Title: **Extensive NEUROG3 occupancy in the human pancreatic endocrine gene regulatory network**

Valérie Schreiber[1,2,3,4,*], Reuben Mercier[1,2,3,4,], Sara Jiménez[,2,3,4,], Tao Ye[1,2,3,4,], Emmanuel García-Sánchez[1,2,3,4,], Annabelle Klein[1,2,3,4,], Aline Meunier[1,2,3,4,], Sabitri Ghimire[1,2,3,4,], Catherine Birck[1,2,3,4,], Bernard Jost[1,2,3,4,], Kristian Honnens de Lichtenberg[5], Christian Honoré[6], Palle Serup[5] and Gérard Gradwohl[1,2,3,4,*]

[1]Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), [2]Institut National de la Santé et de la Recherche Médicale (INSERM) U1258, [3]Centre National de Recherche Scientifique (CNRS) UMR7104, [4]Université de Strasbourg, 1 rue Laurent Fries, 67404 Illkirch, France. [5]Novo Nordisk Foundation Center for Stem Cell Biology (DanStem), University of Copenhagen, Copenhagen N 2200, Denmark. [6]Department of Stem Cell Biology, Novo Nordisk A/S, DK-2760 Måløv, Denmark.

[*] Corresponding authors: Valérie Schreiber or Gérard Gradwohl, 1 Rue Laurent Fries, 67404 Illkirch, France, Phone: 0033 3 88 65 33 12, Fax: 0033 3 88 65 32 01, email: schreibv@igbmc.fr; gradwohl@igbmc.fr.

Highlights

- NEUROG3 CUT&RUN analysis revealed 1268 target genes in human pancreatic endocrine progenitors (PEPs)

- NEUROG3 binding sites overlap with active chromatin regions in PEPs.

- 1/5 of the genes downregulated in *NEUROG3*$^{-/-}$ hESC-derived PEPs are bound by NEUROG3.

- NEUROG3 targets islet specific TFs and regulators of insulin secretion.

- Several T2DM risk allelles lie within NEUROG3 bound regions.

ABSTRACT

Objective: Mice lacking the bHLH transcription factor (TF) Neurog3 do not form pancreatic islet cells, including insulin secreting beta cells, causing diabetes. In human, homozygous mutations of *NEUROG3* manifest with neonatal or childhood diabetes. Despite this critical role in islet cell development, the precise function and downstream genetic programs regulated directly by NEUROG3 remain elusive. We therefore mapped genome-wide NEUROG3 occupancy in human induced pluripotent stem cell (iPSC)-derived endocrine progenitors and determined NEUROG3 dependency of associated genes to uncover direct targets.

Methods: We generated a novel hiPSC line (NEUROG3-HA-P2A-Venus), where NEUROG3 is HA-tagged and fused to a self-cleaving fluorescent VENUS reporter. We used the CUT&RUN technique to map NEUROG3 occupancy and epigenetic marks in pancreatic endocrine progenitors (PEP) differentiated from this hiPSC line. We integrated NEUROG3 occupancy data with chromatin status and gene expression in PEPs and their NEUROG3-dependence. In addition, we searched whether NEUROG3 binds type 2 diabetes mellitus (T2DM)-associated variants at the PEP stage.

Results: CUT&RUN revealed a total of 863 NEUROG3 binding sites assigned to 1268 unique genes. NEUROG3 occupancy was found at promoters as well as at distant cis-regulatory elements frequently overlapping within PEP active enhancers. *De novo* motif analyses defined a NEUROG3 consensus binding motif and suggested potential co-regulation of NEUROG3 target genes by FOXA, RFX or PBX transcription factors. Moreover, we found that 22% of the genes downregulated in *NEUROG3*$^{-/-}$ hESC-derived PEPs are bound by NEUROG3 and thus likely to be directly regulated. NEUROG3 targets include transcription factors known to have important roles in islet cell development or function, such as *NEUROD1, PAX4, NKX2-2, SOX4, MLXIPL, LMX1B, RFX3*, and *NEUROG3* itself. Remarkably, we uncovered that NEUROG3 binds transcriptional regulator genes with enriched expression in human fetal pancreatic alpha (e.g., *IRX1, IRX2*), beta (e.g., *NKX6-1, SMAD9, ISX, TFCP2L1*) and delta cells (*ERBB4*) suggesting that NEUROG3 could control islets subtype programs. Moreover, NEUROG3 targets genes critical for insulin secretion in beta cells (e.g., GCK, ABCC8/KCNJ11, CACNA1A, CHGA, SCG2, SLC30A8 and PCSK1). In addition, we unveiled a panel of ncRNA potentially regulated by NEUROG3. Lastly, we identified several T2DM risk SNPs within NEUROG3 peaks suggesting a possible developmental role of NEUROG3 in T2DM susceptibility.

Conclusion: Mapping of NEUROG3 genome occupancy in PEPs uncovers an unexpectedly broad, direct control of the endocrine gene regulatory network (GRN) and raises novel hypotheses on how this master regulator controls islet and beta cell differentiation.

Keywords: NEUROG3; iPSC, islet progenitors, CUT&RUN, T2DM, SNPs

1. INTRODUCTION

Diabetes results either from an auto-immune destruction of beta cells (Type 1 diabetes) or defective insulin secretion combined with the resistance of the peripheral tissues to insulin action (Type 2 diabetes). These forms of diabetes are considered as polygenic. On the other hand, mutations in single genes can also lead to rare early-onset forms of diabetes, thus defined as monogenic diabetes, for which the prevalence is estimated to be 2-5% of diabetes cases [1]. Monogenic diabetes is classified according to the age of onset and includes Neonatal Diabetes Mellitus (NDM) and Maturity Onset Diabetes of the Young (MODY), where diabetes occurs before 6 months and 25 years, respectively. These rare forms of diabetes result from mutations in genes controlling beta cell development, function, or both, including genes encoding essential transcription factors such as *PTF1A, PDX1, HNF1B, NEUROG3, RFX6,* or *NEUROD1* [1].

Among these genes, the bHLH transcription factor NEUROG3 is the key regulator of endocrine cell fate decision in the embryonic pancreas. In the mouse, all pancreatic islet cells derive from Neurog3-expressing pancreatic endocrine progenitors (PEP) and depend on *Neurog3* [2; 3]. *Neurog3*-deficient newborn mice die within a few days; they are diabetic since they lack insulin-secreting beta cells as well as all other islet cells [3]. In humans, homozygous or compound heterozygous mutations in *NEUROG3* have been identified in patients developing diabetes [4-7]. The pathology declared at various ages from neonatal to childhood, probably reflects differences in how severely NEUROG3 function is compromised. Of note, patients also developed rare forms of congenital malabsorptive diarrhea due to the lack of intestinal endocrine cells, which do not develop in the absence of NEUROG3 [4; 8]. Using pancreatic differentiation of human pluripotent stem cells as a model, it has been shown that *NEUROG3* is required for endocrine cell development [9; 10].

Despite its key function in endocrine commitment, the direct genetic program implemented by NEUROG3 is largely unknown both in mouse and human. Genome wide approaches have been performed to identify *Neurog3*-regulated genes in the mouse embryonic pancreas [11]. However, since the islet lineage is lost in the absence of Neurog3, a comparison of transcriptomes between Neurog3-deficient and control embryos revealed the entire islet transcriptome from endocrine progenitors to mature hormone-expressing cells, not only Neurog3 regulated genes. Direct Neurog3 target candidate genes such as *NeuroD*, *Nkx2-2*, *Insm1*, *Pax4*, *Neurog3* and *Cdkn1a* have been characterized previously using *in vitro* EMSA, Chromatin Immunoprecipitation (ChIP) and transactivation assays [12-16]. Using EMSAs and ChIP-qPCR, direct binding of NEUROG3 to *NKX2-2* and *NEUROG3* regulatory regions in hES-derived pancreatic precursors was recently reported [17]. Nevertheless, genome wide analysis to identify NEUROG3 bound regions and thus the entire panel of potential direct NEUROG3 targets has not been described. Such studies have been hampered by the lack of sensitivity of ChIP-Seq technique combined with the scarcity of NEUROG3-expressing endocrine progenitors.

Here we generated a novel hiPSC cell line where endocrine progenitor cells can be purified, and NEUROG3 is epitope tagged. We used the cleavage under targets and release using nuclease (CUT&RUN)

technique, a method allowing transcription factor profiling from a low cell number [18-20], to identify NEUROG3 bound regions across the genome in hiPSC-derived pancreatic cells. We confirmed previously known NEUROG3 targets validating the experimental approach. Importantly, we identified many NEUROG3 targets that have not been reported before. Comparison with transcriptome data identified NEUROG3 bound genes that are enriched in human fetal pancreatic progenitors and regulated by NEUROG3. Our study uncovers an unexpectedly large panel of direct NEUROG3 targets in human pancreas progenitors, comprising an extensive endocrine GRN that implements NEUROG3 function.

2. MATERIALS AND METHODS

2.1. Culturing of iPSC lines - Wild type SB AD3.1 [21] and AD/N3HAV lines were maintained as undifferentiated hiPSC in mTeSR1 medium (Stem Cell Technology) on 1:30 diluted Matrigel (hESC grade, Corning) coated tissue culture surfaces, with everyday medium change. Cells were split every 3 or 4 days with TrypLE Select (Fisher) and seeded at 1.5-4x10e5 in a Matrigel-coated p35 plate containing 5µM Y27632 (Stem Cell Technologies) (mTeSR+Y) for the first day.

2.2. Generation of the NEUROG3-HA-P2A-Venus line - The SB AD3.1 line [21] was co-transfected with a pX458-plasmid expressing the sg1 guide RNA (Suppl. Table 1) and the Cas9 fused to GFP, and the targeting vector pBSII-KS-hNEUROG3-3HA-2A-3NLS-Venus-pA, both generated in the laboratory. Nucleofection was performed according to the manufacturer instructions (Amaxa), with 8x10e5 SB AD3.1 cells mixed with 2.5 µg each plasmid DNA, and cells were seeded on a p35 with mTeSR1+Y. The following day, cells were harvested with TryPLE, resuspended in PBS containing 2% FCS, 10 µM Y27632 and 1% Penicillin/Streptomycin, sorted by expression of GFP and seeded as 200 and 500 cells in several p35 in mTeSR1+Y. After 12 days, clones were picked by scratching and expanded for banking while genotyping.

2.3. Genotyping - DNA was purified from collected cells using the Nucleospin Tissue XS kit (Macherey-Nagel) according to the manufacturer instructions and genotyped by nested PCR using primers described in Suppl. Figure 1 and Suppl. Table 1. PCR products were purified using the Nucleospin Gel and PCR clean-up kit (Macherey-Nagel) and sequenced with appropriated primers (Suppl. Table 1) at Eurofins Genomics (Ebergberg, Germany).

2.4. Differentiation of hiPSC cells to pancreatic progenitors - Cells were differentiated according to the protocol of Petersen et al. (2017) [21]. At 80-90% confluency, cells were harvested with TryPLE and seeded at 3x10e5 cells/cm2 on Growth Factor Reduced Matrigel-coated 24wells- or 6wells-plates (CellBind Corning) in mTESR+Y. Differentiation was initiated 24 h after seeding. Cells were first rinsed with 1x PBS, then exposed daily to freshly prepared differentiation medium (Suppl. Table 1).

2.5. Flow cytometry analyses - Cells were harvested with TrypLE Select as described above, quenched with 3 volumes of MCDB131-3 medium containing 5 mM Y27632 (M3Y), washed once with PBS and fixed with 4% formaldehyde in PBS for 20 min. After 2 washes with PBS, cells were either stored at +4°C in PBS, BSA 1% for delayed analysis, or permeabilized 30 min with PBS, Triton 0.2%, 5% Donkey serum (permeabilization buffer) then incubated overnight at +4°C with primary antibodies (Suppl. Table 1) diluted in permeabilization buffer. After 2 washes with PBS-Triton 0.1%, 0.2% BSA (PBSTB), cells were incubated for 1-2 hour at RT with fluorophore-conjugated secondary antibodies (Suppl. Table 1) diluted in permeabilization buffer. After 2 washes with PBSTB, cells were resuspended at 1M/ml in PBS, 1% BSA, filtered on 85µm nylon mesh and analyzed on a BD Fortessa LSR II Cell analyser (BD Bioscience).

2.6. Immunofluorescence imaging - Cells were washed twice with PBS, fixed with 4% formaldehyde in PBS for 20 min, permeabilized 30 min with PBS-Triton 0.5% and blocked for 30 min in PBSTB. Cells were incubated with primary antibodies (Suppl. Table 1) diluted in PBSTB overnight at 4°C, washed 3x in PBS-Triton 0.1% and incubated for 1-2 hour at RT with fluorophore-conjugated secondary antibodies (Suppl. Table 1) diluted in PBSTB. Cells were washed twice in PBSTB, nuclei were stained with Dapi 50 ng/mL in PBST. Image acquisition was done on an inverted fluorescence microscope Leica DMIRE2.

2.7. Flow cytometry sorting of Venus+ cells - Cells were harvested with TrypLE Select at day 13 of differentiation, quenched with 3 volumes of M3Y, centrifuged 4 min at 200g, resuspended at 5M/ml in M3Y and sorted using a FACS Fusion/Aria in M3Y at +4°C. Venus+ cells were collected and either used immediately or cryoconserved in Cryostor10 (Stem Cell Technologies) at -80°C.

2.8. CUT&RUN - We followed the protocol of Hainer and Fazzio (2019) [22] with minor modifications. Freshly sorted cells (75,000 for anti NEUROG3, HA and CTRL donkey anti sheep (DAsh) antibodies and 18,000 cells for H3K4me3 antibody) or thawed sorted cells (15,000 for H3K27me3 and rabbit anti mouse control antibodies) were washed once with 1mL cold PBS and resuspended in nuclear extraction buffer (NE, 20mM HEPES-KOH, pH 7.9, 10mM KCl, 0.5mM Spermidine, 0.1% Triton X-100, 20% glycerol, freshly added protease inhibitors). After 3 min spinning at 4°C at 600g, cells were resuspended in 600 µL NE buffer. Concanavalin A beads (Polysciences, 25 µL bead slurry/sample) were washed twice with ice-cold Binding buffer (20mM HEPES-KOH, pH 7.9, 10mM KCl, 1mM CaCl2, 1mM $MnCl_2$) and resuspended in 300 µL Binding buffer. Nuclei were added to beads with gentle vortexing and incubated for 10 min at 4°C with gentle rocking. Bead-bound nuclei were blocked with 1 mL cold Blocking buffer (20mM HEPES, pH 7.5, 150mM NaCl, 0.5mM Spermidine, 0.1% BSA, 2mM EDTA, freshly added protease inhibitors) by gentle pipetting, incubated 5 min at RT and washed in 1mL cold Wash buffer (WB, 20mM HEPES, pH 7.5, 150mM NaCl, 0.5mM Spermidine, 0.1% BSA, freshly added protease inhibitors) and resuspended in 250 µL cold WB. 250 µL of primary antibody (Suppl. Table 1) diluted 1:100 in cold WB were added with gentle vortexing, and samples were incubated overnight with gentle rocking at 4°C. Samples were washed twice in 1 mL cold WB, and resuspended in 250 µL cold WB. When indicated, incubation with a secondary antibody (Donkey anti Sheep IgG, 1:200) was performed for 1h at 4°C in WB under gentle rocking. After 2 washes with 1 ml WB, and resuspension in 250 µl WB, 200 µL of pA-MN (diluted at 1.4 ng/mL in cold WB) was added with gentle vortexing, and samples were incubated with rotation at 4°C for 1 hour. The pA-MN was produced in-house, according to the protocol described by [23] and using the pK19pA-MN plasmid, obtained from Ulrich Laemmli (RRID:Addgene_86973; http://n2t.net/addgene:86973). Samples were washed twice in 1 mL cold WB and resuspended in 150 µL cold WB. Three µL of 100 mM $CaCl_2$ were added upon gentle vortexing to activate the MN. After 30 min of digestion, reactions were stopped by addition of 150 µL 2XSTOP buffer (200mM NaCl, 20mM EDTA, 4mM EGTA, 50ug/mL RNaseA, 40ug/mL glycogen) and DNA fragments were released by passive diffusion during incubation at 37°C for 20 min. After centrifugation for 5 min at 16,000g at +4°C to pellet cells and beads, 3 µL 10% SDS and 2.5 µL Proteinase K 20mg/ml were added to the supernatants, and samples were incubated 10 min at 70°C. DNA purification was done with phenol/chloroform/isoamyl alcohol extraction followed by chloroform extraction using MaxTract tubes (Qiagen). DNA was precipitated with ethanol after addition of 20 µg glycogene, and resuspended in 36.5µl 0.1XTE.

2.9. High throughput sequencing - Illumina sequencing libraries were prepared at the Genomeast facility (IGBMC, Illkirch). CUT&RUN samples were purified using Agencourt SPRIselect beads (Beckman-Coulter). Libraries were prepared from 10 ng of double-stranded purified DNA using the MicroPlex Library Preparation kit v2 (Diagenode) following the manufacturer's protocol with some modifications. Illumina compatible indexes were added through a PCR amplification (3 min at 72°C, 2 min at 85°C, 2 min at 98°C; [20 sec at 98°C, 10 sec at 60°C] x 13 cycles). Amplified libraries were purified and size-selected using Agencourt SPRIselect beads (Beckman Coulter) by applying the following ratio: volume of beads / volume of libraries = 1,4 / 1. The libraries were sequenced on Hiseq 4000 as Paired-End 2x100 base reads following Illumina's instructions.

2.10. Bioinformatics analyses

2.10.1. Data processing - Image analysis and base calling were performed using RTA 2.7.3 and bcl2fastq 2.17.1.14. Reads were trimmed using cutadapt v1.9.1 with option: -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA -m 5 -e 0.1. Paired-end reads were mapped to Homo Sapiens genome (assembly hg38) using Bowtie2 (release 2.3.4.3, parameter: -N 1 -X 1000). Reads overlapping with ENCODE hg38 blacklisted region V2 were removed using Bedtools. Reads were size selected to <120bp and >150bp, since it has been reported that small reads define more precisely TF binding site, whereas larger reads (>150bp) result from sites occupied by nucleosomes [18; 19]. Bigwig tracks were generated using bamCoverage from deepTools for ≤120 bp and ≥150 bp fragments separately. Tracks were normalized with RPKM method. The bin size is 20. ≤120 bp fragments are used for samples obtained with anti NEUROG3 (VLSR28), HA (VLSR27) and the control donkey anti sheep (DAsh, thereafter named CTRL, VLSR29) antibodies and ≥150 bp fragments for samples obtained with anti H3K4me3 (VLSR32), anti H3K27me3 (VLSR44) and the rabbit anti mouse control (RAM, VLSR41) antibodies. Bigwig tracks (reads <120 bp long for NEUROG3, HA and CTRL samples and >150 pb for histone marks) were displayed on the reference genome *hg38* using UCSC genome browser. For simplicity, only the DAsh CTRL is illustrated throughout the manuscript. Heatmaps and K-means clustering was done using seqMINER v1.3.3g [24]. To compare with previously published data obtained from human *in vitro* derived pancreatic endocrine progenitors [25], multipotent progenitors [26] and from adult islets [27], we converted coordinates of bed and bigwig files to *hg19* coordinates using the UCSC Liftover and the bigwigLiftOver tools (https://github.com/milospjanic/bigWigLiftOver), respectively. Genomic tracks were visualized using http://meltonlab.rc.fas.harvard.edu/data/UCSC/SCbetaCellDiff_ATAC_H3K4me1_H3K27ac_WGBS_tracks.txt.

2.10.2. Peak calling - Peak calling was performed with the Sparse Enrichment Analysis for CUT&RUN SEACRv1.3 tool [28] (https://seacr.fredhutch.org), using the norm and stringent mode on the <120bp size selected reads and VLSR29 (DAsh CTRL) as a control for VLSR27 (HA) and VLSR28 (NEUROG3) datasets. To identify overlapping genomic regions, peak coordinates were intersected using the BEDtools 2.22.0 command *intersect interval files* (http://use.galaxeast.fr).

2.10.3. Association of peaks with genomic features and genes - Genomic annotation was first performed using the HOMER v3.4 [29] *annotatePeaks.pl* script with the default settings (promoters-transcription start site (TSS) from –1 kb to +100 bp to the TSS and transcription termination sites (TTS) from –100 bp to +1 kb of the TTS. GREAT 4.0.4 [30] was used to assign NEUROG3/HA peaks to their nearest coding gene(s) using basal settings (each gene is assigned a basal regulatory domain of 5 kb upstream and 1 kb downstream of its TSS. The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than 1,000 kb extension in one direction. Each peak is associated with all genes whose regulatory domain it overlaps). The NEUROG3 peaks or the distal peaks defined by GREAT (>5kb from TSS) were intersected with enhancers regions of hESC-derived endocrine progenitors (EN) lifted over to the *hg38* genome ([25], GSE139816).

2.10.4. Motifs identification and analyses - De novo motif discovery and known motifs enrichment analysis were performed using the HOMER v3.4 [29] *findMotifsGenome.pl* script with default settings (200-bp windows centred on peak summits, motif lengths set to 8, 10 and 12 bp, hypergeometric scoring). For the 6 most significant *de novo* motifs identified, known best match motifs were associated if their Homer score was >0.85. Known co-occurring motifs were manually curated to exclude redundant bHLH motifs. Co-occurance of the *de novo* identified NEUROG3 motif and known RFX6 or FOXA2 motifs was done on the entire peak sequences using the HOMER script *annotatePeaks.pl* with -size given and -m <motif*n*.motif> options.

2.10.5. Functional annotations - Gene functional annotation and clustering was carried out with DAVID v6.8 (https://david.ncifcrf.gov/home.jsp, [31]), using GO Biological Process, GO Cellular Component and KEGG Pathways. Selected terms significantly enriched and sorted by -Log(P-value) are displayed. To identify NEUROG3 transcription factors target genes, the peaks-assigned genes names were intersected by Venny 2.1.0 (https://bioinfogp.cnb.csic.es/tools/venny/) with a list of 1734 TF combining the 1639 human TF identified by [32] with the 1496 human TF taken from the human protein atlas (https://www.proteinatlas.org) (Suppl. Table 2). To identify ncRNA genes at the vicinity of NEUROG3 binding sites, we extended the genomic coordinates of the 2226 *de novo* and 2112 previously annotated LncRNA listed by Akerman et al (2017) [33] by 100kb (or 5kb) in both directions and intersected with *hg19* converted coordinates of NEUROG3 peaks.

2.10.6. Overlap between bound genes and differentially expressed genes - Differential expression analysis between *NEUROG3*$^{-/-}$ hESC line differentiated to day 13 and its wild-type counterpart, from corresponding RNA-seq data [34], was performed using a negative binomial GLM fit and Wald significance test implemented in the Bioconductor package DESeq2 version 1.16.1 [35]. The variables considered for the GLM model were the batch and condition. Differentially expressed genes were defined as those having a Benjamini – Hochberg-adjusted Wald test $P < 0.05$. A total of 319 genes were differentially expressed, from which 312 were downregulated in *NEUROG3*$^{-/-}$ cells. The list of genes significantly enriched in NEUROG3-eGFP^{+} human pancreatic and endocrine progenitors differentiated *in vitro* from hESC cells (2852 genes), compared to NEUROG3-eGFP^{-} cells, was taken from Liu et al, 2014 [36]. Both gene lists were intersected with NEUROG3 bound genes list by Venny 2.1.0. Expression of genes of interest in human fetal pancreas and during *in vitro* differentiation of human embryonic stem cells to pancreatic endocrine cells was examined using https://descartes.brotmanbaty.org [37] and http://hiview.case.edu/public/BetaCellHub/differentiation.php [38], respectively.

2.10.8. Enrichment of T2D-FG associated variants - The NEUROG3 bound EN_enhancers (hg19) and the NEUROG3-bound regions (hg19) were intersected using Bedtools 2.22.0 with the 23,144 genetic variants associated with T2D and glycemic traits (T2D-FG) on 109 loci, compiled by Miguel-Escalada et al [27].

2.10.9. Data availability - Raw data have been deposited in the GEO database under accession code GSE171963. hESC-derived *NEUROG3*$^{-/-}$ [34] and hESC-derived NEUROG3-eGFP^{+} cell [36] RNA-seq data are from E-MTAB-7185 and GSE54879, respectively. hESC-derived endocrine progenitors (EN) data (enhancers, H3K27ac ChIP-seq and RNA-seq, Ref [25]) are from GSE139817.

3. RESULTS AND DISCUSSION

3.1 Identification of NEUROG3 targets in hiPSC-derived endocrine progenitors

To unveil the endocrinogenic program implemented by NEUROG3, we mapped NEUROG3 occupancy across the genome during directed differentiation of hiPSC into beta cells. We first generated an hiPSC line where NEUROG3 is tagged with 3 HA epitopes and fused to a cleavable nuclear VENUS fluorescent reporter (NEUROG3-HA-P2A-Venus) (Figure 1A and Suppl. Figure 1). Using the protocol described by Petersen et al (2017) [21] and adapted from Rezania et al. (2014) [39], we differentiated the NEUROG3-HA-P2A-VENUS hiPS cells along the pancreatic and islet lineage until the pancreatic endocrine progenitor (PEP) stage 5, at day 13 (Figure 1A). By immunofluorescence, we found that NEUROG3-positive cells are indeed co-expressing HA and Venus, as expected (Suppl. Figure 2A-B). Accordingly, FACS analyses showed a correlation between HA and Venus expression (Suppl. Figure 2C). All the Venus+ cells are expressing PDX1 (Suppl. Figure 2A,C), as expected and previously shown with a NEUROG3-eGFP hiPSC line [21]. To map NEUROG3-bound regions, we used the CUT&RUN technique, an alternative to ChIP-seq for low input cell numbers [18; 19]. This technique is based on the recruitment of micrococcal nuclease, fused to protein A (pA-MNase), to antibody-bound sites within the genome in intact nuclei (Figure 1A). The subsequently cleaved fragments are recovered and sequenced. Endocrine progenitors were purified at day 13 (d13) of differentiation (Suppl. Figure 2D), and CUT&RUN experiments were performed on Venus+ cells using anti-NEUROG3 and anti-HA antibodies. To map chromatin states, we also profiled active (H3K4me3) and repressive (H3K27me3) histone marks.

We identified 1873 and 1428 peaks using NEUROG3 and HA antibodies, respectively (Figure 1B). To enhance the stringency of NEUROG3-bound regions, we intersected both datasets, defining NEUROG3 occupancy at 863 common sites (Figure 1B and Suppl. Table 2). These high confidence NEUROG3 binding sites were found at promoters (35%), in introns (30%), and in intergenic regions (31%) (Figure 1C) and were assigned by GREAT to 1268 unique genes, with 573 peaks (66%) assigned to 2 or more genes (Figure 1D). NEUROG3 binding to distal regions (located >5kb from the TSS of their associated gene) was observed for 65% of sites (557 peaks for 1042 genes) (Figure 1E). Remarkably, 90.8% (506 peaks) of these distal NEUROG3 bound regions were located within enhancer regions of hESC-derived endocrine progenitors (EN), as defined through their H3K27 acetylation by Alvarez-Dominguez et al. (2019) [25] (Figure 1F). In agreement, we found that H3K4me3 active histone marks were enriched at the NEUROG3 peaks compared to the H3K27me3 repressive marks (Figure 1G), indicating NEUROG3 binding at active promoters and enhancers. Taken together, we uncovered the NEUROG3 cistrome in PEPs, suggesting that NEUROG3 activates gene transcription by binding both proximal and distal cis-regulatory elements.

3.2 CUT&RUN detects previously identified and novel binding sites in known NEUROG3 targets

To validate the CUT&RUN approach for identifying of NEUROG3 bound regions in PEPs, we first examined previously characterized direct targets. As expected, peaks were identified in *NEUROD1, NKX2-2, PAX4*, *INSM1,* and *NEUROG3* [12; 13; 15-17], some of which at sites already mapped by ChIP-qPCR and/or EMSA and luciferase assays (Figure 1H). Interestingly, we identified two unreported NEUROG3 binding sites upstream of *NKX2-2* gene and upstream of *NEUROG3* TSS (purple arrowheads in Figure 1H). The sites identified for *NEUROG3* were distinct from the one reported previously by ChIP-qPCR [17] but overlapped with the conserved *Neurog3* enhancer region described in the mouse [40], supporting that NEUROG3 regulates its own transcription [12]. The peak assigned to *INSM1* is distantly located >180kb downstream of its TSS, within an intron of the *RALGAPA2* genes. However, this region has been identified as a super-enhancer directly linked to the *INSM1* gene in promoter capture HiC studies performed in adult pancreatic islet [27] (Figure 1H and Suppl. Figure 3) suggesting a function in the regulation of *INSM1* expression. Of note, we found no binding site for the *CDKN1A* gene, shown in the mouse to be directly regulated by NEUROG3 and to promote cell cycle exit in PEP [14]. It is possible that the NEUROG3 target NEUROD1 serves as an intermediate since NEUROD1 was shown to similarly inhibit cell proliferation by directly regulating *Cdkn1a* transcription [41]. Altogether, these data validate the CUT&RUN technique to unravel NEUROG3 bound sites genome-wide and suggest that the expected NEUROG3-driven endocrinogenic programs are activated in hiPSC-derived PEP.

3.3 Consensus NEUROG3 binding motif and co-binding of transcription factors

To determine the motifs enriched in the NEUROG3 binding regions, we performed a *de novo* motifs analysis [29] that revealed a strong enrichment for the RCCATCTGBY E-box type motif (CANNTG) recognized by bHLH transcription factors (Figure 2A). The NEUROG3 recognition motif is similar to NEUROD1 and NEUROG2 binding motifs, in agreement with the strong homology of the bHLH DNA binding domains between NEUROD and NEUROG families. Several additional motifs were found significantly enriched, such as the motif recognized by NFY, FOX, SP/KLF, RFX, and PBX TFs (Figure 2A-C and Suppl. Figure 4). Some TFs of these families have been reported to regulate pancreas development and islet cell differentiation, such as Pbx1 [42], Rfx3 and Rfx6 [43; 44]. Interestingly, the binding of the general NFY factors was reported biased towards regulatory elements with enhancer activity [45]. In agreement with our findings, KLF, FOXA1/A2, RFX, and MEIS1 (a PBX1 related homeobox gene) TFs have recently been predicted to bind to PEP Super Enhancers in a model of Core transcriptional regulatory circuits (CRCs) in the human islet lineage [25]. Of particular interest are the presence of FOX and RFX motifs in NEUROG3 bound regions. Indeed, FOXA1 and FOXA2 have been shown to act as pioneer factors facilitating chromatin access to other TFs at multiple stages during pancreas development [46]. 28.27% of the NEUROG3 peaks harbor a FOXA2 motif (Figure 2B). Moreover, 189 (21.97%) of NEUROG3 binding sites are bound by FOXA2 in *in vitro*-derived pancreatic multipotent progenitor cells (MPC) [26], and 36 of these sites match with *cis*-regulatory modules (CRM), binding sites of multiple TFs essential for early pancreas development, among which 15 are regulator elements of TF genes (Figure 2D and 2E, and see below) [26]. The pioneer activity of FOXA2 that was also described during human *in vitro* pancreatic progenitor differentiation [47] could therefore be required for the subsequent gene activation mediated by NEUROG3 at primed enhancers. The fact that FOXA2 possibly regulates *NEUROG3* (as shown in mouse [40]) and our findings that NEUROG3 binds *FOXA2* (Figure 2E) provides a possible additional regulatory loop between these two TFs. Interestingly, we identified a RFX6 motif in 37.54% of NEUROG3 peaks

(Figure 2B) and revealed the co-occurence of the NEUROG3 motif with the RFX6 motif in 1/5 of the peaks, from which 1/3 had an additionnal FOXA2 motif (Figure 2C). Several NEUROG3 bound genes were indeed previously identified as Rfx6 targets in a mouse beta cell line [43] (Figure 2C and data not shown). Altogether, FOXA2, as well as RFX6, might be important coregulators of the transcription of NEUROG3 direct targets.

3.4 Integration of NEUROG3 occupancy and gene expression in the islet lineage

Gene ontology (GO) analyses revealed that NEUROG3-bound regions are associated with genes retaled to GO terms such as endocrine pancreas development and insulin secretion, in agreement with the expected proendocrine function of NEUROG3 (Figure 3A and Suppl.Table 2). We, therefore, scrutinized NEUROG3 bound genes that are expressed in the islet lineage, expecting these genes to be downregulated in *NEUROG3*$^{-/-}$ cells, or upregulated in NEUROG3-enriched cells. We thus used RNA-seq data comparing the transcriptome of *NEUROG3*$^{-/-}$ versus wild-type hESC line, differentiated to d13 [34] and a previously published list of genes significantly enriched in NEUROG3-eGFP^{+} hESC-derived endocrine progenitors [36]. From the 319 differentially expressed genes in *NEUROG3*$^{-/-}$ cells, 312 were downregulated (Suppl. Table 2) from which 69 (22%) were directly bound by NEUROG3 (Figure 3B-C, Suppl. Table 2). From the 2852 enriched genes in NEUROG3-eGFP^{+} cells [36], 277 were bound by NEUROG3, including 56 that were downregulated in the *NEUROG3*$^{-/-}$ cells (Figure 3B-C, Suppl. Table 2). Many of these genes encode for TFs or proteins known to regulate islet cell differentiation and function (see below). Thus, a total of 290 genes specifically expressed in the endocrine lineage (out of 2988) are bound by NEUROG3, suggesting that NEUROG3 directly regulates the expression of about 10% of islet specific genes. In addition, we compared the NEUROG3 cistrome with the human pancreatic adult islet regulome [27]. We found that 782 (90.6%) NEUROG3 binding sites matched with at least one of the adult islet regulatory elements, with 655 (75.90%) of them localized within active enhancers or promoters (Figure 3D and E). This suggests that most of the genes regulated by NEUROG3 are still active in the adult islets, supporting the hypothesis that the transient expression of NEUROG3 at the PEP stage is required to initiate the endocrinogenic program while other transcription factors sustain the transcription of NEUROG3 targets in mature islets by binding to the same regulatory elements.

3.5 NEUROG3 binds to a subset of islet enriched transcription factors

To better understand how NEUROG3 drives islet cell differentiation, we first examined the TF genes bound by NEUROG3. Among the 1268 NEUROG3 bound genes, 138 encode for TFs (Figure 4A and Suppl. Table 2). From those, 24 were enriched in NEUROG3-eGFP^{+} hESC-derived endocrine progenitors, including 8 genes also downregulated in *NEUROG3*$^{-/-}$ cells. Besides the TF genes already mentioned above (*NKX2-2, NEUROD1, NEUROG3, PAX4, INSM1,* and *FOXA2*), we unraveled several other TFs known to control islet cell development in the mouse or human, including *SOX4, RFX3, ST18 (MYT3), MLXIPL, NKX6.1* and *LMX1B* (Figure 4A and data not shown), suggesting they could to be regulated directedly by NEUROG3. For instance, NEUROG3 binds to a region in intron 1 of *MLXIPL* (Figure 4B) previously shown to be bound by Rfx6 and Nkx2-2 in the mouse [43; 48]. Interestingly, a NEUROG3 binding site was found 33kb upstream of *SOX4* TSS, and three additional peaks were found within the adjacent *CDKAL1* locus (Figure 4C). The later region is likely to act as a distant enhancer to regulate *SOX4* in islet cells, as suggested by promoter capture HiC data [27] and found to be an activated enhancer (H3K27ac enriched) also at the endocrine progenitor stage [25] (Figure 4C). Thus, while Sox4 has been shown to regulate *Neurog3* expression and be required downstream of *Neurog3* to regulate endocrine differentiation in the mouse [49], *SOX4* might, in turn, be a direct target of NEUROG3. Importantly, we found that NEUROG3 binds to intron 2 of *LMX1B,* a transcription factor recently reported to be critical for generating human islet cells downstream of NEUROG3, suggesting direct transcriptional regulation of *LMX1B* by NEUROG3 (Figure 4D) [25]. Intriguingly, a NEUROG3 peak within the *GLIS3* coding sequence (exon 8) was assigned to both *GLIS3* and *RFX3* (Figure 4E). This peak nicely overlapped with an enhancer region at both endocrine progenitor and adult islets stages [25; 27]. In the adult islets, HiC showed that the two genes are spatially linked [27]. Moreover, RFX3 but not GLIS3 is highly expressed at the endocrine progenitor stage (Figure 4E, [25]) and has recently been documented as a human endocrine fate switch gene regulator [38]. Taken together, these data suggest a possible regulation of *RFX3* by NEUROG3 at the endocrine progenitor stage.

In a recent study, Alvarez-Dominguez et al. [25] described Core transcriptional Regulatory Circuits (CRCs) for each stages of in vitro beta cells differentiation, based on interconnected autoregulatory loops betweens TFs. Strikingly, out of the 40 TF genes defining the endocrine progenitors CRCs, we show here that 35% are bound by NEUROG3: *LMX1B, FOXA1, FOXA2, FOXP1, GATA4, INSM1, KLF3, KLF13, NKX2-2, RFX3, SOX4, SOX11, PAX4* and *PBX1* (Figure 4A). Of note, since the definition of CRCs relied on TF recognition motifs, NEUROG3, whose motif was not yet known, could not be integrated into the endocrine progenitors CRCs [25]. Importantly, our data reveal and provide molecular mechanistic insights into the role of NEUROG3 as a possible direct regulator of many TFs of the endocrine CRCs.

We further scrutinized the TFs dataset to examine whether NEUROG3 binds to genes known to control islet subtype development and to unveil novel candidates. We focused on transcription factor genes for which NEUROG3 binding site(s) coincided with endocrine progenitor active enhancer regions [25] and were enriched in developing alpha, beta or delta cells based on recent transcriptomic profiling of the human fetal pancreas [37] (Figure 5A). Importantly, an essential role of NEUROG3 in promoting the beta cell fate is supported by its direct regulation of *Pax4* expression, a critical regulator of beta cell development [50]. In addition to *Pax4*, *Nkx6-1* has been shown to be critical for endocrine progenitors to acquire a beta destiny in the mouse [51]. Supporting a possible direct regulation of *NKX6-1* by NEUROG3, we found a peak at 466kb downstream of *NKX6-1* TSS (Figure 5B). This region overlaps with an endocrine progenitors specific active enhancer region, suggesting that this site might be important for NEUROG3 regulated expression of *NKX6-1* in human islet progenitors. NEUROG3 binding sites were also associated with genes encoding TFs previously reported as markers for beta cells based on their expression, but not yet functionally addressed in endocrine cells development, such as *SMAD9* [52] and *TFCP2L1* [52] (Figure 5C). For *TFCP2L1,* however, the NEUROG3 binding region was not identified as an endocrine progenitor but an adult islet enhancer [27], belonging to an islet-TAD regulating the *GLI2* gene. Whether *TFCP2L1*, *GLI2*, or both, expressed in human fetal beta cells (Figure 5A), regulate beta destiny in a NEUROG3-dependent manner remains to be established. Of note, we additionally discovered *ETS2* and *ISX* as new NEUROG3 targeted TFs and whose expression is enriched in human feta beta cells, suggesting that they could play a role in human beta cell development (Figures 5A and C).

Regarding alpha destiny, no peaks were assigned to *ARX,* which is essential for alpha cell development in the mouse and human [50; 53]. Of note, NEUROG3 was found bound to regions associated with *IRX1* and *IRX2,* which are both enriched in human fetal (Figure 5A) and adult (Figure 5D and [25; 54]) alpha cells. Interestingly, *Irx2* was induced by ectopic *Neurog3* expression in the chick endoderm [11] and downregulated in hPSC-derived human islet cells lacking ARX [53]. Thus *IRX1/2* are attractive, alpha-specific, NEUROG3 direct targets, although their function in alpha cell development remains to be studied.

In contrast to alpha and beta cells, much less is known regarding the regulation of delta cell destiny. We did not find any binding of NEUROG3 associated with the delta transcription factor HHEX. Nevertheless, our analysis pointed to a possible NEUROG3-dependent candidate regulators of delta cell development. Indeed, we identified a NEUROG3 binding site within the first intron of the EGFR family member Erb-B2 Receptor Tyrosine Kinase 4 (*ERBB4*) gene (Figure 5E), that is highly and specifically expressed in human fetal (Figure 5A) and adult [54] delta cells, and whose ligand neuregulin-4 (NGR-4) was found to be essential for the determination of delta cells in mouse [55]. Of note, ERBB4 is cleaved by gamma-secretase to generate an intracellular domain endowed with TF regulatory activity [56]. Furthermore, during human *in vitro* beta cell differentiation, a gamma-secretase inhibitor is added at the endocrine progenitor stage to inhibit Notch signalling and further promote the beta lineage [39]. Whether the concomitant inhibition of ERBB4, by impeding the delta destiny, could favor the beta destiny remains to be tested.

Taken together, mapping NEUROG3 occupancy revelaled an unexpectedly broad direct control of TFs in the endocrine gene regulatory network (GRN).

3.6 NEUROG3 binds to genes involved in islet cell function

As mentioned above, gene ontology analyses revealed that many NEUROG3 bound genes were associated with insulin secretion, suggesting that NEUROG3 could regulate the expression of genes of the hormone secretory machinary. Indeed, NEUROG3 binding was found in genes linking glucose metabolism to electrical activity in beta cells and subsequent insulin secretion [57], such as the glucose sensor gene *GCK* and *ABCC8/KCNJ11* encoding subunits of the ATP-sensitive K+ channel (Figures 6A-C). Interestingly, other K+ (ATP-independent) channel genes (*KCNA3, KCNB2, KCND3, KCNK16, KCNMA1)*, which also contribute to glucose-stimulated insulin secretion, and are expressed in human fetal islet cells [37], were bound by NEUROG3 (Figure 6A and Suppl. Table 2). In the same line, the voltage-dependent Ca2+ channels (*CACNA1A, CACNA1C, CACNA1E, CACNA2D1, CACNB2*) or genes involved in the formation, composition, or release of secretory granules (*CHGA, SCG2, SLC30A8/ZNT8, SLC18A2/VMAT2, RGS16, RGS4, SYT7, SYT13 SYT3, STX2, STXBP1*) or proinsulin processing (*PCSK1, CPE*) (Figures 6A, D-G, and Suppl. Table 2) are associated to NEUROG3 binding sites. We did not find any binding of NEUROG3 to hormone genes. NEUROG3 binding was also identified in the somatostatin receptor genes *SSTR1*, *SSTR2,* and *SSTR5* involved in the paracrine regulation of insulin and glucagon secretion [57] (Figure 6H and Suppl. Table 2).

These findings of NEUROG3 bound genes involved in islet cell function were unexpected due to the transient expression of NEUROG3 in endocrine progenitors. Interestingly, several of these target genes, like *ABCC8/KCNJ11, CACNA1A, SLC30A8,* and *SLC18A2,* are not or weakly expressed in endocrine progenitors compared to more differentiated hESC-derived beta (SC-beta) or adult islet cells (Figures 6C, E and G, [25]). We noticed that some of these genes, (e.g., *ABCC8/KCNJ11* and *SLC18A2,* (Figures 6C and E), are decorated with H3K27me3 at or close to their TSS, suggesting that NEUROG3 could prime these genes at the endocrine progenitor stage, but subsequent binding by other TF could be required for their full activation. Thus, NEUROG3 might not only promote islet destiny in uncommitted pancreatic progenitors, but also control the initiation of later generic endocrine programs in maturing islet and beta cells.

3.7 NEUROG3 occupancy and ncRNA genes

Non-coding RNAs, such as long non-coding RNA (LnRNA) and microRNA (miRNAs) actively contribute to regulating developmental processes, including pancreatic endocrine specification [33; 58]. We found that 588 (68.3%) NEUROG3 binding sites are less than 100kb distant (98 even within 5kb) from at least one of the human LncRNA expressed in islet cells, compiled by Akerman et al. (2017) [33] (Figure 7A and Suppl. Table 2). Among them, HI-LNC66 (nearby *NEUROD1*), HI-LNC103 (*ABCC8*), HI-LNC4389 (*CACNA1A*), HI-LNC832 (*SLC18A2*) or HI-LNC2984 (*GCG*) could thus as well be targeted by NEUROG3 as their nearby genes (Figures 6B-E and Suppl. Table 2). More generally, this opens the possibility that NEUROG3 could regulate LncRNA expression and consequently their regulatory activity on islet function. Using HOMER annotation, we additionally identified 218 binding sites whose nearest TSS belongs to non-coding genes, among which 66 encode long non coding intergenic LincRNAs, 31 miRNA, and 23 antisense RNA (Figure 7A and Suppl. Table 2). Several ncRNA, not described in Akerman et al. (2017) list [33], were antisense to genes already mentioned above as bound and therefore possibly regulated by NEUROG3, such as *GLIS3, ISX, KCND3, KCNMA1,* and *SSTR5*. NEUROG3 could therefore regulate these genes directly or indirectly by regulating the expression of their antisense RNA. In line with this hypothesis, the somatostatin receptor SSTR5-AS1 is downregulated in *NEUROG3*$^{-/-}$ PEP cells, whereas SSTR5 is enriched in NEUROG3-eGFP^{+} PEP cells (Suppl. Table 2). *LINC00261,* nearby to *FOXA2,* is highly expressed during pancreatic differentiation [25; 58] (Figure 2E) and was shown recently to be required for the efficient differentiation of hESCs to insulin-producing cells [58], through the trans-regulation of *PAX4* and *MAFB* TFs, whereas its *cis*-regulatory role on *FOXA2* is debated [58; 59]. *FOXA2* but not *LINC00261* is downregulated in *NEUROG3*$^{-/-}$ PEP (Suppl Table 2). Therefore, whether NEUROG3 regulates *LINC00261,* or *FOXA2*, or both, is an open question that can be extended to other NEUROG3 bound genes with ncRNAs in their vicinity.

3.8 NEUROG3 binding at T2DM risk variants

Genome-wide association studies (GWAS) have identified hundreds of genetic variants associated with increased T2DM susceptibility [60]. It is essential to understand how these T2DM linked SNPs contribute to the disease, which genes they affect and how, whether it is by altering the protein sequence or, most frequently, distal cis-regulatory elements. Miguel-Escalada et al. [27] have compiled a list of 23,154 genetic variants associated with T2D and/or fasting glycemia (T2D/FG SNPs) within 109 loci. We found an overlap between 7 risk loci (harboring 152 SNPs) and NEUROG3 bound endocrine progenitor enhancers (Figures 8A-B and Suppl. Table 2). Moreover, 4 of these risk loci had additional SNPs, not in the abovementioned endocrine enhancers but falling within a NEUROG3-binding site (Figures 8A-B and Suppl. Table 2). A closer examination of the genomic regions of some of these 7 risk loci was performed. The risk alleles rs1799884, rs635299, and rs114152784 lie within NEUROG3-binding sites at the promoter regions of *GCK, SIX5,* and *MDC1*, respectively (Figures 6B, 8C and data not shown) and rs7245708 at the bi-directional promoter of QPCTL and SNRPD2 (Figure 8C). *SIX5*, *QPCTL,* and *SNRPD2* belong to the GIPR GWAS locus, and whereas all 4 genes are variably expressed at the endocrine progenitor stage (Figure 8C and [25]), GIPR is the most highly expressed in the human fetal pancreas [37]. The NEUROG3 binding sites within the *CDKAL1* locus coincide with several T2D-FG SNPs previously assigned to the distal gene *SOX4* (Figure 4C and [27]). The UBE2Z risk region contains several genes expressed at the endocrine progenitor stage and in the fetal pancreas that could be affected by SNPs, such as *ATP5G1, CALCOCO2, TTLL6,* and *UBE2Z* itself (Figure 8D and [37]). A NEUROG3 binding site lies upstream TTLL6 TSS (-1255bp) that is expressed higher at the endocrine progenitor stage than in beta cells (Figure 8D, [25]). In mouse, *TTLL6* expression follows that of NEUROG3 [61] and mouse invalidated for the *Ttll6* gene revealed decreased circulating glucose level (https://www.mousephenotype.org/data/genes/MGI:2683461). Altogether, these data support the potential regulation of *TTLL6* by NEUROG3. Interestingly, the NEUROG3 binding site assigned to the *ARAP1* gene coincided with an islet active enhancer (R11, Figure 8E) belonging to a regulatory region of the *STARD10* and *FCHSD2* genes, highly enriched in T2D/FG variants [27; 62]. The variable region (VR) within this locus was demonstrated to be required for insulin secretion [62]. *STARD10* is already expressed at the endocrine progenitor stage and in the fetal pancreas (Figure 8E and [25; 37]), but whether this expression is regulated by NEUROG3 and this regulation affected by genetic variants remain to be established.

Taken together, the overlap of T2DM SNPs and NEUROG3 bound regions suggests that T2DM susceptibility could arise from impaired, NEUROG3-dependent genetic programs.

4. CONLUSION

Despite the major progress made in the generation of functional beta cells from pluripotent stem cells for a cell therapy in diabetes, directed differentiation protocols lack robustness, and obtaining glucose-responsive cells remains difficult. The overall strategy was to mimic pancreas and islet developmental programs identified essentially in rodents. While the successful production of insulin-producing cells from PSC *in vitro* attests that these programs are remarkably conserved, it is important to acquire additional insights into the gene regulatory networks controlling islet cell development in human to optimize differentiation protocols. Notwithstanding the essential function of NEUROG3 in islet cell development in mouse and human, its downstream direct targets implementing the endocrinogenic program are essentially unknown. Searching and analyzing NEUROG3 binding sites in purified hiPSC-derived PEP, using the

CUT&RUN technique, revelead more than a thousand novel putative direct targets. Importantly, NEUROG3 binding largely overlaps with PEP active enhancers (H3K27ac binding) as defined by others [25], underlining the importance of NEUROG3 in promoting gene expression in PEPs. Our study revealed that NEUROG3 binds to a high number of important islet TFs as well as novel possible transcriptional regulators of islet cell differentiation. Moreover, a plethora of genes involved at several key steps of the insulin secretion pathway is bound by NEUROG3. In addition, we unveiled a panel of ncRNA potentially regulated by NEUROG3. Finally, we reveal that NEUROG3 binding overlaps with a series of T2DM associated SNPs. Taken together, our results suggest that NEUROG3 controls the progression of islet cell differentiation as well as the setting up of the hormone secretory machinery. The pleiotropic functions of NEUROG3 direct targets support the severity of NEUROG3 mutations in mice and humans as well as the potential of NEUROG3 to induce an endocrinogenic program when expressed ectopically. To our knowledge, this is the first genome wide characterization of NEUROG3 occupancy in iPSC-derived PEPs.

Figure Legends

Figure 1. Characterization of the genome-wide binding sites of NEUROG3 in human hiPSC-derived pancreatic endocrine progenitors. (A) Overview of the study: a 5-stage protocol was used to differentiate hiPSC to pancreatic endocrine progenitors using the sequential supplementation of factors indicated. At day 13, Venus+ cells were sorted and used in a CUT&RUN experiment. Inset: schematic representation of the NEUROG3-3HA-P2A-3NLS-Venus allele. (B) Venn diagram showing the number and overlap of peaks identified by CUT&RUN with an anti-NEUROG3 or an anti-HA antibody. (C) Genomic distribution (number and % of peaks) of the 863-high confidence NEUROG3 binding sites. (D) Number of genes assigned by GREAT per NEUROG3 peaks. (E) Distance of NEUROG3 peaks to their gene(s)-associated TSS. (F) Overlap between NEUROG3 distal binding sites (>5kb from TSS) and enhancers regions of hiPSC-derived endocrine progenitors (EN), as defined by [25]. (G) Normalized read density surrounding NEUROG3 peak summit ±5Kb for H3K4me3 and H3K27me3 CUT&RUN data sets. (H) Genome browser tracks showing NEUROG3, HA, H3K4me3, H3K27me3 and the CTRL (Donkey anti Sheep antibody) CUT&RUN data at the *NEUROD1, NKX2-2, NEUROG3, PAX4* and *INSM1*/*RALGAPA2* loci. Identified NEUROG3 peaks are highlighted in light blue. Peaks matching previously reported NEUROG3 binding sites are indicated by blue arrowheads, newly discovered peaks by purple arrowheads, and reported sites not confirmed here by grey arrowheads [12; 13; 15-17].

Figure 2. TF motifs discovery in NEUROG3 binding sites. (A). *De novo* motif discovery ranked by P-value reflecting motif enrichment within peak summit ±100bp. Number (#) and % of target and background sequences harboring each motif is indicated for the 6 most significant motifs identified. Known best matching transcription factors were associated if their HOMER score was >0.85. (B) Selection of known co-occuring motifs ranked by P-value within the entire peak sequences. The full list of the 50 most significantly enriched motifs is shown in Suppl Figure 4. (C) Co-occurence of NEUROG3 *de novo* identified motifs with motifs for RFX6 and/or FOXA2 on the entire peak sequences. Some selected GREAT assigned genes are indicated, and in bold when identified as targets for Rfx6 in mouse beta cell line [43]. (D-E) Regions bound by NEUROG3 at PEP stage that are bound by FOXA2 at the MPC stage. 36 regions coincide with MPC *cis*-regulatory modules (CRM) defined by [26], from which, 15 regulate TF genes (D). (E) Genome browser tracks showing NEUROG3, HA, and the CTRL CUT&RUN data at the *FOXA2* loci. Coordinates are from *hg19*. The position of NEUROG3 binding sites is highlighted in light blue (or in grey, when identified in a single dataset). Data of RNA-seq (EN_RNA), H3K27ac ChIP-seq (EN_K27AC) and the position of enhancers (EN Enhancers) from hESC-differentiated to endocrine progenitors were taken from [25] and http://meltonlab.rc.fas.harvard.edu/data/UCSC/SCbetaCellDiff_ATAC_H3K4me1_H3K27ac_WGBS_tracks.txt. Position of hESC-derived multipotent progenitor cells (MPC) enhancers and *cis*-regulatory modules (CRM, defined as regions bound by at least two TF) are taken from [26]. Data from adult islets (Super-enhancers, Islet regulome, T2D/FG SNP and TAD-like regions) are taken from [27], isletregulome.org and http://epigenomegateway.wustl.edu/.

Figure 3. Integration of NEUROG3 occupancy and islet expression of target candidates. (A) Gene ontology analysis (biological process, KEGG pathways) showing selected significantly enriched terms (Log10 p-value ≥2) related to the 1268 NEUROG3 bound genes. (B) Venn diagram illustrating the overlap of NEUROG3 bound genes and the 312 downregulated genes in *NEUROG3*$^{-/-}$ hESC line differentiated to d13 stage (*NEUROG3*$^{-/-}$ down; [34]) and the 2852 genes enriched in NEUROG3-eGFP^{+} hESC line differentiated to pancreatic endocrine progenitors (NEUROG3-eGFP^{+} enriched; [36]). (C) List of the 69 NEUROG3 target genes downregulated in *NEUROG3*$^{-/-}$ PEP. In italics, the 13 bound genes downregulated in *NEUROG3*$^{-/-}$ PEP but not enriched in NEUROG3-GFP^{+} PEP. (D-E) NEUROG3 binding sites matching with one or more islets regulatory element(s) (islet regulome; [27]). OCR, open chromatin regions, CTCF, CTCF binding sites.

Figure 4. NEUROG3 binding to transcription factors genes. (A) NEUROG3 binds 138 genes encoding human TFs. A selection of TF genes is given. Complete list is given in Suppl Table 2, with the list of 1734 human TFs taken from [32] and https://www.proteinatlas.org. The 24 TF enriched in GFP-NEUROG3+ endocrine progenitors [36], among which 8 are downregulated in *NEUROG3*$^{-/-}$ endocrine progenitors [34] are underlined and in bold, respectively. The 14 TFs belonging to Core Regulatory Circuits (CRCs) in endocrine progenitors as defined by [25] are in red. (B-E) NEUROG3 binding to the *MLXIPL* (B), *SOX4-CDKAL1* (C), *LMXIB* (D) and *RFX3-GLIS3* (E) loci. See Figure 2E for legend description. In (E), RNA-seq data from primary islet beta (Beta-RNA) and alpha (Alpha-RNA) cells are taken from [25].

Figure 5. NEUROG3 binds to transcription factors genes enriched in islet-subtypes. (A) Enriched expression of a selection of transcriptional regulators in islet-subtypes (alpha, beta, delta cells) in the human fetal pancreas (taken from https://descartes.brotmanbaty.org). Expression of INS, GCG and SST is provided to define beta, alpha and delta cells, respectively. (B-D) NEUROG3 binding to transcription factor genes expressed in fetal alpha or beta cells: NKX6-1 (B), *SMAD9, TFCP2L1-GL12, ETS2* and *ISX* (C) for beta cells; *IRX1-IRX2* (D) for alpha cells. (E) NEUROG3 binding to the Receptor Tyrosine Kinase/transcription coactivator *ERBB4* gene expressed in developing delta cells. See Figure 2E for legend description.

Figure 6. NEUROG3 binding to genes regulating glucose-dependent insulin secretion. (A) Schematic representation of insulin secretion upon glucose sensing in a beta cell. The NEUROG3 bound genes are indicated in red. (B-H) NEUROG3 binding to genes involved in glucose-stimulated insuline secretion: (B) GCK; (C) ABCC8-KCNJ11; (D) CACNA1A; (E) SLC18A2; (F) CHGA; (G) SLC30A8 and (H) SSTR2. See Figure 2E for legend description.

Figure 7. NEUROG3 binds to ncRNA genes (A) Distribution of NEUROG3 binding sites and number of associated non coding (nc) genes located within 100kb (or 5kb from the TSS) of human beta-cell enriched

lncRNAs, taken from [33]. HI-LNC, human islet long non coding RNA. (B) Distribution of NEUROG3 binding sites within ncRNA genes, including LINC RNA, miRNA, antisense RNA and other RNA annotated by HOMER. Some selected examples are given, underlined when found enriched in NEUROG3-eGFP$^+$ PEP and in bold when downregulated in *NEUROG3*$^{-/-}$ PEP.

Figure 8. T2D and fasting glycemia (FG)-associated genetic variants located within NEUROG3-bound EN enhancers and NEUROG3-bound regions. (A) Venn diagram illustrating the overlap of NEUROG3 bound EN enhancers (left, [25]) or peaks (right) and the 23,154 T2D-FG SNPs distributed over 109 risk loci compiled by [27]. (B) The 7 risk loci overlapping with NEUROG3-bound enhancers or peaks, with the number of SNPs, NEUROG3-peaks assigned genes, some selected nearby genes and selected SNPs given for each risk loci. (C-E) NEUROG3 binding to the *GIPR-SNRPD2-FBXO46-SIX5* (C)*, UBE2Z-TTLL6* (D) and *ARAP1-STARD10* (E) loci. In E, the inset is a magnification of the region previously described as a regulatory region of the *STARD10* and *FCHSD2* genes [27; 62]. R5-13, regulatory regions; VR, variable region. See Figure 2E for legend description.

AUTHOR CONTRIBUTION

V.S., R.M., E.G.S, A.K., A.M. and S.G. performed iPSC gene editing, differentiations and characterizations. V.S. and R.M. performed the CUT&RUN experiments, B.J. the Illumina sequencing and T.Y, V.S and S.J. the bioinformatics analyses. C.B. produced the pA-MN. C.H. provided the SB AD3.1 line and expertise for iPSC culture. K.H.L and P.S. performed the RNA-seq data for NEUROG3$^{-/-}$ iPSC line and participated in the manuscript redaction. V.S. and G.G. conceived the work, analyzed the data and wrote the manuscript. G.G. obtained financial support.

ACKNOWLEDGMENTS

The authors thank the members of the Gradwohl team and the Genomeast platform (particularly Christelle Thibault-Carpentier and David Rodriguez), Flow cytometry and Cell culture facilities for the sequencing of the CUT&RUN samples, cell sorting and hiPSC maintenance respectively. The authors are grateful to I. Cebola for providing ChIP-seq data and R. Scharfmann for helpful discussions. The Gradwohl lab is funded by the Novo Nordisk Foundation (Challenge Grant NNF14OC0013655). Sequencing was performed by the GenomEast platform, a member of the 'France Génomique' consortium (ANR-10-INBS-0009). This work used the Integrated Structural Biology platform of the Strasbourg Instruct-ERIC center IGBMC-CBI supported by FRISBI (ANR-10-INBS-0005-001). IGBMC is supported by the grant ANR-10-LABX-0030-INRT, a French State fund managed by the Agence Nationale de la Recherche under the frame program Investissements d'Avenir ANR-10-IDEX-0002-02.

CONFLICT OF INTEREST

The authors have declared no competing interest

APPENDIX A. SUPPLEMENTARY DATA

Supplemental Figure 1. Generation of the NEUROG3-HA-P2A-Venus hiPSC line by CRISPR/Cas9. (A) CRISPR/Cas9 mediated targeting of the NEUROG3 locus to knock in the 3HA-P2A-3NLS-Venus cassette in fusion with NEUROG3 coding sequence. Position of the sgRNA and the primers used for genotyping and sequencing is indicated. (B). Genotyping by PCR using the indicated primers. The clone used throughout this study is clone#31. See Suppl. Table 1 for oligonucleotide sequences.

Supplemental Figure 2. Differentiation of NEUROG3-HA-P2A-Venus hiPSCs line to pancreatic endocrine progenitors (PEP) and FACS-sorting of Venus+ cells. (A-B) Immunofluorescence staining for NEUROG3, Venus and PDX1 (A) or HA (B) at PEP stage (day 13 of differentiation). DNA counterstained with Dapi is shown as an inset within the merged images. Scale bar = 50 µM. Arrows point to a selection of cells co-expressing NEUROG3, Venus and PDX1 (A) or NEUROG3, Venus and HA (B). (C) Representative flow cytometry plots showing the percentage of cells expressing NEUROG3-3HA (anti-HA antibody), Venus (anti-GFP anibody) and PDX1 (anti-PDX1 antibody) at day 13 for the differentiated SB AD3.1 and NEUROG3-HA-P2A-Venus hiPSC lines. (D) Representative flow cytometry plot showing the sorted living Venus+ cells at day 13 used for CUT&RUN experiments.

Supplemental Figure 3. NEUROG3 binds to *INSM1-RALGAPA2* locus. Genome browser tracks showing NEUROG3, HA, H3K4me3, H3K27me3 and the CTRL CUT&RUN data at the *INSM1-RALGAPA2* locus. Coordinates are from hg19. The position of NEUROG3 binding sites is highlighted in light blue. Position of endocrine progenitor enhancers (EN Enhancers) were taken from [25]. Data from adult islets (Super-enhancers, Islet regulome, TAD-like regions and promoter capture HiC pc-HiC_islets) are taken from [27], isletregulome.org and http://epigenomegateway.wustl.edu/.

Supplemental Figure 4. List of the 50 most significantly enriched TFs known motifs in NEUROG3 binding sites, in regions defined by the entire peak coordinates.

Supplemental Table 1.

ST1.1. List of oligonucleotides

ST1.2. List of antibodies

ST1.3. hiPSC differentiation protocol and media

Supplemental Table 2.

ST2.1. The 863 NEUROG3 binding sites, identified by CUT&RUN with both the anti NEUROG3 and anti HA antibodies, and annotated with HOMER v3.4

ST2.2. The 863 NEUROG3 binding sites are assigned to 1268 unique genes by GREAT v4.0.4

ST2.3. Co-occurence of NEUROG3, FOXA2 and RFX6 motifs in the NEUROG3 peaks

ST2.4. Gene ontology performed with https://DAVID.org on the 1268 NEUROG3 bound genes

ST2.5. Genes downregulated in NEUROG3$^{-/-}$ hESC line differentiate to pancreatic endocrine progenitors

ST2.6. NEUROG3 bound genes and expression in islet lineage

ST2.7. NEUROG3 bound TFs genes

ST2.8. T2D and FG-associated genetic variants located within NEUROG3-bound EN enhancers and/or NEUROG3-bound regions.

Bibliography

[1] Schwitzgebel, V.M., 2014. Many faces of monogenic diabetes. Journal of Diabetes Investigation 5(2):121-133.

[2] Gu, G., Dubauskaite, J., Melton, D.A., 2002. Direct evidence for the pancreatic lineage: NGN3+ cells are islet progenitors and are distinct from duct progenitors. 129(10):2447-2457.

[3] Gradwohl, G., Dierich, A., LeMeur, M., Guillemot, F., 2000. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. 97(4):1607-1611.

[4] Wang, J., Cortina, G., Wu, S.V., Tran, R., Cho, J.H., Tsai, M.J., et al., 2006. Mutant neurogenin-3 in congenital malabsorptive diarrhea. N.Engl.J.Med. 355(3):270-280.

[5] Rubio-Cabezas, O., Jensen, J.N., Hodgson, M.I., Codner, E., Ellard, S., Serup, P., et al., 2011. Permanent Neonatal Diabetes and Enteric Anendocrinosis Associated With Biallelic Mutations in NEUROG3. Diabetes 60(4):1349-1353.

[6] Pinney, S.E., Oliver-Krasinski, J., Ernst, L., Hughes, N., Patel, P., Stoffers, D.A., et al., 2011. Neonatal diabetes and congenital malabsorptive diarrhea attributable to a novel mutation in the human neurogenin-3 gene coding sequence. The Journal of clinical endocrinology and metabolism 96(7):1960-1965.

[7] Hancili, S., Bonnefond, A., Philippe, J., Vaillant, E., De Graeve, F., Sand, O., et al., 2017. A novel NEUROG3 mutation in neonatal diabetes associated with a neuro-intestinal syndrome. Pediatric diabetes 21:464.

[8] Mellitzer, G., Beucher, A., Lobstein, V., Michel, P., Robine, S., Kedinger, M., et al., 2010. Loss of enteroendocrine cells in mice alters lipid absorption and glucose homeostasis and impairs postnatal survival. The Journal of clinical investigation 120(5):1708-1721.

[9] McGrath, P.S., Watson, C.L., Ingram, C., Helmrath, M.A., Wells, J.M., 2015. The Basic Helix-Loop-Helix Transcription Factor NEUROG3 Is Required for Development of the Human Endocrine Pancreas. Diabetes 64(7):2497-2505.

[10] Zhu, Z., Li, Q.V., Lee, K., Rosen, B.P., González, F., Soh, C.-L., et al., 2016. Genome Editing of Lineage Determinants in Human Pluripotent Stem Cells Reveals Mechanisms of Pancreatic Development and Diabetes. Stem Cell:1-53.

[11] Petri, A., Ahnfelt-Ronne, J., Frederiksen, K.S., Edwards, D.G., Madsen, D., Serup, P., et al., 2006. The effect of neurogenin3 deficiency on pancreatic gene expression in embryonic mice. Journal of Molecular Endocrinology 37(2):301-316.

[12] Smith, S.B., Watada, H., German, M.S., 2004. Neurogenin3 activates the islet differentiation program while repressing its own expression. Molecular Endocrinology 18(1):142-149.

[13] Mellitzer, G., Bonne, S., Luco, R., Van de Casteele, M., Lenne-Samuel, N., Collombat, P., et al., 2006. IA1 is NGN3-dependent and essential for differentiation of the endocrine pancreas. Embo Journal 25(6):1344-1352.

[14] Miyatsuka, T., Kosaka, Y., Kim, H., German, M.S., 2011. Neurogenin3 inhibits proliferation in endocrine progenitors by inducing Cdkn1a. 108(1):185-190.

[15] Huang, H.P., Liu, M., El-Hodiri, H.M., Chu, K., Jamrich, M., Tsai, M.J., 2000. Regulation of the pancreatic islet-specific gene BETA2 (neuroD) by neurogenin 3. Molecular and Cellular Biology 20(9):3292-3307.

[16] Smith, S.B., Gasa, R., Watada, H., Wang, J., Griffen, S.C., German, M.S., 2003. Neurogenin3 and hepatic nuclear factor 1 cooperate in activating pancreatic expression of Pax4. The Journal of biological chemistry 278(40):38254-38259.

[17] Zhang, X., McGrath, P.S., Salomone, J., Rahal, M., McCauley, H.A., Schweitzer, J., et al., 2019. A Comprehensive Structure-Function Study of Neurogenin3 Disease-Causing Alleles during Human Pancreas and Intestinal Organoid Development. Developmental cell 50(3):367-380.e367.

[18] Hainer, S.J., Bošković, A., McCannell, K.N., Rando, O.J., Fazzio, T.G., 2019. Profiling of Pluripotency Factors in Single Cells and Early Embryos. Cell 177(5):1319-1329.e1311.

[19] Skene, P.J., Henikoff, S., 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife 6:576.

[20] Skene, P.J., Henikoff, J.G., Henikoff, S., 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nature Protocols 13(5):1006-1019.

[21] Petersen, M.B.K., Azad, A., Ingvorsen, C., Hess, K., Hansson, M., Grapin-Botton, A., et al., 2017. Single-Cell Gene Expression Analysis of a Human ESC Model of Pancreatic Endocrine Development Reveals Different Paths to &beta;-Cell Differentiation. Stem cell reports:1-37.

[22] Hainer, S.J., Fazzio, T.G., 2019. High-Resolution Chromatin Profiling Using CUT&RUN. Curr Protoc Mol Biol 126(1):e85.

[23] Schmid, M., Durussel, T., Laemmli, U.K., 2004. ChIC and ChEC; genomic mapping of chromatin proteins. Mol Cell 16(1):147-157.

[24] Ye, T., Krebs, A.R., Choukrallah, M.A., Keime, C., Plewniak, F., Davidson, I., et al., 2011. seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 39(6):e35.

[25] Alvarez-Dominguez, J.R., Donaghey, J., Rasouli, N., Kenty, J.H.R., Helman, A., Charlton, J., et al., 2020. Circadian Entrainment Triggers Maturation of Human In Vitro Islets. Cell Stem Cell 26(1):108-122 e110.

[26] Cebola, I., Rodríguez-Seguí, S.A., Cho, C.H.H., Bessa, J., Rovira, M., Luengo, M., et al., 2015. TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. Nature Cell Biology 17(5):615-626.

[27] Miguel-Escalada, I., Bonas-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., et al., 2019. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet 51(7):1137-1148.

[28] Meers, M.P., Tenenbaum, D., Henikoff, S., 2019. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics Chromatin 12(1):42.

[29] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., et al., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38(4):576-589.

[30] McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., et al., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28(5):495-501.

[31] Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44-57.

[32] Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., et al., 2018. The Human Transcription Factors. Cell 175(2):598-599.

[33] Akerman, I., Tu, Z., Beucher, A., Rolando, D.M.Y., Sauty-Colace, C., Benazra, M., et al., 2017. Human Pancreatic beta Cell lncRNAs Control Cell-Specific Regulatory Networks. Cell Metab 25(2):400-411.

[34] de Lichtenberg, K.H., Funa, N., Nakic, N., Ferrer, J., Zhu, Z., Huangfu, D., et al., 2018. Genome-Wide Identification of HES1 Target Genes Uncover Novel Roles for HES1 in Pancreatic Development. BioRxiv doi: https://doi.org/10.1101/335869.

[35] Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol 11(10):R106.

[36] Liu, H., Yang, H., Zhu, D., Sui, X., Li, J., Liang, Z., et al., 2014. Systematically labeling developmental stage-specific genes for the study of pancreatic beta-cell differentiation from human embryonic stem cells. Cell Research 24(10):1181-1200.

[37] Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., et al., 2020. A human cell atlas of fetal gene expression. Science 370(6518).

[38] Weng, C., Xi, J., Li, H., Cui, J., Gu, A., Lai, S., et al., 2020. Single-cell lineage analysis reveals extensive multimodal transcriptional control during directed beta-cell differentiation. Nat Metab 2(12):1443-1458.

[39] Rezania, A., Bruin, J.E., Arora, P., Rubin, A., Batushansky, I., Asadi, A., et al., 2014. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. Nature Biotechnology 32(11):1121-1133.

[40] van Arensbergen, J., Dussaud, S., Pardanaud-Glavieux, C., Garcia-Hurtado, J., Sauty, C., Guerci, A., et al., 2017. A distal intergenic region controls pancreatic endocrine differentiation by acting as a transcriptional enhancer and as a polycomb response element. PLoS ONE 12(2):e0171508.

[41] Mutoh, H., Naya, F.J., Tsai, M.J., Leiter, A.B., 1998. The basic helix-loop-helix protein BETA2 interacts with p300 to coordinate differentiation of secretin-expressing enteroendocrine cells. Genes Dev 12(6):820-830.

[42] Kim, S.K., Selleri, L., Lee, J.S., Zhang, A.Y., Gu, X., Jacobs, Y., et al., 2002. Pbx1 inactivation disrupts pancreas development and in Ipf1-deficient mice promotes diabetes mellitus. Nat Genet 30(4):430-435.

[43] Piccand, J., Strasser, P., Hodson, D.J., Meunier, A., Ye, T., Keime, C., et al., 2014. Rfx6 maintains the functional identity of adult pancreatic β cells. Cell reports 9(6):2219-2232.

[44] Ait-Lounis, A., Bonal, C., Seguín-Estévez, Q., Schmid, C.D., Bucher, P., Herrera, P.L., et al., 2010. The transcription factor Rfx3 regulates beta-cell differentiation, function, and glucokinase expression. Diabetes 59(7):1674-1685.

[45] Andersson, R., Sandelin, A., 2020. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet 21(2):71-87.

[46] Gao, N., LeLay, J., Vatamaniuk, M.Z., Rieck, S., Friedman, J.R., Kaestner, K.H., 2008. Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. Genes & development 22(24):3435-3448.

[47] Lee, K., Cho, H., Rickert, R.W., Li, Q.V., Pulecio, J., Leslie, C.S., et al., 2019. FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation. Cell reports 28(2):382-393 e387.

[48] Churchill, A.J., Gutiérrez, G.D., Singer, R.A., Lorberbaum, D.S., Fischer, K.A., Sussel, L., 2017. Genetic evidence that Nkx2.2 acts primarily downstream of Neurog3 in pancreatic endocrine lineage development. eLife 6.

[49] Xu, E.E., Krentz, N.A.J., Tan, S., Chow, S.Z., Tang, M., Nian, C., et al., 2015. SOX4 cooperates with neurogenin 3 to regulate endocrine pancreas formation in mouse models. Diabetologia 58(5):1013-1023.

[50] Collombat, P., Mansouri, A., Hecksher-Sorensen, J., Serup, P., Krull, J., Gradwohl, G., et al., 2003. Opposing actions of Arx and Pax4 in endocrine pancreas development. Genes Dev 17(20):2591-2603.

[51] Schaffer, A.E., Taylor, B.L., Benthuysen, J.R., Liu, J., Thorel, F., Yuan, W., et al., 2013. Nkx6.1 controls a gene regulatory network required for establishing and maintaining pancreatic Beta cell identity. PLoS Genetics 9(1):e1003274.

[52] Muraro, M.J., Dharmadhikari, G., Grun, D., Groen, N., Dielen, T., Jansen, E., et al., 2016. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst 3(4):385-394 e383.

[53] Gage, B.K., Asadi, A., Baker, R.K., Webber, T.D., Wang, R., Itoh, M., et al., 2015. The Role of ARX in Human Pancreatic Endocrine Specification. PLoS ONE 10(12):e0144100-0144124.

[54] Lawlor, N., Marquez, E.J., Orchard, P., Narisu, N., Shamim, M.S., Thibodeau, A., et al., 2019. Multiomic Profiling Identifies cis-Regulatory Networks Underlying Human Pancreatic beta Cell Identity and Function. Cell reports 26(3):788-801 e786.

[55] Huotari, M.A., Miettinen, P.J., Palgi, J., Koivisto, T., Ustinov, J., Harari, D., et al., 2002. ErbB signaling regulates lineage determination of developing pancreatic islet cells in embryonic organ culture. Endocrinology 143(11):4437-4446.

[56] Han, W., Sfondouris, M.E., Semmes, E.C., Meyer, A.M., Jones, F.E., 2016. Intrinsic HER4/4ICD transcriptional activation domains are required for STAT5A activated gene expression. Gene 592(1):221-226.

[57] Rorsman, P., Ashcroft, F.M., 2018. Pancreatic beta-Cell Electrical Activity and Insulin Secretion: Of Mice and Men. Physiol Rev 98(1):117-214.

[58] Gaertner, B., Carrano, A.C., Sander, M., 2019. Human stem cell models: lessons for pancreatic development and disease. Genes Dev 33(21-22):1475-1490.

[59] Jiang, W., Liu, Y., Liu, R., Zhang, K., Zhang, Y., 2015. The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression. Cell reports 11(1):137-148.

[60] Krentz, N.A.J., Gloyn, A.L., 2020. Insights into pancreatic islet cell dysfunction from type 2 diabetes mellitus genetics. Nat Rev Endocrinol 16(4):202-212.

[61] Scavuzzo, M.A., Hill, M.C., Chmielowiec, J., Yang, D., Teaw, J., Sheng, K., et al., 2018. Endocrine lineage biases arise in temporally distinct endocrine progenitors during pancreatic morphogenesis. Nature Communications 9(1):1607-1621.

[62] Hu, M., Cebola, I., Carrat, G., Jiang, S., Nawaz, S., Khamis, A., et al., 2021. Chromatin 3D interaction analysis of the STARD10 locus unveils FCHSD2 as a regulator of insulin secretion. Cell reports 34(11):108881.

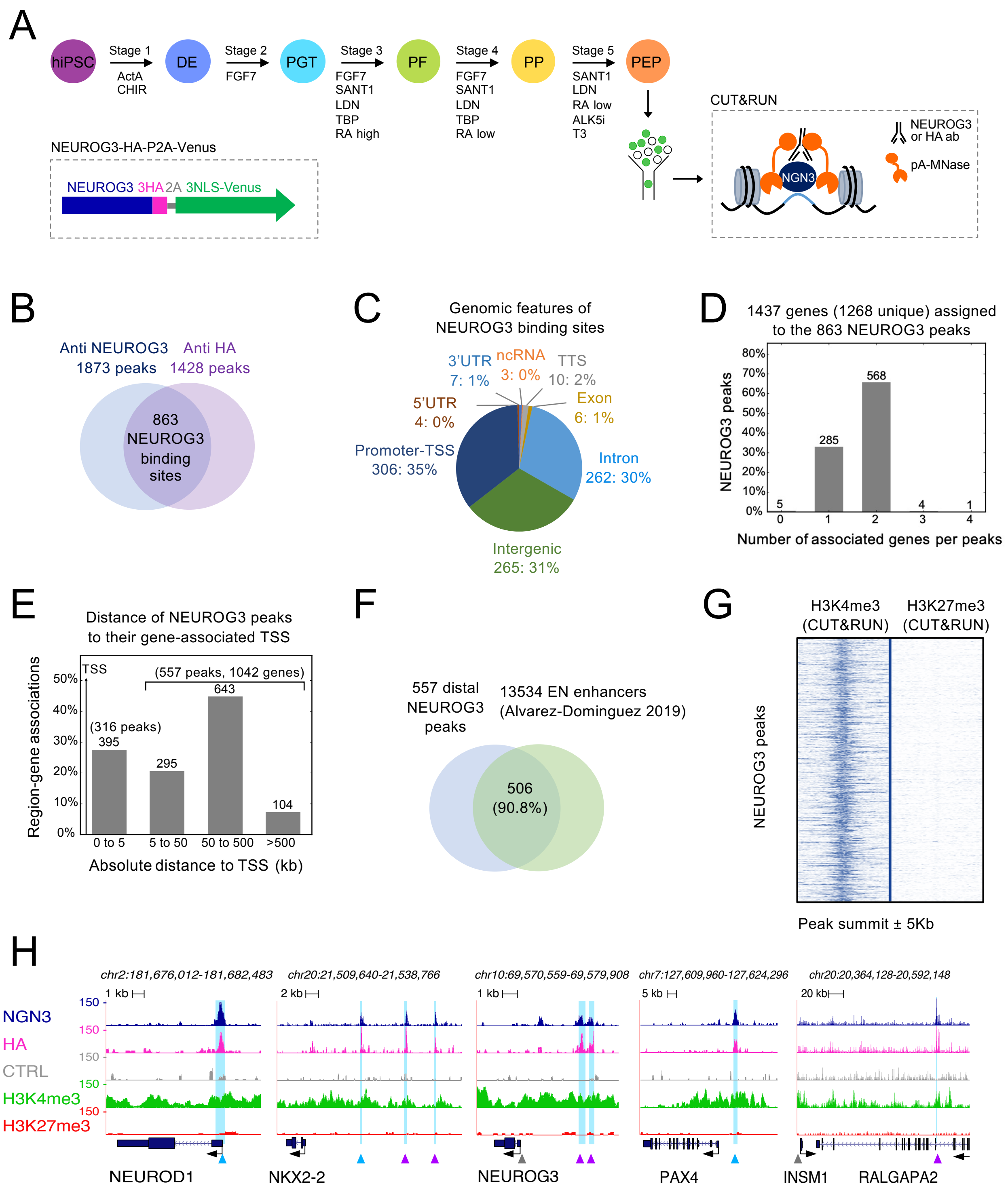
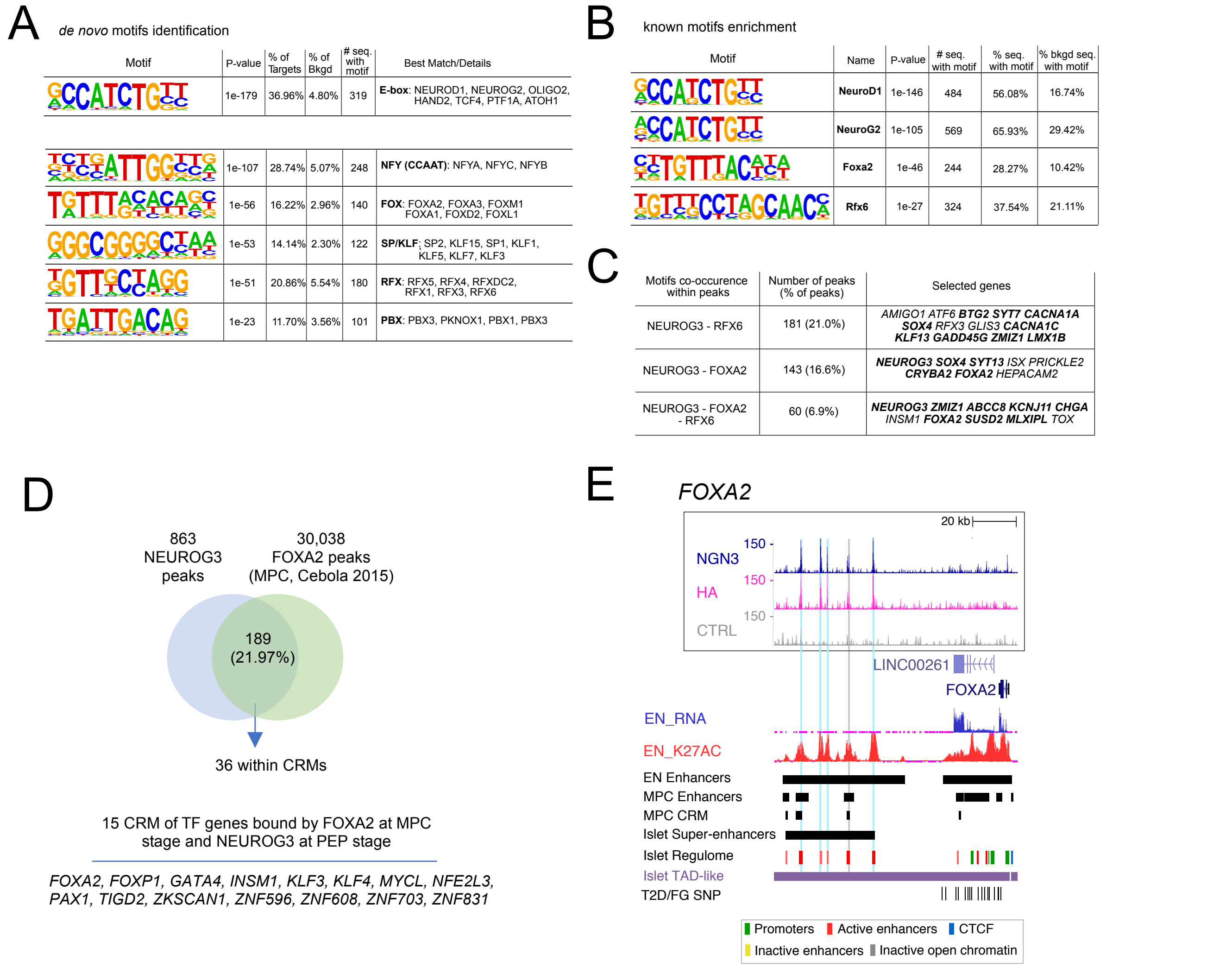

A
hiPSC
Stage 1
ActA
CHIR
DE
Stage 2
FGF7
PGT
Stage 3
FGF7
SANT1
LDN
TBP
RA high
PF
Stage 4
FGF7
SANT1
LDN
TBP
RA low
PP
Stage 5
SANT1
LDN
RA low
ALK5i
T3
PEP
NEUROG3-HA-P2A-Venus
NEUROG3 3HA 2A 3NLS-Venus
CUT&RUN
NGN3
NEUROG3 or HA ab
pA-MNase
B
Anti NEUROG3 1873 peaks
Anti HA 1428 peaks
863 NEUROG3 binding sites
C
Genomic features of NEUROG3 binding sites
3'UTR 7: 1%
ncRNA 3: 0%
TTS 10: 2%
Exon 6: 1%
5'UTR 4: 0%
Promoter-TSS 306: 35%
Intron 262: 30%
Intergenic 265: 31%
D
1437 genes (1268 unique) assigned to the 863 NEUROG3 peaks
NEUROG3 peaks
80%
70%
60%
50%
40%
30%
20%
10%
0%
5
285
568
4
1
0
1
2
3
4
Number of associated genes per peaks
E
Distance of NEUROG3 peaks to their gene-associated TSS
Region-gene associations
50%
40%
30%
20%
10%
0%
TSS
(557 peaks, 1042 genes)
(316 peaks)
395
295
643
104
0 to 5
5 to 50
50 to 500
>500
Absolute distance to TSS (kb)
F
557 distal NEUROG3 peaks
13534 EN enhancers (Alvarez-Dominguez 2019)
506 (90.8%)
G
H3K4me3 (CUT&RUN)
H3K27me3 (CUT&RUN)
NEUROG3 peaks
Peak summit ± 5Kb
H
chr2:181,676,012-181,682,483
chr20:21,509,640-21,538,766
chr10:69,570,559-69,579,908
chr7:127,609,960-127,624,296
chr20:20,364,128-20,592,148
1 kb
2 kb
1 kb
5 kb
20 kb
NGN3
HA
CTRL
H3K4me3
H3K27me3
150
150
150
150
150
NEUROD1
NKX2-2
NEUROG3
PAX4
INSM1
RALGAPA2

Figure 1



A *de novo* motifs identification

Motif	P-value	% of Targets	% of Bkgd	# seq. with motif	Best Match/Details
	1e-179	36.96%	4.80%	319	**E-box**: NEUROD1, NEUROG2, OLIGO2, HAND2, TCF4, PTF1A, ATOH1
	1e-107	28.74%	5.07%	248	**NFY (CCAAT)**: NFYA, NFYC, NFYB
	1e-56	16.22%	2.96%	140	**FOX**: FOXA2, FOXA3, FOXM1 FOXA1, FOXD2, FOXL1
	1e-53	14.14%	2.30%	122	**SP/KLF**; SP2, KLF15, SP1, KLF1, KLF5, KLF7, KLF3
	1e-51	20.86%	5.54%	180	**RFX**: RFX5, RFX4, RFXDC2, RFX1, RFX3, RFX6
	1e-23	11.70%	3.56%	101	**PBX**: PBX3, PKNOX1, PBX1, PBX3

B known motifs enrichment

Motif	Name	P-value	# seq. with motif	% seq. with motif	% bkgd seq. with motif
	NeuroD1	1e-146	484	56.08%	16.74%
	NeuroG2	1e-105	569	65.93%	29.42%
	Foxa2	1e-46	244	28.27%	10.42%
	Rfx6	1e-27	324	37.54%	21.11%

C

Motifs co-occurence within peaks	Number of peaks (% of peaks)	Selected genes
NEUROG3 - RFX6	181 (21.0%)	*AMIGO1 ATF6* ***BTG2 SYT7 CACNA1A SOX4*** *RFX3 GLIS3* ***CACNA1C KLF13 GADD45G ZMIZ1 LMX1B***
NEUROG3 - FOXA2	143 (16.6%)	***NEUROG3 SOX4 SYT13*** *ISX PRICKLE2* ***CRYBA2 FOXA2*** *HEPACAM2*
NEUROG3 - FOXA2 - RFX6	60 (6.9%)	***NEUROG3 ZMIZ1 ABCC8 KCNJ11 CHGA*** *INSM1* ***FOXA2 SUSD2 MLXIPL*** *TOX*

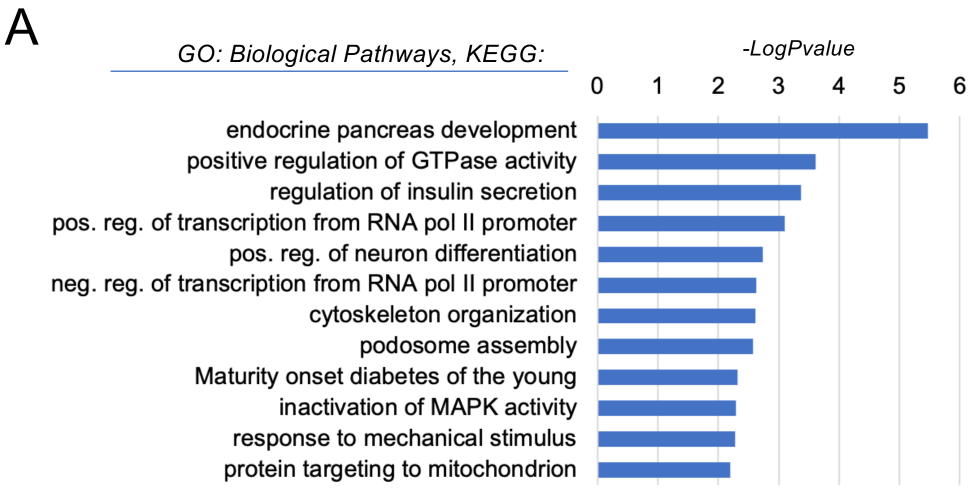
D

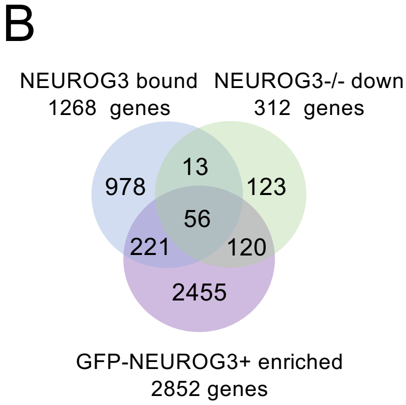
15 CRM of TF genes bound by FOXA2 at MPC stage and NEUROG3 at PEP stage

FOXA2, FOXP1, GATA4, INSM1, KLF3, KLF4, MYCL, NFE2L3, PAX1, TIGD2, ZKSCAN1, ZNF596, ZNF608, ZNF703, ZNF831

E

Figure 2

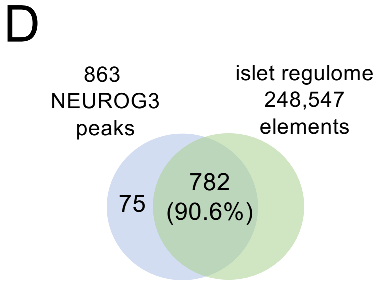




C

69 NEUROG3 target genes downregulated in NGN3-/- pancreatic endocrine progenitors

ABCC8
AMER3
BAIAP3
BTBD17
C1QL1
CD99L2
CDH18
CDH22
CELF3
CELF4
CERKL
CHGA
CHRNA4
CORO2B
CPE
CRYBA2
DDC
FAM92B
FTCD
GABBR2
GAD2
GCK
GNAO1
GPR142
HEPACAM2
HIST1H2BJ
INSM1
KCNA3
KCNB2
KCND3
KCNJ11
KCNK16
KIF19
LDLRAP1
LMX1B
MGAT4A
MLXIPL
NEUROD1
NEUROG3
NKX2-2
NOL4
NPTX2
OPCML
OPRK1
PAX4
PCSK1
PDE2A
PLXNA2
PRPH2
PSCA
RGS16
RGS9
SCG2
SEC11C
SEZ6L
SGCD
SLC18A2
SLC30A8
SPTBN4
SSTR1
SSTR2
ST18
STXBP1
SUSD2
SYNPR
TACR1
THSD4
TNR
UBE2QL1



E

Islet regulome regulatory element	NEUROG3 peaks
Active enhancer I	233
Active enhancer II	80
Active enhancer III	34
Active promoter	310
Inactive enhancer	94
Inactive OCR	33
Strong CTCF	4

788 regulatory elements for 782 NEUROG3 peaks

655 (75.90%) NEUROG3 peaks within islets active enhancers or promoters

Figure 3

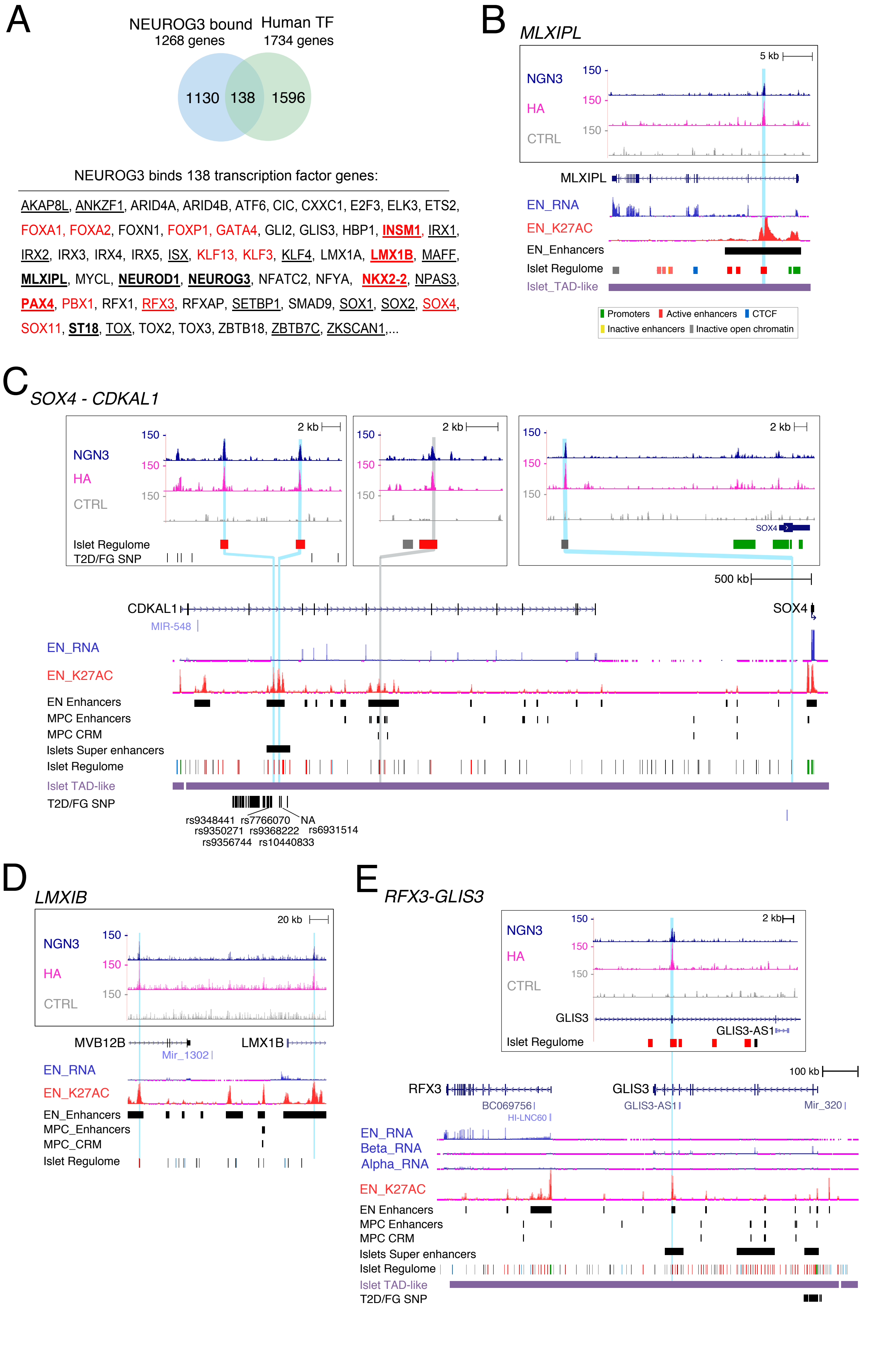

A
NEUROG3 bound
1268 genes
Human TF
1734 genes
1130
138
1596
NEUROG3 binds 138 transcription factor genes:
AKAP8L, ANKZF1, ARID4A, ARID4B, ATF6, CIC, CXXC1, E2F3, ELK3, ETS2,
FOXA1, FOXA2, FOXN1, FOXP1, GATA4, GLI2, GLIS3, HBP1, INSM1, IRX1,
IRX2, IRX3, IRX4, IRX5, ISX, KLF13, KLF3, KLF4, LMX1A, LMX1B, MAFF,
MLXIPL, MYCL, NEUROD1, NEUROG3, NFATC2, NFYA, NKX2-2, NPAS3,
PAX4, PBX1, RFX1, RFX3, RFXAP, SETBP1, SMAD9, SOX1, SOX2, SOX4,
SOX11, ST18, TOX, TOX2, TOX3, ZBTB18, ZBTB7C, ZKSCAN1,...
B
MLXIPL
5 kb
150
NGN3
HA
CTRL
MLXIPL
EN_RNA
EN_K27AC
EN_Enhancers
Islet Regulome
Islet_TAD-like
Promoters
Active enhancers
CTCF
Inactive enhancers
Inactive open chromatin
C
SOX4 - CDKAL1
2 kb
150
NGN3
HA
CTRL
Islet Regulome
T2D/FG SNP
SOX4
500 kb
CDKAL1
SOX4
MIR-548
EN_RNA
EN_K27AC
EN Enhancers
MPC Enhancers
MPC CRM
Islets Super enhancers
Islet Regulome
Islet TAD-like
T2D/FG SNP
rs9348441
rs7766070
NA
rs9350271
rs9368222
rs6931514
rs9356744
rs10440833
D
LMXIB
20 kb
150
NGN3
HA
CTRL
MVB12B
LMX1B
Mir_1302
EN_RNA
EN_K27AC
EN_Enhancers
MPC_Enhancers
MPC_CRM
Islet Regulome
E
RFX3-GLIS3
2 kb
150
NGN3
HA
CTRL
GLIS3
GLIS3-AS1
Islet Regulome
100 kb
RFX3
GLIS3
BC069756
HI-LNC60
GLIS3-AS1
Mir_320
EN_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN Enhancers
MPC Enhancers
MPC CRM
Islets Super enhancers
Islet Regulome
Islet TAD-like
T2D/FG SNP

Figure 4

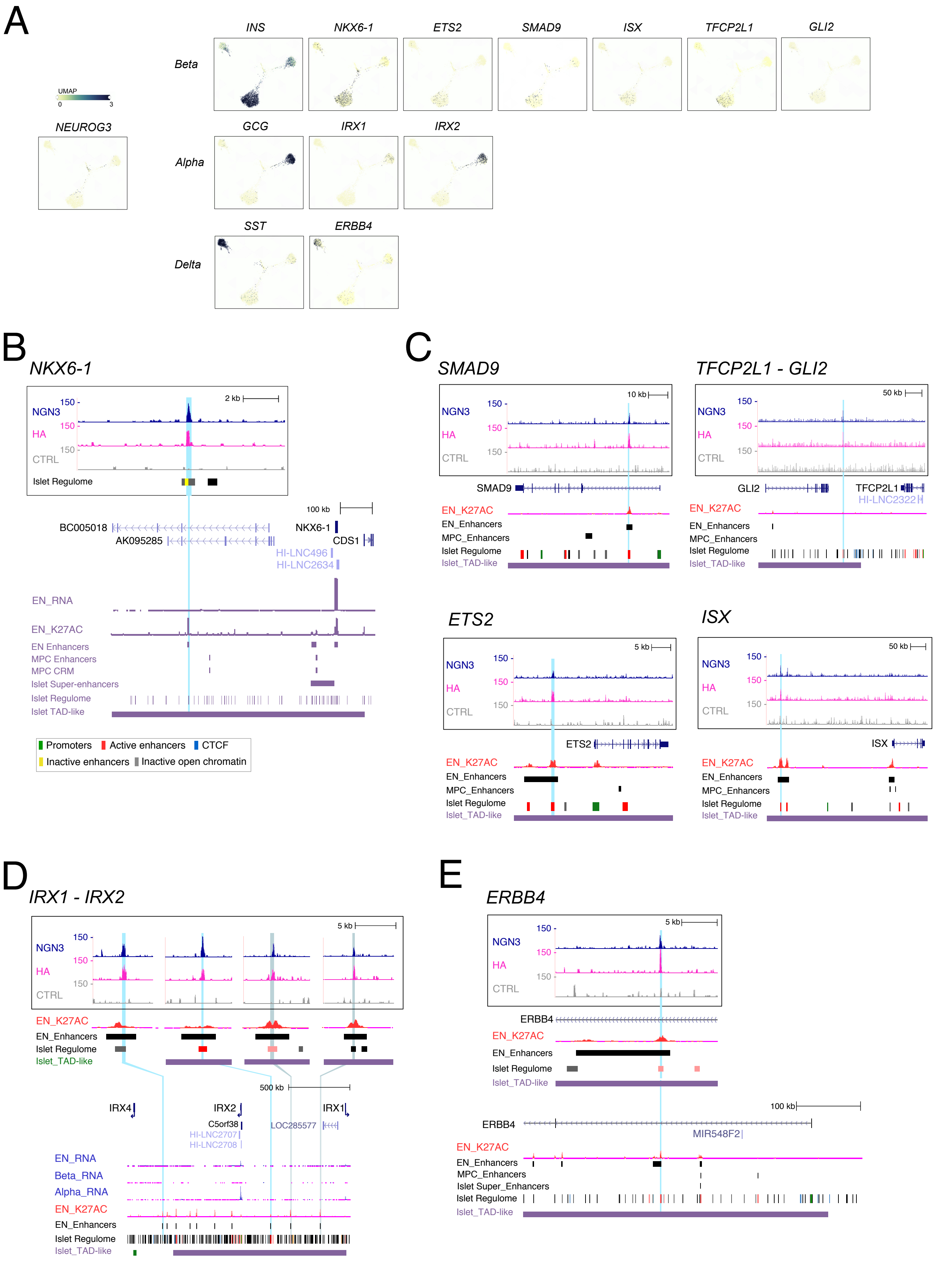

A
INS
NKX6-1
ETS2
SMAD9
ISX
TFCP2L1
GLI2
Beta
UMAP
0
3
NEUROG3
GCG
IRX1
IRX2
Alpha
SST
ERBB4
Delta
B
NKX6-1
2 kb
150 -
NGN3
HA
CTRL
Islet Regulome
100 kb
BC005018
NKX6-1
AK095285
CDS1
HI-LNC496
HI-LNC2634
EN_RNA
EN_K27AC
EN Enhancers
MPC Enhancers
MPC CRM
Islet Super-enhancers
Islet Regulome
Islet TAD-like
Promoters
Active enhancers
CTCF
Inactive enhancers
Inactive open chromatin
C
SMAD9
TFCP2L1 - GLI2
10 kb
50 kb
GLI2
TFCP2L1
HI-LNC2322
EN_K27AC
EN_Enhancers
MPC_Enhancers
Islet Regulome
Islet_TAD-like
ETS2
ISX
5 kb
D
IRX1 - IRX2
500 kb
IRX4
IRX2
C5orf38
HI-LNC2707
HI-LNC2708
LOC285577
IRX1
EN_RNA
Beta_RNA
Alpha_RNA
E
ERBB4
MIR548F2
Islet Super_Enhancers

Figure 5

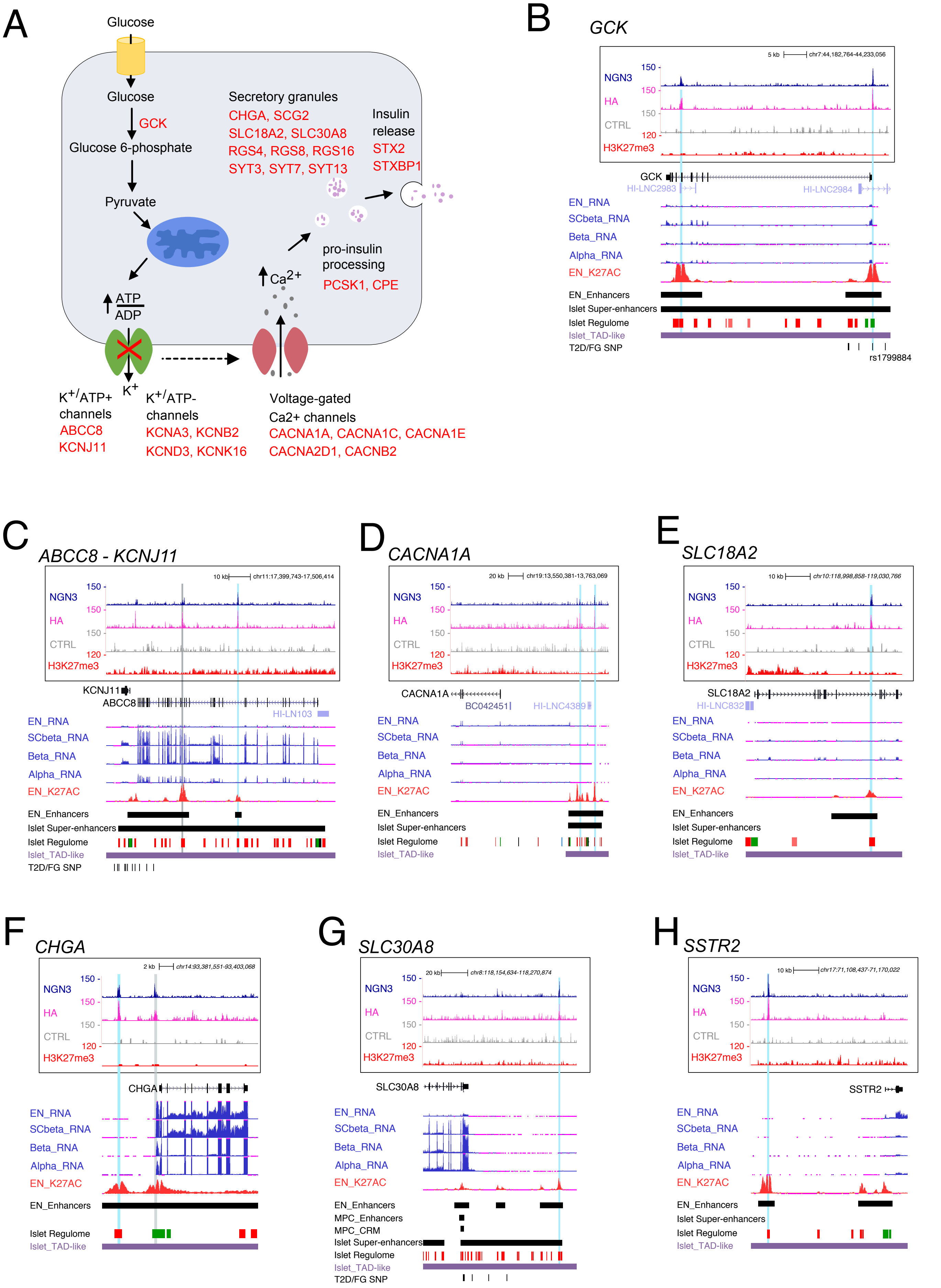

A
Glucose
Glucose
GCK
Glucose 6-phosphate
Pyruvate
ATP
ADP
Secretory granules
CHGA, SCG2
SLC18A2, SLC30A8
RGS4, RGS8, RGS16
SYT3, SYT7, SYT13
Insulin release
STX2
STXBP1
pro-insulin processing
PCSK1, CPE
Ca2+
K+
K+/ATP+ channels
ABCC8
KCNJ11
K+/ATP- channels
KCNA3, KCNB2
KCND3, KCNK16
Voltage-gated Ca2+ channels
CACNA1A, CACNA1C, CACNA1E
CACNA2D1, CACNB2
B
GCK
5 kb
chr7:44,182,764-44,233,056
NGN3
HA
CTRL
H3K27me3
150
150
150
120
GCK
HI-LNC2983
HI-LNC2984
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
Islet Super-enhancers
Islet Regulome
Islet_TAD-like
T2D/FG SNP
rs1799884
C
ABCC8 - KCNJ11
10 kb
chr11:17,399,743-17,506,414
NGN3
HA
CTRL
H3K27me3
150
150
150
120
KCNJ11
ABCC8
HI-LN103
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
Islet Super-enhancers
Islet Regulome
Islet_TAD-like
T2D/FG SNP
D
CACNA1A
20 kb
chr19:13,550,381-13,763,069
NGN3
HA
CTRL
H3K27me3
150
150
150
120
CACNA1A
BC042451
HI-LNC4389
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
Islet Super-enhancers
Islet Regulome
Islet_TAD-like
E
SLC18A2
10 kb
chr10:118,998,858-119,030,766
NGN3
HA
CTRL
H3K27me3
150
150
150
120
SLC18A2
HI-LNC832
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
Islet Super-enhancers
Islet Regulome
Islet_TAD-like
F
CHGA
2 kb
chr14:93,381,551-93,403,068
NGN3
HA
CTRL
H3K27me3
150
150
150
120
CHGA
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
Islet Regulome
Islet_TAD-like
G
SLC30A8
20 kb
chr8:118,154,634-118,270,874
NGN3
HA
CTRL
H3K27me3
150
150
150
120
SLC30A8
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
MPC_Enhancers
MPC_CRM
Islet Super-enhancers
Islet Regulome
Islet_TAD-like
T2D/FG SNP
H
SSTR2
10 kb
chr17:71,108,437-71,170,022
NGN3
HA
CTRL
H3K27me3
150
150
150
120
SSTR2
EN_RNA
SCbeta_RNA
Beta_RNA
Alpha_RNA
EN_K27AC
EN_Enhancers
Islet Super-enhancers
Islet Regulome
Islet_TAD-like

Figure 6

A

NEUROG3 binding sites within 100Kb (5kb) from
LncRNA genes expressed in human islet cells

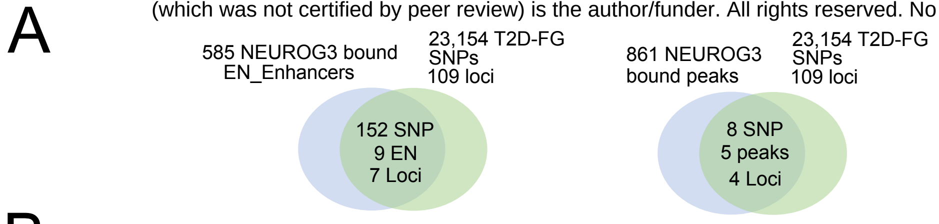
	nb NEUROG3 peaks	nb associated nc genes	selected LncRNA
LINC RNA	15 (5)	10 (5)	LINC01134, LINC01132, LINC00870, LINC00240, LINC00094
HI-LNCRNA	290 (51)	265 (50)	HI-LNC66, HI-LNC103, HI-LNC4389, HI-LNC832, HI-LNC2984
antisense RNA	39 (6)	33 (6)	ILF3-AS1, FBXL19-AS1, MORF4L2-AS1, DACT3-AS1
others	244 (69)	212 (67)	
	588 (131)	520 (128)	

B

NEUROG3 binding sites assigned to ncRNAs genes
by nearest TSS by HOMER

	nb NEUROG3 peaks	nb associated nc genes	selected ncRNA
LINC RNA	72	66	LINC00261, **LINC01108**, LINC00240, LINC00957, LIN02392
miRNA	35	31	<u>MIR3911,</u> MIR4644, MIR1200, MIR8080, MIR548F2
antisense RNA	24	23	**SSTR5-AS1**, GLIS3-AS1, ISX-AS1, KCND3-AS1, KCNMA1-AS1
others	69	67	<u>KC6</u>
	200	187	

Figure 7


A
585 NEUROG3 bound EN_Enhancers
23,154 T2D-FG SNPs 109 loci
152 SNP
9 EN
7 Loci
861 NEUROG3 bound peaks
23,154 T2D-FG SNPs 109 loci
8 SNP
5 peaks
4 Loci

B

T2D-FG risk loci	Nb SNPs within NEUROG3 bound EN	Nb SNPs within NEUROG3 peaks	NEUROG3-peaks assigned genes	Selected other nearby genes	Selected SNPs
CDKAL1	8	-	*CDKAL1, SOX4*		*rs7766070, rs10440833*
ARAP1	26	-	*ARAP1, PDE2A*	*STARD10, FCHSD2*	*rs74333814, rs7129793*
TP53INP1	89	-	*TP53INP1, NDUFAF6*		*rs74633235, rs896848*
PPT2	2	1	*MDC1, TUBB, HSPA1B*	*EHMT2, HSPA1A, SLC44A4*	*rs114152784, rs34103657*
UBE2Z	4	3	*ATP5G1, TTLL6*	*CALCOCO2*	*rs520878, rs200168742*
GCK	4	1	*GCK, YTK6*		*rs1799884, rs2971670*
GIPR	22	3	*SNRPD2-QPCTL-FBXO46-SIX5*	*DMPK, SYMPK*	*rs635299, rs7245708*

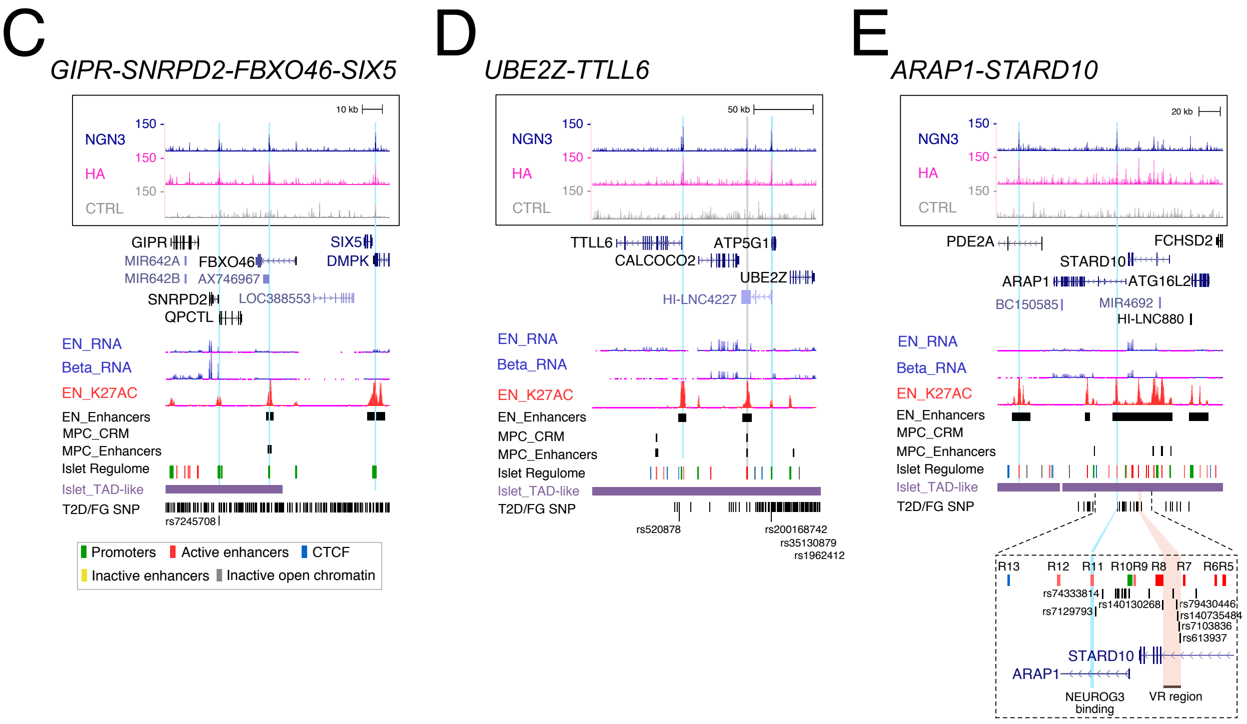
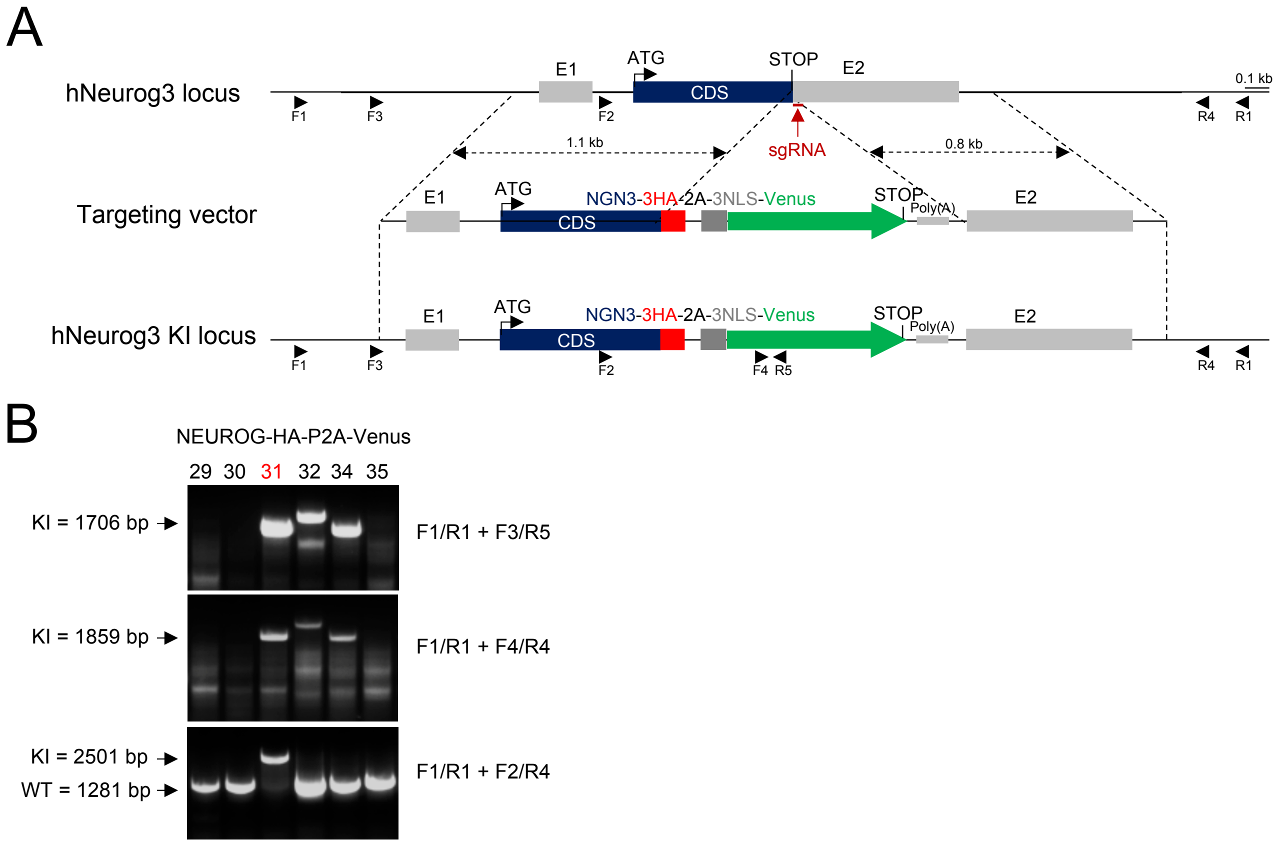
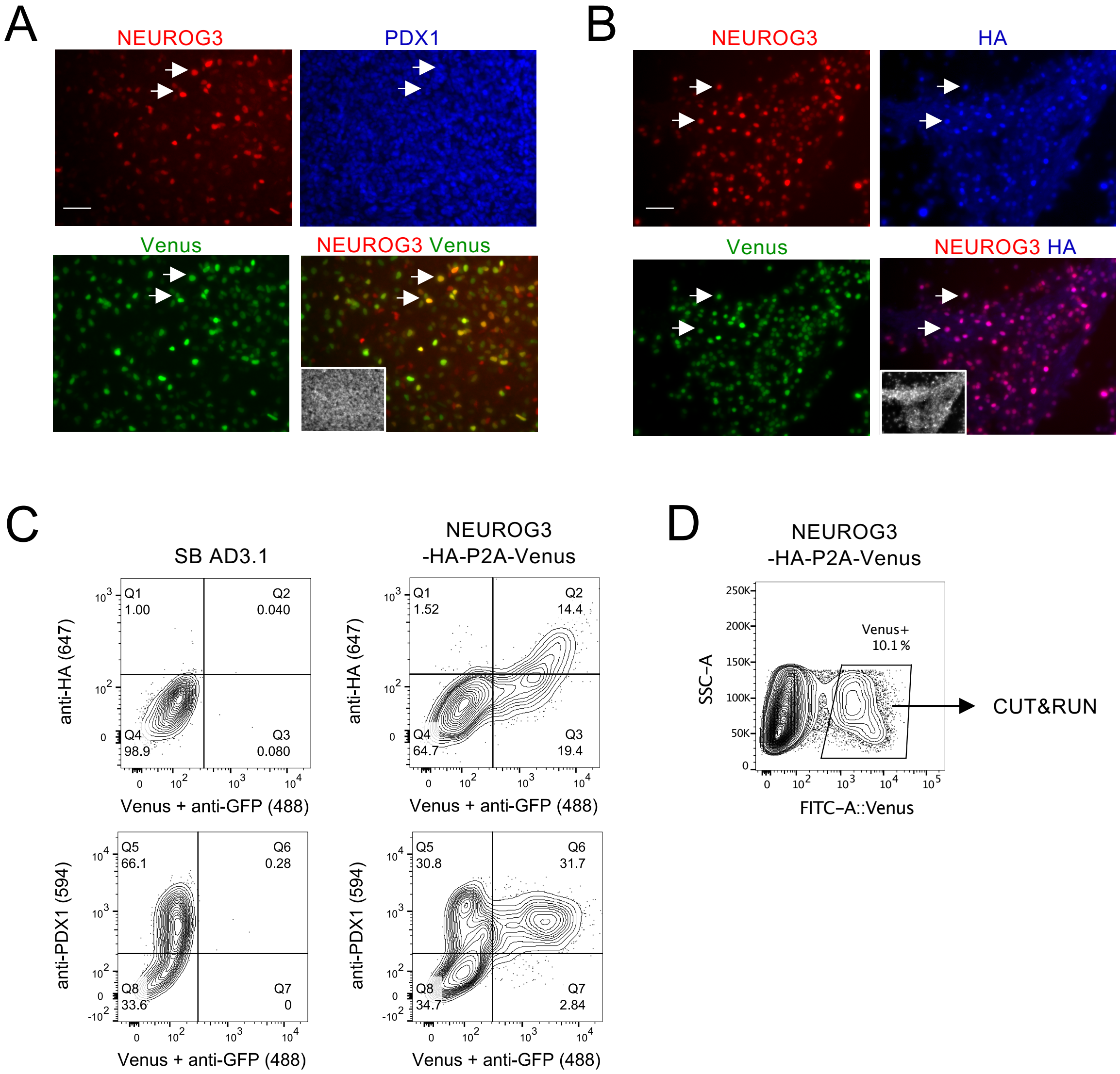


Figure 8

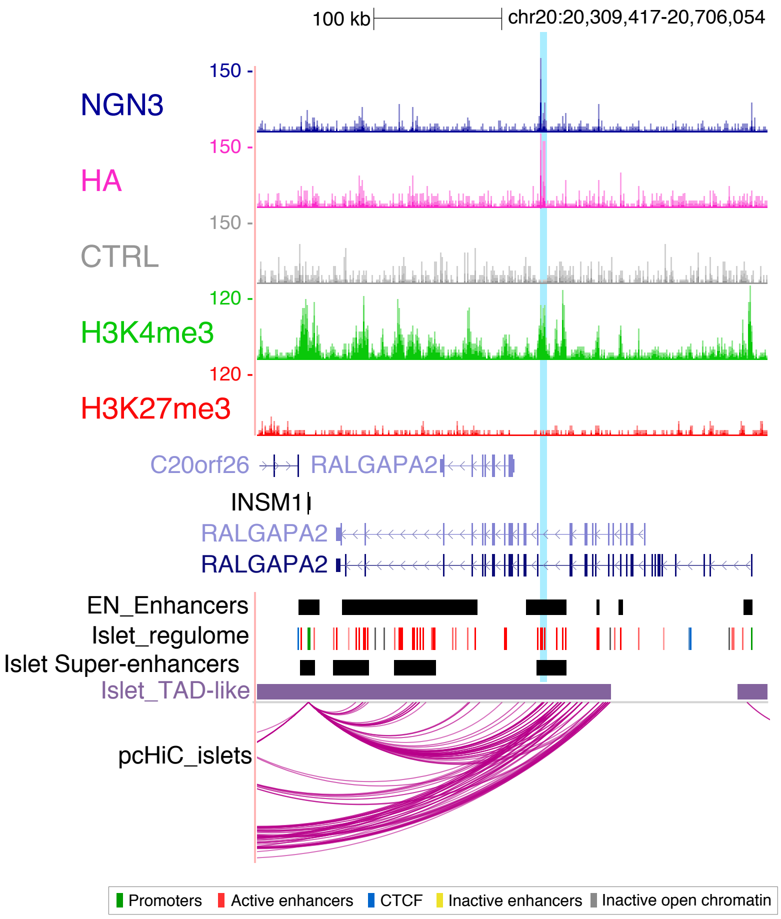


Supplemental Figure 1



Supplemental Figure 2

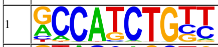
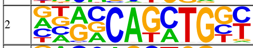
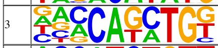
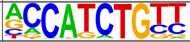
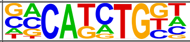
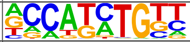
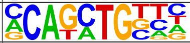
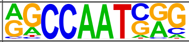
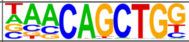
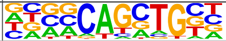
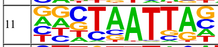
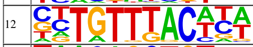
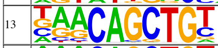
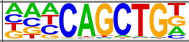
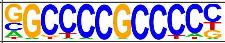
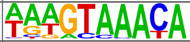
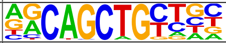
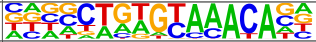
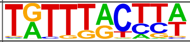
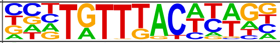
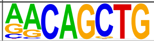
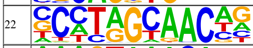
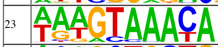
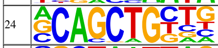
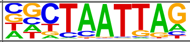
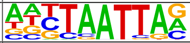
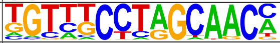
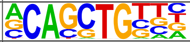
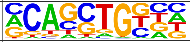
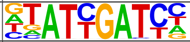
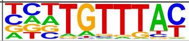
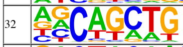
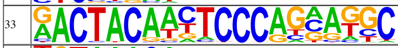
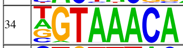
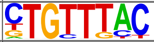
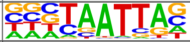
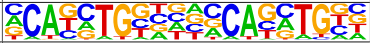
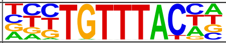
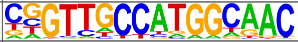
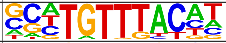
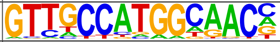
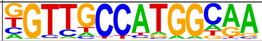
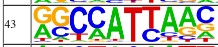
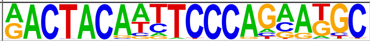
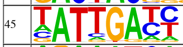
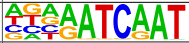
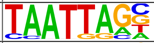
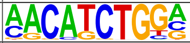
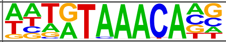
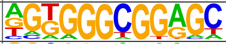
INSM1 - RALGAPA2 (hg19)



Supplemental Figure 3

Enriched known motifs within NEUROG3 peaks (entire peak sequences)

Total Target Sequences = 863, Total Background Sequences = 47348

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer	1e-146	-3.377e+02	0.0000	484.0	56.08%	7926.0	16.74%
2		Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	1e-128	-2.952e+02	0.0000	531.0	61.53%	10734.8	22.67%
3		BHLHA15(bHLH)/NIH3T3-BHLHB8.HA-ChIP-Seq(GSE119782)/Homer	1e-106	-2.442e+02	0.0000	556.0	64.43%	13288.7	28.06%
4		NeuroG2(bHLH)/Fibroblast-NeuroG2-ChIP-Seq(GSE75910)/Homer	1e-105	-2.434e+02	0.0000	569.0	65.93%	13931.3	29.42%
5		TCF4(bHLH)/SHSY5Y-TCF4-ChIP-Seq(GSE96915)/Homer	1e-93	-2.158e+02	0.0000	558.0	64.66%	14313.7	30.23%
6		Olig2(bHLH)/Neuron-Olig2-ChIP-Seq(GSE30882)/Homer	1e-91	-2.111e+02	0.0000	590.0	68.37%	16081.8	33.96%
7		Twist2(bHLH)/Myoblast-Twist2.Ty1-ChIP-Seq(GSE127998)/Homer	1e-86	-1.991e+02	0.0000	585.0	67.79%	16269.7	34.36%
8		NFY(CCAAT)/Promoter/Homer	1e-70	-1.633e+02	0.0000	334.0	38.70%	6538.7	13.81%
9		Tcf21(bHLH)/ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369)/Homer	1e-61	-1.416e+02	0.0000	398.0	46.12%	9721.3	20.53%
10		Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer	1e-60	-1.389e+02	0.0000	551.0	63.85%	17010.2	35.92%
11		En1(Homeobox)/SUM149-EN1-ChIP-Seq(GSE120957)/Homer	1e-46	-1.075e+02	0.0000	423.0	49.02%	12209.6	25.78%
12		Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-46	-1.065e+02	0.0000	244.0	28.27%	4933.6	10.42%
13		Myf5(bHLH)/GM-Myf5-ChIP-Seq(GSE24852)/Homer	1e-45	-1.046e+02	0.0000	296.0	34.30%	6923.3	14.62%
14		Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738)/Homer	1e-39	-9.036e+01	0.0000	398.0	46.12%	11881.5	25.09%
15		Sp1(Zf)/Promoter/Homer	1e-37	-8.741e+01	0.0000	211.0	24.45%	4353.8	9.19%
16		FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-37	-8.579e+01	0.0000	295.0	34.18%	7610.9	16.07%
17		MyoD(bHLH)/Myotube-MyoD-ChIP-Seq(GSE21614)/Homer	1e-36	-8.426e+01	0.0000	302.0	34.99%	7963.7	16.82%
18		Fox:Ebox(Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	1e-36	-8.388e+01	0.0000	271.0	31.40%	6722.3	14.20%
19		FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer	1e-35	-8.185e+01	0.0000	266.0	30.82%	6605.1	13.95%
20		Foxa3(Forkhead)/Liver-Foxa3-ChIP-Seq(GSE77670)/Homer	1e-35	-8.117e+01	0.0000	126.0	14.60%	1823.4	3.85%
21		MyoG(bHLH)/C2C12-MyoG-ChIP-Seq(GSE36024)/Homer	1e-35	-8.095e+01	0.0000	364.0	42.18%	10806.0	22.82%
22		Rfx5(HTH)/GM12878-Rfx5-ChIP-Seq(GSE31477)/Homer	1e-34	-8.028e+01	0.0000	153.0	17.73%	2634.7	5.56%
23		FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-33	-7.642e+01	0.0000	253.0	29.32%	6310.8	13.33%
24		Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465)/Homer	1e-32	-7.476e+01	0.0000	345.0	39.98%	10269.0	21.68%
25		LHX9(Homeobox)/Hct116-LHX9.V5-ChIP-Seq(GSE116822)/Homer	1e-30	-6.950e+01	0.0000	312.0	36.15%	9082.0	19.18%
26		Lhx3(Homeobox)/Neuron-Lhx3-ChIP-Seq(GSE31456)/Homer	1e-29	-6.693e+01	0.0000	334.0	38.70%	10193.5	21.53%
27		Rfx6(HTH)/Min6b1-Rfx6.HA-ChIP-Seq(GSE62844)/Homer	1e-27	-6.235e+01	0.0000	324.0	37.54%	9998.9	21.11%
28		Ptf1a(bHLH)/Panc1-Ptf1a-ChIP-Seq(GSE47459)/Homer	1e-23	-5.499e+01	0.0000	633.0	73.35%	26714.4	56.41%
29		HEB(bHLH)/mES-Heb-ChIP-Seq(GSE53233)/Homer	1e-22	-5.172e+01	0.0000	550.0	63.73%	22178.1	46.83%
30		HNF6(Homeobox)/Liver-Hnf6-ChIP-Seq(ERP000394)/Homer	1e-22	-5.163e+01	0.0000	148.0	17.15%	3316.0	7.00%
31		FOXK1(Forkhead)/HEK293-FOXK1-ChIP-Seq(GSE51673)/Homer	1e-21	-4.930e+01	0.0000	233.0	27.00%	6751.9	14.26%
32		SCL(bHLH)/HPC7-Scl-ChIP-Seq(GSE13511)/Homer	1e-21	-4.896e+01	0.0000	730.0	84.59%	33358.1	70.44%
33		Ronin(THAP)/ES-Thap11-ChIP-Seq(GSE51522)/Homer	1e-21	-4.892e+01	0.0000	50.0	5.79%	467.1	0.99%
34		Foxo3(Forkhead)/U2OS-Foxo3-ChIP-Seq(E-MTAB-2701)/Homer	1e-21	-4.870e+01	0.0000	181.0	20.97%	4668.8	9.86%
35		Foxo1(Forkhead)/RAW-Foxo1-ChIP-Seq(Fan_et_al.)/Homer	1e-20	-4.714e+01	0.0000	391.0	45.31%	14176.2	29.94%
36		DLX2(Homeobox)/BasalGanglia-Dlx2-ChIP-seq(GSE124936)/Homer	1e-19	-4.577e+01	0.0000	291.0	33.72%	9509.7	20.08%
37		Twist(bHLH)/HMLE-TWIST1-ChIP-Seq(Chang_et_al)/Homer	1e-19	-4.530e+01	0.0000	81.0	9.39%	1310.6	2.77%
38		FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-19	-4.480e+01	0.0000	124.0	14.37%	2710.1	5.72%
39		RFX(HTH)/K562-RFX3-ChIP-Seq(SRA012198)/Homer	1e-19	-4.411e+01	0.0000	58.0	6.72%	715.3	1.51%
40		FOXK2(Forkhead)/U2OS-FOXK2-ChIP-Seq(E-MTAB-2204)/Homer	1e-19	-4.380e+01	0.0000	168.0	19.47%	4383.4	9.26%
41		Rfx2(HTH)/LoVo-RFX2-ChIP-Seq(GSE49402)/Homer	1e-18	-4.344e+01	0.0000	60.0	6.95%	776.7	1.64%
42		Rfx1(HTH)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	1e-18	-4.201e+01	0.0000	91.0	10.54%	1697.5	3.58%
43		Nanog(Homeobox)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	1e-18	-4.174e+01	0.0000	647.0	74.97%	28699.0	60.60%
44		GFY-Staf(?,Zf)/Promoter/Homer	1e-17	-4.110e+01	0.0000	53.0	6.14%	642.8	1.36%
45		Hnf6b(Homeobox)/LNCaP-Hnf6b-ChIP-Seq(GSE106305)/Homer	1e-17	-4.107e+01	0.0000	193.0	22.36%	5524.8	11.67%
46		Cux2(Homeobox)/Liver-Cux2-ChIP-Seq(GSE35985)/Homer	1e-17	-4.009e+01	0.0000	119.0	13.79%	2704.9	5.71%
47		Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	1e-17	-3.965e+01	0.0000	212.0	24.57%	6402.7	13.52%
48		ZBTB18(Zf)/HEK293-ZBTB18.GFP-ChIP-Seq(GSE58341)/Homer	1e-17	-3.933e+01	0.0000	174.0	20.16%	4839.3	10.22%
49		FoxL2(Forkhead)/Ovary-FoxL2-ChIP-Seq(GSE60858)/Homer	1e-16	-3.842e+01	0.0000	185.0	21.44%	5338.0	11.27%
50		Sp5(Zf)/mES-Sp5.Flag-ChIP-Seq(GSE72989)/Homer	1e-16	-3.784e+01	0.0000	355.0	41.14%	13134.7	27.74%

Supplemental Figure 4