



Single-cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk

Joshua Chiou^{1,7}, Chun Zeng^{2,3,7}, Zhang Cheng⁴, Jee Yun Han⁴, Michael Schlichting^{2,3}, Michael Miller⁴, Robert Mendez⁴, Serina Huang², Jinzhao Wang^{2,3}, Yinghui Sui^{2,3}, Allison Deogaygay², Mei-Lin Okino², Yunjiang Qiu³, Ying Sun², Parul Kudtarkar², Rongxin Fang³, Sebastian Preissl⁴, Maïke Sander^{2,3,5,8}✉, David U. Gorkin^{3,4,6,8}✉ and Kyle J. Gaulton^{2,5,8}✉

Single-nucleus assay for transposase-accessible chromatin using sequencing (snATAC-seq) creates new opportunities to dissect cell type-specific mechanisms of complex diseases. Since pancreatic islets are central to type 2 diabetes (T2D), we profiled 15,298 islet cells by using combinatorial barcoding snATAC-seq and identified 12 clusters, including multiple alpha, beta and delta cell states. We cataloged 228,873 accessible chromatin sites and identified transcription factors underlying lineage- and state-specific regulation. We observed state-specific enrichment of fasting glucose and T2D genome-wide association studies for beta cells and enrichment for other endocrine cell types. At T2D signals localized to islet-accessible chromatin, we prioritized variants with predicted regulatory function and co-accessibility with target genes. A causal T2D variant rs231361 at the *KCNQ1* locus had predicted effects on a beta cell enhancer co-accessible with *INS* and genome editing in embryonic stem cell-derived beta cells affected *INS* levels. Together our findings demonstrate the power of single-cell epigenomics for interpreting complex disease genetics.

Gene regulatory programs are largely orchestrated by *cis*-regulatory elements that direct the expression of genes in response to specific developmental and environmental cues. Genetic variants associated with complex disease are highly enriched within putative *cis*-regulatory elements¹. The activity of regulatory elements is often restricted to specific cell types and/or cell states, limiting the ability of ATAC-seq and other ‘ensemble’ (or ‘bulk’) epigenomic technologies to map regulatory elements in individual cell types within disease-relevant tissues. To overcome this limitation, new approaches to obtain ATAC-seq profiles from single nuclei allow for the disaggregation of open chromatin from heterogeneous samples into component cell types and subtypes^{2–4}. These developments create opportunities to dissect the molecular mechanisms that underlie genetic risk of disease. However, to date snATAC-seq data from disease-relevant human tissues are limited^{5–8}.

T2D is a multifactorial disease with highly polygenic inheritance⁹. Pancreatic islets are central to the genetic risk of T2D, as evidenced by shared association between T2D risk and quantitative measures of islet function^{10–13} and enrichment of T2D risk variants in islet regulatory sites^{14–18}. Islets consist of multiple endocrine cell types with distinct functions^{19–21} and are heterogeneous^{22–24} in gene expression and other molecular signatures, which likely reflect different functional cell states^{22,25,26}. Heterogeneity in the epigenomes of islet cell types has not been described; however, it is necessary to understand islet regulation and interpret the molecular

mechanisms of noncoding T2D risk variants. In this study, we mapped the accessible chromatin profiles of individual islet cells by using snATAC-seq, defined the regulatory programs of islet cell types and cell states, described their relationship to T2D risk and fasting glycemia and predicted the molecular mechanisms of T2D risk variants.

Results

Islet snATAC-seq reveals 12 distinct cell clusters. We performed snATAC-seq on human pancreatic islets from three donors by using a combinatorial barcoding snATAC-seq approach optimized for use on tissues^{2,4} (Supplementary Table 1). To confirm library quality, we first analyzed the data as ensemble ATAC-seq by aggregating all high-quality mapped reads irrespective of barcode. Ensemble snATAC-seq from all three samples showed the expected insert size distribution (Extended Data Fig. 1a), strong enrichment of signal at transcription start sites (TSS) (Extended Data Fig. 1b) and high concordance of signal with published islet ATAC-seq data^{14,27–29} (Extended Data Fig. 1c).

To obtain high-quality single-cell profiles, we first filtered out barcodes with fewer than 500 reads (Extended Data Fig. 1d), resulting in a total of 18,544 cells across the 3 samples. We then clustered accessible chromatin profiles from these cells, making key modifications to previous approaches⁴ (Extended Data Fig. 2). After filtering out cells with aberrant quality metrics, we retained 15,298

¹Biomedical Graduate Studies Program, University of California, San Diego, La Jolla, CA, USA. ²Department of Pediatrics, Pediatric Diabetes Research Center, University of California, San Diego, La Jolla, CA, USA. ³Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA. ⁴Center for Epigenomics, University of California, San Diego, La Jolla, CA, USA. ⁵Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA, USA. ⁶Present address: Department of Biology, Emory University, Atlanta, GA, USA. ⁷These authors contributed equally: Joshua Chiou, Chun Zeng. ⁸These authors jointly supervised this work: Maïke Sander, David U. Gorkin, Kyle J. Gaulton. ✉e-mail: masander@ucsd.edu; david.gorkin@emory.edu; kgaulton@ucsd.edu

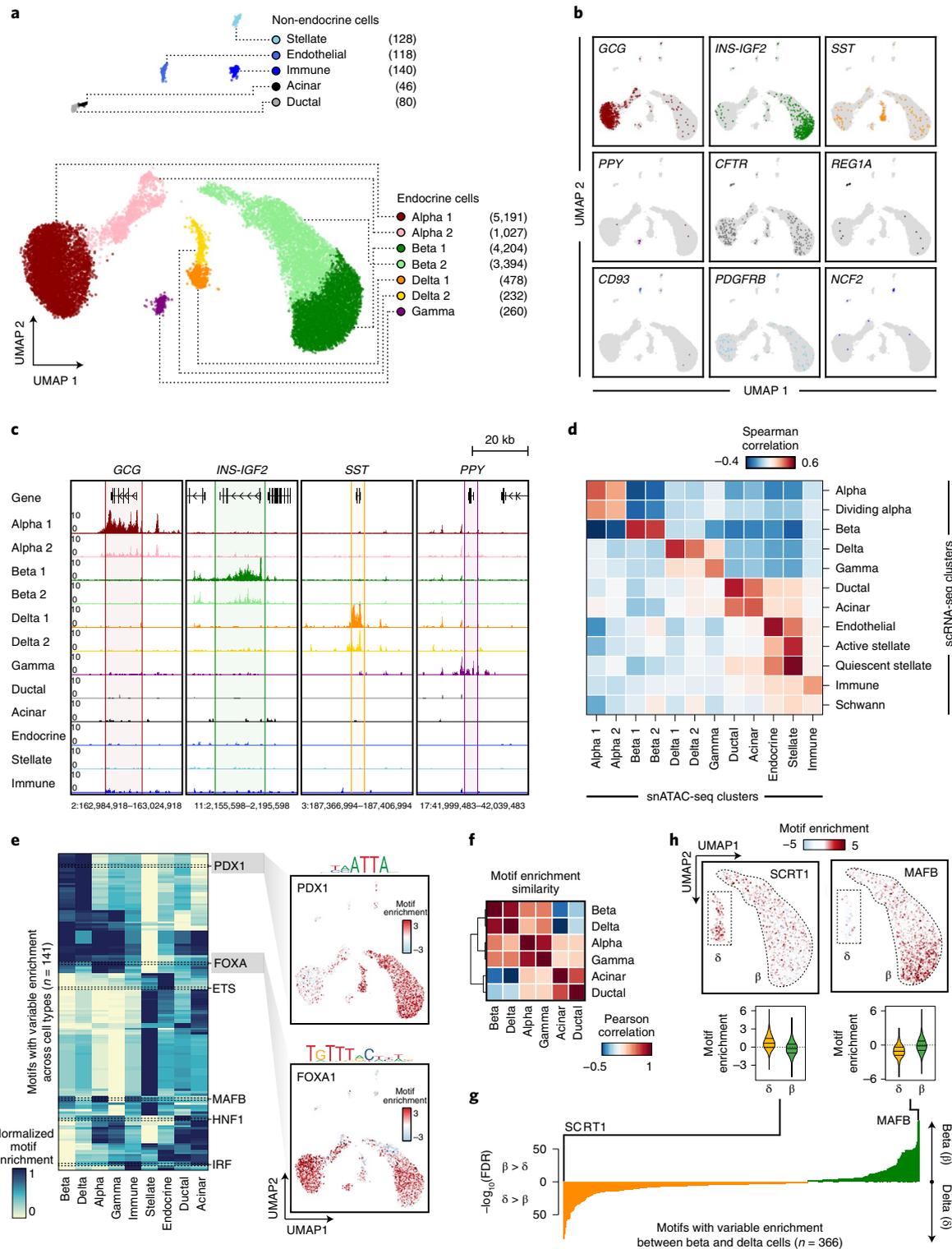


Fig. 1 | Pancreatic islet cell type-accessible chromatin defined using snATAC-seq. a, Clustering of accessible chromatin profiles from 15,298 pancreatic islet cells identifies 12 distinct clusters plotted on UMAP coordinates. The number of cells for each cluster is listed in parentheses next to the cluster label. **b**, Promoter accessibility in a 1-kb window around the TSS for selected marker genes. **c**, Aggregate read density (counts per 1×10^5) at hormone marker genes: *GCG* (alpha); *INS-IGF2* (beta); *SST* (delta); and *PPY* (gamma). **d**, Spearman correlation between *t* statistics of cluster-specific genes based on promoter accessibility (snATAC-seq) and gene expression (scRNA-seq). **e**, Row-normalized chromVAR motif enrichment z-scores for 141 transcription factor sequence motifs with variable enrichment across clusters (left). Cell types with multiple clusters were collapsed into a single cluster (for example, beta 1 + beta 2 into beta). Enrichment z-scores for FOXA1 and PDX1 motifs for each cell projected onto UMAP coordinates (right). **f**, Pearson correlation of transcription factor motif enrichment z-scores between endocrine and exocrine cell types. **g**, FDR-corrected *P* values from two-sided, two-sample *t*-tests of differential chromVAR motif enrichment comparison between delta and beta cells for 366 transcription factor motifs. **h**, Enrichment z-scores for SCRT1 and MAFB motifs in 7,598 beta and 710 delta cells projected onto UMAP coordinates (top) and shown as violin distributions (bottom; the lines represent the median and quartiles).

cells, which mapped to 12 clusters (Fig. 1a). To determine the cell type represented by each cluster, we examined chromatin accessibility at the promoter region of the cognate hormone genes for endocrine cells and known marker genes for non-endocrine cell types. We identified clusters representing beta (*INS-IGF2*/insulin), alpha (*GCG*/glucagon), delta (*SST*/somatostatin), gamma (*PPY*/pancreatic polypeptide), ductal (*CFTR*), acinar (*REG1A*), immune (*NCF2*)³⁰, stellate (*PDGFRB*)³⁰ and endothelial (*CD93*)³¹ cells (Fig. 1b,c). We defined a broader set of marker genes for each cluster by identifying promoters with accessibility most specific to each cluster (Supplementary Table 2). We observed highly specific correlations between marker genes defining cell types in snATAC-seq and islet single-cell RNA sequencing (scRNA-seq)²³ (Fig. 1d and Extended Data Fig. 3a–e).

To characterize the regulatory programs in each cell type, we aggregated reads for cells within each cluster and identified accessible chromatin sites for the cluster using MACS2 (ref. 32). In total, we identified 228,873 accessible chromatin sites merged across the 12 clusters (Supplementary Data 1). We next used chromVAR³³ to identify transcription factor motifs from the JASPAR 2018 database³⁴ enriched within the accessible chromatin of each cell. Analysis of motif enrichments averaged across cells for each cell type revealed cell type-specific patterns of motif enrichment (Fig. 1e and Supplementary Table 3). For example, we observed enrichment of PDX1 in beta and delta cells³⁵, MAF in alpha and beta cells^{36–38}, IRF in immune cells³⁹ and ETS in endothelial cells⁴⁰ (Fig. 1e). Hierarchical clustering of cell types based on motif enrichment revealed that the regulatory programs of beta and delta cells were closely related as were alpha and gamma cells (Fig. 1f), which is consistent with single-cell expression data^{30,41,42}. Motifs highly enriched in delta cell chromatin relative to beta cells included SCRT (SCRT1 $-\log_{10}$ (false discovery rate (FDR)) = 86.49) and ELF transcription factors (ELF5 $-\log_{10}$ (FDR) = 79.41); motifs enriched in gamma cell chromatin relative to alpha cells included HOXA9 $-\log_{10}$ (FDR) = 20.92) and IRF transcription factors (IRF1 $-\log_{10}$ (FDR) = 20.22) (Fig. 1g,h, Extended Data Fig. 4 and Supplementary Tables 4 and 5).

Heterogeneity in islet endocrine cell regulatory programs. A major strength of single-cell approaches is the ability to reveal heterogeneity within a cell type. Indeed, our initial clustering showed that alpha, beta and delta cells segregated into subclusters. We identified gene promoters with variable accessibility between subclusters (Methods and Supplementary Data 2). Notably, *INS* had among the most variable promoter accessibility between beta cell subclusters (*INS-IGF2* beta odds ratio (OR) = 4.74, two-sided Fisher's exact test $P = 1.78 \times 10^{-40}$); therefore, for clarity, we renamed the clusters *INS*^{high} and *INS*^{low} beta cells, respectively (Fig. 1b,c and Fig. 2a). Similarly, *GCG* promoter accessibility was variable between alpha cell subclusters (*GCG* alpha OR = 3.67, $P = 3.45 \times 10^{-22}$), which we renamed *GCG*^{high} and *GCG*^{low} alpha cells; *SST* promoter accessibility was variable between delta cell subclusters (*SST* delta OR = 1.86

$P = 0.02$), which we renamed *SST*^{high} and *SST*^{low} delta cells (Fig. 1b,c and Fig. 2a).

We found significant overlap in the genes that distinguish hormone-high from hormone-low alpha, beta and delta cells by gene set enrichment analysis (GSEA) (Fig. 2b). Genes with increased promoter accessibility in hormone-high states were enriched for hormone secretion and glucose response; in contrast, genes with increased promoter accessibility for hormone-low states were enriched for stress-induced signaling response (Fig. 2a,c and Supplementary Table 6). We also observed enriched transcription factor motifs that distinguished different states (Fig. 2d and Supplementary Table 7). For example, RFX family motifs were enriched in hormone-high but not in hormone-low states (RFX1 mean enrichment *INS*^{high} = 0.36, *INS*^{low} = -0.95, $P = 0$; *GCG*^{high} = 0.52, *GCG*^{low} = -1.16, $P = 7.3 \times 10^{-260}$; *SST*^{high} = 0.76, *SST*^{low} = -1.24, $P = 3.9 \times 10^{-58}$) (Fig. 2d). In contrast, FOS/JUN family motifs were prominently enriched in hormone-low but not hormone-high states (FOS::JUN mean enrichment *INS*^{high} = -1.78, *INS*^{low} = 3.90, $P = 0$; *GCG*^{high} = -2.86, *GCG*^{low} = 5.50, $P = 0$; *SST*^{high} = -0.21, *SST*^{low} = 7.62, $P = 1.1 \times 10^{-121}$) (Fig. 2d). These data reveal epigenomic differences between endocrine cell states among genes involved in hormone production and stress-induced signaling responses and point to an underlying commonality in regulatory networks that govern state-specific functions of endocrine cell types.

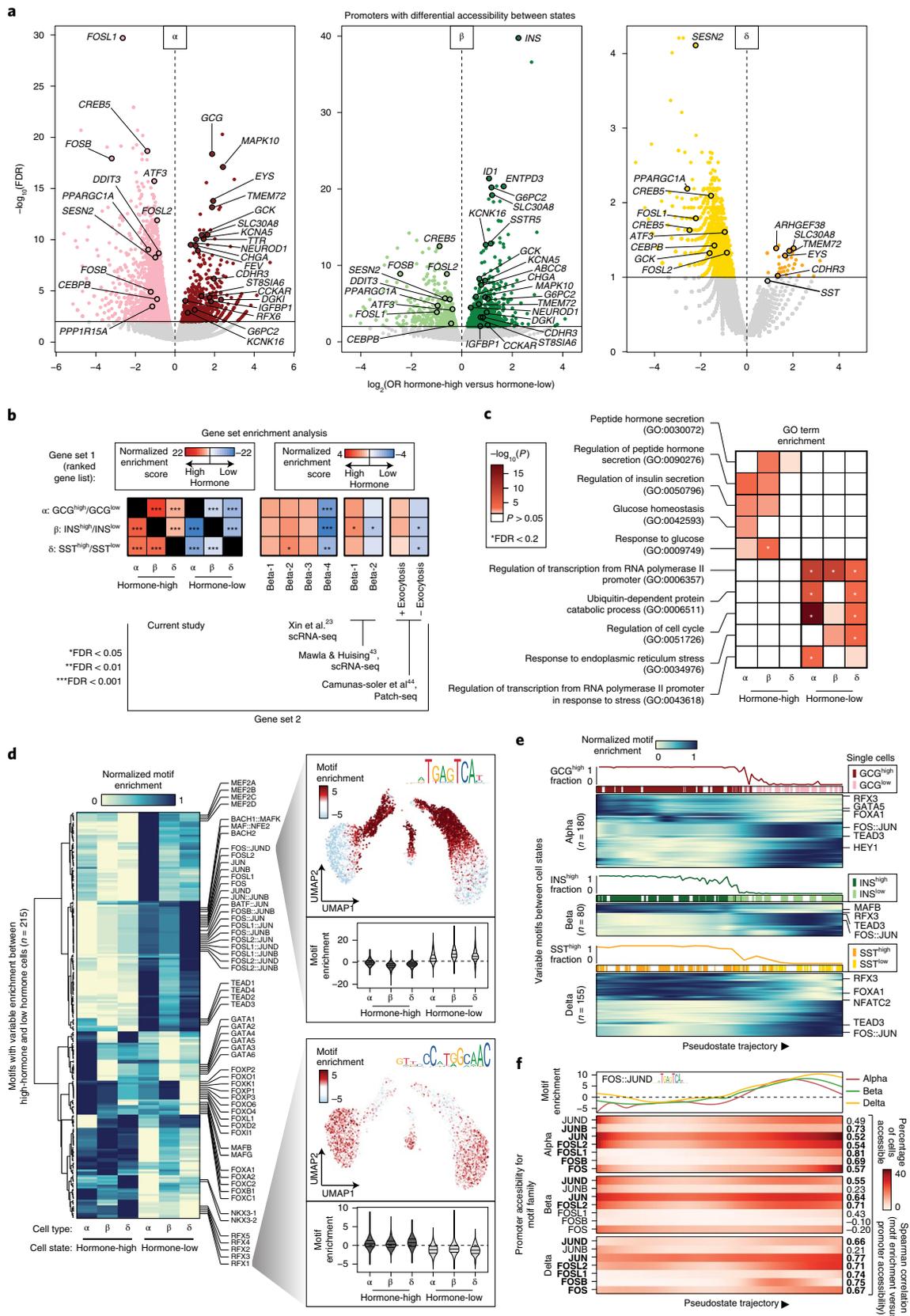
We next sought to determine whether the observed heterogeneity in the epigenomes of endocrine cells correlated with heterogeneity in islet gene expression and function. We first compared our states to beta cell subclusters from a previous scRNA-seq study²³. Genes with increased promoter accessibility in hormone-low cells were enriched in a beta cell subcluster (β -subcluster 4) associated with endoplasmic reticulum stress and protein folding and low *INS* expression, whereas genes with increased promoter accessibility in hormone-high cells were enriched in the other beta cell subclusters (β -subclusters 1–3) (Fig. 2b). We further found enrichment of genes with differential promoter accessibility among gene sets preferentially expressed in beta cell subclusters from a recent scRNA-seq meta-analysis⁴³ (Fig. 2b). Finally, we found significant overlap in genes with differential promoter accessibility between states and genes that correlate with electrophysiological measures of beta cell function from a recent Patch-seq study⁴⁴ (Fig. 2b). Thus, these results provide a link between epigenomic heterogeneity in endocrine cells and heterogeneity in gene expression and electrophysiological function.

To explore potential gradations among endocrine cells as a continuum rather than as binary states^{23,45}, we used Cicero⁸ to order cells from each cell type along trajectories based on chromatin accessibility. We observed cells on a gradient between hormone-high and hormone-low states (Fig. 2e and Extended Data Fig. 5a–c). These trajectories allowed us to examine gene promoter accessibility and transcription factor motif enrichment as a function of pseudostate (Fig. 2e and Extended Data Fig. 5d). Consistent with binary subclusters,

Fig. 2 | Heterogeneity in endocrine cell-accessible chromatin and regulatory programs. **a**, Gene promoters with significantly differential chromatin accessibility between subclusters of alpha (left), beta (middle) and delta cells (right). **b**, Enrichment of gene sets using ranked gene lists from the differential promoter analyses. Panels include genes with differential promoter accessibility between hormone-high and hormone-low states (first subpanel from the left); genes expressed in beta cell subclusters from islet scRNA-seq (second and third subpanels); and genes positively and negatively correlated with exocytosis from islet Patch-seq (fourth subpanel). **c**, Enrichment of GO terms related to glucose response, hormone secretion, stress response and cell cycle among genes with differential promoter accessibility between endocrine cell states. **d**, Row-normalized motif enrichments for 215 transcription factor motifs with variable enrichment across endocrine cell states. Single-cell motif enrichment z-scores for a representative RFX (RFX3) and FOS/JUN (FOS::JUN) motif were projected onto UMAP coordinates (right); the violin plots below show motif enrichment distribution within endocrine cell states (the lines represent the median and quartiles). **e**, Ordering of alpha, beta and delta cells across pseudostate trajectories using high *GCG/INS-IGF2/SST* promoter accessibility as the reference point. Across each trajectory, the percentage of cells in the hormone-high state and the binary cluster call of individual cells are shown above the heatmaps, which show row-normalized motif enrichment for variable motifs between cell states. **f**, Promoter accessibility for genes in the FOS/JUN motif family across pseudostate trajectories. Genes with matching promoter accessibility and motif enrichment patterns ($\rho > 0.5$) are shown in bold.

lineage-specifying genes and transcription factor family motif enrichments, such as RFX and NFAT, decreased along the trajectory from hormone-high to hormone-low cells, whereas motif enrichment in transcription factor families such as FOS/JUN increased

(Fig. 2e). Structurally related transcription factors often have similar motifs; thus, to assign motifs to specific transcription factors we correlated promoter accessibility of transcription factors within the structural subfamily with motif enrichments across the state



trajectory. Motif enrichment for the FOS/JUN family correlated with the promoter accessibility of *FOSL1*, *FOSL2* and *JUN* (Fig. 2f), supporting a role for these specific transcription factors in hormone-low cell regulation.

Islet cell type enrichment for diabetes and glycemia genome-wide association studies. Genetic variants influencing diabetes and fasting glucose level are enriched in pancreatic islet regulatory elements^{14,17,18,46}. Using our islet cell type- and state-resolved accessible chromatin profiles, we determined the enrichment of variants associated with diabetes^{9,47} and related quantitative phenotypes^{11,12,48–51} and other complex traits^{52–59}. We first determined the enrichment of variants in accessible chromatin sites for each cell type and state using stratified linkage disequilibrium (LD) score regression^{60–62}. We observed significant enrichment (FDR < 0.1) of fasting glucose level association in INS^{high} beta cells and T2D association for both INS^{high} and INS^{low} beta cell states (fasting glucose INS^{high} $z = 3.58$ FDR = 0.013; T2D INS^{high} $z = 4.41$ FDR = 0.001, INS^{low} $z = 4.19$ FDR = 0.002) (Fig. 3a). We also observed more nominal evidence for enrichment of T2D association for GCG^{low} alpha cells and both delta cell states, as well as multiple glycemic traits for endocrine cells (Fig. 3a).

In these analyses, we observed differences in enrichment between INS^{high} and INS^{low} beta cells for fasting glucose (Fig. 3a). To further resolve the heterogeneity of genetic association enrichment, we tested enrichment of T2D and fasting glucose as well as several negative control traits within single-cell profiles (Fig. 3b and Extended Data Fig. 6a). We observed marked heterogeneity among beta cells for fasting glucose-associated variants, where INS^{high} cells had significantly stronger enrichment (INS^{high} median $Z = 2.58$, INS^{low} median $Z = 0.68$, $P = 1.19 \times 10^{-225}$) (Fig. 3b). We further calculated the average enrichment for cells binned across pseudostate (Fig. 2), which revealed decreasing enrichment from INS^{high} to INS^{low} beta cells (Fig. 3b). In contrast, for T2D we observed consistent enrichment across INS^{high} and INS^{low} beta cells (INS^{high} median $Z = 0.93$, INS^{low} median $Z = 0.91$, $P = 0.44$) (Fig. 3b). Since many variants affecting fasting glucose also affect T2D^{12,13,63}, we grouped T2D loci associated with fasting glucose (T2D/fasting glucose loci) and tested these loci for enrichment of INS^{high} and INS^{low} beta cell sites using fGWAS⁶⁴. We observed strong enrichment of T2D/fasting glucose loci for INS^{high} beta cells only (INS^{high} beta $\ln(\text{enrichment}) = 4.54$, INS^{low} beta $\ln(\text{enrichment}) = -25.7$), suggesting that the T2D loci affecting glucose levels have state-specific effects (Fig. 3c). For example, at the *ADCY5* locus associated with both fasting glucose and T2D, the rs11708067:A>G variant (posterior probability of association (PPA) = 0.79) overlapped a site more specific to INS^{high} beta cells (Fig. 3d).

Outside of beta cells, we also observed evidence for enrichment of T2D association for variants in chromatin sites for other endocrine cell states including GCG^{low} alpha cells and SST^{high} and SST^{low} delta cells (Fig. 3a). To understand the potential role of these cell states in T2D risk, we performed additional enrichment analysis at known

T2D risk loci using fGWAS⁶⁴. We tested alpha, delta and gamma sites not overlapping beta cell sites for enrichment at fine-mapped T2D loci from the DIAMANTE study. Again, we observed enrichment of T2D association for GCG^{low} alpha cells ($\ln(\text{enrichment}) = 1.75$) as well as SST^{high} and SST^{low} delta cells ($\ln(\text{enrichment}) = 0.86, 1.30$) (Fig. 3e). In further support of the likely role of these cell types in T2D, several fine-mapped risk variants overlapped sites specific to these cell types; for example, rs1111875:C>T (PPA = 0.16) mapped in a delta cell-specific site at the *HHEX* locus (Fig. 3f and Supplementary Table 8).

Given our ability to map both complex trait and transcription factor motif enrichments to single cells, we reasoned that joint analysis could provide insight into transcription factors regulating trait-relevant chromatin. We correlated single-cell fasting glucose and T2D enrichment z -scores with transcription factor motif enrichments from chromVAR³³. Across all 15,298 cells, we observed positive correlation between fasting glucose and T2D enrichment and motifs for beta cell transcription factors such as PDX (Fig. 3g, Extended Data Fig. 6b and Supplementary Table 9). Across the 7,598 beta cells only, we observed the strongest positive correlation between fasting glucose and motifs in transcription factor families enriched for INS^{high} beta cells such as bHLH (NEUROD1 $\rho = 0.21$, $P = 9.43 \times 10^{-78}$) and RFX (RFX1 $\rho = 0.21$, $P = 8.83 \times 10^{-74}$) (Fig. 3g and Supplementary Table 9). For T2D, the strongest positive correlations included motifs for transcription factor families such as RFX (RFX1 $\rho = 0.052$, $P = 3.72 \times 10^{-9}$), NFAT (NFATC2 $\rho = 0.047$, $P = 4.69 \times 10^{-5}$) and MEF2 (MEF2A $\rho = 0.062$, $P = 6.99 \times 10^{-8}$) (Extended Data Fig. 6b and Supplementary Table 9). We further determined whether transcription factor motifs preferentially harbored associated variants directly. For fasting glucose, we identified the strongest enrichment for INS^{high} state-specific transcription factor motifs, most notably RFX (RFX2 $P = 1.3 \times 10^{-10}$) and NEUROD ($P = 0.049$) (Fig. 3h). For T2D, we observed enrichment for transcription factor motifs positively correlated with T2D association including RFX (RFX1 $P = 4.0 \times 10^{-15}$), NFAT (NFATC2 $P = 2.2 \times 10^{-4}$) and MEF2 (MEF2D $P = 1.9 \times 10^{-3}$) (Extended Data Fig. 6c). These motifs remained significantly enriched for T2D when considering only variants predicted to disrupt the motif (RFX1 $P = 1.1 \times 10^{-9}$, NFATC2 $P = 2.2 \times 10^{-4}$, MEF2D $P = 7.7 \times 10^{-3}$).

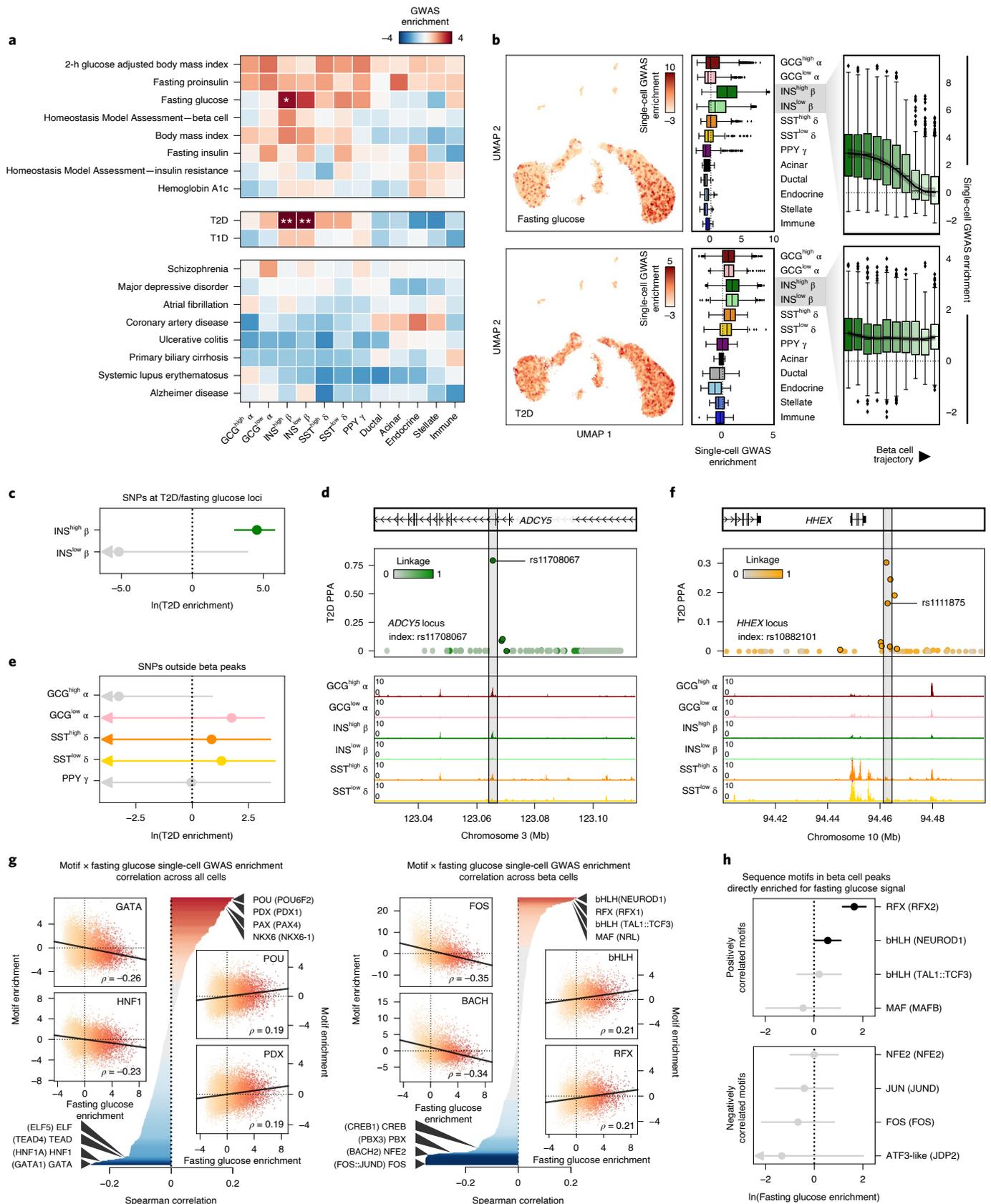
Together these results provide state-resolved insight into the role of beta cells and their transcription factors in both T2D risk and fasting glucose level and implicate other endocrine cell types in T2D risk.

Predictions of variant effects on islet cell type chromatin. Predicting the effects of noncoding genetic variants on regulatory activity is a major challenge, in large part because the sequence vocabularies that encode regulatory function differ between cell types and states. Therefore, we used deltaSVM⁶⁵ to predict the effects of genetic variants from the Haplotype Reference Consortium panel⁶⁶ on chromatin accessibility in each endocrine cell type and

Fig. 3 | Enrichment of islet-accessible chromatin for diabetes and fasting glycemia. **a**, Stratified LD score regression enrichment z -scores for diabetes-related quantitative endophenotypes (top), T1D and T2D diabetes (middle) and control traits (bottom) for islet cell types. **FDR < 0.01, *FDR < 0.1. **b**, Single-cell enrichment z -scores for fasting glucose level and T2D projected onto UMAP coordinates (left), enrichment per cell type (middle) and beta cell enrichment split into 10 trajectory bins (right). The box plot center line, limits and whiskers represent the median, quartiles and 1.5 \times interquartile range, respectively. **c**, Enrichment (estimate \pm 95% confidence interval (CI)) by fGWAS of variants at loci associated with both T2D and fasting glucose (T2D/fasting glucose) within beta cell-accessible chromatin. **d**, The candidate causal T2D variant rs11708067 overlaps an enhancer active in INS^{high} beta cells at the *ADCY5* locus, which is consistent with beta cell enrichment patterns for T2D/fasting glucose loci. **e**, Enrichment (estimate \pm 95% CI by fGWAS) of variants at T2D loci in accessible chromatin for non-beta endocrine cells after removing beta-accessible chromatin. **f**, The candidate causal T2D variant rs1111875 overlaps a delta cell-specific site at the *HHEX* locus. **g**, Correlation between single-cell fasting glucose and transcription factor motif enrichments across all 15,298 cells (left) and 7,598 beta cells (right). Across all cells, fasting glucose had positive correlations with beta-enriched transcription factor families such as PDX, NKX6 and PAX. Within beta cells, fasting glucose had positive correlations with INS^{high} beta-enriched transcription factor families such as RFX, MAF/NRL and FOXA. **h**, Enrichment (effect \pm s.e.) of fasting glucose-associated variants directly overlapping sequence motifs for those either positively or negatively correlated with fasting glucose in beta cells.

cell state. We identified 432,072 variants genome-wide with predicted allelic effects (FDR < 0.1), encompassing between 115,000 and 161,000 variants (7.8–10.9% of tested variants) per cell type or state (Fig. 4a and Supplementary Data 3).

To validate that our predictions captured true allelic effects on islet chromatin accessibility, we first compared alpha and beta cell predictions to direct measurements of allelic effects on chromatin accessibility. We found significant correlations between predicted



allelic effects and allelic imbalance estimates for all alpha and beta cell states (GCG^{high} Spearman $\rho=0.255$, $P=1.20\times 10^{-34}$, GCG^{low} $\rho=0.214$, $P=2.35\times 10^{-7}$, INS^{high} $\rho=0.275$, $P=1.03\times 10^{-34}$, INS^{low} $\rho=0.334$, $P=4.73\times 10^{-38}$) (Fig. 4b). We further validated five likely causal T2D variants predicted to affect beta cell chromatin that had directionally consistent effects on enhancer activity using gene reporter assays in the MIN6 beta cell line (Fig. 4c). We also compared predictions to islet chromatin accessibility quantitative trait loci (caQTLs)²⁸ and observed significant enrichment of caQTLs among variants with predicted effects in alpha or beta cells (observed=38.4%, expected=19.7%, two-sided Fisher's exact test $P=2.78\times 10^{-97}$) (Fig. 4d). When subdividing predictions based on shared, cell type- or state-specific effects, we observed significant enrichment of caQTLs only among shared effect variants (Fig. 4d).

We further characterized genetic variants predicted to have cell type- and state-dependent effects on islet chromatin. Variants with state-specific effects tended to disrupt motifs for transcription factor families such as NEUROD and RFX for hormone-high states ($-\log_{10}(P)=59.0$, 24.5) (Fig. 4e). Similarly, variants with alpha cell- or beta cell-specific effects tended to disrupt motifs for lineage-defining transcription factor families including GATA for alpha cells ($-\log_{10}(P)=21.6$) and NKX6 and PDX1 for beta cells ($-\log_{10}(P)=11.2$, 10.3) (Fig. 4e). To assign motifs to specific transcription factors, we examined promoter accessibility of transcription factors within the structural transcription factor subfamily⁶⁷. Among GATA subfamily members only GATA6 had high promoter accessibility in alpha cells ($GCG^{\text{high}}=1.00$, $GCG^{\text{low}}=0.96$, $INS^{\text{high}}=0.22$, $INS^{\text{low}}=0.14$), suggesting that GATA6 binding is likely disrupted in alpha cells. Similarly, among RFX family members RFX6 had promoter accessibility in hormone-high state cells ($GCG^{\text{high}}=0.92$, $GCG^{\text{low}}=0.70$, $INS^{\text{high}}=0.93$, $INS^{\text{low}}=0.80$) (Fig. 4e).

We evaluated whether our predictions could prioritize lower-frequency (defined as a minor allele frequency (MAF) < 0.05) functional variants involved in T2D risk. We observed enrichment of genome-wide-significant T2D associations among lower-frequency variants with predicted effects in any endocrine cell type compared to background (Fig. 4f). When considering each cell type, we observed enrichment of T2D association among variants with predicted effects in beta and delta cells, even at subgenome-wide-significant P values (Fig. 4f). For example, at the *IGF2BP3* locus, rs78840640:C>G (MAF=0.02) had allelic effects on beta cell chromatin (INS^{high} beta FDR=1.15 $\times 10^{-4}$, INS^{low} beta FDR=6.93 $\times 10^{-3}$) and fine-mapping supported a possible causal role in T2D (PPA=0.33) (Fig. 4g). This variant affected enhancer activity in gene reporter assays where the alternate (and T2D risk) allele G had reduced activity (Fig. 4c). These results reveal that cell type-specific chromatin can provide accurate functional predictions of lower-frequency variants.

Co-accessibility links regulatory variants to target genes.

Defining the genes affected by regulatory element activity is a major challenge because enhancers can regulate gene activity over large, nonadjacent distances⁶⁸. A number of approaches have been

developed to link regulatory elements to target genes^{69,70} but are not typically cell type resolved^{27,71}. Recently, a new approach was developed to link regulatory elements at cell type resolution based on co-accessibility across single cells⁸. Thus, we leveraged single-cell accessible chromatin profiles to define co-accessibility between accessible chromatin sites in alpha, beta and delta cells.

To calibrate the extent to which co-accessibility reflected physical interactions between regulatory elements, we performed a distance-matched comparison between co-accessible sites stratified by co-accessibility threshold and chromatin loops identified from Hi-C and promoter capture Hi-C (pcHi-C) in primary islets^{27,71}. We observed strong enrichment for sites with co-accessibility scores > 0.05 for islet chromatin loops compared to non-co-accessible sites (Fig. 5a and Extended Data Fig. 7a–e). Therefore, we used this threshold (0.05) to define co-accessible sites (Supplementary Data 4–6). Among co-accessible sites were 47,871 (alpha), 46,036 (beta) and 42,234 (delta) distal sites co-accessible with a gene promoter (Extended Data Fig. 7f,g); most (71.9%) were cell type-specific (Extended Data Fig. 8a–c). For example, the *PDX1* promoter was co-accessible with 31 sites in beta and 39 sites in delta cells, including sites that directly coincided with islet pcHi-C, only 7 of which were in alpha cells (Fig. 5b).

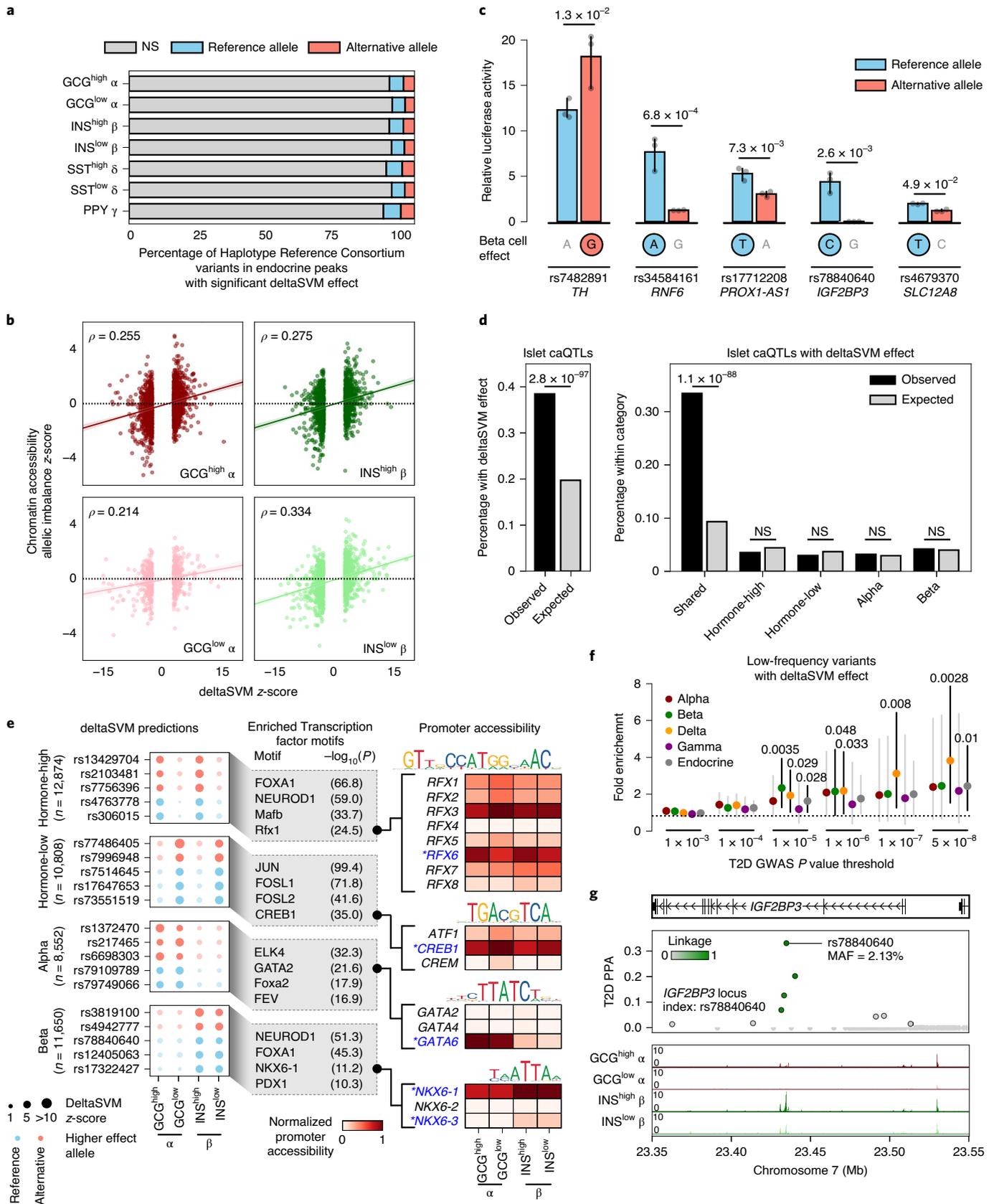
Distal sites with co-accessibility links to gene promoters harbored risk variants for T2D at many loci (Supplementary Table 8). At the *KCNQ1* locus, an islet chromatin site located in intron 3 of *KCNQ1* had beta cell-specific co-accessibility with the *INS* promoter over 500 kilobases (kb) distally and harbored a causal T2D variant rs231361:G>A (PPA=1) (Fig. 5c). The site containing rs231361 was more accessible in INS^{high} compared to INS^{low} beta cells and rs231361 was predicted to have state-specific effects on beta cell chromatin accessibility (INS^{high} beta FDR=0.060; INS^{low} beta FDR=0.40). Furthermore, rs231361 disrupted an *RFX* motif, which itself was enriched in INS^{high} beta cells (Fig. 5c). Published chromosome conformation capture-on-chip (4C) data from the EndoC- β H1 human beta cell line⁷² revealed physical proximity between this site and the *INS* promoter (Extended Data Fig. 9) but there was not similar evidence in Hi-C or other chromosome conformation capture (3C)-based data from EndoC- β H1 and primary islets^{27,71,73}.

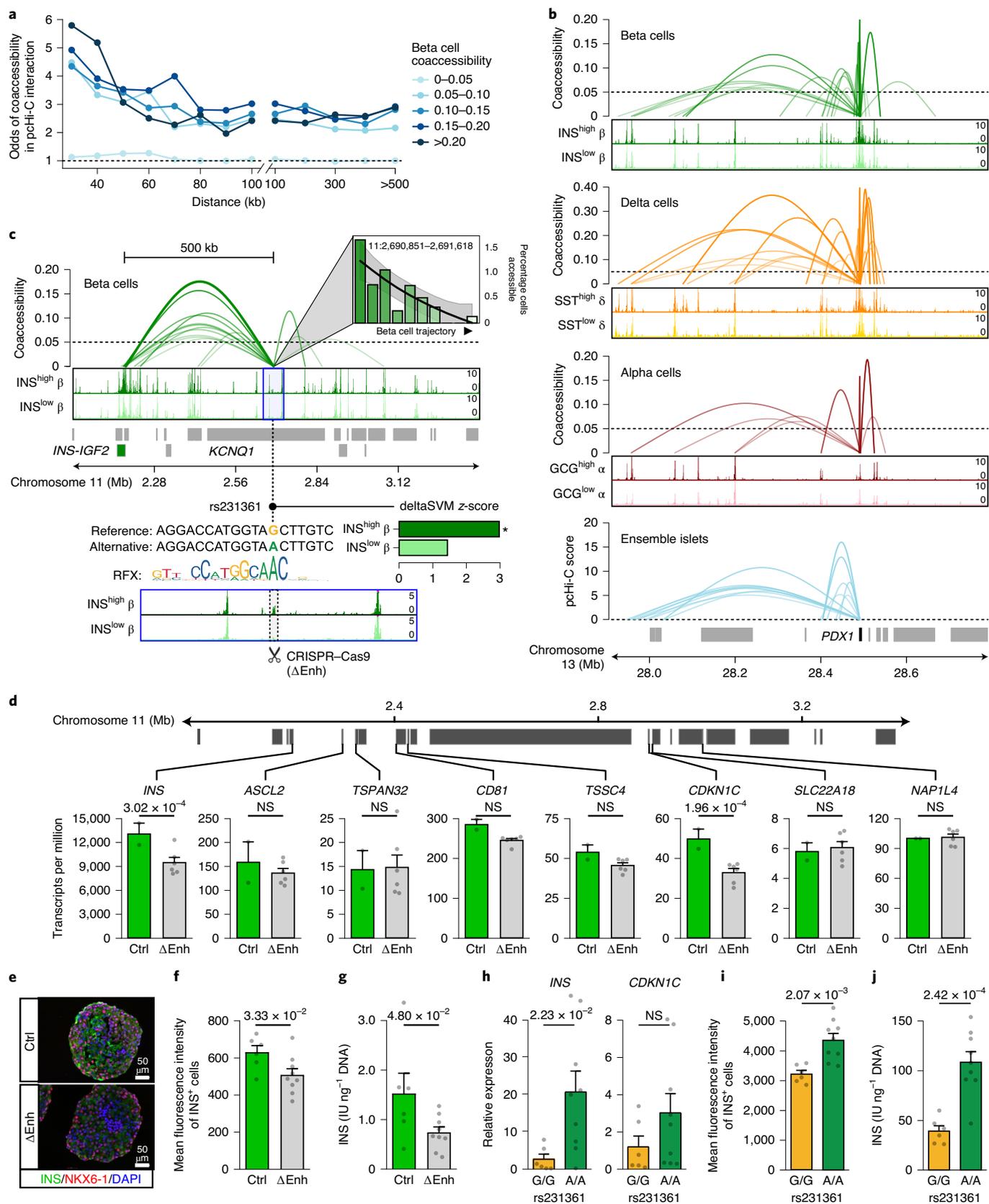
Although we observed physical proximity between rs231361 and the *INS* promoter in beta cells, the absence of a canonical chromatin loop necessitated further validation. Therefore, we deleted a 2.6-kb region flanking the site in human embryonic stem cells (hESCs) by CRISPR-Cas9-mediated genome editing in three biallelic clones (*KCNQ1* ^{Δ Enh}) (Fig. 5c and Extended Data Fig. 10a,b) and differentiated *KCNQ1* ^{Δ Enh} clones and unedited control clones into beta cells by using a modified version of an established protocol⁷⁴. Analysis of cultures at the beta cell stage revealed similar numbers of INS^+ / $NKX6-1^+$ cells in *KCNQ1* ^{Δ Enh} and control clones (47.1 \pm 13.4% versus 56.5 \pm 7.6%) (Extended Data Fig. 10c), suggesting that deletion of the enhancer had no effect on beta cell differentiation. We determined the effects of enhancer deletion on expression of all genes within 2 Mb of the enhancer and observed a significant

Fig. 4 | Genetic variants with islet cell type- and state-specific effects on chromatin accessibility. **a**, Percentage of Haplotype Reference Consortium variants in any endocrine cell type peak ($n=1,411,387$ variants) that had significant deltaSVM effects (FDR < 0.1) for the reference or alternative allele. **b**, Spearman correlation between deltaSVM z-scores and chromatin accessibility allelic imbalance z-scores for variants with predicted effects in alpha and beta states. **c**, Relative luciferase reporter activity (mean \pm 95% CI; $n=3$ replicates) for 5 T2D variants with predicted beta cell effects. The allele with predicted effect is circled. P values by two-sided Student's t -tests. **d**, Enrichment of islet caQTLs for variants with predicted effects in alpha and beta cells (left) and stratified based on shared, cell type- and state-specific effects (right). P values by two-sided Fisher's exact test. NS, not significant. **e**, Examples of variants with predicted effects in alpha and beta cells (left). Transcription factor motif families enriched in sequences surrounding the effect allele relative to the non-effect allele (middle) and promoter accessibility patterns for genes in enriched transcription factor motif families (right). Genes with matching promoter accessibility and transcription factor motif enrichment patterns are highlighted. **f**, Enrichment (estimate \pm 95% CI) of low-frequency and rare variants with predicted effects on islet chromatin at different T2D association thresholds. P values by two-sided binomial test. **g**, The low-frequency T2D variant rs78840640 at the *IGF2BP3* locus with high causal probability (PPA = 0.33) has predicted effects in beta cells.

decrease in the expression of *INS* ($P = 3.02 \times 10^{-4}$; FDR = 0.066) and *CDKN1C* ($P = 1.96 \times 10^{-4}$; FDR = 0.059) in *KCNQ1*^{ΔEnh} compared to control cells and not for other genes (all $P > 0.05$) (Fig. 5d and

Supplementary Table 10). Analysis of *INS* protein levels by immunofluorescence staining, flow cytometry and ELISA further revealed reduced *INS* protein abundance in *KCNQ1*^{ΔEnh} beta cells (Fig. 5e–g).





To determine whether rs231261 itself had distal effects on *INS* regulation in beta cells, we used targeted base editing to generate hESC lines that were homozygous for either the major allele

G (*KCNQ1*^{G/G}, two clones) or the minor (and T2D risk) allele A (*KCNQ1*^{A/A}, three clones) (Extended Data Fig. 10d–f). We then differentiated the *KCNQ1*^{A/A} and *KCNQ1*^{G/G} clones into beta cells by

Fig. 5 | Chromatin co-accessibility links diabetes risk variants to target genes. **a**, Distance-matched odds that co-accessible sites in beta cells overlap islet pcHi-C interactions at different threshold bins. **b**, Single-cell co-accessibility and islet pcHi-C interactions at the *PDX1* promoter. **c**, Enhancer harboring the T2D variant rs231361 shows distal beta cell co-accessibility to the *INS* promoter and other non-promoter sites. The enhancer is accessible in *INS*^{high} but not *INS*^{low} beta and has decreasing accessibility across the beta cell trajectory. rs231361 disrupts an *RFX* motif and has predicted effects in *INS*^{high} beta. *FDR < 0.1. CRISPR-Cas9 deletion of a 2.6-kb region (highlighted, ΔEnh) is shown around the enhancer. **d**, Expression (transcripts per million ± s.e.m.) of genes within 2 Mb of the enhancer in beta cell stage cultures for ΔEnh (*n* = 6; 3 clones × 2 differentiations) and control (*n* = 2; 1 clone × 2 differentiations). *P* values from DESeq2. **e**, Representative immunofluorescence staining for *INS* (green), *NKX6-1* (red) and DAPI staining (blue) in beta cell stage cultures for ΔEnh (*n* = 3 experiments) and control (*n* = 3 experiments). **f,g**, Quantification of *INS* mean fluorescence intensity (**f**) and *INS* content (**g**) in beta cell stage cultures for ΔEnh (*n* = 9; 3 clones × 3 differentiations) and control (*n* = 6; 2 clones × 3 differentiations). **h**, Relative expression of *INS* and *CDKN1C*. **i,j**, Quantification of *INS* mean fluorescence intensity (**i**) and *INS* content (**j**) in beta cell stage cultures for *KCNQ1*^{Δ^{Enh}} (*n* = 9; 3 clones × 3 differentiations) and *KCNQ1*^{G/G} (*n* = 6; 2 clones × 3 differentiations). Data shown are the mean ± s.e.m., *P* values obtained by two-sided Student's *t*-test.

using the same protocol as for *KCNQ1*^{Δ^{Enh}} with additional modifications. We determined the effects of the variant alleles on expression of the two genes *INS* and *CDKN1C* with significant changes in the enhancer deletion (Fig. 5d) using quantitative PCR (qPCR). We observed a significant difference in *INS* expression between alleles (*P* = 0.022) and evidence for a difference in *CDKN1C* expression although this was not significant (*P* > 0.05) (Fig. 5h). Analysis of *INS* protein levels by flow cytometry and ELISA revealed significant differences in *INS* abundance between variant alleles (Fig. 5i,j).

Discussion

Over 400 known risk signals for T2D have been identified, yet only a handful have been characterized molecularly^{16,17,27,75–81}. Our findings provide a road map demonstrating how single-cell accessible chromatin derived from disease-relevant primary tissue can be utilized to define cell types, cell states, *cis*-regulatory elements and genes involved in the genetic basis of T2D and other complex diseases.

The *KCNQ1* locus has a complex contribution to T2D involving at least ten independent signals⁹. Among these was the causal T2D variant rs231361, where genome editing in hESC beta cell models revealed the effects on insulin transcript and protein levels. Chromatin conformation from 4C in EndoC-βH1 cells⁷² revealed physical proximity between the enhancer and *INS* promoter, although we did not find corresponding evidence for an interaction in other 3C-based data from EndoC-βH1 cells⁷³, hESC beta cells or primary islets^{27,71}. Thus, while our results support a possible *cis*-regulatory effect of rs231361 on *INS* expression, we cannot currently rule out that the observed effects instead occur in *trans*, secondary to other effects. We anticipate that future studies to resolve phase between variant alleles and *INS* allelic expression in heterozygous samples will further clarify the nature of these effects.

Single-cell accessible chromatin uncovered heterogeneity in the regulatory programs of endocrine cell types, pointing to transcription factors that likely drive cell state-specific functions. Integrating single-cell heterogeneity with large-scale genetic association data revealed that genetic variants modulating fasting glucose levels likely act through the high insulin-producing beta cell state, whereas genetic risk of T2D is mediated through effects on both the high insulin-producing state and other functional beta cell state(s) likely related to stress and signaling responses. Moreover, given similar heterogeneity in the epigenomes of alpha and delta cells, our results reveal that endocrine cell regulation involves both lineage-specific programs and an additional layer of state-specific programs common across endocrine cell types. Our results also implicate these other endocrine cell types in the genetic risk of T2D independent of beta cells, most prominently delta cells.

Previous studies have characterized heterogeneity in beta cell physiological function, cell surface markers and gene expression^{22,44,82–84}. We found that heterogeneity in the beta cell epigenome mapped to cellular states related to insulin production and stress-related signaling response and was tightly linked to heterogeneity in beta cell gene expression²³ and electrophysiology⁴⁴. However,

there is often not perfect correspondence between subclusters identified by different techniques and/or studies and we anticipate that using multi-omics methods will help to further clarify these differences. Regardless of the technology used, heterogeneity defined from single-cell data is by nature dependent on computational clustering or ordering. Therefore, separating true heterogeneity from other substructure ultimately requires experimental validation, for example, by profiling cell populations sorted using subtype markers. Since the subclusters described in our study have not been linked directly to functional differences, experiments will be necessary to determine the relationship between epigenomic heterogeneity and cellular function. Furthermore, while we observed heterogeneity in endocrine cells from cryopreserved tissue in addition to purified islets, determining the true extent and nature of heterogeneity in vivo will require more extensive studies across a broader range of samples and conditions.

In summary, we have presented a detailed characterization of islet cell type and state regulatory programs. When combined with genetic fine-mapping and whole-genome sequencing, as well as additional cell type-specific data in islets⁸⁵, this resource will greatly enhance efforts to define the molecular mechanisms of T2D risk. More broadly, our study provides a framework for using single-cell chromatin from disease-relevant tissues to interpret the genetic mechanisms of complex disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-021-00823-0>.

Received: 24 May 2019; Accepted: 18 February 2021;

Published online: 01 April 2021

References

1. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
3. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
4. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
5. Litzgenburger, U. M. et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol.* **18**, 15 (2017).
6. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548.e16 (2018).
7. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
8. Pliner, H. A. et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).

9. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
10. Wood, A. R. et al. A genome-wide association study of IVGTT-based measures of first-phase insulin secretion refines the underlying physiology of type 2 diabetes variants. *Diabetes* **66**, 2296–2309 (2017).
11. Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
12. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
13. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
14. Thurner, M. et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *eLife* **7**, e31977 (2018).
15. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
16. Gaulton, K. J. et al. A map of open chromatin in human pancreatic islets. *Nat. Genet.* **42**, 255–259 (2010).
17. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
18. Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
19. van der Meulen, T. et al. Urocortin3 mediates somatostatin-dependent negative feedback control of insulin secretion. *Nat. Med.* **21**, 769–776 (2015).
20. Caicedo, A. Paracrine and autocrine interactions in the human islet: more than meets the eye. *Semin. Cell Dev. Biol.* **24**, 11–21 (2013).
21. DiGruccio, M. R. et al. Comprehensive alpha, beta and delta cell transcriptomes reveal that ghrelin selectively activates delta cells and promotes somatostatin release from pancreatic islets. *Mol. Metab.* **5**, 449–458 (2016).
22. Dorrell, C. et al. Human islets contain four distinct subtypes of β cells. *Nat. Commun.* **7**, 11756 (2016).
23. Xin, Y. et al. Pseudotime ordering of single human β -cells reveals states of insulin production and unfolded protein response. *Diabetes* **67**, 1783–1794 (2018).
24. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
25. Bader, E. et al. Identification of proliferative and mature β -cells in the islets of Langerhans. *Nature* **535**, 430–434 (2016).
26. Liu, J. S. E. & Hebrok, M. All mixed up: defining roles for β -cell subtypes in mature islets. *Genes Dev.* **31**, 228–240 (2017).
27. Greenwald, W. W. et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.* **10**, 2078 (2019).
28. Khetan, S. et al. Type 2 diabetes-associated genetic variants regulate chromatin accessibility in human islets. *Diabetes* **67**, 2466–2477 (2018).
29. Varshney, A. et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl Acad. Sci. USA* **114**, 2301–2306 (2017).
30. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360.e4 (2016).
31. Galvagni, F. et al. CD93 and dystroglycan cooperation in human endothelial cell adhesion and migration. *Oncotarget* **7**, 10090–10103 (2016).
32. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
33. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
34. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D1284 (2018).
35. Wilson, M. E., Scheel, D. & German, M. S. Gene expression cascades in pancreatic development. *Mech. Dev.* **120**, 65–80 (2003).
36. Conrad, E. et al. The MAFB transcription factor impacts islet α -cell function in rodents and represents a unique signature of primate islet β -cells. *Am. J. Physiol. Endocrinol. Metab.* **310**, E91–E102 (2016).
37. Katoh, M. C. MafB is critical for glucagon production and secretion in mouse pancreatic α cells in vivo. *Mol. Cell. Biol.* **38**, e00504-17 (2018).
38. Nishimura, W., Takahashi, S. & Yasuda, K. MafA is critical for maintenance of the mature beta cell phenotype in mice. *Diabetologia* **58**, 566–574 (2015).
39. Ozato, K., Tailor, P. & Kubota, T. The interferon regulatory factor family in host defense: mechanism of action. *J. Biol. Chem.* **282**, 20065–20069 (2007).
40. De Val, S. & Black, B. L. Transcriptional control of endothelial cell development. *Dev. Cell* **16**, 180–195 (2009).
41. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
42. Lawlor, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).
43. Mawla, A. M. & Huising, M. O. Navigating the depths and avoiding the shallows of pancreatic islet cell transcriptomes. *Diabetes* **68**, 1380–1393 (2019).
44. Camunas-Soler, J. et al. Patch-seq links single-cell transcriptomes to human islet dysfunction in diabetes. *Cell Metab.* **31**, 1017–1031.e4 (2020).
45. Trapnell, C. et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotechnol.* **32**, 381–386 (2014).
46. Parker, S. C. J. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110**, 17921–17926 (2013).
47. Aylward, A., Chiou, J., Okino, M.-L., Kadakia, N. & Gaulton, K. J. Shared genetic risk contributes to type 1 and type 2 diabetes etiology. *Hum. Mol. Genet.* <https://doi.org/10.1093/hmg/ddy314> (2018).
48. Strawbridge, R. J. et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624–2634 (2011).
49. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
50. Saxena, R. et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42**, 142–148 (2010).
51. Wheeler, E. et al. Impact of common genetic determinants of hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: a transethnic genome-wide meta-analysis. *PLoS Med.* **14**, e1002383 (2017).
52. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
53. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
54. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
55. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. *Nat. Genet.* **45**, 1452–1458 (2013).
56. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
57. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
58. Nielsen, J. B. et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
59. Cordell, H. J. et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
60. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
61. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
62. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
63. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).
64. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
65. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
66. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
67. Wingender, E., Schoeps, T., Haubrock, M. & Dönitz, J. TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.* **43**, D97–D102 (2015).
68. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
69. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743–755 (2016).
70. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

71. Miguel-Escalada, I. et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* **51**, 1137–1148 (2019).
 72. Jian, X. & Felsenfeld, G. *Insulin* promoter in human pancreatic β cells contacts diabetes susceptibility loci and regulates genes affecting insulin metabolism. *Proc. Natl Acad. Sci. USA* **115**, E4633–E4641 (2018).
 73. Lawlor, N. et al. Multiomic profiling identifies *cis*-regulatory networks underlying human pancreatic β cell identity and function. *Cell Rep.* **26**, 788–801.e6 (2019).
 74. Rezania, A. et al. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat. Biotechnol.* **32**, 1121–1133 (2014).
 75. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
 76. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the *CDC123/CAMK1D* type 2 diabetes GWAS locus. *PLoS Genet.* **10**, e1004633 (2014).
 77. Rusu, V. et al. Type 2 diabetes variants disrupt function of SLC16A11 through two distinct mechanisms. *Cell* **170**, 199–212.e20 (2017).
 78. Carrat, G. R. et al. Decreased *STARD10* expression is associated with defective insulin secretion in humans and mice. *Am. J. Hum. Genet.* **100**, 238–256 (2017).
 79. Claussnitzer, M. et al. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* **156**, 343–358 (2014).
 80. Roman, T. S. et al. A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the *ADCY5* locus. *Diabetes* **66**, 2521–2530 (2017).
 81. Kycia, I. et al. A common type 2 diabetes risk variant potentiates activity of an evolutionarily conserved islet stretch enhancer and increases *C2CD4A* and *C2CD4B* expression. *Am. J. Hum. Genet.* **102**, 620–635 (2018).
 82. Avrahami, D., Klochendler, A., Dor, Y. & Glaser, B. Beta cell heterogeneity: an evolving concept. *Diabetologia* **60**, 1363–1369 (2017).
 83. Modi, H. et al. *Ins2* gene bursting activity defines a mature β -cell state. Preprint at *bioRxiv* <https://doi.org/10.1101/702589> (2019).
 84. Farack, L. et al. Transcriptional heterogeneity of beta cells in the intact pancreas. *Dev. Cell* **48**, 115–125.e4 (2019).
 85. Rai, V. et al. Single-cell ATAC-seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. *Mol. Metab.* **32**, 109–121 (2020).
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- © The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Islet processing and nuclei isolation. We obtained islet preparations for three donors from the Integrated Islet Distribution Program (Supplementary Table 1). Islet preparations were further enriched using zinc-dithizone staining followed by hand picking. For experiments performed with whole pancreas tissue, a cryopreserved tissue sample was obtained from the Network for Pancreatic Organ Donors with Diabetes biorepository. Studies were given exempt status by the institutional review board of the University of California, San Diego.

Generation of snATAC-seq libraries. Combinatorial barcoding snATAC-seq was performed as described previously^{23,4} with several modifications (described in the Supplementary Methods). Libraries were quantified using a Qubit 3 fluorimeter (Thermo Fisher Scientific) and the nucleosomal pattern was verified using a TapeStation (High Sensitivity D1000; Agilent Technologies). Libraries were sequenced on a HiSeq 2500 system (Illumina) using custom sequencing primers, 25% spike-in library and the following read lengths: 50 + 43 + 40 + 50 (Read1 + Index1 + Index2 + Read2).

The pancreas tissue preparation is described in the Supplementary Methods. Droplet-based single-cell ATAC-seq libraries were generated using the Chromium Chip E Single Cell ATAC Kit (PN-1000086; 10x Genomics) and indexes (Chromium i7 Multiplex Kit N, Set A; PN-1000084; 10x Genomics) according to the manufacturer's instructions. Final libraries were quantified using a Qubit fluorimeter and the nucleosomal pattern was verified using a TapeStation. Libraries were sequenced on a NextSeq 500 and HiSeq 4000 sequencer (Illumina) with the following read lengths: 50 + 8 + 16 + 50 (Read1 + Index1 + Index2 + Read2).

Raw data processing and quality control. For each read, we first appended the cell barcode metadata to the read name. The cell barcode consisted of four pieces (P7, I7, I5, P5), which were derived from the index read files. We first corrected for sequencing errors by calculating the Levenshtein distance between each of the four pieces and a whitelist of possible sequences. If the piece did not perfectly match a whitelisted sequence, we took the best matching sequence if it was within two edits and the next matching sequence was at least two additional edits away. If none of these conditions were met, we discarded the read from further analysis. The sequence processing steps are provided in the Supplementary Methods.

Cluster analysis for snATAC-seq. We split the genome into 5-kb windows and removed windows overlapping blacklisted regions (v.2) from the Encyclopedia of DNA Elements (ENCODE)^{86,87}. For each experiment, we created a sparse $m \times n$ matrix containing the read depth for m cells passing read depth thresholds at n windows. Using SCANPY⁸⁸ v.1.4.4.post1, we extracted highly variable windows using mean read depth and normalized dispersion ($\text{min_mean}=0.01$, $\text{min_disp}=0.25$). After normalization to uniform read depth and log-transformation, for each experiment, we regressed out the log-transformed read depth within highly variable windows for each cell. We then performed principal component analysis and extracted the top 50 principal components. We used Harmony²⁴ v.0.1.0 to correct the principal components and remove batch effects across experiments, using donor of origin as a covariate. We used Harmony-corrected components to calculate the nearest 30 neighbors using the cosine metric, which were subsequently used for uniform manifold approximation and projection (UMAP) dimensionality reduction ($\text{min_dist}=0.3$) and Leiden clustering⁸⁹ (resolution=1.5).

We performed iterative clustering to identify and remove cells with abnormal features before the final clustering results (Supplementary Methods). After removing these cells, we ended up with 15,298 cells mapping to 12 clusters. We used chromatin accessibility at windows overlapping promoters for marker hormones to assign cell types for the endocrine islet cell types and chromatin accessibility at windows around marker genes from scRNA-seq to assign cluster labels for non-endocrine islet clusters.

Comparison to bulk and sorted islet ATAC-seq. We processed the sequence data of bulk islet ATAC-seq^{14,27–29} and bulk pancreas from ENCODE⁸⁶ (Supplementary Methods). We calculated the Spearman correlation between normalized read coverages and used hierarchical clustering to assess similarity between bulk islet samples. To check peak call overlap between aggregated snATAC-seq and bulk ATAC-seq, we split peaks based into promoter proximal (± 500 base pairs (bp) from the TSS) and distal peaks based on promoter overlap. For each cluster, we calculated the percentage of aggregate peaks that overlapped merged autosomal bulk peaks and individual sample-level autosomal bulk peaks. We processed the raw sequence data of ATAC-seq from flow-sorted pancreatic cells from two previous studies^{90,91} (Supplementary Methods). We calculated Spearman correlations between read coverages and used hierarchical clustering to assess similarity between sorted and snATAC-seq samples.

Identifying marker peaks of chromatin accessibility. To identify peaks for each cell type, we aggregated reads for all cells within a cluster. We shifted reads aligning to the positive strand +4 bp and reads aligning to the negative strand –5 bp, extended reads to 200 bp and centered the reads. We used MACS2 (ref. ³³) v.2.1.2 to call peaks for each cluster (–nomodel–keep–dup all). We removed peaks that overlapped the ENCODE blacklist v.2 (refs. ^{86,87}). We then used BEDTools⁹² v.2.26.0

to merge peaks from clusters and create a set of merged peaks. We generated a sparse $m \times n$ matrix containing binary overlap between m peaks in the merged set of islet regulatory peaks and n cells. We calculated the t statistics of peak specificity for each cluster by using linear regression models. For each peak and each cluster, we used binary encoding of read overlap with the peak as the outcome and whether a cell was in the cluster as the predictor; we included the log read depth of each cell as a covariate in the model.

Matching islet snATAC-seq with scRNA-seq clusters. To verify that clusters definitions and labels from single-cell chromatin accessibility data matched those from single-cell expression data, we obtained published scRNA-seq data from 12 islet donors without diabetes²³. Because cluster definitions for all cell types were not available, we reanalyzed the data and performed clustering analyses. We used the Spearman correlation between t statistics from islet snATAC-seq and islet scRNA-seq data to verify cluster labels. See the Supplementary Methods for additional details.

Single-cell motif enrichment. We used chromVAR³³ v.1.5.0 to estimate transcription factor motif enrichment z -scores for each cell. First, we constructed a merged peak by cell sparse binary matrix as the input for chromVAR. We corrected for GC bias based on hg19 (BSgenome.Hsapiens.UCSC.hg19) using the addGCbias function. For transcription factor motifs within the nonredundant JASPAR 2018 CORE vertebrate set, we calculated bias-corrected deviation z -scores for each cell. Across all cell types, we selected variable transcription factor motifs ($n = 141$) with variability > 1.2. For each cell type, we then computed the average transcription factor motif enrichment z -score across single cells, collapsing the values of cell types with more than one state. We compared motif enrichment z -scores between single cells by using Benjamini–Hochberg-corrected P values (FDR < 0.01) from two-sample t -tests.

Ordering alpha, beta and delta cells along a pseudostate trajectory. We used Cicero⁸ v.1.3.3 with Monocle 3 to order alpha, beta and delta cells along separate trajectories. Starting from the merged peak by cell sparse binary matrix, we extracted beta cells and filtered out peaks that were not present in beta cells. We then preprocessed the data with latent semantic indexing and continued onto dimensionality reduction, cell clustering and trajectory graph learning using default parameters. We then chose the root state (that is, the start of the trajectory) based on the highest average *INS* promoter accessibility. We repeated the same procedure for beta and delta cells, instead choosing the root state based on *GCG* or *SST* promoter accessibility, respectively.

Comparison of endocrine cell states. To identify transcription factor motifs variable between cell states, we performed a one-sided Student's t -test on motif z -scores between cells in each state. We adjusted P values with the Benjamini–Hochberg procedure and defined motifs with an FDR < 0.05 and absolute difference in z -score > 0.5 as differential. To analyze differential promoter accessibility between cell states, we performed two-sided Fisher's exact tests between hormone-high and hormone-low states for each promoter against the null hypothesis that the promoter had similar accessibility across states. We used the Benjamini–Hochberg-adjusted P values (FDR < 0.01 for alpha or beta cells and FDR < 0.1 for delta cells) to identify gene promoters with differential accessibility across states. Differentially accessible promoters were used to perform gene ontology (GO) term enrichment on biological processes (2018 version) using Enrichr⁹³ v.1.0. We filtered for terms that contained < 150 genes. Details of the trajectory analysis of motifs and promoters are provided in the Supplementary Methods.

We collected gene sets from Xin et al.²³, Mawla & Huising⁴³ and Camunas-Soler et al.⁴⁴ (details in the Supplementary Methods). For each gene list, we performed GSEA⁹⁴ using log₂ ORs from previous Fisher's exact tests. We also performed GSEA using significantly differential promoters from Fig. 2a as the gene sets to assess whether cell states showed concordant differences across cell types.

Genome-wide association study enrichment with aggregate peak annotations. We used stratified LD score regression^{60–62} v.1.0.1 to calculate the enrichment for genome wide association study (GWAS) traits. We obtained GWAS summary statistics for quantitative traits^{11,12,48–51}, diabetes^{9,47} and control traits^{52–59}. To create custom LD score files, we annotated variants using peaks for each cluster as a binary annotation. In addition to the annotations included in the baseline LD model v.2.2, we included LD scores estimated from merged peaks across all clusters as the background. For each trait, we used stratified LD score regression to estimate the enrichment z -scores of each annotation relative to background. We computed one-sided P values for enrichment based on the z -scores and corrected for multiple tests using the Benjamini–Hochberg procedure.

GWAS enrichment with single-cell annotations. We determined genetic enrichment of accessible chromatin profiles in individual cells for fasting glucose level¹², T2D¹⁰ and the control traits major depressive disorder⁵⁶ and systemic lupus erythematosus⁵⁴ GWAS using polyTest⁹⁵. Additional details are provided in the Supplementary Methods.

To identify transcription factors correlated with trait enrichments, we regressed out the log read depth, fraction of reads in peaks and fraction of promoters used from the single-cell enrichments. We calculated the Spearman correlation between the residuals of fasting glucose or T2D enrichment z -scores and motif enrichment z -scores using all cells or only beta cells. We used Bonferroni correction to correct P values for multiple tests. To identify motifs directly enriched for T2D or fasting glucose association in beta cells, we identified all variants mapping in a beta cell site and using individual motif occurrences⁹⁶ predicted the instances of each motif in JASPAR 2018³⁴ using the sequence surrounding each allele. We considered variants disrupting motifs where there was a motif prediction for only one allele. We tested for enrichment of variants in a predicted motif or disrupting the motif for T2D or fasting glucose using polyTest⁹⁵.

GWAS enrichment at known T2D loci using aggregate peak annotations. We identified primary T2D risk signals¹⁰ where the highest causal probability variant was associated with fasting glucose level¹² at genome-wide significance. We annotated variants at each signal with INS^{high} and INS^{low} beta cell sites and tested for enrichment using fGWAS⁹⁴ v.0.3.6 (-fine). For alpha, delta and gamma cells, we retained sites that did not overlap a beta cell site and annotated variants at all T2D risk signals⁷. To exclude the possibility that enrichments could be driven by other relevant tissues, we annotated variants in the liver and adipose ATAC-seq from ENCODE⁹⁶. We tested for enrichment using fGWAS⁹⁴ (-fine) including other tissue annotations in the model. We considered annotations with positive enrichment estimates enriched for T2D risk.

Predicting genetic variant effects on chromatin accessibility. We used deltaSVM⁶⁵ to predict the effects of noncoding variants on chromatin accessibility in each cell type and cell state (Supplementary Methods). From variant z -scores, we calculated P values and FDRs and considered variants significant at an FDR < 0.1.

Luciferase gene reporter assays. We cloned sequences containing reference alleles in the forward orientation upstream of the minimal promoter of firefly luciferase vector pGL4.23 (Promega Corporation) using the KpnI and SacI restriction sites. For rs34584161:A>G, we cloned the alternative allele in the same manner as the reference alleles. For other variants, we introduced the alternative alleles via site-directed mutagenesis using the NEB Q5 Site-Directed Mutagenesis Kit (New England Biolabs). Site-directed mutagenesis and details of the reporter assay are provided in the Supplementary Methods. We normalized firefly activity to *Renilla* activity and compared it to the empty vector; the normalized results were expressed as fold change compared to control per allele. We used a two-sided Student's t -test to compare the two alleles.

Mapping allelic imbalance within clusters. We extracted genomic DNA either from spare islet nuclei (donors 1 and 2) or acinar cells (donor 3). We used the DNeasy Blood & Tissue Kits (QIAGEN) according to the manufacturer's protocol to purify total DNA and performed genotyping on the Infinium Omni2.5-8 Kit v.1.4 (Illumina). Additional details are provided in the Supplementary Methods. Using cluster assignments for each cell, we split mapped reads for each donor into cluster-specific reads. We used WASP⁹⁷ v.0.3.0 to correct for reference mapping bias at heterozygous variants. We then used a two-sided binomial test to assess imbalance assuming a null hypothesis where both alleles were equally likely to be observed. For each variant, we combined imbalance z -scores across donors by using Stouffer's z -score method with sequencing depth as a weight for each sample.

Grouping of predicted effect variants and enrichment for islet caQTLs. We categorized variants with predicted effects on alpha or beta cells based on effects across cell type and states: alpha ($n = 8,552$), beta ($n = 11,650$), hormone-high ($n = 12,874$), hormone-low ($n = 10,808$) and shared ($n = 27,140$). We also determined the concordance in the direction of effect for variants across cell states. For the set of variants with significant effects in each state, we calculated the fraction of variants where the effect allele had a predicted effect in other states. We determined significance using a two-sided binomial test assuming an expected fraction of 50%. We assessed enrichment of predicted effect variants in alpha or beta cell states for islet caQTLs⁹⁸ compared to any caQTL in alpha or beta cell sites using two-sided Fisher's exact tests. We stratified variants with predicted effects by category and assessed enrichment of caQTLs with predicted effects within each category with two-sided Fisher's exact tests.

Transcription factor motif enrichment within predicted effect variant categories. For each cell- or state-resolved category of variants with predicted effects, we extracted 29-bp sequences (± 14 bp around the variant) corresponding to the higher or lower predicted effect allele. We used AME⁹⁶ in MEME v.4.12.0 to predict motif enrichment, using position weight matrices from JASPAR 2018³⁴. We used sequences from the higher or lower effect alleles as the test or background set, respectively. We used TFClass (<http://tfclass.bioinf.med.uni-goettingen.de/>) to group motifs by family. To determine the transcription factors most likely driving the enrichment, we checked normalized promoter accessibility within family members in alpha or beta cells.

Enrichment of predicted variants for lower-frequency variants. We obtained summary statistics for T2D⁹ and performed LD pruning (Supplementary Methods). We identified variants that had significant effects in any endocrine cell type, as well as for each cell type for either state. We created a background set of variants without significant effects in any endocrine cell type (FDR > 0.1). We created thresholds based on T2D association P values (5×10^{-8} , 1×10^{-7} , 1×10^{-6} , 1×10^{-5} , 1×10^{-4} , 1×10^{-3}). For each threshold, we calculated the fold enrichment of predicted effect variants as compared to the background and determined significance using two-sided binomial tests.

Single-cell chromatin co-accessibility. We used Cicero⁹ v.1.3.3 to calculate co-accessibility for pairs of peaks for alpha, beta and delta cells. As in the trajectory analysis, we started from the merged peak by cell sparse binary matrix, extracted beta cells and filtered out peaks that were not present in beta cells. We used `make_cicero_cds` to aggregate cells based on the 50 nearest neighbors. We then used Cicero to calculate co-accessibility using a window size of 1 Mb and a distance constraint of 500 kb. We repeated the same procedure for alpha and delta cells. We used a co-accessibility threshold > 0.05 to define pairs of peaks as co-accessible.

Enrichment of islet Hi-C and pHi-C loops in co-accessible peaks. We obtained islet promoter capture Hi-C (pHi-C) and merged Hi-C loops^{27,71}. For both datasets, we used coordinates for anchors directly from the loop calls. To compare co-accessibility with pHi-C or Hi-C, we used direct overlap of peaks with anchors. For binned thresholds of co-accessibility in 0.05 increments, we calculated distance-matched ORs for alpha, beta or delta cell co-accessible peaks containing pHi-C or Hi-C loops versus non-co-accessible peaks (co-accessibility < 0). We used two-sided Fisher's exact tests to assess significance.

Hi-C library construction and data analysis. We performed in situ Hi-C as described previously using MboI on two batches of hESC-derived beta cells by using the differentiation protocol described below cultured in high (20 mM mg ml⁻¹) or low glucose (5 mM mg ml⁻¹). Hi-C libraries were sequenced to read counts of 1,509,428,732 and 1,918,698,012, respectively. We analyzed Hi-C using Juicer⁹⁸ v.1.5.7 with default settings and visualized with HiGlass⁹⁹ v.1.7.2. Unless otherwise indicated, interaction frequencies were normalized using iterative matrix balancing. We generated virtual 4C tracks by extracting normalized interaction frequencies from an anchor bin of interest from the contact matrix. We performed aggregate peak analysis using the Juicer tools with the settings -u -r 10000.

Annotating fine-mapped diabetes risk variants. We annotated risk signals in compiled fine-mapping data for T2D (additional details in the Supplementary Methods). For each signal, we identified variants that were in the 99% credible set with a PPA > 0.01. We intersected these candidate variants with sites for each islet cell type and cell state and then identified variants with predicted effects on the overlapping cell types/states. We annotated variants based on overlap with sites co-accessible to gene promoters. For target genes linked to diabetes risk variants, we determined enriched gene sets using GSEA⁹⁴ (Supplementary Table 11).

Analysis of *INS* promoter 4C data. We downloaded and reanalyzed published 4C of the *INS* promoter for EndoC- β H1 (ref.⁷²) with 4C-ker¹⁰⁰. We created a reduced genome using 25-bp flanking sequences of BglII cutting sites. For the three replicates, we aligned reads to this reduced genome using Bowtie 2¹⁰¹ v.2.2.9 (-N 0 -5 20). We extracted counts for each fragment after removing self-ligated and undigested fragments and input bedGraph files to R4Cker (no version). We generated normalized counts and called high interaction regions using nearBaitAnalysis ($k = 10$).

CRISPR-Cas9-mediated genome editing in hESCs. H1 hESCs (WA01; purchased from WiCell; National Institutes of Health (NIH) registration no. 0043) were seeded onto Matrigel-coated 6-well plates at a density of 50,000 cells cm⁻² and maintained in mTeSR1 medium (STEMCELL Technologies) for 3–4 d with the medium changed daily. Details for the generation of clonal hESC lines for enhancer deletion and allele-specific editing are provided in the Supplementary Methods. Single-guide RNA oligonucleotides, single-strand DNA donor template and genotyping primers are listed in Supplementary Table 12. hESC research was approved by the University of California, San Diego institutional review board and Embryonic Stem Cell Research Oversight Committee.

Pancreatic differentiation of hESC clones. *KCNQ1* enhancer-deleted hESC lines were differentiated in a suspension-based format using rotational culture with modifications to a published protocol⁹⁴. *KCNQ1* base-edited hESC lines were differentiated in a suspension-based format using rotational culture with modifications to a published protocol¹⁰². Details of the modifications are provided in the Supplementary Methods. For the experimental analysis of clones, we performed flow cytometry, immunofluorescence staining, messenger RNA sequencing, qPCR and insulin content measurement (details of each assay can be found in the Supplementary Methods).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequencing data have been deposited into the National Center for Biotechnology Information Gene Expression Omnibus with accession numbers GSE160472, GSE160473 and GSE163610. Processed data files and annotations for snATAC-seq are available through the Diabetes Epigenome Atlas (<https://www.diabeteseipigenome.org/>). All other data are either found in the article or available upon request to the corresponding author. Source data are provided with this paper.

Code availability

The code for processing and clustering the snATAC-seq datasets is available at https://github.com/kjgaulton/pipelines/tree/master/islet_snATAC_pipeline.

References

86. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
87. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
88. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
89. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
90. Arda, H. E. et al. A chromatin basis for cell lineage and disease risk in the human pancreas. *Cell Syst.* **7**, 310–322.e4 (2018).
91. Ackermann, A. M., Wang, Z., Schug, J., Naji, A. & Kaestner, K. H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* **5**, 233–244 (2016).
92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
93. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
94. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
95. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
96. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
97. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
98. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
99. Kerpedjiev, P. et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
100. Raviram, R. et al. 4C-ker: a method to reproducibly identify genome-wide interactions captured by 4C-seq experiments. *PLoS Comput. Biol.* **12**, e1004780 (2016).
101. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
102. Velazco-Cruz, L. et al. Acquisition of dynamic function in human stem cell-derived β cells. *Stem Cell Rep.* **12**, 351–365 (2019).

Acknowledgements

This work was supported by NIH grant nos. R01DK114650 and U01DK105554 (sub-award) to K.G., grant nos. R01DK068471 and U01DK105541 to M.S. and grant no. U01DK120429 to K.G. and M.S. and by the University of California, San Diego School of Medicine to the Center for Epigenomics. We thank the QB3 Macrolab at University of California, Berkeley for the purification of the Tn5 transposase. We thank K. Jepsen, the University of California, San Diego Institute for Genomic Medicine Genomics Center and S. Kuan for sequencing and B. Li for bioinformatics support. Data from the UK Biobank was accessed under application no. 24058. We thank I. Matta for the preparation of the RNA-seq libraries.

Author contributions

K.J.G., D.U.G. and M. Sander conceived and supervised the research. J.C. performed the analyses of the single-cell and genetic data. C.Z., M. Schlichting and J.W. performed the hESC experiments. Z.C. performed the analyses of the single-cell and Hi-C data. J.Y.H. performed the combinatorial barcoding single-cell assays and genotyping. M.M. performed the 10x single-cell assays. R.M. performed the Hi-C experiments. S.H., A.D. and M.-L.O. performed the reporter experiments. Y.Q. performed the analyses of the 4C data. Y. Sui performed the analyses of the hESC data. Y. Sun and P.K. developed and processed the data for the epigenome database. R.F. contributed to the analyses of the single-cell data. S.P. contributed to the development of the single-cell assays. K.J.G., D.U.G., M. Sander, J.C., C.Z. and Z.C. wrote the manuscript.

Competing interests

K.J.G. does consulting work for Genentech and holds stock in Vertex Pharmaceuticals; neither is related to the work in this study. The other authors declare no competing interests.

Additional information

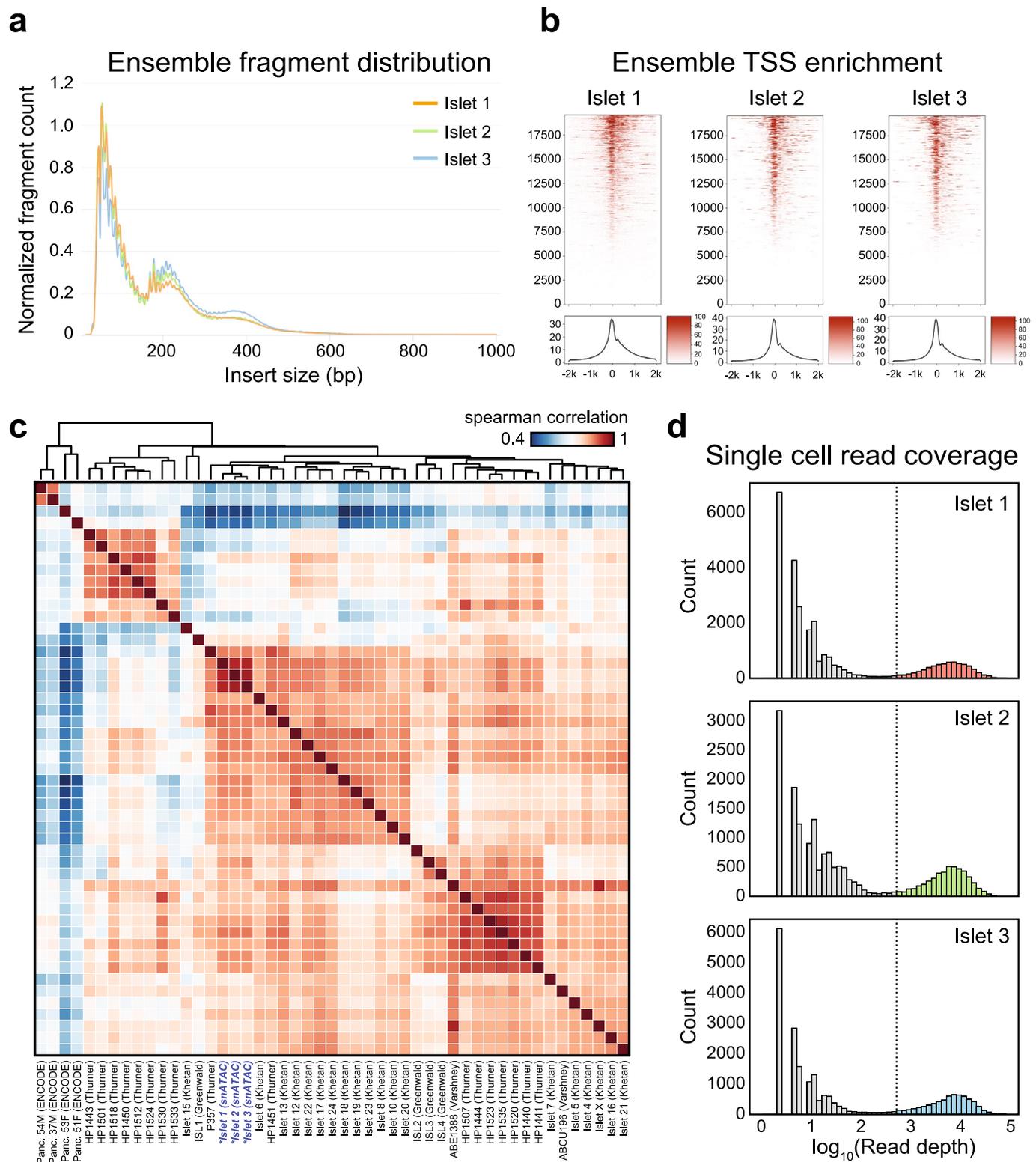
Extended data is available for this paper at <https://doi.org/10.1038/s41588-021-00823-0>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-021-00823-0>.

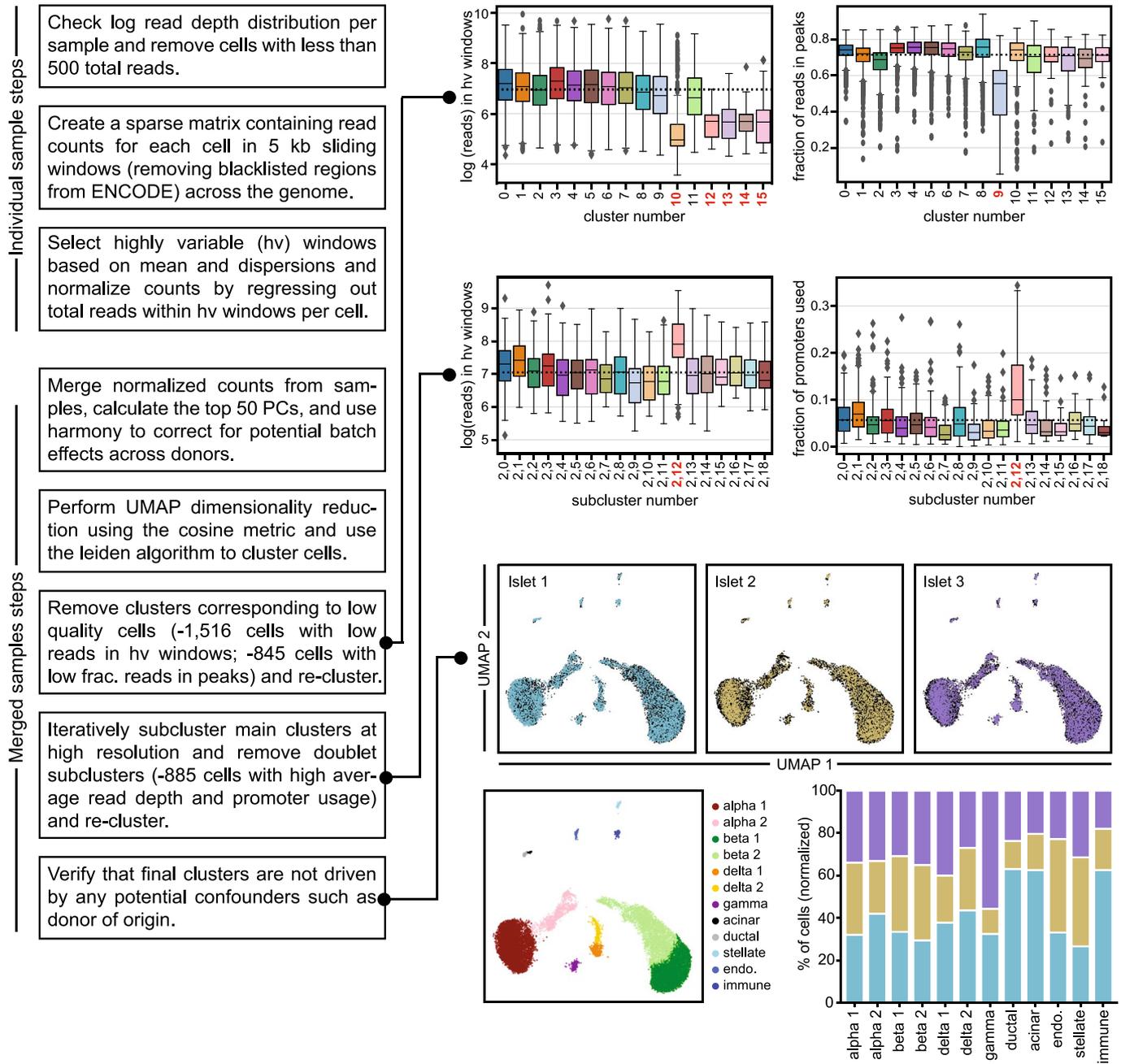
Correspondence and requests for materials should be addressed to M. Sander, D.U.G. or K.J.G.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

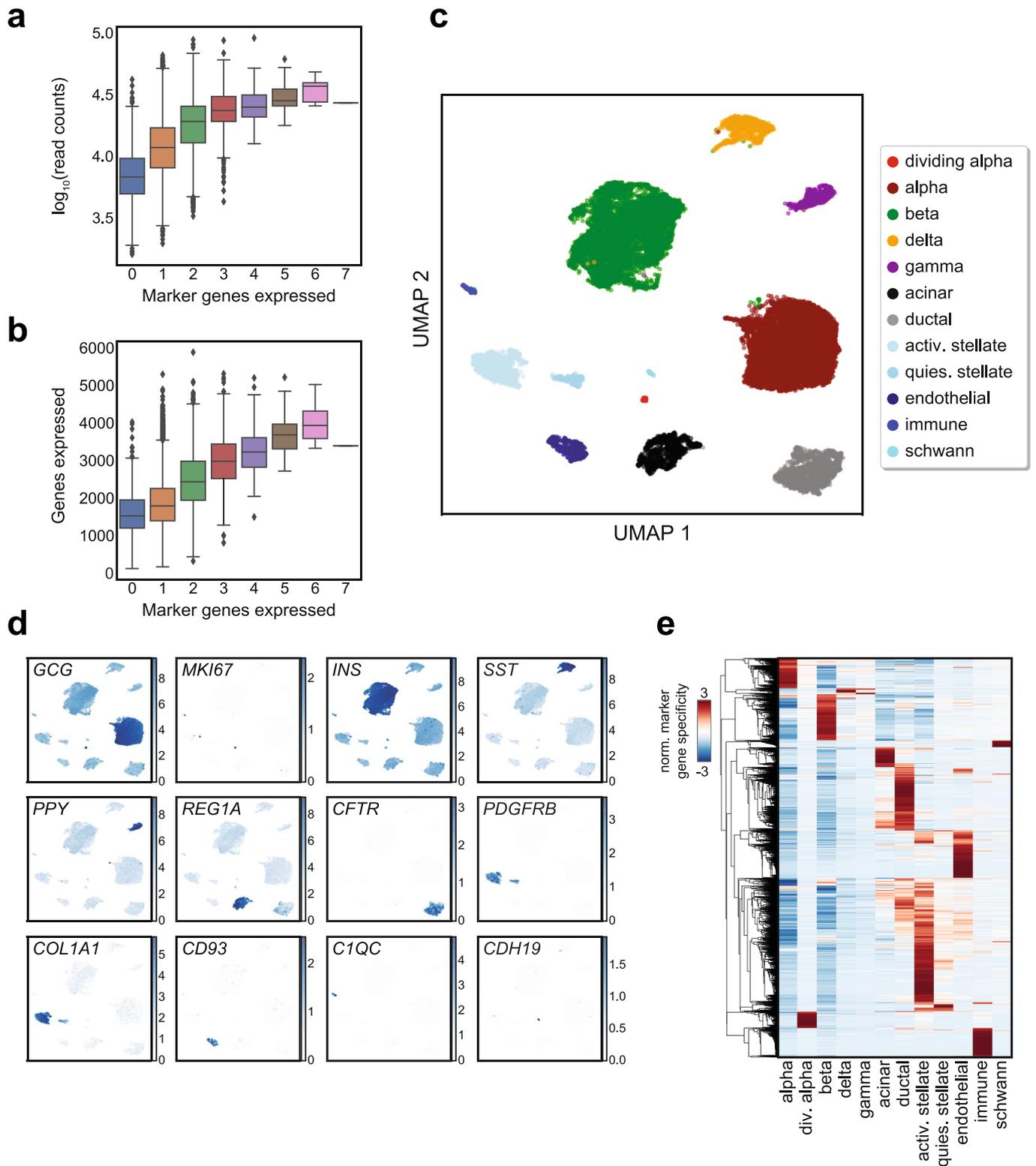
Reprints and permissions information is available at www.nature.com/reprints.



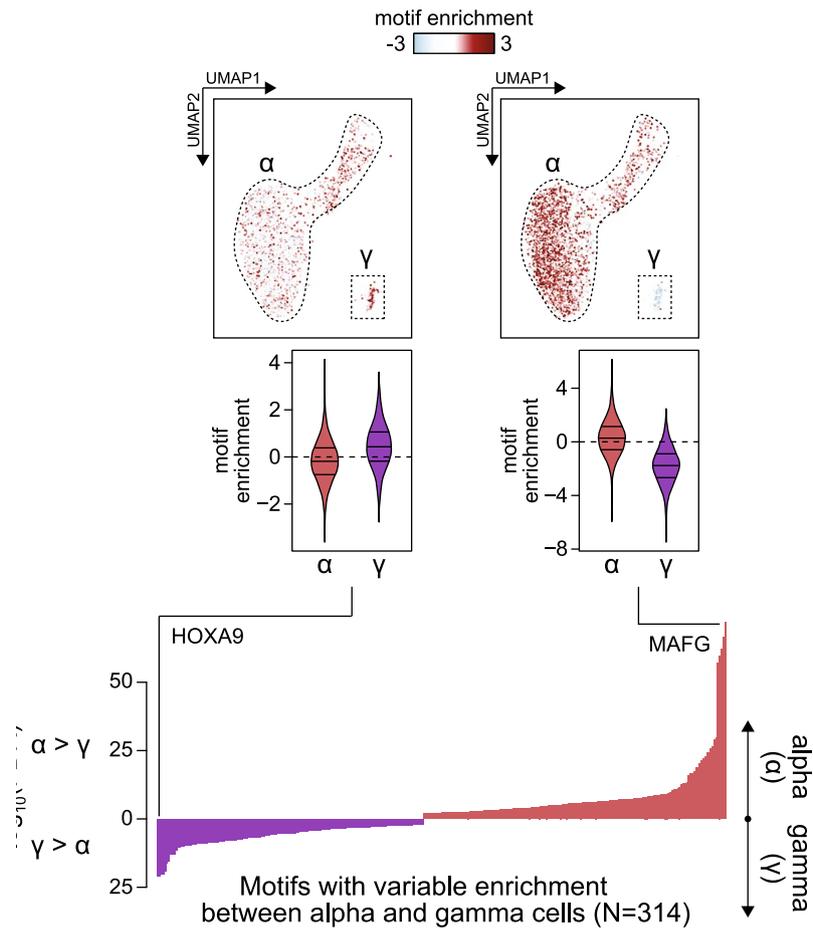
Extended Data Fig. 1 | Quality control metrics and aggregate comparison to bulk islet ATAC. **a**, Insert size distribution for aggregate reads from each snATAC-seq experiment. **b**, Aggregated read coverage from each snATAC-seq experiment in a ± 2 kb window around individual promoters (top) and averaged across all promoters (bottom). **c**, Spearman correlation between normalized read coverage within a merged set of peaks from 3 aggregated islet snATAC-seq, 42 bulk islet ATAC-seq, and 4 bulk pancreas ATAC-seq datasets. Names of samples are from the original sources of the data. **d**, Binned \log_{10} read depth distribution for each experiment.



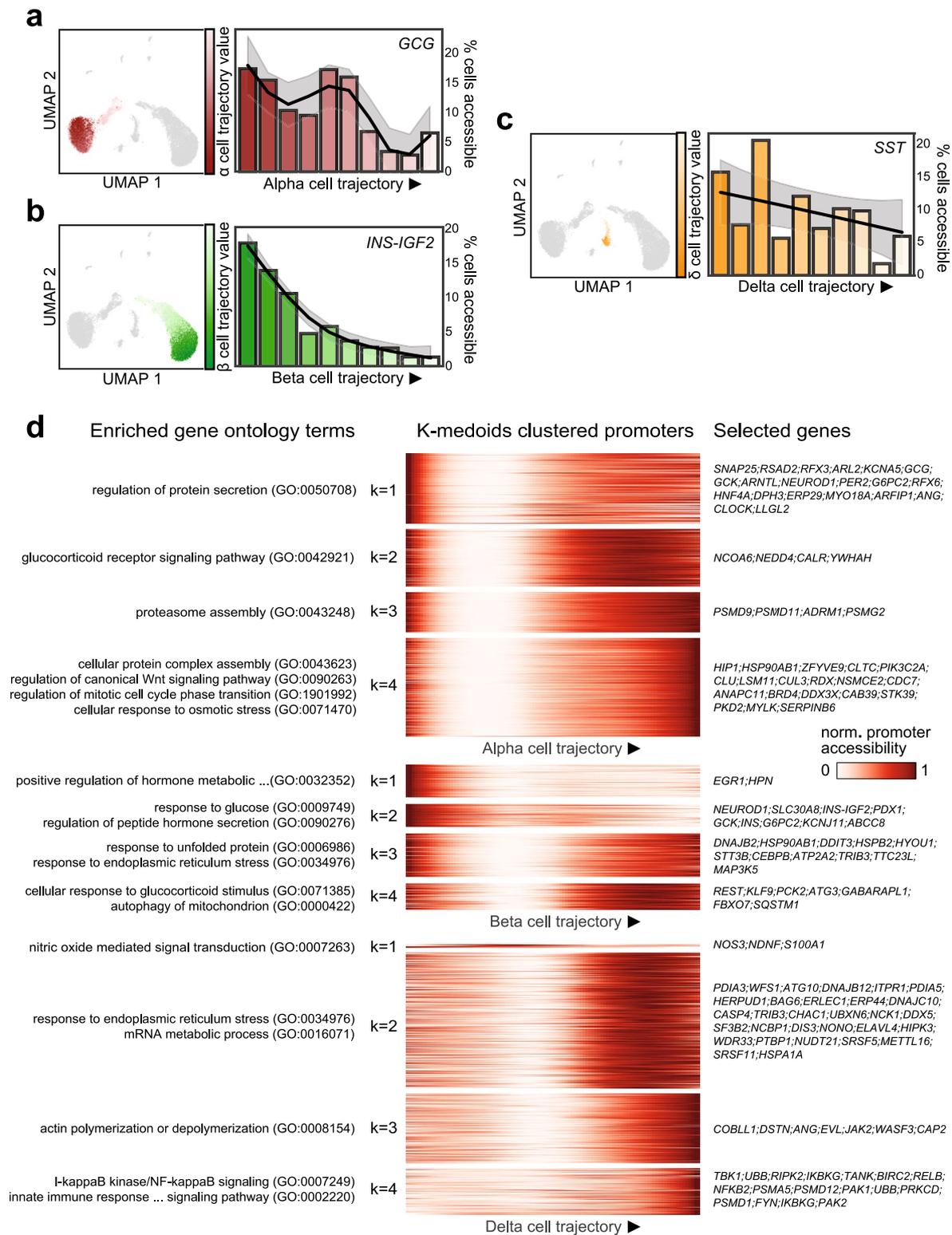
Extended Data Fig. 2 | Flowchart of the snATAC-seq data processing pipeline. **a**, Flowchart summarizing key steps of the snATAC-seq processing pipeline, including the various steps where cells were filtered out. Samples were first processed individually. All samples were then combined using a batch correction method. Clusters corresponding to cells from low quality cells, including those with low read depth in highly variable windows and low fraction of reads in peaks were then removed. After re-clustering, iterative subclustering of the main clusters at high resolution was used to identify and remove doublet subclusters. The final clusters are not driven by potential confounders such as donor of origin. Boxplot center lines, limits, and whiskers represent median, quartiles, and 1.5 IQR respectively.



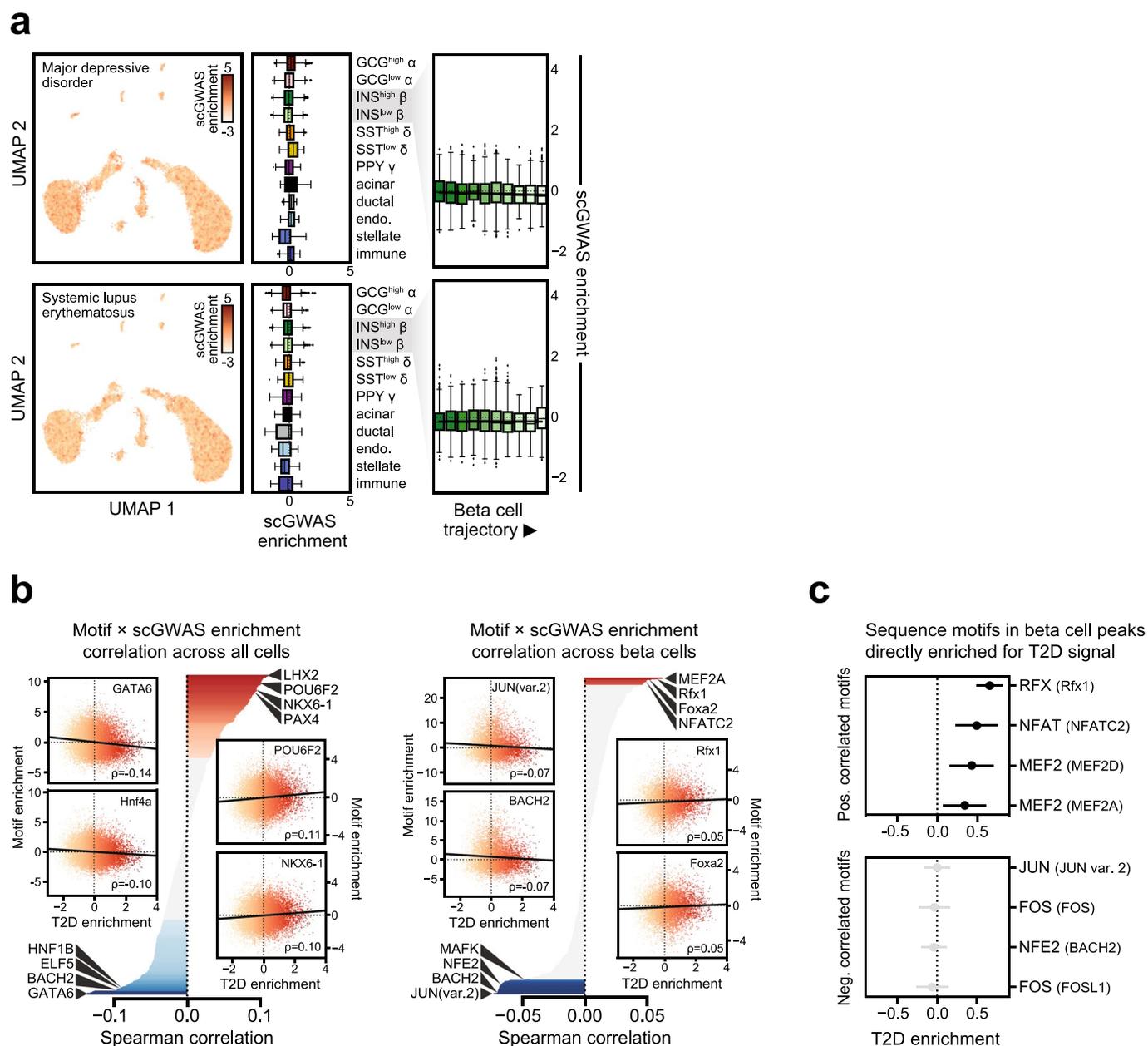
Extended Data Fig. 3 | Analysis of islet single cell gene expression data. **a**, \log_{10} transformed read depth or **(b)** total number of genes expressed compared with number of marker genes expressed per cell from scRNA-seq data. Boxplot center lines, limits, and whiskers represent median, quartiles, and 1.5 IQR respectively. Cells expressing more than one marker gene (defined by mixture models) were marked as doublets and filtered out. **c**, Clusters of islet cells from single cell RNA-seq data plotted on UMAP coordinates. quies. stellate, quiescent stellate. activ. stellate, activated stellate. **d**, Selected marker gene \log_2 (expression) for each cluster plotted on UMAP coordinates. **e**, Row-normalized t-statistics of marker gene specificity showing the most specific genes (t-statistic > 20) for each cluster.



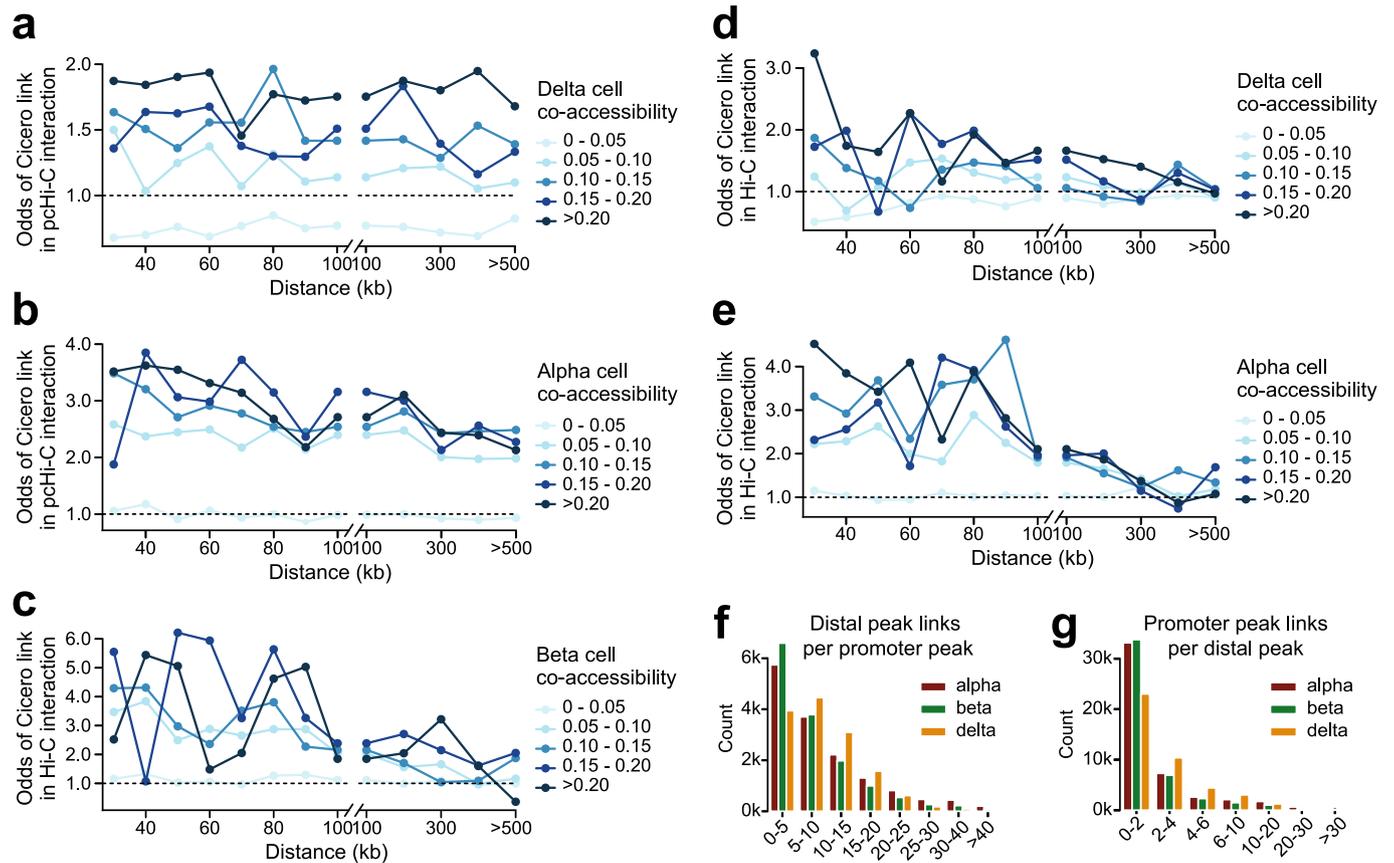
Extended Data Fig. 4 | Comparison of motif enrichment between alpha and gamma cells. Differential enrichment of motifs between alpha cell open chromatin regions and gamma cell open chromatin regions as measured by a 2-sided T-test, with FDR calculated by the Benjamini-Hochberg procedure. Examples are highlighted of motifs enriched in alpha cells and gamma cells, respectively (MAFG, HOXA9). UMAP plots show enrichment z-scores for the indicated motifs in alpha and gamma cells. Violin plots below show the distribution of enrichment z-scores across alpha or gamma cells, where the lines represent median and quartiles.



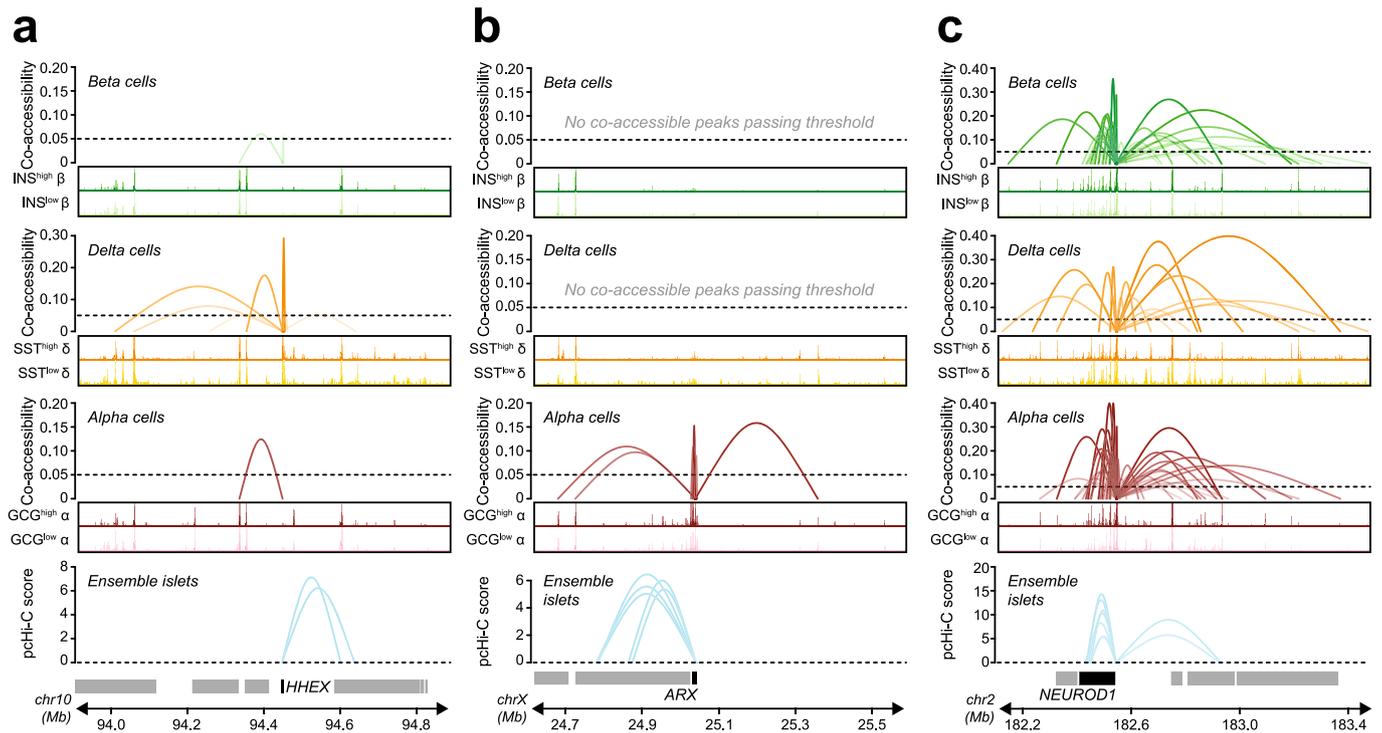
Extended Data Fig. 5 | Differentially accessible promoters across pseudo-states. **a**, Pseudo-state (trajectory) values for alpha cells plotted on UMAP coordinates (left) and percentage of cells with *GCG* promoter accessibility decreases across 10 bins along the alpha (α) cell trajectory (right). **b**, Pseudo-state (trajectory) values for beta (β) cells plotted on UMAP coordinates (left) and percentage of cells with *INS-IGF2* promoter accessibility decreases across 10 bins along the beta cell trajectory (right). **c**, Pseudo-state (trajectory) values for delta (δ) cells plotted on UMAP coordinates (left) and percentage of cells with *SST* promoter accessibility decreases across 10 bins along the beta cell trajectory (right). **d**, Heatmaps showing promoters with dynamic accessibility across trajectories for alpha (top), beta (middle) and delta (bottom) cell trajectories. Gene promoters are clustered into 4 groups for each trajectory with k-medoids clustering. Enriched gene ontology for each k-medoid cluster (left) and selected genes present in at least one enriched gene ontology.



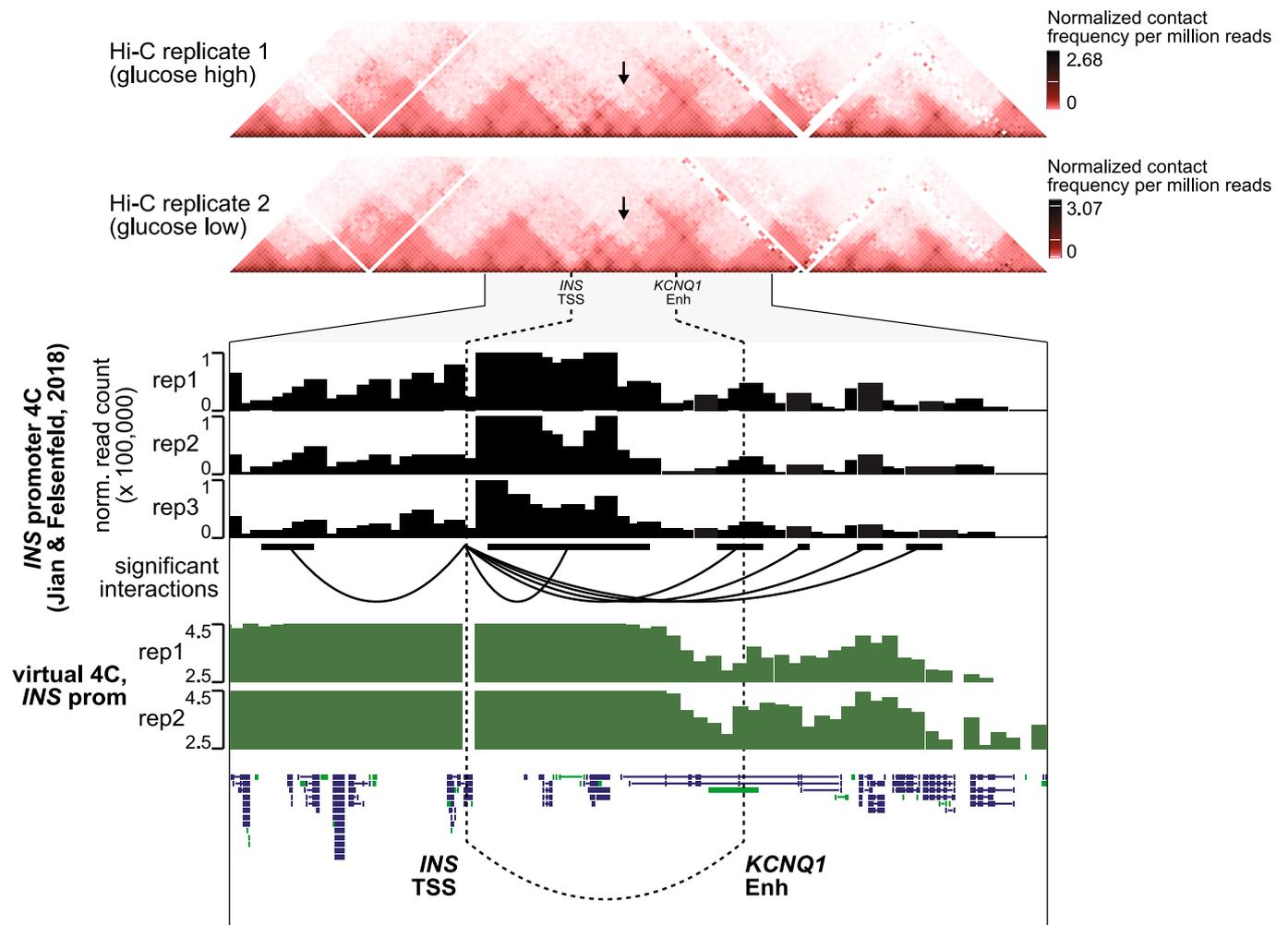
Extended Data Fig. 6 | Single cell GWAS enrichment and correlation with TF motifs. a, Single cell GWAS enrichment z-scores for Major depressive disorder and Systemic lupus erythematosus projected onto UMAP coordinates (left panels), z-score enrichment distribution per cell type and state (middle panels) and z-score enrichment distribution split into 10 bins based on beta cell trajectory values (right panels). Boxplot center lines, limits, and whiskers represent median, quartiles, and 1.5 IQR respectively. **b**, Correlation between single cell GWAS enrichment z-scores for Type 2 Diabetes and chromVAR TF motif enrichment z-scores across either all cells (left) or beta cells (right). Inset scatterplots highlight the top correlated motifs in either direction. **c**, Variants mapping directly in sequence motifs positively correlated with T2D risk in beta cells are enriched for T2D association, whereas variants mapping in motifs negatively correlated with T2D risk in beta cells show no such enrichment. Values represent effect size and SE.



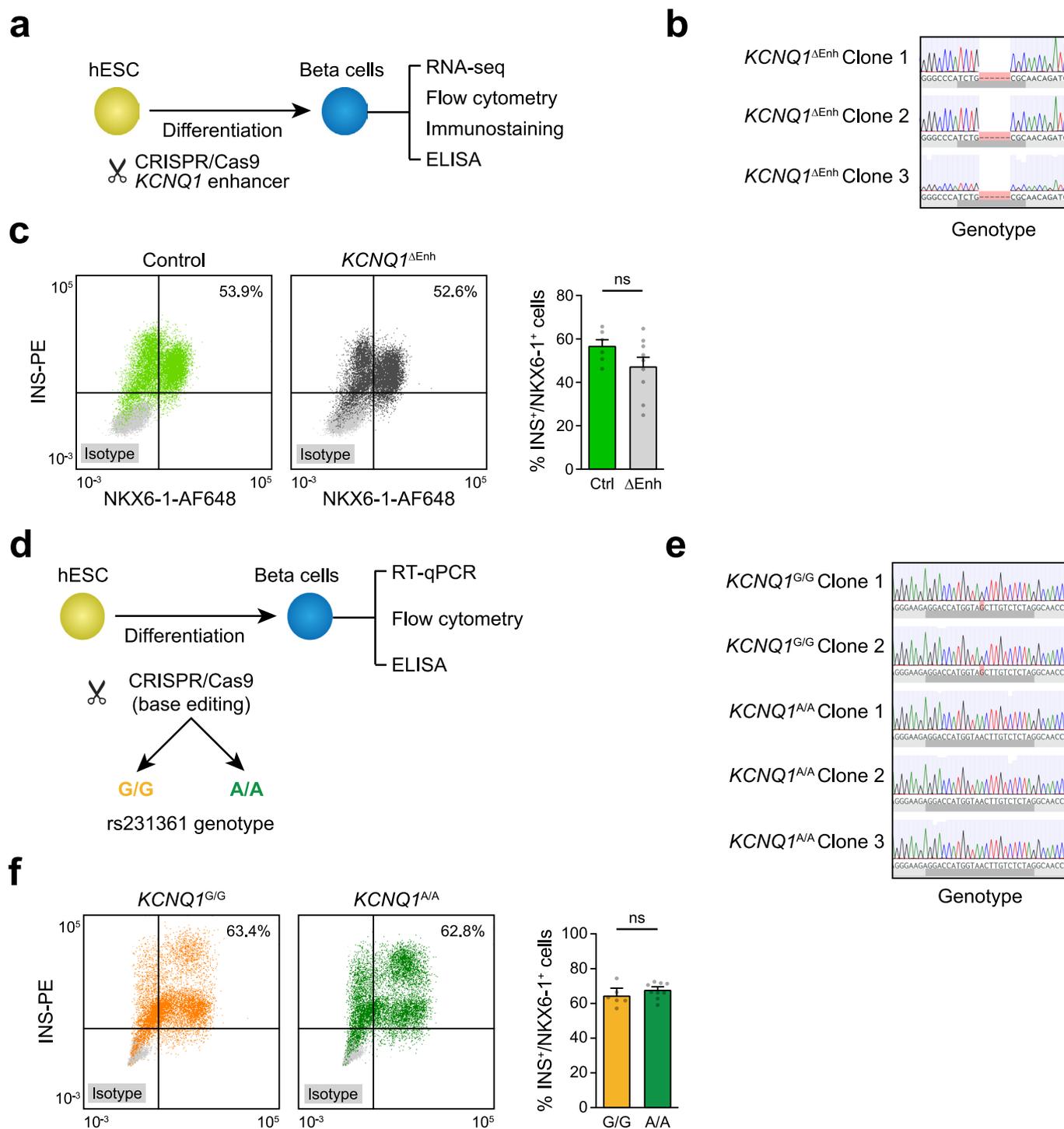
Extended Data Fig. 7 | Single cell co-accessibility analyses in islet cell types. **a**, Distance-matched odds that delta cell co-accessibility links overlap islet pcHi-C chromatin loops at different co-accessibility threshold bins in 0.05 intervals demonstrate that co-accessible links are enriched for chromatin interactions. **b**, Same analysis as in (a) but with alpha cell co-accessibility. **c**, Same analysis as in (a) but with beta cell co-accessibility and Hi-C loops. **d**, Same analysis as in (a) but with delta cell co-accessibility and Hi-C loops. **e**, Same analysis as in (a) but with alpha cell co-accessibility and Hi-C loops. **f**, Number of distal sites linked to each promoter peak for alpha, beta, and delta cells. **g**, Number of promoters linked to each distal site for alpha, beta, and delta cells.



Extended Data Fig. 8 | Cell type-specific and shared co-accessible sites. a, An example of co-accessibility anchored at the promoter for the delta cell identity TF *HHEX*. Co-accessibility for beta, delta, and alpha cells are shown compared to high-confidence pHi-C loops from ensemble islets. Genome browser plots scale: 0-10. **b**, An example of co-accessibility anchored at the promoter for the alpha cell identity TF *ARX*. **c**, An example of shared co-accessibility anchored at the promoter for the shared islet identity TF *NEUROD1*.



Extended Data Fig. 9 | 3D chromatin interactions at the T2D-associated *KCNQ1* locus. Top panels show Hi-C contact matrices from hESC-derived beta cells, visualized at 25 kb resolution. Region shown is chr11:500,000-4,500,000, hg19. Black arrows indicate putative interaction point of *INS* TSS and *KCNQ1* enhancer. Genome browser plot below shows a zoomed view of chr11:1,750,000-3,250,000. Data from 4C-seq anchored on the *INS* promoter in EndoC- β H1 cells (Jian & Felsenfeld⁷²) is shown, as analyzed with the 4C-ker package. Normalized read counts are shown in black from 3 biological replicates. Significant interactions from *INS* promoter are shown as arcs below read counts tracks. Interactions calls from data pooled across 3 replicates are shown here. The region containing the *KCNQ1* enhancer was called as a significant interaction region with *INS* promoter independently in each 4C replicate. Virtual 4C plots in green show log(normalized Hi-C interaction frequency) from *INS* promoter.



Extended Data Fig. 10 | Genome editing of the *KCNQ1* locus in hESCs. a, Schematic of the workflow and **(b)** Sanger sequencing for *KCNQ1* enhancer deletion in three independent hESC clones. **c**, Representative figures of flow cytometry analysis for NKX6-1 and INS comparing control and *KCNQ1*^{ΔEnh} cells (left). Quantification of the percentage of NKX6-1⁺/INS⁺ cells in beta cell stage cultures from control ($n=6$; 2 clones \times 3 differentiations) and *KCNQ1*^{ΔEnh} ($n=9$; 3 clones \times 3 differentiations) cells (right). Values represent mean and SEM. ns, not significant by two-sided Student's T-test without adjustment for multiple comparisons. **d**, Schematic of the workflow and **(e)** Sanger sequencing for two independent *KCNQ1*^{G/G} clones and three *KCNQ1*^{A/A} clones. **f**, Representative figures of flow cytometry analysis for NKX6-1 and INS comparing *KCNQ1*^{G/G} and *KCNQ1*^{A/A} clones (left). Quantification of the percentage of NKX6-1⁺/INS⁺ cells in beta cell stage cultures from *KCNQ1*^{G/G} ($n=6$; 2 clones \times 3 differentiations) and *KCNQ1*^{A/A} ($n=9$; 3 clones \times 3 differentiations) cells (right). ns, not significant by two-sided Student's T-test without adjustment for multiple comparisons. Values represent mean and SEM.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software used.

Data analysis

Software and package versions are also listed in the methods section.

Software: trim_galore (v.0.4.4), bwa (v.0.7.17-r1188), bowtie2 (v.2.2.9), STAR (v2.5.3a), MACS2 (v2.1.2), polyTest (no version), picard (no version), python (v.3.6.6), R (v.3.5.1), LD score regression (v.1.0.0), deltaSVM (no version), samtools (v.1.6), MEME suite (v.4.12.0), PLINK (v.1.90b6.7), NEBaseChanger (v.1.2.8), HTSeq (v.0.9.0), Juicer (v.1.5.7), GSEA (v.4.1.0), fgwas (v.0.3.6), HiGlass (v.1.7.2), GenomeStudio (v.2.0.4), FlowJo (v.10)

Python packages: scanpy (v.1.4), WASP (v.0.3.0), cooler (v.0.8.10), leiden (v.0.7.0), bedtools (v.2.26.0), FlowJo (v.10), GraphPad Prism 7.
R packages: harmony (v.0.1.0), chromVAR (v.1.5.0), enrichR (v. 1.0), mgcv (v1.8.28), fpc (v.2.1.11.1), Cicero (v.1.3.3), R4Cker (no version), DESeq2 (v.1.10.1)

Custom code for snATAC-seq analysis: https://github.com/kjgaulton/pipelines/tree/master/islet_snATAC_pipeline

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data have been deposited into the NCBI Gene Expression Omnibus (GEO) with accession numbers GSE160472, GSE160473, and GSE163610.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. Sample size for islet snATAC-seq (n=3 biological samples) was based on current standard in the field as defined by comparable studies generating reference maps of the islet epigenome (PMID: 31253982, 28193859, 31064983, 32029221, 24413736) Sample size for hESC differentiation experiments (for deletion experiments n=2 individual control clones, n=3 individual deletion clones, 3 independent differentiations; for base editing experiments n=2 individual clones for the reference allele, n=3 individual clones for the alternate allele, 3 independent differentiations) was considered sufficient to ensure reproducibility, and also consistent with comparable studies (PMID 29992946).
Data exclusions	No data exclusions.
Replication	For islet snATAC-seq, 3 biological replicates were used and reproducibility for epigenome data was assessed through correlation with published bulk islet and sorted islet ATAC-seq data, as well as using snATAC-seq data generated from whole pancreas using two independent technologies. All attempts at replication for this study were successful. For flow cytometry, insulin content and RT-qPCR experiment of hESC enhancer deletions and base editing clones, 2 distinct control clones and 3 distinct deletion clones were conducted differentiation for three times with consistent phenotypes. For immunofluorescence staining of enhancer deletion clones, 2 distinct control clones and 3 distinct deletion clones were conducted differentiation for three times with similar results, but only 1 enhancer deletion and 1 control clone were shown as representative in the manuscript. For RNA-seq, 1 control clone and 3 distinct enhancer deletion clones were differentiated 2 times. Two to three biological replicates for RNA-seq was determined to be sufficient based on the research community standards when the studies were performed. https://www.encodeproject.org/about/experiment-guidelines . For Hi-C of hESC-derived beta cells, 1 replicate for each treatment (high and low glucose) was used. There is no requirement of replicates for Hi-C based on the research community standards when the studies were performed. https://www.encodeproject.org/about/experiment-guidelines .
Randomization	No randomization was performed. For snATAC-seq, each sample was assayed separately. For hESC studies, randomization was not possible because cells were genetically modified.
Blinding	No blinding was performed during data collection or analysis. For snATAC-seq, all samples were non-disease and no comparisons were performed that would necessitate blinding. For hESC studies, we did not have the personnel resources to blind this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Mouse anti-Nkx6-1-Alexa Fluor@647 (BD Biosciences, BDB563338, Clone:R11-560)
 Rabbit anti-insulin (C27C9)-PE (Cell Signaling Technology, 8508s), RRID:AB_11179076
 Rabbit anti-IgG-Alexa Fluor@647 (Cell Signaling Technology, 2985s), RRID:AB_1196589
 Mouse anti-IgG1-PE (BD Biosciences, BDB556650), RRID: AB_396514
 Guinea pig anti-insulin (DAKO, A056401-2), RRID: AB_10013624
 Mouse anti-NKX6-1 (LifeSpan BioScience, LS-C143534), RRID: AB_10947571
 Secondary:
 Donkey Anti-guinea pig IgG (H+L)-Alexa Fluor® 488 (Jackson Immuno Research Laboratories, 706-545-148), RRID:AB_2340472
 Donkey Anti-mouse IgG (H+L)- Cy™3 (Jackson Immuno Research Laboratories, 715-165-151), RRID:AB_2315777

Validation

Mouse anti-Nkx6-1-Alexa Fluor@647 (BD Biosciences, BDB563338, Clone:R11-560)
 The manufacturer's website shows validation that the antibody is specific for both mouse and human NKX6.1 protein and cites several studies in which the antibody was used for flow cytometry.

Rabbit anti-insulin (C27C9)-PE (Cell Signaling Technology, 8508s), RRID:AB_11179076
 The manufacturer's website shows validation that the antibody is specific for mouse INS protein and cites several studies in which the antibody was used for flow cytometry. The antibody was validated in-house on human tissue against IFC for insulin.

Rabbit anti-IgG-Alexa Fluor@647 (Cell Signaling Technology, 2985s), RRID:AB_1196589
 The manufacturer's website shows validation that the antibody is not expected to recognize any specific antigen and cites several studies in which the antibody was used for flow cytometry.

Mouse anti-IgG1-PE (BD Biosciences, BDB556650), RRID: AB_396514
 The manufacturer's website shows validation that the antibody is not expected to recognize any specific antigen and cites several studies in which the antibody was used for flow cytometry.

Guinea pig anti-insulin (DAKO, A056401-2), RRID: AB_10013624
 This antibody was validated in-house for IFC using mouse and human tissue (PMID: 21829703; PMID: 23318056). The antibody is highly specific for endocrine cells in the pancreatic islet. The manufacturer's website shows validation that the antibody is specific for both mouse and human INS protein. The antibody is highly cited for IFC according to citeab.com.

Mouse anti-NKX6-1 (LifeSpan BioScience, LS-C143534), RRID: AB_10947571
 The manufacturer's website shows validation that the antibody is specific for both mouse and human NKX6.1 protein.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	H1 hESCs (WiCell), MIN6 (gift from the Jhala lab at UC San Diego)
Authentication	H1 human embryonic stem cells (hESC) experiments were conducted on passage 60-75 cryopreserved cells. hESCs stocks were tested for mycoplasma and confirmed negative at the time of freezing. hESCs were karyotyped quarterly to ensure genomic stability and additionally subjected to Short Tandem Repeat (STR) profiling to authenticate cell line identity. MIN6 cells were authenticated using Western blotting of key marker proteins.
Mycoplasma contamination	All cell lines tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cell aggregates derived from hESCs were allowed to settle in microcentrifuge tubes and washed with PBS. Cell aggregates were incubated with Accutase® at room temperature until a single-cell suspension was obtained. Cells were washed with 1 mL ice-cold flow buffer comprised of 0.2% BSA in PBS and centrifuged at 200 g for 5 min. BD Cytofix/Cytoperm™Plus Fixation/Permeabilization Solution Kit was used to fix and stain cells for flow cytometry according to the manufacturer's instructions.
--------------------	---

Briefly, cell pellets were re-suspended in ice-cold BD fixation/permeabilization solution (300 μ L per microcentrifuge tube). Cells were incubated for 20 min at 4 °C. Cells were washed twice with 1 mL ice-cold 1X BD Perm/Wash™ Buffer and centrifuged at 10 °C and 200 x g for 5 min. Cells were re-suspended in 50 μ L ice-cold 1X BD Perm/Wash™ Buffer containing diluted antibodies, for each staining performed. Cells were incubated at 4 °C in the dark for 1 hr. Cells were washed with 1.25 mL ice-cold 1X BD Wash Buffer and centrifuged at 200 g for 5 min. Cell pellets were re-suspended in 300 μ L ice-cold flow buffer.

Instrument	FACSCanto II (BD Biosciences).
Software	FlowJo (v.10) software https://www.flowjo.com/solutions/flowjo , RRID: SCR_008520
Cell population abundance	Data for 10,000 events in the post-sorted fraction were recorded for each sample. Purity was determined as falling within gated regions determined to exclude negative control (isotype-stained) samples.
Gating strategy	Forward and side scatter (FSC vs SSC) density plots were used to exclude debris. Forward scatter height versus forward scatter area density were used to exclude doublets. Single parameter histograms were used to identify INS+ or NKX6-1+ cells based on the isotype control signal.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.