

A Transcriptome-Wide Association Study (TWAS) Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer

Jun Zhong et.al.

Supplementary Material: Supplementary Methods, Data Access, Tables, Figures, Acknowledgements and Funding Information

Supplementary Methods

Tissue samples, transcriptome datasets and quality control (QC)

For the Laboratory of Translational Genomics (LTG) dataset, 1 μg RNA (RIN scores >7.5) isolated from 95 fresh frozen histologically normal pancreatic tissue samples (all from individuals of European ancestry) with the Ambion mirVana kit, underwent massively parallel sequencing at the National Cancer Institute's CCR Sequencing Facility as previously described. The project was approved by the Institutional Review Boards of each participating institution as well as the NIH. At each participating institution, samples were confirmed to be non-tumorous and contain $\geq 80\%$ epithelial component by histological review by a pathologist and macro-dissected when needed (1). Briefly, RNA sequencing was performed on the Illumina HiSeq 2000 sequencing platform using TruSeq v3 chemistry for paired-end sequencing. The average sequence depth was ~ 300 million mapped reads per sample (1). Alignment to the human reference genome GRCh37/hg19 was performed using STAR v2.4.2a (2) as per the GTEx pipeline, based on the GENCODE v19 gene annotations. Gene-level expression quantification was collapsed to a single transcript model for each gene using an isoform collapsing procedure, comprising the following steps as in the GTEx pipeline (3): (i) exons associated with transcripts annotated as "retained_intron" and "read_through" were excluded; (ii) exon intervals overlapping within a gene were merged; (iii) the intersections of exon intervals overlapping between genes were excluded; (iv) the remaining exon intervals were mapped to their respective gene identifier and stored in a GTF format. After converting gene read counts to transcripts per million (TPM) (4), genes were included in the analysis based on expression thresholds > 0.1

TPM in ≥ 10 samples and > 5 reads in ≥ 10 samples. The expression data were normalized as follows: (i) expression values were quantile-normalized to the average empirical distribution observed across samples; and (ii) for each gene, expression values were inverse quantile-normalized to a standard normal distribution across samples.

For the Genotype-Tissue Expression (GTEx) dataset, RNA-sequencing was performed on the Illumina HiSeq 2000/2500 system, generating gene read counts for 11,688 tissue samples derived from 635 individuals through a rapid autopsy program, as previously described (average sequence depth ~ 100 million mapped reads per sample) (3). Gene read counts for 220 pancreatic tissue samples were obtained through controlled access (phs000424.v7.p2). Based on our ancestry analysis (described below), 174 subjects of European ancestry were available for further analysis. QC and normalization procedures were performed in the same manner as described above for the LTG dataset.

The LTG and GTEx pancreatic transcriptome datasets were combined (at the TPM level) for genes with matching Ensembl gene IDs; genes expressed (at the threshold described above) in only one of the datasets were excluded. We further performed QC and normalization for the combined dataset ($n = 269$) using the approach described above.

Genotype data and quality control

DNA samples for the LTG pancreatic tissue dataset were isolated from blood (Mayo Clinic), histologically normal fresh-frozen pancreatic tissue samples (Penn State) or histologically normal fresh-frozen spleen or duodenum tissue samples (Memorial Sloan Kettering) using the Gentra Puregene Tissue Kit (Qiagen), and genotyped on the Illumina OmniExpress or Omni1M array at the Cancer Genomics Research Laboratory of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH (1). Ancestry was assessed using the GLU struct.admix module (5) and samples with $< 80\%$ European ancestry were excluded. The following quality-control metrics were applied prior to imputation: SNPs with call rates $< 95\%$, minor allele frequency (MAF) < 0.01 or Hardy Weinberg Equilibrium (HWE) $P < 1 \times 10^{-6}$ were excluded using VCFtools; after checking strand, alleles, position, reference/alternative assignments and frequency differences with 1000 Genomes reference panel (1KG, Phase3, version 80) (6), SNPs with allele frequency differences > 0.2 between our data and 1KG, those not available in the 1KG panel, or A/T and G/C variants on ambiguous DNA strand (MAF > 0.4) were excluded.

Samples with call rates $< 90\%$ were excluded using PLINK (7). After genotype quality control, genotypes were imputed using the 1000 Genomes imputation reference dataset via the Michigan Imputation Server (8). Post-imputation variants with MAF < 0.05 , imputation quality score (R^2) < 0.5 or duplicated variants were removed by PLINK (7).

DNA isolation protocols and sample quality-control metrics for the Genotype-Tissue Expression (GTEx, v7) project have been described elsewhere (3). In brief, whole genome sequencing (WGS) was performed by the Broad Institute's Genomics Platform on DNA samples at an average coverage of 30X with Illumina HiSeq 2000/X-Ten. Genotypes derived from WGS for the 635 individuals included in GTEx were obtained via controlled access from dbGaP (phs000424.v7.p2). Samples with $< 80\%$ European ancestry were excluded based on analysis using the Genotyping Library and Utilities (GLU) *struct.admix* module (5), resulting in a total of 174 individuals with both genetic data and gene expression data for pancreatic tissue samples available for analyses. Variants with call rates $< 95\%$ or MAF < 0.01 were excluded using VCFtools (9).

Genetic variants from the GTEx ($n = 8,130,638$ variants) and LTG ($n = 6,475,451$ variants) expression datasets were annotated with rsID numbers based on dbSNP (v150) and the human reference genome GRCh37/hg19 (10) by BCFtools (11) using genome positions and alleles as matching criteria. The two datasets were combined for variants with matching positions and alleles ($n = 5,119,190$ variants), and genotypes unique to only one of the datasets were excluded. For model building using FUSION (12), variants not present in the 1000 Genome European populations (6) were excluded from further analysis. For model building using MetaXcan (13, 14), variants absent in HapMap (15) or with any missing genotypes ($n = 645,774$) were excluded from further analysis.

Covariates controlling for ancestry and experimental confounders in gene expression prediction model building

After removing genomic regions of extended high LD (such as the HLA region) and pruning variants based on LD (using the *plink.prune* module in PLINK with 50 kb and 5 variants for each step size, and $r^2 \geq 0.2$ to exclude variants), we calculated principal components (PCs) using SNPRelate (16). The gene expression values were adjusted for the top 5 PCs, the Probabilistic Estimation of Expression Residuals (PEER) factors (17), as well as gender. The number of PEER

factors was chosen according to the number of samples (n) as per GTEx (3): 15 PEER factors were used for the LTG pancreas dataset, 30 for GTEx pancreas (v7) and 45 for the combined pancreas dataset.

Building pancreatic tissue gene expression prediction models

To build robust gene expression models for pancreatic tissues, we used genome-wide genotype and RNA-seq pancreas transcriptome data from individuals with European ancestry for the LTG, GTEx (v7) and combined LTG + GTEx datasets. The LTG dataset is derived mostly from histologically normal pancreatic tissue samples that are adjacent to pancreatic tumors, whereas the GTEx dataset is derived from individuals who do not have a diagnosis of pancreatic cancer.

Prediction models were computed using the following four linear methods included in FUSION (12) and the best model was selected for the association test: (i) least absolute shrinkage and selection operator (LASSO) (18), a penalized regression method using the L1 norm as the penalty function; (ii) elastic-net regression (Enet) (19), a penalized regression approach with a mixing parameter of 0.5, using the weighted average of the L1 and L2 norms; (iii) best linear unbiased predictor (BLUP) (20), which estimates the effect sizes of all SNPs in the locus jointly using a single variance component; and (iv) Bayesian sparse linear mixed model (BSLMM) (21), which estimates the underlying effect size distribution and then fits all SNPs in the locus jointly. For BLUP and BSLMM, prediction was done over all post-QC SNPs using GEMMA (21). BLUP and BSLMM both perform shrinkage of the SNP weights but not variable selection (12), so all SNPs are included in the predictor. For each gene, variants +/-500kb of the gene boundary, as defined by GENCODE v19 gene annotations, were used to estimate *cis*-SNP-heritability. Only protein-coding genes, long non-coding RNAs (lncRNAs), processed transcripts, immunoglobulin genes and T-cell receptor genes, as defined by GENCODE v19, were extracted for model building. Genes with nominally significant *cis*-SNP-heritability (LRT $P < 0.05$) and cross-validation R^2 for the best performing model of > 0.01 were retained, and the genotypes were used to train TWAS prediction models using Enet, LASSO, BLUP and BSLMM models. Five-fold cross-validation was performed for each model. This resulted in 2,827, 4,992 and 5,902 gene expression prediction models (prediction performance $R^2 \geq 0.01$) in the LTG,

GTEx and the combined LTG + GTEx datasets, respectively. Among these, 41%, 18%, 14% and 27% were derived from LASSO, Enet, BLUP and BSLMM, respectively.

As a complementary analysis, the prediction models for the LTG, GTEx (v7) and combined pancreas datasets were computed independently in S-PrediXcan (14) (a component of MetaXcan (13)) using genome-wide genotype and RNA-seq pancreas transcriptome data from the same European ancestry samples. We performed the following model building pipeline for the LTG and combined (LTG and GTEx) datasets, and obtained the prediction models trained on WGS genotypes (after imputing missing genotypes) for pancreas European GTEx (v7) from PredictDB (<http://predictdb.org/>) (14). Genetically regulated expression for each gene was estimated for SNPs within +/- 1 Mb of the gene boundaries, as defined by GENCODE v19 gene annotations. Only protein-coding genes, long non-coding RNAs (lncRNAs), processed transcripts, immunoglobulin genes and T-cell receptors defined in the GENCODE v19 gene annotation were extracted for model building. The expression prediction model for each gene using the Enet method was implemented in the glmnet R package, with a ridge-lasso mixing parameter of $\alpha = 0.5$ and a penalty parameter lambda chosen through 10-fold cross-validation (13). We retained models with an average Pearson correlation higher than 0.1 between the predicted and observed expression during nested cross validation (equivalent to $R^2 > 0.01$) and the estimated p value < 0.05 . This resulted in gene expression models for 2,440, 4,763 and 5,775 genes from the LTG, GTEx and combined LTG + GTEx datasets, respectively.

To assess statistical power for the TWAS, we simulated gene expression and GWAS summary statistics using genotype data from 22,330 individuals from GWAS (PanScan I+II, PanScan III and PanC4). For the PanC4 GWAS dataset we used the dbGAP dataset (phs000648.v1.p1) that we obtained through controlled access. We randomly selected 100 genes and included their *cis*-SNPs (+/- 1Mb from each gene) from the GWAS. We randomly selected 269 individuals from the GWAS as our expression panel using the `twas_sim` tool (https://github.com/mancusolab/twas_sim). Our simulations had three main parameters: the number/percentage of causal SNPs (1, 1% and 10%) for the expression for a given gene in the *cis* region, the fraction of gene expression variance explained by causal SNPs (H^2 , 0.1, 0.3 and 0.5), and the fraction of phenotypic variance explained by the expression of the gene (R^2 , 0.2×10^{-3} , 0.4×10^{-3} , 0.6×10^{-3} , 0.8×10^{-3} and 1×10^{-3}). To compute power, we fixed H^2 (h^2_g) and then varied causal SNPs at 1, 1% and 10%, then varied R^2 (h^2_{ge}) from 0 to 0.001 and recompute power (100

times per configuration (H^2 , R^2) and then computed how often the TWAS/GWAS tests were statistically significant (TWAS $P < 2.27 \times 10^{-6}$ (0.05/22k); GWAS $P < 5 \times 10^{-8}$). We then repeated this process for different H^2 values (0.1, 0.3 and 0.5) (**Supplementary Figure 4**). In our simulations, we assumed that causal SNPs impacted the phenotype through the given gene and that the eQTL was a causal mechanism for the trait. If eQTLs do not contribute to the trait under study then the power for TWAS is 0, whereas GWAS may have power > 0 because GWAS can find associations that are not mediated through expression.

We also compared model performance across the three expression datasets (LTG, GTEx and LTG+GTEx) and the two TWAS methods (FUSION and MetaXcan) and observed good correlation (Pearson r ranged from 0.60-0.93 for the different gene expression datasets and from 0.87-0.98 when comparing FUSION and MetaXcan) (**Supplementary Figures 1A, 4C and 4D**). For genes that were statistically significant using at least one TWAS method (FUSION or MetaXcan), the correlation was almost perfect (Person $r=0.99$) (**Supplementary Figure 1B**).

Cross-tissue genetically regulated expression models

We downloaded and used publicly available gene expression prediction models for 48 different human tissues ($n=74-421$ samples per tissue; a total 8869 samples from 608 European ancestry individuals) from PredictDB (<http://predictdb.org/>)(14). These models were trained using GTEx (v7) data for European participants only, using PrediXcan. For these models, the genotype data were imputed using the Michigan Imputation Server (8) with the HRC (Version r1.1 2016) as the reference panel (22). Variants for each tissue were filtered based on $MAF \geq 1\%$, imputation $R^2 \geq 0.8$, including only bi-allelic variants with unambiguous strand assignment. Variants were mapped to rsID numbers (dbSNP v150) using genomic location and alleles as matching criteria. The quantification, QC and normalization of gene expression data for each tissue were performed according to the GTEx consortium guideline (3), as per <https://github.com/broadinstitute/gtex-pipeline>. Before training models to predict gene expression, gene expression values were adjusted for the following covariates: sequencing platform (Illumina HiSeq 2000/X Ten), gender, top 3 PCs, and PEER factors. The number of PEER factors was chosen according to the number of samples (N) according to GTEx (v7) protocol: 15 factors for $n < 150$, 30 factors for $150 \leq n < 250$, 45 factors for $250 \leq n < 350$, and 60 factors for $n \geq 350$.

Association analyses between predicted gene expression and pancreatic cancer risk

For prediction models derived from FUSION, summary-level based imputation was performed using the ImpG-Summary algorithm (23) extended to train on the *cis* genetic component of expression (12). Z_{GWAS} is a vector of standardized effect sizes of SNP for a trait at a given *cis* locus (Wald statistics $\hat{\beta}/se(\hat{\beta})$). We imputed the z score of the expression and trait as a linear combination of elements of Z with weights W_{GE} . Given the prediction model weights W_{GE} , GWAS summary Z scores Z_{GWAS} , and SNP-correlation (LD) matrix V ; associations between predicted expression and pancreatic cancer risk were estimated using the following formula (methodological details in ref (12)).

$$z_{\text{TWAS}} = \frac{\mathbf{w}_{\text{GE}} \mathbf{z}_{\text{GWAS}}}{\sqrt{\text{var}(\mathbf{w}_{\text{GE}} \mathbf{z}_{\text{GWAS}})}} = \frac{\mathbf{w}_{\text{GE}} \mathbf{z}_{\text{GWAS}}}{\sqrt{\mathbf{w}_{\text{GE}}' \mathbf{V} \mathbf{w}_{\text{GE}}}},$$

For prediction models derived from MetaXcan, associations between genetically predicted gene expression levels and pancreatic cancer risk were estimated using the following formula (methodological details in ref (22)):

$$z_g \approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{se(\hat{\beta}_l)}$$

where w_{lg} is the weight of SNP l for predicting the expression of gene g , β_l and $se(\beta_l)$ are the association regression coefficient and its standard error for SNP l in GWAS, and σ_l and σ_g are the estimated variances of SNP l and the predicted expression of gene g , respectively. The weights for gene expression predicting SNPs, GWAS summary statistics data, and correlations between predictor SNPs were the input variables included in the MetaXcan analyses. The performance of MetaXcan has been found to be generally consistent with that of PrediXcan, which uses individual-level genetic data (13, 14). The pancreatic cancer GWAS summary statistics were based on 9,040 pancreatic ductal adenocarcinoma (PDAC) cases and 12,496 controls of European ancestry from PanScan and PanC4 (24), details of which have been previously described (5, 25-27). All participating studies obtained informed consent from study participants and Institutional Review Board (IRB) approvals. The PanScan and PanC4 GWAS data are available through dbGAP (accession numbers phs000206.v5.p3 and phs000648.v1.p1,

respectively). We used an FDR corrected P -value threshold of < 0.05 for each analysis (i.e. using the LTG, GTE_x and LTG + GTE_x transcriptome datasets in FUSION and MetaXcan).

Bonferroni correction for multiple testing was also used when indicated (correcting for 2440-5902 tests (see number of tests in each analysis in **Supplementary Figure 2**)).

The estimated inflation of the test statistic was $\lambda=1.227$ (LTG), 1.265 (GTE_x) and 1.204 (LTG+GTE_x) for FUSION, and $\lambda=1.077$ (LTG), 1.169 (GTE_x) and 1.189 (LTG+GTE_x) for MetaXcan. Adjusted to 1,000 case-control pairs, the estimated inflation was $\lambda_{1000}=1.022$ (LTG), 1.025 (GTE_x) and 1.019 (LTG+GTE_x) for FUSION, and $\lambda_{1000}=1.007$ (LTG), 1.016 (GTE_x) and 1.018 (LTG+GTE_x) for MetaXcan. The quantile-quantile (QQ) plots are shown in **Supplementary Figure 3**.

Finally, we used Summary-MulTiXcan (SMulTiXcan) (28) to test associations between predicted gene expression levels and pancreatic cancer risk with cross tissue models. Using univariate S-PrediXcan results and LD information from a reference panel (1KG), SMulTiXcan consists of the following steps: (i) computation of single tissue association results with S-PrediXcan; (ii) estimation of the correlation matrix of predicted gene expression for the models using the LD information from 1KG panel; (iii) discarding components of smallest variation from this correlation matrix to avoid collinearity and numerical problems; (iv) estimation of joint effects from the single-tissue results and expression correlation; (v) discarding suspicious results arising from LD-structure mismatch. The estimated inflation of the test statistic was $\lambda=1.040$ for the SMulTiXcan analysis; after adjusting to 1,000 case-control pairs this was $\lambda_{1000}=1.004$. The quantile-quantile (QQ) plot is shown in **Supplementary Figure 3**.

Summary-based joint/conditional tests

To assess the extent of residual association of a SNP with pancreatic cancer risk after removing genetically predicted expression for genes of interest, conditional SNP association tests using GWAS summary statistics were carried out using FUSION (FUSION.post_process.R) (12). To assess whether associations between genetically predicted gene expression and pancreatic cancer risk were independent of the most statistically significant GWAS-identified association signal at each locus (\pm 1Mb of each TWAS gene), we performed conditional analyses for the GWAS dataset using GCTA-COJO (29). We then reran FUSION and MetaXcan analyses for the pancreas tissue models using updated summary statistics. Note that we were not

able to perform conditional analysis in the LTG pancreas dataset for chr17q12 to confirm results in the GTEx dataset, as the model for *CDK12* did not pass the prediction performance threshold ($R^2 \geq 0.01$). For cross-tissues models, we reran MetaXcan (14) for each tissue after conditioning the GWAS analysis on the most statistically significant SNP (multiple independent SNPs were used for chr5p15.33) at each locus, and then combined analyses were rerun using SMuTiXcan (28).

Transcriptome differences and pathway analyses for TWAS-identified genes

We assessed transcriptome changes associated with high and low expression of TWAS-identified genes in the LTG (1) and GTEx (3) pancreatic datasets. This approach was based on our recent analysis for *NR5A2*, where we observed a high correlation between genes that were differentially expressed in pancreatic tissues from wild type mice when compared to heterozygous KO mice for *Nr5a2* (*Nr5a2*^{+/+} vs. *Nr5a2*^{+/-}), and in human pancreatic tissue samples in the top vs. bottom quartile of *NR5A2* expression using our LTG transcriptome datasets (Figure 1g in reference (30)). This correlation was highly statistically significant for genes that were upregulated in *Nr5a2*^{+/-} vs. *Nr5a2*^{+/+} mice, and expressed at higher levels in the top vs. bottom quartile of human *NR5A2* expression ($P = 1 \times 10^{-31}$ when compared to a random list of genes) but not observed for genes expressed at lower levels in *Nr5a2*^{+/-} mice and the bottom quartile of the human pancreatic tissue samples ($P = 0.58$) (30). We therefore assessed transcriptome changes for each of the genes identified from the TWAS using pancreatic tissue models, by comparing gene expression for samples in the bottom quartile to the top quartile of expression in the LTG ($n = 95$) and GTEx ($n = 174$) pancreatic samples using the EdgeR package in R (31, 32). For each dataset, gene expression counts were scaled for sequencing depth and RNA composition across all samples in each dataset to give normalized counts of the trimmed mean of M-values (TMM). Genes with no reads for > 20% of the samples were not included in this analysis. Normalized reads were then used to determine the top and bottom quartiles of expression for each gene of interest across samples in each dataset. For subsequent gene-based analysis, only those samples in the top and bottom quartile ($n=24$ for LTG and $n=44$ for GTEx quartiles) were used. The raw counts for these selected samples for the filtered genes were re-normalized for sequencing depth to obtain pseudo-counts which were analyzed using the quantile-adjusted conditional maximum likelihood (qCML) method in EdgeR. Differential

expression (\log_2 [bottom/top quartile] and P -values) was then assessed using an exact test. Genes statistically significantly differentially expressed at $FDR < 0.05$ and fold-change > 2 -fold ($|\log_2FC| > 1$) were included in pathway analyses using DAVID (33, 34) to identify enrichment in GO Biological Processes, GO Molecular Functions, and KEGG Pathways.

Data access

Genotype and RNA-seq data for the NCI LTG eQTL dataset are available from the database of Genotypes and Phenotypes (dbGAP, <https://www.ncbi.nlm.nih.gov/gap>) under accession number phs001776.v1.p1. The GTEx dataset (phs000424.v7.p2) as well as the PanScan (phs000206.v5.p3) and PanC4 (phs000648.v1.p1) genome-wide association data are available through dbGaP.

Supplementary Tables and Figures

Supplementary Table 1. Characteristics that differ between the FUSION and MetaXcan TWAS methods*

| Characteristic | FUSION | MetaXcan |
|--|-----------------------------|----------|
| Genotype missing rate tolerance for each variant | 5% | 0% |
| SNPs boundaries from each gene | +/- 500kb | +/- 1Mb |
| Genotype filtering reference | 1000G | HapMap |
| GWAS summary imputation | Yes | No |
| Model training methods | LASSO, Enet, BLUP and BSLMM | Enet |
| Cross-validation (fold) | 5 | 10 |
| Conditional analysis on predicted genes | Yes | No |

*Both programs were used using default TWAS settings.

Supplementary Table 2. Gene expression correlation (Pearson's correlation coefficients (R)) for TWAS genes at 2 genomic loci in the GTEx and LTG datasets.

| Dataset | chr17q12 | | | chr16q23.1 |
|---------|------------------------------|------------------------------|-------------------------------|-------------------------------|
| | <i>CDK12</i> and <i>PNMT</i> | <i>PNMT</i> and <i>PGAP3</i> | <i>CDK12</i> and <i>PGAP3</i> | <i>WDR59</i> and <i>CFDP1</i> |
| GTEx | 0.09 | 0.33 | 0.66 | 0.80 |
| LTG | 0.09 | 0.27 | 0.29 | 0.52 |

Supplementary Table 3. Association results for variants that tag pancreatic cancer risk signals on chr5p15.33 in the PanScan – PanC4 GWAS dataset before and after conditional analysis*

| Variants | GWAS <i>P</i> -value | GWAS <i>P</i> -value after conditioning on rs31490 | GWAS <i>P</i> -value after conditioning on rs2736098 | GWAS <i>P</i> -value after conditioning on rs36115365 | GWAS <i>P</i> -value after conditioning on rs35226131 |
|------------|----------------------|--|--|---|---|
| rs31490 | 1.28E-17 | 1 | 2.47E-08 | 4.04E-07 | 2.51E-15 |
| rs2736098 | 5.80E-14 | 3.31E-05 | 1 | 3.55E-10 | 2.93E-16 |
| rs36115365 | 6.12E-12 | 9.25E-03 | 2.95E-08 | 1 | 1.71E-10 |
| rs35226131 | 2.19E-08 | 3.17E-06 | 8.21E-11 | 4.54E-07 | 1 |

*An additive genetic model was used to perform the association analysis (two-sided test).

Supplementary Table 4. Number of genes significantly differentially expressed (Fold change, FC > 2, FDR < 0.05) in samples in the top (GeneX^{high}) vs. bottom (GeneX^{low}) quartile of expression for each listed gene*

| Gene | GTEx | | LTG | | Shared | |
|----------------------|------------------------|---------------------------|------------------------|---------------------------|------------------------|---------------------------|
| | Higher in top quartile | Higher in bottom quartile | Higher in top quartile | Higher in bottom quartile | Higher in top quartile | Higher in bottom quartile |
| <i>CELA3B</i> | 280 | 665 | 3,097 | 4,056 | 204 | 555 |
| <i>SMC2</i> | 99 | 323 | 1,616 | 1,481 | 57 | 31 |
| <i>SMUG1</i> | 120 | 176 | 1,257 | 3,963 | 16 | 93 |
| <i>BTBD6</i> | 292 | 46 | 4,384 | 2,178 | 259 | 22 |
| <i>HEXA</i> | 3,471 | 2,934 | 183 | 168 | 146 | 55 |
| <i>RCCD1</i> | 393 | 127 | 3,431 | 4,057 | 141 | 64 |
| <i>PNMT</i> | 314 | 432 | 204 | 585 | 72 | 100 |
| <i>CDK12</i> | 142 | 68 | 3,474 | 1,742 | 60 | 32 |
| <i>PGAP3</i> | 97 | 240 | 2,117 | 3,990 | 41 | 164 |
| <i>SUPT4H1</i> | 38 | 122 | 1,947 | 1,599 | 9 | 14 |
| <i>RP11.888D10.3</i> | 68 | 63 | 13 | 37 | 0 | 1 |
| <i>PGPEP1</i> | 16 | 125 | 1,775 | 3,717 | 3 | 93 |
| <i>INHBA</i> | 752 | 87 | 4,962 | 3,250 | 628 | 34 |
| <i>ABO</i> | 112 | 108 | 626 | 2,616 | 7 | 22 |
| <i>PDX1</i> | 21 | 104 | 1,018 | 2,669 | 4 | 33 |
| <i>KLF5</i> | 302 | 68 | 614 | 502 | 104 | 17 |
| <i>WDR59</i> | 164 | 90 | 1,038 | 4,523 | 17 | 78 |
| <i>CFDP1</i> | 45 | 232 | 1,462 | 3,640 | 20 | 136 |

*The number of genes differentially expressed in samples in the top vs. bottom quartile of expression for each gene listed. As we had previously shown for *NR5A2* that this approach was highly consistent for genes that were upregulated in pancreatic tissue samples from heterozygous knockout mice (*Nr5a2*^{+/-} as compared to *Nr5a2*^{+/+}) and those expressed at higher levels in the bottom as compared to top *NR5A2* expression for human samples (*NR5A2*^{low} vs. *NR5A2*^{high} samples in LTG dataset) for genes expressed at higher levels in the bottom quartile of gene expression (columns 2, 4 and 6 above) we focused the pathway analysis on those genes.

Supplementary Table 5. Top pathways enriched in the samples in the bottom vs. top quartile of TWAS gene expression in the GTEx and LTG pancreas datasets*

| TWAS Gene | Category | Term | GTEx | | | LTG | |
|---------------|----------|--|--------------|-----------------|----------|-----------------|----------|
| | | | DE genes (n) | Fold Enrichment | P-value† | Fold Enrichment | P-value† |
| <i>ABO</i> | GO BP | inflammatory response | 18 | 11.4 | 1.1E-10 | 4 | 4.3E-30 |
| <i>ABO</i> | GO BP | cellular response to tumor necrosis factor | 10 | 21.8 | 2.3E-07 | 3.5 | 1.5E-05 |
| <i>ABO</i> | KEGG | TNF signaling pathway | 10 | 15.5 | 3.5E-07 | 2.3 | 9.2E-03 |
| <i>ABO</i> | KEGG | Complement and coagulation cascades | 9 | 21.5 | 4.4E-07 | 3.4 | 1.9E-04 |
| <i>ABO</i> | GO BP | chemokine-mediated signaling pathway | 8 | 27 | 2.3E-06 | 5.9 | 2.9E-11 |
| <i>ABO</i> | GO BP | cellular response to interleukin-1 | 8 | 27 | 2.3E-06 | 3.7 | 7.3E-04 |
| <i>ABO</i> | GO BP | immune response | 14 | 8 | 2.5E-06 | 4.5 | 9.8E-44 |
| <i>ABO</i> | GO BP | acute-phase response | 7 | 43.1 | 2.7E-06 | 3.2 | 2.3E-01 |
| <i>ABO</i> | GO BP | positive regulation of neutrophil chemotaxis | 6 | 65.4 | 3.0E-06 | 8.5 | 4.3E-06 |
| <i>ABO</i> | GO MF | chemokine activity | 7 | 35.5 | 6.1E-06 | 5.6 | 2.6E-06 |
| <i>BTBD6</i> | GO BP | muscle filament sliding | 7 | 99.8 | 8.1E-09 | 2.8 | 9.8E-01 |
| <i>BTBD6</i> | GO MF | alpha-amylase activity | 4 | 435.6 | 3.1E-06 | 16.9 | 1.3E-01 |
| <i>CELA3B</i> | GO BP | inflammatory response | 72 | 6.3 | 3.8E-33 | 3.2 | 6.0E-55 |
| <i>CELA3B</i> | GO BP | immune response | 70 | 5.6 | 1.2E-28 | 3 | 2.7E-50 |
| <i>CELA3B</i> | KEGG | Staphylococcus aureus infection | 24 | 11.5 | 7.9E-17 | 4.5 | 3.5E-21 |
| <i>CELA3B</i> | GO BP | cell adhesion | 57 | 4.1 | 1.4E-16 | 2.7 | 1.2E-40 |
| <i>CELA3B</i> | GO BP | innate immune response | 52 | 4 | 1.6E-14 | 2.4 | 8.6E-25 |
| <i>CELA3B</i> | GO BP | chemotaxis | 28 | 7.7 | 1.2E-13 | 3.2 | 8.7E-16 |
| <i>CELA3B</i> | KEGG | Osteoclast differentiation | 30 | 5.9 | 6.1E-13 | NA | NA |
| <i>CELA3B</i> | GO BP | regulation of immune response | 32 | 6 | 7.4E-13 | 2.7 | 5.3E-14 |
| <i>CELA3B</i> | KEGG | Cytokine-cytokine receptor interaction | 37 | 4.2 | 1.6E-11 | 2.7 | 3.5E-25 |
| <i>CELA3B</i> | GO BP | extracellular matrix organization | 29 | 4.9 | 2.3E-09 | 3.3 | 6.9E-29 |
| <i>CELA3B</i> | GO BP | leukocyte migration | 23 | 6.3 | 3.6E-09 | 3.5 | 1.4E-19 |
| <i>CELA3B</i> | GO BP | chemokine-mediated signaling pathway | 18 | 8.5 | 6.4E-09 | 3.5 | 8.6E-11 |
| <i>CELA3B</i> | GO BP | acute-phase response | 14 | 12 | 1.4E-08 | 3.1 | 5.2E-04 |
| <i>CELA3B</i> | GO BP | neutrophil chemotaxis | 17 | 8.6 | 1.5E-08 | 3.4 | 7.9E-10 |
| <i>CELA3B</i> | GO BP | signal transduction | 77 | 2.2 | 1.6E-08 | 1.8 | 5.6E-28 |
| <i>CELA3B</i> | KEGG | Rheumatoid arthritis | 20 | 5.9 | 2.3E-08 | 3 | 1.4E-12 |
| <i>CELA3B</i> | KEGG | Complement and coagulation cascades | 18 | 6.8 | 2.7E-08 | 3.1 | 6.9E-11 |
| <i>CELA3B</i> | GO BP | cell-cell signaling | 31 | 4.1 | 2.9E-08 | 2.4 | 1.8E-14 |
| <i>CELA3B</i> | GOBP | positive regulation of neutrophil chemotaxis | 11 | 16.7 | 4.6E-08 | 4.9 | 6.5E-07 |
| <i>CELA3B</i> | KEGG | Tuberculosis | 27 | 4 | 6.8E-08 | 2 | 1.1E-07 |
| <i>CELA3B</i> | KEGG | TNF signaling pathway | 21 | 5.1 | 7.1E-08 | 1.9 | 8.3E-04 |

| | | | | | | | |
|----------------------|-------|---|----|-------|---------|-----|---------|
| <i>CELA3B</i> | KEGG | Malaria | 15 | 8 | 7.6E-08 | 3.8 | 6.3E-12 |
| <i>CELA3B</i> | GO BP | cell chemotaxis | 16 | 8.2 | 1.0E-07 | 3.3 | 1.2E-08 |
| <i>CELA3B</i> | GO BP | cellular response to lipopolysaccharide | 20 | 5.9 | 1.6E-07 | 2.8 | 2.2E-10 |
| <i>CELA3B</i> | KEGG | Chemokine signaling pathway | 27 | 3.8 | 1.8E-07 | 2.1 | 1.6E-09 |
| <i>CELA3B</i> | KEGG | Hematopoietic cell lineage | 18 | 5.5 | 3.3E-07 | 3.8 | 9.5E-22 |
| <i>CELA3B</i> | GO BP | interferon-gamma-mediated signaling pathway | 16 | 7.5 | 3.4E-07 | 3.7 | 3.5E-13 |
| <i>CELA3B</i> | KEGG | Leishmaniasis | 16 | 5.9 | 9.9E-07 | 3.3 | 1.5E-12 |
| <i>CELA3B</i> | GO MF | cytokine activity | 24 | 4.7 | 1.0E-06 | 2.5 | 1.3E-10 |
| <i>CELA3B</i> | GO BP | regulation of complement activation | 11 | 12.2 | 1.3E-06 | 3.6 | 1.8E-04 |
| <i>CELA3B</i> | GO MF | receptor activity | 26 | 4.1 | 1.4E-06 | 2.7 | 9.3E-17 |
| <i>CELA3B</i> | KEGG | Amoebiasis | 19 | 4.7 | 1.5E-06 | 2.5 | 8.9E-09 |
| <i>CELA3B</i> | GO BP | adaptive immune response | 21 | 4.7 | 2.4E-06 | 3.2 | 4.5E-20 |
| <i>CELA3B</i> | GO BP | positive regulation of T cell proliferation | 14 | 7.8 | 2.5E-06 | 4 | 4.1E-13 |
| <i>CELA3B</i> | KEGG | Phagosome | 22 | 3.7 | 5.0E-06 | 2.5 | 1.9E-12 |
| <i>CELA3B</i> | KEGG | Cell adhesion molecules (CAMs) | 21 | 3.8 | 5.9E-06 | 3.1 | 8.4E-22 |
| <i>CELA3B</i> | GO BP | cytokine-mediated signaling pathway | 19 | 4.8 | 8.9E-06 | 2.2 | 4.3E-05 |
| <i>CELA3B</i> | GO BP | defense response | 14 | 7 | 9.1E-06 | 3.3 | 6.8E-09 |
| <i>KLF5</i> | GO MF | alpha-amylase activity | 5 | 482.3 | 5.1E-09 | NA | NA |
| <i>PGAP3</i> | GO BP | inflammatory response | 22 | 6 | 1.5E-07 | 3.2 | 7.7E-51 |
| <i>PGAP3</i> | GO BP | immune response | 22 | 5.4 | 5.2E-07 | 3.1 | 6.9E-54 |
| <i>PGAP3</i> | GO BP | chemokine-mediated signaling pathway | 11 | 16 | 7.1E-07 | 3.4 | 2.7E-10 |
| <i>PGAP3</i> | GO MF | receptor binding | 19 | 5.8 | 1.6E-06 | 2.1 | 1.3E-12 |
| <i>PNMT</i> | GO BP | immune response | 35 | 5.1 | 1.7E-11 | 4 | 7.1E-10 |
| <i>PNMT</i> | GO BP | inflammatory response | 27 | 4.4 | 5.0E-07 | 4.2 | 5.7E-10 |
| <i>PNMT</i> | GO MF | chemokine activity | 11 | 14.2 | 1.6E-06 | 4.7 | 3.2E-01 |
| <i>PNMT</i> | KEGG | Rheumatoid arthritis | 14 | 7.3 | 7.9E-06 | 8.4 | 6.8E-11 |
| <i>RP11888D1 0.3</i> | GO BP | keratinization | 7 | 52.1 | 1.0E-06 | NA | NA |
| <i>RP11888D1 0.3</i> | GO MF | structural molecule activity | 10 | 15.2 | 1.3E-06 | NA | NA |
| <i>RP11888D1 0.3</i> | GO BP | keratinocyte differentiation | 7 | 32.9 | 8.4E-06 | NA | NA |
| <i>SMUG1</i> | GO BP | keratinocyte differentiation | 12 | 25.3 | 7.5E-10 | NA | NA |
| <i>SMUG1</i> | GO MF | structural molecule activity | 16 | 11 | 2.3E-09 | NA | NA |
| <i>SMUG1</i> | GO BP | keratinization | 10 | 33.3 | 3.9E-09 | NA | NA |
| <i>SMUG1</i> | GO BP | peptide cross-linking | 9 | 28.8 | 1.4E-07 | NA | NA |
| <i>SUPT4H1</i> | GO BP | keratinocyte differentiation | 11 | 30.8 | 9.5E-10 | NA | NA |
| <i>SUPT4H1</i> | GO BP | keratinization | 9 | 39.9 | 1.2E-08 | NA | NA |
| <i>SUPT4H1</i> | GO MF | structural molecule activity | 13 | 11.8 | 9.1E-08 | NA | NA |
| <i>SUPT4H1</i> | GO BP | peptide cross-linking | 8 | 34 | 5.0E-07 | NA | NA |
| <i>WDR59</i> | GO BP | chemotaxis | 12 | 22 | 4.2E-09 | 3 | 2.8E-12 |

| | | | | | | | |
|--------------|-------|--------------------------------------|----|------|---------|-----|---------|
| <i>WDR59</i> | GO BP | inflammatory response | 16 | 9.5 | 3.3E-08 | 3 | 8.8E-44 |
| <i>WDR59</i> | KEGG | Amoebiasis | 10 | 14.8 | 1.1E-06 | 2.2 | 1.6E-05 |
| <i>WDR59</i> | GO BP | chemokine-mediated signaling pathway | 8 | 25.2 | 6.8E-06 | 3.3 | 2.8E-09 |

GO BP: GO Biological Process; GO MF: GO Molecular Function.

* Pathways with Bonferroni adjusted P -values $< 1 \times 10^{-5}$ in GTEx are shown with corresponding enrichment and P -values in LTG. All tests were two-sided.

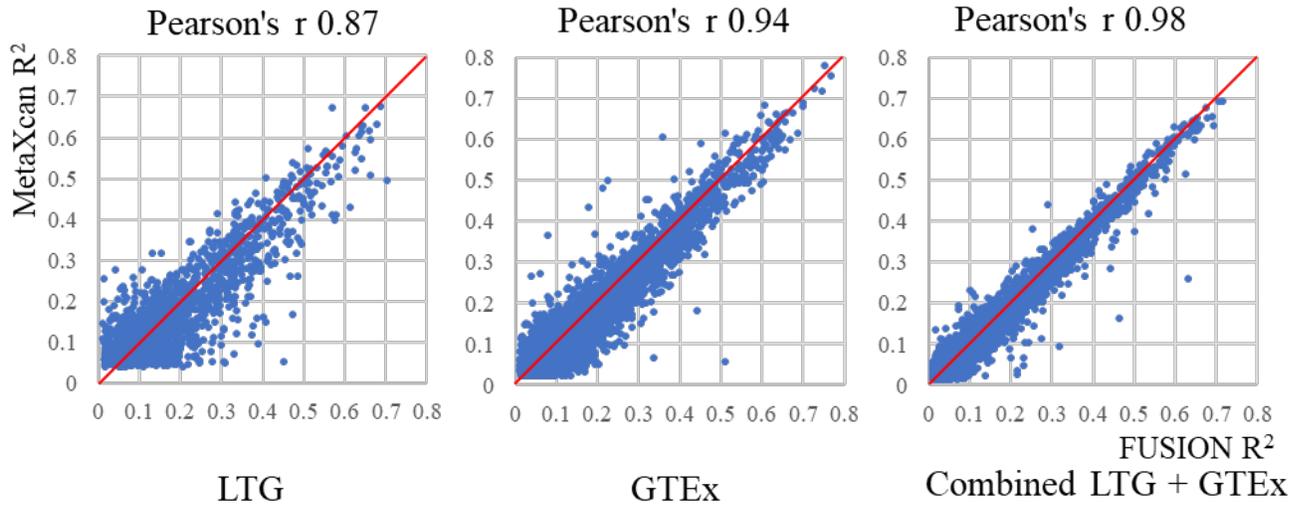
† Fisher Exact test was used to perform the gene enrichment and functional annotation analysis (two-sided test).

Supplementary Table 6. Median gene expression (TMM) for TWAS genes in samples in the top and bottom quartile of LTG and GTEEx pancreatic datasets and % difference for each gene.

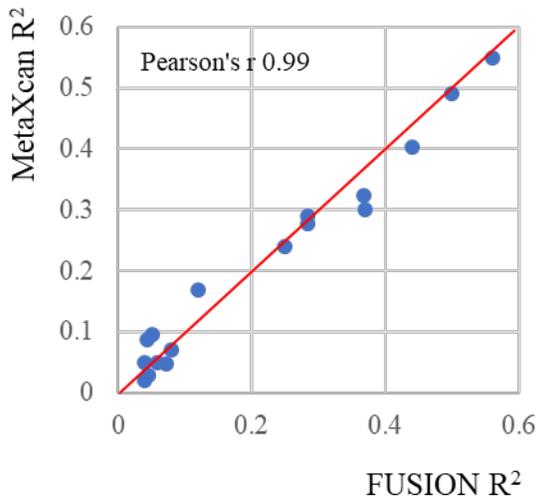
| Gene | GTEEx | | | LTG | | |
|----------------------|--------|--------|--------------|--------|--------|--------------|
| | Top | Bottom | % Difference | Top | Bottom | % Difference |
| <i>CELA3B</i> | 16,980 | 4,122 | 76% | 23,245 | 2,084 | 91% |
| <i>NR5A2*</i> | 130.78 | 81.03 | 38% | 136.25 | 55.74 | 59% |
| <i>PGPEP1</i> | 59.59 | 40.26 | 32% | 39.54 | 19.1 | 52% |
| <i>PGAP3</i> | 42.28 | 25.18 | 40% | 26.67 | 14.8 | 45% |
| <i>WDR59</i> | 42.05 | 30.96 | 26% | 23.85 | 15.72 | 34% |
| <i>CDK12</i> | 28.39 | 20.66 | 27% | 27.89 | 18.92 | 32% |
| <i>CFDP1</i> | 28.09 | 18.38 | 35% | 26.79 | 15.55 | 42% |
| <i>HEXA</i> | 26.26 | 18.88 | 28% | 37.48 | 20.27 | 46% |
| <i>RCCD1</i> | 21.35 | 9.74 | 54% | 10.5 | 4.49 | 57% |
| <i>SUPT4H1</i> | 19.81 | 14.46 | 27% | 20.94 | 14.56 | 30% |
| <i>KLF5</i> | 18.29 | 7.95 | 57% | 26.7 | 8.76 | 67% |
| <i>PDX1</i> | 18.06 | 10.86 | 40% | 16.48 | 7.25 | 56% |
| <i>SMUG1</i> | 15.81 | 10.76 | 32% | 10.63 | 6.07 | 43% |
| <i>BTBD6</i> | 12.95 | 7.62 | 41% | 13.37 | 6.21 | 54% |
| <i>SMC2</i> | 7.32 | 3.76 | 49% | 9.71 | 5.13 | 47% |
| <i>ABO</i> | 6.07 | 1.24 | 80% | 4.39 | 0.81 | 82% |
| <i>INHBA</i> | 1.44 | 0.25 | 82% | 38.02 | 0.54 | 99% |
| <i>PNMT</i> | 1.34 | 0.09 | 94% | 0.82 | 0.04 | 95% |
| <i>RP11.888D10.3</i> | 0.19 | 0.02 | 92% | 0.16 | 0.03 | 81% |

* *NR5A2* is not a TWAS gene in this analysis but shown as a reference for the analysis for differentially expressed gene in the top and bottom quartiles of gene expression (see **Methods**).

A.

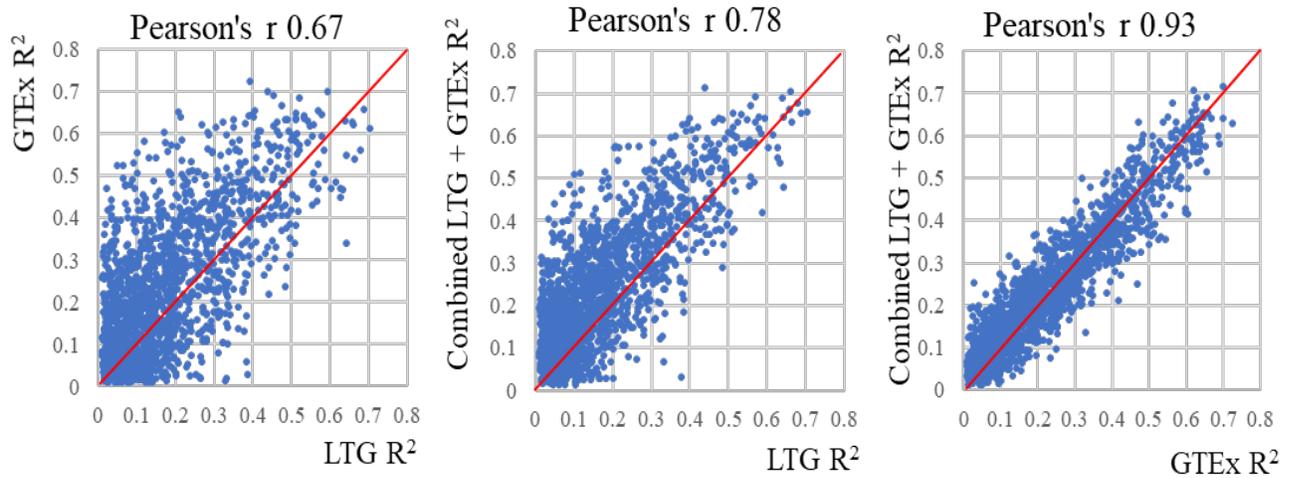


B.

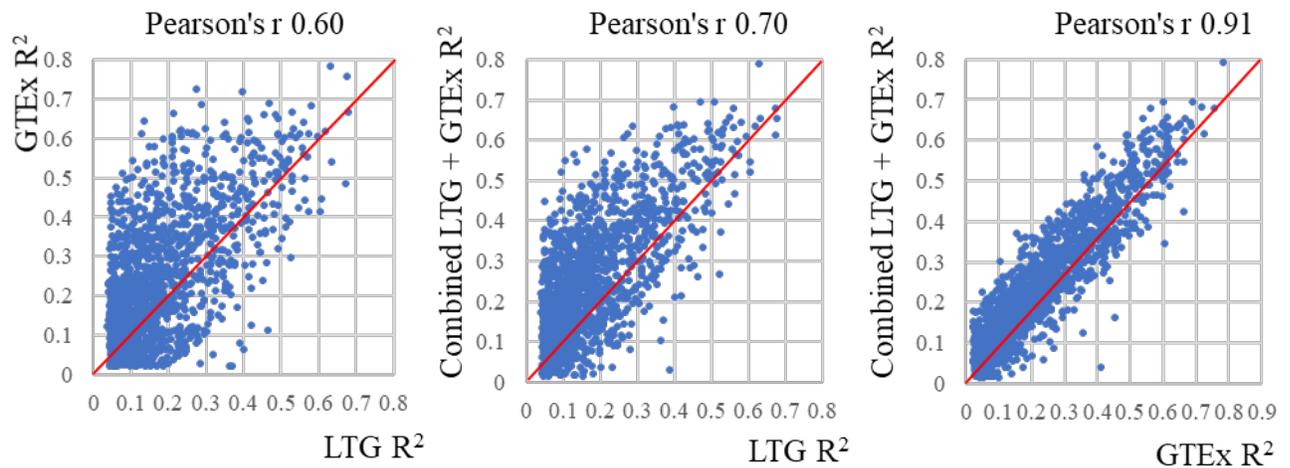


| Gene | FUSION R^2 | MetaXcan R^2 | Dataset |
|----------------------|--------------|----------------|----------|
| <i>CELA3B</i> | 0.06 | 0.05 | GTEX |
| <i>CELA3B</i> | 0.04 | 0.05 | Combined |
| <i>SMC2</i> | 0.04 | 0.02 | Combined |
| <i>SMUG1</i> | 0.28 | 0.29 | GTEX |
| <i>BTBD6</i> | 0.07 | 0.05 | GTEX |
| <i>BTBD6</i> | 0.05 | 0.10 | Combined |
| <i>RCCD1</i> | 0.37 | 0.32 | Combined |
| <i>RCCD1</i> | 0.44 | 0.40 | LTG |
| <i>RCCD1</i> | 0.28 | 0.28 | GTEX |
| <i>PGAP3</i> | 0.25 | 0.24 | GTEX |
| <i>RP11-888D10.3</i> | 0.04 | 0.09 | GTEX |
| <i>SUPT4H1</i> | 0.08 | 0.07 | GTEX |
| <i>ABO</i> | 0.37 | 0.3 | LTG |
| <i>ABO</i> | 0.56 | 0.55 | GTEX |
| <i>ABO</i> | 0.5 | 0.49 | Combined |
| <i>KLF5</i> | 0.05 | 0.03 | GTEX |
| <i>CFDP1</i> | 0.12 | 0.17 | Combined |

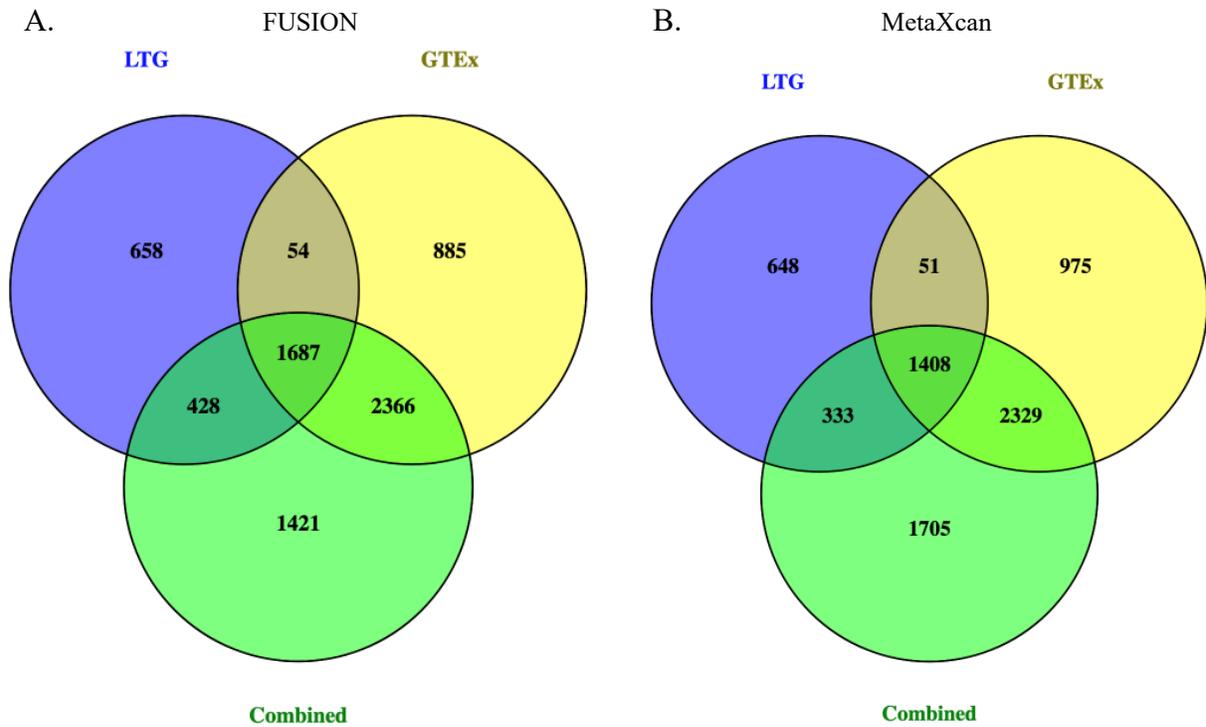
C.



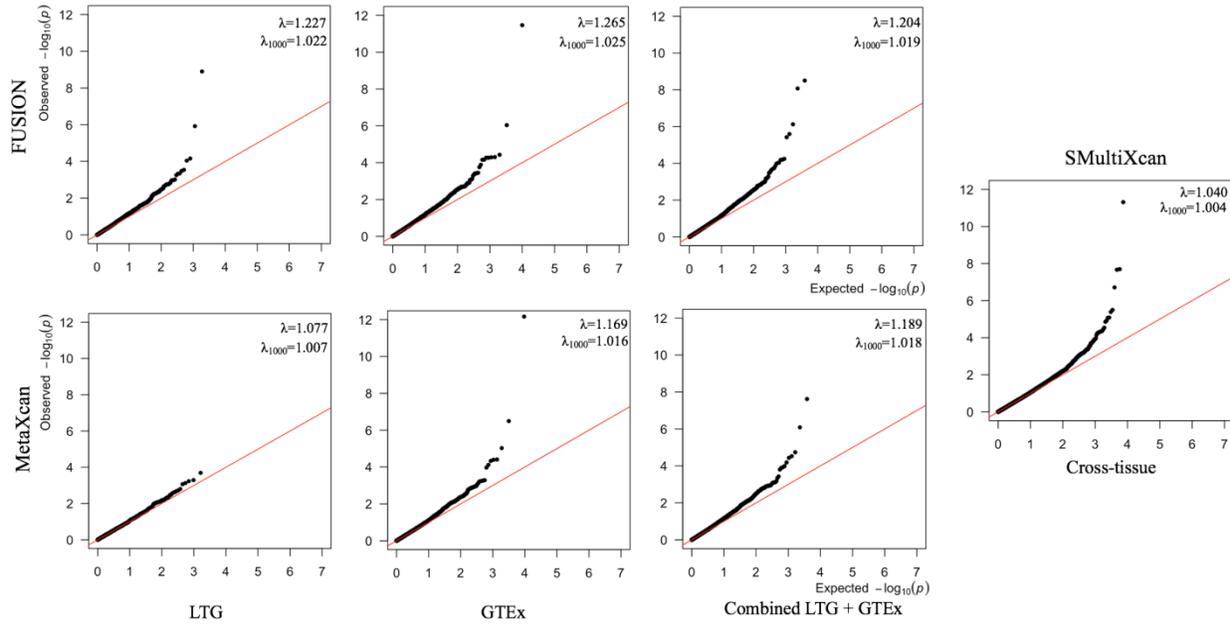
D.



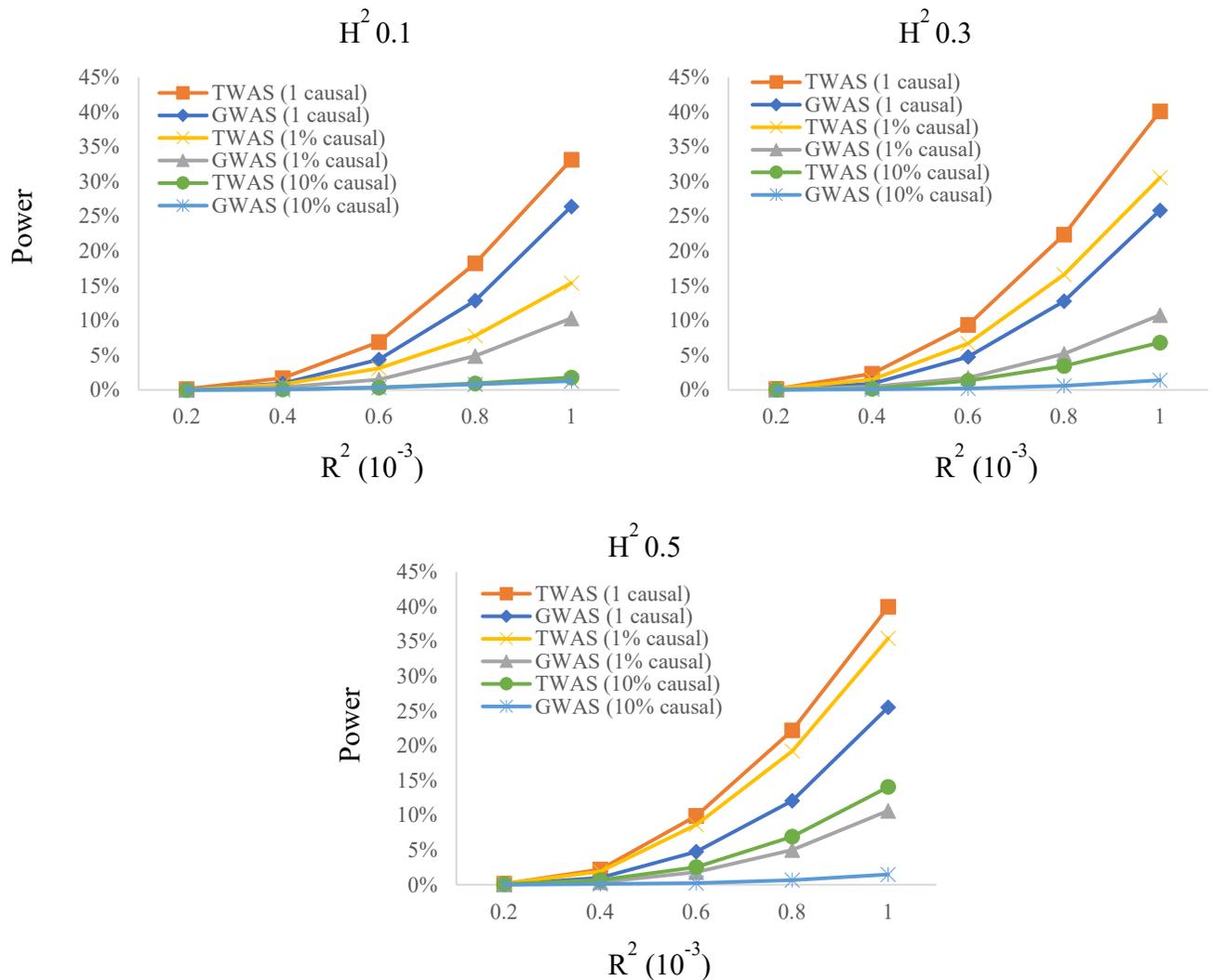
Supplementary Figure 1: Comparison of gene prediction model performance using different TWAS methods and gene expression panels. (A) Pearson correlation for the prediction performance (R^2) of gene models generated by FUSION and MetaXcan. (B) Pearson correlation for genes that were significant ($FDR < 0.05$) in at least one method (FUSION or MetaXcan). The correlation for all prediction models is shown in (C) for FUSION and in (D) for MetaXcan.



Supplementary Figure 2. Intersection of gene expression prediction models among LTG, GTEx and combined (LTG + GTEx) datasets. Numbers represent the number of gene prediction models that pass QC thresholds in one or more of the expression datasets. The overlap for models from the FUSION (A) and MetaXcan (B) TWAS analyses, with cross-validation performance $R^2 > 0.01$, is shown.

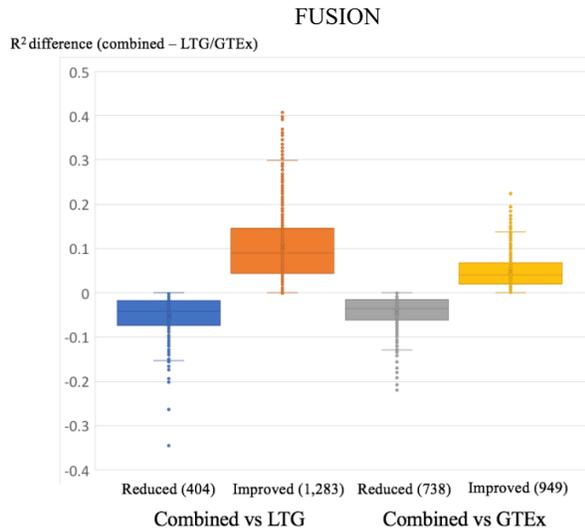


Supplementary Figure 3. Quantile-quantile (QQ) plots for the TWAS analysis using the LTG, GTEx, and combined (LTG + GTEx) expression panels in FUSION and MetaXcan, and cross-tissue expression datasets in SMultiXcan. Shown are lambda values before (λ) and after adjustment to a sample size of 1000 cases and 1000 controls (λ_{1000}) for all seven analyses.

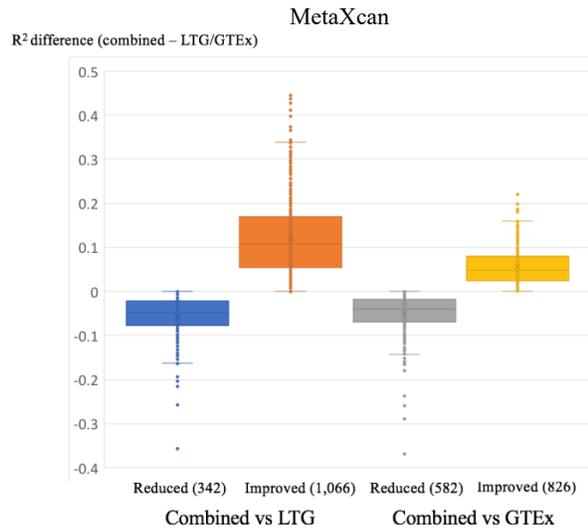


Supplementary Figure 4. Assessing and comparing statistical power for the pancreatic cancer TWAS as compared to the GWAS. Three main parameters were used for this analysis: the number/percentage (1, 1% and 10%) of causal SNPs for the expression in the *cis*-region (\pm 1Mb) for a given gene, the fraction (H^2 , 0.1, 0.3 and 0.5) of expression variance that is explained by causal SNPs, and the fraction (R^2 , 0.2×10^{-3} , 0.4×10^{-3} , 0.6×10^{-3} , 0.8×10^{-3} and 1×10^{-3}) of phenotypic variance explained by the expression of each gene. Colors and shapes correspond to the number/percentage of causal variants simulated for TWAS/GWAS. The expression reference panel included 269 out-of-sample individuals from a total GWAS sample size of 22,330. Power was computed as the fraction of 100 simulations where significant associations were identified.

A.

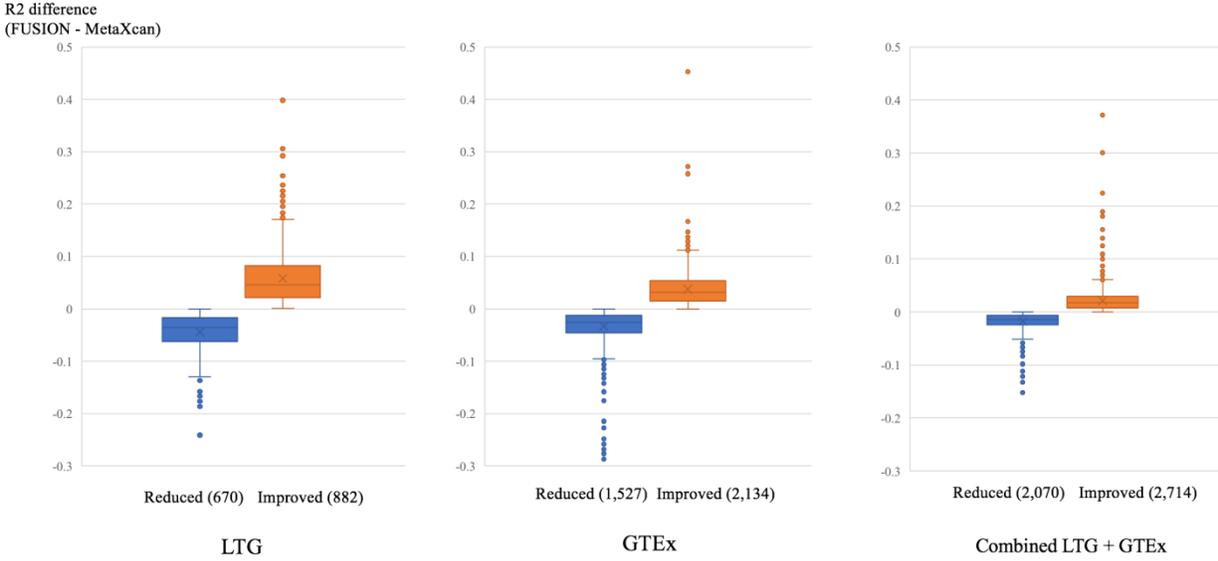


B.

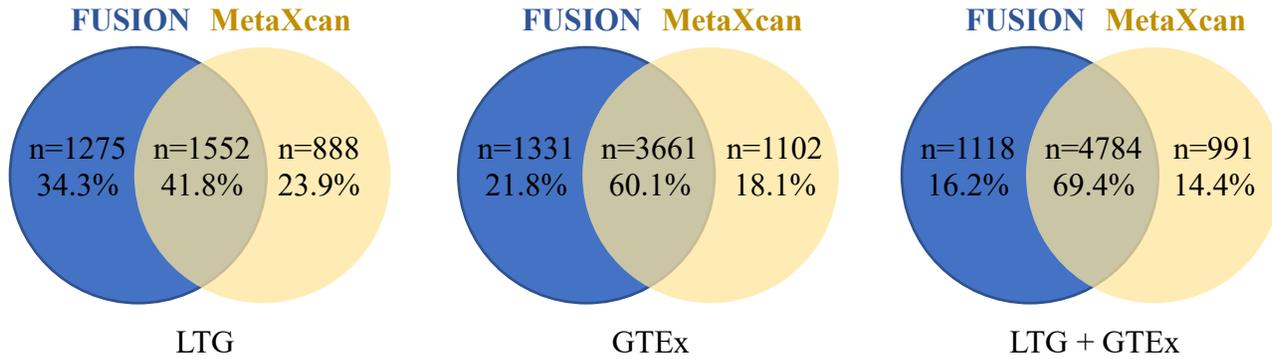


Supplementary Figure 5. Comparison of gene expression prediction models between the individual LTG or GTEX datasets versus the combined LTG + GTEX dataset. The models from FUSION (A) and from MetaXcan (B) with cross-validation $R^2 > 0.01$ and are overlapping among LTG, GTEX and combined LTG + GTEX datasets are shown with information about improved or reduced performance (R^2) in the combined vs. individual datasets. For instance, compared with LTG models, the performance for 404 models were reduced while 1,283 improved in the combined LTG + GTEX dataset. The average differences in R^2 were -0.052, 0.10, -0.044 and 0.048 (FUSION) and -0.057, 0.12, -0.050 and 0.056 (MetaXcan), respectively.

FUSION vs MetaXcan

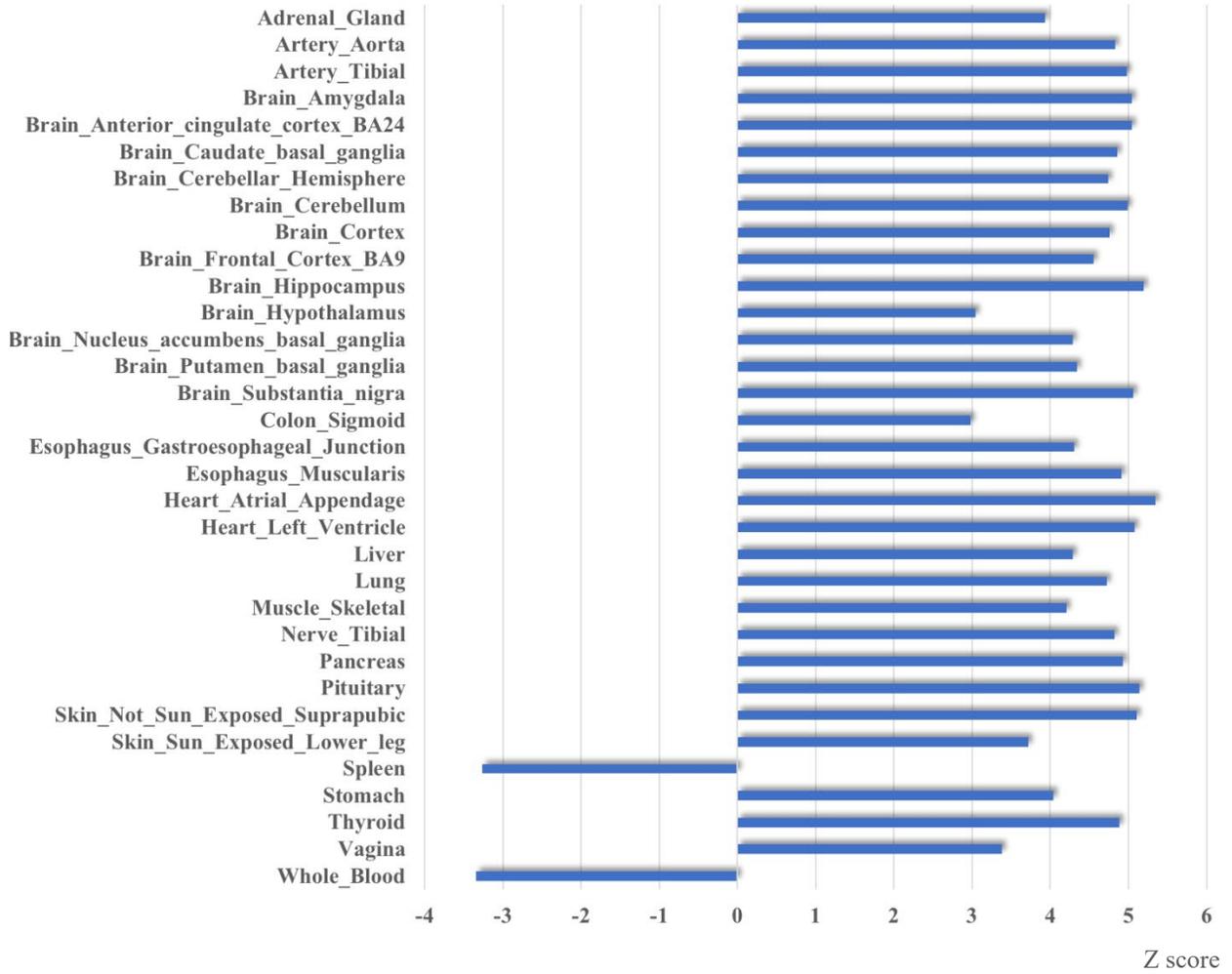


B.



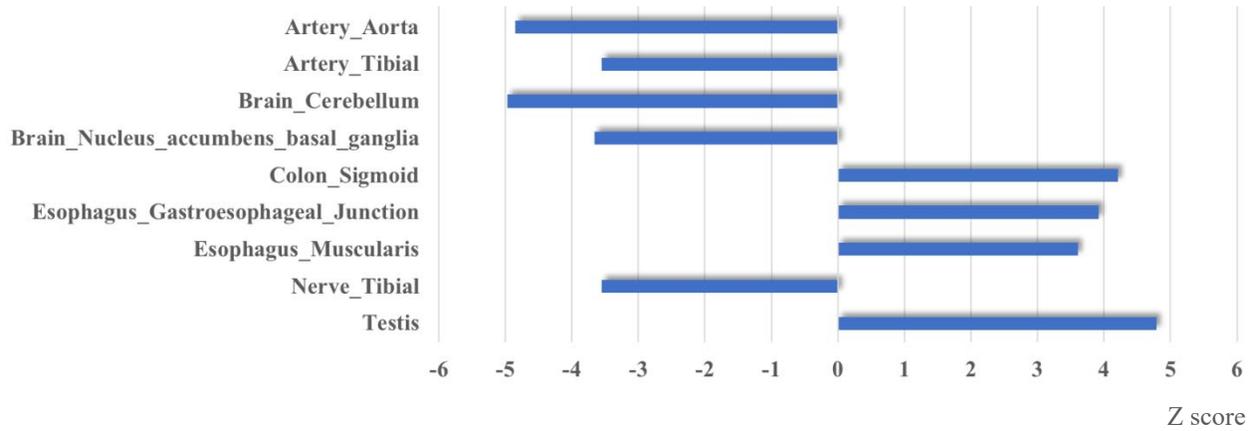
Supplementary Figure 6: Comparing gene expression prediction model performance between FUSION and MetaXcan. **(A)** The number of gene prediction models with cross validation $R^2 > 0.01$ that have better (orange) or worse (blue) performance in FUSION as compared to MetaXcan for LTG, GTEx and the combined LTG+GTEx datasets. Prediction models that have higher prediction performance (R^2) using FUSION as compared to MetaXcan are shown in orange color, while those that have reduced performance in FUSION (i.e. better performance in MetaXcan) are shown in blue. A total of 5,730 gene prediction models had improved performance using FUSION and 4,267 using MetaXcan. The average differences in R^2 were -0.045 and 0.058 (LTG), -0.033 and 0.038 (GTEx), -0.018 and 0.022 (LTG+GTEx), respectively. **(B)** Gene prediction model overlap is shown for the LTG, GTEx and Combined LTG+GTEx datasets between FUSION and MetaXcan. Differences in prediction performance decreases and overlap increases as the transcriptome datasets get larger, indicating improved statistical power in the combined LTG+GTEx dataset. The number of prediction models that are unique to one dataset or shared are listed as numbers (n) and percentages (%).

A.



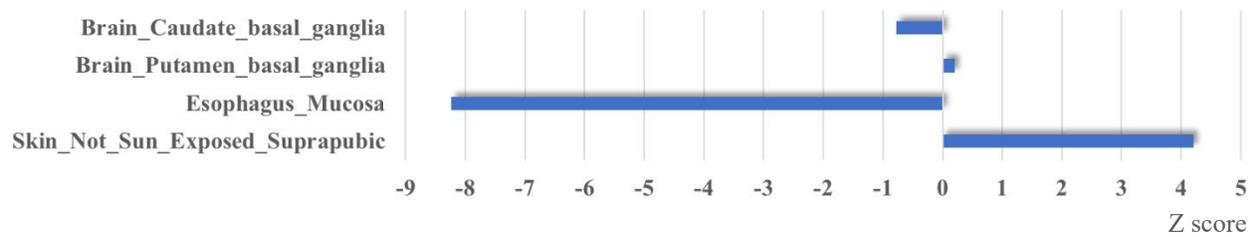
SMC2

B.



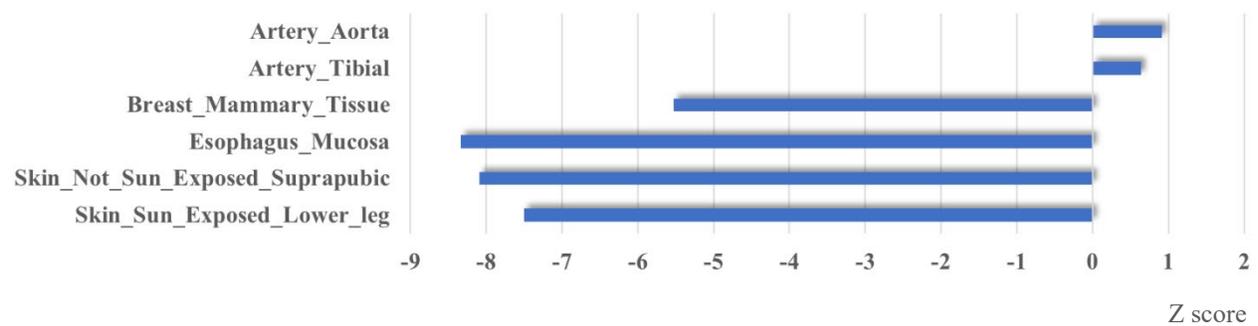
SMC2-AS

C.



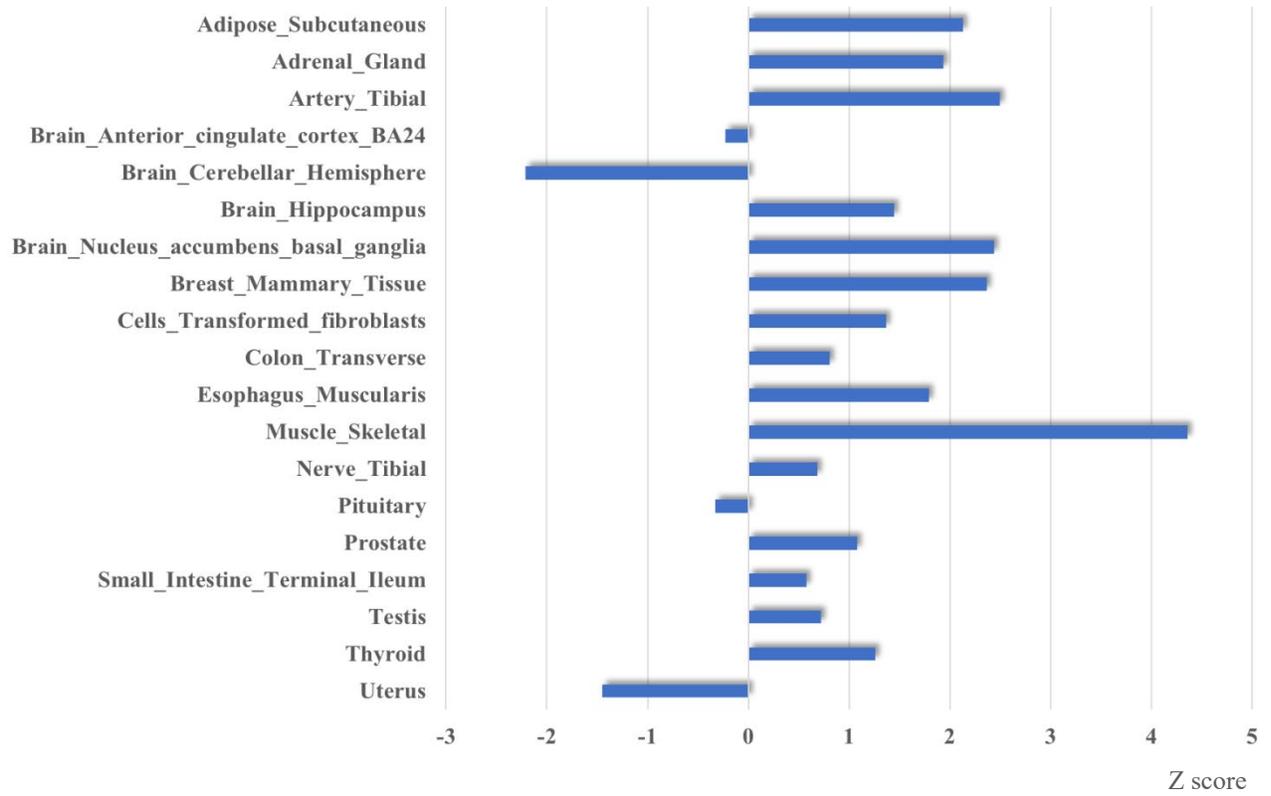
TERT

D.



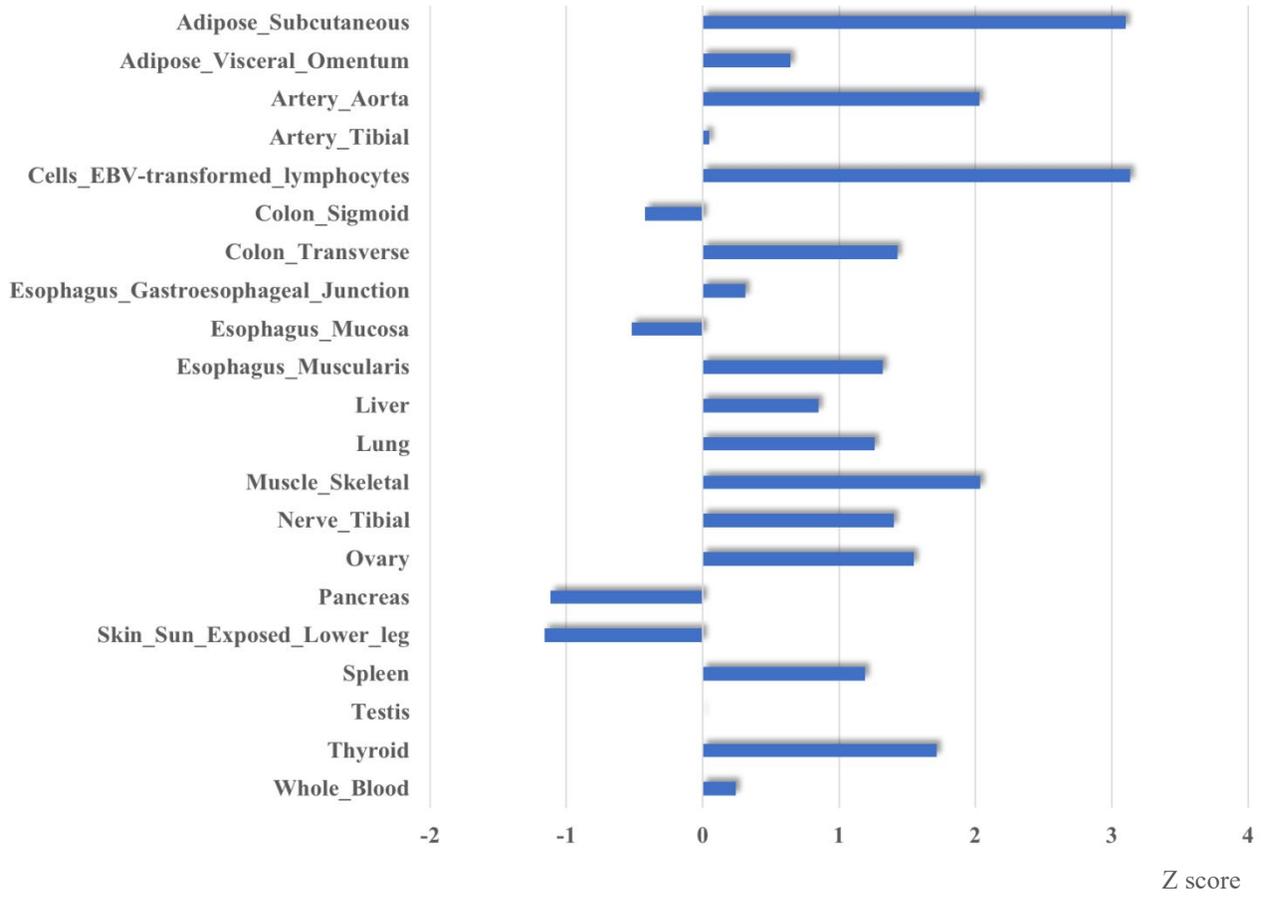
CLPTMIL

E.



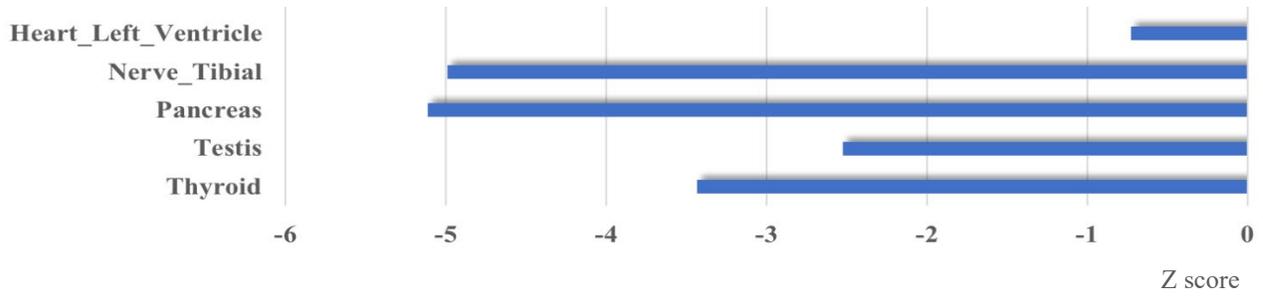
RP11-80H5.9

F.



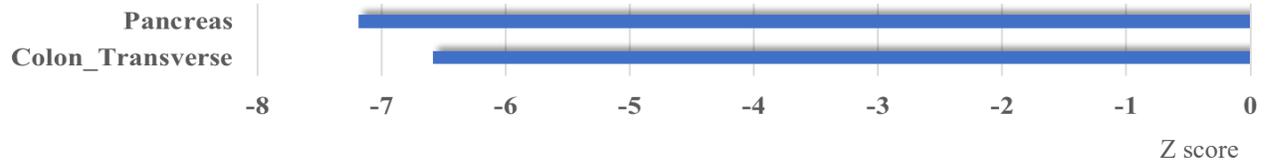
ZDHHC11B

G.



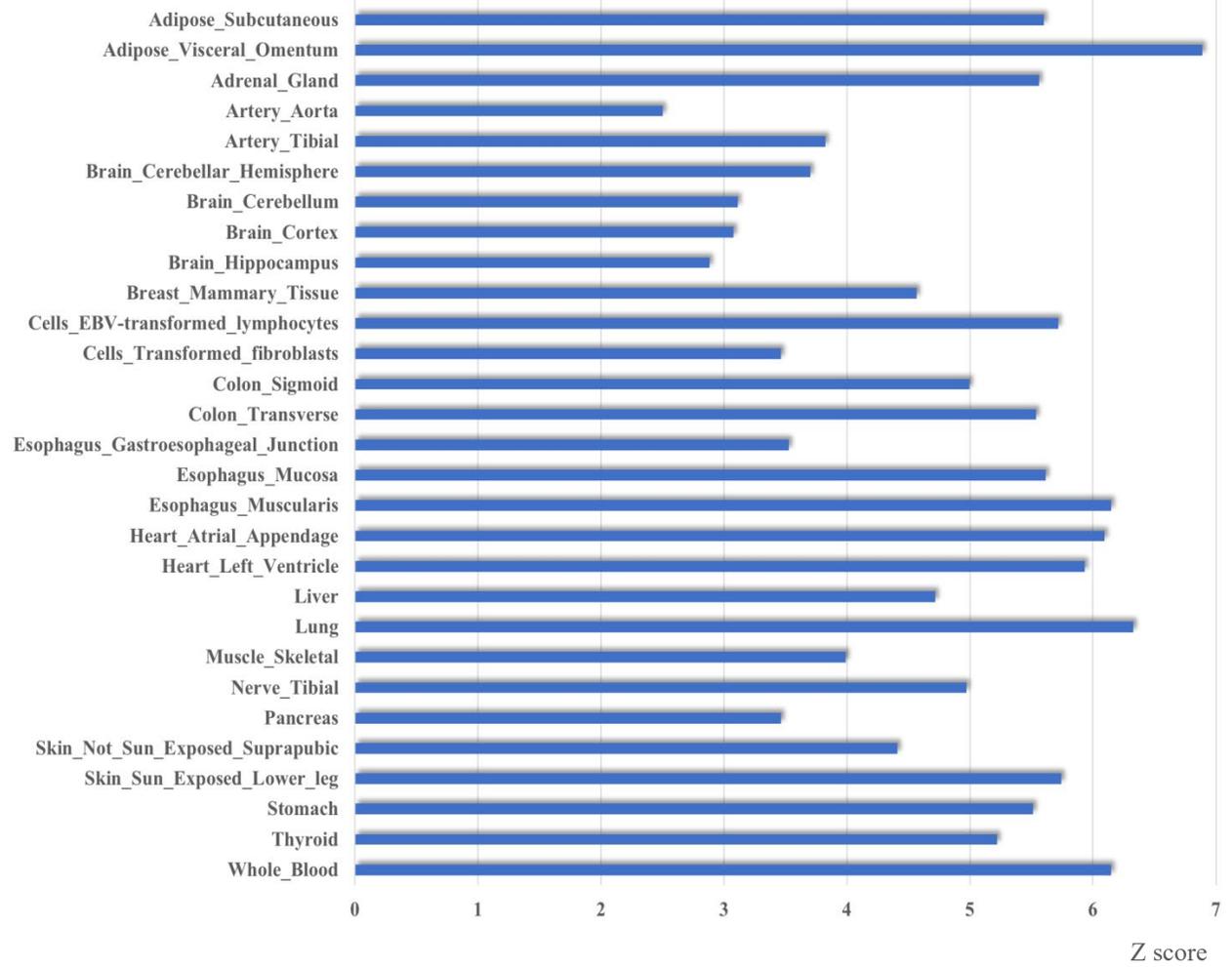
INHBA

H.



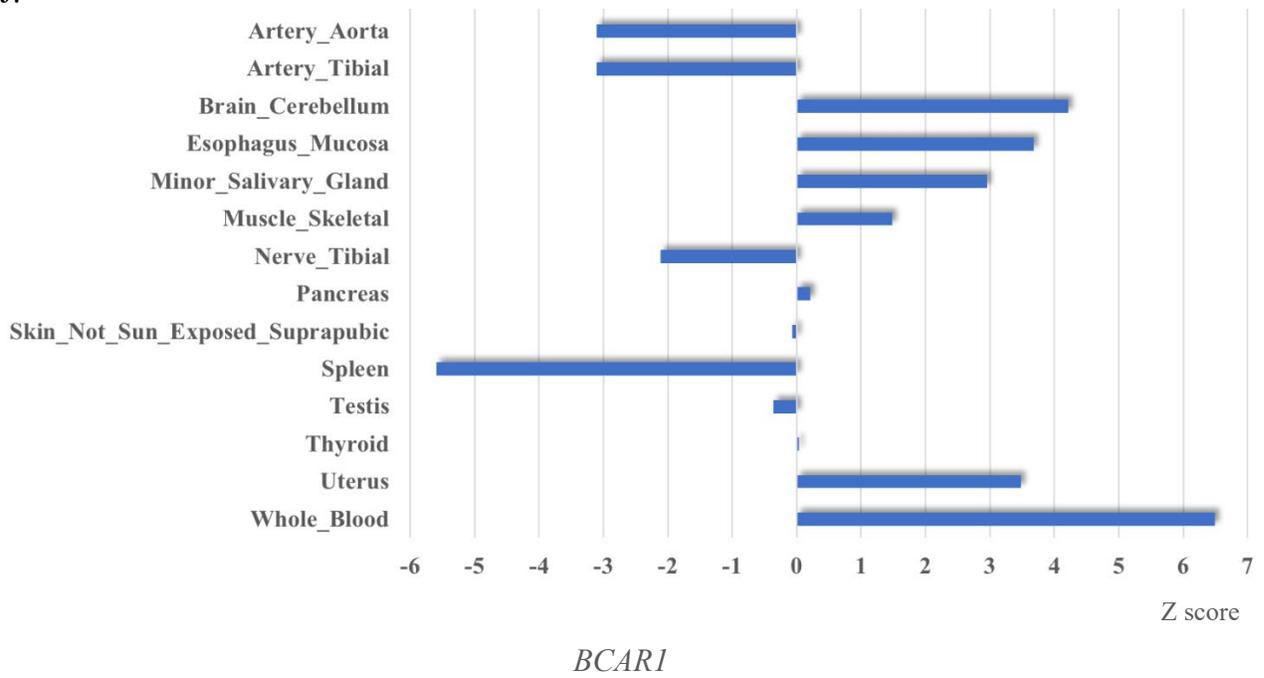
PDX1

I.

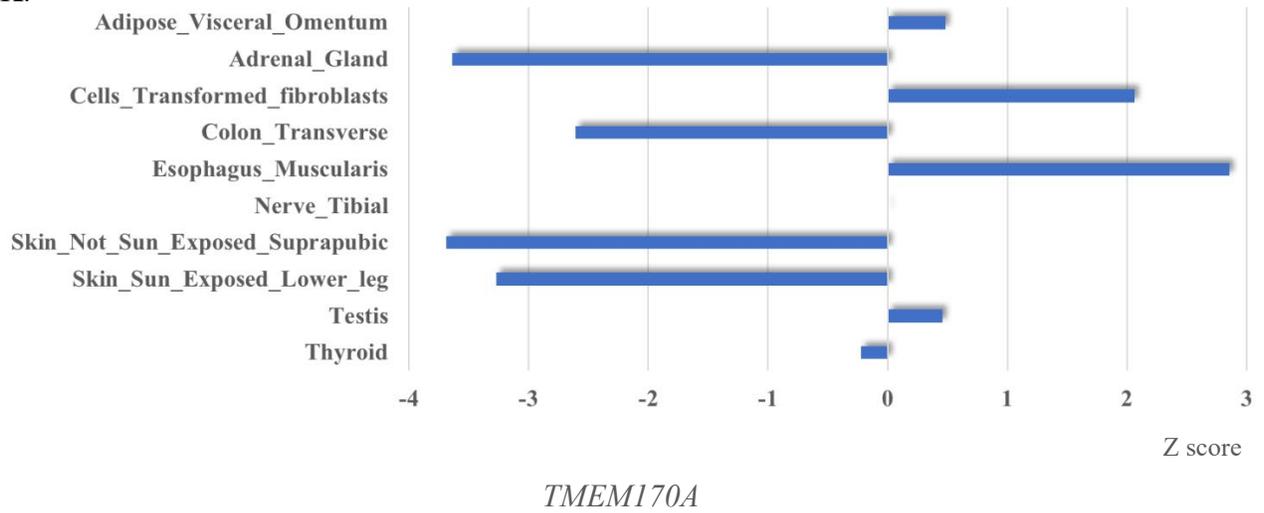


CFDPI

J.

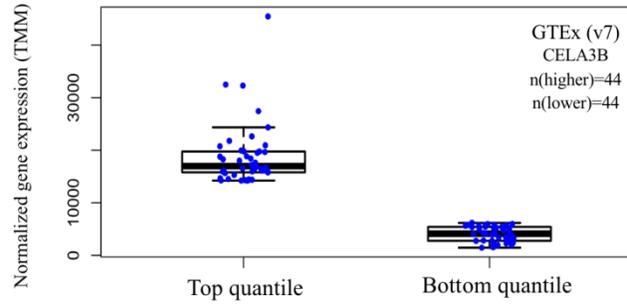


K.

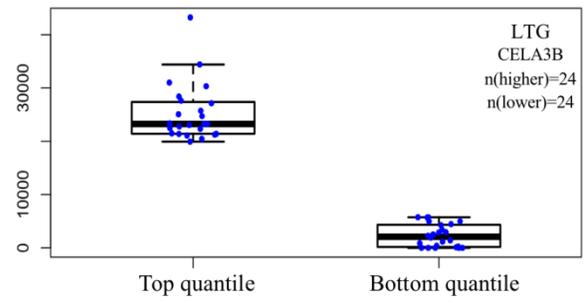


Supplementary Figure 7. TWAS results using cross-tissue expression panels showing effect sizes and direction of effect in different GTEx tissues.

A.



B.



Supplementary Figure 8. *CELA3B* expression in samples in the top and bottom quartiles for the GTEx (A) and LTG (B) transcriptome datasets.

References

1. Zhang M, Lykke-Andersen S, Zhu B, *et al.* Characterising cis-regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. *Gut* 2018;67(3):521-533.
2. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15-21.
3. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* 2017;550(7675):204-213.
4. Conesa A, Madrigal P, Tarazona S, *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;17:13.
5. Wolpin BM, Rizzato C, Kraft P, *et al.* Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat Genet* 2014;46(9):994-1000.
6. Genomes Project C, Auton A, Brooks LD, *et al.* A global reference for human genetic variation. *Nature* 2015;526(7571):68-74.
7. Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81(3):559-75.
8. Das S, Forer L, Schonherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284-1287.
9. Danecek P, Auton A, Abecasis G, *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;27(15):2156-8.
10. Sherry ST, Ward MH, Kholodov M, *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29(1):308-11.
11. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* 2017;33(13):2037-2039.
12. Gusev A, Ko A, Shi H, *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;48(3):245-52.
13. Barbeira AN, Dickinson SP, Bonazzola R, *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 2018;9(1):1825.
14. Gamazon ER, Wheeler HE, Shah KP, *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;47(9):1091-8.
15. International HapMap C, Altshuler DM, Gibbs RA, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467(7311):52-8.
16. Zheng X, Levine D, Shen J, *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 2012;28(24):3326-8.
17. Stegle O, Parts L, Durbin R, *et al.* A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* 2010;6(5):e1000770.
18. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* 1996;58(1):267-288.
19. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2005;67:301-320.
20. Robinson GK. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science* 1991;6(1):15-32.

21. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* 2013;9(2):e1003264.
22. McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279-83.
23. Pasaniuc B, Zaitlen N, Shi H, *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* 2014;30(20):2906-14.
24. Klein AP, Wolpin BM, Risch HA, *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nat Commun* 2018;9(1):556.
25. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, *et al.* Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* 2009;41(9):986-90.
26. Petersen GM, Amundadottir L, Fuchs CS, *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* 2010;42(3):224-8.
27. Childs EJ, Mocchi E, Campa D, *et al.* Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nat Genet* 2015;47(8):911-6.
28. Barbeira AN, Pividori D. M, Zheng J, *et al.* Integrating Predicted Transcriptome From Multiple Tissues Improves Association Detection. *bioRxiv* 2018; <https://doi.org/10.1101/292649>.
29. Yang J, Lee SH, Goddard ME, *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;88(1):76-82.
30. Cobo I, Martinelli P, Flandez M, *et al.* Transcriptional regulation by NR5A2 links differentiation and inflammation in the pancreas. *Nature* 2018;554(7693):533-537.
31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139-40.
32. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 2012;40(10):4288-97.
33. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4(1):44-57.
34. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;37(1):1-13.

Acknowledgements

We thank Laurie Burdett, Aurelie Vogt, Belynda Hicks, Amy Hutchinson, Meredith Yeager and other staff at the National Cancer Institute's Division of Epidemiology and Genetics (DECG) Cancer Genomics Research Laboratory (CGR) for GWAS genotyping. We also thank Bao Tran, Jyoti Shetty and other members of the NCI Center for Cancer Research (CCR) Sequencing Facility for sequencing RNA from histologically normal pancreatic tissue samples (LTG samples). This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH, Bethesda, MD, USA (<http://biowulf.nih.gov>).

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI/Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), GTEx Project March 5, 2014 version Page 5 of 8 Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at <http://www.ncbi.nlm.nih.gov/gap> through dbGaP accession number phs000424.v7.p2. Support for title page creation and format was provided by AuthorArranger, a tool developed at the National Cancer Institute (<https://authorarranger.nci.nih.gov/#/>).

Funding

The work conducted at NCI was supported by the Intramural Research Program (IRP) of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institutes of Health (NIH).

The work conducted at the Vanderbilt University Medical Center was supported in part by R01CA188214 and K99CA218892.

The Melbourne Collaborative Cohort Study cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database.

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. For a list of WHI investigators contributed to WHI science, please visit:

<https://www.whi.org/researchers/SitePages/Principal%20Investigators.aspx>

Cancer incidence data for CLUE were provided by the Maryland Cancer Registry, Center for Cancer Surveillance and Control, Department of Health and Mental Hygiene, 201 W. Preston Street, Room 400, Baltimore, MD 21201, <http://phpa.dhmh.maryland.gov/cancer>, 410-767-4055. We acknowledge the State of Maryland, the Maryland Cigarette Restitution Fund, and the National Program of Cancer Registries of the Centers for Disease Control and Prevention for the funds that support the collection and availability of the cancer registry data.” We thank all the CLUE participants.

The NYU study was funded by NIH R01 CA098661, UM1 CA182934 and center grants P30 CA016087 and P30 ES000260.

The Physicians' Health Study was supported by research grants CA-097193, CA-34944, CA40360, HL-26490, and HL-34595 from the National Institutes of Health, Bethesda, MD USA.

The Women's Health Study was supported by research grants CA-047988, HL-043851, HL080467, and HL-099355 from the National Institutes of Health, Bethesda, MD USA.

Health Professionals Follow-up Study is supported by NIH grant UM1 CA167552 from the National Cancer Institute, Bethesda, MD USA.

Nurses' Health Study is supported by NIH grants UM1 CA186107, and R01 CA49449 from the National Cancer Institute, Bethesda, MD USA.

The PANKRAS II Study in Spain was supported by research grants from Instituto de Salud Carlos III-FEDER, Spain: Fondo de Investigaciones Sanitarias (FIS) ((#PI95/0017, #PI12/00815, #PI13/00082 and #PI15/01573), Red Temática de Investigación Cooperativa en Cáncer (#RD12/0036/0050), and CIBER de Epidemiología (CIBERESP); Ministerio de Ciencia y Tecnología (CICYT SAF 2000-0097); Generalitat de Catalunya (CIRIT - SGR), Spain.

The IARC/Central Europe study was supported by a grant from the US National Cancer Institute at the National Institutes of Health (R03 CA123546-02) and grants from the Ministry of Health of the Czech Republic (NR 9029-4/2006, NR9422-3, NR9998-3, MH CZDRO-MMCI 00209805).

The National Familial Pancreas Tumor Registry at Johns Hopkins University was supported by the NCI Grants P50CA062924 and R01CA97075. Additional support was provided by the Lustgarten Foundation, Susan Wojcicki and Dennis Troper and the Sol Goldman Pancreas Cancer Research Center. The PANC4 GWAS was supported by RO1 CA154823 and federal funds from the National Cancer Institute (NCI), US National Institutes of Health (NIH) under contract number HHSN261200800001E.

The Mayo Clinic Biospecimen Resource for Pancreas Research study is supported by the Mayo Clinic SPORE in Pancreatic Cancer (P50 CA102701).

Funding at Memorial Sloan Kettering was supported by the National Cancer Institute of the National Institutes of Health grant number P30 CA008748.

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10 CA37429 (CD Blanke), and UM1 CA182883 (CM Tangen/IM Thompson) for SELECT.

The PACIFIC Study was supported by RO1CA102765, Kaiser Permanente and Group Health Cooperative.

The Queensland Pancreatic Cancer Study was supported by a grant from the National Health and Medical Research Council of Australia (NHMRC) (Grant number 442302). RE Neale is supported by a NHMRC Senior Research Fellowship (#1060183).

The UCSF pancreas study was supported by NIH-NCI grants (R01CA1009767, R01CA109767-S1 and R0CA059706) and the Joan Rombauer Pancreatic Cancer Fund. Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program; the NCI's SEER Program under contract HSN261201000140C awarded to CPIC; and the CDC's National Program of Cancer Registries, under agreement #U58DP003862-01 awarded to the California Department of Public Health.

The Yale (CT) pancreas cancer study is supported by National Cancer Institute at the U.S. NIH, grant 5R01CA098870. The cooperation of 30 Connecticut hospitals, including Stamford Hospital, in allowing patient access, is gratefully acknowledged. The Connecticut Pancreas Cancer Study was approved by the State of Connecticut Department of Public Health Human Investigation Committee. Certain data used in that study were obtained from the Connecticut Tumor Registry in the Connecticut Department of Public Health. The authors assume full responsibility for analyses and interpretation of these data.