Supplemental Material

Pancreatitis-associated *PRSS1-PRSS2* haplotype alters T cell receptor beta (*TRB*) repertoire more strongly than *PRSS1* expression.

Authors: Dongni Fu, Brandon M Blobner PhD, Phil J. Greer MS, Robert Lafyatis MD, Melena D Bellin MD and David C. Whitcomb MD PhD for the Prospective Autos study, the GREAT1 study and the North American Pancreatitis Study Group

Online Methods.

NAPS2 cohort

The North American Pancreatitis Study II (NAPS2) is a multicenter, cross-sectional cohort study of recurrent acute pancreatitis (RAP) and chronic pancreatitis (CP) in three phases. The first phase established the primary cohort. [1] followed by a replication cohort [2] for genome-wide association studies.[3]

The current study included patients of European Ancestry, as these were the initial cases and controls that were originally genotyped.[3] Additional, well phenotyped controls from the Pittsburgh, PA, USA region were included with the original NAPS2 controls after comparing genotyping methods and minor allele frequencies (MAF) of important variants. The additional controls were from the Center for Craniofacial and Dental Genetics, Department of Oral Biology, and Department of Oral and Craniofacial Sciences School of Dental Medicine (see acknowledgements)

The final cohort of cases (n=475 RAP; n=866 CP, total=1247) and controls (n=5691) were similar in ancestry but the cases were older (29 +/- 19 years controls vs 49 +/- 16 years cases, p<1e-10), male (50.2% controls vs 41.3% cases, p=8e-10) and had a greater body mass index

(BMI)(20 +/1 13 kg/m² controls vs 25 +/- 7.3 kg/m² cases, p=1e-10).

NAPS2 cohort genotyping

The NAPS2 cohort (ClinicalTrials.gov NCT01545167) and NAPS2 GWAS consortium data was used for population-based studies. All blood samples for genomic DNA were obtained from consented patients under IRB-approved protocols. DNA was genotyped for a genome-wide association study (GWAS) using the Illumina HumanOmniExpress BeadChip[3], and HumanCoreExome BeadChip V1 (Illumina, San Diego, CA) as previously described in detail.[3, 4] Detailed demographic information, additional genotyping, imputation and analysis methods have also been described.[3, 4] The initial SNP array was limited to European ancestries, to we chose patients of European ancestries for population genetics studies.

Sample preparation of total RNA-Seq and scRNA-Seq

Pancreatic tissue collection. Pancreas tissue samples from consented and well phenotyped individuals undergoing pancreatic surgeries were obtained at UPMC through the Genomic Resources for Enhancing Available Therapies (GREAT-1, NCT04306939) study, the Prospective Study of Outcomes Following Pancreatectomy and Autologous Islet Transplantation for Treatment of Pancreatic Disease ("Prospective Autos"; IRB# 0609M91887) at the University of Minnesota or the Prospective Observational Study of TPIAT (POST) study (NCT03260387) at the University of Pittsburgh and the University of Minnesota. The tissue from the university of Pittsburgh was used for both total RNA-Seq and scRNA-Seq and included "normal" pancreas as surgical waste during primary resection of tumors of the distal bile duct (n=2, Whipple) and mucinous cystic neoplasm of the tail (n=1).

The remaining samples were from total pancreatectomy with islet autotransplantation (TPIAT). TPIAT is performed for recurrent acute pancreatitis (RAP) or early-established chronic pancreatitis (CP) while pancreatic function remains (especially islet function) so that significant pancreatic tissue is present after resection.

TPIAT samples from Pittsburgh for RAP and duct obstruction with early CP (n=1), hereditary pancreatitis (n=1) idiopathic CP (n=1) and alcohol-associated CP (n=2).
The tissues from Minnesota were from patients undergoing TPIAT for pancreatitis etiologies of hereditary (n=4), idiopathic (n=4) and alcohol-associated (N=1).

Tissue processing: Pieces of fresh pancreatic tissue samples were divided and immediately processed for histology, RNA-seq using RNAlater™ or single cell isolation for scRNA-Seq as previously described.[5]

RNA transcriptomics. Total RNA extraction and library preparation was performed at the University of Pittsburgh Genomics Research Core using the manufacturers recommendations as previously described.[6] Barcoded RNA-seq libraries were sequenced using an Illumina NextSeq 500 (Illumina, SanDiego,Calif) with 75 cycle high output paired end sequencing. Reads were aligned to GRCh38. The specific method of alignment and quality control was the same as previously described.[6]

Single cell RNA-Seq. Samples for scRNA-Seq were minced with sterile scissors, and then incubated in 10ml DMEM/F12 with 1mg/ml collagenase+protease inhibitor (Sigma, cat

#C6079) at 37°C for 20 min. A vortex mixer was used to resuspend cells every 5 minutes to created single cell suspensions as previously described in detail.[5] Cell suspensions were transferred to the University of Pittsburgh Health Sciences Core Research Facilities Single Cell RNA Sequencing laboratory (Robert Lafyatis MD, Director) for single cell cDNA library preparation using droplet-based technology from 10X Genomics®. Libraries were generated according to the 10X Genomics Chromium Single-Cell 3' v2 protocol (https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v2-chemistry). The library was loaded onto an Illumina NextSeq500 with 2x75 cycle paired end sequencing by the University of Pittsburgh Genomics Research Core. Single-cell libraries were sequenced at the UPMC Genome Center (Pittsburgh, PA). Transcriptome data is publicly available at: https://github.com/Whitcomb-Lab/scRNA-Seq_analysis.

Bioinformatic analysis

We previously demonstrated the risk for RAP/CP associated with the *PRSS1-PRSS2* rs10273639_C haplotype.[3] Since rs10273639 is in a non-coding region of the locus (5' to PRSS1 in GRCh37, intronic in GRCh38), we identified a linked SNP within the *PRSS1* coding region as a surrogate for the risk/protective haplotype in transcription analysis. *PRSS1* SNP rs6667_C (p.N246=) is in linkage disequilibrium (LD) with rs10273639 (r^2 = <0.8 to 1.0 depending on cohorts) (Figure 1A). The Integrative Genomics Viewer (IGV)[7] was used for counting *SPINK1* rs17107315 C and T allele transcripts and *PRSS1-PRSS2* rs6667 C and T transcripts in RNA-Seq bam-files.

TOPHAT[8] was used for alignment and expression analysis and Seurat[9] was used for single-cell RNA-Seq data. Cells expressing < 200 and > 3000 genes or > 10% mitochondrial RNA were excluded. For single-cell analysis, a Jackstraw test was performed on the remaining cells to identify the significant principal components. Cell clusters were identified first by feature plots, then by identifying the cluster markers and submitting the cluster markers to CellMarker[10] (http://biocc.hrbmu.edu.cn/CellMarker/) and PanglaoDB[11] (https://panglaodb.se/). Uniform Manifold Approximation and Projection (UMAP)[12] (https://arxiv.org/abs/1802.03426) was performed for dimensional reduction. STRING[13] (https://string-db.org/) was used to conduct pathway analysis.

Quantification and Statistical Analysis

In RNA-Seq, relative expression of *SPINK1* rs17107315 T and C transcripts were analyzed by Welch Two Sample t-test (two-sided), and *PRSS1-PRSS2* rs6667 T and C transcripts were analyzed by Wilcoxon signed ranked test since the distribution of the data was non-normal. The frequency of *SPINK1* rs17107315 TC, TT and *PRSS1-PRSS2* rs10273639 TT, TC and CC in GWAS were performed in the case-control association using the chi-square (χ^2) test. A p-value of less than 0.05 was considered statistically significant. Analysis was performed by R 3.6.2.[14]

UK Biobank Replication

A replication study of the *SPINK1* prevalence and *SPINK1, PRSS1-PRSS2* risk haplotype interaction was performed on tag SNPs in the UK Biobank with pancreatitis patients

phenotyped as previously reported by Ariel Precision Medicine.[15] The UK Biobank cohort is primarily of subjects > 40 years of age and there are fewer early onset and idiopathic pancreatitis patients.

Pancreatitis subjects from the UK Biobank (n=1249) were classified by ICD-10 codes and compared with matched controls (n=23256) as previously described[15]. The MAF of the *SPINK1* p.N34S risk allele is 1.3% in controls and 2.5% in CP+RAP (p-value = 4.095e-06) confirming an association with moderate effect size and replicating the NAPS2 cohort findings.

The effect of the *PRSS1-PRSS2*_rs10273639 genotypes on *SPINK1*_rs17107315TC (risk) was calculated. As with NAPS2, the risk of RAP/CP with *SPINK1*_rs17107315TC on the PRSS1-PRSS2_rs10273639TT (protective) background was increased, but did not reach statistical significance (OR=1.37, CI 0.51-3.00 p=0.37). Unlike NAPS2, the risk of *SPINK1*_rs17107315TC on the rs10273639TC in the UK Biobank did not reach statistical significance (OR=1.39, CI 0.86=2.15; p=0.13), but like NAPS2, the *SPINK1*_rs17107315TC on the *PRSS1-PRSS2*_rs10273639CC (risk) background was significant (OR=2.57, CI 1.60-3.00, p= 9.16e-05), replicating the NAPS2 findings.

The test of interaction between the *SPINK1* and *PRSS1-PRSS2* risk haplotypes in the UK Biobank data was non-significant, as seen in NAPS2 for either the *SPINK1* and *PRSS1-PRSS2* risk haplotypes (3 x 2 Chi-squared: p = 0.11) or the PRSS1-PRSS2 protective (TT) and PRSS1-

PRSS1 risk (TC + CC) genotypes and the SPINK1 TT and TC genotypes (2 x2 Chi-squared; OR 0.67, CI 0.25-1.53, p=0.47).

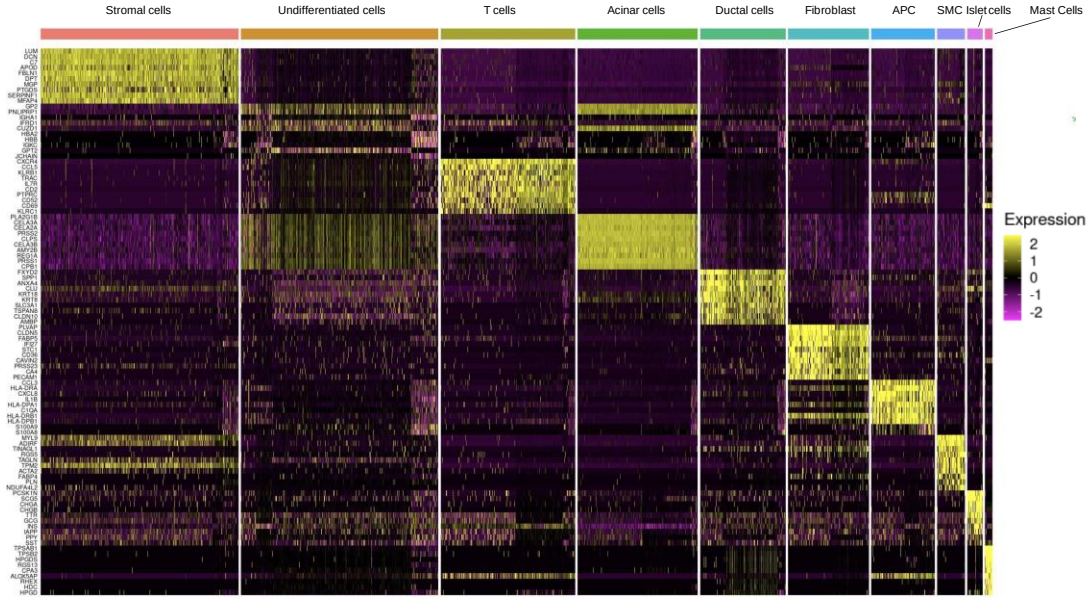
References for Supplemental Information.

References

1. Whitcomb DC, Yadav D, Adam S, et al. Pancreatology 2008;8:520-31.
2. Conwell DL, Banks PA, Sandhu BS, et al. Dig Dis Sci 2017;62:2133-2140.
3. Whitcomb DC, Larusch J, Krasinskas AM, et al. Nature genetics 2012;44:1349-54.
4. Dunbar E, Greer PJ, Melhem N, et al. J Gastroenterol 2020;55:1000-1009.
5. Blobner BM, Saloman JL, Shelton Ohlsen CA, et al. Am J Physiol Gastrointest Liver Physiol 2021;321:G449-G460.
6. Blobner BM, Bellin MD, Beilman GJ, et al. Pancreas 2020;49:1037-1043.
7. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Nat Biotechnol 2011;29:24-6.
8. Trapnell C, Pachter L, Salzberg SL. Bioinformatics 2009;25:1105-11.
9. Stuart T, Butler A, Hoffman P, et al. Cell 2019;177:1888-1902 e21.
10. Zhang X, Lan Y, Xu J, et al. Nucleic Acids Res 2019;47:D721-D728.
11. Franzen O, Gan LM, Bjorkegren JLM. Database (Oxford) 2019;2019.
12. McInnes L, Healy J, Melville J. arXiv 2018:https://arxiv.org/abs/1802.03426.
13. von Mering C, Jensen LJ, Snel B, et al. Nucleic Acids Res 2005;33:D433-7.
14. Team RC. R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing, 2019.
15. Spagnolo DM, Greer PJ, Ohlsen CS, et al. Clin Transl Gastroenterol 2022;19:e00455.

Supplemental Table S1. The enriched KEGG pathways in Cluster 6 versus Acinar cells based on gene expression signatures highly expressed in cluster of cells with high PRSS1/SPINK1 ratios.

Top 125 pathways	Acinar cells	Cluster 6
1	hsa04971:Gastric acid secretion	hsa05144:Malaria
2	hsa04976:Bile secretion	hsa05150:Staphylococcus aureus infection
3	hsa04657:IL-17 signaling pathway	hsa05152:Tuberculosis
4	hsa04964:Proximal tubule bicarbonate reclamation	hsa04145:Phagosome
5	hsa05417:Lipid and atherosclerosis	hsa04612:Antigen processing and presentation
6	hsa05208:Chemical carcinogenesis - reactive oxygen specie	hsa05140:Leishmaniasis
7	hsa04972:Pancreatic secretion	hsa04610:Complement and coagulation cascades
8	hsa05020:Prion disease	hsa05323:Rheumatoid arthritis
9	hsa04915:Estrogen signaling pathway	hsa04620:Toll-like receptor signaling pathway
10	hsa05134:Legionellosis	hsa04621:NOD-like receptor signaling pathway
11	hsa04260:Cardiac muscle contraction	hsa04668:TNF signaling pathway
12	hsa05167:Kaposi sarcoma-associated herpesvirus infection	hsa05142:Chagas disease (American trypanosomiasis)
13	hsa04714:Thermogenesis	hsa05164:Influenza A
14	hsa00190:Oxidative phosphorylation	hsa05321:Inflammatory bowel disease (IBD)
15	hsa04918:Thyroid hormone synthesis	hsa05310:Asthma
16	hsa04141:Protein processing in endoplasmic reticulum	hsa05133:Pertussis
17	hsa04612:Antigen processing and presentation	hsa05143:African trypanosomiasis
18	hsa04668:TNF signaling pathway	hsa05330:Allograft rejection
19	hsa05012:Parkinson disease	hsa04657:IL-17 signaling pathway
20	hsa05415:Diabetic cardiomyopathy	hsa05322:Systemic lupus erythematosus
21	hsa04530:Tight junction	hsa04060:Cytokine-cytokine receptor interaction
22	hsa04010:MAPK signaling pathway	hsa04672:Intestinal immune network for IgA production
23	hsa05418:Fluid shear stress and atherosclerosis	hsa04062:Chemokine signaling pathway
24	hsa04024:cAMP signaling pathway	hsa04979:Cholesterol metabolism
25	hsa05166:Human T-cell leukemia virus 1 infection	hsa05145:Toxoplasmosis



Supplemental Figure S1. Heatmap showing top 10 genes for 10 clusters in scRNA-Seq. The top 5 cell marker genes from each cell type include: **Stromal cells**: LUM, DCN, C7, APOD, FBLN1; **Undifferentiated cells**: *GP2, PNLIPRP1, IGHA1, IFRD1, CUZD1;* **T cells**: *CXCR4, CCL5, KLRB1, TRAC, IL7R;* **Acinar cells**: *PLA2G1B, CELA3A, CELA2A, PRSS2, AMY2B*; **Ductal cells**: *FXYD2, SPP1, ANXA4, CLU, KRT18;* **Fibroblast**: *PLVAP, CLDN5, FABP5, IFI27, STC1:* **APC**: *CCL3, HLA-DRA, CXCL8, IL1B, HLA-DPA1;* **Smooth Muscle Cells:** *MYL9, ADIRF, TINAGL1, RGS5, TAGLN;* **Islet cells**: *PCSK1N, SCG5, CHGA, CHGB, TTR*; **Mast cells**: *TPSAB1, TPSB2, HPGDS, RGS13, CPA3.*