

1 **TIGER: The gene expression regulatory variation landscape of human pancreatic** 2 **islets**

3 Lorena Alonso^{1,*}, Anthony Piron^{2,3,*}, Ignasi Morán^{1,*}, Marta Guindo-Martínez¹, Sílvia Bonàs-
4 Guarch^{4,5}, Goutham Atla^{4,5}, Irene Miguel-Escalada^{4,5}, Romina Royo¹, Montserrat Puiggròs¹,
5 Xavier Garcia-Hurtado^{4,5}, Mara Suleiman⁶, Lorella Marselli⁶, Jonathan L.S. Esguerra⁷, Jean-
6 Valéry Turatsinze², Jason M. Torres^{8,9}, Vibe Nylander¹⁰, Ji Chen¹¹, Lena Eliasson⁷, Matthieu
7 Defrance², Ramon Amela¹, MAGIC¹², Hindrik Mulder¹³, Anna L. Gloyn^{9,10,14,15,16}, Leif
8 Groop^{7,13,17}, Piero Marchetti⁶, Decio L. Eizirik^{2,18,19}, Jorge Ferrer^{4,5,20}, Josep M.
9 Mercader^{21,22,23,1,#}, Miriam Cnop^{2,24,#}, David Torrents^{1,25,#}.

- 10
11 1 - Barcelona Supercomputing Center (BSC), Joint BSC-CRG-IRB Research Program in
12 Computational Biology, Barcelona, Spain
13 2 - ULB Center for Diabetes Research, Université Libre de Bruxelles, Brussels, Belgium
14 3 - Interuniversity Institute of Bioinformatics in Brussels (IB2), Brussels, Belgium
15 4 - Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), The Barcelona
16 Institute of Science and Technology (BIST), Barcelona, Spain
17 5 - Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas
18 (CIBERDEM) Barcelona, Spain
19 6 - Department of Clinical and Experimental Medicine, and AOUP Cisanello University Hospital,
20 University of Pisa, Pisa, Italy
21 7 - Unit of Islet Cell Exocytosis, Lund University Diabetes Centre, Malmö, Sweden
22 8 - Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population
23 Health, University of Oxford, Oxford, UK
24 9 - Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford,
25 Oxford, UK
26 10 - Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine,
27 University of Oxford, Oxford, UK
28 11 - Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical
29 School, Exeter, UK
30 12 - Members of the consortium are provided in Appendix S1
31 13 - Unit of Molecular Metabolism, Lund University Diabetes Centre, Malmö, Sweden
32 14 - Division of Endocrinology, Department of Pediatrics, Stanford University School of Medicine,
33 Stanford, CA, USA
34 15 - NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, UK
35 16 - Stanford Diabetes Research Centre, Stanford University, Stanford, CA, USA
36 17 - Finnish Institute of Molecular Medicine Finland (FIMM), Helsinki University, Helsinki, Finland
37 18 - WELBIO, Université Libre de Bruxelles, Brussels, Belgium
38 19 - Indiana Biosciences Research Institute, Indianapolis, IN, USA
39 20 - Section of Epigenomics and Disease, Department of Medicine, Imperial College London, London,
40 UK
41 21 - Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and
42 MIT, Cambridge, MA, USA
43 22 - Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA,
44 USA
45 23 - Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA
46 24 - Division of Endocrinology, Erasmus Hospital, Université Libre de Bruxelles, Brussels, Belgium
47 25 - Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
48

49 * - These three authors contributed equally to this work.

50 # - These authors jointly directed this work.

51

52

53 Corresponding authors:

54
55 Josep M Mercader
56 Programs in Metabolism and Medical and Population Genetics
57 Broad Institute of Harvard and MIT
58 75 Ames St
59 02142, Cambridge, MA
60 United States of America
61 E-mail: mercader@broadinstitute.org

62
63 and

64
65 Miriam Cnop
66 Université Libre de Bruxelles (ULB)
67 ULB Center for Diabetes Research
68 U.L.B. CP618
69 Route de Lennik 808
70 1070 Brussels, Belgium
71 E-mail: mcnop@ulb.ac.be

72
73 and

74
75 David Torrents
76 Life Science Department
77 Barcelona Supercomputing Center (BSC)
78 Institució Catalana de Recerca i Estudis Avançats (ICREA)
79 C/Jordi Girona 29, Edifici Nexus II
80 08003 Barcelona, Catalunya, Spain
81 Phone: 3493 413 40 74
82 E-mail: david.torrents@bsc.es

83

84 Keywords

85 Expression quantitative trait locus (eQTL), pancreatic islets, RNA-seq, regulatory variation,
86 epigenomics, allele-specific expression, type 2 diabetes.

87

88 **Abstract**

89 GWAS have identified more than 700 genetic signals associated with type 2 diabetes (T2D).
90 To gain insight into the underlying molecular mechanisms, we created the Translational
91 human pancreatic Islet Genotype tissue-Expression Resource (TIGER), aggregating >500
92 human islet RNA-seq and genotyping datasets. We imputed genotypes using 4 reference
93 panels and meta-analyzed cohorts to improve coverage of expression quantitative trait loci
94 (eQTL) and developed a method to combine allele-specific expression across samples
95 (cASE). We identified >1 million islet eQTLs (56% novel), of which 53 colocalize with T2D
96 signals (60% novel). Among them, a low-frequency allele that reduces T2D risk by half
97 increases *CCND2* expression. We identified 8 novel cASE colocalizations, among which
98 an *SLC30A8* T2D associated variant. We make all the data available through the open-
99 access TIGER portal (<http://tiger.bsc.es>), which represents a comprehensive human islet
100 genomic data resource to elucidate how genetic variation affects islet function and translate
101 this into therapeutic insight and precision medicine for T2D.

102

103

Introduction

Diabetes is a complex metabolic disease, characterized by elevated blood glucose levels, that affects more than 463 million people worldwide ¹. Type 2 diabetes (T2D) accounts for >85% of diabetes cases and is strongly related to age, obesity and sedentary lifestyles. Epidemiologic studies forecast a 40% increase in prevalence by 2030 ²⁻⁴. This makes the study and understanding of diabetes a top research and healthcare priority. Progressive pancreatic islet dysfunction is central to the majority of diabetes forms and hereby key to gain insights into the 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 ⁵⁻⁸. The vast majority of variants in these loci do not disrupt protein coding sequences ^{9,10}. 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 insights into the underlying molecular mechanisms and thus, to develop more effective and targeted preventive and therapeutic strategies ¹¹.

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 ¹². The Roadmap Epigenomics Mapping project ¹³ and the International Human Epigenome project ¹⁴ 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 impacted 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 ^{15,16}. Because human islets are difficult to obtain ¹⁷⁻¹⁹, 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 ²⁰. Hence, the development of publicly available resources and tools that include data on islet tissue is essential to translate T2D genetic signals into molecular and physiological mechanisms.

The first studies of eQTL in human islets pinpointed genes that might be influenced by genetic variants and thus possibly mediate T2D risk^{21,22}. 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, a multinational consortium effort, InsPIRE, generated a largest islet eQTL study with a sample size of 420 islet donors, which identified 46 T2D GWAS signals that colocalize with islet eQTL²³.

To further expand the understanding of human islet regulatory genomics, and its role in T2D the Horizon 2020 T2DSysTems consortium (<https://www.t2dsystems.eu/>) gathered the most extensive collection to date of human islet samples with gene expression, epigenomic data, genotypic and phenotypic information, with a total of 514, from which 207 samples 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 catalogue 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 five cohorts (Center for Genomic Regulation, Lund University, University of Oxford/University of Alberta, Edmonton, Università di Pisa and Université Libre de Bruxelles).

The DNA of 307 new samples was isolated, sequenced and genotyped (Suppl. Table S1, Suppl. Methods) and aggregated to be harmonized with existing data from 207 samples. After quality control, filtering of RNA-seq and genotyping array data (Suppl. Methods), 404 human islet samples remained with high quality genotypes and RNA-seq data (Figure 1B).

To fully characterize the genetic variation present in the samples, genotype imputation was performed separately for each cohort using four different reference panels as previously described^{7,24} (1000 genomes²⁵, GoNL²⁶, the Haplotype Reference Consortium²⁷ and

UK10K²⁸). 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, Suppl. Methods). This allowed imputation of >22 million unique high-quality genetic variants across all samples, 10% of which were indels and small structural variants, and more than 1.05 million variants in chromosome X (Figure 1C) (Suppl. Table S2, Suppl. Methods). Notably, this strategy allowed 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 (Figure 1D).

Additionally, we performed RNA-seq in 514 samples 460 of which were retained after stringent quality control, including >52 billion raw short reads. We uniquely aligned more than 48 billion reads (median of 93 million per sample) (Suppl. Table S3), which allowed us to observe >22K genes expressed at >0.5 transcripts per million (TPM) (Suppl. Methods).

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 four cohorts. First, we performed a *cis*-eQTL analysis, using data from each cohort independently (Suppl. Methods). For each analysis, we corrected for known covariates (age, sex and body mass index (BMI)), genetically derived principal components, and PEER factors for hidden confounding factors²⁹. The eQTL results from each of the four cohorts were then meta-analyzed (Figure 2A). This resulted in >1.11 million significant eQTLs in more than 21,115 eGenes (12,802 protein coding genes, 8,313 non-coding) at 5% false discovery rate (FDR) after Benjamini-Hochberg correction for multiple testing³⁰ (Figure 2B). The quantile-quantile plot showed no baseline inflation in the results (Suppl. Figure S1). More than 12% of all significant eQTLs were small indels or larger structural variants, 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 in human immune cell types³¹, and in GTEx, where it was observed that SVs have a stronger effect than SNVs³².

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 assays¹⁰. 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, $p=3\times10^{-166}$) (Suppl. Methods), as well as with strong enhancers¹⁰ (2-fold, $p=1.4\times10^{-16}$), and open-chromatin regions (1.4-fold, $p=3.9\times10^{-45}$) (Figure 2C, Suppl. Figure S2). These results are consistent with eQTL studies in other tissues¹².

Next, we contrasted the TIGER human islet results with the latest GTEx eQTL datasets, which analyzed 54 human tissues including whole pancreas, but not islets¹². Of all significant human islet eQTLs, 64.7% were also significant in at least one other 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 similar to the rest of 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 compared to 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 the TIGER results with the recently published eQTL analysis of 420 islet samples²³. Overall, 56.2% of the significant eQTLs were exclusive to our analysis (not assayed or non-significant in the InsPIRE study²³). Identification of eQTLs driven by low-frequency or rare variants may be more clinically impactful as significant low-frequency variants tend to have larger effects on disease risk and gene expression³³. Notably, the proportion of TIGER exclusive eQTLs increased to 74.6% for low-frequency variants, 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 (Suppl. Figure S3).

Gene ontology analysis of the significant human islet eQTL genes revealed signaling (including G-protein coupled receptor signaling) and metabolic regulation terms, albeit with moderate significance (Suppl. Figure S4). 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 related terms (including “lymphocyte/T-cell activation” and “regulation of immune system process”) (Figure 2E). This suggests that these novel eQTLs affected genes relevant to β -cell physiology, including some related to immune processes with potential relevance for type 1 diabetes³⁴.

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^{5–7}, by performing colocalization analyses using COLOC method³⁵ (Suppl. Methods).

This analysis uncovered 49 eQTL variants associated with expression of 53 genes that significantly colocalized with T2D GWAS loci (Suppl. Table 4), of which 32 are novel (Table

1, Suppl. Figure S5). Interestingly, we identified three low-frequency variants, 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. Among the 49 colocalizing signals (Suppl. Figure S5), rs77864822 (MAF=0.07) minor allele (G) was associated with higher *RMST* expression and decreased T2D risk (OR=0.93, $p=2.2 \times 10^{-8}$). By interrogating the latest GWAS study on glycemic traits³⁶, 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 hours glucose in an oral glucose tolerance test (beta=-0.064, $p=2.4 \times 10^{-4}$; Suppl. Table 4). The variant rs1531583 colocalized with *CPLX1* expression (Figure 3A-C). 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 rs73221115 ($r^2=0.978$ with rs1531583) and rs73221116 ($r^2=0.98$ with rs1531583) had allele-specific H3K27ac binding suggesting that these two variants are the most likely causal variants of the *CPLX1* locus (Figure 3D-E). We also identified significant colocalization between the low-frequency variant rs76895963, known to reduce T2D risk by half³⁷, and increased *CCND2* expression in islets (Figure 3F-G). This variant was also associated with reduced fasting glucose (beta=-0.033, $p=0.0017$), HbA1c (beta=-0.042, $p=3.6 \times 10^{-8}$) and reduced 2 hours glucose in oral glucose tolerance test (beta=-0.095, $p=0.01$, Suppl. Table 4).

An atlas of cASE in human pancreatic islets

Preferential expression of mRNA copies containing one of the two 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.

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: a) environmental and batch effects, which are important confounding factors in eQTL studies³⁸⁻⁴³, b) sample heterogeneity, which is prevalent in human islet samples⁴⁴, and c) *trans* effects, since these would affect the two 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 the overall ASE across the cohort of samples

(Suppl. Methods, Suppl. Figure S6)⁴⁵. 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 towards the reference allele were successfully controlled (see also Suppl. Figure S6B-E).

When comparing cASE genes against a set of non-significant genes (matched by gene expression level, Suppl. Methods), we observed that cASE genes were enriched for islet-specific 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 (Suppl. Methods) 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; Suppl. Figure S7), all of which with strong imbalance towards 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 first 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 two effects.

Of 2,707 cASE significant genes, 2,052 (75.8%) were detected in eQTL analysis, whereas 655 (24.2%) were detected uniquely through cASE (Figure 4F, top panel). The same trend was observed when considering only islet-specific expressed 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 and not eQTL analysis (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 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 between heterozygous and homozygous for the 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, see Suppl. Methods).

This analysis uncovered 256,981 putative regulatory variants for 3,425 genes, including 570 genes that had no significant reporter variants by themselves, but that did reach significance upon stratifying by genotype of regulatory variants (Figure 5C, orange points, see Suppl. Figure S8 for examples). To assay the potential functional impact of the identified reporter variants, we tested for their enrichment in human islet regulatory regions¹⁰, 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, $p=4\times 10^{-87}$) (Suppl. Methods), as well as with strong enhancers¹⁰ (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 1.11M eQTLs, we found 123,748 variants significant by both methods (3,138 with MAF<5%), and a further 133,233 (9,190 with MAF<5%) that 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 regulatory variants for all imbalanced genes in the region and identified 14 colocalized locus-gene pairs (Table 2, Suppl. Figure S9). Of these, 6 had also been identified in eQTL/T2D GWAS colocalization analyses, showing consistency between the two methods. Interestingly, the 8 colocalizations identified by cASE alone suggested that these T2D variants may mediate disease risk by causing an imbalance in allelic expression, rather than altering overall gene expression. 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 (Figure 5F-G) (Suppl. Table 5) for which there was no eQTL colocalization. Thus, our novel 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 6). 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 (Suppl. Methods).

The TIGER website represents homogeneous gene expression levels from 446 RNA-seq pancreatic islet samples corrected for batch and covariate effects (Suppl. Figure S10), and enables comparison with GTEx expression data ¹² (Suppl. Methods).

In addition to the eQTL and cASE results and to provide further functional assessment, we gathered islet regulatory information ^{9,10,46}, methylation marks ^{47,48} and chromatin modification datasets ⁴⁹⁻⁵¹. Further, to enable the translation of genetic variation to disease risk, we also integrated the latest T2D GWASs meta-analysis summary statistics ^{5,7,52,53} (Figure 1A).

The TIGER database currently contains expression and molecular data for >59K Gencode genes (version gencode.v23lift37 ⁵⁴) and >27M 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 that will help visualize and interpret the molecular context of the query, as well as download the data. As a result of these efforts, the TIGER resource has already been used in recent studies ⁵⁵⁻⁵⁷.

As an example, we present the visualization of *MTNR1B*, a gene associated with type 2 diabetes and impaired insulin secretion ⁵⁸. Although, this gene is lowly expressed in pancreatic islets (median 0.25 TPM) by comparison with other GTEx tissues it only shows low expression in testis (median 0.61 TPM) and brain (median 0.06 TPM) but none expression in whole pancreas and other tissues (median 0 TPM), thus highlighting the utility of this resource for studying human islet-specific expression (Figure 6A-B). A T2D risk associated locus has been previously described and fine-mapped ⁵ to a single variant (rs10830963, $p=4.8 \times 10^{-43}$, PP=0.99, Figure 6C, Suppl. Figure S5). Notably, this variant is located within islet H3K27ac peaks, suggesting potential regulatory implications of this variant (Figure 6D). In summary, the close lookup at this locus emphasizes that the TIGER portal can be easily used to interrogate gene expression, epigenomic and genomic variation regulatory landscape, providing an invaluable resource to the research community for the study of complex diseases affecting pancreatic islets.

Discussion

By analyzing the largest dataset to date of pancreatic islets with gene expression and dense genotyping information we have uncovered one million significantly associated variant-gene pairs. Of all 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 proxy for pancreatic islets and vice-versa.

We compared our findings with those obtained in the InsPIRE islet eQTL study that comprised 420 samples²³, of which 207 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 study (Suppl. Table 4). 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 large 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 detected 32 novel T2D colocalizations with low MAF variants, including variants associated with expression of *CCND2*, *RMST*, and *CPLX1*. The variant rs76895963 (MAF 0.02) that upregulates *CCND2*, halves the risk of T2D³⁷ and is potentially implicated in the perinatal development of human β -cells⁵⁹. While the posterior probability of the colocalization was below the threshold of 0.8, the SNP had a clear eQTL with the gene, and a convincing colocalization (see Locus Compare plots, Figure 3G). The variant rs77864822 (MAF=0.07) upregulates *RMST* expression and decreases T2D risk. *RMST* (rhabdomyosarcoma 2 associated transcript) is a reportedly neuron-specific long noncoding RNA involved in neurogenesis⁶⁰; it is well expressed in human islet cells⁶¹ but its function in β -cells is unknown. The variant rs1531583, with the minor T allele associated with increased T2D risk⁵, 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 secretion⁶². GWAS often report as a target the gene closest to the variant, in this case *PCGF3*, for which eQTLs exist in many GTEx tissues. Notably, rs1531583 lies in an intronic region of *PCGF3*, and is an eQTL for this gene in several GTEx tissues. However, we demonstrate here that it is specifically associated with *CPLX1* expression in human islets and not with *PCGF3*, challenging the hypothesis that the closest gene is often the most likely target gene (Figure 3A-E).

The imputation with four reference panels allowed us to analyze different sources of genetic variation, including indels and structural variants. 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 ⁵ did not consider indels, and so we could not include them in our colocalization analysis. In the near future, this approach could be used to fine-map the contribution to disease risk of indels and structural variants.

Capitalizing on this valuable pancreatic islet resource, we also analyzed *cis*-regulation via ASE for the first time. We developed a novel method named 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 the overall gene expression, and thus undetectable by eQTL. We extended the cASE results in colocalization analysis and identified 14 T2D colocalizations. While 6 of them were detected in the eQTL/T2D GWAS colocalization, 8 were novel signals, including *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 allelic expression of their target genes, a novel 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 ^{63–66}. 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 any or both of those SNPs could induce allelic imbalance. Rare loss-of-function variants in *SLC30A8* strongly reduce T2D risk ⁶⁷ by enhancing insulin secretion ⁶⁸. However, the direction of effect of the common coding variants is not known. Our cASE results suggest that imbalanced expression towards 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.

In summary, we generated the largest to date expression regulatory variation resource in human pancreatic islets, a tissue with a central pathogenic role in most if not all types of diabetes. All these 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 towards understanding the molecular underpinnings of complex diseases that impact pancreatic islet biology and provide a path for the identification of novel and personalized drug targets.

451

452 **Data and code availability**

453 The eQTL and cASE results are available for browsing at TIGER (<http://tiger.bsc.es>) and the
454 full summary statistics will also be available for download upon publication.

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

456 Source data used for this study supporting all findings are available within the article and its
457 Supplementary Information files or from the appropriate repositories. Already published
458 genotype, sequence, methylation and expression data was obtained from the European
459 Genome-phenome Archive (EGA; <https://www.ebi.ac.uk/ega/>) under the following accession
460 numbers: EGAD00001001601; EGAD00001003946; EGAD00001003947 and Gene
461 Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under the following
462 accession numbers: GSE50244; GSE76896; GSE53949; GSE35296. Samples used to
463 generate Islet regulome annotations, ChIP-seq and ATAC-seq were taken from EGA
464 repository, under the accession numbers: EGAD00001005203, EGAD00001005202,
465 EGAD00001005204, EGAD00001005201 and their corresponding processed files are
466 available through <https://www.crg.eu/es/programmes-groups/ferrer-lab#datasets>. New
467 genotype, RNA-sequencing and associated metadata from Pisa and CRG samples are
468 being deposited in EGA (EGA number pending).

469 **Acknowledgments**

470 This work has been supported by the European Union's Horizon 2020 research and
471 innovation program T2Dsystems under grant agreement No 667191. L.Alonso was
472 supported by the grant BES-2017-081635 of Severo Ochoa Program, awarded by the
473 Spanish Government. I. Moran was supported by the FJCI-2017-31878 Juan de la Cierva
474 grant, awarded by the Spanish Government. Work in the M.Cnop and D.Eizirik labs was
475 further supported by the Fonds National de la Recherche Scientifique (FNRS), the Brussels
476 Region Innoviris project DiaType and the Walloon Region SPW-EER Win2Wal project
477 BetaSource, Belgium. Eizirik is also supported by a grant from the Welbio-FNRS (Fonds
478 National de la Recherche Scientifique), Belgium, and start-up funds from the Indiana
479 Biosciences Research Institute (IBRI), USA. J.M.Mercader is supported by American
480 Diabetes Association Innovative and Clinical Translational Award 1-19-ICTS-068. J.Chen is
481 supported by an Expanding excellence in England award from Research England. H.Mulder,
482 J.L.S.Esguerra and L.Eliasson are supported by the Swedish Strategic Research
483 Foundation (IRC15-0067).A.L.Gloyn is a Wellcome Trust Senior Fellow in Basic Biomedical
484 Science. This work was funded in Oxford & Stanford by the Wellcome Trust (095101 [ALG],

200837 [A.L.G.], 106130 [A.L.G.], 203141 (A.L.G.) and NIH (U01-DK105535; U01-DK085545) [M.I.M., A.L.G.]. The research was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) [A.L.G.]. I.Miguel-Escalada was supported by EFDS/Novo Nordisk Rising Star Programme. Work in J.F. lab was supported by Imperial College London Research Computing Service, NIHR Imperial Biomedical Research Centre (BRC), and CRG genomics facility, and grants from Ministerio de Ciencia e Innovación (BFU2014-54284-R, RTI2018-095666-B-I00), Medical Research Council (MR/L02036X/1), Wellcome Trust Senior Investigator Award (WT101033), 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 for designing the T2DSysTems logo and Laia Codó for the technical support with the website allocation, Isabelle Millard and Anyisha Musuaya from the ULB Center for Diabetes Research for excellent technical and experimental support.

Data Sources:

The database generated in this project has made use from the following list of publicly available resources: The Genotype Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal [GTEx Analysis V7 - Transcript TPMs] on 04/09/19. FastDMA probe full annotation: Wu D, Gu J, Zhang MQ (2013) FastDMA: An Infinium HumanMethylation450 Beadchip Analyzer. PLoS ONE 8(9): e74275. ENCODE (2012-2016) Open Chromatine Dnase: The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome, Nature (2012). ENCODE Data Coordination Center (DCC). The ENCODE Consortium and the ENCODE production laboratory(s) generating the datasets. Gene Ontology: Ashburner et al. Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9. Online at Nature Genetics. GO Consortium, Nucleic Acids Res., 2017. DIAGRAM 1000G GWAS meta-analysis Stage 1 Summary statistics, Trans-ethnic T2D GWAS meta-analysis and DIAMANTE T2D GWAS meta-analysis. DIAGRAM Consortium. Reactome Pathway database: Reactome <https://reactome.org/download-data/> (Jul 2017). DisGeNET, May 2017. Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno,

Javier García-García, Ferran Sanz, and Laura I. Furlong. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucl. Acids Res.* (2016). doi:10.1093/nar/gkw943. GWAS Catalog version 1.0 release 2020-12-02. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington Z, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, and Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 2017, Vol. 45 (Database issue): D896-D901. Ensembl Variant Effect Predictor version 87.27. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biology* Jun 6;17(1):122. (2016). doi:10.1186/s13059-016-0974-4. RefSeq BUILD.37.3: O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016 Jan 4;44(D1):D733-45. Gencode v23 lift 37 annotation: Frankish A, et al (2018) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2018: Oct24. doi:10.1093/nar/gky955. gnomAD version 2.0.2: The authors would like to thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource. A full list of contributing groups can be found at <http://gnomad.broadinstitute.org/about>. Data on glycaemic traits have been contributed by MAGIC investigators (www.magicinvestigators.org). Members of MAGIC are provided in Appendix S1.

Author Contribution

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 ChIP-seq and RNA-seq datasets, and supervised cASE, which was developed and implemented by I.M. during his PhD in IDIBAPS and Imperial College London. I.M. further applied cASE in the TIGER dataset with collaboration of L.A., M.G-M., S.B-G., M.P., R.A. and J.M.M. A.P. performed eQTL and colocalization analyses with 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 collaboration of R.R. and J.M.M. and performed expression analysis with collaboration of I.M. and 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 with islet regulatory data and analysis. I.M., S.B-G. and J.F. contributed with Imperial and CRG data and analysis. J.L.S.E., L.E., H.M. and L.G. contributed with Lund data and analysis. J-V.T., D.L.E. and M.C. contributed

with ULB data and analysis. M.S., L.M. and P.M. contributed with Pisa data and analysis. M.S., L.M., P.M. contributed with Pisa islet samples. J.L.S.E. contributed with Pisa sample sequencing. V.N. contributed with Pisa sample genotyping. J.M.T., V.N. and A.L.G. contributed with Oxford data and analysis and the genotyping of Pisa samples. X.G-H prepared chromatin immuno-precipitation, RNA and DNA samples and managed 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 with MAGIC data and analysis. J.M.M., M.C. and D.T. supervised the study.

References

1. IDF Diabetes Atlas 9th edition 2019. <https://www.diabetesatlas.org/en/>.
2. Khan, M. A. B. *et al.* Epidemiology of Type 2 diabetes - Global burden of disease and forecasted trends. *J. Epidemiol. Glob. Health* **10**, 107–111 (2020).
3. Saeedi, P. *et al.* 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 (2019).
4. Wild, S., Roglic, G., Green, A., Sicree, R. & King, H. Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care* **27**, 1047–1053 (2004).
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. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
7. Bonàs-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun.* **9**, 321 (2018).
8. Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).
9. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
10. Miguel-Escalada, I. *et al.* Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* **51**,

- 593 1137–1148 (2019).
- 594 11. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* vol. 577 179–
595 189 (2020).
- 596 12. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across
597 human tissues. *Science* (80-.). **369**, 1318–1330 (2020).
- 598 13. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature*
599 *Biotechnology* vol. 28 1045–1048 (2010).
- 600 14. Bujold, D. *et al.* The International Human Epigenome Consortium Data Portal. *Cell*
601 *Syst.* **3**, 496-499.e2 (2016).
- 602 15. Krentz, N. A. J. & Gloyn, A. L. Insights into pancreatic islet cell dysfunction from type
603 2 diabetes mellitus genetics. *Nature Reviews Endocrinology* vol. 16 202–212 (2020).
- 604 16. Eizirik, D. L., Pasquali, L. & Cnop, M. Pancreatic β -cells in type 1 and type 2 diabetes
605 mellitus: different pathways to failure. *Nature Reviews Endocrinology* vol. 16 349–362
606 (2020).
- 607 17. Barovic, M. *et al.* Metabolically phenotyped pancreatectomized patients as living
608 donors for the study of islets in health and diabetes. *Molecular Metabolism* vol. 27
609 S1–S6 (2019).
- 610 18. Burgarella, S., Merlo, S., Figliuzzi, M. & Remuzzi, A. Isolation of Langerhans islets by
611 dielectrophoresis. *Electrophoresis* **34**, 1068–1075 (2013).
- 612 19. Meier, D. T. *et al.* Determination of Optimal Sample Size for Quantification of β -Cell
613 Area, Amyloid Area and β -Cell Apoptosis in Isolated Islets. *J. Histochem. Cytochem.*
614 **63**, 663–673 (2015).
- 615 20. The Pancreas and Its Functions | Columbia University Department of Surgery.
616 <https://columbiasurgery.org/pancreas/pancreas-and-its-functions>.
- 617 21. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic
618 islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U.*
619 *S. A.* **111**, 13924–9 (2014).
- 620 22. van de Bunt, M. *et al.* Transcript Expression Data from Human Islets Links Regulatory
621 Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic
622 Traits to Their Downstream Effectors. *PLOS Genet.* **11**, e1005694 (2015).

- 623 23. Viñuela, A. *et al.* Genetic variant effects on gene expression in human pancreatic
624 islets and their implications for T2D. *Nat. Commun.* **11**, 1–14 (2020).
- 625 24. Guindo-Martínez, M. *et al.* The impact of non-additive genetic associations on age-
626 related complex diseases. *Nat. Commun.* **12**, 2436 (2021).
- 627 25. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–
628 65 (2012).
- 629 26. Boomsma, D. I. *et al.* The Genome of the Netherlands: Design, and project goals.
630 *Eur. J. Hum. Genet.* **22**, 221–227 (2014).
- 631 27. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation.
632 *Nat. Genet.* **48**, 1279–1283 (2016).
- 633 28. Consortium, T. U. The UK10K project identifies rare variants in health and disease.
634 *Nature* **526**, 82–90 (2015).
- 635 29. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of
636 expression residuals (PEER) to obtain increased power and interpretability of gene
637 expression analyses. *Nat. Protoc.* **7**, 500–7 (2012).
- 638 30. Benjamini, Yoav and Hochberg, Y. Controlling the False Discovery Rate: A Practical
639 and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* (1995).
- 640 31. Kundu, K. *et al.* Genetic associations at regulatory phenotypes improve fine-mapping
641 of causal variants for twelve immune-mediated diseases. *bioRxiv* 2020.01.15.907436
642 (2020) doi:10.1101/2020.01.15.907436.
- 643 32. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat.*
644 *Genet.* **49**, 692–699 (2017).
- 645 33. Flannick, J. The Contribution of Low-Frequency and Rare Coding Variation to
646 Susceptibility to Type 2 Diabetes. *Current Diabetes Reports* vol. 19 (2019).
- 647 34. Ramos-Rodríguez, M. *et al.* The impact of proinflammatory cytokines on the β -cell
648 regulatory landscape provides insights into the genetics of type 1 diabetes. *Nat.*
649 *Genet.* **51**, 1588–1595 (2019).
- 650 35. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic
651 Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
- 652 36. Chen, J. *et al.* The Trans-Ancestral Genomic Architecture of Glycaemic Traits. *bioRxiv*

- 653 **14**, 203 (2020).
- 654 37. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants
655 associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298
656 (2014).
- 657 38. Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature*
658 *Genetics* vol. 32 490–495 (2002).
- 659 39. Akey, J. M., Biswas, S., Leek, J. T. & Storey, J. D. On the design and analysis of gene
660 expression studies in human populations [1]. *Nature Genetics* vol. 39 807–808 (2007).
- 661 40. Fare, T. L. *et al.* Effects of atmospheric ozone on microarray data quality. *Anal. Chem.*
662 **75**, 4672–4675 (2003).
- 663 41. Branham, W. S. *et al.* Elimination of laboratory ozone leads to a dramatic
664 improvement in the reproducibility of microarray gene expression measurements.
665 *BMC Biotechnol.* **7**, 8 (2007).
- 666 42. Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite
667 method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*
668 **30**, e15 (2002).
- 669 43. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat.*
670 *Methods* **2**, 345–349 (2005).
- 671 44. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by
672 surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
- 673 45. Newhall, R. A. *et al.* The American Soldier: Adjustment During Army Life. Volume I.
674 *Mississippi Val. Hist. Rev.* **36**, 339 (1949).
- 675 46. Akerman, I. *et al.* Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory
676 Networks. *Cell Metab.* **25**, 400–411 (2017).
- 677 47. Hall, E. *et al.* Sex differences in the genome-wide DNA methylation pattern and
678 impact on gene expression, microRNA levels and insulin secretion in human
679 pancreatic islets. *Genome Biol.* **15**, 522 (2014).
- 680 48. Turner, M. *et al.* Integration of human pancreatic islet genomic data refines
681 regulatory mechanisms at Type 2 Diabetes susceptibility loci. *Elife* **7**, (2018).
- 682 49. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nat. Genet.*

683 **42**, 255–259 (2010).

684 50. Stitzel, M. L. *et al.* Global epigenomic analysis of primary human pancreatic islets
685 provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* **12**, 443–455
686 (2010).

687 51. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome.
688 *Nature* **489**, 57–74 (2012).

689 52. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes
690 in Europeans. *Diabetes* **66**, 2888–2902 (2017).

691 53. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into
692 the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244
693 (2014).

694 54. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse
695 genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

696 55. Saponaro, C. *et al.* 1900-P: HNF1A Deficiency Leads to Perturbed Glucagon
697 Secretion in Humans. *Diabetes* **69**, 1900-P (2020).

698 56. Saponaro, C. *et al.* Interindividual heterogeneity of SGLT2 expression and function in
699 human pancreatic islets. *Diabetes* **69**, 902–914 (2020).

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

702 58. Lyssenko, V. *et al.* Common variant in MTNR1B associated with increased risk of type
703 2 diabetes and impaired early insulin secretion. *Nat. Genet.* **41**, 82–88 (2009).

704 59. Osonoi, S., Ichinohe, H., Kudo, K., Yagihashi, S. & Mizukami, H. 2047-P: Possible
705 Implication of Cyclin D2 in Beta-Cell Proliferation of Human Perinatal Islet. *Diabetes*
706 **69**, 2047-P (2020).

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

709 61. Kaur, S., Mirza, A. H. & Pociot, F. Cell type-selective expression of circular RNAs in
710 human pancreatic islets. *Non-coding RNA* **4**, (2018).

711 62. Abderrahmani, A. *et al.* Complexin I regulates glucose-induced secretion in pancreatic
712 β -cells. *J. Cell Sci.* **117**, 2239–2247 (2004).

713 63. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2
714 diabetes. *Nature* **445**, 881–885 (2007).

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

718 65. Carvalho, S. *et al.* Differential cytolocation and functional assays of the two major
719 human SLC30A8 (ZnT8) isoforms. *J. Trace Elem. Med. Biol.* **44**, 116–124 (2017).

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

724 67. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and
725 24,440 controls. *Nature* **570**, 71–76 (2019).

726 68. Dwivedi, O. P. *et al.* Loss of ZnT8 function protects against diabetes by enhanced
727 insulin secretion. *Nat. Genet.* **51**, 1596–1606 (2019).

728

729

Tables.

Table 1. Novel human pancreatic islet colocalization of expression quantitative trait loci meta-analysis (eQTL) with type 2 diabetes (T2D) genome-wide association studies (GWAS).

Chr	SNP	Gene	COLOC		T2D GWAS					eQTL	
			PP.H4.abf	SNP.PP.H4	EA	EA	NEA	OR	P-value	P-value	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-07	----
1	rs1493694	NBPF7	0,81	0,09	0,11	T	C	1,09	2,1E-16	1,0E-05	?+?+
1	rs340874	RP11-478J18.2	0,98	1,00	0,56	C	T	1,07	5,6E-26	1,3E-06	++++
1	rs4659836	TBCE	0,82	0,12	0,65	A	G	1,04	4,7E-09	2,9E-07	----
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-09	++++
3	rs7640294	SERBP1P3	0,97	0,06	0,56	A	C	1,04	3,0E-08	1,6E-09	++++
4	rs1531583	CPLX1	0,87	0,13	0,046	T	G	1,12	1,2E-12	1,2E-06	++++
4	rs1580278	BDH2	0,81	0,73	0,53	A	C	0,96	2,9E-10	1,1E-09	++++
4	rs58730668	ACSL1	0,89	0,04	0,14	C	T	0,93	1,0E-13	2,5E-05	++++
6	rs6557267	RGS17	0,94	0,08	0,42	T	C	1,04	6,0E-08	8,2E-08	----
8	rs1059592	RP11-582J16.5	0,81	0,12	0,35	A	G	1,03	4,5E-05	4,1E-15	----
8	rs77292833	LRP12	0,84	0,05	0,12	G	C	0,96	1,6E-05	8,1E-08	++++
9	rs10811660	CDKN2B-AS1	0,99	0,48	0,17	A	G	0,85	6,6E-79	1,6E-07	----
9	rs10963924	SAXO1	0,82	0,09	0,43	C	G	1,04	9,2E-10	1,6E-05	----
10	rs827237	PCBD1	0,99	0,19	0,21	T	C	1,04	2,3E-07	2,4E-10	----
11	rs15818	HMBS	0,84	0,06	0,4	G	A	1,03	4,5E-05	2,5E-07	++++
11	rs529623	FXYD2	0,92	0,83	0,52	C	T	0,97	5,8E-06	3,4E-07	++++
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-05	3,0E-11	++++
12	rs76895963	CCND2	0,36	1,00	0,02	G	T	0,62	5,3E-70	1,7E-06	+++?
12	rs77864822	RMST	0,99	0,81	0,07	G	A	0,93	2,2E-08	2,9E-14	++++
12	rs77864822	RP11-528M18.2	0,95	0,17	0,07	G	A	0,93	2,2E-08	3,6E-06	+++
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-06	++++
14	rs10151752	ACTR10	0,86	0,26	0,59	G	A	0,97	7,2E-08	4,0E-06	++++
14	rs1803283	RP11-600F24.7	0,81	0,02	0,65	T	C	1,04	1,4E-07	2,5E-05	-+--
15	rs13737	RP11-817O13.8	0,84	0,10	0,24	T	G	0,96	7,3E-10	2,3E-06	++++
17	rs7218899	USP36	0,96	0,41	0,51	T	C	0,97	1,5E-06	2,4E-10	++++
17	rs8070260	ZNHIT3	0,94	0,13	0,53	G	A	0,97	1,1E-05	4,1E-08	----
18	rs303760	NPC1	0,95	0,08	0,36	T	C	1,03	3,8E-06	2,4E-24	----

Colocalizations not reported in Viñuela et al.²³ 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 type 2 diabetes; the effect alleles (*EA*) with a frequency (*EA*) is shown with the associated effect odd-ratio (*OR*) and the *p-value*. The *GWAS* data is as reported by the *DIAGRAM* consortium⁵. The eQTL *p-value* is reported with the direction of the effect: up- ('+') or down-regulation ('-') *direction* for the effect allele in the four meta-analysis cohorts (order: CRG, Oxford, Lund and Pisa). '?' means that not enough samples are available in the cohort for the minor allele in order to compute a *p-value*.

Table 2. Colocalization of allele specific expression (*cASE*) with type 2 diabetes (T2D) genome-wide association study (GWAS).

Chr	SNP	Gene	COLOC		T2D GWAS					cASE				
			PP.H4.abf	SNP.PP.H4	EAF	EA	NEA	OR	P-value	Reporter variant	Ref	Alt	P-value	Z-score
1	rs1127215	PTGFRN	0.99	0.98	0.42	T	C	0.95	2.3E-13	rs1127656	C	T	8.5E-09	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	0.48c	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-06	-11.1
11	rs529623	FXVD2	0.95	1.00	0.52	C	T	0.97	5.8E-06	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-06	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-07	13.4
18	rs1788762	C18orf8	0.96	0.06	0.64	C	G	0.97	2.3E-06	rs1788820	A	G	3.2E-25	-26.7
18	rs1788762	NPC1	0.96	0.06	0.64	C	G	0.97	2.3E-06	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 type 2 diabetes; the effect alleles (*EA*) with a frequency (*EAF*) is shown with the associated effect odd-ratio (*OR*) and the *p-value*. The GWAS data is as reported by the *DIAGRAM* consortium⁵. 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

Figure legends.

Figure 1: Project overview and genotype imputation. **A)** Overview of the TIGER data portal. **B)** Datasets of the T2DSys consortium and project workflow. **C)** Multi-panel genotype imputation identified 13.1-15.7M autosomal variants (top) and 550-700k chrX variants (bottom), with **D)** a large proportion of low frequency (MAF 1-5%) and rare (<1%) variants, including 10.2% of Structural Variants (SVs), including small indels and large SVs.

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 eQTLs novel in TIGER human islets (green) and previously found in GTEx project: tissues related to T2D aetiology (orange), other tissues (blue); means in dashed lines. Right panel restricted to low minor allele frequency (MAF) variants only. **E)** Gene ontology analysis of the genes of TIGER-specific eQTLs.

Figure 3: Examples of co-localization of pancreatic islets 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).


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 also eQTL significant increased 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 by both (orange).

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 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 by 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 co-localization (right). ABF: Approximate Bayes Factor, PP: Posterior Probability.

Figure 6: TIGER platform example. A) *MTNR1B* normalized log₁₀(TPM) expression in islets; table (top) displays *MTNR1B* normalized TPM expression in each cohort and across the cohorts (bold); histogram (bottom) shows log₁₀(TPM) gene expression distribution in 495 human islets samples, the red dashed line corresponds to *MTNR1B* log₁₀(TPM) expression. B) *MTNR1B* normalized TPM expression in islets vs other GTEx tissues where each boxplot represents one tissue; *MTNR1B* has higher expression in pancreatic islets (black) compared to the whole pancreas (brown), which has almost no expression. C) Table showing the list of variants in a 100Kb window around *MTNR1B* and displaying results from either eQTL or DIAMANTE GWAS data sorted by ascending eQTL *p*-value; the eQTL variant rs10830963 ($p=4.04 \times 10^{-19}$) colocalizes with DIAMANTE ($p=1.50 \times 10^{-43}$). D) 15Kb human islet genomic context of variant rs10830963 (chr11:92708710); islet significant regions (black/blue boxes) and peaks are represented in each track, the blue line corresponds to rs10830963 position.

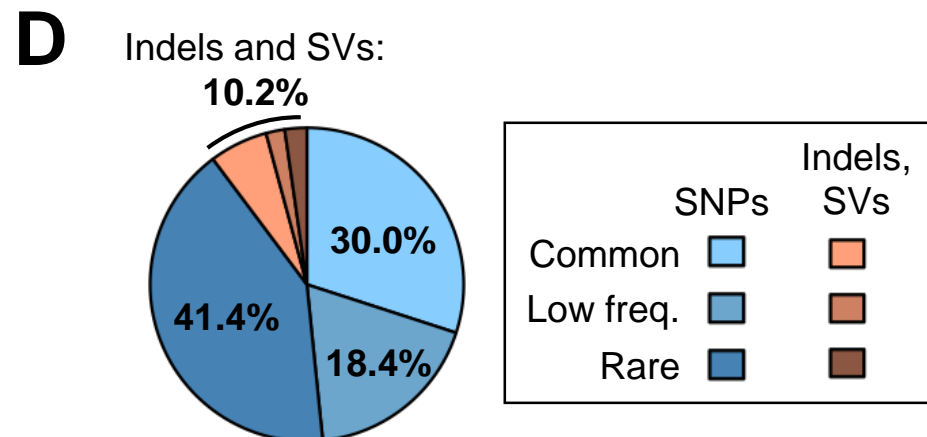
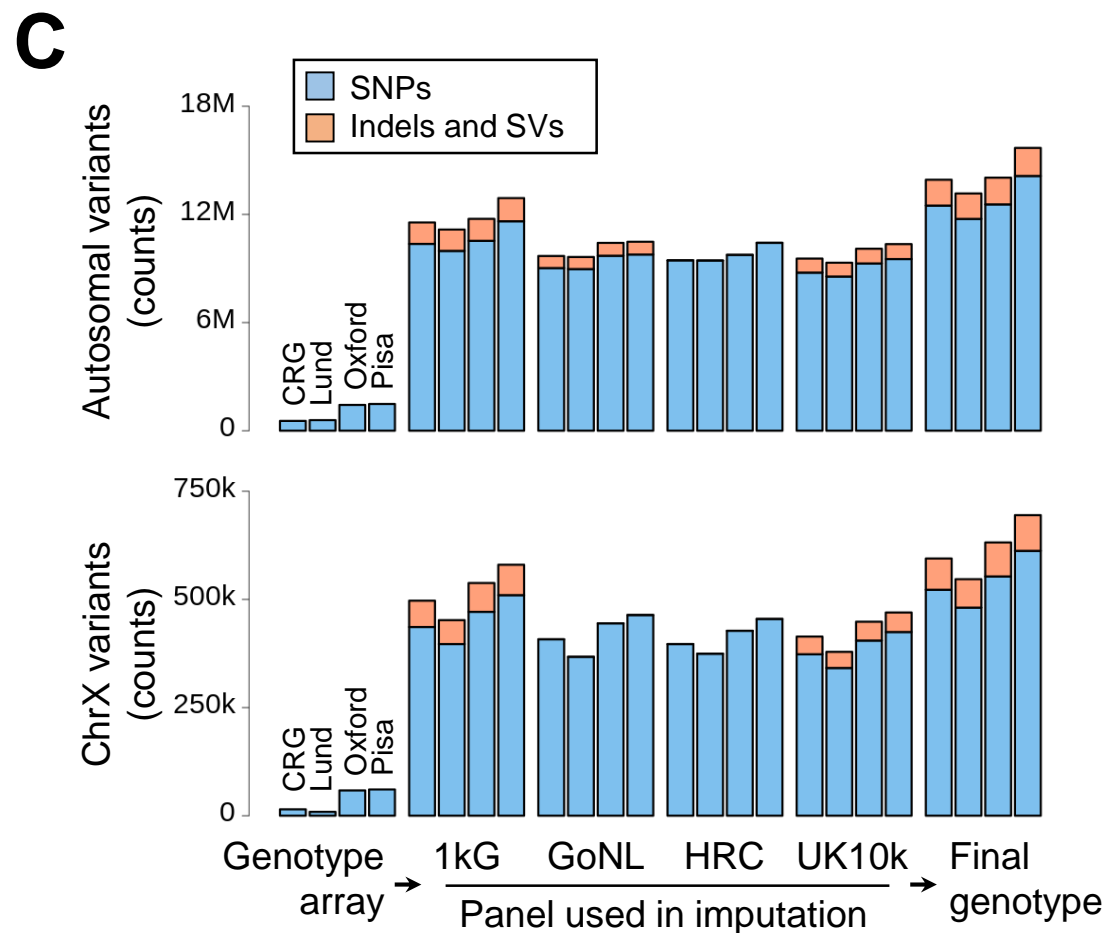
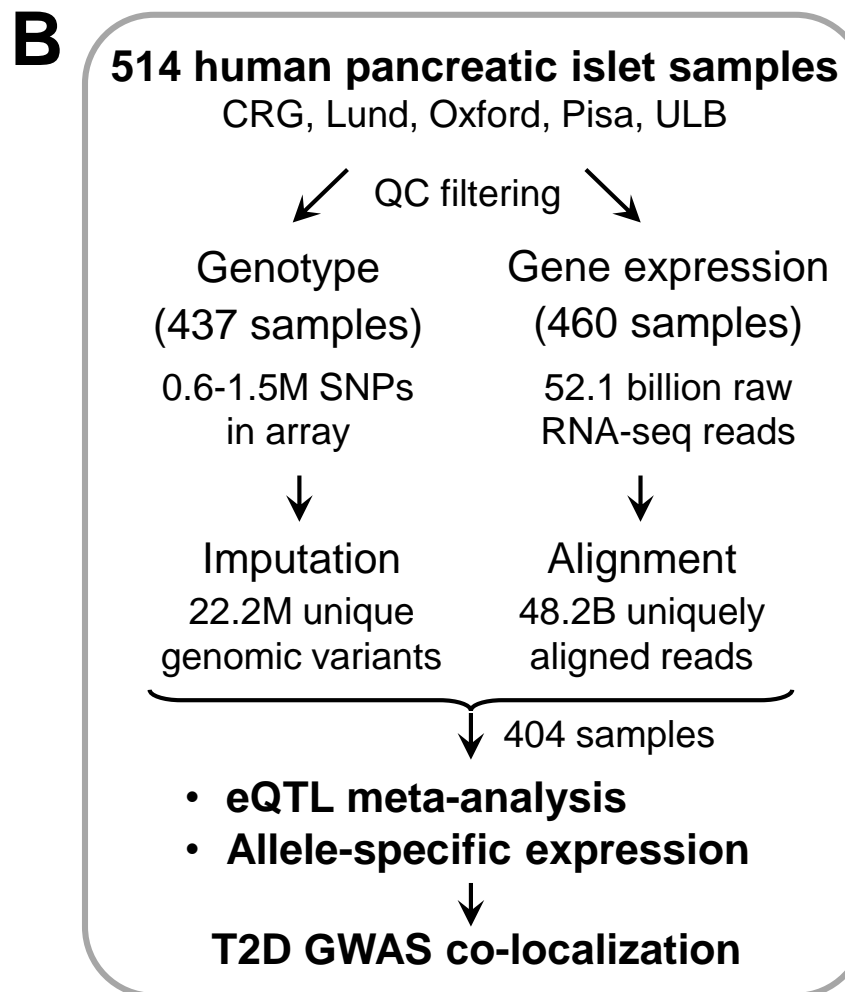
A

 **TIGER Data Portal** tiger.bsc.es

TIGER *cis*-regulation and GWAS co-localization results

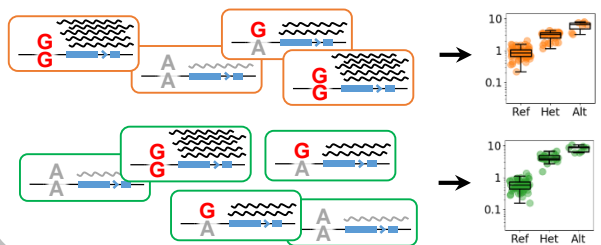
Epigenetic data
(ChIP-seq, ATAC-seq, DNA-methylation)

Public datasets
(GWAS, GTEx, pathway, GO)

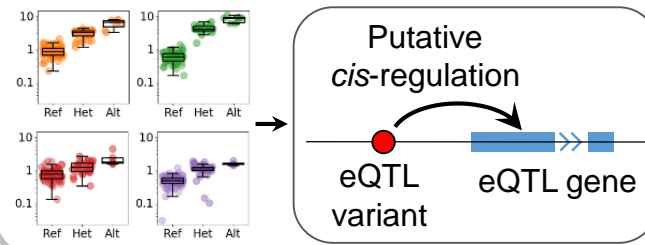
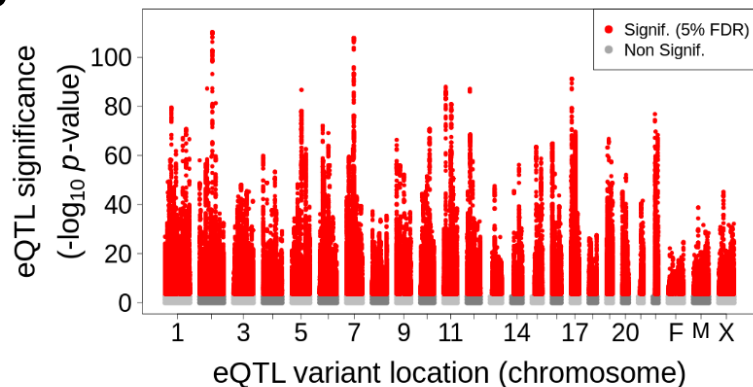
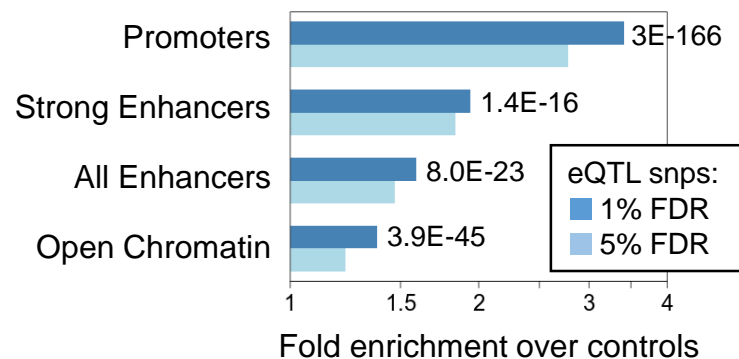
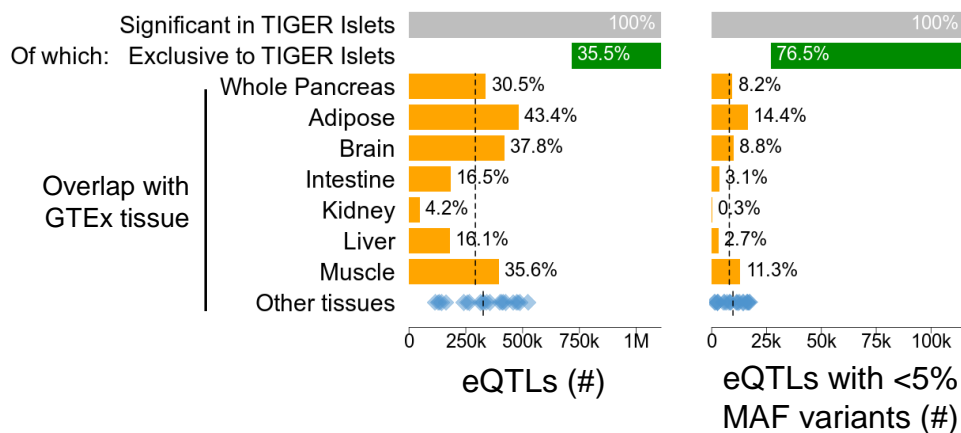
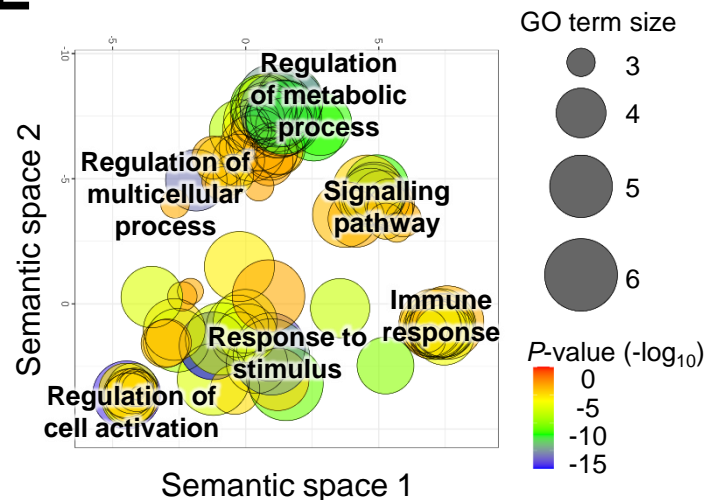


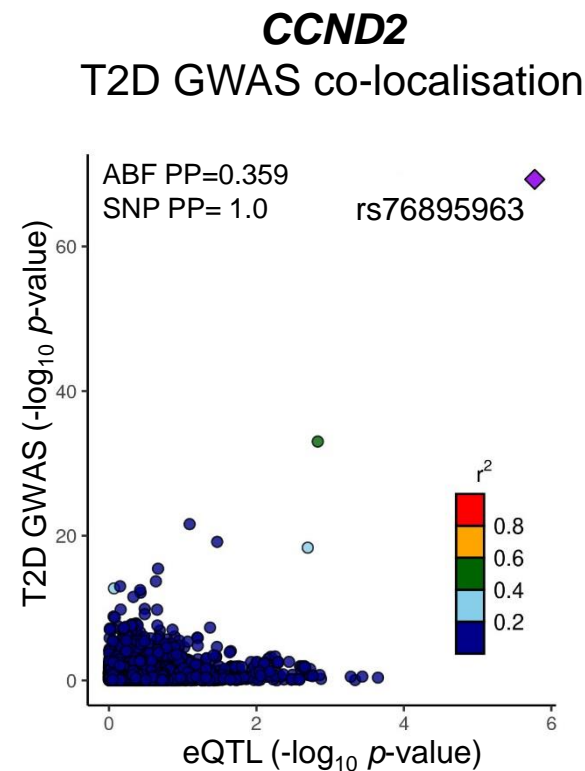
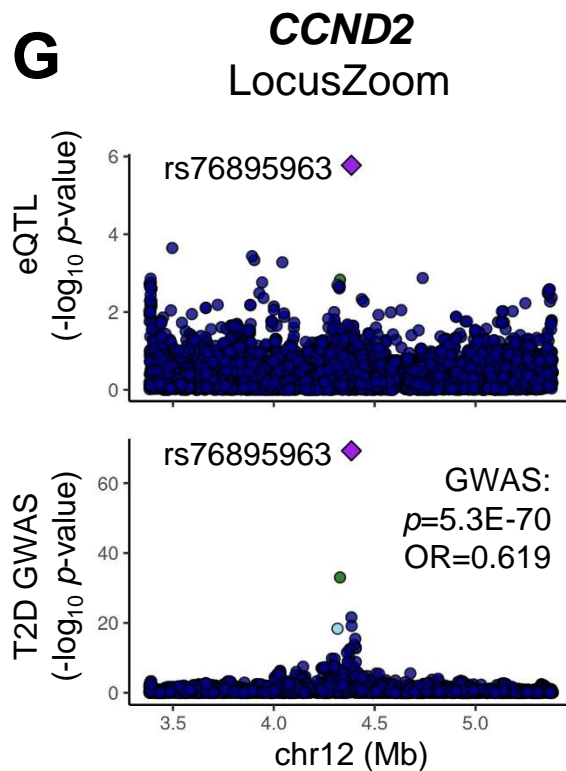
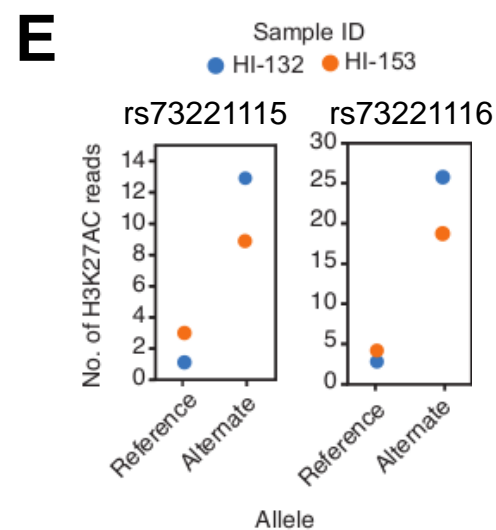
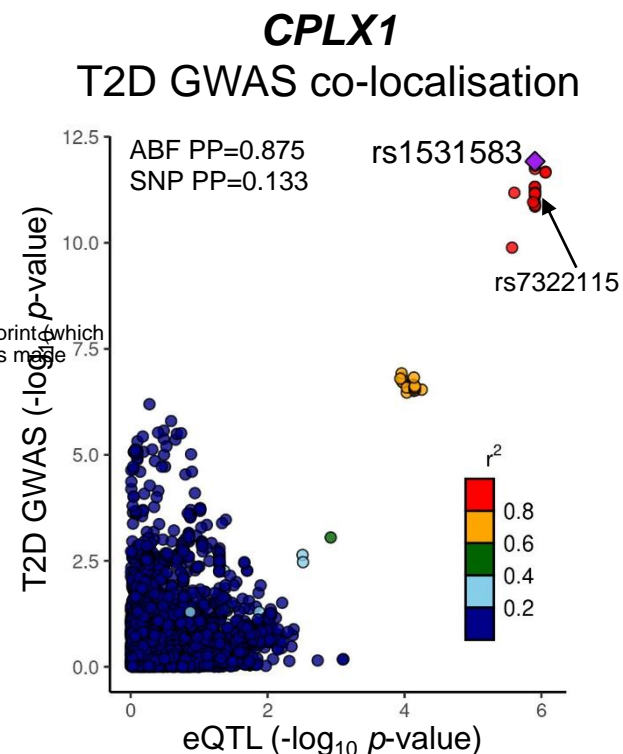
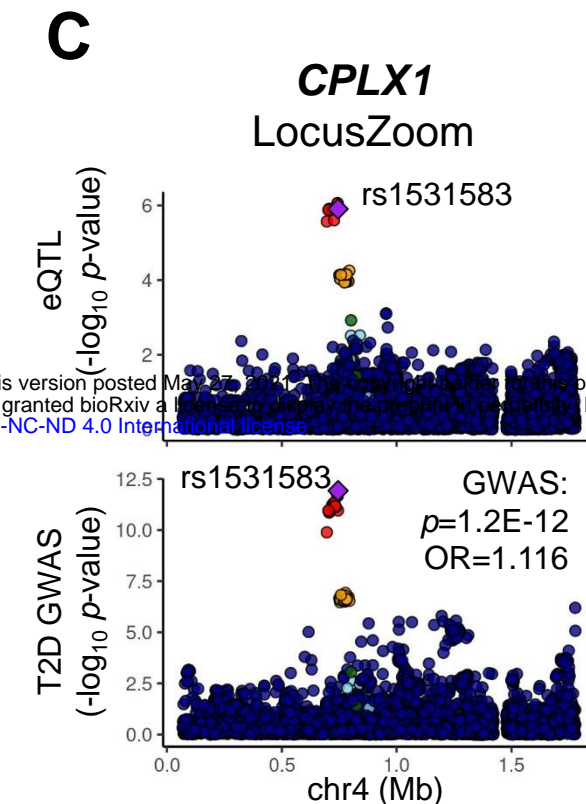
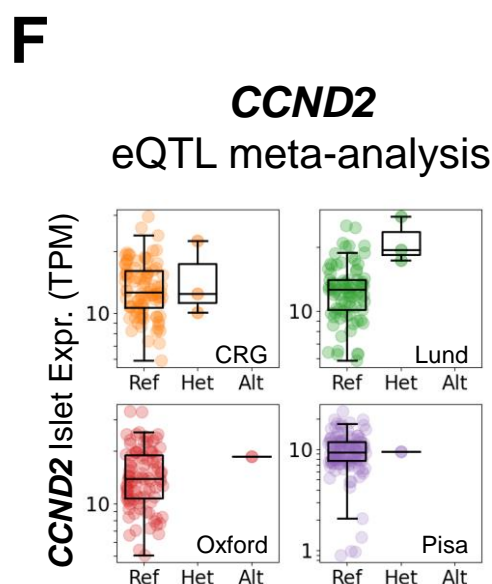
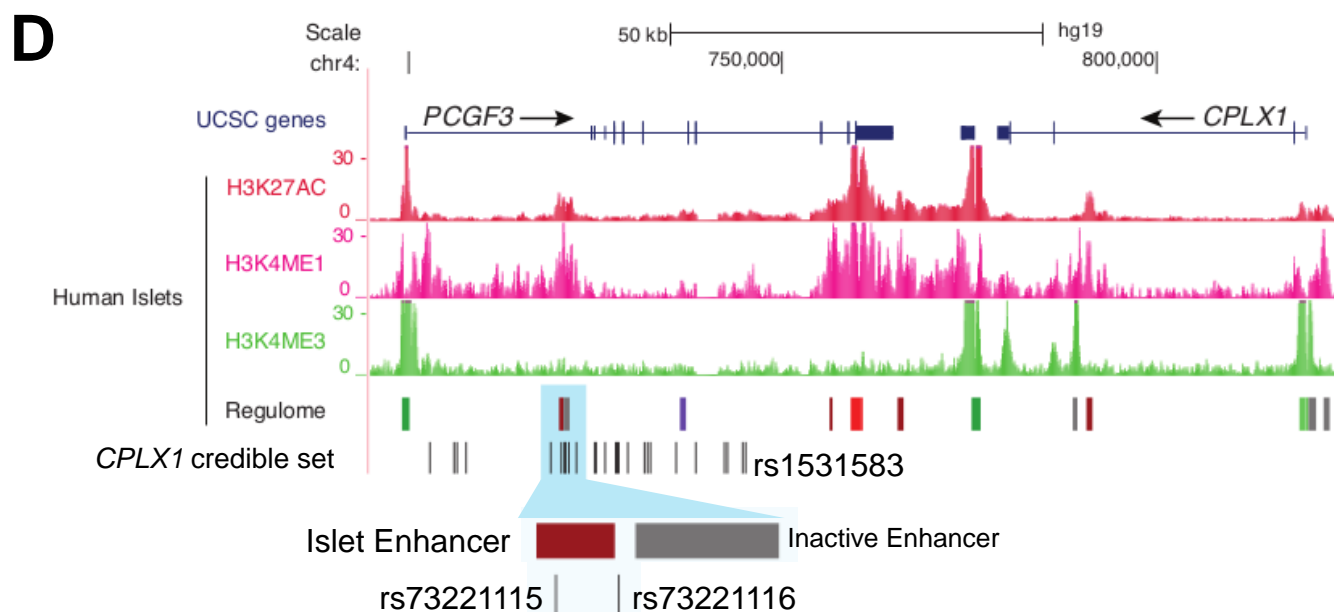
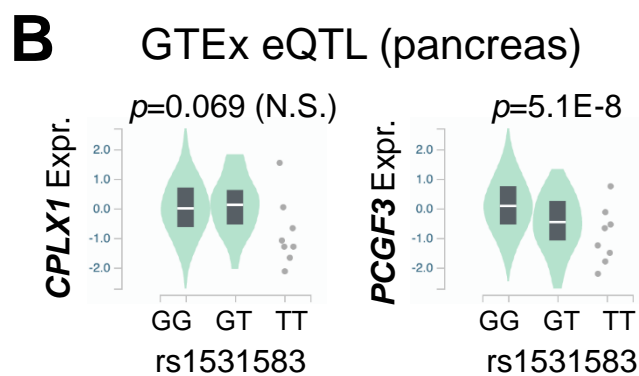
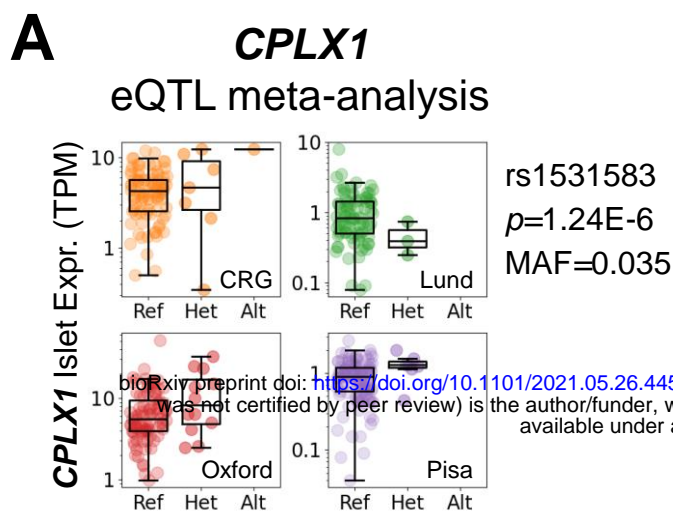
A

Gene expression quantification and eQTL analysis in each cohort



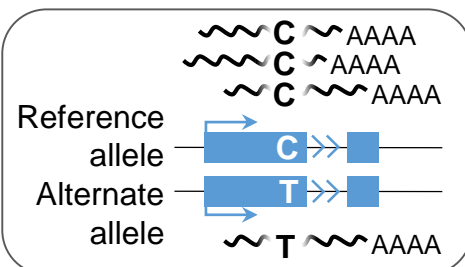
Meta-analysis of all cohorts to uncover genomic *cis*-regulation

**B****C****D****E**



A

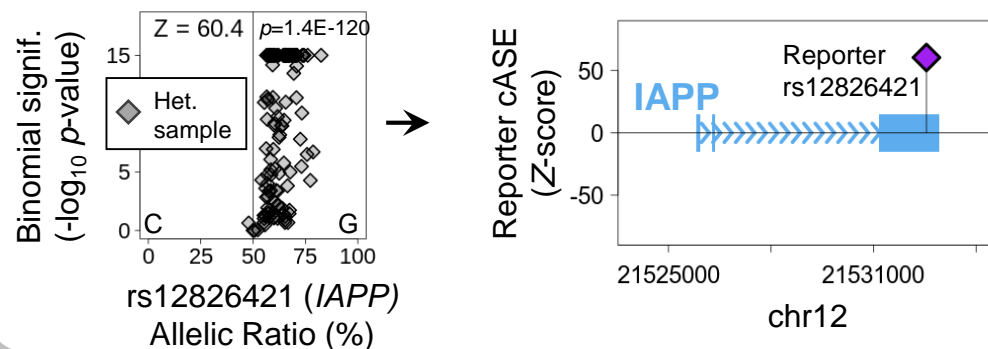
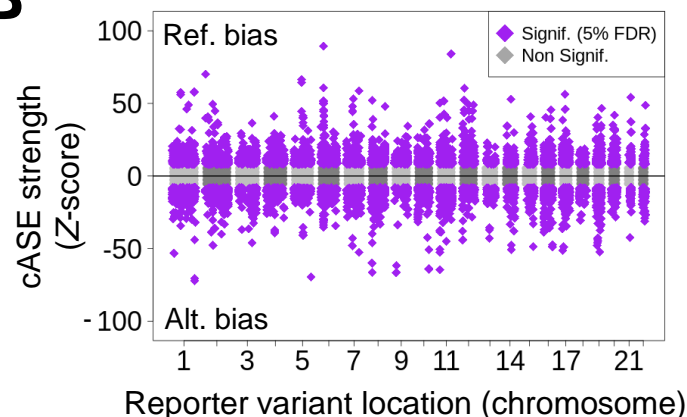
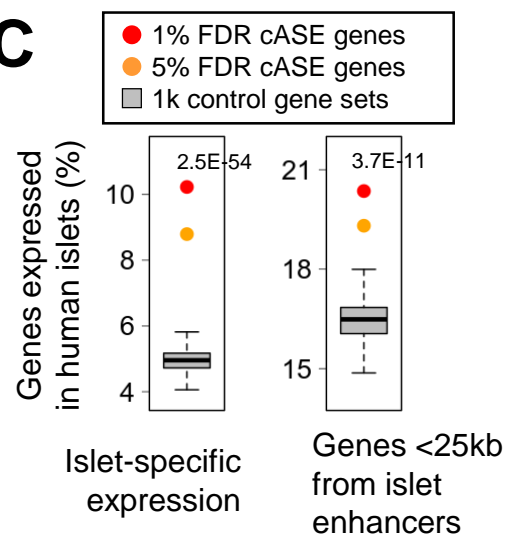
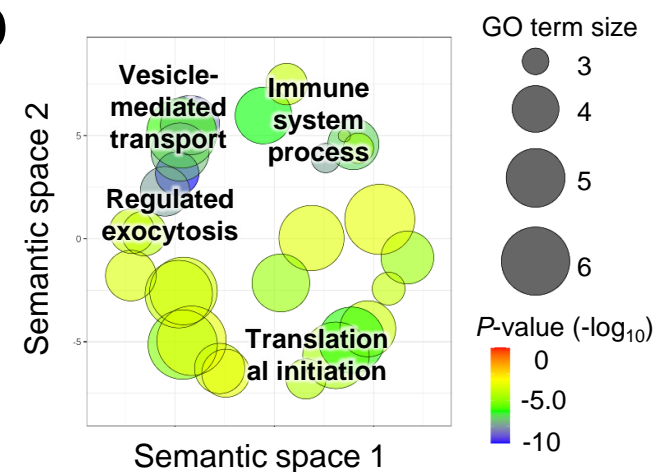
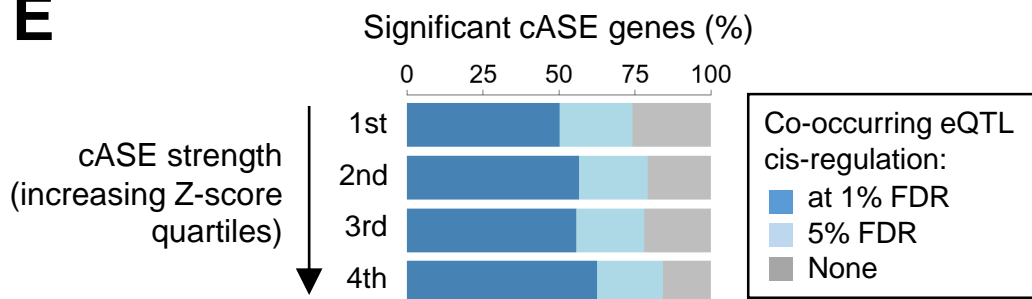
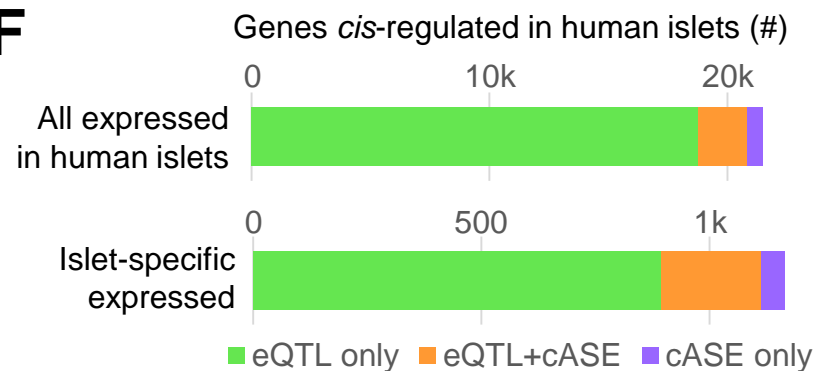
Allele-specific expressed (ASE) genes in individual samples

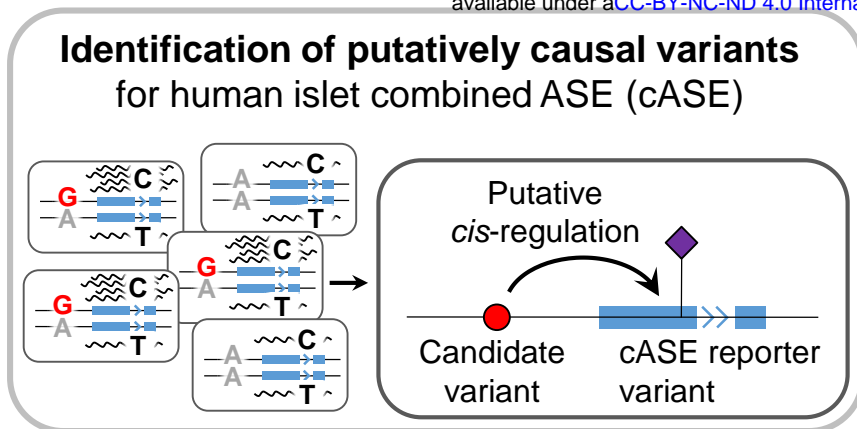
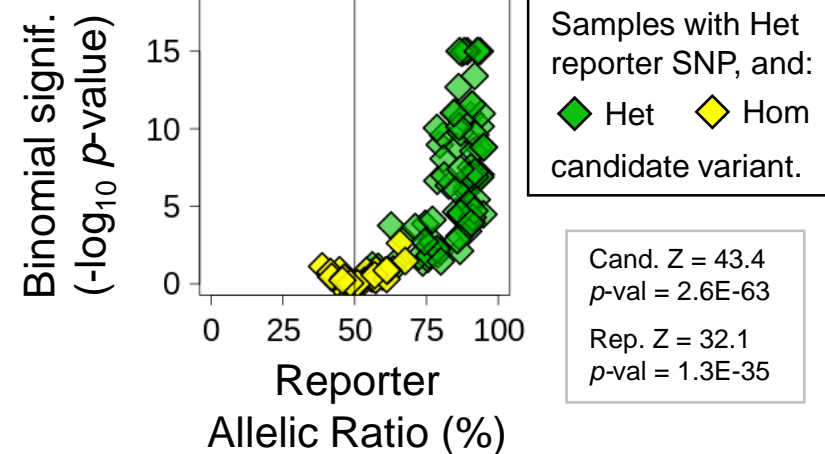
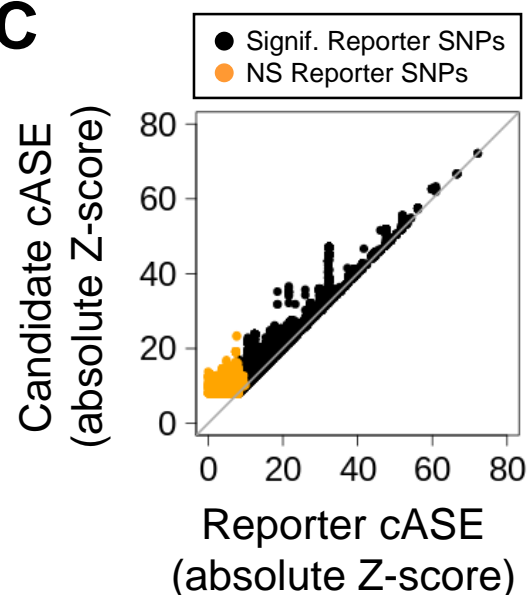
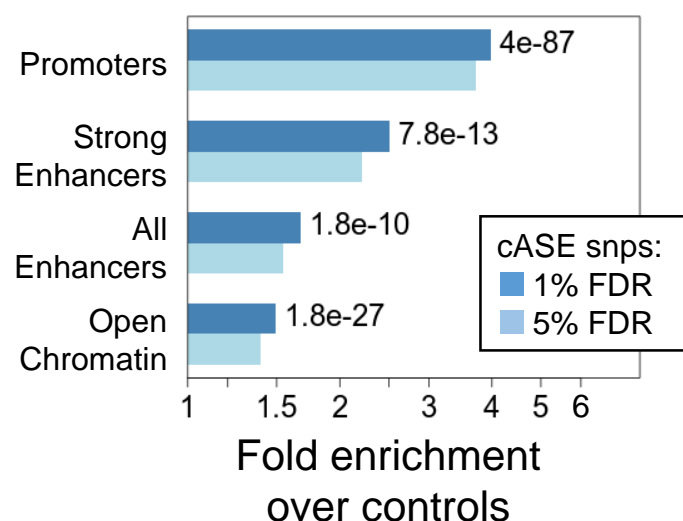
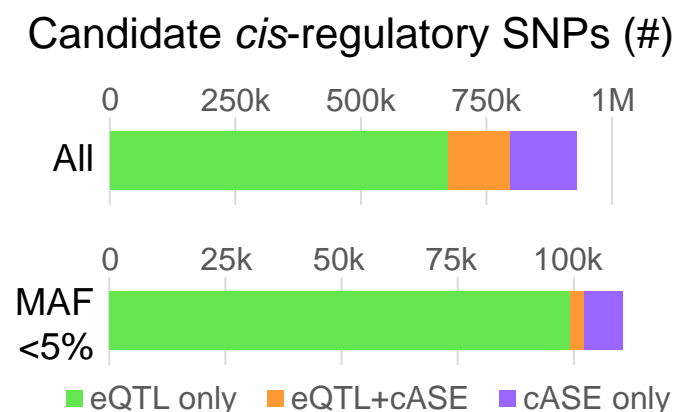
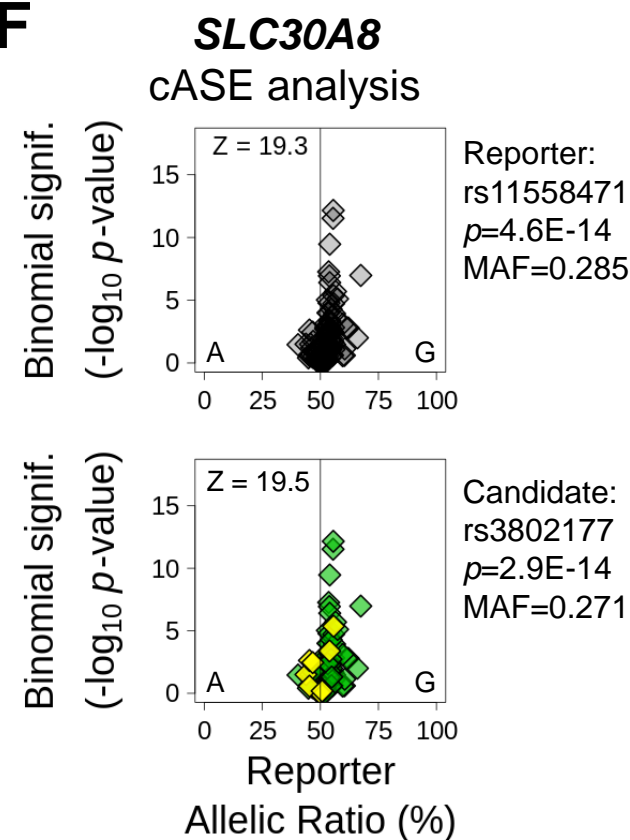
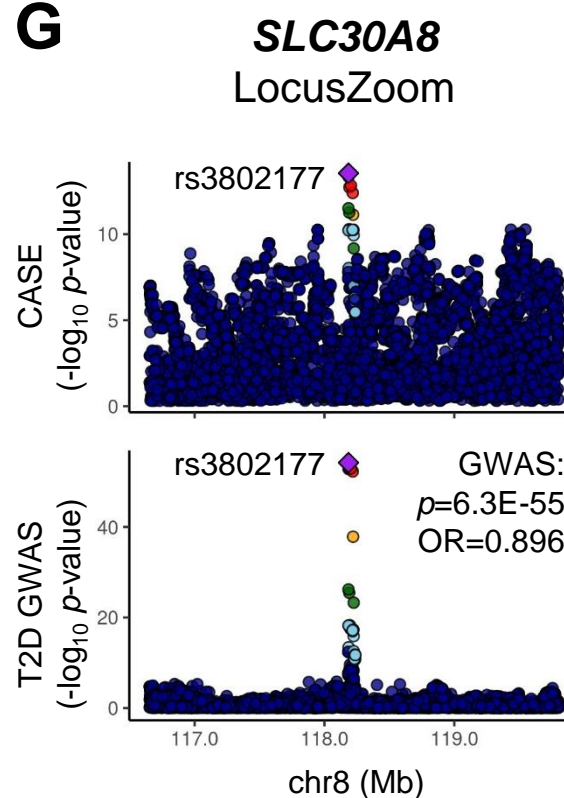


Individual gene expression bias measured by:

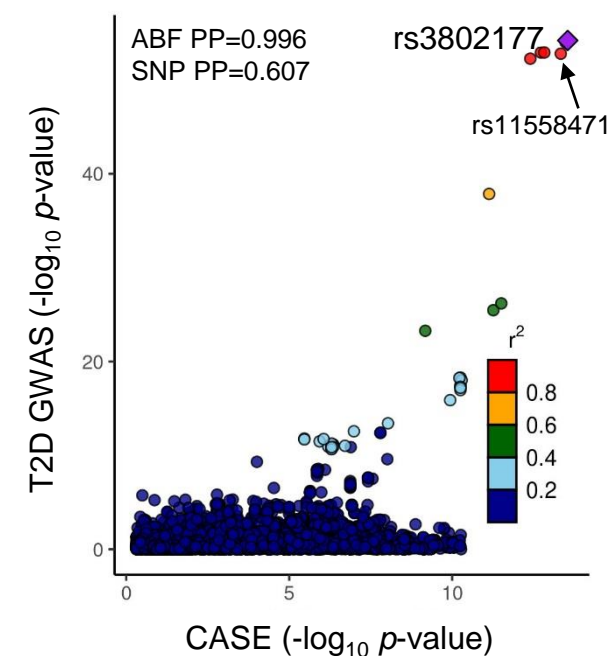
- **Allelic ratio (AR):** % of RNA-seq reads containing the Ref. allele.
- **Statistical significance:** **Binomial p -value.**

Combined allele specific expressed (cASE) genes in ≥ 3 independent samples

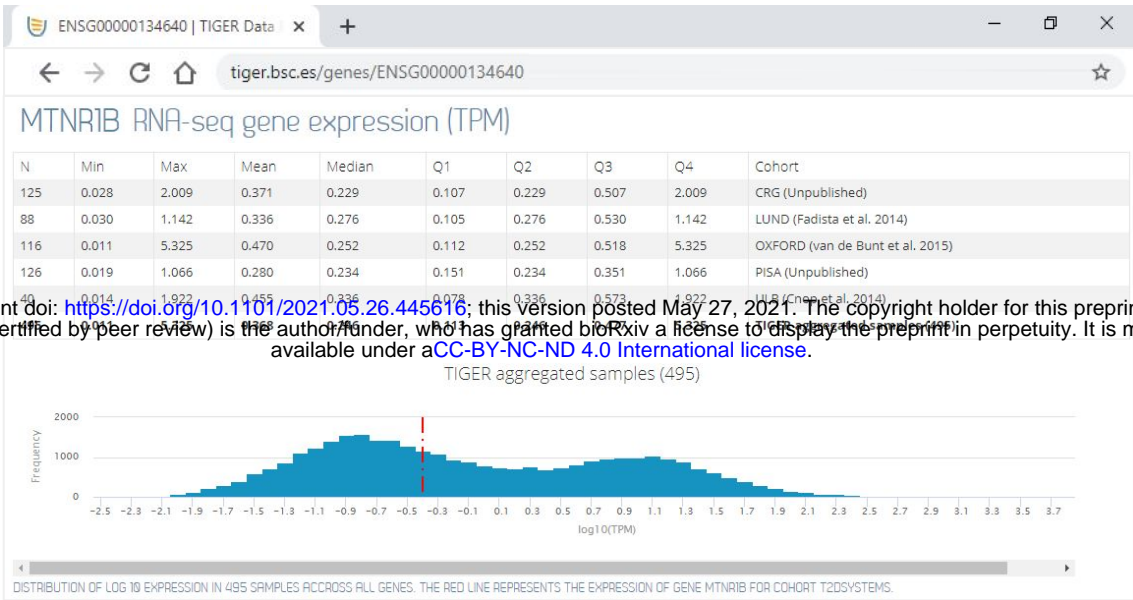
**B****C****D****E****F**

A**B****C****D****E****F****G**

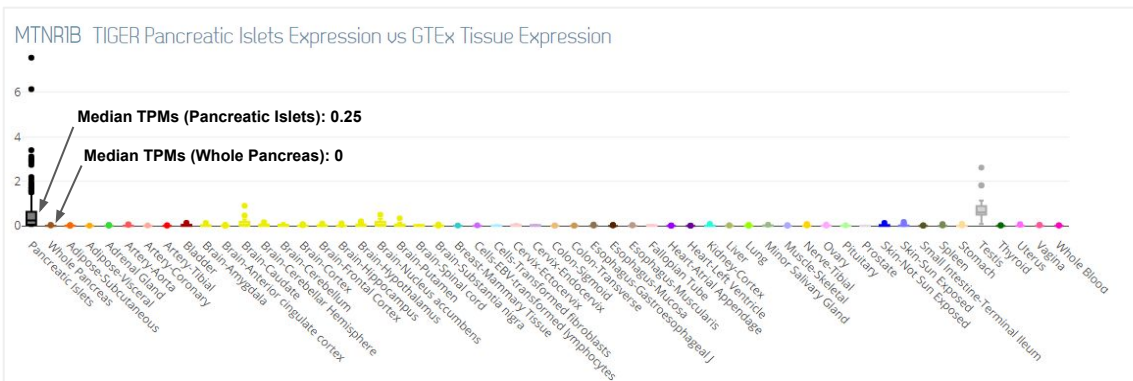
SLC30A8
T2D GWAS co-localisation



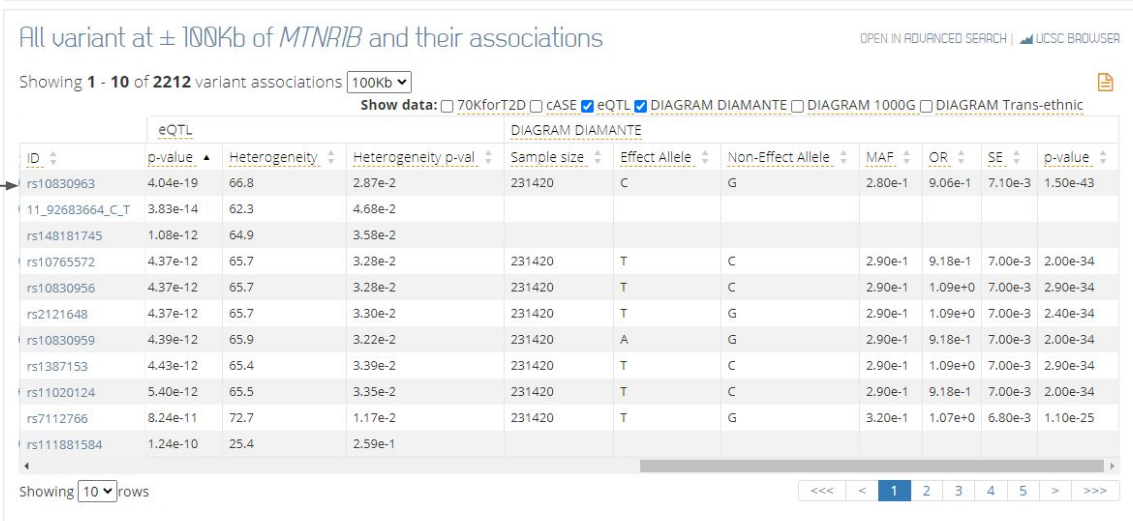
A



B



C



D

