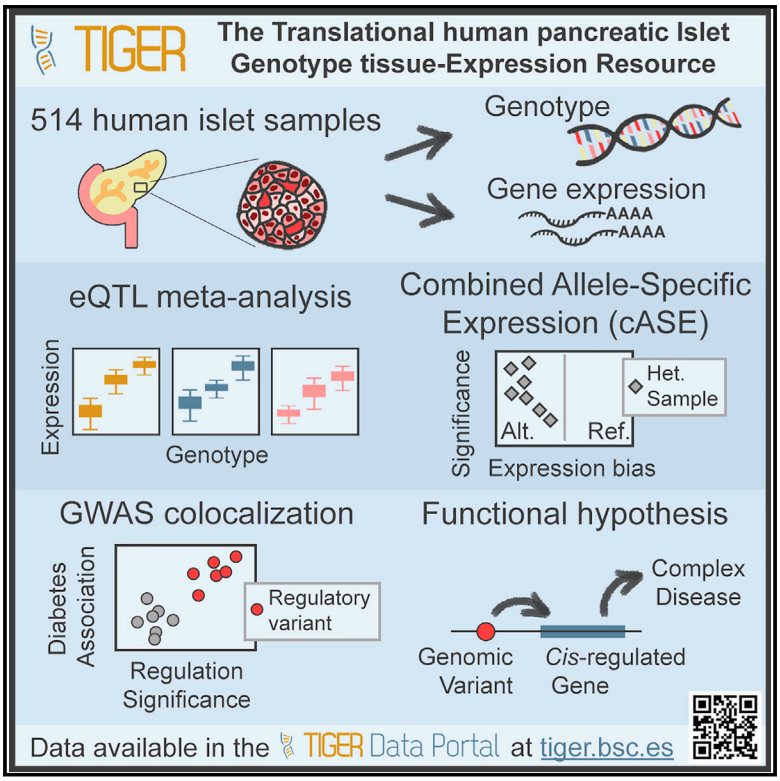
Cell Reports

Resource

TIGER: The gene expression regulatory variation landscape of human pancreatic islets

Graphical abstract



Highlights

- Human pancreatic islets are key drivers of diabetes and related pathophysiology
- TIGER integrates omics and expression regulatory variation in 514 human islet samples
- TIGER expression regulatory variation allows the identification of diabetes effector genes
- The integrated human islet data in TIGER are publicly available through http://tiger.bsc.es

Authors

Lorena Alonso, Anthony Piron, Ignasi Morán, ..., Josep M. Mercader, Miriam Cnop, David Torrents

Correspondence

mercader@broadinstitute.org (J.M.M.), mcnop@ulb.ac.be (M.C.), david.torrents@bsc.es (D.T.)

In brief

Understanding human islet regulatory genetic variation is essential to better understand the pathophysiology of diabetes and related diseases. Here, Alonso, Piron, Moran et al. present a comprehensive characterization of expression regulatory variation in >500 human islet samples and facilitate its access to the scientific community through the TIGER web portal.

Alonso et al., 2021, Cell Reports *37*, 109807
October 12, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.celrep.2021.109807

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

TIGER: The gene expression regulatory variation landscape of human pancreatic islets

Lorena Alonso,[1,25] Anthony Piron,[2,3,25] Ignasi Morán,[1,25] Marta Guindo-Martínez,[1] Sílvia Bonàs-Guarch,[4,5] Goutham Atla,[4,5] Irene Miguel-Escalada,[4,5] Romina Royo,[1] Montserrat Puiggròs,[1] Xavier Garcia-Hurtado,[4,5] Mara Suleiman,[6] Lorella Marselli,[6] Jonathan L.S. Esguerra,[7] Jean-Valéry Turatsinze,[2] Jason M. Torres,[8,9] Vibe Nylander,[10] Ji Chen,[11] Lena Eliasson,[7] Matthieu Defrance,[2] Ramon Amela,[1] MAGIC,[24] Hindrik Mulder,[12] Anna L. Gloyn,[9,10,13,14,15] Leif Groop,[7,12,16] Piero Marchetti,[6] Decio L. Eizirik,[2,17] Jorge Ferrer,[4,5,18] Josep M. Mercader,[1,19,20,21,26,*] Miriam Cnop,[2,22,26,27,*] and David Torrents[1,23,26,*]

[1]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
[2]ULB Center for Diabetes Research, Université Libre de Bruxelles, Brussels 1070, Belgium
[3]Interuniversity Institute of Bioinformatics in Brussels (IB2), Brussels 1050, Belgium
[4]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain
[5]Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM) Barcelona 08013, Spain
[6]Department of Clinical and Experimental Medicine and AOUP Cisanello University Hospital, University of Pisa, Pisa 56126, Italy
[7]Unit of Islet Cell Exocytosis, Lund University Diabetes Centre, Malmö 214 28, Sweden
[8]Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
[9]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK
[10]Oxford Centre for Diabetes, Endocrinology, and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 7LE, UK
[11]Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, Exeter EX4 4PY, UK
[12]Unit of Molecular Metabolism, Lund University Diabetes Centre, Malmö 214 28, Sweden
[13]Division of Endocrinology, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94304, USA
[14]NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford OX3 7DQ, UK
[15]Stanford Diabetes Research Centre, Stanford University, Stanford, CA 94305, USA
[16]Finnish Institute of Molecular Medicine Finland (FIMM), Helsinki University, Helsinki 00014, Finland
[17]WELBIO, Université Libre de Bruxelles, Brussels 1050, Belgium
[18]Section of Epigenomics and Disease, Department of Medicine, Imperial College London, London SW7 2AZ, UK
[19]Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[20]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[21]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[22]Division of Endocrinology, Erasmus Hospital, Université Libre de Bruxelles, Brussels 1070, Belgium
[23]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain
[24]Members of the MAGIC consortium are provided in Appendix S1
[25]These authors contributed equally
[26]Senior author
[27]Lead contact
*Correspondence: mercader@broadinstitute.org (J.M.M.), mcnop@ulb.ac.be (M.C.), david.torrents@bsc.es (D.T.)
https://doi.org/10.1016/j.celrep.2021.109807

SUMMARY

Genome-wide association studies (GWASs) identified hundreds of signals associated with type 2 diabetes (T2D). To gain insight into their underlying molecular mechanisms, we have created the translational human pancreatic islet genotype tissue-expression resource (TIGER), aggregating >500 human islet genomic datasets from five cohorts in the Horizon 2020 consortium T2DSystems. We impute genotypes using four reference panels and meta-analyze cohorts to improve the coverage of expression quantitative trait loci (eQTL) and develop a method to combine allele-specific expression across samples (cASE). We identify >1 million islet eQTLs, 53 of which colocalize with T2D signals. Among them, a low-frequency allele that reduces T2D risk by half increases *CCND2* expression. We identify eight cASE colocalizations, among which we found a T2D-associated *SLC30A8* variant. We make all data available through the TIGER portal (http://tiger.bsc.es), which represents a comprehensive human islet genomic data resource to elucidate how genetic variation affects islet function and translates into therapeutic insight and precision medicine for T2D.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

INTRODUCTION

Diabetes is a complex metabolic disease, characterized by elevated blood glucose levels, that affects >463 million people worldwide. Type 2 diabetes (T2D) accounts for >85% of diabetes cases and is strongly related to age, obesity, and sedentary lifestyle. Epidemiologic studies forecast increases in global prevalence up to 25% by 2030 (Khan et al., 2020; Saeedi et al., 2019; Wild et al., 2004). This makes the study and understanding of diabetes a top research and healthcare priority. Progressive pancreatic islet dysfunction is central to the majority of all types of diabetes and thereby key to gain insight into disease pathophysiology.

Great efforts have been dedicated to uncover the link between genetic variation and complex disease susceptibility through large-scale genetic studies. For T2D, >700 genetic loci have been identified to date (Bonàs-Guarch et al., 2018; Mahajan et al., 2018; Spracklen et al., 2020; Vujkovic et al., 2020). The vast majority of variants in these loci do not disrupt protein coding sequences (Miguel-Escalada et al., 2019; Pasquali et al., 2014). Thus, the mechanisms by which these variants influence predisposition to disease remain to be elucidated. As the number of newly identified risk variants keeps increasing, their functional interpretation constitutes the main bottleneck to gain insight into the underlying molecular mechanisms and, thus, to develop more effective and targeted preventive and therapeutic strategies (Claussnitzer et al., 2020).

To provide functional interpretation of non-coding variation, large international efforts have generated and integrated genomic, transcriptomic, and epigenomic data from a large variety of healthy and diseased samples to build comprehensive and genome-wide maps of functional annotations. Among others, the Genotype-Tissue Expression (GTEx) project uses expression quantitative trait loci (eQTL) analysis to link genetic variation with gene expression across 54 different human tissues (Aguet et al., 2020). The Roadmap Epigenomics Mapping project (Bernstein et al., 2010) and the International Human Epigenome project (Bujold et al., 2016) also provide a broad characterization of epigenomic signatures in a variety of tissues and cell types.

The functional interpretation of genetic variants, which are usually associated with moderate or small effect sizes, requires tools and resources that focus on cells and tissues that are affected in the disease of interest. The islets of Langerhans, which are clusters of specialized endocrine cells that are essential to maintain glucose homeostasis, play a central role in the etiology of T2D (Eizirik et al., 2020; Krentz and Gloyn, 2020). Because human islets are difficult to obtain (Barovic et al., 2019; Burgarella et al., 2013; Meier et al., 2015), large multi-tissue resources such as GTEx do not contain islet data and at best use whole pancreas as a proxy, despite the fact that 97% of the pancreatic tissue consists of exocrine cells that mask islet signals. Hence, the development of publicly available resources and tools that include data on islets is essential to translate T2D genetic signals into molecular and physiological mechanisms.

The first studies of eQTL in human islets pinpointed genes that may be influenced by genetic variants and thus possibly mediate T2D risk (van de Bunt et al., 2015; Fadista et al., 2014). Despite the small number of samples, they identified a few loci linked to differential expression of islet genes, which were enriched in genome-wide association study (GWAS) signals for T2D and related traits. More recently, the InsPIRE Consortium generated a large islet eQTL study with a sample size of 420 islet donors, which identified 46 T2D GWAS signals that colocalize with islet eQTL (Viñuela et al., 2020).

To further expand the understanding of human islet regulatory genomics and its role in T2D, the Horizon 2020 T2DSystems consortium gathered an extensive collection of human islet samples with gene expression, epigenomic data, and genotypic and phenotypic information, with a total of 514 samples, 207 of which were analyzed by the InsPIRE Consortium. In this study, we discovered 40 T2D risk signals that colocalize with eQTL or ASE signals by improving genotype imputation methods and analyses and by developing a new method to combine allele-specific expression (cASE) across samples, knowledge previously unknown.

Importantly, the results from this study are made publicly available to the community through the Translational human pancreatic Islet Genotype tissue-Expression Resource (TIGER, http://bsc.tiger.es) portal (Figure 1A). This portal integrates the newly generated data with publicly available T2D genomic and genetic resources to facilitate the translation of genetic signals into their functional and molecular mechanisms.

RESULTS

A catalog of genetic variation and gene expression in human pancreatic islets

To study gene expression and the effects of genetic variation in human pancreatic islets, we obtained newly generated and published human islet data from 514 organ donors of European background, distributed across 5 cohorts (Center for Genomic Regulation, Lund University, University of Oxford/University of Alberta, Università di Pisa, and Université Libre de Bruxelles) (Method details). The large majority of these samples came from non-diabetic adult donors, and only 30 were from diabetic organ donors (Table S1).

The DNA of 307 samples was isolated, sequenced, and genotyped (Table S1; Method details) and aggregated to be harmonized with the existing data from 207 samples. After quality control, filtering of RNA sequencing (RNA-seq) and genotyping array data (Method details), we had both high-quality genotypes and RNA-seq data for 404 human islet samples (Figure 1B), including 21 from diabetic donors.

To fully characterize the genetic variation present in the samples, genotype imputation was performed separately for each cohort using 4 different reference panels, as previously described (Bonàs-Guarch et al., 2018; Guindo-Martínez et al., 2021), 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015), Genome of the Netherlands (GoNL) (Boomsma et al., 2014), the Haplotype Reference Consortium (McCarthy et al., 2016), and UK10K (Walter et al., 2015). The results were integrated by selecting, for each variant, the imputed genotypes from the reference panel that achieved the best imputation quality (IMPUTE2 info score > 0.7; Method details). We have previously demonstrated that this approach results in increased overall coverage of genetic variation, as well as an

CellPress
OPEN ACCESS

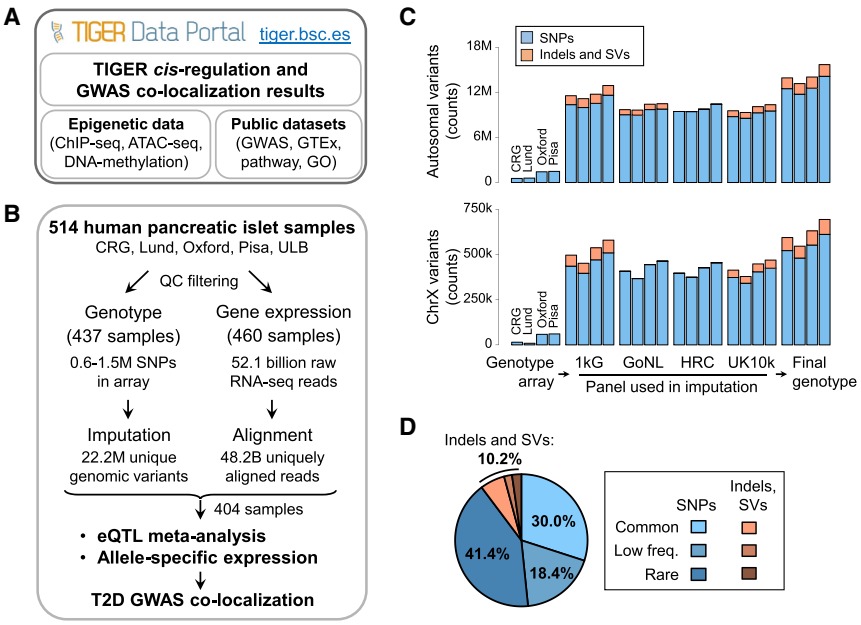


Figure 1. Project overview and genotype imputation

(A) Overview of the TIGER data portal.

(B) Datasets of the T2DSystems Consortium and project workflow.

(C and D) Multi-panel genotype imputation identified 13.1–15.7 million autosomal variants (top) and 550,000–700,000 chrX variants (bottom) (C), with (D) a large proportion of low-frequency (minor allele frequency [MAF] 1%–5%) and rare (<1%) variants, with 10.2% of structural variants (SVs), including small indels and large SVs.

increased number of significant associations, including those that are covered by only one of the reference panels (Guindo-Martínez et al., 2021). This allowed imputation of >22 million unique high-quality genetic variants across all of the samples, 10% of which were indels and small structural variants (SVs), and >1.05 million variants in chromosome X (Figures 1C and 1D; Table S2). Notably, this strategy allowed the accurate imputation of 4 million low-frequency (minor allele frequency [MAF] between 0.05 and 0.01) and 10 million rare (0.01 > MAF > 0.001) variants.

In addition, we performed bulk RNA-seq in 514 human islet samples, 460 of which were retained after stringent quality control, including >52 billion raw short reads. We uniquely aligned >48 billion reads (median of 93 million per sample) (Table S3), which allowed us to observe >22,000 genes expressed at >0.5 transcripts per million (TPM) (Method details).

An atlas of eQTLs in human pancreatic islets

To explore the association between genetic variation and gene expression, we performed an eQTL meta-analysis across 4 cohorts. We performed a *cis*-eQTL analysis in 404 samples, using data from each cohort independently. For each analysis, we corrected for known covariates (age, sex, and body mass index [BMI]), 7 genetic ancestry principal components, and probabilistic estimation of expression residuals (PEER) factors for hidden confounding factors (Stegle et al., 2012). The eQTL results from each of the 4 cohorts were then meta-analyzed (Figure 2A). This resulted in >1.11 million significant eQTLs in >21,115 eGenes (12,802 protein coding genes, 8,313 non-coding) at a 5% false discovery rate (FDR) after Benjamini-Hochberg correction for multiple testing (Benjamini and Hochberg, 1995) (Figure 2B). The quantile-quantile plot showed no baseline inflation in the results. More than 12% of all significant eQTLs were small indels or larger SVs, and this type of variation was the top associated variant for 14% of all genes. This is in line with what has been observed in primary human immune cell types, in which indels comprised 12.5% of the variants in the 95% credible sets for eQTLs (Kundu et al., 2020), and in GTEx, in which SVs were found to have a stronger effect than single nucleotide variants (Chiang et al., 2017).

To assay the potential functional impact of the identified eQTL variants, we tested for their enrichment in human islet regulatory regions, defined by a variety of pancreatic islet chromatin assays (Miguel-Escalada et al., 2019). We observed that eQTL variants overlapped with gene promoters with very strong fold enrichment when compared with a control set of genetic variants (3.1-fold for 1% FDR eQTL variants, $p = 3 \times 10^{-166}$) (Method details), as well as with strong enhancers (Miguel-Escalada et al., 2019) (2-fold, $p = 1.4 \times 10^{-16}$), and open-chromatin regions (1.4-fold, $p = 3.9 \times 10^{-45}$) (Figures 2C and S1). These results are consistent with eQTL studies in other tissues (Aguet et al., 2020).

Next, we contrasted the TIGER human islet results with the latest GTEx eQTL datasets, which comprised 54 human tissues, including whole pancreas, but not islets (Aguet et al., 2020). Of all significant human islet eQTLs, 64.7% were also significant in at least 1 GTEx tissue, whereas 35.3% were exclusive to human islets (Figure 2D, left panel). Only 30.5% of human islet eQTLs were also significant in whole pancreas in GTEx, an overlap that is similar to the rest of the GTEx tissues (26% mean overlap with T2D-related tissues, 29% with other tissues), highlighting that whole pancreas is not a better proxy for pancreatic islets than other tissues. In addition, when considering rare and low-frequency variants, the proportion of TIGER islet exclusive eQTLs increased to 76.5% (Figure 2D, right panel). These observations highlight again the importance of assaying human islets, since a sizeable proportion of the eQTLs cannot be found in other tissues. Interestingly, these observations also held true when we compared TIGER results with recently published InsPIRE eQTLs (Viñuela et al., 2020). Because of its imputation approach, TIGER interrogated a larger number of genomic variants (Figure S2A). Overall, 56.1% of the significant eQTLs were exclusive to our analysis (not assayed or non-significant in InsPIRE; Viñuela et al., 2020) (Figure S2B). Identification of eQTLs driven by low-frequency or rare variants may be more clinically effective, as significant low-frequency variants tend to

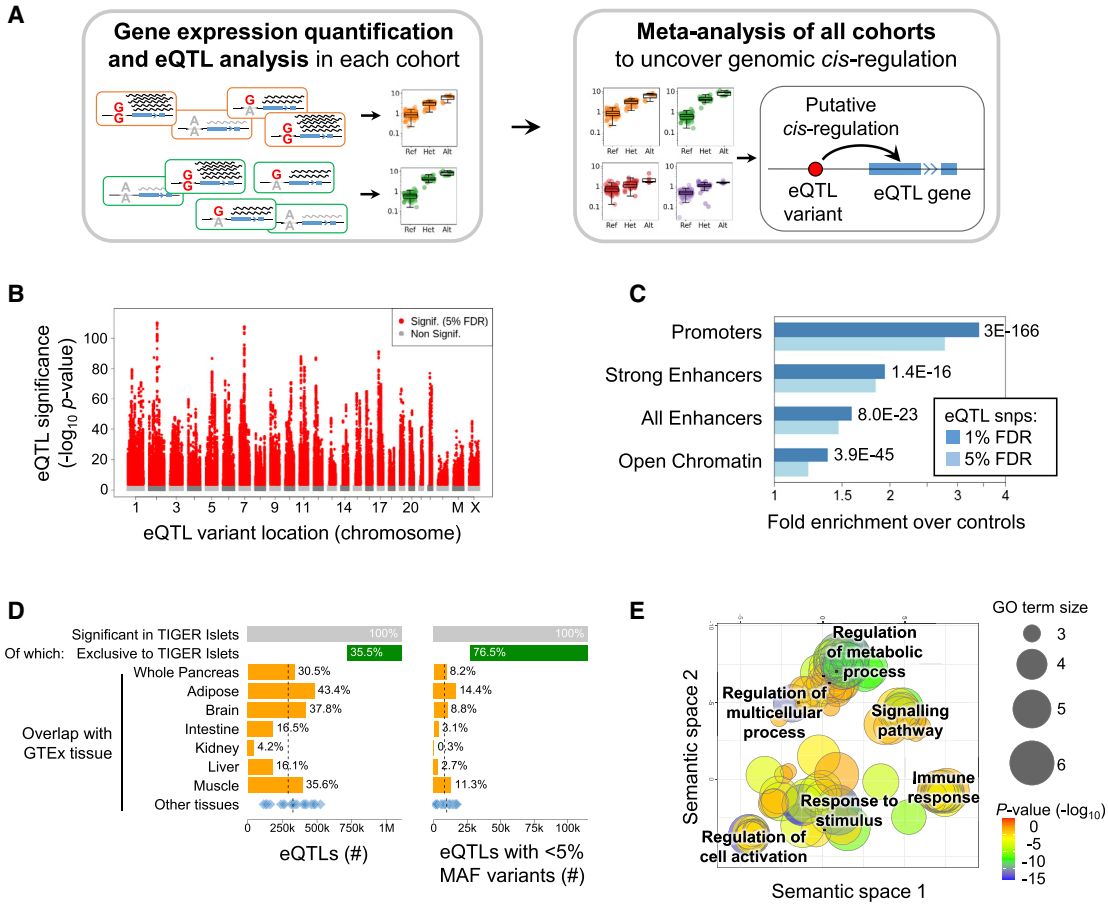


Figure 2. *cis*-eQTL meta-analysis in human pancreatic islets

(A) Overview of the meta-analysis.

(B) Manhattan plot of all eQTLs, including chrX, analyzed with female-only (F) or male-only (M) samples, and jointly (X).

(C) Fold enrichment over controls of significant eQTL variants, in islet regulatory chromatin regions. p values for 1% FDR eQTL enrichments are shown.

(D) Proportion of exclusive eQTLs in TIGER human islets (green) and previously found in GTEx project: tissues related to T2D etiology (orange), other tissues (blue); means in dashed lines. Right panel restricted to low MAF variants only.

(E) Gene Ontology analysis of the genes of TIGER-specific eQTLs.

have larger effects on disease risk and gene expression (Flannick, 2019). Notably, the proportion of TIGER exclusive eQTLs increased to 74.7% for low-frequency variants (Figure S2C), despite similar sample sizes between the studies. Overall, we identified 125,918 low-frequency eQTLs compared to 113,285 low-frequency eQTLs identified in the InsPIRE study (Figure S2C). This resulted in 20,742 eGenes, including the 69% of the 14,881 eGenes described in InsPIRE (Figure S2D). For eQTLs with variants present in both studies, the statistical strength of the association was correlated, as was the direction of effect for those <5% FDR significant in at least 1 of the 2 studies (Figures S2E and S2F). This indicates that the findings in the 2 studies are consistent, even when considering signals that did not reach significance in 1 of the 2.

Gene Ontology analysis of the significant human islet eQTL genes revealed signaling (including G protein-coupled receptor signaling) and metabolic regulation terms (Figure S3). In contrast, comparing TIGER-specific eQTL genes against those also present in GTEx tissues revealed strong enrichment for these terms as well as "response to stimulus" or "regulation of cell activation," and immune system terms (including "lymphocyte/T cell activation" and "regulation of immune system process") (Figure 2E). This suggests that these eQTLs involve β cell physiology genes, including some related to immune processes with potential relevance for T1D (Ramos-Rodríguez et al., 2019).

Islet eQTLs colocalize with T2D GWAS signals

To assess whether the identified eQTLs can help to identify effector transcripts for T2D risk variants, we investigated the intersection between *cis*-eQTLs and known T2D associations (Bonàs-Guarch et al., 2018; Mahajan et al., 2018; Vujkovic et al., 2020) by performing colocalization analyses using *COLOC* (Giambartolomei et al., 2014) (Method details).

This analysis uncovered 49 eQTL variants associated with the expression of 53 genes that significantly colocalized with T2D

CellPress
OPEN ACCESS

Table 1. Human pancreatic islet colocalization of eQTL meta-analysis with T2D GWAS

			COLOC		T2D GWAS					eQTL	
Chr	SNP	Gene	PP.H4.abf	SNP.PP.H4	EAF	EA	NEA	OR	p	p	Direction
1	rs1127215	*PTGFRN*	1.00	0.99	0.42	T	C	0.95	2.3E−13	4.8E−15	−−
1	rs1127215	*CD101*	1.00	0.96	0.42	T	C	0.95	2.3E−13	1.2E−7	−−
1	rs1493694	*NBPF7*	0.81	0.09	0.11	T	C	1.09	2.1E−16	1.0E−5	?+?+
1	rs340874	*RP11-478J18.2*	0.98	1.00	0.56	C	T	1.07	5.6E−26	1.3E−6	++++
1	rs4659836	*TBCE*	0.82	0.12	0.65	A	G	1.04	4.7E−9	2.9E−7	−−
3	rs3887925	*ST6GAL1*	1.00	1.00	0.55	T	C	1.06	1.4E−17	2.1E−13	++++
3	rs3887925	*AC007690.1*	1.00	1.00	0.55	T	C	1.06	1.4E−17	5.2E−9	++++
3	rs7640294	*SERBP1P3*	0.97	0.06	0.56	A	C	1.04	3.0E−8	1.6E−9	++++
4	rs1531583	*CPLX1*	0.87	0.13	0.046	T	G	1.12	1.2E−12	1.2E−6	++++
4	rs1580278	*BDH2*	0.81	0.73	0.53	A	C	0.96	2.9E−10	1.1E−9	++++
4	rs58730668	*ACSL1*	0.89	0.04	0.14	C	T	0.93	1.0E−13	2.5E−5	++++
6	rs6557267	*RGS17*	0.94	0.08	0.42	T	C	1.04	6.0E−8	8.2E−8	−−
8	rs1059592	*RP11-582J16.5*	0.81	0.12	0.35	A	G	1.03	4.5E−5	4.1E−15	−−
8	rs77292833	*LRP12*	0.84	0.05	0.12	G	C	0.96	1.6E−5	8.1E−8	++++
9	rs10811660	*CDKN2B-AS1*	0.99	0.48	0.17	A	G	0.85	6.6E−79	1.6E−7	−−
9	rs10963924	*SAXO1*	0.82	0.09	0.43	C	G	1.04	9.2E−10	1.6E−5	−−
10	rs827237	*PCBD1*	0.99	0.19	0.21	T	C	1.04	2.3E−7	2.4E−10	−−
11	rs15818	*HMBS*	0.84	0.06	0.4	G	A	1.03	4.5E−5	2.5E−7	++++
11	rs529623	*FXYD2*	0.92	0.83	0.52	C	T	0.97	5.8E−6	3.4E−7	++++
11	rs57635800	*HSD17B12*	0.95	0.24	0.29	A	G	1.05	8.5E−13	1.1E−19	−−
12	rs731304	*ABCC9*	0.80	0.19	0.24	A	G	0.97	1.1E−5	3.0E−11	++++
12	rs76895963	*CCND2*	0.36	1.00	0.02	G	T	0.62	5.3E−70	1.7E−6	+++?
12	rs77864822	*RMST*	0.99	0.81	0.07	G	A	0.93	2.2E−8	2.9E−14	++++
12	rs77864822	*RP11-528M18.2*	0.95	0.17	0.07	G	A	0.93	2.2E−8	3.6E−6	+-++
13	rs34584161	*CDK8*	1.00	0.98	0.24	G	A	0.95	2.9E−10	1.3E−17	−−
13	rs488321	*KL*	0.98	0.27	0.83	C	T	0.95	6.8E−10	4.3E−6	++++
14	rs10151752	*ACTR10*	0.86	0.26	0.59	G	A	0.97	7.2E−8	4.0E−6	++++
14	rs1803283	*RP11-600F24.7*	0.81	0.02	0.65	T	C	1.04	1.4E−7	2.5E−5	-+–
15	rs13737	*RP11-817O13.8*	0.84	0.10	0.24	T	G	0.96	7.3E−10	2.3E−6	++++
17	rs7218899	*USP36*	0.96	0.41	0.51	T	C	0.97	1.5E−6	2.4E−10	++++
17	rs8070260	*ZNHIT3*	0.94	0.13	0.53	G	A	0.97	1.1E−5	4.1E−8	−−
18	rs303760	*NPC1*	0.95	0.08	0.36	T	C	1.03	3.8E−6	2.4E−24	−−

Colocalizations not reported in Viñuela et al. (2020). The *R COLOC* package reports the approximate Bayesian factor posterior probability (*PP.H4.abf*) that there is one common causal variant and the posterior probability (*SNP.PP.H4*) that the *SNP* is the associated causal variant. The *GWAS* establishes the link between the *SNP* and T2D; the effect alleles (EA) with a frequency (EAF) are shown with the associated effect odds ratio (OR) and the p value. The *GWAS* data are as reported by the *DIAGRAM* Consortium (Mahajan et al., 2018). The eQTL p value is reported with the direction of the effect: up- ("+") or downregulation ("-") direction for the effect allele in the 4 meta-analysis cohorts (order: CRG, Oxford, Lund, and Pisa). "?" means that not enough samples are available in the cohort for the minor allele to compute a p value.

GWAS loci (Table S4), 32 of which were not previously reported (Table 1; Figure S4; Data S1). Among the 49 colocalizing signals (Data S1), rs77864822 (MAF = 0.07) minor allele (G) was associated with higher *RMST* (rhabdomyosarcoma 2 associated transcript) expression and decreased T2D risk (odds ratio [OR] = 0.93, $p = 2.2 \times 10^{-8}$) (Figure S4A). By interrogating the latest GWAS study on glycemic traits (Chen et al., 2021), we observed that the protective allele was associated with decreased fasting glucose ($\beta = -0.024$, $p = 4 \times 10^{-11}$), reduced HbA1c ($\beta = -0.087$, $p = 4.6 \times 10^{-4}$), and reduced 2-h glucose in an oral glucose tolerance test ($\beta = -0.064$, $p = 2.4 \times 10^{-4}$) (Table S4). Interestingly, we identified two low-frequency variants (Figures 3C and 3G), which may have large effect sizes, that colocalized with gene expression, suggesting a target gene and direction of effect (i.e., whether the genetic variant is associated with increased or decreased gene expression). The variant rs1531583 colocalized with *CPLX1* expression (Figures 3A–3C). Interestingly, the same variant was associated with *PCGF3* but not with *CPLX1* gene expression in whole pancreas in GTEx (Figure 3B), demonstrating once again the importance of performing eQTL in the relevant tissue. A detailed analysis of enhancer chromatin marks in human islets showed that

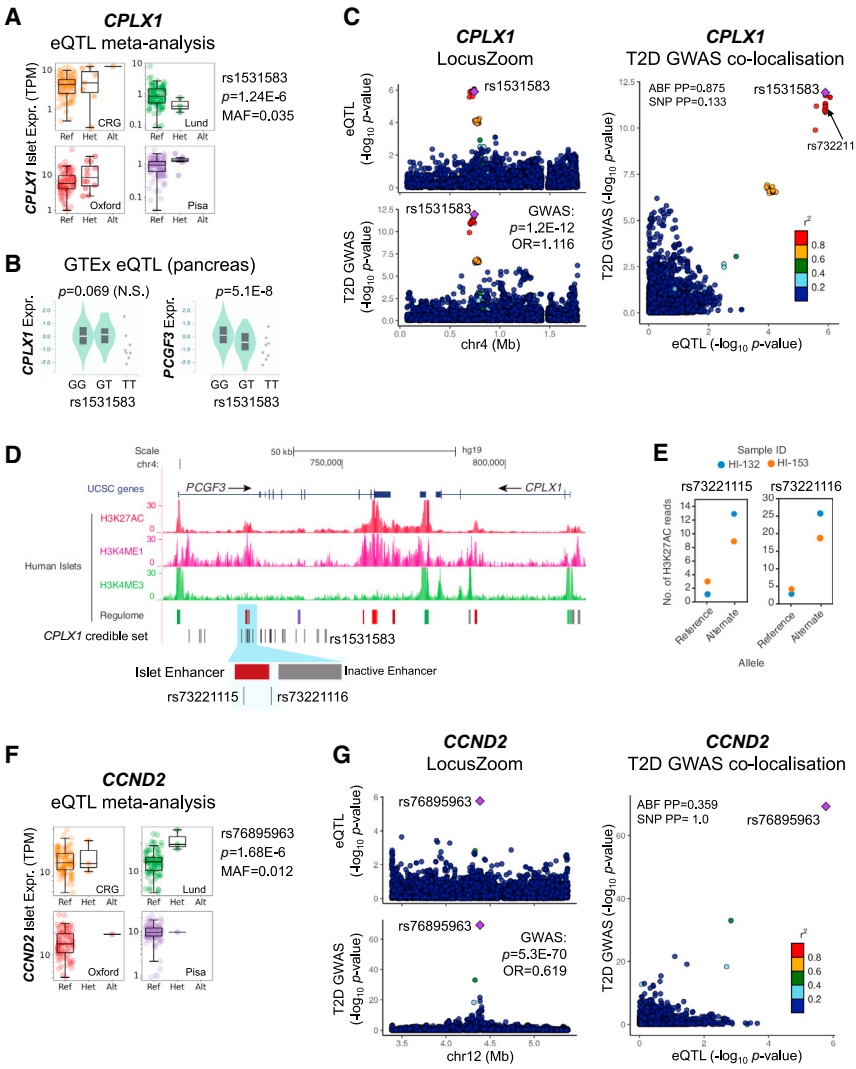


Figure 3. Examples of colocalization of pancreatic islet eQTLs with T2D GWAS

(A) Boxplots representing expression of CPLX1 across different genotypes of variant rs1531583 in each of the cohorts and final meta-analysis results.

(B) rs1531583 was not significant in GTEx whole pancreas for *CPLX1*, but instead it was for *PCGF3* (bottom).

(C) LocusZoom plots of islet eQTL (top) and T2D GWAS (bottom) signals for rs1531583-*CPLX1*, and their co-localization (right). ABF, approximate Bayes factor, PP, posterior probability.

(D) An islet enhancer overlaps with rs73221115 and rs73221116, part of the *CPLX1* credible set of SNPs.

(E) Two human islet samples heterozygous for rs73221115 and rs73221116 showed allelic imbalance in their H3K27ac enhancer chromatin marks.

(F) eQTL meta-analysis of *CCND2* and the low-frequency *cis*-regulatory variant rs76895963.

(G) Co-localization plots for rs76895963-*CCND2*, as in (B).

rs73221115 (r^2 = 0.978 with rs1531583) and rs73221116 (r^2 = 0.98 with rs1531583) had allele-specific H3K27ac binding, suggesting that these 2 variants are the most likely causal variants of the *CPLX1* locus (Figures 3D and 3E). We also identified significant colocalization between the low-frequency variant rs76895963, known to be associated with nearly half reduced T2D risk (Steinthorsdottir et al., 2014), and increased *CCND2* expression in islets (Figures 3F and 3G). This variant was also associated with reduced fasting glucose ($\beta = -0.033$, p = 0.0017), HbA1c ($\beta = -0.042$, p = 3.6 × 10^{-8}), and 2-h glucose in oral glucose tolerance test ($\beta = -0.095$, p = 0.01) (Table S4).

An atlas of cASE in human pancreatic islets

Preferential expression of mRNA copies containing 1 of the 2 alleles of a genetic variant (allele-specific expression [ASE]) can result from *cis*-regulation. However, ASE can occur while the overall amount of expression of a gene remains constant, and therefore this type of regulation cannot be identified by conventional eQTL analysis. While some methods have been developed to identify ASE in gene expression data in single (Edsgärd et al., 2016; Mayba et al., 2014) or multiple samples (Fan et al., 2020; Liang et al., 2021), these methods did not aim to identify candidate *cis*-regulatory variants for the ASE effect.

We implemented a cASE pipeline for the analysis of ASE replicated across multiple samples that differ in age, gender, BMI, and environmental factors, thereby likely to stem from *cis*-regulatory genetic variants (Figure 4A). cASE analysis complements eQTL analysis, and additionally controls for (1) environmental and batch effects, which are important confounding factors in eQTL studies (Akey et al., 2007; Branham et al., 2007; Churchill, 2002; Fare et al., 2003; Irizarry et al., 2005; Yang et al., 2002); (2) sample heterogeneity, which is prevalent in human islets (Leek and Storey, 2007); and (3) *trans* effects, since these would affect the 2 alleles in the same manner and thus cannot result in ASE. cASE combines ASE from each sample into a single *Z* score statistic that summarizes overall ASE across the cohort of samples (Figure S5; Method details,) (Newhall et al., 1949). Variants that preferentially express the reference allele result in a positive *Z* score and vice versa (Figure 4A).

Using this strategy, we identified 2,707 genes with 5,271 reporter variants showing cASE in human islets, at 5% FDR (Figure 4B). The similar number of reference and alternate imbalanced variants (2,606 and 2,589, respectively) showed that alignment biases toward the reference allele were successfully controlled (Figures S5B–S5E).

When comparing cASE genes against a set of non-significant genes (matched by gene expression level, Method details), we observed that cASE genes were enriched for islet-specific

CellPress
OPEN ACCESS

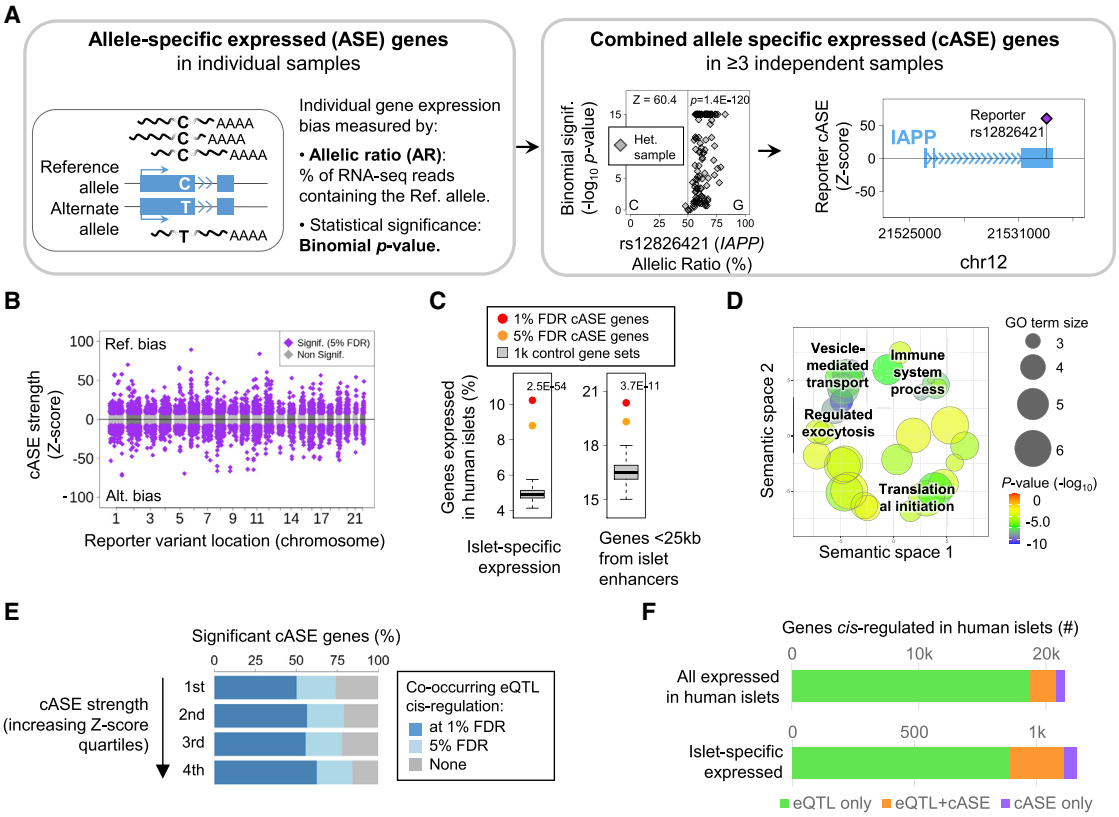


Figure 4. Combined ASE analysis in human islets

(A) Overview of the cASE analysis, with *IAPP* as example of a gene with an imbalanced reporter variant, rs12826421.

(B) Manhattan plot of cASE, positive values refer to reference-biased genes, negative to alternate.

(C) Significant cASE genes are enriched for islet-specific expression and proximity to islet-regulatory regions. p values for 1% FDR eQTL enrichments are shown.

(D) Gene Ontology analysis of cASE significant genes.

(E) In genes with significant cASE, the proportion of those also identified as eGenes grew with increasing cASE magnitude.

(F) Total number of *cis*-regulated genes (top) and of islet-specific expressed (bottom), identified only by the eQTL analysis (green), cASE (purple), and both (orange).

expression (2.1-fold, $p = 2.5 \times 10^{-54}$ at 1% FDR) and preferentially located near islet regulatory regions (1.23-fold, $p = 3.7 \times 10^{-11}$) (Figure 4C). Gene Ontology analysis (Method details) revealed islet-specific terms such as "vesicle-mediated transport" and "regulated exocytosis" (Figure 4D), related to insulin production and secretion in β cells. As a notable example, the islet amyloid polypeptide gene (*IAPP*) was among the most imbalanced cASE genes. *IAPP* had 7 independent reporter SNPs at 1% FDR (Figure 4A, right panel), all of which had strong imbalance toward the reference allele in the >100 independent samples that were heterozygous for the variants. Notably, there were no significant eQTLs for this gene, highlighting the complementarity between the two methods to identify regulatory variation. These findings highlight the potential of cASE to identify genes involved in regulating pancreatic islet physiology.

Given that eQTL and cASE analyses are complementary methods to detect genes affected by *cis*-regulation, we assessed the concordance between each of them. We interrogated the proportion of genes with significant eQTL of all cASE genes across absolute *Z* score quartiles (strength of imbalance) and observed that the proportion of eQTL genes increased with increasing *Z* scores (Figure 4E), indicating that stronger cASE effects were more likely to be also identified in eQTL analysis, and showing a correlation between the 2 effects.

Of 2,707 cASE significant genes, 2,052 (75.8%) were detected in eQTL analyses, whereas 655 (24.2%) were detected uniquely through cASE (Figure 4F, top panel). The same trend was observed when considering only islet-specific genes. Among 270 islet-specific significant eGenes detected by cASE, 218 were also detected by eQTL analysis, while the remaining 52 were exclusively found by cASE (Figure 4F, bottom panel).

Mapping distal cASE variants allows cASE colocalization analysis and implicates additional T2D effector genes

We next developed an approach to identify distal putative cASE regulatory variants by interrogating all of the variants within the same topologically associated domain as the reporter variant (i.e., the variant located in the transcribed gene region). For each candidate regulatory variant, we stratified samples

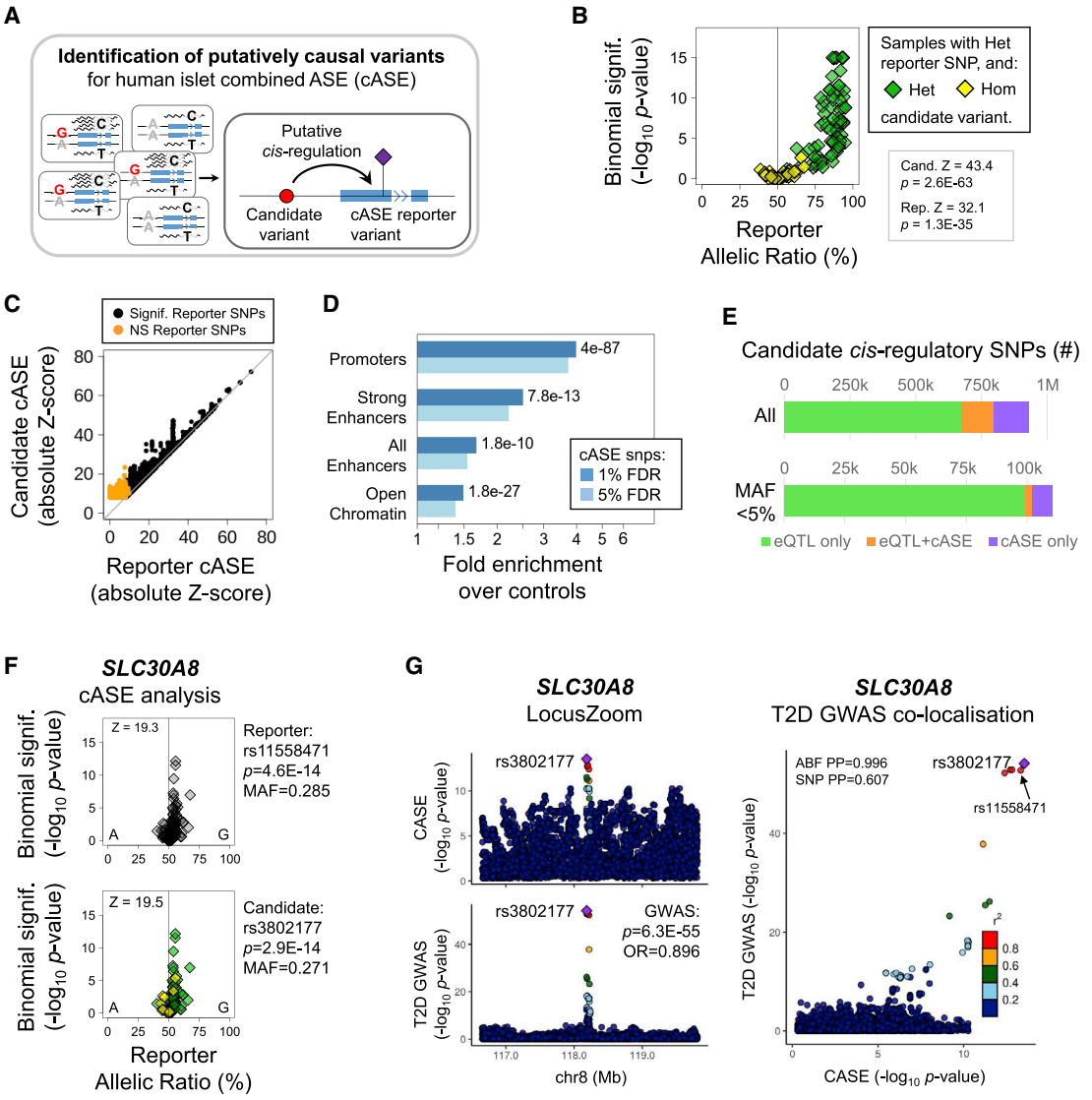


Figure 5. Identification of *cis*-regulatory variants in combined ASE

(A) Overview of the analysis.

(B) An example of *cis*-regulatory variant analysis; the samples Het for the candidate variant (green) have a higher cASE *Z* score for the reporter SNP, while samples that are Hom for the candidate (yellow) do not show significant imbalance for the reporter SNP.

(C) Candidate variants often have stronger *Z* scores than the reporters, including some reporter variants that were non-significant by themselves (orange).

(D) Fold enrichment over controls of significant cASE candidate cis-regulatory variants, in islet regulatory chromatin regions. p values for 1% FDR cASE enrichments.

(E) Total number of candidate cis-regulatory variants (top) and low-frequency variants (bottom) identified by only the eQTL analysis (green), cASE (purple), and both (orange).

(F) cASE analysis for *SLC30A8*, its best reporter SNP (top), and best candidate variant (bottom).

(G) LocusZoom plots of islet cASE (top) and T2D GWAS (bottom) signals for rs3802177-*SLC30A8*, and their colocalization (right).

between the heterozygous and homozygous for the candidate variant. We then recomputed cASE of the reporter variant (i.e., the transcribed variant) for each of the groups (Figure 5A). This approach allowed us to prioritize the candidate variant that had the highest reporter cASE when the candidate regulatory variant was also heterozygous, compared to when the regulatory variant was homozygous (Figure 5B; Method details). This method does not require haplotype phasing since it compares heterozygous versus homozygous and is agnostic to the direction of the association.

This analysis uncovered 256,981 putative regulatory variants for 3,425 genes, including 570 genes that had no significant reporter variant by themselves, but that did reach significance upon stratifying by the genotype of regulatory variants (Figure 5C, orange points). To assay the potential functional impact of the identified reporter variants, we tested for their enrichment in

CellPress
OPEN ACCESS

Table 2. cASE with T2D GWAS

			COLOC		T2D GWAS					cASE				
Chr	SNP	Gene	PP.H4. abf	SNP. PP.H4	EAF	EA	NEA	OR	p	Reporter variant	Ref	Alt	p	*Z* score
1	rs1127215	*PTGFRN*	0.99	0.98	0.42	T	C	0.95	2.3E−13	rs1127656	C	T	8.5E−9	14.6
4	rs10937721	*WFS1*	0.95	0.26	0.59	C	G	1.09	1.6E−40	rs1046320	G	A	3.2E−16	−20.9
8	rs3802177	*SLC30A8*	1.00	0.61	0.31	A	G	0.90	6.3E−55	rs11558471	A	G	2.9E−14	19.5
10	rs2280141	*PLEKHA1*	0.96	0.06	0.48	G	T	0.95	2.0E−13	rs1045216	A	G	1.7E−11	17.2
11	rs35251247	*HSD17B12*	0.95	0.21	0.29	A	G	1.05	8.5E−13	rs11555762	C	T	5.1E−93	52.9
11	rs35251247	*RP11-613D13.5*	0.93	0.07	0.29	A	G	1.05	8.5E−13	rs35251247	G	A	6.8E−12	−17.5
11	rs5215	*KCNJ11*	0.83	0.36	0.63	T	C	0.93	2.0E−26	rs5215	C	T	8.6E−6	−11.1
11	rs529623	*FXYD2*	0.95	1.00	0.52	C	T	0.97	5.8E−6	rs529623	T	C	3.4E−231	84.1
11	rs529623	*RP11-728F11.3*	0.91	0.81	0.52	C	T	0.97	5.8E−6	rs869789	G	A	7.2E−16	20.7
12	rs10879261	*TSPAN8*	0.85	0.08	0.41	G	T	1.05	3.7E−13	rs3763978	C	G	7.2E−11	−16.6
16	rs6600191	*ITFG3*	0.86	0.24	0.18	C	T	0.94	7.0E−13	rs7193384	C	G	1.1E−7	13.4
18	rs1788762	*C18orf8*	0.96	0.06	0.64	C	G	0.97	2.3E−6	rs1788820	A	G	3.2E−25	−26.7
18	rs1788762	*NPC1*	0.96	0.06	0.64	C	G	0.97	2.3E−6	rs1788820	A	G	3.2E−25	−26.7
19	rs3111316	*CALR*	0.99	0.47	0.59	A	G	1.05	1.6E−12	rs1049481	G	T	1.6E−76	−47.9

The *R COLOC* package reports the approximate Bayesian factor posterior probability (PP.H4.abf) that there is one common causal variant and the posterior probability (SNP.PP.H4) that the SNP is the associated causal variant. The GWAS establishes the link between the SNP and T2D; the effect alleles (EA) with a frequency (EAF) are shown with the associated effect OR and the p value. The GWAS data are as reported by the DIAGRAM Consortium (Mahajan et al., 2018). The cASE analysis provides the allelic imbalance for the allele represented by the reporter SNP with a reference allele (Ref) and an alternative allele (Alt), a p value (FDR threshold of 0.006), and a Z score. An increased *Z* score refers to increased expression of the reference allele.

human islet regulatory regions (Miguel-Escalada et al., 2019), observing overlap with gene promoters with very strong fold enrichment when compared with a control set of genetic variants (4-fold for 1% FDR eQTL variants, $p = 4 \times 10^{-87}$) (Method details), as well as with strong enhancers (Miguel-Escalada et al., 2019) (2.5-fold, $p = 7.8 \times 10^{-13}$) and open-chromatin regions (1.5-fold, $p = 1.8 \times 10^{-27}$) (Figure 5D). When comparing these *cis*-regulatory variants with the 1.11 million eQTLs, we found 123,748 variants were significant by both methods (3,138 with MAF <5%), and a further 133,233 (9,190 with MAF < 5%) were identified only by cASE (Figure 5E), showcasing the relevance of this analysis for enriching genetic *cis*-regulatory discovery.

Assigning statistical significance to cASE distal regulatory variants allowed us to test for colocalization between cASE regulatory variants and T2D GWAS variants. For each T2D GWAS locus, we assessed all of the regulatory variants for all of the imbalanced genes in the region and identified 14 colocalized locus-gene pairs (Table 2; Figure S6; Data S2). Of these, 6 had also been identified in eQTL/T2D GWAS colocalization analyses, showing consistency between the 2 methods. Interestingly, the 8 colocalizations identified by cASE alone, *WFS1*, *SLC30A8*, *RP11-613D13.5*, *KCNJ11*, *RP11-728F11.3*, *TSPAN8*, *C18orf8*, and *CALR*, suggested that these T2D variants may mediate disease risk by causing an imbalance in allelic expression, rather than altering overall gene expression (Figure S6). A notable example was the highly significant cASE observed in *SLC30A8* (rs11558471; $p = 2.9 \times 10^{-14}$), which showed colocalization with a well-established T2D-associated variant (Figures 5F and 5G; Table S5) for which there was no eQTL colocalization. Thus, cASE analysis uncovered additional disease-relevant genomic regulation and provides a potential biological mechanism underlying the association.

A web portal to explore regulatory variation and genomic pancreatic islet information

Finally, to provide the research community with a user-friendly open access tool to explore these findings and mine the molecular basis of complex diseases influenced by pancreatic islet biology, we created TIGER (http://tiger.bsc.es) (Figure S7). This portal integrates the results obtained in this study with other public genomic, transcriptomic, and epigenomic pancreatic islet resources, as well as T2D GWAS meta-analysis summary statistics (Method details).

The TIGER website represents homogeneous gene expression levels from 446 RNA-seq pancreatic islet samples corrected for batch and covariate effects, and enables comparison with GTEx expression data (Aguet et al., 2020) (Method details).

In addition to the eQTL and cASE results and to provide further functional assessment, we gathered islet regulatory information (Akerman et al., 2017; Miguel-Escalada et al., 2019; Pasquali et al., 2014), methylation marks (Hall et al., 2014; Thurner et al., 2018), and chromatin modification datasets (Dunham et al., 2012; Gaulton et al., 2010; Stitzel et al., 2010). Furthermore, to enable the translation of genetic variation to disease risk, we integrated the latest T2D GWAS meta-analysis summary statistics (Bonàs-Guarch et al., 2018; Mahajan et al., 2014, 2018; Scott et al., 2017) (Figure 1A).

The TIGER database contains expression and molecular data for 59,625 Gencode genes (version gencode.v23lift37; Frankish et al., 2019) and >26 million variants. The portal allows users to perform both variant and gene-centric queries. The results are displayed in a set of graphical tools and a genomic browser (Down et al., 2011) that help visualize and interpret the molecular context of the query. Each table can be downloaded in csv

format, and the genomic browser integrates tools to search and zoom in on a region, add new tracks, and export the data as publication image. As a result of these efforts, the TIGER resource has already been used in recent studies (Hodson and Rorsman, 2020; Saponaro et al., 2020a, 2020b).

As an example, we present the visualization of *MTNR1B*, a gene associated with T2D and impaired insulin secretion (Lyssenko et al., 2009). This gene is lowly expressed in pancreatic islets (median 0.25 TPM), but virtually absent in whole pancreas and other GTEx tissues (median 0 TPM), except for testis (median 0.61 TPM) and brain (median 0.06 TPM), highlighting the utility of this resource for studying human islet-specific expression (Figures S7A and S7B). A T2D risk-associated locus has been described and fine-mapped (Mahajan et al., 2018) to a single variant (rs10830963, $p = 4.8 \times 10^{-43}$, posterior probability [PP] = 0.99; Figures S4B and S7C). Notably, this variant is located within islet H3K27ac peaks, suggesting potential regulatory implications (Figure S7D). The close-up look at this locus illustrates that the TIGER portal can be easily used to interrogate gene expression and the epigenomic and genomic variation regulatory landscape, providing a very valuable resource to the research community to study complex diseases affecting pancreatic islets.

DISCUSSION

By analyzing a large multi-cohort dataset of pancreatic islets with gene expression and dense genotyping data, we have uncovered 1 million significantly associated variant-gene pairs. Of all of the associations we found, 35.3% were islet specific, highlighting the importance of performing tissue-specific eQTL studies (Figure 2D). Remarkably, 17 human islet eQTLs that colocalized with T2D GWAS signals were not associated with gene expression in any GTEx tissue, including whole pancreas, which emphasizes the fact that pancreas cannot be used as a proxy for pancreatic islets and vice versa.

We compared our findings with those obtained in the InsPIRE islet eQTL study that comprised 420 samples (Viñuela et al., 2020), 207 of which were also included in our study. We observed that 18 (34%) of the 53 eQTLs that colocalized with T2D GWAS signals were also identified in InsPIRE (Table S4). The improved power in our study obtained by the use of integrative approaches, such as combined reference panels genotype imputation and meta-analysis allowed us to detect lower MAF eQTL signals (10.4% with <5% MAF), representing a 7-fold increment of low-frequency eQTL variants compared to this previous islet eQTL study. Importantly, the meta-analyses also allow us to compare the heterogeneity of the associations between cohorts and filter out signals that are not consistent across cohorts, thereby avoiding false positives.

We uncovered 32 T2D colocalizations, 2 of which were led by low MAF variants, including variants associated with the expression of *CCND2*, *RMST*, and *CPLX1*. The variant rs76895963 (MAF = 0.02) that upregulates *CCND2* is associated with a nearly 50% reduced risk of T2D (OR = 0.58) (Mahajan et al., 2018; Steinthorsdottir et al., 2014) and is potentially implicated in the perinatal development of human β cells (Osonoi et al., 2020). While the PP of the colocalization was below the threshold of 0.8, the SNP had a clear eQTL with the gene, and LocusCompare plots showed convincing colocalization (Figure 3G). The variant rs77864822 (MAF = 0.07) upregulates *RMST* expression and decreases T2D risk. *RMST* is a reportedly neuron-specific long non-coding RNA involved in neurogenesis (Ng et al., 2013); it is well expressed in human islet cells (Kaur et al., 2018), but its function in β cells is unknown. The variant rs1531583, with the minor T allele associated with increased T2D risk (Mahajan et al., 2018), upregulates *CPLX1,* encoding complexin-1, again, a reportedly neuron-specific gene. Complexin-1 plays a role in Ca^{2+}-dependent insulin exocytosis in rodent β cells, although it is intriguing that both *CPLX1* silencing and overexpression impaired insulin secretion (Abderrahmani et al., 2004). GWAS often report as a target the gene that is closest to the variant, in this case *PCGF3*. Notably, rs1531583 lies in an intronic region of *PCGF3* and is an eQTL for this gene in several GTEx tissues. In human islets, however, it is specifically associated with *CPLX1* expression and not with *PCGF3*, challenging the hypothesis that the closest gene is often the most likely target gene (Figures 3A–3E).

The imputation with 4 reference panels allowed us to analyze different sources of genetic variation, including indels and SVs. In our study, 12.6% of the eQTL are indels. This stresses the fact that indels are a significant part of the genetic background influencing RNA expression. Unfortunately, the largest available T2D GWAS dataset (Mahajan et al., 2018) did not consider indels, and so we could not include them in our colocalization analyses. In the near future, this approach could be used to fine-map the contribution of indels and SVs to disease risk.

Capitalizing on this valuable pancreatic islet resource, we also analyzed for the first time *cis*-regulation via ASE. We developed a method called cASE, which combines ASE across samples, maximizing the power to detect variants associated with ASE. We identified variants associated with allelic imbalanced expression while not changing overall gene expression, and thus undetectable by eQTL. We extended the cASE results in colocalization analysis and identified 14 T2D colocalizations. Among them, 8 signals non-detected in the eQTL/T2D GWAS colocalization included widely reported T2D-associated signals in *WFS1*, *SLC30A8*, *KCNJ11*, *TSPAN8*, *C18orf8*, and *CALR*. For these, the lead SNP causes allelic imbalance but no overall gene expression change. These findings suggest that a subset of regulatory genetic variants confer disease risk by causing imbalance in the allelic expression of their target genes, a mechanism for which knowledge is lacking. A particular locus of interest was the colocalization for common variant rs3802177 associated with *SLC30A8*. rs3802177 is in strong linkage disequilibrium with rs13266634 T2D-associated variant, widely discussed in the literature (Carvalho et al., 2017; Gupta and Vadde, 2020; Li et al., 2017; Sladek et al., 2007). In our study, both variants had nearly identical p values ($p = 2.9 \times 10^{-14}$ for rs3802177 and $p = 3.3 \times 10^{-14}$ for rs13266634), showing that either or both could induce allelic imbalance. Rare loss-of-function variants in *SLC30A8* strongly reduce T2D risk (Flannick et al., 2019) by enhancing insulin secretion (Dwivedi et al., 2019). However, the direction of effect of the common coding variants is not known. Our cASE results suggest that imbalanced expression toward the rs13266634-T allele is protective for T2D. Since

SLC30A8 loss-of-function decreases risk, these results suggest that the rs13266634-T allele may cause reduced *SLC30A8* function.

This study has a number of limitations. First, there is a substantial overlap of samples between the TIGER and InsPIRE studies. For the variants that were present in both studies, ~70% of TIGER eQTLs were also identified in InsPIRE. The difference in overlapping signals could be due to the lack of power to identify associations or to heterogeneity in the samples or eQTL methodology used. Since TIGER has samples overlapping with InsPIRE, we cannot consider TIGER a replication of InsPIRE results or vice versa. However, results identified in both studies can be considered confirmed. Future efforts should focus on the careful analysis of non-overlapping islet samples from the 2 initiatives. Power will increase further with the integration in TIGER of additional datasets by the human islet community, which we will warmly welcome. A second limitation of this study is that the majority of samples is of European ancestry. Hence, whereas it is a great resource for functional follow up of variants associated with diabetes and related traits, this resource is not useful as a follow-up of variants that are frequent enough only in non-European populations (Mercader and Florez, 2017; Spracklen et al., 2020; Vujkovic et al., 2020). Future human islet omics and genetic studies should focus on collecting data from diverse ancestries. Third, the analysis of pancreatic islet bulk RNA-seq data does not allow the comparison of different cell types that are present in pancreatic islets. Studies using single-cell sequencing will enable the identification of cell-type-specific eQTLs. However, large enough sample sizes of human islet single-cell RNA-seq and paired genotype array datasets are not available yet.

In summary, we generated a large expression regulatory variation resource in human pancreatic islets, a tissue with a central pathogenic role in most, if not all, types of diabetes. The results are available through the TIGER web portal, which constitutes a user-friendly visualization tool that facilitates the exploration of the datasets, democratizing human islet genomic information to all islet researchers and clinicians. We expect that this resource, in combination with the growing number of large-scale genetic and functional studies, will represent a critical step forward toward understanding the molecular underpinnings of complex diseases that affect pancreatic islet biology and provide a path for the identification of novel and personalized drug targets.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Islet sample collection and genotyping
  - Centre for Genomic Regulation (CRG)
  - Lund University Diabetes Centre (Lund)
  - University of Oxford/University of Alberta (Oxford)
  - University of Pisa (Pisa)
  - ULB Center for Diabetes Research (ULB)
- METHOD DETAILS
  - Genotyping quality control
  - Genotype phasing and imputation
  - RNA-seq read mapping
  - Sample concordance verification between genotype and gene expression
  - TIGER web portal development
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - eQTL analysis
  - Identifying variant regulatory enrichments using GREGOR
  - Comparison of TIGER eQTLs with the GTEx and InsPIRE datasets
  - Colocalization analysis
  - Generation of an unbiased set of ASE reporter variants
  - Identification of ASE
  - Assessing cASE using Stouffer's *Z*-score
  - Assessing the significance of cASE *Z*-scores
  - Regulatory enrichment of cASE significant genes
  - Gene ontology analyses and islet-specific expression
  - Identifying candidate SNPs putatively leading to cASE
  - Scaling human islet gene expression values to allow comparisons with the GTEx expression datasets in TIGER

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109807.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation program T2Dsystems under grant agreement no. 667191. L.A. was supported by grant BES-2017-081635 of the Severo Ochoa Program, awarded by the Spanish government. I.M. was supported by the FJCI-2017-31878 Juan de la Cierva grant, awarded by the Spanish government. Work in the Cnop and Eizirik labs was further supported by the Fonds National de la Recherche Scientifique (FNRS), the Brussels Region Innoviris project DiaType, and the Walloon Region SPW-EER Win2Wal project BetaSource, Belgium. D.L.E. is supported by a grant from the Welbio–FNRS, Belgium. P.M., L.G., D.L.E., and M.C. are supported by the Innovative Medicines Initiative 2 Joint Undertaking Rhapsody, under grant agreement no. 115881, which is supported by the European Union's Horizon 2020 research and innovation programme, EFPIA and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0097. J.M.M. is supported by American Diabetes Association Innovative and Clinical Translational Award 1-19-ICTS-068. J.C. is supported by an Expanding Excellence in England Award from Research England. H.M., J.L.S.E., and L.E. are supported by the Swedish Strategic Research Foundation (IRC15-0067). A.L.G. is a Wellcome Trust Senior Fellow in Basic Biomedical Science. This work was funded in Oxford and Stanford by the Wellcome Trust (095101, 200837, 106130, and 203141 [all to A.L.G.]) and the NIH (U01-DK105535 and U01-DK085545 [A.L.G.]). The research was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) (A.L.G.). I.M.-E. was supported by the EFDS/Novo Nordisk Rising Star Programme. Work in the Ferrer lab was supported by the Imperial College London Research Computing Service, the NIHR Imperial BRC, and the Centre for Genomic Regulation (CRG) genomics facility, and grants from Ministerio de Ciencia e

Innovación (BFU2014-54284-R and RTI2018-095666-B-I00), the Medical Research Council (MR/L02036X/1), the Wellcome Trust Senior Investigator Award (WT101033), and the European Research Council Advanced Grant (789055). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. The technical support group from the Barcelona Supercomputing Center is gratefully acknowledged. Finally, we thank the entire Computational Genomics group at the BSC for their helpful discussions and valuable comments on the manuscript. We also acknowledge Cristian Opi and Laia Codó from the Barcelona Supercomputing Center for excellent website design and allocation of technical support and Isabelle Millard and Anyishaï Musuaya from the ULB Center for Diabetes Research for excellent technical and experimental support.

AUTHOR CONTRIBUTIONS

L.A., A.P., I.M., J.F., J.M.M., M.C., and D.T. conceived and planned the main analyses. J.F. provided unpublished allelic chromatin immunoprecipitation sequencing (ChIP-seq) and RNA-seq datasets and supervised cASE, which was developed and implemented by I.M. while he pursued his PhD in IDIBAPS and Imperial College London. I.M. further applied cASE in the TIGER dataset with the collaboration of L.A., M.G.-M., S.B.-G., M.P., R.A., and J.M.M. A.P. performed the eQTL and colocalization analyses with the collaboration of L.A., M.G.-M., S.B.-G., M.D., R.A., and J.M.M. L.A. developed the TIGER portal with the collaboration of R.R. and J.M.M. and performed the expression analysis with the collaboration of I.M., A.P., and J.M.M. I.M., A.P., L.A., J.M.M., D.T., and M.C. wrote and edited the manuscript. G.A. and I.M.-E. contributed the islet regulatory data and analysis. I.M., S.B.-G., and J.F. contributed the Imperial and CRG data and analysis. J.L.S.E., L.E., H.M., and L.G. contributed the Lund data and analysis. J.-V.T., D.L.E., and M.C. contributed the ULB data and analysis. M.S., L.M., and P.M. contributed the Pisa data and analysis. M.S., L.M., and P.M. contributed the Pisa islet samples. J.L.S.E. contributed the Pisa sample sequencing. V.N. contributed the Pisa sample genotyping. J.M.T., V.N., and A.L.G. contributed the Oxford data and analysis and the genotyping of the Pisa samples. X.G.-H. prepared the chromatin immunoprecipitation, RNA, and DNA samples, and managed the CRG data generation. A.L.G., J.L.S.E., P.M., D.L.E., J.F., J.M.M., M.C., and D.T. provided guidance in the design and during the development of the project. D.L.E., M.C., and D.T. worked on the creation of TIGER. J.C. and MAGIC contributed the MAGIC data and analysis. J.M.M., M.C., and D.T. supervised the study.

DECLARATION OF INTERESTS

A.L.G.'s spouse is an employee of Genentech and holds stock options in Roche.

Received: May 28, 2021
Revised: July 23, 2021
Accepted: September 16, 2021
Published: October 12, 2021

REFERENCES

Abderrahmani, A., Niederhauser, G., Plaisance, V., Roehrich, M.E., Lenain, V., Coppola, T., Regazzi, R., and Waeber, G. (2004). Complexin I regulates glucose-induced secretion in pancreatic β-cells. J. Cell Sci. *117*, 2239–2247.

Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Akerman, I., Tu, Z., Beucher, A., Rolando, D.M.Y., Sauty-Colace, C., Benazra, M., Nakic, N., Yang, J., Wang, H., Pasquali, L., et al. (2017). Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory Networks. Cell Metab. *25*, 400–411.

Akey, J.M., Biswas, S., Leek, J.T., and Storey, J.D. (2007). On the design and analysis of gene expression studies in human populations. Nat. Genet. *39*, 807–808, author reply 808–809.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Barovic, M., Distler, M., Schöniger, E., Radisch, N., Aust, D., Weitz, J., Ibberson, M., Schulte, A.M., and Solimena, M. (2019). Metabolically phenotyped pancreatectomized patients as living donors for the study of islets in health and diabetes. Mol. Metab. *27S*, S1–S6.

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

Bernoulli, J. (1899). Wahrscheinlichkeitsrechnung (Ars Conjectandi) (W. Engelmann).

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.

Bonàs-Guarch, S., Guindo-Martínez, M., Miguel-Escalada, I., Grarup, N., Sebastian, D., Rodriguez-Fos, E., Sánchez, F., Planas-Fèlix, M., Cortes-Sánchez, P., González, S., et al. (2018). Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. Nat. Commun. *9*, 321.

Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., van Dijk, F., et al. (2014). The Genome of the Netherlands: design, and project goals. Eur. J. Hum. Genet. *22*, 221–227.

Branham, W.S., Melvin, C.D., Han, T., Desai, V.G., Moland, C.L., Scully, A.T., and Fuscoe, J.C. (2007). Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. BMC Biotechnol. *7*, 8.

Bujold, D., Morais, D.A.L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K.C., Laperle, J., Markovits, A.N., Pastinen, T., et al. (2016). The International Human Epigenome Consortium Data Portal. Cell Syst. *3*, 496–499.e2.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (*D1*), D1005–D1012.

Burgarella, S., Merlo, S., Figliuzzi, M., and Remuzzi, A. (2013). Isolation of Langerhans islets by dielectrophoresis. Electrophoresis *34*, 1068–1075.

Carvalho, S., Molina-López, J., Parsons, D., Corpe, C., Maret, W., and Hogstrand, C. (2017). Differential cytolocation and functional assays of the two major human SLC30A8 (ZnT8) isoforms. J. Trace Elem. Med. Biol. *44*, 116–124.

Chen, J., Spracklen, C.N., Marenne, G., Varshney, A., Corbin, L.J., Luan, J., Willems, S.M., Wu, Y., Zhang, X., Horikoshi, M., et al.; Lifelines Cohort Study; Meta-Analysis of Glucose and Insulin-related Traits Consortium (MAGIC) (2021). The trans-ancestral genomic architecture of glycemic traits. Nat. Genet. *53*, 840–860.

Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. Nat. Genet. *32* (*Suppl*), 490–495.

Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. Nature *577*, 179–189.

Cnop, M., Abdulkarim, B., Bottu, G., Cunha, D.A., Igoillo-Esteve, M., Masini, M., Turatsinze, J.V., Griebel, T., Villate, O., Santin, I., et al. (2014). RNA

sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. Diabetes *63*, 1978–1993.

Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

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

Down, T.A., Piipari, M., and Hubbard, T.J.P. (2011). Dalliance: interactive genome viewing on the web. Bioinformatics *27*, 889–890.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Dwivedi, O.P., Lehtovirta, M., Hastoy, B., Chandra, V., Krentz, N.A.J., Kleiner, S., Jain, D., Richard, A.M., Abaitua, F., Beer, N.L., et al. (2019). Loss of ZnT8 function protects against diabetes by enhanced insulin secretion. Nat. Genet. *51*, 1596–1606.

Edsgärd, D., Iglesias, M.J., Reilly, S.-J., Hamsten, A., Tornvall, P., Odeberg, J., and Emanuelsson, O. (2016). GeneiASE: detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. Sci. Rep. *6*, 21134.

Eizirik, D.L., Pasquali, L., and Cnop, M. (2020). Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. Nat. Rev. Endocrinol. *16*, 349–362.

Fadista, J., Vikman, P., Laakso, E.O., Mollet, I.G., Esguerra, J.L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R.B., et al. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc. Natl. Acad. Sci. USA *111*, 13924–13929.

Fan, J., Hu, J., Xue, C., Zhang, H., Susztak, K., Reilly, M.P., Xiao, R., and Li, M. (2020). ASEP: gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. PLoS Genet. *16*, e1008786.

Fare, T.L., Coffey, E.M., Dai, H., He, Y.D., Kessler, D.A., Kilian, K.A., Koch, J.E., LeProust, E., Marton, M.J., Meyer, M.R., et al. (2003). Effects of atmospheric ozone on microarray data quality. Anal. Chem. *75*, 4672–4675.

Flannick, J. (2019). The Contribution of Low-Frequency and Rare Coding Variation to Susceptibility to Type 2 Diabetes. Curr. Diab. Rep. *19*, 25.

Flannick, J., Mercader, J.M., Fuchsberger, C., Udler, M.S., Mahajan, A., Wessel, J., Teslovich, T.M., Caulkins, L., Koesterer, R., Barajas-Olmos, F., et al.; Broad Genomics Platform; DiscovEHR Collaboration; CHARGE; LuCamp; ProDiGY; GoT2D; ESP; SIGMA-T2D; T2D-GENES; AMP-T2D-GENES (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature *570*, 71–76.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47* (*D1*), D766–D773.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat. Genet. *42*, 255–259.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

Graffelman, J. (2015). Exploring diallelic genetic markers: the HardyWeinberg package. J. Stat. Softw. *64*, 1–23.

Graffelman, J., and Camarena, J.M. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum. Hered. *65*, 77–84.

Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S., Puiggròs, M., Salvoro, C., Miguel-Escalada, I., Carey, C.E., Cole, J.B., Rüeger, S., Atkinson, E., et al.; FinnGen Consortium (2021). The impact of non-additive genetic associations on age-related complex diseases. Nat. Commun. *12*, 2436.

Gupta, M.K., and Vadde, R. (2020). Insights into the structure-function relationship of both wild and mutant zinc transporter ZnT8 in human: a computational structural biology approach. J. Biomol. Struct. Dyn. *38*, 137–151.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47* (*D1*), D853–D858.

Hall, E., Volkov, P., Dayeh, T., Esguerra, J.L.S., Salö, S., Eliasson, L., Rönn, T., Bacos, K., and Ling, C. (2014). Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. Genome Biol. *15*, 522.

Hodson, D.J., and Rorsman, P. (2020). A variation on the theme: SGLT2 inhibition and glucagon secretion in human islets. Diabetes *69*, 864–866.

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G.N., Geoghegan, J., Germino, G., et al. (2005). Multiple-laboratory comparison of microarray platforms. Nat. Methods *2*, 345–350.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. *48* (*D1*), D498–D503.

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

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kaur, S., Mirza, A.H., and Pociot, F. (2018). Cell type-selective expression of circular RNAs in human pancreatic islets. Noncoding RNA *4*, 38.

Khan, M.A.B., Hashim, M.J., King, J.K., Govender, R.D., Mustafa, H., and Al Kaabi, J. (2020). Epidemiology of type 2 diabetes - Global burden of disease and forecasted trends. J. Epidemiol. Glob. Health *10*, 107–111.

Krentz, N.A.J., and Gloyn, A.L. (2020). Insights into pancreatic islet cell dysfunction from type 2 diabetes mellitus genetics. Nat. Rev. Endocrinol. *16*, 202–212.

Kundu, K., Mann, A.L., Tardaguila, M., Watt, S., Ponstingl, H., Vasquez, L., Morrell, N.W., Stegle, O., Pastinen, T., Sawcer, S.J., et al. (2020). Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for twelve immune-mediated diseases. bioRxiv. https://doi.org/10.1101/2020.01.15.907436.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, L., Bai, S., and Sheline, C.T. (2017). HZnT8 (Slc30a8) transgenic mice that overexpress the R325W polymorph have reduced islet Zn2+ and proinsulin levels, increased glucose tolerance after a high-fat diet, and altered levels of pancreatic zinc binding proteins. Diabetes *66*, 551–559.

Liang, Y., Aguet, F., Barbeira, A.N., Ardlie, K., and Im, H.K. (2021). A scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction. Nat. Commun. *12*, 1424.

Liu, B., Gloudemans, M.J., Rao, A.S., Ingelsson, E., and Montgomery, S.B. (2019). Abundant associations with gene expression complicate GWAS follow-up. Nat. Genet. *51*, 768–769.

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

Loh, P.R., Palamara, P.F., and Price, A.L. (2016b). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Lyssenko, V., Nagorny, C.L.F., Erdos, M.R., Wierup, N., Jonsson, A., Spégel, P., Bugliani, M., Saxena, R., Fex, M., Pulizzi, N., et al. (2009). Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat. Genet. *41*, 82–88.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research *45*, 896–901.

Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C.Y., Prokopenko, I., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet. *46*, 234–244.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. *50*, 1505–1513.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.

Marselli, L., Piron, A., Suleiman, M., Colli, M.L., Yi, X., Khamis, A., Carrat, G.R., Rutter, G.A., Bugliani, M., Giusti, L., et al. (2020). Persistent or Transient Human β Cell Dysfunction Induced by Metabolic Stress: Specific Signatures and Shared Gene Expression with Type 2 Diabetes. Cell Rep. *33*, 108466.

Mayba, O., Gilbert, H.N., Liu, J., Haverty, P.M., Jhunjhunwala, S., Jiang, Z., Watanabe, C., and Zhang, Z. (2014). MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. *15*, 405.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Meier, D.T., Entrup, L., Templin, A.T., Hogan, M.F., Samarasekera, T., Zraika, S., Boyko, E.J., and Kahn, S.E. (2015). Determination of Optimal Sample Size for Quantification of β-Cell Area, Amyloid Area and β-Cell Apoptosis in Isolated Islets. J. Histochem. Cytochem. *63*, 663–673.

Mercader, J.M., and Florez, J.C. (2017). The Genetic Basis of Type 2 Diabetes in Hispanics and Latin Americans: Challenges and Opportunities. Front. Public Health *5*, 329.

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B.M., Rolando, D.M.Y., Farabella, I., Morgan, C.C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat. Genet. *51*, 1137–1148.

Newhall, R.A., Stouffer, S.A., Schuman, E.A., DeVinney, L.C., Star, S.A., and Williams, R.M. (1949). The American Soldier: Adjustment During Army Life. Volume I. Mississippi Val. Hist. Rev. *36*, 339.

Ng, S.Y., Bogu, G.K., Soh, B.S., and Stanton, L.W. (2013). The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. Mol. Cell *51*, 349–359.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44* (*D1*), D733–D745.

Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics *32*, 1479–1485.

Osonoi, S., Ichinohe, H., Kudo, K., Yagihashi, S., and Mizukami, H. (2020). 2047-P: Possible Implication of Cyclin D2 in Beta-Cell Proliferation of Human Perinatal Islet. Diabetes *69* (*Suppl 1*).

Pagès, H. (2015). SNPlocs.Hsapiens.dbSNP142.GRCh37: SNP locations for Homo sapiens (dbSNP Build 142). R package version 0.99.5. https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP142.GRCh37.html.

Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. *46*, 136–143.

Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research *45*, 833–839.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45* (*D1*), D833–D839.

Plotly Technologies (2015). Collaborative data science. https://plot.ly.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Ramos-Rodríguez, M., Raurell-Vila, H., Colli, M.L., Alvelos, M.I., Subirana-Granés, M., Juan-Mateu, J., Norris, R., Turatsinze, J.V., Nakayasu, E.S., Webb-Robertson, B.M., et al. (2019). The impact of proinflammatory cytokines on the β-cell regulatory landscape provides insights into the genetics of type 1 diabetes. Nat. Genet. *51*, 1588–1595.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A.A., Ogurtsova, K., et al. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res. Clin. Pract. *157*, 107843.

Saponaro, C., Acosta-Montalvo, A., Anguelova, L., Thevenet, J., Chiral, M., Pasquetti, G., Piron, A., Cnop, M., Gmyr, V., Prehn, J., et al. (2020a). 1900-P: HNF1A Deficiency Leads to Perturbed Glucagon Secretion in Humans. Diabetes *69* (*Suppl 1*).

Saponaro, C., Mühlemann, M., Acosta-Montalvo, A., Piron, A., Gmyr, V., Delalleau, N., Moerman, E., Thévenet, J., Pasquetti, G., Coddeville, A., et al. (2020b). Interindividual heterogeneity of SGLT2 expression and function in human pancreatic islets. Diabetes *69*, 902–914.

Satya, R.V., Zavaljevski, N., and Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-seq readmapping. Nucleic Acids Res. *40*, e127.

Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics *31*, 2601–2606.

Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes *66*, 2888–2902.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature *445*, 881–885.

Solimena, M., Schulte, A.M., Marselli, L., Ehehalt, F., Richter, D., Kleeberg, M., Mziaut, H., Knoch, K.-P., Parnis, J., Bugliani, M., et al. (2018). Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. Diabetologia *61*, 641–657.

Spracklen, C.N., Horikoshi, M., Kim, Y.J., Lin, K., Bragg, F., Moon, S., Suzuki, K., Tam, C.H.T., Tabara, Y., Kwak, S.H., et al. (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. Nature *582*, 240–245.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. *6*, e1000770.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat. Genet. *46*, 294–298.

Stitzel, M.L., Sethupathy, P., Pearson, D.S., Chines, P.S., Song, L., Erdos, M.R., Welch, R., Parker, S.C.J., Boyle, A.P., Scott, L.J., et al.; NISC Comparative Sequencing Program (2010). Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. Cell Metab. *12*, 443–455.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE *6*, e21800.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. *13*, 2129–2141.

Thomas, P.D., Kejariwal, A., Guo, N., Mi, H., Campbell, M.J., Muruganujan, A., and Lazareva-Ulitsky, B. (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res. *34*, W645–W650.

Thurner, M., van de Bunt, M., Torres, J.M., Mahajan, A., Nylander, V., Bennett, A.J., Gaulton, K.J., Barrett, A., Burrows, C., Bell, C.G., et al. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. eLife *7*, e31977.

van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., et al. (2015). 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. PLoS Genet. *11*, e1005694.

Viñuela, A., Varshney, A., van de Bunt, M., Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A.A., Erdos, M.R., Fadista, J., et al. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. Nat. Commun. *11*, 4912.

Vujkovic, M., Keaton, J.M., Lynch, J.A., Miller, D.R., Zhou, J., Tcheandjieu, C., Huffman, J.E., Assimes, T.L., Lorenz, K., Zhu, X., et al.; HPAP Consortium; Regeneron Genetics Center; VA Million Veteran Program (2020). 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. *52*, 680–691.

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

Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Wild, S., Roglic, G., Green, A., Sicree, R., and King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care *27*, 1047–1053.

Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.

Wu, D., Gu, J., and Zhang, M.Q. (2013). FastDMA: An Infinium HumanMethylation450 Beadchip Analyzer. Plos ONE *8*, e74275.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. *30*, e15.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48* (*D1*), D682–D688.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RNA-seq and genotyping array data (in this paper)	Marselli et al., 2020	EGA: EGAS00001005535
RNA-seq and genotyping array data	Fadista et al., 2014	GEO:GSE50244
RNA-seq and genotyping array data	van de Bunt et al., 2015	EGA:EGAD00001001601
RNA-seq data	Cnop et al., 2014	GEO:GSE53949
RNA-seq and genotyping array data	Akerman et al., 2017	EGA:EGAS00001002865
RNA-seq and genotyping array data	Miguel-Escalada et al., 2019; data not shown	EGA pending accession number
Expression array	Solimena et al., 2018	GEO:GSE76896
DNA-methylation	Hall et al., 2014	EGA:EGAD00001003946
Bisulphite sequencing	Thurner et al., 2018	EGA:EGAD00001003947
Cohesin	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
Mediator	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
H3K27ac	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
ATAC-seq	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
Islet regulome annotations, ChIP-seq and ATAC-seq processed files	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
Pancreatic islet enhancer clusters	Pasquali et al., 2014	
H3K4me1	Pasquali et al., 2014	
Long non-coding RNAs (lncRNAs) annotation	Akerman et al., 2017	
Pancreatic islet open chromatin DNase	Stitzel et al., 2010	ENCODE (2012-2016) Open Chromatine Dnase
Pancreatic islet open chromatin DNase	Gaulton et al., 2010	ENCODE (2012-2016) Open Chromatine Dnase
Glycemic traits data	MAGIC investigators (http://magicinvestigators.org.); members of MAGIC are provided in Appendix S1	
70KforT2D GWAS meta-analysis summary statistics	Bonàs-Guarch et al., 2018	http://cg.bsc.es/70kfort2d/
DIAGRAM 1000G GWAS meta-analysis Stage 1 Summary statistics	Scott et al., 2017	https://diagram-consortium.org/downloads.html
DIAGRAM Trans-ethnic T2D GWAS meta-analysis	Mahajan et al., 2014	https://diagram-consortium.org/downloads.html
DIAMANTE T2D GWAS meta-analysis	Mahajan et al., 2018	https://diagram-consortium.org/downloads.html
GTEx Analysis V7 - Transcript TPMs	GTEx Portal	https://www.gtexportal.org/home/
FastDMA probe full annotation	Wu et al., 2013	http://bioinfo.au.tsinghua.edu.cn/member/jgu/fastdma/
Gene Ontology	The Gene Ontology Consortium, 2017	http://geneontology.org/
Reactome	Reactome Pathway database	https://reactome.org/download-data/
DisGeNET, May 2017	Piñero et al., 2016	https://www.disgenet.org/
GWAS Catalog version 1.0 release 2021-06-08	MacArthur et al., 2017	https://www.ebi.ac.uk/gwas/downloads
Ensembl Variant Effect Predictor version 87.27	McLaren et al., 2016	https://m.ensembl.org/info/data/ftp/index.html
RefSeq BUILD.37.3	O'Leary et al., 2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/ARCHIVE/BUILD.37.3

(*Continued on next page*)

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Gencode v23 lift 37 annotation	Frankish et al., 2019	ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_23/GRCh37_mapping/gencode.v23lift37.annotation.gtf.gz
gnomAD version 2.0.2	gnomAD database	https://gnomad.broadinstitute.org/downloads

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Miriam Cnop (mcnop@ulb.ac.be)

Materials availability

This study did not generate new unique reagents.

Data and code availability

RNA-seq and genotyping array data from PISA cohort Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001005535.

Further information about EGA can be found on https://ega-archive.org "The European Genome-phenome Archive of human data consented for biomedical research"(https://www.nature.com/ng/journal/v47/n7/full/ng.3312.html).

RNA-seq and genotyping array data from CRG cohort should be requested through Miguel-Escalada et al. (2019) and coauthor Goutham Atla.

The eQTL and cASE results are available for browsing at TIGER (http://tiger.bsc.es), and the full summary statistics are available for download.

Source data and publicly available resources used for this study supporting all findings are detailed in the key resources table.

The cASE code is available through https://github.com/imoran-BSC/TIGER_cASE.

Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Islet sample collection and genotyping

TIGER data consist of 514 RNA-seq and 485 genotyped array data of deidentified cadaveric human pancreatic islet samples from five research centers: 1) Centre for Genomic Regulation, 2) Lund University Diabetes Centre, 3) University of Oxford/University of Alberta, 4) Department of Endocrinology and Metabolism, University of Pisa and 5) ULB Center for Diabetes Research, Universite Libre de Bruxelles (Table S1). For the latter two centers, islets are prepared from the body and tail of the pancreas.

Centre for Genomic Regulation (CRG)

The DNA of 127 CRG samples was isolated, sequenced, and genotyped using Illumina's Human OmniExpress 12 v1 and 2.5-8 v1.1 chips, as described in Miguel-Escalada et al. (2019). Genotype array was done in 125 samples with Illumina's Genome Studio software providing information on a total of 624k SNPs.

Lund University Diabetes Centre (Lund)

The DNA of 89 Lund samples from cadaver donors of European ancestry provided by the Nordic Islet Transplantation Programme was isolated as described in Fadista et al. (2014). The samples were genotyped using Illumina's HumanOmniExpress 12v1 C chips passing standard quality control metrics providing information on a total of 609k SNPs.

University of Oxford/University of Alberta (Oxford)

The DNA of 118 Oxford samples was isolated from either spleen or the exocrine fraction of the islet isolation using the Tissue DNA Purification Kit. When no other tissue was available, DNA was extracted from human islets using the Trizol fraction remaining after extraction of RNA as described in van de Bunt et al. (2015). The samples were genotyped using Illumina's Human Omni 2.5 exome array following the Illumina Infinium protocol providing information on a total of 2.5M SNPs.

University of Pisa (Pisa)

The DNA of 154 Pisa samples was isolated according to previously described in Marselli et al. (2020) and sequenced. Genotype calling was done in 153 samples with Illumina's Human Omni 2.5 exome array providing information on a total of 2.6M SNPs.

ULB Center for Diabetes Research (ULB)

The 43 ULB samples were isolated in Pisa using collagenase digestion and density gradient purification from beating-heart organ donors with no medical history of diabetes or metabolic disorders. Following islet shipment to Brussels, mRNA was extracted and processed following the RNeasy QIAGEN protocol as described in Cnop et al. (2014).

METHOD DETAILS

Genotyping quality control

PLINK v1.9 (Purcell et al., 2007) was used to do standard quality control of the genotype data, at the variant and sample level (Bonàs-Guarch et al., 2018). At the variant level, we discarded rare variants (Minor Allele Frequency MAF < 0.01) and applied Hardy-Weinberg equilibrium test filtering ($p \leq 1 \times 10^{-6}$) (Graffelman, 2015; Graffelman and Camarena, 2008). Further, we filtered the variants below a missingness threshold of 0.05. At the sample level, we discarded samples presenting a gender discordance between the reported gender in the metadata and the genetic sex, as well as the subjects with at least a 3rd degree of relatedness, those below a missingness threshold of 0.02 and, finally, individuals not clustering within the 4 standard deviations of the first four principal components from the multidimensional scale analysis. The ancestry of the individuals was assessed by principal components analysis comparisons with phase3 1000 Genomes Project populations (The 1000 Genomes Project Consortium, 2015).

After QC this resulted in a total of: 1) 103 individuals, 559,083 SNPs in the CRG cohort, 2) 88 individuals, 596,273 SNPs in the Lund cohort, 3) 102 individuals, 1,487,651 SNPs in the Oxford cohort and 4) 144 individuals, 1,542,765 SNPs in the Pisa cohort.

Genotype phasing and imputation

The autosomal genotypes were phased with Eaglev3 (Loh et al., 2016a, 2016b) using the Human Reference Consortium Project reference panel (McCarthy et al., 2016). The X chromosome was phased without reference panel with SHAPEIT (Delaneau et al., 2011). Then, GUIDANCE (Guindo-Martínez et al., 2021) integrating IMPUTE2 (Marchini et al., 2007) was used for imputation, using 4 reference panels: the 1000 Genomes Project phase 3 (The 1000 Genomes Project Consortium, 2015), the Genome of the Netherlands Project (Boomsma et al., 2014), the Haplotype Reference Consortium Project (McCarthy et al., 2016) and the UK10K Project (Walter et al., 2015), with an IMPUTE2 info score threshold of $\geq$ 0.7. This resulted in a total of 13.7-16.3M SNPs for each cohort separately, that were merged considering the best info score obtained across all panels, resulting in 22,983,795 genotyped and imputed genetic variants with MAF > 0.001.

RNA-seq read mapping

RNA from 514 human donor islet samples was isolated and purified, and was used to construct RNA-seq libraries. These bulk RNA-seq assays generated a total of > 72 billion pair-ended fragments of 75, 76, 100, 101, 125 bp read lengths.

To perform eQTL analysis, we aligned all samples against the transcriptome reference gencode.v23lift37 (Frankish et al., 2019) with STAR v2.4.0 (Dobin et al., 2013), using

- –paired-end –p 8

An alternative mapping strategy was used for RNA-seq read mapping to be used for cASE. Given that the standard reference genome contains only one allele in polymorphic sites, standard RNA-seq read mapping can produce reference-biased alignments, leading to false positives in the study of ASE. To align RNA-seq datasets in an allele unbiased manner, two modified reference genomes were built, defined as a 'masked' and an 'enhanced' genome. The 'masked' reference genome was built by substituting with an 'N' the nucleotide position of each common SNP in dbSNP142 (Pagès, 2015) (MAF > 1%), using the vcf2diploid.jar (Rozowsky et al., 2011) tool. To construct the 'enhanced' reference genome, we modified the scripts developed by Satya et al. (2012) to accommodate RNA-seq reads, which added artificial contigs to the reference genome containing all possible SNP allele combinations. For this step, we used the subset of 4M common SNPs located within gene coordinates in the Ensembl (Yates et al., 2020), RefSeq (O'Leary et al., 2016) and UCSC (Haeussler et al., 2019) annotations, or within previously identified human islet lncRNAs (Akerman et al., 2017) (Figure S5).

STAR v2.2.0 (Dobin et al., 2013) was used to align the RNA-seq datasets against the masked genome, using

- –outFilterMultimapNmax 1–outFilterMismatchNmax 10
- –outSAMstrandField intronMotif–outSAMattributes All

in order to allow up to 10% of nucleotide mismatches, suppress multimapped reads, and make the output compatible with downstream software. Bowtie v2.0.5 (Langmead et al., 2009) was used to align the RNA-seq data against the enhanced genome, using

- –n-ceil L,0,0.03–score-min C,-14,0 -N 1 -X 50000

to allow up to 3 nucleotide mismatches evenly distributed within the read, and long range read pairs. Bowtie2 (Langmead and Salzberg, 2012) was chosen because it does not map the RNA-seq spliced reads, (only the reference allele-containing spliced sequences were present in the enhanced genome) which prevents the generation of allelic alignment bias.

After mapping the RNA-seq datasets to the two modified reference genomes, the outputs of both alignments were combined into one non-redundant set of reads, using the read merging C++ scripts available in our github repository (https://github.com/imoran-BSC/TIGER_cASE, scripts 02 and 03). Reads that aligned to the same genomic positions by both methods were kept, as well as reads mapped only by one of the two methods. In addition, all reads that mapped partially to intronic regions were discarded. The resulting set of reads was named 'unbiased alignment' (Figure S5A). This method successfully eliminated alignment bias in heterozygous positions (Figure S5B), and mapped 86.2% of all RNA-seq reads. When comparing this alignment with one using the standard reference genome and STAR v2.2.0 using a subset of the samples, we recovered an extra 8.5% more reads using the unbiased alignment method (Figure S5C).

Sample concordance verification between genotype and gene expression

To avoid mislabeled samples leading to mismatching errors between genotype-phenotype samples, and to discard samples with poor quality or possible contamination, we used verifyBamID v1.1.3 (Jun et al., 2012) with ''–best,'' applied to the RNA-seq alignments sorted and indexed with samtools v1.1 (Li et al., 2009), and comparing with their genotypes. After these steps, 404 samples with good quality genotype and RNA-seq data and concordance remained for further analysis.

TIGER web portal development

The TIGER web portal (http://tiger.bsc.es) is the comprehensive integration in an ElasticSearch v1.4.4 database of a) T2D GWAS variants identified in 70KforT2D (Bonàs-Guarch et al., 2018), diagram DIAMANTE (Mahajan et al., 2018), diagram Trans-ethnic (Mahajan et al., 2014), diagram 1000G (Scott et al., 2017) T2D meta-analyses or included in the GWAS Catalog v1 release 2021-06-08 (Buniello et al., 2019), b) variant annotation and characterization through Variant Effect Predictor v87.27 (McLaren et al., 2016) and Gnomad v2.0.2 (Karczewski et al., 2020), c) epigenomic marks from islet DNA-methylation sites (Hall et al., 2014; Thurner et al., 2018), chromatin accessibility (Dunham et al., 2012; Gaulton et al., 2010; Stitzel et al., 2010) and CHiP-seq profiles (Miguel-Escalada et al., 2019), d) annotation from Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), lncRNAs (Akerman et al., 2017) and islet regulome (Miguel-Escalada et al., 2019; Pasquali et al., 2014) in a publicly available platform. Genes are referenced to Gencode annotation v23 lift 37 (Frankish et al., 2019) and RefSeq BUILD.37.3 (O'Leary et al., 2016) and enriched with DisGeNET (Piñero et al., 2017) (May 2017) and Reactome Pathway (Jassal et al., 2020) database information. It contains results on gene expression integrating the results of a) gene expression from normalized islet RNA-seq counts, microarrays (Solimena et al., 2018), and the Genotype-Tissue Expression database (GTEx) (Lonsdale et al., 2013), and b) computed eQTL and cASE.

The portal was built upon [ICGC software codebase], the front-end coded in angular v1.5.7 with embedded biodalliance v1.4.4 genomic browser (Down et al., 2011), plotly v1.54.1 (Plotly Technologies, 2015) and highcharts libraries and the back-end coded in Java.

QUANTIFICATION AND STATISTICAL ANALYSIS

eQTL analysis

The *cis*-eQTL analysis of 404 human pancreatic islets for which both RNA-seq and genotyping data remained after QC was performed by cohort with fastQTL v2.0 tool (Ongen et al., 2016). The analysis was run for regions one million base pairs up- or downstream of the transcription start site of each gene using *gencode.v23lift37* (Frankish et al., 2019) version. For each cohort, we corrected for known covariates (age, sex and BMI), 7 genomic ancestry principal components, and 15 PEER v1.3 (Stegle et al., 2010) factors in order to account for hidden confounding factors. For the X chromosome, we used 5 PEER factors and 4 genomic ancestry principal components and the *cis*-eQTL analysis was performed stratified by sex and combined. The full command for *fastQTL* is

```
fastQTL–log 'chr1.log'–vcf. 'chr1.bcf.'–bed 'rsem.bed' -C 'covariates.tsv'–threshold '0.01'–out 'chr1.fastQTL.gz'
```

Age and BMI missing metadata were imputed using the cohort mean.

The by-cohort fastQTL (Ongen et al., 2016) results were then meta-analyzed with METAL (Willer et al., 2010) using the sample size strategy and computing heterogeneity. For the X chromosome, the meta-analysis was run over the 4 cohorts for both sexes together and over the 8 eQTL analysis (4 cohorts, 2 sexes). The full configuration files for METAL are given by:

```
SEPARATOR WHITESPACE
MARKER ensg.snp
ALLELE a0 a1
EFFECT slope
PVALUE pval
```

```
WEIGHT N
PROCESS cohort_CRG
PROCESS cohort_OXFORD
PROCESS cohort_LUND
PROCESS cohort_PISA
OUTFILE metal .tsv
ANALYZE HETEROGENEITY
QUIT
```

Identifying variant regulatory enrichments using GREGOR

To test the eQTL and cASE variants for enrichment in islet regulatory overlaps, we used the Genomic Regulatory Elements and Gwas Overlap algoRithm (GREGOR) (Schmidt et al., 2015), designed to calculate such enrichment while controlling for linkage-disequilibrium between variants, MAF and distance to nearest gene. We used the 1% and 5% FDR set of significant eQTL variants, after selecting them by linkage disequilibrium < 0.2 using PLINKv1.9 (Purcell et al., 2007) with "–indep-pairwise 100k 5 0.2". We tested enrichment against a set of human islet regulatory regions, including gene promoters, enhancers, and open-chromatin derived from ChIP-seq experiments in human islets (Figures 2C and S1) (Miguel-Escalada et al., 2019). Specifically, we used an R^2 threshold of 0.99, a window size of 1,000,000, a min_neighbor_num of 500, and European (EUR) as the population.

Comparison of TIGER eQTLs with the GTEx and InsPIRE datasets

To assess the degree of concordance between the TIGER significant eQTLs and those reported in the GTEx v8 dataset (Aguet et al., 2020), we searched for exact variant-target gene matches among the dataset of significant eQTLs in all 54 GTEx tissues. To analyze the overlap of eQTLs with low-frequency variants, we repeated the analysis, but first filtered the TIGER and GTEx eQTLs to include only those with variants with a MAF < 0.05 in the EUR population of the 1000 genomes phase-3 dataset (The 1000 Genomes Project Consortium, 2015).

To obtain a relevant comparison with the InsPIRE (Viñuela et al., 2020) dataset, we first applied the same multiple-testing correction method used in this study to the full nominal p values of the InsPIRE dataset. The Benjamini-Hochberg corrections for 1 and 5% FDR resulted in the nominal *p-value* thresholds of $p = 8.55 \times 10^{-5}$ and $p = 6.2 \times 10^{-4}$, corresponding to 974,435 and 1,408,891 significant eQTLs. Two eQTLs were considered significant by both methods if they were detected at < 5% FDR in both studies, and had an exact match in both variant and target gene. The low-frequency variant eQTLs were determined as described above.

Colocalization analysis

COLOC 4.0 (Giambartolomei et al., 2014) R package was used for the colocalization analysis of *cis*-eQTL and T2D GWAS. We used the coloc.abf method which implements a variation of the Approximate Bayes Factor computations (Wakefield, 2009). The coloc.abf function was called with two R lists, one for the eQTL and one for the GWAS:

```
list(pvalues = …, N = …, MAF = …, snp = …, type = "quant")
```

with a vector of *p*-values, N the sample size, MAF the minor allele frequency and snp the rsid of the variant.

In order to select regions for colocalization analyses, we selected genes associated with at least one significant eQTL SNP which had been previously reported as a GWAS lead variant (Bonàs-Guarch et al., 2018; Mahajan et al., 2018; Vujkovic et al., 2020). The significant eQTL SNPs were determined based on a 0.05 threshold Benjamini-Hochberg FDR (Benjamini and Hochberg, 1995). Similarly, we used the *p*-values of the cASE analysis to perform colocalization, considering loci with an at least 5% FDR significant signal. The colocalization was run over regions ranging from one million base pairs downstream to one million upstream of the cis-regulatory target gene transcription start site.

The colocalization plots were generated by the locuscompare R package v1.0.0 (Liu et al., 2019) (Data S1 and S2).

Generation of an unbiased set of ASE reporter variants

To identify loci under mappability related allelic biases, a C++ script available in the github repository (https://github.com/imoran-BSC/TIGER_cASE, script 01) was used to generate all possible reads containing both alleles of all possible reporter SNPs. A splice junction database was created using the Ensembl (Yates et al., 2020), RefSeq (O'Leary et al., 2016), UCSC (Haeussler et al., 2019) and human islet lncRNA (Akerman et al., 2017) gene annotations, to take splice junctions into account.

The resulting dataset, consisting of 240M artificial reads, was aligned using the unbiased mapping strategy described above, and the allelic ratios (i.e., the percentage of reference-allele carrying reads) were quantified. Since the same number of reads were purposely generated carrying both alleles, any observed allelic imbalance would derive exclusively from mapping biases. SNPs whose allelic ratio was not between 49%–51% were blacklisted. Additionally, all SNPs located within 100 bps of a common or low-frequency indel present in dbSNP142 (Pagès, 2015) were also blacklisted.

The remaining curated set of 3.97M SNPs were used as bona-fide SNPs for reporting ASE.

Identification of ASE

The number of reads containing the reference and alternate alleles RNA-seq reads overlapping each reporter SNP were quantified using the mpileup command of samtools v1.1 (Li et al., 2009), with the flags "-A -B -d 20000", and the ComputePileupFreqs.pl script (Satya et al., 2012). Sample-specific ASE was assessed calculating the allelic ratio, i.e., the fraction of reads containing the reference allele over the total number of reads. We selected the set of SNPs with at least 3 heterozygous samples with $\geq$ 15 RNA-seq reads (of which $\geq$ 10 non-clonal), resulting in a set of > 170k informative reporter SNPs.

A binomial test (Bernoulli, 1899) was used to assess the significance of ASE for all reporter SNPs, using the number of reads carrying the reference and alternate alleles. To account for any possible remaining alignment bias in the datasets, the median allelic ratio for each possible bi-allelic SNP (AC, AG, AT, CG, CT, GT) across the genome was calculated and used as null, instead of the theoretical 50%. Similarly, the allelic ratios were proportionally adjusted using the sample and nucleotide-pair specific median value.

The resulting *p-value*s were used to calculate a sample-specific 1% and 5% FDR Benjamini-Hochberg (Benjamini and Hochberg, 1995) thresholds, to correct for multiple testing.

Assessing cASE using Stouffer's *Z*-score

To assess cASE in a given heterozygous variant in many independent samples, the Stouffer's Z-score (Newhall et al., 1949) method was used. This method combines independently obtained *p*-values into a Z statistic, which increases in absolute value with significance. The method allows for weighting of independent *p*-values and, additionally, it accounts for a positive or negative direction in the magnitude associated with the *p*-values. Thus, this method allows to differentiate between significant reference and alternate reporter variants, as well as providing a way to account for the variance inherent to differing numbers of informative RNA-seq reads in each reporter.

For each reporter, a Z-score was calculated as follows:

$$Z = \frac{\sum w_i Z_i}{\sqrt{\sum w_i^2}}$$

where w_i was the total read coverage of sample i, and Z_i was the transformed binomial *p*-value p_i:

$$Z_i = \pm \theta^{-1}\left(1 - \frac{p_i}{2}\right)$$

where the sign was positive if the value of the allelic ratio was > 50%, zero if exactly 50%, and negative otherwise, and θ^{-1} was the inverse of the standard normal cumulative distribution function, calculated using the qnorm function in R. A threshold of 10^{-15} was imposed as the minimum possible binomial *p*-value, in order to prevent single events with very significant *p*-values from dominating the Z-score value, while still maintaining their relevance. Therefore, Stouffer's Z-score (Newhall et al., 1949) method accounted for consistency in the overall reference or alternate direction of the allelic bias across samples, and considered all *p*-values into account, regardless of their sample-specific significance.

Z-scores were only calculated if the reporter SNP was heterozygous in 3 or more samples, and only samples with a read coverage of $\geq$ 15 RNA-seq reads, of which $\geq$ 10 non-clonal, were used in the calculation.

Assessing the significance of cASE *Z*-scores

To assess the significance of the obtained Z-scores, we performed 1,000 permutations of the reference/alternate read counts between heterozygous SNPs, and calculated their binomial *p*-values and resulting control Z-scores (https://github.com/imoran-BSC/TIGER_cASE, script 04). To account for the differences in gene expression, all reporter SNPs were distributed in 5 bins: one containing all SNPs with a median coverage of 0 reads, and 4 more bins containing the remaining SNPs according to their read coverage quartile, and the read counts of heterozygous SNPs were only shuffled within their bins. By permuting only the values of the heterozygous SNPs while keeping the reference and alternate homozygous values invariant, the distribution of the number of samples in heterozygosity for each SNP was kept constant.

The resulting null distribution of Z-scores was therefore attributable only to stochasticity, and so for each empiric Z-score, a *p*-value was calculated from this null distribution. The Benjamini-Hochberg method (Benjamini and Hochberg, 1995) was then used to obtain *q*-values from these *p*-values and thus correct for multiple testing.

Regulatory enrichment of cASE significant genes

To calculate these regulatory enrichments, we first generated a null distribution of control genes that were non-significant for cASE but had similar expression levels. First, we separated the cASE significant genes in 4 bins of expression, and randomly selected the same number of non-significant genes of the same expression quartile, 1,000 times. We then calculated, in the 1% and 5% FDR cASE genes and in each of the 1,000 control sets, the proportion of genes that were in the islet-specifically expressed genes list (Miguel-Escalada et al., 2019) (Figure 4C, left). The same procedure was performed to calculate the enrichment for proximity to islet enhancers, by calculating the proportion of genes located at less than 25kb from islet enhancers (Miguel-Escalada et al., 2019). The *p*-values were obtained by approximating these permuted control distributions as Gaussian distributions and deriving a *p*-value using the pnorm R function.

Gene ontology analyses and islet-specific expression

Gene ontology terms in the analyses of eQTL and cASE genes were obtained using the PANTHER (Protein ANalysis THrough Evolutionary Relationships) (Thomas et al., 2003, 2006) classification system.

For eQTL, we analyzed all 5% FDR significant genes versus a background list of all genes expressed in islets (Figure S3), and the list of TIGER exclusive eQTL genes versus a background of all eQTL genes shared with GTEx (Figure 2E).

For cASE, we studied 5% FDR cASE genes versus a background dataset of all genes for which the calculated cASE was non-significant (Figure 4D). The visualization of the syntactic terms was obtained using the REVIGO web tool (Supek et al., 2011).

Identifying candidate SNPs putatively leading to cASE

We aimed to characterize the set of SNPs putatively causal of cASE (referred to as 'candidate SNPs'). To that end, we first identified all variant pairs consisting of a cASE-significant reporter and a candidate variant, as long as both were located within the same topologically associating domain (TAD) (Dixon et al., 2012), plus a boundary leeway of ± 200kbs. Then, we separated the samples using the candidate variant genotype in two groups: those heterozygous (Het), and those homozygous (Hom). Finally, we calculated the reporter Z-score of both sample groups, and selected the candidate variants with significant Z-scores for the Het individuals, which were also non-significant for the Homs (https://github.com/imoran-BSC/TIGER_cASE, script). The underlying hypothesis was that if the candidate variant was homozygous, it was unlikely to be causal.

Putative causal variants were also interrogated for the set of non-cASE significant reporter variants, following the same procedure described above. This produced an additional 1,247 genes that reached cASE significance only after being considered with these putative causal variants.

Scaling human islet gene expression values to allow comparisons with the GTEx expression datasets in TIGER

The RNA-seq expression of human islet samples was measured with RSEM v1.3.0 (Li and Dewey, 2011) in 60,261 transcripts from Gencode database (v23lift37 annotation) (Frankish et al., 2019) using STAR v2.5.3.a (Dobin et al., 2013) and BOWTIE v2.3.2 (Langmead and Salzberg, 2012) hg19 aligned-reads as follows:

```
STAR–runMode genomeGenerate–genomeFastaFiles GRCh37.primary_assembly.genome.fa–sjdbGTFfile gencode.v23lift37.annotation.gtf
rsem-prepare-reference–gtf gencode.v23lift37.annotation.gtf–bowtie2 GRCh37.primary_assembly.genome.fa
rsem-calculate-expression–paired-end–star–paired-end -p 8
```

We obtained measures of raw counts, counts normalized by transcript length (TPM - transcripts per million) and fragment length (FPKM - fragments per kilobase). The batch effects and covariate differences between samples captured in the TPM measures were removed with limma removeBatchEffect function (Ritchie et al., 2015), using the log10 normalized expression of the genes that were expressed in at least 80% of human islet samples. The results of this normalization were evaluated with Spearman correlation, ensuring that there was a correlation above 0.8 between all the samples independently of the cohort after correction.

TPM expression datasets from the 54 tissues available in GTEx (Lonsdale et al., 2013) (20 samples per tissue) were collected, and a decile distribution analysis was performed excluding genes from GTEx samples that miss expression in at least 50% of the samples. Then, TIGER islet expression was scaled to fit these measures according to the following criteria:

(1) Each GTEx decile bin $[D_{G;i},D_{G;i+1}]$ has TPM values in $[T_{G;i},T_{G;i+1}]$, thus the corresponding decilic straight will be: $\boldsymbol{y_G} = (\boldsymbol{T_{G;i+1}} - \boldsymbol{T_{G;i}})\boldsymbol{x} + \boldsymbol{T_{G;i}}$.

(2) Each pancreatic islet decile bin $[D_{G;i},D_{G;i+1}]$ has TPM values in $[T_{PI;i},T_{PI;i+1}]$, thus the corresponding decilic straight will be: $\boldsymbol{y_{PI}} = (\boldsymbol{T_{PI;i+1}} - \boldsymbol{T_{PI;i}})\boldsymbol{x} + \boldsymbol{T_{PI;i}}$.

From Equation (2) one can derive: $x = \frac{y_{PI} - T_{PI;i}}{T_{PI;i+1} - T_{PI;i}}$ (3) thus, allowing the relation between the TPM pancreatic islet values y_{PI} and the TPM GTEx values y_G by replacing (3) in (1): $y_G = \left(\frac{T_{G;i+1} - T_{G;i}}{T_{PI;i+1} - T_{PI;i}}\right) y_{PI} - T_{PI;i}\left(\frac{T_{G;i+1} - T_{G;i}}{T_{PI;i+1} - T_{PI;i}}\right) + T_{G;i}$ the scaling factor.