UNIVERSITY OF CALIFORNIA SAN DIEGO

Multi-omic QTL analysis in pancreatic progenitor cells reveal early developmental insights into adult obesity and diabetes risk

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology with a
Specialization in Biomedical Informatics

by

Jennifer Phuong Nguyen

Committee in charge:

Professor Kelly A. Frazer, Chair
Professor Lucila Ohno-Machado, Co-Chair
Professor Melissa Gymrek
Professor Amit Majithia
Professor Alan Saltiel

2024

The dissertation of Jennifer Phuong Nguyen is approved, and it is acceptable in quality and form of publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my parents, Tina and Paul

My brother, Jimmy

My friends, Armin, Anne, and Vani

And my partner, Arnold

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and guidance of several key individuals. It is a pleasure to thank those who have made this work a possibility.

I am indebted to my dissertation advisor, Dr. Kelly A. Frazer, who welcomed me into her lab when I was a novice researcher, and it is under her mentorship that I have developed into the scientist I am today. Warmest thank you to my committee members for their valuable feedback and guidance on my research and career development as a scientist. I am deeply grateful for my mentors, Drs. Matteo D'Antonio and Agnieszcka D'Antonio-Chronowska, whose patience and vast expertise have been instrumental in my research education. I would also like to extend my thanks to the Frazer Lab and IGM for their tireless efforts in generating the data that was the backbone of this study. To other current and former members of the Frazer Lab, I offer my profound thanks for their encouragement and consistent support. Specifically, I extend my deepest gratitude to Timothy D. Arthur for his guidance and companionship; his creativity and natural gift as a leader have been an inspiration, and I am thankful for his contributions in the completion of this dissertation. I would like to acknowledge my previous mentors (Drs. Lee, Beretta, and Banaszynski) as well as my previous lab colleagues (Chao, Jonghae, Maria, Joon, Luxi, Sara, Dipti, Truong, Jing-Jing, Won-Baek, Rashieda, and Alberto) for their support and long-lasting friendships. Finally, I am thankful for my family, for their love and support throughout this journey, my DW friends, and my partner, for ~~forcing~~ helping me experience me what life is beyond the confines of my room.

Chapter 1, in full, is a reprint of the material as it appears in Nature Communications 2023, Jennifer P. Nguyen, Timothy D. Arthur, Kyohei Fujita, Bianca M. Salgado, Margaret K.R. Donovan, iPSCORE Consortium, Hiroko Matsui, Ji-Hyun Kim, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is an adapted version of a manuscript that is currently in preparation for publication with authors Timothy D. Arthur, Jennifer P. Nguyen, Agnieszka D'Antonio-Chronowska, Jeffrey Jaureguy, Nayara Silva, Benjamin Henson, iPSCORE Consortium, Athanasia D. Panopoulos, Juan Carlos Izpisua Belmonte, Matteo D'Antonio, Graham McVicker, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2012-2016	Bachelor of Science, Biochemistry, University of Texas Dallas
2018-2024	Doctor of Philosophy, Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics, University of California San Diego

PUBLICATION

First Authorships

Mapping genetic effects in early developmental tissues reveals phenotypic and temporal complexity of regulatory variants underlying GWAS loci. In preparation.

eQTL analysis of fetal-like pancreatic progenitors reveal risk loci associated with obesity and diabetes. Nature Communications, 2023, PMID: 37903777

Fine mapping spatiotemporal mechanisms of genetic variants underlying cardiac traits and disease. Nature Communications, 2023, PMID: 36854752

In heart failure reactivation of RNA-binding proteins is associated with the expression of 1,523 fetal-specific isoforms. PLoS Computational Biology, 2022, PMID: 35226669

Co-Authorships

Complex regulatory networks influence pluripotent cell state transitions in human iPSCs. Nature Communications, 2024. In Press.

Association of Human Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. Stem Cell Reports, 2019, PMID: 31668852

ABSTRACT OF THE DISSERTATION

Multi-omic QTL analysis in pancreatic progenitor cells reveal early developmental insights into adult obesity and diabetes risk

by

Jennifer Phuong Nguyen

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego, 2024

Professor Kelly A. Frazer, Chair
Professor Lucila Ohno-Machado, Co-Chair

Adverse events during fetal pancreas development can result in insulin resistance, impaired glucose metabolism, and loss of beta cell function, leading to an increased risk of developing diabetes in adulthood. While current quantitative trait loci (QTL) datasets have been instrumental in characterizing genetic variants associated with diabetes, they only reflect molecular associations present in mature adult tissues. Furthermore, only a fraction

of diabetes-associated loci colocalize with QTLs identified in adult whole pancreas and islet tissues. Given the important role of fetal development in adult diabetes predisposition, interrogating the molecular effects of genetic variation during this crucial period could provide valuable mechanistic insights into the etiology of obesity and diabetes.

First, we conducted an eQTL analysis on 107 RNA-seq samples from iPSC-derived pancreatic progenitor cells (PPC) to map genetic loci associated with gene expression and isoform usage changes during early pancreas development. Colocalization with eQTLs from adult pancreatic tissues identified genetic variants that were either specifically active during early pancreas development, specifically active in the adult pancreatic stage, or shared across both stages but had stage-unique regulatory functions. Colocalization with genome-wide association studies (GWAS) loci revealed developmental-unique eQTLs with potential roles in glucose homeostasis or diabetes, including those associated with *TPD52*, *CDC37L1-DT*, *MEG3*, and *CDH3*.

Second, we conducted chromatin accessibility QTL (caQTL) analysis using matched PPC ATAC-seq samples. We found that caQTL variants were enriched in distal regulatory regions, including CTCF-binding sites and PPC-specific super enhancer regions, and were enriched for motifs of transcription factors expressed in pancreatic progenitors. Colocalization of eQTLs, caQTLs, and GWAS signals identified putative regulatory mechanisms for *TPD52* expression and its impact on fasting glucose levels, as well as *KIT* and its impact on body mass index.

Together, this body of work provides a unique and powerful resource for interrogating the molecular effects of genetic variation during early pancreas development and their potential impact on adult complex pancreatic traits and disease.

Chapter 1: eQTL mapping of fetal-like pancreatic progenitor cells reveals early developmental insights into obesity and diabetes risk

1.1 Abstract

The impact of genetic regulatory variation active in early pancreatic development on adult pancreatic disease and traits is not well understood. Here, we generate a panel of 107 fetal-like iPSC-derived pancreatic progenitor cells (PPCs) from whole genome-sequenced individuals and identify 4065 genes and 4016 isoforms whose expression and/or alternative splicing are affected by regulatory variation. We integrate eQTLs identified in adult islets and whole pancreas samples, which reveal 1805 eQTL associations that are unique to the fetal-like PPCs and 1043 eQTLs that exhibit regulatory plasticity across the fetal-like and adult pancreas tissues. Colocalization with GWAS risk loci for pancreatic diseases and traits show that some putative causal regulatory variants are active only in the fetal-like PPCs and likely influence disease by modulating expression of disease-associated genes in early development, while others with regulatory plasticity likely exert their effects in both the fetal and adult pancreas by modulating expression of different disease genes in the two developmental stages.

1.2 Introduction

Genome-wide association studies (GWAS) have identified hundreds of genetic variants associated with adult pancreatic disease risk and phenotypes [1–4]. However, the majority of these associations map predominantly to non-coding regions of the genome, thereby hindering functional insights into disease processes [5–7]. Previous large-scale expression quantitative trait loci (eQTL) studies have made significant advancements toward understanding how genetic variation affects gene expression in various tissues and

cell types, as well as their contribution to human traits and diseases [8–11]. However, these analyses were conducted in adult tissues and therefore the effects of regulatory variation on gene expression under fetal conditions remain unclear. Moreover, the integration of adult and fetal eQTL datasets would enable the investigation of regulatory plasticity of genetic variants, which refers to changes in variant function under different spatiotemporal contexts [9,12,13]. Understanding how genetic variation affects gene expression during early pancreas development, and how their function changes in adulthood, can expand our understanding of the biological mechanisms underlying adult pancreatic disease and GWAS complex trait loci.

Many lines of evidence from clinical and genomic studies indicate an important role of pancreas development in the health and onset of childhood and adult pancreatic diseases [14–17]. For example, mutations in genes critical to pancreatic development, such as *PDX1*, *HNF4A*, and *HNF1A*, are associated with childhood-onset diabetes [18–20]. Furthermore, type 2 diabetes (T2D)-risk variants map to transcription factors (TFs) that are crucial to pancreas development, including NEUROG3 and HNF1A, and are enriched in accessible pancreatic progenitor-specific enhancers [4,14]. To address the limited availability of fetal pancreatic tissues, protocols have been developed to efficiently guide the differentiation of human induced pluripotent stem cells (iPSCs) into pancreatic progenitor cells (PPCs). This approach serves as a model system to study human pancreas development [21–28]. PPCs demonstrate expression of key transcription factors associated with early pancreas development, including *PDX1*, *NKX6-1*, and *SOX9*, all pivotal for pancreas lineage specification and differentiation [22,29–33]. Additionally, PPCs express developmental signaling pathway, including Notch, WNT, and Hedgehog, that are critical

in pancreas development [23,25,28,34]. While PPCs have provided extensive insights into pancreas developmental biology, they have not yet been utilized to examine the impact of genetic variation on gene expression in the fetal-like pancreas.

In this work, we conduct a large-scale eQTL analysis on 107 PPC samples to map genetic loci associated with gene expression and isoform usage during early pancreas development. We integrate eQTLs from adult pancreatic tissues and identify eQTL loci that display temporal specificity in early pancreas development, as well as eQTL loci that are shared with adult but display regulatory plasticity. Annotation of GWAS risk loci using our spatiotemporally informed eQTL resource reveals causal regulatory variants with developmental-unique effects associated with complex pancreatic traits and disease.

1.3 Study overview

The goal of our study is to understand how regulatory variation active in early pancreas development influences adult pancreatic disease risk and phenotypes (**Figure 1.1a**). We differentiated 106 iPSC lines from the iPSCORE resource [35] derived from 106 whole-genome sequenced individuals to generate 107 PPC samples (one iPSC line was differentiated twice). We characterized the fetal-like pancreatic transcriptome as well as the cellular composition using single-cell RNA-seq (scRNA-seq) of eight PPC samples. Then, we conducted an eQTL analysis on bulk RNA-seq of all 107 samples to identify regulatory variants associated with fetal-like gene expression and isoform usage. To better understand the spatiotemporal context of genetic variants, we integrated eQTLs previously discovered in adult pancreatic islets and whole pancreas samples using colocalization and network analysis. Finally, using our eQTL resource of pancreas tissues (i.e., fetal-like iPSC-derived PPCs, adult islets, adult whole pancreas), we performed GWAS

colocalization and fine-mapping to link developmental regulatory mechanisms and identify putative causal variants underlying pancreatic traits and disease associations.

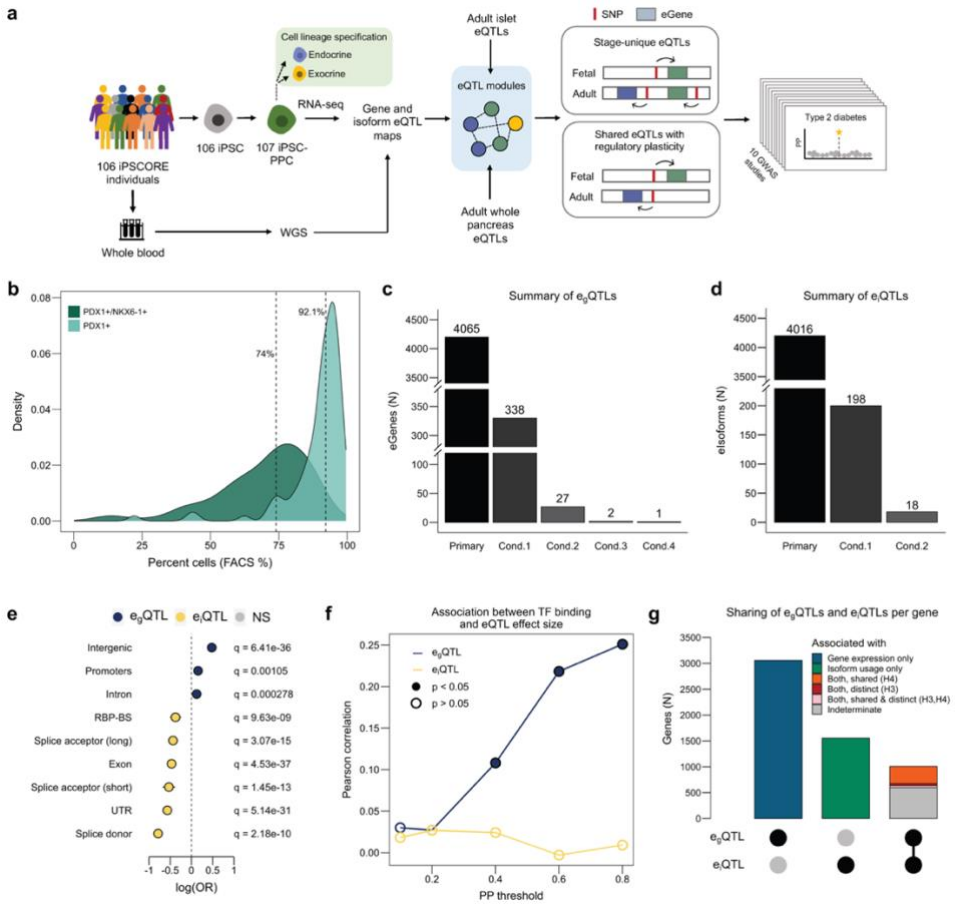


Figure 1.1 Characterization of PPC eQTLs

(a) Study overview created using PowerPoint. (b) Density plots showing the distribution of PDX1+ cells (%; regardless of NKX6-1 status; light green) and PDX1+/NKX6-1+ cells (%; dark green) in the 107 PPC samples. (c) Bar plot showing the number of eGenes with primary and conditional egQTLs. (d) Bar plot showing the number of eIsoforms with primary and conditional eiQTLs. (e) Enrichment (odds ratio, X-axis) of eQTLs in genomic regions (Y-axis) using two-sided Fisher's Exact Tests comparing the proportion of variants with causal posterior probability (PP) ≥ 5% in the genomic regions between egQTLs (blue; n=8,763) and eiQTLs (yellow; n=8,919). (f) Line plot showing Pearson correlations of TF binding score and eQTL effect size at different thresholds of causal PP for egQTLs (blue) and eiQTLs (yellow). Closed points indicate significant correlations (nominal p<0.05) while open points indicate non-significant correlations (nominal (p>0.05). (g) Bar plot showing the number of genes with only egQTLs (blue; n=3,057), only eiQTLs (green; n=1,554), or both. Orange represents genes whose egQTLs colocalized with all their corresponding eiQTLs (PP.H4 ≥ 80%; n=333). Red represents genes whose egQTLs did not colocalize with any of their corresponding eiQTLs (PP.H3 ≥ 80%; n=38), and pink represents genes with both shared and distinct egQTLs and eiQTLs (i.e., an eGene with two eIsoforms may colocalize with one eIsoform but not the other) (n=39). Gray represents genes whose eQTL signals were not sufficiently powered to test for colocalization (PP.H4 < 80% and PP.H3 < 80%; n=598).

1.4 Large-scale differentiation of PPC

We derived 107 PPC samples using iPSC lines reprogrammed from 106 individuals. Differentiation efficiency was assessed using flow cytometry analysis on PDX1 and NKX6-1, which are two markers routinely assayed for early pancreatic progenitor formation. PDX1 marks the specification of cells towards the pancreas lineage (referred to here as "early PPC"; $PDX1^{+}/NKX6\text{-}1^{-}$), while subsequent NKX6-1 expression marks the differentiation and maturation of pancreatic progenitor cells (referred to here as "late PPC"; $PDX1^{+}/NKX6\text{-}1^{+}$) [36]. We observed an 18.1% median percentage of early PPCs ($PDX1^{+}/NKX6\text{-}1^{-}$) across the 107 samples while the median percentage of late PPCs ($PDX1^{+}/NKX6\text{-}1^{+}$) was 74% (range: 9.4%-93.1%) (**Figure 1.1b**). We further found that the median percentage of cells that expressed PDX1+ was more than 90%, confirming that the majority of cells have specified towards the pancreas lineage and that the differentiation procedure was highly efficient (**Figure 1.1b**). Consistent with flow cytometry analysis, scRNA-seq of ten derived PPCs confirmed the presence of both early and late PPCs and that the majority of the cells were late PPCs (**Figure 1.2a-c**; See Methods). Altogether, these results show that the majority of the cells in PPCs were differentiated into late PPCs while a smaller fraction represented a primitive PPC state.

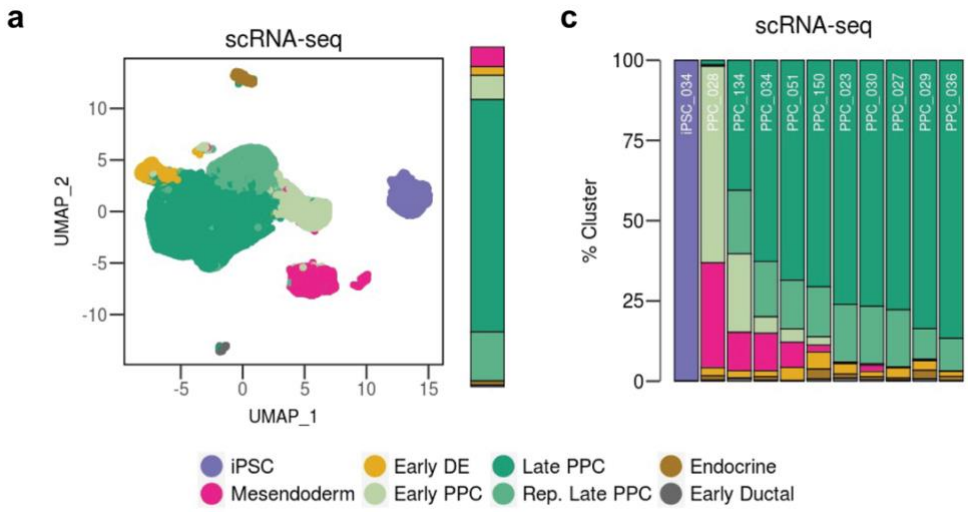
To examine the similarities between PPC and adult pancreatic transcriptomes, we generated bulk RNA-seq for all 107 PPC samples and inferred the pseudotime on each sample, along with 213 iPSCs [35,37], 87 pancreatic islets [38], and 176 whole pancreatic tissues [39]. Pseudotime analysis and comparative expression analysis of early developmental genes showed that the PPC samples corresponded to an early timepoint of pancreas development.

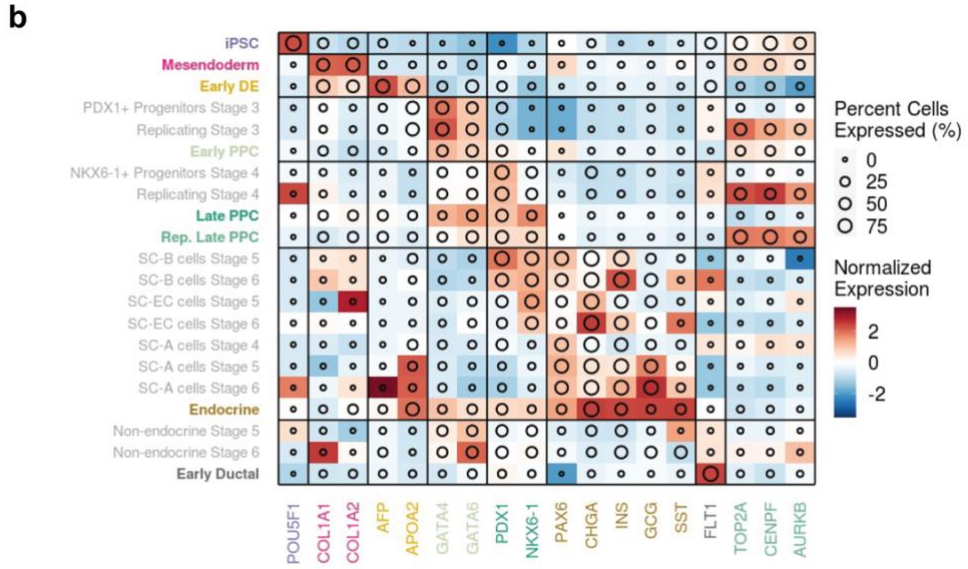
These analyses, combined with the results of previous studies [22,25,27], show that the 107 derived PPCs represent a fetal-like state of pancreatic tissues.

1.5 Gene and Isoform eQTL Discovery

To characterize the effects of genetic variation on the fetal-like PPC transcriptome, we performed an eQTL analysis to map the genetic associations with fetal-like gene expression (e_gQTL) and relative isoform usage (e_iQTL). Considering only autosomal chromosomes, we analyzed a total of 16,464 genes and 29,871 isoforms (corresponding to 9,624 autosomal genes) that were expressed in the fetal-like PPCs. We identified 4,065 (24.7%) eGenes and 4,016 (13.0%) eIsoforms with an e_gQTL or e_iQTL, respectively (FDR < 0.01, **Figure 1.1c-d**). To identify additional independent eQTL signals (i.e., conditional eQTLs) [40], we performed a stepwise regression analysis for each eGene and eIsoform. This analysis yielded 368 e_gQTLs that mapped to 338 eGenes and 216 e_iQTLs that mapped to 198 eIsoforms, totaling to 4,433 independent e_gQTL associations and 4,232 independent e_iQTL associations (**Figure 1.1c-d**). We next predicted candidate causal variants underlying each eQTL (e_gQTL and e_iQTL) association using *coloc* genetic fine-mapping [41] and tested their enrichments in transcribed regions and regulatory elements. We observed an enrichment of e_gQTLs in intergenic and promoter regions while e_iQTLs were enriched in splice sites and RNA-binding protein binding sites (**Figure 1.1e**).

Figure 1.2 Single-cell characterization of PPC samples
We characterized the cellular composition of PPC using scRNA-seq of one iPSC (for PPC034 differentiation) and ten PPC samples with variable percentages of double-positive cells (range: 9.4% - 91.7%). We identified eight distinct cell populations, corresponding to iPSC (*POU5F1*), mesendoderm (*COL1A1/2*), early definitive endoderm (early DE; *AFP*, *APOA2*), early PPC (*GATA*, *GATA6*, *PDX1*), late PPC (*PDX1* and *NKX6-1*), replicating late PPC (*PDX1*, *NKX6-1*, *TOP2A*, *CENPF*, *AURKB*), endocrine (*PAX6*, *CHGA*, *INS*, *GCG*, *SST*), and early ductal (*FLT1*). We observed highly similar gene expression profiles between the cell types identified in PPC and those identified in an ESC-derived PPC (ESC-PPC) reference dataset [27]. (a) UMAP plot of scRNA-seq data from 84,225 single cells from one iPSC and ten PPC samples. Each point represents a single cell color-coded by its assigned cluster. To the right of the UMAP plot, we show relative proportion of cells associated with each cell type (iPSC cells excluded). We show that the vast majority of cells in PPCs were late PPCs. (b) Heatmap comparing the Z-normalized expression of known marker genes between PPC and cells from the reference ESC-PPC study [27]. Color intensity indicates the mean Z-normalized expression across all cell types, and the diameter indicates the percentage of cells expressing the markers above the threshold of 1% of the maximum expression value. Clusters labeled in color correspond to the PPC clusters. Clusters labeled in grey correspond to ESC-PPC clusters [27]. (c) Stacked bar plot showing the relative proportion of cells from each sample assigned to each cluster in scRNA-seq. Color-coding corresponds to the clusters in panel a. Samples with the least number of late PPC cells correspond to those with weaker differentiation efficiency based on FACS, and contain more cells of primitive state compared to the other samples.


a
scRNA-seq
UMAP_2
UMAP_1
c
scRNA-seq
% Cluster
iPSC_034
PPC_026
PPC_134
PPC_034
PPC_051
PPC_150
PPC_023
PPC_030
PPC_027
PPC_029
PPC_036
iPSC
Mesendoderm
Early DE
Early PPC
Late PPC
Rep. Late PPC
Endocrine
Early Ductal


b
iPSC
Mesendoderm
Early DE
PDX1+ Progenitors Stage 3
Replicating Stage 3
Early PPC
NKX6-1+ Progenitors Stage 4
Replicating Stage 4
Late PPC
Rep. Late PPC
SC-B cells Stage 5
SC-B cells Stage 6
SC-EC cells Stage 5
SC-EC cells Stage 6
SC-A cells Stage 4
SC-A cells Stage 5
SC-A cells Stage 6
Endocrine
Non-endocrine Stage 5
Non-endocrine Stage 6
Early Ductal
POU5F1
COL1A1
COL1A2
AFP
APOA2
GATA4
GATA6
PDX1
NKX6-1
PAX6
CHGA
INS
GCG
SST
FLT1
TOP2A
CENPF
AURKB
Percent Cells Expressed (%)
0
25
50
75
Normalized Expression
2
0
-2

We additionally estimated the transcription factor (TF) binding score for each variant using the Genetic Variants Allelic TF Binding Database [42] and found that, at increasing posterior probability (PP, probability that the variant is causal for the association) thresholds, the candidate causal variants underlying e_gQTLs were more likely to affect TF binding compared to those underlying e_iQTLs (**Figure 1.1f**). These results corroborate similar findings from previous studies [10,12,43], showing that the genetic variants underlying e_gQTLs and e_iQTLs primarily affect gene regulation and coding regions or alternative splicing, respectively.

To further characterize the function of genetic variants associated with the fetal-like PPC transcriptome, we examined the distributions of e_gQTLs and e_iQTLs per gene. Of the 5,619 genes whose phenotype was affected by genetic variation, 1,008 were impacted through both gene expression and isoform usage (i.e., had both e_gQTL and e_iQTLs, 17.9%) while 3,057 were impacted through only gene expression (i.e., had only e_gQTLs, 54.4%) and 1,554 through only isoform usage (i.e., had only e_iQTLs, 27.7%, **Figure 1.1g**). For the 1,008 genes with both e_gQTL and e_iQTLs, we performed colocalization with *coloc.abf* [41] to examine whether the same or different genetic variants underpinned their associations. *coloc.abf* [41] employs a Bayesian approach to estimate the PP that each of the five colocalization models best explains the association between two genetic signals: H0) no associations detected in either signal; H1) association detected in only signal 1; H2) association detected in only signal 2; H3) associations detected in both signals but driven by different causal variants, and H4) associations detected in both signals and driven by the same causal variant. We identified 410 (40.7%) genes that had at least one H4 (PP.H4, posterior probability for H4 $\geq$ 80%) or H3 (PP.H3, posterior probability for H3 $\geq$ 80%)

association between their e_gQTL and e_iQTLs, of which the majority (333, 81.2%) had only overlapping signals (all H4), 38 (9.3%) had only non-overlapping signals (all H3), and 39 (9.5%) had both overlapping and non-overlapping e_iQTLs (both H3 and H4; an e_gQTL can colocalize with an e_iQTL corresponding to one isoform but not with another e_iQTL corresponding to a second isoform) (**Figure 1.1g**). The remaining 598 genes had PP.H3 < 80% and PP.H4 < 80% due to insufficient power (**Figure 1.1g**). These findings show that 19.5% (1,008 / 5,169) of genes had both e_gQTLs and e_iQTLs and that their effects were commonly driven by the same causal variants (81.2%) while only a small fraction was driven by different causal variants (9.3%).

Overall, our results show that the majority of genes had either only e_gQTLs or e_iQTLs, indicating that the functional mechanisms underlying these associations are likely independent, where genetic variants affecting alternative splicing do not affect the overall expression of the gene, and vice versa.

1.6 eQTL landscapes of fetal-like PPC and adult pancreatic islets

Studies aimed at identifying and characterizing eGenes have been conducted in both adult human islets and whole pancreatic tissues [8,10,11,38,44]; however, islet tissues have been more thoroughly studied because of their role in diabetes. Therefore, we focused on understanding the similarities and differences between eGenes in the fetal-like PPCs and adult human islets.

We obtained eQTL summary statistics and intersected the 4,211 autosomal eGenes identified in 420 adult islet samples [11] with the 4,065 eGenes in fetal-like PPC. We found that only 1,501 (36.9% of 4,065) eGenes overlapped between the fetal-like PPC and adult islet tissues (**Figure 1.3a**). To determine whether the small overlap was due to gene

expression differences, we calculated how many of the eGenes were expressed in both the fetal-like PPC and adult islets. Of the 4,065 fetal-like PPC eGenes, 88.7% (3,605) were also expressed in the adult islets; likewise, of the 4,211 adult islet eGenes, 78.4% (3,301) were also expressed in the fetal-like PPCs (**Figure 1.3b**). These results suggest that most fetal-like PPC eGenes were expressed but not associated with genetic variation in the adult islet samples, and vice versa.

For eGenes that were present in both the fetal-like PPC and adult islet samples, we next asked whether their expressions were controlled by the same genetic variants. We performed colocalization between e_gQTLs for the 1,501 shared eGenes in the fetal-like PPC and adult islets and found that 795 (52.3%) displayed strong evidence for either H3 or H4 association (PP.H3 or PP.H4 $\geq$ 80%). Of the 795 with an association, 701 (88.2%) had overlapping e_gQTL signals (PP.H4 $\geq$ 80%) while 94 (11.8%) had non-overlapping e_gQTL signals (PP.H3 $\geq$ 80%) (**Figure 1.3c**). These results indicate that most shared eGenes were associated with the same genetic variants controlling their gene expressions in both fetal-like PPC and adult islet tissues, while a subset had non-overlapping genetic variants. Further, we examined the effect sizes of lead variants between adult islets and iPSC-PPC and observed a stronger correlation (r=0.64) for eGenes that were shared between the two tissues compared to those that were not shared (r=0.05) (**Figure 1.3e)**. We identified *SNX29* as an eGene in both fetal-like PPC and adult islets but observed that its expression was associated with distinct eQTL signals approximately 520 kb apart (**Figure 1.3d**). SNX29 is involved in various signaling pathways [45], including TGF-β, ErbB, and WNT signaling pathways, and is predicted to be a causal gene for body-mass index (BMI) and T2D [46].

Taken together, our results show that a minor proportion of fetal-like PPC eGenes (1,501, 37% of 4,065) was shared with adult islets, whereas the majority (2,564 = 4,065–1,501, 63%) were fetal development-specific; and, while most shared eGenes were associated with the same regulatory variants, ~12% were mediated by different eQTLs. These findings indicate that regulatory variants tend to act in a developmental-specific manner, potentially by affecting the binding of key regulatory TFs specific to fetal or adult pancreatic stages.

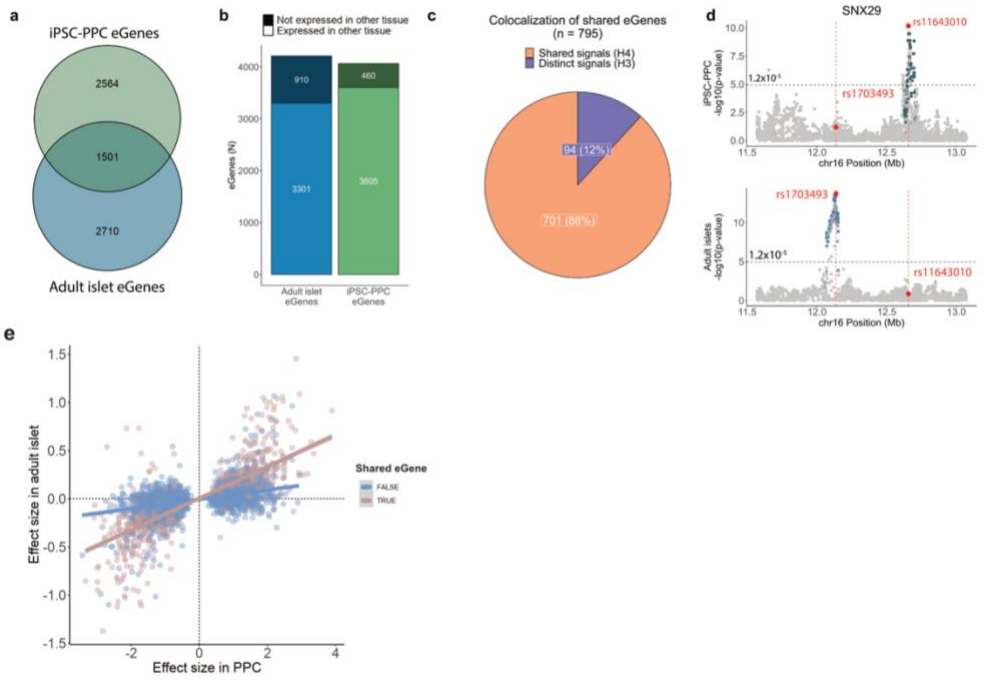


Figure 1.3 Comparison of the genetic architecture underlying gene expression between fetal-like and adult islets

(a) Venn diagram showing the overlap of eGenes between fetal-like PPC and adult islets. (b) Stacked bar plot showing the total number of eGenes detected in adult islets (blue; n=4,211 total) that were expressed in PPC (light blue; n=3,301). Likewise, we show the total number of fetal-like PPC eGenes (green; n=4,065 total) that were expressed in adult islets (light green; n=3,605). These results show that the majority of eGenes were expressed in both tissues, however, a large fraction was influenced by genetic variation in only one of the two tissues. Therefore, the small overlap of eGenes may be due to differences in the genetic regulatory landscape. (c) Pie chart showing the proportion of shared eGenes with distinct genetic loci (PP.H3 ≥ 80%, purple) or shared genetic loci (PP.H4 ≥ 80%, orange). These results show that 12% of the shared eGenes were associated with distinct regulatory variants between fetal-like and adult pancreatic stages. (d) Example of a shared eGene (SNX29) whose expression was associated with distinct egQTL signals (PP.H3 = 90.4%) in fetal-like PPC (green, top panel) and adult islets (blue, bottom panel). The X-axis represents variant positions while the Y-axis shows the -log10(eQTL p-value) for the associations between the genotype of the tested variants and gene expression. For plotting purposes, we assign a single p-value for gene-level significance after Bonferroni-correction (0.05 / number of independent variants tested in fetal-like PPC; horizontal line). Red vertical lines show the positions of the lead variants in fetal-like PPC and adult islets (chr16:12656135:C>G and chr16:12136526:A>G, respectively). (e) Scatter plot showing the correlation of effect sizes of lead variants for shared eGenes (pink) versus non-shared eGenes (blue).

1.7 Developmental stage-unique and shared e_gQTLs

Above, we described eGenes that were unique to either fetal-like PPCs or adult islets, or shared between both. Here, we sought to identify eQTLs (i.e., regulatory variants) that specifically affect gene expression during the pancreas development stage, in the adult stage, or both stages. Because fetal-like PPCs give rise to both endocrine and exocrine cell fates, we included eQTLs from both adult islets [11] and whole pancreas [39] tissues in our analyses. Due to the many different types of eQTLs used in this study, we refer to all eQTLs as a collective unit as "eQTLs", eQTLs that were associated with gene expression as "e_gQTLs" (as defined above), and eQTLs associated with changes in alternative splicing (e_iQTLs, exon eQTLs, and sQTLs) as "e_{AS}QTLs". For simple interpretations, we only describe the results for the analyses conducted on the e_gQTLs below, however, we identified unique and shared PPC e_{AS}QTL associations by conducting the same analyses.

To identify e_gQTLs that shared the same regulatory variants, we performed pairwise colocalization using *coloc.abf* [41] between e_gQTLs in fetal-like PPC, adult islets [11], and adult whole pancreas samples [10]. We considered only e_gQTLs that had at least one variant with causal PP $\geq$ 1% (from genetic fine-mapping [41]), were outside the MHC region, and associated with genes annotated in GENCODE version 34 [47]. From colocalization, we identified 7,893 pairs of e_gQTLs that displayed high evidence of colocalization with PP.H4 $\geq$ 80%, and 8,570 e_gQTLs that did not colocalize with e_gQTLs. Hereafter, we refer to eQTLs that did not colocalize with eQTLs as "singletons" and those that colocalized with another eQTL (PP.H4 $\geq$ 80%; same or different tissue) as "combinatorial" (i.e., the 7,893 pairs of e_gQTLs).

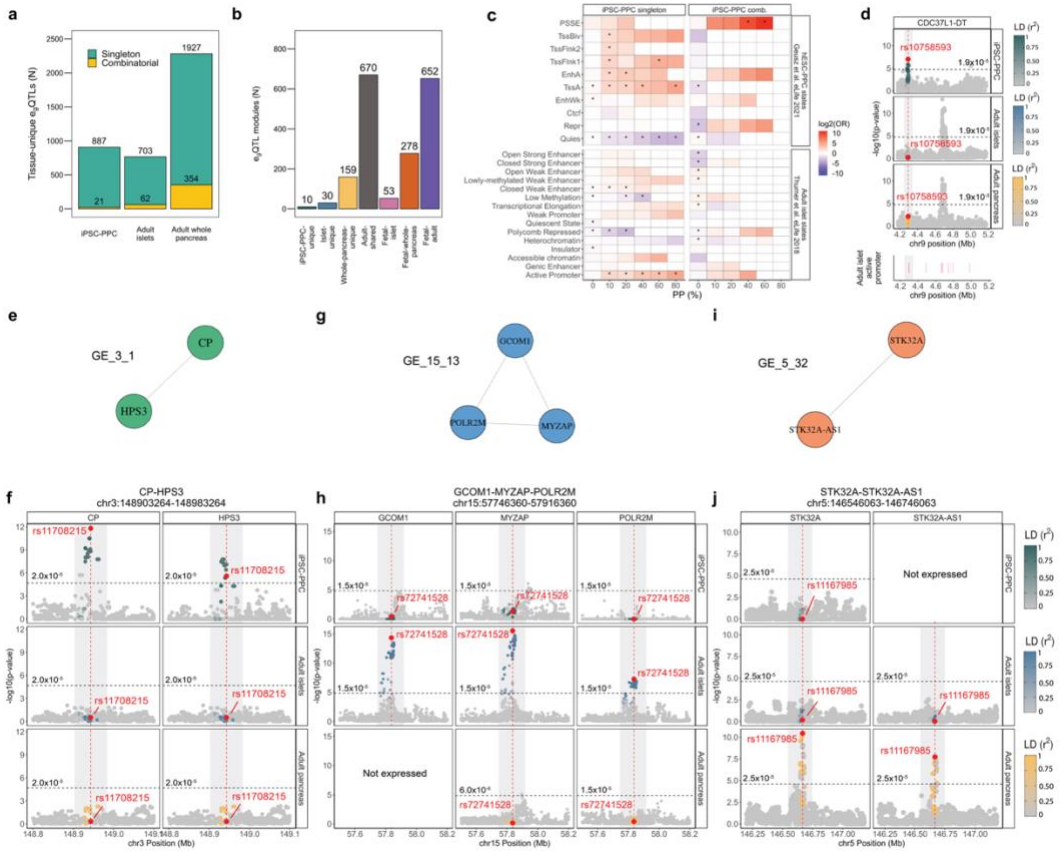
We next sought to identify singleton and combinatorial e_gQTL signals that were unique to PPC or shared between PPC and the adult pancreatic tissues. The singleton PPC e_gQTLs were associated with a single eGene and not active in the adult pancreatic samples, and hence tissue-unique. To ensure that there was no overlap of singleton e_gQTLs with other e_gQTLs in either the fetal-like and adult pancreas tissues, we implemented an LD filter ($r^2 \geq 0.2$ with any e_gQTL within 500 Kb or within 500 Kb if LD could not be calculated; see Methods). We identified 3,517 tissue-unique singleton e_gQTLs (887 PPC + 703 adult islet + 1,927 adult whole pancreas) that were not in LD with any nearby e_gQTLs (**Figure 1.4a**).

To identify tissue-unique combinatorial e_gQTL signals, we created a network using the 7,893 pairs of colocalized e_gQTL associations. We identified 1,852 e_gQTL modules that passed specific criteria for module identification and LD filters (see Methods), of which 939 (50.7%) comprised two e_gQTLs while the remaining 913 (49.3%) had an average of four e_gQTLs per module (range: 3-20 e_gQTLs). In total, we identified 199 (10.7% of 1,852) modules that were tissue-unique, of which 10 were PPC-unique, 30 adult islet-unique, and 159 adult whole pancreas-unique (**Figure 1.4b**), and altogether comprised 21, 62, and 354 e_gQTLs in combinatorial associations, respectively (**Figure 1.4a**). In contrast, the remaining 1,653 (89.3% of 1,852) modules were associated with multiple pancreatic tissues, of which 670 were shared between only adult islet and whole pancreas tissues (referred to as "adult-shared"), 53 were shared between only PPC and adult islets ("fetal-islet"), 278 between only PPC and adult whole pancreas ("fetal-whole-pancreas"), and 652 between all three pancreatic tissues ("fetal-adult") (**Figure 1.4b**). Together, the

983 (53 + 278 + 652) modules shared between PPC and an adult pancreatic tissue were composed of 1,122 PPC, 870 adult islets, and 1,394 adult whole pancreas e_gQTLs.

For e_{AS}QTLs, we observed similar trends in which the majority of fetal-like PPC-unique e_{AS}QTLs were singletons and that combinatorial e_{AS}QTLs were likely shared with the adult pancreas tissues. Altogether, including e_{AS}QTLs, we identified 1,805 PPC eQTLs that were unique to fetal-like PPC, of which 1,518 (887 e_gQTLs + 631 e_{AS}QTLs) functioned as singletons and 287 (21 e_gQTLs + 266 e_{AS}QTLs) in modules; while 1,977 (1,175 e_gQTLs + 802 e_{AS}QTLs) were shared with adult pancreatic tissues, and 4,326 (2,066 e_gQTLs + 2,260 e_{AS}QTLs) failed one or more the stringent filters and were marked as ambiguous.

Figure 1.4. eQTL sharing between PPC, adult islets, and adult whole pancreas
(a) Bar plot showing the number of tissue-unique egQTLs identified in fetal-like PPC, adult islets and adult whole pancreas. (b) Bar plot showing the number of egQTL modules for each annotation. (c) Top panels: Enrichment (odds ratio) of PPC singleton and combinatorial egQTLs in hESC-derived PPC chromatin states 14. Bottom panels: Enrichment (odds ratio) of PPC singleton and combinatorial egQTLs in adult islet chromatin states 48. Enrichment was calculated using a two-sided Fisher's Exact Test comparing the proportion of candidate causal variants overlapping the chromatin states versus a background of randomly selected 20,000 variants at various PP thresholds. Significance was determined by BH-corrected p-values < 0.05 (indicated by asterisk). (d) CDC37L1-DT locus showing an PPC-unique singleton egQTL overlapping an adult islet active promoter region. Lower panel shows the positions of active promoters in the adult islets. (e-f) Example of an "PPC-unique" module (g-h) Example of an "adult islet-unique" module. GCOM1 was not expressed in adult whole pancreas and therefore, was not tested for egQTL association. (i-j) Example of an "adult whole pancreas-unique" egQTL module. STK32A-AS1 was not expressed in PPC and therefore, was not tested for egQTL association. Panels e, g, i display the egQTL modules as networks in which the egQTL associations (nodes) are connected by edges due to colocalization (PP.H4 ≥ 80%). For panels d, f, h, and j, the X-axis represents variant positions while the Y-axis shows the -log10(eQTL p-value) for the associations between the genotype of the tested variants and gene expression. For plotting purposes, we assigned a single p-value for gene-level significance after Bonferroni-correction (0.05 / the number of independent variants tested in fetal-like PPC; horizontal line). Red vertical lines indicate the positions of the lead candidate causal variants underlying the colocalization based on maximum PP.


a
Singleton
Combinatorial
Tissues-unique eQTLs (N)
887
703
1927
21
62
354
iPSC-PPC
Adult islets
Adult whole pancreas
b
eQTL modules (N)
10
30
159
670
53
278
652
c
iPSC-PPC singleton
iPSC-PPC comb.
PSSE
TssBiv
TssFlnk2
TssFlnk1
EnhA
TssA
EnhWk
Ctcf
Repr
Quies
Open Strong Enhancer
Closed Strong Enhancer
Open Weak Enhancer
Lowly-methylated Weak Enhancer
Closed Weak Enhancer
Low Methylation
Transcriptional Elongation
Weak Promoter
Quiescent State
Polycomb Repressed
Heterochromatin
Insulator
Accessible chromatin
Genic Enhancer
Active Promoter
log2(OR)
PP (%)
d
CDC37L1-DT
rs10758593
1.9x10⁻⁶
-log10(p-value)
chr9 position (Mb)
LD (r²)
e
GE_3_1
CP
HPS3
g
GE_15_13
GCOM1
POLR2M
MYZAP
i
GE_5_32
STK32A
STK32A-AS1
f
CP-HPS3
chr3:148903264-148983264
rs11708215
2.0x10⁻⁶
chr3 Position (Mb)
h
GCOM1-MYZAP-POLR2M
chr15:57746360-57916360
rs72741528
1.5x10⁻⁵
Not expressed
6.0x10⁻⁶
chr15 Position (Mb)
j
STK32A-STK32A-AS1
chr5:146546063-146746063
rs11167985
2.5x10⁻⁶
Not expressed
chr5 Position (Mb)

1.8 Characterization of fetal-like PPC-unique e_gQTLs

To functionally characterize the fetal-like PPC tissue-unique singleton and combinatorial e_gQTLs, we calculated their enrichments in chromatin state annotations from human ESC-derived PPCs[14]. At high PP thresholds, we observed the strongest enrichment of singleton e_gQTLs in active promoter (TssA) regions, consistent with their role in regulating the expression of a single gene (**Figure 1.4c**). For combinatorial e_gQTLs, we observed a strong enrichment in PPC-specific stretch enhancer (PSSE) regions at high PP thresholds ($p = 0.001$, OR = 1,345, PP threshold = 60%) (**Figure 1.4c**), consistent with their involvement in the transcriptional regulation of multiple genes. We also evaluated the enrichment of fetal-like PPC-specific singleton and combinatorial e_gQTLs in adult islet chromatin states [48] (**Figure 1.4c**). No meaningful enrichments were observed for fetal-like PPC-unique combinatorial e_gQTLs, but PPC-unique singleton e_gQTLs were enriched in adult promoter regions ($p = 4.1 \times 10^{-4}$, OR = 28.4, PP threshold = 80%). For example, we observed that the PPC-unique singleton e_gQTL in the *CDC37L1-DT* locus overlapped an active promoter region in the adult islet, while in both adult islet and adult whole pancreas, the variants in the same region are not active (**Figure 1.4d**). Overall, these results show that the e_gQTLs annotated as PPC tissue-unique were enriched in regulatory elements consistent with their proposed functions.

Next, we present three examples of tissue-unique e_gQTL modules that further illustrate context-specificity of regulatory variants in the three pancreatic tissues. We identified the e_gQTL module GE_3_1 ("GE" means that this module is associated with gene expression) as a fetal-unique e_gQTL locus (ch3:148903264-148983264) because the underlying genetic variants were associated with *CP* and *HPS3* expression in only fetal-

like PPC while in adult islets and whole pancreas, the variants were not detected as e_gQTLs (**Figure 1.4e-f**). Similarly, GE_15_13 was an adult islets-unique e_gQTL locus (chr15:57746360-57916360) associated with *GCOM1*, *MYZAP*, and *POLR2M* expression, while in the other two pancreatic tissues, the variants were inactive and not associated with gene expression (**Figure 1.4g-h**). Finally, we discovered GE_5_32 as an adult whole pancreas-unique e_gQTL locus (chr5:146546063-146746063) associated with *STK32A* and *STK32A-AS1* expression in only the adult whole pancreas (**Figure 1.4i-j**). Together, these results show that gene regulation varies between fetal-like and adult pancreatic stages, as well as between the two adult tissues, further demonstrating the importance of profiling multiple contexts of the pancreas to delineate molecular mechanisms underlying pancreatic disease.

1.9 Regulatory plasticity in combinatorial e_gQTLs shared between fetal-like and adult pancreatic tissues

Regulatory elements are known to have context-specific gene interactions [49]. To explore this further, we examined the 983 e_gQTL modules shared between fetal-like PPC and adult pancreatic tissues and determined whether the modules were associated with the same or different eGenes between the two stages. We characterized the eGene overlap in five different ways (**Figure 1.5a**): A) 200 (20.3%) e_gQTL modules were associated with same eGene(s) (range: 1-2) between fetal-like PPC and only one of the two adult pancreatic tissues; B) 305 (31.0%) were associated with the expression of the same eGene(s) (range 1-2) in the fetal-like and both adult tissues; C) 350 (35.6%) were associated with 2-12 eGenes, some of which were shared, but at least one eGene was different between the fetal-like and at least one of the adult tissues (referred to as "partial overlap"); D) 88 (9.0%)

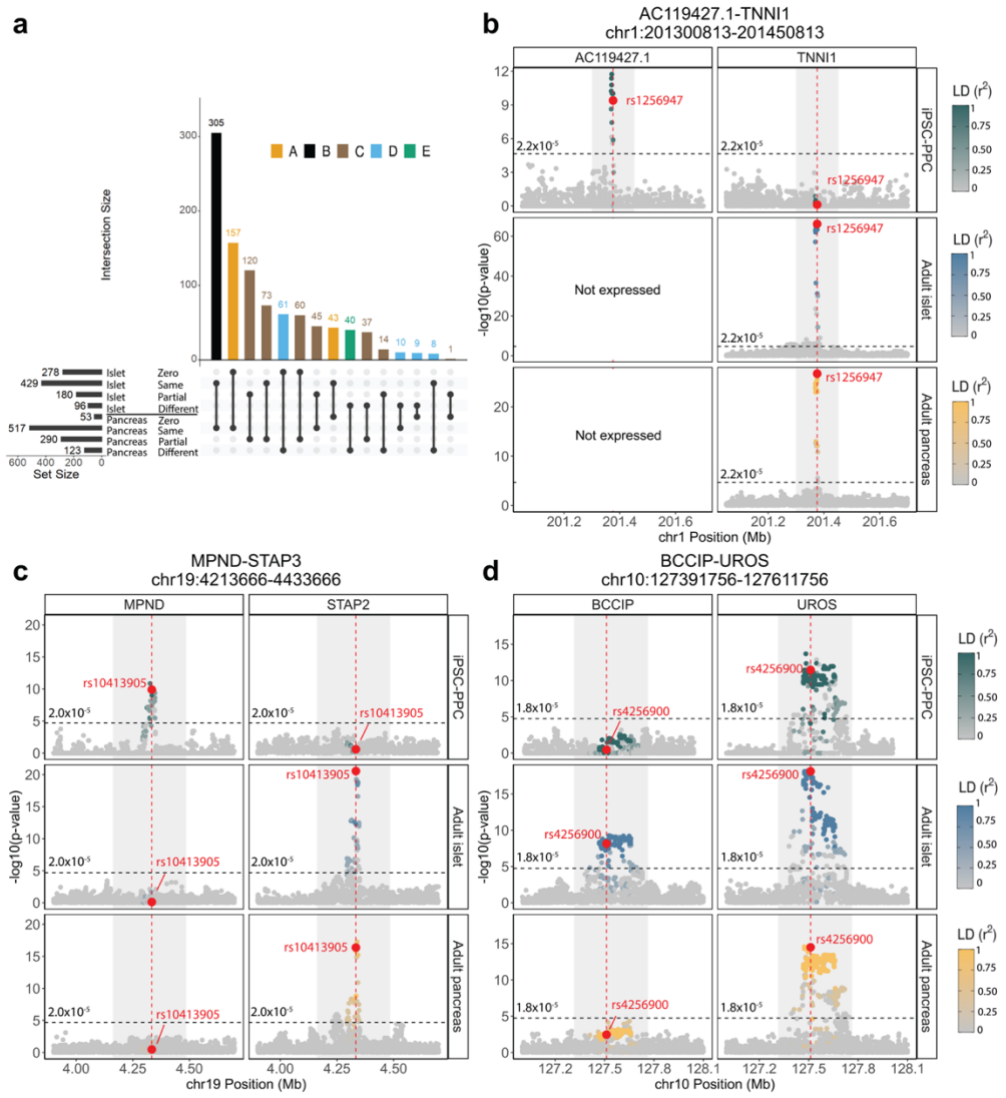
were associated with different eGenes (range: 2-5) between fetal-like PPCs and one of the two adult pancreatic tissues; and E) the remaining 40 (4.1%) were associated with different eGenes (range: 2-7) between the fetal-like and both adult islet and whole pancreas tissues (i.e., there is no overlap of eGenes between the two developmental stages). These data show that 51.3% (505, categories A and B) of the modules shared between the PPCs and adult pancreatic tissues regulated expression of the same genes, while 48.7% (478, categories C-E) displayed spatiotemporal regulatory plasticity.

Here, we illustrate examples of e_gQTL modules in three intervals that display regulatory plasticity between fetal-like and adult states. In the chr1:201300813-201450813 locus, we identified a fetal-adult e_gQTL module (GE_1_163) that comprised e_gQTL associations for different eGenes in PPC and the two adult pancreatic tissues, specifically *AC119427.1* in PPC and *TNNI1* in the two adult tissues (**Figure 1.5b**). Likewise, the chr19:4213666-4433666 locus corresponding to a fetal-adult e_gQTL module (GE_19_90) was associated with *MPND* expression in only PPC but in adult islets and whole pancreas, the underlying variants were associated with *STAP2* expression (**Figure 1.5c**). Finally, the fetal-adult e_gQTL locus (GE_10_11) in chr10:1273918-1276118 affected *UROS* expression in all three pancreatic tissues but in adult islets, the underlying variants also affected *BCCIP* expression (**Figure 1.5d**). Together, these genomic loci illustrate examples of regulatory plasticity observed in genetic variants in which their genotypes incur different impacts on transcriptional activity depending on the life stage of the pancreas.

Altogether, including e_{AS}QTLs, we discovered 655 (478 e_gQTL + 177 e_{AS}QTL modules, categories C-E) shared eQTL loci that displayed regulatory plasticity in which the underlying regulatory variants were associated with one or more different genes and

could thereby affect different biological processes. These 655 shared eQTL loci comprise 1,043 PPC, 934 adult islet, and 1,111 adult whole pancreas eQTL associations.

Figure 1.5 Regulatory Plasticity of eQTLs
(a) Number of e_gQTL modules shared between PPC and at least one adult pancreas tissue categorized by eGene overlap with adult. "Zero" indicates that the module did not contain an e_gQTL in the respective adult tissue. "Same" indicates that the module had e_gQTLs for only the same eGenes in PPC and the adult tissue. "Partial" indicates that the module had e_gQTLs for partially overlapping eGenes between PPC and the adult tissue. "Different" indicates that the module had e_gQTLs for only different eGenes between PPC and the adult tissue. (b-d) Examples of e_gQTL loci demonstrating regulatory plasticity of genetic variation across fetal-like and adult pancreatic stages. Panel b shows a locus strongly associated with *AC119427.1* expression in fetal-like PPC and *TNNI1* expression in adult islet and whole pancreas. Panel c shows a locus associated with *MPND* expression in only fetal-like PPC but *STAP2* expression in both the adult pancreatic tissues. Panel d shows a locus associated with partially overlapping eGenes between the two pancreatic stages (*UROS* in all three pancreatic tissues and *BCCIP* in only adult islets). The X-axis represents variant positions while the Y-axis shows the -log10(eQTL p-value) for the associations between the genotype of the tested variants and gene expression. For plotting purposes, we assigned a single p-value for gene-level significance after Bonferroni-correction (0.05 / the number of independent variants tested in fetal-like PPC; horizontal line). Red vertical lines indicate the positions of the lead candidate causal variants underlying the colocalization based on maximum PP.


a
Intersection Size
305
157
120
73
61
60
45
43
40
37
14
10
9
8
1
A B C D E
278 Islet Zero
429 Islet Same
180 Islet Partial
96 Islet Different
53 Pancreas Zero
517 Pancreas Same
290 Pancreas Partial
123 Pancreas Different
Set Size
b
AC119427.1-TNNI1
chr1:201300813-201450813
AC119427.1
TNNI1
rs1256947
iPSC-PPC
Adult islet
Adult pancreas
Not expressed
-log10(p-value)
LD (r2)
chr1 Position (Mb)
c
MPND-STAP3
chr19:4213666-4433666
MPND
STAP2
rs10413905
chr19 Position (Mb)
d
BCCIP-UROS
chr10:127391756-127611756
BCCIP
UROS
rs4256900
chr10 Position (Mb)

1.10 Associations of spatiotemporal eQTLs with pancreatic traits and disease phenotypes

To better understand the role of regulatory variants associated with complex human traits and disease during early development and adult pancreatic stages, we performed colocalization between GWAS signals and eQTLs (e_gQTL and e_{AS}QTL) detected in fetal-like PPC, adult islets, and adult whole pancreas tissues. For this analysis, we considered GWAS data from ten different studies for two diseases involving the pancreas, including type 1 diabetes (T1D) [3] and type 2 diabetes (T2D) [4], and seven biomarkers related to three traits: 1) glycemic control (HbA1c levels and fasting glucose [FG]) [2,50]; 2) obesity (triglycerides, cholesterol, HDL level, and LDL direct) [50]; and 3) body mass index (BMI) [50].

1.10.1 Singleton eQTLs

Out of the 6,101 singleton eQTLs (3,517 e_gQTLs and 2,584 e_{AS}QTLs) in the fetal-like PPC and two adult pancreatic tissues, we found 118 (1.9%) that displayed strong evidence for colocalization with at least one GWAS signal, including 21 (of 1,518 total singleton eQTLs; 1.4%) fetal-like PPC, 57 (of 2,225; 2.6%) adult islets, and 40 (of 2,358; 1.7%) adult whole pancreas singleton eQTLs (**Figure 1.6a**). Given that some traits were highly correlated with one another[51,52], we observed 38 singleton eQTLs that colocalized with GWAS variants associated with more than one trait (range: 2-6 traits). In total, we identified 183 GWAS loci across the ten traits that colocalized with fetal-like or adult pancreatic singleton eQTLs (each combination of colocalized eQTL-GWAS trait variants was counted as a separate locus). We next identified putative causal variants underlying both eQTL and trait associations using *coloc.abf* [41] and constructed 99% credible sets (i.e.,

set of variants with a cumulative causal PP ≥ 99%; see Methods). Of the total 183 colocalized GWAS loci, we resolved 21 to a single putative causal variant while 63 had between two and ten variants and the remaining 99 had more than ten variants with an average of ~46 variants per locus (**Figure 1.6b**).

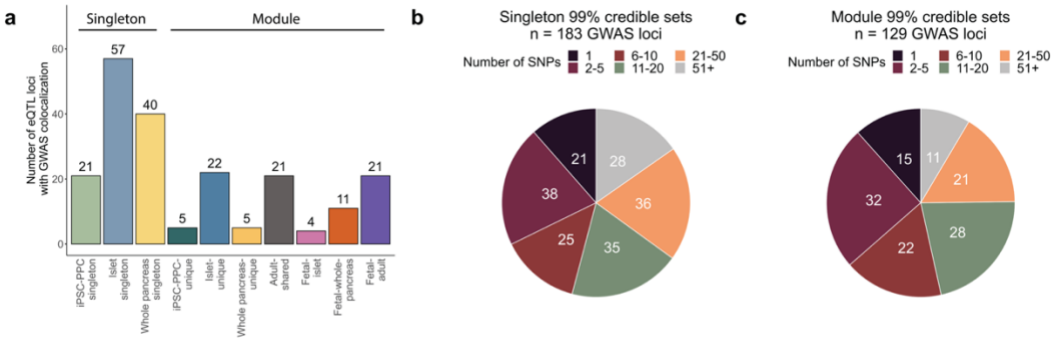


Figure 1.6. Summary of pancreatic GWAS associations
(a) Bar plot showing the number of eQTL loci that colocalized with GWAS variants (PP.H4 ≥ 80%) as a singleton or module. (b) Pie chart showing the number of singleton-colocalized GWAS loci (n=183) color-coded by the number of candidate causal variants identified in their 99% credible sets. (c) Pie chart showing the number of module-colocalized GWAS loci (n=129) color-coded by the number of candidate causal variants identified in their 99% credible sets.

1.10.2 eQTL modules

We next analyzed the combinatorial eQTLs for GWAS colocalization. We considered an eQTL module to overlap with GWAS variants if more than 30% of the eQTLs in the module colocalized with PP.H4 ≥ 80% and the number of H4 associations was twice greater than the number of H3 associations (see Methods). Of the 2,832 (1,852 e_gQTL and 980 e_{AS}QTL) modules, 89 (57 e_gQTL + 32 e_{AS}QTL; 3.1%) colocalized with a total of 129 GWAS loci across the ten traits. Of these 89 GWAS-colocalized modules, 5 were PPC-unique, 36 were shared between both PPC and adult, (4 fetal-islet, 11 fetal-whole-pancreas, and 21 fetal-adult modules), and 48 were associated with only adult (22

islet-unique, 5 whole pancreas-unique, 21 adult-shared) (**Figure 1.6a**). We observed that all 5 PPC-unique eQTL modules corresponded to e_{AS}QTL modules. This finding aligns with multiple studies showing that alternative splicing is more dynamic and extensive in embryonic and fetal stages [53–55]. The 89 modules comprised 49 PPC eQTLs (41 genes), 98 adult islets eQTLs (75 genes), and 71 adult whole pancreas eQTLs (69 genes). To fine-map each of the 129 colocalized GWAS loci, we used the eQTL in the module that resulted in the least number of putative causal variants (see Methods). 15 GWAS loci had a credible set size of one variant, 54 with two to ten variants, and the remaining 60 had more than ten variants and an average of ~32 variants per set (**Figure 1.6c**).

Altogether, these results show complex pancreatic disease and trait GWAS variants colocalized with regulatory variants that were uniquely active in either the fetal-like or adult developmental stages and with regulatory variants shared across the life stage of the pancreas. Furthermore, our data show the utility of using spatiotemporally informed eQTLs for fine-mapping causal variants in GWAS loci.

1.11 Spatiotemporally informed eQTL resource provides mechanistic insights into GWAS signals

To assess the utility of our spatiotemporally informed eQTL resource for interpreting GWAS signals, we initially examined the role of regulatory plasticity in pancreatic disease and traits. We examined the 36 eQTL modules that were shared between fetal-like PPC and the adult pancreatic tissues (i.e., fetal-adult, fetal-islet, and fetal-whole-pancreas) and colocalized with GWAS signals. Thirty of these modules were associated with the same genes (categories A and B), while one was associated with partially overlapping eGenes (category C), and five were associated with entirely different eGenes

(category D and E) (see **Figure 1.5a** for category definitions). These results show that while the function of shared GWAS regulatory variants tended to be conserved across the fetal-like and adult pancreatic stages, a subset (17%, n = 6 of 36 total eQTL modules) were associated with distinct genes between the two stages.

To further assess the utility of our spatiotemporally informed eQTL resource, we next examined GWAS signals that could only be interpreted by including fetal pancreatic eQTLs. We calculated the fraction of GWAS loci that colocalized with only PPC eQTLs, only adult islet eQTLs, and only adult whole pancreas eQTLs. For fair comparisons, we considered only e_gQTLs in this assessment. Of the 191 GWAS loci that colocalized with an e_gQTL, we found that 13% (24 loci) colocalized with 16 PPC-unique e_gQTLs, 25% (47) with 27 adult islet-unique e_gQTLs, and 28% (53) with 46 adult whole pancreas-unique e_gQTLs. The remaining 35% (67) GWAS loci colocalized with 121 e_gQTLs shared between multiple tissues. We next determined how many of the 16 PPC-unique e_gQTLs were active in 48 non-pancreatic tissues in the GTEx study [10]. We calculated LD ($r2 > 0.2$ within 500 Kb or within 500 Kb if LD information was not available) between the lead variants of each of the 16 PPC-unique e_gQTL and each e_gQTL for the non-pancreatic tissues in GTEx. We identified 8 (31% of 16, all singletons) that were independent and exclusive to the fetal-like PPC dataset (i.e., did not have LD). One of these 8 PPC e_gQTLs (*TPD52* e_gQTL) is described below in further detail. These results show that integrating fetal-like PPC eQTLs can help resolve certain GWAS loci that cannot be resolved using only adult datasets. Below, we demonstrate the application of our spatiotemporally informed eQTL resource by providing a detailed description of eight GWAS loci. We propose potential causal mechanisms and offer insights into their spatiotemporal contexts.

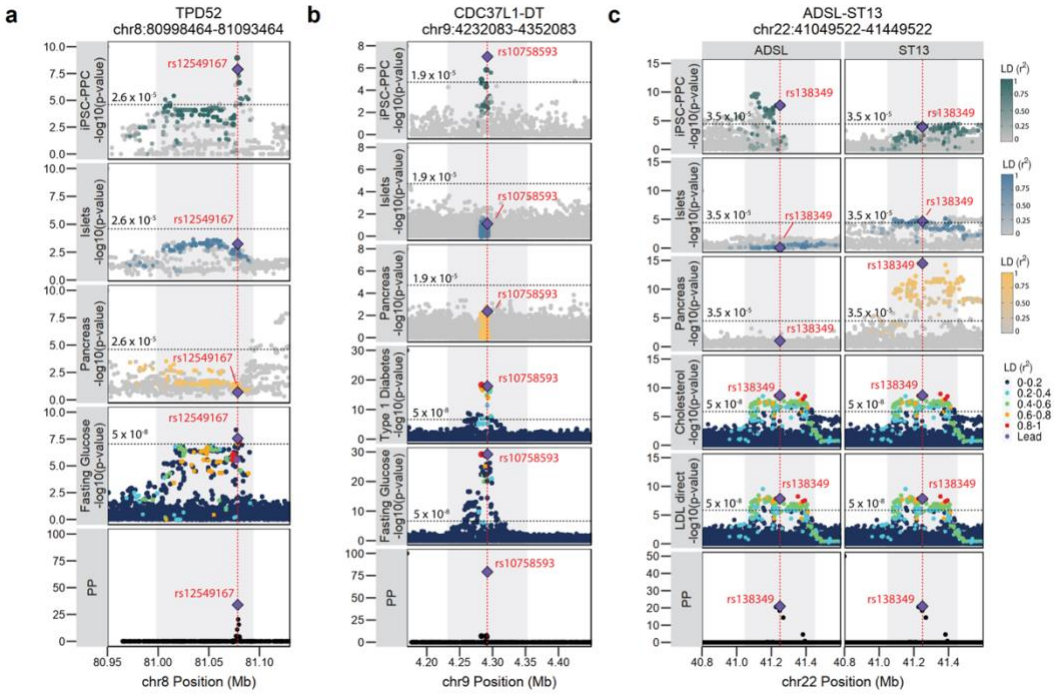


Figure 1.7 Pancreatic GWAS associations with fetal-specific and adult-shared gene Expression

(a) The TPD52 locus is associated with fasting glucose levels and colocalized with a fetal-like PPC-unique singleton e_gQTL with the predicted causal variant identified as rs12549167 (chr8:81078464:C>T, PP = 33.9%). (b) The CDC37L1-DT locus is associated with fasting glucose and type 1 diabetes and colocalized with an PPC-unique singleton e_gQTL with the predicted causal variant identified as rs10758593 (chr9:4292083:G>A, PP = 79.2%). (c) Cholesterol and LDL direct GWAS loci colocalize with a fetal-adult e_gQTL module where the variants are strongly associated with *ADSL* expression in PPC and *ST13* expression in the adult whole pancreas (also weakly associated with *ST13* expression in the adult islets). The predicted causal variant was identified as rs138349 (chr22:41249522:A>G, PP = 21.9%). For plotting purposes, we assigned a single p-value for gene-level significance based on Bonferroni-correction (0.05 divided by the number of independent variants tested in fetal-like PPC; horizontal line). We note that this p-value does not reflect the thresholds used to define a significant eQTL in the original adult studies [10,11]. Therefore, while the *ST13* eQTL in adult islets in panel c is below the horizontal line, it had an FDR < 1% in the original study [11]. In each panel, the X-axis represents variant positions while the Y-axis either shows the -log10(eQTL p-value) for the associations between the genotype of the tested variants and gene expression or the -log10(GWAS p-value) for the associations between the tested variants and the GWAS trait. For GWAS significance, we used -log10($5x10^{-8}$). Red vertical lines indicate the positions of the lead candidate causal variants underlying the colocalization based on maximum PP. For loci that colocalized with multiple GWAS traits, we used the credible set that yielded the smallest number of variants to plot the "PP" fine-mapping panel.

1.11.1 chr8:80998464-81093464

We found that in the chr8:80998464-81093464 locus, a GWAS signal associated with FG levels colocalized with a fetal-like PPC-unique singleton e_gQTL for *TPD52*, also known as tumor protein D52 (effect size = -0.99, PP.H4 = 91.7%) (**Figure 1.7a**). The reported causal variant underlying this GWAS signal is rs12541643 [2]; however, colocalization with our eQTLs identified rs12549167 (chr8:81078464:C>T, PP = 33.9%, r^2 = 0.317 with rs12541643) as the most likely candidate causal variant underlying both *TPD52* expression in fetal-like PPC and FG association. TPD52 is a direct interactor with the AMP-activated protein kinase (AMPK) and negatively affects AMPK signaling. AMPK controls a wide range of metabolic processes and is responsible for maintaining cellular energy homeostasis particularly in tissues associated with obesity, insulin resistance, T2D, and cancer such as muscle, liver, hypothalamus, and the pancreas [56–59]. Dysregulation of AMPK has also been associated with developmental defects in which AMPK activation can lead to fetal malformation [60]. Our findings suggest that decreased expression of *TPD52* during development may influence changes in glucose metabolism and therefore fasting glucose levels during adult stage.

1.11.2 chr9:4232083-4352083

We found that the well-known *GLIS3* GWAS locus associated with FG and T1D-risk [61,62] colocalized with a fetal-like PPC-unique singleton e_gQTL for the lncRNA CDC37L1 divergent transcript (*CDC37L1-DT*; effect size = 1.46; PP.H4 for FG and T1D = 92.4% and 91.2%, respectively, **Figure 1.7b**). Consistent with previous studies [61,62], we identified rs10758593 (chr9:4292083:G>A, PP = 79.2%) as the lead candidate causal variant underlying both eQTL and GWAS associations. Because *GLIS3* plays a critical role

in pancreatic beta cell development and function [58,59,60], it has often been reported as the susceptibility gene for this signal, however it remains unclear what effects rs10758593 has on *GLIS3* expression. Our analysis suggests that another potential gene target of rs10758593, specifically during pancreas development, is *CDC37L1-DT*. While the molecular function of *CDC37L1-DT* is unknown, the gene has been associated with 9p duplication in neurodevelopmental disorders [66]. Furthermore, a recent study observed a significant association between the rs10758593 risk allele and birth weight, indicating a development role played by this locus [67]. Although additional studies are needed to understand the function of *CDC37L1-DT* during pancreas development and in T1D pathology, our analysis indicates that *CDC37L1-DT* may be another candidate susceptibility gene for the variants in the *GLIS3* locus. Assessment of *GLIS3* e_gQTLs in the three pancreatic tissues showed that there was no overlap between the e_gQTLs and GWAS variants.

1.11.3 chr22:41049522-41449522

We found that the GWAS signals associated with cholesterol and LDL direct levels in the chr22:41049522-41449522 locus colocalized with a "fetal-adult" e_gQTL module (module ID: GE_22_63, category E) (**Figure 1.7c**). The module was associated with different eGenes between fetal-like PPC and both adult pancreatic tissues, in which the GWAS variants were associated with *ADSL* expression in PPC (effect size = 0.78) but *ST13* expression in adult whole pancreas (effect size = 0.27) and adult islets (effect size = -0.15, weakly associated). Infants born with ADSL (adenylosuccinate lyase) deficiency suffer from impaired glucose and lipid metabolism, while ST13, also known as Hsc70-interacting protein, is involved in lipid metabolism [68]. Overexpression of *ST13* was found to result in

disordered lipid metabolism in chronic pancreatitis [68]. Although *ST13* was reported to be the candidate causal gene for this locus [69], we determined that the underlying variants may also affect *ADSL* expression but specifically during early pancreas development. Congruent with a previous study [69], our colocalization identified rs138349 (chr22:41249522:A>G, PP = 21.9% for cholesterol and 20.9% for LDL) as the lead candidate causal variant for the e_gQTLs and both cholesterol and LDL GWAS associations. Altogether, annotation of the chr22:41049522-41449522 GWAS locus using our pancreatic eQTL resource suggests that altered expression of *ADSL* during pancreas development and *ST13* in adult pancreatic tissues may contribute to changes in cholesterol and LDL direct levels in adult. Additional studies are required to understand the degree to which *ADSL* and *ST13* are causal for cholesterol and LDL direct levels.

1.11.4 chr10:90001035-90066035

We found a T1D-risk signal in the chr10:90001035-90066035 locus that colocalized with an "adult whole pancreas-unique" e_gQTL module (module ID: GE_10_35) associated with *PTEN* and *LIPJ* expression in adult whole pancreas (effect size = 0.48 and 0.49, respectively) (**Figure 1.8**). Colocalization identified the distal regulatory variant rs7068821 (chr10:90051035:G>T; PP = 85.5%) as the most likely candidate causal variant, which is in LD with the reported index SNP rs10509540 (r^2 = 0.876) in the GWAS catalogue. While *RNLS* was reported to be the susceptibility gene for this locus [70], our analysis suggests that both *PTEN* and *LIPJ* may be candidate causal genes for this locus. Previous studies have shown that knockout of pancreas-specific *PTEN* (PPKO) in mice resulted in enlarged pancreas and elevated proliferation of acinar cells. PPKO mice also exhibited hypoglycemia, hypoinsulinemia, and altered amino metabolism [71]. LIPJ encodes

the lipase family member J and is involved in lipid metabolism [72]. Our findings provide additional biological insight into the chr10:900001035-90066035 T1D locus and support previous studies suggesting a potential causal role of the adult whole pancreas in T1D pathogenesis [3,67].

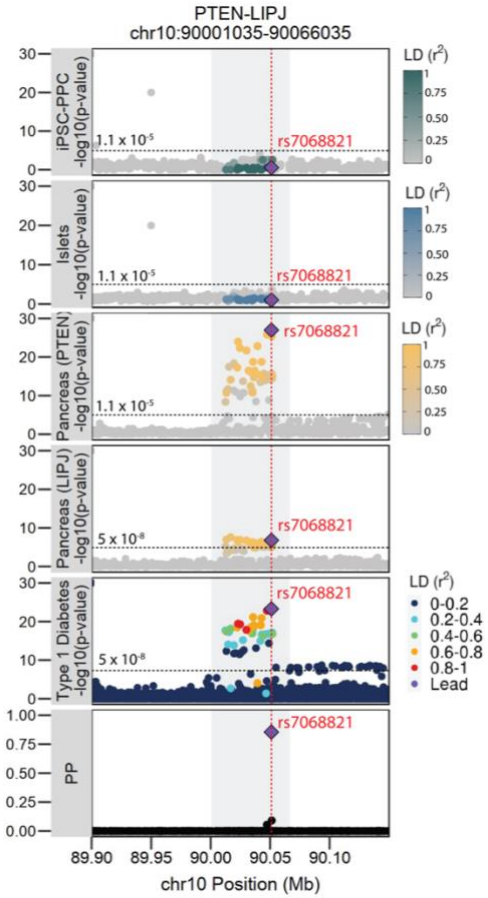


Figure 1.8 *PTEN* and *LIPJ* e_gQTL Associations with GWAS
The GWAS signal associated with T1D-risk in the chr10:90001035-90066035 locus colocalized with an adult whole pancreas-unique e_gQTL module containing e_gQTLs for *PTEN* and *LIPJ*. Each point in these plots represents an individual SNP color-coded by LD with the lead candidate causal variant highlighted in purple. The bottom plot in panel **b** shows the posterior probability (PP) of association for each variant being causal for both e_gQTL and GWAS associations. For plotting purposes, we assigned a single p-value for gene-level significance based on Bonferroni-correction (0.05 divided by the number of variants tested for the gene; horizontal line) for the e_gQTL signals. For PPC and adult islet in panel **b**, we overlaid eQTL associations for *PTEN, LIPJ*, and nearby genes to show that the locus was not associated with gene expression in the tissues. For GWAS signals, we used p-value = $5x10^{-8}$ to indicate genome-wide significance. Red vertical lines indicate the positions of the lead candidate causal variants underlying GWAS and eQTL colocalization based on maximum PP.

1.11.5 chr14:101286447-101326447

The chr14:101286447-101326447 is a well-known GWAS locus associated with T1D and has been reported to affect the lncRNA maternally expressed gene 3 (*MEG3*). While the role of *MEG3* in T1D and T2D pathogenesis has been extensively studied [73–75], the genetic mechanism by which this locus affects *MEG3* expression and therefore, T1D-risk is not well understood. Using our pancreatic eQTL resource, we found that the GWAS signal colocalized with a fetal-like PPC-unique singleton e_{AS}QTL for a *MEG3* isoform (ENST00000522618, PP.H4 = 98%, effect size = 1.3, **Figure 1.9a**). Colocalization with the *MEG3* e_{AS}QTL identified rs56994090 (chr14:101306447:T>C, PP = 100%) as the most likely candidate causal variant, which is concordant with the findings of a previous GWAS study [76]. Given that rs56994090 is located in the novel intron enhancer of *MEG3* [74], we hypothesize that alternative splicing of *MEG3* may alter the enhancer's regulatory function, as previously observed in other lncRNAs [77], and thereby, affect T1D-risk. Altogether, our findings describe a potential causal mechanism for the T1D-risk locus involving differential alternative splicing of *MEG3* specifically during pancreas development.

1.11.6 chr16:684685635-68855635

We determined a known GWAS signal in the chr16:684685635-68855635 locus associated with HbA1c levels [78] colocalized with a fetal-like PPC-unique singleton e_{AS}QTL for the P-cadherin 3 (*CDH3*) isoform ENST00000429102 (effect size = -1.6, PP.H4 = 83.1%) (**Figure 1.9b**). Colocalization using the e_{AS}QTL identified intronic variant rs72785165 (chr16:68755635:T>A, PP = 6.8%) as the most likely candidate causal variant, which is in high LD with the reported GWAS SNP (rs4783565, r^2 = 0.88) [78]. While it remains unclear how alternative splicing of *CDH3* affects HbA1c levels, studies have

shown that chimeric proteins made of cadherin ectodomains, including the P-cadherin CDH3, are important for proper insulin secretion by pancreatic beta cells [79]. Based on our findings, we hypothesize that differential isoform usage of *CDH3* during pancreas development may influence glucose control and therefore, HbA1c levels, in adults.

1.11.7 chr13:30956642-31116642

The GWAS signals associated with T2D and BMI in the chr13:30956642-31116642 locus [80–83] colocalized with the PPC-unique e_{AS}QTL module (module ID: AS_13_2) associated with three *HMGB1* isoforms: ENST00000326004, ENST00000399494 , ENST00000339872, and (effect size = 2.16, -2.26, and -0.85, respectively; HMGB1.1, HGMB1.2, and HMGB1.3, respectively) (**Figure 1.9c**). Our colocalization identified rs3742305 (chr13:31036642:C>G, PP = 49.3%) as a lead candidate causal variant underlying this locus, in which the risk allele (G) was associated with increased usage of ENST00000326004 and decreased usages of ENST0000339872 and ENST00000399494. While a previous study [82] also reported *HMGB1* as the susceptibility gene, the precise mechanism by which rs3742305 affected *HMGB1* expression was unclear. HMGB1, also known as high-mobility group box 1, is an important mediator for regulating gene expression during both developmental and adult stages of life. Deletion of *HMGB1* disrupts cell growth and causes lethal hypoglycemia in mouse pups [84]. In T2D, *HMGB1* promotes obesity-induced adipose inflammation, insulin resistance, and islet dysfunction [84]. Our results suggest that differential usage of *HMGB1* isoforms during pancreas development may affect adult risk of developing obesity and/or T2D.

Altogether, our findings demonstrate the value of our pancreatic eQTL resource to annotate GWAS risk variants with fetal-like and adult temporal and spatial regulatory

information. We show that some causal regulatory variants underlying disease-associated signals may influence adult traits by modulating the expression of genes in early development, while in other cases, they may display regulatory plasticity and exert their effects by modulating the expression of multiple different genes in fetal-like and adult pancreatic stages. Further, we identified an association between whole pancreas and T1D, supporting a potential role of this tissue in diabetes pathogenesis [3].

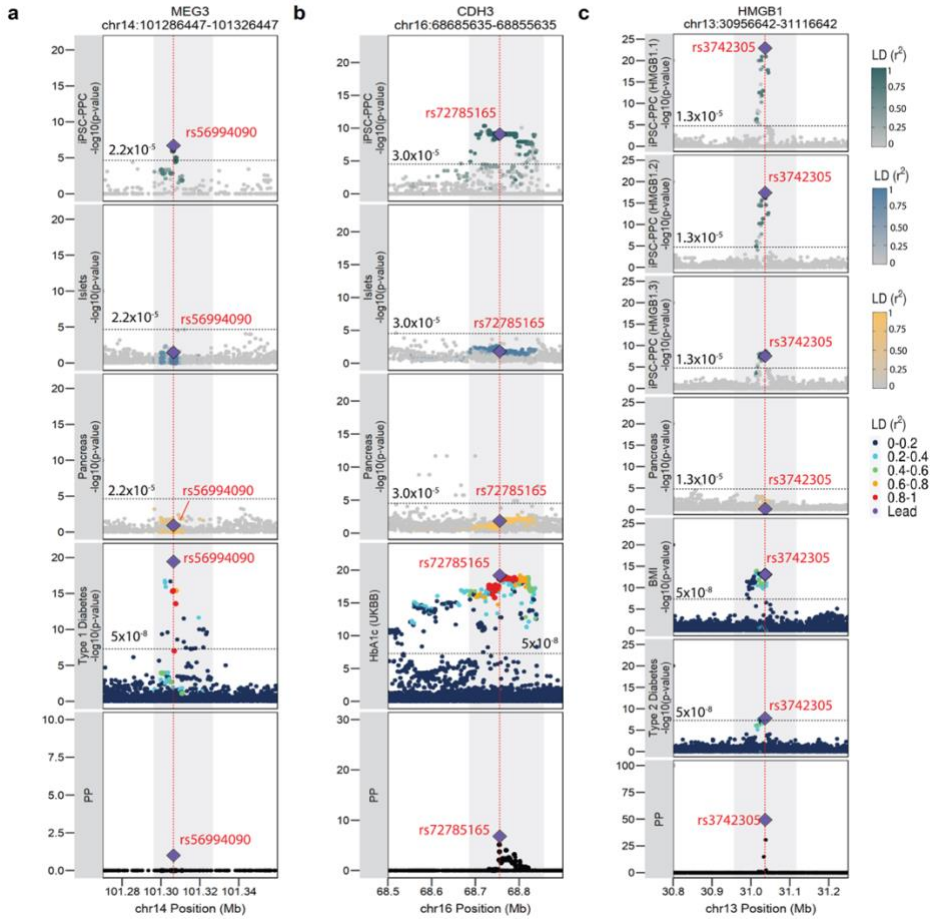


Figure 1.9 Pancreatic GWAS associations with fetal-specific alternative splicing
(a) T1D-risk locus colocalized with a fetal-like PPC-unique singleton e_{AS}QTL for *MEG3* with the predicted causal variant rs56994090 (chr14:101306447:T>C, PP = 100%). (b) GWAS locus associated with HbA1c colocalized with an PPC-unique singleton e_{AS}QTL for *CDH3* with the predicted causal variant rs72785165 (chr16:68755635:T>A, PP = 6.8%). (c) GWAS locus associated with T2D-risk and BMI colocalized with an PPC-unique e_{AS}QTL module (AS_13_2) for differential usage of three *HMGB1* isoforms with a predicted causal variant rs3742305 (chr13:31036642:C>G, PP = 49.3%). In each panel, the X-axis represents variant positions while the Y-axis either shows the -log10(eQTL p-value) for the associations between the genotype of the tested variants and gene expression or the -log10(GWAS p-value) for the associations between the tested variants and the GWAS trait. For GWAS significance, we used -log10($5x10^{-8}$). For eQTL significance, we used a single p-value for gene-level significance after Bonferroni-correction (0.05 / the number of independent variants tested in fetal-like PPC; horizontal line). Red vertical lines indicate the positions of the lead candidate causal variants underlying the colocalization based on maximum PP. For loci that colocalized with multiple GWAS traits, we used the credible set that yielded the smallest number of variants to plot the "PP" fine-mapping panel.

1.12 Discussion

In this study, we leveraged one of the most well-characterized iPSC cohorts comprising >100 genotyped individuals to derive pancreatic progenitor cells and generate a comprehensive eQTL resource for examining genetic associations with gene expression and isoform usage in fetal-like pancreatic cells. We discovered 8,665 eQTLs in the fetal-like PPCs and showed that 60% of eGenes were associated with regulatory variation uniquely active during pancreas development. For the eGenes that were shared with adult, ~12% were associated with different genomic loci, indicating that different regulatory elements may modulate the same gene in fetal-like and adult pancreatic stages. We further identified regulatory variants that displayed early pancreas development-unique function, of which 1,805 were uniquely active in only PPC and 1,043 were active in both developmental and adult contexts but exhibited regulatory plasticity in the genes they regulate. These results concur with previous studies showing that the genetic regulatory landscape changes between fetal tissues and their adult counterparts [86–88], and therefore, highlights the importance of assessing variant function in both fetal and adult tissue contexts. Furthermore, it is widely known that tight regulation of genes during development is essential [89], and our study reflects this in our findings that the majority of developmental-unique eQTLs were restricted to a single eGene.

Finally, we highlighted examples of GWAS associations for which we utilized our spatiotemporally informed eQTL resource to characterize causal risk mechanisms underlying adult pancreatic disease. We showed that some causal regulatory variants underlying GWAS signals identified in the fetal-like PPCs modulate the expression of genes in early development, while others may exert their effects by modulating the

expression of multiple different genes across fetal-like and adult pancreatic stages. Of note, many of the fetal-unique regulatory variants underlying the GWAS signals were e_{AS}QTLs, which is consistent with alternative splicing playing a key role in developing tissues [53–55,90]. Hence, we believe that contribution of alternative splicing differences during fetal pancreas development to complex traits warrants further investigation.

We offer limitations in our study and potential future directions for the field at large. We believe that studies using larger sample sizes are needed to identify additional associations between genetic variation and gene expression in fetal samples. Our eQTL mapping in fetal-like PPC was conducted on much fewer samples compared to the two adult studies that each used ~300-400 samples, rendering our dataset underpowered and not able to capture additional eQTL associations that could be shared with the adult pancreatic tissues. Therefore, several eQTLs we annotated as adult islet-unique or whole pancreas-unique may in reality be shared with fetal pancreas. Further, power differences between the studies may also cause the observed results where there were many singleton eQTLs observed in a single tissue. On the other hand, the eQTLs we annotated as PPC-unique may less likely to be shared, as the signals in the adult datasets are better powered and therefore sufficient for comparing against PPC signals. Additionally, with the rapid generation of eQTL datasets from different tissue contexts [1,2], the development and application of artificial intelligence and machine learning as ways to identify shared eQTL associations between multiple tissues will be extremely useful. While pairwise colocalization and network analysis can identify shared eQTL regulatory loci across a handful of tissues, machine learning approaches could scale these analyses across spatiotemporal contexts of all tissues, thereby providing valuable insights into regulatory

elements that are exclusive to a specific context and also those that exhibit regulatory plasticity across multiple contexts.

In summary, our study provides a valuable resource for discovering causal regulatory mechanisms underlying pancreatic traits and disease across developmental and adult time points of the pancreas. We reveal that disease variants may either display temporal-specificity in which they affect gene expression specifically in one timepoint, or regulatory plasticity, in which they affect gene expression in multiple timepoints but affect different genes. Our findings lay the groundwork for future employment of development contexts for the characterization of disease-associated variants.

1.13 Materials and Methods

Subject Information

We used iPSC lines from 106 individuals recruited as part of the iPSCORE project. There were 53 individuals belonging to 19 families composed of two or more subjects (range: 2-6). Each subject was assigned an iPSCORE_ID (i.e., iPSCORE 4_1), where "4" indicates the family number and "1" indicates the individual number, and a 128-bit universal unique identifier (UUID). The 106 individuals included 68 females and 38 males with ages ranging from 15 to 88 years old at the time of enrollment. Recruitment of these individuals was approved by the Institutional Review Boards of the University of California, San Diego, and The Salk Institute (project no. 110776ZF).

WGS data

Whole-genome sequencing data for the 106 iPSCORE individuals were downloaded from dbGaP (phs001325.v3) as a VCF file [37]. We retained variants with MAF > 5% across all 273 individuals in the iPSCORE resource, that were in Hardy-Weinberg

equilibrium ($p > 10^{-6}$), and that were within 500 Kb of the expressed gene's body coordinates. Specifically, we expanded the coordinates of each of the 16,464 expressed autosomal genes (500 Kb upstream and downstream) and extracted all variants within these regions using *bcftools view* with parameters *--f PASS -q 0.05:minor* [92]. Next, we normalized indels and split multi-allelic variants using *bcftools norm -m-* and removed variants that were genotyped in fewer than 99% of samples using *bcftools filter -i 'F_PASS(GT!="mis") > 0.99* [92]. Finally, we converted the resulting VCF files to text using *bcftools query* [92] and converted the genotypes from character strings (0/0, 0/1, and 1/1) to numeric (0, 0.5, and 1, respectively). This resulted in 6,593,484 total variants used for eQTL mapping.

iPSC Generation

Generation of the 106 iPSC lines has previously been described in detail [35]. Briefly, cultures of primary dermal fibroblast cells were generated from a punch biopsy tissue [93], infected with the Cytotune Sendai virus (Life Technologies) per manufacturer's protocol to initiate reprogramming. Emerging iPSC colonies were manually picked after Day 21 and maintained on Matrigel (BD Corning) with mTeSR1 medium (Stem Cell Technologies). Multiple independently established iPSC clones (i.e. referred to as lines) were derived from each individual. Many of the iPSC lines were evaluated by flow cytometry for expression of two pluripotent markers: Tra-1-81 (Alexa Fluor 488 anti-human, Biolegend) and SSEA-4 (PE anti-human, Biolegend) [35]. Pluripotency was also examined using PluriTest-RNAseq [35]. This iPSCORE resource was established as part of the Next Generation Consortium of the National Heart, Lung and Blood Institute and is available to researchers through the biorepository at WiCell Research Institute

(www.wicell.org; NHLBI Next Gen Collection). For-profit organizations can contact the corresponding author directly to discuss line availability.

Pancreatic Progenitor Differentiation

We performed pancreatic progenitor cell (PPC) differentiation on each of the 106 iPSC lines. One iPSC line was differentiated twice giving a total of 107 differentiations. Each differentiation was assigned a 128-bit universally unique identifier (UUID), and a unique differentiation ID (UDID; "PPCXXX"), where "XXX" represents a numeric integer.

Differentiation Protocol: The iPSC lines were differentiated into PPCs using the STEMdiff™ Pancreatic Progenitor Kit (StemCell Technologies) protocol with minor modifications. Briefly, iPSC lines were thawed into mTeSR1 medium containing 10 µM Y-27632 ROCK Inhibitor (Selleckchem) and plated onto one well of a 6-well plate coated with Matrigel. iPSCs were grown until they reached 80% confluency [94] and then passaged using 2mg/ml solution of Dispase II (ThermoFisher Scientific) onto three wells of a 6-well plate (ratio 1:3). To expand the iPSC cells for differentiation, iPSCs were passaged a second time onto six wells of a 6-well plate (ratio 1:2). When the iPSCs reached 80% confluency, cells were dissociated into single cells using Accutase (Innovative Cell Technologies Inc.) and resuspended at a concentration of 1.85×10^6 cells/ml in mTeSR medium containing 10 µM Y-27632 ROCK inhibitor. Cells were then plated onto six wells of a 6-well plate and grown for approximately 16 to 20 hours to achieve a uniform monolayer of 90-95% confluence (3.7×10^6 cells/well; about 3.9×10^5 cells/cm^2). Differentiation of the iPSC monolayers was initiated by the addition of the STEMdiff™ Stage Endoderm Basal medium supplemented with Supplement MR and Supplement CJ

(2 ml/well) (Day 1, D1). The following media changes were performed every 24 hours following initiation of differentiation (2 ml/well). On D2 and D3, the medium was changed to fresh STEMdiff™ Stage Endoderm Basal medium supplemented with Supplement CJ. On D4, the medium was changed to STEMdiff™ Pancreatic Stage 2-4 Basal medium supplemented with Supplement 2A and Supplement 2B. On D5 and D6, the medium was changed to STEMdiff™ Pancreatic Stage 2-4 Basal medium supplemented with Supplement 2B. From D7 to D9, the medium was changed to STEMdiff™ Pancreatic Stage 2-4 Basal medium supplemented with Supplement 3. From D10 to D14, the medium was changed to STEMdiff™ Pancreatic Stage 2-4 Basal medium supplemented with Supplement 4. On D15, cells were dissociated with Accutase and then collected, counted, and processed for data generation. PPC cells were cryopreserved in CryoStor® CS10 (StemCell Technologies).

PPC Differentiation Efficiency: To evaluate the efficiency of PPC differentiation, we performed flow cytometry on two pancreatic precursor markers, PDX1 and NKX6-1. Specifically, at least 2 x 10^6 cells were fixed and permeabilized using the Fixation/Permeabilized Solution Kit with BD GolgiStop TM (BD Biosciences) following the manufacturer's recommendations. Cells were resuspended in 1x BD Perm/Wash TM Buffer at a concentration of 1 x 10^7 cells/ml. For each flow cytometry staining, 2.5 x 10^5 cells were stained for 75 minutes at room temperature with PE Mouse anti-PDX1 Clone-658A5 (BD Biosciences; Catalog no. 562161; 1:10) and Alexa Fluor® 647 Mouse anti-NKX6.1 Clone R11-560 (BD Bioscience; Catalog no. 563338; 1:10), or with the appropriate class control antibodies: PE Mouse anti-IgG1 κ R-PE Clone MOPC-21 (BD Biosciences; Catalog no. 559320) and Alexa Fluor® 647 Mouse anti IgG1 κ Isotype Clone

MOPC-21 (BD Biosciences; Catalog no. 557732). PE Mouse anti-PDX1 Clone-658A5 and Alexa Fluor® 647 Mouse anti-NKX6.1 Clone R11-560 were validated by the manufacturer to bind to mouse and human PDX-1 and NKX6-1, respectively. Stained cells were washed three times, resuspended in PBS containing 1% BSA and 1% formaldehyde, and immediately analyzed using FACS Canto II flow cytometer (BD Biosciences). The fraction of PDX1- and NKX6-1-positive was calculated using FlowJo software version 10.4.

scRNA-seq

To characterize the cellular composition of the fetal-like PPC samples, we performed single-cell RNA-seq (scRNA-seq) on one iPSC line (from differentiation PPC034) and ten PPC samples with varying percentages of double-positive PDX1+/NKX6-1+ cells based on flow cytometry (range: 9.4-91.7%). Because bulk RNA-seq was generated on cryopreserved cells, we sought to also examine whether cell cryopreservation affects gene expression estimates using scRNA-seq. Therefore, we included both freshly prepared (i.e., not frozen and processed immediately after differentiation) and cryopreserved cells for four PPC samples (PPC029, PPC027, PPC023, PPC034) for scRNA-seq processing.

Sample Collection: Fresh cells from the iPSC line and seven PPC samples were captured individually at D15. Cells from four of these same PPC samples that had been cryopreserved were pooled and captured immediately after thawing (RNA_Pool_1). Cells from an additional three PPC samples were captured only after cryopreservation (RNA_Pool_2).

Library Preparation and Sequencing*:* All single cells were captured using the 10X Chromium controller (10X Genomics) according to the manufacturer's specifications and manual (Manual CG000183, Rev A). Cells from each scRNA-seq sample (one iPSC, seven fresh PPCs, RNA_Pool_1, and RNA_Pool_2) were loaded each onto an individual lane of a Chromium Single Cell Chip B. Libraries were generated using Chromium Single Cell 3' Library Gel Bead Kit v3 (10X Genomics) following manufacturer's manual with small modifications. Specifically, the purified cDNA was eluted in 24 μl of Buffer EB, half of which was used for the subsequent step of the library construction. cDNA was amplified for 10 cycles and libraries were amplified for 8 cycles. All libraries were sequenced on a HiSeq 4000 using custom programs (fresh: 28-8-175 Pair End and cryopreserved: 28-8-98 Pair End). Specifically, eight libraries generated from fresh samples (one iPSC and seven PPC samples) were pooled together and loaded evenly onto eight lanes and sequenced to an average depth of 163 million reads. The two libraries from seven cryopreserved lines (RNA_Pool_1 and RNA_Pool_2) were each sequenced on an individual lane to an average depth of 265 million reads. In total, we captured 99,819 cells. We observed highly correlated cell type proportions between fresh and cryopreserved PPC samples.

scRNA-seq Alignment: We obtained FASTQ files for the ten scRNA-seq samples (one iPSC, seven fresh PPCs, RNA_Pool_1, and RNA_Pool_2) and used CellRanger V6.0.1 (https://support.10xgenomics.com/) with default parameters and GENCODE version 34 hg19 [95] gene annotations to generate single-cell gene counts and BAM files for each of the ten samples.

Dataset Integration and Quality Control: We processed the single-cell gene counts by first aggregating the iPSC and seven fresh PPC samples using the *aggr* function on

CellRanger V6.0.1 with normalization = F. Then, we integrated the aggregated dataset ("aggr") with the two pools of cryopreserved samples (RNA_Pool_1 and RNA_Pool_2) using the standard integration workflow described in Seurat (version 3.2; https://satijalab.org/seurat/archive/v3.2/integration.html). Specifically, for each dataset (aggr, RNA_Pool_1, and RNA_Pool_2), we log-normalized the gene counts using *NormalizeData* (default parameters) then used *FindVariableFeatures* with *selection.method = "vst", nfeatures = 2000,* and *dispersion.cutoff = c(0.5, Inf)* to identify the top 2,000 most variable genes in each dataset. We then used *FindIntegrationAnchors* and *IntegrateData* with *dims = 1:30* to integrate the three datasets. We scaled the integrated data with *ScaleData*, performed principal component analysis with *RunPCA* for *npcs = 30*, and processed for UMAP visualization (*RunUMAP* with *reduction = "pca"* and *dims = 1:30*). Clusters were identified using *FindClusters* with default parameters.

To remove low-quality cells, we examined the distribution of the number of genes per cell and the percentage of reads mapping to the mitochondrial chromosome (chrM) in each cluster. We removed the cluster (11,677 cells) with fewer than 500 genes per cell and more than 50% of the reads mapping to chrM. We re-processed the filtered data (*ScaleData*, *RunPCA, FindClusters, RunUMAP*) and removed another cluster of cells that had the lowest median number of expressed genes (723 versus 2,775) and highest median fraction of mitochondrial reads (34.0% versus 8.39%). After this second filtering step, we retained 84,258 cells.

Demultiplexing Sample Identity: We used Demuxlet [96] to assign pooled cryopreserved cells in RNA_Pool_1 and RNA_Pool_2 (19,136 cells in total) to the correct PPC sample. Specifically, we provided CellRanger BAM files and a VCF file containing

genotypes for biallelic SNVs located at UTR and exon regions on autosomes as annotated by GENCODE version 34 hg19 [95]. We excluded 33 cells that were incorrectly assigned to samples not associated with the pooled sample (i.e., cells from RNA_POOL_1 were predicted to be from other samples not in RNA_Pool_1). 84,225 cells remained for downstream analyses.

Annotation of Cell Type Clusters: We annotated the scRNA-seq clusters by first clustering with three different resolutions (0.5, 0.08, and 0.1). We selected resolution = 0.08 because it best captured the expected PPC cell types based on each cluster's expression for the following gene markers: *POU5F1* (iPSC), *COL1A1*, *COL1A2* (mesendoderm) *AFP, APOA* (early definitive endoderm), *GATA4*, *GATA6, PDX1* (early PPC), *PDX1*, *NKX6-1* (late PPC), *PAX6*, *CHGA*, *INS*, *GCG*, *SST* (endocrine), and *FLT1* (early ductal). We validated our annotations by comparing the PPC clusters to those identified from scRNA-seq of ESC-PPC samples over 4 different stages of differentiation [97] (GSE114412): Stage 3 (Day 6; 7,982 cells), Stage 4 (Day 13; 6,960 cells), Stage 5 (Day 18; 4,193 cells), and Stage 6 (Day 25; 5,186 cells). Specifically, we compared the expression patterns of the gene markers between the clusters using z-normalized mean expression computed on cells expressing at least 1% of maximal expression for the gene, as described in the reference study [97].

Differentially Expressed Genes: To identify differentially expressed genes for each PPC cluster, we used the *FindAllMarkers* function in Seurat [98] with *logfc.threshold = 0.01* and *min.pct = 0.01*. P-values were automatically adjusted by Seurat using Bonferroni correction, and genes with adjusted p-values ≤ 0.05 were considered differentially expressed.

Bulk RNA-seq

Library Preparation and Sequencing: RNA was isolated from total-cell lysates using the Quick-RNA™ MiniPrep Kit (Zymo Research) with on-column DNAse treatments. RNA was eluted in 48 µl RNAse-free water and analyzed on a TapeStation (Agilent) to determine sample integrity. All PPC samples had RNA integrity number (RIN) values over 9. Illumina TruSeq Stranded mRNA libraries were prepared according to the manufacturer's instructions and sequenced on NovaSeq6000 for 101bp paired-end sequencing. All samples except five were sequenced twice to obtain sufficient number of reads.

Data Processing and Quality Control: FASTQ files were obtained for all 107 PPC samples and processed using a similar pipeline described in our previous studies [37,99]. Specifically, RNA-seq reads were aligned with STAR (2.7.3) [100] to the hg19 reference using GENCODE version 34 hg19[95] splice junctions with default alignment parameters and the following adjustments: *-outFilterMultimapNmax 20, -outFilterMismatchNmax 999, -alignIntronMin 20, -alignIntronMax 1000000, -alignMatesGapMax 1000000*. BAM files were sorted by coordinates, and duplicate reads were marked using Samtools (1.9.0) [92]. RNA-seq QC metrics were calculated using Samtools (1.9.0) flagstat [92], Samtools (1.9.0) idxstats [92], and Picard (2.20.1) CollectRnaSeqMetrics [101]. Across all 107 PPC samples, the total read depth ranged from 32.3 M to 160.4 M (mean = 70.7), the median percentage of intergenic bases was 3.31%, the median percentage of mRNA bases was 92.1%, and the median percentage of duplicate reads was 22.2%.

Sample Identity: We obtained common bi-allelic and exonic variants from the 1000 Genomes Phase 3 panel [102] with minor allele frequencies between 45% and 55% and

predicted their genotypes in the 107 bulk RNA-seq samples using *mpileup* and *call* functions in BCFtools (1.9.0) [103,104]. Then, we used the *genome* command in plink [101] to estimate the identity-by-state (IBS) between each pair of bulk RNA-seq and WGS samples. All RNA-seq samples were correctly matched to the subject with PI_HAT > 95%.

Quantification of gene expression and relative isoform usage: We calculated TPM and estimated relative isoform usage for each gene in each RNA-seq sample using RSEM (version 1.2.20) [105] with the following options *–seed 3272015 –estimate-rspd –paired-end –forward-prob*. To identify expressed autosomal genes and isoforms to use for eQTL analyses, we used the same approach previously described [12]. Briefly, autosomal genes were considered expressed if TPM ≥ 1 in at least 10% of samples. To identify expressed isoforms, we required that isoforms had TPM ≥ 1 and usage ≥ 10% in at least 10% of samples and corresponded to expressed genes with at least two expressed isoforms. In total, 16,464 autosomal genes were used for e_gQTL analysis, and 29,871 autosomal isoforms corresponding to 9,624 genes were used for e_iQTL analysis. We quantile-normalized TPM and isoform usage across all 107 samples using the *normalize.quantiles* (preprocessCore) and *qnorm* functions in R (version 4.2.1) to obtain a mean expression = 0 and standard deviation = 1.

Inferring pseudotime using Monocle: We obtained FASTQ files for 213 iPSCs [35,37] (phs000924), 176 adult whole pancreas [8] (phs000424), and 87 adult islets[38] (GSE50398), and processed the data using the same pipeline described above to obtain TPM counts for each gene per sample. We then used Monocle (http://cole-trapnell-lab.github.io/monocle-release/docs/#constructing-single-cell-trajectories) [106] to infer the pseudotime on all of the RNA-seq samples, including the 107 PPCs. Following the standard workflow under

"Constructing Single Cell Trajectories" in the Monocle tutorial, we provided TPM counts for all overlapping autosomal expressed genes in the four tissues as input. Then, we identified differentially expressed genes using *differentialGeneTest*, ordered them (*setOrderingFilter*), and performed dimension reduction analysis using *reduceDimension* with *max_components = 2* and *method = "DDRTree"*. Pseudotime was calculated by rooting time (pseudotime = 0) in the 213 PPCs using the *GM_state* and *orderCells* functions provided in the tutorial.

PCA analysis with iPSCs, adult whole pancreas, and adult islets: We obtained TPM counts (described above) for the 213 iPSCs [37], 176 adult whole pancreas[8], 87 adult islets [38], and the 107 PPCs and performed PCA analysis on the 2,000 most variable genes across the samples using *prcomp* in R (version 4.2.1) with *scale = T* and *center = T*. We observed that the PC clusters corresponded to the iPSCs and each of the three pancreatic tissue types: PPC, adult islets, and adult whole pancreas.

Cellular deconvolution: For each of the eight cell types in scRNA-seq, we selected the top 200 most differentially expressed genes that were unique to the cell type (i.e., not expressed in the other cell types). Replicating late PPCs and late PPCs had many overlapping expressed genes so fewer (n = 16 and 164, respectively) were selected. We obtained the average expression of the signature genes for each cell type using *AverageExpression* in Seurat and provided it as input into CIBERSORTx[107] (https://cibersortx.stanford.edu/) along with bulk TPM matrix. Batch correction and quantile normalization were both disabled. We ran CIBERSORTx[107] deconvolution on absolute mode with at least 100 permutations. The predicted fraction of late PPCs and

replicating late PPCs were compared to FACS measurements of double-positive PDX-1^+/NKX6-1^+ cells.

eQTL Analysis

To investigate the effects of genetic variation on gene expression in PPCs, we performed an expression quantitative trait loci (eQTL) analysis on gene expression and isoform usage. The eQTLs associated with gene expression were defined as e_gQTLs while those associated with relative isoform usage were defined as e_iQTLs.

Covariates for eQTL Mapping*:* We included the following as covariates for eQTL mapping of both gene expression and isoform usage: 1) sex; 2) normalized number of RNA-seq reads; 3) percent of reads that mapped to autosome or sex chromosomes; 4) percent of reads mapped to mitochondrial chromosome; 5) 20 genotype principal components to account for global ancestry; 6) 20 PEER factors to account for transcriptome variability; and 7) kinship matrix to account for genetic relatedness between samples.

Genotype Principal Component Analysis (PCA): Global ancestry was estimated using the genotypes of the 439,461 common variants with minor allele frequency (MAF) between 45 and 55% in the 1000 Genomes Phase 3 Panel[102]. We merged the VCF files for the 106 iPSCORE subjects and the 2,504 subjects in the 1000 Genomes[102] and performed a PCA analysis using *plink --pca*[101].

PEER Factors: We sought to determine the optimal number of PEER factors to use in the eQTL analysis that will result in maximal eGene discovery. To this end, we initially calculated PEER factors on the 10,000 expressed genes with the largest variance across all samples. To limit biases due to the expression levels of each gene, we divided the 16,464 expressed genes into ten deciles based on their average TPM, and selected 50

genes from each decile, for a total of 500 genes. We next performed eQTL analysis on each of the 500 genes using 10 to 60 PEER factors in increments of 10. While 30 PEER factors resulted in the highest percentage of eGenes (14.0%), we opted for using 20 PEER factors because the eQTL analysis had a comparable percentage of eGenes (11.8%) to GTEx tissues with similar sample sizes [10]. Although we observed variable fraction of double-positive PDX1+/NKX6-1+ cells in the PPC samples, we did not include this variable as a covariate because PEER factors 1 and 4 already accounted for this variability.

Kinship Matrix: The kinship matrix was included as a random effects term to account for the genetic relatedness between individuals in our cohort. We constructed the kinship matrix using the same 439,461 variants employed above using the *–make-rel square* function in plink[101].

eQTL Analysis: We performed eQTL analysis using the same method described in our previous study[12]. For each expressed autosomal gene and isoform, we tested variants that were within 500 Kb of the gene body coordinates using the *bcftools query* function. To account for the genetic relatedness between the samples, we performed eQTL mapping using a linear mixed model with the *scan* function in limix (version 3.0.4)[108] that incorporates the kinship matrix as a random effects term. Specifically, eQTL mapping was implemented through the following model:

$$y_i = \beta_{ji} \cdot g_j + \sum_{n=1}^{N} \beta_n \cdot C_n + \mathrm{u} + \epsilon_{ij}$$

Where y_i is the normalized expression value for gene i, β_{ji} is the effect size of genotype of SNP j on gene i, g_j is the genotype of SNP j, β_n is the effect size of covariate n, C_n is a vector of values for covariate n, u is the kinship matrix as a random effect, and ϵ is the

error term for the association between expression of gene i and genotype of SNP j. As described above, we used the following as covariates: 1) sex, 2) normalized number of RNA-seq reads, 3) percent of reads mapped to autosomal or sex chromosome, 4) percent of reads mapped to mitochondrial chromosome, 5) the top 20 genotype PCs (to account to global ancestry), and 6) the top 20 PEER factors (to account for confounders of expression variability).

FDR Correction: To perform FDR correction, we used a two-step procedure described in Huang et al. [109], which first corrects at the gene level and then at the genome-wide level. First, we performed FDR correction on the p-values of all variants tested for each gene or isoform using eigenMT [108], which considers the LD structure of the variants. Then, we extracted the lead eQTL for each gene or isoform based on the most significant FDR-corrected p-value. If more than one variant had the same FDR-corrected p-value, we selected the one with the largest absolute effect size as the lead eQTL. For the second correction, we performed an FDR-correction on all lead variants using the Benjamini-Hochberg method (q-value) and considered only eQTLs with q-value ≤ 0.01 as significant.

Conditional eQTLs: To identify additional independent eQTLs (i.e., conditional eQTLs) for each eGene and eIsoform, we performed a step-wise regression analysis in which the genotype of the lead eQTL was included as a covariate in the model and the eQTL mapping procedure (regression and multiple test correction) was re-performed. We repeated this analysis to discover up to five additional associations for each eGene and eIsoform. Conditional eQTLs with q-values ≤ 0.01 were considered significant.

Functional characterization of PPC eQTLs

Fine-mapping of eQTL Associations: To define a credible set of candidate causal variants for each eQTL association, we performed genetic fine-mapping using the *finemap.abf* function in *coloc* (version 5.1.0, R) [41]. This Bayesian method converts p-values of all variants tested for a specific gene to posterior probabilities (PP) of association for being the causal variant. Variants with PP $\geq$ 1% are available on Figshare: https://figshare.com/projects/Large-scale_eQTL_analysis_of_PPC/156987. eQTLs not present in the table do not have variants with PP $\geq$ 1% (i.e., all variants were estimated to have PP < 1%).

Genomic enrichments of e_gQTLs and e_iQTLs: For each independent eQTL association, we obtained candidate causal variants whose PP $\geq$ 5% and determined their overlap with each of the following genomic annotations using *bedtools intersect*: short splice acceptor sites ($\pm$ 50bp), long splice acceptor sites ($\pm$ 100bp), splice donor sites ($\pm$ 50bp), UTR, intron, exon, intergenic, promoters, and RNA-binding protein binding sites (RBP-BS). RBP-BS were downloaded from a published dataset that utilized enhanced CLIP to identify binding sites of 73 RBPs [110]. We considered only binding sites with irreproducible discovery rate (IDR) threshold of 0.01, indicating that these sites were reproducible across multiple biological samples. Enrichment of candidate causal variants for genomic regions was calculated using a Fisher's Exact Test comparing the proportion of SNPs that overlap each annotation between e_gQTLs and e_iQTLs. P-values were corrected using the Benjamini-Hochberg method and were considered significant if their FDR-corrected p-value $\leq$ 0.05 (Figure 1E).

Quantification of allele-specific binding of transcription factors using GVATdb: To annotate each candidate causal variant by their effects on transcription factor (TF)

binding, we used the Genetic Variants Allelic TF Binding Database (GVATdb) to estimate the TF binding impact score associated with each variant and each of the 58 PPC-expressed TF available on the database and with a AUPRC > 0.75 indicating a high-confidence deltaSVM model. We estimated the score using the instructions and reference files provided on the GVATdb GitHub repository (https://github.com/ren-lab/deltaSVM). The software required a list of SNPs as input along with hg19 reference files provided in the GVATdb repository. The output provides the deltaSVM score [111] for each variant-TF pair, indicating whether the variant results in a promotion ("Gain"), disruption ("Loss"), or no change ("None") in TF binding. deltaSVM scores for each variant-TF pair are available on Figshare: https://figshare.com/projects/Large-scale_eQTL_analysis_of_PPC/156987.

Correlation between eQTL effect size and binding affinity of transcription factors: To determine whether e_gQTLs were more likely to affect TF binding compared to e_iQTLs, we performed a Spearman Correlation Analysis between deltaSVM score and eQTL effect size on candidate causal variants with PP ≥ 10%, 20%, 40%, 60% and 80%. We considered nominal p-value ≤ 0.05 as significant.

Colocalization between PPC gene and isoform eQTLs: To determine the overlap of genetic variants between e_gQTLs and e_iQTLs for the same gene, we performed Bayesian colocalization using the *coloc.abf* function in *coloc* (version 5.1.0, R) [41], where each pair of signals was given a summary PP that each of the following five hypotheses was true: H0) no association was detected in both signals, H1) an association was detected only in signal 1, H2) an association was detected only in signal 2, H3) an association was detected in both signals but the underlying causal variants are different, and H4) an association was detected for both signals and the underlying causal variants are the same. We filtered the

results by requiring that each colocalization used the number of overlapping variants (called "nsnps" in the *coloc.abf* output) ≥ 500. We considered two eQTL signals to be shared if the PP for H4 (called "PP.H4.abf" in *coloc.abf* output; hereafter referred to as PP.H4) ≥ 80%. Conversely, two signals were considered distinct if the PP for H3 (called "PP.H3.abf" in *coloc.abf* output; hereafter referred to as PP.H3) ≥ 80%. eQTL associations with PP.H4 < 80% and PP.H3 < 80% were due to insufficient power in one or both eQTL signals. As input into *coloc.abf*, we provided p-values, minor allele frequency, and sample size.

Genomic enrichment of overlapping e_gQTL and e_iQTL signals compared to non-overlapping: To test the enrichment of overlapping e_gQTLs and e_iQTLs in genomic regions compared to non-overlapping signals, we determined the overlap of candidate causal variants with PP ≥ 1% in each genomic annotation using *bedtools intersect* and compared the proportion of variants overlapping each annotation against a background set of 20,000 random variants using a Fisher's Exact Test as previously described [10]. For overlapping eQTLs, we used the candidate causal variants predicted in the *coloc.abf* output. Enrichments with nominal p-value < 0.05 were considered significant.

Downloading eQTL summary statistics for adult pancreatic tissues

We downloaded complete eQTL summary statistics for gene and exon associations for 420 adult human islets from the InSPIRE Consortium (https://zenodo.org/record/3408356) [11], and gene and splicing associations for 305 adult whole pancreas from the GTEx Data Portal for GTEx Analysis version 8 [10] (https://console.cloud.google.com/storage/browser/gtex-resources). All GTEx SNPs were converted to hg19 using the UCSC liftOver Bioconductor package in R

(https://www.bioconductor.org/help/workflows/liftOver/). Lead SNPs for conditional associations in the adult islets and whole pancreas datasets were downloaded from their respective studies (complete statistics were not readily available). Due to the different types of eQTLs used in this study that are associated with changes in alternative splicing (e_iQTLs, exon eQTLs, and sQTLs), hereafter we refer to this collective unit as "e_{AS}QTLs".

Comparing eGenes between fetal-like PPC and adult islets

To identify eGenes that were shared between PPC and adult islet tissues, we compared the 4,065 eGenes in PPC and the 4,211 eGenes in adult islets that complete summary statistics were available for. Specifically, we used the *intersect* function in R to identify eGenes that overlapped between the two tissues and *setdiff* function in R to identify eGenes that did not overlap. Similarly, using the *intersect* function in R, we compared the 22,266 expressed genes in adult islet tissues with the 4,065 eGenes in PPC to identify the proportion of PPC eGenes that were expressed in adult islets, and vice versa with the 17,098 expressed genes in PPC and 4,211 eGenes in adult islets. The 22,266 expressed genes in adult islet tissues were obtained from the complete summary statistics uploaded by the previous study in https://zenodo.org/record/3408356.

Comparing eQTLs present in fetal-like PPC and adult pancreatic tissues

Colocalization between PPC and adult eQTLs: To identify eQTLs whose effects were driven by the same causal signals in PPC and adult pancreatic tissues (islets and whole pancreas), we performed Bayesian colocalization using the *coloc.abf* function in *coloc* (version 5.1.0, R) [41]. Specifically, for each PPC and adult eQTL, we tested its overlap with nearby eQTLs within 3 Mb from the gene body coordinates. eQTLs with no overlapping variants would automatically not be tested. Then, we filtered the results by requiring that

each colocalization used the number of overlapping variants (called "nsnps" in the *coloc.abf* output) ≥ 500. As described above, we considered two eQTL signals to be shared if PP.H4 ≥ 80% or distinct if PP.H3 ≥ 80%. eQTL associations with PP.H4 < 80% and PP.H3 < 80% were due to insufficient power in one or both eQTL signals.

Because we, and others, have shown that e_gQTLs are functionally different from e_{AS}QTLs (e_iQTLs, exon eQTLs, and splicing eQTLs), we performed colocalization for e_gQTLs and e_{AS}QTLs independently (i.e., colocalization of e_gQTL was performed only with another e_gQTL and an e_{AS}QTL only with another e_{AS}QTL).

Fine-mapping of adult eQTL associations: Similarly for PPC eQTLs, we identified candidate causal variants using the *finemap.abf* function in *coloc* (version 5.1.0, R). This Bayesian method converts p-values of all variants tested for a specific gene to a PP value for being the causal variant.

For all downstream analyses beyond this point, we used only PPC, adult islets, and adult whole pancreas eQTLs that had at least one candidate causal variant with PP ≥ 1%, were outside of the MHC region, and were annotated in GENCODE version 34 hg19, to ensure that our analyses were sufficiently powered and the multiple datasets were comparable.

Identifying tissue-unique singleton eQTLs: To identify tissue-unique singleton eQTLs, we obtained all eQTLs that did not colocalize with another eQTL and examined their LD with all other eQTLs of the same phenotype (e_gQTLs or e_{AS}QTLs) using their most likely candidate causal variants based on the highest PP from fine-mapping (*finemap.abf*). If the candidate causal variant was not genotyped in the 1000 Genomes Phase 3 panel, then we used the next top candidate causal variant. We repeated this process

until we found a variant that was in the 1000 Genomes or no more variants remained with causal PP ≥ 1%. Because complete summary statistics were not available for the adult conditional eQTLs, we used their lead variants publicly available from their respective studies to account for the presence of multiple causal variants in the genomic region. LD was calculated using *plink --r2 square --keep-allele-order --make-bed* [101] and the 1000 Genomes Phase 3 panel [102]. We considered two eQTLs to be in LD if their candidate causal variants were within 500 Kb and had $r^2 \geq 0.2$. If LD could not be measured, because one of the variants was not genotyped in the 1000 Genomes, then we used distance as a metric for LD, where if the variants were within 500 Kb of each other, we considered them to be in LD. Singleton eQTLs that were found to be in LD with another eQTL (regardless of tissue) were re-annotated as "ambiguous" and excluded from downstream analyses. Otherwise, we kept their annotations as tissue-unique singletons.

Identifying eQTL modules: eQTL modules were identified by first creating a network using the *graph_from_data_frame* function in igraph (version 1.3.4, R) [112] where the input was a data frame containing all pairs of colocalized eQTLs (nsnps ≥ 500 and PP.H4 ≥ 80%) as binary edges. We created networks for each chromosome and phenotype (gene expression and alternatively splicing) independently, totaling to 44 networks (22 chromosomes x 2 phenotypes = 44 networks). Then, we performed community detection analysis using the *cluster_leiden* function with *--objective_function = "modularity"*, *n_iterations = 500, resolution = 0.3* to identify modules of eQTLs. Upon examining them in depth, we observed that 5% of the modules contained at least one H3 association (PP.H3 ≥ 80%) between a pair of eQTLs, indicating that signals within a module were predicted to have distinct genetic variants despite being assigned to the same module. Therefore, to

filter for modules that contained eQTLs likely to share the same causal variants, we required that at least 30% of all eQTL pairs had a H4 association and that the number of H4 "edges" was twice the number of H3 "edges" (number of H4 edges / number of H3 edges ≥ 2). For example, a module with four eQTLs would have six possible pairwise combinations, and to be considered a validated module, we required at least two H4 edges and no more than one H3 edge. Modules that did not pass these thresholds were annotated as "module_failed" and excluded from downstream analyses. Module IDs were assigned such that the first term indicates the phenotype the module was associated with ("GE" for gene expression or "AS" for alternative splicing), the second term indicates the chromosome number, and the third term indicates a unique integer. For example, "GE_1_32" indicates that this module is associated with changes in gene expression, located in in chromosome 1, and assigned the number 32.

Identifying tissue-unique and tissue-sharing eQTL modules: Combinatorial eQTLs were defined in this study as an eQTL having at least one H4 association (PP.H4 ≥ 80%) with another eQTL either in the same or different tissue. These combinatorial eQTLs were then connected to form a module, which we identified using the network analysis described above. We then categorized each module based on the activity of eQTLs in the three pancreatic tissues, having a total of seven module categories (Figure 3B):

1. PPC-unique: contains eQTLs in **only** PPC
2. Adult islet-unique: contains eQTLs in **only** adult islets
3. Adult whole pancreas-unique: contains eQTLs in **only** adult whole pancreas
4. Adult-shared: contains eQTLs in adult islets **and** adult whole pancreas
5. Fetal-islet: contains eQTLs in PPC **and** adult islets

6. Fetal-whole-pancreas: contains eQTLs in PPC **and** adult whole pancreas
7. Fetal-adult: contains eQTLs in **all** three pancreatic tissues

We next examined the eQTLs modules for LD with eQTLs in other tissues to confirm tissue specificity. Similar to the analysis described above for identifying tissue-unique singletons, we calculated LD using *plink --r2 square --keep-allele-order --make-bed* [101] and the 1000 Genomes Phase 3 panel [102] between the eQTLs' most likely candidate causal variants (based on the highest PP; PP $\geq$ 1%). We considered two eQTLs to be in LD if they had an $r^2 \geq 0.2$ and were within 500 Kb of each other. If LD could not be calculated because candidate causal variants were not genotyped in the 1000 Genomes Phase 3 panel, then we used distance as a metric for LD and considered two eQTLs to be in LD if their candidate causal variants were within 500 Kb. To account for the presence of multiple causal variants in the genomic region, we included the lead variants from the adult islet and whole pancreas conditional eQTLs in the LD comparisons to prevent misclassification of tissue-unique eQTLs.

For each of the module categories, we required that the following were true to be considered for downstream analyses:

1. PPC-unique: contains eQTLs in only PPC, and **all** eQTLs were not in LD with eQTLs in adult islets **and** adult whole pancreas
2. Adult islet-unique: contains eQTLs in only adult islets, and **all** eQTLs were not in LD with eQTLs in adult whole pancreas **and** PPC
3. Adult whole pancreas-unique: contains eQTLs in only adult whole pancreas, and **all** eQTLs were not in LD with eQTLs in adult islets **and** PPC

4. Adult-shared: contains eQTLs in only adult islets and adult whole pancreas, and **all** eQTLs were not in LD with eQTLs in PPC
5. Fetal-islet: contains eQTLs in PPC and adult islets, and **all** eQTLs were not in LD with eQTLs in adult whole pancreas
6. Fetal-whole-pancreas: contains eQTLs in PPC and adult whole pancreas, and **all** eQTLs were not in LD with eQTLs in adult islets
7. Fetal-adult: contains eQTLs in **all** three pancreatic tissues.

For any module that did not meet the above requirements, we annotated the eQTLs in the module "ambiguous" and excluded for downstream analysis. Hereafter, we refer the eQTL associations in tissue-unique modules (categories 1-3) as tissue-unique combinatorial eQTLs and those in categories 5-7 as eQTLs shared between both fetal-like and adult stages.

Enrichment of fetal-like PPC-unique singleton and combinatorial eQTLs in chromatin states

We obtained chromatin state maps for human embryonic stem cell-derived pancreatic progenitor cells and adult islets from previously published studies [14,113]. Because e_gQTLs were likely to affect non-coding regulatory elements (**Figure 1E**), we only considered them in this analysis and excluded e_{AS}QTLs. Enrichments for PPC-unique singleton and combinatorial e_gQTLs were calculated using a Fisher's Exact Test by comparing the proportion of fine-mapped variants (from *finemap.abf*) of the e_gQTLs at different thresholds of PP from 0-0.8 at 0.1 intervals to a background set of 20,000 randomly selected variants. Enrichments were Benjamini-Hochberg-corrected. Corrected p-values ≤ 0.05 were considered significant.

1.13.13 Overlap of eGenes in shared modules between fetal-like PPC and adult pancreatic tissues

For the modules shared between both fetal-like and the two adult pancreatic tissues (categories 5-7; described above), we compared the eGenes associated with: 1) PPC e_gQTLs versus adult islet e_gQTLs; and 2) PPC e_gQTLs versus adult whole pancreas e_gQTLs. For e_{AS}QTLs, we compared the genes mapping to: 1) each isoform in PPC versus exon in adult islets; and 2) each isoform in PPC versus splice interval in adult whole pancreas. From these comparisons, we assigned each module an "islet_egene_overlap" label and an "whole_pancreas_egene_overlap" label, where "zero" indicates that the module did not contain an eQTL in the adult tissue, "same" indicates that the module contained eQTLs corresponding to only the same eGenes in PPC and adult, "partial" indicates that the module contained eQTLs corresponding with partially overlapping eGenes between PPC and adult, and "different" indicates that the module contained eQTLs corresponding to only different genes. For example, if a module was annotated with "zero" for islet_egene_overlap and "same" for whole_pancreas_egene_overlap, this meant that the module did not contain an eQTL from adult islet and had only eQTLs associated with the same eGenes between PPC and adult whole pancreas. These annotations also meant that this module was in the "fetal-whole-pancreas" category (i.e, only contained eQTLs from PPC and adult whole pancreas).

Complex Trait GWAS Associations

Colocalization of eQTLs with GWAS associations: We obtained GWAS summary statistics from ten different studies: 1) type 1 diabetes [3], 2) type 2 diabetes [113], 3) body mass index [50], 4) triglycerides [50], 5) HDL cholesterol [50], 6) LDL direct [50], 7) cholesterol [50], 8)

glycated hemoglobin A1C (HbA1c) levels from the MAGIC Consortium [114], 9) HbA1c levels from the Pan-UKBB Study [50], and 10) fasting glucose [114]. All of the data, except for type 1 diabetes, were provided in hg19 coordinates, therefore we converted the coordinates from hg38 to hg19 using the liftOver package in R [115]. We sorted and indexed each file using *tabix* [92]. For each trait, we performed colocalization between GWAS variants and all filtered significant eQTLs (see bolded section above) in the three pancreatic tissues with the *coloc.abf* function in *coloc* (version 5.1.0, R) [41] using p-values, MAF, and sample size as inputs. Then, we filtered results based on whether the lead candidate causal variant underlying both GWAS and eQTL association (from *coloc.abf* output) is genome-wide significant for GWAS association (p-value $\leq 5x10^{-8}$) and the number of overlapping variants used to test for colocalization (nsnps) ≥ 500. eQTLs were considered to share a genetic signal with GWAS if PP.H4 $\geq$ 80% or have distinct signals with GWAS if PP.H3 $\geq$ 80%. For eQTL modules, we required that at least 30% of the eQTLs in the module colocalized with GWAS (PP.H4 $\geq$ 80%) and that the number of H4 associations is twice the number of H3 associations (number of H4 associations / number of H3 associations $\geq$ 2).

GWAS 99% Credible Sets: For each GWAS locus, we constructed 99% credible sets with the predicted candidate causal variants underlying both eQTL and GWAS associations (from *coloc.abf* output). If the GWAS locus colocalized with a singleton eQTL, the credible sets were constructed using the output of the eQTL's colocalization with GWAS. If the GWAS locus colocalized with an eQTL module, we constructed credible sets for each of the pairwise eQTL-GWAS colocalization and retained the eQTL that resulted in the least number of candidate causal variants. If multiple eQTLs had the

same number of variants in their credible set, we considered the eQTL with the highest PP.H4 for GWAS colocalization. 99% credible sets were constructed by first sorting the variants by descending order of causal PP and obtaining the least number of variants that resulted in a cumulative PP ≥ 99%.

LD with non-pancreatic GTEx tissues: We downloaded the lead SNPs for all significant e_gQTLs (including their conditionals) for the 48 non-pancreatic tissues in the GTEx dataset version 8 [10] and converted their genomic positions to hg19 using UCSC liftOver [115]. We then calculated their LD with the lead SNPs of the 16 PPC-unique e_gQTLs that colocalized with GWAS using *plink* [101] *--tag-kb 500 --tag-r2 0.2 --show-tags all* and the 1000 Genomes [102] as the reference panel. We considered two eQTLs to be in LD if their lead SNPs were within 500 Kb and had $r^2 > 0.2$. If LD could not be calculated because the SNP was not genotyped in the reference panel, we used distance as a metric in which we considered two eQTLs to be in LD if their lead SNPs were within 500 Kb.

1.14 Data Availability

The PPC scRNA-seq and bulk RNA-seq data generated in this study have been deposited in the GEO database under accession codes GSE152610 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152610] and GSE182758 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182758], respectively. The WGS data used in this study for iPSCORE individuals were obtained as a VCF file from phs001325.v3 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001325]. The reference gene annotation file for aligning bulk RNA-seq data of PPC were obtained from GENCODE release version 34 in GRCh37 as a GTF file [https://www.gencodegenes.org/human/release_34.html]. The bulk RNA-seq

data for iPSC, adult islet, and adult whole pancreas samples used in PCA and pseudotime analyses were obtained from phs000924 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000924], GSE50398 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50398], and phs000424 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424], respectively. eQTL summary statistics for adult whole pancreas and islet samples were obtained from the GTEx Data Repository [https://console.cloud.google.com/storage/browser/gtex-resources] and a previously published study [11] [https://zenodo.org/record/3408356], respectively. GWAS summary statistics were obtained from the Pan UK BioBank resource [https://pan.ukbb.broadinstitute.org/], the MAGIC (Meta-Analyses of Glucose and Insulin-related traits) Consortium [https://magicinvestigators.org/downloads/; https://doi.org/10.1038/s41588-021-00852-9] the DIAMANTE Consortium [https://diagram-consortium.org/downloads.html; http://doi.org/10.1038/s41588-018-0241-6], and a previously published study [3] [http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90014001-GCST90015000/GCST90014023/]. Full summary statistics for PPC eQTLs, supplemental data, and processed scRNA-seq have been deposited in Figshare: https://figshare.com/projects/Large-scale_eQTL_analysis_of_PPC/156987.

1.15 Code Availability

Scripts for processing RNA-seq and scRNA-seq data and performing downstream analyses are publicly available at /https://github.com/jenniferngp/iPSC_PPC_eQTL_Project (version 1.0.0 of the release).

1.16 Acknowledgements

This work was supported by the National Library Training Grant T15LM011271 (J.P.N., M.K.R.D., T.D.A.) and the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) F31DK131867 (J.P.N.), U01DK105541, DP3DK112155 and P30DK063491 (K.A.F.). Additional support was also received from the National Heart, Lung and Blood Institute (NHLBI) F31HL158198 (T.D.A.) and the Dongguk University Research Fund of 2023 (J.H.K.). We thank Drs. Maike Sanders and Kyle Gaulton for their advice on experimental design and analyses. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant S10OD026929.

1.17 Author Information

iPSCORE Consortium: Lana Ribeiro Aguiar[5], Angelo D. Arias[4], Timothy D. Arthur[2,3], Paola Benaglio[4], W. Travis Berggren[7], Juan Carlos Izpisua Belmonte[8], Victor Borja[5], Megan Cook[5], Matteo D'Antonio[2,4,5], Agnieszka D'Antonio-Chronowska[4], Christopher DeBoever[1], Kenneth E. Diffenderfer[7], Margaret K.R. Donovan[1,2], KathyJean Farnam[5], Kelly A. Frazer[4,5], Kyohei Fujita[4], Melvin Garcia[5], Olivier Harismendy[2], Benjamin A. Henson[5], David Jakubosky[2,3], Kristen Jepsen[5], He Li[4], Hiroko Matsui[5], Naoki Nariai[4], Jennifer P. Nguyen[1,2], Daniel T. O'Connor[9], Jonathan Okubo[5], Athanasia D. Panopoulos[8], Fengwen Rao[9], Joaquin Reyna[5], Bianca M. Salgado[5], Nayara Silva[4], Erin N. Smith[4], Josh Sohmer[5], Shawn Yost[1], William W. Young Greenwald[1]

[7] Stem Cell Core, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[8] Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[9] Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA

Author Contributions

K.A.F. and iPSCORE consortium members conceived the study. A.D.C., B.M.S., K.F., and iPSCORE consortium members performed the differentiations and generated molecular data. J.P.N., M.K.R.D. and H.M. performed quality check on scRNA-seq and RNA-seq samples. J.P.N., T.D.A., J.H.K. performed the computational analyses. K.A.F. and iPSCORE consortium members oversaw the study. J.P.N., M.D. and K.A.F. prepared the manuscript.

Chapter 1, in full, is a reprint of the material as it appears in Nature Communications 2023, Jennifer P. Nguyen, Timothy D. Arthur, Kyohei Fujita, Bianca M. Salgado, Margaret K.R. Donovan, iPSCORE Consortium, Hiroko Matsui, Ji-Hyun Kim, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 2: Investigating the genetic regulatory mechanisms underlying gene expression changes during early pancreas development

2.1 Abstract

Most GWAS loci are presumed to affect gene regulation, however, only 43% colocalize with expression quantitative trait loci (eQTLs). To address this colocalization gap, we conduct multi-omic QTL analyses using bulk gene expression and chromatin accessibility data generated from over 100 samples of iPSC-derived pancreatic progenitor cells (PPC) to characterize the regulatory effects of genetic variation associated with obesity and diabetes risk. We identify over 12,000 caQTLs and expand the number of eQTLs to 8,000 with a newly updated analysis. caPeaks in PPC are enriched for pancreatic developmental motifs and are preferentially located in CTCF binding sites and enhancer states. Colocalization with GWAS signals associated with 8 metabolic traits and two pancreatic diseases (type 1 and type 2 diabetes) identified 222 GWAS loci that colocalized with PPC QTLs, implicating 102 genes and 144 regulatory elements with putative biological roles in pancreatic traits and disease. We further identify 8% caQTLs and 15% eQTLs that were specifically active to fetal development and show that the genes associated with fetal-unique eQTLs are subject to stronger evolutionary constraint and are strongly depleted for GWAS colocalization. This study provides a unique, valuable, and comprehensive resource for the scientific community to study regulatory variation active during early pancreas development.

2.2 Introduction

Genetic variants identified by genome-wide association studies (GWAS) have been shown to be enriched in non-coding regions of the genome, suggesting that these variants may exert their effects on phenotypes through disruption of gene regulatory mechanisms. Towards understanding the underlying mechanisms of these associations, extensive efforts have been made over the past decade to generate comprehensive expression quantitative trait loci (eQTL) maps for diverse adult tissues and cell types. While eQTL discoveries have shed light on biological mechanisms for many GWAS loci, approximately 60% or more remains unexplained, i.e. do not colocalize with eQTLs. This discrepancy between GWAS and eQTL associations can be explained by their preferences for different types of genetic variants. Specifically, GWAS hits tend to be distal and are associated with complex regulatory landscapes, while eQTLs are preferentially biased towards gene promoters and are involved in more straightforward mechanisms. Moreover, genes that play a significant role in disease tend to be subject to more intense selective pressure on their variants, thus making their eQTL discoveries more challenging. Incorporating additional molecular phenotypes alongside eQTLs, especially in the context of fetal development, may help address these challenges, as well as clarify on the regulatory mechanisms underlying both gene expression and phenotype variation.

In this chapter, we aim to leverage assay for transposase-accessible chromatin (ATAC-seq) to profile accessible chromatin regions in 109 iPSC-derived pancreatic progenitor cells (PPC) samples (**Table 2.1**). This assay captures both distal and proximal cis-regulatory elements (CREs) and can reveal information about transcription factors that are involved in downstream gene regulation through footprinting analyses. Additionally, this cohort contains 107 PPC samples that were previously used to map eQTLs. This

enables us to conduct accurate comparative analyses to elucidate the genetic functional consequences on chromatin accessibility, gene expression, and therefore disease predisposition.

Table 2.1 Table describing the number of samples and subjects for each data type generated for the PPC cohort

Data Type	Cell Type	N Samples	N Subjects
scRNA-seq	iPSC	1	1
scRNA-seq	PPC	10	9
snATAC-seq	PPC	7	7
Bulk RNA-seq	PPC	107	106
Bulk ATAC-seq	PPC	109	108

2.2 Characterization of single nuclei accessible chromatin

To characterize the cellular heterogeneity in accessible chromatin in PPC, we performed snATAC-seq on seven PPC samples, all of which had matched scRNA-seq (**Table 2.1**). After filtering low-quality nuclei, we retained 26,026 nuclei and detected 326,021 peaks across the PPC epigenome (**Figure 2.1a**). Integration and clustering analyses using Signac[116] detected nine cell populations in snATAC-seq, five of which comprised of one larger cluster. Like in scRNA-seq {ref}, this large cluster contained the majority of the nuclei (n = 23,976, 92.1%). To determine the cell identity of each nine populations, we estimated TF motif activity levels using chromVar and compared them to their expression levels in scRNA-seq (**Figure 2.1b**). We identified: mesendoderm (high motif activity levels for TFAP2A/B), early definitive endoderm (GRHL1/2), early PPC (GATA4/6, CDX1/4, PDX1), a sub-population of late PPC (named late PPC 1; PDX1, NKX6-1, HNF1A/B, and residual expression of early PPC markers), a second and slightly

more advanced sub-population of late PPC (named late PPC 2; PDX1, NKX6-1, HNF1A/B), replicating late PPC (cell-cycle mediators FOS/JUN, BACH2), endocrine (MAFA, NKX2-2, NEUROD1), and ductal precursors (named early ductal; ETV1, ETS1/2). Unlike scRNA, snATAC detected endocrine precursors, expressing both HES1, an endocrine-exocrine specification mediator[117], and several endocrine developmental markers (PAX4/6, RFX1/3). In general, we found that 86.7% (n=22,574) of the nuclei mapped to either early PPC, late PPC, or replicating late PPC, consistent with scRNA[118].

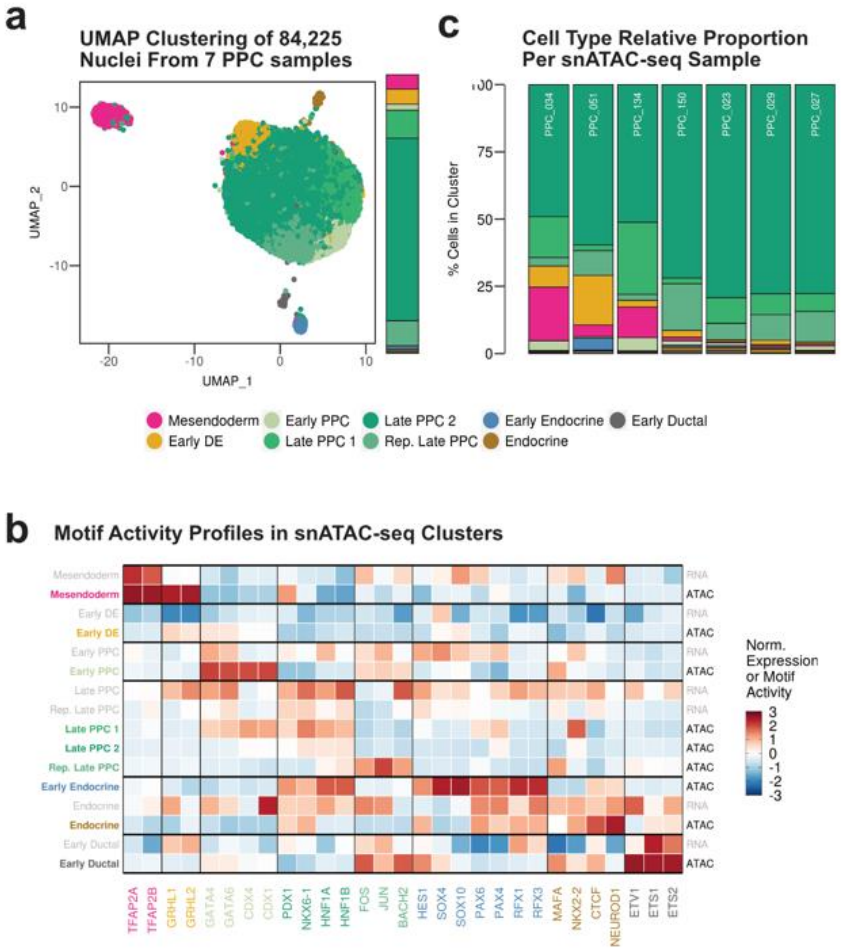


Figure 2.1 Characterization of PPC snATAC-seq
(a) UMAP plot of snATAC-seq data from 26,026 single nuclei from seven PPC samples. Each point represents a single nuclei color-coded by its assigned cluster. (b) Heatmap comparing the z-normalized motif activity scores from chromVAR for pancreatic-associated transcription factors in the snATAC-seq clusters from panel a. Also shown are the normalized expression of the pancreatic-associated transcription factors in the scRNA-seq clusters from Nguyen et al. Clusters labeled in italicized color correspond to the snATAC-seq clusters in panel a. Clusters labeled in grey correspond to the scRNA-seq clusters in Nguyen et al. (c) Stacked bar plot showing the fraction of cells from each sample assigned to each cluster in snATAC-seq. Color-coding corresponds to the clusters in panel a.

We next estimated the relative proportions of nuclei that was contributed by each PPC sample. We found that, like in scRNA-seq[118], each sample contained >70% of PPC (early and late; mean=87.2%, range=71.1-97.4%) (**Figure 2.1c**). Comparison with scRNA-seq also revealed a strong correlation in the relative proportions between the two datasets (**Figure 2.2**). Altogether, our results are consistent with scRNA-seq that PPC contain limited heterogeneity at both the transcriptome and open chromatin level.

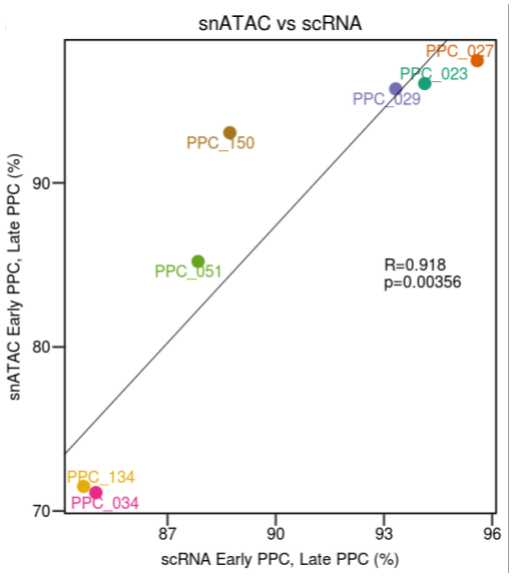


Figure 2.2. Correspondence of relative cell type fractions between scRNA and snATAC
We next determined whether the cell type fractions in snATAC-seq corresponds to scRNA-seq. Because snATAC-seq was not able to detect early PPC and early DE, we reasoned that these cells were included in the main PPC cluster. Therefore, we computed the fraction of cells that are early DE, early PPC, late PPC, and replicating late PPC in scRNA-seq and compared it to the PPC fraction in snATAC-seq. We observed a significant correlation between scRNA-seq and snATAC-seq ($r = 0.918$, $p = 0.00356$).

2.4 Chromatin accessibility profiles of PPC reflects a developmental-specific regulatory landscape

To determine the genome-wide location of cis-regulatory elements (CREs) in PPC, we performed bulk ATAC-seq sequencing in 109 PPC samples, differentiated from 108

iPSC lines from 108 iPSCORE donors (one iPSC line was differentiated twice) (**Table 2.1**). We identified consensus peaks by calling peaks from 24 high-quality reference samples from unrelated individuals (**Figure 2.3**, see Methods). We identified a total of 289,980 ATAC-seq peaks (i.e., regions of accessible chromatin) across the genome, of which 193,428 were used for downstream analyses after removing peaks on sex chromosome or with low accessibility (TMM < 1 in at least 20% of the samples). Across the 109 samples, chromatin accessibility profiles were highly correlated with one other with Spearman correlations ranging from 0.81 to 0.98, indicating that ATAC-seq profiles were consistent and reproducible across the PPC samples.

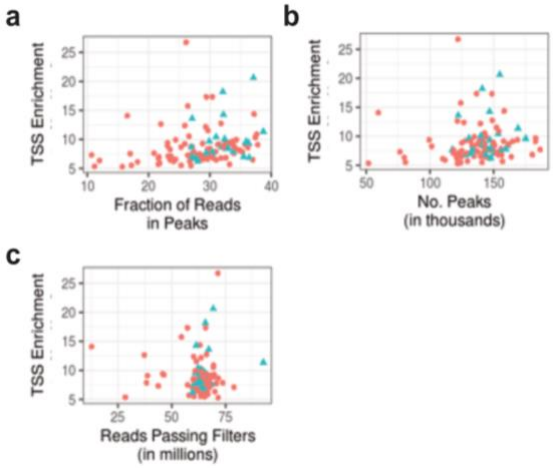


Figure 2.3 Selection of High-Quality Reference Samples for Peak-Calling
Scatter plots showing the transcription start site enrichment (TSSE; calculated by the ATACseqQC R package) of the samples plotted against (a) the fraction of reads in peaks (FRiP), (b) the number of ATAC-seq peaks per sample, and (c) the number of reads passing filters.

To assess the regulatory potential of the CREs identified in ATAC-seq, we performed a motif enrichment analysis comparing PPC ATAC-seq peaks with those identified in adult islets[119]. We observed that the most enriched motif found in the PPCs was a CTCF motif ($p=10^{-4793}$) (**Figure 2.4a**). Transcription factors relevant to pancreas

development were also enriched, including FOXA2 ($p=10^{-2483}$), TEAD1/3 ($p=10^{-1931}$ and 10^{-1932}), GATA2 ($p=10^{-1758}$), PDX1 ($p=10^{-946}$), and NKX6-1 ($p=10^{-465}$) (**Figure 2.4a**). We next assessed the overlap between PPC ATAC-seq peaks and PPC chromatin states[14] and found that the majority of the peaks that were highly shared across the 109 samples overlap primarily either active promoter or enhancer regions (**Figure 2.4b**). As we increase the TMM threshold, we observed that highly expressed peaks overlap primarily active promoter regions. Together, these results show that high chromatin profiles of ATAC-seq of PPC capture cis-regulatory elements that are important in gene regulation during early pancreas development.

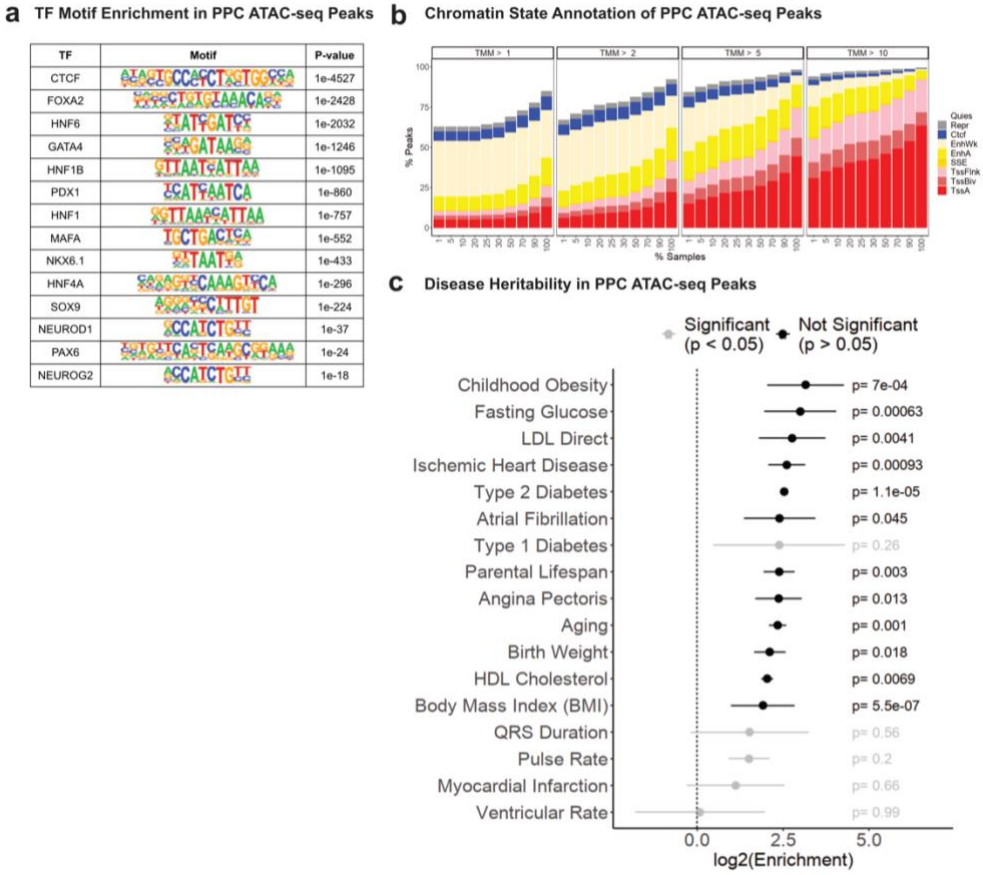


Figure 2.4 Accessible Chromatin Profiles of Pancreatic Progenitor Cells
(a) TF motif enrichment in ATAC-seq peaks in PPC compared to those identified in adult pancreatic islets. (b) Overlap with PPC chromatin states[14] at different TMM thresholds and percentage of samples that express the peaks. We observed that highly shared and expressed peaks tend to represent active regulatory regions compared to peaks with low accessibility and less shared among samples. (c) log2(Enrichment) of GWAS variants in PPC ATAC-seq peaks. Colored enrichments represent significance ($p<0.05$).

2.5 Accessible chromatin of PPC is enriched for trait heritability

GWAS have shown that trait-associated variants were enriched in non-coding chromatin, suggesting that these variants may overlap distal regulatory elements. To assess whether CREs in PPCs were enriched for trait-associated variants, we applied LD Score Regression[120,121] to test the enrichment of heritability in PPC ATAC-seq peaks for

developmental-, longevity-, and pancreas-associated traits (see Methods). As negative controls, we included cardiac-related traits, including angina pectoris, QRS duration, pulse rate, acute myocardial infarction, and ventricular rate.

As expected, we observed strong enrichments (p-value < 0.05) for variants associated with pancreas-associated traits, including childhood obesity, fasting glucose, and type 2 diabetes (**Figure 2.4c**). Variants associated with birth weight was also found to be significantly enriched in PPC ATAC-seq peaks, suggesting a potential developmental role of pancreas development. Interestingly, we observed enrichments for ischemic heart disease and atrial fibrillation. This observation may be explained by that diabetes is a risk factor for heart dysfunction and may share common risk variants (**Figure 2.4c**). Finally, enrichments for heritability of several cardiac phenotypes, including QRS duration, pulse rate, myocardial infarction, and ventricular, was not significant, validating that active regulatory elements during pancreas development are not highly associated cardiac phenotypes and disease. Together, these results indicate that regulatory elements active during pancreas development are enriched for genetic heritability of adult pancreatic traits and disease.

2.6 Chromatin accessibility QTL analysis identifies regulatory variation in PPC

To identify regulatory variants associated with accessible chromatin during early pancreas development, we tested the association between the genotypes for common SNPs (MAF>5%) and chromatin accessibility for 193,428 ATAC-seq peaks, using a linear mixed model to account for the genetic relatedness between the subjects (**Figure 2.5a**). We included the following variables as covariates in the model: 1) iPSC passage to account

variabilities explained by iPSC passage number, 2) genotype principal components to account for global ancestry, and 3) PEER factors to account for hidden biological and technical confounders of molecular variability. For each expressed ATAC-seq peak, we tested the association for only variants within 100 kb of the peak boundaries. Because a genomic locus can harbor multiple causal variants, we performed conditional analyses to identify additional independent caQTLs.

In total, we identified 12,236 caQTLs corresponding to 10,313 peaks (caPeaks) (**Figure 2.5b**). Of the 10,313 caPeaks, 1,718 (16.7%) had more than one independent signals, suggesting that accessibility of these regions may be affected by multiple independent genetic variants. One mechanism by which genetic variants can disrupt chromatin accessibility is through alteration of transcription factor (TF) binding sites (TFBS). To identify potential TFBS in each peak, we performed TF footprinting analysis using TOBIAS[78] along with motifs for 512 PPC-expressed TFs in the JASPAR 2020 database[122]. We identified 3,667 (30% of 12,236) caPeaks that harbored at least one putative TFBS. Of note, we observed that caPeaks were more likely to harbor a TFBS compared to non-caPeaks (two-sided Fisher's Exact Test [FET] odds ratio = 1.6, p-value=1.3×10^{-90}). In addition, motif enrichment analyses revealed a strong enrichment of caPeaks for motifs of TFs relevant to pancreas development compared to non-caPeaks. These TFs include FOXA1 ($p=10^{-38}$), FOXA2 ($p=10^{-28}$), NKX6-1 ($p=10^{-13}$), and PDX1 ($p=10^{-8}$). Together, these results suggest that a potential mechanism of caQTLs in PPC could involve the disruption of binding of key developmental TFs.

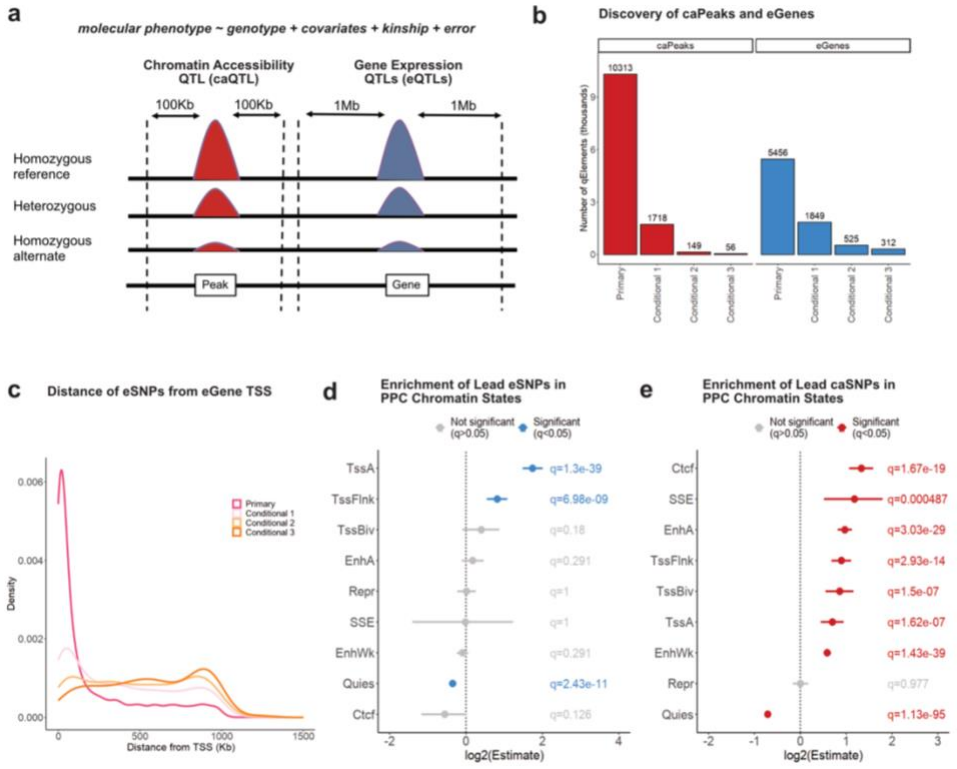


Figure 2.5 Chromatin Accessibility QTLs.
(a) Schematic depicting genotype effects on chromatin accessibility (red) and gene expression (green) detected by caQTL and eQTL analyses, respectively. (b) Number of caPeaks and eGenes discovered along with their conditionals. (c) Distribution of the distance between lead eSNPs from their corresponding eGene TSS's. (d) Enrichment of lead eSNPs in PPC chromatin states. (e) Enrichment of lead caSNPs in PPC chromatin states). Enrichments were tested using a two-sided Fisher's Exact Test. Enrichments with p-value < 0.05 were considered significant.

2.7 QTL modules provide insights into putative biological roles of regulatory variants

One challenge in elucidating the mechanisms of GWAS loci is that the implicated variants often lie far from the genes they influence. To identify putative target genes for caQTLs, we leveraged the RNA-seq data generated in Chapter 1 for matched PPC samples (n=107). To maintain consistency with the hg38 genome build of the caQTLs, we re-

mapped the eQTLs using the same linear mixed model and identical covariates employed in the discovery of caQTLs. In this updated analysis, we used a 1 Mb window from the gene body instead of the 500 kb window that was used previously. Including results from conditional analyses, we identified a total of 8,142 eQTLs for 5,456 genes, increasing the number of eQTLs by 3,709 (83% increase from 4,433) and the number of eGenes by 1,391 (34% increase from 4,065). We observed that the majority of the new eQTLs corresponded to conditional eQTLs. Given that conditional eQTLs tend to be more distal to the TSS[123] (**Figure 2.5c**), these results suggest that increasing the window of variant testing may improve the detection of more distal eQTLs}.

We next assessed the enrichment of eQTLs and caQTLs in PPC chromatin states[14]. We observed a significant enrichment of lead eSNPs in active Tss (two-sided FET p=1.3e-39) and flanking Tss (two-sided FET p=6.9x10^{-9}) regions (**Figure 2.5d**), while on the other hand, lead caSNPs were most enriched in CTCF binding sites (two-sided FET OR = 2.5, p = 7.4x10^{-20}), followed by PPC-specific stretch enhancer (SSE) regions (two-sided FET OR = 2.3, p = 0.000447), active enhancers (two-sided FET OR = 2.0, p = 1.0x10^{-29}), and Tss flanking (two-sided FET OR = 1.9, p=1.6x10^{-14}), bivalent (two-sided FET OR = 1.8, p = 1.0x10^{-8}), and active (two-sided FET OR = 1.6, p = 1.0x10^{-7}) regions (**Figure 2.5e**). These results confirm that eQTLs and caQTLs identify different types of regulatory variation. Specifically, eQTLs identify regulatory variants that are proximal to promoter regions while caQTLs identify both proximal and distal regulatory variants but have a stronger preference towards the latter.

We next sought to identify shared genetic associations between caQTLs and eQTLs. We performed pairwise Bayesian colocalization and identified a total of 2,264

colocalizations (posterior probability that both signals are shared; PP.H4 ≥ 80%) between 1,173 caQTLs and 1,129 eQTLs. We also identified singleton QTLs that did not colocalize with any other QTL, including 11,206, caQTLs and 7,289 eQTLs (**Figure 2.6a**). To identify QTLs that were associated with multiple qElements (i.e., a caPeak or an eGene), we created networks by loading the colocalized QTL pairs as edges and identified 790 QTL modules composed of two or more QTLs. 331 QTL modules were associated with only two qElements while the remaining 459 were associated with three or more qElements (**Figure 2.6a**). The two largest QTL modules comprised of 9 qElements, one was associated with 9 caPeaks while the other was associated with 8 caPeaks and one eGene (**Figure 2.6a**). We next annotated each QTL module according to the types of QTLs it contained. 60% modules (n=474) comprised of at least one caQTL and at least one eQTL ("caQTL-eQTL" in **Figure 2.6b**), while 24% (n=193) comprised of only caQTLs ("caQTL" in **Figure 2.6b**) and 15.5% (n=123) comprised of only eQTLs ("eQTL" in **Figure 2.6b**). The observation that about a fourth of the QTL modules only contain caQTLs suggest that these analyses capture complex regulatory variation in enhancers and other elements that is missed by eQTL analyses. We next assessed the correlation between effect sizes of the lead colocalized SNPs (i.e., SNP with the highest causal PP for both caQTL and eQTL associations) between 740 caPeak-eGene pairs within the same modules and observed a strong correlation between the two molecular phenotypes (Pearson correlation $r = 0.46$, $p = 1.01 \times 10^{-53}$) (**Figure 2.6c**).

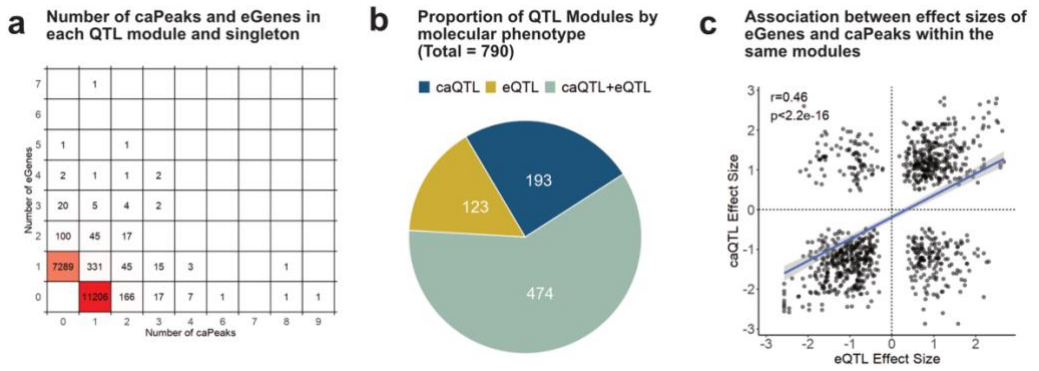


Figure 2.6 QTL modules represent complex regulatory loci
(a) Distribution of eGenes and caPeaks across QTL modules (in white) and singleton QTLs (colored in red). For example, there were 11,206 caQTL singletons, and there were 166 QTL modules associated with 2 caPeaks and 0 eGenes. (b) Proportion of QTL modules based on the QTL types they are associated with. 474 QTL modules comprised of both a caQTL and an eQTL, 123 QTL modules comprised of only eQTLs, and 193 QTL modules comprised of only caQTLs. (c) Correlation of effect sizes between eGenes and caPeaks within the same modules.

2.8 Fetal-unique regulatory variants are under high evolutionary constraint and tend to be distal to their eGenes.

Given that PPCs resemble the fetal-like stage, we next sought to identify QTLs that were unique to the PPCs and not active in adult tissues. We calculated LD between lead fine-mapped variants for all 20,378 PPC QTLs (12,236 caQTLs, 8,142 eQTLs) and adult QTLs from the following datasets: 1) adult primary and conditional eQTLs from 49 GTEx tissues[10], 2) adult caQTLs from QTLbase2[124], 3) adult pancreatic islet eQTLs[11], 4) adult pancreatic islet caQTLs[119], and 5) adult haQTLs from QTLbase2[124].

We considered a PPC QTL to be fetal-unique if the lead fine-mapped variant was not in LD ($r^2 < 0.2$ within 500 kb) with an adult QTL. Of the total 20,378 QTLs, we identified 1,995 (16.3%) caQTLs and 634 (7.8%) eQTLs that were potentially fetal-unique (**Figure 2.7a**). Of these, 57 (18 eQTLs, 39 caQTLs) were in modules while 2,572 (616 eQTLs, 1,956 caQTLs) were singletons.

Because distal regulatory elements, such as enhancers, are often more context-specific compared to proximal regulatory elements[125], we sought to determine whether fetal-unique eQTLs were more likely to be distal or proximal compared to eQTLs that were shared with adult. We observed that fetal-unique eQTLs were more likely to be further away from the eGene TSS compared to shared eQTLs (one-sided Wilcoxon Test p-value = 2.0×10^{-61}) (**Figure 2.7b**). Given that conditional eQTLs were more distal to the gene promoter compared to primary eQTLs, we also assessed whether conditional eQTLs were more likely to be fetal-unique compared to primary eQTLs. We observed a strong enrichment (two-sided FET odds ratio = 0.34, p-value = 2.2×10^{-16}) of conditional eQTLs (4.0%) for detecting fetal-unique eQTLs compared to primary eQTLs (1.6%).

Given the precise regulation of embryonic development and its significant for adult health and disease [126], we hypothesized that fetal-unique eQTLs may be under higher evolutionary constraint compared to adult-shared eQTLs. We annotated 3,934 eGenes with the probability of loss function tolerance (pLI) from the gnomAD database and removed 185 eGenes that had both a fetal-unique and adult-shared eQTL. This removal resulted in 193 fetal-unique and 3556 adult-shared eGenes remaining. Using the remaining eGenes, we observed that fetal-unique eGenes have higher pLI scores compared to adult-shared eGenes (one-sided Wilcoxon Test p-value = 0.041) (**Figure 2.7c**). Given that distal regulatory variation tend to have lower effect sizes compared to those proximal to the gene, we compared the absolute effect sizes between fetal-unique eQTLs and adult-shared eQTLs and observed that fetal-unique eQTLs have significantly lower effect sizes (p = 0.002).

Together, these results show that in the iPSCORE fetal-like tissues, regulatory variation that are uniquely active during early fetal development exhibit distinct characteristics compare to those shared with adult tissues. Particularly, fetal-unique eQTLs tend to be distal to their eGenes, have weaker genetic effects on gene expression, and tend to be detected as conditional QTLs, compared to adult-shared QTLs. Further, we show that fetal-unique eGenes are associated with higher evolutionary constraint, suggesting regulatory variation during fetal development may have a stronger impact on phenotypes and potentially disease compared to those shared with adult.

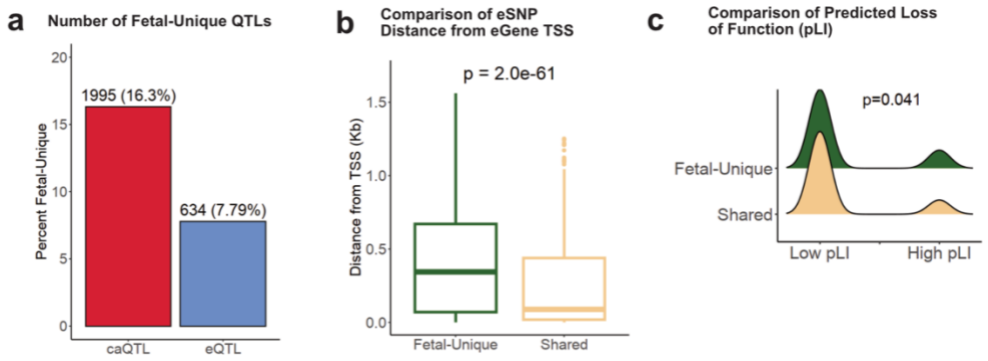


Figure 2.7 Fetal-unique eQTLs are associated with high evolutionary constraint
(a) Bar plot showing the percentage of fetal-unique eQTLs and caQTLs in PPC, (b) Box plot showing the distribution of distance from the TSS between fetal-unique and shared eQTLs. (c) Ridge plot showing that fetal-unique eGenes were associated with higher pLI scores compared to adult-shared eGenes.

2.9 Multi-omic QTLs in PPC are associated with complex pancreatic traits

To identify PPC QTLs that were associated with adult complex traits and diseases, we performed Bayesian colocalization between all 20,378 PPC QTLs (12,236 caQTLs, 8,142 eQTLs) and 2,392 genome-wide significant GWAS signals associated with the following 10 pancreas-related traits: childhood obesity, aging, parental lifespan, birth weight, fasting glucose, type 1 diabetes, type 2 diabetes, LDL direct, HDL cholesterol, and

body mass index. Specifically, we conducted pairwise colocalization between each genome-wide significant GWAS signals and each QTL that had a minimum of 50 overlapping variants.

In total, 210 (8.0%) colocalized with at least one PPC QTL (**Figure 2.8a-b**). 79% (n=165) of the 210 loci colocalized with QTLs associated with only one molecular phenotype while 22% colocalized with both a caQTL and an eQTL. HDL cholesterol exhibited the highest number of colocalizations (n=63), where 23 loci colocalized with only eQTLs, 30 with only caQTLs, and 10 with both a caQTL and an eQTL. Of note, 45% of the GWAS loci colocalized with only a caQTL, indicating that inclusion of epigenomic QTLs resulted in a 1.8-fold increase in the number of GWAS loci annotated with a molecular phenotype. The discrepancy between caQTL and eQTL overlap can be attributed to the different types of variants they identify, where caSNPs are more likely to capture distal regulatory variants while eSNPs are more likely to capture those closer to the promoter (**Figure 2.5d-e**).

a

		Number of GWAS loci that colocalized with…				
Trait	**Total Signals**	**Any QTL**	**eQTL only**	**caQTL only**	**Both**	**Fetal-unique**
Childhood Obesity	12	2	1	1	0	0
Parental Lifespan	20	3	0	2	1	0
Aging	26	3	0	2	1	0
Birth Weight	53	3	1	1	1	0
Fasting Glucose	80	6	4	1	1	0
Type 1 Diabetes	103	11	6	4	1	0
Type 2 Diabetes	221	33	11	13	9	0
LDL Direct	278	25	8	10	7	0
HDL Cholesterol	657	63	23	30	10	1
Body Mass Index	847	61	17	30	14	2
Total	**2297**	**210**	**71**	**94**	**45**	**3**

b

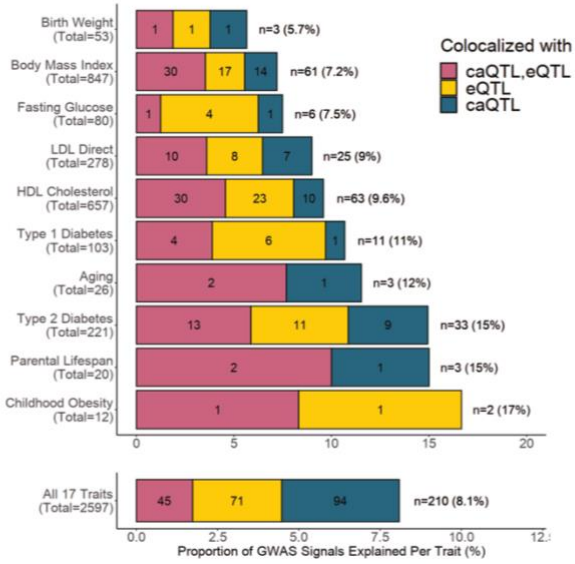


Figure 2.8 PPC QTLs are associated with GWAS variants
(a) Table describing the number of GWAS signals that colocalized with different types of PPC QTLs. (b) Percentage of GWAS loci colocalized with QTLs per trait and color-coded by QTL type.

Given that QTL modules provide mechanistic insights into shared regulatory variation across multiple caPeaks and/or eGenes, we sought to annotate each of the 210 GWAS loci with the QTL modules and singletons they colocalized with. Because a QTL can colocalize with multiple GWAS traits, and a GWAS locus can colocalize with multiple QTLs, we do not observe a one-to-one correspondence. For example, type 2 diabetes had 33 GWAS loci that colocalized with QTLs (**Figure 2.8b**), and these QTLs correspond to

11 QTL modules and 25 QTL singletons (**Figure 2.9**). In total, 158 singletons and 38 QTL modules colocalized with a GWAS locus. Of these QTL modules, 27 (71%) corresponded to modules that were associated with both molecular phenotypes, while the remaining 11 (19%) corresponded to modules associated with only one molecular phenotype. On the other hand, of the 158 singletons, there was an equal distribution of caQTL singletons (87, 55%) and 71 eQTL singletons (45%).

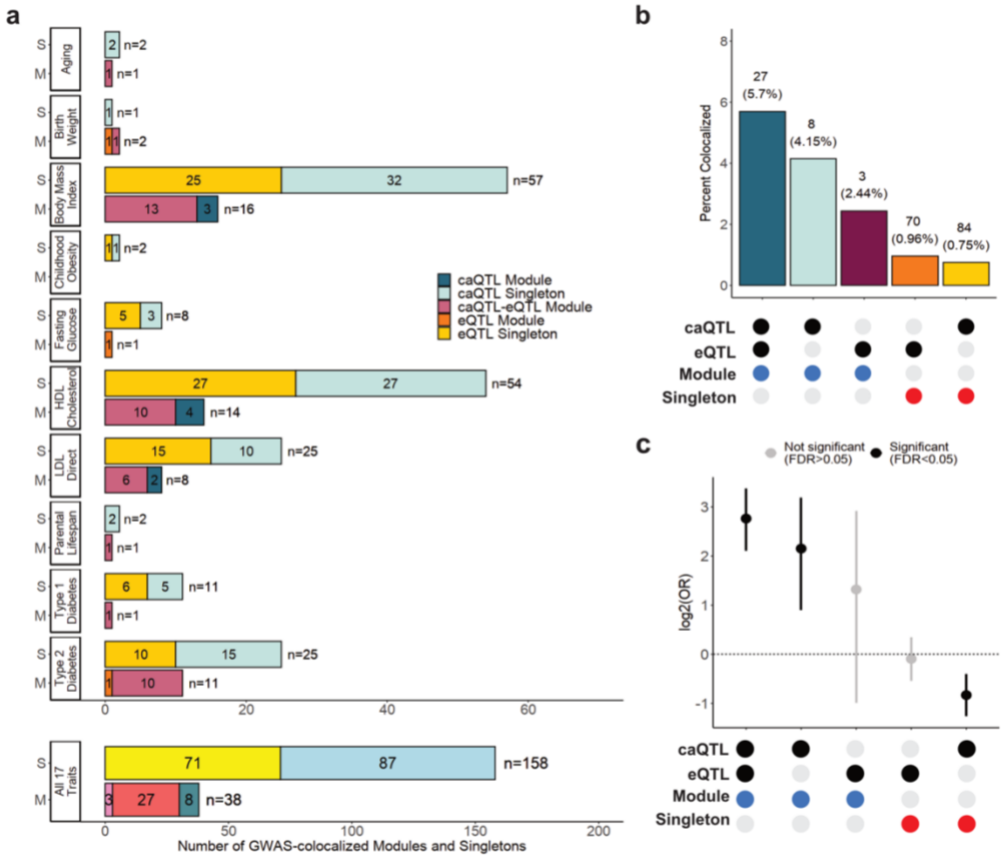


Figure 2.9 Genetic variants with regulatory complexity is enriched for GWAS colocalization
(a) Bar plot showing the number of QTL modules and singletons that colocalized with each GWAS trait, color-coded by the molecular phenotypes the QTL was associated with. (b) Bar plot showing the percentage of each QTL category that colocalized with GWAS. (c) Enrichment of each QTL category for colocalization with GWAS. Significance was tested using a two-sided Fisher's Exact Test, where significance was determined by Bonferroni-corrected p-value < 0.05.

Given these results, we next sought to assess the enrichment of GWAS colocalization for QTLs in these five categories: caQTL-eQTL module, caQTL only module, eQTL only module, eQTL singleton, and caQTL singleton. Notably, we observed that caQTL-eQTL modules were most enriched for GWAS colocalization (two-sided FET odds ratio = 6.8, $p = 2.7x10^{-13}$), followed by caQTL only modules (two-sided FET odds ratio = 4.4, $p = 7.2x10^{-4}$). In addition, eQTL singletons were not enriched, and caQTL singletons were depleted, suggesting that many of the singleton associations may not be relevant to traits and disease. Altogether, these findings show that regulatory variation during early pancreas development is associated with adult complex traits and disease. Further, multi-phenotype QTL modules were enriched for GWAS colocalization, indicating that regulatory variation that affects multiple phenotypes may more likely influence disease, further supporting the use of multi-omic approaches to characterize GWAS mechanisms.

2.9 Fetal-unique PPC caQTL is associated with a BMI GWAS locus

Three GWAS loci colocalized with a fetal-unique PPC QTL, two of which corresponded to body mass index and one corresponded to HDL cholesterol. These GWAS loci each colocalized with a PPC singleton caQTL. Notably, we observed that fetal-unique QTLs were depleted for GWAS colocalization compared to those that were shared in adult stage (**Figure 2.10**). This result suggests that the majority of GWAS loci can be largely explained by adult molecular phenotypes while a handful can be explained by fetal development. We identified the chr4:54544847-54744637 locus that was associated with body mass index and colocalized with a fetal-unique caQTL singleton in the PPC (ppc_atac_peak_197599). Fine mapping identified rs13140079 as the lead variant

(chr4:546310000:C>T, PP = 22.4%) for this locus, which strongly associated with normalized chromatin accessibility of ppc_atac_peak_197599. rs13140079 is located 27 kb upstream of *KIT* that encodes a tyrosine kinase receptor involved in c-KIT signaling, which is important for many cellular processes, including cell proliferation, survival, migration, metabolism and plays a role in pancreas beta cell development and survival[127]. In particular, mice with a mutation in *Kit* have been shown to develop[127] early onset of diabetes due to impaired glucose tolerance, decreased insulin secretion, and a marked reduction in beta cell mass[128,129]. Our findings suggest that a fetal-unique regulatory element (ppc_atac_peak_197599) located 27 kb upstream of *KIT* may disrupt c-KIT signaling during early pancreas development and therefore influence metabolism and body mass index later in life.

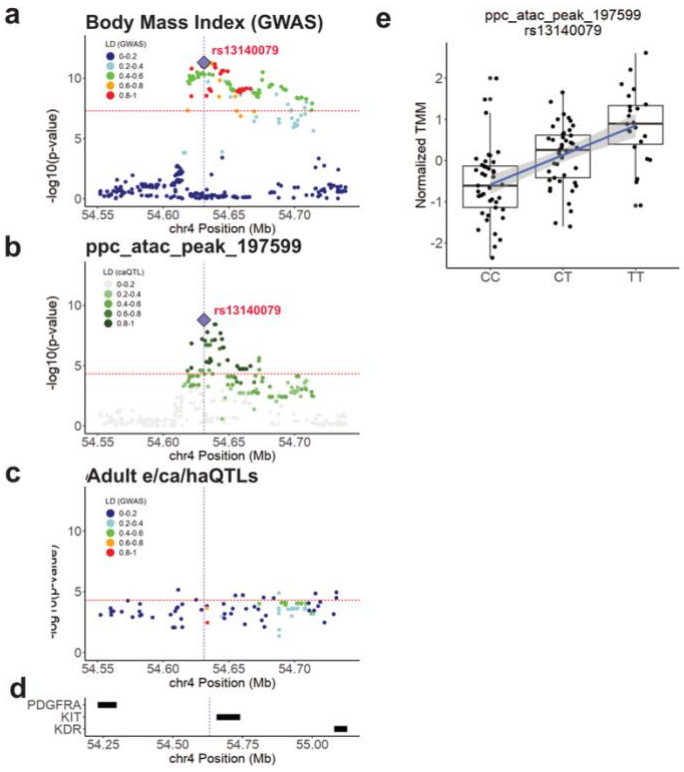


Figure 2.10 A BMI Locus is associated with a fetal-unique PPC caQTL
(a) A body mass index GWAS locus that colocalized with a PPC fetal-unique caQTL singleton. Panel (a) shows the -log10(p-values) from the GWAS body mass index. Panel (b) shows the -log10(p-values) for the PPC caQTL. Panel (c) shows the -log10(p-values) for all adult eQTLs, caQTLs, and haQTLs. Panel (d) are the hg38 coordinates for genes within the GWAS locus. Red horizontal lines indicate genome-wide significance thresholds for GWAS (p = 5e-08) and e/ca/haQTL (p = 5e-05) for plotting purposes. Each variant was colored according to their LD with the lead fine-mapped variant (purple diamond; rs13140079, chr4:546310000:C>T, PP = 22.4%) using the 1000 Genomes Phase 3 Panel (Europeans only) as reference. (e) plot showing the association between the lead fine-mapped variant (rs13140079, chr4:546310000:C>T, PP = 22.4%) and normalized chromatin accessibility of ppc_atac_peak_197599.

2.10 Discussion

Previous studies examining expression quantitative trait loci (eQTLs) have been instrumental in elucidating the biological underpinnings of GWAS loci associated with complex human traits and diseases[10–12,44,118]. However, such eQTL studies have only been able to explain 43% of the loci identified by GWAS. This modest coverage may be

attributed by two key factors: first, regulatory variation may exert their effects in a developmental-specific manner[9,12,13,118]; and second, eQTLs tend to be biased towards promoter regions while GWAS hits tend to be in non-coding distal regions[130]. Given these two limitations, it is important to consider multi-layered genomic approaches that captures not only variants in close proximity to promoter regions but also those at more remote, distal regions of the genome.

To complement our previous eQTL analyses, we have generated high-quality maps of chromatin accessibility for the same set of ~100 samples (with the addition of two samples). We show that these maps capture regulatory sequences that are specific to pancreas development, including motifs of FOXA1, PDX1, and NKX6-1. caQTL analyses identify 12,236 caQTLs that are associated with the changes in accessibility of 10,313 caPeaks, which are enriched for motifs of pancreas developmental TFs compared to non-caPeaks.

To address the first limitation that regulatory variation may act in a developmental-specific manner, we sought to identify candidate caQTLs and eQTLs that may be unique to fetal development. In total, we identified candidate 1,995 caQTLs and 634 eQTLs that are unique to fetal development. Interestingly, we observed distinct characteristics between fetal-unique QTLs compared to those that were shared with adult. In particular, we found that fetal-unique eQTLs were more distal to their eGenes compared to shared eQTLs. Further, fetal-unique eGenes exhibited higher pLI scores compared to shared eQTLs, suggesting that these fetal-unique genes may be subject to stronger evolutionary constraint compared to adult-shared eGenes. Despite the large number of fetal-unique QTLs, we observed that only three caQTL singletons colocalized with GWAS variants, indicating a

strong depletion of fetal-unique QTLs for colocalization with GWAS. These results suggest that the majority of the genetic effects on adult phenotypes may largely occur at the adult stage and may be minimally contributed by fetal development. Future validation analyses such as mashr[91] to assess effect size differences between contexts are needed to better prioritize and evaluate QTLs that are unique to fetal development.

Next, we assessed the utility of epigenomic QTLs in the functional characterization of GWAS loci. Chromatin state enrichment analyses showed that eQTLs and caQTLs capture distinct types of genetic variants, where eQTLs largely capture variants in the promoter region while caQTLs largely capture variants in distal regulatory regions. Network analyses has uncovered distinct QTL modules composed exclusively of caQTLs, highlighting regulatory sites detected through only caQTLs but not eQTLs. Moreover, the integration of caQTLs with eQTLs resulted in a substantial 1.8-fold increase in the number of GWAS loci colocalization as well as improved the detection of GWAS-associated QTLs. Specifically, QTL modules that were associated with multiple phenotypes were more likely to colocalize with GWAS variants compared to QTL modules and singletons that were associated with a single molecular phenotype. Together, these results underscore the value of utilizing a multi-omics approach to elucidate the genetic mechanisms underlying obesity and diabetes associations.

This dataset provides a unique and valuable resource for studying regulatory variation underlying gene expression and disease during early pancreas development. We identified 12,236 caQTLs and 8,142 eQTLs, 8-16% of which were specifically active during early development. We show that caQTLs identified additional regulatory loci that are missed by eQTLs due to their distinct properties to detect distal regulatory variation.

Further, inclusion of caQTLs increased the numbers of GWAS colocalizations by 1.8-folds. In total, we identified 112 genes and 155 regulatory elements that have potential associations with obesity and diabetes risk and may be strong candidates for future functional studies.

2.11 Methods

1. Subject Information

This study involves 109 PPC samples from 108 iPSCORE subjects[35,118]. Of these 109 PPC samples, 107 (from 106 subjects) has matching RNA-seq samples[118]. Of the 109 iPSCORE individuals, 55 belong to 20 families composed of two or more subjects (range: 2-6 subjects). Each subject was assigned a Universal Unique Identifier (UUID) and an iPSCORE_ID (i.e, iPSCORE_4_1) which donates family (4) and individual number (1). Sex, age, and self-reported race/ethnicity were recorded at the time of enrollment. We previously estimated the ancestry of each subject by comparing their genomes to those of individuals in the 1000 Genomes Project (KGP)[35]. Recruitment of individuals was approved by the Institutional Review Boards of the University of California, San Diego, and The Salk Institute (project no. 110776ZF). The iPSC lines in the iPSCORE resource are available to non-profit organizations through WiCell Research Institute (www.wicell.org). For-profit organizations can contact the corresponding author directly to discuss availability of iPSC lines as well as differentiated cell types.

2. Library Generation

2.1 ATAC-seq

All ATAC-seq samples were processed in the same manner using a modified version of the Buenrostro et al. protocol [131] as previously described[132]. Briefly, frozen nuclear pellets of $1x10^5$ PPC cells were thawed on ice and tagmented in total volume of 25µl in permeabilization buffer containing digitonin (10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM $MgCl^2$, 0.01% digitonin) and 2.5µl of Tn5 from Nextera DNA Library Preparation Kit (Illumina) for 45-75min at 37°C in a thermomixer (500 RPM shaking). We included a double size selection step during purification using AMPure XP DNA beads (Beckman Coulter). To eliminate confounding effects due to index hopping, all libraries within a pool were indexed with unique pairs of i7 and i5 barcodes. Libraries were amplified for 12 cycles using NEBNext® High-Fidelity 2X PCR Master Mix (NEB) in total volume of 25µl in the presence of 800nM of barcoded primers (400nM each) custom synthesized by Integrated DNA Technologies (IDT) and sequenced with a combination of 100 bp paired-end and 150 bp paired-end reads on an Illumina HiSeq4000.

2.2 snATAC-seq

A total of 7 PPC samples were used for snATAC-seq generation (**Table 2.1**). Cells from seven cryopreserved PPCs samples were captured for snATAC-seq immediately after thawing. All seven samples have matched scRNA-seq. Cells from four cryopreserved PPC samples were pooled (ATAC_Pool_1) and cells from the other 3 PPC samples were pooled (ATAC_Pool_2) prior to capture. Nuclei from two pools were isolated according to the manufacturer's recommendations (Manual CG000169, Rev B), transposed, and captured as independent samples according to the manufacturer's recommendations (Manual CG000168, Rev B). All single nuclei were captured using the 10x Chromium controller (10x Genomics) according to the manufacturer's specifications and manual (Manual

CG000168, Rev B). Cells for each sample were loaded on the individual lane of a Chromium Chip E. Libraries were generated using Chromium Single Cell ATAC Library Gel & Bead Kit (10x Genomics) following manufacturer's manual (Manual CG000168, Rev B). Sample Index PCR material was amplified for 11 cycles. Libraries were sequenced using a custom program (50-8-16-50 Pair End) on HiSeq 4000. Specifically, two libraries from seven cryopreserved PPC samples (ATAC_Pool_1 and ATAC_Pool_2) were each sequenced on an individual lane.

3 Data Processing

3.1 WGS

We downloaded the VCF in hg19 for 273 iPSCORE individuals from dbGaP (phs001325.v3), phased the 273 WGS with the Michigan Imputation Server using the 1000 Genomes 30x GRCh38 as the reference panel[133–135], and performed liftOver to hg38 using CrossMap[136] and the hg38 reference genome from UCSC (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/).

3.2 RNA-seq

All RNA-seq samples were processed in a uniform manner. Libraries that were sequenced more than once were merged by concatenating the FASTQ files. The reads were aligned onto the hg38 human reference genome downloaded from Gencode version 44 [47,95] using STAR 2.7.10b (https://github.com/alexdobin/STAR) with the following parameters: --outSAMattributes All --outSAMunmapped Within --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000. PCR duplicates were marked with Picard

(https://github.com/broadinstitute/picard) (v3.1.0) and counted using samtools flagstat [137] (v1.17). Number and percentage of mapped reads were calculated using samtools flagstat [137] (v1.17). Percentage of intergenic and mRNA bases were determined using Picard (v3.1.0) CollectRnaSeqMetrics. Gene TPM expression and read counts were calculated using RSEM[138] (v1.3.3) with gene annotations from Gencode version 44[47,139] (hg38) and the following parameters: --seed 3272015 --estimate-rspd --forward-prob 0 --paired-end. RNA-seq samples were examined for quality using GTEx standards[10]. Specifically, we required that samples met the following metrics: 1) the number of mapped reads > 10 million; 2) percent of intergenic bases < 30; 3) percent of mRNA bases > 70; 4) percent of duplicate reads < 30; 5) percent mapped reads > 85%; 6) number of reads passing filters > 25M, and 7) matched via a sample identity check to the correct subject with PI_HAT > 90%.

For eQTL mapping, gene expression values were normalized and filtered using the same procedure as GTEx[10]. Specifically, 1) read counts were TMM normalized across all genes using edgeR[140] (v3.38.4) with functions *DGEList*, *calcNormFactors*, and *cpm*; 2) autosomal genes were selected and filtered based on expression thresholds of 0.1 TPM in 20% of samples and 6 reads (unnormalized) in 20% of samples; 3) TMM expression values for each gene were inverse normal transformed across samples using *rank* and *qnorm* in R v4.2.1 and used as input for eQTL analyses. This resulted in 20,738 genes used for eQTL mapping.

3.3 ATAC-seq

All ATAC-seq samples were processed in a uniform manner using the same procedure as the ENCODE (https://github.com/ENCODE-DCC/atac-seq-pipeline). Illumina adapters were removed from the reads using cutadapt[141]. Reads were aligned using BWA MEM (https://bio-bwa.sourceforge.net/bwa.shtml) onto the hg38 human reference genome from UCSC (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/). Multi-mapped reads were randomly assigned using ENCODE's custom script (assign_multimappers.py) (https://github.com/ENCODE-DCC/atac-seq-pipeline). Using samtools [137], reads that were either unmapped, not in primary alignment, failed Illumina QC metrics, or had an unmapped mate were removed (samtools view -F 1804). Properly paired reads with mapping quality 30 were retained (samtools view -f 2 -q 30). Duplicates were marked by Picard and then removed with samtools. Mitochondrial reads were also excluded from downstream analyses. Filtered BAM files were converted to bed files (bedtools bamtobed) and shifted for Tn5 bias, and then used to call narrow peaks using MACS2 [142] with ENCODE default parameters (https://github.com/ENCODE-DCC/atac-seq-pipeline): –shift 75 –extsize 150 -q 0.01 –nomodel -B –SBMR –keep-dup all. Peaks overlapping blacklisted regions were removed. ATAC-seq samples were examined for quality and excluded if they did not pass one of the following metrics: 1) non-redundant fraction (NRF) > 0.9; 2) PCR-bottlenecking coefficient 1 (PBC1) > 0.9; 3) PCR-bottlenecking coefficient 2 (PBC2) > 3; 4) percent of mapped reads > 0.95; 5) fraction of reads in peaks (FRIP) > 10; 6) TSS enrichment (TSSE) > 4, and 7) matched via a sample identity check to the correct subject with PI_HAT > 90%. Across all ATAC-seq samples the number of read pairs passing filters ranged from 6.4 million to 46.2 million with an average of 31.4 million.

To identify consensus peaks for each tissue, we selected high quality reference samples using the following filters: 1) 25 < FRiP < 45; 2) 5 < TSSE < 25; and 3) 75,000 < number of peaks < 200,000) from unrelated individuals in different families with two or more individuals in the iPSCORE collection. If multiple samples from the same family passed these filters, we selected the sample with the highest TSSE, which resulted in 24 PPC reference samples. For each reference sample, we removed short peaks (<150 bp), then concatenated and merged peaks across all the reference samples for each tissue, resulting in 289,980 PPC reference peaks used for downstream analyses. For each ATAC-seq sample, we then used featureCounts (v2.0.6)[143] to count the number of reads in each reference peak.

For caQTL mapping, we first TMM-normalized the reference ATAC-seq peak counts across samples using the *calcNormFactors* and *cpm* functions in in the edgeR package v3.38.4[140]. We then removed ATAC-seq peaks on sex chromosomes or with low accessibility (TMM < 1 in at least 20% of the samples), resulting in,428 PPC ATAC-seq peaks used for caQTL mapping.

3.3 snATAC-seq

For two snATAC-seq samples (ATAC_Pool_1 and ATAC_Pool_2), we retrieved FASTQ files and used CellRanger V2.0.0 (https://support.10xgenomics.com/) to align files to the hg19 genome using *cellranger-atac count* with default parameters. NarrowPeaks were called using the MACS2 command *macs2 callpeak --keep-dup all --nomodel --call-summits* on the BAM files merged from the two pooled samples and detected 288,813 peaks. Peaks called on ambiguous chromosomes or the mitochondrial genome were

removed, leaving 280,079 peaks remaining. Using these peaks, each snATAC-seq sample was reanalyzed using *cellranger-atac reanalyze* to generate single-nuclei peak counts for each sample. To integrate the two snATAC-seq datasets for downstream analyses, we performed Signac[116] integration by first applying normalization (*RunTFIDF*) and linear dimensional reduction (*FindTopFeatures* and *RunSVD*) on each sample dataset. We then identified a random subset of 20,000 peaks and computed a set of integration anchors between the samples (*FindIntegrationAnchors* for 2,000 anchors) The two snaATAC-seq was integrated using *IntegrateData* and 2-30 most significant dimensions calculated from dimension reduction analyses. Finally, on the integrated dataset, dimension reduction was applied (*RunSVD* for 30 singular values), and single cells were visualized using UMAP (*RunUMAP* on 2:30 dimensions). Clusters were identified using a SNN-graph method using *FindNeighbors* and *FindClusters*. To remove low quality cells, we removed cells that satisfy one of the following critera: 1) the number of peak region fragments < 2,000 or > 20,000, 2) the percentage of reads in peaks < 40%, 3) nucleosome signal > 1.5, or 4) TSS enrichment score < 2.5. Furthermore, we removed cells that do not visually belong to a cluster (i.e. cells that are scattered between two distinct clusters). We performed iterative clustering until we do not observe significant outliers of single cells. After filtering, 25,564 nuclei remained and clustering resolutions of 0.1, 0.15, and 0.2 were tested.

To reassign pooled nuclei back to the original subject from two snATAC-seq samples (ATAC_Pool_1 and ATAC_Pool_2), we applied Demuxlet[96] to the two samples using the same set of reference variants as stated above.

3.4 Sample Identity

Sample identity was performed as previously described[12,35,37,99,118,132] . Briefly, genotypes were called from BAM files of each molecular dataset for common variants with minor allele frequency (MAF) > 45% and < 55% using bcftools[137] mpileup and call, and then compared to WGS genotypes using plink[144] –genome, which calculates IBD between each pair of samples. Samples that matched the correct subject with PI_HAT > 90% passed sample identity check.

4. Analyses

4.1 snATAC cluster annotation

To determine the cell types within the integrated snATAC-seq dataset, we used chromVAR {28825706} within the Signac[116] pipeline to identify transcription factor motifs from the JASPAR 2020[122] database that are enriched for accessible chromatin for each cluster. Specifically, we used the *RunChromVAR* function in Signac[116] and the hg19 reference (BSgenome.Hsapiens.UCSC.hg19) to compute a deviation z-score for each motif in each cell. To annotate the cell types, we examined the motif activities of transcription factors with known developmental or pancreatic functions: TFAP2A/B (mesendoderm), GATA4/6 (PDX1+ progenitors), HNF4A, FOXA1/2, PDX1, NKX6-1 (NKX6-1+ progenitors), PAX4/6, RFX1/3, HNF1A, MAFA, NKX2-2, NEUROD1 (endocrine), and ETV1, ETS1, ETS2 (early ductal). To validate our annotations, we compared the motif activities to their gene expression in scRNA-seq using the same z-normalization method. We examined the motif activity profiles at resolutions 0.1, 0.15, and 0.20, and reasoned that because subclusters within the predominant cluster expressed both PDX1 and NKX6-

1 but at varying levels, we collapsed these clusters into NKX6-1+ progenitors. Resolution 0.1 was used for downstream analyses.

4.2 Bulk ATAC-seq Homer Motif Enrichment

Motif enrichment was performed using the Homer findMotifsGenome.pl with parameters *hg38 --size given*. For Figure 2.5a, enrichment was performed using adult islet ATAC-seq peaks from Khetan et al.[119]. To identify motifs enriched in PPC caPeaks, we performed enrichment analyses using non-caPeaks as background with the same parameters listed above.

4.3 TF Binding Prediction

The TOBIAS[145] algorithm leverages distribution of reads across the genome for a given sample, therefore to profile TF occupancy, we ran TOBIAS to predict binding at 1,012 motifs across ATAC-seq peaks for each tissue, independently. We first merged BAM files for the reference samples used to establish reference peaks for each tissue. We followed the standard workflow in the TOBIAS tutorial (https://github.com/loosolab/TOBIAS). Briefly, for each merged reference BAM file, we applied *ATACorrect* to correct for cut site biases introduced by the Tn5 transposase within the ATAC-seq peaks, using the following parameters: --genome hg38.fa (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/) and –blacklist hg38-blacklist.v2.bed (https://github.com/Boyle-Lab/Blacklist/blob/master/lists/hg38-blacklist.v2.bed.gz). Next, we calculated footprints scores with *ScoreBigwig*, using the corresponding narrowPeak file for each tissue as input. To identify the predicted transcription factor binding sites, we ran *BINDetect* with 747 motifs from JASPAR 2020[122]

using hg38 fasta file and respective narrowPeak file as the genome and regions. For downstream analyses, we used 748 predicted TFBSs (the 747 JASPAR motifs. The tables for predicted TFBSs at JASPAR and HOCOMOCO motifs are deposited on Figshare (see original publication).

4.4 LD Score Regression

To estimate the enrichment of heritability for developmental and adult GWAS traits in ATAC-seq in PPCs, we considered the following 17 traits: fasting glucose, chronic ischemic heart disease, birth weight, type 2 diabetes, LDL direct levels, HDL cholesterol levels, angina pectoris, type 1 diabetes, body mass index, ventricular rate, pulse rate, atrial fibrillation and flutter, QRS duration, and acute myocardial infarction, and childhood obesity. GWAS summary statistics for the 17 traits were downloaded from the 1) UK Biobank (https://pan.ukbb.broadinstitute.org/downloads/index.html) for angina pectoris, atrial fibrillation, body mass index, HDL cholesterol, ischemic heart disease, LDL direct, myocardial infarction, pulse rate, QRS duration, and ventricular rate, 2) the Early Growth Genetic Consortium (http://egg-consortium.org/) for childhood obesity[146] and birth weight[147], 3) the Meta-Analyses of Glucose and Insulin-related Traits Consortium (http://magicinvestigators.org/downloads/) for fasting glucose[148], 4) a previous study[149] for type 1 diabetes, 5) the DIAGRAM Consortium (https://diagram-consortium.org) for type 2 diabetes[150], and 5) the Edinburgh Data Share (https://datashare.ed.ac.uk/handle/10283/3209; https://datashare.ed.ac.uk/handle/10283/3599) for longevity GWAS for parental lifespan[151] and aging[152]. All of the data, except for type 1 diabetes, were provided in hg19 coordinates.

To convert coordinates from hg19 to hg38, we used the liftOver software downloaded from UCSC (https://genome-store.ucsc.edu/). Then, we sorted and indexed each GWAS summary statistics file using tabix[137].

We performed LD Score Regression (LDSC, v1.0.1)[120,121] using the HapMap3 variants that the developers found to be optimal for the analysis. First, we annotated each HapMap3 variant with a binary label (1/0) indicating whether the variant overlapped the ATAC-seq peak in PPC. Then, we estimated LD scores for each annotation with ldsc.py –l2 using 1000 Genomes Phase 3 reference files in hg38 available at broad-alkesgroup-public-requester-pays/LDSCORE/GRCh38/plink_files.tgz. Finally, we tested for heritability enrichment with ldsc.py –h2 using regression weights downloaded from broad-alkesgroup-public-requester-pays/LDSCORE/GRCh38/weights.tgz and baseline annotations (v.1.2) from broad-alkesgroup-public-requester-pays/LDSCORE/GRCh38/baseline_v1.2.tgz. Annotations were enriched for trait heritability if p-values (Enrichment_p) < 0.05.

4.5 Quantitative Trait Loci (QTL) Mapping

4.5.1 WGS Variant Selection

For all QTL analyses, we used single nucleotide polymorphisms (SNPs) that met following criteria across the 273 individuals (see **3.1 WGS**): 1) passed Illumina QC; 2) in Hardy-Weinberg equilibrium ($p > 0.000001$); 3) genotyped in at least 99% of the individuals; and 4) had MAF > 0.05. 5,536,303 variants remained.

4.5.2 Kinship Matrix

To account for genetic relatedness between samples, we performed LD pruning on the 5,536,303 variants using plink[144] 1.90b6.21 (-indep-pairwise 50 5 0.2). We then used the 323,697 LD-pruned variants to construct a kinship matrix for the 273 iPSCORE individuals using plink[144] 1.90b6.21 (–make-rel square).

4.5.3 Global Ancestry: Genotype Principal Component Analysis

We performed genotype principal component analysis (PCA) across all 273 individuals in the iPSCORE Collection. First, we intersected the 323,697 LD-pruned variants above with 1000 Genomes[133–135] single nucleotide polymorphisms (SNPs). Then, using plink[144] 1.90b6.21 (--pca-cluster-names AFR EUR AMR EAS SAS –pca), we performed PCA excluding 1000 Genome subjects without super-population information. We determined that the first five genotype PCs for QTL analysis were sufficient and captured the majority of the variability that were due to global ancestry. The ancestries reported for the 108 subjects in this study, were assigned in a previous study describing the iPSCORE Collection[35].

4.5.4 PEER Factor Calculation

To account for hidden technical and biological confounders that influence gene expression variability, we used Probabilistic Estimation of Expression Residuals (PEER)[153] to estimate a set of latent factors for each molecular data type (RNA-seq, ATAC-seq). We used the top 2,000 most variable genes/peaks to calculate a maximum number of PEER factors that is equivalent to ~25% of the samples, which in this case is 30, as recommended by the original developers[153]. As previously described[12,118], to determine the number of PEER factors to use for QTL discovery, we piloted QTL mapping on a random set of 1,000

genes or 4,000 peaks using varying numbers of PEER factors as covariates (listed in Table S2; Table S4) and selected the least number of PEERs that resulted in maximum eGene and caPeak discovery. For eQTLs, we used 22 PEER factors as covariates. For caQTLs, we used 20 PEER factors as covariates. We found that the variance captured by PEER factors was correlated with known biological and technical factors recorded for each sample. In particular, we observed that the top PEER factors across all the molecular data types were highly correlated with sequencing quality, differentiation efficiency and sex.

4.5.5 QTL Covariates

For all QTL analyses, we included the following as general covariates: sex, iPSC passage number, the first five genotype PCs to control for global ancestry and PEER factors to account for hidden confounders of molecular phenotype variability (see **4.4.4 PEER Factor Calculation**).

4.5.6 QTL Mapping

QTL analysis was performed on each of the 8 iPSCORE molecular datasets independently using a linear mixed model (LMM) with the kinship matrix as a random effect to account for the genetic relatedness between samples. First, using *rank* and *qnorm* in R (v4.2.1), we inverse normal transformed the TMM gene expression/peak accessibility or acetylation values across the samples. Genes within 1 Mb and peaks within 100 kb of the MHC region[154] (chr6:28,510,120-33,480,577) were removed due to the complex LD structure in the interval. For the elements (i.e. genes and peaks) outside the MHC region, we used bcftools[155] query to obtain the genotypes for all the variants within 1 Mb for genes or 100 kb for ATAC-seq peaks and H3K27ac ChIP-seq peaks. Then, we applied the scan

function in limix (v3.0.4) (https://github.com/limix/limix) to run the following linear mixed model:

$$Y_i = \beta_j X_{ij} + \sum_{m=5}^{M} \gamma_m PC_{im} + \sum_{n=1}^{N} \gamma_n PEER_{in} + \sum_{p=1}^{P} \gamma_p C_{ip} + u_i + \in_{ij}$$

Where Y_i is the normalized expression value for sample i, β_j is the effect size (fixed effect) of SNP j, X_{ij} is the genotype of sample i at SNP j, M is the number of genotype principal components used (M = 5 for all QTL analyses), γ_m is the effect size of the mth genotype principal component, PC_{im} is the value of the mth genotype principal component for the individual associated with sample i, N is the number of PEER factors (See **4.4.4 PEER Factor Calculation),** γ_n is the effect size of the nth PEER factor, $PEER_{in}$ is the value of the nth PEER factor for sample i, P is the number of covariates used (P = 1 for all QTL analyses corresponding to iPSC passage number) , γ_p is the effect size of the pth covariate, C_{ip} is the value of the pth covariate for sample i, u_i is a vector of random effects for the individual associated with sample i defined from the kinship matrix, and $\in_{ij}$ is the error term for individual i at SNP j.

4.5.7 FDR correction

We used a two-step procedure described in Huang et al.[109], which first corrects at the gene level and then at the genome-wide level. First, we performed FDR correction on the p-values of all independent variants tested for each gene or isoform using eigenMT[156], which considers the LD structure of the variants. Then, we extracted the lead eQTL for each gene or isoform based on the most significant FDR-corrected p-value. If more than one variant had the same FDR-corrected p-value, we selected the one with the largest absolute effect size as the lead eQTL. For the second correction, we performed FDR-

correction on all lead variants using Benjamini-Hochberg (q-value). We considered only eQTLs with q-value < 0.05 as significant.

4.5.8 Conditional QTL mapping

To identify additional independent QTL associations for a gene or peak (i.e., conditional QTLs), we performed stepwise regression analysis in which we re-performed QTL analysis with the genotype of the lead eQTL as a covariate. We repeated the procedure to discover up to five conditional associations. For each iteration, we performed the two-step procedure described above and considered conditional eQTLs with q-values < 0.05 as significant.

4.6 Chromatin state enrichment

We tested for the enrichment of caQTLs and eQTLs in PPC chromatin states[14] using two-sided Fisher's Exact Test with the contingency table for the following two classifications: 1) if the QTL's lead variant overlaps the chromatin state, 2) if the QTL was significant (q-value < 0.05, see **4.5.7 FDR Correction**). We considered enrichments to be significant if the Benjamini-Hochberg-corrected p-value < 0.05.

4.7 Identification of Fetal-Unique QTLs

To assess whether the QTLs discovered were fetal-unique, we examine their overlap with publicly available adult QTLs. We downloaded all caQTLs and haQTL associations from QTLbase2[124] (http://www.mulinlab.org/qtlbase) that were generated from adult tissues. To reduce memory storage, we filtered for significant variants with $p < 1x10^{-5}$, resulting in 29,762 caVariants and 244,563 haVariants. For eQTLs, we downloaded all 276,116 lead eQTLs from the GTEx database (version 8) for 49 adult tissues

(https://www.gtexportal.org/home/downloads/adult-gtex#qtl). We considered an iPSCORE QTL as fetal-unique if their lead fine-mapped variant was not in LD with any adult QTL ($r^2 < 0.2$ within 500 kb) regardless of molecular phenotype. If the QTL was in a module, we required that all QTLs in the module were not in LD with any adult QTL ($r^2 < 0.2$ within 500 kb) in order to be considered fetal-unique. LD between variants was calculated with plink[144] --tag-r2 0.2 --tag-kb 500 using the 1000 Genomes Panel 3[133–135] (Europeans) as reference. For variants not in the reference panel and therefore LD could not be calculated, if the QTL was within 500 kb of an adult QTL it was not considered to be fetal-unique.

4.8 Identification of QTL Modules

4.8.1 Intra-Tissue QTL Colocalization

To identify QTLs within the same tissue across the three molecular data types that shared causal variants, we performed pairwise Bayesian colocalization using the *coloc.abf* function in *coloc* v5.2.2[41] between the six pairwise combinations of QTLs (eQTL-eQTL, caQTL-caQTL, haQTL-haQTL, caQTL-eQTL, haQTL-eQTL, and caQTL-haQTL; PPCs do not have H3K27ac ChIP-seq, therefore combinations with haQTLs were not analyzed). First, for each tissue, we created a bed file containing the coordinates of windows tested for each QTL (i.e. for each eGene, we included 1 Mb upstream and downstream; for each caPeak, we included 100 kb upstream and downstream; and for each haPeak coordinates, we included 100 kb upstream and downstream). Then, we identified overlapping regions between QTLs, removing overlaps of the same element (i.e., QTL A overlaps with QTL A). We then performed colocalization (*coloc.abf*) on each QTL pair, considering only pairs

that shared at least 50 biallelic SNPs. Two QTL signals were considered shared if the PP.H4 ≥ 0.8. We observed 2,609 instances where multiple QTLs for the same qElement colocalized (PP.H4 ≥ 0.8) with the same QTL of a different qElement. This likely is due to the fact that conditional QTLs are present in primary QTL analyses, therefore affecting the distribution of P-values. In these cases, we selected the colocalization with the highest PP.H4 for each QTL-element pair. For example, if eGene1_Primary and caPeak1_Primary PP.H4 = 95% and eGene1_Primary and caPeak1_Condition1 PP.H4 = 80%, we would have selected the eGene1_Primary and caPeak1_Primary pair.

4.8.2 Network Analysis

For each tissue, we loaded each pair of colocalized QTLs as edges into an *igraph*[112] (v1.3.2) (https://igraph.org) network. To identify modules, we clustered the colocalized QTL networks using the *cluster_louvain* function, then divided each module into a subgraph, using the induced subgraph function, and calculated each subgraph's modularity, using the *modularity* function. If a subgraph exhibited high modularity (> 0.3), the subgraph was recursively clustered using the *cluster_louvain* function and divided into multiple modules. In total, there were 1,883 QTL modules composed of two or more QTLs, and 18,495 singleton QTLs that did not colocalize with another QTL.

4.9 GWAS associations with QTLs

4.9.1 GWAS-QTL Colocalization

For each QTL, we performed pairwise colocalization with GWAS variants (see **4.3 LD Score Regression** for list of GWAS summary statistics) using effect size and variance as input into the *coloc.abf* function in *coloc*[41] (v5.2.2). For a QTL to colocalize with GWAS

variants, we required that the all of the following criteria were satisfied: 1) had at least 50 overlapping variants; 2) PP.H4 ≥ 80%; 3) the lead putative causal variant is genome-wide significant for GWAS association (p-value ≤ 5×10^{-8}); 4) the lead putative causal variant is significant for QTL association (p-value ≤ 5×10^{-5}); and 5) the lead putative causal variant had causal PP ≥ 1%. For QTL modules, we required that at least one of the QTLs in the module to colocalize with GWAS with PP.H4 ≥ 80%.

4.9.2 GWAS-QTL candidate causal variants

For each GWAS-QTL colocalization, *coloc*[41] outputs the causal PP for each variant that was tested during colocalization. We assigned a lead candidate causal variant for each pair by taking the variant with the highest causal PP. For modules, we assigned the variant with the highest causal PP among the QTLs that colocalized with the GWAS signal. As an example, for a module with four QTLs, two QTLs can colocalize with the GWAS signal, each having their own lead candidate causal variant. To assign a single candidate causal variant for the module, we assigned the one from the QTL that had the maximum causal PP.

4.9.3 Fraction of GWAS Signals colocalized with QTLs

To determine the fraction of GWAS loci explained by QTLs, we calculated the number of independent genome-wide significant signals for each of the 17 GWAS studies. Specifically, we first filtered for variants that were above the genome-wide significant threshold of $p < 5 \times 10^{-8}$. Then, we applied LD pruning using plink[144] with the following parameters: *--indep-pairwise 500 50 0.01*, where 500 is the variant count window, 50 is the step count, and 0.01 is the LD threshold. This command outputs a list of 2,597

independent variants (i.e., not in LD) that each represents an independent genome-wide significant GWAS signal. After identifying the 2,597 independent GWAS signals, we then sought to identify the subset that had colocalized to a QTL (either a singleton or module) (PP.H4 $\geq$ 80%). Using 1000 Genomes Phase 3 (Europeans only) as reference, we calculated LD between the lead candidate causal variant from the GWAS-QTL colocalization (see **4.8.2 GWAS-QTL candidate causal variants)** and the LD-pruned GWAS variants using *plink –tag-kb 500 –tag-kb 0.7*[144]. If the variant was in high LD ($r^2 \geq$ 0.7 within 500 kb) with a pruned GWAS variant, then we assigned the QTL module or singleton to that GWAS signal. For variants absent from the reference panel, and therefore LD could not be calculated, we assigned the QTL module or singleton to the nearest GWAS signal. Using QTLs from all three tissues in the original manuscript, we observed that 110 of the 2,597 (1.3%) GWAS signals were in LD with multiple independent QTL modules or singletons (range 2-9 QTL modules/singletons per signal). For 68 of these signals, the QTL modules or singletons resulted in the same lead candidate causal variant (i.e., GWAS colocalization with QTL module/singleton A resulted in the same lead candidate causal variant as colocalization with QTL module/singleton B), suggesting that independent QTL modules or singletons from different tissues colocalized with the same GWAS signal. For the remaining 65 signals, fine-mapping identified different lead candidate causal variants, suggesting that there are multiple causal variants underlying the GWAS signal, consistent with previous observations[157,158].

4.9.4 Enrichment of QTL Modules with GWAS Variants

We annotated each GWAS-colocalized QTL module and singleton based on the molecular phenotypes the QTLs were associated with. For example, if a QTL module was composed of both caQTL(s) and eQTL(s), we annotated the module as "caQTL-eQTL". For singleton QTLs, we included them as either eQTL, caQTL, or haQTL. Across all GWAS-colocalized QTL modules and singletons, we observed the following molecular phenotype combinations: 1) caQTL-haQTL-eQTL, 2) haQTL-eQTL, 3) caQTL-eQTL, 4) caQTL-haQTL, 5) eQTL, 6) caQTL, 7) haQTL. Enrichment of each of these combinations for GWAS variants was calculated using a Fisher's Exact test, where the contingency table consisted of two classifications: 1) if the QTL module or singleton was annotated with the classification, and 2) if the QTL module or singleton colocalized with at least one GWAS trait using the criteria described in **4.8.1 GWAS-QTL colocalization**. A molecular phenotype combination was considered enriched for GWAS if the corrected p-value with Bonferroni's Method < 0.05.

2.12 Code Availability

Scripts for processing FASTQ-files and performing downstream analyses are publicly available at https://github.com/frazer-lab/iPSCORE_QTL_Resource.

2.13 Data Availability

FASTQ sequencing data for 109 PPC ATAC-seq have been deposited into GEO GSE197140. WGS data for iPSCORE subjects were downloaded as a VCF file from phs001325.v3. GWAS summary statistics were obtained from the Pan UK BioBank resource (https://pan.ukbb.broadinstitute.org/), the MAGIC (Meta-Analyses of Glucose and Insulin-related traits) Consortium (https://magicinvestigators.org/downloads/;

https://doi.org/10.1038/s41588-021-00852-9), the DIAMANTE Consortium (https://diagram-consortium.org/downloads.html; http://doi.org/10.1038/s41588-018-0241-6), and a previously published studies[149,151,152]. Full QTL summary statistics, phenotype matrices, element coordinates, and TFBS predictions have been deposited in Figshare.

2.14 Author information

iPSCORE Consortium, University of California, San Diego, La Jolla, CA, 92093, US

Angelo D. Arias, Timothy D. Arthur, Paola Benaglio, Victor Borja, Megan Cook, Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Christopher DeBoever, Margaret K.R. Donovan, KathyJean Farnam, Kelly A. Frazer, Kyohei Fujita, Melvin Garcia, Olivier Harismendy, David Jakubosky, Kristen Jepsen, Isaac Joshua, He Li, Hiroko Matsui, Naoki Nariai, Jennifer P. Nguyen, Daniel T. O'Connor, Jonathan Okubo, Fengwen Rao, Joaquin Reyna, Lana Ribeiro Aguiar, Bianca Salgado, Nayara Silva, Erin N. Smith, Josh Sohmer, Shawn Yost, William W. Young Greenwald

Contributions

TDA, JPN, and KAF conceived the study. JPN, TDA, BH, and JJ performed the computational analyses. KAF, MD, GM, JCIB, and iPSCORE consortium members oversaw the study. ADC, NS, and members of the iPSCORE Consortium performed the differentiations and generated molecular data. TDA, JPN, and KAF prepared the manuscript.

Acknowledgements

This work was supported by the National Library Training Grant T15LM011271, the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) F31DK131867, U01DK105541, DP3DK112155 and P30DK063491, the National Heart, Lung and Blood Institute (NHLBI) F31HL158198 and U01HL107442, and the National Human Genome Research Institute (NHGRI) RM1HG011558 and R41HG008118. Additional support was also received from a California Institute for Regenerative Medicine grant GC1R-06673-B, NSF-CMMI division award 1728497. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant S10OD026929.

Chapter 2, in part, is an adapted version of a manuscript that is currently in preparation for publication with authors Timothy D. Arthur, Jennifer P. Nguyen, Agnieszka D'Antonio-Chronowska, Jeffrey Jaureguy, Nayara Silva, Benjamin Henson, iPSCORE Consortium, Athanasia D. Panopoulos, Juan Carlos Izpisua Belmonte, Matteo D'Antonio, Graham McVicker, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

References

1. Broad Genomics Platform, DiscovEHR Collaboration, CHARGE, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*. 2019;570(7759):71-76. doi:10.1038/s41586-019-1231-2

2. Chen J, Spracklen CN, Marenne G, et al. The trans-ancestral genomic architecture of glycemic traits. *Nat Genet*. 2021;53(6):840-860. doi:10.1038/s41588-021-00852-9

3. Chiou J, Geusz RJ, Okino ML, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature*. 2021;594(7863):398-402. doi:10.1038/s41586-021-03552-w

4. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet*. 2018;50(11):1505-1513. doi:10.1038/s41588-018-0241-6

5. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473(7345):43-49. doi:10.1038/nature09906

6. Maurano MT, Humbert R, Rynes E, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*. 2012;337(6099):1190-1195. doi:10.1126/science.1222794

7. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330. doi:10.1038/nature14248

8. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-213. doi:10.1038/nature24277

9. Kim-Hellmuth S, Aguet F, Oliva M, et al. Cell type–specific genetic regulation of gene expression across human tissues. *Science*. 2020;369(6509):eaaz8528. doi:10.1126/science.aaz8528

10. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776

11. Viñuela A, Varshney A, van de Bunt M, et al. Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat Commun*. 2020;11(1):4912. doi:10.1038/s41467-020-18581-8

12. D'Antonio M, Nguyen JP, Arthur TD, et al. Fine mapping spatiotemporal mechanisms of genetic variants underlying cardiac traits and disease. *Nat Commun*. 2023;14(1):1132. doi:10.1038/s41467-023-36638-2

13. Strober BJ, Elorbany R, Rhodes K, et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*. 2019;364(6447):1287-1290. doi:10.1126/science.aaw0040

14. Geusz RJ, Wang A, Chiou J, et al. Pancreatic progenitor epigenome maps prioritize type 2 diabetes risk genes with roles in development. *eLife*. 2021;10:e59067. doi:10.7554/eLife.59067

15. Dabelea D, Pettitt DJ. Intrauterine Diabetic Environment Confers Risks for Type 2 Diabetes Mellitus and Obesity in the Offspring, in Addition to Genetic Susceptibility. *Journal of Pediatric Endocrinology and Metabolism*. 2001;14(8). doi:10.1515/jpem-2001-0803

16. Bhattacharya A, Freedman AN, Avula V, et al. Placental genomics mediates genetic associations with complex health traits and disease. *Nat Commun*. 2022;13(1):706. doi:10.1038/s41467-022-28365-x

17. Petersen MBK, Gonçalves CAC, Kim YH, Grapin-Botton A. Recapitulating and Deciphering Human Pancreas Development From Human Pluripotent Stem Cells in a Dish. In: *Current Topics in Developmental Biology*. Vol 129. Elsevier; 2018:143-190. doi:10.1016/bs.ctdb.2018.02.009

18. Colclough K, Bellanne-Chantelot C, Saint-Martin C, Flanagan SE, Ellard S. Mutations in the Genes Encoding the Transcription Factors Hepatocyte Nuclear Factor 1 Alpha and 4 Alpha in Maturity-Onset Diabetes of the Young and Hyperinsulinemic Hypoglycemia. *Human Mutation*. 2013;34(5):669-685. doi:10.1002/humu.22279

19. Hansen L, Urioste S, Petersen HV, et al. Missense Mutations in the Human Insulin Promoter Factor-1 Gene and Their Relation to Maturity-Onset Diabetes of the Young and Late-Onset Type 2 Diabetes Mellitus in Caucasians*. *The Journal of Clinical Endocrinology & Metabolism*. 2000;85(3):1323-1326. doi:10.1210/jcem.85.3.6421

20. Sanyoura M, Philipson LH, Naylor R. Monogenic Diabetes in Children and Adolescents: Recognition and Treatment Options. *Curr Diab Rep*. 2018;18(8):58. doi:10.1007/s11892-018-1024-2

21. Ameri J, Borup R, Prawiro C, et al. Efficient Generation of Glucose-Responsive Beta Cells from Isolated GP2 + Human Pancreatic Progenitors. *Cell Reports*. 2017;19(1):36-49. doi:10.1016/j.celrep.2017.03.032

22. Gonçalves CA, Larsen M, Jung S, et al. A 3D system to model human pancreas development and its reference single-cell transcriptome atlas identify signaling pathways required for progenitor expansion. *Nat Commun*. 2021;12(1):3144. doi:10.1038/s41467-021-23295-6

23. Nostro MC, Sarangi F, Yang C, et al. Efficient Generation of NKX6-1+ Pancreatic Progenitors from Multiple Human Pluripotent Stem Cell Lines. *Stem Cell Reports*. 2015;4(4):591-604. doi:10.1016/j.stemcr.2015.02.017

24. Pagliuca FW, Millman JR, Gürtler M, et al. Generation of Functional Human Pancreatic β Cells In Vitro. *Cell*. 2014;159(2):428-439. doi:10.1016/j.cell.2014.09.040

25. Rezania A, Bruin JE, Arora P, et al. Reversal of diabetes with insulin-producing cells derived in vitro from human pluripotent stem cells. *Nat Biotechnol*. 2014;32(11):1121-1133. doi:10.1038/nbt.3033

26. Russ HA, Parent AV, Ringler JJ, et al. Controlled induction of human pancreatic progenitors produces functional beta-like cells *in vitro*. *EMBO J*. 2015;34(13):1759-1772. doi:10.15252/embj.201591058

27. Veres A, Faust AL, Bushnell HL, et al. Charting cellular identity during human in vitro β-cell differentiation. *Nature*. 2019;569(7756):368-373. doi:10.1038/s41586-019-1168-5

28. Sean DLO, Liu Z, Sun H, et al. *Single-Cell Multi-Omic Roadmap of Human Fetal Pancreatic Development*. Developmental Biology; 2022. doi:10.1101/2022.02.17.480942

29. Seymour PA. Sox9: A Master Regulator of the Pancreatic Program. *Rev Diabet Stud*. 2014;11(1):51-83. doi:10.1900/RDS.2014.11.51

30. Seymour PA, Freude KK, Tran MN, et al. SOX9 is required for maintenance of the pancreatic progenitor cell pool. *Proc Natl Acad Sci USA*. 2007;104(6):1865-1870. doi:10.1073/pnas.0609217104

31. Aigha II, Abdelalim EM. NKX6.1 transcription factor: a crucial regulator of pancreatic β cell development, identity, and proliferation. *Stem Cell Res Ther*. 2020;11(1):459. doi:10.1186/s13287-020-01977-0

32. Oliver-Krasinski JM, Stoffers DA. On the origin of the β cell. *Genes Dev*. 2008;22(15):1998-2021. doi:10.1101/gad.1670808

33. Van Hoof D, D'Amour KA, German MS. Derivation of insulin-producing cells from human embryonic stem cells. *Stem Cell Research*. 2009;3(2-3):73-87. doi:10.1016/j.scr.2009.08.003

34. Ramond C, Beydag-Tasöz BS, Azad A, et al. Understanding human fetal pancreas development using subpopulation sorting, RNA sequencing and single-cell profiling. *Development*. Published online January 1, 2018:dev.165480. doi:10.1242/dev.165480

35. Panopoulos AD, D'Antonio M, Benaglio P, et al. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports*. 2017;8(4):1086-1100. doi:10.1016/j.stemcr.2017.03.012

36. Jin W, Jiang W. Stepwise differentiation of functional pancreatic β cells from human pluripotent stem cells. *Cell Regen*. 2022;11(1):24. doi:10.1186/s13619-022-00125-8

37. DeBoever C, Li H, Jakubosky D, et al. Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell*. 2017;20(4):533-546.e7. doi:10.1016/j.stem.2017.03.009

38. Fadista J, Vikman P, Laakso EO, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci USA*. 2014;111(38):13924-13929. doi:10.1073/pnas.1402665111

39. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653

40. Jansen R, Hottenga JJ, Nivard MG, et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum Mol Genet*. 2017;26(8):1444-1451. doi:10.1093/hmg/ddx043

41. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Williams SM, ed. *PLoS Genet*. 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383

42. Yan J, Qiu Y, Ribeiro dos Santos AM, et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature*. 2021;591(7848):147-151. doi:10.1038/s41586-021-03211-0

43. Garrido-Martín D, Borsari B, Calvo M, Reverter F, Guigó R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun*. 2021;12(1):727. doi:10.1038/s41467-020-20578-2

44. van de Bunt M, Manning Fox JE, Dai X, et al. Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. Stranger BE, ed. *PLoS Genet*. 2015;11(12):e1005694. doi:10.1371/journal.pgen.1005694

45. Chen JH, Zhao Y, Khan RAW, et al. SNX29, a new susceptibility gene shared with major mental disorders in Han Chinese population. *The World Journal of Biological Psychiatry*. 2021;22(7):526-534. doi:10.1080/15622975.2020.1845793

46. Anderson D, Cordell HJ, Fakiola M, et al. First genome-wide association study in an Australian aboriginal population provides insights into genetic risk factors for body mass index and type 2 diabetes. *PLoS One*. 2015;10(3):e0119333. doi:10.1371/journal.pone.0119333

47. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*. 2019;47(D1):D766-D773. doi:10.1093/nar/gky955

48. Thurner M, van de Bunt M, Torres JM, et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. *eLife*. 2018;7:e31977. doi:10.7554/eLife.31977

49. Bolt CC, Lopez-Delisle L, Hintermann A, et al. Context-dependent enhancer function revealed by targeted inter-TAD relocation. *Nat Commun*. 2022;13(1):3488. doi:10.1038/s41467-022-31241-3

50. Pan-UKB team. Published online 2020. https://pan.ukbb.broadinstitute.org

51. Dimas AS, Lagou V, Barker A, et al. Impact of Type 2 Diabetes Susceptibility Variants on Quantitative Glycemic Traits Reveals Mechanistic Heterogeneity. *Diabetes*. 2014;63(6):2158-2171. doi:10.2337/db13-0949

52. Grarup N, Sandholt CH, Hansen T, Pedersen O. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia*. 2014;57(8):1528-1541. doi:10.1007/s00125-014-3270-4

53. D'Antonio M, Nguyen JP, Arthur TD, et al. In heart failure reactivation of RNA-binding proteins is associated with the expression of 1,523 fetal-specific isoforms. Zhang Z, ed. *PLoS Comput Biol*. 2022;18(2):e1009918. doi:10.1371/journal.pcbi.1009918

54. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*. 2017;18(7):437-451. doi:10.1038/nrm.2017.27

55. Mazin PV, Khaitovich P, Cardoso-Moreira M, Kaessmann H. Alternative splicing during mammalian organ development. *Nat Genet*. 2021;53(6):925-934. doi:10.1038/s41588-021-00851-w

56. Brun T, Jiménez-Sánchez C, Madsen JGS, et al. AMPK Profiling in Rodent and Human Pancreatic Beta-Cells under Nutrient-Rich Metabolic Stress. *IJMS*. 2020;21(11):3982. doi:10.3390/ijms21113982

57. Minokoshi Y, Alquier T, Furukawa N, et al. AMP-kinase regulates food intake by responding to hormonal and nutrient signals in the hypothalamus. *Nature*. 2004;428(6982):569-574. doi:10.1038/nature02440

58. Shaw RJ, Lamia KA, Vasquez D, et al. The Kinase LKB1 Mediates Glucose Homeostasis in Liver and Therapeutic Effects of Metformin. *Science*. 2005;310(5754):1642-1646. doi:10.1126/science.1120781

59. Yamauchi T, Kamon J, Minokoshi Y, et al. Adiponectin stimulates glucose utilization and fatty-acid oxidation by activating AMP-activated protein kinase. *Nat Med*. 2002;8(11):1288-1295. doi:10.1038/nm788

60. Wu Y, Viana M, Thirumangalathu S, Loeken MR. AMP-activated protein kinase mediates effects of oxidative stress on embryo gene expression in a mouse model of diabetic embryopathy. *Diabetologia*. 2012;55(1):245-254. doi:10.1007/s00125-011-2326-y

61. Grant SFA, Qu HQ, Bradfield JP, et al. Follow-Up Analysis of Genome-Wide Association Data Identifies Novel Loci for Type 1 Diabetes. *Diabetes*. 2009;58(1):290-295. doi:10.2337/db08-1022

62. the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Morris AP, Voight BF, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*. 2012;44(9):981-990. doi:10.1038/ng.2383

63. Kang HS, Kim YS, ZeRuth G, et al. Transcription Factor Glis3, a Novel Critical Player in the Regulation of Pancreatic β-Cell Development and Insulin Gene Expression. *Mol Cell Biol*. 2009;29(24):6366-6379. doi:10.1128/MCB.01259-09

64. Kang HS, Takeda Y, Jeon K, Jetten AM. The Spatiotemporal Pattern of Glis3 Expression Indicates a Regulatory Function in Bipotent and Endocrine Progenitors during Early Pancreatic Development and in Beta, PP and Ductal Cells. Blondeau B, ed. *PLoS ONE*. 2016;11(6):e0157138. doi:10.1371/journal.pone.0157138

65. Yang Y, Chang BH, Chan L. Sustained expression of the transcription factor GLIS3 is required for normal beta cell function in adults. *EMBO Mol Med*. 2013;5(1):92-104. doi:10.1002/emmm.201201398

66. Sams EI, Ng JK, Tate V, et al. From karyotypes to precision genomics in 9p deletion and duplication syndromes. *Human Genetics and Genomics Advances*. 2022;3(1):100081. doi:10.1016/j.xhgg.2021.100081

67. Aylward A, Chiou J, Okino ML, Kadakia N, Gaulton KJ. Shared genetic risk contributes to type 1 and type 2 diabetes etiology. *Human Molecular Genetics*. Published online November 7, 2018. doi:10.1093/hmg/ddy314

68. Cao R chang, Yang W jun, Xiao W, et al. St13 protects against disordered acinar cell arachidonic acid pathway in chronic pancreatitis. *J Transl Med*. 2022;20(1):218. doi:10.1186/s12967-022-03413-8

69. Graham SE, Clarke SL, Wu KHH, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*. 2021;600(7890):675-679. doi:10.1038/s41586-021-04064-3

70. Barrett JC, Clayton DG, Concannon P, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*. 2009;41(6):703-707. doi:10.1038/ng.381

71. Tong Z, Fan Y, Zhang W, et al. Pancreas-specific Pten deficiency causes partial resistance to diabetes and elevated hepatic AKT signaling. *Cell Res*. 2009;19(6):710-719. doi:10.1038/cr.2009.42

72. Wong H, Schotz MC. The lipase gene family. *Journal of Lipid Research*. 2002;43(7):993-999. doi:10.1194/jlr.R200007-JLR200

73. Chang W wei, Zhang L, Yao X ming, et al. Upregulation of long non-coding RNA MEG3 in type 2 diabetes mellitus complicated with vascular disease: a case–control study. *Mol Cell Biochem*. 2020;473(1-2):93-99. doi:10.1007/s11010-020-03810-x

74. Kameswaran V, Bramswig NC, McKenna LB, et al. Epigenetic Regulation of the DLK1-MEG3 MicroRNA Cluster in Human Type 2 Diabetic Islets. *Cell Metabolism*. 2014;19(1):135-145. doi:10.1016/j.cmet.2013.11.016

75. Kameswaran V, Golson ML, Ramos-Rodríguez M, et al. The Dysregulation of the *DLK1 - MEG3* Locus in Islets From Patients With Type 2 Diabetes Is Mimicked by Targeted Epimutation of Its Promoter With TALE-DNMT Constructs. *Diabetes*. 2018;67(9):1807-1815. doi:10.2337/db17-0682

76. Onengut-Gumuscu S, Chen WM, Burren O, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet*. 2015;47(4):381-386. doi:10.1038/ng.3245

77. Chen J, Liu Y, Min J, et al. Alternative splicing of lncRNAs in human diseases. *Am J Cancer Res*. 2021;11(3):624-639.

78. Wheeler E, Leong A, Liu CT, et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med*. 2017;14(9):e1002383. doi:10.1371/journal.pmed.1002383

79. Parnaud G, Lavallard V, Bedat B, et al. Cadherin engagement improves insulin secretion of single human β-cells. *Diabetes*. 2015;64(3):887-896. doi:10.2337/db14-0257

80. Pulit SL, Stoneman C, Morris AP, et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet*. 2019;28(1):166-174. doi:10.1093/hmg/ddy327

81. Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet*. 2021;53(10):1415-1424. doi:10.1038/s41588-021-00931-x

82. Vujkovic M, Keaton JM, Lynch JA, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet*. 2020;52(7):680-691. doi:10.1038/s41588-020-0637-y

83. Zhu Z, Guo Y, Shi H, et al. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J Allergy Clin Immunol*. 2020;145(2):537-549. doi:10.1016/j.jaci.2019.09.035

84. Calogero S, Grassi F, Aguzzi A, et al. The lack of chromosomal protein Hmg1 does not disrupt cell growth but causes lethal hypoglycaemia in newborn mice. *Nat Genet*. 1999;22(3):276-280. doi:10.1038/10338

85. Wang Y, Zhong J, Zhang X, et al. The Role of HMGB1 in the Pathogenesis of Type 2 Diabetes. *J Diabetes Res*. 2016;2016:2543268. doi:10.1155/2016/2543268

86. Chen C, Yu W, Tober J, et al. Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. *Cell Reports*. 2019;29(12):4200-4211.e7. doi:10.1016/j.celrep.2019.11.065

87. Gorkin DU, Barozzi I, Zhao Y, et al. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature*. 2020;583(7818):744-751. doi:10.1038/s41586-020-2093-3

88. Zhang K, Hocker JD, Miller M, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*. 2021;184(24):5985-6001.e19. doi:10.1016/j.cell.2021.10.024

89. Ong C, Corces VG. Enhancers: emerging roles in cell fate specification. *EMBO Rep*. 2012;13(5):423-430. doi:10.1038/embor.2012.52

90. Su CH, D D, Tarn WY. Alternative Splicing in Neurogenesis and Brain Development. *Front Mol Biosci*. 2018;5:12. doi:10.3389/fmolb.2018.00012

91. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet*. 2019;51(1):187-195. doi:10.1038/s41588-018-0268-8

92. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):1-4. doi:10.1093/gigascience/giab008

93. Israel MA, Yuan SH, Bardy C, et al. Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature*. 2012;482(7384):216-220. doi:10.1038/nature10821

94. D'Antonio-Chronowska A, D'Antonio M, Frazer K. In vitro Differentiation of Human iPSC-derived Cardiovascular Progenitor Cells (iPSC-CVPCs). *Bio-Protocol*. 2020;10(18):1-43. doi:10.21769/bioprotoc.3755

95. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*. 2012;22(9):1760-1774. doi:10.1101/gr.135350.111

96. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;36(1):89-94. doi:10.1038/nbt.4042

97. Veres A, Faust AL, Bushnell HL, et al. Charting cellular identity during human in vitro β-cell differentiation. *Nature*. 2019;569(7756):368-373. doi:10.1038/s41586-019-1168-5

98. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411-420. doi:10.1038/nbt.4096

99. D'Antonio-Chronowska A, Donovan MKR, Young Greenwald WW, et al. Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. *Stem Cell Reports*. 2019;13(5):924-938. doi:10.1016/j.stemcr.2019.09.011

100. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

101. Shaun Purcell CC. PLINK 1.9.0.

102. The 1000 Genomes Project Consortium, Corresponding authors, Auton A, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

103. Danecek P, McCarthy SA, HipSci Consortium, Durbin R. A Method for Checking Genomic Integrity in Cultured Cell Lines from SNP Genotyping Data. *PLoS One*. 2016;11(5):e0155014. doi:10.1371/journal.pone.0155014

104. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509

105. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. Published online 2011. doi:10.1201/b16589

106. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381-386. doi:10.1038/nbt.2859

107. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol*. 2019;37(7):773-782. doi:10.1038/s41587-019-0114-2

108. Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. *Nat Methods*. 2015;12(8):755-758. doi:10.1038/nmeth.3439

109. Huang QQ, Ritchie SC, Brozynska M, Inouye M. Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Research*. 2018;46(22):e133-e133. doi:10.1093/nar/gky780

110. Van Nostrand EL, Pratt GA, Shishkin AA, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*. 2016;13(6):508-514. doi:10.1038/nmeth.3810

111. Lee D, Gorkin DU, Baker M, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet*. 2015;47(8):955-961. doi:10.1038/ng.3331

112. Csardi, Gabor N Tamas. The igraph software package for complex network research. *InterJounal*. 2006;Complex Systems:1695.

113. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*. 2018;50(11):1505-1513. doi:10.1038/s41588-018-0241-6

114. Chen J, Spracklen CN, Marenne G, et al. The trans-ancestral genomic architecture of glycemic traits. *Nature Genetics*. 2021;53(6):840-860. doi:10.1038/s41588-021-00852-9

115. Bioconductor Package Maintainer. liftOver: Changing genomic coordinate systems with rtracklayer::liftOver. Published online 2022.

116. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. *Nat Methods*. 2021;18(11):1333-1341. doi:10.1038/s41592-021-01282-5

117. Kopinke D, Brailsford M, Shea JE, Leavitt R, Scaife CL, Murtaugh LC. Lineage tracing reveals the dynamic contribution of Hes1+ cells to the developing and adult pancreas. *Development*. 2011;138(3):431-441. doi:10.1242/dev.053843

118. Nguyen JP, Arthur TD, Fujita K, et al. eQTL mapping in fetal-like pancreatic progenitor cells reveals early developmental insights into diabetes risk. *Nat Commun*. 2023;14(1):6928. doi:10.1038/s41467-023-42560-4

119. Khetan S, Kursawe R, Youn A, et al. Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes*. 2018;67(11):2466-2477. doi:10.2337/db18-0393

120. Finucane HK, Bulik-Sullivan B, ReproGen Consortium, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47(11):1228-1235. doi:10.1038/ng.3404

121. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Bulik-Sullivan BK, Loh PR, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295. doi:10.1038/ng.3211

122. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*. Published online November 8, 2019:gkz1001. doi:10.1093/nar/gkz1001

123. Dobbyn A, Huckins LM, Boocock J, et al. Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. *Am J Hum Genet*. 2018;102(6):1169-1184. doi:10.1016/j.ajhg.2018.04.011

124. Huang D, Feng X, Yang H, et al. QTLbase2: an enhanced catalog of human quantitative trait loci on extensive molecular phenotypes. *Nucleic Acids Res*. 2023;51(D1):D1122-D1128. doi:10.1093/nar/gkac1020

125. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. *Genome Biol*. 2021;22(1):108. doi:10.1186/s13059-021-02322-1

126. Calkins K, Devaskar SU. Fetal origins of adult disease. *Curr Probl Pediatr Adolesc Health Care*. 2011;41(6):158-176. doi:10.1016/j.cppeds.2011.01.001

127. Huang Z, Ruan HB, Xian L, et al. The stem cell factor/Kit signalling pathway regulates mitochondrial function and energy expenditure. *Nat Commun*. 2014;5:4282. doi:10.1038/ncomms5282

128. Krishnamurthy M, Ayazi F, Li J, et al. c-Kit in early onset of diabetes: a morphological and functional analysis of pancreatic beta-cells in c-KitW-v mutant mice. *Endocrinology*. 2007;148(11):5520-5530. doi:10.1210/en.2007-0387

129. Feng ZC, Riopel M, Popell A, Wang R. A survival Kit for pancreatic beta cells: stem cell factor and c-Kit receptor tyrosine kinase. *Diabetologia*. 2015;58(4):654-665. doi:10.1007/s00125-015-3504-0

130. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet*. 2023;55(11):1866-1875. doi:10.1038/s41588-023-01529-1

131. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29.9. doi:10.1002/0471142727.mb2129s109

132. Arthur TD, Nguyen JP, D'Antonio-Chronowska A, et al. Complex regulatory networks influence pluripotent cell state transitions in human iPSCs. *Nat Commun*. 2024;15(1):1664. doi:10.1038/s41467-024-45506-6

133. Byrska-Bishop M, Evani US, Zhao X, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 2022;185(18):3426-3440.e19. doi:10.1016/j.cell.2022.08.004

134. Zheng-Bradley X, Streeter I, Fairley S, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience*. 2017;6(7):1-8. doi:10.1093/gigascience/gix038

135. Lowy-Gallego E, Fairley S, Zheng-Bradley X, et al. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res*. 2019;4:50. doi:10.12688/wellcomeopenres.15126.2

136. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. 2014;30(7):1006-1007. doi:10.1093/bioinformatics/btt730

137. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008

138. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. doi:10.1186/1471-2105-12-323

139. Frankish A, Diekhans M, Jungreis I, et al. GENCODE 2021. *Nucleic Acids Res*. 2021;49(D1):D916-D923. doi:10.1093/nar/gkaa1087

140. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

141. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j*. 2011;17(1):10. doi:10.14806/ej.17.1.200

142. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137

143. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

144. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

145. Bentsen M, Goymann P, Schultheis H, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun*. 2020;11(1):4267. doi:10.1038/s41467-020-18035-1

146. Bradfield JP, Vogelezang S, Felix JF, et al. A trans-ancestral meta-analysis of genome-wide association studies reveals loci associated with childhood obesity. *Hum Mol Genet*. 2019;28(19):3327-3338. doi:10.1093/hmg/ddz161

147. Horikoshi M, Beaumont RN, Day FR, et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature*. 2016;538(7624):248-252. doi:10.1038/nature19806

148. Chen J, Spracklen CN, Marenne G, et al. The trans-ancestral genomic architecture of glycemic traits. *Nat Genet*. 2021;53(6):840-860. doi:10.1038/s41588-021-00852-9

149. Chiou J, Geusz RJ, Okino ML, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature*. 2021;594(7863):398-402. doi:10.1038/s41586-021-03552-w

150. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet*. 2018;50(11):1505-1513. doi:10.1038/s41588-018-0241-6

151. Timmers PR, Mounier N, Lall K, et al. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife*. 2019;8:e39856. doi:10.7554/eLife.39856

152. Timmers PRHJ, Wilson JF, Joshi PK, Deelen J. Multivariate genomic scan implicates novel loci and haem metabolism in human ageing. *Nat Commun*. 2020;11(1):3570. doi:10.1038/s41467-020-17312-3

153. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*. 2012;7(3):500-507. doi:10.1038/nprot.2011.457

154. Dilthey AT. State-of-the-art genome inference in the human MHC. *Int J Biochem Cell Biol*. 2021;131:105882. doi:10.1016/j.biocel.2020.105882

155. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008

156. Davis JR, Fresard L, Knowles DA, et al. An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am J Hum Genet*. 2016;98(1):216-224. doi:10.1016/j.ajhg.2015.11.021

157. Hormozdiari F, Zhu A, Kichaev G, et al. Widespread Allelic Heterogeneity in Complex Traits. *Am J Hum Genet*. 2017;100(5):789-802. doi:10.1016/j.ajhg.2017.04.005

158. Abell NS, DeGorter MK, Gloudemans MJ, et al. Multiple causal variants underlie genetic associations in humans. *Science*. 2022;375(6586):1247-1254. doi:10.1126/science.abj5117