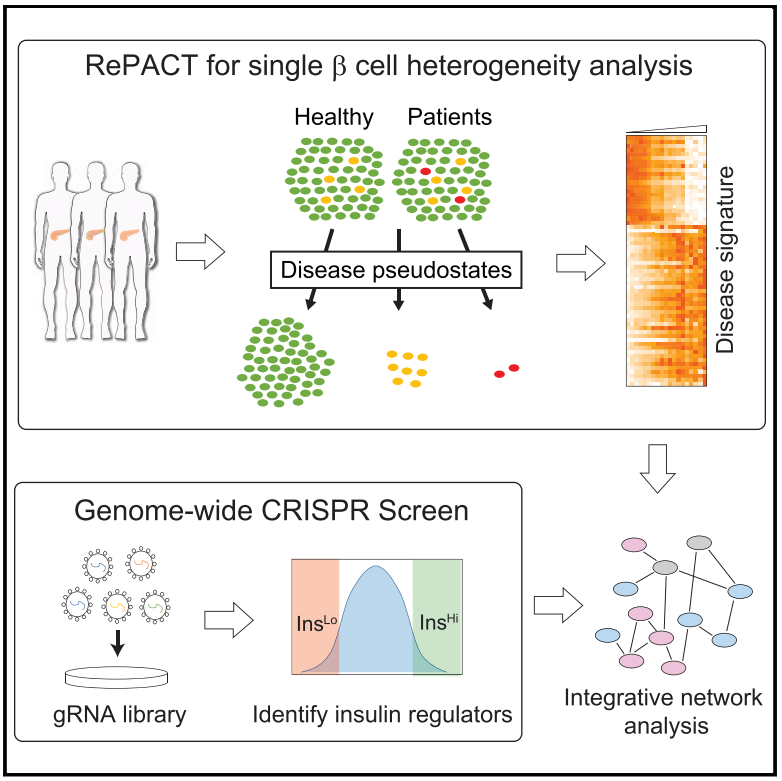
Resource

Cell Reports

Single-Cell Heterogeneity Analysis and CRISPR Screen Identify Key β-Cell-Specific Disease Genes

Graphical Abstract



Highlights

- Transcriptome of 39,905 single islet cells from healthy, obese, and T2D human donors
- Obesity and diabetes cause distinct heterodetic profiles in pancreatic β cells
- RePACT can identify disease signatures using single-cell data from very few donors
- Functional annotation of disease signature genes using genome-wide CRISPR analysis

Authors

Zhou Fang, Chen Weng, Haiyan Li, ..., Yanxin Pei, Fulai Jin, Yan Li

Correspondence

fxj45@case.edu (F.J.),
yxl1379@case.edu (Y.L.)

In Brief

Fang et al. found that β cells from healthy, obese, and diabetic donors have a distinct cellular heterogeneity pattern, which allows sensitive identification of disease signature genes from a small number of donors. Combined with results from a genome-wide CRISPR screen, they further annotated signature genes with insulin regulatory functions.

Fang et al., 2019, Cell Reports *26*, 3132–3144
March 12, 2019
https://doi.org/10.1016/j.celrep.2019.02.043

CellPress

OPEN ACCESS
CellPress

Cell Reports
Resource

Single-Cell Heterogeneity Analysis and CRISPR Screen Identify Key β-Cell-Specific Disease Genes

Zhou Fang,[1,8] Chen Weng,[1,8] Haiyan Li,[1] Ran Tao,[2] Weihua Mai,[1,3] Xiaoxiao Liu,[1] Leina Lu,[1] Sisi Lai,[1] Qing Duan,[4] Carlos Alvarez,[1,5] Peter Arvan,[6] Anthony Wynshaw-Boris,[1] Yun Li,[4] Yanxin Pei,[2] Fulai Jin,[1,7,*] and Yan Li[1,9,*]

[1]Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA
[2]Center for Cancer and Immunology Research, Brain Tumor Institute, Children's National Medical Center, Washington, D.C. 20010, USA
[3]Department of Neurology, the Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, Guangdong Province 519000, China
[4]Department of Biostatistics, Department of Genetics, Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA
[5]The Biomedical Sciences Training Program (BSTP), School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA
[6]Division of Metabolism, Endocrinology, and Diabetes, University of Michigan Medical Center, Ann Arbor, MI 48109, USA
[7]Department of Population and Quantitative Health Sciences, Department of Electrical Engineering and Computer Science, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH 44106, USA
[8]These authors contributed equally
[9]Lead Contact
*Correspondence: fxj45@case.edu (F.J.), yxl1379@case.edu (Y.L.)
https://doi.org/10.1016/j.celrep.2019.02.043

SUMMARY

Identification of human disease signature genes typically requires samples from many donors to achieve statistical significance. Here, we show that single-cell heterogeneity analysis may overcome this hurdle by significantly improving the test sensitivity. We analyzed the transcriptome of 39,905 single islets cells from 9 donors and observed distinct β cell heterogeneity trajectories associated with obesity or type 2 diabetes (T2D). We therefore developed RePACT, a sensitive single-cell analysis algorithm to identify both common and specific signature genes for obesity and T2D. We mapped both β-cell-specific genes and disease signature genes to the insulin regulatory network identified from a genome-wide CRISPR screen. Our integrative analysis discovered the previously unrecognized roles of the cohesin loading complex and the NuA4/Tip60 histone acetyltransferase complex in regulating insulin transcription and release. Our study demonstrated the power of combining single-cell heterogeneity analysis and functional genomics to dissect the etiology of complex diseases.

INTRODUCTION

Pancreatic islets provide the endocrine function of the pancreas and are comprised of at least five hormone-producing cell types: α cells (secreting glucagon, *GCG*), β cells (insulin, *INS*), γ/PP cells (pancreatic polypeptide, *PPY*), δ cells (somatostatin, *SST*), and ε cells (ghrelin, *GHRL*). Malfunction of pancreatic islets, especially β cells, is associated with obesity and type 2 diabetes (T2D). However, because obesity and T2D are two highly related diseases (i.e., many patients have both diseases), understanding the commonality and differences between the two diseases at cellular level is challenging. Conventional transcriptome analysis requires pure endocrine cell subpopulations from a large number of patients to achieve statistical significance, which can be prohibitively difficult.

Single-cell RNA sequencing (scRNA-seq) technologies allow transcriptome profiling in individual cells and are revolutionizing the analyses of rare or complex tissues, including pancreatic islets (Baron et al., 2016; Dorajoo et al., 2017; Lawlor et al., 2017; Li et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Wang et al., 2016; Xin et al., 2016). However, although these previous studies have clearly demonstrated the great potential of scRNA-seq in islet biology, limitations remain. First, most of these studies only analyzed up to a few hundred cells from each donor, therefore had low sensitivity mapping rare cell subpopulations. Second, due to the limited availability of human islet samples, especially from diabetic patients, the sensitivity of detecting disease relevant gene signatures is low. Finally, the cellular functions of identified signature genes remain to be validated.

In this study, we used Drop-Seq (Macosko et al., 2015) to generate massively parallel single-cell transcriptome data from thousands of islet cells. The improved throughput allowed us to map the heterogeneity of endocrine cell subpopulations sensitively. Importantly, even though there is no α or β cell subpopulation correlated with obesity or T2D, we observed obvious trajectories of continuous cellular heterogeneity associated with diseases. We therefore developed a general single-cell analysis algorithm named RePACT (regressing principle components for the alignment of continuous trajectory) and demonstrated that it is feasible to identify both common and specific signature genes associated with obesity and T2D with a limited supply of human islets. Additionally, we also used an unbiased genome-wide

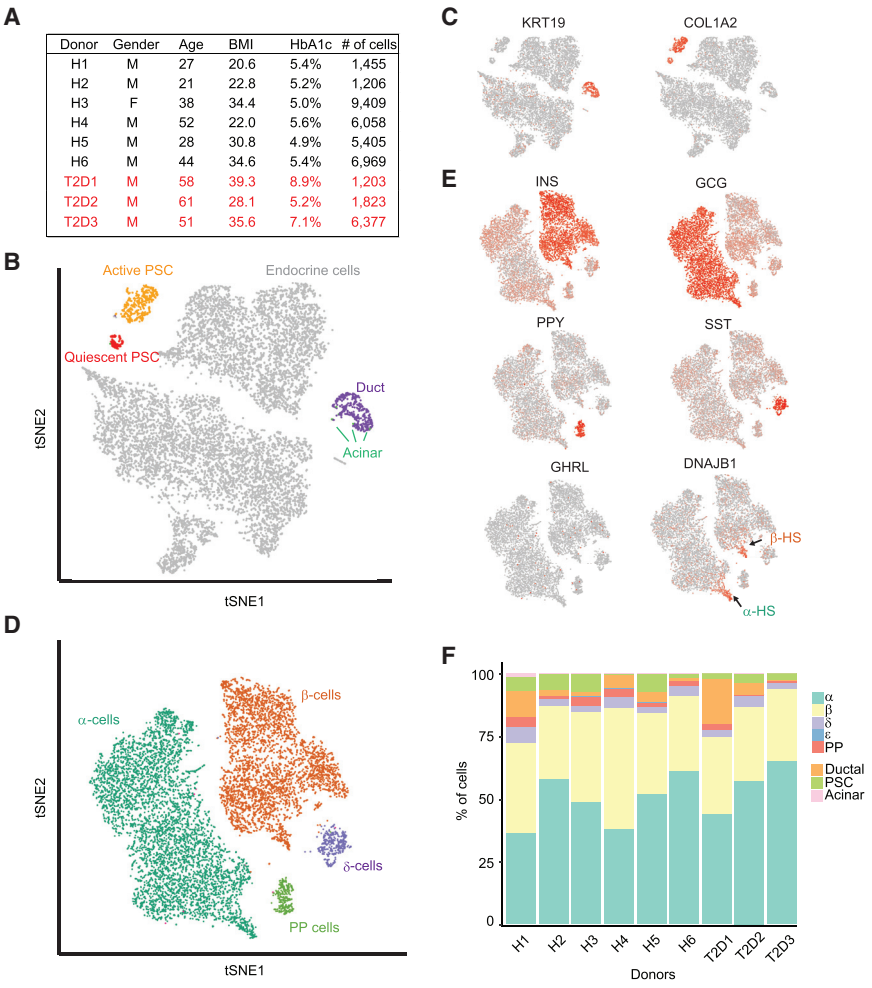


Figure 1. Single Islet Cell Transcriptomes Generated by Drop-Seq

(A) Table of donor information.

(B) Two-dimensional t-SNE plot of the top 11,697 STAMPs with non-endocrine cells highlighted in color.

(C) Expression levels of *KRT19* (duct marker) and *COL1A2* (PSC marker) were overlaid onto the t-SNE plot in (B).

(D) Two-dimensional t-SNE plot of distinct endocrine cell types.

(E) Expression levels of endocrine cell markers and DNAJB1 are overlaid onto the t-SNE plot in (D).

(F) Bar graphs demonstrating the percentage of all cell types in each donor.

CRISPR screen to validate the insulin regulatory functions of disease signature genes. Our integrative analyses uncovered previously uncharacterized β-cell-specific insulin regulator modules, including the *MAU2-NIPBL* cohesin loading complex, and the NuA4/TIP60 histone acetyltransferase (HAT) complex. Taken together, our study provides a general strategy for systematically characterizing disease genes in pancreatic islets as well as other complex tissues.

RESULTS

Drop-Seq Analysis of Human Islet Samples

We prepared Drop-Seq libraries with fresh human islet samples from 6 healthy (3 overweighed with BMI >30) and 3 T2D donors (2 overweighed). In total, we obtained transcriptome data from 39,905 single cells (1,206–9,409 cells from each donor, Figure 1A) and used a very stringent clustering-based analysis pipeline to determine the types of 28,026 "clean" cells without ambiguity (Figure S1; Data S1). When projecting the cells to a two-dimensional t-distributed stochastic neighbor embedding (tSNE) plot, we observed a clear distinction between endocrine cells and a few non-endocrine cell types, mainly pancreatic ductal cells (PDCs) marked by several keratin genes (KRTs), and pancreatic stellate cells (PSCs) marked by collagen genes (Figures 1B and 1C). We observed very few acinar cells marked by *REG1A* and *PRSS1* genes, which were identified as PCA outliers but failed to form a distinct cluster in t-SNE due to the scarcity (n = 108, Figures 2A–2D). We further performed a second-round unsupervised clustering with the endocrine cells and distinguished four major endocrine clusters, which are recognized as α, β, δ, and PP cells based on the enrichment of corresponding marker genes (Figures 1D and 1E). We could not observe a distinct cluster of ε cells in tSNE due to the extreme scarcity of this cell type in our samples: only 13 of the 28,026 "clean" cells express the ε cell hormone gene *GHRL* (Figures 2A–2D). Taken together, all of the samples contain 10%–20% non-endocrine cells (Figure 1F), consistent with an estimated 80%–90% islet purity, and ~90% of endocrine cells in every donor are α or β cells (Figure 1F).

Gene Signatures of Non-endocrine Cell Types

We first used a negative binomial model to define the non-endocrine cell marker genes (STAR Methods), including a number of transcription factors (TFs) that may function as master cell type regulators (Figures 2E and 2F; a complete gene list is included in Data S2). As expected, PSCs express collagen genes; ductal cells express keratins, *CFTR*, and *TSPAN8*; and exocrine acinar cells express *REG1A*, *REG1B*, *REG3A*, and a number of digestive enzymes (Figures 2E and 2G). Interestingly, we also observed duct-exclusive expression of inflammation genes such as *CCL2*, *CXCL2*, *MMP7*, and *DEFB1* (Figures 2E and 2G). It is not clear whether these duct-expressed inflammatory genes may contribute to disease initiation or development, although the elevated levels of cytokines in pancreatitis and pancreatic cancer is well documented.

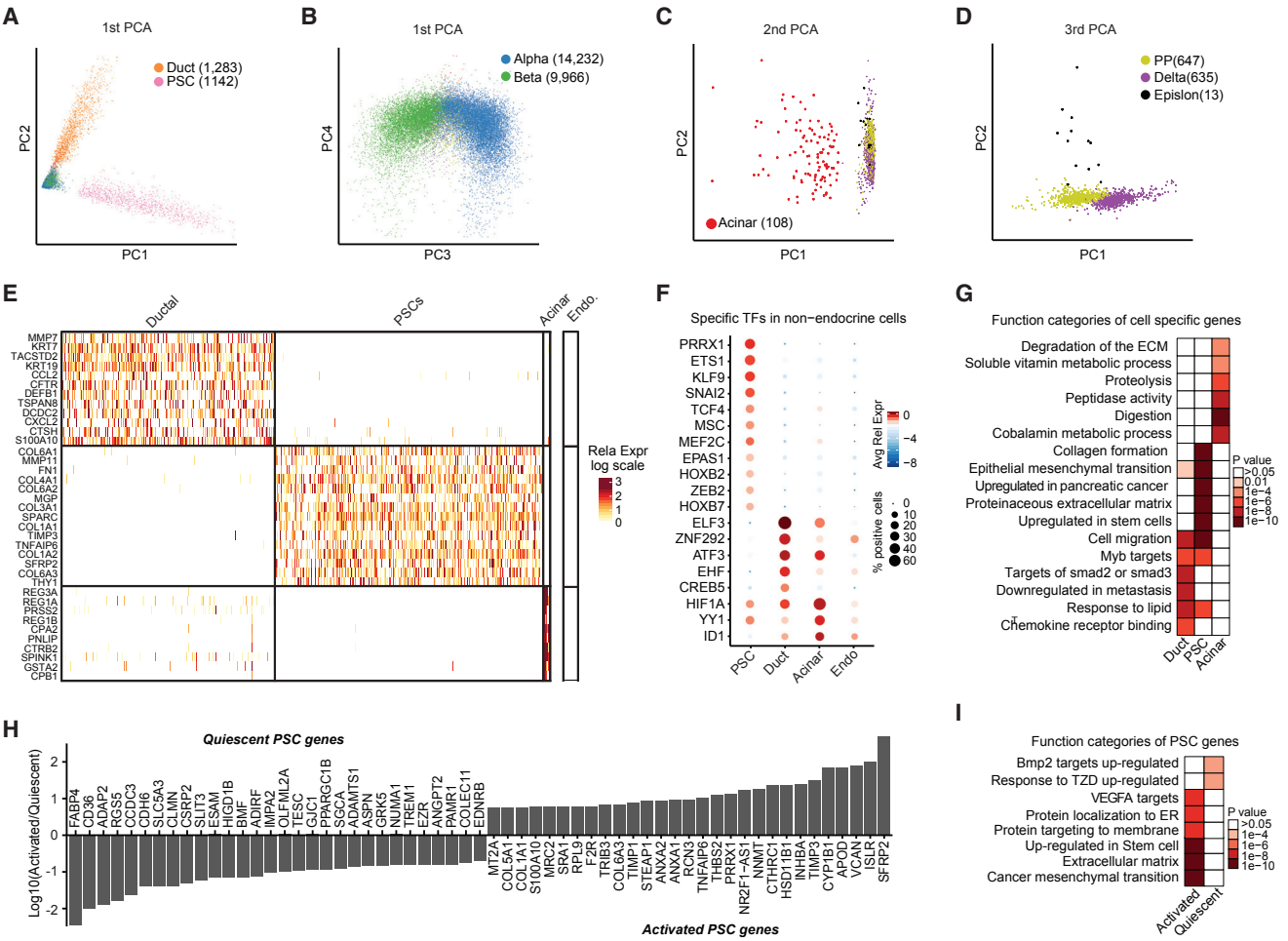


Figure 2. Non-endocrine Cell Populations and Their Marker Genes

(A–D) PCA analysis of all STAMPs with unique cell type assignment. (A) After 1st round PCA, PC1, and PC2 distinguish ductal cells and PSCs. (B) PC3 and PC4 distinguish α and β cells (ductal cells and PSCs are masked). (C) PC1-PC2 in 2nd round PCA (after removing ductal cells, PSCs, α and β cells) distinguished the acinar cells. (D) A 3rd round PCA can further distinguish ε, δ, and PP populations after removing acinar cells.

(E) Heatmaps demonstrating the non-endocrine cell marker genes. The rightmost column shows the average expression of all endocrine cells.

(F) Bubble plot showing the expression patterns of top non-endocrine cell-type-specific TFs. The size of each bubble indicates the percentage of single cells with detectable transcripts. The color indicates one gene's average transcripts number in the cell population.

(G) Gene Set Enrichment Analysis (GSEA) of each set of non-endocrine cell-type-specific genes.

(H) Bar graph showing the top genes specifically expressed in quiescent versus activated PSCs.

(I) GSEA results of activated PSC-specific genes.

We observed two clusters of PSCs representing activated and quiescent populations; ~70% of PSCs are activated (Figure 1B). Both populations contain cells from multiple donors (comparing Figures 1B with S2A). It is known that under quiescent state, the PSCs contain abundant lipid droplets in the cytoplasm, and pancreatic injury or other pathogenic insults can activate PSCs, leading to increased migration and proliferation, more production of extracellular matrix (ECM) proteins, metalloproteinase (MMPs), and tissue inhibitors of metalloproteinase (TIMPs) (Omary et al., 2007). Indeed, we found that the activated PSCs expressed more collagen, TIMPs, and stem cell or cancer signature genes but lost the expression of lipid-processing or adipogenesis genes such as *FABP4* and *ADIRF* (Figures 2H and 2I).

Gene Signatures of Endocrine Cell Types

We next defined hundreds of genes specifically expressed in various endocrine cell types (Figure 3A; Data S2). As expected, β-cell-specific genes are associated with insulin secretion and diabetes (Figure 3B). We identified a number of endocrine-specific transcription factors with dramatic specificity between different endocrine cell types (Figure 3C). For example, *PAX6*, *NEUROD1*, and *INSM1* are expressed in all four, and *MAFB* is expressed in three endocrine cell types. Our top list includes many classically known cell identity master transcription factors, such as *PDX1*, *MAFA*, and *NKX6-1* in β cells, *ARX* and *IRX2* in α cells, and *HHEX* in δ cells (Figure 3C). Consistent with a few other single-cell results (Baron

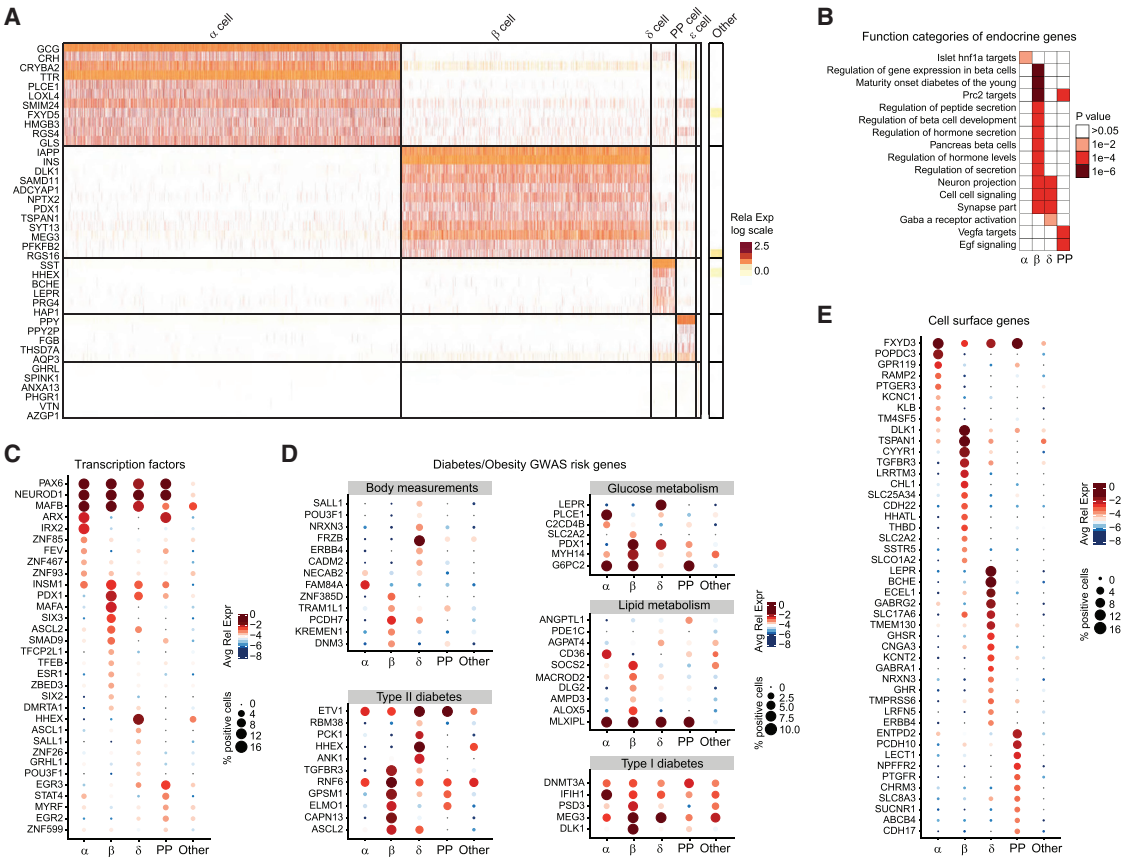


Figure 3. Endocrine Cell-Type-Specific Genes

(A) Heatmaps demonstrating the endocrine cell marker genes. The column on the right shows the average expression of all non-endocrine cells.

(B) GSEA results of four major cell-type-specific genes. The ε cell was not included in the analysis due to its extreme scarcity.

(C–E) Bubble plots demonstrating examples of endocrine cell-type-specific TFs (C), endocrine cell-type-specific GWAS risk genes (D), and endocrine cell-type-specific cell surface proteins (E).

et al., 2016; Lawlor et al., 2017; Segerstolpe et al., 2016), we also observed dual specificity of *PDX1* (β and δ) and *ARX* (α and PP). Some signaling-dependent TFs are also cell-type-specific, such as *ESR1* (β cell) and *EGR2*, *EGR3*, and *STAT4* (PP cell) (Figure 3C). These signal-dependent TFs may have important physiological functions. For example, studies in both human and mouse have shown that the activation of estrogen receptor α (*ESR1*) protects β cells from apoptosis and preserves functional β cell mass in diabetes (Tiano and Mauvais-Jarvis, 2012).

Diabetes and obesity are two complex multigene disorders contributed by many tissues. We collected over 1,000 GWAS risk genes (Data S3) and reasoned that knowing their expression pattern in endocrine cells, especially β cells, would be helpful for the understanding the disease etiology. We found 163 GWAS risk genes specifically expressed in one or more endocrine cell types (Figure 3D; Data S3) including some well-known β cell genes *RNF6*, *PDX1*, *SLC2A2*, *MEG3*, and *DLK1*. Interestingly, a number of risk genes are δ-cell-specific, including *HHEX*, *LEPR*, *ERBB4*, etc., suggesting a particularly important role of δ cells to diabetes or obesity (Figure 3D).

Last, we examined the endocrine-specific cell surface genes due to their potential as cellular markers, or targets for pharmacological intervention (Figure 3E; Data S3). Interestingly, we noticed that δ cells specifically express several important hormone receptor genes, including *LEPR* (receptor of leptin), *GHSR* (receptor of Ghrelin), *GHR* (receptor of growth hormone), *ERBB4* (receptor of EGF), and two GABA receptors (*GABRG2* and *GABRA1*) (Figure 3E). Receptor activity is actually one of the top functional categories for the δ-cell-specific genes (Figure 3B). These results suggested a key regulatory role of δ cells by integrating multiple cell signaling (Segerstolpe et al., 2016).

Identifying α or β Cell Subpopulations

Previous studies, including several recent single-cell RNA-seq analyses, have investigated β cell heterogeneity but reached discrepant conclusions regarding the existence of β cell subpopulations, partly due to the low cell numbers analyzed (Gutierrez et al., 2017). We posited that Drop-Seq would be more sensitive in resolving distinct cell populations with greatly improved throughput. In this study, we pooled α or β cells from all donors and used principle component analysis (PCA) to identify cell

subpopulations as outliers (Figures S2B and S2C) then projected the cells onto t-SNE plots to confirm their presence in multiple donors (Figures S2D and S2E). Marker genes for these subpopulations are listed in Data S2.

The most obvious α cell subpopulation consists of a small number of proliferating α cell expressing *TOPA1*, *CENPF*, and *AURKB* (Figures 2F and S2B). Importantly, proliferating α cell is a reproducible population that can be found from multiple donors including H1, H3, H4, H5, H6, and T2D3 (Figure S2E, cells in dark green). Notably, the proliferating α cells were also reported in another recent single-cell study but with low sensitivity due to limited cell throughput (Segerstolpe et al., 2016). Our study has therefore confirmed the existence of this rare α cell population, which may also play an important role in β cell replenishment (Thorel et al., 2010).

The biggest subpopulations are α-HS and β-HS cells expressing the same set of heat shock genes including *DNAJA1*, *DNAJB1*, *HSPA*s, and *HSPB*s, etc. (Figures 1E, 2F, 2G, and S2B–S2E). However, both α-HS and β-HS cells are mainly found in donor H3 and T2D3 (Figure S2E). Many types of stresses can activate heat shock genes (HSPs). Reduced expression of HSPs is associated with diabetes, and upregulation of HSPs may provide a cytoprotective effect to β cells (Hooper and Hooper, 2009). Interestingly, a recent single-cell study also observed a correlation between stress gene expression and aging (Enge et al., 2017). Last, we also identified two other minor cell clusters (α-KCNQ1OT1 and β-KCNQ1OT1), both of which were exclusive to donor H5 (Figure S2). More donors are necessary to evaluate the physiological relevance of these individual-specific subpopulations.

Importantly, we did not find any of these subpopulations correlate with the BMI or T2D status of the donors, leading to a conclusion that disease-associated effect on α or β cells does not create distinctive cell subpopulations. Therefore, we excluded the minor cell populations in the following disease association analysis and focused on the continuous changes of α or β transcriptome associated with disease status.

Obesity and T2D Cause Different Single-Cell Transcriptome Heterogeneity

We next investigated the molecular differences between cells from healthy, obese, and T2D donors. A few recent single-cell studies have explored this question (Lawlor et al., 2017; Segerstolpe et al., 2016; Xin et al., 2016). These studies used conventional statistical models comparing cells from normal and patients and looked for differentially expressed genes correlated with disease states. However, these models do not account for cellular heterogeneity and treat all cells from the same donor equally, which may result in low sensitivity for the following reasons: (1) disease-causing cells can be too rare to be detected when a majority of cells are normal (Figure 4A); and (2) the difference between normal and patient cells can be small and masked by large individual variance. In both scenarios, many human samples become necessary to increase the statistical power.

Here, we propose a strategy to improve the sensitivity to identify disease signature genes by dissecting disease-associated single-cell heterogeneity. Our strategy is based on a fundamentally different concept assuming the presence of disease relevant cellular heterogeneity from the same donor. Specifically, obesity or T2D may cause different transcriptome heterogeneity at single-cell level. Theoretically, by ranking and directly comparing cells at normal, transitional, and disease states, it is possible to improve the sensitivity even if very few donors are available (Figure 4A). Figure 4B shows scatterplots of all the β cells in the space of top three principle components (PCs). Remarkably, although cells from different donors do not segregate in the PCA plots, there is a clear continuous shift correlated with BMI or T2D status. Importantly, the two disease conditions show different trajectories, indicating different effects on the transcriptome.

RePACT: A Sensitive Approach to Identify Disease-Relevant Gene Signatures from Single-Cell Data

We therefore developed RePACT (regressing principle components for the assembly of continuous trajectory), to identify genes associated with disease relevant cellular heterogeneity. In this algorithm, we first performed PCA to reduce the dimension of transcriptome data. Next, we used regression analysis to draw two optimal trajectory lines reflecting the obesity- or T2D-relevant variation. Figure 4B is an example when we only used the top three PCs to plot the trajectory lines. In this study, we actually used top 10 PCs as predictors (STAR Methods). The numeric projection of each cell on the obesity or T2D trajectory (BMI index and T2D index) served as a measurement of the degree to which the cell has transformed during disease development. We then binned the cells into a number of pseudo-states according to the index values (Figure 4C). By comparing cells from different pseudo-states, RePACT greatly improved the statistical power to identify gene signatures for obesity or T2D status in α or β cell (Figures 4D and S3A–S3C; STAR Methods). To evaluate the robustness of RePACT method, we called top 200 T2D signature genes in α or β cells using data from all three T2D donors and then compared them to the results when only two T2D donors were included (dropout analyses). We found that 91%–95% of the T2D signature genes can be recovered from at least two of the three dropout analyses (Figures S3D–S3E), proving that the signature genes identified from RePACT were highly reproducible.

RePACT Identifies Common and Specific β Cell Gene Signatures Associated with Obesity and T2D

We identified 1,368 T2D trajectory genes and 1,188 obesity trajectory genes in β cells (Figure 4D; Data S4). It is known that obesity increases diabetes risk, and as expected, the two signatures shared many common genes (Figure 4E). For example, *GAPDH*, *IAPP*, *SPP1*, and CPE are downregulated in both obesity and T2D trajectories (Figure 4D). *GAPDH* is a key enzyme for glycolysis and glucose metabolism. *IAPP*, *SPP1*, and *CPE* are all well-known diabetes or obesity risk genes. Particularly, *IAPP* forms islet amyloid, which is linked to cellular toxicity in T2D (Clark and Nilsson, 2004); *SPP1* mediates the obesity-induced macrophage infiltration into adipose tissue and insulin resistance (Nomiyama et al., 2007); and *CPE* is important for proinsulin processing and its mutation leads to obesity and hyperproinsulinemia in mouse (Naggert et al., 1995). With the single-cell trajectory data, we also determined the transcription dynamics of 149

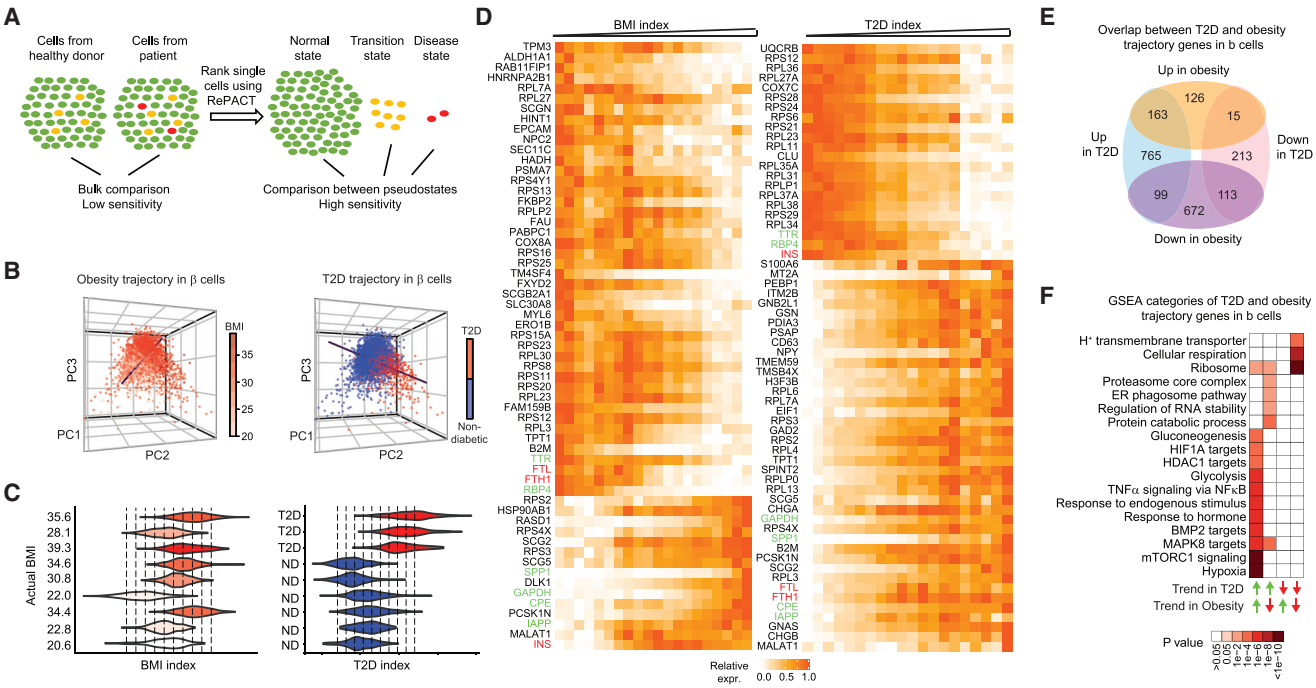


Figure 4. A Sensitive RePACT Algorithm to Identify T2D or Obesity Signature Genes in β Cells

(A) Schematic demonstrating how RePACT may improve the sensitivity identifying disease relevant genes with small sample size.

(B) All β cells are plotted in the 3D space of the top 3 principle components. Left: cells are colored based on the BMI of donor. Right: cells were colored based on whether the donor was a T2D patient.

(C) Left: comparison of the BMI index values of cells from each donor; the color of the violin plots represents the actual BMI of each donor. Right: comparison of the T2D index distribution of each donor. The color of the plots represents the T2D status of the donors. Vertical dash lines demonstrate how cells are binned into pseudostates.

(D) Heatmaps demonstrating the top obesity trajectory genes (left) and T2D trajectory genes (right). Each row in the heatmaps represent the transcriptional changes from low-index pseudostates to high-index pseudostates.

(E) Venn diagram showing the overlap between obesity and T2D signature genes identified from RePACT analyses.

(F) Function categories enriched among genes with agreeing or opposite tends.

GWAS risk genes (Figure S4A), which may help the dissection of etiology of obesity or T2D.

Most interestingly, we found many genes specific to one trajectory but not the other, including genes with opposite trends in the two trajectories (Figure 4E). The best example is probably the insulin gene (*INS*) itself, which is upregulated in obesity (consistent with the hyperinsulinemia in obesity) but downregulated in T2D (consistent with the β cell dysfunction in T2D) (Marchetti et al., 2008; Templeman et al., 2017). On the contrary, two ferritin genes (*FTL* and *FTH1*) are downregulated in obesity but upregulated in T2D (Figure 4D). Ferritin is the major intracellular iron storage protein. Clinical results have shown that low-serum iron concentration is associated with obesity (Lecube et al., 2006; Nead et al., 2004), while iron overload is a risk factor for T2D (Simcox and McClain, 2013).

Another interesting observation is that although only 15 obesity-upregulated genes are downregulated in T2D (including *INS*), many more (99) obesity-downregulated genes shift their trends in to upregulation in T2D (Figure 4E). We therefore performed a gene set enrichment analysis (GSEA) on genes with common or opposite trends in obesity or T2D (Figure 4F). For example, consistent with the connection to hypoxia in both T2D and BMI (Ye, 2009), we found upregulation of hypoxia, glycolysis, and HIF1A target genes, but downregulation of aerobic respiration pathways in both trajectories (Figure 4F). Surprisingly, the proteasome genes or pathway are elevated in T2D but downregulated in obesity trajectory (Figure 4F). Polymorphisms at proteasome genes have been associated to obesity (Kupca et al., 2013), and proteasome activity is weakened in obese liver that can induce endoplasmic reticulum (ER) stress and insulin resistance (Otoda et al., 2013). However, a recent study showed that in β cells, the inhibition of proteasome activity improves insulin production (Weisberg et al., 2016). Our results therefore suggest an important mechanism that is differentially presented in obesity and T2D governing β cell function.

A Genome-Wide Gene Deletion Analysis to Identify Insulin Regulators

The primary functions of β cell are to produce, store, and secrete insulin. Although we have identified a number of signature genes in β cells with Drop-Seq, the remaining question is whether these genes affect β cell functions. To address this problem, we decided to perform an unbiased genome-wide CRISPR screen

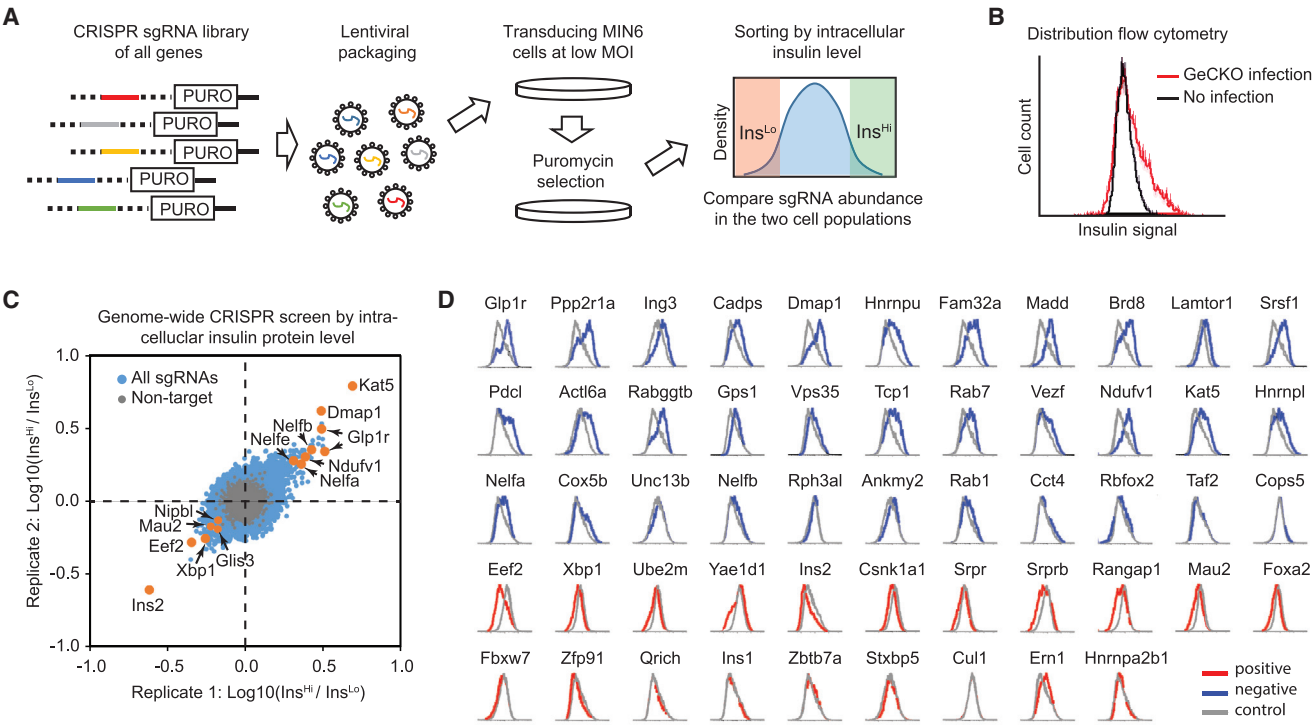


Figure 5. Genome-wide CRISPR Screen of Insulin Regulators

(A) Schematic of genome-wide CRISPR screening in the MIN6 cells.

(B) Distribution of intracellular insulin intensity before and after GeCKO viral library infection.

(C) Scatterplot showing the enrichment of sgRNAs in two replicated CRISPR screens. Top hits are highlighted in orange.

(D) Validation of top insulin regulators with individual cloned sgRNAs. Grey lines, cells with control sgRNAs; blue, sgRNAs of negative regulators; red, sgRNAs of positive regulators.

to identify insulin regulators. We choose to use the MIN6 insulinoma cell because it is one of the most robust β cell lines for the study of glucose sensing and insulin secretion (Skelin et al., 2010). Although MIN6 is a mouse cell line, and the functions of screen hits eventually need to be test in native human islets, we reasoned that most of the mouse insulin-regulatory genes should have conserved functions with their human orthologs. Therefore, the genome-wide mouse CRISPR data should still serve the purpose of providing an additional layer of information about gene functions.

Briefly, we packaged a lentiviral library using the mouse GeCKOv2 single guide RNA (sgRNA) library that contains ~130 K sgRNAs (6 sgRNAs per protein-coding gene in the mouse genome), and transduced >100 million MIN6 cells at low MOI so that most cells were infected with only one virus. After 7 days of puromycin selection expansion of transduced cells, we fixed and permeabilized the cells and sorted ~200 million cells based on the intracellular insulin protein level (Figure 5A; STAR Methods). Flow cytometry analysis demonstrated an increased variation of the insulin signal among the infected cells, suggesting that the screen has successfully targeted both positive and negative insulin regulators (Figure 5B). We collected top (Ins^{Hi}) and bottom (Ins^{Lo}) 10% cells and amplified the sgRNA sequences from integrated viral DNA for next generation sequencing. By comparing the abundance of each sgRNA in the two cell populations, we can determine the effect of target genes on insulin level in the cells.

We performed two independent screens and called hit sgRNAs only if they are reproducible between two replicates. Based on their significance and reproducibility, (Figure 5C; STAR Methods), we classified hit sgRNAs into five tiers (tier one sgRNAs have $p < 0.001$ in both replicates, false discover rate [FDR] = 0.1) and called 373 hit genes with at least one tier-one sgRNA. The design of sgRNA redundancy (6 sgRNAs per gene) also allowed us to evaluate the off-target risk for all hit genes from the screen. We found that most tier-one hit genes (223 of 373, or 59.8%) are supported by two or more sgRNAs (Data S5). However, the more likely reason for a sgRNA to fail the significance test is its under-representation in the GeCKO sgRNA library: the concentration of most abundant sgRNAs can be over 100-fold higher than the low-abundance ones (Figure S5E). Therefore, in this study, we choose to report all tier 1 genes, but also divide tier 1 hit genes into sub tiers to indicate if they are supported by multiple sgRNAs: 1A (119 genes, ≥3 supporting sgRNAs, FDR = 0.001), 1B (104 genes, 2 supporting sgRNAs, FDR = 0.038), and 1C (150 genes, 1 supporting sgRNA, FDR = 0.28). Note that there is a significant off-target risk for Tier 1C genes because they only have one supporting sgRNA (STAR Methods). Proteins from the same complexes, such as *Nelfa-Nelfb-Nelfe* and *Mau2-Nipbl*, were often identified together, suggesting that our CRISPR screen is sensitive and

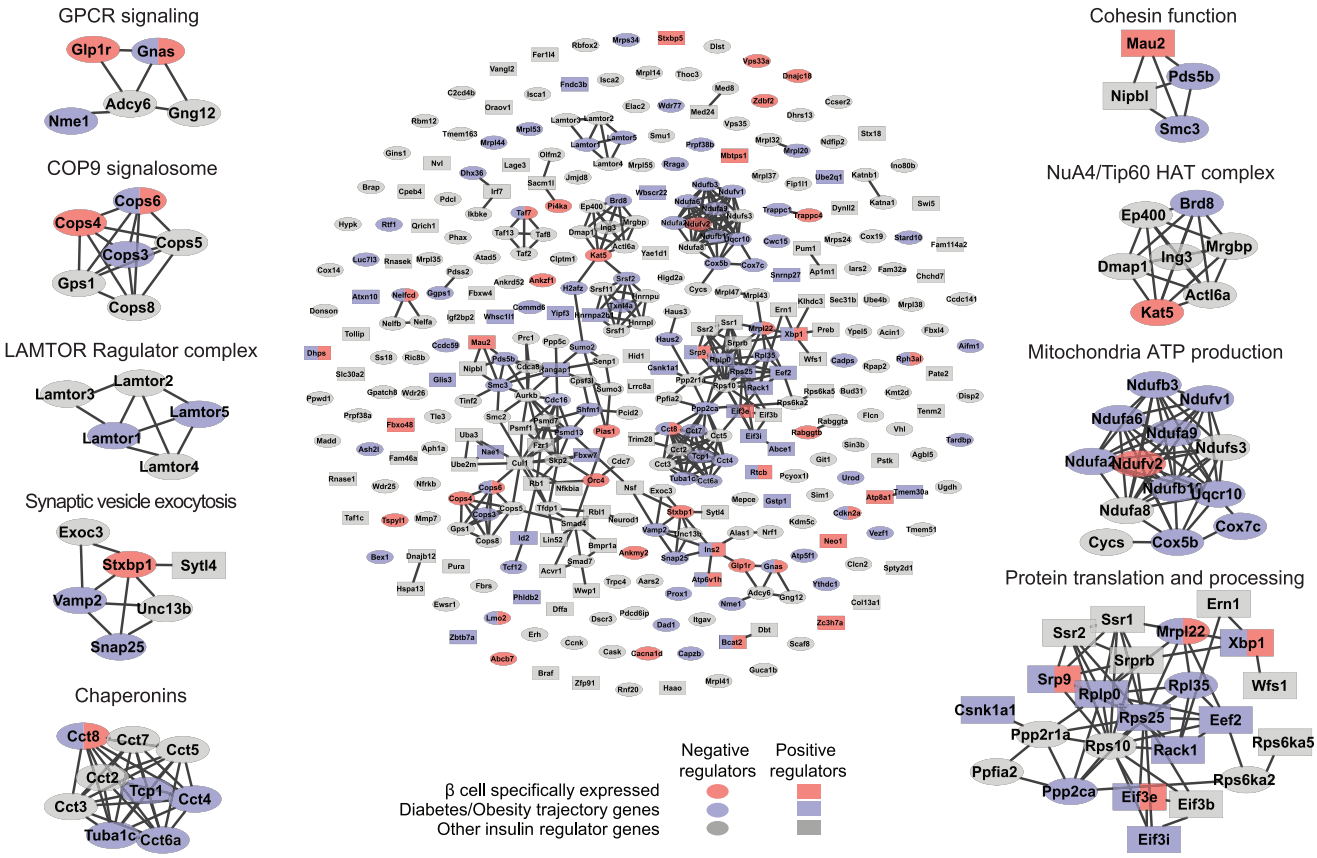


Figure 6. The Network of Insulin Regulators

Top insulin regulator genes from CRISPR screen were analyzed by search tool for the retrieval of interacting genes or proteins (STRING). Oval shape, negative insulin regulators; rectangle, positive regulators. The β cell signature genes identified from Drop-Seq data are highlighted in color: purple, diabetes or obesity trajectory genes; red, β-cell-specific genes. Enlarged are the nine subnetworks with signature genes.

robust (Figure 5C). We called a gene "positive regulator" if its' sgRNA was enriched in Ins^{Lo} population, meaning that the gene increased the intracellular insulin amount. As expected, the strongest positive insulin regulator was *Ins2*, the mouse ortholog of human insulin gene *INS*. Conversely, we also called "negative regulators" from sgRNAs enriched in the Ins^{Hi} population (Figure 5C).

Multiple mechanisms may regulate the intracellular insulin levels. Among the top hits (Figure 5C), we found transcriptional regulators (such as *Glis3* and *Nelfs*), translational regulators (such as *Eef2*), and post-translational regulators (such as unfolded protein response regulator gene *Xbp1*) (Lee et al., 2011). Importantly, our screen also identified regulators of insulin secretion, especially genes involved in glucose induced insulin secretion (GSIS). For example, *Glp1r* (Trujillo and Nuffer, 2014), a known GSIS regulator, was identified as one of the top negative regulators (Figure 5C). Notably, although 74 of 373 insulin regulators are also "fitness genes" (leading to slow growth upon deletion) (Hart et al., 2015), the distribution of fitness genes among positive (17 of 117, 15%) or negative insulin regulators (57 of 256, 22%) is not biased ($p = 0.094$, Fisher's exact test) suggesting that slow-growing cells do not cause a bias in our screen.

To validate the hits from the screen, we also infected MIN6 cells with individually cloned sgRNAs and used flow cytometry to measure the intracellular insulin levels (STAR Methods). We verified 13 out of 20 (65%) positive regulators and 30 out of 33 (91%) negative regulators (Figure 5D). We further test 20 of the verified genes with newly designed sgRNAs and re-validated 19 of them, indicating a low off-target rate (Figure S5F). It is worth noting that our genome-wide CRISPR analysis tends to report higher fold changes than individual experiments. This is mostly likely because in the individual experiments, a significant fraction of cells did not obtain loss-of-function mutations, whereas the selection of Ins^{Hi} and Ins^{Lo} populations in the screen enriches the cells with successful gene knockout events. As a result, the pooled screen had a higher sensitivity, which may also explain why some CRISPR hits were not validated in the individual experiments.

Integrative Analysis of Single-Cell and CRISPR Data Revealed Disease-Relevant Insulin-Regulating Modules

We performed STRING analysis (Szklarczyk et al., 2015) on the 373 tier-one insulin regulators and drew all the functional or physical associations between these genes (Figure 6; STAR Methods). Compared to Drop-Seq data, we identified 100 insulin regulators up- or downregulated in the diabetes or obesity trajectory, including *Cdkn2a*, *Cox7c*, *Glis3*, and *Xbp1* (Figure 6

and S4B). Therefore, these transcriptional changes may play a causal role in the disease development. We also identified 40 insulin regulators specifically expressed in β cell including *Glp1r*. These genes may have important selective functions in β cells (Figure 6; Data S5). Seventeen β-cell-specific insulin regulators also change their expression in diabetes or obesity trajectories (*Atp6v1h*, *Bcat2*, *Cct8*, *Cdkn2a*, *Cops6*, *Dhps*, *Eif3e*, *Gnas*, *Ins2*, *Lmo2*, *Mrpl22*, *Nelfcd*, *Rph3al*, *Rtcb*, *Srp9*, *Taf7*, and *Xbp1*) (Figure 6).

The network analysis also expanded the repertoire of disease relevant insulin regulators. Our analysis highlighted nine notable modules of insulin regulators; each module contains at least one diabetes or obesity signature genes in β cell, and all of them represent multi-protein complexes with cooperative functions (Figure 6). The identification of a whole protein complex from the CRISPR screen is strong evidence that the complex is key for β cell function. Alteration of signature genes in these complexes may influence their normal functions and contribute to the disease development. For example, the largest module is the protein translation and processing network. Most of the genes in this module are positive insulin regulators, presumably because they positively regulate insulin protein production and maturation. Negative regulators in this module, such as PP2A protein phosphatase members *Ppp2r1a* and *Ppp2ca*, may function as inhibitors of protein translation or processing.

Six protein modules appear to regulate insulin release (Figure 6): (1) GPCR signaling complex (GPCR signal such as *Glp1r* amplifies insulin secretion); (2) mitochondria ATP production module (important for ATP production, therefore ATP-dependent insulin release); (3) synaptic vesicle exocytosis module (insulin release); (4) COP9 signalosome (protein ubiquitination and ubiquitin-dependent endocytosis); (5) LAMTOR Ragulator complex (late endosome or lysosome scaffold proteins); and (6) chaperonins (promoting insulin secretion by assisting protein folding). Interestingly, nearly all genes in these modules are negative insulin regulators, i.e., once these genes are deleted, insulin accumulates in the cells due to the impaired secretion. The only positive regulator in these modules is *Sytl4*, a known inhibitor of exocytosis. Consistently, another exocytosis inhibitor *Stxbp5* is also a positive insulin regulator (Figure 6). It is interesting that the endocytic modules are also negative insulin regulators, because this suggests that insulin granules require not only the exocytosis process for secretion, but also the endocytosis process to ensure vesicle recycle or membrane recapture (MacDonald and Rorsman, 2007). Misregulation of either of these two trafficking pathways will lead to the failure of sustainable insulin release.

Our CRISPR screen also identified two previously unrecognized insulin regulators modules: the cohesin function module and the NuA4/Tip60 histone acetyltransferase (HAT) complex module (Figure 6). The β cell functions of these two complexes are unknown, so we performed additional analyses to gain further insights into the underlying mechanisms (Figure 7).

Mau2-Nipbl Cohesin Loading Complex Regulates Insulin Gene Transcription

Cohesin is a multifunction protein complex. During S phase and M phase, cohesin tethers two sister chromatids together and is essential for proper chromosome segregation in mitosis (Peters et al., 2008). In the G1 phase, cohesin is important for long-range chromosome interactions at promoters, enhancers, and CCCTC binding factor (CTCF)-occupied insulator elements. Our CRISPR screen identified two cohesin proteins, *Smc3* and *Pds5b*, as negative regulators, but intriguingly, cohesin loading factors *Mau2* and *Nipbl* were identified as positive insulin regulators (Figure 6). In the mammalian genome, *Nipbl* mainly binds to promoters and enhancers, while the strongest peaks of cohesin are at intergenic CTCF sites (Busslinger et al., 2017; Kagey et al., 2010; Zuin et al., 2014). Therefore, the cohesin loading complex may play a direct role regulating gene transcription independent of the CTCF-bound cohesin or mitotic cohesin (Zuin et al., 2014). Consistent with these findings, we found that *Mau2* and *Nipbl* bind to *Ins2* promoter, and the transcription level of *Ins2*, rather than *Ins1*, was lower upon *Mau2* or *Nipbl* deletion (Figures 7A and 7B). Our single-cell data demonstrated that in human islets, *MAU2* is expressed at higher levels in β cells than in other cell types (Figure 7C). Interestingly, from the 171 human pancreas samples collected by the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2015), we found a significant positive correlation between the transcription of *MAU2* and *INS* gene (Figure 7D). Taken together, our results suggested a β-cell-specific function of cohesin loading factors in regulating insulin gene transcription. These results may also shed light on the mechanism of Cornelia de Lange syndrome (CdLS), which has been linked to protein mutations in NIPBL and cohesin complex (Liu and Baynam, 2010; Liu et al., 2009).

NuA4/Tip60 HAT Complex Regulates Insulin Secretion

We were interested in the β cell function of NuA4/Tip60 HAT complex module because *Kat5* and *Dmap1* were the strongest hits from our CRISPR screen (Figure 5C), suggesting a direct insulin regulatory function. The best-characterized function of the NuA4/Tip60 complex is transcriptional activation by acetylation of histones H4 and H2A. However, this *trans*-activity cannot explain the increase of intracellular insulin level upon deletion of NuA4/Tip60 complex members *Kat5*, *Dmap1*, and *Brd8* (Figures 5D and 7F). We therefore tested the possibility that the NuA4/Tip60 complex may control insulin secretion. First, we observed compromised GSIS when treating either MIN6 cells or primary human islets with NU9056, a *KAT5*-specific acetyltransferase inhibitor (Coffey et al., 2012). Acetate can enhance GSIS, presumably by increasing the cellular level of acetyl-CoA, the substrate of all acetyl-transferases including *KAT5* (Figure 7E) (Shimazu et al., 2010). Furthermore, deletion of NuA4/Tip60 complex proteins (*Kat5*, *Dmap1*, and *Brd8*) as well as known insulin secretion regulators (*Glp1r*, *Gnas*, and *Stxbp1*) all led to accumulation of intracellular insulin, lower baseline insulin secretion, and GSIS defect (Figures 7F–7H). As a control, deletion of positive regulator *Mau2* and *Nipbl* did not affect glucose responses. Interestingly, the baseline insulin secretion from *Mau2* and *Nipbl* knockout cells increased (Figure 7G); the reason remains elusive because knocking out these two proteins may cause widespread effects on gene transcription, genome architecture, and cell cycle. Taken together, these data strongly suggest that the acetyltransferase activity of NuA4/Tip60 complex is key for insulin secretion from β cell. It will be interesting

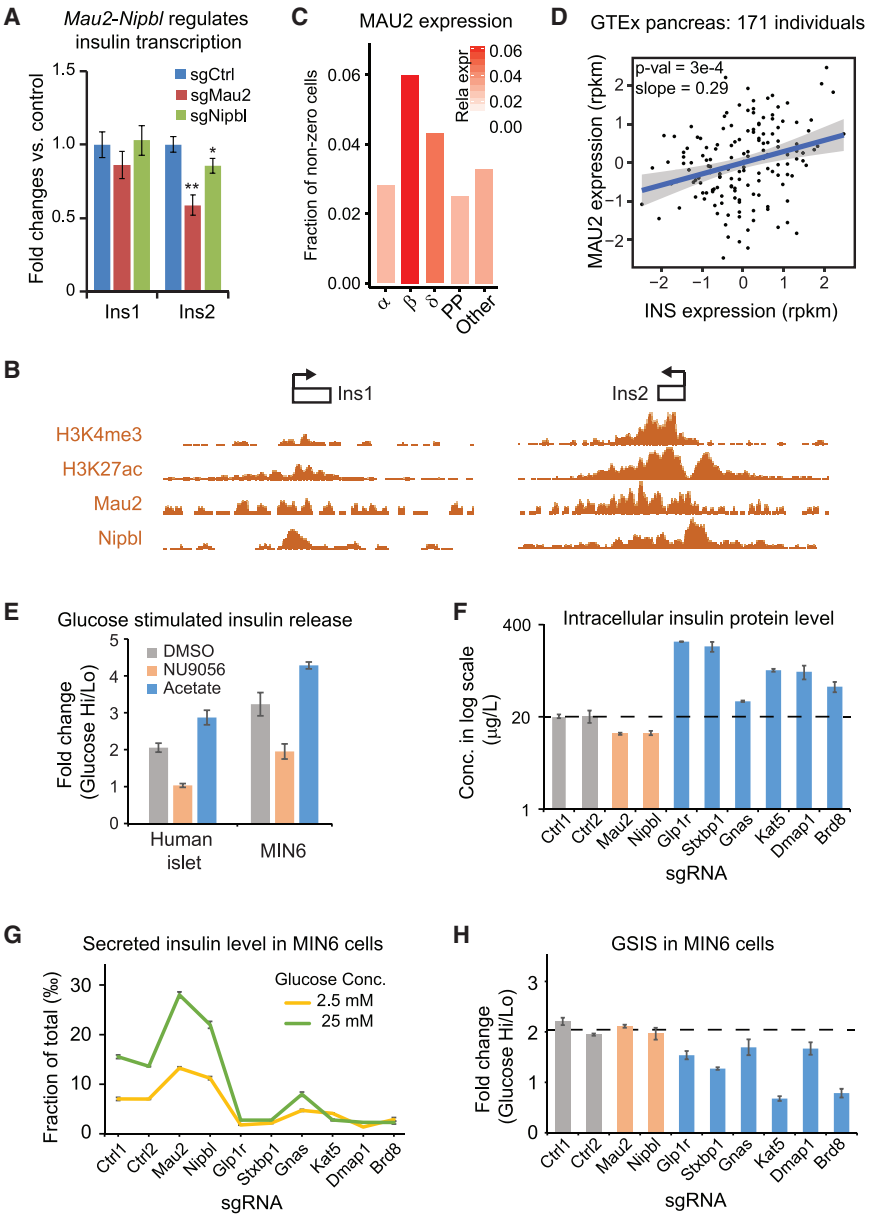


Figure 7. *Mau2-Nipbl* and NuA4/Tip60 Complex Are Insulin Regulators

(A) Fold change of *Ins1* and *Ins2* genes in mouse MIN6 cells after CRISPR deletion of *Mau2* and *Nipbl*. Expression levels are normalized to control sgRNA. **p < 0.01; *p < 0.05, t test.

(B) MIN6 cells chromatin immunoprecipitation sequencing (ChIP-seq) data with H3K4me3, H3K27ac, Mau2, and Nipbl at *Ins1* and *Ins2* loci.

(C) Drop-Seq demonstrates β-cell-specific expression of the *MAU2* gene.

(D) Correlation between human *MAU2* and *INS* gene in the GTEx cohort of 171 human pancreas tissue.

(E) GSIS is measured by the fold change of extracellular insulin levels between low glucose and high glucose challenge: human islet (1.6 mM/16.6 mM); MIN6 (1.25 mM/25 mM). The cells were pretreated with DMSO, 5 nM NU9056, or 1 mM acetate for 24 h as indicated.

(F) ELISA quantification of intracellular insulin levels in CRISPR knockout MIN6 cells. Gray, control sgRNAs; orange, positive regulators; blue, negative regulators.

(G) Insulin secretion from MIN6 knockout cells are measured as fraction of total intracellular insulin levels.

(H) GSIS are computed as secreted insulin levels between low and high glucose conditions (2.5 mM/25 mM). All error bars in this figure are SD from triplicated assays.

We also observed a few rare cell subpopulations that are relatively individual-specific; more donors are necessary to obtain a comprehensive picture of endocrine cell heterogeneity.

Highly Sensitive RePACT Algorithm Identifies Disease-Specific Cellular Changes

Obesity and T2D are two highly related diseases that share many common characteristics. Understanding the commonality and differences between the two diseases can be challenging because many patients have both diseases. To address this challenge, we developed RePACT, a general single-cell analysis algorithm to identify disease relevant genes with high sensitivity. Similar to several published single-cell analysis methods in differentiation systems (Trapnell et al., 2014), the key step of RePACT is to determine a series of pseudo states best reflecting disease-relevant variance. By doing this, RePACT can discern a continuous disease trajectory that is too subtle to be detected from bulk analysis (Figure 4A). In this study, with merely nine donors, RePACT defined different trajectories for obesity and T2D and further identified numerous common and specific signature genes in β cells. We believe that the use of RePACT in future single-cell studies with more donors will significantly improve the sensitivity and robustness. The RePACT strategy is applicable to elucidate the downstream targets of NuA4/Tip60 complex responsible for insulin release.

DISCUSSION

Higher Throughput Single-Cell Analysis

We have performed an in-depth analysis of 39,905 single islet cell transcriptome from nine human donors. To our knowledge, this is the largest single-cell transcriptome dataset in primary human islets. The data of identified cell types and their signature genes will serve as a valuable resource to understand islet cell differentiation and disease development. Due to improved throughput, we achieved high sensitivity and robustness to detect rare cell subpopulations, such as the proliferating α cells.

to the studies of many other human diseases, especially when the sample availability is an issue.

Integration of Genome-wide CRISPR Analyses to Reveal Disease-Contributing Genes and Pathways

Recent development of CRISPR-based genome editing has enabled pooled genome-wide screens in mammalian cells (Gilbert et al., 2014; Konermann et al., 2015; Parnas et al., 2015; Shalem et al., 2014; Wang et al., 2014). In this study, in order to reveal the functions of β cell signature genes from Drop-Seq data, we also performed an unbiased CRISPR screen for insulin regulators. Simply using resting-state intracellular insulin level as a readout, our screen identified not only transcriptional and translational regulators, but also numerous genes that control energy production, protein folding, vesicle trafficking, etc., suggesting that insulin production and release is controlled by a highly complex network. Importantly, the integrative analysis of single-cell transcriptome and CRISPR screen data highlighted potential causal genes in diabetes or obesity disease development, among which we have identified the *Mau2-Nipbl* cohesin loading complex and the NuA4/Tip60 HAT complex as two previously unrecognized insulin regulating protein modules. We further revealed that the *Mau2-Nipbl* complex regulates insulin gene transcription, and a surprising role of the NuA4/Tip60 HAT complex in regulating insulin protein release. Our study provides a general strategy for systematically characterizing diabetes genes in pancreatic islets, as well as disease genes in other complex tissues.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human islets and cell lines
- METHOD DETAILS
  - Transformation and amplification of GeCKO library
  - Cloning individual sgRNAs
  - Virus packaging
  - Measuring virus titers
  - Lentiviral infection
  - Flow cytometry and cell sorting
  - Library preparation
  - Glucose-stimulated insulin secretion
  - RNA extraction and real-time PCR
  - Preparation of human pancreatic islet single cell
  - Drop-Seq
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Drop-Seq reads processing
  - Distinguish cell barcodes with single cell transcriptomes
  - Down-sampling sequencing data
  - Cell type identification using unsupervised clustering
  - Cell type identification of low-depth STAMPs
  - Differential expression analysis
  - Public databases
  - RePACT
  - Raw CRISPR screen sequencing data analysis
  - Guide-RNA level CRISPR screen data analysis
  - Gene-level CRISPR screen data analysis
  - STRING network analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, four tables and five data files and can be found with this article online at https://doi.org/10.1016/j.celrep.2019.02.043.

ACKNOWLEDGMENTS

This work was supported by grants from NIH (R01DK113185 to Yan Li and R01HG009658 to F.J.), Mt. Sinai Health Care Foundation (OSA510114 to Yan Li and OSA510113 to F.J.), a pilot award from Clinical and Translational Science Collaborative (CTSC) at Case Western Reserve University (VSN639001 to Yan Li), and a pilot award from SFARI (401625 to F.J.).

AUTHOR CONTRIBUTIONS

Yan Li and F.J. conceived the project. Z.F., C.W., H.L., W.M., L.L., S.L., and C.A. performed the experiments. R.T., and Y.P. performed high-throughput cell sorting. C.W., Z.F., X.L., Q.D., and Yun Li performed bioinformatic analysis. F.J., Yan Li, C.W., and Z.F. wrote the manuscript. P.A. and A.W. also contributed to the manuscript writing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 29, 2017
Revised: May 3, 2018
Accepted: February 12, 2019
Published: March 12, 2019

REFERENCES

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. *3*, 346–360.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093.

Busslinger, G.A., Stocsits, R.R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., and Peters, J.M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. Nature *544*, 503–507.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Clark, A., and Nilsson, M.R. (2004). Islet amyloid: a complication of islet dysfunction or an aetiological factor in Type 2 diabetes? Diabetologia *47*, 157–169.

Coffey, K., Blackburn, T.J., Cook, S., Golding, B.T., Griffin, R.J., Hardcastle, I.R., Hewitt, L., Huberman, K., McNeill, H.V., Newell, D.R., et al. (2012). Characterisation of a Tip60 specific inhibitor, NU9056, in prostate cancer. PLoS ONE *7*, e45539.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

da Cunha, J.P., Galante, P.A., de Souza, J.E., de Souza, R.F., Carvalho, P.M., Ohara, D.T., Moura, R.P., Oba-Shinja, S.M., Marie, S.K., Silva, W.A., Jr., et al. (2009). Bioinformatics construction of the human cell surfaceome. Proc. Natl. Acad. Sci. USA *106*, 16752–16757.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dorajoo, R., Ali, Y., Tay, V.S.Y., Kang, J., Samydurai, S., Liu, J., and Boehm, B.O. (2017). Single-cell transcriptomics of East-Asian pancreatic islets cells. Sci. Rep. *7*, 5024.

Enge, M., Arda, H.E., Mignardi, M., Beausang, J., Bottino, R., Kim, S.K., and Quake, S.R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell *171*, 321–330.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

Gutierrez, G.D., Gromada, J., and Sussel, L. (2017). Heterogeneity of the Pancreatic Beta Cell. Front. Genet. *8*, 22.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526.

Hooper, P.L., and Hooper, P.L. (2009). Inflammation, heat shock proteins, and type 2 diabetes. Cell Stress Chaperones *14*, 113–115.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kanarek, N., Keys, H.R., Cantor, J.R., Lewis, C.A., Chan, S.H., Kunchok, T., Abu-Remaileh, M., Freinkman, E., Schweitzer, L.D., and Sabatini, D.M. (2018). Histidine catabolism is a major determinant of methotrexate sensitivity. Nature *559*, 632–636.

Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera, Mdel.C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat. Biotechnol. *32*, 267–273.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

Korkmaz, G., Lopes, R., Ugalde, A.P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. *34*, 192–198.

Kupca, S., Sjakste, T., Paramonova, N., Sugoka, O., Rinkuza, I., Trapina, I., Daugule, I., Sipols, A.J., and Rumba-Rozenfelde, I. (2013). Association of obesity with proteasomal gene polymorphisms in children. J. Obes. *2013*, 638154.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M.L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. Genome Res. *27*, 208–222.

Lecube, A., Carrera, A., Losada, E., Hernández, C., Simó, R., and Mesa, J. (2006). Iron deficiency in obese postmenopausal women. Obesity (Silver Spring) *14*, 1724–1730.

Lee, A.H., Heidtman, K., Hotamisligil, G.S., and Glimcher, L.H. (2011). Dual and opposing roles of the unfolded protein response regulated by IRE1alpha and XBP1 in proinsulin processing and insulin secretion. Proc. Natl. Acad. Sci. USA *108*, 8885–8890.

Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu, X.S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. *15*, 554.

Li, J., Klughammer, J., Farlik, M., Penz, T., Spittler, A., Barbieux, C., Berishvili, E., Bock, C., and Kubicek, S. (2016). Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. EMBO Rep. *17*, 178–187.

Liu, J., and Baynam, G. (2010). Cornelia de Lange syndrome. Adv. Exp. Med. Biol. *685*, 111–123.

Liu, J., Zhang, Z., Bando, M., Itoh, T., Deardorff, M.A., Clark, D., Kaur, M., Tandy, S., Kondoh, T., Rappaport, E., et al. (2009). Transcriptional dysregulation in NIPBL and cohesin mutant human cells. PLoS Biol. *7*, e1000119.

Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. *17*, 75.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45* (*D1*), D896–D901.

MacDonald, P.E., and Rorsman, P. (2007). The ins and outs of secretion from pancreatic beta-cells: control of single-vesicle exo- and endocytosis. Physiology (Bethesda) *22*, 113–121.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Marchetti, P., Dotta, F., Lauro, D., and Purrello, F. (2008). An overview of pancreatic beta-cell defects in human type 2 diabetes: implications for treatment. Regul. Pept. *146*, 4–11.

Muraro, M.J., Dharmadhikari, G., Grun, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J., et al. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst. *3*, 385–394.

Naggert, J.K., Fricker, L.D., Varlamov, O., Nishina, P.M., Rouille, Y., Steiner, D.F., Carroll, R.J., Paigen, B.J., and Leiter, E.H. (1995). Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. Nat. Genet. *10*, 135–142.

Nead, K.G., Halterman, J.S., Kaczorowski, J.M., Auinger, P., and Weitzman, M. (2004). Overweight children and adolescents: a risk group for iron deficiency. Pediatrics *114*, 104–108.

Nomiyama, T., Perez-Tilve, D., Ogawa, D., Gizard, F., Zhao, Y., Heywood, E.B., Jones, K.L., Kawamori, R., Cassis, L.A., Tschöp, M.H., and Bruemmer, D. (2007). Osteopontin mediates obesity-induced adipose tissue macrophage infiltration and insulin resistance in mice. J. Clin. Invest. *117*, 2877–2888.

Omary, M.B., Lugea, A., Lowe, A.W., and Pandol, S.J. (2007). The pancreatic stellate cell: a star on the rise in pancreatic diseases. J. Clin. Invest. *117*, 50–59.

Otoda, T., Takamura, T., Misu, H., Ota, T., Murata, S., Hayashi, H., Takayama, H., Kikuchi, A., Kanamori, T., Shima, K.R., et al. (2013). Proteasome dysfunction mediates obesity-induced endoplasmic reticulum stress and insulin resistance in the liver. Diabetes *62*, 811–824.

Parnas, O., Jovanovic, M., Eisenhaure, T.M., Herbst, R.H., Dixit, A., Ye, C.J., Przybylski, D., Platt, R.J., Tirosh, I., Sanjana, N.E., et al. (2015). A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell *162*, 675–686.

Peters, J.M., Tedeschi, A., and Schmitz, J. (2008). The cohesin complex and its roles in chromosome biology. Genes Dev. *22*, 3089–3114.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.M., Andréasson, A.C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., et al. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab. *24*, 593–607.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelson, T., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., and Zhang, F. (2014).

Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shifrut, E., Carnevale, J., Tobin, V., Roth, T.L., Woo, J.M., Bui, C.T., Li, P.J., Diolaiti, M.E., Ashworth, A., and Marson, A. (2018). Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. Cell *175*, 1958–1971.

Shimazu, T., Hirschey, M.D., Huang, J.Y., Ho, L.T., and Verdin, E. (2010). Acetate metabolism and aging: An emerging connection. Mech. Ageing Dev. *131*, 511–516.

Simcox, J.A., and McClain, D.A. (2013). Iron and diabetes risk. Cell Metab. *17*, 329–341.

Skelin, M., Rupnik, M., and Cencic, A. (2010). Pancreatic beta cell lines and their applications in diabetes mellitus research. ALTEX *27*, 105–113.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Templeman, N.M., Skovsø, S., Page, M.M., Lim, G.E., and Johnson, J.D. (2017). A causal role for hyperinsulinemia in obesity. J. Endocrinol. *232*, R173–R183.

Thorel, F., Népote, V., Avril, I., Kohno, K., Desgraz, R., Chera, S., and Herrera, P.L. (2010). Conversion of adult pancreatic alpha-cells to beta-cells after extreme beta-cell loss. Nature *464*, 1149–1154.

Tiano, J.P., and Mauvais-Jarvis, F. (2012). Importance of oestrogen receptors to preserve functional β-cell mass in diabetes. Nat. Rev. Endocrinol. *8*, 342–351.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Trujillo, J.M., and Nuffer, W. (2014). GLP-1 receptor agonists for type 2 diabetes mellitus: recent developments and emerging agents. Pharmacotherapy *34*, 1174–1186.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wang, Y.J., Schug, J., Won, K.J., Liu, C., Naji, A., Avrahami, D., Golson, M.L., and Kaestner, K.H. (2016). Single-Cell Transcriptomics of the Human Endocrine Pancreas. Diabetes *65*, 3028–3038.

Weisberg, S., Leibel, R., and Tortoriello, D.V. (2016). Proteasome inhibitors, including curcumin, improve pancreatic β-cell function and insulin sensitivity in diabetic mice. Nutr. Diabetes *6*, e205.

Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. (2008). DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. *36*, D88–D92.

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A.J., Yancopoulos, G.D., Lin, C., and Gromada, J. (2016). RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. Cell Metab. *24*, 608–615.

Ye, J. (2009). Emerging role of adipose tissue hypoxia in obesity and insulin resistance. Int. J. Obes. *33*, 54–66.

Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., and Wei, W. (2014). High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature *509*, 487–491.

Zuin, J., Franke, V., van Ijcken, W.F., van der Sloot, A., Krantz, I.D., van der Reijden, M.I., Nakato, R., Lenhard, B., and Wendt, K.S. (2014). A cohesin-independent role for NIPBL at promoters provides insights in CdLS. PLoS Genet. *10*, e1004153.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rat monoclonal Insulin (Clone 182410)	R&D Systems	Cat# MAB1417; RRID:AB_2126533
Anti-Scc4 antibody [EPR14390] (Mau2) 100ul	Abcam	Cat# ab183033; RRID:AB_2783830
Rabbit anti-NIPBL Antibody	Bethyl	Cat# A301-778A; RRID:AB_1211233
Bacterial and Virus Strains		
NEB 10-beta electro-competent cells	NEB	C3020K
Biological Samples		
Human islets	PRODO Laboratories	N/A
Chemicals, Peptides, and Recombinant Proteins		
NU 9056	TOCRIS	4903
Critical Commercial Assays		
Human Insulin ELISA	Mercodia	10-1113-01
Mouse Insulin ELISA	Mercodia	10-1247-01
Deposited Data		
GSE101207	GEO	GEO: GSE101207
Experimental Models: Cell Lines		
293T cells	ATCC	CRL-3216
MIN6 cells	ATCC	CRL-11506
Oligonucleotides		
sgRNA oligos (See Table S3)	IDT	N/A
PCR primers (See Table S4)	IDT	N/A
Recombinant DNA		
Plasmid: LentiCRISPR v2	Addgene	#52961
Plasmid: GeCKO Library	Addgene	#1000000052
Plasmid: pCMV-VSVG	Addgene	#8454
Plasmid: pCMV-dR8.91	Gift from Bing Ren's Lab	N/A
Software and Algorithms		
MAGeCK R package	Li et al., 2014	https://sourceforge.net/p/mageck/wiki/Home/
Cluego, Cytoscape package	Bindea et al., 2009	http://www.ici.upmc.fr/cluego/
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
FASTX-Toolkit	Hannon Lab	http://hannonlab.cshl.edu/fastx_toolkit/index.html
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR
SEURAT	Butler et al., 2018	https://satijalab.org/seurat/
scran	Lun et al., 2016	http://bioconductor.org/packages/release/bioc/html/scran.html
STRING	Szklarczyk et al., 2015	https://string-db.org
RePACT	This study	https://github.com/chenweng1991/RePACT

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yan Li (yxl1379@case.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human islets and cell lines

Human Islets were purchased from Prodo Laboratories, Inc. and cultured in PIM medium with Human AB Serum (Prodo Laboratories) in 6-well plates with ultralow attachment surface (Corning). MIN6 cells (female) were cultured in high glucose DMEM containing GlutaMax (Invitrogen), 1 mM Sodium pyruvate (Invitrogen), and 50 μM β-mercaptoethanol. 293T cells (female) were cultured in high glucose DMEM (Invitrogen) with 10% Fetal Bovine Serum (VWR Scientific). All cells were cultured at 37°C in Forma Steri-Cycle i160 CO2 Incubator (ThermoFisher) with 5% CO2.

METHOD DETAILS

Transformation and amplification of GeCKO library

Mouse GeCKOv2 CRISPR sgRNA libraries A and B were purchased from Addgene (Addgene #1000000052). 10ng libraries were used to transform NEB 10-beta electro-competent cells in 0.1cm electroporation cuvettes (Bio-rad) with MicroPulser (Bio-rad) at 5 kV for 2 ms. Transformation efficiencies were determined by series dilution of the cells and plating the cells on LB agar plates. Transformed cells with high efficiencies (> 10^8 cfu/10ng plasmids) are seeded into 500mL LB medium and cultured overnight. Plasmids were prepared from the cells using Maxi preparation kits (QIAGEN).

Cloning individual sgRNAs

LentiCRISPR v2 plasmid was purchased from Addgene (Addgene #52961). 500ng LentiCRISPR v2 plasmid is linearized with BsmBI (NEB) at 55°C for 4 hours and purified with 1% agarose gel. The sgRNA sequences are listed in Table S3. For one sgRNA, we designed two oligos complementary to each other in the following format: 5′-CACCGXXXXXXXXXXXXXXXXXXX-3′ and 5′-AAAC YYYYYYYYYYYYYYYYYYYC-3′ (X 20-mers and Y 20-mers are complementary target sequences). To anneal the complimentary oligos, 1uL from each of the two oligos (100 uM), 1uL 10x T4 ligation buffer (NEB), 6.5uL H2O and 0.5uL T4 PNK (NEB) were mixed together and incubated at 37°C for 30 minutes followed by incubation in 95°C for 5 minutes. After 95°C incubation, shut off the block heater and let the reaction cooled down naturally to room temperature. Annealed oligos were then diluted at 1:200 dilution for use. To clone the individual sgRNAs, 50ng linearized vector, 1uL diluted oligo complex, 5uL 2X quick ligase buffer (NEB) are mixed and add water to 9uL. 1uL quick ligase (NEB) were next added into the mixture and the reaction is performed at 25°C for 10 minutes. 5uL ligation products were transformed immediately into Stbl3 bacteria following standard transformation protocol.

Virus packaging

24 hours before transfection, 500 million 293T cells were split into 50 10cm plates so that cells reach 60% confluence the next day. Each plate was co-transfected with 4μg GeCKOv2 library (or LentiCRISPR v2 plasmid expressing single sgRNAs), 2μg delta V8.91, and 2μg pCMV-VSVG. For each plate, plasmids and 21μg polyethylenimine (PEI) were pre-mixed in 500uL Optium-MEM (Invitrogen) and incubated at room temperature for 10 minutes. Meanwhile, 293T cells are switched into 6mL fresh Optium-MEM. Plasmids mixture were added to the cells after incubation for transfection. 6 hours after transfection, each plate of 293T cells were switched to 10 mL fresh complete medium. 3 days after transfection, cell medium containing viral particles was harvested, filtered through 0.45μM Millipore filters. For individual CRISPR experiments, the crude viral supernatant were applied directly to infect cells. For large-scale viral infection, the viral supernatant was concentrated by centrifuging for 90 minutes at 25,000 rpm in 4°C. Virus pellets were washed once with ice-cold PBS before re-suspended in PBS with 1mM EDTA.

Measuring virus titers

Virus titers were determined by adding 1uL of a series of diluted virus to 1 million MIN6 cells in 12-well plates, with two wells for each viral dilution. To minimize proliferation, MIN6 cells were cultured in the medium with 1% FBS. 4 days after infection, for each viral concentration, one well of cells were selected with 4 μg/ml puromycin (Sigma-Aldrich) for two days. Living cells and dead cells were collected and count the viability of the cells. Total cell numbers for each dilution were also compared with the infected MIN6 cells without puromycin.

Lentiviral infection

For individual knockout, cells were seeded into 12 well plates (0.1 million per well) together with 300 μL crude viral supernatant. For large-scale viral infection, 120 million cells were seeded into three multilayer plates (Nest, #731002, 870 cm2) together with concentrated GeCKO library virus; the ratio between cells and virus were calculated so that 60% of cells can be infected (MOI 0.92). 24 hours after transduction, cells were washed once with PBS and then cultured in fresh medium. 4 days after transduction, cells were selected with 4 μg/ml puromycin for two days. After selection, a portion of cells were fixed as input control and the rest cells were expanded. For large scale screening, ~200 million cells were fixed in permeabilization/fixation buffer (BD Biosciences) after expansion.

Flow cytometry and cell sorting

For flow cytometry, cells were seeded into 24 well plates at a density of 0.5 million/cm2. On the second day, for each well, cells were trypsinated and washed once with PBS. Cells were fixed in 1% formaldehyde (Sigma) for 20 minutes and permeabilized in Perm/Wash buffer (BD Biosciences) for 40 minutes at room temperature. Cells were then stained with Insulin antibodies, rat IgG (R&D Systems) at 1:100 dilution in 50 ul Perm/Wash buffer in 4°C overnight. On the second day, cells were washed once with Perm/Wash buffer and stained with PE-conjugated anti-rat IgG secondary antibodies at 1:200 dilution in 50 ul Perm/Wash buffer at room temperature for 30 minutes in the dark. Cells were then washed once with PBS and analyzed in BD LSR II Flow Cytometer (BD Biosciences).

Library preparation

Genomic DNA was extracted from cells using Dneasy Blood & Tissloue Kit (QIAGEN). All sgRNA expressing cassettes within the genomic DNA are amplified from genomic DNA using mutual GeCKO primers (See Table S3). For each PCR amplification reaction, no more than 3 μg genomic DNA were amplified with Herculase II (Agilent) polymerase for 15 cycles. For each sample, all the PCR products were pooled together and purified with Aline PCR Clean DX magnetic beads. Purified PCR products were then ligated to illumine TruSeq adapters. Ligation products were purified before second-round PCR amplification using TruSeq D&E primers (Illumina) for 10 cycles. The resulting DNA libraries were then sequenced with standard TruSeq sequencing primers.

Glucose-stimulated insulin secretion

For MIN6 cells, 0.5 million cells per well were seeded into 12-well plates. For fresh islet samples, approximately 200 islets were seeded into 24-well plate with ultralow attachment surface. We changed fresh medium 24 hours before glucose challenge. Four hours before glucose challenge, cells were starved in low glucose Krebs' buffer (NaCl 12.8 mM, KCl 0.48 mM, KH2PO4 0.12 mM, MgSO4 1.2 mM, CaCl2 0.25 mM, NaHCO3 5mM, HEPES 10 uM, 0.5% BSA, 1.6 mM glucose) for 4 hours before challenged with high/low (16.6 mM/1.6 mM) glucose in fresh Krebs' buffer for up to 6 rounds alternatively. In each round, the supernatant was harvested for insulin ELISA analysis. Cells were washed once with PBS before next round stimulation. After GSIS, cells were lysed in RIPA buffer (Thermal Fisher) to measure intracellular insulin level if necessary.

RNA extraction and real-time PCR

We used TRIzol (Invitrogen) for RNA preparation. RNA was pretreated with DNase I before reverse transcription with M-MLV reverse transcriptase (Invitrogen) following standard protocol. For quantitative PCR, 1 ul cDNA, 10 ul 2X PerfeCTa SYBR Green SuperMix (Quanta Biosciences), 8 ul H2O, and 1 ul premixed real-time PCR primers were mixed, and PCR reactions were performed in PTC-200 Thermal Cycler with Chromo4 Fluorescence detector (MJ Research). Relative gene expression were determined with ddCt methods.

Preparation of human pancreatic islet single cell

To dissociate islets into single cells, cells were washed once in HBSS (Sigma-aldrich, #6648) and incubated in Accutase (Innovative Cell Technologies, #AT104) at 37°C for 20-25 min. The islets were broken up gently with a 5 mL pipette every 5 min. When > 95% of the islets were digested into single cells, PIM(S) medium were added to neutralize the Accutase. Single cells were washed again with HBSS and resuspended in HBSS at the concentration of 2×10^5/mL for Drop-Seq analysis.

Drop-Seq

We performed Drop-Seq following the protocol as previously described (Macosko et al., 2015). Briefly, three pump-controlled syringes with cell suspension (200,000 cells/mL), barcoded beads in lysis buffer (200,000 beads/mL), and droplet generation oil were connected to a microfluidic device under microscope supervision. During droplets generation, we set the cell and bead flow speed at 4,000uL/hr, and the oil speed at 15,000uL/hr. The droplets were collected into 50mL falcon tubes (usually less than 5mL). Under this setting, most droplets had at most one beads or one cell. Following droplet breakage, we performed 1st strand cDNA synthesis on beads following SMART-PCR protocol (Macosko et al., 2015). Finally, the resulting full-length cDNA library were prepared for sequencing.

For mixing species Drop-seq experiments, we mixed equal number of HEK293 and 3T3 cells and load them to Drop-Seq at different cell density. We found that at 200 cells/uL, less than 1% STAMPs were from more than one cell (doublets) (Figure S1F). However, the rate of doublets usually increases when handling human tissues, most likely due to the incomplete digestion. We have therefore taken a few data filtering steps to remove potential doublets computationally (further discussed below).

QUANTIFICATION AND STATISTICAL ANALYSIS

Drop-Seq reads processing

We performed raw reads processing following the instructions described in the original Drop-Seq publication (Macosko et al., 2015). The sequenced Drop-Seq libraries yield 50-base paired-end reads (PE50). However, since only the first 20bp of read 1 is informative (base 1-12 cell barcode, base 13-20 UMI), we trimmed base 21-50 of read 1 before further analysis. We first remove all data with the

quality score of read 1 (base 1-20) lower than 10. Read 2 was trimmed at 3′ end to remove ployA tails of at least 6 bases, and trimmed at 5′ if template switching oligo (TSO) adaptor sequence appears. Clean reads were then aligned to hg19 or mm9 using STAR with default settings. We only keep uniquely mapped reads on gene exons. We next filtered out PCR duplicates with the same coordinates, cell barcode, and UMI. We then grouped the reads by cell barcode, and generated the digital UMI-count matrix after counting transcripts for every genes with every cell barcode.

Distinguish cell barcodes with single cell transcriptomes

We defined STAMPs (single cell transcriptome attached to microparticles) as cell barcodes with significantly more reads than background. Under the Drop-Seq experimental settings, only 2∼5% of beads are co-encapsulated with cells. Therefore, most cell barcodes only have a small number of transcripts from mRNA contamination during the bead breakage step. In order to distinguish STAMPs from empty beads, we examined the density plots of transcript counts for all cell barcodes (Figure S1B). In all experiments, we observed a major peak from empty beads and a fat right tail representing STAMPs with single cell transcriptomes. We therefore took a simple approach by calculating mean (μ) and standard deviation (σ) of the major peak assuming a Gaussian distribution. Any cell barcode with more than $\mu + 2 * \sigma$ transcripts were called as STAMPs.

Down-sampling sequencing data

Because the Drop-Seq libraries have different sequencing depth, we observed variable sensitivity in detecting transcripts / genes from each library (Figures S1C and S1D), which causes bias during the clustering or comparative analyses. We therefore took a down-sampling approach to normalize the sequencing depth. We first run raw data processing as described above using full data and estimate the total STAMP numbers for each donor. For down sampling, we only took a portion of reads from every library so that the average per-STAMP sequencing depth are similar. New UMI-count matrices were generated again for all donors after down sampling. We found that the normalization of sequencing depth resulted in cleaner clusters in t-SNE plot (Figure S1E). We only used the down-sampled data matrices when different donors need to be compared.

Cell type identification using unsupervised clustering

We designed a pipeline to determine the cell types of most STAMPs with high confidence using unsupervised clustering methods (Figure S1A). First, we performed initial clustering analysis with the 11,920 top STAMPs with at least 1,000 transcripts after down sampling. It has been previously estimated that in human islets, < 0.1% endocrine cells are positive with more than one marker hormones (*INS*, *GCG*, *PPY*, *SST*, and *GHRL*). We therefore first filtered out 890 STAMPs (out of 12,810, or 6.9%) expressing two hormones (Figure S1G) before clustering analysis. In this step, one STAMP is considered as doublets if it has two hormone genes with 15 transcripts. As mentioned above, the percentage of doublets is significantly greater than estimated from species mix experiment because single cells from tissues are more inclined to adhere with each other than cultured cells.

For clustering, we first ranked top 10,000 genes based on average expression level among all cells; then grouped them into 10 bins with 1,000 genes each. Coefficient of variation (CV) was calculated for every gene within each bin. From every bin, we pick top 50 genes with highest CV as informative genes. All together, we picked 500 informative genes for clustering analysis. We used *Seurat* package for clustering analysis with default parameters. In *Seurat*, PCA was performed with the 500 informative genes. Using PC1 to PC10, cells were embedded in a K-nearest neighbor (KNN graph). Smart local moving algorithm (SLM) was applied to group cells into communities. PC1 to PC10 were used as input to visualize cell clusters in two-dimensional t-SNE space. In order to define cell type, we used Seurat *FindMarker* function to find marker genes of each cell cluster, and defined cell types based on our knowledge and literatures. We performed the first-round clustering to classify non-endocrine cells (ductal cells, active PSCs and quiescent PSCs, Figure 1B) and the second round to distinguish the endocrine cell types (α, β, δ, PP cells, Figure 1E). Acinar and ε cells are not distinguishable in t-SNE plot due to scarcity, but can be clearly recognized from PCA plots (Figures 2A–2D). We also noticed a very small number of STAMPs (223, or 1.8%) expressing hormone genes inconsistent with their cell type classification (> 15 transcripts) (Figure S1H), which were also filtered as possible doublets after clustering analysis (Figure S1A). Finally, we successfully assigned unique cell types to 11,697 STAMPs with high confidence (Data S1). We used the same method for other clustering analyses in this work.

Cell type identification of low-depth STAMPs

Finally, we classified the low-transcript STAMPs using the knowledge obtained from clustering the top STAMPs as training dataset (Figure S1A). As mentioned above, we performed PCA clustering of the training dataset using 500 informative genes. From the PCA results, we took 32 significant principle components (PCs) as ''knowledge'' learned from training set. The 32 PCs are linear combinations of the 500 informative genes, and compose a virtual 32-dimensional space. Each cell type should form a cluster in the space. We next calculated the arithmetic centers of 8 cell types from training dataset (ductal, acinar, PSC, α, β, δ, ε, PP cells), and built spheres for all cell types centered at their arithmetic mean in the 32-dimentional space. We also computed the Euclidian distance between every cell to the center of its cell type, and empirically defined the radius of each ''cell type sphere'' as 80 percentile of all the distances in this cell type. For any low depth STAMP, we also took the same 500 informative genes, computed its projection onto the 32 PCs from training data, and the distances between the STAMP and the centers of all cell type spheres. If one STAMP is located exclusively in one cell type's ''sphere,'' we will annotate the STAMP to that cell type. We also performed several filtering steps

similar to training set, and successfully classified 16,329 additional STAMPs (Figure S1A). Lastly, we used 2-dimensional PCA plots and visually confirmed the correctness of cell type assignment (Figures 2A–2D).

Differential expression analysis

Negative binomial (NB) distribution was often used in differentially expression (DE) analysis, for counts data with over-dispersion. Here, we assume that for any gene in a given cell, the transcripts number, UMI, can be modeled using NB distribution.

$$log\widehat{UMI} = \beta_0 + \beta_C C + \beta_D D + \log(sf)$$

$\widehat{UMI}$ is the expected value of UMI;

β_0 is the intercept, and β_C and β_D is the slope for C and D;

C stands for cell type, which is a categorical variable;

D stands for donor, which is a categorical variable. This variable was used to regress out the donor specific effect;

sf stands for size factor, which is use to normalized the single cell transcriptome. It mainly corrects the sequencing depth of each cell (total transcript counts of a cell). However, the size factor can be biased due to the dropout zeros, therefore needs further correction. In this study, we calculated sf using the *computeSumFactor* function in a Bioconductor package *scran* (Lun et al., 2016).

We performed pairwise comparison between any two cell types based on the negative binomial model described above. For every gene, we perform the regression analysis using the generalized linear model function *glm.nb* in the R package *MASS*. The p value of pairwise cell type specificity of any gene is provided by the function as the significance of β_C. The p values of all genes were further adjusted with Bioconductor package *qvalue* for to obtain q-values. We also computed two fold changes between the average transcript counts. Differentially expressed genes are defined when q-value < 0.05, and ranked by fold changes.

Public databases

GWAS data were downloaded from GWAS catalog (MacArthur et al., 2017). We hand-picked 80 diabetes related traits and classified them into 5 categories: body measurements, glucose metabolism, lipid metabolism, Type I diabetes, and Type II diabetes. In total, 1,050 genes were retrieved. The list of transcription factors (TFs) is obtained from transcriptional factor prediction database (DBD) (Wilson et al., 2008). The list of cell surface protein was downloaded from a previous study (da Cunha et al., 2009). All these gene lists used in this study are included in Data S3.

RePACT

We developed RePACT (Regressing Principle components for the Assembly of Continuous Trajectory) as a general method to sensitively identify disease relevant gene signatures using single cell data. The key step is to find the best trajectory to rank single cells (e.g., β cells) reflecting the change of disease status. In this study, we used RePACT to study obesity (denoted by a continuous BMI variable) and T2D (denoted by a dichotomous variable T2D).

The first step of RePACT is dimension reduction. For example, we took β cells from all donors and perform PCA analysis. A number of principle components (PCs) will be identified and ranked by the percentage of variance they can explain. Each PC is a linear combination of genes. This allows us to convert every cell's transcriptome from a high-dimensional vector (e.g., a vector of 10,000 genes) into a low-dimensional vector (e.g., a vector of top 10 PCs). In this study, we empirically picked top 10 PCs, while RePACT can be conveniently run with more or fewer PCs.

We next performed linear regression for the continuous variable BMI:

$$\widehat{\mathrm{BMI}} = \beta_0 + \sum_{i=1}^{10} \beta_i * PC_i$$

$\widehat{\mathrm{BMI}}$ is the expected BMI value;

β_0 is the intercept, and β_i is the slope of each PCs;

PC_i is the i th principle component of a cell.

With the regressed β values, we computed the BMI-index for every cell, which is the predicted BMI value from the model. The BMI-index was used to rank the cells, and its value indicates how far a cell is transformed toward obesity status.

For dichotomous variable (True / False) indicating whether a cell is from T2D donor, we used a logistic regression model.

$$\mathrm{logit}(\widehat{p}_{T2D}) = \beta_0 + \sum_{i=1}^{10} \beta_i * PC_i$$

$\widehat{p}_{T2D}$ is a number between 0 and 1 indicating the risk of a cell been from T2D donor;

β_0 is the intercept, and β_i is the slope of each PCs;

PC_i is the i th principle component of a cell.

With the regressed β values, we computed the T2D-index for every cell, which is the predicted $\mathrm{logit}(\widehat{p}_{T2D})$ value from the model. The T2D-index was used to rank the cells, and its value indicates how far a cell is transformed toward T2D status.

In order to identify genes associated with obesity or T2D trajectory. We grouped all cells into 20 bins with equal BMI- or T2D-index intervals; every bin contains hundreds of single cells. For every gene, we then calculated the average transcript counts from cells in each bin and obtained a vector of 20 values. A simple linear regression was performed between the average transcript counts and the index values of the bins with p value. The p values of all genes were adjusted with Bioconductor package *qvalue* to obtain q-values. Genes with q-value less than 0.05 were called significant trajectory genes.

Raw CRISPR screen sequencing data analysis

Raw FASTQ files were analyzed with FASTX-Toolkit (Hannon's Lab). sgRNA sequences were aligned to GeCKO library sequences with Bowtie2.

Guide-RNA level CRISPR screen data analysis

We performed two independent screens to evaluate the reproducibility of every sgRNA. Partly because late-passaging MIN6 cells grows slow and may also start to lose their insulin secretion ability, we performed two independent viral-transduction to reduce the time needed for cell expansion. It should be noted that this "viral-transduction replication" strategy is in contrast to "split-plate replication," in which the viral infected cells are split into multiple cultures after expansion for replicate experiments (Figure S5A). Most CRISPR screen papers we examined, e.g., (Hart et al., 2015; Kanarek et al., 2018; Shalem et al., 2014; Zhou et al., 2014), used the "split-plate" strategy; "viral-transduction replication" strategy was also used in several (but fewer) papers (Koike-Yusa et al., 2014; Korkmaz et al., 2016; Shifrut et al., 2018).

Statistically, the two replication strategies have drastically different error-structures. For our "viral-transduction replication" strategy, we have to consider the viral infection variability between the two experiments. For example, assuming the sequencing depth in all libraries are the same, a significant sgRNA may show 2-fold change in one replicate experiment between InsLo and InsHi population at 200 versus 100 reads with $p < 0.001$; in another replicate the same sgRNA may merely show 20 versus 10 reads with marginal p value due to viral infection variability. The proper way to handle this type of variation is to do pairwise comparison for every sgRNA in each experiment separately; then examine if the enriched/depleted sgRNA is reproducible between two experiments. We performed the following steps for sgRNA level analysis:

Step 1

For every sgRNA in each of the replicates, we used binomial tests to calculate the p values for the difference of abundance between InsHi and InsLo cell populations (with MAGeCK package inR). The p values of all sgRNAs in the two replicates are listed in Data S5. We setup three significance levels for p values (< 0.05, < 0.01, or < 0.001); in step 2 we will keep a sgRNA only if it is significantly enriched or depleted in both replicates.

Step 2

To determine how reproducible a sgRNA is, we further classified all sgRNAs into 5 groups based on their p values in two replicates. Tier 1 sgRNA have both p values < 0.001; Tier 2 sgRNA have one p value < 0.001 and the other p value between 0.01 and 0.001; Tier 3 sgRNAs have both p values between 0.01 and 0.001; Tier 4 sgRNAs have one p value between 0.01 and 0.001 and the other p value between 0.05 and 0.01; Finally, Tier 5 sgRNAs have both p values between 0.05 and 0.01. The reproducibility of Tier 1 sgRNAs by p value and fold change are shown in Figures S5B and S5C. It is clear that tier 1 sgRNA are separated from most spots, and importantly distant from non-target sgRNA controls in the sgRNA library.

Step 3

To further assess the FDR of the sgRNAs, we randomly shuffled the data from all sgRNA in each replicated screen for 100 times, and compute the FDR of tier 1 through tier 5 sgRNAs with different criteria (Table S1). According to this analysis, our tier 1 sgRNAs achieved FDR of 0.1; sgRNAs in other tiers can be also valuable although to a lesser extent.

Gene-level CRISPR screen data analysis

Since an off-target sgRNA can be also reproducible between replicate screens, it is important to perform gene level analysis examine the effect of different sgRNAs targeting the same gene. The GeCKO library has 6 sgRNAs for each gene, which provide redundancy for reproducibility test at gene level. However, many guides may not achieve statistical significance due to variation in library construction, viral transduction, cell sorting procedure, sequencing depth, etc. Additionally, some guide may not be potent enough or abundant enough in the library to cause detectable effects, especially when the number of sorted cells is relatively small. In this study, as mentioned before, we have to harvest cells as early as possible after virus transduction because late-passaging MIN6 cells grows slower and may also start to lose their insulin secretion ability, limiting the number of cells we can analyze. Therefore, our philosophy for gene level analysis is: (i) we certainly prefer genes with multiple supporting sgRNA; (ii) we can also accept hit genes with neutral guide RNAs; (iii) we prohibit a hit gene to have contradictory sgRNAs.

The following rules were applied for gene-level analysis:

- Tier 1-5 sgRNAs from Table S1 were all considered "supporting" sgRNA.
- Tier 1 hit gene must have at least one Tier 1 sgRNA; Tier 2 hit gene must have at least one Tier 2 sgRNA. We did not report tier 3-5 genes since their best sgRNAs are weak; in fact, we only used Tier 1 hit genes for follow-up analysis in this paper.

- A hit gene cannot have contradictory sgRNAs. i.e., a hit gene cannot be classified as ''enrichment'' and ''depletion'' by two guides at the same time.
- After filtering, the rest Tier 1 and 2 hit genes were further classified into sub-tiers based on the number of supporting sgRNAs (Table S2).
- On randomly shuffled data (the p values of sgRNAs are shuffled in each experiment), we simulate all the aforementioned gene-level rules to estimate the FDR for hit genes (Table S2). As expected, the FDR of genes supported by 2 or more sgRNAs is extremely low.

We summarized the results of all tier 1 and tier 2 genes in the ''collapse_to_gene'' sheet of Data S5. As expected, genes supported by tier 1 sgRNAs are also more likely to be supported by multiple sgRNAs (Figure S5D). The strong FDRs of Tier 1A,1B, and 2A genes clearly indicate that genes with multiple supporting sgRNAs are high-confidence hits. We chose to report Tier 1C genes in this paper because for these genes, we are fairly certain about their guide RNA (tier 1, $p < 0.001$ in both replicates). However, the off-target risk of these genes must be acknowledged.

STRING network analysis

Gene interactions were analyzed using Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (https://string-db.org/cgi/input.pl). Network cosmetics were reconstructed using Cytoscape 3.4.0. Only interactions with scores greater than 0.85 were revealed.

DATA AND SOFTWARE AVAILABILITY

Raw data files of Drop-seq and Chip-seq are accessible at GEO: GSE101207; RePACT package is available at https://github.com/chenweng1991/RePACT.