Open access • Posted Content • DOI:10.1101/2021.01.13.426472

Large-scale genetic association and single cell accessible chromatin mapping defines cell type-specific mechanisms of type 1 diabetes risk — Source link

Joshua Chiou, Ryan J Geusz, Mei-Lin Okino, Jee Yun Han ...+10 more authors

Institutions: University of California, San Diego, University of Ulm

Published on: 15 Jan 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Genome-wide association study, Epigenomics and Cell type

Related papers:

- Interpreting type 1 diabetes risk with genetics and single-cell epigenomics
- A comprehensive integrated post-GWAS analysis of Type 1 diabetes reveals enhancer-based immune dysregulation.
- Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes
- Single-Cell Epigenomics and Functional Fine-Mapping of Atherosclerosis GWAS Loci.
- From Genetic Association to Molecular Mechanisms for Islet-cell Dysfunction in Type 2 Diabetes.

Share this paper:

View more about this paper here: https://typeset.io/papers/large-scale-genetic-association-and-single-cell-accessible-4ags8eufnh

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 # Large-scale genetic association and single cell accessible chromatin
2 # mapping defines cell type-specific mechanisms of type 1 diabetes risk

3

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

7 1. Biomedical Sciences Graduate Program, University of California San Diego, La Jolla CA
8 92093
9 2. Department of Pediatrics, Pediatric Diabetes Research Center, University of California San
10 Diego, La Jolla CA 92093
11 3. Center for Epigenomics, Department of Cellular and Molecular Medicine, University of
12 California San Diego, La Jolla CA 92093
13 4. Department of Internal Medicine I, Ulm University, Ulm, Germany
14 5. Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla
15 CA 92093
16 6. Institute for Genomic Medicine, University of California San Diego, La Jolla CA 92093

17

18 # Corresponding authors:

19 Kyle J Gaulton
20 9500 Gilman Drive, #0746
21 Department of Pediatrics
22 University of California San Diego
23 858-822-3640
24 kgaulton@ucsd.edu

25

26 Joshua Chiou
27 9500 Gilman Drive, #0746
28 Biomedical Sciences Graduate Program
29 University of California San Diego
30 510-449-8870
31 joshchiou@ucsd.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

32 **ABSTRACT**

33 Translating genome-wide association studies (GWAS) of complex disease into mechanistic
34 insight requires a comprehensive understanding of risk variant effects on disease-relevant cell
35 types. To uncover cell type-specific mechanisms of type 1 diabetes (T1D) risk, we combined
36 genetic association mapping and single cell epigenomics. We performed the largest to-date
37 GWAS of T1D in 489,679 samples imputed into 59.2M variants, which identified 74 novel
38 association signals including several large-effect rare variants. Fine-mapping of 141 total signals
39 substantially improved resolution of causal variant credible sets, which primarily mapped to non-
40 coding sequence. To annotate cell type-specific regulatory mechanisms of T1D risk variants, we
41 mapped 448,142 candidate *cis*-regulatory elements (cCREs) in pancreas and peripheral blood
42 mononuclear cell types using snATAC-seq of 131,554 nuclei. T1D risk variants were enriched in
43 cCREs active in CD4+ T cells as well as several additional cell types including pancreatic exocrine
44 acinar and ductal cells. High-probability T1D risk variants at multiple signals mapped to exocrine-
45 specific cCREs including novel loci near *CEL, GP2* and *CFTR*. At the *CFTR* locus, the likely
46 causal variant rs7795896 mapped in a ductal-specific distal cCRE which regulated *CFTR* and the
47 risk allele reduced transcription factor binding, enhancer activity and *CFTR* expression in ductal
48 cells. These findings support a role for the exocrine pancreas in T1D pathogenesis and highlight
49 the power of combining large-scale GWAS and single cell epigenomics to provide insight into the
50 cellular origins of complex disease.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

51 **INTRODUCTION**

52 Type 1 diabetes (T1D) is a complex autoimmune disease characterized by the loss of insulin-
53 producing pancreatic beta cells and subsequent hyperglycemia[1], where the triggers of
54 autoimmunity and disease onset remain poorly understood. T1D has a strong genetic component,
55 most prominently at the major histocompatibility complex (MHC) locus but including 60 additional
56 risk loci identified in genome-wide and targeted array association studies[2–6]. T1D associated
57 variants at risk loci are largely non-coding, and intersection of T1D associated variants with
58 epigenomic data has identified an enrichment of risk variants within lymphoid enhancers[2].
59 However, due to limited sample sizes, incomplete variant coverage, and the limited cell type
60 resolution of existing epigenomic maps, the causal variants and cellular mechanisms of action of
61 T1D risk loci are largely unresolved.

62

63 **RESULTS**

64 **Comprehensive discovery and fine mapping of T1D risk signals**

65 To discover novel risk loci and improve fine mapping of causal variants for T1D, we performed a
66 genome-wide association study (GWAS) of 18,803 T1D cases and 470,876 controls of European
67 ancestry from 9 country-of-origin and array-matched cohorts (**Supplemental Table 1**). After
68 applying uniform quality-control measures (**Supplemental Figure 1**), where we removed low-
69 quality genotypes, individuals of non-European ancestry, or controls with other autoimmune
70 diseases, we imputed genotypes into the TOPMed r2 panel and tested for T1D association[7].
71 Through meta-analysis, we combined association results for 59,244,856 variants across cohorts
72 and observed 80 loci reaching genome-wide significance ($P<5\times10^{-8}$), including 30 loci previously
73 unreported in T1D risk (**Figure 1a**, **Supplemental Figure 2, Supplemental Table 2**). Previous
74 studies have identified independent association signals at multiple T1D loci[2], and we reasoned
75 that our increased sample size would uncover additional independent signals. Through iterative
76 conditional analyses, we discovered 52 secondary signals at locus-wide significance ($P<1\times10^{-5}$),
77 of which 44 were previously unknown (**Supplemental Figure 3, Supplemental Table 2**). Over
78 40% (36/89) of loci contained more than one independent signal; for example, the known *BACH2*
79 locus and novel *BCL11A* locus each had three signals (**Figure 1b**), and at the *IL2RA* locus we
80 identified six independent signals, three of which were novel (**Supplemental Figure 3**).

81 The TOPMed r2 panel enables more accurate imputation of rare variants over previous reference
82 panels, and in our study, we identified five novel T1D-associated variants with minor allele

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

83 frequency (MAF) less than 0.005 and large effects on disease risk (**Supplemental Table 2,**
84 **Supplemental Figure 4**). Among these rare variants, rs541856133 (MAF=.0015, OR=2.97)
85 mapped to a non-coding region directly upstream of *CEL,* which has been implicated previously
86 as the cause of maturity-onset diabetes of the young with pancreatic exocrine dysfunction
87 (MODY8)[8]. We also identified a novel protein-coding protective variant at *IFIH1* (p.Asn160Asp,
88 rs75671397, MAF=.002, OR=0.32), which was conditionally independent of the known protein-
89 coding variant signals in this gene. The three additional rare T1D risk variants mapped to non-
90 coding regions at the 16q23 (rs138099003, MAF=.0015, OR=2.29), *SH2B*3 (rs762349492,
91 MAF=.0018, OR=1.99), and *TOX* (rs192456638, MAF=.0045, OR=1.80) loci (**Supplemental**
92 **Table 2, Supplemental Figure 4**).

93 We next sought to fine map causal variants of T1D signals using a Bayesian approach[9]. In total
94 we considered 141 signals including 89 primary and 52 conditional signals at known and novel
95 loci excluding the MHC locus due to complex LD structure (**Figure 1c**). We defined linkage
96 disequilibrium (LD)-based credible sets for the 141 signals, using new index variants at known
97 loci where applicable. For each signal, we then used approximate Bayes factors[9] to calculate the
98 posterior probability of association (PPA) for each variant and defined credible sets of variants
99 that summed up to 99% cumulative PPA (**Supplemental Table 3**). Compared to previous
100 efforts[2,10], our fine-mapping resolution was drastically improved based on two complementary
101 measures: 1) fewer number of credible set variants per signal (median 24 variants) and 2) a
102 greater number of variants with high causal probabilities (**Figure 1d**). At nearly half of all T1D
103 signals (49%; 69/141) the credible set contained 20 or fewer variants, and 25% (35/141)
104 contained a single variant explaining the majority of the posterior probability (>50% PPA). Among
105 credible set variants, 23 variants with PPA>1% were nonsynonymous changes, including several
106 at novel loci p.Arg471Cys in *AIRE* (PPA=.99), p.Val11Ile in *BATF3* (PPA=.081), p.Ala91Val in
107 *PRF1* (PPA=0.038), and p.Val131Phe in *CD3G* (PPA=.028) (**Supplemental Table 4**).

108 Given our comprehensive genome-wide T1D genetic association and fine-mapping data, we used
109 these data to derive insight into disease pathophysiology. We therefore broadly characterized
110 relationships between T1D and other complex traits and diseases by performing genome-wide
111 genetic correlation analyses using LD score regression. As expected, T1D had significant
112 (FDR<.10) positive correlations with autoimmune diseases including rheumatoid arthritis (r_g=0.43,
113 FDR=7.34×10^{-5}), systemic lupus erythematosus (r_g=0.36, FDR=2.52×10^{-7}), celiac disease
114 (r_g=0.28, FDR=1.11×10^{-3}), and autoimmune vitiligo (r_g=0.30, FDR=2.02×10^{-5}), as well as a
115 negative correlation with ulcerative colitis (r_g=-0.17, FDR=2.94×10^{-3}) (**Supplemental Figure 5**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

116 Among other traits, we observed significant positive correlations with metabolic traits and
117 diseases such as fasting proinsulin (r_g=0.18, FDR=8.91×10^{-2}) and fasting insulin level, (r_g=0.18,
118 FDR=6.85×10^{-3}), coronary artery disease (r_g=0.12, FDR=6.85×10^{-3}) and type 2 diabetes (r_g=0.10,
119 FDR=4.39×10^{-3}), and positive correlations with pancreatic diseases such as pancreatic cancer
120 (r_g=0.25, FDR=7.40×10^{-2}) and chronic pancreatitis (r_g=0.13, FDR=3.84×10^{-1}), although the latter
121 estimate was not significant. These results demonstrate relationships between genetic effects on
122 T1D risk and a diversity of traits including autoimmune, pancreatic and metabolic disease.

123 **Defining cell type-specific *cis*-regulatory programs in T1D-relevant tissues**

124 The large majority of T1D risk signals map to non-coding regions and likely affect gene
125 regulation[2]. In order to annotate gene regulatory programs affected by T1D risk variants, we
126 generated a reference map of cell type-specific accessible chromatin using single nucleus ATAC-
127 seq (snATAC-seq) assays of T1D-relevant tissues including peripheral mononuclear blood cells
128 (PBMC), purified pancreatic islets, and whole pancreas tissue from non-diabetic donors
129 (**Supplemental Table 5**). To cluster cells obtained from these assays, we used a modified version
130 of our previous pipeline[11] that included rigorous quality control, removal of potential doublets, and
131 removal of potential confounding effects between different donors, tissues, and technologies to
132 group 131,554 chromatin accessibility profiles into 28 clusters (**Figure 2a, Supplemental Figure**
133 **6**). We assigned cell type identity to each cluster using the chromatin accessibility profiles of gene
134 bodies for known marker genes, and identified cells representing lymphoid, myeloid, endocrine,
135 exocrine, endothelial, and stellate cell types (**Figure 2a-b**). Within lymphoid and myeloid cells,
136 there were clusters representing both peripheral blood cells as well as tissue resident cells in the
137 pancreas based on both marker gene accessibility and tissue-of-origin profiles (**Figure 2a-b,**
138 **Supplemental Figure 6**). For example, we observed accessibility at *C1QB* marking pancreatic
139 tissue-resident macrophages, at *REG1A* marking pancreatic acinar cells, and at *CFTR* marking
140 pancreatic ductal cells (**Figure 2b**). We also observed distinct patterns of chromatin accessibility
141 at marker genes between different clusters of the same cell type allowing us to further discriminate
142 specific sub-types such as *FOXP3* for regulatory T cells relative to other T cells and *TCL1A* for
143 naïve B cells relative to memory B cells (**Figure 2b**).

144 To characterize the regulatory programs of each cell type and cell state, we aggregated reads
145 from cells within each cluster and called accessible chromatin sites representing candidate *cis*-
146 regulatory elements (cCREs). Across all 28 clusters, we identified a total of 448,142 cCREs and
147 an average of 77,812 cCREs per cluster (**Supplementary Data 1**). To further define regulatory
148 programs defining the identity of each cell type, we calculated the relative accessibility of each

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

149 cCRE across all clusters and identified 25,436 cell type-specific cCREs with accessibility patterns
150 specific to a given cluster **(Figure 2c, Supplementary Data 2).** To confirm that cell type-specific
151 cCREs regulated key processes involved in cellular identity, we identified gene ontology (GO)
152 terms enriched for each set of cell type-specific cCREs using GREAT[12]. GO terms significantly
153 enriched in cell type-specific cCREs represented highly specialized cellular processes, for
154 example inflammatory response for pancreatic tissue-resident macrophages ($P=6.09\times10^{-12}$),
155 extracellular matrix organization for activated stellate cells ($P=1.47\times10^{-41}$), transepithelial water
156 transport for ductal cells ($P=1.26\times10^{-21}$) and digestion for acinar cells ($P=1.18\times10^{-11}$) (**Figure 2c,**
157 **Supplementary Table 6**).

158 We next decoded the regulatory logic underlying cCRE activity for each cell type. First, we
159 identified candidate transcription factors (TFs) regulating cCRE activity by identifying sequence
160 motifs enriched in accessible chromatin of each cell type using chromVAR[13]. There were 290
161 motifs in JASPAR[14] with evidence for variable enrichment across cell types (**Supplementary**
162 **Table 7**). Enriched motifs included TF families with lineage-specific enrichment such as SPI in
163 myeloid and B cells, ETS in T cells, and FOXA in pancreatic endocrine and exocrine cells[15–17]
164 (**Figure 2d**). We also identified motifs enriched in specific cell types such as NR5A in acinar
165 cells[18], HNF1 in ductal cells[19], and EBF in B cells[20] (**Figure 2d**), as well as motifs for TF families
166 enriched in specific states within a cell type, such as POU2 in memory B cells[21], TCF7 in naïve
167 CD4+ T cells[22], and RUNX in adaptive NK cells[23] (**Figure 2d**). Second, we defined cell type-
168 resolved links between distal cCREs and putative target gene promoters using co-accessibility
169 across single cells with Cicero[24]. Considering all cell types, we observed a total of 1,028,428 links
170 between distal cCREs and gene promoters (**Supplemental Data 3**), where 145,138 distinct distal
171 cCREs were linked to at least one promoter. In many cases, co-accessible links were highly cell
172 type-specific; for example, multiple distal cCREs were co-accessible with the *AQP1* promoter in
173 ductal cells and the *CEL* promoter in acinar cells, none of which were identified in other cell types
174 (**Figure 2e**). Together these results identify candidate transcriptional regulators and target genes
175 of distal cCREs in pancreatic and immune cell types.

176 Annotating fine-mapped T1D risk variants with cell type-specific regulatory programs

177 We reasoned that our cell type-resolved regulatory maps would enable deeper insight into
178 pancreatic and blood cell types involved in T1D pathogenesis. We therefore determined
179 enrichment of variants associated with T1D as well as other complex diseases[25–42] and qualitative
180 endophenotypes[43–52] for cCREs using stratified LD score regression[53]. For T1D, the most
181 significant enrichment was for variants in CD4+ T cell cCREs (naïve CD4+ T $Z=4.54$,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

182 FDR=1.26×10^{-3}; activated CD4+ T Z=3.83, FDR=5.88×10^{-3}; regulatory T Z=3.26, FDR=1.35×10^{-2}) (**Figure 3a**). Notably, we did not observe evidence for enrichment in resident immune cells in
184 the pancreas (pancreatic CD8+ T cell Z=0.46, FDR=0.93; pancreatic tissue-resident macrophage
185 Z=-1.02, FDR=1.0). Outside of immune cell types, pancreatic ductal cell cCREs had the strongest
186 T1D enrichment, although this estimate was not significant (ductal Z=0.46, FDR=0.93). Other
187 immune-related diseases were also enriched within lymphocyte cCREs, although Crohn's
188 disease was also enriched for monocytes and conventional dendritic cell cCREs (**Figure 3a**). As
189 expected, type 2 diabetes and glycemic traits were strongly enriched in pancreatic endocrine cell
190 cCREs, but interestingly, glycemic traits such as glucose levels at 2 hours post-OGTT were also
191 enriched in pancreatic acinar and ductal cell cCREs (**Figure 3a**). Together these results
192 demonstrate that T1D associated variants are broadly enriched for CD4+ T cell cCREs, and
193 highlight other complex traits and diseases enriched for pancreatic and immune cell type cCREs.

194 Despite the strong enrichment of T1D-associated variants in CD4+ T cells, less than half of fine-
195 mapped T1D signals overlapped a CD4+ T cell cRE, suggesting that additional cell types
196 contribute to T1D risk. In order to identify additional disease-relevant cell types, we used an
197 orthogonal approach to test for enrichment of T1D variants within the subset of cCREs specific to
198 each cell type (from **Figure 2c; see Methods**). As expected, T1D variants genome-wide were
199 enriched in cCREs specific to CD4+ T cells (activated CD4+ T log enrich=4.14, 95% CI=0.97-
200 5.37) as well as pancreatic beta cells (log enrich=3.64, 95% CI=1.23-4.90) (**Figure 3b**).
201 Interestingly, T1D variants were also enriched in cCREs specific to plasmacytoid dendritic cells
202 (log enrich=4.08, 95% CI=2.09-5.16), classical monocytes (log enrich=4.04, 95% CI=2.74-4.92),
203 and pancreatic acinar and ductal cells (ductal log enrich=3.43, 95% CI=1.07-4.71, acinar log
204 enrich=2.74, 95% CI=0.66-4.02) (**Figure 3b**). We further enumerated the contribution of these
205 cell types to T1D risk by determining the cumulative posterior probability (cPPA) of fine-mapped
206 variants overlapping cell type-specific cCREs after removing variants overlapping a more
207 probable cell type (**see Methods**). Among broad annotation categories, distal cCREs harbored
208 the most cumulative risk (cPPA=24.3, N_{vars}=291), followed by coding exons (cPPA=7.98, N=34)
209 and promoters (cPPA=6.63, N=55) (**Figure 3c**). When breaking down distal cCREs by cell type
210 categories, CD4+ T cells had the most cumulative risk (cPPA=9.7, N=112), followed by exocrine
211 cells (acinar and ductal; cPPA=6.2, N=51), monocytes (cPPA=3.1, N=54), and then endocrine
212 cells (cPPA=2.3, N=33) (**Figure 3c**).

213 Given insight into cell types contributing to T1D risk, we next annotated individual T1D signals in
214 cCREs for these cell types. Over 75% (109/141) of T1D signals contained at least one fine-

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

215 mapped variant (with PPA>.01) overlapping a cCRE, and at 83% (90/109) of these signals the
216 cCRE was further co-accessible with at least one gene promoter (**Supplementary Table 8**). For
217 each T1D signal, we calculated the cPPA of fine-mapped variants overlapping cCREs for disease-
218 enriched cell types. At 58 T1D signals a fine-mapped variant overlapped a CD4+ T cell cCRE,
219 and signals with the highest cPPA in CD4+ T cells included the *CD2, IL2RA, PRF1* and *IKZF4*
220 loci (**Figure 3d**). We also identified T1D signals with high cPPA in pancreatic acinar and ductal
221 (exocrine) cCREs and monocyte cCREs, many of which were cell type-specific (**Figure 3d**). For
222 example, three variants at the *GP2* locus accounted for .951 of the PPA and mapped in an acinar-
223 specific cCRE co-accessible with the promoter of *GP2*, which encodes the major membrane
224 glycoprotein of pancreatic zymogen granules (**Figure 3e**). Similarly, rs72802342 at the *BCAR1*
225 locus (PPA=.30) mapped in an acinar-specific cCRE co-accessible with the *CTRB1* and *CTRB2*
226 promoters (**Figure 3f**). We observed similar predicted mechanisms in acinar cells at the *RNLS*
227 and *COBL* loci, as well as the novel *CEL* locus, where rs541856133 (PPA=.99) mapped in a
228 region of broad acinar-specific accessibility although not in a cCRE directly (**Supplementary**
229 **Figure 7a-c**). At *CTLA4*, variant rs3087243 (PPA=.99) mapped in an acinar-specific cCRE,
230 although the region around the variant was also broadly accessible in regulatory T cells, in line
231 with the specialized function of *CTLA4* in regulatory T cells[54] (**Supplementary Figure 7d**).
232 Exocrine cCREs harboring T1D risk variants at these loci were also largely specific relative to
233 previous studies of accessible chromatin from stimulated immune cells[55] and cytokine-stimulated
234 islets[56] except for *CTLA4* which mapped in a stimulated immune site (**Supplemental Table 8**).

235 **Risk variant at novel T1D locus has pancreatic ductal cell-specific effects on *CFTR***

236 As another example of an exocrine-specific T1D signal, at the *CFTR* locus fine-mapped variant
237 rs7795896 (PPA=0.60) mapped in a distal cCRE highly specific to pancreatic ductal cells
238 upstream of the *CFTR* gene (**Figure 4a**). Furthermore, the cCRE harboring rs7795896 had ductal
239 cell-specific co-accessibility with the *CFTR* promoter in addition to several other genes (**Figure**
240 **4a**). Recessive mutations in *CFTR* cause cystic fibrosis (CF) which is often comorbid with exocrine
241 pancreas insufficiency and CF-related diabetes (CFRD)[57]. Furthermore, carriers of *CFTR*
242 mutations often develop chronic pancreatitis[58]. As *CFTR* has not been previously implicated in
243 T1D, we sought to validate the mechanism of this locus. First, we determined whether rs7795896
244 had allele-specific activity using luciferase reporter and gel shift assays in Capan-1 cells, an
245 established model of ductal cell function[59]. We observed both significantly reduced enhancer
246 activity ($P=3.35\times10^{-2}$, **Figure 4b**) and reduced protein binding for the T1D risk allele (**Figure 4c**).
247 The variant mapped in a predicted sequence motif for the ductal cell-specific transcription factor

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

248 HNF1B (**Supplemental Table 6**) and overlapped a HNF1B ChIP-seq site previously identified in
249 ductal cell models (**Supplemental Figure 8**).

250 To determine whether the enhancer harboring rs7795896 regulated the expression of *CFTR* in
251 ductal cells, we used CRISPR interference (CRISPRi) to repress the activity of the enhancer
252 ($CFTR^{Enh}$) in Capan-1 cells using two independent guide RNAs. As positive and negative controls,
253 we inactivated the *CFTR* promoter ($CFTR^{Prom}$) and used a non-targeting guide RNA, respectively.
254 RNA-seq analysis revealed a significant reduction in *CFTR* expression after enhancer inactivation
255 ($CFTR^{Enh}$ $\log_2$(FC)=-0.40, P=2.41x10^{-3}), whereas expression of other genes co-accessible with
256 the enhancer was unchanged (**Figure 4d**), identifying *CFTR* as a target gene of this enhancer.
257 We next determined whether risk variants affected *CFTR* expression directly using pancreas
258 eQTL data from GTEx[60]. Out of 13 genes tested by GTEx for association with these variants, only
259 *CFTR* had evidence for an eQTL (P=4.31×10^{-4}), and this eQTL was statistically colocalized with
260 the T1D signal (PP_{shared}=91.4%) (**Figure 4e**). The T1D risk allele C was also associated with
261 decreased *CFTR* expression, consistent with effects on enhancer activity and TF binding. To
262 evaluate whether the *CFTR* eQTL signal in whole pancreas tissue was driven by ductal cells, we
263 used MuSiC[61] to estimate cell type proportions in each GTEx pancreas RNA-seq sample (**Figure**
264 **4f, Supplemental Figure 9**). We then re-calculated eQTL association including estimated cell
265 type proportion for each sample as an interaction term in the model, and only ductal cells had
266 significant association (P=2.37×10^{-4}) (**Figure 4g**).

267 As *CFTR* has been implicated in risk of pancreatic cancer[62] and pancreatitis[63], we finally asked
268 whether rs7795896 was significantly associated with these phenotypes in the UK biobank[64],
269 FinnGen, and other GWAS[28–31]. The T1D risk allele (C) was associated with increased risk of
270 pancreatitis (chronic pancreatitis OR=1.15, P=3.18×10^{-3}; acute pancreatitis OR=1.07, P=1.15×10^{-}
271 ^{2}), pancreatic cancer (OR=1.10, P=7.85×10^{-2}), and other pancreatic diseases which includes
272 pancreatitis and pancreatic cysts (OR=1.13, P=4.72×10^{-5}) (**Figure 4h**). In contrast, rs7795896 did
273 not show evidence for association with other autoimmune diseases (all P>.05), supporting that it
274 likely does not affect intrinsic immune cell function. Together our findings support a model in which
275 non-coding variants regulating the activity of genes such as *CFTR* in the exocrine pancreas
276 contribute to risk of T1D as well as pancreatic disease (**Figure 4i**).
277

278 ## DISCUSSION

279 Population-based association studies of complex disease are a powerful tool for genetic discovery
280 and, when coupled with cell type-resolved epigenome maps, can help reveal the cellular origins

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

281 of disease. Our results represent the largest genome-wide study of T1D genetics to date, more
282 than doubling the set of known risk signals, and provide a comprehensive resource for
283 interrogating T1D risk mechanisms. Integration of these data with cell type-specific accessible
284 chromatin maps both confirmed the prominent role of CD4+ T cells and implicated additional cell
285 types in disease risk notably pancreatic acinar and ductal cells. T1D risk variants mapped to
286 genes with specialized function in acinar and ductal cells such as *CFTR, GP2* and *CEL,* none of
287 which have been previously implicated in T1D. Observational studies have reported exocrine
288 pancreas abnormalities in T1D at disease onset[65] as well as in autoantibody positive individuals[66]
289 and first-degree relatives of T1D[67], but it was unknown whether this was contributing causally to
290 disease[68,69]. Studies in zebrafish, mice and humans have demonstrated that reduced *CFTR* leads
291 to CFRD via intra-islet inflammation and immune infiltration rather than intrinsic defects of beta
292 cell function, and immune infiltration in the exocrine pancreas has been suggested to contribute
293 to T1D pathogenesis[70–72]. We therefore hypothesize a causal role for gene regulation in exocrine
294 cells in T1D, potentially mediated through immune infiltration and inflammation, which may
295 provide novel avenues for therapeutic discovery in T1D.

296

297 **METHODS**

298 **Genotype quality control and imputation**

299 We compiled individual-level genotype data and summary statistics of 18,803 T1D cases and
300 470,876 controls of European ancestry from public sources (**Supplementary Table 1**), where
301 T1D case cohorts were matched to population control cohorts based on genotyping array
302 (Affymetrix, Illumina Infinium, Illumina Omni, and Immunochip) and country of origin where
303 possible (US, British, and Ireland). For the GENIE-UK cohort, because we were unable to find a
304 matched country of origin control cohort, we used individuals of British ancestry (defined by
305 individuals within 1.5 interquartile range of CEU/GBR subpopulations on the first 4 PCs from PCA
306 with European 1000 Genomes Project samples) from the University of Michigan Health and
307 Retirement study (HRS). For non-UK Biobank cohorts, we first applied individual and variant
308 exclusion lists (where available) to remove low quality, duplicate, or non-European ancestry
309 samples and failed genotype calls for each cohort. For control cohorts, we also used phenotype
310 files (where available) to remove individuals with type 2 diabetes or autoimmune diseases.

311 We then applied a uniform processing pipeline and used PLINK[73] to remove variants based on (i)
312 low frequency (MAF<1%), (ii) missing genotypes (missing>5%), (iii) violation of Hardy-Weinberg
313 equilibrium (HWE $p<1\times10^{-5}$ in control cohorts and HWE $p<1\times10^{-10}$ in case cohorts), (iv) substantial

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

314 differences in allele frequency compared to the Haplotype Reference Consortium r1.1 reference
315 panel[74], and (v) allele ambiguity (AT/GC variants with MAF>40%). We further removed individuals
316 based on (i) missing genotypes (missing>5%), (ii) sex mismatch with phenotype records
317 (het_{chrX}>.2 for females and het_{chrX}<.8 for males), (iii) cryptic relatedness through identity-by-
318 descent (IBD>.2), and (iv) non-European ancestry through PCA with 1000 Genomes Project[75] (>3
319 interquartile range from 25th and 75th percentiles of European 1KGP samples on the first 4 PCs)
320 (**Supplementary Figure 1**). For the affected sib-pair (ASP) cohort genotyped on the Immunochip,
321 we retained only one T1D sample from each family selected at random. For the GRID case and
322 1958 Birth control cohorts genotyped on the Immunochip, a portion of the cases overlapped the
323 T1DGC or 1958 Birth cohorts genotyped on a genome-wide array. We thus used sample IDs from
324 the phenotype files to remove these samples from the GRID and 1958 Birth cohorts and verified
325 that no samples were duplicated between the Immunochip and genome-wide array datasets by
326 checking IBD values. We combined data for matched case and control cohorts based on
327 genotyping array and country of origin for imputation. We used the TOPMed Imputation Server[76,77]
328 to impute genotypes into the TOPMed r2 panel[7] and removed variants based on low imputation
329 quality (R^2<.3). Following imputation, we implemented post-imputation filters to remove variants
330 based on potential genotyping or imputation artifacts based on empirical R^2 (genotyped variants
331 with empirical R^2<.5 and all imputed variants in at least low LD (r^2>.3) with them).

332 For the UK Biobank cohort, we downloaded imputed genotype data from the UK Biobank v3
333 release which were imputed using a combination of the HRC and UK10K + 1000 Genomes
334 reference panels. We used phenotype data to remove individuals of non-European descent. We
335 then used a combination of ICD10 codes to define 1,458 T1D cases (T1D diagnosis and insulin
336 treatment within a year of diagnosis, no T2D diagnosis). We defined controls as 362,257
337 individuals without diabetes (no T1D, T2D, or gestational diabetes diagnosis) or other
338 autoimmune diseases (systemic lupus erythematosus, rheumatoid arthritis, juvenile arthritis,
339 Sjögren syndrome, alopecia areata, multiple sclerosis, autoimmune thyroiditis, vitiligo, celiac
340 disease, primary biliary cirrhosis, psoriasis, or ulcerative colitis). We removed variants with low
341 imputation quality (R^2<.3).

342 For the FinnGen cohort, we downloaded GWAS summary statistics for type 1 diabetes
343 (E4_DM1_STRICT) from FinnGen freeze 2. This phenotype definition excluded individuals with
344 type 2 diabetes from both cases and controls.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

345 **Association testing, meta-analysis, and detection of conditional signals**

346 We tested low-frequency and common variants (MAF>.001%) for association to T1D with firth
347 bias reduced logistic regression using EPACTS (https://genome.sph.umich.edu/wiki/EPACTS) for
348 non-UK Biobank cohorts or SAIGE[64] for the UK Biobank, using genotype dosages adjusted for
349 sex and the first four ancestry PCs. We then combined association results across matched
350 cohorts through inverse-variance weighted meta-analysis. We used the liftOver utility to convert
351 GRCh38/hg38 into GRCh37/hg19 coordinates for all cohorts except for the UK biobank. We
352 removed variants that were unable to be converted, were duplicated after coordinate conversion,
353 or were located on different chromosomes after conversion. In total, our association data
354 contained summary statistics for 59,244,856 variants. To evaluate the extent to which genomic
355 inflation was driven by the polygenic nature of T1D or population stratification, we used LD score
356 regression to compare the LDSC intercept to lambda genomic control (GC). We observed an
357 intercept of 1.08 (SE=.03) compared to a lambda GC of 1.21, suggesting that the majority of the
358 observed inflation was driven by polygenicity rather than population stratification.

359 We used a threshold of $P<5\times10^{-8}$ to define genome-wide significance for primary signals, and we
360 defined novel loci as those statistically independent ($r^2<.01$) from reported index variants from
361 previous T1D association studies. For all cohorts except for FinnGen, we performed exact
362 conditional analyses on lead index variants to identify conditionally independent signals and used
363 a locus-wide threshold of $P<1\times10^{-5}$ to define significance. For genomic regions with multiple
364 known signals within close proximity, we conditioned on index variants from both signals. We
365 iterated through this process for each locus until there were no remaining significant signals at
366 the locus-wide threshold.

367 **Fine mapping of distinct association signals**

368 We constructed LD-based genetic credible sets of variants for 141 signals at 89 known and novel
369 loci excluding the MHC locus for complex LD structure and *ICOSLG*, for which we were unable
370 to find imputed proxy variants in our dataset. For the main signals at known loci, we defined
371 credible set variants by taking all variants in at least low LD ($r^2>.1$) with newly identified index
372 variants within a 5 Mb window. For both novel and conditional signals, we used the most
373 significant variant at the signal and the same credible set definition. We used effect size and
374 standard error estimates to calculate approximate Bayes factors[9] (ABF) for each variant; at
375 signals with multiple distinct association signals, we derived values from the corresponding
376 conditional analysis. We then calculated the posterior probability of association (PPA) for each
377 variant by dividing its ABF by the sum of ABF for all variants in the signal's credible set. To derive

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

378 99% credible sets for each signal, we sorted variants for each signal by descending PPA and
379 retained variants that added up to a cumulative PPA>0.99. To verify that variant coverage across
380 different imputation panels did not affect fine mapping, we calculated the effective sample size for
381 all credible set variants. There were only 9 credible set variants in total with <50% of the maximum
382 effective sample size, all of which had PPA<.01, and we did not further filter these variants.

383 **GWAS correlation analyses**

384 We used LD score regression (version 1.0.1) to estimate genome-wide genetic correlations
385 between T1D and immune diseases[25–31,41,42], other diseases[32–40,64,78,79], and non-disease traits[43–
386 50,80–88], using European subsets of GWAS where applicable. For acute pancreatitis, chronic
387 pancreatitis, and pancreatic cancer, we used inverse variance weighted meta-analysis to combine
388 SAIGE analysis results from the UK biobank[64] (PheCodes 577.1, 577.2, and 157) and FinnGen
389 (K11_ACUTPANC, K11_CHRONPANC, C3_PANCREAS_EXALLC). We used pre-computed
390 European 1000 Genomes LD scores to calculate correlation estimates (r_g) and standard errors.
391 We then corrected p-values for multiple tests using FDR correction, considering traits with FDR<.1
392 as significant. We also performed genetic correlation analyses using a version of the T1D meta-
393 analysis excluding the Immunochip cohorts and observed highly similar results.

394 **Generation of snATAC-seq libraries**

395 Combinatorial indexing single cell ATAC-seq (snATAC-seq/sci-ATAC-seq). snATAC-seq was
396 performed as described previously[89,90] with several modifications as described below. For the islet
397 samples, approximately 3,000 islet equivalents (IEQ, roughly 1,000 cells each) were resuspended
398 in 1 mL nuclei permeabilization buffer (10mM Tris-HCL (pH 7.5), 10mM NaCl, 3mM $MgCl_2$, 0.1%
399 Tween-20 (Sigma), 0.1% IGEPAL-CA630 (Sigma) and 0.01% Digitonin (Promega) in water) and
400 homogenized using 1mL glass dounce homogenizer with a tight-fitting pestle for 15 strokes.
401 Homogenized islets were incubated for 10 min at 4°C and filtered with 30 µm filter (CellTrics). For
402 the pancreas samples, frozen tissue was pulverized with a mortar and pestle while frozen and
403 immersed in liquid nitrogen. Approximately 22 mg of pulverized tissue was then transferred to an
404 Eppendorf tube and resuspended in 1 mL of cold permeabilization buffer for 10 minutes on a
405 rotator at 4°C. Permeabilized sample was filtered with a 30µm filter (CellTrics), and the filter was
406 washed with 300 µL of permeabilization buffer to increase nuclei recovery.

407 Once permeabilized and filtered, nuclei were pelleted with a swinging bucket centrifuge (500 x g,
408 5 min, 4°C; 5920R, Eppendorf) and resuspended in 500 µL high salt tagmentation buffer (36.3 mM
409 Tris-acetate (pH = 7.8), 72.6 mM potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and
410 counted using a hemocytometer. Concentration was adjusted to 4500 nuclei/9 µl, and 4,500 nuclei

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

411 were dispensed into each well of a 96-well plate. Glycerol was added to the leftover nuclei
412 suspension for a final concentration of 25 % and nuclei were stored at -80°C. For tagmentation,
413 1 µL barcoded Tn5 transposomes[90] were added using a BenchSmart™ 96 (Mettler Toledo),
414 mixed five times and incubated for 60 min at 37°C with shaking (500 rpm). To inhibit the Tn5
415 reaction, 10 µL of 40 mM EDTA were added to each well with a BenchSmart™ 96 (Mettler Toledo)
416 and the plate was incubated at 37°C for 15 min with shaking (500 rpm). Next, 20 µL 2 x sort buffer
417 (2 % BSA, 2 mM EDTA in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All wells
418 were combined into a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using a SH800
419 (Sony), 20 nuclei were sorted per well into eight 96-well plates (total of 768 wells) containing
420 10.5 µL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma))[90]. Preparation of sort
421 plates and all downstream pipetting steps were performed on a Biomek i7 Automated Workstation
422 (Beckman Coulter). After addition of 1 µL 0.2% SDS, samples were incubated at 55 °C for 7 min
423 with shaking (500 rpm). We added 1 µL 12.5% Triton-X to each well to quench the SDS and
424 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix (NEB). Samples were PCR-amplified (72 °C
425 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
426 wells were combined. Libraries were purified according to the MinElute PCR Purification Kit
427 manual (Qiagen) using a vacuum manifold (QIAvac 24 plus, Qiagen) and size selection was
428 performed with SPRI Beads (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one
429 more time with SPRI Beads (Beckmann Coulter, 1.5x). Libraries were quantified using a Qubit
430 fluorimeter (Life technologies) and the nucleosomal pattern was verified using a TapeStation
431 (High Sensitivity D1000, Agilent). The library was sequenced on a HiSeq2500 sequencer
432 (Illumina) using custom sequencing primers, 25% spike-in library and following read lengths: 50
433 + 43 + 40 + 50 (Read1 + Index1 + Index2 + Read2).

434 Droplet-based 10X single cell ATAC-seq (scATAC-seq). 10X scATAC-seq protocol from 10x
435 Genomics was followed: Chromium SingleCell ATAC ReagentKits UserGuide (CG000209, Rev
436 A). Cryopreserved PBMC samples were thawed in 37°C water bath for 2 min and followed 'PBMC
437 thawing protocol' in the UserGuide. After thawing cells, the pellets were resuspended again in 1
438 mL chilled PBS (with 0.04% PBS) and filtered with 50 µm CellTrics (04-0042-2317, Sysmex). The
439 cells were centrifuged (300g, 5 min, 4°C) and permeabilized with 100 µl of chilled lysis buffer
440 (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween-20, 0.1% IGEPAL-CA630,
441 0.01% digitonin and 1% BSA). The samples were incubated on ice for 3 min and resuspended
442 with 1mL chilled wash buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween-
443 20 and 1% BSA). After centrifugation (500g, 5 min, 4°C), the pellets were resuspended in 100 µL
444 of chilled Nuclei buffer (2000153, 10x Genomics). The nuclei concentration was adjusted between

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

445 3,000 to 7,000 per µl and 15,300 nuclei which targets 10,000 nuclei was used for the experiment.
446 For pancreas tissue (pulverized as described above), approximately 31.7 mg of pulverized tissue
447 was transferred to a LoBind tube (Eppendorf) and resuspended in 1 mL of cold permeabilization
448 buffer (10mM Tris-HCL (pH 7.5), 10mM NaCl, 3mM $MgCl_2$, 0.1% Tween-20 (Sigma), 0.1%
449 IGEPAL-CA630 (Sigma), 0.01% Digitonin (Promega) and 1% BSA (Proliant 7500804) in water)
450 for 10 min on a rotator at 4°C. Permeabilized nuclei were filtered with 30 µm filter (CellTrics).
451 Filtered nuclei were pelleted with a swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R,
452 Eppendorf) and resuspended in 1 mL Wash buffer (10mM Tris-HCL (pH 7.5), 10mM NaCl, 3mM
453 $MgCl_2$, 0.1% Tween-20, and 1% BSA (Proliant 7500804) in molecular biology-grade water). Nuclei
454 wash was repeated once. Next, washed nuclei were resuspended in 30 µL of 1X Nuclei Buffer
455 (10X Genomics). Nuclei were counted using a hemocytometer, and finally the nuclei
456 concentration was adjusted to 3,000 nuclei/µl. 15,360 nuclei were used as input for tagmentation.
457
458 Nuclei were diluted to 5 µl with 1X Nuclei buffer (10x Genomics) and, mixed with ATAC buffer
459 (10x Genomics) and ATAC enzyme (10x Genomics) for tagmentation (60 min, 37°C). Single cell
460 ATAC-seq libraries were generated using the (Chromium Chip E Single Cell ATAC kit (10x
461 Genomics, 1000086) and indexes (Chromium i7 Multiplex Kit N, Set A, 10x Genomics, 1000084)
462 following manufacturer instructions. Final libraries were quantified using a Qubit fluorimeter (Life
463 technologies) and the nucleosomal pattern was verified using a TapeStation (High Sensitivity
464 D1000, Agilent). Libraries were sequenced on a NextSeq 500 and HiSeq4000 sequencer
465 (Illumina) with following read lengths: 50 + 8 + 16 + 50 (Read1 + Index1 + Index2 + Read2).
466

467 **Single cell chromatin accessibility data processing**

468 Prior to read alignment, we used trim_galore (version 0.4.4) to remove adapter sequences from
469 reads using default parameters. We aligned reads to the hg19 reference genome using bwa
470 mem[91] (version 0.7.17; parameters: '-M -C') and removed low mapping quality (MAPQ<30),
471 secondary, unmapped, and mitochondrial reads using samtools[92]. To remove duplicate
472 sequences on a per-barcode level, we used the MarkDuplicates tool from picard (parameters:
473 'BARCODE_TAG'). For each tissue and snATAC-seq technology, we used log-transformed read
474 depth distributions from each experiment to determine a threshold separating real cell barcodes
475 from background noise. We used 500 total reads (passing all filters) as the cutoff for combinatorial
476 barcoding snATAC and between 2,300 and 4,000 total reads, as well as at least 0.3 fraction of
477 reads in peaks for 10x snATAC-seq experiments (**Supplemental Figure 5a**).
478

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

479 **Single cell chromatin accessibility clustering**

480 We identified snATAC-seq clusters using a previously described pipeline with a few modifications.
481 For each experiment, we first constructed a counts matrix consisting of read counts in 5 kb
482 windows for each cell. Using scanpy[93], we normalized cells to a uniform read depth and log-
483 transformed counts. We extracted highly variable (*hv*) windows (parameters: 'min_mean=.01,
484 min_disp=.25') and regressed out the total log-transformed read depth within *hv* windows (usable
485 counts). We then merged datasets from the same tissue and performed PCA to extract the top
486 50 PCs. We used Harmony[94] to correct the PCs for batch effects across experiments, using
487 categorical covariates such as donor-of-origin (all tissues), biological sex (PBMCs), and snATAC-
488 seq assay technology (pancreas). We used the corrected components to construct a 30 nearest
489 neighbor graph using the cosine metric, which we used for UMAP dimensionality reduction
490 (parameters: 'min_dist=.3') and clustering with the Leiden algorithm[95] (parameters:
491 'resolution=1.5').

492 Prior to combining cells across all tissues, we performed iterative clustering to identify and remove
493 cells with aberrant quality metrics. First, we identified and remove clusters of cells with lower
494 quality metrics (islets: 948, pancreas: 2,588, PBMCs: 5,268 cells removed total), including lower
495 usable counts or fraction of reads in peaks. Next, after removing the low-quality cells and
496 repeating the previous clustering steps, we sub-clustered the resulting main clusters at high
497 resolution (parameters: 'resolution=3.0') to identify sub-clusters containing potential doublets
498 (islets: 886, pancreas: 4,495, PBMCs: 5,844 cells removed total). We noted that these sub-
499 clusters tended to have higher average usable counts, promoter usage, and accessibility at more
500 than one marker gene promoter. After removing 20,029 low-quality or potential doublet cells, we
501 performed one final round of clustering using experiments from all tissues, including tissue-of-
502 origin as another covariate. We further removed 672 cells mapping to improbable cluster
503 assignments (islet or pancreatic cells in PBMC clusters or vice versa). After all filters, we ended
504 up with 131,554 cells mapping to 28 distinct clusters with consistent representation across
505 samples from the same tissue (**Supplemental Figure 5b**). We cataloged known marker genes
506 for each cell type and assessed gene accessibility (sum of read counts across each gene body)
507 to assign labels to each cluster.

508

509 **Single cell chromatin accessibility analyses**

510 We identified chromatin accessibility peaks with MACS2[96] by calling peaks on aggregated reads
511 from each cluster. In brief, we extracted reads from all cells within a given cluster, shifted reads
512 aligned to the positive strand by +4 bp and reads aligned to the negative strand by -5 bp, and

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

513 centered the reads. We then used MACS2 to call peaks (parameters: '--nomodel --keep-dup-all')
514 and removed peaks overlapping ENCODE blacklisted regions[97]. We then merged peaks from all
515 28 clusters with bedtools[98] to create a consistent set of 448,142 regulatory elements for
516 subsequent analyses.

517 To compare accessible chromatin profiles from snATAC-seq to those from bulk ATAC-seq on
518 FACS purified cell types, we reprocessed published ATAC-seq data from sorted pancreatic[99] and
519 unstimulated immune cells[55]. We created pseudobulk profiles from the snATAC-seq data for each
520 donor and cluster, retaining those that contained information from at least 50 cells. We then
521 extracted read counts in the 448,142 merged peaks for all sorted and pseudobulk profiles. We
522 used PCA to extract the top 20 principal components and used UMAP for dimensionality reduction
523 and visualization (parameters: 'min_dist=.5, n_neighbors=80').

524 To identify cluster-specific peaks, we used logistic regression models for each peak treating each
525 cell as an individual data point. For each model, we used cluster assignment and covariates such
526 as donor-of-origin and the log usable count as predictors and binary accessibility of the peak as
527 the outcome to calculate t-statistics (t-stats) for specificity. For a given cluster, we defined cluster-
528 specific peaks by taking the top 1000 peaks with the highest t-stats, after first filtering out peaks
529 which also had high t-stats for other clusters (peak t-stat>90th percentile of all t-stats for the given
530 cluster in more than 2 other clusters). We then used GREAT[12] to annotate peaks and summarize
531 linked genes in the form of gene ontology terms for the set of cluster-specific peaks as compared
532 to all merged peaks.

533 We estimated TF motif enrichment z-scores for each cell using chromVAR[13] (version 1.5.0) by
534 following the steps outlined in the user manual. First, we constructed a sparse binary matrix
535 encoding read overlap with merged peaks for each cell. For each merged peak, we estimated the
536 GC content bias based on the hg19 human reference genome to obtain a set of matched
537 background peaks. To ensure a motif enrichment value for each cell, we did not apply any
538 additional filters based on total reads or the fraction of reads in peaks. Next, using 580 TF motifs
539 within the JASPAR 2018 CORE vertebrate (non-redundant) set[14], we computed GC bias-
540 corrected enrichment z-scores (chromVAR deviation scores) for each cell. To extract highly
541 variable TF motifs, we computed the enrichment variability of each motif across all cells and used
542 the median as the cutoff. For each cluster, we then computed the average TF motif enrichment
543 z-score across all cells in the cluster.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

544 We used Cicero[24] (version 1.3.3) to calculate co-accessibility scores between pairs of peaks for
545 each cluster. As in the single cell motif enrichment analysis, we started from a sparse binary
546 matrix. For each cluster, we only retained merged peaks that overlapped peaks from the cluster.
547 Within each cluster, we aggregated cells based on the 50 nearest neighbors and used cicero to
548 calculate co-accessibility scores, using a 1 Mb window size and a distance constraint of 500 kb.
549 We then defined promoters as ±500 bp from the TSS of protein coding transcripts to annotate co-
550 accessibility links between distal and promoter peaks.

551 **GWAS enrichment analyses**

552 We used LD score regression[100] to calculate genome-wide enrichment z-scores for 32 diseases
553 and traits including T1D. We obtained GWAS summary statistics for autoimmune and
554 inflammatory diseases (immune-related)[25–31,41,42], other diseases[32–40], and quantitative
555 endophenotypes[43–52], and where necessary, we filled in variant IDs and alleles. Using the
556 'munge_sumstats.py' script, we converted summary statistics to the standard format for LD score
557 regression. For each cluster, we used overlap with chromatin accessibility peaks as a binary
558 annotation for variants. We also created a background annotation using merged peaks across all
559 clusters. Then, we computed annotation-specific LD scores by following the instructions for
560 creating partitioned LD scores. We used stratified LD score regression[53] to estimate enrichment
561 coefficient z-scores for each annotation relative to the background, which we defined as merged
562 peaks across all clusters combined with the annotations in the baseline-LD model (version 2.2).
563 Based on the enrichment z-scores, we computed one-sided p-values to assess significance and
564 corrected for multiple tests using the Benjamini-Hochberg procedure[101]. We also calculated
565 GWAS enrichment z-scores for T1D using a version of the meta-analysis excluding the
566 Immunochip cohorts and observed highly similar enrichment results. We used fgwas to estimate
567 enrichment within cell type-specific cCREs using 2000 variants per window.

568 **Annotating cell type mechanisms of variants at fine mapped signals**

569 We first annotated fine mapped variants with PPA>1% using broad genomic annotations. We
570 defined "coding" as coding exons of protein coding genes, "promoter" as ±500 bp from the TSS
571 of protein coding transcripts, and "distal" as peaks in any cell type that did not overlap promoter
572 regions. We then assigned variants to each group without replacement, in the priority
573 coding>promoter>distal. To then further breakdown distal variants, we assigned clusters to cell
574 type groups (CD4 T cell: naïve CD4 T, activated CD4 T, regulatory T; CD8 T cell: naïve CD8 T,
575 activated CD8 T, pancreatic CD8 T; NK cell: adaptive and cytotoxic NK; B cell: naïve and memory
576 B; monocyte/ MΦ: classical and non-classical monocyte, pancreatic macrophage; dendritic:

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

577 conventional and plasmacytoid dendritic; other cell: megakaryocyte, endothelial, activated and
578 quiescent stellate; exocrine: acinar and ductal; endocrine: alpha, beta, delta, and gamma) and
579 created merged peak annotations for each group. We then assigned variants to each cell type
580 group without replacement, prioritizing groups in order based on their cumulative PPA.

581 **Luciferase reporter assay**

582 To test for allelic differences in enhancer activity at rs7795896, we cloned human DNA sequences
583 (Coriell) containing the reference or alternate allele upstream of the minimal promoter in the
584 luciferase reporter vector pGL4.23 (Promega) in the forward direction using the restriction
585 enzymes SacI and KpnI. We then created a construct containing the alternate allele using the
586 NEB Q5 SDM kit (New England Biolabs). The primer sequences used were:
587

588 Cloning FWD_P1 TAGCGGTACCTAATGGGAAATCATGCCAACC
589 Cloning FWD_P2 AATAGAGCTCATGTGTGTGTGCTGGGATGT
590

591 We grew Capan-1 cells (ATCC) to approximately 70% confluency in 6-well dishes according to
592 ATCC culture recommendations. We co-transfected cells with either the experimental or empty
593 vector and pRL-SV40. We then lysed cells 48 hours post transfection and assayed them using
594 the Dual-Luciferase Reporter System (Promega). We normalized Firefly activity to Renilla activity
595 and expressed normalized results as fold change compared to the luciferase activity of the empty
596 vector. We used a two-sided t-test to compare the luciferase activity between the two alleles.
597

598 **Electrophoretic mobility shift assay**

599 We ordered 5' biotinylated and unlabeled (cold) oligos with the reference and alternate alleles
600 from Integrated DNA Technologies. We annealed oligos with an equivalent volume of equimolar
601 complementary oligo in a binding buffer containing 10mM Tris pH 8.0, 50mM NaCl, and 1mM
602 EDTA at 95ºC for 5 minutes and cooled them gradually to room temperature before further use.
603

604 C oligo: (5' biotin)CAATTAGATGTAACTCATTAACATTAGAAAAA
605 T oligo: (5' biotin)CAATTAGATGTAACTTATTAACATTAGAAAAA
606

607 We carried out binding reactions using the LightShift Chemiluminescent EMSA kit (Thermo
608 Fisher) according to manufacturer's instructions with the following adjustments: 100 fmol of
609 biotinylated probe per reaction and 20 pmol of non-biotinylated "cold" probe in competition

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

610 reactions. We used approximately 16 ug of nuclear protein extract from Capan-1 cells purified
611 using NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Fisher) per binding
612 reaction.
613

614 **CRISPR inactivation of enhancer element**

615 We maintained HEK293T cells in DMEM containing 100 units/mL penicillin and 100 mg/mL
616 streptomycin sulfate supplemented with 10% fetal bovine serum (FBS). To generate CRISPRi
617 expression vectors, we designed guide RNA sequences to target the enhancer containing
618 rs7795896 or the *CFTR* promoter. These guides, as well as a non-targeting control, were placed
619 downstream of the human U6 promoter in the pLV hU6-sgRNA hUbC-dCas9-KRAB-T2a-Puro
620 backbone (Addgene, #71236). The guide RNA sequences were:

rs7795896 enhancer guide 1	GTAGTTGGCTTCCTCAGTAAG
rs7795896 enhancer guide 2	GAACAGTATGATTTACGTAA
CFTR promoter	GCGCCCGAGAGACCATGCAG
Non-targeting control	GTGACGTGCACCGCGGTGTG

621

622 We generated high-titer lentiviral supernatants by co-transfection of the resulting plasmid and
623 lentiviral packaging constructs into HEK293T cells. Specifically, we co-transfected CRISPRi
624 vectors with the pCMV-R8.74 (Addgene, #22036) and pMD2.G (Addgene, #12259) expression
625 plasmids into HEK293T cells using a 1mg/mL PEI solution (Polysciences). We collected lentiviral
626 supernatants at 48 hours and 72 hours after transfection and concentrated lentiviruses by
627 ultracentrifugation for 120 minutes at 19,500 rpm using a Beckman SW28 ultracentrifuge rotor at
628 4°C.

629 We obtained Capan-1 pancreatic ductal adenocarcinoma cell lines from ATCC and cultured them
630 using Iscove's Modified Dulbecco's Media with 20% fetal bovine serum, 100 units/mL penicillin,
631 and 100 mg/mL streptomycin sulfate. 24 hours prior to infection, we passaged cells into a 6-well
632 plate at a density of 650,000 cells per well. The following day, we added fresh media containing
633 5ug/mL polybrene and 5uL/mL concentrated CRISPRi lentivirus to each well. We incubated the
634 cells at 37ºC for 30 minutes and then spun them in a centrifuge for 1 hour at 30ºC at 950 × g. 6
635 hours later, we replaced viral media with fresh base culture media and left the cells to recover.
636 After 48 hours, we replaced media daily with the addition of 2ug/mL puromycin for a further 72

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

637 hours. We then harvested infected cells and isolated RNA using the RNeasy® Micro Kit (Qiagen)
638 according to the manufacturer instructions.

639 **Differential analysis of CRISPR inactivation experiments**

640 We used STAR (version 2.7.3a) to map reads to the hg19 genome using ENCODE standard
641 options (parameters: '--outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin
642 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax
643 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000'). We then
644 used featureCounts (version 1.6.4) to count the number of uniquely mapped reads mapping to
645 genes in GENCODE v19 (parameters: '-Q 30 -p -B -s 2 --ignoreDup'). We used DESeq2 to
646 evaluate differential mRNA expression between either the *CFTR* enhancer (pooled data from both
647 guides), or promoter inactivation versus the non-targeting guide.

648 **Colocalization and deconvolution of the pancreas *CFTR* eQTL**

649 We obtained GTEx consortium release v7[60] eQTL summary statistics for pancreas tissue from
650 220 samples and used effect size and standard error estimates to calculate Bayes factors[9] for
651 each variant. Where a T1D-associated variant had evidence for a pancreas eQTL, we considered
652 all variants in a 500kb window around the T1D GWAS index variant, and used the coloc[102]
653 package to calculate the probability that the variants driving T1D association and eQTL signals
654 were shared. We considered signals as colocalized based on the probability that they were shared
655 (PP_{shared}>.9).

656 We downloaded and re-processed a published pancreas single cell RNA-seq dataset[103] of 12 islet
657 donors. After re-processing and generating a counts matrix with the 10x Genomics cellranger
658 (version 3.0.0) pipeline, we first used scanpy[93] and filtered out 1) cells with <500 genes expressed,
659 2) cells with >20% mitochondrial reads, or 3) genes expressed in <3 cells. To ensure clustering
660 would not be affected by read depth, we normalized the total counts per cell to 10k and
661 subsequently log-normalized the resulting counts. We identified highly variable genes (hvgs)
662 based on mean expression and dispersion with (parameters: 'min_mean=.005, max_mean=6,
663 min_disp=.1'). We then extracted counts for hvgs and regressed out the total read count within
664 the hvgs. After dimensionality reduction with PCA, we used harmony[94] with default parameters to
665 correct for batch effects due to donor. We used the top 30 corrected PCs for graph-based
666 clustering with the leiden algorithm[95] (parameters: 'resolution=1.25') and visualization on reduced
667 dimensions with UMAP[104] (parameters: 'min_dist=.3'). To assign cell types to each cluster, we
668 used well-established marker genes from literature and labelled 18,279 cells.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

669 We used MuSiC[61] to estimate the proportions of major pancreatic cell types (acinar, duct, stellate,
670 alpha, beta, delta, gamma) in each pancreas sample from the GTEx v7 release. As input, we
671 used raw count matrices of the islet scRNA-seq and GTEx v7 pancreas samples and cell type
672 labels from the analysis of the former dataset. For each cell type, we used the proportion as an
673 interaction term and constructed linear models of CFTR expression (TMM normalized) as a
674 function of the interaction between genotype dosage and cell type proportion, accounting for
675 covariates used by GTEx including sex, sequencing platform, 3 genotype PCs, and 28 inferred
676 PCs from the expression data. From the original 30 inferred PCs, we excluded inferred PCs 2 and
677 3 because they were highly correlated (Spearman's ρ>.7) with acinar cell proportion.

678 **Phenotype associations at *CFTR* variant**

679 We tested for association of the T1D index variant rs7795896 at *CFTR* to pancreatic and
680 autoimmune disease phenotypes. For acute pancreatitis, chronic pancreatitis, and pancreatic
681 cancer, we used inverse variance weighted meta-analysis to combine SAIGE analysis results
682 from the UK biobank[64] (PheCodes 577.1, 577.2, and 157) and FinnGen (K11_ACUTPANC,
683 K11_CHRONPANC, C3_PANCREAS_EXALLC). As mutations that cause cystic fibrosis (CF)
684 map to this locus, which are risk factors for pancreatitis and pancreatic cancer, we determined
685 the impact of the most common CF mutation F508del/rs199826652 on the association results for
686 rs7795896. For T1D, we tested for association of rs7795896 conditional on F508del/rs199826652
687 in all cohorts except for FinnGen and observed no evidence for a difference in T1D association.
688 For pancreatitis and pancreatic cancer, we identified F508del/rs199826652 carriers in UK
689 Biobank and repeated the association analysis for these phenotypes in UK biobank data after
690 removing these individuals and observed no evidence of a change in the effect of rs7795896.

691

692 ## CODE AVAILABILITY

693 Code used for processing snATAC-seq datasets and clustering cells is available at
694 https://github.com/kjgaulton/pipelines/tree/master/T1D_snATAC_pipeline.

695

696 ## DATA AVAILABILITY

697 Summary statistics and fine mapping credible sets for T1D GWAS will be available in the GWAS
698 catalog and in the T1D Knowledge Portal (http://t1d.hugeamp.org). Raw data files for snATAC-
699 seq will be deposited to GEO, and processed data files for snATAC-seq will be available through
700 the Diabetes Epigenome Atlas (https://www.diabetesepigenome.org/).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

701

702 **ACKNOWLEDGEMENTS**

703 This work was supported by NIH grants DK112155, DK120429 and DK122607 to K.J.G and M.S.,
704 and T32 GM008666 to R.G. We thank Samantha Kuan in the Ren Lab at the LICR for assistance
705 with sequencing.

706 nPOD: This research was performed with the support of the Network for Pancreatic Organ donors
707 with Diabetes (nPOD; RRID:SCR_014641), a collaborative type 1 diabetes research project
708 sponsored by JDRF (nPOD: 5-SRA-2018-557-Q-R) and The Leona M. & Harry B. Helmsley
709 Charitable Trust (Grant #2018PG-T1D053). The content and views expressed are the
710 responsibility of the authors and do not necessarily reflect the official view of nPOD. Organ
711 Procurement Organizations (OPO) partnering with nPOD to provide research resources are listed
712 at http://www.jdrfnpod.org/for-partners/npod-partners/.

713 DCCT/EDIC: The Diabetes Control and Complications Trial (DCCT) and its follow-up the
714 Epidemiology of Diabetes Interventions and Complications (EDIC) study were conducted by the
715 DCCT/EDIC Research Group and supported by National Institute of Health grants and contracts
716 and by the General Clinical Research Center Program, NCRR. The data from the DCCT/EDIC
717 study were supplied by the NIDDK Central Repositories.

718 GENIE: The Genetics of Nephropathy, an International Effort (GENIE) study was conducted by
719 the GENIE Investigators and supported by the National Institute of Diabetes and Digestive and
720 Kidney Diseases (NIDDK). The data from the GENIE study reported here were supplied by the
721 GENIE investigators from the Broad Institute of MIT and Harvard, Queens University Belfast and
722 the University of Dublin.

723 GoKinD: The Genetics of Kidneys in Diabetes (GoKinD) Study was conducted by the GoKinD
724 Investigators and supported by the Juvenile Diabetes Research Foundation, the CDC, and the
725 Special Statutory Funding Program for Type 1 Diabetes Research administered by the National
726 Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data [and samples] from
727 the GoKinD study were supplied by the NIDDK Central Repositories. This manuscript was not
728 prepared in collaboration with Investigators of the GoKinD study and does not necessarily reflect
729 the opinions or views of the GoKinD study, the NIDDK Central Repositories, or the NIDDK.

730 T1DGC: This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium
731 (T1DGC), a collaborative clinical study sponsored by the National Institute of Diabetes and
732 Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and the Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418. The UK case series collection was additionally funded by the JDRF and Wellcome Trust and the National Institute for Health Research Cambridge Biomedical Centre, at the Cambridge Institute for Medical Research, UK (CIMR), which is in receipt of a Wellcome Trust Strategic Award (079895). The data from the T1DGC study were supplied by dbGAP. This manuscript was not prepared in collaboration with Investigators of the T1DGC study and does not necessarily reflect the opinions or views of the T1DGC study or the study sponsors.

T1DGC (ASP/UK GRID): This research was performed under the auspices of the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and Juvenile Diabetes Research Foundation International (JDRF).

WTCCC: This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

UK Biobank: Data from the UK Biobank was accessed under application 24058.

FinnGen: We want to acknowledge the participants and investigators of the FinnGen study.

CSGNM: We thank the participants of the Trinity Student Study. This study was supported by the Intramural Research Programs of the National Institutes of Health, the National Human Genome Research Institute, and the Eunice Kennedy Shriver National Institute of Child Health and Development.

NIMH Schizophrenia Controls: Funding support for the Genome-Wide Association of Schizophrenia Study was provided by the National Institute of Mental Health (R01 MH67257, R01 MH59588, R01 MH59571, R01 MH59565, R01 MH59587, R01 MH60870, R01 MH59566, R01 MH59586, R01 MH61675, R01 MH60879, R01 MH81800, U01 MH46276, U01 MH46289 U01 MH46318, U01 MH79469, and U01 MH79470) and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The datasets used for the analyses

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

764 described in this manuscript were obtained from the database of Genotypes and Phenotypes
765 (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number
766 phs000021.v3.p2. Samples and associated phenotype data for the Genome-Wide Association of
767 Schizophrenia Study were provided by the Molecular Genetics of Schizophrenia Collaboration
768 (PI: Pablo V. Gejman, Evanston Northwestern Healthcare (ENH) and Northwestern University,
769 Evanston, IL, USA).

770 Neurodevelopmental Genomics: Support for the collection of the data for Philadelphia
771 Neurodevelopment Cohort (PNC) was provided by grant RC2MH089983 awarded to Raquel Gur
772 and RC2MH089924 awarded to Hakon Hakonarson. Subjects were recruited and genotyped
773 through the Center for Applied Genomics (CAG) at The Children's Hospital in Philadelphia
774 (CHOP). Phenotypic data collection occurred at the CAG/CHOP and at the Brain Behavior
775 Laboratory, University of Pennsylvania.

776 eMERGE Network: Group Health Cooperative/University of Washington – Funding support for
777 Alzheimer's Disease Patient Registry (ADPR) and Adult Changes in Thought (ACT) study was
778 provided by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). A gift
779 from the 3M Corporation was used to expand the ACT cohort. DNA aliquots sufficient for GWAS
780 from ADPR Probable AD cases, who had been enrolled in Genetic Differences in Alzheimer's
781 Cases and Controls (Walter Kukull, PI, R01 AG007584) and obtained under that grant, were made
782 available to eMERGE without charge. Funding support for genotyping, which was performed at
783 Johns Hopkins University, was provided by the NIH (U01HG004438). Genome-wide association
784 analyses were supported through a Cooperative Agreement from the National Human Genome
785 Research Institute, U01HG004610 (Eric B. Larson, PI). Mayo Clinic – Samples and associated
786 genotype and phenotype data used in this study were provided by the Mayo Clinic. Funding
787 support for the Mayo Clinic was provided through a cooperative agreement with the National
788 Human Genome Research Institute (NHGRI), Grant #: UOIHG004599; and by grant HL75794
789 from the National Heart Lung and Blood Institute (NHLBI). Funding support for genotyping, which
790 was performed at The Broad Institute, was provided by the NIH (U01HG004424). Marshfield Clinic
791 Research Foundation – Funding support for the Personalized Medicine Research Project (PMRP)
792 was provided through a cooperative agreement (U01HG004608) with the National Human
793 Genome Research Institute (NHGRI), with additional funding from the National Institute for
794 General Medical Sciences (NIGMS) The samples used for PMRP analyses were obtained with
795 funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health
796 Policy grant number D1A RH00025, and Wisconsin Department of Commerce Technology

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

797 Development Fund contract number TDF FYO10718. Funding support for genotyping, which was
798 performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Northwestern
799 University – Samples and data used in this study were provided by the NUgene Project
800 (www.nugene.org). Funding support for the NUgene Project was provided by the Northwestern
801 University's Center for Genetic Medicine, Northwestern University, and Northwestern Memorial
802 Hospital. Assistance with phenotype harmonization was provided by the eMERGE Coordinating
803 Center (Grant number U01HG04603). This study was funded through the NIH, NHGRI eMERGE
804 Network (U01HG004609). Funding support for genotyping, which was performed at The Broad
805 Institute, was provided by the NIH (U01HG004424).

806 Vanderbilt University - Funding support for the Vanderbilt Genome-Electronic Records (VGER)
807 project was provided through a cooperative agreement (U01HG004603) with the National Human
808 Genome Research Institute (NHGRI) with additional funding from the National Institute of General
809 Medical Sciences (NIGMS). The dataset and samples used for the VGER analyses were obtained
810 from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and
811 by the Vanderbilt CTSA grant UL1RR024975 from NCRR/NIH. Funding support for genotyping,
812 which was performed at The Broad Institute, was provided by the NIH (U01HG004424).
813 Assistance with phenotype harmonization and genotype data cleaning was provided by the
814 eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for
815 Biotechnology Information (NCBI). The datasets used for the analyses described in this
816 manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP
817 accession number phs000360.v3.p1.

818 This manuscript was not prepared in collaboration with investigators of these studies and does
819 not necessarily reflect the opinions or views of the DCCT/EDIC, GENIE, GoKinD, T1DGC,
820 WTCCC, studies or study groups, the NIDDK Central Repositories, the NIH, or the study
821 sponsors.

822

823 AUTHOR CONTRIBUTIONS

824 K.J.G and J.C. designed the study and wrote the manuscript. J.C. performed the genetic
825 association and single cell accessible chromatin analyses. R.G., M.O. and S.Huang performed
826 molecular experiments of enhancer and variant function. J.Y.H and M.M. generated single cell
827 accessible chromatin data. P.B. and K.K. contributed to data analysis. D.U.G and S.P. supervised
828 the generation of single cell accessible chromatin and contributed to data interpretation and

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

829 analyses. M.S. supervised experiments related to enhancer function and contributed to data
830 interpretation. S.Heller and A.K. contributed to interpretation of experimental data.

831

832 FIGURE LEGENDS

833 **Figure 1. Genome-wide association and fine mapping identifies novel signals for T1D risk.**
834 (a) Manhattan plot showing genome-wide T1D association p-values (-log10 transformed). Novel
835 loci are colored in red and labeled based on the nearest gene, and index variants have larger
836 radii and are circled. The dotted line indicates genome-wide significance ($P=5\times10^{-8}$). (b) Locus
837 plots showing independent association signals at the known *BACH2* locus (left) and the novel
838 *BCL11A* locus (right). For conditional signals, the variants used for conditional analysis are
839 indicated under the title in parentheses. Variants are colored (known=blue, novel=red) based on
840 linkage disequilibrium (r^2) with the index variant for each signal. The dotted line indicates the
841 genome-wide significance threshold ($P=5\times10^{-8}$) for the main signal and the locus wide
842 significance threshold ($P=1\times10^{-5}$) for the conditional signals. (c) Breakdown of 141 independent
843 T1D risk signals after conditional fine-mapping analyses. Among these were 89 main signals at
844 59 known loci (excluding the MHC region) and 30 novel loci, and 52 conditional signals including
845 43 at known loci and 9 at novel loci. (d) Breakdown of the number of signals per locus (top),
846 number of 99% credible set variants per signal from fine mapping (middle), and the number of
847 variants with posterior probability of association >1% (bottom).

848 **Figure 2. Comprehensive reference map of 131,554 single cell chromatin accessibility**
849 **profiles from T1D-relevant tissues.** (a) Clustering of accessible chromatin profiles from 131,554
850 cells from single cell experiments of peripheral blood mononuclear cells, whole pancreas tissue,
851 and purified pancreatic islets. Cells are plotted on the first two UMAP components and colored
852 based on cluster assignment. Clusters are grouped into categories of cell types, and the number
853 of cells in each cluster are shown next to its corresponding label. (b) Dot plot (top) of relative gene
854 accessibility (chromatin accessibility reads across gene bodies, averages for each cluster and
855 scaled from 0-100 across columns/clusters) showing examples of marker genes used to identify
856 cluster labels. Circle sizes are scaled according to the relative gene accessibility value. Genome
857 browser tracks (bottom) showing aggregated chromatin accessibility profiles in a 50 kb window
858 around selected marker genes. (c) Relative peak accessibility for 25,436 cluster-specific peaks
859 across all 28 clusters (left), and enriched gene ontology terms with GREAT for peaks specific to
860 pancreatic macrophages, activated stellate, ductal, and acinar cells (right). (d) Single cell motif

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

861 enrichment z-scores for TFs showing specificity for cell lineage (SPI – myeloid and B cells, ETS
862 – T cells, FOXA – pancreatic), cell type (NR5A – acinar, HNF1 – ductal, EBF – B cells), and cell
863 state (POU2 – memory B, TCF7 – naïve CD4 T, RUNX – adaptive NK) . The sequence logo for
864 the enriched motif is displayed to the left of each UMAP plot. (e) Examples of cell type-specific
865 co-accessibility between the promoter of AQP1 and distal sites in ductal cells (left,
866 chr7:30,000,000-31,100,000, scale: 0-10 CPM) and the promoter of CEL and distal sites in acinar
867 cells (right, chr9:135,800,000-136,000,000, scale: 0-10 CPM).

868 **Figure 3. Cell type-specific enrichment and mechanisms of T1D risk variants.** (a) Relative
869 LD score regression enrichment z-scores (enrichment relative to background genomic
870 annotations including a merged set of all peaks) for autoimmune and inflammatory diseases (top),
871 other diseases (middle), and non-disease quantitative endophenotypes (bottom) for cCREs active
872 in pancreatic and blood cell types and states. ***FDR<.001 **FDR<.01 *FDR<.1. (b) T1D
873 enrichment within cell type-specific cCREs. Labeled clusters have a positive enrichment estimate.
874 Points represent log-transformed fgwas enrichment estimates and lines represent 95%
875 confidence intervals. (c) Breakdown of cumulative fine mapping probability (PPA) (left) and fine
876 mapped variants (right). Variants and their probabilities are assigned without replacement to
877 annotations from top to bottom. Variants are first broken down by genomic annotations (top), and
878 variants overlapping a distal peak are further broken down by cell type groups (bottom). CD4 T
879 cell: naïve CD4 T + activated CD4 + regulatory T; exocrine: acinar + ductal; endocrine: GCG^{high}
880 alpha + GCG^{low} alpha + INS^{high} beta + INS^{low} beta + SST^{high} delta + SST^{low} delta + gamma;
881 monocyte/MΦ: classical monocyte + non-classical monocyte + pancreatic macrophage; NK cell:
882 cytotoxic NK + adaptive NK; B cell: naïve B + memory B; CD8 T cell: naïve CD8 T + activated
883 CD8 T + pancreatic CD8 T; other cell: megakaryocytes + activated stellate + quiescent stellate +
884 endothelial; dendritic: conventional dendritic + plasmacytoid dendritic. (d) Signals with the highest
885 cumulative PPA for cell type groups with at least 2.5 cumulative PPA. (e) The *GP2* signal contains
886 3 variants (rs4238595, rs8060932, and rs8060932) in a distal peak upstream of the *GP2* promoter
887 (top, chr16:20,300,000-20,380,000). These variants are linked to *GP2* through co-accessibility in
888 acinar cells and account for the majority of the causal probability (cumulative PPA=.98) for the
889 signal (middle). Genome browser tracks (bottom) show that chromatin accessibility at both the
890 peak and the *GP2* promoter is highly specific to acinar cells. (f) The top variant at the
891 *CTRB1/2/BCAR1* signal rs72802342 (middle) overlaps a distal peak co-accessible with the
892 *CTRB2 and CTRB1* promoters in acinar cells (top: chr16:75,220,000-75,260,000, hg19). Genome
893 browser tracks (bottom, scale: 0-15) show that chromatin accessibility at the *CTRB1* and *CTRB2*

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

894 promoters are highly specific to acinar cells. Fine mapped variants are colored based on linkage
895 disequilibrium to the index variant. Variants contained in the 99% credible set are circled in black.

896 **Figure 4. Fine-mapped variant at the *CFTR* locus mediates T1D risk through distal**
897 **regulation of *CFTR* in pancreatic ductal cells.** (a) The *CFTR* locus contains a single fine-
898 mapped variant (rs7795896) in a distal cCRE linked to the promoter of *CFTR* and several other
899 genes through co-accessibility (top; region shown: chr7:116,490,000-117,860,000). The cCRE is
900 located approximately 33 kb upstream of the *CFTR* promoter. Zoomed-in view (chr7:117,040,000-
901 117,140,000, scale: 0-5 CPM) of fine mapped variants (middle) and genome browser tracks
902 (bottom) at this locus show that the cCRE is highly specific to ductal cells. (b) Luciferase reporter
903 assay in Capan-1 cells transfected with pGL4.23 minimal promoter plasmids containing
904 rs7795896 in the forward orientation. Relative luciferase units represent Firefly:Renilla ratios
905 normalized to control cells transfected with the empty vector. P-values are from a two-tailed
906 Student's t-test. (c) Electrophoretic mobility shift assay (EMSA) with nuclear extract from Capan-
907 1 cells using probes from both alleles of rs7795896. Bands with specific binding are labeled. (d)
908 CRISPR interference-mediated inactivation of the distal site containing rs7795896 (*CFTR*^{iEnh}; 2
909 guide RNAs; 3 replicates; n=6 total) or the *CFTR* promoter (*CFTR*^{iProm}; n=3 replicates) in CAPAN-
910 1 cells. Differential analysis of genes with promoters co-accessible with the peak show that CFTR
911 expression is significantly reduced in both *CFTR*^{iProm} and *CFTR*^{iEnh} cells. Data are shown as
912 transcripts per million (TPM). Error bars show 95% confidence interval and datapoints underlying
913 each boxplot are shown. (e) Bayesian colocalization showing that the T1D risk signal (top) and
914 *CFTR* pancreas eQTL from GTEx v7 (bottom) are likely driven by the same causal variant.
915 Variants are colored based on the linkage disequilibrium to the index variant. Variants in the 99%
916 credible set are circled in black. (f) Heatmap showing the average expression (normalized counts,
917 scaled from 0-1 across cell types) of marker genes of different pancreatic cell types from single
918 cell RNA-seq. *CFTR* expression is highly specific to ductal cells. (g) Deconvolution of the *CFTR*
919 pancreas eQTL using *in-silico* cell type proportion estimation and re-analyses of GTEx pancreas
920 data using interaction analyses shows that the eQTL signal only has a significant interaction with
921 ductal cell proportion. (h) Forest plot showing association of pancreatic disease traits in a meta-
922 analysis of UK Biobank and FinnGen data for rs7795896 compared to association of autoimmune
923 traits from large European GWAS. (i) Variants regulating genes with specialized function in the
924 exocrine pancreas influence risk of type 1 diabetes. At the *CFTR* locus, a variant reducing ductal
925 cell enhancer activity and *CFTR* expression increases risk of T1D and other pancreatic disease,
926 and we hypothesize that these effects are mediated through inflammation and immune infiltration
927 in the exocrine pancreas.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

928 **REFERENCES**

929 1. Katsarou, A. *et al.* Type 1 diabetes mellitus. *Nat. Rev. Dis. Primer* **3**, 17016 (2017).

930 2. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence
931 for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386
932 (2015).

933 3. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci
934 affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).

935 4. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional
936 type 1 diabetes loci. *Nat. Genet.* **40**, 1399–1401 (2008).

937 5. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases
938 of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

939 6. Bradfield, J. P. *et al.* A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts
940 Identifies Multiple Associated Loci. *PLOS Genet.* **7**, e1002293 (2011).

941 7. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
942 *bioRxiv* 563866 (2019) doi:10.1101/563866.

943 8. Raeder, H. *et al.* Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic
944 exocrine dysfunction. *Nat. Genet.* **38**, 54–62 (2006).

945 9. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values.
946 *Genet. Epidemiol.* **33**, 79–86 (2009).

947 10. Aylward, A., Chiou, J., Okino, M.-L., Kadakia, N. & Gaulton, K. J. Shared genetic risk
948 contributes to type 1 and type 2 diabetes etiology. *Hum. Mol. Genet.* (2018)
949 doi:10.1093/hmg/ddy314.

950 11. Chiou, J. *et al.* Single cell chromatin accessibility reveals pancreatic islet cell type- and
951 state-specific regulatory programs of diabetes risk. *bioRxiv* 693671 (2019)
952 doi:10.1101/693671.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

953 12. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions.
954 *Nat. Biotechnol.* **28**, 495–501 (2010).
955 13. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
956 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*
957 **14**, 975–978 (2017).
958 14. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription
959 factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D1284 (2018).
960 15. Chen, H. *et al.* PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene* **11**,
961 1549–1560 (1995).
962 16. Kaestner, K. H. The FoxA factors in organogenesis and differentiation. *Curr. Opin.*
963 *Genet. Dev.* **20**, 527–532 (2010).
964 17. Eyquem, S., Chemin, K., Fasseu, M. & Bories, J.-C. The Ets-1 transcription factor is
965 required for complete pre-T cell receptor function and allelic exclusion at the T cell receptor
966 beta locus. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15712–15717 (2004).
967 18. Hale, M. A. *et al.* The nuclear hormone receptor family member NR5A2 controls aspects
968 of multipotent progenitor cell formation and acinar differentiation during pancreatic
969 organogenesis. *Dev. Camb. Engl.* **141**, 3123–3133 (2014).
970 19. De Vas, M. G. *et al.* Hnf1b controls pancreas morphogenesis and the generation of
971 Ngn3+ endocrine progenitors. *Dev. Camb. Engl.* **142**, 871–882 (2015).
972 20. O'Riordan, M. & Grosschedl, R. Coordinate regulation of B cell differentiation by the
973 transcription factors EBF and E2A. *Immunity* **11**, 21–31 (1999).
974 21. Corcoran, L. M. & Karvelas, M. Oct-2 is required early in T cell-independent B cell
975 activation for G1 progression and for proliferation. *Immunity* **1**, 635–645 (1994).
976 22. Issuree, P. D. *et al.* Stage-specific epigenetic regulation of CD4 expression by
977 coordinated enhancer elements during T cell development. *Nat. Commun.* **9**, 3594 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

978 23. Rapp, M. *et al.* Core-binding factor β and Runx transcription factors promote adaptive
979 natural killer cell responses. *Sci. Immunol.* **2**, (2017).
980 24. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell
981 Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).
982 25. Ji, S.-G. *et al.* Genome-wide association study of primary sclerosing cholangitis identifies
983 new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat.*
984 *Genet.* **49**, 269–273 (2017).
985 26. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate
986 and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat.*
987 *Genet.* **47**, 1457–1464 (2015).
988 27. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary
989 biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
990 28. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug
991 discovery. *Nature* **506**, 376–381 (2014).
992 29. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of
993 multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
994 30. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune
995 gene expression. *Nat. Genet.* **42**, 295–302 (2010).
996 31. Jin, Y. *et al.* Genome-wide association studies of autoimmune vitiligo identify 23 new risk
997 loci and highlight key pathways and regulatory variants. *Nat. Genet.* **48**, 1418–1424 (2016).
998 32. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional
999 pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
1000 33. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using
1001 high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513
1002 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1003 34. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new
1004 loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
1005 35. Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with
1006 bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
1007 36. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine
1008 the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
1009 37. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum
1010 disorder. *Nat. Genet.* **51**, 431–444 (2019).
1011 38. Watson, H. J. *et al.* Genome-wide association study identifies eight risk loci and
1012 implicates metabo-psychiatric origins for anorexia nervosa. *Nat. Genet.* **51**, 1207–1214
1013 (2019).
1014 39. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological
1015 insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
1016 40. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses
1017 of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
1018 41. López-Isac, E. *et al.* GWAS for systemic sclerosis identifies multiple risk loci and
1019 highlights fibrotic and vasculopathy pathways. *Nat. Commun.* **10**, 4955 (2019).
1020 42. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and
1021 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–1456
1022 (2015).
1023 43. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body
1024 mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649
1025 (2018).
1026 44. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with
1027 adult disease. *Nature* **538**, 248–252 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1028 45. Jiang, X. *et al.* Genome-wide association study in 79,366 European-ancestry individuals
1029 informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun.* **9**, 260 (2018).
1030 46. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies
1031 genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**,
1032 659–669 (2012).
1033 47. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2
1034 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide
1035 meta-analysis. *PLoS Med.* **14**, e1002383 (2017).
1036 48. Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic
1037 signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**,
1038 1294–1303 (2015).
1039 49. Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at
1040 menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49**, 834–841
1041 (2017).
1042 50. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals
1043 identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
1044 51. Saxena, R. *et al.* Genetic variation in GIPR influences the glucose and insulin responses
1045 to an oral glucose challenge. *Nat. Genet.* **42**, 142–148 (2010).
1046 52. Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants
1047 associated with fasting proinsulin levels and provides new insights into the pathophysiology
1048 of type 2 diabetes. *Diabetes* **60**, 2624–2634 (2011).
1049 53. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide
1050 association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
1051 54. Wing, K. *et al.* CTLA-4 control over Foxp3+ regulatory T cell function. *Science* **322**, 271–
1052 275 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1053 55. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse
1054 human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
1055 56. Ramos-Rodríguez, M. *et al.* The impact of proinflammatory cytokines on the β-cell
1056 regulatory landscape provides insights into the genetics of type 1 diabetes. *Nat. Genet.* **51**,
1057 1588–1595 (2019).
1058 57. Gibson-Corley, K. N., Meyerholz, D. K. & Engelhardt, J. F. Pancreatic Pathophysiology
1059 in Cystic Fibrosis. *J. Pathol.* **238**, 311–320 (2016).
1060 58. Sharer, N. *et al.* Mutations of the cystic fibrosis gene in patients with chronic pancreatitis.
1061 *N. Engl. J. Med.* **339**, 645–652 (1998).
1062 59. Namkung, W. *et al.* Ca2+ activates cystic fibrosis transmembrane conductance
1063 regulator- and Cl- -dependent HCO3 transport in pancreatic duct cells. *J. Biol. Chem.* **278**,
1064 200–207 (2003).
1065 60. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues.
1066 *Nature* **550**, 204–213 (2017).
1067 61. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution
1068 with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 1–9 (2019).
1069 62. McWilliams, R. R. *et al.* Cystic fibrosis transmembrane conductance regulator (CFTR)
1070 gene mutations and risk for pancreatic adenocarcinoma. *Cancer* **116**, 203–209 (2010).
1071 63. Noone, P. G. *et al.* Cystic fibrosis gene mutations and pancreatitis risk: relation to
1072 epithelial ion transport and trypsin inhibitor gene mutations. *Gastroenterology* **121**, 1310–
1073 1319 (2001).
1074 64. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness
1075 in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
1076 65. Virostko, J. *et al.* Pancreas Volume Declines During the First Year After Diagnosis of
1077 Type 1 Diabetes and Exhibits Altered Diffusion at Disease Onset. *Diabetes Care* **42**, 248–
1078 257 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1079 66. Campbell-Thompson, M., Wasserfall, C., Montgomery, E. L., Atkinson, M. A. & Kaddis,
1080 J. S. Pancreas organ weight in individuals with disease-associated autoantibodies at risk for
1081 type 1 diabetes. *JAMA* **308**, 2337–2339 (2012).
1082 67. Campbell-Thompson, M. L. *et al.* Relative Pancreas Volume Is Reduced in First-Degree
1083 Relatives of Patients With Type 1 Diabetes. *Diabetes Care* **42**, 281–287 (2019).
1084 68. Campbell-Thompson, M., Rodriguez-Calvo, T. & Battaglia, M. Abnormalities of the
1085 Exocrine Pancreas in Type 1 Diabetes. *Curr. Diab. Rep.* **15**, 79 (2015).
1086 69. Campbell-Thompson, M. L. *et al.* The influence of type 1 diabetes on pancreatic weight.
1087 *Diabetologia* **59**, 217–221 (2016).
1088 70. Hart, N. J. *et al.* Cystic fibrosis-related diabetes is caused by islet loss and inflammation.
1089 *JCI Insight* **3**, (2018).
1090 71. Navis, A. & Bagnat, M. Loss of cftr function leads to pancreatic destruction in larval
1091 zebrafish. *Dev. Biol.* **399**, 237–248 (2015).
1092 72. Valle, A. *et al.* Reduction of circulating neutrophils precedes and accompanies type 1
1093 diabetes. *Diabetes* **62**, 2072–2077 (2013).
1094 73. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based
1095 linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
1096 74. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat.*
1097 *Genet.* **48**, 1279–1283 (2016).
1098 75. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
1099 *Nature* **526**, 68–74 (2015).
1100 76. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**,
1101 1284–1287 (2016).
1102 77. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation.
1103 *Bioinforma. Oxf. Engl.* **31**, 782–784 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1104 78. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial
1105 fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
1106 79. Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through
1107 genome-wide analyses of UK Biobank data. *Nat. Genet.* **51**, 230–236 (2019).
1108 80. Taal, H. R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head
1109 circumference. *Nat. Genet.* **44**, 532–538 (2012).
1110 81. Teumer, A. *et al.* Genome-wide analyses identify a role for SLC17A4 and AADAT in
1111 thyroid hormone regulation. *Nat. Commun.* **9**, (2018).
1112 82. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci
1113 associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
1114 83. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat
1115 distribution. *Nature* **518**, 187–196 (2015).
1116 84. Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies
1117 new risk loci and functional pathways. *Nat. Genet.* **51**, 394–403 (2019).
1118 85. Cousminer, D. L. *et al.* Genome-wide association and longitudinal analyses reveal
1119 genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Hum. Mol.*
1120 *Genet.* **22**, 2735–2747 (2013).
1121 86. Felix, J. F. *et al.* Genome-wide association analysis identifies three new susceptibility
1122 loci for childhood body mass index. *Hum. Mol. Genet.* **25**, 389–403 (2016).
1123 87. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat.*
1124 *Genet.* **45**, 1274–1283 (2013).
1125 88. van der Valk, R. J. P. *et al.* A novel common variant in DCST2 is associated with length
1126 in early life and height in adulthood. *Hum. Mol. Genet.* **24**, 1155–1168 (2015).
1127 89. Cusanovich, D. A. *et al.* Multiplex Single Cell Profiling of Chromatin Accessibility by
1128 Combinatorial Cellular Indexing. *Science* **348**, 910–914 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1129 90. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse
1130 forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439
1131 (2018).
1132 91. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
1133 transform. *Bioinforma. Oxf. Engl.* **26**, 589–595 (2010).
1134 92. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*
1135 **25**, 2078–2079 (2009).
1136 93. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression
1137 data analysis. *Genome Biol.* **19**, 15 (2018).
1138 94. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
1139 Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
1140 95. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-
1141 connected communities. *Sci. Rep.* **9**, 1–12 (2019).
1142 96. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137
1143 (2008).
1144 97. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
1145 human genome. *Nature* **489**, 57–74 (2012).
1146 98. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1147 features. *Bioinformatics* **26**, 841–842 (2010).
1148 99. Arda, H. E. *et al.* A Chromatin Basis for Cell Lineage and Disease Risk in the Human
1149 Pancreas. *Cell Syst.* **7**, 310-322.e4 (2018).
1150 100. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
1151 polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
1152 101. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
1153 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1154 102. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic
1155 association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
1156 103. Xin, Y. *et al.* Pseudotime Ordering of Single Human β-Cells Reveals States of Insulin
1157 Production and Unfolded Protein Response. *Diabetes* db180365 (2018) doi:10.2337/db18-
1158 0365.
1159 104. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for
1160 Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).
1161

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1

a

RUNX3 TFAP2E PGM1 CD2 NOTCH2 PTPRC BATF3 BCL11A SEPT2 4p14 IRF2 IRF4 ITGB8 CFTR TOX COLEC10 AGO2 CEL GATA3 PRF1 CD6 EMSY CD3D FLI1 AKAP11 GP2 16q23 NEURL4 PRR15L AIRE

$-\log_{10}$(P-value)

Chromosome

b

BACH2 (known locus) linkage (r^2) 0 1

BCL11A (novel locus) linkage (r^2) 0 1

Main signal 5×10^{-8}

Cond. signal 1 (rs6908626) 1×10^{-5}

Cond. signal 2 (rs6908626, rs7754251) 1×10^{-5}

Cond. signal 1 (rs13035645) 1×10^{-5}

Cond. signal 2 (rs13035645, rs77146844) 1×10^{-5}

$-\log_{10}$(P-value)

BACH2 90.6 Chr.6 (Mb) 91.4

BCL11A 60.0 Chr.2 (Mb) 61.5

c

at known loci
at novel loci
89 main signals
141 independent signals
59 30 9 43
52 cond. signals
at novel loci
at known loci

d

Number of signals per locus:

1	2	3	4+
53	25	8	3

Number of credible set variants per signal:

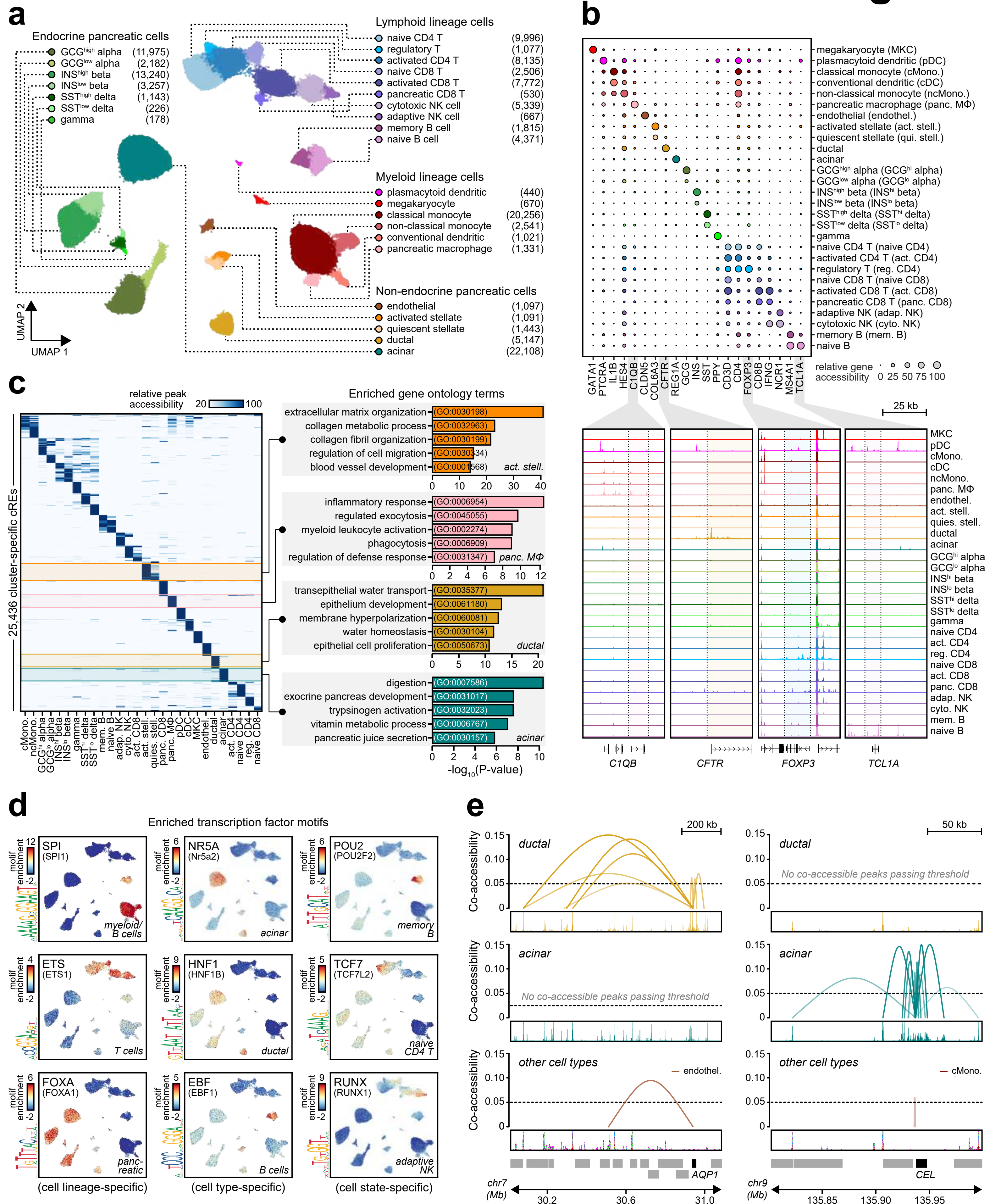
1-2	3-5	6-20	21-50	51-100	101+
11	19	39	34	23	15

Number of variants with PPA > 1%:

>50	30-50	20-30	10-20	1-10%
35	34	69	134	1,701

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3

a

GWAS enrichment
-5 5

Immune-related diseases

Type 1 diabetes
Atopic dermatitis
Primary sclerosing cholangitis
Systemic lupus erythematosus
Primary biliary cirrhosis
Systemic sclerosis
Rheumatoid arthritis
Ulcerative colitis
Celiac disease
Autoimmune vitiligo
Crohn's disease

Other diseases

Alzheimer's disease
Type 2 diabetes
Chronic kidney disease
Coronary artery disease
Bipolar disorder
Major depressive disorder
Autism spectrum disorder
Anorexia nervosa
Schizophrenia

Non-disease traits

Height
Birth weight
Vitamin D
Fasting insulin
Hemoglobin A1c (HbA1c)
Age at menopause
Age at menarche
Intelligence
Body mass index (BMI)
Fasting proinsulin
Fasting glucose
2hr glucose adj. BMI

naive CD4, act. CD4, reg. CD4, naive CD8, act. CD8, panc. CD8, naive B, mem. B, adap. NK, cytot. NK, INS^{hi} beta, INS^{lo} beta, GCG^{hi} alpha, GCG^{lo} alpha, SST^{hi} delta, SST^{lo} delta, gamma, cMono., ncMono., cDC, pDC, panc. MΦ, MKC, act. stell., qui. stell., endothel., acinar, ductal

b

T1D log(enrichment) in cluster-specific cREs

act. CD4, pDC, cMono., INS^{hi} beta, ductal, acinar

c

distal, coding, prom., break-down

CD4+ T cell, exocrine, monocyte/MΦ, endocrine, NK cell, B cell, other cell, CD8+ T cell, dendritic

Cumulative PPA across annotation

Fine mapped variants with PPA>1%

d

distal cumulative PPA 0 1

CD4+ T cell, exocrine, monocyte/MΦ, endocrine

CD2 (main), IL2RA (cond.5), IL2RA (main), PRF1, IKZF4 (cond.2), RAD51B, C1QTNF6 (cond.2), CAMSAP2, SIRPG (main), BACH2 (main), CTLA4 (main), GP2, RNLS (main), CFTR, BCAR1 (main), IL10 (main), FLI1, AGO2, TNFAIP3(cond.1), IKZF1(cond.1), DLK1(cond.1), TOX(main), COLEC10, IFIH1 (cond.3), BACH2 (cond.2)

e

10 kb

Co-accessibility

exocrine

acinar, ductal

PPA

rs4238595, rs8062123, rs8060932

linkage (r^2) 0 1

GP2, *PDILT*, *UMOD*

exocrine (acinar, ductal); endocrine (alpha, delta, beta, gamma); CD4+ T cell (naive / act. CD4, reg. CD4); CD8+ T cell (naive / act. CD8, panc. CD8); NK cell (cyto. NK, adap. NK); B cell (naive B, mem. B); monocyte/MΦ (cMono / ncMono, panc. MΦ); dendritic (pDC, cDC); other cell (endothel., MKC, quies. / act. stell.)

Chromosome 16 (Mb)

f

5 kb

Co-accessibility

exocrine

acinar, ductal

PPA

rs72802342, rs55993634, rs8056814

linkage (r^2) 0 1

CTRB2, *CTRB1*

exocrine (acinar, ductal); endocrine (alpha, delta, beta, gamma); CD4+ T cell (naive / act. CD4, reg. CD4); CD8+ T cell (naive / act. CD8, panc. CD8); NK cell (cyto. NK, adap. NK); B cell (naive B, mem. B); monocyte/MΦ (cMono / ncMono, panc. MΦ); dendritic (pDC, cDC); other cell (endothel., MKC, quies. / act. stell.)

Chromosome 16 (Mb)

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426472; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4

a

ductal

Co-accessibility

200 kb

CAPZA2 WNT2 NAA38 ST7 ASZ1 CTTNBP2 CFTR

T1D PPA

rs7795896

linkage (r^2)

ASZ1 CFTR

exocrine: acinar, ductal
endocrine: alpha, delta, beta, gamma
CD4+ T cell: naive / act. CD4, reg. CD4
CD8+ T cell: naive / act. CD8, panc. CD8
NK cell: cyto. NK, adap. NK
B cell: naive B, mem. B
monocyte/MΦ: cMono / ncMono, panc. MΦ
dendritic: pDC, cDC
other cell: endothel., MKC, quies. / act. stell.

Chromosome 7 (Mb)

b

relative luciferase units

3.35×10^{-2}

non-risk (T) risk (C) empty vector

rs7795896

c

biotin probe
nuclear extract
competitor

specific binding

free biotin probe

non-risk (T) risk (C) probe allele

rs7795896

d

Control
$CFTR^{Enh}$
$CFTR^{Prom}$

CAPZA2 ST7 CFTR CTTNBP2 NAA38

1.26×10^{-2} ns ns 3.07×10^{-96} 2.41×10^{-3} ns ns ns ns

TPM

e

T1D GWAS
CFTR locus

GTEx v7 Pancreas
CFTR eQTL

PP_{shared} = 91.4%

log(Bayes Factor)

Chromosome 7 (Mb)

f

average expression

GCG, MKI67, INS, SST, PPY, REG1A, CFTR, PLVAP, COL1A1, RGS5, C1QC, CD3D, CDH19

alpha, div. alpha, beta, delta, gamma, acinar, ductal, endothelial, act. stell., qui. stell., macrophage, t-cell, schwann

g

CFTR expression (rs7795896)

pancreas

deconvoluted: ductal, acinar, delta, gamma, alpha, beta, stellate

$-\log_{10}$(P-value)

h

Phenotype	rs7795896	P-value
Type 1 diabetes		1.37×10^{-15}
Pancreatic diseases		4.72×10^{-5}
Chronic pancreatitis		3.18×10^{-3}
Acute pancreatitis		1.15×10^{-2}
Pancreatic cysts		1.73×10^{-2}
Pancreatic cancer		7.85×10^{-2}
Celiac disease		7.85×10^{-1}
Ulcerative colitis		6.31×10^{-1}
Rheumatoid arthritis		7.10×10^{-1}
Autoimmune vitiligo		5.62×10^{-1}

OR (95% CI)

i

Risk variant

Ductal cells: Enhancer activity, CFTR

Exocrine pancreas: Inflammation, Immune infiltration

Pancreatic diseases: Type 1 diabetes, Pancreatitis, Pancreatic cancer