bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Title: Single-cell gene expression and chromatin accessibility profiling of human pancreatic islets at basal and stimulatory conditions nominates mechanisms of type 1 diabetes genetic risk

Authors: Ricardo D'Oliveira Albanus[1†]*, Xuming Tang[2]*, Henry J. Taylor[3]*, Nandini Manickam[1], Michael Erdos[4], Narisu Narisu[3], Yuling Han[2], Peter Orchard[1], Arushi Varshney[1,5], Chengyang Liu[6], Ali Naji[6], HPAP Consortium[7], Francis S. Collins[3]**, Shuibing Chen[2]**, Stephen C. J. Parker[1,5,8]**

[1] Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[2] Department of Surgery, Weil Cornell Medicine, New York, NY, USA

[3] Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

[5] Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

[6] Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[7] The Human Pancreas Analysis Program (RRID:SCR_016202)

[8] Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[†] Current affiliation: Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO, USA

* Equally contributing authors

** Corresponding authors

Abstract

Type 1 diabetes (T1D) is a complex autoimmune disease characterized by the loss of pancreatic islet beta cells. The mechanisms of T1D genetic risk remain poorly understood. Here, we present a multi-omic integrative study of single-cell/nucleus molecular profiles of gene expression and chromatin accessibility in the same biological samples from healthy and beta-cell autoantibody$^+$ (AAB+) human pancreatic islets to characterize mechanisms of islet-mediated T1D genetic risk. We additionally performed single-cell/nucleus multi-omic profiling of healthy islets under two stimulatory conditions used as *in vitro* models of T1D (cytokine cocktail and CVB4 infection) to evaluate how environmental exposures recapitulate multi-omic signatures of T1D. In total, we analyzed 121,272 cells/nuclei across 34 libraries, identifying 10 distinct cell types. We identified cell-type-specific and disease-associated *cis*-regulatory elements and nominated likely target genes. We provide evidence that T1D genetic risk is mediated through multiple pancreatic cell populations, including islet endocrine cells (beta, alpha, gamma, and delta), exocrine acinar and ductal cells, and immune cells. Finally, we identified three independent T1D risk variants acting through pancreatic islet endocrine cells at the *TOX*, *RASGRP1*, and *DLK1/MEG3* loci. Together, this work improves our understanding of how non-coding genetic variants encode T1D risk through a complex interplay of different cell types in the pancreas.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

Introduction

Type 1 diabetes (T1D) is a complex autoimmune disease that accounts for 5-10% of all diagnosed diabetes cases (*1*). The primary manifestation of this disease is the targeting of endocrine beta cells by the immune system, likely mediated by T-cells, which leads to beta-cell loss and insulin deficiency (*2*). Advances in genotyping and imputation enabled increased power and accuracy for genome-wide association studies (GWAS) of T1D genetic risk (*3*, *4*). However, despite these substantial developments, the molecular mechanisms of T1D genetic risk are still poorly understood.

It is widely accepted that immune cells are the primary mediators of T1D genetic risk (*2*), which is supported by the strong genetic association of the major histocompatibility complex (MHC) in T1D GWAS (*3*, *5*). However, increasing evidence suggests that other cell types, including pancreatic islets, also contribute to T1D etiology and genetic risk (*3*, *4*, *6*). For example, one proposed mechanism for T1D risk variants acting through beta cells is to modulate their propensity for immune-mediated apoptosis (*7*). Two recent studies using functional genomics at the single-cell level helped clarify some of the biology driving T1D genetic risk and contributing to T1D progression (*4*, *8*). Both studies identified a role for non-immune cell types in the pancreas, particularly acinar and ductal cells, in mediating T1D genetic signals (*4*) or contributing to T1D onset and progression (*8*). In addition, one of these studies reported that *cis*-regulatory elements active in beta cells are significantly enriched to overlap T1D GWAS variants (*4*), indicating that beta cells mediate T1D genetic risk. Therefore, one crucial question that remains unanswered is how genetic variants acting through other pancreatic cell types, particularly beta cells, contribute to T1D onset and progression. Answering this question will be critical to help guide the development of novel T1D therapies.

Due to the scarcity of pancreatic tissue samples obtained from T1D donors and the limitation that disease progression leads to beta-cell destruction, several *in vitro* models of T1D using healthy pancreatic tissue have been developed to understand the early mechanisms of T1D in the pancreas. These models include treating primary islet cultures with a cytokine cocktail (TNF-α, IFN-γ and IL-1β) or infecting islets with Coxsackievirus B4 (CVB4) virus (*9*, *10*), which simulate the stressed environment beta cells are exposed to during T1D. However, the cell-specific molecular pathways underlying these experimental perturbations and to what extent these pathways mimic T1D have not been extensively characterized.

Here, we performed single-cell resolution multi-omic integration of high-throughput molecular profiles of paired gene expression and chromatin accessibility from the same biological samples obtained from healthy and T1D human pancreatic islets. We characterized mechanisms of T1D genetic risk, focusing on identifying variants acting through islet endocrine cells. In addition, we characterized two experimental models of T1D in islets to determine how they recapitulate the molecular aspects of T1D. Finally, we identify three independent T1D risk variants which likely mediate T1D genetic risk through islet endocrine cells. Our work identifies how all pancreatic cell populations partially mediate T1D genetic risk. Together, this work improves our understanding of how non-coding genetic variants encode T1D risk through a complex interplay of immune and pancreatic cell types.

Results

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

Identifying islet cell types by co-clustering gene expression and chromatin accessibility profiles

We performed gene expression (single-cell RNA-seq; scRNA-seq) and chromatin accessibility (single-nucleus ATAC-seq; snATAC-seq) on human pancreatic islets from healthy (n=8) and auto-antibody positive donors (AAB+; n=3). Given that procuring pancreatic tissue from affected donors is difficult, we aimed to investigate whether two established experimental models of T1D in human islets recapitulate AAB+ molecular profiles at the cell-specific epigenomic and transcriptomic levels. To this end, we additionally performed scRNA-seq and snATAC-seq on islets from a subset of healthy donors (n=3) under cytokine stimulation (TNF-α, IFN-γ and IL-1β) and CVB4 infection (**Figure 1A**, **Supplementary Table 1**). After stringent quality control (QC; Methods), we profiled 121,272 cells (49,897 snATAC-seq nuclei and 71,375 scRNA-seq cells; **Supplementary Figures 1 and 2**, **Supplementary Table 2**). We performed joint clustering of the molecular profiles across samples and modalities (n=34 libraries) using Seurat (*11*). We identified ten major distinct cell types based on the gene expression of known marker genes and the chromatin accessibility of their gene bodies (**Figure 1B-D**, **Supplementary Figure 2**). The identified cell types represent the endocrine (beta, alpha, delta, and gamma cells), exocrine (acinar and ductal), stellate (activated and quiescent), endothelial, and immune lineages. Cell type representation ranged from 1.4% (immune) to 35% (ductal) of all cells. We profiled 41,569 islet endocrine cells and nuclei, corresponding to 34.3% of all profiled cells and nuclei. Alpha cells were the most abundant endocrine cells (n=21,151), followed by beta (n=15,577), delta (n=2,703), and gamma cells (n=2,138). All cell types were well-represented across samples and modalities, and we did not identify any sample- or modality-specific clusters after QC (**Figure 1C, Supplementary Figure 2**). Importantly, we observed during the initial QC steps that the ambient RNA contamination (RNA “soup”) was a source of technical variation across libraries and could lead to misinterpretation of results if not correctly accounted for (Methods, **Supplementary Figure 3**). This is in line with a recent study indicating that ambient RNAs can confound single-cell analyses (*12*).

Transcriptional changes in experimental models of T1D recapitulate disrupted pathways in T1D

Aiming to identify pathways and regulatory programs associated with T1D, we first performed differential expression analyses across disease states and experimental perturbations. We accounted for biological and technical covariates that could influence results to quantify differential expression across conditions accurately. After adjusting for technical variation, we detected thousands of differentially expressed genes (DEGs) at 5% false discovery rate (FDR) across all cell types and conditions combined (ranging from 24 to 1,663 per cell type and condition, median = 476; **Figure 2A**). We observed the largest transcriptional changes associated with disease state (AAB+ vs. controls) relative to the perturbations (cytokines and CVB4) in the islet endocrine cells (beta, alpha, delta, and gamma), while the endothelial cells had stronger transcriptional changes under cytokine stimulation. On the other hand, the immune cells had the most comparable levels of transcriptional changes across disease state and experimental perturbations, consistent with immune cell types being highly responsive to environmental conditions. We observed lower transcriptional changes associated with CVB4

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

infection compared to cytokine stimulation in all cell types, which motivated us to investigate if CVB4 infection efficiency differed across samples. Indeed, we observed differences in the number of detectable CVB4 mRNAs in each CVB4-treated sample (**Supplementary Figure 4**). This variability may explain why the CVB4 infection DEG effect sizes were generally smaller. Together, these results are consistent with T1D inducing global changes in the pancreatic transcriptional landscape. However, these transcriptional changes are more pronounced in islet endocrine and immune cells compared to other pancreatic cells.

Aiming to better understand if the experimental perturbations recapitulated functional aspects of T1D in pancreatic cells, we performed pathway enrichment analyses using the DEGs from disease state and perturbations. The DEGs in AAB+ were generally not the same as the perturbations for most cell types (DEG $\log_2$FC Spearman’s ρ ranging across conditions from -0.12 to 0.88, median = 0.19; **Supplementary Figure 5**). However, we found overall high concordance between the pathway enrichments for nominally significant enrichments in AAB+ compared to cytokine stimulation and CVB4 infection in beta cells and other endocrine cells (**Figure 2B-D**, **Supplementary Figure 6-7**). These findings suggest cytokine stimulation and CVB4 infection affect similar pathways in beta cells compared to T1D, albeit regulating different genes within those pathways. Overall, the islet endocrine cells had the highest agreement between disease state and experimental perturbations at the level of pathway enrichments (**Figure 2D**). These results indicate that these experimental models recapitulate aspects of T1D in islet cells. However, these experimental models perturb different pathways than those associated with disease state at other pancreatic cell types. Therefore, these experimental models may not be the most suitable for studying T1D in cell types other than islets.

Transcription factors regulating the epigenomic landscape of pancreatic cells

To characterize the epigenomic landscape of the different pancreatic cell types, we used the BMO tool (*13*) to predict bound transcription factor (TF) sites using a non-redundant collection of 540 motifs and calculated their chromatin information patterns. The observed chromatin information patterns reflect the impact of specific TFs in organizing local chromatin architecture and establishing cell identity (*13*) (**Figure 3A-B**). We identified common and cell-type-specific TFs driving the epigenomic landscape for each cell type (**Figure 3C)**. The TFs CTCF, AP-1, and NFE2 consistently scored highest in chromatin information across cell types (**Supplementary Table 3)**, likely reflecting their constitutive roles in chromatin organization (*14*, *15*). On the other hand, a subset of TF families had a higher impact on chromatin organization in a cell-specific manner. These TF families include RFX in endocrine cells, HNF in exocrine cells, and SPI1 (PU.1) in immune cells (**Figure 3C**). All these TF families have been extensively characterized as cell fate determinants and play functional roles in their respective lineages (*16*–*18*) and, therefore, underscore the specificity of our epigenomic analyses. Importantly, we observed changes in the underlying chromatin organization associated with a subset of TFs when comparing conditions (**Figure 3D**). The IRF motif family was associated with increased chromatin organization in beta cells under cytokine treatment, consistent with previous studies showing that cytokines stimulation induces IRF-1 activation in beta cells and subsequent apoptosis (*19*, *20*). Similarly, cytokine treatment induced changes in chromatin organization at the SPI1, MAF, and ETS family TF motifs in immune cells, which are well-known mediators of cytokine response in these cells (*21*, *22*). Notably, the chromatin organization changes in AAB+

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

cells were less pronounced than the environmental perturbations. In agreement with the scRNA-seq results, these chromatin accessibility results indicate that the experimental models of T1D differ from disease in that they associate with more acute changes in cellular state.

Enrichment of T1D GWAS variants nominates cell types likely mediating T1D genetic risk

In order to investigate the mechanisms involved in T1D genetic risk, we used fGWAS (*23*) to calculate the enrichment of the accessible chromatin of the different cell types captured by our snATAC-seq experiments using the summary statistics of a recent T1D GWAS (*4*). As expected, we observed the highest T1D GWAS enrichment in the immune cluster (log enrichment = 2.78; **Figure 4A**). The other significantly enriched cell types were acinar, quiescent stellate, beta, ductal, and alpha (log enrichments ranging from 1.53 to 2.12). These results indicate that multiple pancreatic cell types, including islet endocrine cells, contribute to T1D genetic risk. These enrichments, however, likely represent the baseline (unperturbed) state of these cells and, therefore, provide an incomplete picture of T1D genetics. To contextualize these results, we tested the enrichment of accessible chromatin using the summary statistics of type 2 diabetes and fasting glucose from the DIAMANTE (*24*) and MAGIC (*25*) GWAS studies. We observed the strongest enrichments for these two traits in accessible chromatin regions from beta cells and other islet endocrine cell types (**Figure 4A**), which is consistent with previous studies (*26*, *24*, *27*, *25*).

We investigated the context-specific roles of the studied cell types in T1D predisposition. To this end, we used fGWAS to calculate the enrichment of T1D GWAS summary statistics in differentially accessible regions (DARs) across disease states and experimental perturbations. Because of data sparseness and inflation of *p*-values associated with differential analyses in single-cell data (*28*), we developed a stringent effect-size-based approach for detecting DARs in our snATAC-seq data (**Supplementary Figure 8**, Methods). As expected, DARs for AAB+ and cytokine treatment in immune cells were more highly enriched for T1D GWAS than non-DARs (**Figure 4B**). In addition, the enrichment point estimates increased as we used more stringent DAR thresholds. This result is consistent with a substantial component of T1D genetic risk encoded by responsive elements in immune cells, such as the MHC locus (*4*). We also observed a similar trend in DARs for CVB4 infection in immune cells, but it did not reach significance, likely due to the difference in CVB4 infection efficiency across replicates (**Supplementary Figure 4**). Interestingly, we found AAB+ DARs in beta cells more enriched for T1D GWAS than non-DARs. Similar to the previous results in immune cells, the enrichment point estimates for the beta-cell DARs increased with more stringent DAR thresholds (**Figure 4B**). This result indicates that the environmentally responsive regulatory elements in beta cells also mediate T1D genetic risk and, therefore, indicate a role for islet endocrine cells in mediating T1D progression.

Regulatory elements in beta and other islet endocrine cells mediate T1D genetic risk

Next, we aimed to understand regions and regulatory elements that are responsible for driving the observed T1D GWAS enrichments in pancreatic cells. To this end, we developed a novel approach to quantify the relative contributions of each cell type to T1D genetic risk and prioritize candidate cell types mediating genetic risk at a given locus. This approach is based on the cell-

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

type-specific chromatin accessibility levels at each variant in a T1D genetic credible set, weighted by the posterior probability of association (PPA) of the variant (Methods). As a proof of concept, the three independent GWAS signals at the *INS* locus were prioritized to act through beta cells (**Figure 4C**). A broader analysis of all 136 T1D GWAS signals showed that genetic risk is partitioned across all the cell types analyzed in this study (**Figure 4D**). Immune cells contribute to most of the T1D genetic risk, as expected. However, we observed multiple signals prioritized to act through pancreatic endocrine (beta, alpha, delta, gamma), exocrine (acinar, ductal), stellate, and endothelial cells. Importantly, we identified several signals with beta- or islet-specific accessibility, indicating that these genetic signals are mediated by islet endocrine cells in the pancreas. These islet endocrine loci include the three independent signals at the *INS* locus, the primary and secondary signals at *DLK1/MEG3*, and the signals at *TOX*, *RASGRP1*, and *GLIS3* (**Figure 4D**).

We next attempted to prioritize T1D risk loci likely acting through beta or other islet endocrine cells for functional validation. In addition to the PPA-weighted chromatin accessibility for each locus, we accounted for the number of variants in the 99% credible set (CS) and the PPA distribution across variants to nominate candidate loci where functional validation experiments were feasible. We prioritized loci with either a few variants in the 99% CS or loci where the PPA distribution was highly skewed towards a small number of variants. In addition, we used CICERO (*29*) to calculate co-accessibility between variant-harboring regulatory elements and gene promoters to help identify candidate target genes. To further reduce the search space for candidate variants, we performed functional fine-mapping (FFM) with fGWAS using a joint model accounting for the chromatin accessibility peaks from cell types enrichmed for T1D GWAS (Methods). Using these criteria, we nominated the main signals at *TOX* (99% CS size = 28) and *RASGRP1* (99% CS size = 66) and the secondary signal *DLK1/MEG3* (10 variants with PPA > 0.01; 99% CS size = 2,053) as the most compelling candidate loci likely acting through beta or islet endocrine cells (**Figures 5A-C**).

At the *TOX* locus, our FFM analyses prioritized rs367116 and rs1947178, with the latter being the lead variant at the locus. The intronic beta-cell regulatory element containing rs1947178 was co-accessible with the *TOX* promoter region (CICERO co-accessibility = 0.065), making *TOX* the candidate gene for this locus (**Figure 5A**). At the *RASGRP1* locus, FFM prioritized rs55728265, which is in strong linkage disequilibrium ($r^2 = 0.93$) with the lead variant, rs35134214. The regulatory element harboring rs55728265 overlaps the *RASGRP1* promoter region and was not co-accessible with any other promoter, making *RASGRP1* the candidate gene at this locus (**Figure 5B**). The lead variant at this locus (rs35134214) did not overlap ATAC-seq peaks in pancreatic cell types, therefore highlighting the validity of using FFM approaches to prioritize genetic signals. At the *DLK1/MEG3* locus, our FFM analyses prioritized the lead variant for the primary signal (rs56994090), despite this variant not overlapping any features used in the FFM model (**Figure 5C**). We also prioritized the primary variant at the secondary signal at *DLK1/MEG3* (rs3783355; PPA = 0.56) because it had a 7-fold higher PPA compared to the second highest variant in the 99% CS (rs10145648; PPA = 0.08) and overlapped a highly accessible chromatin region in beta, alpha, and ductal cells. Interestingly, we observed increased co-accessibility between the regulatory element harboring rs3783355 and the *DLK1* and *MEG3* promoter regions in AAB+ and cytokine-stimulated beta cells

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

compared to healthy beta cells (*MEG3*–rs3783355 CICERO score = 0.013 for cytokine; *DLK1*–rs3783355 CICERO scores 0.002 and 0.144 for healthy and cytokine, respectively). These results suggest that the regulatory element harboring rs3783355 acts in a context-dependent manner to mediate T1D risk in pancreatic islet endocrine cells.

T1D risk variants are predicted to disrupt islet endocrine cells regulatory elements

We next attempted to characterize the functional mechanisms through which the variants of interest at the *TOX*, *RASGRP1*, and *DLK1/MEG3* loci act to mediate T1D risk. We aimed to characterize the impact of the risk and non-risk alleles associated with these variants. Because we had genotype information for 10 of the donors, we calculated the cell type-specific ATAC-seq allelic bias at each heterozygous SNP with enough coverage (**Supplementary Figure 9A-B**). In parallel, we trained a predictive model of sequence features associated with chromatin accessibility in beta cells using LS-GKM and DeltaSVM (*30*, *31*) to predict beta-cell allelic effects associated with any base-pair change in the genome (Methods; **Supplementary Figure 9C-D**). We used the observed allelic bias to validate our predictive model. The predicted allelic effects from the model were highly concordant (87.1% effect size direction agreement) with the observed allelic effects (ATAC-seq allelic bias) at heterozygous SNPs, indicating that the model correctly captured allelic regulatory changes associated with increased chromatin accessibility in beta cells (**Figure 5D**). The predictions from the model trained in beta cells had a higher agreement with the observed allelic effects calculated using the entire dataset (92.6% effect size direction agreement), which we attribute to increased power when combining data across all cell types. Alternatively, this also can be interpreted as the model trained in beta cells also capturing sequence features associated with chromatin accessibility more broadly.

To further gain information from our predictive model, we applied GkmExplain (*32*) to the variants of interest and predicted the regulatory effects associated with each allele within the entire sequence context around the variants (**Figure 5E**). At the *TOX* locus, the risk allele at the lead variant, rs1947178 (risk = A; non-risk = G), was predicted to increase chromatin accessibility. The predicted impact for the risk allele at rs1947178 was also higher than that of the FFM-nominated SNP, rs367116 (risk = C; non-risk = T). At the *RASGRP1* locus, the lead variant, rs35134214 (risk = CTG; non-risk = C), was predicted to increase accessibility. Conversely, the *RASGRP1* FFM-nominated SNP, rs55728265 (risk = T; non-risk = C), was predicted to decrease accessibility. While we did not observe any ATAC-seq peaks at rs35134214, we cannot discard that this variant mediates T1D genetic risk through other cell types not assayed in this study. Finally, at the *DLK1/MEG3* locus, we predicted stronger effects in chromatin accessibility associated with the risk allele at the secondary signal lead variant, rs3783355 (risk = G; non-risk = A) compared to the lead variant at the primary signal (rs56994090; risk = T, non-risk = C). Consistent with the predicted effects in dysregulating chromatin accessibility, we identified multiple predicted bound TF motifs overlapping these risk variants, including PAX4 and HNF4 (*RASGRP1*), ITGB2, and ZBTB6 (*DLK1/MEG3*), and CPHX (*TOX*) (**Supplementary Table 4**). Together, these results implicate rs1947178 (*TOX*), rs55728265 (*RASGRP1*), and rs3783355 (*DLK1/MEG3*) as likely causal variants mediating T1D genetic risk through islet cell types.

Discussion

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

After decades of research, T1D genetic risk is widely accepted to be driven by variants disrupting the endogenous pathways that inhibit self-reactivity, which in turn increase autoimmune responses (*1*, *2*). We have integrated epigenomic and transcriptomic profiles of human pancreas samples from healthy and AAB+ donors to better understand how T1D risk variants act across the different cell types in the pancreas and lead to changes in gene regulation. Rather than being mediated by one or a few cell types, we find that T1D genetic risk variants overlap active regulatory elements in every pancreatic cell type analyzed in this study. Our findings are consistent with the increasing evidence linking non-immune cells to mediating T1D risk (*3*, *4*, *6*). In particular, our work identifies three genes expressed in beta cells and other islet cell types as putative causal genes for three independent T1D risk variants: *DLK1/MEG3*, *TOX*, and *RASGRP*. Our prioritization of the *DLK1/MEG3* and *TOX* loci as mediated through islet endocrine cells is supported by a previous scATAC-seq study, which observed a higher overlap of high-PPA variants in these loci with beta-cell regulatory elements (*4*). Our work expands on these findings by predicting rs1947178 and rs3783355 as causal variants at these loci and further prioritizing rs55728265 at the *RASGRP1* locus as an additional variant mediating T1D genetic risk through islet endocrine cells.

While the role of immune cells mediating T1D genetic risk is generally understood, it is still unclear how other pancreatic cell types contribute to T1D risk. One hypothesis is that risk variants at these other cell types lead to disease predisposition by promoting the recruitment of self-reactive T-cells or creating a harsher cellular microenvironment that further predisposes beta-cell death. Support for this hypothesis is provided by a previous snRNA-seq study from healthy, AAB+, and T1D human pancreas, which suggested that T1D ductal cells may help promote CD4$^+$ T cell tolerance through the expression of MHC molecules and other surface receptors (*8*). Our work indicates that the immune cells indeed have the highest individual contribution to T1D genetic risk. However, this contribution is relatively small compared to all the other cell types combined. In addition to multiple variants acting through islet endocrine cells, we identified a role for acinar, stellate, endothelial, and to a lesser degree, ductal cells as likely mediators of T1D genetic risk. This unexpected finding agrees with and expands on other studies of T1D at the single-cell level identifying the contributions of other pancreatic cell types to T1D genetic risk and onset (*4*, *8*). Therefore, an important question for future studies is understanding how T1D risk variants act through non-immune cell types, particularly beta cells.

Among the active areas of T1D research is developing experimental models to understand disease biology using healthy islets. In this study, we characterized the molecular profiles of healthy islets challenged with cytokine stimulation or CVB4 infection. These experimental models inherently disturb healthy islets in a time window several orders of magnitude smaller than the disease duration (hours vs. years). However, we found similarities in the transcriptomic and epigenomic profiles associated with these experimental perturbations. Furthermore, we observe that these experimental models most strongly perturb different genes compared to T1D. However, these perturbed genes participate in several of the same pathways observed in islets from affected donors, which supports the use of these experimental models to understand T1D biology. Our results suggest that while general agreement exists between the downstream pathways, some experimental models may be more appropriate for studying specific aspects of T1D biology (*e.g.,* cytokines triggered differentiation pathways and CVB4 infection triggered

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

more stress responses). Therefore, more in-depth studies are required to explore the full gamut of protocols associated with these T1D experimental models, such as different cytokine combinations, to determine the most appropriate experimental approach to model specific aspects of T1D biology.

Among the limitations of this study is that we jointly analyze pre-diabetic (AAB+ without symptomatic presentation) and diabetic donors due to the low sample size. While our results suggest that this is a valid approach to detecting disease-relevant biology, this design would miss molecular signatures associated with different stages of the disease. In particular, one can hypothesize that the beta cells that survive in T1D donors are transcriptionally different from the beta cells from the pre-diabetic donors and develop molecular characteristics to make them more resistant to immune targeting. Therefore, separately studying beta cells from T1D donors is an important future direction that can provide essential clues for new therapeutic strategies.

Acknowledgments

We are especially grateful to the donors and their families for making this work possible. This work was funded by the NIH (5U01DK127777 to S.C. and S.C.J.P.; 1-ZIA-HG000024 to F.S.C.). Human Pancreas Analysis Program for Type 1 Diabetes (2U01DK112217-02A1). We thank the genomics resources core facility at Weill Cornell Medicine and NHGRI Genomic Core for genotyping SNP array of the samples. Islet tissue samples were obtained through Human Pancreas Analysis Program (HPAP). Pancreas and islets illustrations were modified from Servier Medical Art and obtained through Bioicons (https://bioicons.com).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

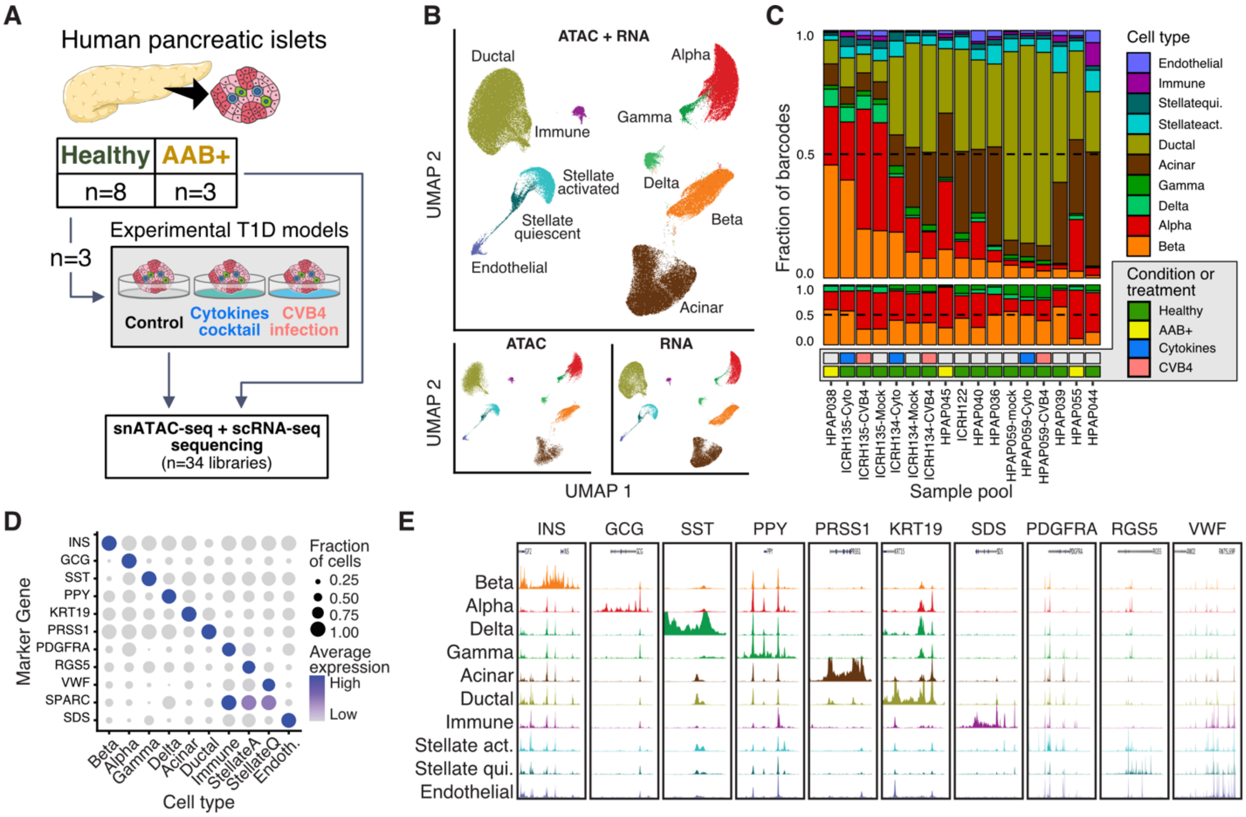


Figure 1: Study overview. A) Experimental design for multi-omic library generation. **B**) Uniform Manifold Approximation and Projection (UMAP) representation of the fully integrated dataset. Bottom panel is the same data faceted by modality. **C**) Overview of the representation of all cell types (top), islet endocrine cell types (middle), and conditions (bottom) across the combined scRNA-seq and snATAC-seq libraries for each sample pool. **D**) scRNA average expression values for marker genes across the cell types identified via joint modality clustering. **E**) Normalized aggregate ATAC-seq signal tracks across marker genes for each cell type.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

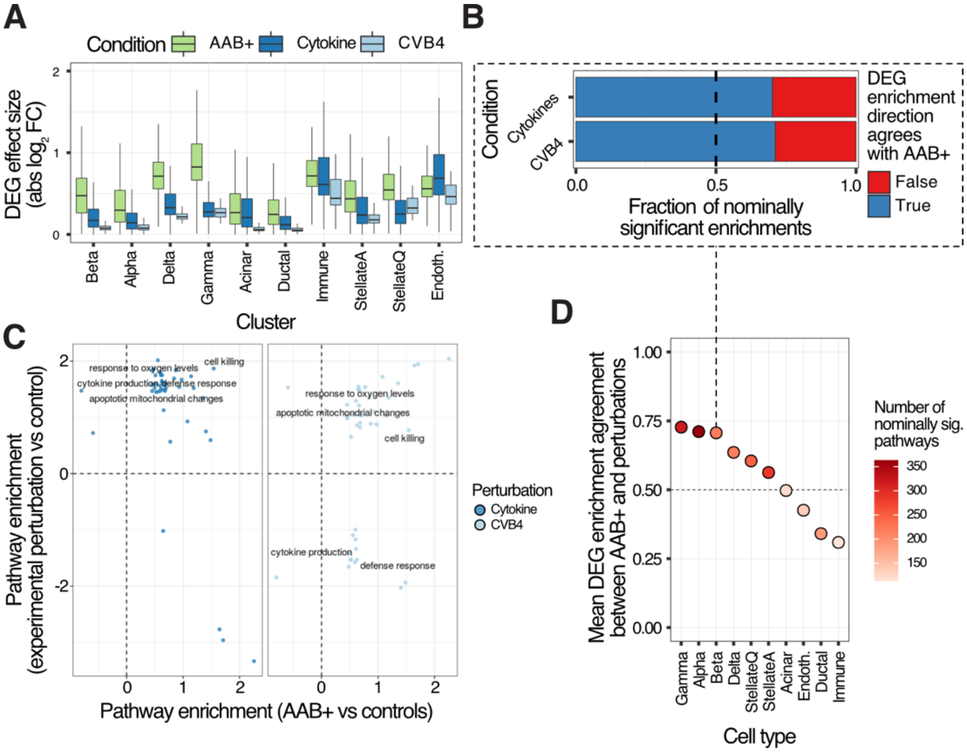


Figure 2: Transcriptomic changes associated with T1D and experimental models. A) Differentially expressed gene (DEG) effect size comparison across cell types and conditions. **B**) Beta cells DEG pathway enrichment effect size direction agreements between experimental models of T1D and AAB+ cells. **C**) Significantly enriched pathways across AAB+ and experimental models (summary of significant terms using rrvigo). **D**) Average pathway effect size direction agreement per cell type between AAB+ and experimental models for nominally significant terms in at least one condition.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

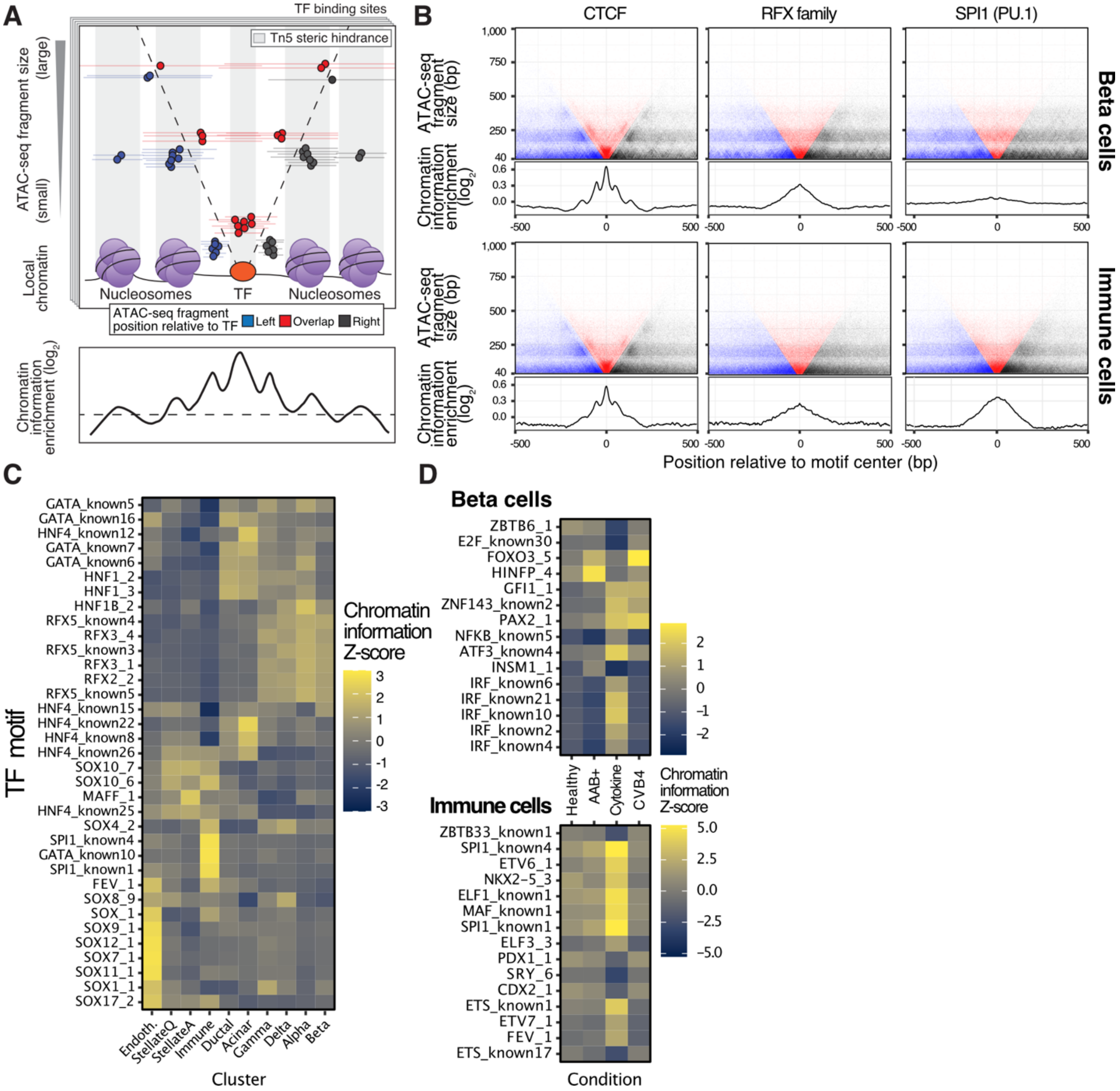


Figure 3: TF regulatory landscape of pancreatic cell types. A) Chromatin information enrichment calculation overview (adapted from (*13*)). **B**) V-plots showing aggregate ATAC-seq fragment midpoints distribution around predicted bound sites for three TFs (top facets) and their associated chromatin information enrichment (bottom facets) in beta cells and immune cells. **C**) Chromatin information Z-scores for a subset of TFs across all cell types indicate differential regulatory activity. **D**) Similar to C, but directly comparing across conditions for beta cells (top) and immune cells (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

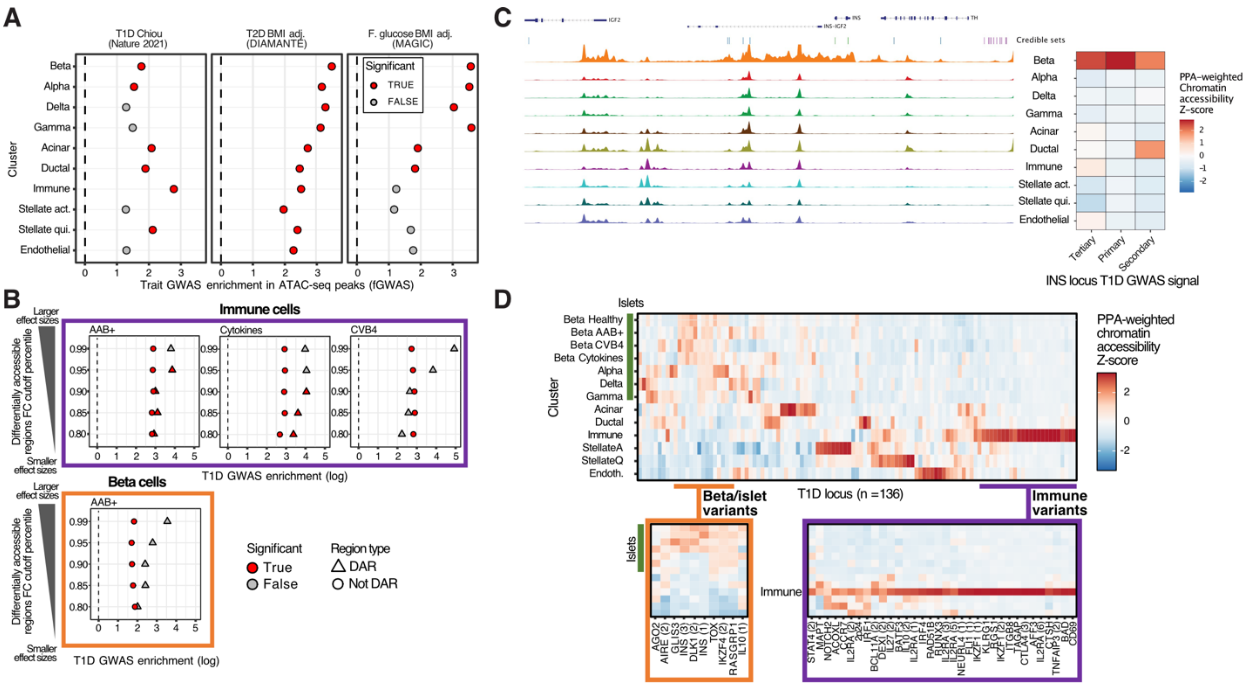


Figure 4: The regulatory landscape associated with T1D genetics in pancreatic cells. A) fGWAS enrichments for GWAS summary statistics of three traits in accessible chromatin regions from each cell type in our data. **B**) fGWAS enrichments for T1D summary statistics in immune and beta cells across progressively stringent thresholds to identify differentially accessible regions (DARs) and their non-significant counterparts. **C**) Example of our PPA-weighted chromatin accessibility score strategy to identify cell types likely mediating three independent T1D GWAS signals at the INS locus. **D**) PPA-weighted chromatin accessibility scores across all T1D loci and cell types and candidate loci likely mediated by islet and immune cell types.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

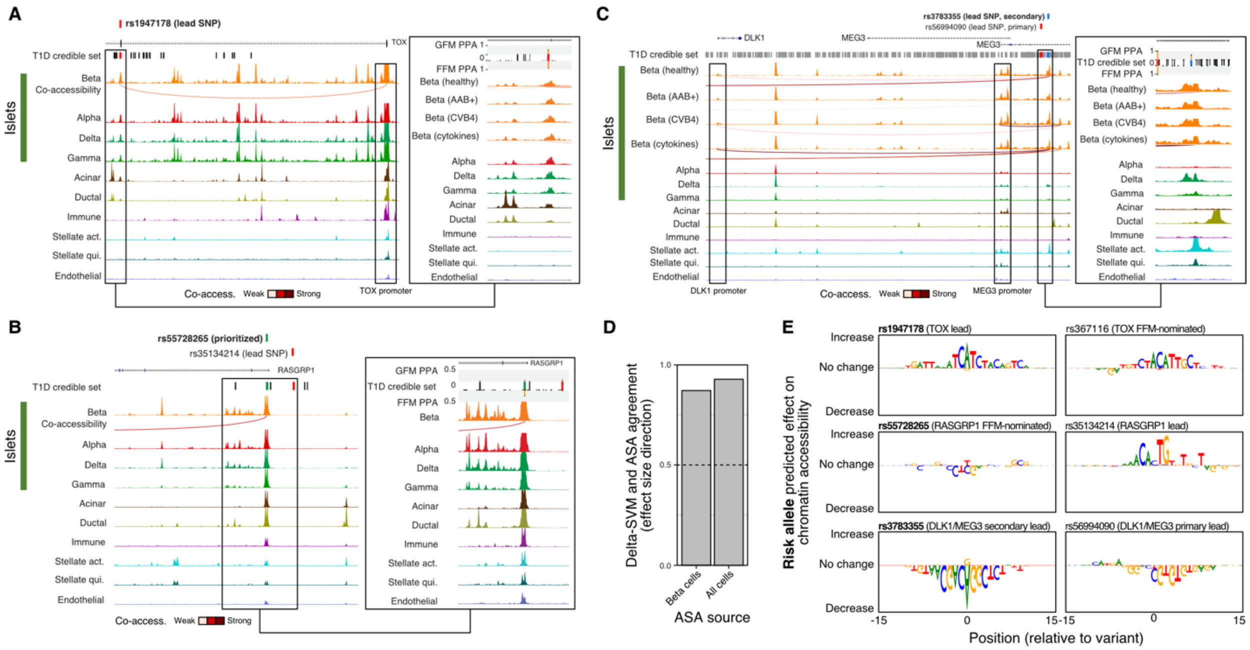


Figure 5: Genetic variants mediating T1D risk in islet endocrine cells. T1D signals at the *TOX* (**A**), *DLK1/MEG3* (**B**), and *RASGRP1* (**C**) loci. Left panels represent the broad locus overview, and the insets highlight the regions and variants of interest and their associated genetic and functional fine-mapping PPA values. For simplicity, only beta-cell co-accessibility tracks are shown. **D**) Agreement between predicted and observed ATAC-seq allelic imbalance (allele-specific accessibility; ASA) in beta cells and all cells using a predictive model trained in beta cells. **E**) Predicted regulatory impact of T1D risk variants of interest in beta cell chromatin accessibility using GkmExplain.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

Methods

Tissue processing and sample preparation

Human pancreatic islets were isolated in the Human Islet Core at the University of Pennsylvania following the requirements of the Clinical Islet Transplantation consortium procedure. The pancreatic islets were grown in CIT culture medium and maintained in a humidified incubator with 5% CO2 at 37℃. Single-cell RNA-seq and single-nucleus ATAC-seq were performed using 10X Chromium platform at genomics resources core facility at Weill Cornell Medicine.

Single-nucleus ATAC-seq processing

Single-nucleus ATAC-seq data was processed using the Parker Lab snATAC-seq pipeline (https://github.com/porchard/snATACseq-NextFlow). Briefly, after performing adapter trimming with cta (v. 0.12; https://github.com/ParkerLab/cta), reads were aligned to the hg19 reference genome using bwa mem (v. 0.7.15-r1140; (*33*)) using *-I 200,200,5000* to avoid large fragments being artificially assigned low MAPQ scores. Barcode sequences were corrected for sequence mismatches by calculating the Hamming distance between all barcodes and fixing all barcodes with a Hamming distance smaller or equal to 2 to a barcode sequence in the 10X Genomics barcode list. After mapping, we identified barcodes using Picard MarkDuplicates (v. 2.8.1; https://broadinstitute.github.io/picard). We used ataqv (https://github.com/ParkerLab/ataqv (*34*)) to obtain barcode-level QC metrics, such as the number of high-quality autosomal alignments (HQAA) and transcription start site (TSS) enrichment. For downstream analyses, we retained only barcodes with HQAA ≥ 5,000, TSS enrichment between 3 and 20, and no more than 15% of all reads originating from a single autosome. The latter metric helps to remove barcodes associated with low-integrity nuclei. Doublets were flagged and removed using ArchR (v. 0.9.5) (*35*). Because the ambient signal (soup) from the snATAC-seq library is mainly from chrM, which was filtered for our analyses, we did not perform ambient DNA correction. For integration with the scRNA-seq data (described below), we generated count matrices for each library encoding the number of ATAC-seq fragments overlapping promoter (5 Kb upstream of most upstream transcription start site) and gene body regions of autosomal, protein-coding genes using bedtools (v2.26.0).

Single-cell RNA-seq

Single-cell RNA-seq data were processed with the Parker Lab snRNA-seq pipeline (https://github.com/porchard/snRNAseq-NextFlow). Reads were aligned to the hg19 reference genome and GENCODE v19 (*36*) using STARsolo (STAR v. 2.5.4 (*37*)). Barcode sequences were corrected for mismatches using the same approach as in the snATAC-seq data. We then calculated QC metrics for each barcode (number of UMIs, % mitochondrial reads, etc.). We selected for downstream analyses barcodes that had at least 1,000 UMIs and were called non-empty (1% FDR) by EmptyDrops (*38*). For each library, we calculated the % mitochondrial reads rank distribution and identified the inflection (knee) using the uik function of the inflection package in R (*39*). We only kept barcodes with % mitochondrial reads smaller than the inflection value, ranging from 6.6% to 20.2%. Doublets were flagged and removed using DoubletFinder (v2.0.2) (*40*) with default parameters. After removing doublets and barcodes that failed QC, we

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

used DecontX (Celda v1.2.4) (*41*) to control for ambient RNA (soup RNAs). We performed a first-pass clustering of the barcodes that passed QC using Seurat (**Supplementary Figure 1**) to identify broad cell identities. We then used the first-pass clustering information with DecontX with stringent parameters (delta 1 = 10 and delta 2 = 20) to obtain the ambient-subtracted count matrices for each library. We used the ambient-subtracted count matrices of autosomal, protein-coding genes for downstream analyses.

Sample genotyping

Samples were genotyped using the Illumina Infinium 2.5M exome chip (InfiniumOmni2-5Exome-8v1.3_A2). The genotyping call rates for the 16 samples ranged from 99.0% to 99.7%. The SNP probe sequences were remapped to GRCh37 and all problematic SNPs were discarded. This process resulted in a total of 2,522,105 SNPs with genotypes. Next, SNPs that have genotype missingness in >=2 out of our samples and duplicate SNPs with the same genomic coordinates with another one were removed. Further, we merged our genotypes with that of the 1000G phase 3v5 samples (*42*). Subsequently, the SNPs with HWE p-value < 1e-4, and palindromic SNPs (A/T, or G/C SNPs) with MAF>0.4 in the merged data set were removed. Phasing was performed on the joint data set of 1,609,033 SNPs using Eagle (v2.4) (*43*). Genotypes were imputed using 1000 genomes phase 3 panel in the Michigan Imputation Server using Minimac4 (v1.5.7) (*44*) and the 1000G phase 3v5 (GRCh37) reference panel. No sex discrepancy was found by assessing the SNP genotypes using verifybamID (*45*) with the reported gender. Sample ICRH135 did not have sufficient DNA for genotyping and was dropped from the genetic analyses.

CVB4-hg19 alignments

In order to quantify CVB4 infection efficiency, we aligned scRNA-seq and snATAC-seq reads to a hybrid hg19-CVB4 genome, where the CVB4 genome (GenBank AF311939.1) is appended to hg19 as a separate chromosome. Similarly, we built a hybrid GTF file with the human genes and the CVB4 genome as an additional gene. We generated STAR and bwa indices for the hybrid hg19-cvb4 genome and mapped reads using the same pipeline described below. To quantify the CVB4 infection efficiency, we counted the fraction of reads mapping to the CVB4 portion of the hybrid genome. To independently confirm that our pipeline worked as expected, we used SANDY (https://github.com/galantelab/sandy) to generate hybrid paired-end reads from both genomes using the command *sandy genome* with flag *--id=" %i.%U__read=%c:%t-%n__mate=%c:%T-%N__length=%r"* and verified that the snATAC-seq and scRNA-seq pipelines aligned these simulated reads to the correct coordinates on both assemblies.

Cross-modality integration of snATAC-seq and scRNA-seq profiles

In order to integrate all 34 libraries, we used Seurat (v.4.0.3)(*11*). After exhaustively testing different pipelines, we obtained the best results for this dataset using Seurat's standard workflow. After running the principal component analysis (PCA) step, we extracted the first 30 PC embeddings for each barcode and calculated the Spearman correlation with technical variables (sequencing depth, % mitochondrial reads, etc.) to identify PCs driven by technical aspects. We used PCs 1,3-30 for the FindNeighbors and RunUMAP steps because PC 2 was

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

correlated with sequencing depth. We used options resolution=1, algorithm=2, n.start=1000, and n.iter=1000 for FindClusters and parameters n.neighbors=50 and n.epochs=500 for RunUMAP. This approach yielded 30 clusters in the integrated data. We next identified and removed clusters that could not be unambiguously assigned to any cell type (*i.e.,* loaded on more than one cell-type-specific marker) or had aberrant QC metrics. After filtering these low-quality barcodes, we iteratively merged the remaining clusters based on similar gene express/accessibility patterns to obtain the final cluster assignments used in this study. A subset of the snATAC barcodes assigned to the UMAP region corresponding to the acinar cells could not be unambiguously classified as acinar cells and was removed. This resulted in a higher fraction of scRNA-seq barcodes in the acinar cluster compared to the other clusters. Despite the relatively smaller fraction of acinar snATAC-seq barcodes, the number of barcodes was still higher than most clusters and, therefore, did not substantially affect our chromatin accessibility analyses for the acinar cells.

Peak calling

We generated BAM files for each cluster by combining data from all barcodes in that cluster (pseudo-bulk analyses). We also generated BAM files for each cluster/library combination. We used MACS2 (v. 2.1.1.20160309) to call summits on each cluster bam file, and we extended each summit by 150 bp in both directions. The set of extended summits called on the cluster-level bam file (all libraries combined) was labeled as the primary summit list. We assessed the reproducibility of each extended summit in the primary list using bedtools intersect (v2.26.0) to count the number of intersections in the per-library extended summits. We retained for downstream analyses the extended summits from the primary list that 1) overlapped extended summits from at least two different libraries and 2) did not overlap any regions with known mappability issues.

Differential gene expression analyses

For each cell type, we tested for association of gene expression with AAB+ status (*i.e.*, T1D or pre-T1D) using MAST v1.14.0 (*46*). We filtered lowly expressed genes (DecontX-corrected counts ≥ 1 in ≤ 5 cells across all samples and cell types) using the pp.filter_genes function with min_cells=5 from scanpy v1.5.1 (*47*), retaining 16,844 genes. To account for variable sequencing depth across cells, we normalized the DecontX-corrected counts for the remaining genes by the total number of counts per cell, scaled to counts per 10,000 (CP10K; pp.normalise_per_cell function in scanpy), and log-transformed the CP10K expression matrix (ln[CP10K+1]; scanpy's pp.log1p function). Using the ln[CP10K+1] normalized counts as input, we modeled the gene expression for each cell type using MAST's zlm function with default parameters. We included disease status, donor ID, sex, age, body mass index (BMI), and proportion of donor cells identified as alpha cells (which is a proxy of islet content and accounts for any differences in background RNA persisting after DecontX correction; **Supplementary Figure 3**), and cell complexity (the number of genes detected per cell (*46*, *48*)) as fixed effect covariates. Age, BMI, and alpha cell proportion were standardized to unit variance (mean-centered and scaled). For each model, we performed the likelihood ratio test (LRT; implemented in MAST's summary function with logFC=TRUE and doLRT=T1D status) to test for association between gene expression and disease status. Finally, we controlled for the number of tests

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

performed across all cell types using the Benjamini-Hochberg procedure (*49*) and LRT-derived p-values.

Gene set enrichment

We tested for gene sets enriched in the differential expression results for each cell type using the fgseaMultilevel function from fGSEA v1.16.0 (*50*) with eps=1x10^{-10}, scoreType='std', and the rest as default parameters. We used z-scores derived from the $\log_2$ FCs as implemented in MAST to pre-rank the genes. We tested gene sets found in the following databases, which were downloaded via the molecular signatures database (MSigDB) v7.2 (*51*, *52*) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (*53*), BioCarta pathways (*54*), and Gene Ontology (GO) biological processes (August 2020 release) (*55*). We controlled for the number of tests performed per cell type using a Bonferroni correction. To simplify GO terms in visualizations, we used rrvigo (https://ssayols.github.io/rrvgo).

Transcription factor binding prediction and chromatin information analyses

We used BMO and our previously described chromatin information analysis pipeline (*13*) available at https://github.com/ParkerLab/BMO/tree/pre-1.1 to predict bound TF motifs and estimate the impact of TFs in their local chromatin architecture. Briefly, we used the hg19 motif scans from a non-redundant position weight matrices collection corresponding to 540 TF motifs (described in (*13*)). For each cell type pseudo-bulk snATAC-seq BAM file, we calculated the distribution of ATAC-seq fragments overlapping each TF motif instance and the number of co-occurring motifs from the same TF motif within 100 bp to use as input for BMO. BMO predicts TFs using a simple premise that highly accessible motif clusters will be more likely bound by TFs, as the vast majority of TFs cannot induce open chromatin based on DNA sequence alone (*13*). BMO fits two negative binomial distributions for the ATAC-seq signal and the number of co-occurring motifs per motif instance and calculates the probability of a given motif instance being bound based on the combined p-value for these two distributions.

Chromatin information for each TF motif was estimated using the feature V-plot information content enrichment (f-VICE) score described in our previous study (*13*). Briefly, we generated V-plots (aggregate ATAC-seq fragment midpoint distributions around TF binding sites) for non-overlapping (within 500 bp) BMO-predicted bound instances of a given TF motif (**Figure 3B**, top plots). We then calculated the chromatin information (f-VICE score) for each motif by quantifying the $\log_2$ information content enrichment at TF-adjacent (-25 to +25 from motif) and TF-proximal (-70 to -50 and 50 to 70 bp from motif) regions compared to a randomly shuffled ATAC-seq midpoint distribution (**Figure 3B**, bottom signal tracks). These regions are expected to have high information content when the TF induces nucleosome phasing. We then normalized f-VICE scores for each cell type by calculating the residuals of the linear model f-VICE ~ $\log_{10}$(total fragments) + $\log_{10}$(total co-occurring motifs), which controls for the abundance and overall accessibility of the predicted bound instances for each TF motif.

In order to compare chromatin information across conditions (**Figure 3D**), we calculated the f-VICE scores separately for the pseudo-bulk snATAC-seq BAM files obtained from each cell type and donor combination (*i.e.,* Donor 1 beta cells, Donor 2 beta cells, etc.). First, we calculated f-VICEs separately per donor and cell type to avoid confounding by the different number of nuclei.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

We then converted each donor and cell type normalized f-VICE distribution into Z-scores. Finally, we calculated the median Z-score for each TF motif to obtain a single value for a TF motif per condition and cell type. For visualizing this data in **Figures 3C-D** heatmaps, we calculated row-wise (per motif) Z-scores.

Differential accessibility analyses

We used DESeq2 (1.3.2) to perform differential accessibility analyses. We used as input the pseudo-bulk counts from each library for the reproducible extended summits called on each cluster. For the AAB+ versus healthy comparisons, we controlled for age, sex, BMI, median TSS enrichment, and $\log_{10}$(HQAA). We scaled and centered age and BMI. For the CVB4 and cytokine versus control comparisons, we opted for a paired design that accounted for donor ID and median TSS enrichment per library, but not age and BMI due to collinearity. Because of statistical instability observed in single-cell approaches for differential analyses in this dataset, we designed an alternative approach to calculate significance based on effect sizes. For each comparison, we removed features with a mean number of reads < 3 and divided the remaining features into 50 equally spaced bins of mean chromatin accessibility using the chop_evenly function from the Santoku R package (https://github.com/hughjonesd/santoku). We removed regions with $\log_2$ fold-change > 10, as these likely represented technical artifacts from low ATAC-seq coverage. For each of the 50 chromatin accessibility bins, we identified the features in the 80th, 85th, 90th, 95th, and 99th percentiles of absolute $\log_2$ fold-change, which were used for the fGWAS enrichments described below. A summary of this approach is included in **Supplementary Figure 8**.

Co-accessibility analyses

Co-accessibility between accessible regions were calculated for each cell type separately by condition using CICERO (*29*) with default parameters. We generated count matrices for each pseudo-bulk BAM file representing a cell type and condition (e.g. healthy beta cells) for the accessible regions of that cell type (reproducible extended summits, described above). We used as input for CICERO the count matrix and the corresponding UMAP coordinates of each barcode. We annotated the resulting connections based on whether each connected peak overlapped a T1D credible set SNP or a gene TSS from GENCODE V19.

GWAS enrichments and functional fine-mapping using fGWAS

We calculated GWAS enrichments in features of interest using fGWAS (commit 0b6533d) (*23*). For the GWAS enrichments of the accessible regions per cluster, we ran fGWAS with the *-print* flag using as input the summary statistics from each GWAS study and a reproducible list of extended summits per cluster. For the DARs T1D GWAS enrichments, we used similar steps as above. However, instead of splitting the genome into windows of 5,000 variants based on their order of occurrence (fGWAS default), we generated a bed file of custom 5,000 variant windows where the window corresponding to each T1D loci was centered on the lead variant of the primary signal using the flag *-bed*. The remaining genomic windows were either left unchanged or shortened in case they overlapped a T1D locus chunk. This step was necessary due to the sparseness of the genomic territory covered by DARs. For the functional fine-mapping, we

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

assigned a 0 or 1 value for each T1D variant encoding whether they overlapped a reproducible extended summit in each cell type. We ran fGWAS using the option *-fine* and including all clusters with significant enrichment in the T1D GWAS.

PPA-weighted chromatin accessibility Z-scores

To identify which cell types likely mediate T1D genetic risk in each locus, we developed an approach based on the chromatin accessibility for each cell type at the locus. First, we extended each variant in the genetic fine-mapping credible sets (calculated by Chiou *et al.*) by 50 bp in each direction. Next, we counted how many snATAC-seq reads overlapped the extended variant region in the pseudo-bulk data from each cell type. We then normalized the snATAC-seq signal by the sequencing depth and multiplied it by the genetic fine-mapping PPA. When two or more variants overlapped in the extended region, we calculated the ATAC-seq signal for the merged region and used the highest PPA. We retained for analysis only loci where at least one credible set variant overlapped a reproducible (minimum of 2 samples) ATAC-seq broad peak. We then summed each locus's PPA-weighted chromatin accessibility values to obtain a single score per cell type. Finally, we applied a Z-score transformation for each locus across cell types.

GWAS variants regulatory impact prediction

We used LS-GKM (*30*) to train a predictive model of 11-mers for each cell type using as positive regions the extended summits. We used the genNullSeqs function from the gkmSVM R package (*56*) to obtain the negative set of GC- and repeat-content matched regions per cell type. To predict the regulatory impact of the SNPs of interest, we used GkmExplain (*32*) using as input the ±25 bp flanking each allele and calculated the predicted importance scores for each base. In order to validate the LS-GKM model, we separately calculated the ATAC-seq allelic imbalance at heterozygous SNPs and compared it to the Delta-SVM scores for each allele. Using the genotype data from each donor, we used WASP (v. 0.2.1, commit 5a52185; python version 2.7)(*57*) to diminish reference bias using the same mapping and filtering parameters described for the initial mapping and filtering. Duplicates were removed using WASP's rmdup_pe.py script. To avoid double-counting alleles, overlapping read pairs were clipped using bamUtil clipOverlap (v. 1.0.14; http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap). We counted the number of reads containing each allele for each heterozygous autosomal SNP, using only bases with a base quality of at least 10. We further split each donor's BAM file per cell type to calculate allelic imbalance per cell type separately and for the entire library. We used a two-tailed binomial test that accounted for reference allele bias to evaluate the significance of the allelic bias at each SNP. The observed allelic bias was then correlated with the Delta-SVM score, which was obtained by scoring the 11-mers centered on the REF and ALT alleles for the 1,000 Genomes (Phase 3). We used all SNPs with an absolute Delta-SVM score ≥ 2 to compare with the observed allelic imbalance.

Genome visualizations

We used pyGenomeTracks (version 3.7) (*58*) to generate genome visualizations of snATAC-seq signals, co-accessible regions, and GWAS variants.

GWAS data

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

T1D summary statistics were downloaded from the EBI Catalog (accession number GCST90012879)

Data availability

All data will be deposited in GEO upon publication.

Code availability

All code used for this manuscript is publicly available at (http://github.com/ParkerLab/albanus_2020_nih_islets_sn_t1d). We use snakemake (*59*) to facilitate reproducibility.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

References

1. M. Mobasseri, M. Shirmohammadi, T. Amiri, N. Vahed, H. Hosseini Fard, M. Ghojazadeh, Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. *Health Promot. Perspect.* **10**, 98–115 (2020).

2. L. A. DiMeglio, C. Evans-Molina, R. A. Oram, Type 1 diabetes. *Lancet Lond. Engl.* **391**, 2449–2462 (2018).

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

4. J. Chiou, R. J. Geusz, M.-L. Okino, J. Y. Han, M. Miller, R. Melton, E. Beebe, P. Benaglio, S. Huang, K. Korgaonkar, S. Heller, A. Kleger, S. Preissl, D. U. Gorkin, M. Sander, K. J. Gaulton, Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature*. **594**, 398–402 (2021).

5. J. A. Noble, A. M. Valdes, Genetics of the HLA region in the prediction of type 1 diabetes. *Curr. Diab. Rep.* **11**, 533–542 (2011).

6. L. Granlund, A. Hedin, M. Wahlhütter, P. Seiron, O. Korsgren, O. Skog, M. Lundberg, Histological and transcriptional characterization of the pancreatic acinar tissue in type 1 diabetes. *BMJ Open Diabetes Res. Care*. **9**, e002076 (2021).

7. J. Størling, F. Pociot, Type 1 Diabetes Candidate Genes Linked to Pancreatic Islet Cell Inflammation and Beta-Cell Apoptosis. *Genes*. **8**, E72 (2017).

8. M. Fasolino, G. W. Schwartz, A. R. Patil, A. Mongia, M. L. Golson, Y. J. Wang, A. Morgan, C. Liu, J. Schug, J. Liu, M. Wu, D. Traum, A. Kondo, C. L. May, N. Goldman, W. Wang, M. Feldman, J. H. Moore, A. S. Japp, M. R. Betts, HPAP Consortium, R. B. Faryabi, A. Naji, K. H. Kaestner, G. Vahedi, Single-cell multi-omics analysis of human pancreatic islets reveals novel cellular states in type 1 diabetes. *Nat. Metab.* **4**, 284–299 (2022).

9. D. V. Serreze, C. Wasserfall, E. W. Ottendorfer, M. Stalvey, M. A. Pierce, C. Gauntt, B. O'Donnell, J. B. Flanagan, M. Campbell-Thompson, T. M. Ellis, M. A. Atkinson, Diabetes acceleration or prevention by a coxsackievirus B4 infection: critical requirements for both interleukin-4 and gamma interferon. *J. Virol.* **79**, 1045–1052 (2005).

10. F. Brozzi, T. R. Nardelli, M. Lopes, I. Millard, J. Barthson, M. Igoillo-Esteve, F. A. Grieco, O. Villate, J. M. Oliveira, M. Casimir, M. Bugliani, F. Engin, G. S. Hotamisligil, P. Marchetti, D. L. Eizirik, Cytokines induce endoplasmic reticulum stress in human, rat and mouse beta cells via different mechanisms. *Diabetologia*. **58**, 2307–2316 (2015).

11. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell*. **184**, 3573-3587.e29 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

12. E. Caglayan, Y. Liu, G. Konopka, Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*, S0896-6273(22)00815–7 (2022).

13. R. D'Oliveira Albanus, Y. Kyono, J. Hensley, A. Varshney, P. Orchard, J. O. Kitzman, S. C. J. Parker, Chromatin information content landscapes inform transcription factor and DNA interactions. *Nat. Commun.* **12**, 1307 (2021).

14. Y. Fu, M. Sinha, C. L. Peterson, Z. Weng, The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. *PLoS Genet.* **4**, e1000138 (2008).

15. C. E. Malnou, F. Brockly, C. Favard, G. Moquet-Torcy, M. Piechaczyk, I. Jariel-Encontre, Heterodimerization with different jun proteins controls c-Fos intranuclear dynamics and distribution. *J. Biol. Chem.* **285**, 6552–6562 (2010).

16. K. A. Patel, J. Kettunen, M. Laakso, A. Stančáková, T. W. Laver, K. Colclough, M. B. Johnson, M. Abramowicz, L. Groop, P. J. Miettinen, M. H. Shepherd, S. E. Flanagan, S. Ellard, N. Inagaki, A. T. Hattersley, T. Tuomi, M. Cnop, M. N. Weedon, Heterozygous RFX6 protein truncating variants are associated with MODY with reduced penetrance. *Nat. Commun.* **8** (2017), doi:10.1038/s41467-017-00895-9.

17. T. Nammo, K. Yamagata, R. Hamaoka, Q. Zhu, T. E. Akiyama, F. J. Gonzalez, J. Miyagawa, Y. Matsuzawa, Expression profile of MODY3/HNF-1alpha protein in the developing mouse pancreas. *Diabetologia*. **45**, 1142–1153 (2002).

18. M. Fernandez Garcia, C. D. Moore, K. N. Schulz, O. Alberto, G. Donague, M. M. Harrison, H. Zhu, K. S. Zaret, Structural Features of Transcription Factors Associating with Nucleosome Binding. *Mol. Cell* (2019), doi:10.1016/j.molcel.2019.06.009.

19. D. Pavlovic, M. C. Chen, C. A. Gysemans, C. Mathieu, D. L. Eizirik, The role of interferon regulatory factor-1 in cytokine-induced mRNA expression and cell death in murine pancreatic beta-cells. *Eur. Cytokine Netw.* **10**, 403–412 (1999).

20. C. Gysemans, H. Callewaert, F. Moore, M. Nelson-Holte, L. Overbergh, D. L. Eizirik, C. Mathieu, Interferon regulatory factor-1 is a key transcription factor in murine beta cells under immune attack. *Diabetologia*. **52**, 2374–2384 (2009).

21. B. Koh, M. M. Hufford, D. Pham, M. R. Olson, T. Wu, R. Jabeen, X. Sun, M. H. Kaplan, The ETS Family Transcription Factors Etv5 and PU.1 Function in Parallel To Promote Th9 Cell Development. *J. Immunol.* **197**, 2465–2472 (2016).

22. C. Imbratta, H. Hussein, F. Andris, G. Verdeil, c-MAF, a Swiss Army Knife for Tolerance in Lymphocytes. *Front. Immunol.* **11**, 206 (2020).

23. J. K. Pickrell, Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

24. A. Mahajan, D. Taliun, M. Thurner, N. R. Robertson, J. M. Torres, N. W. Rayner, A. J. Payne, V. Steinthorsdottir, R. A. Scott, N. Grarup, J. P. Cook, E. M. Schmidt, M. Wuttke, C. Sarnowski, R. Mägi, J. Nano, C. Gieger, S. Trompet, C. Lecoeur, M. H. Preuss, B. P.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

Prins, X. Guo, L. F. Bielak, J. E. Below, D. W. Bowden, J. C. Chambers, Y. J. Kim, M. C. Y. Ng, L. E. Petty, X. Sim, W. Zhang, A. J. Bennett, J. Bork-Jensen, C. M. Brummett, M. Canouil, K.-U. Ec Kardt, K. Fischer, S. L. R. Kardia, F. Kronenberg, K. Läll, C.-T. Liu, A. E. Locke, J. Luan, I. Ntalla, V. Nylander, S. Schönherr, C. Schurmann, L. Yengo, E. P. Bottinger, I. Brandslund, C. Christensen, G. Dedoussis, J. C. Florez, I. Ford, O. H. Franco, T. M. Frayling, V. Giedraitis, S. Hackinger, A. T. Hattersley, C. Herder, M. A. Ikram, M. Ingelsson, M. E. Jørgensen, T. Jørgensen, J. Kriebel, J. Kuusisto, S. Ligthart, C. M. Lindgren, A. Linneberg, V. Lyssenko, V. Mamakou, T. Meitinger, K. L. Mohlke, A. D. Morris, G. Nadkarni, J. S. Pankow, A. Peters, N. Sattar, A. Stančáková, K. Strauch, K. D. Taylor, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, J. Tuomilehto, D. R. Witte, J. Dupuis, P. A. Peyser, E. Zeggini, R. J. F. Loos, P. Froguel, E. Ingelsson, L. Lind, L. Groop, M. Laakso, F. S. Collins, J. W. Jukema, C. N. A. Palmer, H. Grallert, A. Metspalu, A. Dehghan, A. Köttgen, G. R. Abecasis, J. B. Meigs, J. I. Rotter, J. Marchini, O. Pedersen, T. Hansen, C. Langenberg, N. J. Wareham, K. Stefansson, A. L. Gloyn, A. P. Morris, M. Boehnke, M. I. McCarthy, Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

25. J. Chen, C. N. Spracklen, G. Marenne, A. Varshney, L. J. Corbin, J. Luan, S. M. Willems, Y. Wu, X. Zhang, M. Horikoshi, T. S. Boutin, R. Mägi, J. Waage, R. Li-Gao, K. H. K. Chan, J. Yao, M. D. Anasanti, A. Y. Chu, A. Claringbould, J. Heikkinen, J. Hong, J.-J. Hottenga, S. Huo, M. A. Kaakinen, T. Louie, W. März, H. Moreno-Macias, A. Ndungu, S. C. Nelson, I. M. Nolte, K. E. North, C. K. Raulerson, D. Ray, R. Rohde, D. Rybin, C. Schurmann, X. Sim, L. Southam, I. D. Stewart, C. A. Wang, Y. Wang, P. Wu, W. Zhang, T. S. Ahluwalia, E. V. R. Appel, L. F. Bielak, J. A. Brody, N. P. Burtt, C. P. Cabrera, B. E. Cade, J. F. Chai, X. Chai, L.-C. Chang, C.-H. Chen, B. H. Chen, K. N. Chitrala, Y.-F. Chiu, H. G. de Haan, G. E. Delgado, A. Demirkan, Q. Duan, J. Engmann, S. A. Fatumo, J. Gayán, F. Giulianini, J. H. Gong, S. Gustafsson, Y. Hai, F. P. Hartwig, J. He, Y. Heianza, T. Huang, A. Huerta-Chagoya, M. Y. Hwang, R. A. Jensen, T. Kawaguchi, K. A. Kentistou, Y. J. Kim, M. E. Kleber, I. K. Kooner, S. Lai, L. A. Lange, C. D. Langefeld, M. Lauzon, M. Li, S. Ligthart, J. Liu, M. Loh, J. Long, V. Lyssenko, M. Mangino, C. Marzi, M. E. Montasser, A. Nag, M. Nakatochi, D. Noce, R. Noordam, G. Pistis, M. Preuss, L. Raffield, L. J. Rasmussen-Torvik, S. S. Rich, N. R. Robertson, R. Rueedi, K. Ryan, S. Sanna, R. Saxena, K. E. Schraut, B. Sennblad, K. Setoh, A. V. Smith, T. Sparsø, R. J. Strawbridge, F. Takeuchi, J. Tan, S. Trompet, E. van den Akker, P. J. van der Most, N. Verweij, M. Vogel, H. Wang, C. Wang, N. Wang, H. R. Warren, W. Wen, T. Wilsgaard, A. Wong, A. R. Wood, T. Xie, M. H. Zafarmand, J.-H. Zhao, W. Zhao, N. Amin, Z. Arzumanyan, A. Astrup, S. J. L. Bakker, D. Baldassarre, M. Beekman, R. N. Bergman, A. Bertoni, M. Blüher, L. L. Bonnycastle, S. R. Bornstein, D. W. Bowden, Q. Cai, A. Campbell, H. Campbell, Y. C. Chang, E. J. C. de Geus, A. Dehghan, S. Du, G. Eiriksdottir, A. E. Farmaki, M. Frånberg, C. Fuchsberger, Y. Gao, A. P. Gjesing, A. Goel, S. Han, C. A. Hartman, C. Herder, A. A. Hicks, C.-H. Hsieh, W. A. Hsueh, S. Ichihara, M. Igase, M. A. Ikram, W. C. Johnson, M. E. Jørgensen, P. K. Joshi, R. R. Kalyani, F. R. Kandeel, T. Katsuya, C. C. Khor, W. Kiess, I. Kolcic, T. Kuulasmaa, J. Kuusisto, K. Läll, K. Lam, D. A. Lawlor, N. R. Lee, R. N. Lemaitre, H. Li, Lifelines Cohort Study, S.-Y. Lin, J. Lindström, A. Linneberg, J. Liu, C. Lorenzo, T. Matsubara, F. Matsuda, G. Mingrone, S. Mooijaart, S. Moon, T. Nabika, G. N. Nadkarni, J. L. Nadler, M. Nelis, M. J. Neville, J. M. Norris, Y. Ohyagi, A. Peters, P. A. Peyser, O. Polasek, Q. Qi, D. Raven, D. F. Reilly, A. Reiner, F. Rivideneira, K. Roll, I. Rudan, C. Sabanayagam, K. Sandow, N. Sattar, A. Schürmann, J. Shi, H. M. Stringham, K. D. Taylor, T. M. Teslovich, B. Thuesen, P. R. H. J. Timmers, E. Tremoli, M. Y. Tsai, A. Uitterlinden, R. M. van Dam, D. van Heemst, A. van Hylckama Vlieg, J. V. van Vliet-Ostaptchouk, J.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

Vangipurapu, H. Vestergaard, T. Wang, K. Willems van Dijk, T. Zemunik, G. R. Abecasis, L. S. Adair, C. A. Aguilar-Salinas, M. E. Alarcón-Riquelme, P. An, L. Aviles-Santa, D. M. Becker, L. J. Beilin, S. Bergmann, H. Bisgaard, C. Black, M. Boehnke, E. Boerwinkle, B. O. Böhm, K. Bønnelykke, D. I. Boomsma, E. P. Bottinger, T. A. Buchanan, M. Canouil, M. J. Caulfield, J. C. Chambers, D. I. Chasman, Y.-D. I. Chen, C.-Y. Cheng, F. S. Collins, A. Correa, F. Cucca, H. J. de Silva, G. Dedoussis, S. Elmståhl, M. K. Evans, E. Ferrannini, L. Ferrucci, J. C. Florez, P. W. Franks, T. M. Frayling, P. Froguel, B. Gigante, M. O. Goodarzi, P. Gordon-Larsen, H. Grallert, N. Grarup, S. Grimsgaard, L. Groop, V. Gudnason, X. Guo, A. Hamsten, T. Hansen, C. Hayward, S. R. Heckbert, B. L. Horta, W. Huang, E. Ingelsson, P. S. James, M.-R. Jarvelin, J. B. Jonas, J. W. Jukema, P. Kaleebu, R. Kaplan, S. L. R. Kardia, N. Kato, S. M. Keinanen-Kiukaanniemi, B.-J. Kim, M. Kivimaki, H. A. Koistinen, J. S. Kooner, A. Körner, P. Kovacs, D. Kuh, M. Kumari, Z. Kutalik, M. Laakso, T. A. Lakka, L. J. Launer, K. Leander, H. Li, X. Lin, L. Lind, C. Lindgren, S. Liu, R. J. F. Loos, P. K. E. Magnusson, A. Mahajan, A. Metspalu, D. O. Mook-Kanamori, T. A. Mori, P. B. Munroe, I. Njølstad, J. R. O'Connell, A. J. Oldehinkel, K. K. Ong, S. Padmanabhan, C. N. A. Palmer, N. D. Palmer, O. Pedersen, C. E. Pennell, D. J. Porteous, P. P. Pramstaller, M. A. Province, B. M. Psaty, L. Qi, L. J. Raffel, R. Rauramaa, S. Redline, P. M. Ridker, F. R. Rosendaal, T. E. Saaristo, M. Sandhu, J. Saramies, N. Schneiderman, P. Schwarz, L. J. Scott, E. Selvin, P. Sever, X.-O. Shu, P. E. Slagboom, K. S. Small, B. H. Smith, H. Snieder, T. Sofer, T. I. A. Sørensen, T. D. Spector, A. Stanton, C. J. Steves, M. Stumvoll, L. Sun, Y. Tabara, E. S. Tai, N. J. Timpson, A. Tönjes, J. Tuomilehto, T. Tusie, M. Uusitupa, P. van der Harst, C. van Duijn, V. Vitart, P. Vollenweider, T. G. M. Vrijkotte, L. E. Wagenknecht, M. Walker, Y. X. Wang, N. J. Wareham, R. M. Watanabe, H. Watkins, W. B. Wei, A. R. Wickremasinghe, G. Willemsen, J. F. Wilson, T.-Y. Wong, J.-Y. Wu, A. H. Xiang, L. R. Yanek, L. Yengo, M. Yokota, E. Zeggini, W. Zheng, A. B. Zonderman, J. I. Rotter, A. L. Gloyn, M. I. McCarthy, J. Dupuis, J. B. Meigs, R. A. Scott, I. Prokopenko, A. Leong, C.-T. Liu, S. C. J. Parker, K. L. Mohlke, C. Langenberg, E. Wheeler, A. P. Morris, I. Barroso, Meta-Analysis of Glucose and Insulin-related Traits Consortium (MAGIC), The trans-ancestral genomic architecture of glycemic traits. *Nat. Genet.* **53**, 840–860 (2021).

26. A. Varshney, L. J. Scott, R. P. Welch, M. R. Erdos, P. S. Chines, N. Narisu, R. D. Albanus, P. Orchard, B. N. Wolford, R. Kursawe, S. Vadlamudi, M. E. Cannon, J. P. Didion, J. Hensley, A. Kirilusha, L. L. Bonnycastle, D. L. Taylor, R. Watanabe, K. L. Mohlke, M. Boehnke, F. S. Collins, S. C. J. Parker, M. L. Stitzel, Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci.* **114**, 2301–2306 (2017).

27. M. S. Udler, J. Kim, M. von Grotthuss, S. Bonàs-Guarch, J. B. Cole, J. Chiou, Christopher D. Anderson on behalf of METASTROKE and the ISGC, M. Boehnke, M. Laakso, G. Atzmon, B. Glaser, J. M. Mercader, K. Gaulton, J. Flannick, G. Getz, J. C. Florez, Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med.* **15**, e1002654 (2018).

28. G. Gibson, Perspectives on rigor and reproducibility in single cell genomics. *PLoS Genet.* **18**, e1010210 (2022).

29. H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D. Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina, A. C. Adey, F. J. Steemers, J. Shendure, C. Trapnell, Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell*. **71**, 858-871.e8 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

30. D. Lee, LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics*. **32**, 2196–2198 (2016).

31. D. Lee, D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, M. A. Beer, A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).

32. A. Shrikumar, E. Prakash, A. Kundaje, GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics*. **35**, i173–i182 (2019).

33. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **26**, 589–595 (2010).

34. P. Orchard, Y. Kyono, J. Hensley, J. O. Kitzman, S. C. J. Parker, Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. *Cell Syst.* **10**, 298-306.e4 (2020).

35. J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, W. J. Greenleaf, ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

36. A. Zadissa, J. M. Gonzalez, S. Searle, J. van Baren, V. Boychenko, D. Barrell, R. Guigo, B. Pei, N. Walters, F. Kokocinski, G. Despacio-Reyes, J. Rajan, T. J. Hubbard, B. L. Aken, M. Kellis, S. Balasubramanian, R. Harte, A. Tanzer, M. Kay, I. Ezkurdia, C. Steward, G. Mukherjee, J. M. Rodriguez, T. Hunt, A. Valencia, J. Chrast, D. Haussler, M. Brent, M. Diekhans, T. Derrien, M. Tress, E. Tapanari, M. Lin, I. Barnes, A. Frankish, G. Saunders, A. Bignell, M. Gerstein, C. Howald, J. Harrow, A. Reymond, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

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

38. participants in the 1st Human Cell Atlas Jamboree, A. T. L. Lun, S. Riesenfeld, T. Andrews, T. P. Dao, T. Gomes, J. C. Marioni, EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).

39. D. Christopoulos, Introducing Unit Invariant Knee (UIK) As an Objective Choice for Elbow Point in Multivariate Data Analysis Techniques. *SSRN Electron. J.* (2016), doi:10.2139/ssrn.3043076.

40. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-337.e4 (2019).

41. S. Yang, S. E. Corbett, Y. Koga, Z. Wang, W. E. Johnson, M. Yajima, J. D. Campbell, Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

42. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

43. P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, A. L Price, Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

44. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

45. G. Jun, M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke, H. M. Kang, Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).

46. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, R. Gottardo, MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

47. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

48. C. S. Smillie, M. Biton, J. Ordovas-Montanes, K. M. Sullivan, G. Burgin, D. B. Graham, R. H. Herbst, N. Rogel, M. Slyper, J. Waldman, M. Sud, E. Andrews, G. Velonias, A. L. Haber, K. Jagadeesh, S. Vickovic, J. Yao, C. Stevens, D. Dionne, L. T. Nguyen, A.-C. Villani, M. Hofree, E. A. Creasey, H. Huang, O. Rozenblatt-Rosen, J. J. Garber, H. Khalili, A. N. Desch, M. J. Daly, A. N. Ananthakrishnan, A. K. Shalek, R. J. Xavier, A. Regev, Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell*. **178**, 714-730.e22 (2019).

49. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

50. G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, A. Sergushichev, “Fast gene set enrichment analysis” (preprint, Bioinformatics, 2016), , doi:10.1101/060012.

51. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, Molecular signatures database (MSigDB) 3.0. *Bioinforma. Oxf. Engl.* **27**, 1739–1740 (2011).

52. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

53. M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, M. Tanabe, KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

54. D. Nishimura, BioCarta. *Biotech Softw. Internet Rep.* **2**, 117–120 (2001).

55. S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. C. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Z. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C. J. Mungall, M. A. Haendel, P. N. Robinson, The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).

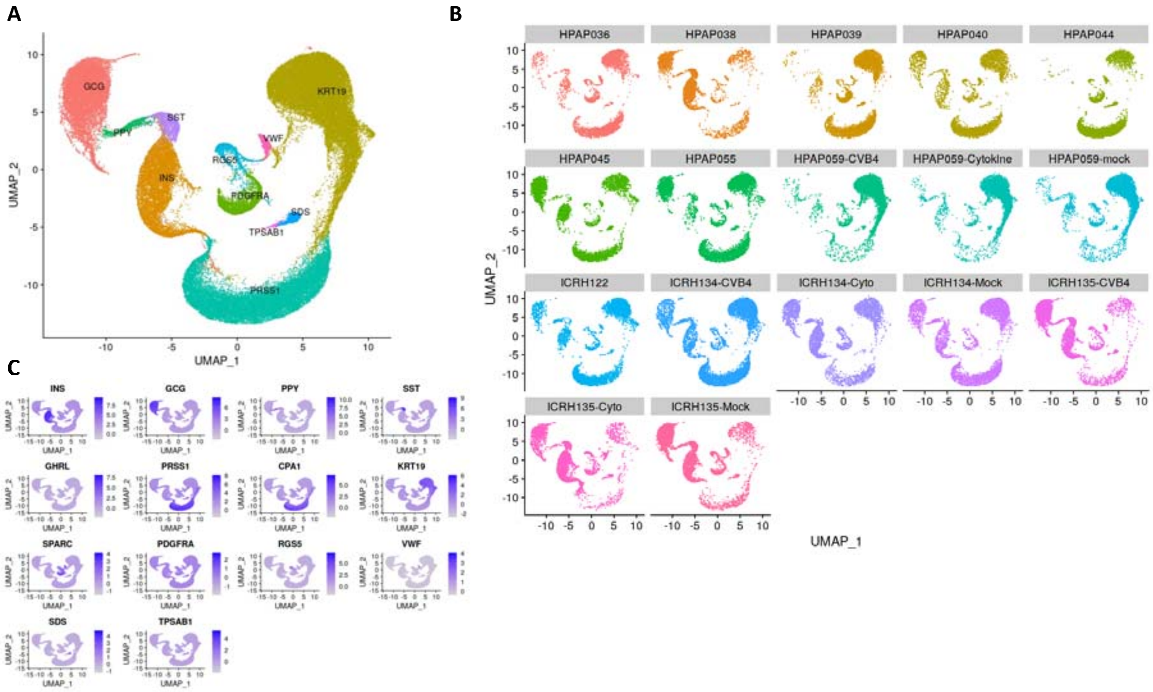
56. M. Ghandi, M. Mohammad-Noori, N. Ghareghani, D. Lee, L. Garraway, M. A. Beer, gkmSVM: an R package for gapped-kmer SVM. *Bioinforma. Oxf. Engl.* **32**, 2205–2207 (2016).

57. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*. **12**, 1061–1063 (2015).

58. L. Lopez-Delisle, L. Rabbani, J. Wolff, V. Bhardwaj, R. Backofen, B. Grüning, F. Ramírez, T. Manke, pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinforma. Oxf. Engl.* **37**, 422–423 (2021).

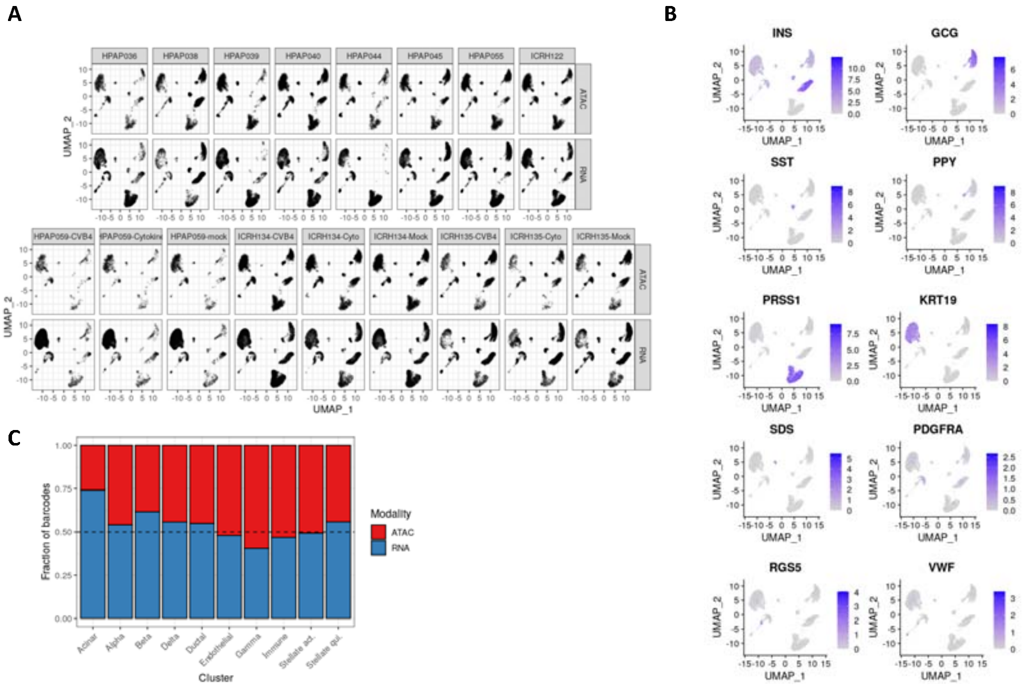
59. J. Köster, S. Rahmann, Snakemake--a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **28**, 2520–2522 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.



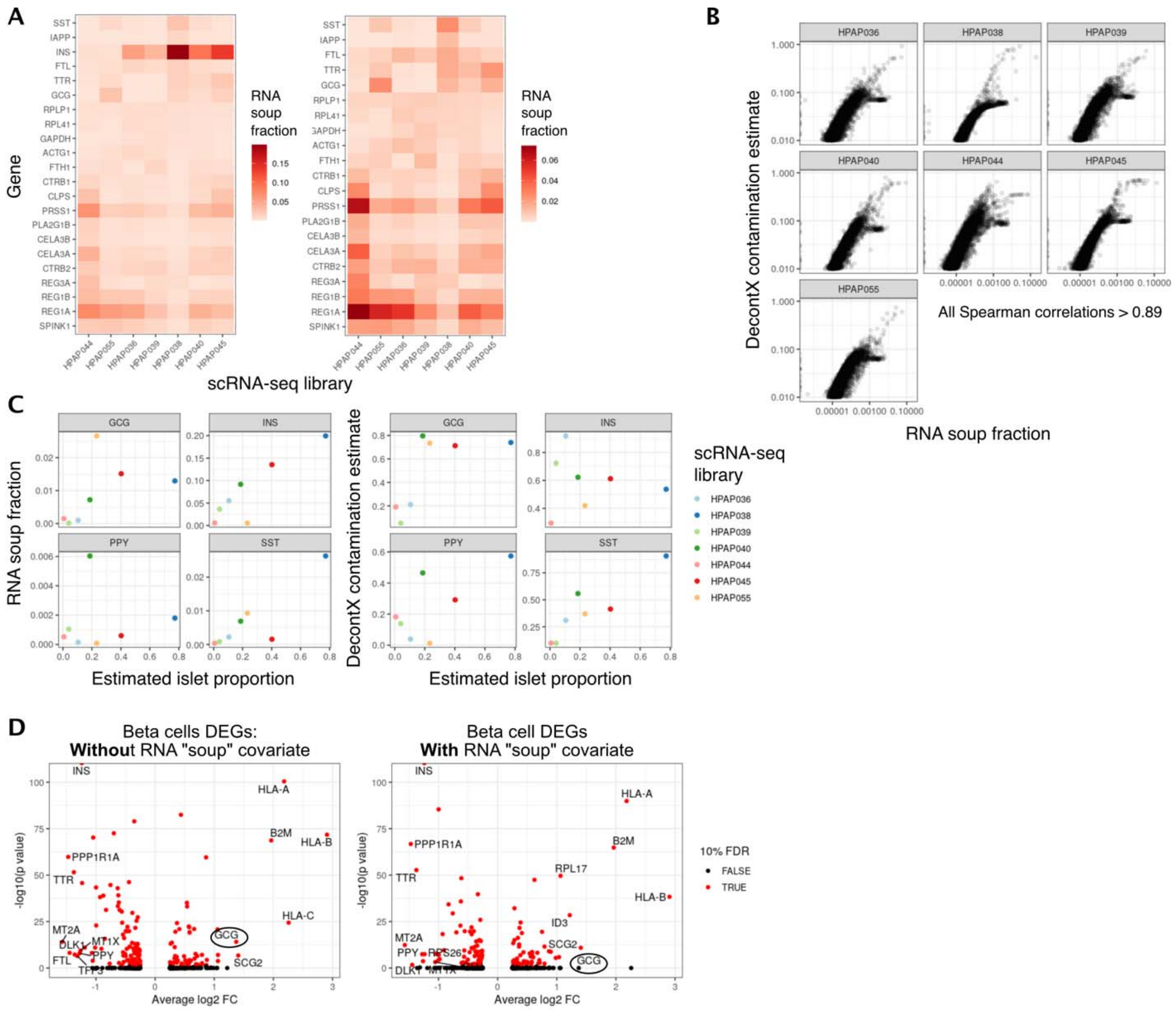
Supplementary Figure 1: A) UMAP representation of the first-pass scRNA-seq-only integration and clustering used as input for DecontX (B). UMAP representation split by samples. C) Marker gene expression in the first-pass scRNA-seq clustering.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.



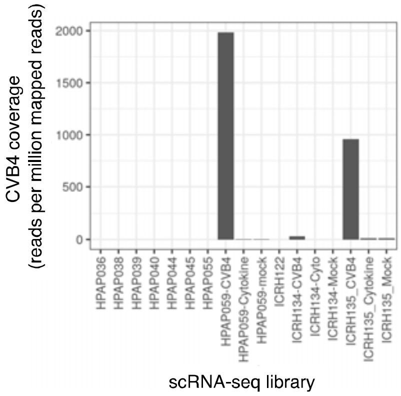
Supplementary Figure 2: A) UMAP representation of integrated scRNA-seq and snATAC-seq data faceted by sample (columns) and modality (rows) . B) Marker gene expression across clusters. C) Distribution of ATAC and RNA barcodes for each cell type.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

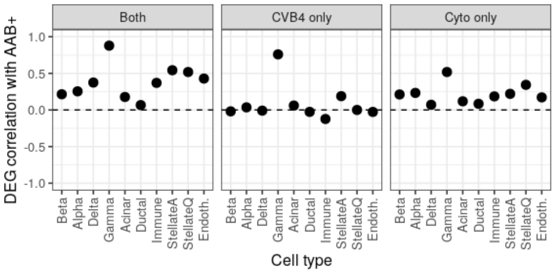


Supplementary Figure 3: A) Estimated ambient RNA ("RNA soup") composition for a subset of scRNA-seq libraries, obtained by combining all barcodes with less than 10 UMIs (i.e. empty droplets). Right plot is the same as left, but without INS for visibility. B) Agreement of the RNA contamination estimated by DecontX to ambient RNA fraction estimated directly from empty droplets. Clusters of off-diagonal genes correspond to ribosomal proteins. C) Comparison of ambient RNA fraction for each gene in the facets to the estimated islet proportion (fraction of barcodes assigned to the islet clusters) per library. D) DEGs in beta cells between HPAP055 (AAB+) versus controls with and without a covariate accounting for ambient RNA. HAPAP055 has a higher fraction of alpha cells compared to the other samples, which leads to higher levels of GCG in the ambient RNA. This, in turn, leads to erroneous assignment of GCG as a DEG (left plot, black circle). This technical artifact is mitigated once we include the estimated alpha cells proportion in the sample as a proxy of ambient RNA (right plot, black circle).

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.

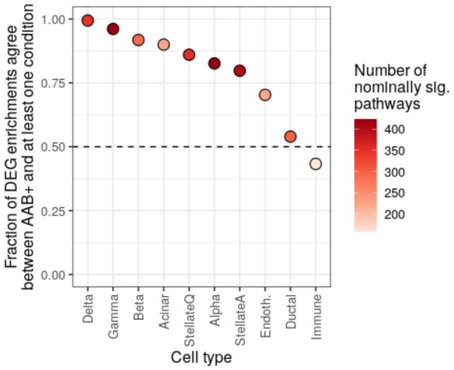


bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



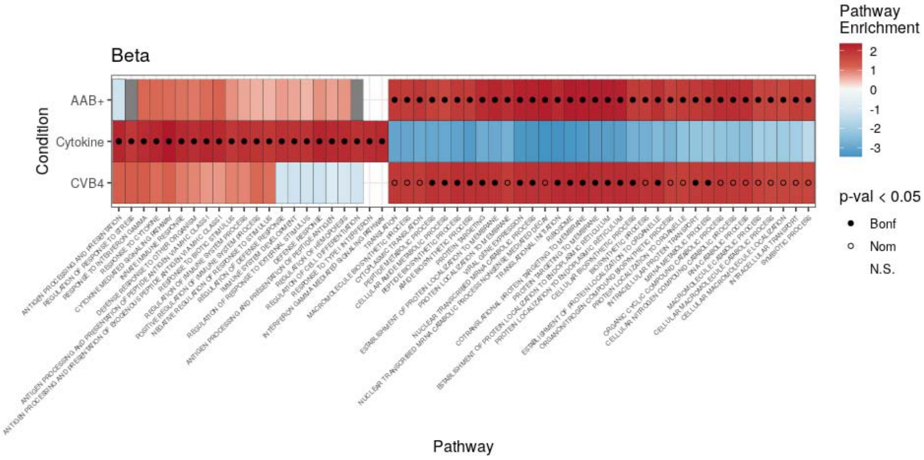
Supplementary Figure 5: DEG effect size correlation (Spearman) of nominally significant genes between AAB+ and other conditions across cell types.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.



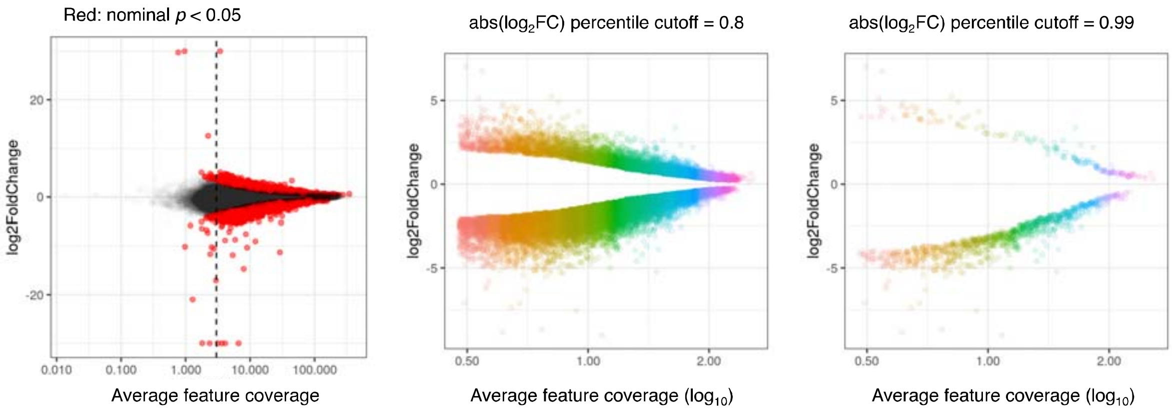
Supplementary Figure 6: Pathway enrichments agreements between DEGs in AAB+ versus other conditions across all cell types.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.



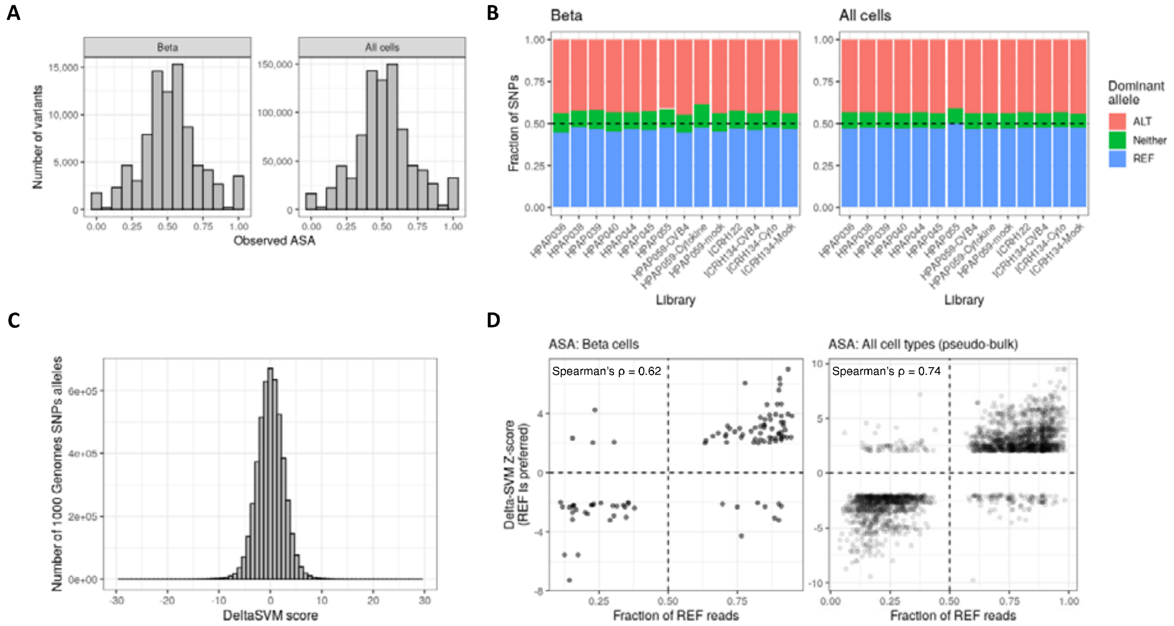
Supplementary Figure 7: All terms significantly enriched in beta cells differentially expressed genes across conditions, used as input for the rrvgo analyses in Figure 2C.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.



Supplementary Figure 8: Example DAR significance calculation using effect sizes. Each color in the rainbow plots in the middle and right panels correspond to one of the 50 ATAC-seq signal bins.

bioRxiv preprint doi: https://doi.org/10.1101/2022.11.12.516291; this version posted November 14, 2022. 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.



Supplementary Figure 9: A and B) Allele-specific accessibility (ASA) distribution in beta cells and all cells for all heterozygous SNPs to estimate reference bias in WASP. C) DeltaSVM score distribution for all heterozygous SNPs. D) Effect size comparison between SNPs with significant ASA and DeltaSVM scores.