## Electronic Supplementary Material: Methods

*Mouse maintenance and islet isolations*

Mice were housed in micro-isolator cages and maintained according to the guidelines of the Canadian Council on Animal Care. All protocols were approved by the UBC Animal Care Committee. Hand-picked pancreatic islets were isolated as previously described [1].

*Chromatin immunoprecipitation sequencing*

For islet ChIPs, islets from at least ten adult (8-10 week old) ICR mice of mixed sexes were purified (~2,000 islets, or 1-2x$10^6$ cells) for each ChIP experiment by collagenase digestion, filtration through a 70µm filter, and subsequent hand picking. For liver ChIPs, homogenised perfused livers from a minimum of three adult mice of mixed sexes were used. In each case ChIP was performed essentially as described [2,3] using 3µg of anti-NEUROD1 (Santa Cruz), anti-MAFA (Abcam), anti-H3K9me3 (Millipore), or anti-H3K27me3 (Millipore). DNA from at least triplicate pooled ChIP experiments was purified by 8% PAGE to obtain 100-300 bp fragments and sequenced on an Illumina GA2x or HiSeq 2000 sequencer at the Canada's Michael Smith Genome Sciences Centre. 36 bp sequence reads were aligned to the NCBI37/mm9 genome using Eland or Burrows-Wheeler Aligner (BWA) [4]. Peaks were identified using FindPeaks4 [5] and thresholded at an estimated false discovery rate of 0.01, and regions that overlapped peaks from an islet or liver input DNA negative control [3] sample were removed. Data were deposited under GEO accession GSE30298.

*Islet RNA-sequencing library preparation and sequencing*

RNA-seq was performed essentially as previously described [6]. Total RNA from islets from C57BL/6J mice was converted into a library of template molecules using the TruSeq Sample

Prep Kit (Illumina, San Diego, CA, USA) according to manufacturer's protocol. Libraries were expanded with the Illumina Cluster generation protocol and sequenced at the Canada's Michael Smith Genome Sciences Centre using a HiSeq 2000 sequencer. 75 bp paired-end sequencing was performed with two independent indexed libraries, each from separate islet preparations.

*Identification of putative enhancer loci*

The model-based probabilistic algorithm PING [7] was used to identify H3K4me1-marked nucleosome positions from sonicated H3K4me1 ChIP-seq data in pancreatic islets. These predictions were filtered to remove both low confidence nucleosome calls, and calls generated from low read numbers that were adjacent to highly enriched H3K4me1-marked nucleosomes [7]. From this we identified 251,705 high-confidence nucleosome calls, which together demarcated 251,684 loci flanked on both sides by H3K4me1-enriched nucleosomes. We then eliminated 171,483 loci with a flanking nucleosome spacing <250 bp or >850 bp, as these are unlikely to be functional enhancer elements (ESM Fig. 8). Next, we eliminated 61,149 loci found within intragenic regions, as H3K4me1 is known to be enriched across the gene bodies of transcribed genes independent of its function at enhancer loci [8]. We approximate that this lead to the elimination of roughly two thousand real enhancer loci (based on comparing the numbers of loci identified 5' versus 3' to known TSSs, and assuming roughly equal numbers of enhancers should be predicted on both sides). However, it was our goal to identify a list of loci with as low a rate of false positives as possible, and thus it was essential to eliminate these loci in order to remove the large number of false positives their inclusion would have represented. Finally, we eliminated loci that were within ±2 kb of any Ensembl Transcript NCBIM.37 transcriptional start site (TSS), or were enriched in H3K4me3, in order to ensure we eliminated all possible promoter regions [9,10]. This left 16,835 H3K4me1-marked nucleosome-flanked enhancer loci.

We identified additional candidate enhancer loci using genome-wide transcription factor binding data from mouse islets for PDX1 and FOXA2 [3], and for MAFA and NEUROD1. In total, after thresholding at a false discovery rate of 0.01 (see above) we identified 13,770 PDX1-bound, 6,176 FOXA2-bound, 3,638 MAFA-bound, and 6,568 NERUOD1-bound loci. 24,405 unique loci were bound by at least one of these factors. 9,605 of these had flanking H3K4me1-marked nucleosomes typical of active enhancers [3]. The remaining 14,800 loci had inappropriate nucleosome spacing (ESM Fig. 8), were within nucleosomal DNA, were not associated with sufficient H3K4me1 enrichment to allow accurate nucleosome position predictions, or were not associated H3K4me1 enrichment at all. Our previous observations suggested that such loci are largely inactive in regulating gene expression [3] and thus these loci were removed from consideration. It is worth noting, however, that active transcription factor bound loci that were associated with lower levels of H3K4me1 enrichment, which would have generated low scoring nucleosome calls, may have been eliminated by these criteria. This is because the low scoring flanking nucleosomes would have been thresholded out, leaving the transcription factor binding site incorrectly associated with more distal higher scoring nuclesomes. This may have resulted in these loci being eliminated if these higher scoring nucleosomes were too far apart. Lowering the nucleosome score threshold to prevent this was not found to be practical, as this lead to the identification of 'false' nucleosomes in regions of high H3K4me1 enrichment, which in turn would have lead to the elimination of many transcription factor bound loci associated with high levels of H3K4me1. Further, as noted above, it was our goal to generate a list of enhancer loci with as low a rate of false positives as possible, and we therefore felt it was an acceptable compromise to only keep transcription factor bound loci flanked by high scoring H3K4me1 based nucleosome predictions, as these are the most high confidence loci, and also likely the most active. Eliminating loci within ±2 kb of an Ensembl Transcript NCBIM.37 TSS, or enriched in H3K4me3, left 8,569 PDX1-, MAFA-, NEUROD1-, or FOXA2-bound (PMNF) enhancer regions. 3,181 of these loci were also identified using our H3K4me1-marked nucleosome predictions; while the remaining 5,388 loci were not initially identified primarily because they were intronic or because different nucleosome spacing thresholds were used (ESM Fig. 8). In all cases the boundries of the putative loci were defined by the mid-points of the flanking nucleosomes (ESM Table 2).

*Association of enhancer regions to genes, mapping transcription factors to enhancers, and determination of gene expression levels*

Enhancer regions were associated with genes by identifying the closest annotated Ensembl Transcript NCBIM.37 gene within 200 kb with H3K4me1, H3K4me3, H3K9me3, or H3K27me3 reads present in a 2 kb window around its TSS. Sites were considered to be in promoter regions if they fell within 2 kb of an Ensembl Transcript NCBIM.37 TSS. A transcription factor was considered to occupy a given enhancer if its peak summit was within identified enhancer boundaries. ESC and liver gene expression levels were determined from previously generated data deposited under GEO accession number GSM929718 and SRX17602 respectively. Islet gene expression levels were determined using the islet RNA-seq data described above. Islet specificity of a gene was determined using data from 203 SAGE libraries [11,12] by comparing the expression of the gene in the islet library with the number of other libraries the gene is expressed in, combined with its mean expression in non-islet libraries [3]. Enriched GO or KEGG terms were identified using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [13,14].

### *Detection of orthologous regions in humans*

To identify orthologous regions in humans we used the UCSC Batch Coordinate Conversion utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to obtain hg18 coordinates for our permissive enhancer regions. We then compared these regions with locations of open chromatin in human islets as identified by DNaseI-seq [15] and/or by moderate stringency Formaldehyde-Assisted Isolation of Regulatory Elements-sequencing (FAIRE-seq) peaks [16].

### *DNA sequence motif discovery*

Enriched motifs in the islet specific enhancers (ISEs) versus the non-specific enhancers (NSEs), and vice versa, were determined by first extracting their sequences from the NCBI37 (mm9) UCSC genome browser. Next, both sets of sequences were scanned [17] with each of the PWMs from Uniprobe [18], JASPAR [19] and TRANSFAC v12.1 [20] using a PWM score *p*-value cutoff of 0.0001. Any PWM whose binding sites were found in less than 7.5% of the sequences was removed. Finally, the enrichment *p*-value was computed, using a Fisher exact test (one-sided for enrichment in the islet-specific set). Statistical Analysis of Metagenomic Profiles (STAMP) [21,22] was used to determine the similarities of the enriched motifs and Molecular Evolutionary Genetics Analysis (MEGA) 5 was used to generate phylograms.

### *Identification of novel transcripts in islets*

Contigs whose alignments overlapped no annotated gene in this database were then filtered to remove contigs with less than two exons and with a mean exonic coverage of less than 6 reads per base. The remaining transcript contigs were further filtered to remove any that overlapped annotated exons in the Ensembl NCBIM.37, Refseq or UCSC mm9 transcript databases. The coding potential of the remaining transcripts was determined using PhyloCSF [23] using an eight-way multispecies alignment. Transcripts with a PhyloCSF score below 100 were considered non-coding [24].

## Supplementary References:

1. Li DS, Yuan YH, Tu HJ, Liang QL, Dai LJ (2009) A protocol for islet isolation from mouse pancreas. Nat Protoc 4: 1649-1652.
2. Wederell E, Bilenky M, Cullum R, Thiessen N, Dagpinar M, et al. (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res 36: 4549-4564.
3. Hoffman BG, Robertson G, Zavaglia B, Beach M, Cullum R, et al. (2010) Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. Genome Research 20: 1037-1051.
4. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
5. Fejes A, Robertson G, Bilenky M, Varhol R, Bainbridge M, et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24: 1729-1730.
6. Kim H, Toyofuku Y, Lynn FC, Chak E, Uchida T, et al. (2010) Serotonin regulates pancreatic beta cell mass during pregnancy. Nat Med 16: 804-808.
7. Zhang X, Robertson AG, Woo S, Hoffman BG, Gottardo R (2012) Probabilistic Inference for Nucleosome Positioning with MNase-based or Sonicated Short-read Data. PloS ONE 7: e32095.
8. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.
9. Heintzman N, Stuart R, Hon G, Fu Y, Ching C, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.
10. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.
11. Hoffman BG, Zavaglia B, Witzsche J, Ruiz de Algara T, Beach M, et al. (2008) Identification of transcripts with enriched expression in the developing and adult pancreas. Genome Biol 9: R99.
12. Siddiqui AS, Khattra J, Delaney AD, Zhao Y, Astell C, et al. (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. Proc Natl Acad Sci USA 102: 18485-18490.
13. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
14. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.
15. Stitzel ML, Sethupathy P, Pearson DS, Chines PS, Song L, et al. (2010) Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility Loci. Cell Metab 12: 443-455.

16. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, et al. (2010) A map of open chromatin in human pancreatic islets. Nature Genetics 42: 255-259.
17. Li L (2009) GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. J Comput Biol 16: 317-329.
18. Robasky K, Bulyk ML (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic acids research 39: D124-128.
19. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic acids research 38: D105-110.
20. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108-110.
21. Mahony S, Auron PE, Benos PV (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. PLoS Comput Biol 3: e61.
22. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res 35: W253-258.
23. Lin MF, Jungreis I, Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27: i275-282.
24. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development 25: 1915-1927.