Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D

Ana Viñuela et al.

Most signals detected by genome-wide association studies map to non-coding sequence and their tissue-specific effects influence transcriptional regulation. However, key tissues and cell-types required for functional inference are absent from large-scale resources. Here we explore the relationship between genetic variants influencing predisposition to type 2 diabetes (T2D) and related glycemic traits, and human pancreatic islet transcription using data from 420 donors. We find: (a) 7741 cis-eQTLs in islets with a replication rate across 44 GTEx tissues between 40% and 73%; (b) marked overlap between islet cis-eQTL signals and active regulatory sequences in islets, with reduced eQTL effect size observed in the stretch enhancers most strongly implicated in GWAS signal location; (c) enrichment of islet cis-eQTL signals with T2D risk variants identified in genome-wide association studies; and (d) colocalization between 47 islet cis-eQTLs and variants influencing T2D or glycemic traits, including DGKB and TCF7L2. Our findings illustrate the advantages of performing functional and regulatory studies in disease relevant tissues.
**Results**

**Characterization of genetic regulation of gene expression in islets.** We combined islet RNA-Seq with dense genome-wide genotype data from 420 individuals. Data from 196 of these individuals have been reported previously. We aggregated, and then jointly mapped and reprocessed, all samples (median sequence-depth per sample ~60 M reads) to generate exon- and gene-level quantifications, using principal component (PC) methods to correct for technical and batch variation. We detected 6039 eGenes (FDR < 1%) with at least two independent exon-eSNPs, 1783 cis-eQTLs in 1291 of the eGenes (21.3% of all eGenes), giving a total of 7741 eQTLs (Supplementary Data 1–4). At the 1291 eGenes with at least two independent exon-eSNPs, most primary eSNPs mapped closer to the canonical TSS than secondary eSNPs (Wilcoxon test P = 6.3 × 10^-30), however, there were 503 (39.0%) of these genes for which the second eSNP identified during stepwise conditional analysis was more proximal to the TSS (Supplementary Fig. 2).

**Tissue-specific regulatory variation in islets.** For many complex traits of biomedical interest, the value of targeting the specific cell types of interest for dedicated eQTL discovery—as opposed to relying on existing eQTL data from more accessible tissues—remains unclear. To examine this, we considered the degree to which the set of 7741 exon-level islet eQTLs overlapped eQTLs detected in 44 tissues (N > 70) from version 6p of GTEx. To allow direct comparison with InsPIRE, we reprocessed GTEx data to generate exon-level eQTLs ("Methods" section). Of the 6039 islet eGenes, 5% (337) had no significant eQTLs, in either exon- or gene-level analyses in any of the 44 tissues examined in GTEx (Supplementary Data 5). We used p-value enrichment analysis (pi) to measure the proportion of islet eQTLs shared with other GTEx tissues, generating estimates ranging from 0.40 (hypothalamus) to 0.73 (adipose). By comparing the p-value distributions across tissues, this analysis can detect evidence of sharing which does not depend on arbitrary statistical thresholds. We detected the expected positive linear relationship between pi measures and sample sizes for the respective tissues in GTEx (Fig. 1a). However, pi enrichment only reached 0.65 and 0.57 (respectively) for skeletal muscle (n = 361), and whole blood (n = 338), the tissues with the largest representation in this version of GTEx (see "Influence of sample size" in Supplementary Discussion). Whole pancreas, often naively used as a surrogate for the T2D-relevant islet component, represented an imperfect proxy for islet (pi = 0.65 with islets). This is not purely a consequence of low sample sizes: tissues with slightly larger sample sizes showed both higher (pi = 0.71, artery) and lower sharing of eQTLs (pi = 0.64, skin no sun exposed). These data demonstrate that there is a component of tissue-specific genetic regulation that could, at these sample sizes, only be detected in islet, illustrating the value of extending current expression profiling efforts to additional tissues and cell types of particular biomedical importance.

**Cellular heterogeneity.** The human islets analyzed in this, and other, studies include a mixture of cell types, including the hormone-producing α, β, and δ-cells, and a variable amount of adherent exocrine material. From the perspective of T2D pathogenesis, the transcriptomes of the endocrine compartment are of most interest. However, the eQTLs identified could have their origins from any of the cellular components. We used a number of approaches to address interpretative challenges resulting from this cellular heterogeneity.
First, we performed tissue deconvolution analysis to estimate the proportion of exocrine contamination: these analyses were performed prior to the PC adjustment used to generate the main results and used reference expression signatures for exocrine pancreas, β-cells, and islet non-β-cells. The last two of these were generated from a subset (n = 26) of the islet preparations after FAC sorting using the zinc-binding dye Newport Green18 ("Methods" section and in Supplementary Discussion). Estimates of the proportion of exocrine pancreas contamination ranged from 1.8 to 91.8% (median 33.5%): these were significantly correlated (Spearman correlation, \( p = 0.406, P = 1.4 \times 10^{-11} \)) with independent estimates of exocrine content obtained at islet collection by dithizone staining (n = 232; Supplementary Fig. 3). Within the islet endocrine fraction, median estimates of β-cell (58.8%, IQR 43.2–66.9%) and non-β-cell (41.2%, 33.1–56.8%) fractions are in agreement with estimates from morphometric assessment23. In 37 samples from donors annotated as having T2D, median estimates of β-cell composition were lower than those from non-diabetic donors (n = 330; linear model, \( P = 3.3 \times 10^{-2} \), Supplementary Fig. 3). This provides independent support, based on transcriptomic signatures, of evidence, from morphometric and physiological studies, that the functional mass of β-cells is reduced in T2D24,25.

Of the 420 InsPIRE samples, β-cell-enriched transcriptomes were available for 26 following FAC sorting. With this limited sample size, the only eQTL reaching significance, and then only at a less stringent threshold of FDR < 5% (Supplementary Data 6) was at ADORA2B (\( P = 3.8 \times 10^{-10}, \text{slope} \equiv -1.20 \)): this signal was also detected in InsPIRE islets (\( P = 3.9 \times 10^{-31}, \text{slope} \equiv -0.65 \)) and GTEx pancreas (\( P = 1.6 \times 10^{-16}, \text{slope} \equiv -0.73 \); Supplementary Fig. 4 and Supplementary Data 7). By comparing the \( p \)-value distributions of the eQTLs in islets vs β-cells22, we estimate that 46% (\( \tau_{\beta} = 0.46 \)) of islet eQTLs are active in β-cells (Fig. 1b). By reevaluating significance for eQTLs in β-cell association results using the 7741 independent significant SNP-exon pairs, we were able to replicate 227 islet eQTLs in β-cells (FDR < 1%, Supplementary Data 8). Genes with cell-type-specific regulatory effects were sought by testing for interactions between genotype and cellular fraction estimates, controlling for technical variables ("Methods" section). We identified 18 islet cis-eQTLs with a genotype-by-β-cell proportion interaction and eight with a genotype-by-exocrine cell proportion interaction (FDR < 1%, Supplementary Data 9, 10 and 11).

We conclude that a substantial proportion of the regulation of gene expression detected in pancreatic islets derives from cell-type-specific effects. Ongoing efforts to develop a single-cell view of islet transcriptional signatures should inform these analyses, although the limited sample size of current studies26–30 and the paucity of genotype information means they offer little direct insight into the relationship between genetic variation and cell-type-specific transcript abundance.

**Functional properties of islet genetic regulatory signals.** Using previously published islet chromatin states derived from histone modification data15, we observed a significant enrichment of islet eSNPs in active islet chromatin states, including active TSS (fold enrichment = 3.84, \( P = 5.5 \times 10^{-206} \)), active enhancers (fold enrichment > 1.73, \( P < 4.8 \times 10^{-94} \) between active/inactive enhancer states), and stretch enhancers (fold enrichment = 1.57, \( P = 2.7 \times 10^{-13} \)), with concomitant depletion of eSNPs in repressed and quiescent states (fold enrichment < 0.66; Supplementary Fig. 5 and Supplementary Table 1). This recapitulates the enrichment observed for T2D GWAS signals within active islet chromatin (Supplementary Fig. 6 and Supplementary Table 2)11,13,31,32. Next, we examined the relationship between the chromatin context of islet eSNPs and their effect sizes (Fig. 2a): not only that the active TSS chromatin states showed high enrichment of eSNPs, but that...
the eSNPs that overlap active TSS chromatin states had larger effects than those in repressed or weak-repressed polycomb states (Wilcoxon rank-sum test \( P = 3.9 \times 10^{-2} \)).

A major portion (65.7%) of the islet-active TSS chromatin state territory is occupied by islet ATAC-seq peaks. When we focused the analysis on eSNPs within islet ATAC-seq peaks (Supplementary Fig. 7 and Supplementary Table 3), those within stretch enhancers (islet-specific enhancer chromatin state segments >3 kb (ref. 29)) had smaller effects than those in active TSSs (Wilcoxon rank-sum test \( P = 3.4 \times 10^{-3} \); Fig. 2b). Given the inverse relationship between eQTL effect size and the number of samples required to identify significant association, one corollary is that eSNPs in tissue-specific stretch enhancers, which have smaller effect sizes compared to those in more ubiquitous TSSs, are likely to require larger sample sizes for eQTL discovery.

We previously reported enrichment of selected TF footprint motifs at islet eSNPs\(^2\). Here, with a larger eSNP catalog, we sought to determine how eSNP effect size and target gene expression directionality is associated with base-specific TF-binding preferences. Using published TF footprint data (in vivo-predicted TF motif binding) from human islet ATAC-seq analyses\(^3\), we partitioned eSNPs into two equally sized bins (absolute slope \( \geq \) 3; Fig. 2c and Supplementary Data 12).
assay (MPRA) data from HepG2 and K562 cell lines, whether eSNP alleles matching the base preference at TF footprint motifs have a consistent directional impact on gene expression. We defined a motif directionality fraction score (ranging from repressive [0] to activating [1]) for each TF footprint motif ("Methods" section). Of the 109 motifs reported as consistently activating or repressive across HepG2 and K562 cell lines that were also present in our study, 16.5% (n = 18) showed skewed activator preference in islets (<10% FDR; Fig. 2d, Supplementary Fig. 8 and Supplementary Table 4). The activator motifs we identified include many ETS family members that have a known preference for transcriptional activation.

Our analyses demonstrate the value of contrasting tissue-specific stretch enhancers with more ubiquitous TSS states to delineate the role of underlying chromatin on function, and illustrate how the integration of eQTL information with ATAC-seq and high-resolution TF footprinting reveals the in vivo activities of these upstream regulators.

Islet eQTLs are enriched among T2D and glycemic GWAS variants. Diverse lines of evidence emphasize the contribution of reduced pancreatic islet function to the development of T2D, with many T2D GWAS loci acting primarily through reductions in insulin secretion. To examine the relationships between tissue-specific regulation of gene expression and T2D predisposition alleles, we focused on 403 lead GWAS SNPs with the strongest associations to T2D in Europeans (as reported in Mahajan et al.) and 56 variants significantly associated with T2D-relevant continuous glycemic traits, including fasting glucose and β-cell function (HOMA-B) in nondiabetic individuals (Supplementary Data 13). For comparison, we included 55 GWAS variants implicated in T1D predisposition. To determine the extent to which the GWAS variants were selectively enriched for islet eQTL associations, we extracted exon-level eQTL information for each of these variants from InSIRE and the 44 GTEx tissues. We compared observed effect size estimates to those derived from a null distribution of 15,000 random eSNPs, matched to the GWAS SNPs with respect to the number of SNPs in linkage disequilibrium (LD), distance to TSS, number of nearby genes, and minor allele frequency (MAF; "Methods" section). Figure 3 shows the enrichment in eQTL effect sizes at T2D/glycemic GWAS-associated variants for five tissues implicated in T2D pathogenesis (subcutaneous adipose tissue, skeletal muscle, liver, islets, and plus hypothalamus), with pancreas and whole blood for comparison.

We detected nominally significant enrichment for islet eQTLs for variants associated with continuous glycemic traits (normalized enrichment score (NES) = 1.27; P = 3.7 × 10^-3; Supplementary Fig. 9 and Supplementary Data 14): of all the tissues considered, islets generated the most significant enrichment for this phenotype. Islet cis-eQTLs were also enriched amongst the 403 T2D variants (NES = 1.09; p = 5.3 × 10^-3), as were eQTLs from pancreas, skeletal muscle, adipose, and eight other GTEx tissues. In a subset of 43 variants (from the 403) considered to be mediated through defects in insulin secretion (thereby implicating islet dysfunction), the degree of enrichment was increased (NES = 1.15, p = 3 × 10^-2). No evidence of enrichment of islet eQTL signals was seen for T1D-risk variants (NES = 1.01; p = 0.47), consistent with the consensus that genetic risk for T1D is largely mediated through immune mechanisms. These data reveal tissue-specific patterns of genetic regulatory impact for variants at T2D- and glycemic-trait loci similar to the mechanistic inferences generated by physiological analysis of those signals. They also highlight the importance of matching the tissue origin of the transcriptomic data used for mechanistic inference to the tissue-specific impact of each GWAS signal on disease predisposition.

Identifying effector transcripts for T2D and glycemic traits. Previous studies have identified 27 GWAS signals displaying apparent overlap between islet eQTLs and the T2D/glycemic GWAS signals, but not all of these signals have been evaluated with respect to the statistical evidence for colocalization and not all coincident signals have replicated despite ostensibly similar designs and power. Here, we took the opportunity offered by increases in the sample sizes available for both T2D GWAS and islet cis-eQTL analyses, to undertake a systematic reanalysis to identify effector transcripts mediating the activity of T2D and glycemic traits GWAS variants. There are multiple methods for evaluating the evidence for colocalization, but these make different assumptions and often lead to discrepant results. We focused on the colocalization evidence provided by two complementary algorithms: COLOC, which assesses differences in regression coefficients of variants associated to two traits, and RTC, which assesses the differences in ranking of SNPs associated with one trait after conditioning on the most associated SNP for the other.

We detected evidence for colocalization (using either method) for islet eQTLs at 46 GWAS loci (47 independent signals, Supplementary Figs. 10 and 11). Of those, signals at 22 loci were supported by both methods (constituting 23 signals, given 2 signals at DGKB): 21 of these signals reflect associations with T2D, and 6 with glycemic traits (4 were signals for both T2D and glycemic traits, Supplementary Data 15). Amongst this set of 23 signals, we confirmed colocalization with T2D or glycemic associations at several previously reported cis-eQTL signals, adding to the evidence for ADCYS, HMG20A, CAMKID, and DGKB, as candidate effector transcripts. At 13 other signals, including CLUAP1, EIF3C, and RNF6, we observed islet cis-eQTL colocalization not reported before (Supplementary Data 15). At the remaining 24 (of 47) signals, colocalization was supported by either RTC or COLOC but not both, 19 of these reflecting associations with T2D, 8 with glycemic traits (3 overlapping). This included 7 previously reported islet cis-eQTL signals (including those at NKKX6-3, IGF2BP2, or KHLH42), plus 17 signals never reported before (including those at ADRA2A, PDE6B, and SLC12A8) (Supplementary Table 5).

For example, previous efforts to characterize the mechanism of action at the TCF7L2 locus have demonstrated that the fine-mapped T2D-risk allele at rs7903146 influences chromatin accessibility and enhancer activity in islets, but evidence linking these events to TCF7L2 expression was not previously detected in genome-wide eQTL studies. Our data reveal that the C allele of rs7309146 increases islet expression of the last (3′) exon of TCF7L2 (eQTL slope = 0.21, P = 1.9 × 10^-7, Fig. 3b). A directionally consistent signal (p < 0.05) was seen for 15 more of the 18 exons expressed. The same eQTL signal was also detected in the smaller β-cell-specific analysis (n = 26; eQTL slope = 0.72; p = 1.0 × 10^-7). The association between rs7903146 and TCF7L2 expression was restricted to islets, consistent with evidence that nondiabetic carriers of the TCF7L2 risk-allele display markedly reduced insulin secretion. Recent studies have proposed ACSL5 as a possible effector transcript at this locus, but we found no support—in any tissue—that rs7903146 influences ACSL5 expression.

Of the 27 previously reported signals of GWAS/islet eQTL overlap (Supplementary Table 5), 12 were not observed in our exon-eQTL-based analysis. Potential explanations include low expression level of targeted genes, and between-study differences in analytical approach and significance thresholds. For example,
the primary eQTL for MTNR1B has shown consistent islet cis-eQTL signals in previous studies \textsuperscript{17,46}, but the low level of expression of MTNR1B in islets meant that the exon-level read coverage fell below our threshold for inclusion. However, in the complementary gene-level analyses, there remained strong evidence of colocalization between the lead T2D variant (rs10830963) and MTNR1B expression ($p = 5.3 \times 10^{-21}$; Supplementary Data 2). At ZMIZ1, the previously reported cis-eQTL was nominally significant (rs185040218; $p = 3.0 \times 10^{-5}$), but did not reach the 1% FDR threshold for inclusion in colocalization testing.

At some loci, complex, but divergent, patterns of association between the eQTL and T2D GWAS signals challenged the assumptions of these colocalization methods. At the MAP3K11 locus, for example, the association plots indicate two independent islet eQTL signals for LTBP3 (rs11227223 and rs1194077), but only the latter signal colocalizes with the T2D GWAS signal at rs1783541 (Supplementary Fig. 10). RTC detects this as colocalization as it controls for additional eQTL signals, but this was not possible with COLOC that assumed a single eQTL variant to be active.

We further attempted to characterize eGenes that overlapped T2D/glycemic GWAS signals by assessing the impact of changes in glycemic status on islet expression. We used data from a recent analysis of islets recovered from diabetic and nondiabetic donors, focusing on transcripts that showed acute changes in expression when exposed to glucose levels in culture that contrasted with those to which they had been habituated \textsuperscript{47} (Supplementary Data 16). Islet eGenes, such as STARD10, WARS, SIX3, NKK6-3, and KLHL42, which may be of particular interest in that their
expression in islets is regulated both by T2D-associated variation and by acute changes in glucose exposure.

**Experimental validation at DGKB.** The DGKB locus features three independent T2D GWAS signals. In the T2D meta-analysis of Mahajan et al.\(^5\), these are represented by the lead SNPs rs2191349, rs17168486, and rs2908334. In the current study, we also report three independent islet cis-eQTLs influencing DGKB expression represented by the lead eSNPs rs10278505, rs17168486, and rs10231021. There is colocalization of these associations at two signals: we refer to these as the 5’ signal (rs17168486) and the 3’ signal (rs10231021), which is in perfect LD with rs2191349 in Europeans (\(r^2 = 1, D’ = 1\); Fig. 4a). There was no evidence for colocalization of the third GWAS signal (located on the 3’ end of the region (15.2 Mb)) and the third eSNP (located in the far 5’ of the region (14.3 Mb)): these are not discussed further. For both colocalizing GWAS signals, the T2D-risk allele is associated with increased islet expression of DGKB, showing no evidence of colocalization.

**Fig. 4 Functional assessment of DGKB eQTL locus.** A. We show the two of the three independent islet eQTL signals that colocalize with identified independent GWAS variants near the DGKB gene locus (lead SNP rs17168486 referred at as the 5’ signal and lead SNP rs10231021 referred to as the 3’ signal). These signals colocalize with two independent T2D GWAS signals shown in b, where rs17168486 is referred to as the 5’ signal and lead SNP rs2191349 referred to as the 3’ signal. LD information was not available for SNPs denoted by (×). The third GWAS variant and the third eSNP are not shown as both are located outside this region and in opposite location with respect to DGKB, showing no evidence of colocalization. C. Normalized DGKB gene expression levels relative to the T2D-risk-allele dosage at the 3’ eQTL for DGKB lead SNP rs10231021, eQTL p-value adjusted to the beta distribution is shown. D. Genome browser view of the region highlighted in purple in a and b that contains the 3’ DGKB eQTL and T2D GWAS signals. Two regulatory elements (element 1 highlighted in green, element 2 highlighted in blue) overlapping ATAC-seq peaks in islet β-cells (islet single nuclei ATAC-seq\(^4\)) and bulk islets (islet track represents one islet sample from Varshney et al.\(^5\)) were cloned into a luciferase reporter assay construct for functional validation. All ATAC-seq tracks are normalized to 10 M reads and scaled from 0–15. E. Log 2 luciferase assay activities (normalized to empty vector) are shown for in rat (832/13), mouse (MIN6), and human (EndoC-bh1) β-cell lines for the element 2 (cloned in the forward orientation), highlighted in blue in d. The risk haplotype shows significantly higher (\(p < 0.05\)) activity than the non-risk haplotype in 832/13 and MIN6, consistent with the eQTL direction shown in c. P-values were determined using unpaired two-sided t-tests. F. EMSA for probes with risk and non-risk alleles at the four SNPs overlapping the regulatory element validated in e, using nuclear extract from MIN6 cells. Filled arrows, allele-specific binding; open arrows, non-allele-specific binding of proteins to probes. Source data are provided as a Source data file.
For functional follow-up of the 3′ signal (Fig. 4a, b), we considered seven variants that mapped to ATAC-Seq peaks in bulk islets and islet β-cells (islet single nuclei ATAC-seq) and were in high LD ($r^2 > 0.8$) with rs2191349 (Fig. 4d). Three (rs7798124, rs7798360, and rs7781710, Fig. 4d, element 1) overlap an ATAC-seq peak shared across islets/β-cells, skeletal muscle, and the lymphoblastoid cell line GM12878 (ref. 56); four others (rs10228796, rs10258074, rs2191348, and rs2191349, Fig. 4d, element 2) lie in a smaller but more islet/β-cell-specific peak 8145 bp away. We cloned these putative regulatory elements into luciferase reporter constructs and performed transcriptional reporter assays in three widely used cellular models of the β-cell (human EndoC-βH1 (ref. 57), rat INS-1-derived 823/13, and mouse MIN6 (ref. 58); see “Methods” section). Element 1 showed consistently high enhancer activity across all three lines (when cloned in the forward orientation), but no allelic differences consistent with the eQTL direction of effect (Supplementary Fig. 12). Element 2 showed reduced luciferase expression in all three β-cell lines when in forward orientation with respect to DGKB (Fig. 4c). The T2D-risk haplotype showed higher expression than the non-risk haplotype in 832/13 ($p = 1.9 \times 10^{-4}$) and MIN6 ($p = 1.1 \times 10^{-6}$; Fig. 4e), which is consistent with the eQTL direction (Fig. 4c). Equivalent data for the human EndoC-BH1 cell line was directionally consistent but not significant (Fig. 4e). Luciferase assays using element 2 in reverse orientation also showed consistent trends across the cell lines, reaching significance in 832/13 (Supplementary Fig. 13). In electrophoretic mobility shift assays (EMSA) performed using MIN6 nuclear extract, three element 2 variants (rs10228796, rs2191348, and rs2191349) showed allele-specific binding to nuclear proteins (Fig. 4f, filled arrows), supporting a functional regulatory role for all these variants. Given that the three variants are in a relatively small region and show allelic differences in binding to proteins, one explanation is that a complex of TFs binds to this regulatory element, with the T2D-risk alleles also alleviating the regulatory element repression in a direction consistent, with the observed effects of the 3′ signal on DGKB expression (Fig. 4c).

At the 5′ eQTL, we focused attention on rs17168486, which was both the lead SNP for islet cis-expression and T2D association and located in an islet ATAC-seq peak (Supplementary Fig. 14a). However, luciferase reporter constructs found no consistent allelic effects on transcriptional activity (Supplementary Fig. 14b).

### Discussion

We have used transcriptome sequencing in 420 human islet preparations to address issues of general relevance to the mechanistic interpretation of noncoding association signals detected by GWAS. We report an increase in the catalog of eQTLs from pancreatic islets from ~4000 published in Varshney et al.15 to >7000 and document the degree to which RNA-Seq of a disease-relevant tissue missing from a reference set (e.g., GTEx) provides a more complete survey of islet eQTLs. We used this information to extend the number of association signals for T2D and related glycemic traits from 27 loci overlapping with islet eQTLs to 46 (47 signals, including 23 signals supported by two different methods), identifying candidate effector transcripts at several loci. We also explored how cellular heterogeneity (both within the tissue of interest, and reflecting contamination with cells not of direct relevance) can complicate the interpretation of GWAS signal colocalization. We integrated our eQTL catalog with islet epigenomic data to reveal effect size heterogeneity attributable to local chromatin context and to infer in vivo TF directional activities.

Analyses of the physiological association patterns and regulatory annotation enrichment signals of T2D-risk alleles indicate that many, though by no means all, act through the islet9,10,12,14,53. A major motivation behind development of this enhanced catalog of islet eQTLs was to support identification of effector transcripts mediating the downstream consequences of these noncoding alleles. At DGKB, for example, evidence that both the T2D signals colocalize with islet eQTLs with directionally consistent impacts on DGKB expression lends credibility to a causal role for DGKB in T2D predisposition.

However, it is important to emphasize that robust inference from the coincidence of eQTLs and GWAS signals is difficult. First, the expression data in our study was derived from human islets cultured in basal glycemic conditions: eQTL signals restricted to a subset of the cells within those islets would have been hard to detect, and the same is true for genes whose expression is dependent on stimulation. Since not all T2D loci act through the mature islet, some of the eQTLs detected may reflect tissue-specific regulation that is not germane to the development of the diabetic phenotype.

Second, confident assignment of colocalization can be difficult. There are multiple algorithms to assess the evidence that two association signals are likely to reflect the same causal variants, but agreement between them is incomplete39. An additional challenge arises from the complex architecture of many GWAS signals, such that conditional decomposition is required before colocalization across multiple overlapping signals can be accurately assigned54. This is especially important when the sets of GWAS and cis-eQTL signals at a given locus are not completely overlapping, since obvious colocalization at one of the contributing signals can be masked by differences in the overall shape of the association signals that confounds simplistic analysis.

Third, recent studies have shown that functionally constrained genes—which are depleted for misense or loss-of-function variants—are also less likely to have eQTLs, indicating uniform intolerance of both regulatory and coding variation55–57. Complementary studies focusing on regulatory elements have shown that large, cell-specific stretch enhancers harbor smaller effect size eQTLs than ubiquitous promoter regions58 and that genes with more cognate enhancer sequence are depleted for eQTLs57. Our finding that islet eQTLs that map to the islet stretch enhancers most frequently implicated in GWAS regions had smaller eQTL effect sizes is consistent with these observations.

Finally, it is critical to emphasize that, even when colocalization has been demonstrated between a GWAS variant and a tissue-appropriate eQTL signal, this does not constitute proof that the eGene concerned mediates disease predisposition. Causal relationships other than variant-to-gene-to-disease are possible, including the possibility the variant has horizontally pleiotropic effects on each59. Growing understanding of the extent of shared local regulatory activity and regulatory pleiotropy makes such an alternative explanation all the more credible60. It is best to regard genes highlighted by coincident GWAS and eQTL signals as candidate effector transcripts, and to proceed to experimental approaches that enable direct tests of causality. These may involve perturbing the gene across a range of disease-relevant cell lines and animal models, and determining the impact on phenotypic readouts that represent reliable surrogates of disease pathophysiology.

### Methods

#### Cohort characteristics

The samples from 420 donors included 189 males and 231 females, with an age range of 16–81 years (median = 54 years, 11 not available). Of the 420 individuals, 37 were identified as diabetic. BMI information was available for 334 individuals (median, 26.3 kg/m²), while HbA1c measurements were available for 198 (median, 5.8%). Due to the historical nature of some of the
samples used in this study, QCs information about the pancreatic islet isolation was limited: 234 samples listed their purity (median, 75%) and 26 samples listed their viability (median, 93.3%). No other biological information was available in the historical records. This information is available in the covariate files included in theEGA submission.

Pancreatic islet sample collection and processing. Samples collection and processing is summarized in Supplementary Fig. 1.

The processing of the Genome samples from Nica et al.18 were originally done using Affymetrix libraries with 49-bp paired-end reads; however, in order to allow joint analysis with the other available datasets for this study, mRNA samples were reprocessed using a 100-bp paired-end sequencing protocol. The library preparation and sequencing followed custom Illumina TruSeq protocols for next generation sequencing as described in the original publication62. All procedures followed ethical guidelines at the University Hospital in Geneva. The 89 Lund samples from Fadista et al.18 were jointly processed with 102 new islet samples that were processed uniformly following the same protocol. These islet samples were obtained from 191 cadaver donors of European ancestry from the Nordic Islet Transplantation Programme (http://www.nordicislets.org). Purity of islets was assessed by dithiothreitol staining, measurement of the shared sGAPDH expression and an estimate of the contribution of the exocrine and endocrine tissue were assessed as previously described63. Total RNA was isolated with the AllPrep DNA/RNA Mini Kit following the manufacturer’s instructions (Qiagen), sample preparation was performed using Illumina’s TruSeq RNA Sample Preparation Kit according to manufacturer’s instructions. The template was selected using a paired-end 101 bp protocol on the HiSeq2000 platform (Illumina). Illumina Casava v1.8.2 software was used for base calling. All procedures were approved by the ethics committee at Lund University.

The Oxford dataset included samples collected in Oxford and Edmonton that were jointly sequenced in Oxford are included in this set of samples. Islet sample procurement, mRNA processing, and sequencing procedure has been described in van de Bunt et al.17. To the 117 samples previously published (78 from Edmonton and 39 from Oxford), 57 samples were added and processed following similar protocols as before (27 from Edmonton and 30 from Oxford). Briefly, freshly isolated human islets were collected at the Oxford Centre for Islet Transplantation in Oxford, or the Alberta Diabetes Institute IsletCore (www.isletcore.ca) in Edmonton, Canada. Additional islets were obtained from the Alberta Diabetes Institute IsletCore’s long-term cryopreserved biobank. Freshly isolated islets were processed for RNA and DNA extraction after 1–3 days in culture in CMRL media. Cryopreserved islets thawed as described in ‘Methods’ (Supplementary Fig. 1) and RNA was extracted from human islets using Trizol (Ambion, UK or Sigma-Aldrich, Canada). To clean remaining media from the islets, samples were washed three times with phosphate-buffered saline (Sigma-Aldrich, UK). After the final cleaning step 1 mL Trizol was added to the cells. The cells were lysed by pipetting immediately to ensure rapid inhibition of RNase activity and incubated at room temperature for 10 min. Lysates were then transferred to clean 1.5 mL RNase-free centrifuge tubes (Applied Biosystems, UK). RNA quality (RIN score) was determined using an Agilent Bioanalyzer 2100 (Agilent, UK); samples with RNA integrity number (RIN) > 6.5 were prepared for Illumina with custom 8 bp indexes64. Libraries were multiplexed (three samples per lane), clustered using TruSeq PE Cluster Kit v3, and paired-end sequenced (100 nt) in a single lane, using the Illumina HiSeq2000 platform. All procedures were approved by the Human Research Ethics Board at the University of Alberta (Pro00013094), the University of Oxford’s Oxford Tropical Research Ethics Committee (OxTREC Ref. 2-19), or the Oxfordshire Regional Ethics Committee B (REC reference: 09/H0605/2). All organ donors provided informed consent for use of pancreatic tissue in research.

The USA samples from Varshney et al.15 were originally processed as follows: 39 islet samples from organ donors were received from the Integrated Islet Distribution Program, the National Disease Research Interchange, and Prodo-Labs. Total RNA from 2000–3000 islet equivalents was extracted and purified using Trizol (Life Technologies). RNA quality was confirmed with Bioanalyzer 2100 (Agilent, CA) samples with RNA integrity number > 6.5 were prepared for mRNA sequencing. We added External RNA Control Consortium spike-in controls (Life Technologies) to one microgram of total RNA. PolyA+ stranded mRNA RNA-Seq libraries were generated for each islet using the TruSeq-stranded mRNA kit according to manufacturer’s protocol (Illumina). Each islet RNA-Seq library was barcoded, pooled into 12-parallel batches, and sequenced over multiple lanes of HiSeq2000 to obtain an average depth of 100 million 2 × 101 bp sequences. All procedures followed ethical guidelines at the National Institutes of Health (NIH).

B-cell sample collection and processing. To the 11 FAC-sorted beta-cell populations samples previously published19, we added 15 more samples that were processed following the same protocols. Briefly, islets were dispersed into single cells, stained with the Hoechst 33342, and sorted into 12 fractions (in the way and labels to be further described in the number of the four laboratories before performing a standard eQTL analysis against non-permutated genotypes (and matched PC’s for genotypes), using different numbers of
PCs for expression. Significant eQTLs beyond a 5% FDR in any of these analyses are considered false positives due to technical differences across laboratories. Our results for 25 PCs for expression were sufficient to minimize the number of false positives due to batch effects originating from differences in processing of the islet samples. In addition, we performed an eQTL discovery controlling for 1, 5, 10, 20, 30, 40, and 50 PCs for expression, as well as gender, 4 PCs derived from genotype data, and a variable defining the laboratory of origin (coded as: OXF, LUN, and UCL). For the evaluation of GWAS results, we conclude that controlling for 25 PCs for expression was optimal as this controlled for differences across and within laboratories, while maximizing the discovery of eQTLs.

**eQTL analysis.** eQTL analysis for islets and β-cells were performed using fastqTL19 on 420 islet samples and 26 FAC-sorted β-cell samples with available genotypes. cis-eQTL analysis was restricted to SNPs in a 1 MB window upstream and downstream of the gene. For each gene, and each SNP, we derived MAF > 1%. For the analysis of β-cell samples, we used a filter of MAF > 5%. Exon-level eQTLs identified best exon–SNP association per gene (using the group flag), while gene-level eQTLs used gene quantification and identified the best gene–SNP association. Variables included in the linear models were the first 4 PCs for genotypes, the first 25 PCs for expression, gender, and a variable identifying the laboratory of origin of the samples. Significance for the SNP–gene association was assessed using 1000 permutations per gene, correcting p-values with a beta approximation distribution.25 Genome-wide multiple testing correction was performed using the q-value correction22 implemented in largeQvalue26.

Results of this joint analysis were highly correlated with those obtained from a fixed-effects meta-analysis of the four component studies, indicating appropriate control for the technical differences between the studies (Supplementary Fig. 16).

To discover multiple independent eQTLs, we applied a stepwise regression procedure that has also been described in Brown et al.25. We tested significant eGenes (FDR < 5%) for associations in any exon per gene. The maximum correct p-value (correcting for multiple testing across the SNPs and exons) over these genes was taken as the gene-level threshold. The next stage proceeded iteratively for each gene and threshold. A cis-scan of the window was performed in each iteration, using 1000 permutations and correcting for all previously discovered SNPs. If the beta-adjusted p-value for the max significant exon–SNP sets of SNP (fixed association) was not significant at the gene-level threshold, the forward stage was complete and the procedure moved on to the backward step. If this p-value was significant, the best association was added to the list of discovered eQTLs as an independent signal and the forward step proceeded to the next iteration. The exon-level cis-eQTLs associated with the best SNP; however, any remaining exons were dropped, otherwise the best association from the scan was chosen as the variant that represented the signal best in the full model.

**GTeX eQTLs.** We identified exon-level eQTLs for 44 GTeX tissues using fastqTL19 following the same procedure as for the islet eQTLs. Covariates included the following as fixed effect variables for ISVs for ISVs: 1) for expression with ≤154 samples; 30 PCs for samples between 155 and 254 samples; and 35 PCs for samples with >254 samples. Independent eQTLs from exons were identified as described for islet eQTLs. The proportion of shared eQTLs between islet and β-cell eQTLs and the eQTLs from GTeX tissues were identified using H.22.

**Tissue deconvolution.** To identify the contribution of β-cells, non-β-cells and exocrine (non-islet) cells to overall gene expression measured in islets, we performed an expression deconvolution analysis. Expression profiles from GTeX whole pancreas was used as a model for the exocrine component of expression27 while FAC-sorted expression profiles from β-cell and non-β-cells from Nica et al.18 were used to identify the fraction of expression derived from islet cells. First, we performed differential expression analysis of (a) exocrine vs whole islet samples; (b) β-cell vs whole islet samples; and (c) non-β-cell vs whole islet samples. The top 500 genes from each analysis were combined, and a deconvolution matrix of log2-transformed median expression values was prepared for each cell type. Next, deconvolution was performed using the Bioconductor package DeconRNASeq27. Deconvolution values per sample are included in the covariates file, together with the expression values in the EGA submission.

**Genotype-by-cell type associations.** Genotype-by-cell-type-specific regulatory effects were identified by testing for interactions between SNPs and cellular fraction estimates. We performed the analysis using a linear model and residuals from expression in gene quantifications. Residuals were extracted using a linear mixed model controlling for fixed effect variables (batch effects, islets purity, GC methylation level, and merge), and random variables (tag/primers and date of sequencing). The merge variable identified samples that were sequenced multiple times, with a final set of reads merged from multiple files. Significance for the genotype-by-cell interactions were evaluated using FDR and 100 permutations in each analysis: genotype-by-hybrid proportions, genotype-by-non-β-cell proportions, and genotype-exocrine cell proportions.

**Enrichment of eQTLs in T2D and glycemic GWAS.** Within each tissue, we asked if the magnitude of the eQTL effect for a given set of GWAS SNPs were larger than expected for a randomly selected matched set of SNPs (as described below). We performed enrichment analysis separately for each trait and tissue type using the following procedure. For each trait, we used lead GWAS variants from the following tissue types: T2D (all; n = 2772 associated variants of ≤20 SNPs) that likely act via β-cell action (n = 43); glycemic traits (fasting glucose and β-cell function (HOMA-B) in nondiabetic individuals; n = 56); and for comparison T1D (n = 55)36 (Supplementary Data 13)36,62. As INSPIRE pancreatic islet and GTeX tissue-based estimates of the variant effects, we used eQTL beta coefficients for each GWAS lead variant for a given trait, we identified the eQTL with the largest absolute effect size estimate among all the tested exons (max individual SNP beta). Across the lead GWAS variants, we took the median of the max individual SNP betas (observed median of the maxes). To generate a null distribution of the medians of the maxes, we repeated the analysis 15,000 times. For each replicate, we matched each GWAS SNP to a SNP present in the eQTL data based on the number of SNPs in LD, distance to TSS, number of nearby genes and MAP, and calculated the median of the maxes (null median of maxes). To form a distribution of the effect size enrichment we divided the observed median of the maxes by the null medians of the maxes for each tissue and factor, we defined the maximum size enrichment distribution as the enrichment estimate. We estimated the one-sided 95% confidence intervals as the fifth percentiles of the effect size enrichment distribution. We calculated the one-sided p-value for enrichment as the proportion of replicates with enrichment values <1.

**Colocalization of islet eQTL with T2D GWAS.** Colocalization of GWAS variants and eQTLs was performed using both COLOC30 and RTC35. For the analysis using COLOC, all variants within regions tagged as independent were considered as GWAS variants and tagged variants were tested for colocalization, using default parameters from the software. We analyzed summary statistics from T2D GWAS35 and fasting glucose35. GWAS variants and eSNPs pairs were considered to colocalize if the COLOC score for shared signal was >0.9. RTC analysis was also performed using defaults parameters from the RTC analysis of GWAS variants for fasting glucose (Supplementary Data 13). Associations between GWAS and gene expression were considered as colocalized if RTC score was >0.9. The plots showing colocalization of GWAS and eQTLs were generated using LocusCompare34.

**Chromatin states, islet ATAC-seq, and TF footprints.** We used a previously published 13 chromatin state model that included pancreatic islets along with other diverse tissues18. Briefly, these chromatin states were generated from cell-based ATAC-seq data for H3K4me1, H3K27ac, H3K4me3, H3K36me3, and H3K4me4, and input control from a diverse set of publically available data17–20 using the ChromHMM program41. Chromatin states were learned jointly from 33 cell/tissues that passed QC by applying the ChromHMM (version 1.10) hidden Markov model algorithm at 200-bp resolution to five chromatin marks and input25. ChromHMM with a replication of 100 states for each model, which in the most accurately captured information from higher-state models and provided sufficient resolution to identify biologically meaningful patterns in a reproducible way. As reported previously19, stretch enhancers were defined as contiguous enhancer chromatin state (active enhancers 1 and 2, genic enhancer, and weak enhancer) segments longer than 3 kb, whereas typical enhancers were enhancer state segments smaller than the median length of 800 bp (ref. 31).

We downloaded raw ATAC-seq data (fastq files) for a total of 33 islet samples from nondiabetic donors: 14 from ref. 82, 17 from ref. 12, and two from ref. 15. We processed these data uniformly by trimming all reads to 36 bp, mapping to hg19 (version 17.2.1) with the bwa-mem (version 0.7.15-r1140)83, removing duplicates using Picard (http://broadinstitute.github.io/picard), and pruning reads to retain properly paired and mapped reads (samtools view -F 256 -F 1024 -F 2048 -q 30). Since these 33 samples were obtained from different studies, we sought to obtain the set of peaks that were reproducible. We subsampled each ATAC-seq sample bam file to the minimum read depth across samples of 27,994,993 reads, merged these subsampled bam files across all samples and called peaks, using (1) the merged bam file that uniformly represents each sample and (2) each individual sample bam file. We used MACS2 (https://github.com/tadoi/MACS), version 2.1.0, with flags -g hs –nomodel –shift -100–extsize 200 -B–keep-dup all, to call peaks and filtered out regions blocked by the ENCODE germline Human database due to poor mapability (wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed). We then selected peaks from the merged bam file that were reproducibly called across the majority (at least 17) of the 33 individual samples, resulting in 64,129 peaks that we used for downstream analyses.

**TF footprint motifs are occurrences of TF motifs (obtained from databases of DNA-binding motifs for several TFs) in accessible chromatin regions (identified
from assays such as ATAC-seq). We downloaded previously published islet TF footprint motifs\(^{25}\), which were generated in a haplotype-aware manner using ATAC-seq and genotyping data from the phased, imputed genotypes for two islet samples using Vc2Idiploid81 v0.2.6a and DNA-binding motif information for 1995 publicly available TF motifs\(^{84-49}\). We subset the footprint motifs selecting occurrences within the new set of reproducible ATAC-seq peaks described above.

**Filtering eQTL SNPs.** Since low MAF SNPs, due to low power, can only be identified as significant eQTL SNP (eSNPs) with high eQTL effect sizes (slope or the beta from the linear regression), we observed that absolute effect size varies with respect to the most preferred base in the motif. For each TF footprint motif, we determined the motif directionality analysis for the selected footprints to overlap in eSNPs with same parameters as described above (Supplementary Data 12).

**eSNP effect size distribution in chromatin states.** We identified the islet eQTL SNPs (after LD pruning and MAF filtering as described above) occurring in chromatin states or ATAC-seq peaks within chromatin states using BEDTools intersect\(^{45}\). Similar to the enrichment calculation procedure, we considered a unique eQTL overlap if the lead eSNP or a proxy SNP with LD (\(r^2\)) > 0.99 occurred in these regions. We considered the effect size as the slope or the beta from the linear regression for eQTL overlaps in this region. \(P\) values were calculated using the Wilcoxon rank-sum test in R (ref. \(^{49}\)).

**MF motif directionality analysis.** We selected TF footprint motifs that were significantly enriched (after Bonferroni correction accounting for 1995 total motifs) to overlap islet eQTL (considering LD (< 0.2) pruned lead eSNPs or their \(r^2\) > 0.99 proxy eSNPs as a locus) with at least ten eQTL locus overlaps, resulting in \(N = 329\) TF motifs. We determined the overlap position of the TF with each TF footprint motif. We considered instances where the eSNP overlapped the MF motif at a position with information content > 0.7 and either the eSNP effect or the non-eSNP allele was the most preferred base in the motif. For each TF footprint motif and eSNP overlap, we rekeyed the direction of effect on the target gene being positive or negative with respect to the most preferred base in the motif. For each TF motif, we compiled the fraction of instances where the SNP allele that was most preferred in the TF footprint motif (i.e., base with highest probability in the motif) associated with increased expression of the associated gene. We refer to this metric as the motif directionality fraction where fractions near 1 suggest activating and fractions near 0 suggest repressive effects toward the target gene expression. Motif directionality fraction near 0.5 suggests no activity preference or context dependence.

We compared our results to a previously published study that quantified transcription activating or repressive activities based on massively parallel reported assays in HepG2 and K562 cells\(^{34}\) (Supplementary Fig. 8). We found that the motif directionality measures metrics derived from MPRAs performed in HepG2 and K562 cell line\(^{34}\) (Supplementary Fig. 8). We then considered 109 motifs from our analyses that were reported to have significant (\(p < 0.05\)) activating or repressive activities in both HepG2 and K562. With the null expectation of the motif directionality fraction being equal to 0.5, i.e., TF binding equally likely to increase or decrease target gene expression, we used a binomial test to identify TFs that show significant deviation from the null (\(N = 18\) at FDR < 0.10, Supplementary Data 12).

**Cell culture and transcriptional reporter assays.** MIN6 mouse insulinoma beta-cells\(^{27}\) were cultured in Dulbecco’s modified Eagle’s Medium (DMEM; Sigma-Aldrich, St. Louis, Missouri/USA) with 10% fetal bovine serum, 1 mM sodium pyruvate, and 0.1 mM beta-mercaptoethanol. INS-1 832/13 rat insulinoma (a gift from C. Newg, Duke University, Durham, North Carolina/USA) were grown in RPMI-1640 medium (Corning, New York/USA) supplemented with 10% fetal bovine serum, 10 mM HEPES, 2 mM L-glutamine, 1 mM sodium pyruvate, and 0.05 mM beta-mercaptoethanol. EndoC-H1 cells (Endocell) were grown according to Ravassard et al.\(^{48}\) in DMEM (Sigma-Aldrich), 5.6 mM/L glucose with 2% BSA fraction V fatty acid free (Roche Diagnostics), 50 mM/L 2-mercaptoethanol, 10 mM/L nicotinamide (Calbiochem), 5.5 μg/mL transferrin (Sigma), 6.7 ng/ml insulin (Sigma-Aldrich), 100 μM penicillin, and 100 μg/mL streptomycin. Cells were grown on coating consisting of 1% matrigel (Sigma). We maintained cell lines at 37 °C and 5% CO2.

To test haplotypes for allele-specific effects on transcriptional activity, we PCR amplified a 765-bp genomic region (element 1) containing variants: rs7598698, rs7598636, and rs7587110, and a second 592-bp genomic region (element 2) containing variants: rs10228796, rs10258074, rs2191349, and rs2191348 from DNA of human lymphoblastoid cells (Supplementary Table 6). We cloned the PCR amplicons into the multiple cloning site of the Firefly luciferase reporter vector pGL4.23 (Promega, Madison, Wisconsin/USA) in both orientations, as described previously\(^{27}\). Vectors are designated as “forward” or “reverse” based on the PCR-amplonation protocol with respect to DGKB gene. We isolated and verified the sequence of five independent clones for each haplotype in each orientation. For the 5′ eQTL a 250 bp construct containing the rs17168486 SNP (Origene) was subcloned into the Firefly luciferase reporter vector pGL4.23 (Promega) in both orientations.

We plated the MIN6 (200,000 cells) or 832/13 (300,000 cells) in 24-well plates 24 h before transfections and plated the EndoC-H1 cells (140,000 cells) 48 h prior to transfection. We co-transfected the pGL4.23 constructs with pRL-TK Renilla luciferase reporter vector (Promega) in duplicate into MIN6 or 832/13 cells, and in tripliicate into EndoC-H1 cells. We performed luciferase reporter assays in both HepG2 and K562 (ref. \(^{90}\)). We then performed islet eQTL enrichment:

\[
\text{islet eQTL enrichment} = \frac{n_{\text{filtered}}}{n_{\text{unfiltered}}} < 0.8 \text{ for the EUR population using PLINK}\text{ and 1000 genomes enrichment (ref. } 90)\text{.}
\]

We subset the footprint motifs selecting.

\[
\text{motif directionality analysis} = \frac{\text{fraction of instances where the SNP allele that was most preferred in the TF footprint motif (i.e., base with highest probability in the motif) associated with increased expression of the associated gene}}{\text{total number of instances where the motif directionality analysis was calculated}}.
\]

**Data availability**

All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon request.
References
1. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44, 50–54 (2012).

2. Mahajan, A. et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet. 46, 234 (2014).

3. Scott, R. A. et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat. Genet. 44, 991 (2012).

4. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. Nature 536, 41–47 (2016).

5. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. 50, 1505–1513 (2018).

6. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes. Diabetes Eur. Diabetes 50, 1724–1731 (2011).

7. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. 50, 956–967 (2018).

8. Agúndez, F. et al. Genetic effects on gene expression across human tissues. Nature 550, 294 (2017).

9. Small, K. S. et al. Regulatory variants at KLF14 in pancreatic islets in health and type 2 diabetes. Cell Metab. 24, 593–607 (2016).

10. Wang, Y. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 29, 351–364 (2019).

11. B. Lavello, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. Genome Res. 27, 208–222 (2017).

12. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. Cell 171, 321–330 e14 (2017).

13. Bonnycastle, L. L. et al. Single-cell transcriptomics from human pancreatic islets: sample preparation matters. Biol. Methods Protoc. 4, bps02019 (2019).

14. Abraham, B. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl Acad. Sci. USA 115, 17921–17926 (2018).

15. Pasquale, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. 46, 136–143 (2014).

16. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. 33, 364–376 (2015).

17. Ernst, J. et al. Genome-scale high-resolution mapping of activating and repressive nucleosomes in regulatory regions. Nat. Biotechnol. 34, 1180–1190 (2016).

18. Manni, A. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat. Genet. 44, 659–669 (2012).

19. Klaproth, B. et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes 60, 2624–2634 (2011).

20. Urdh, M. S. et al. Type 2 diabetes genetic loci informed by multi-traits associations point to disease mechanisms and subtypes: a soft clustering analysis. PLoS Med. 15, e1002654 (2018).

21. Ongen, M. et al. Fine mapping of type 2 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat. Genet. 47, 381–386 (2015).

22. Kanduri, C., Bock, C., Gundersen, S., Hovig, E. & Sandve, G. K. Colocalization analyses of genomic elements: approaches, recommendations and challenges. Bioinformatics 35, 1615–1624 (2019).

23. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383 (2014).

24. Ongen, M. et al. Estimating the causal tissues for complex traits and diseases. Nat. Genet. 49, 1676–1683 (2017).

25. Carrat, G. R. et al. Decreased STARD10 expression is associated with defective insulin secretion in humans and mice. Cell 167, 1505–1517 (2016).

26. Zhou, Y. et al. TCF7L2 is a master regulator of insulin production and processing. Hum. Mol. Genet. 23, 6419–6431 (2014).

27. Xia, Q. et al. The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. Diabetologia 59, 2360–2368 (2016).

28. Tuomi, T. et al. Increased melanin signaling is a risk factor for type 2 diabetes. Cell Metab. 23, 1067–1077 (2016).

29. Ottosson-Lakso, E. et al. Glucose-induced changes in gene expression in human pancreatic islets: causes or consequences of chronic hyperglycemia. Diabetes 66, 3013–3028 (2017).

30. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. Nat. Genet. 50, 559–567 (2018).

31. R. V. et al. Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. Mol. Metab. 32, 109–121 (2020).

32. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 7, 1213–1218 (2010).

33. Wii, P. et al. A genetically engineered human pancreatic beta cell line exhibiting glucose-inducible insulin secretion. J. Clin. Investig. 121, 3589–3597 (2011).

34. Miyazaki, J. et al. Establishment of a pancreatic beta cell line that retains glucose-inducible insulin secretion: special reference to expression of glucose transporter isoforms. Endocrinology 127, 126–132 (1990).
53. Vaise, C. et al. Melanocortin-4 receptor mutations are a frequent and heterogeneous cause of morbidity and obesity. J. Clin. Investig. 106, 253–262 (2000).
54. Liu, R., Glaudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. Nat. Genet. 51, 768–769 (2019).
55. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
56. Varshney, A. et al. Cell specific regulatory motifs in ENCODE TF binding experiments. Genetica 121, 757–772 (2019).
57. Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene function in complex disease. Am. J. Hum. Genet. 106, 215–233 (2020).
58. Varnheuy, A. et al. Cell specificity of human regulatory annotations and their genetic effects on gene expression. Genetics 211, 549–562 (2019).
59. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. 48, 709 (2016).
60. Deloue, A. et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. Science 364, eaat8266 (2019).
61. Taneira, J. et al. A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. Cell Metab. 122, 132–144 (2020).
62. Manning Fox, J. E. et al. Human islet function following 20 years of cryogenic preservation. J. Clin. Investig. 106, 253–262 (2000).
63. Lyon, J. et al. Research-focused isolation of human islets from donors with and without diabetes. Diabetes. 47, 1225–1235 (2008).
64. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for a heterogeneous cause of morbid obesity. J. Clin. Investig. 106, 253–262 (2000).
65. Murgia, S. et al. DNA-binding specificity of gene regulatory proteins. Bioinformatics 29, 2646–2647 (2013).
66. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.5997 (2013).
67. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Nat. Biotech. 211, 215–233 (2012).
68. Delaneau, O., Marchini, J. & Genomes Project, C. Integrating sequence and imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
69. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
70. Duggal, G., Nandi, P. & Hariharan, I. BWA-MEM: fast and accurate alignment by filtration. Nat. Methods 9, 1185–1188 (2012).
71. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1766–1774 (2012).
72. Wang, X. & Goldstein, D. B. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
73. Brown, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).
74. t Hoen, P. A. C. et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotech. 31, 1015–1022 (2013).
75. Brown, A. A. LargeQvalue: A program for calculating FDR estimates with large datasets. Preprint at https://doi.org/10.1101/100704 (2015).
76. Brown, B. W. et al. Predicting causal variants affecting expression using whole-genome sequencing and RNA-seq from multiple human tissues. Nature 491, 56 (2012).
77. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
78. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
79. Brown, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).
80. Howie, B. N. et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotech. 31, 1015–1022 (2013).
81. Brown, A. A. LargeQvalue: A program for calculating FDR estimates with large datasets. Preprint at https://doi.org/10.1101/100704 (2015).
82. Brown, B. W. et al. Predicting causal variants affecting expression using whole-genome sequencing and RNA-seq from multiple human tissues. Nature 491, 56 (2012).
83. Li, H. LargeQvalue: A program for calculating FDR estimates with large datasets. Preprint at https://doi.org/10.1101/100704 (2015).
84. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44, D110–D112 (2016).
85. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44, D110–D112 (2016).
86. Jolma, A. et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013).
87. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascope 4, 7 (2015).
88. Schmidt, E. M. et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 31, 2601–2606 (2015).
89. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible and accurate suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
90. R Core Team. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria 2020) https://www.R-project.org/.
91. Foegarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Molhle, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMKID type 2 diabetes GWAS locus. PLoS Genet. 10, e1004633 (2014).
Nordisk, Pfizer, Roche, Sanofi Aventis, Servier, and Takeda. As of June 2019, he is an employee of Genentech, and a holder of Roche stock. M.v.d.B. is an employee of Novo Nordisk A/S, although all experimental work was carried out under employment at the University of Oxford.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18581-8.

Correspondence and requests for materials should be addressed to A.V. or M.I.M.

Peer review information Nature Communications thanks the anonymous reviewers for their contributions to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

---

Ana Viñuela1,2,3,4,22✉, Arushi Varshney5,22, Martijn van de Bunt6,7,8,22, Rashmi B. Prasad9,22, Olof Asplund9, Amanda Bennett6, Michael Boehnke10, Andrew A. Brown1,2,3,11, Michael R. Erdos12, João Fadista9,13,14, Ola Hansson9,14, Gad Hatem9, Cédric Howald1,2,3, Apoorva K. Iyengar15, Paul Johnson6, Ulrika Krus9, Patrick E. MacDonald16, Anubha Mahajan6,21, Jocelyn E. Manning Fox16, Narisu Narisu12, Vibe Nylander7, Peter Orchard17, Nikolay Oskolkov9, Nikolaos I. Panousis1,2,3, Anthony Payne6, Michael L. Stitel18,19, Swarooparani Vadlamudi15, Ryan Welch10, Francis S. Collins12, Karen L. Mohlke15, Anna L. Gloyn6,7,8,20, Laura J. Scott10, Emmanouil T. Dermitzakis1,2,3, Leif Groop9,14, Stephen C. J. Parker5,17 & Mark I. McCarthy6,7,8,21✉

1Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. 2Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland. 3Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. 4Biosciences Institute, Faculty of Medical Sciences, Newcastle University, NE1 4EP Newcastle, UK. 5Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA. 6Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK. 7Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 7LE, UK. 8Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford OX3 7LE, UK. 9Lund University Diabetes Centre, Department of Clinical Sciences, Lund University, Skåne University Hospital, Malmö, Sweden. 10Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA. 11Population Health and Genomics, University of Dundee, Dundee, Scotland DD1 9SY, UK. 12Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. 13Department of Epidemiology Research, Statens Serum Institut, Copenhagen, DK 2300, Denmark. 14Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland. 15Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA. 16Department of Pharmacology and Alberta Diabetes Institute, University of Alberta, Edmonton, Alberta, Canada. 17Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. 18The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA. 19Department of Genetics and Genome Sciences, Institute for Systems Genomics, University of Connecticut, Farmington, CT 06032, USA. 20Department of Pediatrics, Division of Endocrinology, Stanford School of Medicine, Stanford University, Stanford, CA, USA. 21Present address: Human Genetics, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA. 22These authors contributed equally: Ana Viñuela, Arushi Varshney, Martijn van de Bunt, Rashmi B. Prasad. ✉email: ana.vinuela@newcastle.ac.uk; mccarthy.mark@gene.com