Article

Interpreting type 1 diabetes risk with genetics and single-cell epigenomics

https://doi.org/10.1038/s41586-021-03552-w

Received: 21 June 2020

Accepted: 14 April 2021

Published online: 19 May 2021

Check for updates

Joshua Chiou[1,7]✉, Ryan J. Geusz[1], Mei-Lin Okino[2], Jee Yun Han[3], Michael Miller[3], Rebecca Melton[1], Elisha Beebe[2], Paola Benaglio[2], Serina Huang[2], Katha Korgaonkar[2], Sandra Heller[4], Alexander Kleger[4], Sebastian Preissl[3], David U. Gorkin[3,8], Maike Sander[2,5,6] & Kyle J. Gaulton[2,6]✉

Genetic risk variants that have been identified in genome-wide association studies of complex diseases are primarily non-coding[1]. Translating these risk variants into mechanistic insights requires detailed maps of gene regulation in disease-relevant cell types[2]. Here we combined two approaches: a genome-wide association study of type 1 diabetes (T1D) using 520,580 samples, and the identification of candidate *cis*-regulatory elements (cCREs) in pancreas and peripheral blood mononuclear cells using single-nucleus assay for transposase-accessible chromatin with sequencing (snATAC–seq) of 131,554 nuclei. Risk variants for T1D were enriched in cCREs that were active in T cells and other cell types, including acinar and ductal cells of the exocrine pancreas. Risk variants at multiple T1D signals overlapped with exocrine-specific cCREs that were linked to genes with exocrine-specific expression. At the *CFTR* locus, the T1D risk variant rs7795896 mapped to a ductal-specific cCRE that regulated *CFTR*; the risk allele reduced transcription factor binding, enhancer activity and *CFTR* expression in ductal cells. These findings support a role for the exocrine pancreas in the pathogenesis of T1D and highlight the power of large-scale genome-wide association studies and single-cell epigenomics for understanding the cellular origins of complex disease.

Type 1 diabetes is a complex autoimmune disease that is characterized by the loss of insulin-producing pancreatic beta cells[3], but the triggers of autoimmunity and disease onset remain poorly understood. T1D has a strong genetic component, most prominently at the major histocompatibility complex (MHC) locus, but including 59 additional risk loci[4–6]. Risk variants for T1D are largely non-coding, and intersection of risk variants with epigenomic data has shown that these variants are enriched within lymphoid enhancers[4]. However, owing to limited sample sizes, incomplete variant coverage, and limited cell-type resolution of existing epigenomic maps, the causal variants and cellular mechanisms of action of T1D risk loci are largely unresolved.

Discovery and fine mapping of T1D loci

We performed a genome-wide association study (GWAS) of 18,942 patients with T1D and 501,638 control participants of European ancestry from 9 cohorts (Supplementary Table 1). After applying uniform quality control (Supplementary Fig. 1), we imputed genotypes into the TOPMed reference panel and tested for association with T1D[7]. Through meta-analysis, we combined the association results for 61,947,369 variants and identified 81 loci that reached genome-wide significance ($P < 5 \times 10^{-8}$), including 48 of 59 known loci and 33 loci that were previously unreported, to our knowledge (Fig. 1a, Supplementary Fig. 2, Supplementary Table 2). At 92 total loci (59 known and 33 novel), we identified 44 independent signals, of which 36 were previously unreported (Fig. 1b, Supplementary Fig. 3). Nearly a third of loci (32%; 29 of 92) contained more than one signal; for example, the *PTPN2* and *BCL11A* loci had three signals each (Extended Data Fig. 1).

We fine-mapped causal variants for 136 T1D signals (92 main and 44 independent signals; Fig. 1b). We obtained the posterior probability of association (PPA) for tested variants and defined 99% credible sets for each signal (Supplementary Table 3, Supplementary Data 1). Compared to a previous study[8], fine-mapping resolution was improved on the basis of credible set size and maximum posterior probability (Supplementary Fig. 4). The median credible set size was 31 variants, where nearly a quarter (24%; 32 of 136) contained 5 or fewer variants, and 28% (38 of 136) contained a single variant with a PPA of less than 0.50 (Fig. 1c). Credible sets at 15% of signals (21 of 136) contained a nonsynonymous variant with PPA > 0.01, including the novel loci *AIRE* p.Arg471Cys (PPA = 0.99), *BATF3* p.Val11Ile (PPA = 0.078), *PRF1* p.Ala91Val (PPA = 0.28), and *INPP5B* p.Gly250Cys (PPA = 0.055) (Supplementary Table 4).

The TOPMed reference panel enables more accurate imputation of rare variants. We identified four novel variants with a minor allele frequency (MAF) of below 0.005 and large effects on T1D (Extended Data

[1]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA. [2]Department of Pediatrics, Pediatric Diabetes Research Center, University of California San Diego, La Jolla, CA, USA. [3]Center for Epigenomics, Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA. [4]Department of Internal Medicine I, Ulm University, Ulm, Germany. [5]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA. [6]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA. [7]Present address: Internal Medicine Research Unit, Pfizer Worldwide Research, Cambridge, MA, USA. [8]Present address: Department of Biology, Emory University, Atlanta, GA, USA. ✉e-mail: joshua.chiou@pfizer.com; kgaulton@health.ucsd.edu

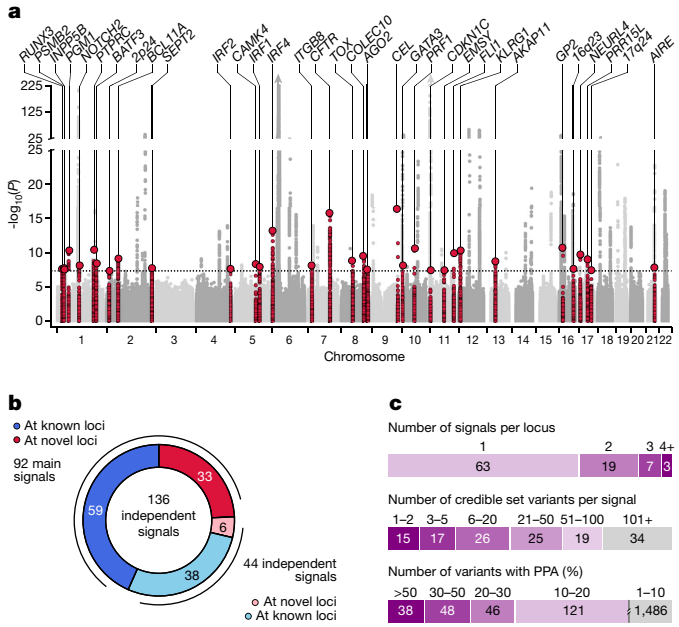


Fig. 1 | Genome-wide association and fine mapping identify T1D risk signals. a, Genome-wide T1D association (two-sided $-\log_{10}$-transformed *P* values from meta-analysis of $n = 520{,}580$ samples, unadjusted for multiple comparisons). Previously unidentified loci are coloured red (±250 kb of the index variant) and labelled with the nearest gene. Dotted line indicates genome-wide significance ($P = 5 \times 10^{-8}$). **b**, Breakdown of 136 T1D risk signals, including 92 main signals (59 known and 33 novel) and 44 independent signals (38 at known and 6 at novel loci). **c**, Number of signals per locus (top), 99% credible set variants from fine mapping (middle), and variants with posterior probability of association (PPA) at various thresholds (bottom).

Fig. 2a). Among these, rs541856133 (MAF = 0.0015, odds ratio (OR) = 3.01, 95% confidence interval (CI) = 2.33–3.89) mapped directly upstream of *CEL*, a gene that has been implicated in maturity-onset diabetes of the young type 8 (MODY8)[9]. We also identified a novel protein-coding protective variant at *IFIH1* (p.Asn160Asp, rs75671397, MAF = 0.002, OR = 0.35, 95% CI = 0.22–0.55) independent of known signals in this gene. Two additional non-coding risk variants mapped to *SH2B3* (rs570074821, MAF = 0.0019, OR = 1.89, 95% CI = 1.37–2.61) and *CAMK4* (rs72663304, MAF = 0.0013, OR = 2.54, 95% CI = 1.72–3.76) (Extended Data Fig. 2b).

We characterized genetic correlations (r_g) between T1D and other complex traits and diseases. Consistent with previous reports[4,10], T1D had significant (false discovery rate (FDR) < 0.10) positive correlations with autoimmune diseases such as rheumatoid arthritis ($r_g = 0.44$, FDR = 7.52×10^{-5}) and systemic lupus erythematosus ($r_g = 0.35$, FDR = 5.05×10^{-7}), and a negative correlation with ulcerative colitis ($r_g = -0.18$, FDR = 1.95×10^{-3}) (Extended Data Fig. 3). We also observed positive correlations with metabolic traits such as fasting insulin level ($r_g = 0.18$, FDR = 4.04×10^{-3}), coronary artery disease ($r_g = 0.12$, FDR = 1.23×10^{-2}), and type 2 diabetes (T2D; $r_g = 0.10$, FDR = 1.95×10^{-3}), and with pancreatic diseases such as pancreatic cancer ($r_g = 0.25$, FDR = 1.11×10^{-1}), although the latter was just above significance. These results demonstrate relationships between genetic effects on T1D and autoimmune, metabolic and pancreatic disease.

Pancreas and immune cell gene regulation

The majority of T1D risk is likely to affect gene regulation[4]. To annotate T1D risk variants, we generated an accessible chromatin reference map using snATAC–seq of peripheral blood and pancreas from donors without diabetes (Supplementary Table 5). We grouped chromatin accessibility profiles from 131,554 cells into 28 clusters (Fig. 2a, Supplementary Fig. 5a–c) and assigned cell-type identities using chromatin accessibility at marker genes (Supplementary Table 6). For example, chromatin accessibility at *C1QB* marked pancreas tissue-resident macrophages, *REG1A* marked acinar cells, and *CFTR* marked ductal cells (Extended Data Fig. 4a). We also observed patterns of chromatin accessibility at marker genes for cell subtypes, such as *FOXP3* for regulatory T cells (Extended Data Fig. 4a). To relate cell-type-resolved accessible chromatin to gene expression, we created a single-cell RNA sequencing (scRNA-seq) reference map of peripheral blood and pancreas. We assigned cell-type identities for 90,495 cells to 29 clusters, which identified similar cell types and proportions to snATAC–seq (Extended Data Fig. 5a–c).

To characterize *cis*-regulatory programs, we aggregated reads from cells within each snATAC–seq cluster and identified accessible chromatin peaks that represented cCREs. There were 448,142 cCREs across all 28 clusters and an average of 77,812 cCREs per cluster (Supplementary Data 2). We also aggregated reads from cells within each scRNA-seq cluster to derive normalized expression (Supplementary Data 3). To delineate regulatory programs that specified each cell type, we identified 25,436 cCREs with accessibility patterns that were most specific to each cluster (Fig. 2b, Supplementary Data 4). Genes within 100 kb of cCREs specific to a cell type had more specific expression for that cell type than genes that were close to cCREs specific to other cell types (Supplementary Fig. 6). Cell-type-specific cCREs were also enriched for gene ontology (GO) terms that represent highly specialized cellular processes (Fig. 2b, Supplementary Table 7).

We defined transcriptional regulators of cCRE activity by assessing transcription factor (TF) motif enrichment (Supplementary Data 5). Enriched TF motifs included those with lineage, cell-type, and cell-state specificity (Extended Data Fig. 4b). As TFs within subfamilies often have similar motifs, we grouped TFs into subfamilies to identify TFs with matching expression and motif enrichment patterns (Supplementary Table 8). For example, the FOXA subfamily TF genes *FOXA2* and *FOXA3* were specifically expressed in pancreatic endocrine and exocrine cells, the HNF1 subfamily TF gene *HNF1B* was specifically expressed in ductal cells, and the ROR subfamily TF gene *RORC* was specifically expressed in memory $CD8^+$ T cells (Extended Data Fig. 4b, Supplementary Table 8).

As the target genes of cCRE activity are largely unknown, we identified cell-type-resolved co-accessibility links between distal (non-promoter) cCREs and putative target gene promoters. Across all cell types, we observed 1,028,428 links (with co-accessibility more than 0.05) between distal cCREs and gene promoters (Supplementary Data 6). Co-accessible links were often cell-type-specific; for example, distal cCREs were co-accessible with the *AQP1* promoter in ductal cells and the *CEL* promoter in acinar cells (Extended Data Fig. 4c). In nearly every cell type, target genes that were co-accessible with distal cCREs were more likely than matched genes to be expressed in that cell type (Supplementary Fig. 7).

Cell-type annotation of T1D risk variants

We measured the enrichment of variants associated with T1D and other complex traits and diseases for cell-type cCREs. For T1D, the most significant enrichment was in T cell cCREs (naive T cell, Z-score (Z) = 5.57, FDR = 2.26×10^{-5}; memory $CD8^+$ T cell, $Z = 4.80$, FDR = 4.67×10^{-4}; activated $CD4^+$ T cell, $Z = 4.62$, FDR = 6.74×10^{-4}; cytotoxic $CD8^+$ T cell, $Z = 4.49$, FDR = 1.09×10^{-3}; regulatory T cell, $Z = 3.26$, FDR = 7.23×10^{-3}) and adaptive natural killer cells ($Z = 3.50$, FDR = 9.93×10^{-3}; Extended Data Fig. 6). Notably, we did not observe enrichment in pancreas-resident immune cells ($CD8^+$ T cell, $Z = 0.65$, FDR = 1.0; macrophage, $Z = -0.56$, FDR = 1.0). For other immune-related diseases, variants were primarily enriched within lymphocyte cCREs, whereas those associated with T2D and glycaemic traits were enriched in pancreatic endocrine, acinar, and ductal cCREs (Extended Data Fig. 6). These results show that T1D variants are broadly enriched for T cell cCREs and highlight other traits that are enriched for pancreatic and immune cell cCREs.

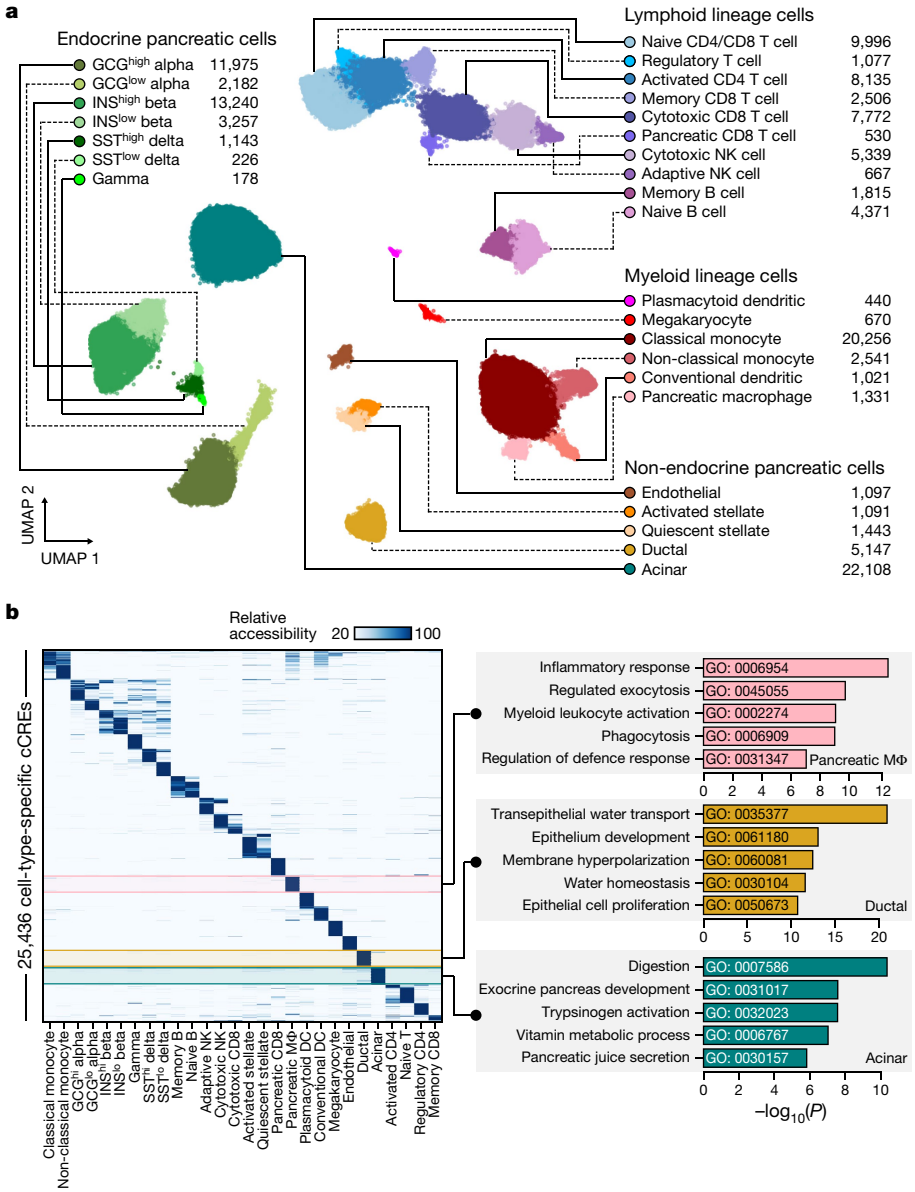


Fig. 2 | Reference map of single-cell chromatin accessibility from T1D-relevant tissues. a, Leiden clustering of single-cell accessible chromatin profiles from $n = 131{,}554$ cells. Cells are plotted on the first two uniform manifold approximation and projection (UMAP) components, clusters are grouped into categories of cell types, and the number of cells per cluster is shown to the right of each list. **b**, Relative accessibility (row-normalized) for the 25,436 cCREs that were most specific to each cluster (left), and enriched gene ontology terms for cCREs that were specific to pancreatic macrophages, ductal cells and acinar cells (right). MΦ, macrophage; DC, dendritic cell; NK, natural killer.

Despite the strong enrichment of T1D-associated variants in T cells, many T1D signals did not overlap a T cell cCRE, which suggests that additional cell types contribute to T1D risk. To identify additional disease-relevant cell types, we used an orthogonal approach to test for enrichment of T1D variants within the subset of cell-type-specific cCREs. As expected, T1D-associated variants were enriched in cCREs specific to T cells and beta cells (activated $CD4^+$ T cell, ln(enrich) = 4.25, 95% CI = 1.11–5.43; cytotoxic $CD8^+$ T cell, ln(enrich) = 4.04, 95% CI = 0.20–5.20; INS^{high} beta cell, ln(enrich) = 3.58, 95% CI = 0.95–4.84) (Fig. 3a). Notably, T1D variants were also enriched in cCREs specific to plasmacytoid dendritic cells (pDC) (ln(enrich) = 4.00, 95% CI = 1.96–5.10), classical monocytes (ln(enrich) = 3.78, 95% CI = 2.23–4.74), acinar cells (ln(enrich) = 3.35, 95% CI = 1.59–4.46) and ductal cells (ln(enrich) = 3.28, 95% CI = 0.18–4.69) (Fig. 3a).

Given insight into key T1D-relevant cell types, we next annotated T1D signals in cCREs for these cell types. More than 75% of T1D signals (103 of 136) contained at least one variant (PPA > 0.01) that overlapped with a cCRE, and at 65% of these signals (67 of 103) the cCRE was co-accessible with a gene promoter (Supplementary Table 9). Variants with high probabilities (PPA > 0.50) were significantly more likely than other credible set variants to map in a cCRE (OR = 3.9, 95% CI = 1.9–7.8, $P = 1.9 \times 10^{-4}$), and these cCREs were more likely to be co-accessible with a promoter (OR = 6.1, 95% CI = 1.3–55.9, $P = 7.1 \times 10^{-3}$). For each signal, we calculated the cumulative posterior probability (cPPA) of credible set variants overlapping distal cCREs in each disease-enriched cell type. Numerous T1D signals had high cPPA in T cell cCREs and not in other disease-relevant cell types (Fig. 3b). We also found T1D signals that had high cPPA in acinar and ductal (exocrine) cell, beta cell, monocyte and pDC cCREs, several of which were highly cell-type-specific (Fig. 3b). For each signal, we further annotated genes within 1 Mb that were expressed in the same cell type and co-accessible with cCREs (Fig. 3b, Supplementary Table 9).

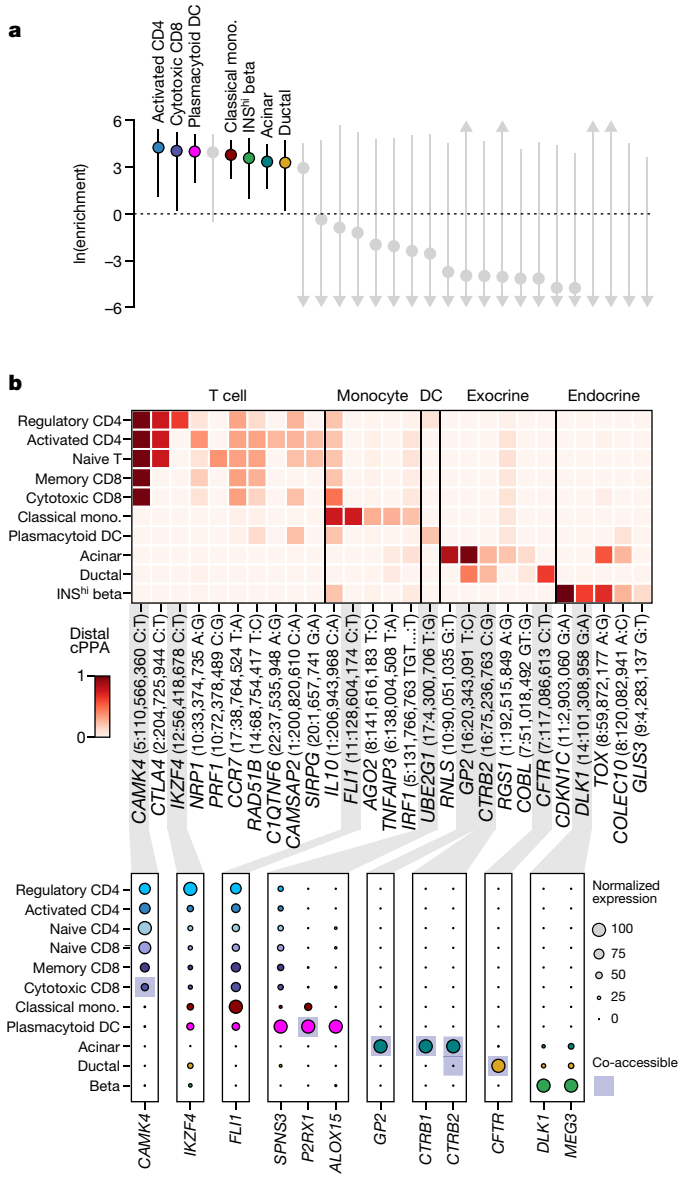


Fig. 3 | Cell-type-specific enrichment and mechanisms of T1D risk variants. **a**, T1D ln(enrichment) within cell-type-specific cCREs. Labelled cell types have positive enrichment and 95% CI lower bound above 0. Data are ln(enrichment) ± 95% CI from fgwas. **b**, T1D signals with highest cPPA in cCREs for disease-enriched cell types (more than 0.20 cPPA for T cells and monocytes, more than 0.10 cPPA for other groups), and more than 0.01 cPPA away from the next closest group (top). Column-normalized expression for genes with transcripts per million (TPM) above 1 in the highlighted cell type(s) and within ±500 kb of the index variant. Genes that are co-accessible with cCREs that contain risk variants are annotated in rectangles (bottom). Mono., monocyte. Index variants shown in parentheses.

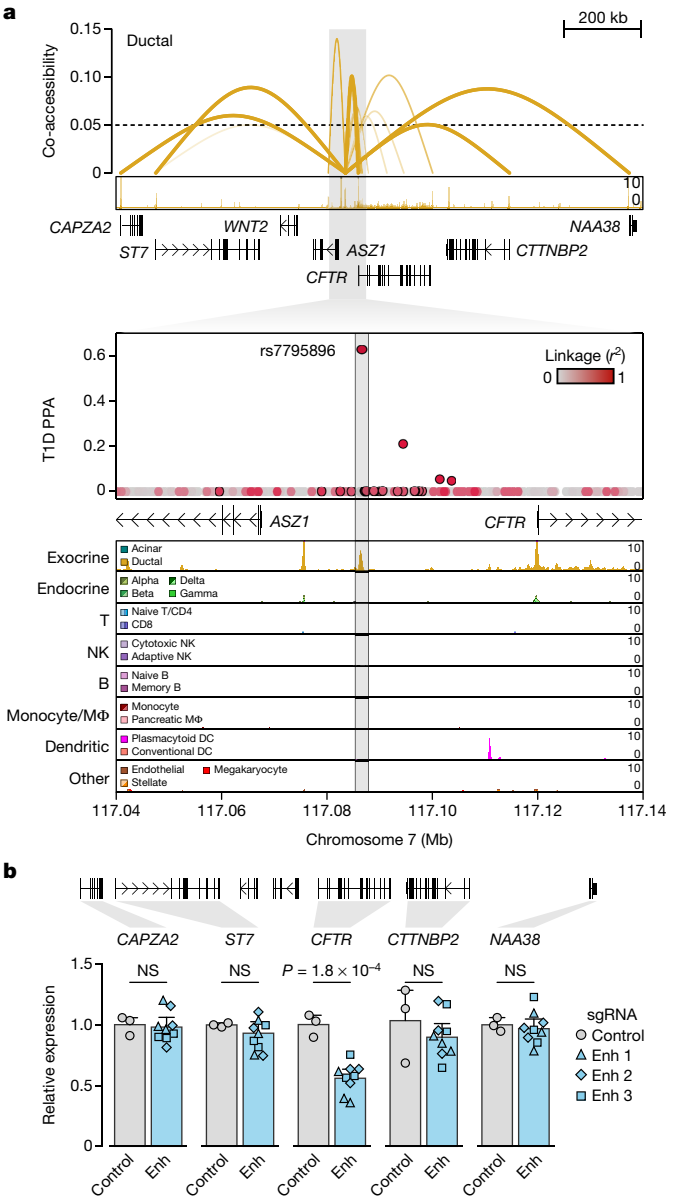


Fig. 4 | Fine-mapped T1D variant regulates *CFTR* in pancreatic ductal cells. **a**, Variant rs7795896 at the *CFTR* locus mapped in a cCRE that is co-accessible with *CFTR* and other genes. Zoomed-in view (top) shows that the cCRE is ductal cell-specific. **b**, Expression of genes that are co-accessible with the distal cCRE in Capan-1 cells with CRISPR-inactivated enhancer (Enh; $n = 9$, 3 single-guide RNAs (sgRNAs) × 3 biological replicates) compared to non-targeting control sgRNA ($n = 3$ biological replicates). Data are mean ± 95% CI. *P* values by two-sided ANOVA; NS, not significant.

Multiple T1D signals had high cPPA specifically in pancreatic exocrine cells and were linked to genes with exocrine-specific expression. At the *GP2* locus, three variants accounted for 0.951 PPA and mapped in an acinar-specific cCRE that was co-accessible with the promoter of *GP2*, which had acinar-specific expression (Fig. 3b, Extended Data Fig. 7a). Similarly, rs72802342 at the *BCAR1* locus (PPA = 0.30) mapped in an acinar-specific cCRE that was co-accessible with the promoters of *CTRB1* and *CTRB2*, both of which had acinar-specific expression (Fig. 3b, Extended Data Fig. 7b). Other signals such as *CEL* had similar exocrine-specific profiles (Supplementary Fig. 8a–c). Exocrine cCREs at T1D loci were also largely specific relative to accessible chromatin in stimulated immune and islet cells (Supplementary Table 10).

T1D variant affects *CFTR* in ductal cells

The *CFTR* locus contained the fine-mapped variant rs7795896 (PPA = 0.63) in a distal cCRE that is specific to ductal cells and co-accessible with the *CFTR* promoter, in addition to other genes (Fig. 4a). Recessive mutations in *CFTR* cause cystic fibrosis, which is often comorbid with exocrine pancreas insufficiency and cystic fibrosis-related diabetes (CFRD)[11]. Furthermore, carriers of *CFTR* mutations often develop chronic pancreatitis[12]. As *CFTR* has not been

implicated in T1D, we sought to validate the mechanism of this locus. The T1D risk allele of rs7795896 significantly reduced enhancer activity (594-bp sequence two-sided ANOVA, $P = 1.15 \times 10^{-2}$, Extended Data Fig. 8a; 180-bp sequence two-sided *t*-test, $P = 3.35 \times 10^{-2}$, Extended Data Fig. 8b) and reduced protein binding (bound fraction rs7795896 C allele = 0.007, T allele = 0.081; Extended Data Fig. 8c, Supplementary Fig. 9) in Capan-1 cells. The variant mapped in a sequence motif for HNF1B, albeit in a position predicted to minimally affect binding, and overlapped a HNF1B chromatin immunoprecipitation with sequencing (ChIP–seq) site that was previously identified in ductal cells[13] (Extended Data Fig. 8d).

To determine whether the enhancer that contains rs7795896 regulates *CFTR* in ductal cells, we used CRISPR interference (CRISPRi) to inactivate enhancer activity ($CFTR^{Enh}$) in Capan-1 cells (Supplementary Table 11). As positive and negative controls, we inactivated the *CFTR* promoter ($CFTR^{Prom}$) and used a non-targeting guide RNA, respectively. Quantitative PCR (qPCR) showed that there was a significant reduction in *CFTR* expression after enhancer inactivation (two-sided ANOVA, $P = 1.77 \times 10^{-4}$), whereas expression of other genes at the locus was unchanged (Fig. 4b, Extended Data Fig. 8e). We tested whether risk variants affected *CFTR* expression using pancreas expression quantitative trait locus (eQTL) data from the GTEx Consortium[14]. Out of 13 tested genes, only *CFTR* had evidence for an eQTL ($P = 4.31 \times 10^{-4}$), which was colocalized with the T1D signal (shared posterior probability (PP_{shared}) = 91.8%) (Extended Data Fig. 9a). Among candidate variants for which eCAVIAR provided evidence for driving the shared signal (colocalization posterior probability (CLPP) > 0.01), only rs7795896 mapped in a cCRE. The T1D risk allele of rs7795896 was associated with decreased *CFTR* expression, consistent with its effects on enhancer activity and TF binding. We recalculated the eQTL association to include estimated pancreas cell-type proportion as an interaction term, and only ductal cells showed a significant association ($P = 2.37 \times 10^{-4}$) (Extended Data Fig. 9b–d).

As *CFTR* has been implicated in pancreatic cancer[15] and pancreatitis[16], we investigated whether rs7795896 was associated with these phenotypes in the UK Biobank and FinnGen. The T1D risk allele was associated with increased risk of pancreatitis (chronic pancreatitis OR = 1.15, $P = 3.18 \times 10^{-3}$; acute pancreatitis OR = 1.07, $P = 1.15 \times 10^{-2}$) and other pancreatic diseases (OR = 1.13, $P = 4.72 \times 10^{-5}$) (Extended Data Fig. 10a). By contrast, rs7795896 was not associated with other autoimmune diseases (all $P > 0.05$). T1D signals that were associated with increased risk of pancreatic disease had significantly higher cPPA in exocrine cCREs compared to other signals (two-sided Student's *t*-test, $P = 0.027$) and showed no difference in T cell cCREs ($P = 0.36$). Together, our findings support a model in which variants that regulate *CFTR* and other genes in the exocrine pancreas increase the risk of T1D and pancreatic diseases (Extended Data Fig. 10b).

High-resolution mapping of both genetic variants that influence T1D risk and cell-type-specific *cis*-regulatory programs in T1D-relevant tissues enabled us to gain insight into disease mechanisms. Risk variants at multiple loci mapped to genes with specialized functions in exocrine cells. Although our results support the idea that variants in exocrine cCREs mediate T1D risk, fine mapping has not resolved a single variant at most loci. Risk variants in exocrine-specific cCREs may also function in other cell types in the context of development, environmental changes, or disease progression. Continued fine mapping in trans-ethnic cohorts with systematic evaluation of variant function in relevant cell types will further clarify risk mechanisms. Furthermore, as co-accessible links represent correlations that require both sites to vary in their accessibility, future studies will benefit from linking changes in chromatin to gene expression directly through single-cell multi-omics.

Observational studies have reported exocrine pancreas abnormalities at the onset of T1D[17], but it was unknown whether this caused the disease[18]. Genomic studies have also identified changes in exocrine cells in patients with T1D[19,20]. Abnormalities in the exocrine pancreas have been considered secondary to other disease processes in T1D, such as beta cell loss causing reduced insulinotropic effects on exocrine cells or viral infection leading to exocrine inflammation. By contrast, our findings provide evidence that exocrine cells intrinsically contribute to T1D. Reduced *CFTR* leads to CFRD via intra-islet inflammation and immune infiltration, and immune infiltration in the exocrine pancreas has been suggested to contribute to T1D[21–23]. Other implicated genes encode proteins that are secreted from acinar cells and have been linked to risk of pancreatic disease[24–26], and may contribute to an inflammatory state. We therefore hypothesize that gene regulation in the exocrine pancreas has a causal role in T1D, which may provide new avenues for therapeutic discovery.

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/s41586-021-03552-w.

1. Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Katsarou, A. et al. Type 1 diabetes mellitus. *Nat. Rev. Dis. Primers* **3**, 17016 (2017).
4. Onengut-Gumuscu, S. et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
5. Barrett, J. C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
6. Bradfield, J. P. et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7**, e1002293 (2011).
7. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
8. 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.* ddy314 (2018).
9. Raeder, H. et al. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat. Genet.* **38**, 54–62 (2006).
10. Ramos, P. S. et al. A comprehensive analysis of shared loci between systemic lupus erythematosus (SLE) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS Genet.* **7**, e1002406 (2011).
11. Gibson-Corley, K. N., Meyerholz, D. K. & Engelhardt, J. F. Pancreatic pathophysiology in cystic fibrosis. *J. Pathol.* **238**, 311–320 (2016).
12. Sharer, N. et al. Mutations of the cystic fibrosis gene in patients with chronic pancreatitis. *N. Engl. J. Med.* **339**, 645–652 (1998).
13. Diaferia, G. R. et al. Dissection of transcriptional and cis-regulatory control of differentiation in human pancreatic cancer. *EMBO J.* **35**, 595–617 (2016).
14. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
15. McWilliams, R. R. et al. Cystic fibrosis transmembrane conductance regulator (CFTR) gene mutations and risk for pancreatic adenocarcinoma. *Cancer* **116**, 203–209 (2010).
16. Noone, P. G. et al. Cystic fibrosis gene mutations and pancreatitis risk: relation to epithelial ion transport and trypsin inhibitor gene mutations. *Gastroenterology* **121**, 1310–1319 (2001).
17. Virostko, J. et al. Pancreas volume declines during the first year after diagnosis of type 1 diabetes and exhibits altered diffusion at disease onset. *Diabetes Care* **42**, 248–257 (2019).
18. Campbell-Thompson, M., Rodriguez-Calvo, T. & Battaglia, M. Abnormalities of the exocrine pancreas in type 1 diabetes. *Curr. Diab. Rep.* **15**, 79 (2015).
19. Camunas-Soler, J. et al. Patch-seq links single-cell transcriptomes to human islet dysfunction in diabetes. *Cell Metab.* **31**, 1017–1031 (2020).
20. Fasolino, M. et al. Multiomics single-cell analysis of human pancreatic islets reveals novel cellular states in health and type 1 diabetes. Preprint at https://doi.org/10.1101/2021.01.28.428598 (2021).
21. Hart, N. J. et al. Cystic fibrosis-related diabetes is caused by islet loss and inflammation. *JCI Insight* **3**, e98240 (2018).
22. Valle, A. et al. Reduction of circulating neutrophils precedes and accompanies type 1 diabetes. *Diabetes* **62**, 2072–2077 (2013).
23. Navis, A. & Bagnat, M. Loss of *cftr* function leads to pancreatic destruction in larval zebrafish. *Dev. Biol.* **399**, 237–248 (2015).
24. Lin, Y. et al. Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. *Nat. Commun.* **11**, 3175 (2020).
25. Wolpin, B. M. et al. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat. Genet.* **46**, 994–1000 (2014).
26. Johansson, B. B. et al. The role of the carboxyl ester lipase (CEL) gene in pancreatic disease. *Pancreatology* **18**, 12–19 (2018).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Genotype quality control and imputation

We compiled individual-level genotype data and summary statistics from 18,942 individuals with T1D and 501,638 control individuals of European ancestry from public sources (Supplementary Table 1), where T1D case cohorts were matched to population control cohorts on the basis of genotyping array (Affymetrix, Illumina Infinium, Illumina Omni, and Immunochip) and country of origin where possible (USA, UK, and Ireland). For the GENIE-UK cohort, because we were unable to find a matched country of origin control cohort, we used individuals of British ancestry (defined by individuals within 1.5 interquartile range of CEU/GBR subpopulations on the first four principal components (PCs) from principal component analysis (PCA) with European 1000 Genomes Project samples) from the University of Michigan Health and Retirement study (HRS). For non-UK Biobank cohorts, we first applied individual and variant exclusion lists (where available) to remove low-quality, duplicate, or non-European ancestry samples and failed genotype calls for each cohort. For control cohorts, we also used phenotype files (where available) to remove individuals with T2D or autoimmune diseases.

We then applied the HRC imputation preparation program (version 4.2.9) and used PLINK[27] (version 1.90b6.7) to remove variants based on (i) low frequency (MAF < 1%), (ii) missing genotypes (missing >5%), (iii) violation of Hardy–Weinberg equilibrium (HWE $P < 1 \times 10^{-5}$ in control cohorts and HWE $P < 1 \times 10^{-10}$ in case cohorts), (iv) difference in allele frequency >0.2 compared to the Haplotype Reference Consortium r1.1 reference panel[28], and (v) allele ambiguity defined as AT/GC variants with MAF >40%[29]. We further removed individuals for (i) missing genotypes (missing >5%), (ii) sex mismatch with phenotype records ($hom_{chrX} > 0.2$ for females and $hom_{chrX} < 0.8$ for males), (iii) cryptic relatedness through identity-by-descent (IBD >0.2), and (iv) non-European ancestry through PCA with 1000 Genomes Project[30] (>3 interquartile range from 25th and 75th percentiles of European 1KGP samples on the first four PCs) (Supplementary Fig. 1). Lists of independent variants for IBD and PCA calculations were generated using PLINK ('--indep 50 5 2'). For the affected sib-pair (ASP) cohort genotyped on the Immunochip, we retained only one T1D sample from each family selected at random. For the GRID case and 1958 birth control cohorts genotyped on the Immunochip, a portion of the cases overlapped the T1DGC or 1958 birth cohorts genotyped on a genome-wide array. We thus used sample IDs from the phenotype files to remove these samples from the GRID and 1958 birth cohorts and verified that no samples were duplicated between the Immunochip and genome-wide array data sets by checking IBD. We combined data for matched case and control cohorts based on genotyping array and country of origin for imputation. We used the TOPMed Imputation Server[31] to impute genotypes into the TOPMed r2 panel[7] and removed variants based on low imputation quality ($R^2 < 0.3$). Following imputation, we implemented post-imputation filters to remove variants based on potential genotyping or imputation artefacts based on empirical R^2 (genotyped variants with empirical $R^2 < 0.5$ and all imputed variants in at least low linkage disequilibrium; LD, $r^2 > 0.3$).

For the UK Biobank cohort, we downloaded imputed genotype data from the UK Biobank v3 release that were imputed using a combination of the HRC and UK10K + 1000 Genomes reference panels. We removed data for individuals who had withdrawn participation from the UK Biobank. We used phenotype data to remove individuals of non-European descent. To resolve duplicate samples represented in both the UK Biobank and other cohorts on different genotyping arrays, we calculated IBD between samples in the UK Biobank and cohorts of UK origin, removing duplicated samples from the UK Biobank (IBD > 0.9). Following these filters, we then used a combination of ICD10 (International Classification of Diseases 10th Revision) codes to define 1,445 T1D cases (T1D diagnosis, insulin treatment within a year of diagnosis, no T2D diagnosis). We defined controls as 362,050 individuals without diabetes (no T1D, T2D, or gestational diabetes diagnosis) or other autoimmune diseases (systemic lupus erythematosus, rheumatoid arthritis, juvenile arthritis, Sjögren syndrome, alopecia areata, multiple sclerosis, autoimmune thyroiditis, vitiligo, coeliac disease, primary biliary cirrhosis, psoriasis, or ulcerative colitis). We removed variants with low imputation quality ($R^2 < 0.3$).

For the FinnGen cohort, we downloaded GWAS summary statistics for T1D (T1D_STRICT) from FinnGen freeze 3 (http://r3.finngen.fi/). This phenotype definition excluded individuals with T2D from both cases and controls.

Association testing and meta-analysis

We tested variants with MAF $>1 \times 10^{-5}$ for association to T1D with Firth bias reduced logistic regression using EPACTS (https://genome.sph.umich.edu/wiki/EPACTS) for non-UK Biobank cohorts or SAIGE[32] (version 0.38) for the UK Biobank, using genotype dosages adjusted for sex and the first four ancestry PCs. For the UK Biobank we used SAIGE as it is designed to run on biobank-scale cohorts and with highly imbalanced ratios of cases to controls. For FinnGen, we used association results from the freeze 3 release that were generated using SAIGE. Before meta-analysis, we used liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert GRCh37/hg19 into GRCh38/hg38 coordinates for the UK Biobank. We then combined association results across matched cohorts through inverse-variance weighted meta-analysis. We used liftOver to convert GRCh38/hg38 back into GRCh37/hg19 coordinates for the meta-analysis. We removed variants that could not be converted, were duplicated after coordinate conversion, or were located on different chromosomes after conversion. In total, our association data contained summary statistics for 61,947,369 variants. To evaluate the extent to which genomic inflation was driven by the polygenic nature of T1D or population stratification, we used LD score regression[33] to compare the LDSC intercept to lambda genomic control (GC). We observed an intercept of 1.07 (s.e. = 0.03) compared to a lambda GC of 1.20, suggesting that the majority of the observed inflation was driven by polygenicity rather than population stratification.

Stochastic search and fine mapping of independent signals

We identified 59 loci (excluding the MHC locus) with T1D risk variants that had been reported in previous genetic studies of T1D[4–6,34], and considered a locus in our study known if the most associated variant mapped within 500 kb of a previously reported T1D variant. We defined 33 novel loci where a variant reached genome-wide significance ($P < 5 \times 10^{-8}$), and both mapped at least 500 kb away and was not in LD ($r^2 < 0.01$) with a previously reported T1D variant. At 92 (59 known and 33 novel) loci, we defined the 'index' variant as the variant with strongest T1D association at the locus.

For all 92 loci, we used a 1-Mb window around the index variant as the region for fine mapping using FINEMAP[35] (version 1.4). For each region, we first filtered for variants with MAF > 0.0005 and constructed pairwise LD matrices with PLINK[27] ('--r --square --keep-allele-order') using the TOPMed2-imputed cohorts with genome-wide coverage (DCCT-EDIC, GENIE-ROI, GENIE-UK, GoKinD, T1DGC, WTCCC1-T1D and their respective control cohorts). We then applied FINEMAP using these matrices to conduct shotgun stochastic search and Bayesian fine mapping using the default prior ('--sss --n-causal-snps 10 --prob-cred-set 0.99 --prior-std 0.05'). We selected the number of independent signals (causal variants) for each region based on the configuration with the highest FINEMAP posterior probability and used 99% credible sets from the FINEMAP output for the resulting signals. We calculated the effective sample size for all credible set variants, and no credible set variant with PPA > 0.01 had <50% of the maximum effective sample

size. We compared fine-mapping results to a previous fine-mapping dataset[8]. At 56 signals in common to both studies, we calculated the number of variants in the 99% credible set and the probability of the most likely causal variant.

GWAS correlation analyses

We used LD score regression[29,33] (version 1.0.1) to estimate genome-wide genetic correlations between T1D and immune diseases[36–44], other diseases[45–55], and non-disease traits[56–74], using European subsets of GWAS where applicable. For acute pancreatitis, chronic pancreatitis, and pancreatic cancer, we used inverse variance weighted meta-analysis to combine SAIGE analysis results from the UK Biobank[32] (PheCodes 577.1, 577.2, and 157) and FinnGen r3 (K11_ACUTPANC, K11_CHRONPANC, C3_PANCREAS_EXALLC). We used pre-computed European 1000 Genomes LD scores to calculate correlation estimates (r_g) and standard errors. We then corrected *P* values for multiple tests using FDR correction and considered FDR < 0.1 as significant. We also performed genetic correlation analyses using a version of the T1D meta-analysis excluding the Immunochip cohorts and observed highly similar results.

Generation of snATAC–seq libraries

Combinatorial indexing single-cell ATAC–seq (snATAC–seq). We performed snATAC–seq as described previously[75–77] with several modifications as described below. For the islet samples, approximately 3,000 islet equivalents (IEQ, roughly 1,000 cells each) were resuspended in 1 ml nuclei permeabilization buffer (10 mM Tris-HCL (pH 7.5), 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20 (Sigma), 0.1% IGEPAL-CA630 (Sigma) and 0.01% Digitonin (Promega) in water) and homogenized using a 1-ml glass dounce homogenizer with a tight-fitting pestle for 15 strokes. Homogenized islets were incubated for 10 min at 4 °C and filtered with a 30-µm filter (CellTrics). For the pancreas samples, frozen tissue was pulverized with a mortar and pestle while frozen and immersed in liquid nitrogen. Approximately 22 mg of pulverized tissue was then transferred to an Eppendorf tube and resuspended in 1 ml cold permeabilization buffer for 10 min on a rotator at 4 °C. Permeabilized sample was filtered with a 30-µm filter (CellTrics), and the filter was washed with 300 µl permeabilization buffer to increase nucleus recovery.

Once permeabilized and filtered, nuclei were pelleted with a swinging bucket centrifuge (500*g*, 5 min, 4 °C; 5920R, Eppendorf) and resuspended in 500 µl high-salt tagmentation buffer (36.3 mM Tris acetate (pH 7.8), 72.6 mM potassium acetate, 11 mM Mg acetate, 17.6% DMF) and counted using a haemocytometer. Concentration was adjusted to 4,500 nuclei per 9 µl, and 4,500 nuclei were dispensed into each well of a 96-well plate. Glycerol was added to the leftover nucleus suspension for a final concentration of 25% and nuclei were stored at −80 °C. For tagmentation, 1 µl barcoded Tn5 transposomes were added using a BenchSmart 96 (Mettler Toledo), mixed five times and incubated for 60 min at 37 °C with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µl of 40 mM EDTA was added to each well with a BenchSmart 96 (Mettler Toledo) and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 µl 2 × sort buffer (2% BSA, 2 mM EDTA in PBS) was added using a BenchSmart 96 (Mettler Toledo). All wells were combined into a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using an SH800 (Sony), 20 nuclei were sorted per well into eight 96-well plates (total of 768 wells) containing 10.5 µl EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma)). Preparation of sort plates and all downstream pipetting steps were performed on a Biomek i7 Automated Workstation (Beckman Coulter). After addition of 1 µl 0.2% SDS, samples were incubated at 55 °C for 7 min with shaking (500 rpm). We added 1 µl 12.5% Triton-X to each well to quench the SDS and 12.5 µl NEBNext High-Fidelity 2 × PCR Master Mix (NEB). Samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72 °C 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were combined. Libraries were purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRI Beads (Beckmann Coulter, 0.55× and 1.5×). Libraries were purified one more time with SPRI Beads (Beckmann Coulter, 1.5×). Libraries were quantified using a Qubit fluorimeter (Life technologies) and the nucleosomal pattern was verified using a TapeStation (High Sensitivity D1000, Agilent). The library was sequenced on a HiSeq2500 sequencer (Illumina) using custom sequencing primers, 25% spike-in library and the following read lengths: 50 + 43 + 40 + 50 (Read1 + Index1 + Index2 + Read2).

Droplet-based 10x single-cell ATAC–seq (snATAC–seq). The 10x single-cell ATAC–seq protocol from 10x Genomics was followed: Chromium SingleCell ATAC ReagentKits UserGuide (CG000209, Rev A). Cryopreserved PBMC samples were thawed in a 37 °C water bath for 2 min according to the 'PBMC thawing protocol' in the UserGuide. After cells were thawed, the pellets were resuspended in 1 ml chilled PBS (with 0.04% PBS) and filtered with 50 µm CellTrics (04-0042-2317, Sysmex). The cells were centrifuged (300*g*, 5 min, 4 °C) and permeabilized with 100 µl chilled lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween-20, 0.1% IGEPAL-CA630, 0.01% digitonin and 1% BSA). The samples were incubated on ice for 3 min and resuspended with 1 ml chilled wash buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween-20 and 1% BSA). After centrifugation (500*g*, 5 min, 4 °C), the pellets were resuspended in 100 µl chilled Nuclei buffer (2000153, 10x Genomics). The nucleus concentration was adjusted to between 3,000 and 7,000 per µl and 15,300 nuclei (which targets 10,000 nuclei) were used for the experiment. For pancreas tissue (pulverized as described above), approximately 31.7 mg of pulverized tissue was transferred to a LoBind tube (Eppendorf) and resuspended in 1 ml cold permeabilization buffer (10 mM Tris-HCL (pH 7.5), 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20 (Sigma), 0.1% IGEPAL-CA630 (Sigma), 0.01% Digitonin (Promega) and 1% BSA (Proliant 7500804) in water) for 10 min on a rotator at 4 °C. Permeabilized nuclei were filtered with a 30-µm filter (CellTrics). Filtered nuclei were pelleted with a swinging bucket centrifuge (500*g*, 5 min, 4 °C; 5920R, Eppendorf) and resuspended in 1 ml wash buffer (10 mM Tris-HCL (pH 7.5), 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, and 1% BSA (Proliant 7500804) in molecular biology-grade water). The nucleus wash was repeated once. Next, washed nuclei were resuspended in 30 µl 1× nuclei buffer (10x Genomics). Nuclei were counted using a haemocytometer, and finally the nucleus concentration was adjusted to 3,000 nuclei per µl. We used 15,360 nuclei as input for tagmentation.

Nuclei were diluted to 5 µl with 1× nuclei buffer (10x Genomics) and mixed with ATAC buffer (10x Genomics) and ATAC enzyme (10x Genomics) for tagmentation (60 min, 37 °C). Single-cell ATAC–seq libraries were generated using the Chromium Chip E Single Cell ATAC kit (10x Genomics, 1000086) and indexes (Chromium i7 Multiplex Kit N, Set A, 10x Genomics, 1000084) following the manufacturer's instructions. Final libraries were quantified using a Qubit fluorimeter (Life Technologies) and the nucleosomal pattern was verified using a TapeStation (High Sensitivity D1000, Agilent). 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).

Single-cell chromatin accessibility data processing

Before read alignment, we used trim_galore (version 0.4.4) to remove adaptor sequences from reads using default parameters. For combinatorial barcoding data, we aligned reads to the hg19 reference genome using bwa mem[78] (version 0.7.17-r1188; '-M -C') and removed low mapping quality (MAPQ < 30), secondary, unmapped, and mitochondrial reads using samtools[79] (version 1.10). To remove duplicate sequences on a per-barcode level, we used the MarkDuplicates tool from picard ('BARCODE_TAG'). For droplet-based 10x data, we used Cell Ranger ATAC (version 1.1.0) to process, align, and remove duplicate reads. For each tissue and snATAC–seq technology, we used log-transformed read depth distributions from each experiment to determine a threshold

separating real cell barcodes from background noise. We used >500 total reads for combinatorial barcoding snATAC–seq and >2,300–4,000 total reads, as well as >0.3 fraction of reads in peaks, for 10x snATAC–seq experiments (Supplementary Fig. 7a).

Single-cell chromatin accessibility clustering

We identified snATAC–seq clusters using a previously described pipeline with a few modifications[75]. For each experiment, we first constructed a counts matrix consisting of read counts in 5-kb windows for each cell. Using scanpy[80] (version 1.4.4.post1), we normalized cells to a uniform read depth and log-transformed counts. We extracted highly variable (hv) windows ('min_mean = 0.01, min_disp = 0.25') and regressed out the total log-transformed read depth within hv windows (usable counts). We then merged datasets from the same tissue and performed PCA to extract the top 50 PCs. We used Harmony[81] (version 1.0) to correct the PCs for batch effects across experiments, using categorical covariates including donor-of-origin, biological sex, and snATAC–seq assay technology. We used the corrected components to construct a 30 nearest neighbour graph using the cosine metric, which we used for UMAP dimensionality reduction ('min_dist = 0.3') and clustering with the Leiden algorithm[82] ('resolution = 1.5').

Before combining cells across all tissues, we performed iterative clustering to identify and remove cells with low fraction of reads in peaks (using preliminary peaks called from data in bulk) or low usable counts (islets: 948, pancreas: 2,588, PBMCs: 5,268 cells removed in total). Next, after removing low-quality cells and repeating the previous clustering steps, we sub-clustered the resulting main clusters at high resolution ('resolution = 3.0') to identify sub-clusters containing potential doublets (islets: 886, pancreas: 4,495, PBMCs: 5,844 cells removed in total). We noted that these sub-clusters tended to have higher average usable counts, promoter usage, and accessibility at more than one marker gene promoter. After removing 20,029 low-quality or potential doublet cells, we performed a final round of clustering using experiments from all tissues, including tissue-of-origin as another covariate. We further removed 672 cells that mapped to improbable cluster assignments (islet or pancreatic cells in PBMC clusters or vice versa). After all filters, we ended up with 131,554 cells mapping to 28 distinct clusters with consistent representation across samples from the same tissue (Supplementary Fig. 7b). We catalogued known marker genes for each cell type using a combination of literature search and PanglaoDB[83] (Supplementary Table 6) and assessed gene accessibility (sum of read counts across each gene body) to assign labels to each cluster.

Single-cell gene expression clustering

We compiled publicly available scRNA-seq data sets of peripheral blood (10x Genomics; v1 Chemistry: 3k, 6k, and 33k; v2 Chemistry: 4k and 8k; v3 Chemistry: 5k and 10k; v3.1 Chemistry: 5k, 10k single indexed, and 10k dual indexed) and pancreatic islets[84]. We re-processed each dataset using Cell Ranger RNA (version 4.0.0) with the GRCh37 reference genome and removed cells with <500 genes expressed (non-zero counts). We extracted hv genes for PBMCs and pancreatic islets separately and merged both lists to obtain a single set of hv genes. For each sample, we used count matrices as input for scanpy[80] (version 1.4.4.post1), normalized counts for each cell to uniform read depth, log-transformed the normalized counts, and regressed out the log total counts for hv genes. We then merged all datasets and extracted the top 100 PCs using PCA. We used Harmony[81] (version 1.0) to correct PCs for covariates including the experiment, donor, tissue, and biological sex. We constructed a 30 nearest neighbour graph using the cosine metric, performed UMAP dimensionality reduction ('min_dist = 0.3'), and clustered with the Leiden algorithm[82] ('resolution = 1.25'). We performed iterative clustering to remove 10,014 low quality and 5,286 potential doublet cells, leaving 90,495 cells for the cell-type-resolved expression reference map. We used a combination of literature search and PanglaoDB[83] (Supplementary Table 6) to assign labels to each cluster. For each cell type, we normalized aggregated reads from individual cells to derive TPM for each gene.

Cataloguing cell-type-resolved cCREs

We identified chromatin accessibility peaks with MACS2[85] (version 2.1.2) by calling peaks on aggregated reads from each cluster. In brief, we extracted reads from all cells within a given cluster, shifted reads aligned to the positive strand by +4 bp and reads aligned to the negative strand by −5 bp, and centred the reads. We then used MACS2 to call peaks (' --nomodel --keep-dup-all') and removed peaks that overlapped ENCODE blacklisted regions[2,86]. We then merged peaks from all 28 clusters with bedtools[87] (version 2.26.0) to create a consistent set of 448,142 cCREs for subsequent analyses.

To compare accessible chromatin profiles from snATAC–seq to those from bulk ATAC–seq on FACS-purified cell types, we reprocessed published ATAC–seq data from sorted pancreatic[88] and unstimulated immune cells[89]. We created pseudobulk profiles from the snATAC–seq data for each donor and cluster, retaining those that contained information from >50 cells. We then extracted read counts in the 448,142 cCREs for all sorted and pseudobulk profiles. We used PCA to extract the top 20 principal components and used UMAP for dimensionality reduction and visualization ('min_dist = 0.5, n_neighbours = 80').

Defining cell-type-specific cCREs

To identify cCREs with accessibility levels most specific to each cluster, we used logistic regression models for each cCRE treating each cell as an individual data point. We performed separate regressions for each cluster, with binary cluster assignment and the covariates donor-of-origin and the log usable count as predictors and binary accessibility of the peak as the outcome, to calculate chromatin accessibility (CA) *t*-statistics. For a given cluster, we defined cCREs with activity most specific to that cluster by taking the top 1,000 cCREs with the highest CA *t*-statistics, after first filtering out cCREs that also had high CA *t*-statistics for other clusters (cCRE cell type CA *t*-statistics >90th percentile in >2 other cell types). The cCREs were all significant after Bonferroni correction for the number of peaks ($P < 1.1 \times 10^{-7}$) except for pancreatic $CD8^+$ T cells ($n = 428$ after correction), regulatory T cells ($n = 347$) and memory $CD8^+$ T cells ($n = 175$). We then used GREAT[90] (version 3) to annotate gene ontology terms that were enriched in each set of cell-type-specific cCREs compared to a background of all cCREs.

To assess whether cell-type-specific cCREs tended to be close in proximity to genes with cell-type-specific expression, we defined 100-kb windows around the midpoint of each cell-type-specific cCRE and annotated genes with overlapping TSSs. For each cell type that had a corresponding cluster in scRNA-seq, we compared whether genes around cell-type-specific cCREs for that cell type had higher gene expression specificity scores than the rest of the cell type-specific cCREs using two-sided Welch's *t*-tests. We collapsed cell-type-specific cCREs for cell types with more than one state in snATAC–seq but only one state in scRNA-seq.

Comparing single-cell chromatin accessibility and gene expression clusters

To compare cell types from snATAC–seq and scRNA-seq, we first derived gene expression *t*-statistics for each gene using linear regression models separately for each cluster of log-transformed read count as a function of binary cluster assignment, donor-of-origin, and log sequencing depth, treating cells as individual data points. For each gene, we also used chromatin accessibility *t*-statistics for promoter cCREs (see 'Defining cell-type-specific cCREs'). For each scRNA-seq cluster, we extracted the top 100 most specific genes based on the gene expression *t*-statistic. Using a merged list of the most specific genes across all clusters, we compared gene expression and promoter accessibility *t*-statistics using Pearson correlation.

Single-cell motif enrichment

We estimated TF motif enrichment *z*-scores for each cell using chromVAR[91] (version 1.5.0) by following the steps outlined in the user manual. First, we constructed a sparse binary matrix encoding read overlap with merged peaks for each cell. For each merged peak, we estimated the GC content bias to obtain a set of matched background peaks. To ensure a motif enrichment value for each cell, we did not apply any additional filters based on total reads or the fraction of reads in peaks. Next, using 580 TF motifs within the JASPAR 2018 CORE vertebrate (non-redundant) set[92], we computed GC bias-corrected enrichment *z*-scores (chromVAR deviation scores) for each cell. For each cell type, we considered a TF motif enriched if the average *z*-score across cells was greater than 2. We used the TFClass database[93] (http://tfclass.bioinf.med.uni-goettingen.de/) to group enriched TF motifs into structural sub-families. We determined the expression of all TFs within the sub-family in each cell type identified in scRNA-seq and considered TFs expressed in a cell type with TPM > 1.

Single-cell co-accessibility

We used Cicero[94] (version 1.3.3) to calculate co-accessibility scores between pairs of peaks for each cluster. As in the single-cell motif enrichment analysis, we started from a sparse binary matrix. For each cluster, we only retained merged peaks that overlapped peaks from the cluster. Within each cluster, we aggregated cells based on the 50 nearest neighbours and used Cicero to calculate co-accessibility scores, using a 1-Mb window size and a distance constraint of 500 kb. We then defined promoters as ±500 bp from the TSS of protein-coding transcripts from GENCODE v19[95] to annotate co-accessibility links between gene promoters and distal cCREs (non-promoter cCREs).

To assess whether genes with co-accessible links between the promoter and distal cCREs (co-accessible genes; co-accessibility score >0.05) were expressed more often than non-co-accessible genes (co-accessibility score <0) within each cell type, we separated co-accessible links into bins based on the distance between the gene promoter and distal cCRE. Within each bin, we then compared the fraction of genes expressed in the cell type (TPM > 1 from scRNA-seq) between co-accessible and non-co-accessible genes using two-sided Fisher's exact tests. We collapsed co-accessible links for cell types with more than one state in snATAC–seq but only one state in scRNA-seq (alpha, beta, and delta cells). No comparison was made for pancreatic $CD8^+$ T cells, which did not have a corresponding cluster in scRNA-seq.

GWAS enrichment analyses

We used stratified LD score regression[33,96,97] (version 1.0.1) to calculate genome-wide enrichment *z*-scores for 32 diseases and traits including T1D. We obtained GWAS summary statistics for autoimmune and inflammatory diseases (immune-related)[36–44], other diseases[45–53], and quantitative endophenotypes[56–65], and where necessary, we filled in variant IDs and alleles. Using 'munge_sumstats.py', we converted summary statistics to the LD score regression standard format. For each cluster, we considered overlap with chromatin accessibility peaks as a binary annotation for variants. Then, we computed annotation-specific LD scores by following the instructions for creating partitioned LD scores. We estimated enrichment coefficient *z*-scores for each annotation relative to the background annotations in the baseline-LD model (version 2.2). Using the enrichment *z*-scores, we computed two-sided *P* values to assess significance and corrected for multiple tests using the Benjamini–Hochberg procedure. We also calculated GWAS enrichment *z*-scores for T1D using a version of the meta-analysis excluding the Immunochip cohorts and observed highly similar enrichment results.

From the full GWAS summary statistics, we first extracted variants with MAF > 0.05 and calculated approximate Bayes factors[98] for each variant, assuming prior variance in allelic effects = 0.04. We then used fgwas[99] (version 0.3.6) to estimate T1D enrichment for common variants (MAF > 0.05) within cell-type-specific cCREs using an average window size of 1 Mb and also including annotations for coding exons, 3′/5′UTR regions and 1 kb upstream of the TSS from GENCODE in each model. We considered cell-type annotations enriched where ln(95% CI lower bound) > 0 and depleted where ln(95% CI upper bound) < 0.

Annotating cell-type mechanisms of variants at fine mapped signals

We compared the proportion of credible set variants with PPA > 0.50 that overlapped a cCRE compared to other credible set variants using a two-sided Fisher's exact test. Among credible set variants in cCREs, we further compared the proportion of credible set variants with PPA > 0.50 in a cCRE that was co-accessible with a gene promoter compared to other credible set variants using a two-sided Fisher's exact test.

For each T1D signal, we calculated the cumulative posterior probability of all credible set variants that overlapped cCREs that were active in T cells, monocytes, plasmacytoid dendritic cells, beta cells, acinar cells and ductal cells. For each signal that overlapped cCREs, we annotated genes within 1 Mb of the index variant that were (i) expressed in the same cell type(s) (TPM >1 from scRNA-seq) and (ii) co-accessible with a cCRE harbouring a credible set variant with PPA > 0.01.

Luciferase reporter assays

We tested for allelic differences in enhancer activity at rs7795896 using multiple constructs. First, we cloned a 180-bp sequence of human DNA (Coriell) containing the reference or alternate allele into the luciferase reporter vector pGL4.23 (Promega) in the forward direction using the restriction enzymes SacI and KpnI. Second, we cloned a larger 594-bp sequence of human DNA (Coriell) containing the rs7795896 reference allele that corresponded to the coordinates of the ductal-specific cCRE into pGL4.23 in the forward direction using the restriction enzymes SacI and KpnI. We introduced the alternate allele via SDM using the NEB Q5 Site Directed Mutagenesis kit (New England Biolabs) on 1 ng plasmid containing the reference allele and primers designed using the NEBaseChanger v.1.2.8 software. Sequence identity for all plasmids was confirmed with Sanger sequencing using the RV3 primer. Cloning primers were designed using Primer3 version 0.4.0. Primer sequences for cloning and SDM are listed in Supplementary Table 11.

We obtained Capan-1 cells from ATCC, and cells were authenticated by ATCC using karyotyping, morphology and PCR-based approaches. Cells tested negative for mycoplasma contamination. We grew Capan-1 cells, a model for ductal cells[100], to approximately 70% confluency according to ATCC culture recommendations in 6-well or 24-well plates and fed complete growth medium the day before transfection. For the 180-bp construct, 2,500 ng experimental or empty (pGL4.23) vector was co-transfected with 50 ng pRL-SV40 per sample using Lipofectamine 3000 (Invitrogen) into Capan-1 cells grown in a 6-well plate. For the 594-bp construct, 500 ng experimental or empty vector was co-transfected with 10 ng pRL-TK per sample using Lipofectamine 3000 (Invitrogen) into Capan-1 cells grown in a 24-well plate. The experiment was also repeated using 50 ng pRL-TK per sample. For all experiments, samples were assayed 48 h after transfection using the Dual-Glo Luciferase Assay System (Promega). We normalized Firefly:Renilla ratios with respect to the empty vector and used either two-sided, two-way ANOVA or two-sided Student's *t*-test to compare luciferase activity between the two alleles.

Electrophoretic mobility shift assay

We ordered double-stranded 5′ biotinylated and corresponding unlabelled (cold) oligonucleotides of 16 bp centred on rs7795896 with the reference and alternate alleles from Integrated DNA Technologies. Oligo sequences are listed in Supplementary Table 11. We performed EMSA using the LightShift Chemiluminescent EMSA kit (Thermo Fisher) according to the manufacturer's instructions with the following adjustments: 100 fmol of biotinylated duplex probe per reaction, and 20 pmol

of the same-allele non-biotinylated duplex 'cold' probe in competition reactions (200 × molar excess of the biotin probe). We used the NE-PER Nuclear Protein and Cytoplasmic Extraction Reagents (Thermo Fisher) kit to extract nuclear protein from Capan-1 cells and used 2 µl nuclear extract per binding reaction, corresponding to approximately 5–15 µg nuclear protein per reaction. We quantified bound and free probe (unbound) band intensity using ImageJ (v.1.53) and calculated the ratio of bound to unbound intensity. We then averaged bound ratios for replicates of each allele and compared ratios between alleles.

CRISPR inactivation of enhancer element

We obtained HEK293T cells from ATCC, and cells were authenticated by ATCC using karyotyping, morphology and PCR-based approaches. Cells tested negative for mycoplasma contamination. We maintained HEK293T cells in DMEM containing 100 units/ml penicillin and 100 mg/ml streptomycin sulfate supplemented with 10% fetal bovine serum. To generate CRISPRi lentiviral expression vectors, we designed guide RNA sequences to target the enhancer containing rs7795896 or the *CFTR* promoter. These guide RNAs, as well as a non-targeting control guide RNA, were placed downstream of the human U6 promoter in the pLV hU6-sgRNA hUbC-dCas9-KRAB-T2a-Puro backbone (Addgene, plasmid #71236). Targeting guide RNAs were designed using Benchling and selected to maximize both on-target binding[101] and guide specificity[102]. The non-targeting control guide RNA was selected from a previously validated genome-wide library[103]. Guide RNA sequences and targeted regions are listed in Supplementary Table 11. Higher scores indicate greater on-target binding and specificity.

We generated high-titre lentiviral supernatants by co-transfection of the resulting plasmid and lentiviral packaging constructs into HEK293T cells. Specifically, we co-transfected CRISPRi vectors with the pCMV-R8.74 (Addgene, #22036) and pMD2.G (Addgene, #12259) expression plasmids into HEK293T cells using a 1 mg/ml PEI solution (Polysciences). We collected lentiviral supernatants at 48 and 72 h after transfection and concentrated lentiviruses by ultracentrifugation for 120 min at 19,500 rpm using a Beckman SW28 ultracentrifuge rotor at 4 °C. Lentiviral titres were subsequently determined using a qPCR Lentivirus Titer Kit (Abm Bio), and aliquots were stored at −80 °C.

We obtained Capan-1 pancreatic ductal adenocarcinoma cell lines from ATCC and cultured them using Iscove's modified Dulbecco's medium with 20% fetal bovine serum, 100 units/ml penicillin, and 100 mg/ml streptomycin sulfate. Twenty-four hours before transduction, we passaged cells into a 12-well plate at a density of 100,000 cells per well. The following day, we added fresh medium containing 8 µg/ml polybrene and concentrated CRISPRi lentivirus at an MOI of 40 to each well. For each condition (1 non-targeting guide RNA, 3 enhancer-targeting guide RNAs, and 1 promoter-targeting guide RNA) we transduced 3 wells for a total of 15 wells. We additionally included 3 wells of mock-transduced cells without lentivirus. We incubated the cells at 37 °C for 30 min and then spun them in a centrifuge for 1 h at 30 °C at 950*g*. Six hours later, we replaced viral medium with fresh base culture medium for cell recovery. After 48 h, we replaced medium daily with the addition of 1 µg/ml puromycin for an additional 72 h, at which point all mock-transduced cells were killed. We reduced the concentration of puromycin to 0.5 µg/ml and cultured cells with daily medium changes for an additional week before passaging each cell line into a 48-well plate at a density of approximately 100,000 cells per well. The following morning, we harvested cells from each condition and isolated RNA using the RNeasy Micro Kit (Qiagen) according to the manufacturer's instructions.

For qRT–PCR, we performed cDNA synthesis using the iScript cDNA Synthesis Kit (Bio-Rad) and 250 ng of isolated RNA per reaction. We performed qRT–PCR reactions in triplicate with 5 ng of template cDNA per reaction using a CFX96 Real-Time PCR Detection System and the iQ SYBR Green Supermix (Bio-Rad). We used PCR of the TATA binding protein (TBP) coding sequence as an internal control, quantified relative expression via double delta CT analysis, and compared relative expression using two-sided ANOVA (enhancer inactivation versus non-targeting control) or a two-sided Student's *t*-test (promoter inactivation versus non-targeting control). Genes with C_t values greater than 34 were considered as not expressed. We also evaluated changes in expression of the puromycin resistance gene and the dCAS9 gene as additional controls. For eukaryotic genes, each primer pair was designed to span an exon–exon junction. Primers used for qPCR are listed in Supplementary Table 11.

Colocalization and deconvolution of the pancreas *CFTR* eQTL

We obtained GTEx v7[14] eQTL summary statistics for pancreas tissue from 220 samples and used effect size (beta) and standard error estimates from the regression model for *CFTR* expression to calculate approximate Bayes factors[98] for each variant, assuming prior variance in allelic effects = 0.04. We considered all variants in a 500-kb window around the T1D index variant at *CFTR* (rs7795896) tested in both the GWAS and eQTL data sets and used coloc[104] (version 4.0.4) to calculate the probability that the variants driving T1D and eQTL signals were shared, using prior probabilities $PP_{T1D} = 1 \times 10^{-4}$, $PP_{eQTL} = 1 \times 10^{-4}$, and $PP_{shared} = 1 \times 10^{-5}$. We considered the T1D and *CFTR* eQTL signals to be colocalized on the basis that the probability that they were shared (PP_{shared}) > 0.9. We applied eCAVIAR[105] (version 2.2) using variants in a 500-kb window that were tested for both T1D and eQTL association using LD calculated from EUR samples in 1000 Genomes[30] and considered variants with CLPP > 0.01 to be candidate causal variants for a shared signal.

We used MuSiC[106] (version 0.1.1) to estimate the proportions of major pancreatic cell types (acinar, duct, stellate, alpha, beta, delta, gamma) in each GTEx v7 pancreas sample. As input, we used raw count matrices from scRNA-seq of pancreatic cell types with labels from the gene expression reference map and GTEx v7 pancreas samples. For each cell type, we used the proportion as an interaction term and constructed linear models of TMM-normalized *CFTR* expression as a function of the interaction between genotype dosage and cell-type proportion, accounting for covariates used by GTEx including sex, sequencing platform, 3 genotype PCs, and 28 inferred PCs from the expression data. From the original 30 inferred PCs, we excluded inferred PCs 2 and 3 because they were highly correlated with acinar cell proportion (Spearman's $\rho > 0.7$). No remaining PCs were highly correlated (Spearman's $\rho < 0.3$) with the proportions of other cell types.

Phenotype associations at T1D signals

We tested association of the T1D index variant at *CFTR* (rs7795896) for pancreatic and autoimmune disease phenotypes. For acute pancreatitis, chronic pancreatitis, and pancreatic cancer, we used inverse variance weighted meta-analysis to combine SAIGE analysis results from the UK Biobank[32] (PheCodes 577.1, 577.2, and 157) and FinnGen (K11_ACUTPANC, K11_CHRONPANC, C3_PANCREAS_EXALLC). As mutations that cause cystic fibrosis (CF), which are risk factors for pancreatitis and pancreatic cancer, map to this locus, we determined the impact of the most common CF mutation F508del/rs199826652 on the association results for rs7795896. For T1D, we tested for association of rs7795896 conditional on F508del/rs199826652 in all cohorts except for FinnGen and observed no evidence for a difference in T1D association. For pancreatitis and pancreatic cancer, we identified F508del/rs199826652 carriers in the UK Biobank and repeated the association analysis for these phenotypes in UK Biobank data after removing these individuals and observed no evidence of a change in the effect of rs7795896.

We identified T1D signals where the risk allele had at least nominal association ($P < 0.05$) with increased risk of acute pancreatitis, chronic pancreatitis, or pancreatic cancer. We then tested whether these T1D signals had a difference in cPPA in exocrine cell cCREs or T cell cCREs compared to other T1D signals using a two-sided Student's *t*-test.

Human participant ethics

Genotype data obtained from dbGAP, WTCCC, and the UK Biobank were used in accordance with approved research plans for these data as obtained from the respective data repositories. Tissue samples for pancreas and peripheral blood were obtained from external biorepositories nPOD and Hemacare, and all individuals gave consent for the use of tissue samples. All genotype data and tissue samples were de-identified before being obtained and all studies were approved by the Institutional Review Board of UCSD.

Reporting summary

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

Data availability

Full summary statistics for the T1D GWAS have been deposited into the NHGRI-EBI GWAS catalogue with accession number GCST90014023 and can be downloaded from http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90014001-GCST90015000/GCST90014023/. Sequencing data for snATAC–seq have been deposited into the NCBI Gene Expression Omnibus (GEO) with accession number GSE163160. Data obtained from the TFClass database are available at http://tfclass.bioinf.med.uni-goettingen.de/ and from the PanglaoDB database at https://panglaodb.se/. Source data are provided with this paper.

Code availability

Code used for processing snATAC–seq data sets and clustering cells is available at https://github.com/kjgaulton/pipelines/tree/master/T1D_snATAC_pipeline.

27. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. **81**, 559–575 (2007).
28. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet*. **48**, 1279–1283 (2016).
29. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet*. **47**, 1236–1241 (2015).
30. 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
31. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet*. **48**, 1284–1287 (2016).
32. Zhou, W. et al. Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet*. **50**, 1335–1341 (2018).
33. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet*. **47**, 291–295 (2015).
34. Evangelou, M. et al. A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genet. Epidemiol*. **38**, 661–670 (2014).
35. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
36. Ji, S.-G. et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat. Genet*. **49**, 269–273 (2017).
37. 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).
38. 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).
39. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
40. 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).
41. Dubois, P. C. A. et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet*. **42**, 295–302 (2010).
42. Jin, Y. et al. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat. Genet*. **48**, 1418–1424 (2016).
43. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet*. **47**, 1449–1456 (2015).
44. López-Isac, E. et al. GWAS for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. *Nat. Commun*. **10**, 4955 (2019).
45. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet*. **51**, 404–413 (2019).
46. 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).
47. 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).
48. Stahl, E. A. et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet*. **51**, 793–803 (2019).
49. 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).
50. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet*. **51**, 431–444 (2019).
51. Watson, H. J. et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat. Genet*. **51**, 1207–1214 (2019).
52. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
53. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet*. **51**, 957–972 (2019).
54. Nielsen, J. B. et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet*. **50**, 1234–1239 (2018).
55. Tachmazidou, I. et al. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nat. Genet*. **51**, 230–236 (2019).
56. 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).
57. Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).
58. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet*. **27**, 3641–3649 (2018).
59. Jiang, X. et al. Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun*. **9**, 260 (2018).
60. 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).
61. Day, F. R. et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet*. **47**, 1294–1303 (2015).
62. Day, F. R. et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet*. **49**, 834–841 (2017).
63. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet*. **50**, 912–919 (2018).
64. 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).
65. 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).
66. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet*. **42**, 441–447 (2010).
67. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
68. Cousminer, D. L. et al. Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Hum. Mol. Genet*. **22**, 2735–2747 (2013).
69. Taal, H. R. et al. Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat. Genet*. **44**, 532–538 (2012).
70. Teumer, A. et al. Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. *Nat. Commun*. **9**, 4455 (2018).
71. Jansen, P. R. et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet*. **51**, 394–403 (2019).
72. van der Valk, R. J. P. et al. A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Hum. Mol. Genet*. **24**, 1155–1168 (2015).
73. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet*. **45**, 1274–1283 (2013).
74. Felix, J. F. et al. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Hum. Mol. Genet*. **25**, 389–403 (2016).
75. Chiou, J. et al. Single cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. *Nat. Genet*. **53**, 455–466 (2021).
76. 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).
77. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
78. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
79. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
80. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. **19**, 15 (2018).
81. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
82. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep*. **9**, 5233 (2019).
83. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).

84. 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).
85. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol*. **9**, R137 (2008).
86. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep*. **9**, 9354 (2019).
87. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
88. Arda, H. E. et al. A chromatin basis for cell lineage and disease risk in the human pancreas. *Cell Syst*. **7**, 310–322 (2018).
89. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet*. **51**, 1494–1505 (2019).
90. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol*. **28**, 495–501 (2010).
91. 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).
92. 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**, D260–D266 (2018).
93. 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).
94. Pliner, H. A. et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
95. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. **22**, 1760–1774 (2012).
96. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet*. **47**, 1228–1235 (2015).
97. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
98. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol*. **33**, 79–86 (2009).
99. 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).
100. Namkung, W. et al. Ca^{2+} activates cystic fibrosis transmembrane conductance regulator- and Cl^--dependent HCO3− transport in pancreatic duct cells. *J. Biol. Chem*. **278**, 200–207 (2003).
101. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol*. **34**, 184–191 (2016).
102. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol*. **31**, 827–832 (2013).
103. Horlbeck, M. A. et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* **5**, e19760 (2016).
104. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. **10**, e1004383 (2014).
105. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet*. **99**, 1245–1260 (2016).
106. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun*. **10**, 380 (2019).

Acknowledgements This work was supported by NIH grants DK112155, DK120429 and DK122607 to K.J.G. and M.S., and T32 GM008666 to R.J.G. We thank S. Kuan for assistance with sequencing. Additional acknowledgements for each cohort are listed in the Supplementary Information.

Author contributions K.J.G. and J.C. designed the study and wrote the manuscript. J.C. performed genetic association and single-cell genomics analyses. R.J.G. performed molecular experiments on enhancer function. M.-L.O. and S. Huang performed molecular experiments on variant function. R.M. and E.B. contributed to analyses of single-cell gene expression. J.Y.H. and M.M. generated single-cell accessible chromatin data. P.B. and K.K. contributed to single-cell motif enrichment analysis. D.U.G. and S.P. supervised the generation of single-cell accessible chromatin data and contributed to data interpretation and analyses. M.S. supervised experiments related to enhancer function and contributed to data interpretation. S. Heller and A.K. contributed to the design and interpretation of enhancer experiments.

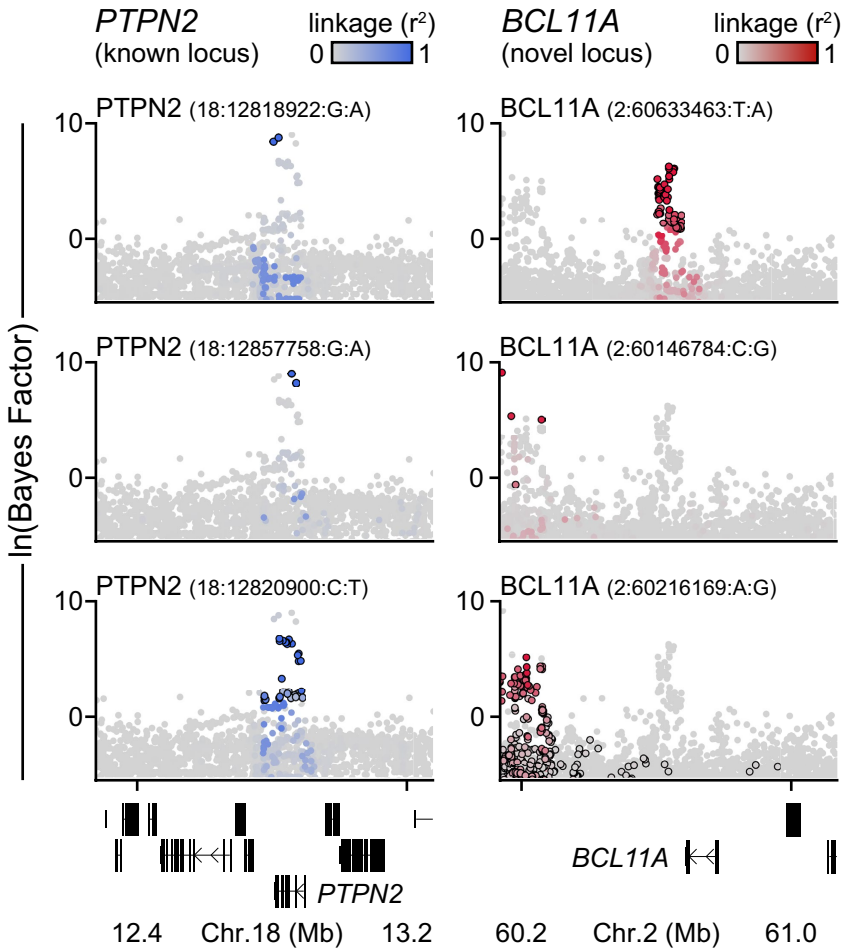
Competing interests K.J.G is a consultant for Genentech and holds stock in Vertex Pharmaceuticals; neither is related to the work in this study. The other authors declare no competing interests.

Additional information

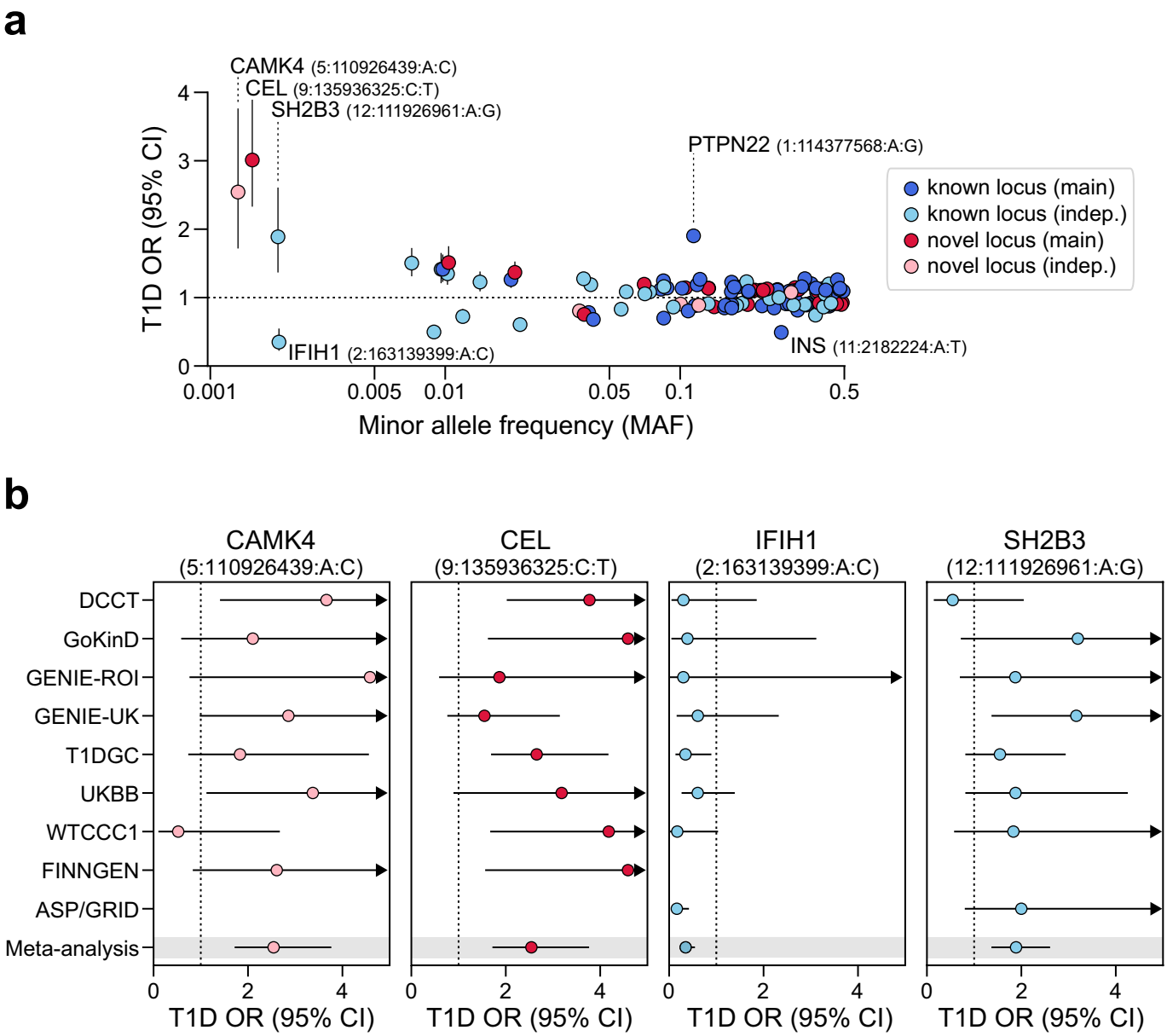
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03552-w.

Correspondence and requests for materials should be addressed to J.C. or K.J.G.

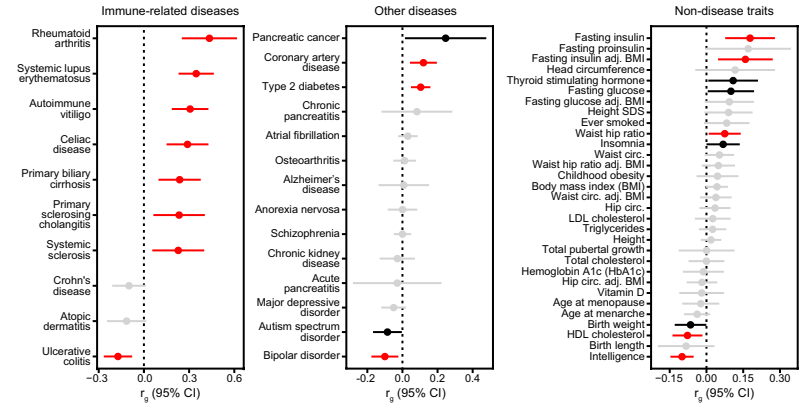
Peer review information *Nature* thanks Jason Torres, Anna Hutchinson, Chris Wallace and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



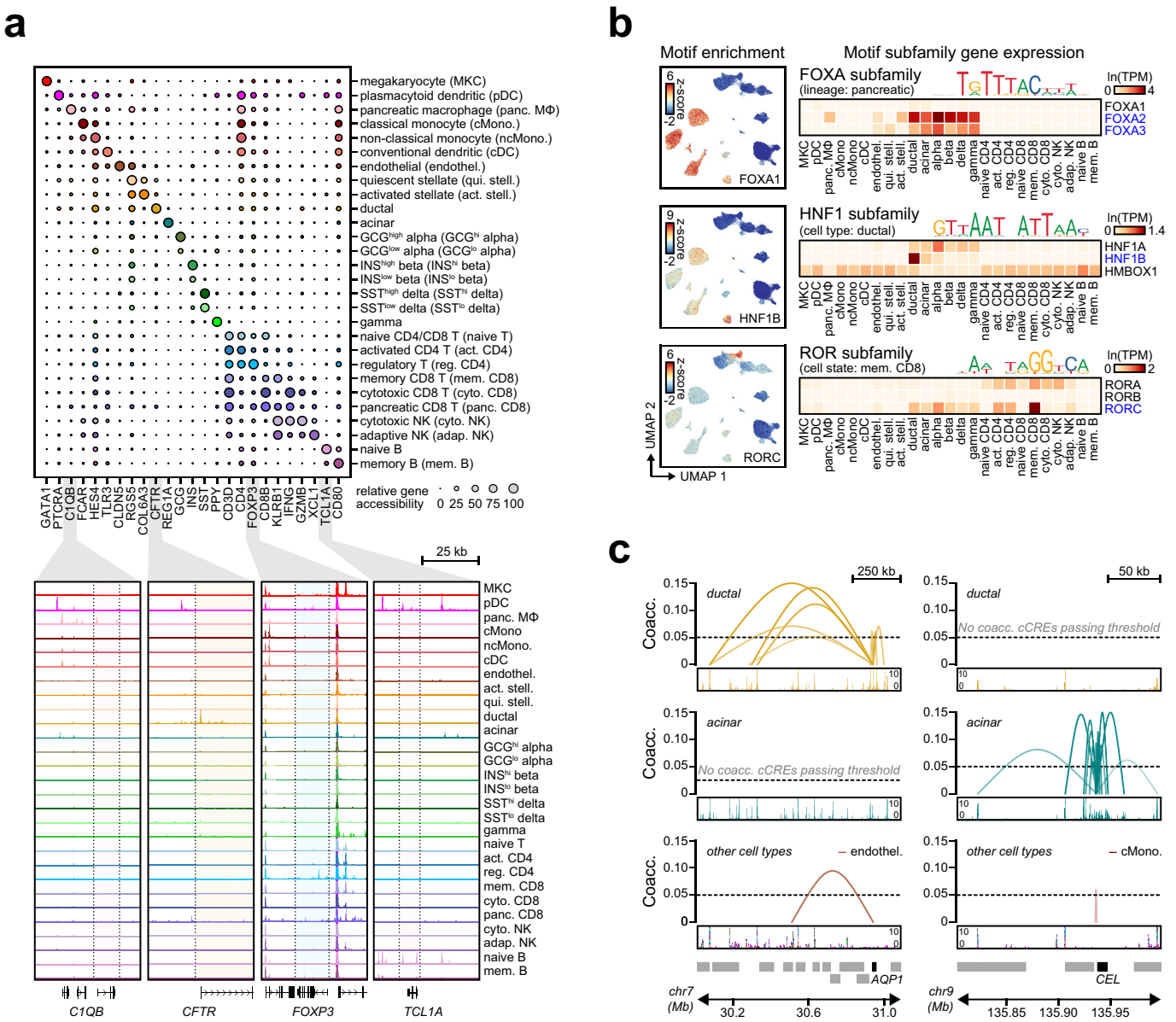
Extended Data Fig. 1 | Independent association signals at T1D risk loci. Bayes factors (natural log-transformed) for independent association signals at the known *PTPN2* locus (left) and the novel *BCL11A* locus (right). Variants are coloured by linkage disequilibrium (r^2) with the index variant for each signal.



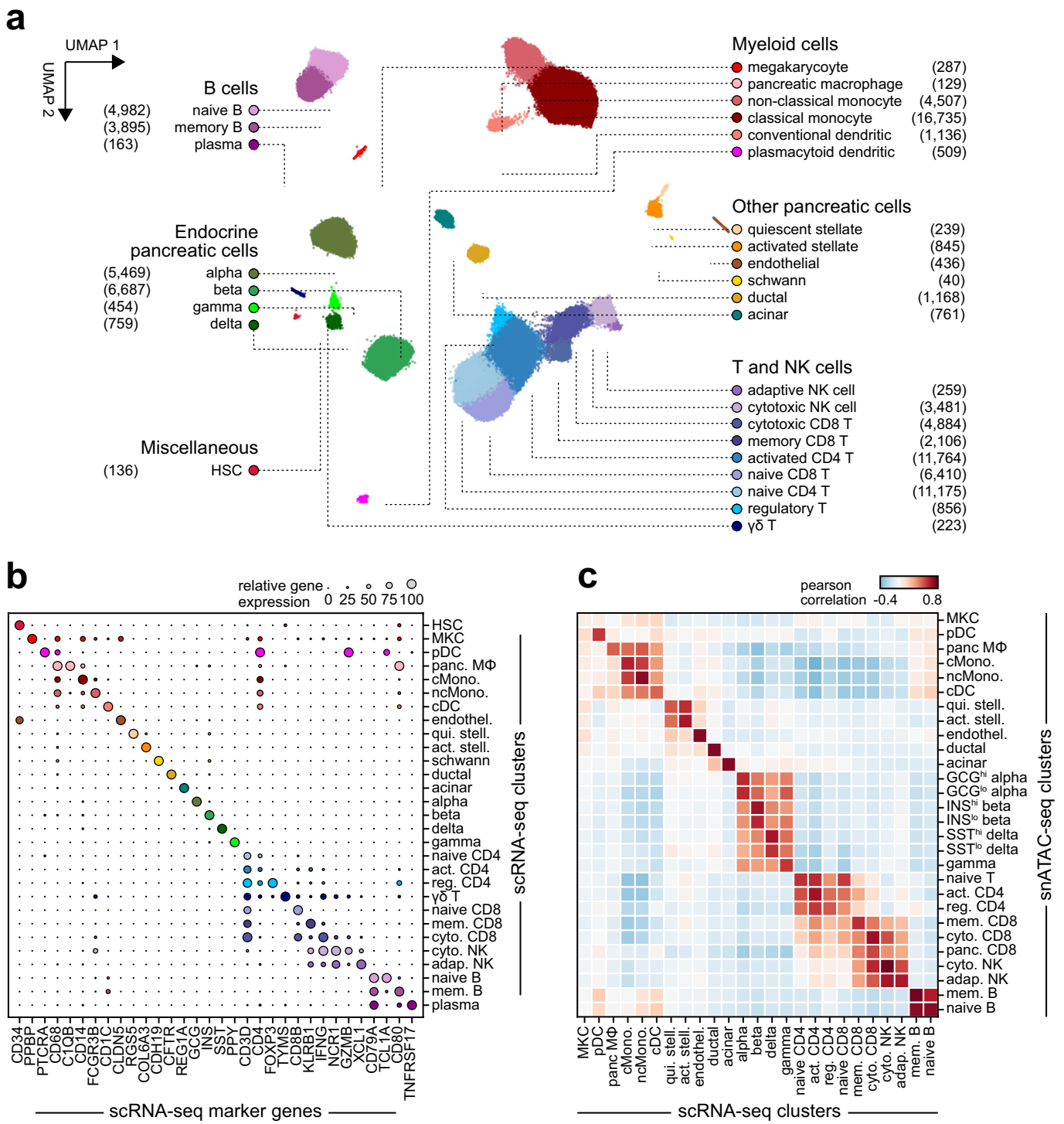
Extended Data Fig. 2 | Rare variants with large effects on T1D risk. a, The relationship between minor allele frequency and T1D odds ratios (OR) for index variants at 136 T1D signals. Signals with common index variants and larger effect size estimates (PTPN22 1:114,377,568 A:G and INS 11:2,182,224 A:T) or rare index variants (MAF < 0.005) are labelled. Points and lines represent estimates for OR and 95% CI. **b**, Comparison of OR across cohorts for rare variants. Missing values indicate that the variant was not tested in the cohort. Points and lines represent estimates for OR and 95% CI.



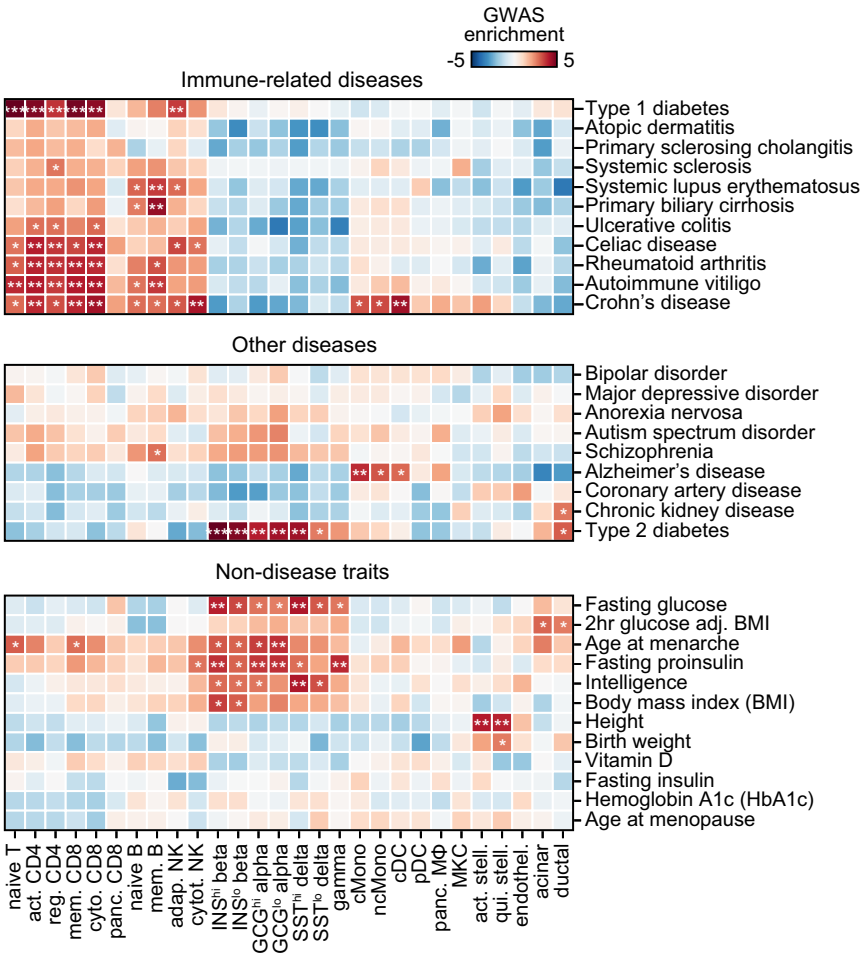
Extended Data Fig. 3 | Genetic correlations between T1D and other traits. Genetic correlations between T1D and immune-related diseases (left), other diseases (middle), and non-disease traits (right); adj., adjusted; circ., circumference. Two-sided *P* values are adjusted for multiple comparisons with false discovery rate (FDR). Colours indicate significance: red, correlation is significant after FDR correction (FDR < 0.1); black, correlation is nominally significant ($P < 0.05$) but not significant after FDR correction; grey, correlation is not significant. Points and lines represent genetic correlation estimates and 95% CI.



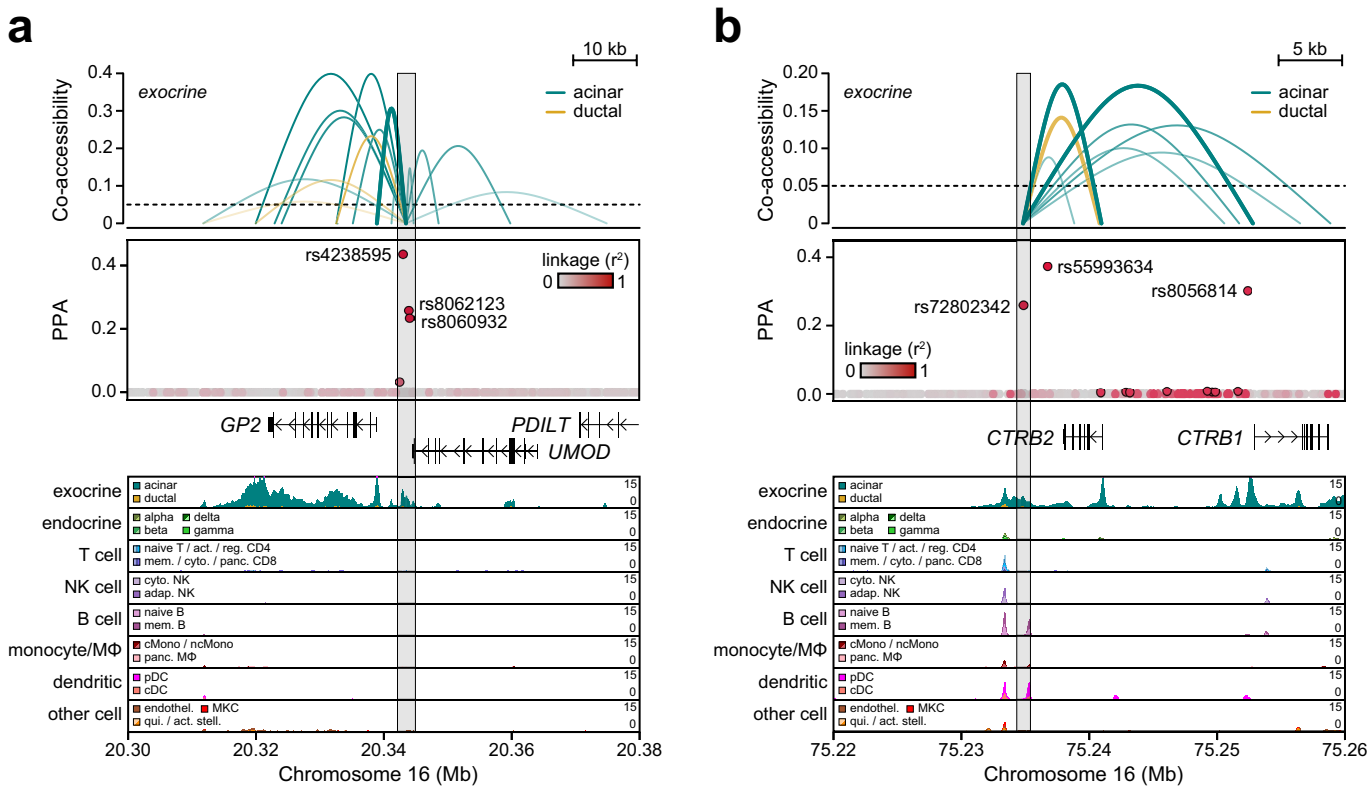
Extended Data Fig. 4 | Annotations derived from single-cell chromatin accessibility of T1D-relevant tissues. a, Relative gene accessibility (column-normalized chromatin accessibility reads in gene bodies) showing examples of marker genes used to identify cluster labels. Aggregated chromatin accessibility profiles in a 50-kb window around selected marker genes (bottom). **b**, Single-cell motif enrichment *z*-scores (left) and expression of motif subfamily members (right) for examples of TFs with lineage-, cell-type-, or cell-state-specific motif enrichment and expression. TFs with matching motif enrichment and expression are highlighted. **c**, Co-accessibility between *AQP1* and cCREs in ductal cells (left) and *CEL* and cCREs in acinar cells (right).



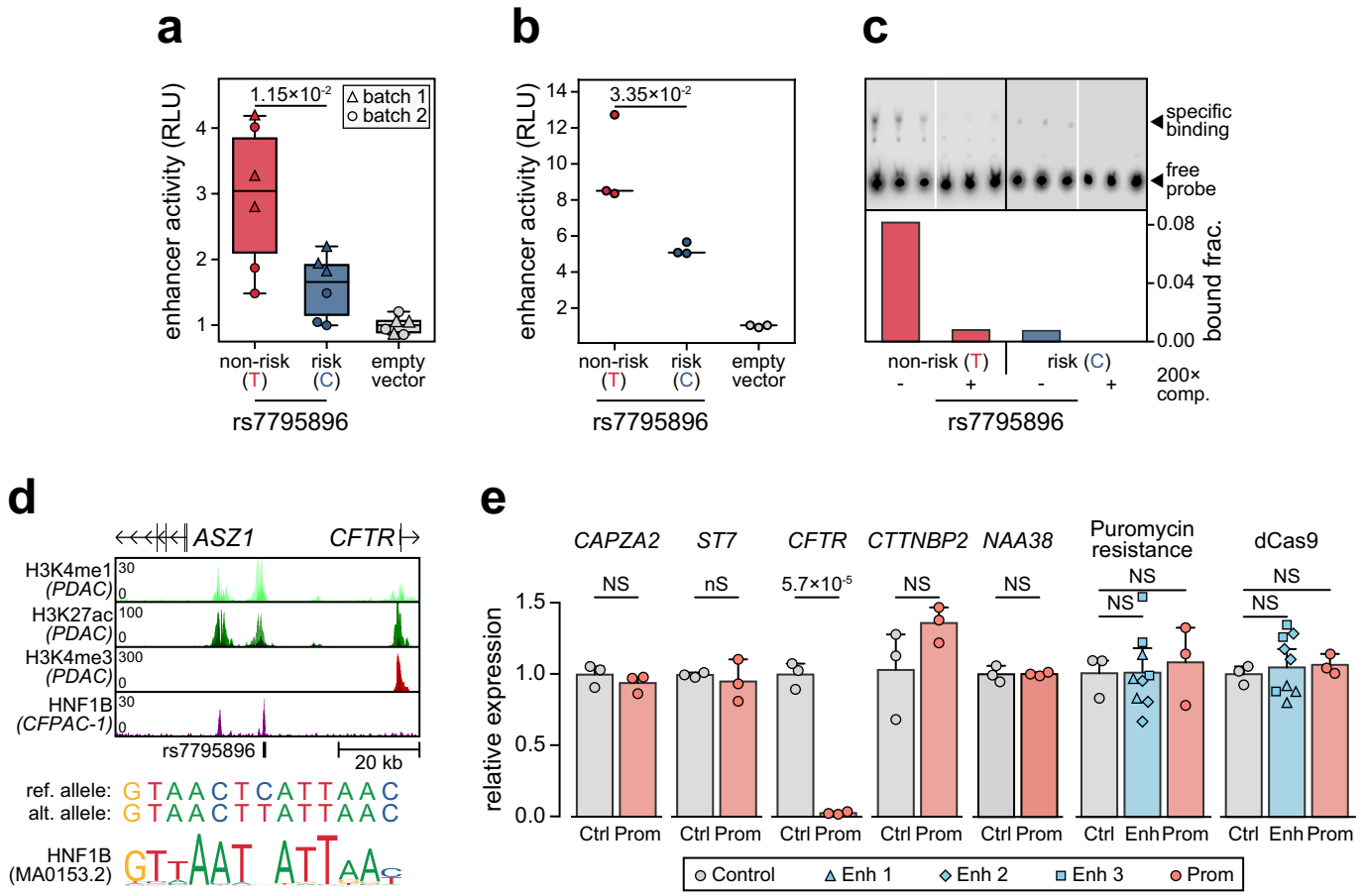
Extended Data Fig. 5 | Single-cell RNA-seq reference map of PBMCs and pancreatic islets. a, Clustering of 90,495 expression profiles from scRNA-seq experiments of peripheral blood mononuclear cells and pancreatic islets from published studies. Cells are plotted on the first two UMAP components and coloured according to cluster assignment. The number of cells in each cluster is shown next to its corresponding label. HSC, haematopoietic stem cell; γδ T, gamma delta T cell; pDC, plasmacytoid dendritic cell. **b**, Relative gene expression (average expression for all cells within a cluster and scaled from 0–100 across clusters) showing examples of marker genes used to assign cluster labels. **c**, Pearson correlation coefficient between gene expression and promoter accessibility specificity scores using a list containing the top 100 most specific genes for each scRNA-seq cluster found in snATAC–seq.



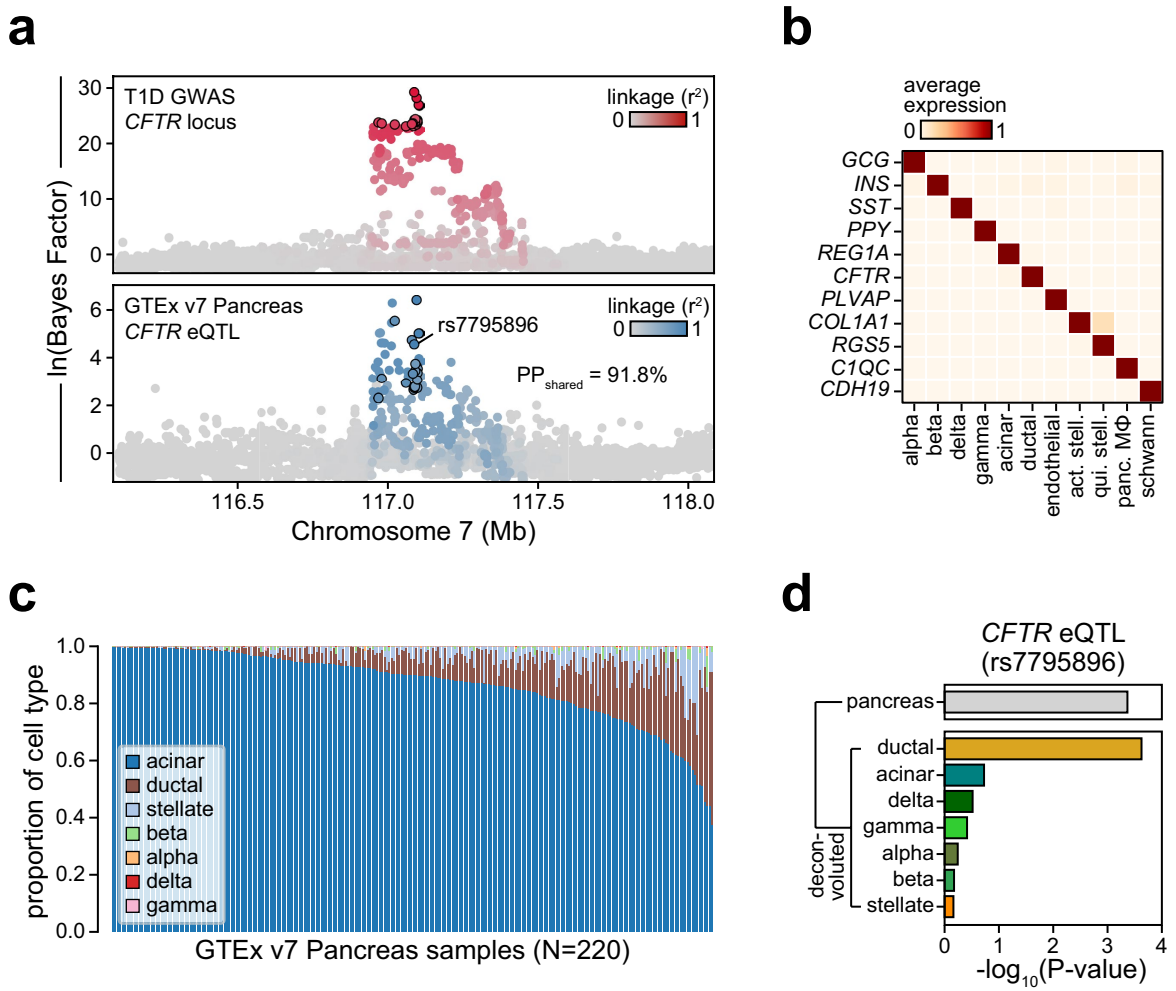
Extended Data Fig. 6 | GWAS enrichment for T1D compared to other diseases and traits. Stratified LD score regression coefficient *z*-scores for autoimmune and inflammatory diseases (top), other diseases (middle), and non-disease quantitative endophenotypes (bottom) for cCREs that are active in immune and pancreatic cell types. Two sided *P* values were calculated from *z*-scores and multiple test correction was performed using FDR. ***FDR < 0.001, **FDR < 0.01, *FDR < 0.1.



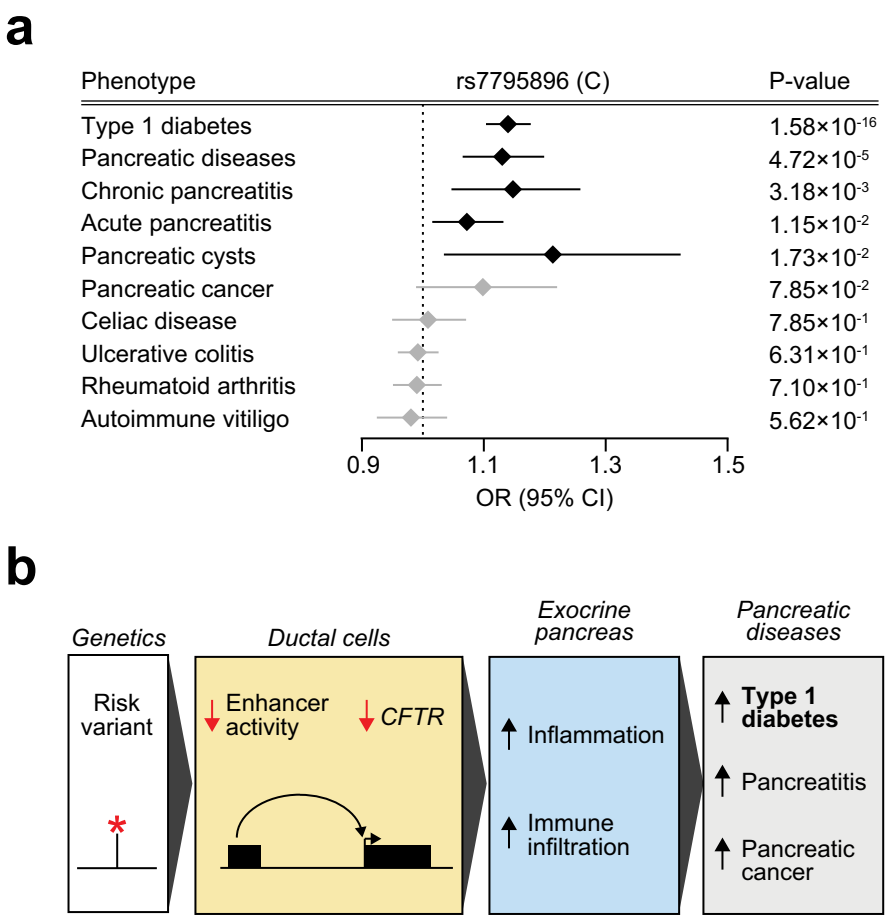
Extended Data Fig. 7 | Fine-mapped variants linked to exocrine-specific genes. a, The *GP2* locus contains three variants in a distal cCRE that is co-accessible with the GP2 promoter in acinar cells, which account for the majority of the causal probability (cPPA = 0.98). Chromatin accessibility at both the distal cCRE and the *GP2* promoter is highly specific to acinar cells. **b**, Variant rs72802342 at the *CTRB1/2/BCAR1* locus overlaps a distal cCRE that is co-accessible with the *CTRB2* and *CTRB1* promoters in acinar cells. Chromatin accessibility at the *CTRB1* and *CTRB2* promoters is highly specific to acinar cells. Variants contained in the 99% credible set are circled in black.



Extended Data Fig. 8 | rs7795896 has allelic effects on ductal enhancer activity. a, Relative luciferase units (RLU) for reporter containing 594-bp sequence surrounding rs7795896 in Capan-1 cells ($n = 6$; 2 batches × 3 transfections). Centre line, median; box limits, 25th and 75th percentiles; whiskers extend to 1.5 × the interquartile range from the 25th and 75th percentiles. *P* value by two-sided, two-way ANOVA. **b**, Luciferase reporter assay in Capan-1 cells transfected with pGL4.23 minimal promoter plasmids containing rs7795896 in the forward orientation. Relative luciferase units (RLU) represent Firefly:Renilla ratios normalized to control cells transfected with the empty pGL4.23 vector. *P* value by two-sided Student's *t*-test. **c**, Electrophoretic mobility shift assay with nuclear extract from Capan-1 cells using probes for rs7795896 alleles, with or without 200× unlabelled competitor probe (200× comp.). Quantification of the bound fraction (specific binding/free probe). Data are from $n = 1$ experiment. **d**, rs7795896 overlaps histone marks of active enhancers (H3K4me1, H3K27ac; region: chr7:117,050,000–117,125,000, hg19) but not promoters (H3K4me3) in pancreatic ductal adenocarcinoma (PDAC) cell lines (Capan-1, Capan-2, and CFPAC-1). rs7795896 overlaps a ChIP–seq peak for the transcription factor HNF1B in CFPAC-1 cells and a predicted HNF1B motif. **e**, Relative expression for genes in a 2-Mb window around rs7795896 with non-zero expression and the puromycin resistance and dCas9 genes. Ctrl, control, $n = 3$ biological replicates; Enh, enhancer, $n = 9$, 3 sgRNAs × 3 biological replicates; Prom, promoter, $n = 3$ biological replicates. Data are mean ± 95% CI. *P* values by two-sided Student's *t*-test (Prom versus Ctrl) or two-sided ANOVA (Enh versus Ctrl); NS, not significant.



Extended Data Fig. 9 | rs7795896 affects *CFTR* expression levels in ductal cells. a, Bayesian colocalization of T1D signal and *CFTR* pancreas eQTL. Variants in the T1D credible set are circled. **b**, Expression of pancreatic cell type marker genes from scRNA-seq. **c**, Proportions of selected pancreatic cell types estimated by MuSiC for 220 bulk pancreas RNA-seq samples from the GTEx v7 release using single-cell expression profiles. **d**, $-\log_{10}$-transformed two-sided uncorrected *P* values from linear regression interaction between dosage and cell-type proportion for the *CFTR* pancreas eQTL.



Extended Data Fig. 10 | Relationship between T1D and other pancreatic diseases. a, rs7795896 GWAS association for T1D (from full meta-analysis), pancreatic disease, and autoimmune disease. Points and lines represent OR estimates and 95% CI. Two-sided *P* values from GWAS meta-analysis are unadjusted for multiple comparisons. **b**, Variants that regulate genes with specialized exocrine pancreas function influence T1D risk, and we hypothesize that these effects are mediated through inflammation and immune infiltration.

nature research

Corresponding author(s): KJG, JC

Last updated by author(s): Mar 16, 2021

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 our Editorial Policies 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 was used for data collection.
Data analysis	GWAS analyses: HRC imputation preparation [version 4.2.9] (https://www.well.ox.ac.uk/~wrayner/tools/) EPACTS [version 3.2.6] (https://github.com/statgen/EPACTS/) SAIGE [version 0.38] (https://github.com/weizhouUMICH/SAIGE/) PLINK [version 1.90b6.7] (https://www.cog-genomics.org/plink/) FINEMAP [version 1.4] (http://www.christianbenner.com/) liftOver [no version] (https://genome.ucsc.edu/cgi-bin/hgLiftOver) LD score regression [version 1.0.1] (https://github.com/bulik/ldsc) fgwas [version 0.3.6] (https://github.com/joepickrell/fgwas) coloc [version 4.0.4] (https://github.com/chr1swallace/coloc) eCAVIAR [version 2.2] (http://genetics.cs.ucla.edu/caviar/) MuSiC [version 0.1.1] (https://github.com/xuranw/MuSiC) Single cell analyses: trim_galore [version 0.4.4] (https://github.com/FelixKrueger/TrimGalore) bwa [version 0.7.17-r1188] (https://github.com/lh3/bwa) samtools [version 1.10] (https://github.com/samtools/samtools) picard [no version] (https://broadinstitute.github.io/picard/) scanpy [version 1.4.4.post1] (https://github.com/theislab/scanpy) Harmony [version 1.0] (https://github.com/immunogenomics/harmony) PanglaoDB [no version] (https://panglaodb.se/)

MACS2 [version 2.1.2] (https://github.com/macs3-project/MACS)
bedtools [version 2.26.0] (https://github.com/arq5x/bedtools2)
GREAT [version 3] (http://great.stanford.edu/public/html/)
chromVAR [version 1.5.0] (https://github.com/GreenleafLab/chromVAR)
Cicero [version 1.3.3] (https://github.com/cole-trapnell-lab/cicero-release)
Cell Ranger ATAC [version 1.1.0] (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest)
Cell Ranger RNA [version 4.0.0] (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest)
Clustering pipeline (https://github.com/kjgaulton/pipelines/tree/master/T1D_snATAC_pipeline)
TFClass [no version] (http://tfclass.bioinf.med.uni-goettingen.de/)

Validation experiments:
NEBaseChanger [version 1.2.8] (https://nebasechanger.neb.com/)
Primer3 [version 0.4.0] (https://bioinfo.ut.ee/primer3/)
ImageJ [version 1.53] (https://imagej.nih.gov/ij/)
Benchling [no version] (https://www.benchling.com/)

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 and 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

Full summary statistics for the T1D GWAS have been deposited into the NHGRI-EBI GWAS catalog with accession number GCST90014023 and can be downloaded from http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90014001-GCST90015000/GCST90014023/. Sequencing data for snATAC-seq have been deposited into the NCBI Gene Expression Omnibus (GEO) with accession number GSE163160 and are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163160. Data obtained from the TFClass database are available at http://tfclass.bioinf.med.uni-goettingen.de/ and from the PanglaoDB database at https://panglaodb.se/.

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

Life sciences study design

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

Sample size	No statistical methods were used to pre-determine sample sizes. For the T1D GWAS, we compiled the largest sample size to date (N~520k) imputed using the most comprehensive reference panel (TOPMed r2) at the time. For the single nucleus ATAC-seq experiments, we compiled at least 4 biological replicates for each tissue sampled, including whole pancreas (n=4), purified pancreatic islets (n=4), and peripheral blood mononuclear cells (n=8). These sample sizes were determined based on similar studies of single cell epigenomics (PMID: 29434377, 29706549, 30858613, 33106633), and which enable the definition of reference maps of accessible chromatin. For datasets used in the single cell RNA-seq analyses, we compiled publicly available datasets for pancreatic islets (n=12) and peripheral blood mononuclear cells (n=5), which enable to definition of reference maps of gene expression. For CRISPRi experiments, sample size (n=3, 3 biological replicates × 1 sgRNA for non-targeting control; n=9, 3 biological replicates × 3 sgRNAs for enhancer; n=3, 3 biological replicates × 1 sgRNA for promoter) was determined based on feasibility and on sample size of similar studies (PMID: 26501517, 27708057). For luciferase reporter assays, we selected sample size (594 bp construct n=6, 3 biological replicates × 2 batches; 180 bp construct n=3, 3 biological replicates × 1 batch) based on the minimum for statistical comparison. For the electrophoretic mobility shift assay, we did not have biological replicates (n=1).
Data exclusions	For the T1D GWAS, we excluded samples that were low-quality (based on missing genotypes or sex mismatch with phenotype records), had cryptic relatedness with other samples, or had non-European ancestry through PCA with 1000 Genomes Project. For the UK biobank specifically, we excluded samples with withdrawn consent and duplicate samples genotyped for other cohorts with other genotyping arrays. For CRISPRi experiments, we excluded a technical replicate of one qPCR experiment due to a empty/null reading of a control gene from the machine. For single cell and reporter experiments, no samples or data were excluded. All exclusions were based on pre-specified criteria.
Replication	We used all available data for the T1D GWAS analysis, and no additional replication was performed. For single cell assays no additional replication was performed for this study. For CRISPRi, luciferase and EMSA assays the experiments were all re-performed for the revised manuscript and the results reproduced the findings reported in the original submission, and no additional replication was performed.
Randomization	Randomization is not relevant to the GWAS, as we used pre existing genotype data, or for the single cell assays as no group comparisons were

Randomization	performed. For the CRISPRi assays cells were randomly allocated into groups for transduction and the subsequent experiments were performed for samples across all groups together, for luciferase assays cells were randomly allocated into groups for transfection and depending on the construct the experiments were either performed all together or in multiple batches containing each group, and for the EMSA only one independent sample was used and the entire experiment was performed together.
Blinding	Blinding is not relevant for the GWAS as we used pre-existing genotype data or for the single cell assays where no group comparisons were performed. For CRISPRi, luciferase and EMSA assays blinding of the experiments themselves was not feasible due to personnel constraints, and for data collection the values were quantitative and did not require subjective interpretation.

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
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Capan-1 (ATCC® HTB-79™) were sourced from ATCC: https://www.atcc.org/Products/All/HTB-79.aspx HEK293T (ATCC® CRL-3216™) were sourced from ATCC: https://www.atcc.org/products/all/crl-3216.aspx
Authentication	Cell lines were authenticated by ATCC using karyotyping, morphology and PCR-based approaches to profile species specific variants and STRs to rule out intra- and inter-species contamination
Mycoplasma contamination	All cell lines tested negative for Mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None.