved serum-free protocol by *Jiang et al.* which involved activin A, retinoic acid (RA), basic fibroblast growth factor (bFGF) and nicotinamide to promote pancreatic differentiation resulted in islet-like structures with distinct insulin-, glucagon-, and somatostatin-positive mono-hormonal cells (Jiang et al., 2007b). The protocol, composed of four stages (definitive endoderm induction, pancreatic endoderm formation, endocrine induction, islet-like cluster maturation) generated about 24% *PDX1*-positive cells and 4% C-peptide-positive cells. These cells were insulin-producing and responsive to fluctuations in glucose levels in a suspension cell culture system but showed low levels of response when cultured in adherence.

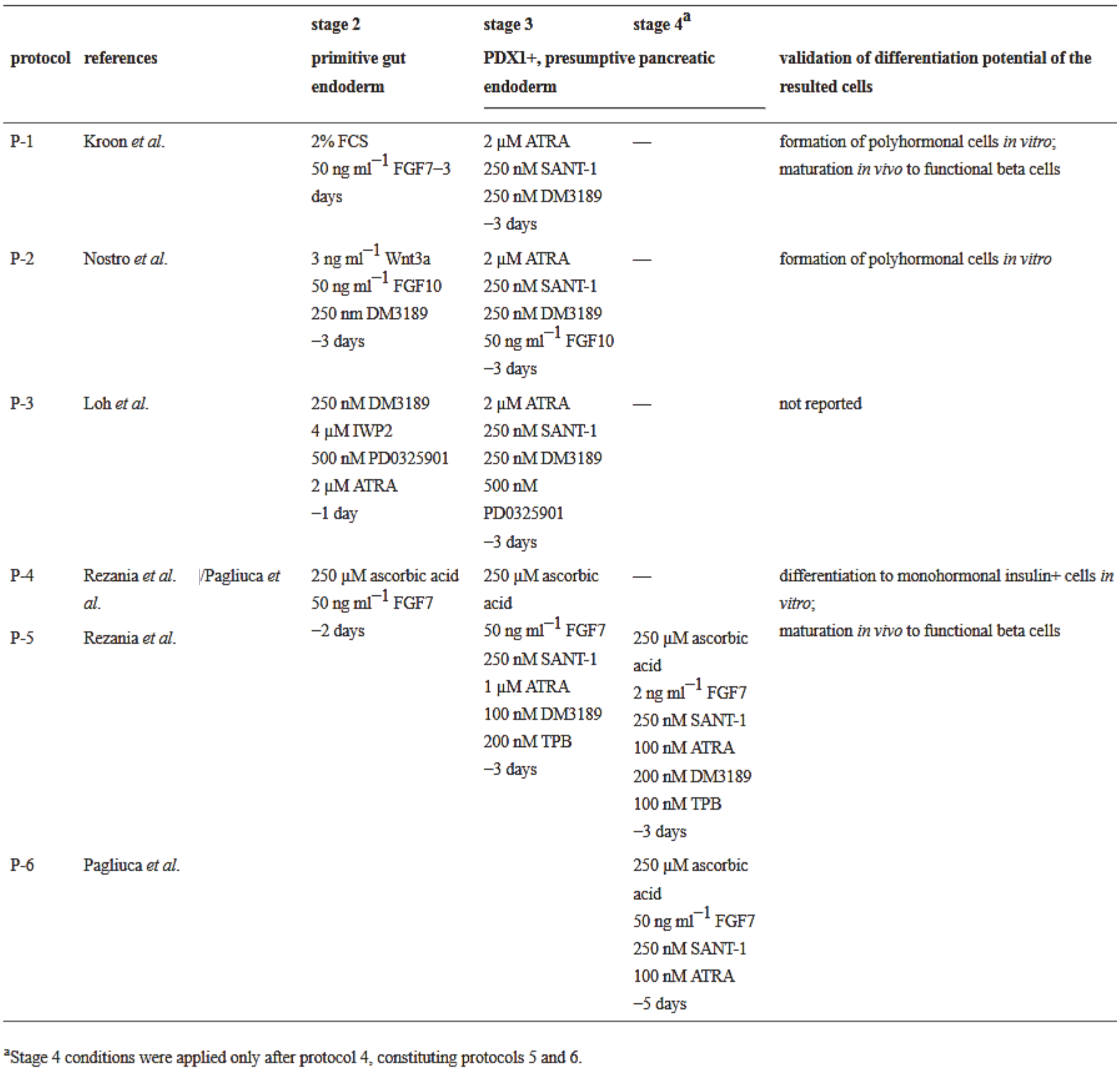
These studies were succeeded by numerous modified variations of pancreatic differentiation protocols (Jiang et al., 2007a, Kroon et al., 2008, Zhang et al., 2009, Cai et al., 2010, Nostro et al., 2011, Loh et al., 2014, Pagliuca et al., 2014, Rezania et al., 2014, Russ et al., 2015, Cho et al., 2012). With the introduction of such a wide variety of different protocols coupled with the ability to generate hPSC lines from healthy individuals or patients with different genetic backgrounds, it is not unexpected that reports on variations in differentiation efficiencies due to the use of different protocols cell lines have arisen (Osafune et al., 2008, Chin et al., 2009). A recent study that closely compared protocol variations at both the DE (Table 2) and pancreatic progenitor (Table 3) stages, and variation in lineage propensity among hPSC lines reported varying differentiation efficiencies between the tested hiPSC and hESC (H9) lines (Rostovskaya et al., 2015). Interestingly, the different protocols specifying pancreatic progenitors yielded no significant difference between the hPSC lines. Furthermore, it was also reported that certain protocols displayed higher endodermal and pancreatic differentiation efficiencies than others (Rostovskaya et al., 2015); two protocols that were the most recently published fared the best for pancreatic specification, generating over 90% *PDX1*-positive cells (Rezania et al., 2014, Pagliuca et al., 2014).

Although there have been substantial improvements in pancreatic differentiation protocols over the years, several challenges hindering the complete generation of functional β-cells that fully mimics those *in vivo* still remains. Transplantation of these hPSC-derived pancreatic progenitors into immunocompromised mice often resulted in the formation of teratomas, indicating the presence of pluripotent cells and incomplete differentiation. Efficient and consistently reproducible generation of pure pancreatic lineages derivatives is, therefore, key toward driving research in human β-cell biology, drug testing, disease modelling, development of cell replacement therapy and other applications. Understanding the molecular mechanisms promoting β-cell specification, including the studying the transcriptional regulatory networks of β-cell specification, will greatly contribute to the advancement of this field of research.



protocol	references	stage 1		
DE-1	Loh *et al.*	100 ng ml^{-1} Activin A 100 nM PI103 3 μM Chiron 10 ng ml^{-1} FGF2 3 ng ml^{-1} BMP4 –1 day	100 ng ml^{-1} Activin A 100 nM PI103 20 ng ml^{-1} FGF2 250 nM DM3189 –2 days	
DE-2	Touboul *et al.*	100 nM PI103 100 ng ml^{-1} Activin A 20 ng ml^{-1} FGF2 10 ng ml^{-1} BMP4 –3 days		
DE-3	Rezania *et al.*	100 ng ml^{-1} GDF8 3 μM Chiron –1 day	100 ng ml^{-1} GDF8 0.3 μM Chiron –1 day	100 ng ml^{-1} GDF8 –1 day
DE-4	D'Amour *et al.*	100 ng ml^{-1} Activin A 25 ng ml^{-1} Wnt3a –1 day	0.2% FBS 100 ng ml^{-1} Activin A –2 days	
DE-5	Cheng *et al.*	100 ng ml^{-1} Activin A 40 ng ml^{-1} Wnt3a –1 day	0.5 ng ml^{-1} BMP4 10 ng ml^{-1} bFGF 100 ng ml^{-1} Activin A 10 ng ml^{-1} VEGF –4 days	

Table 2. Comparison of definitive endoderm differentiation protocols for hPSCs. Table adapted from Rostovskaya et al., 2015 showing the different conditions of various protocols that were developed and published by independent groups.



protocol	references	stage 2 primitive gut endoderm	stage 3 PDX1+, presumptive pancreatic endoderm	stage 4^{a}	validation of differentiation potential of the resulted cells
P-1	Kroon *et al.*	2% FCS 50 ng ml^{-1} FGF7–3 days	2 µM ATRA 250 nM SANT-1 250 nM DM3189 –3 days	—	formation of polyhormonal cells *in vitro*; maturation *in vivo* to functional beta cells
P-2	Nostro *et al.*	3 ng ml^{-1} Wnt3a 50 ng ml^{-1} FGF10 250 nm DM3189 –3 days	2 µM ATRA 250 nM SANT-1 250 nM DM3189 50 ng ml^{-1} FGF10 –3 days	—	formation of polyhormonal cells *in vitro*
P-3	Loh *et al.*	250 nM DM3189 4 µM IWP2 500 nM PD0325901 2 µM ATRA –1 day	2 µM ATRA 250 nM SANT-1 250 nM DM3189 500 nM PD0325901 –3 days	—	not reported
P-4	Rezania *et al.* /Pagliuca *et al.*	250 µM ascorbic acid 50 ng ml^{-1} FGF7 –2 days	250 µM ascorbic acid 50 ng ml^{-1} FGF7 250 nM SANT-1 1 µM ATRA 100 nM DM3189 200 nM TPB –3 days	—	differentiation to monohormonal insulin+ cells *in vitro*; maturation *in vivo* to functional beta cells
P-5	Rezania *et al.*			250 µM ascorbic acid 2 ng ml^{-1} FGF7 250 nM SANT-1 100 nM ATRA 200 nM DM3189 100 nM TPB –3 days	
P-6	Pagliuca *et al.*			250 µM ascorbic acid 50 ng ml^{-1} FGF7 250 nM SANT-1 100 nM ATRA –5 days	

[a] Stage 4 conditions were applied only after protocol 4, constituting protocols 5 and 6.

Table 3. Comparison of pancreatic progenitor differentiation protocols for hPSCs. Table adapted from Rostovskaya et al., 2015 showing the different conditions of various pancreatic progenitor differentiation protocols that were developed and published by independent groups.

1.2.2 Transcription factors associated with pancreas development

Most biological processes are regulated on a transcriptional level. In mammalian cells, the transcription of genes is regulated by several regulatory proteins known as transcription factors (TFs). TFs recognise specific DNA sequences near the target gene, often in regulatory promoter regions that are located upstream of the transcriptional start site (TSS), and can either activate or repress these promoter regions. During islet cell development, TFs play an integral role in directing cell fates by regulating the transcriptional network controlling pancreatic specification and ultimately mature function (Figure 4). Some of the TFs that play a vital role in promoting β-cell function and identity are described below.

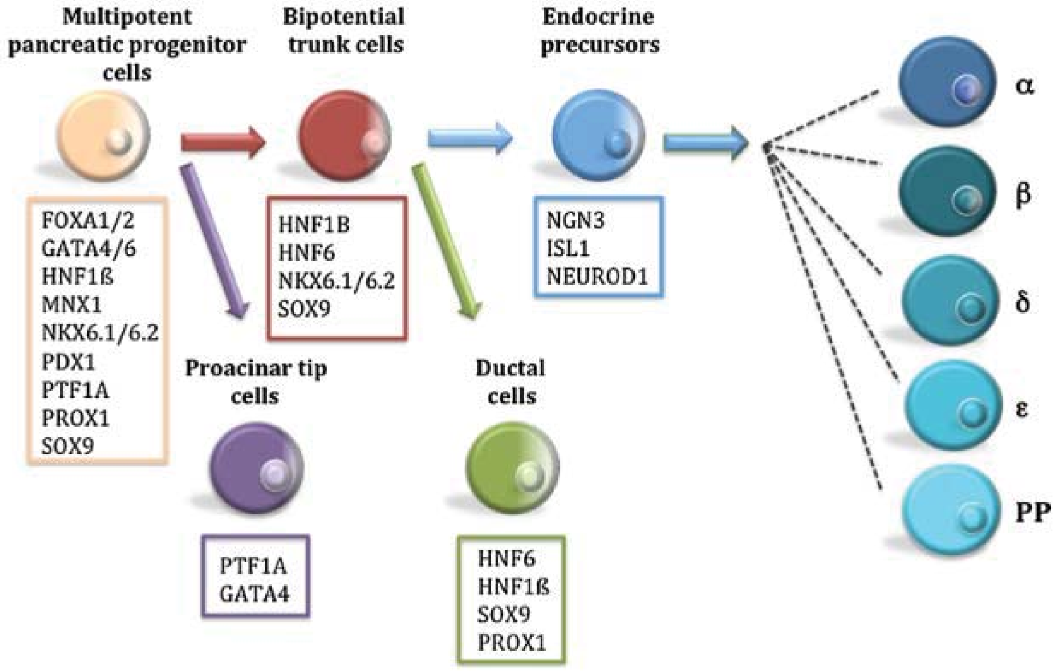


Figure 4. Key transcription factors controlling lineage specification during pancreatic development. Highlighted in this diagram are the key transcription factors known to have an integral role at each stage of pancreatic development. Adapted from (Cano et al., 2014).

SRY (sex determining region Y)-box 17 (*SOX17*)

The Sox family of TFs has a conserved DNA-binding HMG (high mobility group) domain (Bowles et al., 2000), and its early expression is necessary for endoderm formation. In *Xenopus*, *Xsox17* plays important roles in endoderm formation (Hudson et al., 1997, Clements and Woodland, 2000). It has also been shown to be a direct regulator of *FoxA1* and *FoxA2* (Sinner et al., 2004). β-catenin, an intracellular signal transducer in the Wnt signalling pathway, physically interacts with *Sox17* and promotes transcriptional activation of its target genes (Sinner et al., 2004).

Sox17/*SOX17* expression in the mouse and human is similar; it is expressed as an early endoderm marker immediately before 4 weeks post conception in human and e6.0 in mouse, then excluded from pancreatic cells 1 week or 2.5 days later in human and mouse respectively (Kanai-Azuma et al., 2002, Piper et al., 2004, Jennings et al., 2013). *Sox17* is required for the induction of *Pdx1* expression and *Sox17*$^{-/-}$ mice are deficient in the formation of the DE, although they form some foregut but not midgut or hindgut (Kanai-Azuma et al., 2002). Furthermore, constitutive expression of *SOX17* in hESCs produced stable definitive endoderm progenitors, while absence of *SOX17* blocked endoderm differentiation (Seguin et al., 2008).

Forkhead Box A2 (*FOXA2*)

Winged helix/forkhead transcription factors such as *FoxA2* have been reported to be necessary for DE formation (Dufort et al., 1998). In early human development, *FOXA2* is consistently expressed from week 4 (Lyttle et al., 2008, Jeon et al., 2009, Jennings et al., 2013). This is similar in the mouse where *FoxA2* is expressed throughout pancreatic development, and remains expressed in all mature pancreatic cell types of both mice and humans (reviewed in (Pan and Wright, 2011, Cano et al., 2014)). *FoxA1* and *FoxA2* both regulate the expression of key pancreatic gene *Pdx1* by co-occupying multiple regulatory domains in the *PDX1* gene, although this has not been verified in human (Gao et al., 2008, Pan and Wright, 2011). Compound conditional ablation of both *FoxA1* and *FoxA2* in mice resulted in complete loss of *Pdx1* expression and severe pancreatic hypoplasia (Gao et al., 2008). Forkhead genes of the FOXA class have also been proposed to interact with GATA factors during DE formation (Bossard and Zaret, 1998, Cirillo et al., 2002). *FoxA2* is also strongly expressed in DE derivatives such as the liver (Ang et al., 1993). In contrast to *Sox17*$^{-/-}$ mice, *FoxA2*$^{-/-}$ mice can form the hindgut but not the foregut and midgut (Dufort et al., 1998).

Hepatocyte nuclear factor 1 homeobox beta (*HNF1B*)

HNF1B is highly expressed in humans around 7 weeks post conception and remains expressed throughout pancreatic development (Jeon et al., 2009). In the mouse, *Hnf1B* is expressed in the foregut endoderm prior to the onset of *Pdx1* expression (e8), and later restricted to the epithelial trunk domain and exocrine ducts (Coffinier et al., 1999, Maestro et al., 2003, Haumaitre et al., 2005). *Hnf1B*$^{-/-}$ mice die before gastrulation due to defective visceral endoderm formation, but when the embryonic lethality was rescued by tetraploid aggregation, *Hnf1B*$^{-/-}$ mice displayed absence of the ventral pancreatic bud and an extremely reduced and transient dorsal bud that leads to pancreas agenesis by e13.5 (Haumaitre et al., 2005). In humans, heterozygous mutations in the *HNF1B* gene are associated with MODY5 (Nishigori et al., 1998, Lindner et al., 1999, Bingham et al., 2000, Horikawa et al., 1997). This is in contrast to the mouse where only homozygous mutations produced diabetes in mice.

PDX1

PDX1 is expressed in all pancreatic precursor cells and has been shown to be critically important for early pancreatic development (Bernardo et al., 2009). In humans, *PDX1* is broadly expressed at around 4 weeks post conception, peaking at a later stage when its expression becomes restricted to β-cells (Lyttle et al., 2008, Jennings et al., 2013). In mice, *Pdx1* is first expressed in the primitive gut tube at e8.5, marking the pre-pancreatic endoderm. *Pdx1* is not expressed exclusively in the pancreas and by e10.5, its expression has been reported in parts of the posterior foregut including the stomach, duodenum and bile duct (Guz et al., 1995, Offield et al., 1996). *Pdx1* high expression then becomes restricted mostly to endocrine cells in the pancreas just before birth (Guz et al., 1995, Offield et al., 1996, Stoffers et al., 1999). Growth of the pancreatic epithelium in *Pdx1*$^{-/-}$ mice is arrested around e10.5 despite the presence of initial budding (Jonsson et al., 1994, Offield et al., 1996, Ahlgren et al., 1996). Microarray analyses performed on e10.5 *Pdx1*$^{-/-}$ mutant mice embryos found downregulation of several TFs including *Nkx6-1* and *Ptf1A*, supporting its critical role in the pancreatic transcriptional network (Svensson et al., 2007). However, the direct regulation of *Pdx1* to *Nkx6-1* and *Ptf1A* has yet to be established, though this has been shown for other TFs such as *Gata4*, *FoxA2* and *Hnf1B* (Rojas et al., 2009, Oliver-Krasinski et al., 2009). In humans, homozygous inactivating mutations of the *PDX1* gene result in pancreatic agenesis, known as MODY4 (Stoffers et al., 1997).

PTF1A

PTF1A expression in the human foetal pancreas is only detectable by quantitative real time polymerase chain reaction (qRT-PCR) around mid-gestation, when acinar cells are formed (Jeon et al., 2009). From studies in mice, the separation process of pro-acinar tip cells and trunk cells is established by an antagonistic relationship between TFs *Nkx6-1* and *Ptf1A* (Schaffer et al., 2010). Thus, *Nkx6-1* and *Ptf1A* have important roles in specifying progenitors toward an endocrine or acinar fate, respectively. From being broadly expressed in the dorsal and ventral pancreatic buds, *PTF1A* is progressively restricted to the pro-acinar tip cells, while *NKX6-1* and other TFs such as *SOX9*, and *HNF1B* are localised to the trunk (Obata et al., 2001, Jeon et al., 2009, Schaffer et al., 2010). In mice, *Ptf1A* is first expressed in the pancreatic epithelium at e9.5. By e13.5, its expression is restricted to acinar precursor cells as the tip and trunk domains become segregated. In contrast to *Pdx1*, *Ptf1A* expression is expressed only in the pancreas during development (Kawaguchi et al., 2002). *Ptf1A*$^{-/-}$ mice died shortly after birth and displayed a complete absence of exocrine pancreatic tissue (Krapp et al., 1998). In humans, mutations in the *PTF1A* enhancer may lead to pancreatic agenesis (Sellick et al., 2004, Weedon et al., 2014).

GATA factors

The GATA family identified in vertebrates is composed of six zinc-finger TFs, each playing important roles in the specification and differentiation of multiple cell types (Arceci et al., 1993, Molkentin, 2000, Patient and McGhee, 2002). All members of the GATA family contains a highly conserved DNA binding domain consisting of two zinc fingers that recognise and bind to the motif WGATAR, in which W indicates A/T and R indicates A/G, in the regulatory sequences of target genes (Ko and Engel, 1993). The six GATA members are known as *Gata1* (Evans and Felsenfeld, 1989, Tsai et al., 1989), *Gata2* (Yamamoto et al., 1990, Lee et al., 1991, Dorfman et al., 1992), *Gata3* (Yamamoto et al., 1990, Ho et al., 1991, Joulin et al., 1991, Ko et al., 1991), *Gata4* (Arceci et al., 1993, Kelley et al., 1993), *Gata5* (Laverriere et al., 1994), and *Gata6* (Laverriere et al., 1994). Based on their expression patterns in restricted tissues, the GATA members have been divided into two subfamilies: *GATA1-3* and *GATA4-6* (Molkentin, 2000). *Gata1-3* are prominently expressed in hematopoietic stem cells where they regulate lineage-specific gene expression in T-lymphocytes, erythroid cells, and megakaryocytes (reviewed in (Orkin, 1998)). *Gata4 -6* are expressed in various mesoderm- and endoderm-derived tissues such as the heart, liver, lung, pancreas and gut where they play critical roles in regulating tissue-specific gene expression (Arceci et al., 1993, Kelley et al., 1993, Laverriere et al., 1994, Morrisey et al., 1996a, Suzuki et al., 1996, Morrisey et al., 1997). Of these GATA family members, only *Gata4* and *Gata6* have been shown to be expressed in the pancreas and have a role in pancreatic development (Decker et al., 2006, Carrasco et al., 2012). In the mouse embryo, *Gata4* and *Gata6* overlap in the foregut endoderm at e9.5, including the pre-pancreatic endoderm (Molkentin, 2000). As embryonic development proceeds, *Gata4* and *Gata6* expression diverges to be expressed in acinar cells at e16.5, and endocrine islets at e14.5 respectively (Ketola et al., 2004).

Gata4 null mice display severe developmental abnormalities, resulting in embryonic lethality between e7.0 and e9.5 (Molkentin et al., 1997). Tetraploid embryo complementation experiments were able to rescue these defects, enabling the generation of clonal embryonic e9.5 *Gata4* $^{-/-}$ embryos directly from embryonic stem cells (Watt et al., 2004). Similarly, *Gata6* null mice die after implantation because of defects in visceral endoderm function and extraembryonic development (Morrisey et al., 1998). This early embryonic deficiency associated with *Gata6* null mice could also be rescued with tetraploid embryo complementation by providing *Gata6*-null embryos with a wild-type extraembryonic endoderm (Zhao et al., 2005). Thus, although the early embryonic lethality associated with the *Gata4*$^{-/-}$ and *Gata6*$^{-/-}$ mice has precluded loss-of-function analyses in the pancreas, *in vivo* mouse studies using tetraploid complementation and a transgenic Gata-engrailed fusion protein have suggested that Gata4 and/or Gata6 contribute to the regulation of pancreas development (Decker et al., 2006, Watt et al., 2007).

As mentioned earlier, *GATA4* is expressed during early human pancreatic budding between 4 to 5 weeks post conception, then becomes reduced in pancreatic progenitors, remaining mainly in mature acinar cells, an expression pattern similar to that of mice. In humans, the precise expression pattern of *GATA6* during pancreatic development from has not been closely studied. Recently, studies established a critical regulatory role for *GATA4* and *GATA6* in human pancreas formation, and reported that heterozygous mutations in *GATA4* or *GATA6* can lead to pancreatic agenesis (Lango Allen et al., 2012, Shaw-Smith et al., 2014, Bonnefond et al., 2012). Heterozygous mutations in *GATA4* and *GATA6* have also been associated with congenital heart defects (Garg et al., 2003, Lango Allen et al., 2012, Bonnefond et al., 2012). Strikingly, this is not the case in mice. Heterozygous or homozygous inactivation of either *Gata4* or *Gata6* does not impair pancreas formation, but simultaneously inactivation of both three or four *Gata4* and *Gata6* alleles in the pancreatic progenitor domain leads to pancreatic agenesis and loss of *Pdx1* expression, indicating a functional redundancy for these TFs during pancreas development in mice (Carrasco et al., 2012, Xuan et al., 2012).

SOX9

SOX9, a member of the SRY/HMG box family, is found in *PDX1*-positive cells in early human pancreas by about 4 weeks post conception and in the mouse is expressed in the *Pdx1* domain from e9.5 (Seymour et al., 2007, Lynn et al., 2007, McDonald et al., 2012, Jennings et al., 2013). Although *SOX9* expression is absent in subsequent endocrine cells and restricted to pancreatic duct cells (Jennings et al., 2013), it plays integral roles in maintaining the pancreatic progenitor pool, supporting endocrine cell differentiation, and co-localising with and regulating the expression of other important TFs such as *FOXA2* and *NGN3* (Seymour et al., 2007, McDonald et al., 2012). In mice, conditional inactivation of *Sox9* in the *Pdx1* domain results in severe pancreatic hypoplasia (Seymour et al., 2007). In addition, the *Sox9*$^{+/-}$ mice display a similar phenotype to *SOX9* haploinsufficiency in humans, where failed maintenance of endocrine progenitors result in islet hypoplasia (Sosa-Pineda et al., 1997, Piper et al., 2002, Seymour et al., 2008).

NKX6-1

As mentioned earlier, *NKX6-1* expression in humans is detected after 4 weeks post conception, once *SOX17* is excluded from the pancreatic buds. Its expression then becomes restricted to β-cells by 14-16 weeks (Brissova et al., 2005, Jennings et al., 2013). Similarly in rodent, *Nkx6.1* expression is broadly expressed in the early stages of pancreatic development, then gradually becomes restricted to β-cells (Sosa-Pineda et al., 1997). *NKX6-1*$^{-/-}$ mice exhibit a severe reduction in β-cells, and failure of conditional *Nkx6.1* mutants to express *Pdx1* reveal its role in specifying endocrine precursors toward β-cell lineage (Sander et al., 2000, Henseleit et al., 2005, Schaffer et al., 2013). In human T2D islets, there is a reduced expression of *NKX6-1* (Guo et al., 2013).

NGN3

The expression pattern of *NGN3* in human has been described in an earlier section (1.1.1 Development of the human pancreas). *NGN3*$^{-/-}$ mice fail to generate pancreatic endocrine cells and die postnatally from severe hyperglycemia (Gradwohl et al., 2000). In humans, it has been reported that a rare biallelic *NGN3* null mutation resulted in PNDM with no histologically detectable islets, but detectable C-peptide levels suggest the presence of some β-cells (Rubio-Cabezas et al., 2011).

1.2.3 Transcription factor *GATA6*

The human (Suzuki et al., 1996) and mouse (Narita et al., 1996, Morrisey et al., 1996a) *GATA6*/*Gata6* gene was first described in 1996. In humans, the *GATA6* gene is located on human chromosome 18 q11.1–q11.2 (Suzuki et al., 1996). Initially, the *GATA6* cDNA was reported to have an open reading frame (ORF) from nucleotide residues 348 to 1697 extending from an initiator methionine codon at 716 bp, encoding a predicted protein size of 45.3 kDa composed of 449 amino acids (MYQ-ORF, Figure 5) (Suzuki et al., 1996, Huggon et al., 1997). It was subsequently discovered that translation of the *GATA6* gene can initiate from two alternative initiator methionine codons, giving rise to two protein isoforms (Brewer et al., 1999). In this study, a longer potential ORF encoding a protein of 595 amino acids, which commences at a more upstream, “in-frame” putative initiator methionine codon at 278 bp was revealed (MALT-ORF, Figure 5) (Brewer et al., 1999). Both methionine codons are within a theoretically favourable context for translation initiation (Kozak, 1981, Cavener and Ray, 1991, Kozak, 1997) and are located within exon 2, out of the 7 exons of the *GATA6* gene (Figure 5). Both isoforms possess an N-terminal transactivation domain and two zinc finger domains, both of which are essential for activity (Takeda et al., 2004). It has been reported that the two isoforms differ in their transactivation potential; full length *GATA6* which expresses both isoforms and an altered *GATA6* which only produces the longer isoform had the highest transactivation potentials (Brewer et al., 1999). However, deletion of the extended N-terminal 146 amino acids reduced transactivation potential by approximately 50%, and deletion of the region proximal to the zinc finger domains resulted in very little transactivation activity (Brewer et al., 1999).

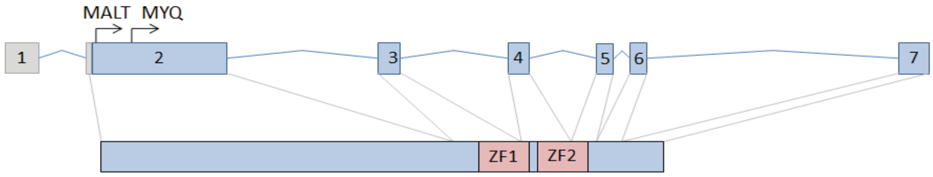


Figure 5. Corresponding cDNA transcript and protein product of *GATA6*. Top row represents cDNA, noncoding (grey) and coding (blue regions) of exons 1-7. Two initiator methionine codons (MALT and MYQ) are indicated in exon 2. Bottom row represents the full length isoform of the *GATA6* protein, including locations of the two zinc finger DNA binding domains (red boxes, ZF1 and ZF2).

The distribution of *GATA6* transcripts in embryonic tissue appeared to be high in the heart and lungs, and absent in brain, liver, or kidney (Suzuki et al., 1996). In adult tissues, *GATA6* transcripts were expressed at high levels in the heart, ovary, lung, and pancreas, low levels in the liver and spleen, and absent in the brain, placenta, skeletal muscle, thymus, prostate, testes, small intestine, colon, or leukocytes. Distribution of *GATA4* differed slightly as was not detected in either adult or embryonic lung or in adult spleen, but present in testes. Of note, *GATA6* and *GATA4* expression overlapped in the adult pancreas and heart (Suzuki et al., 1996).

In mice, the *Gata6* gene maps to a region of chromosome 18 that shows homology to human chromosome 18 (Narita et al., 1996). The *Gata6* cDNA was initially reported to include a 1332 bp ORF encoding a 444 amino acid polypeptide with a predicted protein size of 45 kDa (Morrisey et al., 1996a). Similarly to human, the mouse *Gata6* gene encodes a longer polypeptide in addition to the one described earlier, with both methionine initiator codons located within exon 2, resulting in two protein isoforms (Brewer et al., 1999). The extended N-terminal sequence comprises 147 amino acid residues. *Gata6* distribution overlaps with *Gata4* in the adult mouse, and is abundant in the heart, lung, stomach, small intestine, large intestine and ovary, has lower levels in the liver and is absent in the brain, kidney, or skeletal muscle (Narita et al., 1996). Importantly, during mouse development, *Gata4* and *Gata6*, but not *Gata5*, are expressed in overlapping domains within the primitive and foregut endoderm, including the regions that give rise to liver and pancreas (Bossard and Zaret, 1998, Decker et al., 2006).

The recent genome sequencing of 27 neonatal diabetic patients with pancreatic agenesis or severe pancreas hypoplasia that revealed 56% of the patients had spontaneous heterozygous mutations in the *GATA6* gene sparked a potential interest in the *GATA6* gene (Lango Allen et al., 2012). This study associated *GATA6*, on top of previously identified genes such as *PDX1* and *PTF1A* whose inactivation causes pancreatic agenesis in humans, to a potential role in pancreas morphogenesis.

1.3. Disease modelling of pancreatic agenesis

hPSC technology is a ground-breaking step toward modelling human disease in a controlled laboratory setting. hiPSCs can be derived from healthy volunteers or patients, thus hiPSC technology allows cellular models of disease to be formed from differentiated human pluripotent stem cells. Although animal models have proven invaluable in uncovering fundamental biology, inherent differences between human and rodent biology lead to limitations in the ability of animal systems to recapitulate human disease. Several such examples are human islets comprise a lower proportion of β-cells, and a higher proportion of α- and δ- cells compared to mouse islets (Brissova et al., 2005), and a different islet architecture between the two species, with human β-cells being dispersed among α- and δ- cells, while mouse β-cell maintaining a core surrounded by the four other endocrine cell types (Figure 6) (Cabrera et al., 2006). In rodents, insulin is encoded by two genes (*INS1* and *INS2*), whereas in humans, insulin is only encoded by one gene (*INS*) (Melloul et al., 2002). Indeed, discordant phenotypes in the mouse compared to human caused by *GATA6* mutations have been noted as discussed in an earlier section (1.2.2). As such, this precludes the use of animal models in modelling the disease.

The advancement of hPSC technology coupled with the availability of genome editing tools (discussed in the next section) provide a valuable opportunity to accurately model pancreatic agenesis as a disease and investigate the role of TF *GATA6* in pancreatic development and elucidate how *GATA6* mutations can impair the formation of the human pancreas.

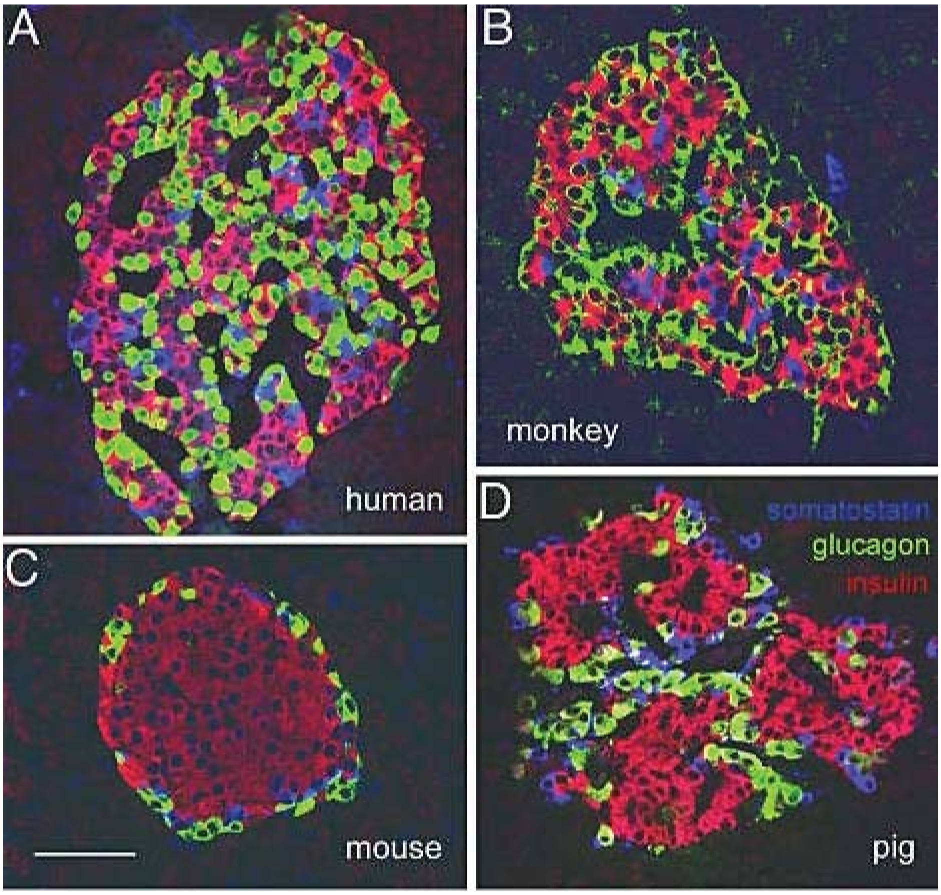


Figure 6. Confocal microscopy images of α-, β- and δ- cells within Islets of Langerhans show striking interspecies differences. (A, B) In human and monkey islets, insulin-immunoreactive β-cells (red), glucagon-immunoreactive α-cells (green), and somatostatin-immunoreactive δ-cells (blue) cells were all found randomly distributed. **(C)** In mouse islets, insulin-containing cells were located in the core, and glucagon- and somatostatin-containing cells in the circumference. **(D)** In pig islets, α-, β- and δ- cell distributions are similar to that of the mouse but appear to be formed in smaller units. Scale bar, 50 µm. Adapted from (Cabrera et al., 2006).

1.3.1 Genome editing tools

The concept of genetic engineering was first introduced in 1972 by Paul Berg's lab in the form of recombinant DNA, where scientists successfully combined genes from different species (Jackson et al., 1972). Over the years, there has been tremendous progress in the development of methods not only to manipulate DNA, but also to generate vector systems and optimise their delivery into cells. The success of the Human Genome Project opened up many doors toward a deeper understanding of how the nucleotide sequence of human nuclear DNA relates to pathology of hereditary as well as multifactorial diseases. It also enabled the study of functional elements within the human genome, such as transcription factors. In order to establish relationships between gene function and disease, two strategies are often used: repression by switching off a gene i.e. knockdown or knockout and activation by overexpressing a gene.

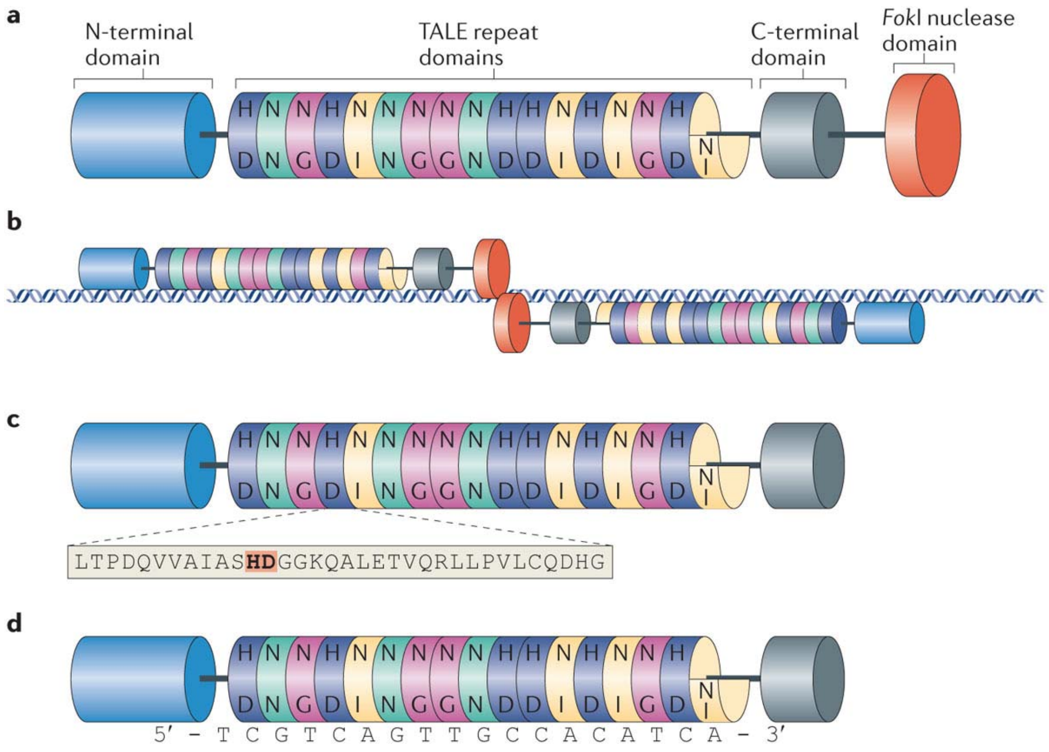
In 1996, a study reporting precise genome editing using a zinc finger protein domain coupled with the FokI endonuclease domain was published (Kim et al., 1996). For the first time, it was possible to perform site-specific nuclease cutting of DNA at strictly defined sites *in vitro*. Zinc finger nuclease (ZFN) thus became the basis for editing cultured cells including pluripotent stem cells, plants and animals (Bibikova et al., 2002, Townsend et al., 2009, Zhang et al., 2010, Lombardo et al., 2011, Provasi et al., 2012, Torikai et al., 2012). However, the technology brought several disadvantages such as the high complexity and cost of assembling the DNA-binding protein domains, low efficiency and potential off-target effects. This drove the discovery of two more genome editing tool which succeeded the ZFN: transcription activator-like effector nuclease (TALEN) from *Xanthomonas* bacteria (Boch et al., 2009, Moscou and Bogdanove, 2009, Christian et al., 2010, Miller et al., 2011) and RNA-guided DNA endonuclease Cas9 from the type II bacterial adaptive immune system clustered regulatory interspaced short palindromic repeats (CRISPR/Cas) (Cong et al., 2013, Mali et al., 2013). As CRISPR technology was only just emerging when the project started, TALEN was the genome editing tool that was used in this project, and will be described in detail below.

1.3.2 Transcription activator-like effector nuclease (TALEN)

The TALEN system was developed based on the study of the bacteria *Xanthomonas* genus, which act as pathogens to crop plants by secreting naturally occurring effector proteins (transcription activator-like effectors, TALEs) that support bacterial virulence, proliferation, and dissemination (Boch and Bonas, 2010). The TALE proteins bind to genomic DNA to alter transcription in host cells, thereby facilitating pathogenic bacterial colonisation.

TALEN proteins contain a DNA-binding TALE repeat domains composed of a series of 33 to 35 amino acid repeat domains each recognising a single nucleotide base (Figure 7d), flanked by an amino (N)-terminal domain and a carboxy (C)-terminal domain that is fused to a *Fok*I restriction endonuclease domain (Figure 7a). In order for the TALEN to recognise a specific sequence on the double-stranded DNA, two TALEN proteins are required, commonly called the left and right TALEN arms, and they each recognise a particular sequence on the forward or reverse strand of the DNA (Figure 7b).

Each TALE repeat domain has an almost identical amino acid sequence, except for two hypervariable residues typically found at positions 12 and 13 of the repeat domain, which determines which nucleotide base the individual TALE repeat domain will recognise (Figure 7c) (Boch et al., 2009). These hypervariable residues are also known as the repeat variable di-residues (RVDs). The RVDs NN, NI, HD and NG code for the recognition of nucleotide base guanine, adenine, cytosine and thymine, respectively (Figure 7d). The last repeat that binds a nucleotide at the 3’-end of the recognition site consists only of 20 amino acid residues, and is therefore called a half-repeat. Subsequent improvements were made to increase binding and specificity of these residues to their respective nucleotide. For example, the RVDs NK was reported to be more specific for guanine than NN, although it also recognises adenine (Moscou and Bogdanove, 2009, Miller et al., 2011) and displayed less activity as compared to NN (Streubel et al., 2012). The RVDs NH were suggested to be more specific than the NN repeat but with lower activity (Streubel et al., 2012, Cong et al., 2012). It was also reported that RVDs HD and NH bind their preferred nucleotides strongly, while NI and NG bind to their preferred nucleotides relatively weaker (Streubel et al., 2012). Furthermore, a second generation scaffold greatly increased *in vivo* modification efficacy (Bedell et al., 2012).


a
N-terminal domain
TALE repeat domains
C-terminal domain
*Fok*I nuclease domain
H N N H N N N N N H H N H N N H N
D N G D I N G G N D D I D I G D I
b
c
H N N H N N N N N H H N H N N H N
D N G D I N G G N D D I D I G D I
LTPDQVVAIAS**HD**GGKQALETVQRLLPVLCQDHG
d
H N N H N N N N N H H N H N N H N
D N G D I N G G N D D I D I G D I
5′ - T C G T C A G T T G C C A C A T C A - 3′

Nature Reviews | Molecular Cell Biology

Figure 7. Overview of TALEN proteins. (a) Schematic of one arm of a fully constructed TALEN protein comprising of an amino (N)-terminal and carboxy (C)-terminal domain that are required for DNA-binding, a non-specific *Fok*I endonuclease domain and TALE repeat domains forming an array ending with a truncated half repeat. **(b)** A pair of TALEN arms, namely the left and right TALEN arm, binds and cleaves as dimers on a specific target site of the double-stranded DNA, resulting in a double-stranded break of the DNA. Cleavage by the *Fok*I nuclease domains occurs within a spacer region that lies between binding sites of the left and right TALEN arms. **(c)** The amino acid sequence of a single TALE repeat domain is highly conserved and is similar in all the various domains except for the 12th and 13th amino acid known as the hypervariable residues or repeat variable di-residues (RVDs) (highlighted in orange and bold text). **(d)** Each RVD confers specificity to a single nucleotide and is arranged in the order of its target sequence during construction of the TALEN protein. The TALE array is responsible for binding to DNA at a specific site. Preceding the first base bound by a TALE repeat at the 5’ end is a thymine. Adapted from (Joung and Sander, 2013).

TALEN construction can be challenging due to the nature of the nearly identical repeat sequences of the TALE repeat domains, and the assembly of numerous domains. Many groups have devised platforms for TALEN assembly to facilitate a simple and efficient construction process and these can be broadly grouped into three categories: standard restriction enzyme and ligation-based cloning (Sander et al., 2011, Huang et al., 2011), 'Golden Gate' cloning (Engler et al., 2008, Engler et al., 2009), and solid-phase assembly (Reyon et al., 2012). In this project, the standard restriction enzyme and ligation-based cloning platform was used. This method utilises an archive of plasmids encoding single or multiple TALE repeat domains and join them in a parallel hierarchical fashion via restriction digestion and ligation reactions. The 'Golden Gate' cloning platform is a multi-fragment ligation strategy and allows for 3 to 10 TALE repeat domains to be simultaneously ligated in a particular linear order into a plasmid vector (Weber et al., 2011). Solid-phase assemblies such as Fast Ligation-based Automatable Solid-phase High-throughput (FLASH) assembly, Iterative Capped Assembly (ICA) are automated, high-throughput methods for assembling numerous TALE repeat arrays (Reyon et al., 2012, Briggs et al., 2012, Wang et al., 2012).

1.3.3 Nuclease-mediated mutations

Simultaneous introduction of the left and right TALEN arms into cells often lead to site-specific DNA double-stranded breaks (DSB). The *Fok*I endonuclease domain is crucial for the successful cleavage of the double-stranded DNA by recognising a non-palindromic DNA sequence and making a double-stranded cut outside of that sequence, commonly within a region known as a spacer, resulting in a 5′ overhang (Hiroyuki and Susumu, 1981). The spacer must be an appropriate length of around 16 nucleotides to permit dimer formation. In order to cleave the DNA, each of the *Fok*I domain within the left or right TALEN arm must dimerise during adjacent and independent binding events of each arm onto the site-specific DNA in the correct orientation (Vanamee et al., 2001). The need for two DNA binding events to occur and for the *Fok*I domains to form a heterodimer pair prior to DNA cleavage improves specificity and reduces off-target effects via the elimination of unwanted homodimers (Miller et al., 2007, Szczepek et al., 2007).

Numerous studies have established that normal cellular repair of DSB occur through two pathways; non-homologous end-joining (NHEJ) or homologous recombination (HR) (Rudin et al., 1989, Plessis et al., 1992, Rouet et al., 1994, Choulika et al., 1995, Lieber, 2008, Jackson and Bartek, 2009). It was subsequently realised that these two highly conserved cell repair pathways can be exploited to introduce targeted mutations in a wide variety of cell types and species after nuclease-induced DSBs have taken place (Figure 8) (Bibikova et al., 2001, Bibikova et al., 2002, Bibikova et al., 2003). The NHEJ repair pathway is error-prone and often lead to mutations containing insertions and/or deletions (indel) of variable length originating from the site of the DSB, thus resulting in frameshift mutations that can lead to the knockout of gene function (Bibikova et al., 2002). Alternatively, if a double-stranded DNA 'donor template' is supplied in combination with the pair of TALEN arms, HR of a nuclease-induced DSB can be used to introduce precise nucleotide substitutions or insertions by repairing the DSB with the information encoded on this template (Moehle et al., 2007).

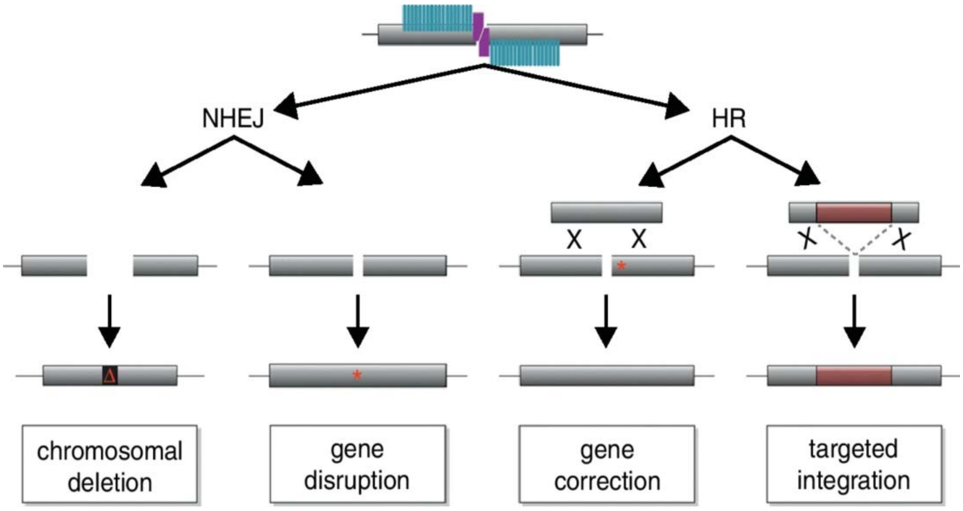


Figure 8. TALEN-mediated genome editing. After a TALEN-induced double-stranded break in the DNA occurs, the cell may undergo one of two highly conserved repair pathways, NHEJ or HR, to repair the cleaved DNA. In the absence of a donor template, the cell undergoes the error-prone NHEJ pathway by ligating the DNA. Nucleotide insertions and/or deletions (asterisk) will commonly be introduced, disrupting the open reading frame and possibly resulting in a premature stop codon. In the presence of a donor template, the cell undergoes the HR pathway which can be used to either correct a mutation (asterisk) in the genome or to target integration of a transgene into a specific site. Adapted from (Mussolino and Cathomen, 2012).

Genome editing tools such as the TALEN technology provide tremendous potential for experimental, biotechnological and therapeutic purposes. Such applications include gene disruptions in model organisms, cell-based disease modelling such as in hPSCs, and gene corrections (allele editing). In cell-based disease modelling, the impact of a specific gene disruption and of specific sequence variants on gene function can be studied closely and be directly associated with diseases. This enables the generation of isogenic cell lines so any possible effects on the disease phenotype under investigation that may be caused by genetic background variations can be excluded (Figure 9).

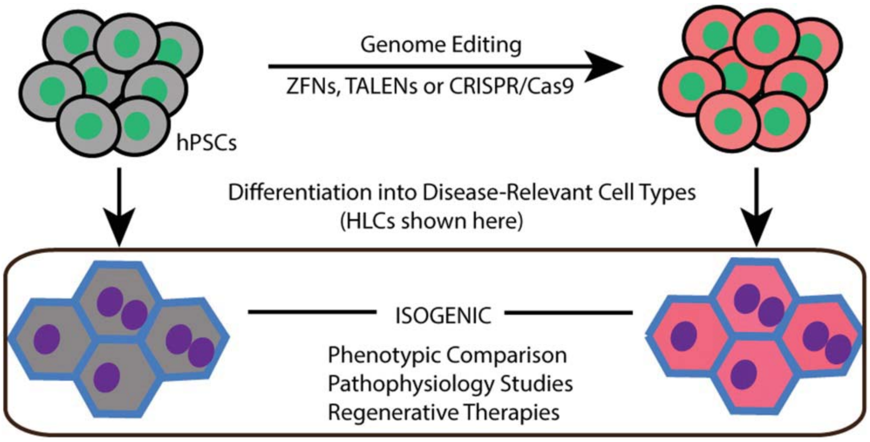


Figure 9. hPSC-based disease modelling. Wild-type hPSCs are targeted by genome editing tools such as TALEN to generate isogenic cell lines. Wild-type and mutated hPSCs are then differentiated into disease-relevant cell types followed by phenotypic comparison and pathophysiological studies to determine the direct association of the gene to the disease phenotype. Adapted from (Yu and Cowan, 2016).

1.4. Objectives of the project

Advancement in genome sequencing technologies in the recent years has provided a major unexpected discovery in this field. The recent study published by Allen et al. where 15/27 (56%) patients with pancreatic agenesis and exocrine pancreatic insufficiency requiring enzyme replacement therapy, born to non-diabetic parents harboured *de novo* heterozygous inactivating mutations in *GATA6* established a key role for the transcription factor *GATA6* in human pancreatic development (Lango Allen et al., 2012). This study is the basis for my project. Human genetics, therefore, has established that *GATA6* is an essential regulator of human pancreas development, but it does not shed light on the underlying molecular mechanism, nor does it define the precise cell types or developmental stages in which the essential role takes place. The role of *GATA6* in the development of the pancreas has been well studied in the mouse, but this is not the case in humans. From mouse studies, it is known that *GATA6* is expressed in the developing pancreas and is an important regulator of pancreas development. Thus, the overall objective of my project is to elucidate the role *GATA6* in the development of the human pancreas. Knowledge gained from this project could potentially contribute to therapies for neonatal diabetes.

The first objective of this study is to perform directed differentiation of hPSCs into the pancreatic lineage using a fully defined culture system. The second objective is to obtain *GATA6* patient lines, reprogram them to obtain patient-derived *GATA6* mutant lines, and perform directed differentiation into the pancreatic lineage to assay the effect of *GATA6* mutations on the development of the pancreas. The third objective is to perform disease modelling of pancreatic agenesis by generating *GATA6* heterozygous and homozygous mutations in hPSC lines via TALEN as a genome editing tool. The fourth objective is to perform phenotypic comparisons between these TALEN-generated *GATA6* mutant hPSCs and their respective isogenic control hPSC lines. The final objective is to define the molecular mechanisms of *GATA6* by investigating the transcriptional networks controlled by *GATA6* through RNA-sequencing and identifying its interacting partners through ChIP-sequencing.