UNRAVELING THE MOLECULAR PHYSIOLOGY OF THE β-CELL: GENOME WIDE ANALYSIS OF BINDING SITES FOR THE TRANSCRIPTION FACTOR PDX1

by

Michael Beach

Honours B.Sc. Trinity Western University, 2006

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

in

The Faculty of Graduate Studies
(Interdisciplinary Oncology)

The University of British Columbia
(Vancouver)

July 2009

ABSTRACT

The selected expression of the genome determines distinct cell types, properties, and conditions. In the pancreatic β-cell, our knowledge of how this is regulated and maintained is incomplete. Deciphering the molecular physiology of the β-cell is critical to develop improvements for expanding pools of donor islets for transplantation, the most promising curative option for sufferers of diabetes.

Genomic regulation is controlled primarily by transcription factors, of which pancreatic duodenal homeobox 1 (Pdx1) plays a critical role in both the developing and mature pancreas. As such, I begin to unlock the molecular physiology of the β-cell by identifying the binding sites of Pdx1 in pancreatic islets on a genome-wide scale through the use of chromatin immunoprecipitation followed by sequencing (ChIP-Seq). This provides the best picture of Pdx1 binding that has ever been assembled. Moreover, I identify a highly co-occurring relationship between Pdx1 and pre-B-cell leukemia homeobox 1 (Pbx1) in adult islets.

The coupling of this data with other genome-wide analyses will prove invaluable to discovering novel transcriptional complexes and the genes they regulate. It will also contribute to the creation of an islet transcriptional network, thereby greatly enhancing our knowledge of β-cell regulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ATP	Adenosine TriPhosphate
ChIP	Chromatin Immunoprecipitation
EDTA	Ethylenediaminetetraacetic Acid
EM	Enrichment Maximization
ERα	Estrogen Receptor Alpha
ES cell	Embryonic Stem Cell
FACS	Fluorescence Activated Cell Sorting
GADEM	A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery
GCK	Glucokinase
GLUT	Glucose Transporter
GMAT	Genome-wide Mapping Technique
HBSS	Hanks Balanced Salt Solution
HNF	Hepatocyte Nuclear Factor
IAPP	Islet Amyloid Polypeptide
ISL1	Islet-1
KD	Knockdown
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAFA	V-maf Musculoaponeurotic Fibrosarcoma Oncogene Homolog A
MIN6	Mouse Insulinoma 6
MODY	Mature Onset Diabetes of the Young

NEUROD1	Neurogenic Differentiation 1
NGN3	Neurogenin 3
NKX	NK Homeobox
PAX	Paired Box
PBS	Phosphate Buffered Saline
PBX1	Pre-B-cell-Leukemia Homeobox 1
PCR	Polymerase Chain Reaction
PDX1	Pancreatic Duodenal Homeobox 1
PET	Paired End DiTag
PWM	Position Weight Matrix
RT	Reverse Transcription
SABE	Serial Analysis of Binding Enrichment
SACO	Serial Analysis of Chromatin Occupancy
SAGE	Serial Analysis of Gene Expression
STAGE	Sequence Tag Analysis of Genomic Enrichment
TBST	Tris-Buffered Saline Tween-20
TE	Trypsin-EDTA
TSS	Transcriptional Start Site
UCSC	University of California, Santa Cruz
WHO	World Health Organization

ACKNOWLEDGEMENTS

Special thanks to Dr. Brad Hoffman for his mentorship and training, as well as other members of the Helgason lab: Bo Zavaglia and Joy Witzsche, and my supervisory committee: Dr. Pamela Hoodless, Dr. Cheryl Helgason, Dr. Dixie Mager, and Dr. Sylvia Ng. Islet isolations at the Verchere lab were performed by Galina Soukhatcheva. ChIPs at the Genome Sciences Centre were performed by Balgit Kamoh. Motif Discovery analysis courtesy Gordon Robertson and Leping Li.

CHAPTER 1. INTRODUCTION

1.1 Pancreas Development, Structure, and Function

Germ layer formation at gastrulation establishes the endoderm, the germ layer from which the pancreas develops. Subsequently, distinct morphological events accompanied by specific onsets of gene expression culminate in pancreas formation. The point of pancreas determination is termed the primary transition, which occurs shortly after the onset of FoxA2 expression in the endoderm. As the embryo begins to rotate, FoxA2 induces Pdx1 expression, driving cells towards the pancreatic fate[1]. Dorsal and ventral pancreatic buds begin to form and Nkx6.1 and NeuroD1 become expressed in the epithelium. Expansion of the epithelium occurs before the secondary transition, when terminal differentiation of islet and exocrine cells occurs. At this point, insulin or exocrine genes experience a 100-fold activation, with Pdx1, Nkx6.1, and Nkx2.2 becoming β-cell restricted[1]. Finally, at isletogenesis, endocrine cells group into the islets and exocrine acinars form.

The role of the pancreas is twofold, as its exocrine cells produce and secrete digestive enzymes into the intestine, while endocrine cells release hormones into the bloodstream that are crucial to maintain homeostatic body metabolism. Exocrine cells compose the majority of the pancreas, and the enzymes they produce enter the duodenum through the ampulla of Vater[2] (also termed major duodenal papilla[3]) where the common bile duct and main pancreatic duct join. With a pyramidal shape and basal nuclei, exocrine cells possess an abundance of rough ER and many secretory vesicles to release their digestive enzymes[4]. Additionally, hydrogen carbonate is also produced by the

exocrine pancreas to neutralize the hydrochloric acid produced in the stomach[5]. Thus, the exocrine pancreas has a critical role in nutrient digestion in the small intestine. While this exocrine function of the pancreas is of utmost importance, the endocrine role of the pancreas in metabolic homeostasis through the cells grouped in its islets has been a major research focus.

1.2 Islet Structure and Function

In the pancreas, endocrine cells comprise only about two percent of the total pancreas mass[6]. Nevertheless, this relatively small population of cells is absolutely critical for normal metabolic maintenance. Embedded within the exocrine tissue, endocrine cells are found in clusters termed Islets of Langerhans. The islet is composed of several types of endocrine cells, and while their exact percentage contribution to each islet is variable, general proportions are agreed upon. The majority of the cells in islets are β-cells. These account for 60-80%[7, 8, 9] of the cell mass and release the hormone insulin. Typically, α-cells are the next most abundant. These cells release glucagon and comprise anywhere from 10-28% of the islet[7, 8, 9]. The remaining cell types are typically less abundant and are as follows: somatostatin producing δ-cells 2-10%[7, 8, 9, 10], pancreatic polypeptide producing PP-cells 3-19%[7, 8, 9, 10], and ghrelin producing ε-cells 1%[9, 11]. Being released directly into the bloodstream, these hormones target primarily the liver, muscle, and fat cells[5], as they play major roles in metabolic homeostasis. Consequently, islets receive a rich arterial blood supply via a unique capillary system that allows them to receive ten times the amount of blood per mass compared to exocrine cells[12]. Islet capillaries are also larger and contain fenestrae that increase permeability[13] and aid in

insulin uptake following its release from the islet[14]. As the most dominant cell type in the islet, and the source of insulin, the β-cell plays the most significant role in metabolic maintenance through careful regulation of blood glucose levels.

1.3 Insulin Release and Glucose Regulation

Insulin release has been long understood to occur in two phases[15]. The first phase is a rapid response to increasing blood glucose levels and lasts approximately 2-4 minutes before decreasing to a plateau at 10-15 minutes. A more gradual process, the second phase of insulin release lasts 2-3 hours during which a steady state of insulin levels is achieved. Sensing of glucose by the β-cell does not occur via a glucose receptor, but rather through the metabolic products of glucose that trigger a molecular response culminating in insulin secretion[5]. This process begins with glucose entering the β-cell through the channel protein GLUT2[16]. In the cytosol, glucokinase phosphorylates glucose[17] preventing it from exiting the cell though GLUT2. Highly efficient oxidative metabolism breaks down glucose to CO_2 and H_2O resulting in an increase in ATP levels through oxidative phosphorylation via the mitochondrial electron transport chain. At this point, signal transduction moves from metabolic to electric, as membrane-bound potassium channels close in response to the increased levels of cytosolic ATP[18]. Closure of these channels causes depolarization of the plasma membrane, eliciting the opening of voltage-gated calcium channels and allowing calcium ions to flood the cytosol[19]. This rise in internal calcium levels stimulates the cortical actin network to disband, permitting insulin containing granules to fuse with the cell membrane and release insulin[20]. Moving through the circulation, insulin stimulates glucose uptake, acting primarily at striated

muscle tissue and adipose tissue. Upon binding of insulin to the insulin receptor, the glucose transporter GLUT4 moves to the cell surface and facilitates entry of glucose into the cell[21]. Metabolism of glucose produces ATP to meet the energy demands of the cell, or results in production of high potential energy storage molecules such as glycogen.

1.4 Diabetes Mellitus

In the absence of proper insulin controlled regulation, blood glucose levels become abnormally high, indicative of the disease diabetes mellitus. In the year 2000, the WHO reported 171 million cases of diabetes worldwide with the incidence continuing to rise particularly in developed countries[22]. The core symptoms of diabetes include frequent urination, increased fluid uptake due to thirst, and increased appetite. If allowed to progress untreated, severe conditions can include diabetic coma, blindness, loss of limbs, renal failure, and death. Both hereditary and environmental factors significantly contribute to the progression of diabetes.

In type I diabetes, the β-cells of the pancreas are destroyed by T-cell mediated autoimmune attack[23]. While individuals remain responsive to insulin, the severe reduction in β-cell numbers results in insufficient production of the hormone for the demands of the body. Conversely, type II diabetes stems from diminished insulin sensitivity leading to insulin resistance. Central obesity is a major risk factor for development of type II diabetes, and for this reason exercise is often prescribed as treatment and can restore insulin sensitivity.

While environmental factors play a significant role in the development of diabetes, there are also major contributing genetic factors. This has been particularly

well characterized in a third form of diabetes, MODY (mature onset diabetes of the young). While not all contributing genes are known, several have been well established, most of which have been termed MODY factors. MODY genes typically have an autosomal dominant mode of inheritance, and their mutation disrupts insulin production. Depending on the gene mutation, MODY is categorized as MODY 1 through 8. The genes belonging to each category are shown below:

Table 1 – Summary of MODY Genes

MODY 1	Hnf4a
MODY 2	Gck
MODY 3	Hnf1a
MODY 4	Pdx1
MODY 5	Hnf1b
MODY 6	NeuroD1
MODY 7	Kruppel like factor 11
MODY 8	Bile salt dependent lipase

Compared to type I and type II diabetes, the MODY forms are extremely rare. However, regardless of the type, no form of diabetes has a cure. While the disease can be well managed through careful monitoring of blood glucose and insulin administration, this is only therapeutic in nature. The most likely *curative* option for individuals suffering from diabetes is islet transplantation. As such, there is a significant amount of research being focused on how to maximize islet transplant success and how to expand islet pools *in vitro* for transplant purposes.

1.5 Expanding Islet Pools and Islet Transplant

Islet transplantation accounts for only a small percentage of the total transplant procedures being performed in British Columbia. In 2008, out of 266 transplants in BC, only 15 were pancreatic islets[24]. Despite this being the only curative option for persons suffering from diabetes, there are two main reasons why so few transplants are being done: 1) graft survival in islet transplants is not long lasting, with only 33% of recipients claiming insulin independence after 2 years[25] and 2) there is a huge shortage of available tissue for transplant. This has motivated the majority of islet research to focus on how to improve islet graft survival, or how to increase islet survival and proliferation in culture and/or differentiate stem cells into insulin producing cells suitable for transplant. Both types of research are of critical importance for islet transplantation to evolve into a true curative therapy for diabetes.

The latter of these research focuses, expansion of islet pools and stem cell differentiation, is of vital importance because currently, islets from several deceased donors must be harvested to perform a single transplant. Moreover, once in culture, survival of islets is poor, and proliferation of the cells does not readily occur[26]. To date, there has been some success in overcoming this barrier, but further refinements are required[27]. Consequently, advances in expanding isolated islet populations are invaluable to provide more tissue for transplant. Similarly, stem cell research addresses this same problem through the generation of β-cells from early lineage precursors. This could also address the problem of graft rejection given that tissue could be differentiated directly from stem cells of the patient. While such endeavours have yielded insulin-producing cells[28], they are not true β-cells and are not as of yet suitable for transplant use[29].

Whether the goal is to enhance existing islet survival and proliferation, or produce β-cells from a stem cell antecedent, it is clear that a better understanding of the molecular physiology of the β-cell is needed to augment these efforts. This is because a cell's properties are determined by the information carried in its genome, the selected expression of which serves to define distinct cell types and conditions[30]. This controlled expression of genomic information is regulated by transcription factors. Therefore, a comprehension of transcription factor binding and networks can aid in better understanding how a given cell type employs its genome to arrive at and maintain its final functions.

1.6 Transcription Factor Biology

Transcription factors are proteins that possess DNA binding domains allowing them to directly bind to DNA and regulate transcription through activation and/or repression[31]. These factors are significant contributors to controlled expression of the genome, along with microRNAs[30]. In addition to the DNA binding domain, transcription factors can also contain trans-activating domains that serve as binding sites for other proteins acting as coregulators. This allows multiple transcription factors to associate and form complexes for highly controlled genomic regulation. Transcription factors are grouped into families based on the structure of their DNA binding domain. Pdx1, for example, is grouped in the homeodomain protein family.

1.7 Key Transcription Factors of the Endocrine Pancreas

Most transcription factors known to have essential roles in the pancreatic β-cell have been identified based on their roles developmentally, or from their direct influence on insulin regulation. In pancreatic islets, critical transcription factors include but are not limited to: FoxA2 (Hnf3β), Hnf4α, Hnf1α, Hnf1β, Nkx2.2, Nkx6.1, NeuroD1, Ngn3, Pax4, Pax6, Isl1, Mafa, Pbx1, and Pdx1.

The importance of FoxA2 rests primarily on its developmental role as an activator of Pdx1[32], whose activity is often mediated by Pbx1[33, 34]. Similar to FoxA2, Hnf1α[35] and Hnf1β[36] are also regulators of Pdx1, in addition to themselves being MODY genes. It has been suggested that these transcription factors, as well as Hnf4α, act cooperatively with Pdx1 in the adult β-cell to drive expression of essential β-cell specific genes[1].

While the above Hnf family members serve to both regulate and act with Pdx1, the Nkx family members are suspected targets of Pdx1 that are also crucial transcription factors in β-cells[1, 37]. Knockout studies of Nkx2.2 reveal that while endocrine cells differentiate normally, the β-cells are unable to activate the insulin gene and expression of Nkx6.1 is also lost[38]. Nkx6.1 is specific to the β-cells of the pancreas and is essential for β-cell formation. Loss of Nkx6.1 expression results in pancreases showing normal development of islet cells with the exception of the mature β-cell, which is completely absent[39]. This observation has led to speculation that Nkx6.1 serves to repress genes that confer α-cell fate, thereby stabilizing the β-cell phenotype.

Before the stabilization of endocrine cell type can be conferred, the overall endocrine fate must first be selected; this is accomplished through Ngn3. Forced expression of Ngn3 in pancreatic ductal cells has been shown to activate an endocrine

program[40]. Similarly, transfection of endodermal ES-cells with Ngn3 induces insulin gene transcription as well as expression of other endocrine type factors[41]. It is believed that this occurs as a result of Ngn3 activation of another key pancreas transcription factor, NeuroD1.

NeuroD1 is a basic helix-loop-helix factor that binds to E-box elements of the β-cell insulin gene promoter in a complex with Pdx1 and Mafa, although it is also expressed in all other endocrine cell types of the pancreas[1, 42]. Despite its ubiquitous islet expression, it does not appear to be necessary for endocrine differentiation as NeuroD1 knockout mice successfully produce all islet cell types. However, upon islet formation, these same mice develop diabetes due to β-cell apoptosis and a reduction of islet cell numbers[43].

Several other transcription factors have been identified as critical to pancreas development and/or function as their altered expression gives rise to definitive phenotypes. Isl1 was one of the first genes identified as having a role in pancreas development. The dorsal pancreatic bud fails to develop in Isl1 deficient mice, and in the ventral bud glucagon expressing cells are absent[44]. Distinct phenotypes are also observed in Pax4 and Pax6 inactivated mice. In both cases, mice die shortly after birth of similar yet opposite causes. In Pax4 knockout mice, both β-cells and δ-cells are completely absent while α-cells persist. Conversely, Pax6 knockout mice show the opposite trend in endocrine cell type presence. When both Pax4 and Pax6 are inactivated, no pancreas endocrine cell types are observed[45]. The factors that regulate Isl1, Pax4, and Pax6 expression in the pancreas are not well understood.

In almost every case, the abovementioned transcription factors have been identified as crucial to the pancreas as a result of clear presentation of phenotypes. In addition, a uniting thread of a relationship with the pancreatic master regulator Pdx1 is apparent. Therefore due to its overarching role, Pdx1 represents an ideal starting candidate to decipher the molecular physiology of the β-cell.

1.8 Pdx1 and the Endocrine Pancreas

In addition to the suspected relationship of Pdx1 to many other pancreas critical transcription factors, Pdx1 was also the first gene identified to be independently required for pancreas development in mice and humans[46, 47]. Its expression begins at E8.5 in the definitive endoderm, where it drives pancreatic fate, and more specifically β-cell differentiation[1]. Consequently, knockout of Pdx1 results in embryonic lethality as pancreas formation does not progress past the initial budding stages. In the absence of Pdx1, the undifferentiated cells of the pancreas fail to expand after dorsal and ventral budding occurs. Therefore, the onset of Pdx1 expression at this stage is typically regarded as the beginning of pancreagenesis. Pdx1 protein distribution remains homogenous until the secondary transition, when exocrine cells down regulate Pdx1 while endocrine cells up regulate Pdx1 resulting in a 100-fold difference in expression. In the mature islet, Pdx1 is restricted to β-cells and a small set of δ-cells[48].

In addition to its essential role in development, Pdx1 also maintains vital importance in the adult. The most recognizable gene that Pdx1 regulates is insulin. For this reason, Pdx1 is a MODY factor. Binding of Pdx1 at the insulin promoter has been reported to occur at two distinct E-box elements upstream of the transcriptional start

site[49]. Here, the protein is thought to form a transcriptional complex with NeuroD1 and Mafa[42]. While the insulin promoter possesses potential binding sites for a variety of transcription factors[50], the binding of Pdx1 to these elements is not only confirmed but indispensable, as the loss of even one of the elements results in insulin deficiency. Recently, a short-range DNA looping model of Pdx1 regulation at the insulin gene has been proposed that results in distal enhancer regions being brought into close proximity to the transcriptional start site[42]. In this model, only a single true Pdx1 binding site exists, with the second binding site indirectly linked through NeuroD1.

Insulin, though arguably the most important, is not the only critical β-cell gene regulated by Pdx1. It has also been shown to activate Gck[51], Glut2[52], IAPP[53], Mafa[54], and its own promoter[35]. From this, it seems that Pdx1 functions not only as a master regulator of pancreas development, but also a master regulator of β-cell function in the adult. Mouse models confirm the importance of Pdx1 in mature islets. Since the Pdx1 knockout is embryonic lethal, our best insight into how the absence of Pdx1 affects the adult β-cell comes from conditional knockout studies. These mice show reduced insulin secretion as well as reduced expression of Glut2, thereby substantiating *in vivo* the importance of Pdx1 in adulthood[55].

In both the embryo as well as the adult, the transcriptional activity of Pdx1 is moderated, at least in part, by Pbx1. The formation of Pdx/Pbx heterodimers has been shown to occur *in vitro*, and has been hypothesized to play a role in refining Pdx1 activity in exocrine versus endocrine cell types[1]. Developmentally, the importance of the Pdx/Pbx interaction has been demonstrated through generation of Pdx1 mice with a mutated Pbx1 interaction domain. In these mice, the quantity and organization of

endocrine cells is severely impaired, suggesting a critical role for Pdx/Pbx complexes in expansion of precurser cell populations[33]. The significant nature of this interaction seems to continue into the mature β-cell, as mice heterozygous for Pdx1 and Pbx1 mutant alleles develop more severe diabetes and hypoinsulinemia than single mutants of either gene[34].

The importance of Pdx1 to both pancreas development and adult function is unquestionable. However, only a handful of Pdx1 target genes, though absolutely critical, are known. Consequently, on a genome-wide scale there is still very little known about how Pdx1 is operating to maintain β-cell function.

1.9 Chromatin Immunoprecipitation and Platforms for Sequencing

The chromatin immunoprecipitation (ChIP) procedure is a valuable tool for identifying transcription factor binding at target sites. Transcription factors are crosslinked to DNA from isolated cells and the membranes lysed to release the chromatin. Sonication pulses are used to shear the DNA into small fragments that are subsequently incubated with an antibody directed against the transcription factor of interest. To isolate antibody bound DNA, protein G beads are added which bind the antibody-transcription factor-DNA complex, allowing for isolation, elution, and retrieval of only those DNA fragments bound by the transcription factor of interest. A schematic of the ChIP procedure is displayed in Figure 1.

Classically, ChIP-DNA has been assessed through polymerase chain reaction (PCR) on a site-by-site basis, which requires prior suspicion of a site of interest to warrant testing for binding enrichment. However, advancements of array and sequencing technologies have made identifying large numbers of novel binding sites more feasible

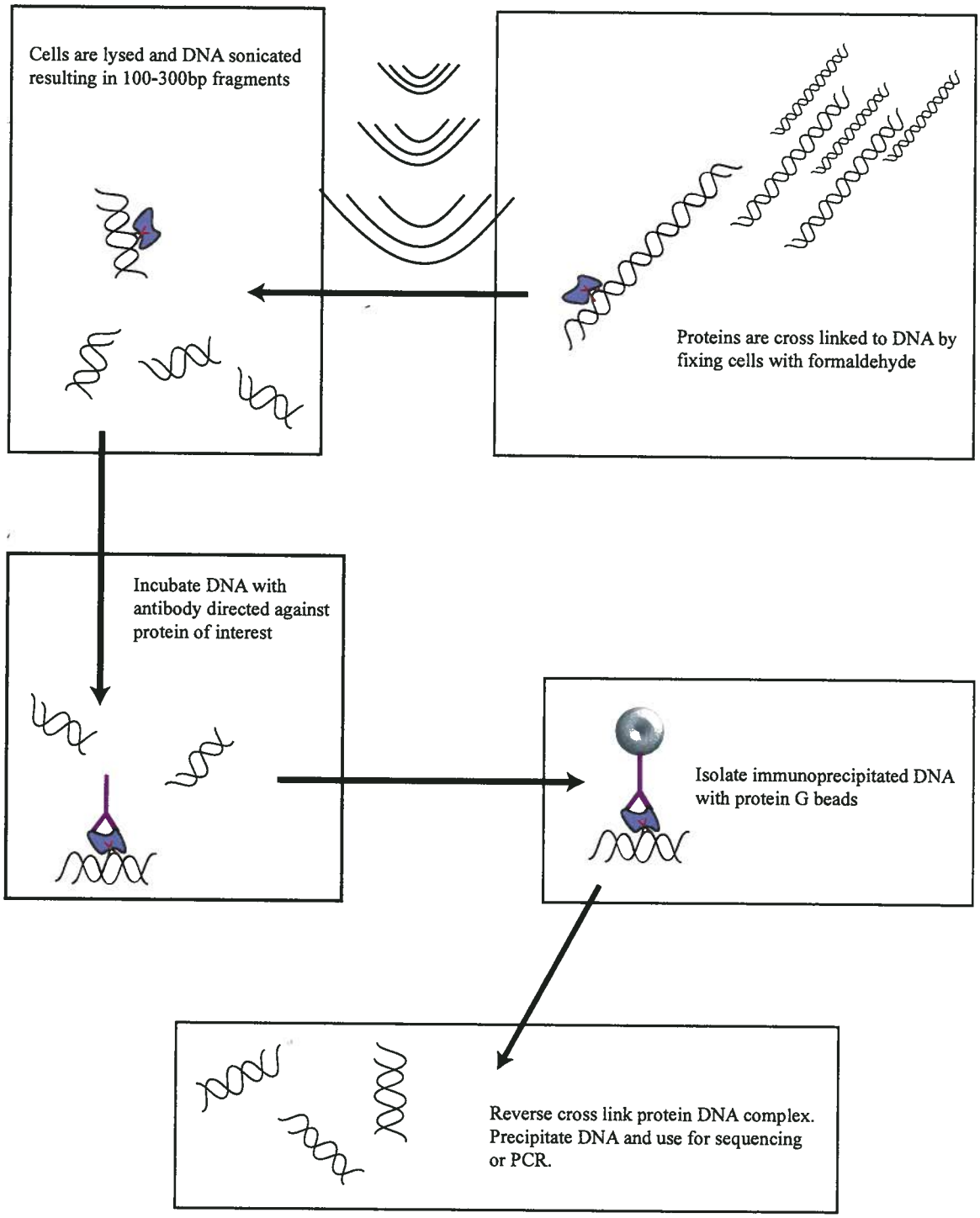


Figure 1 - Chromatin Immunoprecipitation. The steps of the ChIP procedure are depicted leading to the isolation of transcription factor bound DNA to be sequenced.

and increasingly cost effective. Over the last several years, there has been much variation in the precise technology employed to identify DNA fragments isolated by ChIP. Initial hybridization of ChIP DNA to promoter microarrays (ChIP-Chip) has proven extremely cost effective in identifying transcription factor binding regions[56, 57]. However, these studies are limited insofar as they bias their results solely to promoter regions, thereby failing to account for the majority of genomic sequence. An attempt to address this shortcoming was first made through the use of Sanger sequencing in ChIP-SACO[58], ChIP-SABE[59], ChIP-STAGE[60], and GMAT[61] studies. Nevertheless, these methods proved to be extremely cost limiting, and as such failed to present as reasonable options for identification of anything other than the most noteworthy of binding sites. More recently, ChIP-PET has made use of Roche 454 parallel pyrosequencing to identify binding sites of p53[62], Oct4[63], Nanog[63], and ERα[64]. While these studies marked significant improvement over previous techniques, ChIP-PET still cannot reach a cost-effective sequencing depth necessary to scrutinize an entire genome. It has only been with the emergence of flow cell sequencing technologies that our ability to confidently and cost-effectively identify binding sites at a genome wide level has truly emerged through the ChIP-Seq method[65].

The use of flow cell sequencing in ChIP-Seq allows for tens of millions of DNA fragments to be sequenced in a single run on parallel lanes. Currently, Roche 454 and Illumina represent the two most commonly used flow cell sequencers. For the purposes of ChIP-Seq, the Illumina device is superior to that offered by Roche 454 due to its ability to generate ten times the number of DNA sequences at approximately one tenth the cost. These flow cell technologies require as little as 10ng of input DNA for

sequencing. Comparatively, ChIP-ChIP procedures require 4-5μg of material[65]. Moreover, as a constantly advancing technology, the cost associated with flow cell sequencing is continually lessening. The improving cost-effectiveness of ChIP-Seq is noteworthy, as the main competing methodology, despite its aforementioned bias and limitations, continues to be ChIP-ChIP due to its low cost. In fact, a ChIP-ChIP study of Pdx1 binding in an insulinoma NIT-1 cell line has been published previously[56]. However, with this study being cell line based in addition to encompassing the inferiorities of ChIP-ChIP as compared to ChIP-Seq, our work sought to provide a far superior representation of genome-wide Pdx1 binding through ChIP-Seq in primary tissue, pancreatic islets.

Hypothesis, Aims, and Objectives

Curative options for diabetes, an increasingly prevalent worldwide disease characterized by an inability to regulate blood glucose levels, find the most substantial promise in islet transplant. A major limitation to islet transplantation is the scarcity of tissue, a shortcoming that can be addressed if islet pools can be either expanded or derived from stem cell precursors. To manipulate these cells, a much clearer understanding of the molecular physiology of the β-cell is required. A cell's properties are defined by the selective expression of its genome, which is controlled largely by transcription factors, and in the β-cell the foremost of these is Pdx1. Therefore, to begin to develop a truly in depth knowledge of the molecular workings of the β-cell, the purpose of this work is to attempt to characterize the genome-wide nature of Pdx1 binding in the pancreatic islet through the use of ChIP-Seq. I hypothesize that Pdx1 plays a major role in the β-cell transcriptional network, that a substantial percentage of its binding occurs at DNA regions distal to transcriptional start sites, and that much of its binding is facilitated by cooperative partners.

CHAPTER 2. MATERIALS AND METHODS

2.1 Tissue Culture

The mouse insulinoma adherent cell line MIN6 was maintained in 10cm tissue culture dishes (BD Biosystems) at a minimum 40% confluency and incubated at 37°C and 5.2% CO_2 in high glucose Dulbecco's Modified Eagles Medium (DMEM) (StemCell Technologies) containing 10% Fetal Bovine Serum (FBS) (Invitrogen) and 1% L-Glutamine (Invitrogen). Cells were passaged once per week at a confluency of 80-100% using Trypsin-EDTA (TE) (Invitrogen) and a centrifugation speed of 1200rpm for 5 minutes.

2.2 Mouse Colony

C57B1/6J and ICR mice were maintained in the Animal Resource Centre at the BC Cancer Research Centre in Vancouver according to the guidelines of the Canadian Council on Animal Care and protocols approved by the Animal Care Committee of UBC.

2.3 Western Blotting

A single well of a 24-well plate (BD Biosystems) of adherent MIN6 cells was harvested using TE, centrifuged at 1200rpm for 5 minutes to pellet cells, washed with 1mL of ice-cold 1X Phosphate Buffered Saline (PBS) (StemCell Technologies), and centrifuged again to obtain the clean cell pellet. 100μL of Radio Immuno Precipitation Assay (RIPA) lysis buffer (75mM NaCl, 1mM ethylenediaminetetraacetic acid [EDTA], 50mM Tris-HCl pH 7.25, 0.5% Triton X-100, Protease Inhibitor (PI) @ 1/100) was added and the tube incubated on ice for a minimum of 10 minutes. The lysate was heated

at 96°C for 5 minutes, and placed on ice. A pre-cast polyacrylamide gel (Invitrogen) was loaded into the running dock and 3-(N-morpholino) propanesulfonic acid sodium dodecyl sulfate (MOPS-SDS) running buffer added (Invitrogen). Precision Plus Protein Ladder (BioRad) and MIN6 lysate were added to independent wells and the gel run at 150V for 1 hour. A transfer membrane was submerged in methanol (Sigma Aldrich) for 30 seconds, removed, and submerged in NuPAGE Transfer Buffer (Invitrogen) containing 10% methanol. The gel was removed from its casting tray and the transfer apparatus assembled using the soaked transfer membrane. Protein transfer to the membrane was carried out by running at 35V for 1 hour in NuPAGE Transfer Buffer. Following transfer, the membrane was removed and blocked with 5mL Tris Buffered Saline Tween-20 (TBST) containing 5% milk powder for 1 hour at 4°C. Blocking solution was removed and 5mL of new blocking solution containing primary Pdx1 antibody (Upstate-Chemicon) at 1μL/1000μL was added to the membrane and incubated overnight at 4°C on a rocking platform. The next day, the membrane was washed 3X for 10 minutes with TBST after which 5mL blocking solution containing secondary antibody at 1μL/10,000μL was added and the membrane incubated for 1 hour on a rocking platform at room temperature. Washes with TBST were done 3X for 10 minutes each, after which a 1:1 mix of Detection Reagent 1 and Detection Reagent 2 (Amersham) were added to the membrane which was subsequently taken for exposure and film (Kodak Chemiluminescent BioMax Light) development in a dark room.

2.4 Islet Isolations

To isolate mouse pancreatic islets, C57Bl/6J (Jackson labs) and ICR mice aged 6 to 8 weeks were sacrificed via CO_2 asphyxiation and a midline incision made to expose the inner abdominal and thoracic cavities. Liver lobes were folded upwards to reveal the gall bladder and common bile duct running to the duodenum. A clamp was placed at the major duodenal papilla, the point of connection between the common bile duct and the duodenum, preventing fluid flow to the intestine and limiting it exclusively to the pancreas. Using a 26-gauge needle, 3mL of chilled collagenase (Sigma Aldrich) at 1000 units/mL in 1X Hanks Balanced Salt Solution (HBSS) (Invitrogen) was injected through the common bile duct to perfuse the pancreas. The swelled pancreas was scraped away from the intestine and placed in a 50mL Falcon tube. As multiple pancreases were collected, they were distributed such that each 50mL Falcon tube contained two pancreases and an additional 6mL of collagenase solution was added to each tube. The tubes were immediately placed in a 37°C water bath for 15-20 minutes to facilitate tissue digestion. Next, a transfer pipette was used to mechanically disrupt the contents of each tube until the mixture became homogenous. To stop digestive activity, 20mL of ice-cold 1X HBSS containing 0.25% Bovine Serum Albumin (BSA) (Roche) and 0.1M $CaCl_2$ was added and the tubes placed on ice. The tubes were centrifuged for 1 minute at 1120rpm, the supernatant was poured off, the remaining pellet was washed with 20mL of HBSS, and again centrifuged at 1120rpm for 1 minute. This wash was repeated at least three times, or until the supernatant appeared clear. Pellets were resuspended in 20mL HBSS and exocrine tissue was removed by filtering the solution through a pre-wetted 70μM nylon mesh filter (Fisher Scientific). The contents of the filter were washed with HBSS

into a 10cm petri dish (BD Biosystems) and placed under a stereomicroscope where islets were handpicked into a microcentrifuge tube using a 20µl pipette. Once a clean prep of islets was obtained, a single-cell suspension was created by adding 400uL of Enzyme-free Cell Dissociation Buffer (Gibco) and incubating the tube at room temperature for 12-15 minutes. During this time, islets were gently pipetted up and down every 3 minutes to facilitate dissociation. After a single-cell suspension was acquired, cells were centrifuged at 1200rpm for 1 minute, supernatant was removed, the cell pellet washed with 1mL 1X PBS, and centrifuged again at 1200rpm. The resultant cell pellet was then ready to be used for subsequent experiments. Islet isolation images are shown in Figure 2.

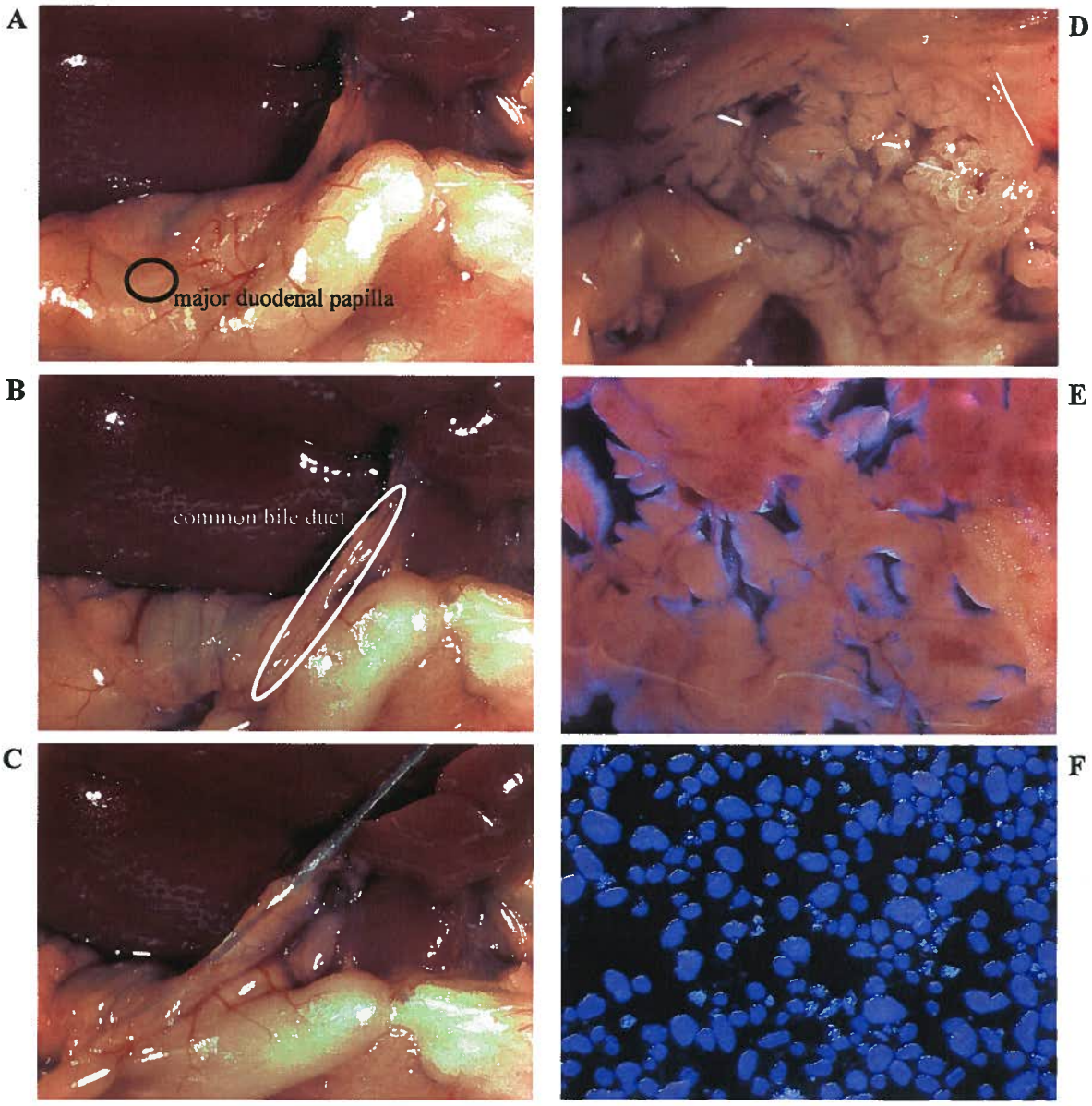


Figure 2 - Islet Isolations. Panels A through F show images of pancreatic islet isolation. In panel A, the major duodenal papilla is labelled marking the site of clamp placement. The bile duct is labelled in panel B, marking the location of syringe insertion seen in panel C. The perfused pancreas is clearly seen in D and magnified in E. Following digestion, washes, and filtration a clean preparation of islets is obtained through pipette picking (F).

2.5 Chromatin Immunoprecipitation

ChIP experiments were carried out in a manner similar to published previously[66]. MIN6 cells or a single-cell suspension of islets were collected and washed in 1mL 1X PBS and centrifuged at 1200rpm for 2 minutes. The cell pellet was resuspended in 1360μL of 1X PBS and 38μL of 37% formaldehyde (Fisher Scientific) was added to crosslink the cells. This fixation was carried out for 10 minutes on a rotating platform at room temperature, after which 175μL of 1M glycine (Invitrogen) was added and the suspension rotated for another 5 minutes to stop the fixation. Cells were centrifuged at 4000rpm for 2 minutes, washed in 1mL 1X PBS, and again centrifuged at 4000rpm for 2 minutes. To lyse the cellular membrane, 500μL of cold ChIP cellular lysis buffer (10mM Tris-Cl pH8.0, 10mM NaCl, 3mM $MgCl_2$, 0.5% NP-40, PI @ 1/100) was added to the cell pellet and the solution dounce homogenized for 10 strokes. The resulting suspension was then incubated on ice for at least 5 minutes and centrifuged at 13,200rpm for 3 minutes. The nuclear membrane was lysed to release the chromatin by adding 100μL of cold ChIP nuclear lysis buffer (1% SDS, 5mM EDTA, 50mM Tris-Cl pH8.0, PI @ 1/100) to the cell pellet and resuspending the cells by passing them through a 26-gauge needle for 5 strokes. Shearing of the resulting chromatin was accomplished through sonication (S3000 Ultrasonic Cell Disruptor Processor, Fisher) of the solution as follows: 10 minutes total sonication time, 1 minute on followed by 30 seconds off, in an ice water bath at 50% output power. Undissolved debris was pelleted and removed by centrifuging at 13,200rpm for 10 minutes and moving the supernatant to a new tube. $1/20^{th}$ of this supernatant was removed to a new tube and ChIP nuclear lysis buffer added to a final volume of 200μL. To this, 8μL of 5M NaCl was added and the tube incubated at 65°C

overnight to reverse crosslink the sample as an input control. To the 95µl of remaining supernatant, 42.5µL of ChIP nuclear lysis buffer and 7.5µL of ChIP spike buffer (10X concentrate of ChIP dilution buffer – 0.01% SDS, 1.1% Triton X-100, 167mM NaCl, 16.7mM Tris-Cl pH8.0, PI @ 1/100) were added making the total volume 150µL. 20µL of Protein G agarose beads (Pierce) were then added to pre-clear the solution by mixing on a rotating platform at 4°C for 1 hour. Beads were spun down at 13,200rpm for 30 seconds and the supernatant transferred to siliconized tubes. 3µg of Pdx1 (Upstate – Chemicon) antibody was added to the supernatant and for each ChIP reaction a separate tube of 20µL of protein G beads were added to 1mL of ChIP dilution buffer supplemented with 1mg/mL BSA, and 0.1mg/mL salmon sperm DNA (Invitrogen) to block the beads. Both the supernatant and the beads were incubated overnight at 4°C on a rotating platform.

The next day, the beads were centrifuged at 13,200rpm for 30 seconds and the supernatant was removed. The antibody mixture was added to an aliquot of blocked beads and placed back on the rotating platform at 4°C for 3 hours. Beads were centrifuged at 13,200rpm for 30 seconds, supernatant removed, and beads washed as follows: 5 minutes in low salt buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-Cl pH8.0, 150mM NaCl), 5 minutes in high salt buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-Cl pH8.0, 500mM NaCl), 5 minutes in LiCl buffer (0.25M LiCl, 1% NP-40, 1% Deoxycholate, 1mM EDTA, 10mM Tris-Cl pH8.0), and 2 washes for 5 minutes each in TE Buffer (10mM EDTA, 10mM Tris-Cl pH8.0). Following these washes, 150µL of elution buffer (1% SDS, 0.1M $NaHCO_3$) was added to the beads, the solution transferred to a fresh tube, and incubated at 50°C on a rotating platform for 1

hour. Beads were centrifuged at 13,200rpm for 30 seconds and the supernatant containing eluted chromatin was transferred to a new tube. An additional 50μL of elution buffer was added to the beads and they were again centrifuged and the supernatant removed and combined with the initial 150μL. To reverse crosslink the eluted chromatin in the ChIP sample, 8μL of 5M NaCl was added and the tube incubated overnight at 65°C on a rotating platform. The DNA from the input sample reverse crosslinked from the previous day was extracted via phenol-chloroform extraction. Similarly, DNA from the ChIP sample was also phenol-chloroform extracted the following day.

2.6 Phenol-Chloroform Extractions

To extract the DNA from input and ChIP samples, Buffer Saturated Phenol (Invitrogen) was combined with chloroform (Fisher Scientific) in a 1:1 ratio. An equivalent volume of this mixture was added to the sample to be extracted and the tube shaken vigorously to mix. After letting stand for 5 minutes to allow phase separation to begin, the sample was centrifuged at 13,200rpm for 10 minutes. The uppermost aqueous phase was removed and transferred to a new tube where a 3X volume of ice-cold 100% ethanol was added and the tube let stand for 30 minutes to precipitate the DNA. Following precipitation, the tube was centrifuged at 13,200rpm for 10 minutes and the supernatant aspirated leaving the invisible DNA pellet. This pellet was resuspended in 20μL DNase RNase free water (Invitrogen).

2.7 Illumina Sequencing of ChIP DNA and Peak Building

Chromatin of 100-300bp was selected by running the sample on a 12% PAGE gel, excising all material found in that size range, and purifying using a Spin-X filter column (Costar) and ethanol precipitation by Baljit Kamoh at the Genome Sciences Centre (GSC). Subsequently, the isolated DNA was sequenced using the Illumina genome analyzer[67] located at the GSC. Briefly, PCR amplification of the DNA was performed using ligated adapters to the size selected fragments for use as primers. The resultant PCR products were affixed to a flow cell where "bridge" amplification was employed to produce clonal clusters of identical DNA fragments. To sequence these fragments, a primer homologous to the ligated adapters was annealed and sequence by synthesis performed using reversibly terminated fluorescently labelled nucleotides. Following each cycle of nucleotide addition, the flow cell image was captured using fluorescence microscopy. At the end of the sequencing run, the combined images were used to make base calls providing sequence information for the affixed fragments. The Illumina sequencing method is displayed in Figure 3.

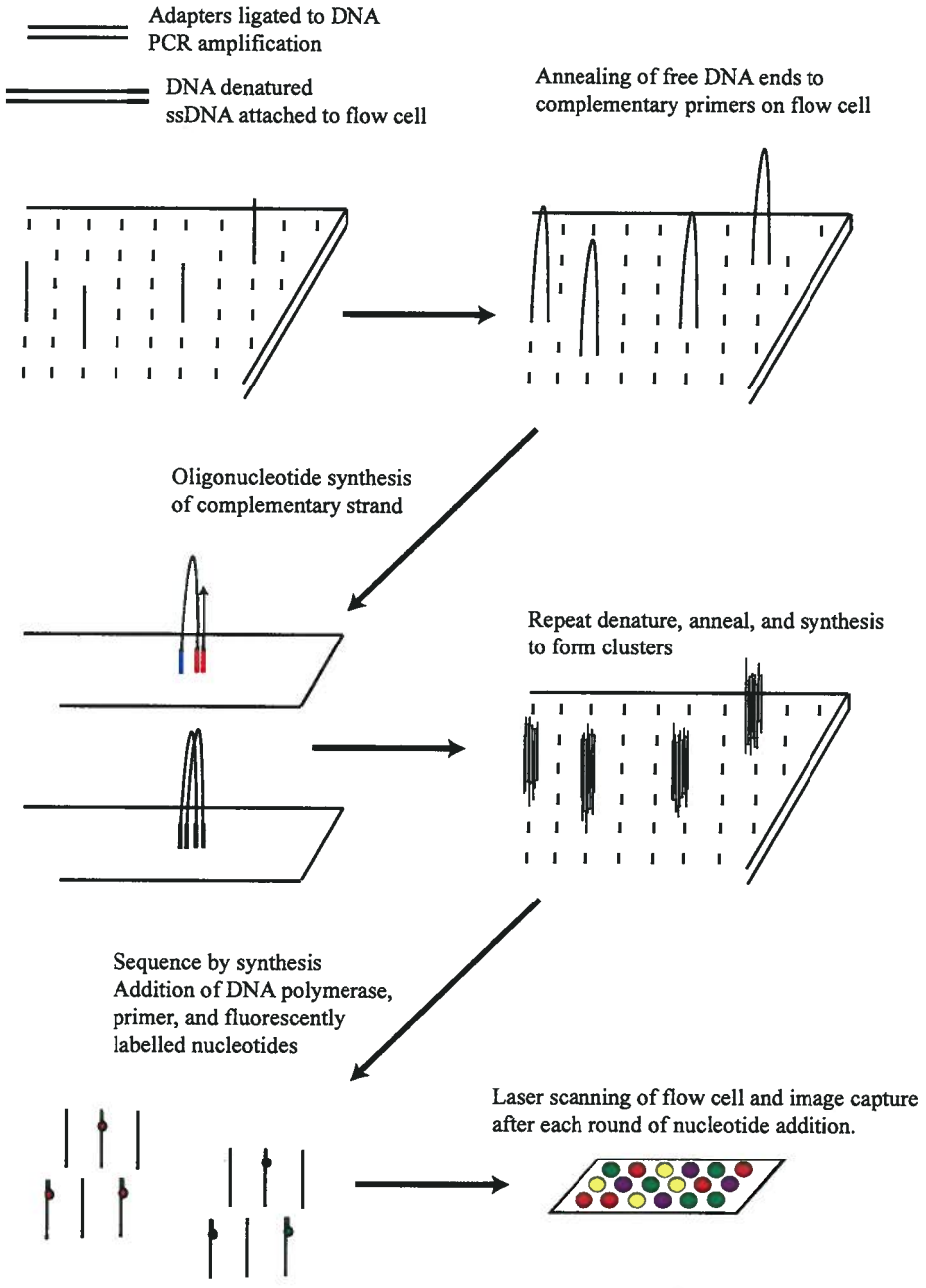


Figure 3 - Illumina Flow Cell Sequencing by Synthesis.

Peaks were constructed from sequenced DNA using the computational tool FindPeaks3.1[68]. The FindPeaks algorithm is utilized to analyze short-read sequencing experiments to identify areas of enrichment and produce a "wig" file that can be uploaded to the UCSC genome browser website. Sequence reads are aligned to the genome and regions of protein-DNA interaction have an enriched concentration of reads compared to an islet input control background model. Sites of enrichment between the protein of interest and the genomic DNA are defined as peaks. A representation of the peak building process is depicted in Figure 4.

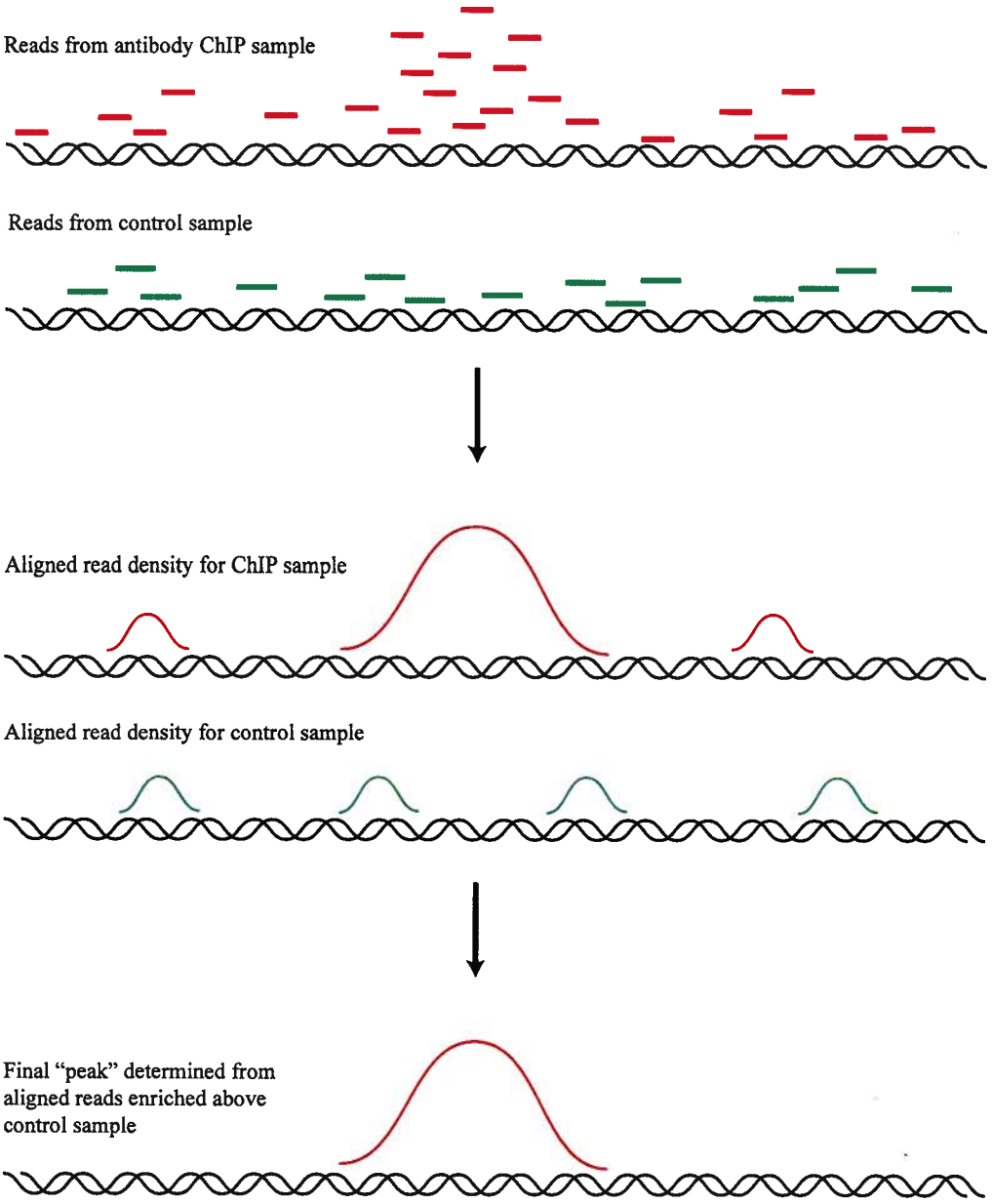


Figure 4 - Constructing Peaks from ChIP-Seq Data. Sequenced reads from ChIP DNA are aligned to the genome. Read density is compared against a control background sample to determine areas of read density enrichment. Where read density is greater than the background control employed, a "peak" is defined.

2.8 qPCR

Reactions were set up with the following components: 4μL SYBRFast (Applied Biosystems), 0.5μL ChIP DNA, 1μL primer mix at 10μM of both forward and reverse, and 4.5μL dH_2O. Reaction plates (Applied Biosystems) were run on a 7500 Fast Real Time PCR System (Applied Biosystems) with cycle conditions of 95°C for 20 seconds, followed by 40 cycles of 95°C for 3 seconds and 60°C for 30 seconds.

2.9 Islet siRNA Transfection

Islets were extracted from C57Bl/6J mice and a single cell suspension created to plate islet cells to 24-well plates at an average confluency of 100,000 cells per well. Cells were cultured overnight at 37°C, 5.2% CO_2 in Royal Park Memorial Institute (RPMI) (StemCell Technologies) media containing 10% FBS and 1% L-Glutamine. Pdx1 and control siRNAs (Dharmacon) were prepared to 2μM solutions in 1X siRNA Buffer (Dharmacon). For each well, 20μL of targeted siRNA was combined with 5μL siGLO indicator (Dharmacon) and 25μL OPTI-MEM serum free media (Invitrogen). In a separate tube, 2μL of DharmaFECT4 transfection reagent was combined with 48μL OPTI-MEM and tubes incubated at room temperature for 5 minutes. The contents of both tubes were combined and incubated at room temperature for an additional 20 minutes and added to each well along with fresh RPMI media. Transfected cells were cultured for 48 hours and harvested for Fluorescence Activated Cell Sorting (FACS).

2.10 FACS

Islet cells were harvested into PBS and dead cells stained with 7-amino actinomycinD (7AAD) at 1/100. Sorting was performed on the BD FACS Vantage SE DiVa in the Terry Fox Lab Flow Cytometry Unit at the BCCRC. Cells were gated to remove 7AAD positives and doublets, while cells positive for siGLO were sorted directly into Trizol (Invitrogen).

2.11 RNA Isolation and RT

Cells from FACS were placed into Trizol and a 1/5 volume of chloroform was added and the tube shaken vigorously. Following a 2-minute incubation at room temperature, samples were centrifuged at 13,200rpm for 10 minutes, supernatants removed, and RNA extracted via manufacturer's protocol using an RNEasy Kit (Qiagen). RNA Pellets were suspended in 20μL DNase RNase free water and a small portion used for subsequent reverse transcription (RT), with the remainder being used for Tag-Seq-lite library construction (section 2.12).

RT was performed as follows. 1μL of islet RNA was added to 1μL 10X DNase 1 reaction buffer (Invitrogen), 1μL Amp grade DNase 1 @ 1U/μL (Invitrogen), and DNase RNase free water to a final volume of 10μL. Tubes were incubated for 15 minutes at room temperature and DNase 1 inactivated by addition of 1μL of 25mM ethylenediaminetetraacetic acid (EDTA) (Invitrogen). Following a 10 minute incubation at 65°C, 250ng of random primers (Invitrogen) and 1μL of 10mM dNTP mix (Invitrogen) were added and the mixture heated for an additional 5 minutes at 65°C. After letting the tube sit on ice for 1 minute, the following were added: 4μL 5X First Strand Buffer

(Invitrogen), 1μL 0.1M dithiothreitol (DTT) (Invitrogen), 1μL RNaseOUT Recombinant RNase Inhibitor (Invitrogen), and 1μL SuperScript III RT @ 200U/μL (Invitrogen). Contents were pipetted up and down and incubated at 25°C for 5 minutes. Incubation temperature was increased to 50°C for an additional 60 minutes after which the reaction was inactivated by again increasing the temperature to 70°C for another 15 minutes. The resultant cDNA was subsequently used for qPCR analysis as outlined in 2.8.

2.12 Tag-Seq-lite

Tag-Seq-lite library construction was performed by the Genome Sciences Centre as described previously[69]. First strand cDNA was synthesized from 40ng of DNAse1 treated islet RNA (control or siPdx1 treated) with Superscript III Reverse Transcriptase (Invitrogen) and amplified by 20 cycles of PCR based on SMART (Switching Mechanism At the 5' end of RNA Transcripts) cDNA synthesis to generate full-length cDNA (Clontech). Subsequently, 500ng of cDNA was digested with the anchoring enzyme NlaIII and ligated to an Illumina specific adapter containing a recognition site for the type IIS tagging enzyme Mme1 as well as sequencing and PCR primers. After digestion with Mme1 and SAP (Shrimp Alkaline Phosphatase) treatment to dephosphorylate the DNA, a second Illumina adapter containing a 2bp 3' overhang was ligated. The resultant "tags" flanked by adapters were amplified via PCR using Phusion polymerase with the following cycling conditions: 98°C for 30 seconds, followed by 13 cycles of 98°C for 10 seconds, 60°C for 30 seconds, and 72°C for 15 seconds, and then 72°C for 5 minutes. PCR products were purified by running samples on a 12% PAGE gel, excising the 85bp band, and purified using a Spin-X filter column and ethanol

precipitation. Quality assessment and DNA amount were determined using an Agilent DNA 1000 series II assay (Agilent) and DNA then diluted to 10nM. DNA was sequenced using the Illumina Genome Analyzer and 17bp Serial Analysis of Gene Expression (SAGE) tags extracted from the resulting reads. The process of Tag-Seq-lite is depicted in Figure 5.

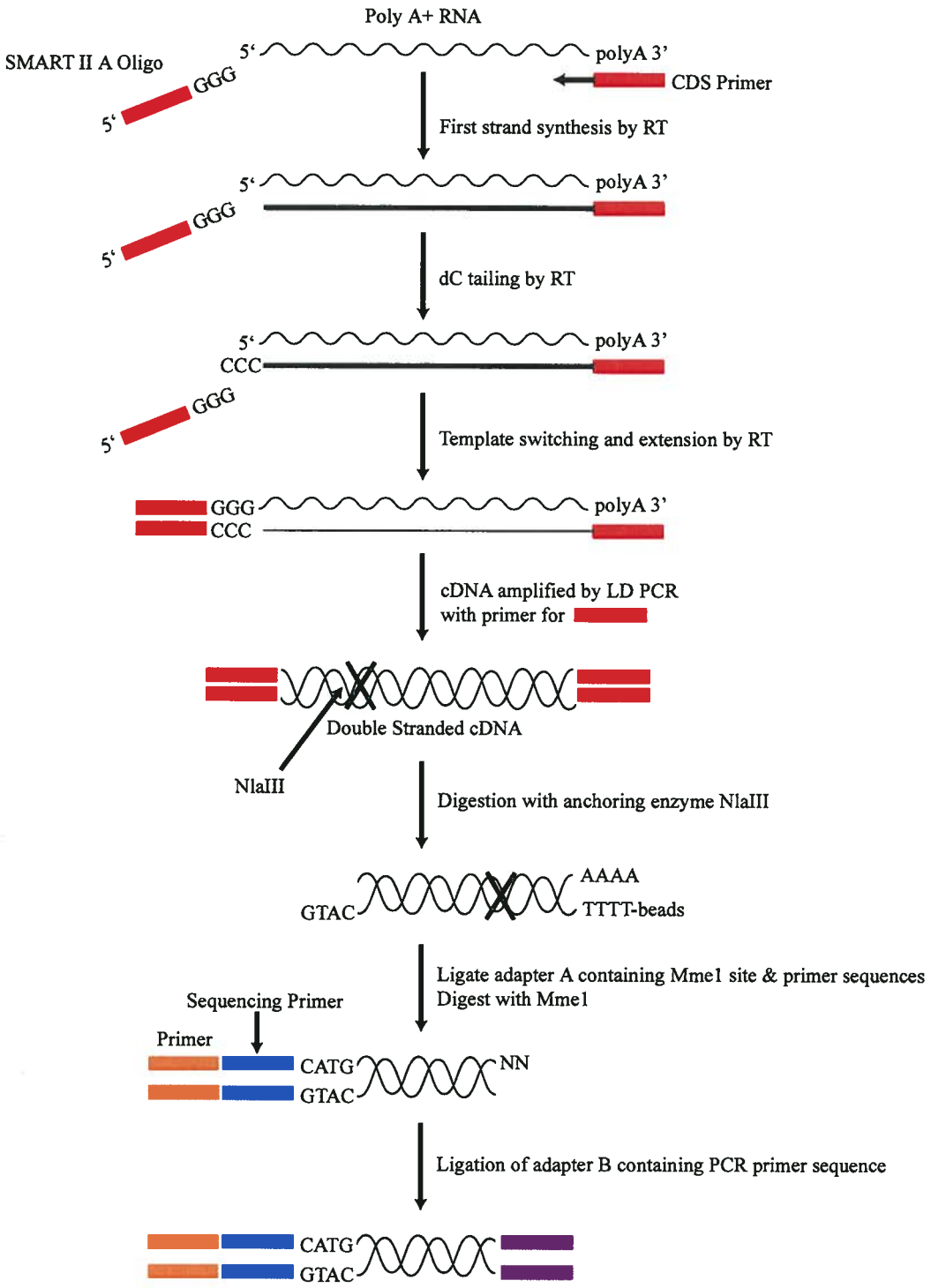


Figure 5 - Tag-Seq Lite Library Construction. The process of Tag-Seq is depicted. Following the final step, 13-17 cycles of PCR are performed to amplify the DNA which is then purified on a PAGE gel and sequenced via Illumina.

2.13 Tag-Seq Library Bioinformatics

Tags were mapped to Refseq genes using Discovery Space 4.0[70]. Tags with a count greater than 5 were included in the analysis. To account for multiple tags mapping to the same Refseq accession, the counts for all tags mapping to the same Refseq were combined, providing a Refseq and an associated count. Counts were normalized based on library size and expressed as counts per million. Normalized counts of genes were subsequently compared between the control and Pdx1 siRNA library to determine which genes were significantly down and up regulated in the Pdx1 siRNA library compared to the control.

2.14 Seeded Motif Discovery

Seeded motif discovery was performed by Gordon Robertson and Leping Li. GADEM[71] (A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery) addresses large sequence sets and identifies highly prevalent motifs based on a user specified threshold. A modified version of GADEM was used that employed an initial "seed" position weight matrix (PWM) provided by the user. A motif is deemed significantly present if its E-value, produced by both its p-value and the number of all possible motif-length segments in the search space, falls below this threshold.

Pdx1 and Pbx1 binding motifs were identified from ChIP-Seq data using the seed PWMs IPF1_Q4_01, TRANSFAC M101013 for Pdx1, and PBX1_02, TRANSFAC M00124 for Pbx1. Threshold was established by setting the p-value limit to $5e^{-4}$, and the GADEM run provided Pdx1-like and Pbx1-like motifs.

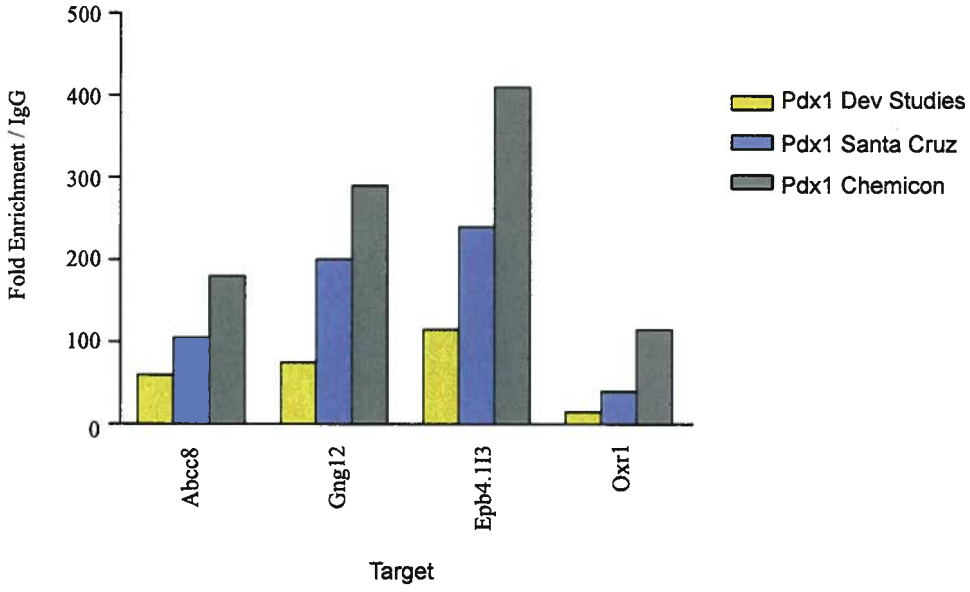
CHAPTER 3. RESULTS

3.1 Pdx1 ChIP-Seq Library Construction

3.1.1 Identification of a ChIP Quality Pdx1 Antibody and Pdx1 Targets

The first step in a successful ChIP experiment is to identify a suitable antibody. Therefore, to identify the best Pdx1 antibody candidate for use in ChIP, antibodies directed against Pdx1 were purchased from Developmental Studies Hybridoma Bank, Chemicon (Upstate), and SantaCruz. The initial comparative ChIP trials were performed using MIN6 cells and enrichment at several ChIP-ChIP identified Pdx1 targets[56] was tested using qPCR. Pdx1 ChIPs were performed using a fully confluent 10cm plate of cells and enrichments established based on comparison against a control IgG ChIP. Figure 6a shows that while all antibodies produced enrichment of Pdx1 targets, in every case the best performance was observed using the Chemicon antibody, followed by SantaCruz and Developmental Studies. The most highly enriched target, Epb4.1l3, was selected as a positive control to assess the degree of success of future ChIPs performed in islets. To ensure Chemicon Pdx1 antibody fidelity, a Western Blot was performed. Figure 6b shows that a single clear band at the expected size of roughly 35kDa, corresponding to Pdx1 protein, was observed. Therefore, based on these results, the Chemicon Pdx1 antibody was selected for use in all future experiments.

A



B

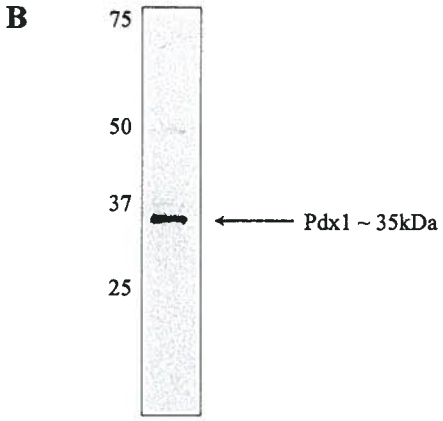


Figure 6 - Identification of a ChIP Quality Pdx1 Antibody. (A) ChIP-qPCR fold enrichments compared to IgG of Pdx1 targets following ChIP with the Pdx1 antibodies from Developmental Studies, Santa Cruz, and Chemicon. The high levels of enrichment of targets over IgG controls indicate ChIPs were all successful. For all targets the best ChIP enrichments are clearly displayed using the Pdx1 antibody from Chemicon. Maximal enrichment at the Epb4.1l3 target identify it as a good positive control for future ChIPs. (B) Western Blot for Pdx1 performed with the Chemicon Pdx1 antibody using MIN6 cell lysate. Presence of a strong and clean band at the expected size of ~35kDa shows the high specificity of the Chemicon antibody.

3.1.2 Collection of Islet Pdx1 ChIP DNA

To obtain primary tissue for Pdx1 ChIPs, seven islet isolations were performed by Galina Soukhatcheva at the Verchere lab at the Child and Family Research Institute, from a total of fifty-six C57Bl/6J mice, over the course of three months. Each islet isolation yielded a minimum of one thousand islets, which were immediately taken as far as the stop fixation step of the ChIP protocol. To maintain ChIP procedural consistency with other ChIP libraries, the fixed cells were delivered to the Genome Sciences Centre (GSC) and ChIPs performed by the GSC gene expression pipeline by Balgit Kamoh. Islets from the first three isolations were used to optimize the ChIP protocol for the creation of a single-cell suspension as well as ideal chromatin shearing conditions. Following this optimization, islets from the remaining four isolations produced chromatin that was of the correct size range of 100-300bp and showed enrichment at the Epb4.1I3 positive target for Pdx1. Figure 7a shows the agilent size range profiles of the sheared DNA from each of the four replicates, and Figure 7b displays qPCR enrichment for the abovementioned Epb4.1I3 target. Taken together, these results indicate that islet chromatin has been sheared sufficiently, and that Pdx1 ChIPs have been successful. Subsequently, the chromatin from these four successful ChIPs was pooled and used for Illumina sequencing and library creation.

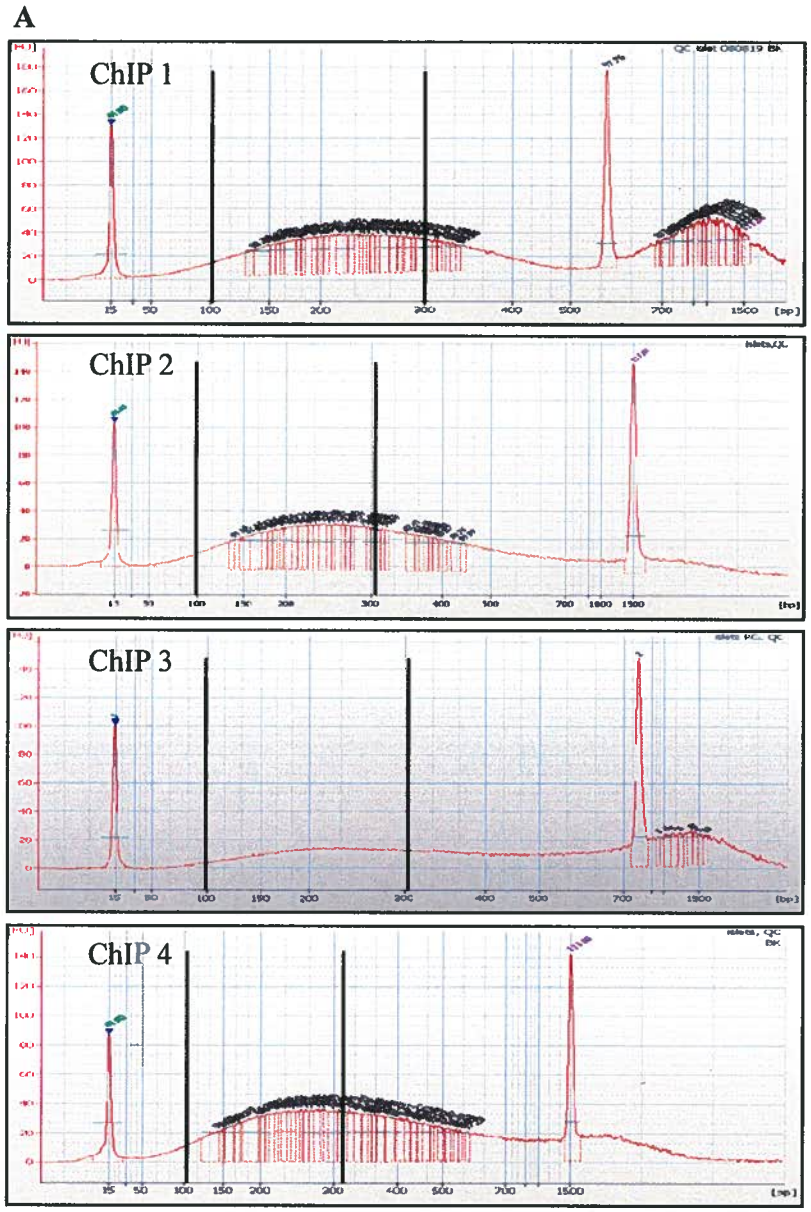


Figure 7 - Validating the Islet Pdx1 ChIP DNA.
(A) Agilents displaying the size range profiles of the sonicated islet DNA going into each ChIP are shown. The 100-300bp size ranges are labeled between the black bars, confirming presence of chromatin in the desired size range.
(B) ChIP-qPCR fold enrichments of the positive Epb4.1l3 target are shown for each of the four successful ChIPs performed in islets. ChIP 1 enrichment was calculated against an IgG control. ChIPs 2 to 4 were compared against an input DNA control to maximize DNA amount going into the Pdx1 ChIPs.

3.2 Pdx1 ChIP-Seq Library Results and Validation

3.2.1 Statistics and Visualization of Pdx1 ChIP-Seq Peaks

In total, 7 lanes of Pdx1 ChIP material was sequenced using Illumina Flow Cell technology at the GSC, resulting in 62.1 million reads and a mapping efficiency of 24% to the mm8 mouse genome assembly. Following peak building using FindPeaks3.1 and the establishment of a peak height threshold of 11, the number of Pdx1 peaks was 13,448. To visually assess the data, the generated "wig" file was loaded into the UCSC genome browser to scan for peaks at known Pdx1 binding sites. Figure 8 shows the UCSC screenshots of several previously identified Pdx1 binding sites: Ins1, Ins2, Pdx1, Gck, IAPP, and Glut2. It clearly illustrates that Pdx1 ChIP-Seq peaks are located at most expected sites. Additionally, scanning of ChIP-Seq data in UCSC revealed Pdx1 sites at suspected, but previously unidentified, genes including Isl1, Nkx2.2, Nkx6.1, and Pax6 (Appendix). Peak to gene associations were performed in the Galaxy Genome Browser (http://main.g2.bx.psu.edu/) by mapping peaks to the closest Refseq transcriptional start site either up or downstream. A peak was defined as being associated with a gene if the closest transcriptional start site was within 50kb of the peak. Using this method, 5560 genes possessed a Pdx1 peak.

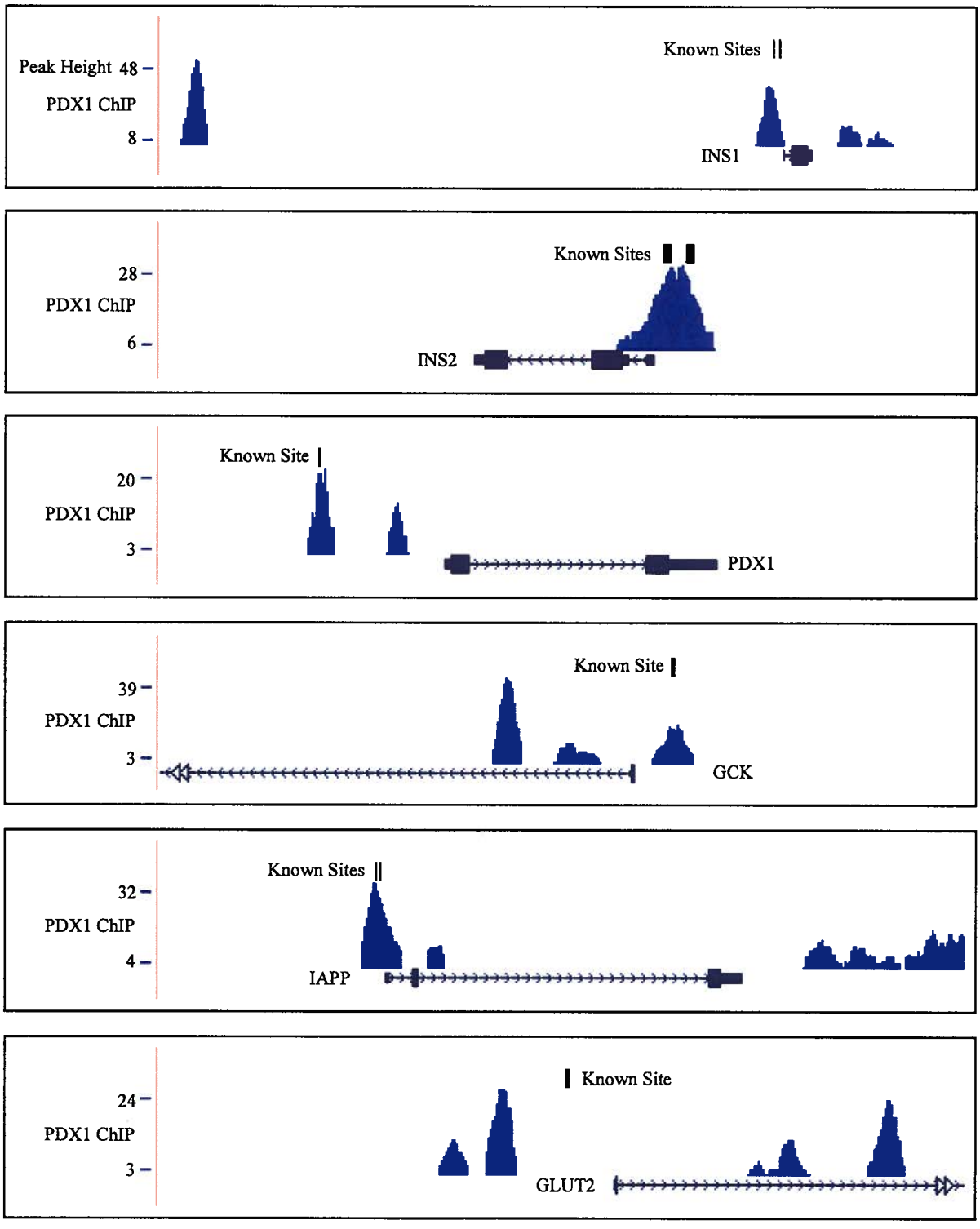
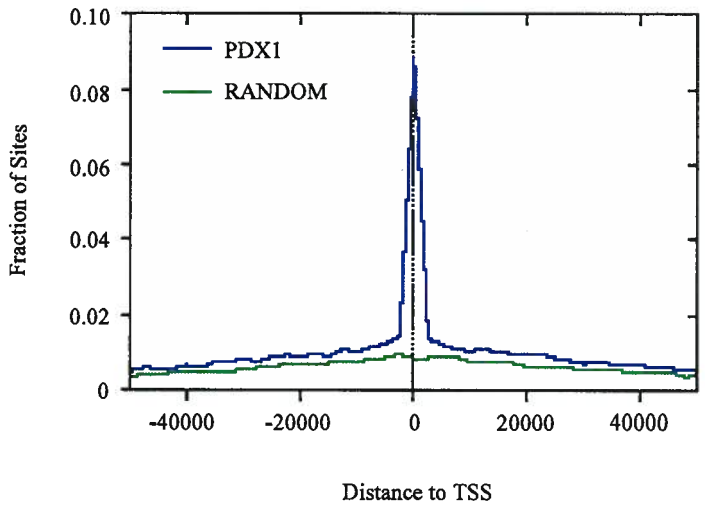


Figure 8 - UCSC Screenshots of Pdx1 ChIP-Seq at Known Sites. The previously Pdx1 identified binding sites at the Ins1, Ins2, Pdx1, Gck, IAPP, and Glut2 genes are shown. The known binding site(s) are labeled as black vertical dashes, ChIP-Seq peaks are shown in blue. ChIP-Seq peaks are present at all known sites with the exception of Glut2.

Because previous ChIP-Seq data had revealed an abundance of transcription factor binding sites located distally from transcriptional start sites, I examined the distribution of Pdx1 peaks to determine if a similar trend was present in this data. The fraction of Pdx1 peaks was plotted against distance to the closest transcriptional start site (Figure 9a). Compared to a random distribution of sites, Pdx1 peaks were highly centred at transcriptional start sites. Using Galaxy Genome Browser, peaks were overlapped with various genomic regions: promoters, enhancers, exons, introns, and regions >10kb from the TSS. This yielded a distribution of Pdx1 peaks as follows: 11% promoter (0-1kb upstream of the TSS), 8% enhancer (1-10kb upstream of the TSS), 4% exons, 27% introns, and 49% >10kb (Figure 9b). This type of distribution is consistent with previous ChIP-Seq studies[72], and illustrates that an abundance of sites were overlooked in ChIP-Chip experiments due to their bias towards promoter and enhancer regions only.

A



B

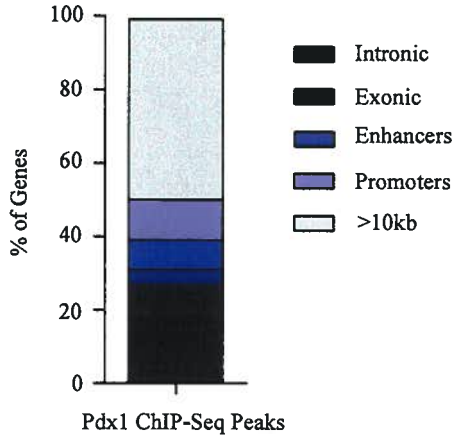
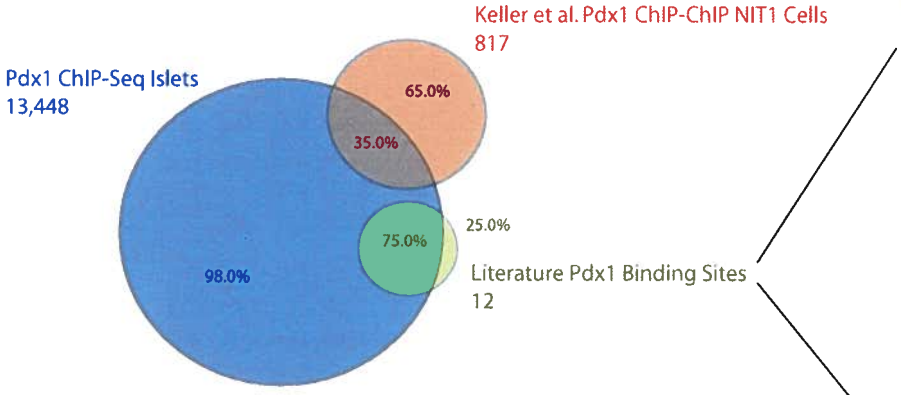


Figure 9 - Distribution of Pdx1 ChIP-Seq Peaks. (A) Histogram of the fraction of sites occurring relative to the position of the TSS. Pdx1 ChIP-Seq peaks are centred around TSS. (B) Distribution of peaks into gene regions. Peaks are found in each region as follows: >10kb away - 49%, promoters - 11%, enhancers - 8%, exonic - 4%, and intronic - 27%.

3.2.2 Validation of the Pdx1 ChIP-Seq Library

To validate the Pdx1 ChIP-Seq data, Pdx1 peaks were compared against known Pdx1 binding sites and previously published genome-wide binding data from ChIP-Chip studies performed in NIT-1 insulinoma cells[56]. Figure 10 shows this comparison. A 35% overlap between Pdx1 ChIP-Seq data and ChIP-Chip data was observed, and 75% of known Pdx1 binding sites are accounted for in the ChIP-Seq dataset while previous ChIP-ChIP data fails to identify these well-established binding sites. A table detailing and referencing the known sites is also displayed in Figure 10.

Additional validation was carried out using ChIP-qPCR to assess enrichment of 35 peaks identified from the ChIP-Seq data, as well as four negative targets. For these, four replicate Pdx1 ChIPs, as well as control IgG ChIPs, were performed on islets isolated by me from ICR mice. Islets were isolated from ten mice yielding at least one thousand islets for each of four replicate ChIPs, and enrichment of the positive Epb4.1l3 target was confirmed. The four ChIPs were pooled and qPCR reactions setup in quadruplicate to determine the enrichments shown in Figure 11a. All tested ChIP-Seq target sites were enriched over the negative controls. Importantly, a positive correlation was observed between ChIP-Seq peak height and ChIP-qPCR fold enrichment (Figure 11b).



Gene	Reference	ChIP Peak
Ins1 A1	German et al. 1995	Yes
Ins1 A3/4	German et al. 1995	Yes
Ins2 A1	German et al. 1995	Yes
Ins2 A3/4	German et al. 1995	Yes
Pdx1	Gerrish et al. 2001	Yes
Gck	Shelton et al. 1992	Yes
Glut2	Waeber et al. 1996	No*
IAPP A1	Carty et al. 1997	Yes
IAPP A2	Carty et al. 1997	Yes
Sst TSE1	Leonard et al. 1993	No**
Sst TSE2	Leonard et al. 1993	No**
Mafa	Raum et al. 2006	Yes

* We do identify Pdx1 binding peaks at Glut2 but not at the exact location outlined by Waeber.
** The absence of Pdx1 binding peaks at Sst in our data is not unexpected as Sst is expressed in delta cells which compose a minimal % of islet mass.

Figure 10 - Comparison of ChIP-Seq Data with ChIP-ChIP and Known Binding Sites. Islet Pdx1 ChIP-Seq data was compared against Pdx1 ChIP-ChIP data as well as a list of known binding sites. The Venn diagram shows that 35% of the binding regions identified in the ChIP-ChIP study are also accounted for in the ChIP-Seq data. It also displays that the total number of sites identified was much greater in the ChIP-Seq study (13,448) versus the ChIP-ChIP study (817). Of known Pdx1 binding sites identified from a literature survey, 75% have ChIP-Seq peaks at the exact location. Conversely, the ChIP-ChIP data fails to directly identify any of these sites. A summary of the known sites is provided in the expanded table, as well as explanations for the absence of Pdx1 ChIP-Seq peaks at those sites not identified.

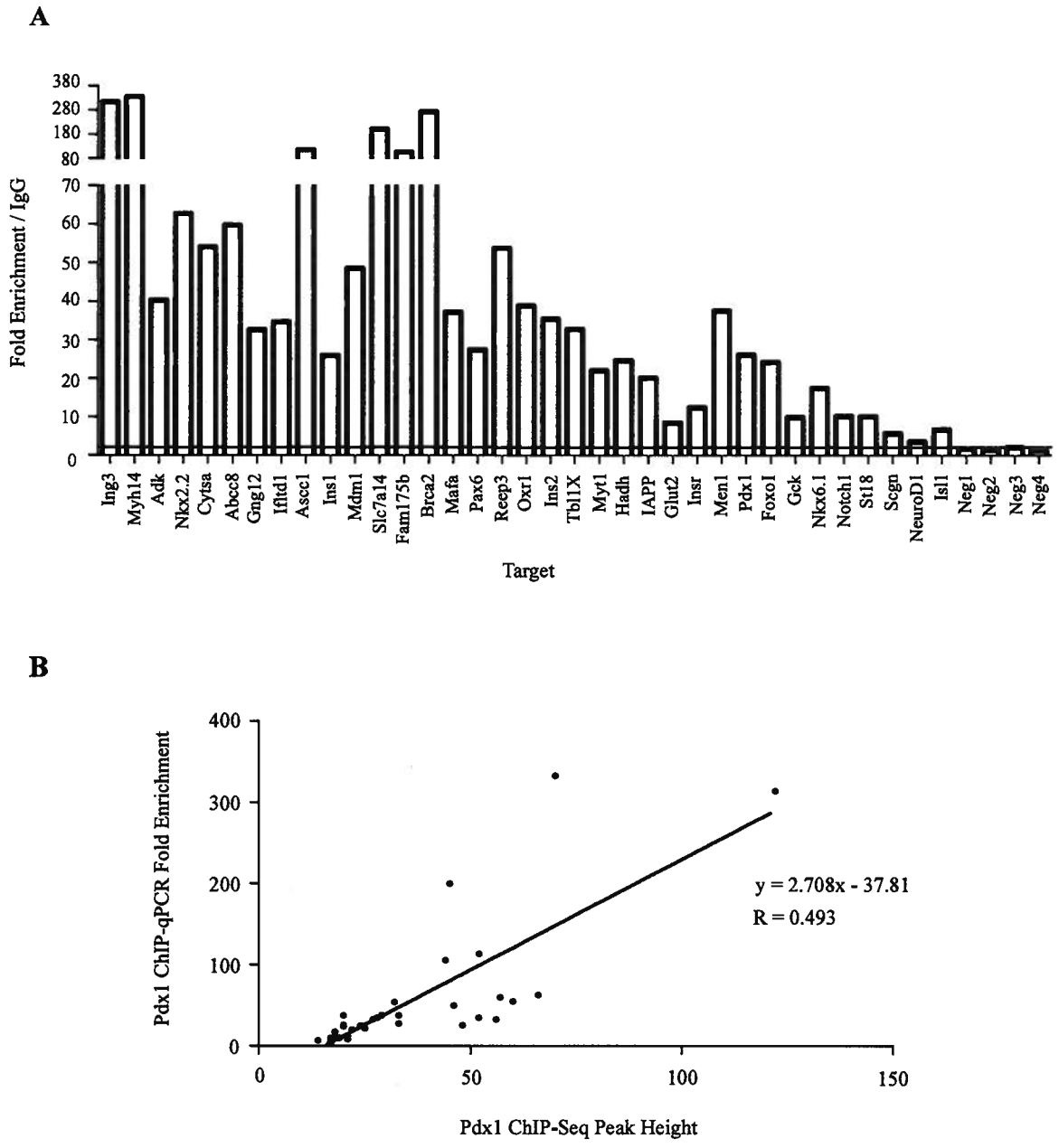


Figure 11 - ChIP-Seq Peaks are Validated Via ChIP-qPCR. (A) Pdx1 ChIP-qPCR results for Pdx1 targets identified in ChIP-Seq data. 35/35 peaks were validated compared to 0/4 negative control targets using isolated ICR islets. qPCR reactions were performed in quadruplicate on pooled material from 4 independent ChIPs. (B) The line of best fit on the scatter plot shows the positive correlation observed between ChIP-Seq peak height and ChIP-qPCR fold enrichment.

3.2.3 Validation Through siPdx1 Tag-Seq Library Construction

To determine genes most highly impacted by direct Pdx1 binding, gene expression libraries were created from islets treated with either control or Pdx1 siRNA. Following siRNA knockdown of Pdx1 or cyclophilin control, islet cells were harvested and sorted via FACS. Cells positive for knockdown were labelled green due to the presence of siGLO and collected directly into Trizol. FACS gating and results from the collection are shown in the appendix. A total of ~30,000 cells were collected for each condition. This approximately corresponded to a 10% transfection efficiency for both collections. RNA was extracted from the sorted cells and a portion used for RT-PCR to confirm Pdx1 knockdown (Figure 12a), while the remainder was sent to the GSC for Tag-Seq expression library construction as outlined in section 2.12. Tag mapping and comparison of the cyclophilin and Pdx1 siRNA libraries were performed using Discovery Space[70] and up and down regulated genes determined. 655 genes were up regulated, while 488 were down regulated, in the siPdx1 library as compared to the control. Relative expression levels of known Pdx1 positively regulated genes such as Ins1, Ins2, Pdx1, Glut2, IAPP, and Gck, displayed reduced expression in the knockdown library, corroborating Pdx1 knockdown was having a quantifiable effect on its targets (Figure 12b). Hence, gene lists were compared against the Pdx1 ChIP-Seq data to determine what portion possessed a Pdx1 peak. Figure 12c shows that 36% of unaltered genes, 39% of up regulated genes, and 45% of down regulated genes had a Pdx1 ChIP-Seq peak association.

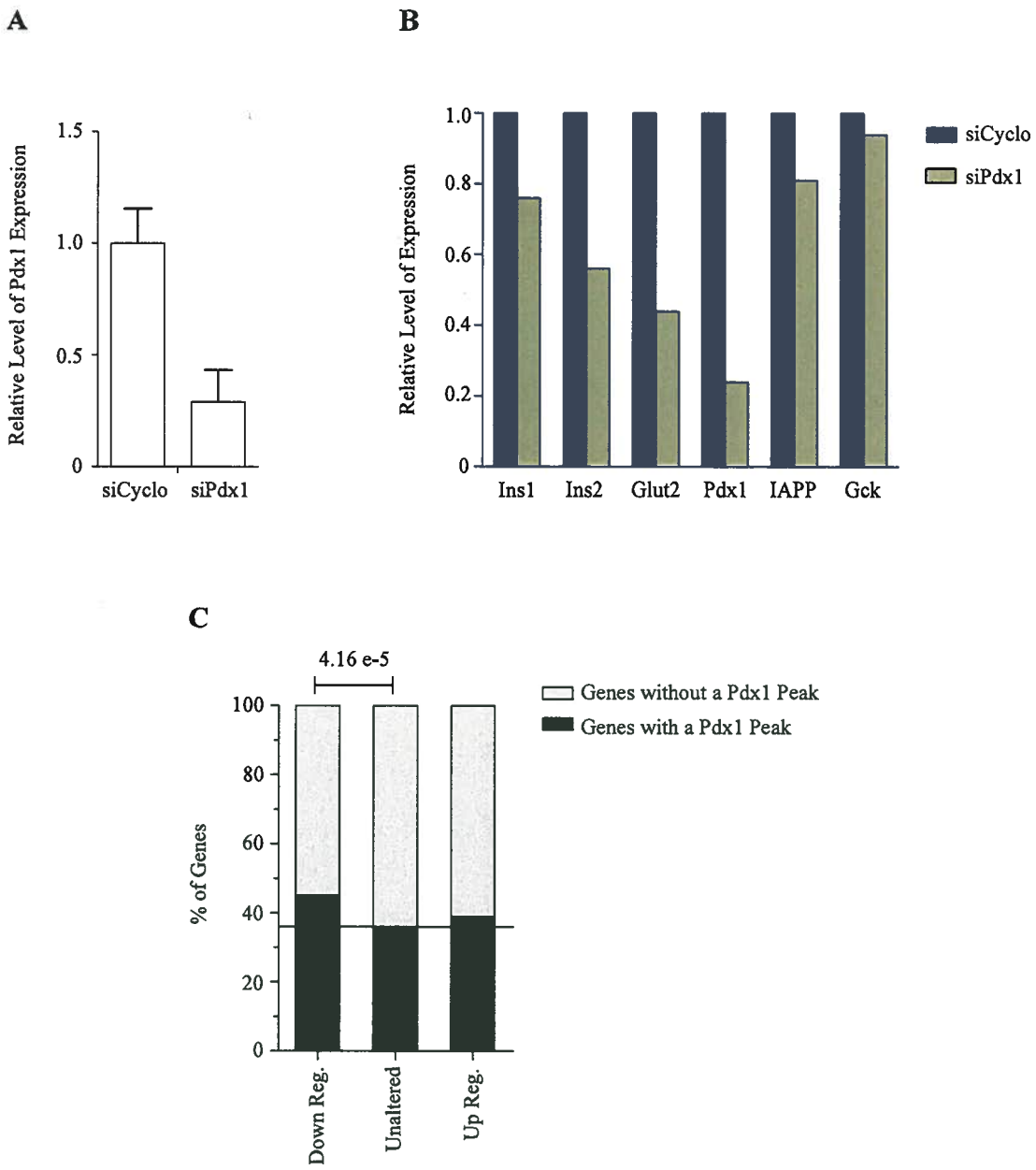


Figure 12 - Down Regulated siPdx1 Tag-Seq Genes Are Significantly Represented in ChIP-Seq Data and Include Expected Genes. (A) RT-PCR from RNA collected from FACSorted siCyclo and siPdx1 islets confirms that Pdx1 expression is decreased. (B) The relative expression levels of known Pdx1 positively regulated genes reveal the expected decreased expression in the siPdx1 Tag-Seq library. (C) Unaltered, down, and up regulated genes were compared against ChIP-Seq genes. Using Fisher's exact test, down regulated genes in the siPdx1 Tag-Seq library are significantly more likely to possess a Pdx1 ChIP-Seq peak.

3.2.4 KEGG Pathways of Pdx1 Genes

All genes that were associated with a Pdx1 ChIP-Seq peak were analyzed against all Refseq genes using WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt/index.php) to determine significantly represented KEGG pathways. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge base for analyzing gene functions in terms of gene networks and molecules. Significant KEGG pathways are shown in Table 2, several of which are expected and critical in the β-cell including: Insulin Signalling Pathway, MODY, and Type II Diabetes Mellitus.

To elucidate the gene pathways on which Pdx1 had the most impact, KEGG pathways were determined for the down regulated genes of the siPdx1 Tag-Seq library that possessed a ChIP-Seq peak. The obtained pathways are shown in Table 3 and again include expected pathways such as MODY and Type II Diabetes Mellitus.

Based on the validative studies performed on the Pdx1 ChIP-Seq data, it was clear that the library was a quality representation of Pdx1 binding in islets.

Table 2 - Significantly Over-Represented KEGG Pathways of all Genes with a Pdx1 ChIP-Seq Peak

KEGG Pathway	Observed	Expected	R Value	P Value
Regulation of actin cytoskeleton	65	53.5359	1.2141	0.0395
MAPK Signalling Pathway	96	72.9778	1.3155	0.0011
Focal adhesion	61	50.4364	1.2094	0.0483
Wnt signalling pathway	54	39.7292	1.3592	0.00577
Insulin Signalling Pathway	48	37.1933	1.2906	0.0246
MODY	14	6.1989	2.2585	0.000558
VEGF Signalling Pathway	32	19.442	1.6459	0.000959
Apoptosis	32	21.6961	1.4749	0.00786
Colorectal Cancer	32	22.8232	1.4021	0.018
Pancreatic Cancer	30	20.2873	1.4788	0.00945
Type II Diabetes Mellitus	20	12.3978	1.6132	0.0107
Cell Cycle	38	29.8674	1.2723	0.0515

Table 3 - Significantly Over-Represented KEGG Pathways of siPdx1 Tag-Seq Down Regulated Genes that have a ChIP-Seq Peak

KEGG Pathway	Observed	Expected	R Value	P Value
Regulation of actin cytoskeleton	13	1.8372	7.076	7.47 e -7
Cytokine-cytokine receptor interaction	5	0.5846	8.5529	0.00131
MODY	4	0.501	7.984	0.00497
Focal adhesion	7	1.9207	3.6445	0.00641
Colorectal cancer	5	1.0021	4.9895	0.00744
Ca signalling pathway	5	1.1691	4.2768	0.0123
mTOR signalling pathway	3	0.334	8.982	0.0125
Dorso-ventral axis formation	3	0.4175	7.1856	0.0189
Pancreatic cancer	4	0.9186	4.3545	0.0238
DRPLA	2	0.167	11.976	0.0320
Type II Diabetes Mellitus	3	0.5846	5.1317	0.0361
Glioma	3	0.6681	4.4903	0.0469

Observed - Number of genes found in dataset of interest
Expected - Number of genes expected to be found based on background
R Value - Ratio of observed to expected
P Value - Probability of result

3.3 Pdx1 ChIP-Seq Library Analysis

3.3.1 Pdx1 and Pbx1 Binding Motif Identification

Because transcription factors frequently bind DNA in complexes, I wanted to examine the sequences contained under Pdx1 ChIP-Seq peaks for nucleotide sequence binding motifs to deduce possible co-regulators acting with Pdx1. Pdx1 ChIP-Seq peaks were scanned for binding motifs similar to the classic Transfac Pdx1 DNA binding motif as well as the Transfac Pbx1 DNA binding motif to see if peaks were enriched for these nucleotide sequences. Pbx1 was selected due to its well-documented embryonic co-regulatory role with Pdx1.

To perform this analysis, Gordon Robertson and Leping Li used the GADEM motif discovery tool (outlined in section 2.14) on Pdx1 ChIP-Seq sequences based on 11bp Pdx1 and 15bp Pbx1 sequences from Transfac. This returned a Pdx1-like motif that occurred in roughly 45% of peaks (Figure 13a), and a Pbx1-like motif that occurred in roughly 43% of peaks (Figure 13b). Taken together, at least one of the identified motifs was present in 63.8% of peaks. Interestingly, the Pbx1-like motif appeared to be a heterodimer comprising core Pbx1 and Pdx1 binding sequences. Because Pdx1 and Pbx1 Transfac motifs contained similar core base pair sequences, some sequences were identified by both independent motif discovery runs. To determine which were duplicates, a histogram of the distance between site types was created (Figure 13c). Sites separated by a distance of –6bp (Pbx1-like relative to Pdx1-like) were identified by both types of motif discovery runs. Consequently, Pdx1-like sites that had Pbx1-like sites located at a distance –6bp were removed from the Pbx1-like list.

A

PDX1-like MONOMER

B

PBX:PDX HETERODIMER

C

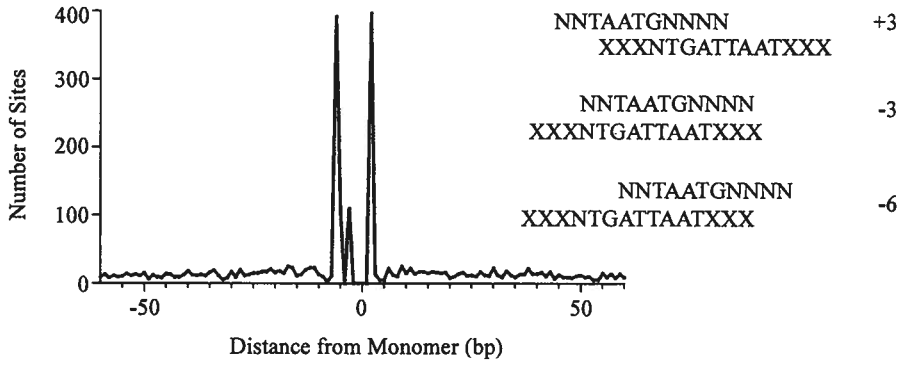


Figure 13 - Seeded Motif Discovery of Pdx1 ChIP-Seq Data Returns Pdx1-like and Pbx1-like Motifs. (A) Using a Pdx1 seed, a Pdx1-like motif (monomer) is found in 45% of peaks. (B) Using a Pbx1 seed, a Pbx1-like motif (heterodimer) is found in 43% of peaks. Taken together, 64% of peaks contain at least one of the two site types. (C) Relative distance of heterodimers from monomers show primary distributions of +3, -3, and -6 base pairs. Alignments of motifs reveals that at -6bp, the same sequences were called sites by both seeded motif runs.

3.3.2 Validation and Analysis of Pbx1 Containing Peaks

In order to explore the relationship of Pdx1 and Pbx1, experimental confirmation of Pbx1 binding at sites identified via motif discovery was needed. The terms monomer and heterodimer were used to describe site type. Monomer sites were those identified from the Pdx1 based motif discovery run, and were indicative of Pdx1 binding alone to DNA. Heterodimer sites were those identified by Pbx1 based motif discovery, and were indicative of a Pdx1 and Pbx1 complex binding to DNA. To confirm Pbx1 binding at peaks containing a heterodimer site, ChIP-qPCR was performed in MIN6 cells with an antibody directed against Pbx1 (SantaCruz), and heterodimer as well as monomer targets tested for enrichment. The results shown in Figure 14a reveal that Pbx1 is enriched at heterodimer containing peaks while at monomer containing peaks it is not.

To determine if Pbx1 is necessary for Pdx1 binding at heterodimer sites, I used siRNA to knockdown Pbx1 expression in MIN6 cells after which Pdx1 binding was tested by ChIP. As a control, siCyclo was transfected alongside siPbx1 and Pdx1 ChIPs also performed. The results of this experiment, shown in Figure 14b, revealed that while Pdx1 binding at target sites was reduced through knockdown of Pbx1, the degree of binding reduction was not greater in peaks containing heterodimer sites than peaks containing monomer sites. Taken with the Pbx1 ChIP result, this indicates that while Pbx1 is binding at heterodimer sites, its impact on Pdx1 binding is no greater at heterodimers as compared to monomers.

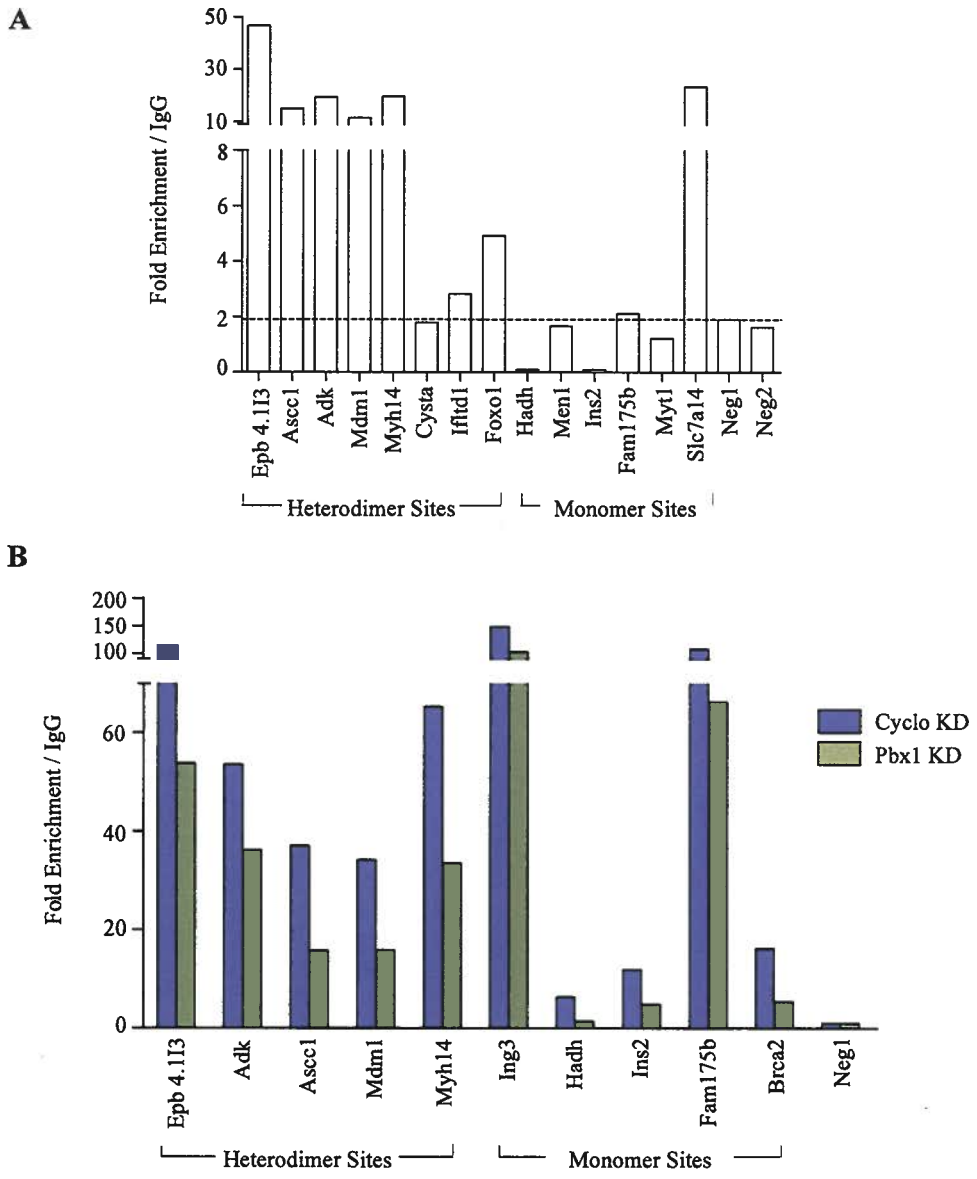


Figure 14 - Pbx1 has no greater affect on Pdx1 binding at heterodimer sites compared to monomer sites. (A) ChIP-qPCR was performed in MIN6 cells with Pbx1 antibody. Primers for heterodimer sites and monomer sites were tested and enrichments confirm Pbx1 binding at heterodimers but absence at monomers with the exception of the Slc7a14 site. (B) ChIP-qPCR was performed using Pdx1 antibody in MIN6 cells subjected to Pbx1 knockdown or a control knockdown to determine if Pbx1 was necessary for Pdx1 binding at heterodimer sites.

We next investigated whether the presence of a monomer or heterodimer affected gene expression and/or gene specificity of the nearest gene. To do so, genes were placed into the following categories:

Table 4 – Monomer and Heterodimer Gene Categories

Monomer	Gene has peak(s) with Pdx1-like motif only
Dimer	Gene has peak(s) with Pbx1-like motif only
Mono + Di	Gene has peaks with both motifs, but motifs never occur in same peak
Mono : Di	Gene has at least one peak where motifs co-occur

For expression analysis, an existing Tag-Seq library of gene expression constructed using wild-type islets was used as the basis of gene expression. A gene was defined as expressed if its count in the Tag-Seq library was greater than five. The presence of monomer and heterodimer sites was not found to have a significant impact on whether a gene was expressed or not (Figure 15a). Moreover, the absolute expression level of those genes that were expressed was also not affected (Figure 15b). An examination of the specificity of the genes in each category was also performed using SAGE libraries generated through the Mouse Atlas of Gene Expression project (www.mouseatlas.org). Genes were assigned a score quantifying their specificity to islets based on the following formula[73]:

$$\text{Specificity} = \frac{Mr \times 3^{\text{Log}_{10}(Ac)}}{3^{\text{Log}_{10}(Lc)}}$$

Where Mr is the ratio of the counts of the tag in the library of interest (islet) over the mean of the counts of the tag in all other libraries, Ac is the absolute count of the tag in the library of interest, and Lc is the number of libraries the tag is found in. The relative specificity scores for the genes represented in each category is depicted in Figure 15c. Interestingly, when genes possess both a monomer as well as a dimer they are far more likely to be islet specific, likely because a far greater proportion of high specificity genes are found in these categories (Figure 15d), where high specificity is defined as a score greater than 2, moderate a score between 0.2 and 2, and low a score less than 0.2.

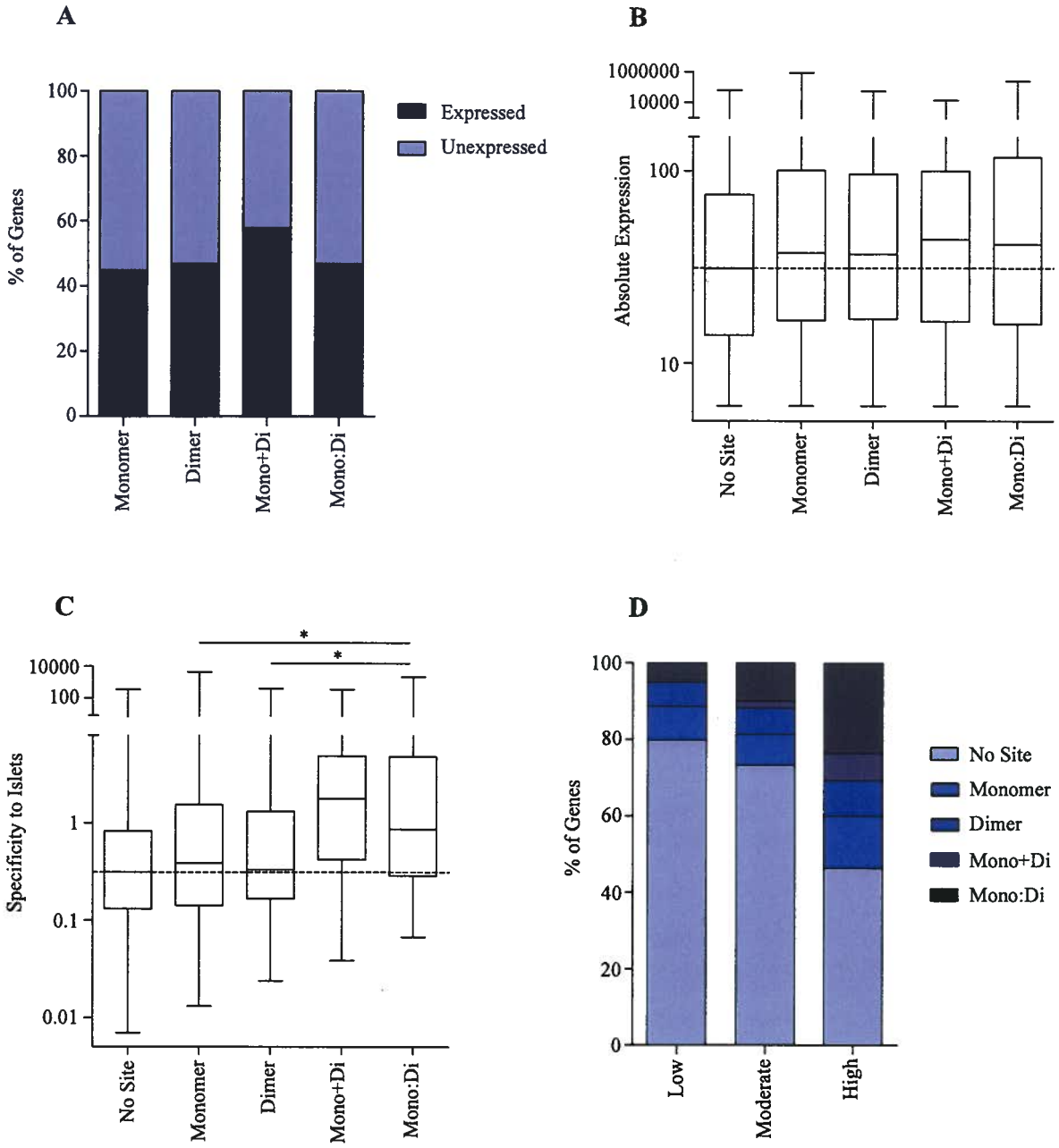


Figure 15 - Analysis of Heterodimer and Monomer Containing Peaks. Expression and specificity analysis was performed comparing genes with varying types of monomer and heterodimer site distributions. Site type was not seen to have an effect on whether or not a gene was expressed (A), or on its relative level of expression (B). However, genes that contained both a monomer and a heterodimer site were significantly more likely to be islet specific (C) due to a greater percentage of high specificity genes belonging to this group (D).

CHAPTER 4. DISCUSSION

The aim of this study was to identify the genome-wide binding of Pdx1 in pancreatic islets. This entire body of work was dependent on the first step of identifying a ChIP grade Pdx1 antibody. The reason for this was twofold; first, the ChIP-Seq procedure requires isolation of sufficient amounts of DNA for sequencing to be successful, and second, all isolated DNA is used for sequencing. An antibody that is used for ChIP-Seq purposes must therefore bind its target protein with high affinity to provide sufficient DNA, and must also be highly specific for only the target protein of interest so that sequenced DNA is a reliable representation of regions bound by the transcription factor. These criteria were fulfilled by the Pdx1 antibody purchased from Chemicon (Figure 6), and was largely expected given that this antibody had also been used for ChIP-ChIP experiments[56].

Once constructed, quality checks of the Pdx1 ChIP-Seq library were conducted using several approaches. The most basic tactic, but also the most reliable, was to scan the generated data for peaks at binding sites that had been well documented in the literature. The UCSC visualizations in Figure 8, as well as the table shown in Figure 10, exemplify the reliability of the library based on this approach. Of the 12 known sites that were surveyed, only 3 did not show Pdx1 peaks in our data. However, 2 of these sites were for somatostatin, a gene expressed in the delta cells of the islet which make up only 2-10% of islet mass. Therefore, the DNA contribution from these cell types into the ChIPs would have been extremely small, and would not have provided enough input for the binding site to be enriched. The other site not identified was for the glucose transporter Glut2. However, the UCSC gene depiction of Glut2 shown in Figure 8

clearly reveals that several Pdx1 peaks are actually present for the gene. This suggests that Pdx1 binding may actually be occurring elsewhere in the Glut2 gene region and perhaps not at the previously reported site[52]. Hence, all well-known Pdx1 binding sites were either present in the ChIP-Seq data, or if not, readily explainable. In addition to those binding sites shown in Figure 8, peak scanning in UCSC also revealed Pdx1 peaks at several distal enhancer regions of genes suspected, but not yet confirmed, to be regulated directly by Pdx1[1]. Most notably, a novel binding site at the Ins1 gene was identified, as well as distal sites upstream of the Nkx2.2 and Nkx6.1 promoters. Binding sites at the Isl1 and Pax6 promoters were also observed, which is significant as factors regulating these β-cell critical transcription factors remain largely unknown (appendix figure A1). These observations provided confidence that while the mapping efficiency of the Pdx1 library (24%) was lower than has been previously reported for ChIP-Seq libraries (FoxA2 liver – 33%[72]), the data is a valid representation of Pdx1 binding. Furthermore, the ChIP-qPCR (Figure 11) that was performed to validate ChIP-Seq peaks provides additional strong evidence that the binding sites are real. Nevertheless, improvements in the quality of our ChIP-Seq data would likely be possible if a greater amount of starting DNA was contributed to each ChIP replicate. This is because DNA input amount is a major limiting factor to ChIP success. With islets contributing so few cell numbers in comparison to other studied tissues such as liver, cell numbers are the most significant limitation facing islet ChIP studies.

As an additional measure of quality, our ChIP-Seq data was compared directly against ChIP-ChIP data previously published for Pdx1[56]. While the 35% overlap that we observe (Figure 10) is lower than previously reported comparisons of ChIP-Seq and

ChIP-Chip[72], there are several reasons why it is not unexpected. The previous ChIP-Chip study had been performed in a NIT-1 cell line whereas our data was generated using primary tissue. In addition, only putative promoter and enhancer elements were included in the ChIP-Chip study. This is a major caveat as it ignores major portions of the genome, which as evidenced in the analysis of peak distribution in gene regions (Figure 9), account for a significant portion of Pdx1 binding and suggest enhancer elements may in fact be functioning further upstream of transcriptional start sites than previously thought. Moreover, a glaring concern exists with the ChIP-Chip data in that it identifies none of the well-known Pdx1 binding sites. This calls into question the reliability of this previous work given that extremely significant targets such as Ins1, Ins2, IAPP, and Gck fail to be recognized. From all of this, it is clear that we have constructed a more comprehensive, accurate, and biologically relevant documentation of genome-wide Pdx1 binding than previously shown.

The generated Tag-Seq libraries of gene expression (control and Pdx1 knockdown) were meant to serve as both a validative tool for the ChIP library as well as to begin to provide insight into those genes that are *most* highly influenced by altered Pdx1 expression. In the analysis of these libraries, multiple tag types had to be combined that mapped to the same gene. This was done because although the libraries are cDNA based, multiple tag types can result from alternative transcripts and errors in enzyme cutting during library construction. While the altered expression levels of known Pdx1 targets such as insulin and Glut2 demonstrate the changes one would expect from diminished Pdx1 expression (Figure 12), genes such as IAPP and Gck (also known Pdx1 targets) show more moderate changes in expression levels. This coincides with what has

been shown in Pdx1 conditional knockout studies where the most notable changes include severe impairment of insulin production as well as diminished Glut2 expression[55]. Therefore, significantly altered genes in the siPdx1 library are likely highly responsive to changes in Pdx1 expression. Based on this, one would expect a substantial portion of these genes to be represented in the Pdx1 ChIP-Seq library. Though this is observed for the down regulated gene set compared to unaltered genes, the same cannot be said for those genes up regulated. Moreover, though statistically significant when compared to unaltered genes, the down regulated genes still only show a 45% correlation with ChIP-Seq. This is lower than desired, and suggests several changes in islet cell gene expression could be based on the stresses of culture and FACSorting rather than knockdown of Pdx1, or that the changes stem from indirect effects of Pdx1 knockdown. Additionally, while substantial knockdown of Pdx1 mRNA was observed at roughly 70% reduction, the possibility exists that even low levels of protein are sufficient to maintain regulation at several of its targets, or that at 48 hours post siRNA transfection, original protein levels had not had sufficient time to drop. Consequently a true knockout library of Pdx1 may be necessary to holistically address its effects. Nevertheless, despite these caveats, the statistical significance between down regulated versus unaltered genes and their correlation to Pdx1 ChIP-Seq does allow for several insights. The most obvious of these is that Pdx1 is clearly functioning most often as an activator. Were it having significant repressive effects, one would have expected to see a greater portion of up regulated genes with ChIP-Seq peaks; this is not the case. Moreover, to determine in which pathways Pdx1 was having the largest activational role, KEGG pathway analysis of all ChIP-Seq genes, as well as those that were down regulated and in possession of a

Pdx1 ChIP-Seq peak, was performed. Tables 2 and 3 show KEGG pathways that are significantly over represented correspond to signalling pathways where Pdx1 is expected to have major influence, such as MODY and Type II diabetes. These are present in both KEGG analyses as one would anticipate given that they are the most impactful subjects of Pdx1. Of most interest in addition to these expected pathways, we observe that several pathways related to cell cycle are also represented, such as those pertaining to pancreatic or colorectal cancer, apoptosis, glioma, and cell cycle itself. While not all of these are present in both KEGG analyses, their varied presence between both establishes a highly probable involvement of Pdx1 in aspects of β-cell cycle.

A relationship between Pdx1 and Pbx1 is known to exist in both the embryo as well as the adult[33, 34]. This association has a profound role in expansion and organization of the developing pancreas, while in the adult a more precise control of insulin regulation has been reported through the Pdx1/Pbx1 affiliation. Given the role of these heterodimers in proliferative function during development, coupled with the observed over representation of KEGG pathways related to cell cycle from Pdx1 ChIP-Seq genes, the Pdx1/Pbx1 adult relationship may similarly drive cell cycle related processes. This suspicion arose as a result of the unexpectedly high percentage of Pdx1 ChIP-Seq peaks that upon motif discovery analysis showed the presence of a Pdx1/Pbx1 heterodimer binding site (Figure 13). To begin to address this, we first examined Pbx1 binding at these sites and the dependency of Pdx1 on such binding. Since these heterodimers have been reported to bind DNA with up to ten times the affinity of Pdx1 alone[33], it was hypothesized that reduction of Pbx1 would significantly alter Pdx1 binding at heterodimer type sites while singly bound Pdx1 sites would be relatively unaffected.

Despite the confirmation of Pbx1 binding at heterodimer sites and not at monomer sites (Figure 14a), the dependency of Pdx1 on Pbx1 was not found to be any greater at heterodimer sites as compared to monomers (Figure 14b). This seems to contradict the notion that Pdx1/Pbx1 binds DNA with 10x affinity at target sites. In addition, expression and specificity analysis of genes with various distributions of site type did not reveal any changes in expressivity, while a positive correlation between specificity and possession of multiple sites was observed (Figure 15). The most likely explanation for these results is that since heterodimer sites still contain the core TAAT motif required for Pdx1-DNA binding, Pdx1 is still capable of binding these regions without dimerization with Pbx1. However, this does not negate the possibility that Pdx1 binding at these sites may not be able to fully drive gene expression without Pbx1. The increased specificity of genes with multiple sites finds explanation in that the greater number of transcription factor binding events occurring for a given gene within a given tissue, are indicative of that gene being highly specific for that tissue. For example, insulin, a highly specific β-cell gene, would have the most transcription factor binding events in β-cells than in any other tissue. Hence, specificity and the number of transcription factors binding are directly proportional.

The work surrounding Pdx1/Pbx1 performed in this study focused on confirming the presence and requirement for Pbx1 at heterodimer sites. While the former was shown, this work revealed that Pbx1 is not essential for Pdx1 binding. As a next step, expression analysis could be performed on heterodimer versus monomer regulated genes following Pbx1 knockdown (siRNA KD-qPCR) to determine if without Pbx1, Pdx1

cannot drive expression at heterodimer sites. If so, Pbx1, though not essential for Pdx1 binding, would be essential for Pdx1 activation of genes at heterodimer sites.

Coming full circle, this bears significance in that the majority of the cell cycle related genes identified from our KEGG analyses possess Pdx1 ChIP-Seq peaks containing heterodimer as opposed to monomer sites. Genes with Pdx1 peaks known to have roles in β-cell proliferation include: Ccnd1, Ccnd2, p15, p21, E2F1, Men1, Rb, p53, Insulin, FoxO, NFAT, Stat5, and Pdx1[74]. Of these 13 genes, 8 have heterodimer sites associated with them. Therefore, an analysis of the necessity for Pbx1 at heterodimer sites for Pdx1 mediated expression would be a logical direction in which to take this work in order to extract more insight into how Pdx1 may be functioning at cell cycle related genes.

As a genome-wide dataset, there also exists much value in coupling this Pdx1 ChIP-Seq information with future genome-wide studies for both other transcription factors and markers of DNA methylation. Since a transcription factor complex involving NeuroD1, Mafa, and Pdx1 is already known to form at the insulin gene, ChIP-Seq studies of NeuroD1 and Mafa would prove extremely beneficial to compile with this Pdx1 dataset to determine genome-wide sites of transcriptional complex formation in islets. Additionally, considering this binding information in the context of DNA methylation status would enable us to determine functionality of sites. Furthermore, construction of an embryonic Pdx1 library at E8.5 (the onset of Pdx1 expression) would allow for comparison of Pdx1 gene regulation developmentally on a genome-wide scale.

CONCLUSION

The purpose of this study was to characterize Pdx1 binding in pancreatic islets on a genome-wide scale. This was clearly accomplished by utilizing the ChIP-Seq strategy to produce the most extensive dataset of Pdx1 binding generated to date. Novel binding sites at genes of high interest include Ins1, Nkx2.2, Nkx6.1 and Isl1. Additionally, a highly occurring relationship with Pbx1 is identified and the binding of Pbx1 confirmed at suspected sites. Given the reported role of Pdx/Pbx in cell expansion and organization in islet precurser populations, as well as the fact that Pdx1 ChIP-Seq and Tag-Seq genes show significant representation in cell cycle pathways, future investigation into the Pdx/Pbx adult role in proliferation is warranted.

The expansion of islet populations for use in transplant for diabetic patients will find substantial improvement as the molecular physiology of the β-cell continues to be exposed. This work begins to address this need, and coupled with future studies of a similar nature, will prove significant in unlocking the workings of β-cell function through the identification of genome-wide islet transcriptional complexes, transcriptional networks, and changes in transcription factor action in the embryo versus the adult.

REFERENCES

1) Jensen, Jan. **2004** Gene Regulatory Factors in Pancreatic Development. *Developmental Dynamics.* 229: 176-200.

2) Avisse, C, Flament, JB, and Delattre, JF. **2000** Ampulla of Vater. Anatomic, embryologic, and surgical aspects. *Surg Clin North Am.* 80: 201-212.

3) Fukuda, Akihisa et al. **2006** Loss of the Major Duodenal Papilla results in brown pigment biliary stone formation in Pdx1 Null mice. *Gastroenterology.* 130: 855-867.

4) Slack, JMW. **1995** Developmental biology of the pancreas. *Development.* 121: 1569-1580.

5) Suckale, Jakob and Solimena, Michele. **2008** Pancreas islets in metabolic signaling – focus on the B-cell. *Nature Precedings.* 2: 12 pgs.

6) Rahier, J, Goebbels, RM, and Henquin, JC. **1983** Cellular Composition of the Human Diabetic Pancreas. *Diabetologia.* (5)24: 366-371.

7) Adeghate, E and Donath, T. **1991** Morphometric and immunohistochemical study on the endocrine cells of pancreatic tissue transplants. *Experimental and Clinical Endocrinology.* 98: 193-199.

8) Stefan, Y et al. **1982** Quantitation of endocrine cell content in the pancreas of nondiabetic and diabetic humans. *Diabetes.* 31: 694-700.

9) Elayat, AA, el-Naggar, MM, and Tahir, M. **1995** An immunocytochemical and morphometric study of the rat pancreatic islets. *Journal of Anatomy.* 186: 629-637

10) Adeghate, E. **1999** Distribution calcitonin-gene-related peptide, neuropeptide-Y, vasoactive intestinal polypeptide, cholecystokinin-8, substance P and islet peptides in the pancreas of normal and diabetic rat. *Neuropeptides.* 33: 227-235.

11) Wierup, N et al. **2002** The ghrelin cell: a novel developmentally regulated islet cell in the human pancreas. *Regul Pept.* 107: 63-69.

12) Henderson, JR, and Moss, MC. **1985** A Morphometric study of the endocrine and exocrine capillaries of the pancreas. *Experimental Physiology.* 70: 347-356.

13) Levick, JR, and Smaje, LH. **1987** An anlysis of the permeability of a fenestra. *Microvascular research.* (2)33: 233-256.

14) Bendayan, M. **1993** Pathway of Insulin in pancreatic tissue on its release by the B-cell. *American Journal of Physiology.* 264: G187-G194.

15) Curry, DL, Bennett, LL, and Grodsky, GM. **1968** Dynamics of insulin secretion by the perfused rat pancreas. *Endocrinology.* (3)83: 572-584

16) Thorens, B, et al. **1988** Cloning and functional expression in bacteria of a novel glucose transporter present in liver, intestine, kidney, and beta-pancreatic islet cells. *Cell.* (2)55: 281-290.

17) Iynedjian, P.B. **1993** Mammalian glucokinase and its gene. *Journal of Biochemistry.* 293: 1-13.

18) Ashcroft, F.M. and Gribble, F.M. **2000** New windows on the mechanism of action of K_{ATP} channel openers. *Trends in Pharmacological Sciences.* (21)11: 439-445.

19) Yang, S.N. and Berggren, P.O. **2006** The Role of Voltage Gated Calcium Channels in Pancreatic β-cell Physiology and Pathophysiology. *Endocrine Reviews.* (6)27: 621-676.

20) Rutter, G.A. et al. **2006** Insulin secretion in health and disease: genomics, proteomics and single vesicle dynamics. *Biochemical Society Transactions.* 34: 247-250

21) Pessin, J.E. and Saltiel, A.R. **2000** Signalling pathways in insulin action: molecular targets of insulin resistance. *Journal of Clinical Investigation.* (2)106: 165-169.

22) Statistics courtesy World Health Organization.

23) Rother, K.I. **2007** Diabetes Treatment – Bridging the divide. *New England Journal of Medicine.* (15)356: 1517-1526.

24) Statistics courtesy BC Transplant

25) Collaborative Islet Transplant Registry 2006 Annual Report

26) Nielson, J.H. et al. **1999** Beta Cell Proliferation and Growth Factors. *Journal of Molecular Medicine.* 77: 62-66.

27) Hayek, A. and Beattie, G.M. **2002** Alternatives to unmodified human islets for transplantation. *Curr. Diab. Rep.* 2: 371-376.

28) Moriscot, C. et al. **2005** Human bone marrow mesenchymal stem cells can express insulin and key transcription factors of the endocrine pancreas developmental pathway upon genetic and/or microenvironmental manipulation in vitro. *Stem Cells.* 23: 594-604

29) Rother, K.I. and Harlan, D.M. **2004** Challenges facing islet transplantation for the treatment of type I diabetes mellitus. *Journal of Clinical Investigation.* (7)114: 877-883.

30) Hobert, Oliver. **2008** Gene Regulation by Transcription Factors and MicroRNAs. *Science.* 319: 1785-1786.

31) Latchman, D.S. **1997** Transcription Factors: An Overview. *Internationl Journal of Biochemistry and Cell Biology.* (12)29: 1305-1312.

32) Wu, K.L. et al. **1997** Hepatocyte nuclear factor 3 beta is involved in pancreatic beta cell specific transcription of the pdx1 gene. *Molecular Cellular Biology.* 17: 6002-6013.

33) Dutta, S. et al. **2001** Pdx:Pbx Complexes are required for normal proliferation of pancreatic cells during development. *Proc Natl Acad Sci USA.* 98: 1065-1070

34) Kim, S.K. et al. **2002** Pbx1 Inactivation disrupts pancreas development and in Ipf-1 deficient mice promotes diabetes mellitus. *National Genetics.* 30: 430-435

35) Gerrish, K et al. **2001** The role of hepatic nuclear factor 1 alpha and pdx-1 in transcriptional regulation of the pdx-1 gene. *Journal of Biological Chemistry.* 276: 47775-47784.

36) Harries, L.W. **2006** Alternate mRNA processing of the hepatocyte nuclear factor genes and its role in monogenic diabetes. *Expert Review of Endocrinology and Metabolism.* 1: 715-726.

37) Watada, H. et al. **2000** Transcriptional and Translational Regulation of beta cell differentiation factor Nkx6.1. *Journal of Biological Chemistry.* 275: 34224-34230

38) Sussel, L. et al. **1998** Mice lacking the homeodomain transcription factor Nkx2.2 have diabetes due to arrested differentiation of pancreatic beta cells. *Development.* 125: 2213-2221.

39) Sander, M. et al. **2000** Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of beta cell formation in the pancreas. *Development.* 127: 5533-5540.

40) Heremans, Y. et al. **2002** Recapitulation of embryonic neuroendocrine differentiation in adult human pancreatic duct cells expressing neurogenin 3. *Journal of Cell Biology.* 159: 303-312.

41) Vetere, A. et al. **2003** Neurogenin3 triggers beta cell differentiation of retinoic acid derived endoderm cells. *Journal of Biochemistry.* 371: 831-841.

42) Babu, D.A. et al. **2008** Pdx1 and Beta2/NeuroD1 participate in a transcriptional complex that mediates short range DNA looping at the insulin gene. *Journal of Biological Chemistry.* (13)283: 8164-8172.

43) Kristinsson, S.Y. et al. **2001** MODY in Iceland is associated with mutations in Hnf1a and a novel mutation in NeuroD1. *Diabetologia.* 44: 2098-2103.

44) Ahlgren, U. et al. **1997** Independent requirement of Isl1 in formation of pancreatic mesenchyme and islet cells. *Nature.* 385: 257-260.

45) Sosa-Pineda, B. et al. **1997** The Pax4 gene is essential for differentiation of insulin producing beta cells in the mammalian pancreas. *Nature.* 386: 399-402.

46) Jonsson, J. et al. **1994** Insulin Promoter factor 1 is required for pancreas development in mice. *Nature.* 371: 606-609

47) Stoffers, D.A. et al. **1997** Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nature Genetics.* 15: 106-110.

48) Oster, A. et al. **1998** Rat endocrine pancreatic development in relation to two homeobox gene products (Pdx-1 and Nkx6.1). *Journal of Immunohistochemistry and Cytochemistry.* 46: 707-715.

49) German, M. et al. **1995** The insulin gene promoter. A Simplified nomenclature. *Diabetes.* (8)44: 1002-1004.

50) Ohneda, K. et al. **2000** Regulation of Insulin Gene Transcription. *Cell and Developmental Biology.* 11: 227-233

51) Shelton, K.D. et al. **1992** Multiple elements in the upstream glucokinase promoter contribute to transcription in insulinoma cells. *Molecular and Cellular Biology.* (10)12: 4578-4589.

52) Waeber, G. et al. **1996** Transcriptional Activation of the Glut2 gene by the IPF1/STF1/IDX1 homeobox factor. *Molecular Endocrinology.* 10: 1327-1334.

53) Carty, M.D. et al. **1997** Identification of cis and trans active factors regulating human islet amyloid polypeptide gene expression in pancreatic beta cells. *Journal of Biological Chemistry.* 272: 11986-11993.

54) Raum et al. **2006** FoxA2, Nkx2.2, and Pdx1 Regulate Islet B-Cell Specific mafa expression through conserved sequences located between base pairs -8118 and -7750 upstream from the transcription start site. *Molecular and Cellular Biology.* 26: 5735-5743.

55) Brissova, M. et al. **2002** Reduction in pancreatic transcription factor Pdx1 impairs glucose stimulated insulin secretion. *Journal of Biological Chemistry.* 277: 11225-11232.

56) Keller, D.M. et al. **2007** Characterization of pancreatic transcription factor Pdx1 Binding sites using promoter microarray and serial analysis of chromatin occupancy. *The Journal of Biological Chemistry.* 282: 32084-32092.

57) Wu, J. et al. **2006** ChIP-chip comes of age for genome wide functional analysis. *Cancer Research.* 66: 6899-6902.

58) Impey, S. et al. **2004** Defining the CREB regulation: a genome-wide analysis of transcription factor regulatory regions. *Cell.* 119: 1041-1054.

59) Chen, J. and Sadowski, I. **2005** Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes by using serial analysis of binding elements. *Prc Natl Acad Sci USA.* 102: 4813-4818.

60) Bhinge, A.A. et al. **2007** Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Research.* 17: 910-916.

61) Roh, T.Y. and Zhao, K. **2008** High resolution genome wide mapping of chromatin modifications by GMAT. *Methods in Molecular Biology.* 387: 95-108

62) Wei, C.L. et al. **2006** A global map of p53 transcription factor binding sites in the human genome. *Cell.* 124: 207-219.

63) Loh, Y.H. et al. **2006** The Oct4 and Nanog Transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics.* 38: 431-440.

64) Lin, C.Y. et al. **2007** Whole genome cartography of estrogen receptor alpha binding sites. *PLoS Genetics.* 3: E87

65) Hoffman, B.G. and Jones, S.J.M. **2009** Genome-wide identification of DNA protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing (ChIP-Seq). *Journal of Endocrinology.* 201: 1

66) Robertson, G. et al. **2007** Genome-wide profiles of STAT1 DNA association using chromatin Immunoprecipitation and massively parallel sequencing. *Nature Methods.* 4: 651-657

67) Mardis, E.R. **2008** The impact of next-generation sequencing technology on genetics. *Trends in Genetics.* 24(3): 133-141.

68) Fejes, A.P. et al. **2008** FindPeaks3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics.* 24(15): 1729-1730.

69) Morrissy, A.S. et al. **2009** Next-generation tag sequencing for cancer gene expression profiling. *Genome Research.* 19(6).

70) Robertson, N. et al. **2007** Discovery Space: an interactive data analysis application. *Genome Biology.* 8(1): R6.

71) Li, L. **2009** GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *Journal of Computational Biology.* 16(2): 317-329.

72) Wederell, E.D. et al. **2008** Global analysis of in vivo FoxA2 binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids research.* 36: 4549-4564.

73) Formula modified from: Hoffman, B.G. et al. **2008** Expression of Groucho/TLE proteins during pancreas development. *BMC Developmental Biology.* 8.

74) Heit, J. et al. **2006** Intrinsic Regulators of Pancreatic Beta Cell Proliferation. *Annual review of Cell and Developmental Biology.* 22: 311-338.

APPENDIX

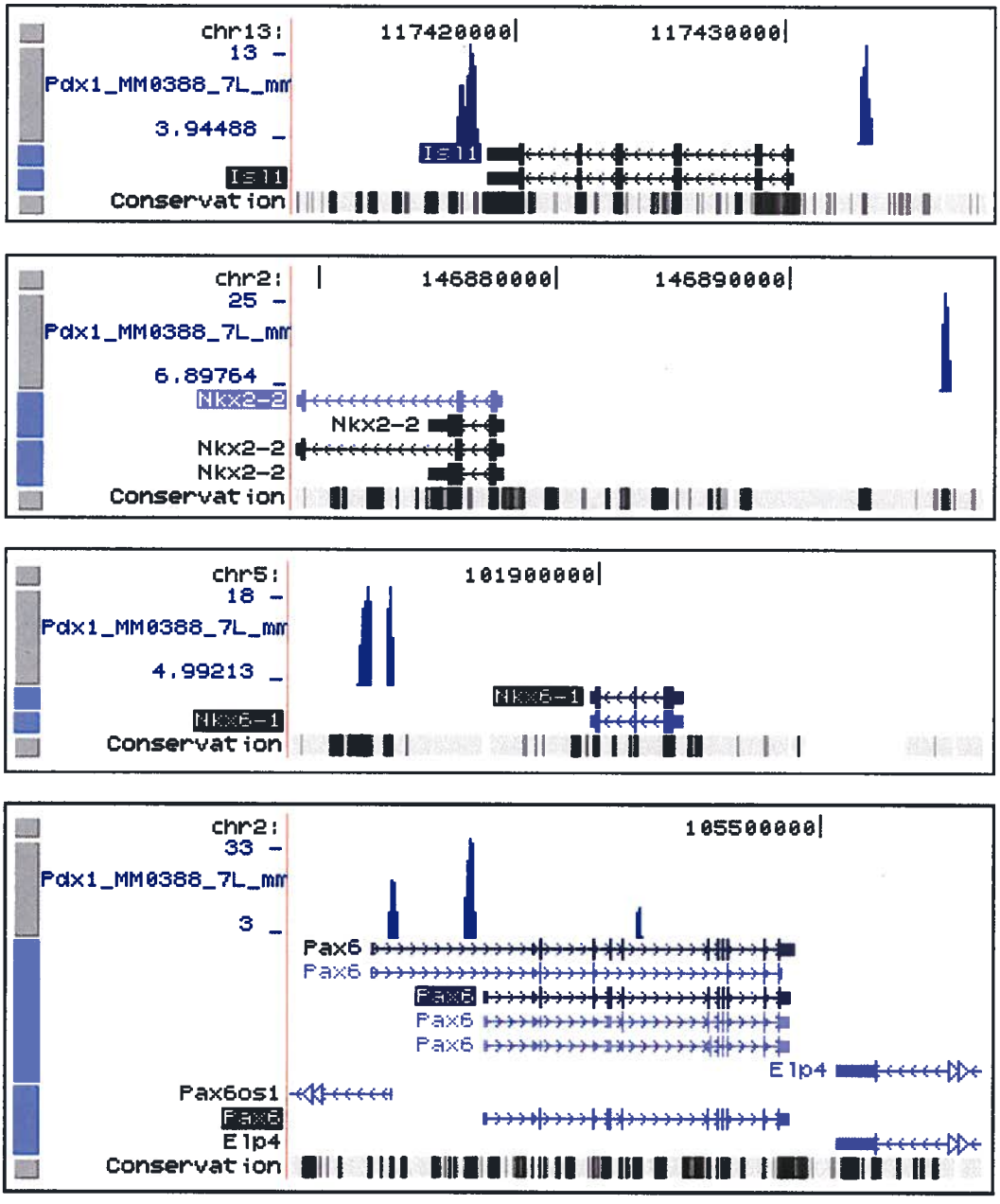


Figure A1: Additional UCSC screenshots of interest of Pdx1 ChIP-Seq binding sites

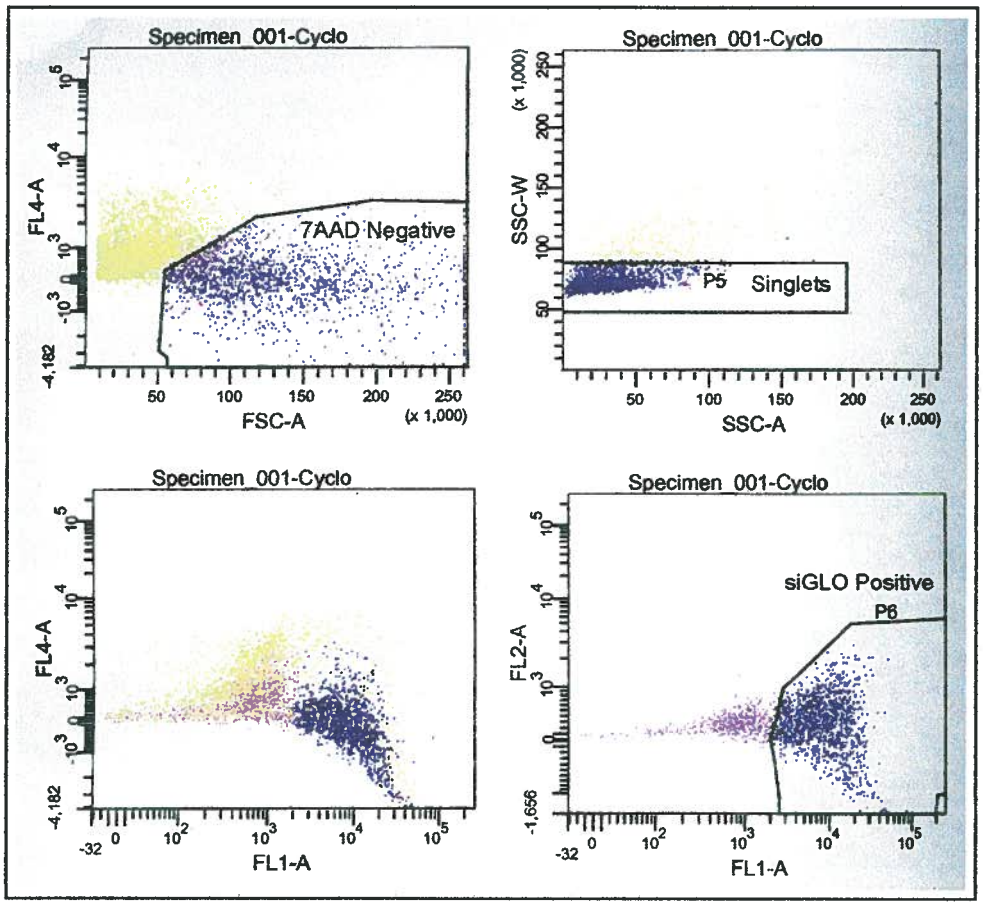


Figure A2: FACSorted siCYCLO islets

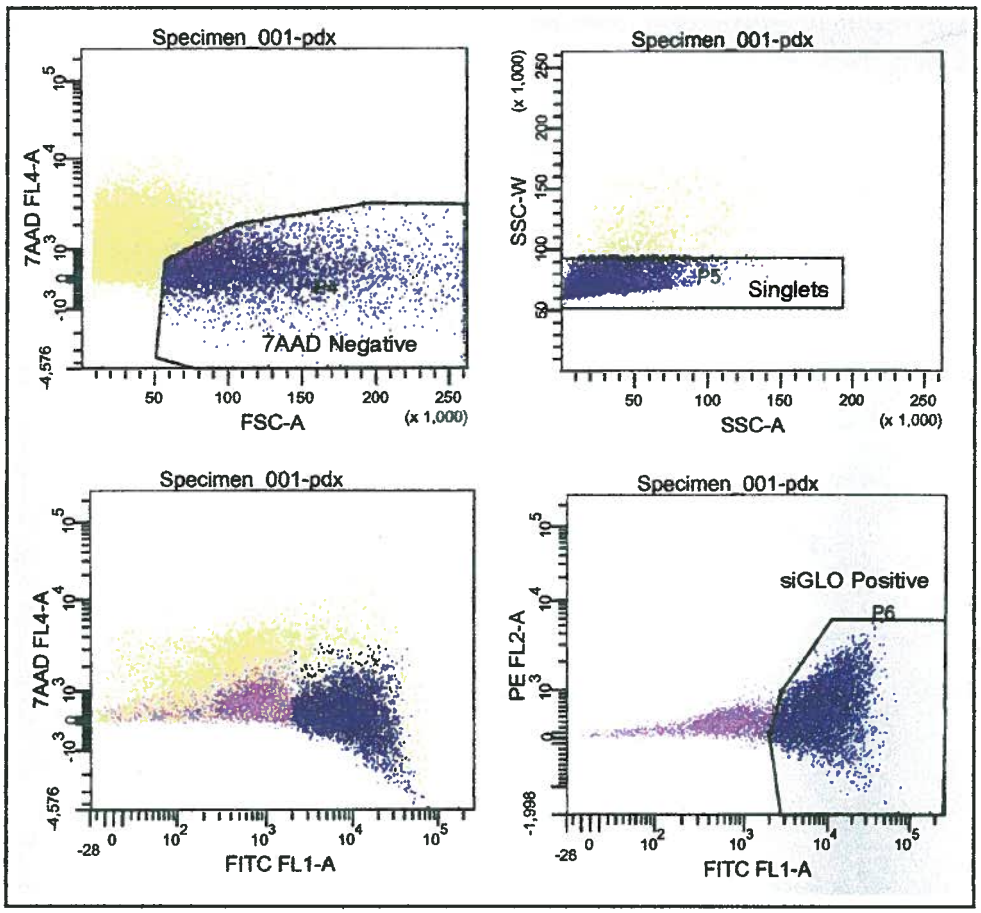


Figure A3: FACSorted siPdx1 islets

THE UNIVERSITY OF BRITISH COLUMBIA

ANIMAL CARE CERTIFICATE

Application Number: A05-1741

Investigator or Course Director: Cheryl D. Helgason

Department: Surgery

Animals:

Mice icrTac:ICR 900
Mice Pdx1-GFP Transgenic 150
Mice NGN3GFP transgenic 150

Start Date: **January 1, 2006** **Approval Date: April 1, 2009**

Funding Sources:

Funding Agency: Genome Canada
Funding Title: Dissecting Gene Regulatory Networks in Mammalian Organogenesis

Unfunded title: N/A

The Animal Care Committee has examined and approved the use of animals for the above experimental project.

This certificate is valid for one year from the above start or approval date (whichever is later) provided there is no change in the experimental procedures. Annual review is required by the CCAC and some granting agencies.

A copy of this certificate must be displayed in your animal facility.

Office of Research Services and Administration
102, 6190 Agronomy Road, Vancouver, BC V6T 1Z3
Phone: 604-827-5111 Fax: 604-822-5093