Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Pelin Ünal

born in: Sarıoğlan, Türkiye

Oral Examination: 24 September 2025

Identification of Novel Germline Genetic Variants Associated with Pancreatic Cancer Risk by Using Genome-Wide Studies

Supervisor: Dr. Federico Canzian

Referees: Prof. Dr. Stefan Wiemann

Prof. Dr. Cristian Schaaf

Prof. Dr. Michael Hoffmeister

Dr. Priya Chudasama

ACKNOWLEDGEMENTS

I would like to acknowledge my PhD supervisor, Federico Canzian, for welcoming me to his research group even though I knew nothing about the research topic at the time. He was the first person who believed in me. He gave me opportunity, space and support to grow and earn the Dr. title. This thesis would not be complete without his ideas, his supervision, his corrections, his saying yes to my crazy project ideas, letting me test them. I have gained precious experiences in his guidance. There were times I made significant mistakes, but he has a way of never making you feel inadequate, only encouraged and motivated. Thanks to him, I felt independent researcher but never alone and never without guidance.

I would like to acknowledge Angelica Macauda for being my first friend in Heidelberg and a supporting colleague. She was there for me in every step. She contributed with insightful ideas and solutions to my thesis. Her guidance and friendship were keeping me from quitting my PhD when I couldn't even install the STAAR R package. She is the person who taught me the basics of genetic epidemiology and how to enjoy good food.

I would like to acknowledge Angelika Stein for making our team feel like family. Every time there was a problem, she was there to solve it and make the day enjoyable with her stories. Without her help, I would not be able to survive the German bureaucracy. I am grateful her presence in my life. I will be inspired by her all my life.

I would like to acknowledge Murat Güler for being my dearest friend, PhD body and problem-solver during my whole PhD time. With his knowledge and talent, I was able to see my mistakes and resolve them in time. Every conversation with him brought me new perspectives and insights. He was there for me and with me while I completed this thesis. I am deeply grateful for his invaluable contributions for this thesis.

I would like to acknowledge Daniele Campa, Manuel Gentiluomo, and the entire Pisa team for their support with my methods and analyses. I am deeply grateful for their warm welcome into their research group and their invaluable assistance. I would also like to thank Ye Lu for sharing her results and contributing to my projects.

I would like to acknowledge Emma Barreto Melian and Yongping Chen for providing valuable feedback on my thesis draft. I am also grateful to all past and present members of the Genomic Epidemiology (C055) group for their contributions throughout my PhD journey and for sharing so many enjoyable laughs along the way.

I am also deeply grateful to my best friends Mehmet, Aysu, Simge, Ecem, Boğaç, as well as my friends in Heidelberg, who became like family during my PhD years. Their warmth, kindness, and endless support turned a foreign place into a home and made this journey not only bearable but truly memorable.

To my partner, Emil, whose patience, love, and belief in me carried me through moments of doubt. I am endlessly thankful to him for always being by my side.

Finally, I wish to express my deepest gratitude to my family, especially to my parents Fatma and Nuri. Although I have been living away from them since I was 14 years old, they never allowed me to feel alone, not even for a moment.

Above all, I am deeply thankful to every single person who has touched my life and shaped my journey. This work carries a piece of each of you within it. This thesis is more than a collection of scientific findings; it is a testament to the people who believed in me when I doubted myself, and to the kindness, guidance, and encouragement I received along the way.

"Hayatta en hakiki mürşit ilimdir."

"In life, the truest guide is science."

– Mustafa Kemal Atatürk

ABSTRACT

Pancreatic cancer, whose most common form is pancreatic ductal adenocarcinoma (PDAC), is one of the most aggressive and fatal cancers, with minimal progress in early detection and treatment. While genome-wide association studies (GWAS) have identified several common germline variants, a large proportion of PDAC heritability remains unexplained. In this thesis project, I applied post-GWAS and secondary analysis strategies to uncover novel genetic risk factors, specifically focusing on regulatory regions, rare variants, and genes involved in cognition network.

In the first part of the study, I explored polymorphisms in transcription factor binding sites (TFBS) and enhancers. Through meta-analyses across independent cohorts, I identified rs2472632 near *CCDC34* as a candidate regulatory variant, supported by enhancer signals and transcription factor motif overlap. This finding highlights a potential mechanism of PDAC risk originating from non-coding genomic elements. The second part of the project focused on genome-wide rare variant association analyses. Using both annotation-weighted and unweighted gene-based tests, I discovered significant associations in several genes, including *RIPK2*, *PTPRT*, *CLU*, and *PARD3*, that are functionally linked to cell signaling, immune regulation, and epithelial structure. These genes provide promising leads for understanding the role of rare germline variation in PDAC. Finally, I investigated a novel hypothesis connecting cognition-related genes to PDAC risk. I identified significant associations in genes such as *RP11-255M2.3, LGR4*, *RORA* and *LINC00907* across two large datasets. This supports the intriguing possibility that neurogenetic factors may contribute to cancer susceptibility through stress response or neuroendocrine regulation.

By combining large-scale datasets with advanced post-GWAS methods, this thesis reveals overlooked layers of genetic risk in PDAC. Despite challenges such as dataset heterogeneity and replication of rare variants, the findings offer new perspectives and lay the groundwork for future functional and translational studies in cancer genomics.

Keywords: Pancreatic ductal adenocarcinoma, germline variation, genome-wide association study, rare variants, regulatory polymorphisms, transcription factor binding sites, epistasis, post-GWAS analysis, cognition-related genes, cancer genetics.

ABSTRACT

Bauchspeicheldrüsenkrebs, dessen häufigste Form das duktale Adenokarzinom des Pankreas (PDAC) ist, gehört zu den aggressivsten und tödlichsten Krebserkrankungen, mit bislang nur geringen Fortschritten bei Früherkennung und Behandlung. Während genomweite Assoziationsstudien (GWAS) mehrere häufige Keimbahnvarianten identifiziert haben, bleibt ein großer Teil der PDAC-Erblichkeit ungeklärt. In diesem Dissertationsprojekt wendete ich Post-GWAS- und Sekundäranalyse-Strategien an, um neue genetische Risikofaktoren zu identifizieren, wobei ich mich auf regulatorische Regionen, seltene Varianten und Gene des Kognitionsnetzwerks konzentrierte.

Im ersten Teil der Studie untersuchte ich Polymorphismen in Transkriptionsfaktor-Bindungsstellen und Enhancern. Durch Metaanalysen in unabhängigen Kohorten identifizierte ich rs2472632 nahe *CCDC34* als potenzielle regulatorische Variante, die durch Enhancer-Signale und die Überlappung von Transkriptionsfaktor-Motiven gestützt wird. Dieser Befund weist auf einen möglichen Mechanismus hin, bei dem nicht-kodierende genomische Elemente das PDAC-Risiko beeinflussen. Im zweiten Teil führte ich genomweite Assoziationsanalysen seltener Varianten durch. Mit annotationsgewichteten und ungewichteten genbasierten Tests entdeckte ich signifikante Assoziationen in *RIPK2*, *PTPRT*, *CLU* und *PARD3*, die funktionell mit Zell-Signalwegen, Immunregulation und epithelialer Struktur verknüpft sind. Diese Gene liefern vielversprechende Ansatzpunkte für das Verständnis der Rolle seltener Keimbahnvarianten bei PDAC. Abschließend verfolgte ich eine neuartige Hypothese, die Gene mit Bezug zur Kognition mit dem PDAC-Risiko verbindet. Ich identifizierte bedeutende Assoziationen in *LGR4*, *RP11-255M2.3*, *RORA* und *LINC00907* in zwei großen Datensätzen. Diese Ergebnisse unterstützen die Möglichkeit, dass neurogenetische Faktoren über Stressantworten oder neuroendokrine Regulation zur Krebsanfälligkeit beitragen könnten.

Durch die Kombination großer Datensätze mit modernen Post-GWAS-Methoden deckte ich übersehene genetische Risikofaktoren auf und schuf eine Grundlage für zukünftige funktionelle und translationale Studien in der Krebsgenomik.

Schlüsselwörter: Pankreaskarzinom, Keimbahnvariation, genomweite Assoziationsstudie, seltene Varianten, regulatorische Polymorphismen, Transkriptionsfaktor-Bindungsstellen, Epistase, post-GWAS-Analyse, kognitionsbezogene Gene, Krebsgenetik

Publications from this thesis

Ünal, Pelin et al. "Polymorphisms in transcription factor binding sites and enhancer regions and pancreatic ductal adenocarcinoma risk." *Human genomics* vol. 18,1 12. 2 Feb. 2024, doi:10.1186/s40246-024-00576-x

[Submitted] Ünal, Pelin et al. "Genome-wide search for rare germline genetic variants associated with pancreatic ductal adenocarcinoma risk."

[In preparation] Ünal, Pelin et al. "The modern human-specific cognitive genes and pancreatic cancer risk."

CONTENTS

ILLUSTRATIONS

TABLES

ABBREVIATIONS

1KGP: 1000 Genomes Project
A: Adenine
ABC: Activity-By-Contact
ACAT: Aggregated Cauchy Association Test
ANS: Autonomic Nervous System
aPC: annotated Principal Component
ARTP: Adaptive Rank Truncated Product
BBJ: BioBank Japan
BCF: Binary Call Format
BGEN: Binary GENotype file format
C: Cytosine
CAGE: Cap Analysis of Gene Expression
CASSI: Combined Association Score for SNP Interactions
CI: Confidence Interval
cMAC: cumulative Minor Allele Count
CNV: Copy Number Variation
CP: Chronic Pancreatitis
DHS: DNase Hypersensitivity Sites
dbGaP: database of Genotypes and Phenotypes
dbSNP: Single Nucleotide Polymorphism Database
DKFZ: Deutsches Krebsforschungszentrum (German Cancer Research Center)
DNA: DeoxyriboNucleic Acid
EPIC: European Prospective Investigation into Cancer and Nutrition
ESTHER: Epidemiological investigations on chances of preventing, recognizing early and optimally treating CHronic diseases in an Elderly Rural population
eQTL: expression Quantitative Trait Loci
FAVOR: Functional Annotation of Variants Online Resource
FDR: False Discovery Rate
FPC: Familial Pancreatic Cancer
G: Guanine
GDS: Genomic Data Structure
GLM: Generalized Linear Model
GRR: Genotype Relative Risk
GTEx: Genotype-Tissue Expression
GWAS: Genome-Wide Association Studies
GWER: Genome-Wide Error Rate
HRC: Haplotype Reference Consortium
HWE: Hardy-Weinberg Equilibrium
IBD: Identical By Descent
IPMN: Intraductal Papillary Mucinous Neoplasms
IV: Instrumental Variable

JaPAN: Japan Pancreatic Adenocarcinoma Network (or Research)
LD: Linkage Disequilibrium
lncRNAs: long non-coding RNAs
MAC: Minor Allele Count
MAF: Minor Allele Frequency
meQTL: methylation Quantitative Trait Loci
mRNA: messenger RiboNucleic Acid
miRNA: microRNA
ML: Machine Learning
MR: Mendelian Randomization
NCC: National Cancer Center
NGS: Next-Generation Sequencing
OR: Odds Ratio
PAAD: Pancreatic Adenocarcinoma
PanC4: Pancreatic Cancer Case-Control Consortium
PanScan: Pancreatic Cancer Screening and Analysis Consortium
PANDoRA: Pancreatic Disease Research
PCA: Principal Component Analysis
PBWT: Positional Burrows-Wheeler Transform
PC: Principal Component
PCR: Polymerase Chain Reaction
PDAC: Pancreatic Ductal AdenoCarcinoma
pLoF: putative Loss-Of-Function
pLoF_ds: putative Loss-Of-Function or disruptive missense
PNEN: Pancreatic Neuroendocrine Neoplasms
PNI: Perineural Invasion
PRS: Polygenic Risk Score
PWM: Position Weight Matrices
QC: Quality Control
QQ: Quantile-Quantile
RFLP: Restriction Fragment Length Polymorphism
RNA: RiboNucleic Acid
SNP: Single Nucleotide Polymorphism
sQTL: splicing Quantitative Trait Loci
SD: Standard Deviation
SE: Standard Error
SKAT: Sequence Kernel Association Test
STAAR: Set-based Test for Association using Annotation infoRmation
T: Thymine
TF: Transcription Factor
TFBS: Transcription Factor Binding Site
TOPMed: Trans-Omics for Precision Medicine
tRNA: transfer RiboNucleic Acid
VEP: Variant Effect Predictor
VCF: Variant Call Format
WES: Whole Exome Sequencing
WGS: Whole Genome Sequencing

1 INTRODUCTION

1.1 PANCREATIC CANCER

The pancreas is a heterocrine gland located deep in the abdominal cavity, playing a crucial role in regulating energy metabolism and consumption. It consists of two structurally and functionally distinct components: the exocrine pancreas, which includes digestive enzyme-secreting acinar cells and bicarbonate-secreting ductal cells, and the endocrine pancreas, which contains the hormone-secreting islets of Langerhans(Kleeff et al., 2016; Q. Zhou & Melton, 2018). These components are affected by different diseases, while diabetes and rare pancreatic neuroendocrine tumors arise from the endocrine islets, pancreatitis and pancreatic cancers originate in the exocrine pancreas(Q. Zhou & Melton, 2018).

Pancreatic cancer develops when abnormal DNA mutations in the pancreas cause cells to grow and divide uncontrollably, leading to tumor formation(US Preventive Services Task Force et al., 2019). It is one of the most aggressive and lethal malignancies, with a notoriously poor prognosis(Earl et al., 2020; Luo et al., 2020). By the time of diagnosis, the disease is often at an advanced stage and has frequently metastasized. Clinically, pancreatic cancer is a general term for malignant tumors originating in the epithelial cells of glandular structures within the pancreatic ducts, specifically referred to as pancreatic ductal adenocarcinoma (PDAC), which accounts for more than 90% of all pancreatic cancers(Aier et al., 2019; Jin & Bai, 2020).

Several factors contribute to the poor prognosis of pancreatic cancer. The disease is typically diagnosed at an advanced stage due to nonspecific or absent symptoms, a lack of reliable tumor markers, and challenges in early detection through imaging, given the pancreas's anatomical location. Furthermore, pancreatic cancer exhibits highly

aggressive behavior, including perineural and vascular invasion, as well as early distant metastases, making curative surgical resection impossible for most patients. The disease is also resistant to conventional treatments, including chemotherapy, radiotherapy, and molecularly targeted therapies. Genetic and epigenetic alterations and an immunosuppressive tumor microenvironment worsen its complexity.

Population wide screening for pancreatic cancer is impractical due to relatively low frequency in the general population and high costs of screening testing. To address these challenges, there is an urgent need to develop an effective pancreatic cancer risk assessment model to identify individuals at high risk. Such a model could enhance our understanding of pancreatic carcinogenesis and disease progression, facilitate early detection, and improve preventive strategies, ultimately reducing pancreatic cancer incidence and mortality.

1.1.1 Epidemiology of pancreatic cancer

Although survival rates for many cancers have improved significantly in recent years, advances in the diagnosis and treatment of pancreatic cancer have been slow. The five-year survival rate after diagnosis remains at approximately 13% (https://gco.iarc.fr/en, accessed March 10, 2025). Pancreatic cancer is the seventh leading cause of cancer-related deaths worldwide. While incidence rates vary significantly across different regions, global trends indicate that pancreatic cancer diagnoses are rising. It is projected to become the second leading cause of cancer-related mortality in Western countries by 2030(Rahib et al., 2014). In nations with a high Human Development Index, pancreatic cancer incidence is approximately five times higher than in developing countries(Huang et al., 2021). Western Europe and North America report age-standardized incidence rates of 8.5 and 8.0 cases per 100,000 individuals, respectively, compared to only 1.3 cases per 100,000 in Southeast Asia(Ilic & Ilic, 2022). **Figure 1** presents the global distribution of age-standardized incidence rates, and **Figure 2** shows the corresponding mortality rates for pancreatic cancer in 2022, highlighting markedly higher burdens in developed regions.

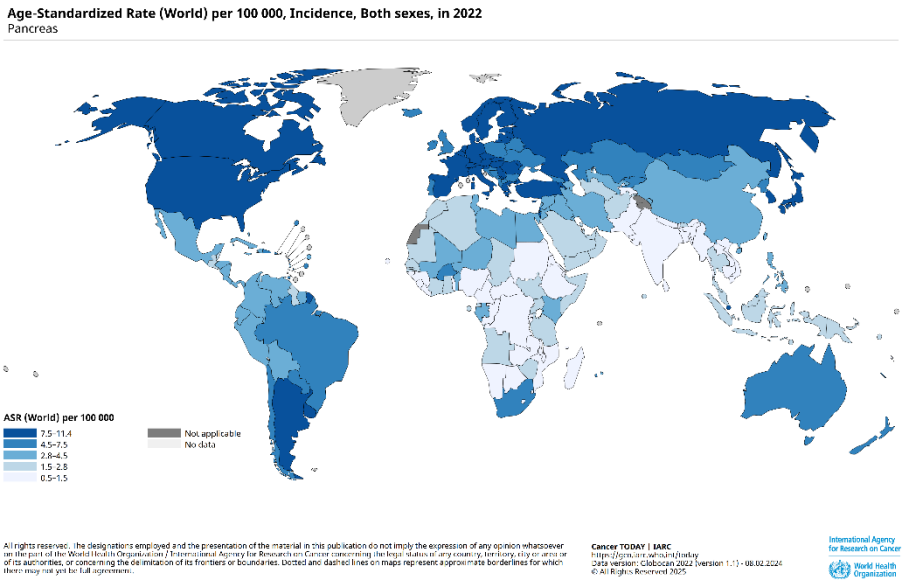


Figure 1. Global Age-Standardized Incidence Rate for Pancreatic Cancer.
Both Sexes, 2022. Figure obtained from the Global Cancer Observatory (https://gco.iarc.fr/today), accessed on 10 March 2025.

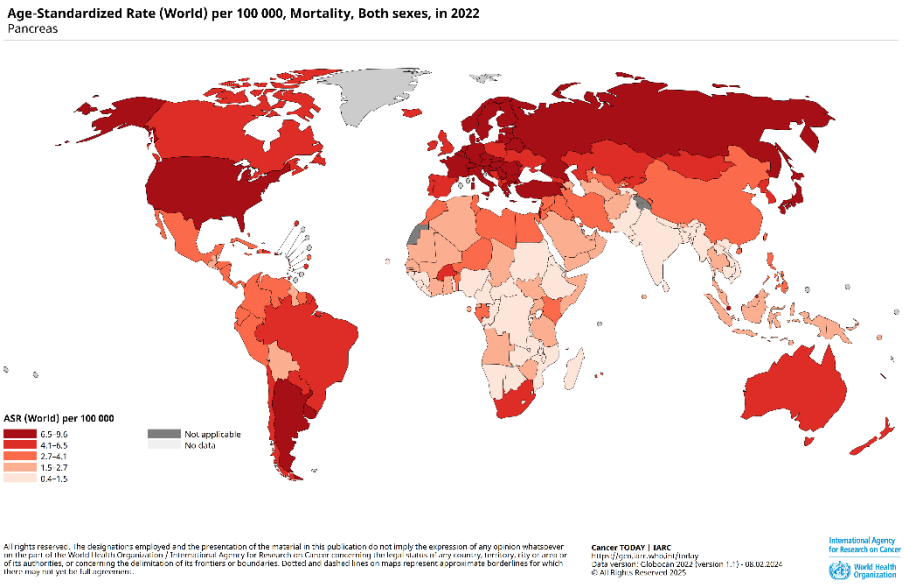


Figure 2. Global Age-Standardized Mortality Rate for Pancreatic Cancer.
Both Sexes, 2022. Figure obtained from the Global Cancer Observatory (https://gco.iarc.fr/today), accessed on 10 March 2025.

In the United States, pancreatic cancer incidence has increased by 0.5% per year since 2010, with an overall lifetime risk of pancreatic cancer estimated at 1.7% by age 75

years(National Cancer Institute. Surveillance, Epidemiology and End Results Program. Available at: www.seer.cancer.gov. Accessed February 10, 2025).

In 2022, there were 510,992 new cases of pancreatic cancer globally, resulting in 467,409 deaths (from https://gco.iarc.fr/today, accessed on 10 March 2025). More than 90% of these cases were PDAC. The disease is slightly more common in males than females, with a male-to-female ratio of 1.4 to 1.0. Common metastatic sites include the liver, lymph nodes, lungs, and peritoneum(Park et al., 2021). The poor five-year survival rate is primarily due to late-stage diagnosis. Only 20% of patients present with early-stage, surgically resectable disease, whereas 50% are diagnosed with metastatic cancer, and the remaining 30% have locally advanced disease with extensive vascular involvement(Siegel et al., 2021). Even among patients who undergo surgical resection, the five-year survival rate remains low, ranging between 15% and 25%.

Pancreatic cancer is rare before the age of 40, but its incidence and mortality rates increase significantly with age(Bosetti et al., 2014; Lynch et al., 2009; Maisonneuve & Lowenfels, 2015). Active surveillance is recommended for people considered at high risk(Goggins et al., 2020), but currently no effective screening method exists. A study by Canto et al. found that approximately 7% of individuals considered at high risk for pancreatic cancer were diagnosed within a 16-year follow-up period. The median time from initial screening to diagnosis was 4.8 years (interquartile range: 1.6-6.9 years). Moreover, long-term surveillance of high-risk individuals has led to earlier detection of operable tumors, improving patient outcomes(Canto et al., 2018). However, research by Corral et al. suggests that endoscopic screening for high-risk individuals has limited efficiency, estimating that 135 high-risk individuals must be screened to detect one case with high-risk pancreatic lesions(Corral et al., 2019). Further research is essential to identify more effective screening markers and strategies for high-risk populations, which could improve early detection and ultimately reduce pancreatic cancer mortality.

1.1.2 Risk Factors

The complete list of pancreatic cancer risk factors remains unknown; however, several modifiable and non-modifiable risk factors are associated with its development. These factors influence the occurrence, progression, and invasion of pancreatic cancer, as summarized below.

1.1.2.1 Modifiable risk factors

Modifiable risk factors include smoking, alcohol consumption, dietary habits, pancreatitis, obesity and infection. Additionally, socioeconomic status is associated with prognosis of pancreatic cancer.

Smoking: Cigarette smoking has the strongest positive association with pancreatic cancer risk(Ghadirian et al., 2003). Patients who smoked prior to diagnosis had a 40% increased hazard for death compared to non-smokers. Nicotine and carcinogens in tobacco smoke directly promote tumor growth, alter tumor-stroma interactions, and enhance myeloid-derived suppressor cell infiltration(Yuan et al., 2017).

Alcohol Consumption: Heavy alcohol consumption (more than three drinks per day) is a key risk factor for chronic pancreatitis (CP), a condition linked to an increased risk of pancreatic cancer. In contrast, low-to-moderate intake is not significantly associated with risk of pancreatic cancer(Rawla et al., 2019).

Dietary Factors: Diets rich in fruits, vegetables, and plant-based foods reduce pancreatic cancer risk, whereas those high in meat, especially red and high-temperature cooked meat, increase risk due to heterocyclic amines and mutagenic activity(Tsai & Chang, 2019)(Stolzenberg-Solomon et al., 2007).

Pancreatitis: Pancreatic cancer burden is significantly linked to acute and chronic pancreatitis(Raimondi et al., 2010). Acute pancreatitis, an inflammatory disease of the exocrine pancreas, may be an early symptom of pancreatic cancer(S. Li & Tian, 2017). CP is a significant risk factor for pancreatic cancer, with relative risks reported to range from 7.6 to 68.1(Kim et al., 2023).

Obesity: Recognized as a strong but modifiable risk factor, obesity is associated with increased pancreatic cancer incidence and poorer outcomes(Xu et al., 2018).

Infection: *Helicobacter pylori* is linked to the etiology of multiple malignancies, including pancreatic cancer. In Western countries, an estimated 4% to 25% of pancreatic cancer cases may be attributable to *H. pylori* infection(Maisonneuve & Lowenfels, 2015). The association is stronger in Europe and East Asia compared to North America(Xiao et al., 2013).

Socioeconomic Status: Socioeconomic disparities impact pancreatic cancer outcomes. Lower-income individuals and racial minorities often experience reduced access to surgery and aggressive treatment, leading to poorer survival rates(W. Sun et al., 2020)(Noel & Fiscella, 2019). In China, urban populations show higher morbidity

and mortality compared to rural populations, likely due to lifestyle and environmental factors(B.-Q. Li et al., 2017).

1.1.2.2 Non modifiable risk factors

Non-modifiable risk factors include age, sex, ethnicity, ABO blood group, diabetes mellitus, and family history of pancreatic cancer.

Age: Pancreatic cancer is predominantly diagnosed in older adults, with 90% of cases occurring in individuals over 55, most commonly between 70-80 years old(McMenamin et al., 2017).

Sex: The incidence of pancreatic cancer is lower in women than in men, particularly in developed countries, potentially due to protective effects of higher steroid levels in women(Klein, 2021)(Masoudi et al., 2017).

Ethnicity: In the U.S., African Americans have a higher pancreatic cancer incidence than Caucasians, while Asian Americans and Pacific Islanders have the lowest rates (Midha et al., 2016). The highest incidence rates are in Europe (7.7/100,000) and North America (7.6/100,000), while Africa has the lowest (2.2/100,000)(Rawla et al., 2019).

ABO Blood Group: Blood group antigens influence pancreatic cancer risk. Individuals with blood types A, AB, or B have a higher risk compared to type O(Risch et al., 2013)(X. Li et al., 2018; Midha et al., 2016).

Family History: Pancreatic cancer has a strong familial component, with first-degree relatives of affected individuals at increased risk(Hruban et al., 2010; Wood & Hruban, 2012). Approximately 80% of pancreatic cancer cases result from sporadic mutations, while a smaller proportion is due to inherited genetic mutations. A Mayo Clinic study identified pathogenic variants in 12% of pancreatic cancer patients with a family history of PDAC. Genetic testing for multiple susceptibility genes is recommended for pancreatic cancer patients with familial history to guide genetic counseling and early diagnosis(Chaffee et al., 2018).

Diabetes Mellitus: Epidemiological studies confirm an increased pancreatic cancer risk among individuals with diabetes(Gardner et al., 2014). Type 1 diabetes increases risk 5-10 times in patients with a disease duration of over 10 years. Similarly, type 2 diabetes is both a risk factor and an early manifestation of pancreatic cancer, with diabetes often preceding cancer diagnosis by several years(Bosetti et al., 2014). Moreover, type 3c diabetes (pancreatogenic diabetes), which arises from pancreatic exocrine dysfunction, is frequently observed in pancreatic cancer patients and may

further complicate diagnosis(Hart et al., 2016). Further research into the relationship between diabetes and pancreatic cancer may enhance early detection strategies.

1.2 GENETIC EPIDEMIOLOGY

Genetic epidemiology was initially defined as the study of the "etiology, distribution, and control of disease in groups of relatives and inherited causes of disease in populations"(Morton N E, 1982). Over time, this definition evolved to emphasize the "role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations"(Khoury et al., 1993). Therefore, the primary goal of genetic epidemiology is to determine whether a genetic component contributes to a disease observed in the population, while considering the influence of environmental factors as well.

1.2.1 Molecular Genetics and Variants

Scientists deciphered the genetic code quickly at the amino acid encoding level and are now successfully identifying and characterizing the genes that are associated with disease. Mutations involving one or more bases (e.g., A to G or T to C) can alter the nucleotide sequence, potentially leading to changes in the amino acid sequence. These changes may result in either substitutions or frameshifts in the genetic code, ultimately affecting the structure and function of the resulting protein changes that may contribute to disease(Reid-Lombardo & Petersen, 2010).

Genetic variation arises from mutations in DNA across the human genome. When these changes do not cause severe effects that are selected against, they may be passed down through generations. Variants that occur in more than 1% of the population are referred to as polymorphisms. Genetic variation can take several forms (**Figure 3**), with commonly studied types including insertions and deletions (indels), copy number variations (CNVs), and single nucleotide polymorphisms (SNPs)(Nussbaum, 2015).

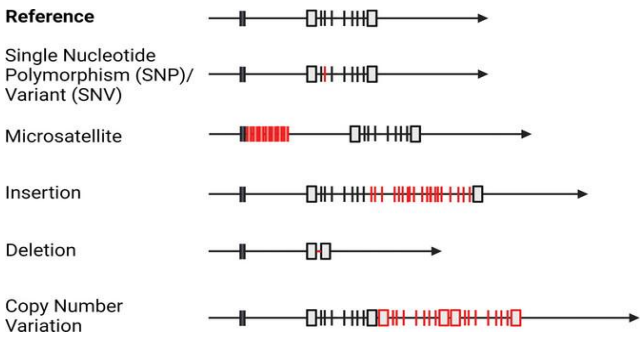


Figure 3. Illustration of Genetic Variations.
Concept based on Kockum et al., 2023.

SNPs are the most common type of genetic variation. For instance, in a healthy individual, a segment of DNA may read GCTACTGCC, while in someone with a disease, it might read GCTACTACC. Each SNP has two possible forms, known as alleles, which can differ at a specific location (locus) within a gene. SNPs can be homozygous (AA or TT) or heterozygous (AT). While the more common allele in a population is called the wild-type allele, the less common allele is called the minor allele. The frequency of the less common allele in a population is referred to as the minor allele frequency (MAF) and this can vary between populations. Genes consist of exons (coding regions) and introns (non-coding regions). Exons are spliced to form mature mRNA. The genetic code translates mRNA into proteins, with start and stop codons marking translation boundaries. Synonymous SNPs code for the same amino acid, non-synonymous SNPs may affect protein function. Only ~1% of the genome strictly codes for proteins; much of the rest plays a role in regulating gene expression. Understanding both genetic and epigenetic changes is crucial for studying complex diseases(Reid-Lombardo & Petersen, 2010).

1.2.2 Population Genetics

Population genetics is fundamentally based on the principle of Hardy-Weinberg Equilibrium (HWE), which describes the expected distribution of genotypes in a large, randomly mating population that is not affected by selection, mutation, migration, or population structure. For a biallelic locus with allele frequencies p (for allele A) and q = 1-p (for allele B), the expected genotype frequencies under HWE are: AA = p^2, AB = 2pq, and BB = q^2. These frequencies remain stable across generations unless one of the equilibrium conditions is violated, in which case deviations may indicate factors such as selection, inbreeding, or technical errors in genotyping process(Panoutsopoulou & Wheeler, 2018).

In terms of genetic similarity, alleles can be identical by descent (IBD), meaning they are inherited from a shared ancestor. IBD is particularly useful for assessing relatedness, with an excess indicating possible kinship or inbreeding(Thompson, 2013). The kinship coefficient quantifies the probability that a gene from one individual is IBD with a gene from another, while the inbreeding coefficient represents the likelihood that both alleles in an individual are IBD(Vigeland, 2020).

Alleles of genetic variants located close together on the same chromosome tend to be inherited together due to limited recombination, a phenomenon known as linkage(Teare, 2011). Linkage disequilibrium (LD) refers to the non-random association of alleles at nearby loci and is commonly quantified by squared correlation (r^2), which ranges from 0 (no correlation) to 1 (perfect correlation)(Hill & Robertson, 1968; Pritchard & Przeworski, 2001). Haplotypes are specific combinations of alleles along a chromosome that are inherited together. When SNPs within a block are in perfect linkage disequilibrium ($r^2 = 1$), only two haplotypes are typically observed, whereas lower LD allows for more haplotype diversity in the population. Regions with high LD are known as haplotype blocks(Cardon & Abecasis, 2003).

1.2.3 Genetic Association Studies

Genetic association studies aim to uncover the relationship between genetic variants and disease risk by analyzing how the presence of specific alleles or genotypes differs between affected individuals (cases) and unaffected individuals (controls)(Panoutsopoulou & Wheeler, 2018). These studies are guided by genetic models, which describe how genotypes at a SNP influence disease risk. For a SNP with two alleles (A and B), the three possible genotypes – AA, AB, and BB – can be modeled using penetrance functions under different assumptions: recessive, dominant, additive, or multiplicative. These models define the probability of disease given a genotype and are used to estimate genotype relative risks (GRRs). For example, under an additive model, disease risk increases linearly with each additional risk allele. In contrast, in a recessive model, risk increases only when two copies of the risk allele are present(Lewis & Knight, 2012).

To evaluate associations between SNPs and disease, researchers often use retrospective case-control studies that compare allele or genotype frequencies between cases and controls(Clarke et al., 2011). The odds ratio (OR) quantifies the strength of association: an OR greater than 1 indicates increased risk, while an OR less than 1 suggests a protective effect. The beta value represents the estimated effect size and its direction—positive for risk, negative for protection. Standard error (SE) reflects the uncertainty

in the beta estimate, and is used to calculate the confidence interval (CI), which provides a range where the true effect size is likely to lie. Narrow CIs indicate more precise estimates. These measures collectively inform the strength, direction, and reliability of associations. In contrast, prospective cohort studies track individuals over time and estimate relative risk by genotype(Grimes & Schulz, 2002).

To explore genetic factors linked to disease, researchers generally use two main approaches: genome-wide association studies (GWAS) and candidate gene association studies. Although both rely on comparing individuals with and without the disease, they differ in their scope and the assumptions behind them.

GWAS are hypothesis-generating studies that scan the entire genome using hundreds of thousands of tag SNPs to pinpoint loci potentially linked to disease. This method does not require prior knowledge of disease biology, making it well-suited for discovering novel susceptibility genes. Given their wide scope, GWAS require large sample sizes, usually more than 1,000 cases and controls, to achieve sufficient statistical power, owing to the very large number of statistical tests being performed. These studies often follow a two-stage design, where SNPs identified in an initial discovery cohort are subsequently tested in an independent replication cohort to validate the results(Uffelmann et al., 2021; Visscher et al., 2017).

In contrast, candidate gene studies are hypothesis-driven, focusing on pre-selected genes believed to be involved in the disease based on biological reasoning. These studies require fewer markers and smaller sample sizes, making them more cost-efficient(Patnala et al., 2013). However, their narrow scope increases the risk of missing important associations in genes or pathways not initially considered.

Together, these strategies, combined with appropriate genetic models and rigorous statistical testing, form the foundation of modern efforts to map the genetic architecture of complex diseases. Building on these approaches, GWAS have transformed our ability to identify genetic risk loci across the genome, while post-GWAS analyses such as genotype imputation, meta-analysis, and functional annotation refine these discoveries and help uncover biological mechanisms. Despite these advances, challenges remain in resolving the contribution of rare and non-coding variants, understanding gene-gene interactions, and translating associations into meaningful insights for disease biology and risk prediction. Addressing these gaps requires integrating diverse analytical strategies and large-scale genomic datasets, which together offer an opportunity to deepen our understanding of the genetic architecture of complex diseases such as pancreatic cancer.

1.2.4 Genetic Data Quality Control

Genotyping determines an individual's genotypes by identifying alleles at specific genome positions, expressed as allele carriage (presence/absence) or dosage (number of copies). In other words, genotype data include information on genetic variants, usually SNPs, measured across the genome for each individual.

Historically, genotypes were inferred through phenotypes, evolving from restriction fragment length polymorphisms (RFLPs) and Polymerase Chain Reaction (PCR)-based methods to high-throughput technologies like DNA microarrays and, more recently, next-generation sequencing (NGS) including whole genome sequencing (WGS) and whole exome sequencing (WES). These advancements have greatly increased the resolution and scale of genotyping.

The choice of genotyping technology depends on factors such as the scope of the study (targeted vs. genome-wide), number of markers, sample size, cost, and computational resources. While genome-wide studies benefit from genotyping arrays or NGS, targeted studies use allele-specific PCR, TaqMan assays, or pyrosequencing.

Quality control (QC) of genotype data procedures help eliminate low-quality samples or markers(variants) and reduce the risk of spurious associations in downstream analyses(M. H. Wang et al., 2019). Low-quality samples and genotype calling errors can lead to an increased rate of false-positive and false-negative results. Because removing a marker may have a greater impact than excluding a single sample, it is advisable to perform QC first at the individual level and then at the variant level, to reduce the risk of discarding a true causal association.

In addition to QC steps, quality assurance is important. Quality assurance refers to the procedures implemented earlier in the data production process to maximize the likelihood that the majority of the data will successfully pass quality control. Examples of quality assurance steps include verifying that a robust study design and appropriate sampling protocols were applied; ensuring that high-quality DNA samples were collected; confirming that DNA extraction and preparation protocols were properly implemented; taking care to balance the handling of cases and controls; routinely monitoring genotyping call rates to identify and address any poorly manufactured chips promptly; and recording detailed information on batch allocation, plate assignment, and genotyping dates to allow later inspection for potential systematic trends(Zondervan & Cardon, 2007).

1.3 FINE MAPPING AND FUNCTIONAL ANNOTATION

Whilst thousands of genetic variants have been associated with human traits, identifying the subset of those variants that are causal requires a further step known as fine-mapping. Fine-mapping seeks to determine the specific genetic variant(s) responsible for the observed association signal within a genomic region, typically identified by a GWAS. The initial GWAS identifies SNPs significantly associated with a trait, but these signals are often due to LD with the true causal variant. Therefore, fine-mapping is essential to distinguish truly causal variants from correlated, non-functional ones(Hutchinson et al., 2020; Schaid et al., 2018).

The basic fine-mapping approach is computationally efficient and uses GWAS summary statistics, but traditionally assumes a single causal variant per associated region, an assumption now widely recognized as biologically unrealistic(Wellcome Trust Case Control Consortium et al., 2012). Multiple causal variants often exist within a locus, and failing to account for this can obscure the true genetic architecture of complex traits. Fine-mapping is typically initiated after genome-wide significant SNPs are identified, focusing on regions of interest often visualized using Manhattan and LocusZoom plots(Pruim et al., 2010). These plots highlight lead or index SNPs, but the lead SNP is not necessarily the causal one due to the use of tag-SNPs and limitations in statistical power.

Zaykin et al. demonstrated that true causal variants may not produce the smallest *p*-values in a region, particularly for traits with small effect sizes(Zaykin & Zhivotovsky, 2005). This finding underscores the need for cautious interpretation of GWAS hits and emphasizes the role of fine-mapping. In population-based studies, fine-mapping exploits LD, non-random associations of alleles at different loci. When loci are closely located on the chromosome, recombination is less likely to separate them, leading to haplotypes. LD is measured as the difference between observed and expected haplotype frequencies, often using Pearson correlation between minor allele counts(Devlin & Risch, 1995).

Using LD for fine-mapping assumes that recombination over generations erodes LD, concentrating association signals near the causal variant. However, evaluating one SNP at a time can mislead due to complex LD patterns, as illustrated in the APOE region and Alzheimer's disease. In this case, extensive LD obscures whether APOE or nearby genes are responsible(Guerreiro & Hardy, 2012; Martin et al., 2000). Moreover, LD is influenced by multiple factors including mutation rates, selection, population structure, and demography, indicating that pairwise LD patterns alone may not suffice for fine-mapping.

The accuracy of fine-mapping is influenced by several factors: the number and effect sizes of causal SNPs, LD structure, sample size, SNP density, and whether causal variants are directly measured. Phenotype definitions can also affect power, particularly if the trait captures more genetically-driven variation (e.g., severe forms of disease)(Hutchinson et al., 2020; Schaid et al., 2018). Functional genomic annotation has become increasingly valuable for prioritizing candidate SNPs from fine-mapping, with major resources such as Gene Ontology(Rhee et al., 2008), GENCODE(Harrow et al., 2012), ENCODE(ENCODE Project Consortium, 2004), FANTOM5(Andersson et al., 2014), and the Roadmap Epigenomics Project(Roadmap Epigenomics Consortium et al., 2015) enabling annotation of majority of the genome.

Protein-coding annotations describe how variants may impact protein structure or function (e.g., exonic location, splice sites). Tools such as CADD and REVEL integrate multiple prediction methods to score coding SNPs for potential pathogenicity(Eilbeck et al., 2017; Mudge & Harrow, 2016). Non-protein-coding annotations address the regulatory genome, where most GWAS signals lie. These include enhancers, promoters, transcription factor binding sites, and epigenetic markers(ENCODE Project Consortium et al., 2007). Regulatory impact scores from tools like FIRE(N. M. Ioannidis et al., 2017), RegulomeDB(Boyle et al., 2012), and CADD(Kircher et al., 2014) help assess non-coding variant function.

Annotation can be integrated into fine-mapping either subjectively or statistically. Bayesian models can incorporate annotations into prior probabilities that a SNP is causal. This has shown modest improvements in narrowing credible sets in simulation and real-data applications, although the gains are often limited. One reason is that annotation databases are incomplete or not specific enough to the disease context. Nevertheless, annotation can be helpful when association signals are moderate or when multiple causal variants exist in high-LD regions(Kichaev et al., 2014; Pickrell, 2014; Sveinbjornsson et al., 2016).

Another layer of functional evidence comes from integrating GWAS with gene expression data. Over 90% of trait-associated variants fall in non-coding regions, often enriched in regulatory elements, and are more likely to be expression quantitative trait loci (eQTL). This supports the hypothesis that genetic variants often affect traits via gene expression regulation(Maurano et al., 2012; Nicolae et al., 2010). Related molecular QTLs highlight additional regulatory mechanisms. For example, splicing QTLs (sQTLs) associate genetic variants with alternative splicing events that generate different transcript isoforms, while methylation QTLs (meQTLs) link variants to DNA methylation changes that can activate or repress gene expression. Together with eQTLs, these approaches provide complementary insights into how genetic variation influences multiple layers of gene regulation and contributes to complex traits(Bykova et al., 2022).

Statistical approaches to integrate eQTL and GWAS data include causal inference tests, Bayesian co-localization, and Mendelian randomization. These methods attempt to test whether the same SNP influences both gene expression and the trait. However, a major challenge is tissue specificity, many diseases involve multiple tissues, and gene expression varies across tissues. The GTEx project addresses this by providing genotype and expression data across dozens of tissue types(Consortium, 2017; Giambartolomei et al., 2014; Hormozdiari et al., 2016; Millstein et al., 2009).

In conclusion, fine-mapping refines the list of candidate causal variants from GWAS hits by leveraging statistical, functional, and expression data. Despite current limitations, such as assumptions in modeling, incomplete annotations, and tissue complexity, fine-mapping remains a critical step toward translating GWAS findings into biological insights and therapeutic targets.

1.4 GENETIC SUSCEPTIBILITY TO PANCREATIC CANCER

The first GWAS for pancreatic cancer was conducted in 2009 as part of the PanScan consortium(L. Amundadottir et al., 2009). Subsequent GWAS, including those from PanScan and the Pancreatic Cancer Case-Control Consortium, expanded the list of risk loci, identifying variants in genes involved in transcriptional regulation, telomere maintenance, and metabolic pathways relevant to pancreatic carcinogenesis(Klein et al., 2018; Petersen et al., 2010; Wolpin et al., 2014; M. Zhang et al., 2016).

GWAS have identified several key loci associated with pancreatic cancer risk (**Table 1**). One of the most consistently replicated loci is *ABO* (9q34), which encodes the blood group antigen and has been repeatedly linked to pancreatic cancer risk. Other significant loci include *NR5A2* (1q32.1), a nuclear receptor involved in pancreatic development and metabolism, and *TERT* (5p15.33), a key regulator of telomere length, which is a well-established risk locus for multiple cancers. Additionally, *HNF4G* (8q21.13), a transcription factor crucial for pancreatic function, and *LINC-PINT* (7q32.3), a long non-coding RNA potentially involved in tumorigenesis, have been identified as pancreatic cancer risk loci. Despite these discoveries, GWAS-identified variants explain only about 4.1% of pancreatic cancer heritability, suggesting that many risk-associated variants remain undiscovered(Klein et al., 2018). Furthermore, the replication of GWAS findings varies across populations; while some loci discovered in European ancestry populations are also found in Asians, others appear to be population-specific(Chang et al., 2018; Lin et al., 2020; C. Wu et al., 2012).

Beyond these common variants, rare high-penetrance germline mutations play a crucial role in pancreatic cancer susceptibility, particularly in familial pancreatic cancer (FPC) and hereditary cancer syndromes. Several genes have been implicated in hereditary pancreatic cancer risk, including *STK11*, *CDKN2A*, *BRCA1/2*, *PRSS1*, *SPINK1*, *CFTR*, *ATM*, *PALB2*, and mismatch repair genes. These genes are associated with various hereditary syndromes, such as Peutz-Jeghers syndrome (*STK11*), familial melanoma (*CDKN2A*), hereditary breast-ovarian cancer (BRCA1/2), and Lynch syndrome (*MLH1*, *MSH2*, *MSH6*, *PMS2*) (Ghiorzo, 2014; Rawla et al., 2019). Among the most well-characterized genes, *BRCA2* (13q13.1) mutations confer a 5- to 10-fold increased risk of pancreatic cancer, particularly in HBOC families(Bartsch et al., 2002; Roberts et al., 2016). Similarly, *PALB2* (16p12.2), a *BRCA2*-interacting partner, is associated with a high pancreatic cancer risk in familial pancreatic cancer families(Jones et al., 2009). *ATM* (11q22.3) mutations, which impair DNA damage repair, are found in approximately 2.4% of familial pancreatic cancer cases(Roberts et al., 2012). *CDKN2A* (9p21.3) mutations result in a 13- to 22-fold increased risk of pancreatic cancer, particularly among melanoma-prone families(Goldstein et al., 2004). *STK11* (19p13.3), the causative gene in Peutz-Jeghers syndrome, is associated with an exceptionally high, 132-fold increased risk of pancreatic cancer(Boardman et al., 1998; Matsubayashi, 2011). Additionally, mismatch repair genes *MLH1*, *MSH2*, *MSH6*, and *PMS2*, known for their role in Lynch syndrome, also contribute to pancreatic cancer susceptibility, with *MLH1* mutations conferring the highest risk(Kastrinos et al., 2009). These findings underscore the importance of both common and rare germline variants in pancreatic cancer genetics, highlighting the need for further research into their functional impact and potential use in risk prediction.

Due to the stringent significance threshold required in GWAS (typically p-value $< 5 \times 10^{-8}$ to correct for multiple testing), many true associations may not reach genome-wide significance. Secondary analyses of GWAS data can help uncover additional risk loci by leveraging different analytical strategies.

Secondary analyses of GWAS data play a crucial role in uncovering additional pancreatic cancer risk loci and refining our understanding of genetic susceptibility. Fine-mapping and functional annotation help identify potential causal variants within GWAS loci by assessing regulatory and coding elements that may influence gene expression or protein function(Fang et al., 2017). For example, expression quantitative trait loci (eQTLs) have been identified where SNPs regulate nearby gene expression, helping to explain associations with PDAC. Non-coding RNAs, including long non-coding RNAs (lncRNAs) and microRNAs (miRNAs), have also gained attention, as genetic variation in these elements may contribute to cancer risk(Corradi et al., 2021; Lu et al., 2021). These functional insights provide a biological context for GWAS signals and help prioritize variants for downstream analyses.

Mendelian Randomization (MR) uses genetic variants as proxies to infer causal relationships between potential risk factors, such as obesity or diabetes, and pancreatic cancer , providing insights into disease etiology(Carreras-Torres et al., 2018; Lu et al., 2020). Pleiotropy scans focus on SNPs that have been previously associated with other diseases or traits, allowing for the identification of novel pancreatic cancer risk variants. For example, *HNF1A*-rs7310409, initially linked to type 2 diabetes, was later found to be associated with pancreatic cancer(Pierce & Ahsan, 2011, p. 1). Additionally, pathway-based analyses examine genetic variation within specific biological pathways, such as inflammation-related genes or metabolic regulators, to identify new associations that may contribute to pancreatic cancer risk(Gentiluomo, Peduzzi, et al., 2019). Together, these secondary analyses enhance the power of GWAS by uncovering previously undetected associations and providing a deeper understanding of pancreatic cancer genetics.

Complementary to these efforts, polygenic risk scores (PRSs) have emerged as a way to estimate an individual's likelihood of developing a specific trait by aggregating the effects of a large set of causal SNPs. PRS are calculated as the weighted sum of risk alleles for a given phenotype in each individual, with weights derived from effect size estimates of the most powerful GWAS on that trait. Studies have shown that PRS can achieve substantially greater predictive power compared to using only a small number of genome-wide significant SNPs, and several tools have been developed to facilitate their calculation(Choi et al., 2020). While PRSs for pancreatic cancer remain in early development(Galeotti et al., 2021), they hold promise for improving risk stratification and informing targeted prevention strategies in high-risk populations.

The estimated number of independent susceptibility variants for pancreatic cancer is in the order of a few thousands according to a method to estimate the degree of polygenicity(Y. D. Zhang et al., 2020); however, only 31 independent loci (p-value $< 5 \times 10^{-8}$) have been discovered so far. This discrepancy highlights that a substantial proportion of heritability remains unexplained, suggesting that many additional risk variants, particularly rare and non-coding ones, await discovery(Childs et al., 2015; Y. Zhang et al., 2018). To address these gaps, secondary analyses of GWAS data that integrate functional annotations and advanced analytical frameworks are essential for identifying novel risk loci and for elucidating the complex genetic architecture of pancreatic cancer. Recent progress in whole-genome sequencing, genotype imputation, and regulatory variant annotation offers unprecedented opportunities to explore underrepresented regions of the genome and to investigate how rare and non-coding variants contribute to disease susceptibility. Building on these advances, this thesis seeks to refine our understanding of pancreatic cancer genetics by investigating rare and regulatory germline variation, aiming to uncover additional susceptibility loci and to provide insights into their potential biological relevance.

Table 1. Established PDAC risk loci from GWAS.

Chromosomal band	CHR	POS (hg38)	rsID	EA	Closest gene	OR[95%CI]	*p*-value	PMID
Loci reaching genome-wide significance (*p*-value < 5×10^{-8}) in at least one published study								
1p36.33	1	959193	rs13303010	G	*NOC2L*	1.26[1.19-1.35]	8.36×10^{-14}	29422604
1q32.1	1	200038304	rs3790844	G	*NR5A2*	0.81[0.76-0.86]	7.62×10^{-16}	29422604
2p24.1	2	21004340	rs183117027	C	*APOB*	2.34[1.72-3.16]	4.21×10^{-8}	30206226
2p14	2	67412637	rs1486134	G	*ETAA1*	1.14[1.09-1.19]	3.36×10^{-9}	26098869
3q28	3	189790682	rs9854771	A	*TP63*	0.9[0.86-0.94]	4.54×10^{-8}	29422604
5p15.33	5	1293971	rs2736098	T	*TERT*	0.84[0.79-0.88]	6.86×10^{-15}	29422604
5p13.1	5	39394887	rs2255280	G	*DAB2*	0.81[0.76-0.87]	4.18×10^{-10}	22158540
7p14.1	7	40827064	rs17688601	A	*SUGCT*	0.88[0.84-0.93]	1.11×10^{-8}	29422604
7p12.3	7	47448971	rs73328514	T	*TNS3*	0.83[0.77-0.88]	4.35×10^{-8}	29422604
7q32.3	7	130995762	rs6971499	C	*LINC-PINT*	0.81[0.76-0.87]	7.41×10^{-14}	29422604
8p21.3	8	21909370	rs2242241	A	*DOK2*	1.85[1.50-2.27]	4.34×10^{-9}	30206226
8q21.13	8	75558169	rs2941471	G	*HNF4G*	0.89[0.85-0.93]	6.60×10^{-10}	29422604
8q24.21	8	127707639	rs10094872	T	*CASC11*	1.15[1.10-1.20]	3.22×10^{-9}	27579533
9q34	9	133273813	rs505922	C	*ABO*	1.27[1.22-1.31]	7.35×10^{-27}	29422604
9p21.3	9	22077544	rs1412832	T	*CDKN2B-AS1/ANRIL*	1.11[1.07-1.15]	5.25×10^{-9}	36451333
10q26.11	10	118519432	rs12413624	T	*SLC25A18P1*	1.23[1.16-1.31]	5.12×10^{-11}	22158540
12q13.13	12	52902133	rs2035875	A	*KRT8*	1.11[1.08-1.15]	7.14×10^{-10}	34216462
13q12.2	13	27919860	rs9581943	A	*PDX1*	1.15[1.12-1.19]	5.12×10^{-14}	29422604
13q22.1	13	73334709	rs9573163	G	*AL162376.1*	1.26[1.18-1.34]	5.14×10^{-13}	22158540
16p12.13	16	20317344	rs78193826	C	*GP2*	1.46[1.29-1.66]	4.28×10^{-9}	32581250
16q23.1	16	75229763	rs7190458	A	*BCAR1*	1.46[1.30_1.65]	1.13×10^{-10}	25086665
17q12	17	37718512	rs4795218	A	*HNF1B*	0.88[0.84-0.92]	1.32×10^{-8}	29422604
17q24.3	17	72405335	rs7214041	T	*LINC00511*	1.25[1.19-1.30]	9.49×10^{-15}	29422604
18q21.32	18	59211042	rs1517037	T	*GRP*	0.86[0.80-0.91]	3.28×10^{-8}	29422604

Chromosomal band	CHR	POS (hg38)	rsID	EA	Closest gene	OR[95%CI]	*p*-value	PMID
19p13.3	19	4892075	rs2656937	T	*ARRDC5*	1.5[1.31-1.57]	3.98×10^{-9}	32671597
19p13.12	19	14464085	rs34309238	C	*PKN1*	1.77[1.48-2.12]	5.35×10^{-10}	30206226
19p12	19	21650212	rs66562280	C	*LOC400682*	1.56[1.36-1.78]	1.00×10^{-10}	32671597
21q21.3	21	29345416	rs372883	C	*BACH1*	0.79[0.75-0.84]	2.24×10^{-13}	22158540
21q22.3	21	42358786	rs1547374	G	*TFF1*	0.79[0.74-0.84]	3.71×10^{-13}	22158540
22q12.1	22	28904318	rs16986825	T	*ZNRF3*	1.18[1.12_1.25]	1.20×10^{-8}	25086665
22q13.32	22	48533757	rs5768709	G	*FAM19A5*	1.25[1.17-1.34]	1.41×10^{-10}	22158540
Loci approaching genome-wide significance in at least one published study								
1p13.2	1	112503773	rs351365	T	*WNT2B*	0.9[0.86-0.95]	6.32×10^{-6}	29422604
6p25.3	6	1339954	rs9502893	G	*AL499606.1*	1.29[1.17-1.43]	3.30×10^{-7}	20686608
7q36.2	7	153928758	rs6464375	A	*DPP6*	3.73[2.24-6.21]	4.41×10^{-7}	20686608
7q36	7	155827039	rs167020	A	*SSH*	1.37[1.22-1.54]	1.76×10^{-7}	19648918
9q31.1	9	104125300	rs2417487	A	*SMC2*	1.09[1.05-1.12]	5.70×10^{-7}	29422604
9q34.2	9	133474633	rs3124761	T	*SLC2A6*	1.17[1.10-1.25]	5.17×10^{-7}	30972876
11p14.1	11	27374328	rs2472632	A	*LGR4*	1.09[1.05-1.14]	5.52×10^{-8}	38308339
12p11.21	12	32283475	rs708224	A	*BICD1*	1.32[1.19-1.47]	3.30×10^{-7}	20686608
15q14	15	36362396	rs8028529	C	*LOC105370767*	1.38[1.22-1.56]	3.55×10^{-7}	19648918
18p11.21	18	13366863	rs12456874	G	*LDLRAD4*	1.38[1.22-1.57]	6.00×10^{-7}	23180869
20q13.12	20	44458008	rs6073450	A	*C20orf62*	1.11[1.06-1.15]	9.21×10^{-7}	26098869

CHR: Chromosome. POS (hg38): Position of the variant on the hg38 genome build. rsID: dbSNP identifier of the variant. EA: Effect allele. OR [95% CI]: Odds ratio with 95% confidence interval. *p*-value: Statistical significance of the association. Closest gene: Nearest gene to the SNP. Chromosomal band: Cytogenetic location of the variant. PMID: PubMed ID of the source study.

2 AIMS OF THE STUDY

This thesis sets out to advance our understanding of the genetic architecture of PDAC the most common form of pancreatic cancer by addressing critical gaps left by previous studies. While GWAS have uncovered numerous common risk loci, these explain only a small proportion of PDAC heritability. Rare variants, gene-gene interactions, and non-coding regulatory elements are emerging as key contributors to disease susceptibility, yet they remain poorly characterized in this context.

To tackle these challenges, this work pursues three interrelated aims:

1. To investigate genetic variants in regulatory regions such as transcription factor binding sites and enhancers that might change epigenetic profiles on the genome, thereby increasing PDAC risk.

2. To explore the contribution of rare germline variants to PDAC susceptibility, leveraging large-scale genotype datasets and stringent quality control frameworks.

3. To research the connection between variants in cognitive-related genes/pathways and PDAC risk directly or indirectly.

By systematically addressing these questions, this thesis aims to move beyond the "low-hanging fruit" of GWAS discoveries, uncovering novel risk loci and interaction networks that may illuminate the biological mechanisms underlying PDAC. Ultimately, these insights could help lay the groundwork for improved risk prediction, functional validation, and precision medicine strategies in pancreatic cancer.

3 MATERIALS AND METHODS

3.1 STUDY POPULATIONS

3.1.1 PanScan-PanC4

The PanScan I study included 1856 PDAC cases and 1890 controls, recruited from 12 prospective cohort studies and 8 case-control studies within the Pancreatic Cancer Cohort Consortium (PanScan)(L. Amundadottir et al., 2009; Petersen et al., 2010; Wolpin et al., 2014; C. Wu et al., 2012) and the Pancreatic Cancer Case-Control Consortium (PanC4)(Childs et al., 2015). Participants were primarily of European ancestry, and principal component analysis (PCA) was applied to adjust for population structure. Cases were histologically confirmed as PDAC, and controls were matched by age, sex, and geographic region. Genotyping was performed using the Illumina HumanHap550 Infinium II chip, with QC measures including genotype completion rate filtering (>98%), Hardy-Weinberg equilibrium in controls, sex mismatches, cryptic relatedness, and population stratification.

The PanScan II study expanded on PanScan I, including 1618 PDAC cases and 1682 controls selected using similar criteria. Genotyping was performed on the Illumina Human 610-Quad chip, with stringent QC ensuring high call rates (>98%) and minimal genotype errors. Imputation was conducted using the 1000 Genomes (Phase I, Dec 2013) and HapMap 3 reference panels to improve variant resolution. PCA was again used to correct for population stratification, and individuals with excess heterozygosity, cryptic relatedness, or non-European ancestry were excluded to maintain a homogeneous sample.

The PanC4 study included 4164 PDAC cases and 3792 controls, pooling data from several institutions including Johns Hopkins Hospital, Mayo Clinic, MD Anderson Cancer Center, and Memorial Sloan-Kettering Cancer Center. Participants were drawn from seven independent case-control studies and were of self-reported European ancestry, confirmed by PCA. Genotyping was conducted using the Illumina HumanOmniExpressExome-8v1 array at the Johns Hopkins Center for Inherited

Disease Research. Extensive QC measures excluded individuals with excessive allele sharing, genotype missingness >2%, or evidence of population outliers. Only autosomal SNPs with call rates >98%, Hardy-Weinberg equilibrium *p*-values $>10^{-6}$, and MAF >0.005 were retained. The final dataset comprised individuals who passed all QC steps and were predominantly of European ancestry, allowing robust genetic association analyses.

The genotyping data from PanScan I, PanScan II, and PanC4 were downloaded from the database of Genotypes and Phenotypes (dbGaP) website (study accession numbers: phs000206.v5.p3 and phs000648.v1.p1, project reference: #12644).

3.1.2 East Asian GWAS

The East Asian GWAS meta-analysis included data from three cohorts: the Japan Pancreatic Cancer Research (JaPAN) consortium, the National Cancer Center (NCC), and BioBank Japan (BBJ). In total, the analysis comprised 2039 pancreatic cancer cases and 32,592 controls of Japanese ancestry. Genotyping was conducted using Illumina HumanCoreExome arrays for the JaPAN study, Illumina HumanHap550 or Human610-Quad arrays for the NCC study, and Illumina HumanOmniExpressExome or OmniExpress arrays for the BBJ study. Each dataset was imputed by using the 1000 Genomes Phase 3 v5 reference panel. Post-imputation quality control excluded variants with MAF < 0.01 or imputation INFO score< 0.5. After filtering, 7,914,378 variants remained for association testing in the meta-analysis(Lin et al., 2020). Summary statistics were obtained via the website http://www.aichi-med-u.ac.jp/JaPAN/current_initiatives-e.html.

3.1.3 PANDoRA

The Pancreatic disease research (PANDoRA) consortium(Campa et al., 2023) brings together 29 research groups across 14 countries, primarily in Europe, to investigate the genetic basis of pancreatic diseases. It includes a diverse set of experts – ranging from biologists and geneticists to clinicians such as oncologists and gastroenterologists – making it uniquely equipped to tackle the complex nature of pancreatic conditions. The consortium focuses largely on PDAC but also includes rarer forms such as pancreatic neuroendocrine neoplasms (PNEN), intraductal papillary mucinous neoplasms (IPMN), and chronic pancreatitis (CP). All cases have a clinically confirmed diagnoses, with detailed clinical data including age, sex, and overall survival, and

biological samples (primarily blood) available for DNA extraction in the Genomic Epidemiology Group at DKFZ.

The PANDoRA has collected data on 19,240 individuals (8,456 controls and 10,784 cases), including PDAC, other exocrine pancreatic cancers, PNEN, IPMN and CP, with the largest cohorts coming from Italy and Germany. The controls are drawn from the general population, blood donors, or non-oncological hospital patients.

3.1.4 UK Biobank

The UK Biobank is a prospective cohort study comprising individuals from across the United Kingdom. Approximately 500,000 participants, aged between 40 and 69 years at enrolment, were recruited between March 2006 and July 2010. During this period, demographic information, medical data, and blood samples were collected. Genetic data from 488,377 individuals was available under application number 66591.

Genotyping was performed centrally by the UK Biobank using either the UK BiLEVE Axiom Array (807,411 markers; N = 49,950) or the UKB Axiom Array (825,927 markers; N = 438,427). After applying genetic variant and sample-level quality control, the final dataset included 488,377 participants with 805,426 SNPs. The UK Biobank team performed phasing using SHAPEIT3 and imputation using IMPUTE4 with the merged UK10K and 1000 Genomes Phase 3 reference panels(Huang et al., 2015). All preprocessing steps, including filtering based on Hardy-Weinberg equilibrium and call rates, were completed centrally by the UK Biobank(Bycroft et al., 2017).

Within this dataset, I identified pancreatic cancer cases from the cancer registry using ICD-10 code C25 (malignant neoplasm of the pancreas), excluding cases classified as endocrine pancreatic tumors (C25.4; Data-Field 40006). I included only individuals with pancreatic cancer as their primary diagnosis. For controls, I selected participants with no diagnosis of any cancer based on cancer registry data, hospital records, or self-report, and restricted the sample to individuals of European ethnic background. This resulted in 1080 PDAC cases and 287,545 controls included in the analysis.

As covariates, I obtained age (Data-Field 21022 for controls, 40008 for cases), genetic sex (Data-Field 31), genotyping array (Data-Field 22000), and the first ten principal components derived from genetic data (Data-Field 22009). I updated the age of controls by subtracting the birth year from the most recent update year provided for each UK Biobank recruitment center.

3.2 Data Quality Control Steps

3.2.1 Per-Individual Quality Control

Missingness: Low DNA concentration can cause poor genotype call rates, leading to false positives by distorting genotype frequencies, or false negatives by masking real signals and reducing power. Missingness is a major concern in QC because it can bias results if it is informative – that is, not randomly distributed across genotypes but enriched in specific ones. Case-control studies are particularly vulnerable due to separate handling of cases and controls, while cohort studies with quantitative traits are generally less affected(Clayton et al., 2005). Samples with >5% missing genotypes should be excluded. Genotype calling tools (e.g., BEAGLE, CHIAMO) can help assess quality by flagging genotypes with low posterior probabilities (<90%)(Howie et al., 2012).

Sex discordance: Identify individuals with mismatched genetic and reported sex by calculating X-chromosome homozygosity. Sex assignment based on the X chromosome can be quantified using the inbreeding coefficient (F). Females typically show F near 0 (Hardy-Weinberg Equilibrium), while males show F near 1 (absence of heterozygosity). Intermediate F values suggest contamination, population differences, or X-chromosome mosaicism, particularly in older cell line DNA. Individuals with intermediate or anomalous F values should be excluded from further analysis(Weale, 2010).

Duplicates and relatives: Cryptic relatedness occurs when individuals are more closely related than expected by chance, often indicating hidden family relationships or sample duplicates. Such relatedness introduces correlation structures into genotype data that can bias association analyses, leading to false positives or negatives, especially if relatedness differs between cases and controls. Therefore, it is standard practice to identify and exclude individuals with excessive relatedness to avoid downstream bias. IBD estimates are used to detect closely related or duplicate individuals: one sample from each pair with IBD > 0.98 (suggesting duplicates) should be removed. A stricter threshold, such as IBD > 0.1875, can be applied to retain only unrelated individuals(M. H. Wang et al., 2019; Weale, 2010).

Ancestry outliers: Individuals with distant ancestry can introduce systematic allele frequency differences unrelated to the disease phenotype, leading to false associations. To minimize this bias, outlier individuals must be identified and removed. PCA is commonly used for this purpose. PCA normalizes the genotype matrix and transforms

it into principal components (PCs), where PC1 captures the largest variance, followed by PC2, and so on. Plotting individuals along the first two PCs allows visual detection of clustering patterns: a homogeneous cohort should form a single major cluster. Outliers are typically defined as individuals located more than six standard deviations (SDs) away from the population mean along one or more top PCs(Patterson et al., 2006; Price et al., 2006). Identification and removal should be performed iteratively, excluding one outlier at a time until no further outliers remain.

3.2.2 Per-Variant Quality Control

SNPs with high missingness: A high proportion of missing genotypes at a SNP often indicates technical difficulties in genotyping and an increased risk of genotyping errors. To ensure data quality, SNPs with a completion rate smaller than 95% should be excluded from analyses when using SNP arrays. For targeted genotyping approaches, a more lenient threshold of 75% is typically applied.

Differential missingness in cases vs. controls: When cases and controls are sampled from different sources or genotyped separately, systematic differences in sample quality can arise, leading to biased missingness patterns. SNPs showing large differences in missing data rates between cases and controls should be identified and removed to minimize the risk of spurious associations.

Hardy-Weinberg Equilibrium (HWE) violations: in a large and homogeneous population with random mating, genotype frequencies are expected to follow Hardy-Weinberg proportions. Deviations from HWE can be detected using a chi-square goodness-of-fit test. In GWAS, significant deviations from HWE often indicate genotyping errors and are used to flag problematic SNPs. To avoid removing SNPs that deviate due to true disease associations, HWE tests are typically performed only in controls. A common threshold for exclusion is a p-value between 10^{-5} and 10^{-7}, depending on the study(Anderson et al., 2010; Wittke-Thompson et al., 2005).

3.2.3 GWAS Analysis

Beyond population structure, clinical and environmental confounders, such as age and sex, use must be controlled in GWAS. Generalized linear models (GLMs) provide a flexible framework for this adjustment.

The GLM is specified as:

$$g(\mu) = \beta_0 + \Sigma \beta j \, Xj + uG + \Sigma \gamma k \, PCk$$

where μ = E(Y) is the expected value of the phenotype, g() is the link function, Xj are covariates (e.g., age, sex), G is the genotype at the test SNP, and PCk are principal components accounting for population structure. For binary traits (e.g., case/control), the logit link function yields logistic regression. For continuous traits, the identity link reduces the model to classical linear regression.

In addition to adjusting for known confounders, hidden population structure must also be addressed to avoid spurious associations.

To correct for population stratification, PCA is widely used. The EIGENSTRAT method adjusts genotype and phenotype data by regressing out ancestry effects estimated from principal components(Price et al., 2006). Alternatively, top PCs (usually 10-20) can be included as covariates in generalized linear models.

Even in carefully designed studies, residual population stratification can inflate variance and generate spurious associations. It can be detected post-analysis by examining Quantile-Quantile (QQ) plots, where deviations from the diagonal line indicate inflation. The genomic control method quantifies this inflation using the genomic inflation factor (λ), calculated as the median of observed chi-squared statistics divided by the expected median (0.456 for a χ^2_1 distribution). Values of λ close to 1 are expected; thresholds such as $\lambda < 1.1$ are commonly applied(Armitage, 1955).

In a single statistical test, the type I error rate (α) is the probability of incorrectly rejecting the null hypothesis when it is true. Typically, α is set at 0.05, allowing a 5% chance of a false positive per test. However, given the large number of tests performed in genetic studies, especially in genome-wide analyses, adjusting significance thresholds to account for multiple comparisons is essential.

The simplest and most widely used adjustment is the Bonferroni correction, which controls the family-wise error rate by dividing the desired α level by the number of tests conducted(Lunetta, 2008). This method assumes independence between tests, making it overly conservative in the presence of correlations among SNPs due to linkage disequilibrium. In GWAS, where hundreds of thousands to millions of variants are tested, a commonly accepted genome-wide significance threshold is *p*-value $< 5 \times 10^{-8}$.

3.2.4 Post-GWAS Analysis

Replication: Replicating a genotype-phenotype association is the "gold standard" for confirming that an observed signal is genuine. Most loci for complex diseases have modest effects, making it unlikely that a single study alone can establish association without replication. Replication studies should be sufficiently powered, conducted in independent datasets, involve the same phenotype and population, test the same SNP, observe the same effect direction, and produce a stronger combined *p*-value.

Replication differs from validation, where findings are tested under altered conditions such as different ethnic backgrounds or phenotypes; thus, validation results may vary due to random or systematic differences(Igl et al., 2008). SNPs with moderate to strong statistical significance, common minor allele frequencies, and larger effect sizes are more likely to replicate, partly because common variants provide more stable frequency estimates across studies.

Genotype Imputation: Genotype imputation infers genotypes at untyped variants, including SNPs, short indels, or missing genotypes, by leveraging the LD structure from phased haplotype reference panels, typically derived from whole-genome sequencing. Standard tools such as Impute5(Rubinacci et al., 2020), Minimac4(Das et al., 2016), and Beagle5.4(Browning et al., 2018) apply the Li and Stephens haplotype model(N. Li & Stephens, 2003), which represents an individual's genome as a mosaic of recombination and minor mutations from other haplotypes.

These tools use Hidden Markov Models, where observed genotypes of unknown phase represent the visible states, and unobserved phased haplotypes represent the hidden states. Benchmarking studies report comparable imputation accuracy, achieving over 97% sensitivity for common variants (MAF > 5%) and over 99% for rare variants (MAF < 5%)(De Marino et al., 2022). Beagle5.4 excels at imputing common variants, while Impute5 and Minimac4 perform better for low-frequency and rare variants. Computationally, Beagle5.4 demonstrates the shortest runtime, while Minimac4 has the lowest memory requirements. Both Minimac4 and Impute5 can alleviate computational burden through pre-phasing, where target sample genotypes are phased before imputation.

Reference panels have also evolved alongside methodological advances. HapMap(International HapMap 3 Consortium et al., 2010) and the 1000 Genomes Project (1KGP)(1000 Genomes Project Consortium et al., 2015) initially provided widely used reference datasets, with the 1KGP Phase III panel offering over 80 million variants from 2504 individuals across 26 populations. Due to differences in haplotype structure across ancestries, using a reference panel matched to the study population is

crucial for maximizing imputation accuracy. The Haplotype Reference Consortium (HRC)(McCarthy et al., 2016), comprising over 30,000 samples and 40 million variants, has largely replaced earlier panels, offering improved performance especially for samples with undetected admixture or rare haplotypes.

In addition, the TOPMed Imputation Reference Panel provides a highly diverse resource based on whole-genome sequencing data from 133,597 individuals from the Trans-Omics for Precision Medicine (TOPMed) program. Available through the Michigan Imputation Server, the current release (version r3) includes over 445 million variants across 22 autosomes and the X chromosome, enhancing imputation quality for studies involving diverse ancestries(Taliun et al., 2019).

Genotype imputation servers now enable researchers to perform imputation remotely without needing local reference panels or advanced computational infrastructure. The Michigan Imputation Server, powered by Minimac3(Das et al., 2016), and the Sanger Imputation Service(McCarthy et al., 2016), using positional Burrows-Wheeler transform (PBWT)(Durbin, 2014), both provide secure cloud-based imputation with access to multiple reference panels, including 1KGP, HRC and TOPMed.

Meta-analysis: Meta-analysis addresses a key challenge in genetic association studies by combining results across multiple studies and assessing inter-study heterogeneity(Y. H. Lee, 2015). Early GWAS showed that common variants associated with complex diseases usually have modest effects, often with odds ratios below 1.2 for binary traits or explained variances below 1% for quantitative traits. Increasing discovery success for loci with even smaller effects requires larger sample sizes, which has been enabled by the rise of disease-specific consortia and large shared datasets. Meta-analyses combine results from multiple GWAS, typically focusing on single-marker, single-phenotype analyses, and have become critical for novel gene discovery and functional variant identification. In effect, meta-analysis primarily relies on two strategies: combining effect sizes using fixed-effects or random-effects models, or combining *p*-values (or weighted Z-scores)(Laird & Mosteller, 1990).

3.2.5 Additional Association Tests

Epistasis (Interaction): Epistasis refers to the interaction between two genes or SNPs that together influence the expression of a specific phenotype, with this interaction going beyond a simple additive effect of the individual genetic contributions(Cordell, 2002). These interactions are commonly illustrated using penetrance tables, typically in a 3x3 format representing all possible allelic combinations at two loci and their corresponding phenotypic effects. Theoretically,

many different interaction patterns can emerge. For a binary penetrance model, classifying genotypes as high-risk or low-risk, across two loci, there are 512 (2^9) potential configurations. Even after accounting for symmetry and removing redundant models, 50 unique patterns remain(Evans et al., 2006). The distribution of individuals across genotypes is roughly constrained by HWE, meaning genotype frequencies are determined by the minor allele frequency at each locus(Edwards, 2008).

Importantly, loci involved in an interaction may show no detectable effect when considered individually. In so-called 'pure' epistatic models, neither locus exhibits a main effect on its own, and both must be analyzed together to reveal the interaction(Urbanowicz et al., 2012). Various detection strategies have been proposed—some applying exhaustive pairwise testing, others incorporating filtering steps to reduce the number of loci tested and, consequently, the computational burden. Broadly, the methods fall into statistical approaches, such as GLMs and contingency tables, and into data mining and machine learning (ML) strategies. Critical aspects such as computational runtime, the ability to adjust for covariates, and the capacity to detect higher-order interactions are also key considerations(Russ et al., 2022).

Initial epistasis detection efforts relied on straightforward statistical methods, including χ^2 tests(Carrasquillo et al., 2002) and GLMs to evaluate interaction effects(Macgregor & Khan, 2006; Millstein et al., 2005). These foundational approaches remain in use and are implemented in several tools. For instance, CASSI and PLINK both offer a Z-score-based test known as fast-epistasis, though they differ slightly in their logistic regression frameworks; PLINK also includes the Wan Log-Linear method and BOOST.

Gene-Based Tests: Aggregation tests assess the cumulative effect of multiple genetic variants within a gene or region, increasing power when several variants contribute to a disease or trait(S. Lee et al., 2014a). Many methods exist, but regression-based approaches are commonly used due to their ability to adjust for covariates. These methods can be grouped into four main classes: burden tests, adaptive burden tests, variance-component tests, and omnibus tests.

Burden tests collapse multiple variants into a single genetic score and test its association with the trait. A simple version counts the number of minor alleles across variants(B. Li & Leal, 2008; Madsen & Browning, 2009; Morris & Zeggini, 2010). These tests assume all rare variants are causal and affect the trait in the same direction (after weighting), an assumption that, if violated, can significantly reduce power(S. Lee, Wu, et al., 2012; Neale et al., 2011).

Adaptive burden tests address this by allowing for both protective and risk alleles and including non-causal variants. For example, the aSum test estimates the direction of

effect for each variant before performing the burden test. These tests are more flexible but often require unstable estimates for rare variants and computationally intensive permutations to determine significance(Basu & Pan, 2011; Han & Pan, 2010).

Variance-component tests (e.g., C-alpha, SKAT) evaluate the distribution of genetic effects rather than collapsing them, typically within a random-effects framework. The sequence kernel association test (SKAT), for instance, models the joint effect of variants and accommodates interactions. However, for binary traits with small sample sizes or low MACs, large-sample approximations can misestimate type I error rates. To correct this, moment-based methods or permutation approaches are used(Neale et al., 2011; Pan, 2009; M. C. Wu et al., 2010, 2011).

Omnibus tests combine burden and variance-component tests to balance their respective strengths. For instance, SKAT-O and similar methods adjust for data-driven selection of the optimal test(Derkach et al., 2013; S. Lee, Wu, et al., 2012; J. Sun et al., 2013). While combined tests may be less powerful than one method alone if its assumptions are fully met, they offer robust performance when little is known about the underlying genetic architecture.

3.3 SOFTWARE AND TOOLS

The software and bioinformatic tools used in this thesis are:

PLINK(Purcell et al., 2007) is an open-source C/C++ software package specifically developed for whole-genome association studies. PLINK is optimized to efficiently handle datasets involving hundreds of thousands of markers across thousands of individuals, providing robust tools for data management, quality control, and association testing. In addition to core functionality, such as summary statistics and logistic regression, PLINK supports advanced analyses including population stratification correction and estimation of IBD. I performed all PLINK analyses using versions 1.9 and 2.0.

MAGMA (Multi-marker Analysis of GenoMic Annotation)(Leeuw et al., 2015a) tool tests the joint association of multiple SNPs within predefined genomic regions while accounting for LD among markers. It addresses common limitations of existing methods, such as reduced power due to linkage disequilibrium, difficulty detecting multi-marker associations, and reliance on computationally intensive permutation testing. It implements a multiple regression framework to model the joint effects of SNPs within genes, improving statistical power and maintaining a correct type I error rate. The gene-set analysis builds on the gene-level results, using a flexible regression-

based approach that allows testing of predefined gene sets as well as continuous gene properties. MAGMA is designed for efficiency and flexibility, making it suitable for large-scale polygenic trait analysis and for exploring complex biological mechanisms underlying genetic associations. I carried out all MAGMA analyses using version 1.10.

R is a free and open-source statistical computing environment, for data processing, visualization, and statistical analysis. R provides a flexible programming framework widely used in bioinformatics and genomics due to its extensive library ecosystem, reproducibility, and ability to handle large and complex datasets. Its comprehensive statistical capabilities and integration with packages such as "data.table", "ggplot2", and "dplyr" made it suitable for performing downstream analysis and generating publication-quality figures. R's scripting environment also facilitated automation and reproducibility across multiple stages of the analysis pipeline. I used R versions 4.2.0 and 4.3.0 as required by package dependencies(https://www.r-project.org/).

I used an LSF-based compute cluster (provided by the German Cancer Research Center, DKFZ) to run command line tools, to process data and submit the job scripts.

bgenix(Band & Marchini, 2018) is a command-line tool optimized for rapid indexing and extraction of genotype data stored in BGEN file format, which is commonly used for large-scale imputed datasets. It enables efficient access to variant or sample-specific data without requiring full file decompression, making it particularly suitable for handling high-density genotype datasets such as those from the UK Biobank. I used the bgenix version 1.1.7 provided by DKFZ cluster.

BCFtools(Danecek et al., 2021) is a versatile software suite for processing files in variant call format (VCF) and binary variant call format (BCF). It supports a wide range of operations including filtering, merging, querying, and indexing, and is commonly used for quality control and data manipulation in genomic workflows. Its efficient C-based implementation and support for standard genomic data formats make it a foundational tool in large-scale variant analysis. I used the BCFtools version 1.19 provided by DKFZ cluster.

REGENIE(Mbatchou et al., 2021) is a two-step whole-genome regression method designed for efficient analysis of large-scale genetic data in both binary and quantitative trait association studies. The first step fits a whole-genome ridge regression model to account for population structure and relatedness, while the second step performs single-variant association testing using logistic or linear regression. REGENIE is optimized for speed and memory efficiency, allowing it to scale to hundreds of thousands of samples and millions of variants. I used the REGENIE version 4.0.

METAL(Willer et al., 2010) is a widely adopted open-source software developed for conducting meta-analyses of GWAS using summary-level data. It provides an efficient

and scalable framework to combine association results across multiple studies, using either inverse-variance weighting or sample size-based methods. METAL is particularly effective for synthesizing findings from different cohorts while accounting for effect sizes, standard errors, and direction of association. I use the METAL version 2011-03-25.

CASSI (Combined Association Score for SNP Interactions)(Ueki & Cordell, 2012) is a command-line tool developed for detecting epistatic interactions by computing a joint association score for SNP pairs. It implements a fast and scalable statistical framework to detect gene-gene interactions without requiring pre-filtering or interaction models that assume additive effects. CASSI is designed to work efficiently with large-scale case-control studies and has been applied to genome-wide interaction scans to prioritize SNP-SNP pairs for downstream analysis. I used CASSI version 2.51.

3.4 FUNCTIONAL ANNOTATION APPROACHES

I conducted functional annotation to interpret the biological relevance of identified variants and to prioritize candidate genes by using various data resources (**Table 2**). I annotated variants reaching genome-wide or suggestive significance using the dbSNP (Single Nucleotide Polymorphism Database) and VEP (Variant Effect Predictor) to determine their predicted functional consequences, such as missense, synonymous, or regulatory effects.

To assess potential regulatory roles, I used eQTL and splicing quantitative trait loci (sQTL) data from The Genotype-Tissue Expression (GTEx) portal to identify variants with possible gene expression effects across relevant tissues, particularly pancreatic and gastrointestinal tissues. I performed pathway and gene-set enrichment analyses using Enrichr to identify biological pathways and functional networks associated with PDAC risk loci. I constructed protein-protein interaction networks using STRING to assess the connectivity of prioritized genes within relevant signaling pathways. These analyses provided a comprehensive functional interpretation of the genetic findings and helped identify potential mechanistic links between genetic variation and PDAC pathogenesis.

I used functional annotation tools such as cMAC (cumulative Minor Allele Count) and CADD (Combined Annotation Dependent Depletion) to assess variant deleteriousness. I applied predictive tools like LINSIGHT and FATHMM-XF to evaluate the impact of noncoding and coding variants, while I used PolyPhen, SIFT, and pLoF (putative Loss-of-function variants) to assess structural and functional variant effects. I mapped regulatory regions using DHS (DNase Hypersensitivity Sites),

CAGE (Cap Analysis of Gene Expression), and GWER (Genome-Wide Error Rate) to improve the identification of rare variant contributions to complex traits and diseases.

I also used additional databases – HaploReg, RegulomedB, Expression Atlas, The Human Protein Atlas, and TNMplot – to link top-associated variants to potential functional mechanisms, such as transcription factor binding changes, chromatin state alterations, and differential expression. Finally, I used the Ensembl website to analyze the regions near significant SNPs and identify surrounding regulatory elements.

Table 2. Data resources used in this study.

Data resource	Description/Usage	Link
Genes - NCBI	Gene information database.	https://www.ncbi.nlm.nih.gov/gene
Single Nucleotide Polymorphism Database (dbSNP)	Database for SNPs and other genetic variations.	https://www.ncbi.nlm.nih.gov/snp
Ensembl Variant Effect Predictor (VEP)	Tool for annotating and analyzing genetic variants.	https://www.ensembl.org/info/docs/tools/vep/index.html
Ensembl BioMart	Data mining tool for accessing Ensembl datasets.	https://www.ensembl.org/biomart/martview/b2b5e62aedb12c7a899bc857b25a9bb9
Functional Annotation of Variants Online Resource (FAVOR)	Tool for functional annotation of variants.	https://favor.genohub.org/
Open Targets	Platform integrating data to identify therapeutic targets.	https://www.opentargets.org/
RefSeqGene	Reference sequences for well-characterized genes.	https://www.ncbi.nlm.nih.gov/refseq/rsg
UniProt	Comprehensive protein sequence and functional information database.	https://www.uniprot.org/
Protein Variation (ProtVar v1.2)	Tool contextualizing human missense variations.	https://www.ebi.ac.uk/Tools/dbfetch/dbfetch?db=ProtVar
The Genotype-Tissue Expression (GTEx) Portal	Resource for studying tissue-specific gene expression and regulation.	https://gtexportal.org/home/
Enricher	Gene set enrichment analysis tool.	https://maayanlab.cloud/Enrichr/
STRING	Protein-Protein Interaction Networks Functional Enrichment Analysis	https://string-db.org/
GeneCards: The Human Gene Database	Comprehensive gene-centric database.	https://www.genecards.org/
Gene Expression - CZ CELLxGENE Discover	Resource for single-cell and bulk gene expression datasets.	https://cellxgene.cziscience.com/

This table provides open-access addresses of the datasets and resources utilized in the analysis.

4 PROJECT 1: TFBS AND ENHANCER POLYMORPHISMS IN PDAC

4.1 INTRODUCTION

Much research has focused on genetic variants in protein-coding regions, as their potential impact on proteins is relatively easy to predict. However, the majority of risk variants are found in non-coding regions. Unlike coding variants, non-coding variants do not directly alter amino acid sequences or affect protein functions such as DNA binding, catalytic activity, or ligand-receptor interactions. Instead, their potential impact lies in differential gene expression.

To identify novel risk loci, secondary analyses have been conducted on existing GWAS data, and additional cases and controls have been genotyped from independent populations which is an approach that has also been successfully applied to PDAC. For example, previous studies have explored the association of SNPs in long non-coding RNAs (lncRNAs)(Corradi et al., 2021), microRNAs(Lu et al., 2021), eQTLs(Pistoni et al., 2021), and specific pathway-related genes(Gentiluomo, Lu, et al., 2019; Gentiluomo, Puchalt García, et al., 2019; Walsh et al., 2019) in PDAC, leading to the discovery of several novel germline risk loci.

In this project, I focused on genetic germline variants located in two major types of regulatory regions: enhancers and transcription factor (TF) binding sites (TFBSs) (**Figure 4**). Enhancers are cis-acting DNA sequences that can enhance gene transcription, playing a crucial role in regulating tissue-specific gene expression(Pennacchio et al., 2013). They typically function independently of orientation and can act at varying distances from their target promoters. TFBSs represent another key class of non-coding regulatory regions, often clustering in short sequences of 5-30 nucleotides within promoters(Wray et al., 2003).

Variants within these regulatory regions can significantly influence gene regulation by altering the binding affinity of TFs. Since TFs recognize and bind specific DNA sequences to regulate target gene expression, polymorphic variants in TFBSs and enhancers have the potential to disrupt TF binding and, consequently, modify gene

expression(Johnston et al., 2019). In this research, I aimed to determine whether variants in these regulatory regions serve as germline cancer susceptibility variants in PDAC by leveraging GWAS data.

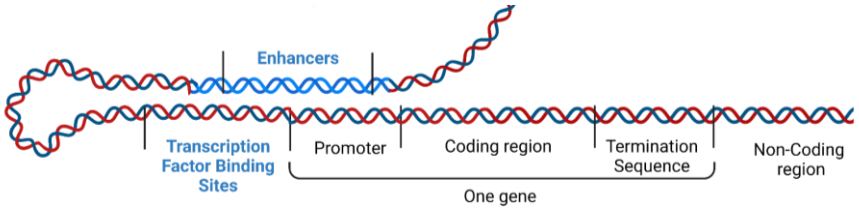


Figure 4. Visual explanation of the targeted regulatory regions.

Figure reproduced from Ünal et al., 2024, licensed under CC BY 4.0.

4.2 MATERIALS AND METHODS

4.2.1 Study populations

For the discovery phase, Dr. Ye Lu (previous doctoral researcher in the Genomic Epidemiology Group at DKFZ) utilized genotyping data from PanScan I, PanScan II, and PanC4. After merging these three datasets (hereafter referred to as PanScan-PanC4), she performed genotype imputation using the TOPMed imputation panel (version TOPMed-r2), followed by rigorous quality control measures. She excluded individuals with cryptic relatedness (PI_HAT > 0.2), gender mismatches, or those failing quality thresholds. She filtered variants based on a MAF < 0.01, completion rate and call rate < 98%, low imputation quality (INFO score < 0.7), and HWE violations (p-value < 1×10^{-5}). To assess population structure, she conducted PCA using PLINK version 2.0, merging study genotypes with those from the 1000 Genomes Project (phase 3). She excluded individuals (N = 439) who did not cluster with European ancestry samples from the 1000 Genomes dataset from further analysis. After applying these filters, she retained 7,509,345 variants across 14,266 individuals (7205 PDAC cases and 7061 controls).

To refine the list of candidate variants, I incorporated summary statistics from an East Asian GWAS, which included 7,914,378 variants.

For the replication phase, I selected an initial set of 11,370 individuals from the PANDoRA consortium to replicate the previously selected variants. I selected PDAC cases and healthy controls from 11 European countries (Greece, Italy, Germany, the Netherlands, Denmark, the Czech Republic, Hungary, Poland, Ukraine, Lithuania, and the UK). Controls were patients from the general population without any pancreatic disease at recruitment, individuals hospitalized for reasons other than cancer, or blood donors, and were recruited in the same geographical regions as the cases. For this project, I included only PANDoRA subjects with self-declared European ancestry.

I obtained genotype and covariate data of additional controls from the Netherlands and Germany respectively from the 'European Prospective Investigation into Cancer and Nutrition' (EPIC)(Riboli et al., 2002) and the 'Epidemiological investigations on chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population' (ESTHER)(Löw et al., 2004).

4.2.2 SNP selection

To improve the likelihood of identifying associations with PDAC risk, I first filtered SNPs from the PanScan-PanC4 dataset (prepared by Dr. Ye Lu) using a significance threshold of p-value $< 10^{-4}$. This threshold was arbitrary, but chosen in order to pre-select a relatively large number of SNPs that had good chance of reaching genome-wide significance upon genotyping of additional cases and controls. To avoid missing potential SNP-TF interactions, I additionally included SNPs in LD ($r^2 > 0.6$) with the significant SNPs, based on LD estimates from European populations (excluding Finnish) using LDlink (https://ldlink.nci.nih.gov). Expanding the SNP list with strong LD pairs resulted in a total of 7815 SNPs.

For TFBS SNPs, I retrieved annotations from the SNP2TFBS database(Kumar et al., 2017), which includes binding affinity predictions for 200 TFs. The database reports a binding score difference between alleles for each SNP-TF interaction. The binding score is the estimation of the effects of SNPs on TF binding across the genome using position weight matrices (PWM), which model the sequence specificity of TF binding. I extracted all predicted SNP-TF interactions from the SNP2TFBS database (accessed on 15 December 2020), yielding a total of 2,281,137 variant-TF interactions involving 1,900,881 unique SNPs.

I intersected the set of 7815 SNPs with 1,900,881 unique TFBS SNPs from the SNP2TFBS database, resulting in 905 unique overlapping SNPs. To ensure a focused and high-confidence subset for downstream analysis, I applied stringent filtering

criteria: retaining only one SNP per TF and limiting the selection to SNPs within at least in the 80th percentile of the TFBS score distribution.

For enhancer SNPs, I utilized a published dataset(Nasser et al., 2021) that employed the activity-by-contact (ABC) model, which predicts enhancer-target gene interactions across 131 human cell types and tissues. I extracted enhancer regions and their associated target genes specifically from normal pancreatic tissue and the PANC-1 pancreatic cancer cell line, resulting in 55,967 enhancer regions. Next, I mapped SNPs to these enhancer regions using the UCSC Genome Browser(Kent et al., 2002) tools, identifying 1,190,420 SNPs located in enhancers.

To evaluate the association between TFBS and enhancer SNPs with PDAC risk, I first assessed their significance in the PanScan-PanC4 discovery dataset. I then compared the effect alleles of the SNPs between the PanScan-PanC4 data and the East Asian GWAS summary statistics. If the effect alleles did not match, I flipped the BETA values in the East Asian dataset by multiplying them by -1 to harmonize the direction of effect. Next, I performed a meta-analysis between PanScan-PanC4 and East Asian GWAS summary statistics using METAL software. For each SNP, I calculated the OR and 95% CI using the formula: $OR = exp(BETA)$ and $95\%\, CI = exp(BETA \pm 1.96 \times SE)$. I excluded SNPs in known pancreatic cancer risk loci and SNPs in LD with them ($r^2 > 0.6$).

For the replication phase, I selected SNPs based on the following criteria:

- p-value < 10^{-5} in the meta-analysis,
- Significant association (p-value < 0.05) in both PanScan-PanC4 and East Asian GWAS summary statistics.

Initially, I selected the following six SNPs: rs10025845, rs11241697, rs2472632, rs11032793, rs11624002 and rs2232079. Additionally, I included SNP rs17358295 despite its moderate significance in the meta-analysis (p-value = 2.7×10^{-2}) because it showed the strongest association in the East Asian GWAS summary statistics (p-value = 4.6×10^{-3}). Finally, I selected the top significant SNPs in TFBSs and enhancers for genotyping analysis in the PANDoRA population.

4.2.3 Sample preparation and genotyping

I conducted the entire sample preparation and genotyping process DKFZ. DNA extractions were carried out from the whole blood of participants in the PANDoRA consortium using a Qiagen DNA extraction kit (Hilden, Germany). I distributed approximately 5ng DNA samples in 384-well plates. To minimize batch effects, I randomized the order of DNA samples across plates to ensure an equal representation of case and control subjects in each batch.

I performed genotyping via allele-specific PCR-based TaqMan technology (ThermoFisher, Applied Biosystems, Waltham, MA), utilizing pre-designed TaqMan SNP Genotyping Assays for the seven selected SNPs (**Table 3**). PCR reactions were prepared in 384-well plates using the following per-well volumes: 1.9 μL TaqMan Genotyping Master Mix, 1.9 μL nuclease-free water, and 0.2 μL SNP assay mix. I carried out thermal cycling on the ViiA™ 7 Real-Time PCR System (Applied Biosystems) with the following conditions: 95 °C for 15 minutes, followed by 40 cycles of 95 °C for 20 seconds and 60 °C for 1 minute. I analyzed genotypes using the ViiA™ 7 RUO Software, version 1.2.2.

Table 3. Details of Selected SNPs for TaqMan Genotyping Assays.

rsID	GENE	EA	NEA	EUR_freq	Assay_ID	VIC	FAM
rs10025845	*LINC01258*	A	G	0.41	C__30294143_10	A	G
rs11241697	*CEP120*	C	T	0.41	C__459532_20	C	T
rs2472632	*LGR4*	A	C	0.36	C__1890263_10	A	C
rs11032793	*EHF*	C	T	0.1	C__32033322_10	C	T
rs17358295	*EHF*	G	T	0.11	C__34396575_10	G	T
rs11624002	*EVL*, *DEGS2*	T	C	0.16	C__1329809_20	C	T
rs2232079	*FERMT1*	T	C	0.11	C__15857320_10	C	T

rsID: Reference SNP ID. GENE: Associated gene. EA: Effect allele. NEA: Non-effect allele. EUR_freq: Effect allele frequency in European populations (1000 Genomes Project). Assay_ID: TaqMan SNP Genotyping Assay identification code. VIC/FAM: Fluorescent dye labels used for allele discrimination.

4.2.4 Statistical analysis

I extracted the genotypes of the seven selected SNPs from the EPIC and ESTHER cohorts, comprising 1162 control samples from the Netherlands and Germany. After genotyping 10,208 PANDoRA samples for the same SNPs, I combined these with the control samples. I then applied QC steps to exclude samples with poor genotyping quality. The QC steps included exclusion of samples missing age, sex, or diagnosis information; non-PDAC cases; samples from countries that contributed only cases but no controls; duplicate samples; and those with a genotyping completion rate below 80% across all seven SNPs. I also tested for deviations from HWE in the control group for each variant.

With post-QC PANDoRA samples, I performed a logistic regression test in R, adjusting for age, sex, and country of origin. I used the country of origin as a surrogate for ethnic groupings. Since GWAS data were not available for PANDoRA, I could not apply principal component adjustments. I used the following R script to perform logistic regression with "stats" R package:

```
""

library(stats)

result <- glm(phenotype ~ SNP + age + sex + country,

                    data = data, family = binomial())

summary(result)

""
```

Finally, I conducted a meta-analysis across all three populations, incorporating data from a total of 56,079 individuals using the "metafor" R package(Viechtbauer, 2010).

```
""

library(metafor)

meta_results <- metagen(TE= data$BETA, seTE= data$SE, studlab = data$Study,

                    method.tau = "SJ", sm="OR",

                    byvar = data$SNP, data=data)

""
```

I based the choice of fixed-effect or random-effect models on heterogeneity levels (fixed-effect: $I^2 < 50\%$, random-effect: $I^2 \geq 50\%$). To correct for multiple testing, I applied LD pruning, with a threshold of $r^2 > 0.6$ to exclude SNPs representing the same association. I applied a Bonferroni-corrected significance threshold of 1.94×10^{-5} (0.05/2575 independent tests).

4.3 RESULTS

First Meta-Analysis Identifies Seven Regulatory SNPs Associated with PDAC Risk Across Populations

I first intersected the TFBS SNP list from SNP2TFBS database with 7815 SNPs (*p*-value $< 10^{-4}$, LD r2>0.6) from the PanScan-PanC4 dataset. Separately, I intersected SNPs located in enhancer regions with 2575 SNPs associated with PDAC risk (*p*-value $< 10^{-4}$) from the same dataset. These intersections resulted in 778 SNPs in TFBS regions and 84 SNPs in enhancer regions. Next, I conducted a meta-analysis on 673 TFBS SNPs and 12 enhancer SNPs, which were present in both the PanScan-PanC4 and East Asian GWAS summary statistics. After excluding known PDAC loci, this analysis identified 46 SNPs in TFBS regions (**Table 5**) and 12 in enhancers (**Table 6**) that showed an association with PDAC risk (*p*-value < 0.05).

Table 5. Meta-analysis of the TFBS SNPs in both PanScan-PanC4 and East Asian GWAS datasets.

CHR	POS	rsID	PanScan-PanC4				East Asian GWAS				metafor/R results				
			EA	OR	BETA	*p*-value	EA	OR	BETA	*p*-value	*p*val_meta	BETA	SE	OR	95%CI
1	64542612	rs1260767	C	0.9379	-0.0641	5.66×10^{-3}	C	1.042	0.0414	6.22×10^{-1}	1.12×10^{-2}	-0.0567	0.0223	0.9449	[0.9044; 0.9872]
1	99990826	rs12725145	A	0.948	-0.0534	2.22×10^{-2}	A	0.967	-0.034	3.99×10^{-1}	1.63×10^{-2}	-0.0485	0.0202	0.9526	[0.9156; 0.9911]
1	119022906	rs1229120	A	1.097	0.0926	1.90×10^{-3}	A	0.983	-0.0175	8.56×10^{-1}	3.41×10^{-3}	0.0831	0.0284	1.0867	[1.0279; 1.1489]
1	189882958	rs79447077	C	0.6267	-0.4673	1.08×10^{-3}	C	0.846	-0.167	1.44×10^{-1}	3.65×10^{-2}	-0.3041	0.1454	0.7378	[0.5548; 0.9810]
1	230500424	rs74230136	C	0.8529	-0.1591	1.93×10^{-3}	C	1.003	0.0032	9.85×10^{-1}	3.07×10^{-3}	-0.1454	0.0491	0.8647	[0.7853; 0.9520]
1	230844273	rs2493133	C	0.9197	-0.0837	3.07×10^{-4}	C	0.981	-0.0194	6.87×10^{-1}	6.11×10^{-4}	-0.0716	0.0209	0.9309	[0.8936; 0.9698]
2	80810170	rs216622	G	1.073	0.0705	1.99×10^{-3}	G	1.006	0.006	8.85×10^{-1}	5.63×10^{-3}	0.0554	0.02	1.0569	[1.0163; 1.0992]
2	191708138	rs11900452	C	0.9071	-0.0975	1.16×10^{-4}	C	0.971	-0.0294	5.58×10^{-1}	2.10×10^{-4}	-0.0837	0.0226	0.9197	[0.8799; 0.9613]
3	54016257	rs6445621	A	0.9343	-0.068	4.01×10^{-3}	A	0.92	-0.083	6.08×10^{-1}	3.46×10^{-3}	-0.0683	0.0234	0.934	[0.8922; 0.9778]
3	79780667	rs6798360	G	0.9535	-0.0476	4.87×10^{-2}	G	0.97	-0.0301	4.28×10^{-1}	3.68×10^{-2}	-0.0426	0.0204	0.9583	[0.9208; 0.9974]
3	178087477	rs74567109	C	1.199	0.1815	4.07×10^{-4}	C	1.114	0.1084	4.00×10^{-1}	3.22×10^{-4}	0.1715	0.0477	1.1871	[1.0812; 1.3034]
4	1575055	rs28373580	G	0.9137	-0.0903	2.08×10^{-2}	G	0.941	-0.0613	6.14×10^{-1}	1.85×10^{-2}	-0.0876	0.0372	0.9162	[0.8518; 0.9854]
4	18013639	rs34497096	C	0.8594	-0.1515	4.71×10^{-5}	C	0.933	-0.0689	3.73×10^{-1}	5.07×10^{-5}	-0.1359	0.0335	0.8729	[0.8174; 0.9322]
4	38431377	rs10025845	A	1.086	0.0825	3.76×10^{-4}	A	1.093	0.0893	3.43×10^{-2}	3.77×10^{-5}	0.0841	0.0204	1.0877	[1.0451; 1.1321]
4	182163172	rs74450484	G	1.201	0.1832	1.65×10^{-4}	G	1.065	0.0634	3.71×10^{-1}	3.11×10^{-4}	0.1447	0.0401	1.1557	[1.0683; 1.2502]
5	1378416	rs11954950	T	1.091	0.0871	1.70×10^{-3}	T	0.967	-0.0335	7.25×10^{-1}	3.38×10^{-3}	0.0778	0.0265	1.0809	[1.0261; 1.1385]
5	93697645	rs29963	A	0.9178	-0.0858	2.77×10^{-4}	A	0.959	-0.0421	3.17×10^{-1}	2.53×10^{-4}	-0.0753	0.0206	0.9274	[0.8908; 0.9656]
5	122763204	rs11241697	C	0.9174	-0.0862	2.46×10^{-4}	C	0.896	-0.1101	4.68×10^{-3}	4.18×10^{-6}	-0.0926	0.0201	0.9116	[0.8763; 0.9482]
6	7370058	rs9406025	C	0.882	-0.1256	8.34×10^{-4}	C	0.956	-0.0448	2.62×10^{-1}	3.08×10^{-2}	-0.0864	0.04	0.9173	[0.8481; 0.9921]
6	33804400	rs34945102	A	0.9264	-0.0764	6.96×10^{-3}	A	0.981	-0.0194	8.34×10^{-1}	8.25×10^{-3}	-0.0716	0.0271	0.9309	[0.8828; 0.9817]
6	93072687	rs16885729	C	0.8525	-0.1596	7.21×10^{-5}	C	0.952	-0.0488	2.70×10^{-1}	4.66×10^{-2}	-0.1058	0.0532	0.8996	[0.8105; 0.9984]
6	117734746	rs9489152	A	1.08	0.077	7.18×10^{-4}	A	1.032	0.0313	4.23×10^{-1}	8.89×10^{-4}	0.0654	0.0197	1.0676	[1.0272; 1.1096]
7	155624226	rs288756	A	1.079	0.076	8.96×10^{-4}	A	1.049	0.0481	2.39×10^{-1}	5.07×10^{-4}	0.0694	0.02	1.0718	[1.0307; 1.1146]
7	155651102	rs1731839	T	1.102	0.0971	4.18×10^{-5}	T	1.046	0.0447	2.33×10^{-1}	3.96×10^{-5}	0.0822	0.02	1.0857	[1.0439; 1.1291]
8	92379327	rs4236765	A	1.101	0.0962	4.47×10^{-5}	A	1.258	0.2299	2.83×10^{-1}	2.95×10^{-5}	0.0978	0.0234	1.1028	[1.0533; 1.1546]
9	14125509	rs2382436	C	1.131	0.1231	4.82×10^{-5}	C	1.276	0.2438	3.55×10^{-1}	3.54×10^{-5}	0.1247	0.0302	1.1328	[1.0678; 1.2017]
9	22067554	rs10811649	C	0.9139	-0.09	1.25×10^{-4}	C	0.935	-0.0674	1.21×10^{-1}	3.88×10^{-5}	-0.0849	0.0206	0.9186	[0.8822; 0.9565]
10	23502417	rs72802160	A	1.087	0.0834	9.94×10^{-4}	A	1.016	0.0157	7.89×10^{-1}	1.71×10^{-3}	0.0728	0.0232	1.0755	[1.0277; 1.1256]
10	73490681	rs61852058	A	1.107	0.1017	5.77×10^{-4}	A	1.026	0.0257	6.76×10^{-1}	1.00×10^{-3}	0.0875	0.0266	1.0915	[1.0360; 1.1499]

CHR	POS	rsID	PanScan-PanC4				East Asian GWAS				metafor/R results				
			EA	OR	BETA	*p*-value	EA	OR	BETA	*p*-value	*p*val_meta	BETA	SE	OR	95%CI
10	91349331	rs11185773	C	0.898	-0.1076	4.25×10^{-4}	C	0.952	-0.0497	1.92×10^{-1}	3.61×10^{-4}	-0.085	0.0238	0.9185	[0.8767; 0.9624]
10	91386496	rs11185819	C	0.899	-0.1065	1.53×10^{-4}	C	0.952	-0.0489	2.16×10^{-1}	1.45×10^{-4}	-0.0871	0.0229	0.9166	[0.8763; 0.9587]
11	27404856	rs11029989	C	1.11	0.1044	2.36×10^{-5}	C	1.054	0.0522	1.80×10^{-1}	1.69×10^{-5}	0.0895	0.0208	1.0936	[1.0499; 1.1391]
11	34653023	rs11032793	C	1.17	0.157	4.27×10^{-5}	C	1.374	0.3175	3.77×10^{-2}	7.90×10^{-6}	0.1666	0.0373	1.1812	[1.0980; 1.2708]
12	38653362	rs1906263	C	0.9295	-0.0731	1.49×10^{-3}	C	1.073	0.07	5.36×10^{-1}	2.80×10^{-3}	-0.0674	0.0226	0.9348	[0.8944; 0.9771]
12	38738984	rs11183418	A	1.079	0.076	9.64×10^{-4}	A	1.036	0.0357	4.78×10^{-1}	9.47×10^{-4}	0.0691	0.0209	1.0715	[1.0285; 1.1163]
12	56011457	rs11609481	T	1.089	0.0853	2.12×10^{-4}	T	1.014	0.0139	8.16×10^{-1}	3.97×10^{-4}	0.076	0.0215	1.0789	[1.0345; 1.1253]
12	129781336	rs10773631	C	0.9199	-0.0835	2.93×10^{-4}	C	0.818	-0.2013	9.62×10^{-2}	1.10×10^{-4}	-0.0876	0.0227	0.9161	[0.8763; 0.9577]
14	73942963	rs7157889	A	0.9171	-0.0865	2.11×10^{-3}	A	1.184	0.1685	3.55×10^{-1}	3.74×10^{-3}	-0.0806	0.0278	0.9226	[0.8736; 0.9742]
14	74442857	rs17782328	A	0.9151	-0.0887	5.59×10^{-3}	A	0.99	-0.0104	8.91×10^{-1}	9.08×10^{-3}	-0.077	0.0295	0.9259	[0.8738; 0.9810]
14	100616812	rs11624002	T	1.13	0.1222	$8.20\times10\text{-}^{5}$	T	1.073	0.0708	1.84×10^{-1}	4.91×10^{-5}	0.1091	0.0269	1.1153	[1.0581; 1.1757]
16	7532188	rs10852684	C	1.071	0.0686	2.59×10^{-3}	C	1.032	0.0316	5.47×10^{-1}	2.74×10^{-3}	0.0627	0.0209	1.0647	[1.0219; 1.1093]
17	45965966	rs11079800	T	0.9117	-0.0924	2.56×10^{-4}	T	0.972	-0.0283	4.67×10^{-1}	5.35×10^{-4}	-0.0734	0.0212	0.9292	[0.8914; 0.9686]
18	60756884	rs751894	T	0.9255	-0.0774	6.88×10^{-3}	T	0.934	-0.0682	2.14×10^{-1}	2.98×10^{-3}	-0.0754	0.0254	0.9273	[0.8823; 0.9747]
18	65003602	rs1455742	A	1.069	0.0667	4.02×10^{-3}	A	0.979	-0.021	7.38×10^{-1}	1.04×10^{-2}	0.0561	0.0219	1.0577	[1.0133; 1.1040]
19	18439383	rs12985909	C	0.9144	-0.0895	9.64×10^{-5}	C	0.964	-0.0363	3.73×10^{-1}	1.25×10^{-4}	-0.0767	0.02	0.9262	[0.8906; 0.9632]
20	6064731	rs2232079	T	0.8765	-0.1318	5.31×10^{-4}	T	0.831	-0.1852	3.94×10^{-2}	6.47×10^{-5}	-0.1399	0.035	0.8694	[0.8117; 0.9312]

CHR: Chromosome. POS: Position of the variant on the hg19 genome build. rsID: SNP identifier. EA: Effect allele. NEA: Non-effect allele. N: Sample size. OR: Odds ratio. BETA: Regression coefficient. *p*-value: Statistical significance of the association. *p*val_meta: *p*-value of the meta-analysis. SE: Standard error of the regression coefficient. 95%CI: 95% confidence interval of the effect estimate.

Table 6. Meta-analysis of the enhancer SNPs in both PanScan-PanC4 and East Asian GWAS datasets.

CHR	POS	rsID	PanScan-PanC4					East Asian GWAS					meta/R results				
			EA	NEA	N	BETA	*p*-value	EA	NEA	BETA	*p*-value	N	*p*val_meta	BETA	SE	OR	95%CI
1	22283583	rs6675669	A	G	15792	-0.1406	6.01×10^{-5}	A	G	-0,0257	5.22×10^{-1}	34631	1.22×10^{-1}	-0.085	0.055	0.92	[0.8248; 1.0229]
1	22302082	rs11591033	C	T	15609	-0.1579	5.12×10^{-5}	T	C	-0,0156	8.19×10^{-1}	34631	3.26×10^{-1}	-0.081	0.082	0.92	[0.7846; 1.0838]
1	22302509	rs11581262	T	C	15605	-0.1589	4.65×10^{-5}	T	C	0,0145	8.31×10^{-1}	34631	3.19×10^{-1}	-0.0821	0.082	0.92	[0.7839; 1.0826]
4	148401190	rs6841581	A	G	15812	-0.1403	2.70×10^{-5}	A	G	-0,0498	2.23×10^{-1}	34631	2.43×10^{-2}	-0.0983	0.044	0.91	[0.8320; 0.9873]
11	27395875	rs2472632	A	C	15802	0.1080	4.35×10^{-6}	A	C	0,0497	2.01×10^{-1}	34631	4.37×10^{-6}	0.0924	0.02	1.1	[1.0544; 1.1409]
11	27423086	rs7952482	T	C	15647	0.1106	6.61×10^{-5}	T	C	0,0327	5.16×10^{-1}	34631	1.38×10^{-4}	0.0926	0.024	1.1	[1.0460; 1.1504]
11	34653771	rs17358295	G	T	15521	0.1587	1.64×10^{-5}	T	G	-0,3904	4.60×10^{-3}	30631	2.67×10^{-2}	0.2345	0.106	1.26	[1.0275; 1.5555]
14	100616812	rs11624002	T	C	15699	0.1222	8.20×10^{-5}	T	C	0,0708	1.84×10^{-1}	34631	4.91×10^{-5}	0.1091	0.027	1.12	[1.0581; 1.1757]
17	37831613	rs1565920	G	A	15807	0.0989	4.32×10^{-5}	A	G	-0,0165	6.58×10^{-1}	34631	1.09×10^{-1}	0.063	0.039	1.07	[0.9861; 1.1503]
19	18439383	rs12985909	C	T	15804	-0.0895	9.64×10^{-5}	T	C	0,0363	3.73×10^{-1}	34631	1.25×10^{-4}	-0.0767	0.02	0.93	[0.8906; 0.9632]
19	18439520	rs35999725	T	C	15469	-0.1099	5.59×10^{-5}	T	C	0,1925	5.72×10^{-3}	34631	8.17×10^{-1}	0.0341	0.147	1.03	[0.7751; 1.3813]
21	15755182	rs34742635	A	G	15809	0.1007	8.83×10^{-5}	A	G	0,0147	8.53×10^{-1}	34631	1.48×10^{-4}	0.0926	0.024	1.1	[1.0458; 1.1507]

CHR: Chromosome. POS: Position of the variant on the hg19 genome build. rsID: SNP identifier. EA: Effect allele. NEA: Non-effect allele. N: Sample size. OR: Odds ratio. BETA: Regression coefficient. *p*-value: Statistical significance of the association. *p*val_meta: *p*-value of the meta-analysis. SE: Standard error of the regression coefficient. 95%CI: 95% confidence interval of the effect estimate.

After excluding known PDAC risk loci and SNPs in high LD ($r^2 > 0.6$), I selected the top 4 significant TFBS SNPs (rs10025845, rs11241697, rs11032793, rs2232079) and the top 3 significant enhancer SNPs (rs2472632, rs17358295, rs11624002) for genotyping analysis in PANDoRA. I present a schematic overview of SNP selection in **Figure 5**.

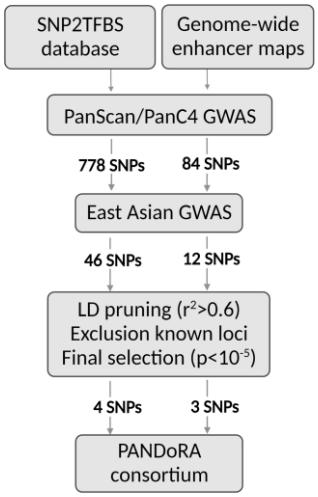


Figure 5. Summary of the variant selection workflow.
Figure reproduced from Ünal et al., 2024, licensed under CC BY 4.0.

After QC steps to the initial set of 11,370 samples, the following exclusions were made: 1096 non-PDAC samples, 38 samples with missing information, 1582 samples with a genotyping completion rate below 80%, 500 duplicate samples, and 972 samples from countries that contributed only cases but no controls. In total, 4188 samples were excluded, resulting in a final dataset of 7182 individuals, including 3392 PDAC cases and 3790 controls.

In this project, I used a final sample size consisting of 12,636 PDAC cases and 43,443 controls, as summarized in **Table 4**.

Table 4. Characteristics of the study populations.

	PanScan-PanC4	**East Asian GWAS**	**PANDoRA**
Total	14,266	34,631	7182
Cases	7205	2039	3392
Controls	7061	32,592	3790
Male%	54.80%	57.00%	54.20%
Median Age (25-75% CI)	65.5 (55-75)	65.0	62.2 (53-71)
Number of variants*	7,509,345	7,914,378	7

*The number of variants remaining after quality control of GWAS data. For PANDoRA, the number of variants genotyped in the replication phase.

A meta-analysis was performed using the logistic regression results for the seven SNPs from the PANDoRA dataset, PanScan-PanC4, and an East Asian GWAS. The associations between these seven SNPs and PDAC risk across the three populations are presented in **Table 7**. The overall meta-analysis results for the four study-wide significant SNPs are shown in **Figure 6**.

Table 7. Associations of the selected SNPs with PDAC risk.

rsID	Locus	EA	NEA	EUR_freq	PanScan-PanC4		East Asian GWAS		PANDoRA		Overall meta-analysis	
					p-value	OR [95%-CI]	*p*-value	OR [95%-CI]	*p*-value	OR [95%-CI]	*p*-value	OR [95%-CI]
rs10025845	4p14	A	G	0.41	3.76×10^{-4}	1.08 [1.03; 1.13]	3.43×10^{-2}	1.09 [1.00; 1.18]	2.00×10^{-2}	1.08 [1.01; 1.16]	**1.32×10^{-5}**	1.08 [1.05; 1.12]
rs11241697	5q23.2	C	T	0.41	2.46×10^{-4}	0.91 [0.87; 0.96]	4.68×10^{-3}	0.89 [0.89; 0.96]	1.54×10^{-1}	1.08 [0.96; 1.22]	3.67×10^{-1}	0.95 [0.83; 1.06]
rs2472632	11p14.1	A	C	0.36	4.35×10^{-6}	1.11 [1.06; 1.16]	2.01×10^{-1}	1.05 [0.97; 1.13]	1.00×10^{-3}	1.13 [1.05; 1.21]	**5.52×10^{-8}**	1.10 [1.06; 1.13]
rs11032793	11p13	C	T	0.1	4.27×10^{-5}	1.17 [1.08; 1.26]	3.77×10^{-2}	1.37 [1.01; 1.85]	5.23×10^{-1}	0.97 [0.91; 1.04]	2.69×10^{-1}	1.10 [0.93; 1.27]
rs17358295	11p13	G	T	0.11	1.64×10^{-5}	1.17 [1.09; 1.25]	4.60×10^{-3}	1.47 [1.13; 1.93]	1.10×10^{-1}	1.09 [0.97; 1.22]	**6.10×10^{-7}**	1.16 [1.10; 1.22]
rs11624002	14q32.2	T	C	0.16	8.20×10^{-5}	1.13 [1.06; 1.20]	1.84×10^{-1}	1.07 [0.96; 1.19]	4.94×10^{-1}	1.03 [0.94; 1.13]	1.04×10^{-4}	1.09 [1.04; 1.14]
rs2232079	20p12.3	T	C	0.11	5.31×10^{-4}	0.87 [0.81; 0.94]	3.94×10^{-2}	0.83 [0.69; 0.99]	4.00×10^{-1}	0.94 [0.83; 1.07]	**6.38×10^{-6}**	0.88 [0.83; 0.93]

rsID: Reference SNP ID. Locus: Genomic region or gene where the SNP is located. EA: Effect allele. NEA: Non-effect allele. EUR_freq: Effect/minor allele frequency in European populations of the 1000 Genomes Project. *p*-value: Statistical significance of the association. OR [95%-CI]: Odds ratio with 95% confidence interval, indicating effect size and direction. Values in bold are study-wise significant (p-value $< 1.94 \times 10^{-5}$).

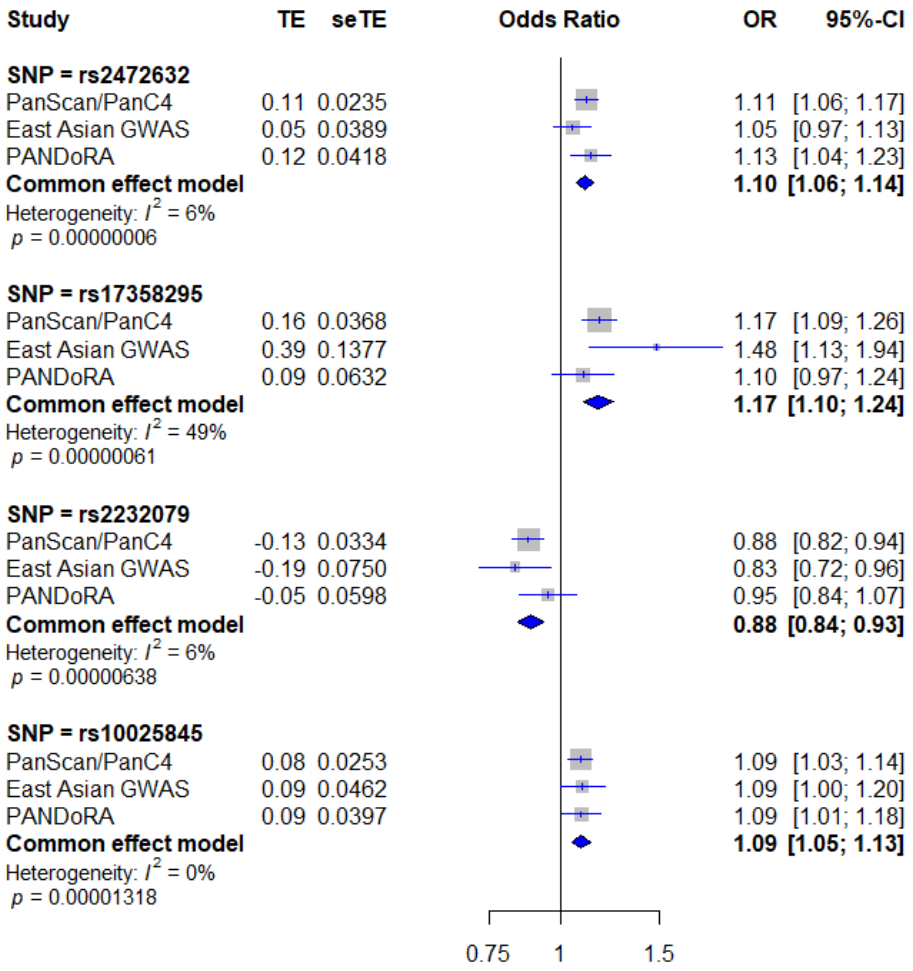


Figure 6. Summary of overall meta-analysis results for the four study-wise significant variants.

Functional Genomic Analyses Highlight rs2472632(A) as a Regulatory Variant Linked to *CCDC34*, *LGR4*, and *LIN7C* Expression

The rs2472632(A) variant was found to be associated with increased PDAC risk at a near genome-wide significance level (p-value = 5.52×10^{-8}) after the meta-analysis of all three populations, with a consistent effect direction across all datasets. This enhancer variant is predicted to influence the expression of the *CCDC34* gene in normal pancreas tissue (ABC score = 0.016; see original paper for ABC score computation) (**Figure 7a**). In pancreatic adenocarcinoma (PAAD) tumor tissues, *CCDC34* was significantly overexpressed compared to normal pancreas tissue (p-value

= 1.22 × 10^{-19}) according to the TNMplot database, which integrates RNA-seq data from TCGA and GTEx (**Figure 7b**). Additionally, rs2472632(A) is located in an intronic region of the *LGR4* gene, which was found to be 3.1 times more highly expressed in PDAC tissues than in normal pancreas tissues according to the Expression Atlas database (False Discovery Rate, FDR-adjusted *p*-value = 1.92 × 10^{-11}). The sQTL analysis from GTEx indicates that rs2472632(A) is associated with an alternative splicing mechanism of *LGR4* mRNA (*p*-value = 2.12 × 10^{-23}) (**Figure 7c**). Furthermore, eQTL analysis in GTEx suggests that rs2472632(A) is associated with higher expression of LIN7C in pancreatic tissue (*p*-value = 8.62 × 10^{-5}) (**Figure 7d**).

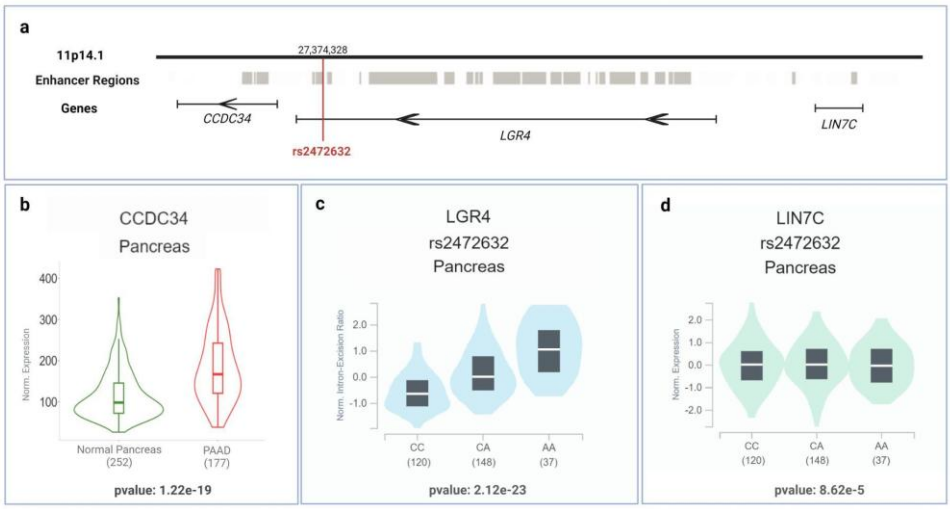


Figure 7. Functional and genomic context of rs2472632.

a) Genomic location of the SNP rs2472632. b) *CCDC34* gene expression in pancreas tissue from normal vs Pancreatic Adenocarcinoma patients in the TNMplot database. c) Violin plot of *LGR4* - rs2472632 sQTL analysis results in GTEx database. d) Violin plot of *LIN7C*-rs2472632 eQTL analysis in GTEx database. Figure reproduced from Ünal et al., 2024, licensed under CC BY 4.0.

Furthermore, rs2472632(A) was found to be associated with reduced protein levels of DEFA5, also known as HD5, according to data from the GWAS Catalog (*p*-value = 1.0×10^{-25}, Study accession: GCST90247261). The enhancer variant rs17358295(G) (*p*-value = 6.1×10^{-7}) and the TFBS variant rs2232079(T) (*p*-value = 6.38×10^{-6}) were both significantly linked to PDAC risk in the overall meta-analysis.

Additional Variants Show Consistent Risk Associations and Regulatory Potential Across Genomic Datasets

While rs17358295(G) was associated with an increased risk, rs2232079(T) was linked to a decreased risk across all three studied populations. The rs17358295 variant is located within an enhancer region that interacts with 20 genes in normal pancreatic tissue, with ABC scores ranging from 0.015 to 0.130 (**Table 8**).

The TFBS variant rs10025845(A) was nominally associated with PDAC risk (p-value < 0.05) in all three populations and reached study-wide significance following the overall meta-analysis (p-value $= 1.35 \times 10^{-5}$). Notably, the minor allele (G) of this variant creates a binding site for the transcription factor *YY1*, whereas the effect allele (A) is predicted by SNP2TFBS to lack any transcription factor binding. Additionally, rs10025845(A) overlaps with *LINC01258*, a long intergenic non-coding RNA gene.

The enhancer variant rs11624002(T) also demonstrated an association with PDAC risk in the meta-analysis; however, it did not meet the Bonferroni-corrected significance threshold. This variant resides in an enhancer region near the protein-coding gene *DEGS2* and has an ABC score of 0.018.

Finally, two SNPs, rs11241697(C) and rs11032793(C), which were predicted to alter transcription factor binding patterns, did not show statistically significant associations in the overall meta-analysis. Additionally, these variants exhibited high heterogeneity ($I^2 > 75\%$), indicating potential variability in effect sizes across the studied populations.

Table 8. ABC scores for selected enhancer variants and their target gene predictions.

rsID	chr:pos	activity_base	TargetGene	TargetGeneTSS	TargetGene PromoterActivity Quantile	distance	ABC.Score	CellType
rs2472632	11:27395874	1.383414	*CCDC34*	27384795	0.67731	10886.5	0.01608	pancreas-Roadmap
rs17358295	11:34653770	21.0354	*TRIM44*	35684299	0.85168	1030535	0.017435	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*APIP*	34937958	0.97947	284194	0.02534	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*FJX1*	35639734	0.53149	985970	0.01993	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*LDLRAD3*	35965530	0.53842	1311766	0.017692	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*ABTB2*	34379555	0.71574	274209	0.01995	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*COMMD9*	36310999	0.56282	1657235	0.017618	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*C11orf74*	36616042	0.61374	1962278	0.017263	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*NAT10*	34127110	0.82907	526654	0.019511	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*PDHX*	34937676	0.98248	283912	0.025339	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*FBXO3*	33796071	0.68124	857693	0.01613	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*FBXO3-AS1*	33796244	0.6616	857520	0.01613	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*LOC100507144*	35159579	0.42048	505815	0.019495	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*PAMR1*	35547579	0.42447	893815	0.021493	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*CD44*	35160416	0.68091	506652	0.016231	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*CAT*	34460471	0.93784	193293	0.02068	body_of_pancreas-ENCODE
rs17358295	11:34653770	21.0354	*EHF*	34642587	0.76275	11177	0.129516	body_of_pancreas-ENCODE
rs17358295	11:34653770	16.83716	*EHF*	34642587	0.84394	11173	0.110963	pancreas-Roadmap
rs17358295	11:34653770	16.83716	*ELF5*	34535347	0.43038	118413	0.026981	pancreas-Roadmap
rs17358295	11:34653770	16.83716	*APIP*	34937958	0.98566	284198	0.017036	pancreas-Roadmap
rs17358295	11:34653770	16.83716	*PDHX*	34937676	0.99165	283916	0.017036	pancreas-Roadmap
rs17358295	11:34653770	16.83716	*CAT*	34460471	0.95979	193289	0.015009	pancreas-Roadmap
rs11624002	14:100616811	2.479308	*DEGS2*	1.01E+08	0.4221	9342.5	0.018028	pancreas-Roadmap

rsID: SNP identifier. chr:pos: Chromosome and position of the variant (hg19). activity_base: Enhancer activity score prior to 3D contact weighting. TargetGeneTSS: Transcription start site of the predicted target gene. TargetGenePromoterActivityQuantile: Relative promoter activity level of the target gene across cell types. distance: Genomic distance between enhancer and target gene TSS. ABC.score: Activity-by-Contact score estimating enhancer–gene regulatory potential. CellType: Cell type or tissue used for ABC prediction. The table is adapted from (Nasser et al., 2021).

4.4 DISCUSSION

The majority of SNPs associated with disease risk are located outside protein-coding regions(French & Edwards, 2020), making the interpretation of GWAS results challenging. This remains true even after fine mapping around associated loci(Farh et al., 2015). Most disease-associated variants influence gene expression by altering functional DNA elements. Recently developed tools provide improved predictions for the functional characteristics of non-coding variants. In this project, I leveraged advanced functional annotation to conduct a comprehensive association analysis between non-coding variants and PDAC susceptibility.

The overall meta-analysis identified a variant, rs2472632(A), that nearly reached genome-wide significance for PDAC risk. This enhancer variant is predicted to target the *CCDC34* gene, an oncogene upregulated in bladder(Gong et al., 2015), cervical(L.-B. Liu et al., 2018), colorectal(Geng et al., 2018), and pancreatic(Qi et al., 2017) cancers. Studies using TCGA and GTEx datasets, as reported by Qi et al. and the TNMplot database, demonstrated significantly increased *CCDC34* mRNA expression in PAAD compared to normal pancreatic tissue, correlating with poor overall survival. However, no functional studies have yet elucidated the gene's direct role in PDAC development.

Additionally, rs2472632 is located within an intronic region of *LGR4*, a known activator of the Wnt/β-catenin signaling pathway(Glinka et al., 2011). While crucial in development, aberrant activation of this pathway predisposes individuals to various cancers, including melanoma(Fan et al., 2017), multiple myeloma(van Andel et al., 2017), ovarian cancer(Z. Wang et al., 2020), and thyroid carcinoma(Kang et al., 2017). In pancreatic cancer, Wnt/β-catenin signaling contributes to apoptosis resistance, promoting disease progression(Modi et al., 2016). *LGR4*, widely expressed in pancreatic tissue, modulates cellular responses to Wnt ligands(Mazerbourg et al., 2004). According to sQTL analysis, rs2472632 appears to reside at a splicing site of an alternative exon of *LGR4*, suggesting a potential role in PDAC risk through alternative RNA splicing. Moreover, eQTL analysis links rs2472632 to *LIN7C* expression, a membrane trafficking protein implicated in metastasis(Onda et al., 2007). Notably, deletions on 11p14.1, the chromosomal region harboring this enhancer variant, have been associated with attention-deficit hyperactivity disorder (ADHD), developmental delay, and obesity(Shinawi et al., 2011).

Furthermore, rs2472632(A) is associated with reduced HD5 peptide levels. HD5, a defensin protein, exhibits potent antimicrobial activity by forming pores in pathogen

membranes(Jung et al., 2017). Paneth cells release HD5 in response to bacterial and cholinergic stimuli, regulating acute and chronic inflammation(Selsted & Ouellette, 2005; Yang & Shen, 2021). Studies have detected HD5 protein in PDAC tissues via immunohistochemistry(Tobi et al., 2013). In a rat model, chemically induced pancreatic duct damage resulted in increased HD5 levels, suggesting a protective role against tissue injury(Cunha et al., 2014). I hypothesise that individuals carrying the rs2472632(A) allele might exhibit lower HD5 expression, reducing protection against pancreatitis, a known PDAC risk factor. Additionally, a study in a healthy Japanese population found significantly lower fecal HD5 concentrations in individuals over 70 years old, supporting the idea that reduced HD5 levels may contribute to disease susceptibility(Shimizu et al., 2022).

Although the precise mechanism linking rs2472632 to PDAC risk remains unclear, multiple lines of evidence highlight the locus's potential functional relevance, warranting further investigation.

I also identified two additional loci significantly associated with PDAC risk: rs10025845(A) and rs2232079(T). The minor allele (A) of rs10025845 is not predicted to create any TF binding site, whereas the G allele generates a *YY1* binding site. *YY1* has been implicated as a tumor suppressor in PDAC(J.-J. Zhang et al., 2014), suggesting that rs10025845(A) may impair its tumor-suppressive function. Meanwhile, rs2232079 is located within a *PAX5* TF binding site and the promoter region of *FERMT1*, which encodes Kindlin-1. Kindlin-1 is overexpressed in pancreatic cancer cell lines but expressed at low levels in normal pancreatic epithelium and fibroblasts(Mahawithitwong et al., 2013).

According to enhancer mapping data, rs17358295(G) is located within an enhancer interacting with 20 genes. The highest ABC score (0.130) for this variant corresponds to the *EHF* gene, whose role in cancer remains largely unexplored. Some studies suggest that EHF overexpression contributes to metastasis, proliferation, and poor survival in cancer patients(J. Liu et al., 2019; L. Wang et al., 2020; T. Zhou et al., 2022).

The strengths of this part of the thesis include its large sample size encompassing multiple ethnicities and the comprehensive evaluation of two major functional classes of polymorphisms. These factors facilitated the identification of a germline variant approaching genome-wide significance for PDAC risk. However, the project has limitations: TFBS SNP data relied on in silico predictions, enhancer SNP functional predictions were based solely on positional data, and experimental validation of findings is lacking. Future studies should aim to validate these associations through functional assays.

5 PROJECT 2: RARE GERMLINE VARIANTS IN PDAC

5.1 INTRODUCTION

Estimates of pancreatic cancer heritability based on GWAS of common germline variants in European populations range from 9.8% to 18%(Chen et al., 2019; Childs et al., 2015; Lu et al., 2014). The unexplained portion of heritability may be attributed to rare germline variants, particularly if individuals carry multiple disease-associated rare variants(Bodmer & Tomlinson, 2010; Wainschtein et al., 2019), or to an interplay between rare and common risk variants that collectively contribute to disease susceptibility.

Despite extensive research on common germline variants, the role of uncommon variants remains largely unexplored. This is primarily due to the limited coverage of GWAS for variants with MAF < 1%, which constitute the majority of human genetic variation, and the high cost of whole-genome or whole-exome sequencing, which restricts large-scale analysis of these variants (**Figure 8**).

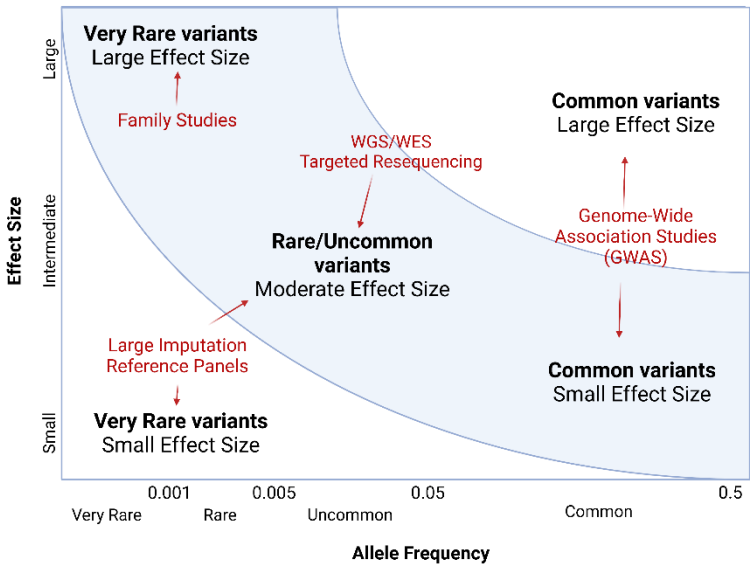


Figure 8. Allelic frequencies and effect sizes for genetic variants associated with cancer risk.

A subset of rare germline variants (defined here as MAF < 1%) exhibit high penetrance in familial pancreatic cancer and some hereditary genetic syndromes, which are established risk factors for pancreatic cancer. Family linkage analysis and multigene panel testing have identified 17 pancreatic cancer risk genes(L. T. Amundadottir, 2016; reviewed in Gentiluomo et al., 2022).

However, detecting low-penetrance rare variants has been challenging due to the lack of rare variant catalogs for genotyping array design, insufficient cohort sizes, and limited methods for accurately predicting variant function(Momozawa & Mizukami, 2021). Advances in whole-genome/exome sequencing, genotyping technologies, and genotype imputation methods have now enabled large-scale rare variant analysis(Weissenkampen et al., 2019).

A cost-effective approach to investigating rare germline variants in large-scale association studies involves genotyping via genome-wide SNP arrays and imputing missing genotypes using reference panels. Genotype imputation is widely used in GWAS to reconstruct common variant genotypes. With larger reference panels, it is now possible to impute rare variants with high confidence (assessed by INFO scores or squared correlations)(Das et al., 2018; Huang et al., 2015; Mitt et al., 2017).

Rare variant association studies require different statistical approaches than those used for common variants, as statistical power decreases with lower MAF(B. Li & Leal, 2008). Set-based tests, which aggregate rare variants within a genetic region to enhance power, are commonly used. These include burden tests, SKAT(M. C. Wu et al., 2011),

SKAT-O(S. Lee, Emond, et al., 2012), and the aggregated Cauchy association test (ACAT)(Y. Liu et al., 2019). Using such approaches, numerous rare deleterious variants and genes have been linked to various traits(Clay et al., 2024; Jurgens et al., 2022; Mueller et al., 2023; Schijven et al., 2024).

Rare variant set-based tests improve statistical power by leveraging computational methods and functional annotations. The burden test aggregates rare variants into a single score for association testing, while SKAT evaluates the distribution of test statistics across aggregated scores, allowing for variable effect directions. STAAR (variant-Set Test for Association using Annotation infoRmation) integrates functional annotations, weighting variants based on biological relevance, and its extension, STAAR-O, combines multiple association tests (including burden tests, SKAT, and ACAT-V) to enhance power. ACAT transforms *p*-values into Cauchy variables and computes a weighted sum for significance testing.

In this project, I conducted a genome-wide search for rare germline variants associated with PDAC risk by leveraging the largest published cohorts on PDAC—PanScan I, PanScan II, PanC4, and UK Biobank—along with imputation using large reference panels. I first conducted single-variant association tests, followed by gene-based burden tests using MAGMA(Leeuw et al., 2015b), summing rare variants within each gene window, both with and without functional annotation score weights. Finally, I performed set-based analyses of functionally annotated rare variants using the "STAARpipeline" (release v0.9.6) R package(X. Li et al., 2020; Z. Li et al., 2022).

5.2 METHODS

5.2.1 Genetic Data, Imputation and Quality Control

5.2.1.1 PanScan-PanC4

I used the R package "plinkQC"(Meyer, 2020) to perform pre-imputation quality control (pre-QC) on the PanScan I, PanScan II, and PanC4 genotype datasets separately, resulting in 3038, 3643, and 7710 samples, respectively. The pre-QC steps involved removing samples with cryptic relatedness (IBD > 0.185, which is midway between the expected IBD for third- and second-degree relatives), sex mismatches, non-European ancestry, heterozygosity rates ±3 SD from the mean, and missingness

rates >3%. Additionally, I excluded genetic variants that violated HWE (p-value < 1×10^{-6} in controls), had a MAF < 0.005, or a call rate < 90%.

I performed imputation separately on the pre-QCed genotype datasets (restricted to autosomal chromosomes) using the TOPMed imputation reference panel (version TOPMed-r3) via the TOPMed imputation server. After imputation, the PanScan I, PanScan II, and PanC4 datasets were merged to form a single dataset, referred to hereafter as PanScan-PanC4.

To ensure data quality post-imputation, I applied post-QC steps to the merged genotype dataset. Specifically, I excluded genetic variants with low imputation quality (INFO score, $R^2 \leq 0.5$) with BCFtools, deviations from HWE (p-value < 1×10^{-6} in controls), and a missingness rate >2% with PLINK version 1.9. For sample-level QC, I used KING software(Manichaikul et al., 2010) to identify first- and second-degree relatives via integrated relationship inference (IBD > 0.185) using the ' *--related* ' command. From each pair of related individuals, I retained the sample with the higher genotyping rate. After all QC steps, 137 samples were excluded, leaving a final dataset of 14,254 samples, including 7239 PDAC cases.

5.2.1.2 UK Biobank

I accessed imputed genotype data (BGEN format) from the UK Biobank 500k March 2018 release, provided in the hg19 genome assembly. I extracted the genetic data for the selected subjects by using the BGENIX tool and applied post-QC steps with PLINK (version 2.0), excluding samples with sex mismatches, heterozygosity rates ±3 SDs from the mean, and missingness rates >3%. Additionally, I removed genetic variants with INFO scores ≤ 0.5, strong deviations from HWE (p-value < 1×10^{-6} in controls), and missingness rates >2%. After these filtering steps, 1021 cases and 285,731 controls remained.

To optimize computational efficiency in subsequent analyses, I randomly selected 10,000 controls from the post-QCed control set, ensuring that their age distribution matched that of the cases. I then converted the variant coordinates from hg19 to hg38 using the UCSC Liftover and Ensembl Assembly converter, achieving a 98% conversion rate. Finally, I updated the variant positions in the dataset using PLINK version 2.0 via the ' *--update-map* ' command.

5.2.2 Bioinformatics Tools for Rare Variant Association Test

For rare variant analysis, MAGMA aggregates rare variants within a gene into burden scores, which are then analyzed instead of individual variants. The tool also allows for variant-specific weighting, enhancing flexibility in association testing.

Another method I applied is the STAARpipeline, a robust and scalable framework for rare variant association testing. This method integrates multiple variant categories and functional annotations using a weighted scheme, while accounting for population structure and relatedness, making it particularly effective for large-scale biobank and whole-genome sequencing studies.

STAARpipeline introduces annotation principal components to summarize in silico functional annotations of rare variants. It leverages data from the Functional Annotation of Variants Online Resource (FAVOR)(H. Zhou et al., 2023), which provides comprehensive functional annotation data from diverse sources, along with precomputed annotation principal components. In the FAVOR database, each variant receives a score based on 11 functional annotations: CADD, FATHMM-XF, LINSIGHT, aPC.EpigeneticActive, aPC.EpigeneticRepressed, aPC.EpigeneticTranscription, aPC.Conservation, aPC.LocalDiversity, aPC.Mappability, aPC.TF, and aPC.Protein.

The STAAR framework applies the STAAR-O (Omnibus Test), which combines *p*-values from multiple annotation-weighted variant set tests using the ACAT method. In addition to the STAAR-O *p*-value, the pipeline reports results from SKAT, ACAT-V, and Burden tests. Furthermore, for each individual variant, it conducts a score test and outputs a corresponding *p*-value.

5.2.3 Statistical Analysis

5.2.3.1 Single-Variant Association Analysis

To evaluate associations of individual genetic variants in PanScan-PanC4 and UK Biobank, I performed a score test using the "STAARpipeline-Tutorial" R package (release v0.9.6), adjusting for age, sex, and 10 PCs. The score test was chosen for its computational efficiency and suitability for hypothesis testing in large datasets. This

method assesses whether a genetic variant influences disease risk by testing the null hypothesis that the variant has no effect, using the first derivative of the likelihood function at the null parameter value.

I followed the STAARpipeline tutorial as described by the authors, with detailed instructions available on GitHub (xihaoli/STAARpipeline-Tutorial, release v0.9.6). The analysis was conducted using R/4.3.0 on an LSF-based compute cluster provided by the German Cancer Research Center. The following tools were used:

- "SeqArray R" package(Zheng et al., 2017) to convert PLINK files to Genomic Data Structure (GDS) format
- "FAVORannotator" R package (CSV version 1.0.0) for variant annotation
- "STAARpipeline" R package for variant-level and gene-level association analyses
- "STAARpipelineSummary" R package to summarize and visualize results, excluding known risk variants and their LD pairs (see **Table 1**)

To validate the score test results, I employed REGENIE to perform a whole-genome regression score test and PLINK version 2.0 to conduct a logistic regression score test. All tests were adjusted for age, sex, genotyping arrays, and 10 PCs. The QQ plots stratified by allele frequency were generated using the "qqman" R package (v0.1.9). Genomic inflation factors (λ) are shown in **Supplementary Figure 1** for the PanScan-PanC4 dataset and **Supplementary Figure 2** for the UK Biobank dataset. The lambda score was computed using the formula:

$$z = qnorm(pvalue/2)$$

$$\lambda = qchisq(0.5, df = 1) \,/\, median(z2, na.rm = T)$$

Finally, I used PLINK to calculate haplotype-based r^2 and D' scores between rare variant pairs using the ' *--ld* ' flag.

5.2.3.2 Annotation-unweighted Gene-based Analysis

I analyzed all 19,264 protein-coding genes located on autosomal chromosomes. For each gene, I defined a genomic region extending 5 kb upstream of the first known exon and 2 kb downstream of the last known exon. Within this window, I included all rare variants associated with each gene.

To prepare the post-QC files, I extracted only rare variants ($0.01 \geq MAF \geq 1 \times 10^{-5}$) from each dataset separately (PanScan-PanC4 and UK Biobank) and converted them into plink file format. I followed the recommended software guidelines to conduct gene-based association analyses, adjusting for age, sex, genotyping arrays, and 10 PCs as covariates.

For gene-based analysis, I used MAGMA with the flags: ' *--gene-model snp-wise=mean* ' and ' *--burden all* '. This approach calculates the SKAT value using an inverse variance weight, based on a burden score that sums all rare variants within a gene window.

Next, I performed a fixed-effect meta-analysis to combine results from PanScan-PanC4 and UK Biobank using the following MAGMA command: ' *--meta genes*= '. This step merged the ".genes.out" files generated from each dataset's gene-based analysis. For statistical significance, I applied Bonferroni correction, setting the threshold at: $0.05/19{,}264 = 2.59 \times 10^{-6}$.

5.2.3.3 Annotation-weighted Gene-based Analysis

I extracted 11 annotation scores (CADD, FATHMM.XF, LINSIGHT, aPC.EpigeneticActive, aPC.EpigeneticRepressed, aPC.EpigeneticTranscription, aPC.Conservation, aPC.LocalDiversity, aPC.Mappability, aPC.TF, aPC.Protein) for each variant from the FAVOR database using the "STAARpipelineSummary" R package (version 0.9.7). For gene-based analysis, I used MAGMA to calculate burden scores weighted by each annotation via the ' *--burden weights* ' flag. I then combined the resulting *p*-values using the ACAT method, generating a single *p*-value per gene for both PanScan-PanC4 and UK Biobank datasets. Variants lacking annotation scores were automatically excluded from the gene-based test. Finally, I conducted a fixed-effect meta-analysis to merge the results across datasets using MAGMA.

Additionally, I performed a functionally annotated rare variant set-based analysis using "STAARpipeline" R package, stratifying variants into the following categories:

- Gene-centric coding: pLoF, pLoF_ds, missense, disruptive_missense, synonymous
- Gene-centric noncoding: downstream, upstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS
- ncRNA genes
- 2 kb dynamic windows covering the whole genome

The significance threshold for protein-coding genes with five different coding functional categories is $0.05/(20{,}000 \times 5) = 5.0 \times 10^{-7}$, for protein-coding genes with seven different noncoding functional categories is $0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$, and for ncRNA genes, it is $0.05/20{,}000 = 2.50 \times 10^{-6}$. I chose 20,000 as a correction factor to compute these thresholds according to the recommendations for the "STAARpipeline" R package manual.

5.2.4 Gene Set Enrichment Analysis

I conducted gene-set enrichment analysis using the Enrichr tool for two sets of significant genes: 20 genes from the annotation-unweighted meta-analysis and 15 genes from the annotation-weighted meta-analysis.

Enrichr is a web-based tool(Kuleshov et al., 2016) that enables enrichment analysis across 35 gene-set libraries by uploading a gene list. The results are reported using three statistical approaches: (1)Fisher's exact test *p*-value, (2)Benjamini-Hochberg adjusted *p*-value (correcting for multiple hypothesis testing) and (3)Combined score (integrating Fisher's exact test *p*-value with a z-score reflecting deviations from expected ranks).

5.3 RESULTS

Extensive Rare Variant Coverage Achieved in Two Large PDAC Case-Control Cohorts

In PanScan-PanC4, I had 7239 cases and 7015 controls, imputed with a diverse ancestry-based reference panel to yield over 24M variants, of which approximately 17M were rare ($0.01 \geq \text{MAF} \geq 1 \times 10^{-5}$). In the final UK Biobank dataset after post-QC, I had 1021 cases and 10,000 controls, imputed with European ancestry-based reference panels to over 33M variants, of which 24M were rare (**Table 9**). The number of overlapping rare variants between two datasets was over 10M.

Table 9. Study population characteristics.

Study	Cases	Controls	Case:Control	Ancestry	Imputation	N of rare variants*
PanScan-PanC4	7239	7015	1:1	European	TOPMed r3 reference panel (60% of non-European ancestry)	16,995,568
Age (Median)	65.4	65				
Male (%)	54.4	53				
UK Biobank	1021	10000	1:10	European	UK10K and 1000 Genomes phase 3 reference panels (European ancestry)	24,724,599
Age (Median)	67.5	68.1				
Male (%)	60.8	52.6				

*Total number of rare variants ($0.01 \geq MAF \geq 1 \times 10^{-5}$) on chr 1-22 after the imputation and quality control steps. The number of overlapping rare variants between both datasets was over 10M.

Rare Variants Show Limited Replication Across Datasets in Single-Variant Association Tests

According to the score test results for individual variant associations, after exclusion of known PDAC loci, 46 rare variants (18 independent variants) reached genome-wide significance (p-value $<5 \times 10^{-8}$) in the PanScan-PanC4 dataset. However, none of them showed significance in both whole-genome regression and logistic regression tests (**Supplementary Table 1**). Similarly, in the UK Biobank, 52 rare variants (29 independent variants) reached genome-wide significance with the score test. Only the rare variant rs369984058(C) showed a strong association in the additional tests, with a p-value of 7.48×10^{-8} (OR = 17.39) in whole-genome regression and a p-value of 4.41×10^{-11} (OR = 9.39) in logistic regression (**Supplementary Table 2**). Nevertheless, this variant was not present in the PanScan-PanC4 dataset. The rare variant rs188205290(T), which is in LD with rs369984058(C) (r2=0.74), is present in the PanScan-PanC4 dataset but showed no significant association in any test (p-value = 0.379). The results from one dataset were not replicated in the other dataset.

Gene-Based Analysis Identifies Distinct Sets of Significant Genes in PanScan-PanC4 and UK Biobank

I conducted a gene-based analysis on rare variants in both datasets. For the PanScan-PanC4 dataset, rare variants were mapped to 17,662 protein-coding genes. The analysis identified 51 genes with statistical significance exceeding the Bonferroni-corrected significance threshold of 2.59×10^{-6}. The most significant genes were *MPZL1* (p-value

= 9.23 × 10^{-10}), *DYSF* (*p*-value = 1.63 × 10^{-9}), *GPC5* (*p*-value = 5.33 × 10^{-9}) and *RBFOX1* (*p*-value = 1.40 × 10^{-8}). In the UK Biobank, rare variants were mapped to 19,155 protein-coding genes and the analysis revealed one significant gene, *ST7L* (*p*-value = 1.04 × 10^{-6}) (**Supplementary Table 3**).

A fixed-effect meta-analysis of the PanScan-PanC4 and UK Biobank gene-based analysis results was performed via MAGMA. The meta-analysis of annotation-unweighted gene *p*-values identified 20 genes that reached Bonferroni-corrected statistical significance: *MPZL1*, *PTPRT*, *MED23*, *NUBP2*, *ST7L*, *ZNF624*, *PARD3*, *RBFOX1*, *TMEM63C*, *PYROXD2*, *ARHGAP44*, *LOC105375678*, *OSGIN2*, *ZMYM1*, *RIPK2*, *LOC105373764*, *WDR92*, *SRGAP3*, *ENPP3* and *GPC6* (**Supplementary Table 3**).

Functional Annotation Weighting Reveals Additional Genes Not Detected in Unweighted Analysis

Using MAGMA, I incorporated 11 functional annotation scores from the FAVOR database to weigh the burden scores. This weighted analysis was independently performed for both datasets. In the PanScan-PanC4, rare variants with annotation scores were mapped to 17,634 genes and the annotation-weighted analysis highlighted 54 significant genes after combining *p*-values using the ACAT method. Notably, 15 of these genes were not significant in the initial annotation-unweighted gene-based analysis but reached significance after incorporating annotation weights, such as *SLC6A11*, *PRRX1*, *MPRIP*, *SLC30A5*, and *CLU* (**Supplementary Table 3**).

For the UK Biobank, rare variants with annotation scores were mapped to 19,122 protein-coding genes and the gene *ST7L* was significant after annotation-weighted analysis with *p*-value 2.55 × 10^{-6}.

Gene-Based Meta-Analysis Highlights Consistently Associated Genes Across Datasets

The meta-analysis between annotation-weighted gene *p*-values revealed 15 significant genes: *PTPRT*, *MPZL1*, *WDR92*, *RBFOX1*, *PYROXD2*, *MED23*, *ST7L*, *OSGIN2*, *PARD3*, *RIPK2*, *GPC6*, *NUBP2*, *SCGB1D2*, *PNO1* and *SRGAP3*. Notably, the genes *SCGB1D2* and *PNO1* were significant only after the inclusion of annotation weights. **Table 10** summarizes the top significant genes identified from the meta-analysis, as well as those with annotation weights.

Seven genes that were significant in the annotation-unweighted gene-based analysis - *TMEM63C*, *ZNF624*, *ARHGAP44*, *LOC105375678*, *ZMYM1*, *ENPP3* and *LOC105373764* - lost their significance during the meta-analysis of the annotation-weighted gene-based analysis.

Table 10. Significant results from gene-based analyses with MAGMA tool with/without functional annotations from meta-analysis.

			Annotation-unweighted gene-based analysis						***Annotation-weighted gene-based analysis***					
			PanScan-PanC4		UK Biobank		Meta-analysis		PanScan-PanC4		UK Biobank		Meta-analysis	
GENE	**CHR**	**Region**	**NSNP**	***p*-value**	**NSNP**	***p*-value**	**NSNP**	***p*-value**	**NSNP**	***p*-value**	**NSNP**	***p*-value**	**NSNP**	***p*-value**
ZMYM1	1:35054786-35117859	1p34.3	442	1.42×10^{-5}	609	1.27×10^{-2}	526	**1.92×10^{-6}**	404	3.89×10^{-5}	525	1.28×10^{-2}	465	3.47×10^{-6}
ST7L	1:112515804-112625838	1p13.2	807	6.71×10^{-3}	1091	**1.04×10^{-6}**	949	**3.01×10^{-7}**	736	8.18×10^{-3}	990	**2.55×10^{-6}**	863	**9.23×10^{-7}**
MPZL1	1:167716950-167793919	1q24.2	539	**9.23×10^{-10}**	720	1.01×10^{-1}	630	**4.25×10^{-8}**	490	**1.01×10^{-9}**	616	1.32×10^{-1}	553	**1.36×10^{-7}**
WDR92	2:68128149-68162560	2p14	299	1.85×10^{-4}	352	1.86×10^{-3}	326	**2.22×10^{-6}**	277	2.59×10^{-5}	336	4.59×10^{-3}	307	**3.82×10^{-7}**
PNO1	2:68152873-68177959	2p14	232	6.47×10^{-5}	287	7.68×10^{-3}	260	**3.82×10^{-6}**	214	1.38×10^{-5}	275	1.89×10^{-2}	245	**2.51×10^{-6}**
LOC105373764	2:177697404-177730138	2q31.2	282	9.43×10^{-3}	285	8.92×10^{-6}	284	**2.15×10^{-6}**	267	1.06×10^{-2}	271	2.47×10^{-5}	269	5.99×10^{-6}
SRGAP3	3:8978591-9367929	3p25.3	2748	1.43×10^{-5}	3836	1.47×10^{-2}	3292	**2.32×10^{-6}**	2586	2.60×10^{-5}	3663	1.50×10^{-2}	3125	**2.57×10^{-6}**
MED23	6:131571966-131633239	6q23.2	394	9.24×10^{-5}	546	2.42×10^{-4}	470	**1.60×10^{-7}**	355	2.11×10^{-4}	500	7.72×10^{-4}	428	**6.44×10^{-7}**
ENPP3	6:131632235-131749413	6q23.2	633	3.59×10^{-5}	990	7.95×10^{-3}	812	**2.39×10^{-6}**	587	1.17×10^{-4}	929	1.49×10^{-2}	758	1.42×10^{-5}
RIPK2	8:89752733-89793064	8q21.3	253	**8.17×10^{-8}**	330	1.49×10^{-1}	292	**1.92×10^{-6}**	235	**4.26×10^{-8}**	308	3.00×10^{-1}	272	**1.87×10^{-6}**
OSGIN2	8:89896868-89929868	8q21.3	215	**2.15×10^{-6}**	287	3.43×10^{-2}	251	**1.63×10^{-6}**	189	2.43×10^{-5}	261	3.02×10^{-3}	225	**1.12×10^{-6}**
LOC105375678	8:101209031-101323227	8q22.3	773	1.09×10^{-4}	1249	2.05×10^{-3}	1011	**1.49×10^{-6}**	715	1.23×10^{-4}	1183	2.89×10^{-3}	949	2.93×10^{-6}
PARD3	10:34107560-34820325	10p11.21	5287	3.97×10^{-6}	6663	1.44×10^{-2}	5975	**8.04×10^{-7}**	4833	8.26×10^{-6}	6246	1.53×10^{-2}	5540	**1.46×10^{-6}**
PYROXD2	10:98381565-98420221	10q24.2	375	2.56×10^{-4}	480	6.60×10^{-4}	428	**1.12×10^{-6}**	358	2.64×10^{-4}	449	2.74×10^{-4}	404	**5.73×10^{-7}**
SCGB1D2	11:62236630-62246812	11q12.3	67	5.09×10^{-4}	113	2.93×10^{-1}	90	2.36×10^{-3}	64	4.16×10^{-3}	97	6.99×10^{-5}	81	**2.08×10^{-6}**
GPC6	13:93221825-94410020	13q31.3	8119	**1.07×10^{-7}**	10966	1.53×10^{-1}	9543	**2.43×10^{-6}**	7517	**2.05×10^{-7}**	10327	1.08×10^{-1}	8922	**1.88×10^{-6}**
TMEM63C	14:77176759-77261495	14q24.3	529	**4.18×10^{-7}**	804	5.85×10^{-2}	667	**1.09×10^{-6}**	513	**1.96×10^{-6}**	722	2.31×10^{-2}	618	3.36×10^{-6}
NUBP2	16:1777923-1791191	16p13.3	164	2.58×10^{-5}	202	1.21×10^{-3}	183	**2.28×10^{-7}**	153	8.84×10^{-5}	198	1.81×10^{-3}	176	**1.94×10^{-6}**
RBFOX1	16:5234782-7715340	16p13.3	31067	**1.40×10^{-8}**	42211	1.90×10^{-1}	36639	**1.01×10^{-6}**	29502	**8.33×10^{-9}**	40406	1.93×10^{-1}	34954	**5.62×10^{-7}**
ARHGAP44	17:12784512-12993643	17p12	1388	3.72×10^{-5}	2071	4.78×10^{-3}	1730	**1.39×10^{-6}**	1264	6.78×10^{-5}	1926	1.00×10^{-2}	1595	4.85×10^{-6}
ZNF624	17:16597870-16660626	17p11.2	404	9.83×10^{-6}	430	5.58×10^{-3}	417	**5.27×10^{-7}**	365	1.29×10^{-4}	398	5.92×10^{-3}	382	4.02×10^{-6}
PTPRT	20:42070752-43194917	20q13.11	8065	**3.67×10^{-7}**	11552	1.57×10^{-2}	9809	**1.37×10^{-7}**	7532	**4.30×10^{-7}**	11021	1.33×10^{-2}	9277	**1.03×10^{-7}**

CHR: Chromosome. START/STOP: Annotation boundaries of the gene on that chromosome in the hg38 genome assembly. Region: Genomic location of the gene. NSNP: Number of rare SNPs annotated to the gene for the gene-based test; for meta-analyses, this reflects the aggregated number of SNPs across both datasets rather than just the overlapping SNPs. *p*-value: Gene-based *p*-value; bold values indicate significance below the Bonferroni-corrected threshold of 2.59×10^{-6}.

Gene Set Enrichment Analysis Links Significant Genes to Biological Traits and Pathways

I uploaded the lists of genes that were significantly associated in annotation-unweighted and annotation-weighted meta-analysis steps (**Figure 9a**) to the "Enrichr" web-based tool and subsequently downloaded the results for each library. Notably, both gene-set enrichment analyses (**Figure 9b,c**) highlighted the following: the pathway 'CagA phosphorylation-independent signaling' with enrichment of the genes *RIPK2* and *PARD3*; the trait 'body mass index' with the genes *PTPRT*, *MED23*, *RBFOX1*, *PARD3*, *ENPP3* and *GPC6*; the trait 'general cognitive ability' with the genes *PTPRT*, *RBFOX1*, *PARD3*, and *PYROXD2*; and the 'arginine levels' enriched with the genes *MED23* and *PYROXD2*.

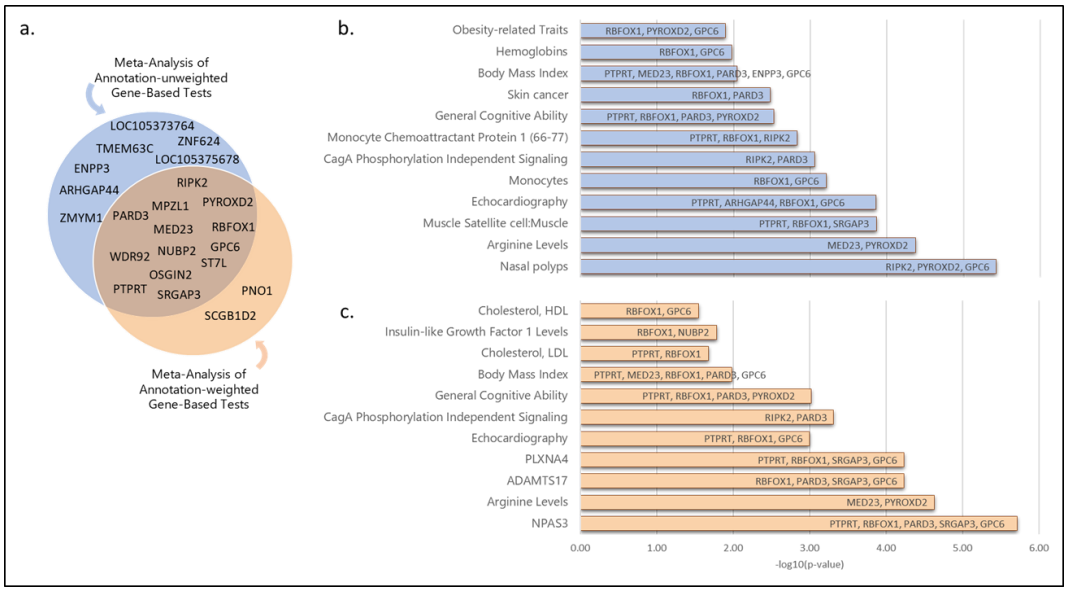


Figure 9. Shared genes and enriched pathways from Gene-Based Meta-Analyses.
(a) The Venn diagram illustrates the genes that were significant in both meta-analyses. The bar plots show top significant results (p-value < 0.05) from gene-set enrichment analyses, based on genes identified in the annotation-unweighted (b) and annotation-weighted (c) gene-based meta-analyses. The x-axis represents the -log10(p-value) for relevant pathways, ontologies, diseases, or cell lines, with the corresponding genes labeled on the bars.

Gene-Centric Analysis Identifies Genes Harboring Rare Coding and Noncoding Variants and Associated with PDAC Risk

In the PanScan-PanC4 dataset, the *RIPK2* gene with four synonymous rare variants showed a strong association (p-value = 4.77×10^{-8}) with the disease risk. Additionally, the *MEP1B* gene with five synonymous rare variants became significant with a p-value of 4.91×10^{-7}, which was just below the significance threshold (**Figure 10a,b**).

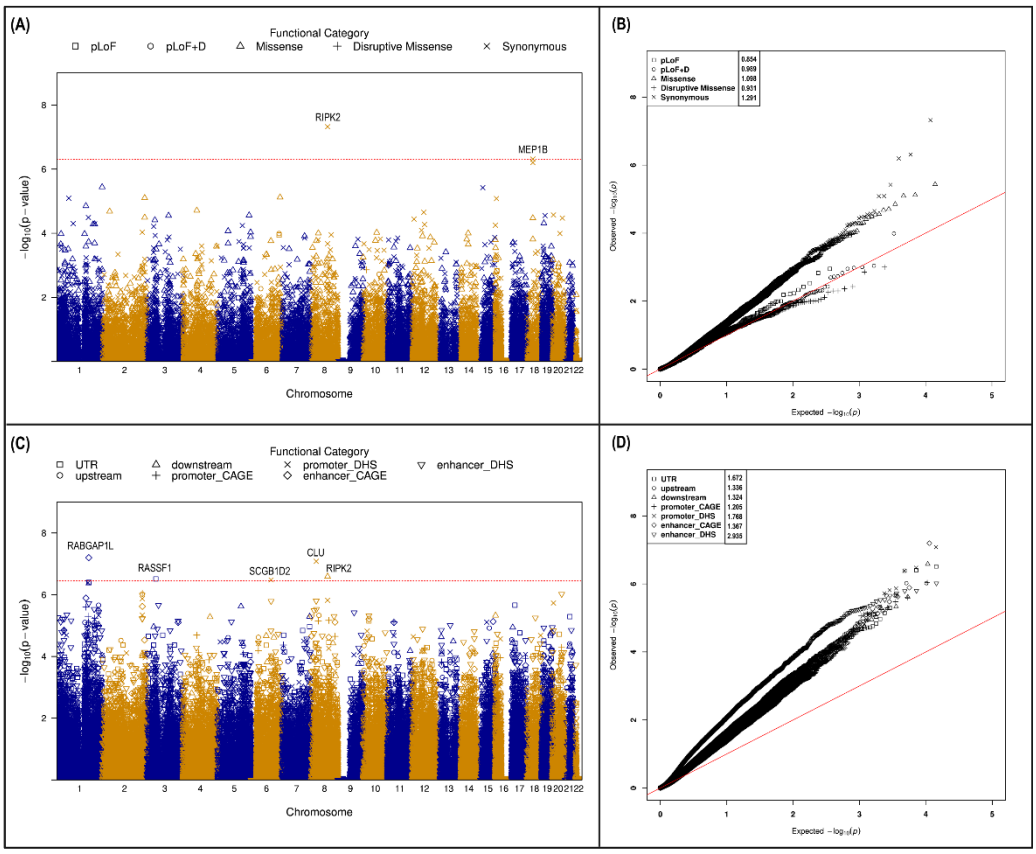


Figure 10. Gene-Centric Manhattan and QQ Plots for Rare Variants in PanScan-PanC4.

Manhattan plot (A) and QQplot (B) of the gene-centric analysis for coding rare variants on each functional category with a *p*-value threshold of 5.0×10^{-7}. Manhattan plot (C) and QQplot (D) of the gene-centric analysis for noncoding rare variants on each functional category with a *p*-value threshold of 2.5×10^{-6}. The inflation scores (lambda) for each category are reported in the QQ plots next to the category indicators.

The gene-centric analysis for noncoding functional categories (UTR, upstream, downstream, promoter_CAGE, promoter_DHS, enhancer_CAGE, and enhancer_DHS) by aggregating rare variants revealed 5 protein-coding genes in the PanScan-PanC4 dataset: the *RABGAP1L* gene with 20 rare variants in the enhancer region (p-value = 6.33×10^{-8}), the *RASSF1* gene with 7 rare variants in the UTR (p-value = 3.09×10^{-7}), the *CLU* gene with 35 rare variants in the promoter region (p-value = 8.27×10^{-8}), the *SLC22A16* gene with 35 rare variants in the promoter region (p-value = 3.34×10^{-7}), the *RIPK2* gene with 4 rare variants in the downstream region of the gene (p-value = 2.59×10^{-7}) (**Figure 10c,d**).

In addition to noncoding functional categories, I examined noncoding RNA gene regions. Three RNA genes showed significance in PanScan-PanC4: the lncRNA gene *CADM3-AS1* with 29 rare variants (p-value = 5.37×10^{-7}), the *RNU6-307P* gene (pseudogene) with two rare variants (p-value = 4.27×10^{-8}), and the gene *LINC00326* with 56 rare variants (p-value = 3.67×10^{-7}) (**Table 11**).

In the UK Biobank dataset, the gene-centric analysis of coding rare variants indicated four significantly associated genes in functional categories: *ZNF79* with six missense rare variants (p-value = 3.05×10^{-7}), *C10orf88* with four synonymous rare variants (p-value = 2.10×10^{-9}), and *CCDC81* with four synonymous rare variants (2.59×10^{-9}) (**Figure 11a,b**). Similarly, seven genes were significant in the gene-centric analysis of rare noncoding variants in the UK Biobank (**Figure 11c,d**). Four of those genes showed significance in different noncoding functional categories. For example, the *SEPTIN8* gene with 13 rare enhancer variants were determined via the CAGE method (p-value = 5.14×10^{-13}), and with 74 rare enhancer variants were determined via the DHS method (p-value = 3.59×10^{-12}), and with 39 rare UTR variants (p-value = 2.75×10^{-12}). Additionally, **Supplementary Table 4** presents the inflation factors for each category of gene-centric analysis displayed in the QQ plots in **Figure 10b,d** and **Figure 11b,d**.

Table 11 presents a comprehensive overview of all the significant results from the gene-centric coding and noncoding analyses across various functional categories and the top significantly associated rare variants in those gene regions. Notably, this project uncovered numerous genes with significant p-values across different functional categories, without overlapping significant genes between datasets.

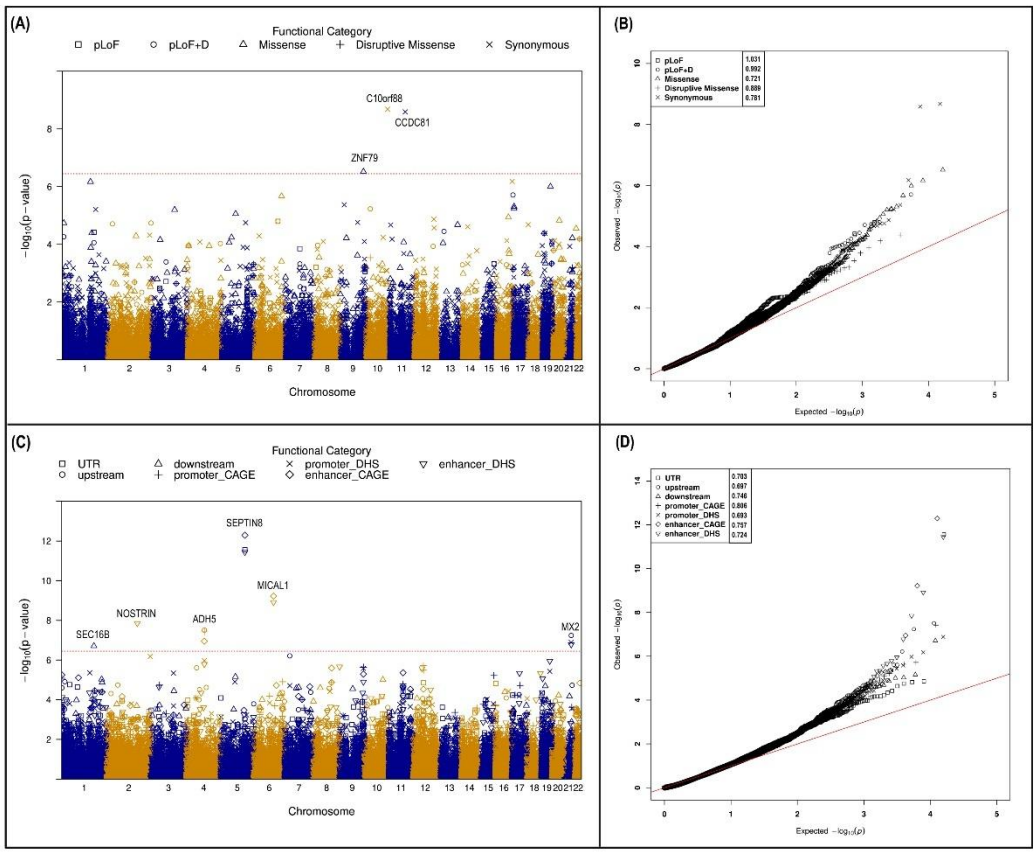


Figure 11. Gene-Centric Manhattan and QQ Plots for Rare Variants in UK Biobank.

Manhattan plot (A) and QQplot (B) of the gene-centric analysis for coding rare variants on each functional category with a *p*-value threshold is 5.0×10^{-7}. Manhattan plot (C) and QQplot (D) of the gene-centric analysis for noncoding rare variants on each functional category with a *p*-value threshold of 2.5×10^{-6}. The inflation scores (lambda) for each category are reported in the QQ plots next to the category indicators.

Table 11. All statistically significant results from gene-centric coding and gene-centric noncoding analysis for functional categories.

GENE	CHR	Category	NSNP	STAAR-O	Dataset	Index SNP	*p*-value	Allele	MAF
SEC16B	1	downstream	2	1.94×10^{-7}	UK Biobank	rs187848861	1.73×10^{-6}	C	0.0015
RABGAP1L	1	enhancer_CAGE	20	6.33×10^{-8}	PanScan-PanC4	rs12072529	2.10×10^{-8}	C	0.0075
CADM3-AS1	1	ncRNA	29	5.37×10^{-7}	PanScan-PanC4	rs2746046	2.21×10^{-7}	C	0.0085
RNU6-307P	1	ncRNA	2	4.27×10^{-8}	PanScan-PanC4	rs74128530	5.32×10^{-8}	G	0.0071
NOSTRIN	2	enhancer_DHS	66	1.42×10^{-8}	UK Biobank	rs202204045	2.37×10^{-10}	T	0.0014
AC012499.1	2	ncRNA	14	1.77×10^{-7}	UK Biobank	rs73979531	4.43×10^{-9}	T	0.0011
RASSF1	3	UTR	7	3.09×10^{-7}	PanScan-PanC4	rs116422509	6.32×10^{-5}	C	0.0027
ADH5	4	upstream	3	3.10×10^{-8}	UK Biobank	rs11568816	1.73×10^{-8}	A	0.0006
ADH5	4	promoter_CAGE	5	3.77×10^{-8}	UK Biobank	rs11568816	1.73×10^{-8}	A	0.0006
ADH5	4	enhancer_CAGE	8	1.12×10^{-7}	UK Biobank	rs11568816	1.73×10^{-8}	A	0.0006
SEPTIN8	5	enhancer_CAGE	13	5.14×10^{-13}	UK Biobank	rs190772633	1.72×10^{-14}	C	0.0007
SEPTIN8	5	UTR	39	2.75×10^{-12}	UK Biobank	rs190772633	1.72×10^{-14}	C	0.0007
SEPTIN8	5	enhancer_DHS	74	3.59×10^{-12}	UK Biobank	rs190772633	1.72×10^{-14}	C	0.0007
MICAL1	6	enhancer_CAGE	22	6.07×10^{-10}	UK Biobank	rs770823329	7.41×10^{-12}	A	0.0010
MICAL1	6	enhancer_DHS	57	1.22×10^{-9}	UK Biobank	rs770823329	7.41×10^{-12}	A	0.0010
SLC22A16	6	promoter_DHS	14	3.34×10^{-7}	PanScan-PanC4	rs7747400	6.61×10^{-8}	A	0.0060
LINC00326	6	ncRNA	56	3.67×10^{-7}	PanScan-PanC4	rs9483534	1.58×10^{-6}	A	0.0055
RIPK2	8	synonymous	4	4.77×10^{-8}	PanScan-PanC4	rs16900617	1.39×10^{-7}	G	0.0046
RIPK2	8	downstream	4	2.59×10^{-7}	PanScan-PanC4	rs10504883	1.39×10^{-7}	T	0.0046
CLU	8	promoter_DHS	35	8.27×10^{-8}	PanScan-PanC4	rs545243	1.78×10^{-6}	T	0.0026
ZNF79	9	missense	6	3.05×10^{-7}	UK Biobank	rs551465890	4.81×10^{-6}	G	0.0002
C10orf88	10	synonymous	4	2.10×10^{-9}	UK Biobank	rs144056903	1.46×10^{-7}	C	0.0006
CCDC81	11	synonymous	4	2.59×10^{-9}	UK Biobank	rs149825851	4.28×10^{-10}	A	0.0003
MEP1B	18	synonymous	5	4.91×10^{-7}	PanScan-PanC4	rs112703073	1.33×10^{-7}	T	0.0024
MX2	21	upstream	17	5.81×10^{-8}	UK Biobank	rs551514117	1.65×10^{-9}	G	0.0008
MX2	21	promoter_DHS	35	1.30×10^{-7}	UK Biobank	rs551514117	1.65×10^{-9}	G	0.0008
MX2	21	enhancer_DHS	49	1.67×10^{-7}	UK Biobank	rs551514117	1.65×10^{-9}	G	0.0008

CHR: Chromosome. Category: Functional category. NSNP: Number of variants from the given category in the gene. cMAC: Cumulative minor allele count. STAAR-O: *p*-values from functionally annotated set-based tests. Index SNP: Most significant variant from the gene and category. *p*-value: Functional annotation-weighted score test *p*-value of the rare variant. Allele: Alternative allele of the SNP, representing the effect allele. MAF: Minor allele frequency of the SNP in the dataset.

Sliding Window Analysis Detects Additional Rare Variant Enrichment Signals in Genomic Regions

Additionally, a dynamic window analysis considering only rare variants revealed seven regions across the genome that reached study-wide significance (**Table 12**). In the UK Biobank, the region chr2:178898793-178910108, overlapping with the lncRNA *ENSG00000287149* gene, with 80 rare variants is strongly associated with the set-test *p*-value = 2.08×10^{-8}. Similarly, the genomic region chr12:90046377-90050704, which falls within the RNA gene *ATP2B1-AS1* and contains 40 rare variants, had a *p*-value of 3.33×10^{-14}.

For PanScan-PanC4, the region chr17:63491605-63504266, covering the whole *ACE1* protein coding gene and partially the *ENSG00000264813* novel protein coding gene, which contains 100 rare variants in the set-test, had *p*-value = 1.78×10^{-10}. Another study-wise significant region was chr8:89764416-89772302 with 50 rare variants in the set. The *RIPK2* protein-coding gene falls within the coordinates of this genomic window. The set-test *p*-value was 1.48×10^{-9}.

Table 12. The all-significant dynamic window analysis results from PanScan-PanC4 and UK Biobank.

CHR	START	STOP	Region	NSNP	Burden	SKAT	STAAR_O	DATASET
2	178898793	178910108	2q31.2	80	3.16×10^{-6}	1.10×10^{-2}	2.08×10^{-8}	UK Biobank
7	82569894	82603847	7q21.11	300	2.79×10^{-9}	3.28×10^{-6}	9.28×10^{-9}	UK Biobank
8	89764416	89772302	8q21.3	50	4.86×10^{-10}	6.02×10^{-8}	1.48×10^{-9}	PanScan-PanC4
12	90046377	90050704	12q21.33	40	3.96×10^{-3}	1.91×10^{-14}	3.33×10^{-14}	UK Biobank
13	66066629	66115570	13q21.32	280	2.61×10^{-6}	4.34×10^{-9}	1.70×10^{-8}	PanScan-PanC4
17	63478430	63483943	17q23.3	50	8.81×10^{-4}	6.44×10^{-9}	1.51×10^{-8}	PanScan-PanC4
17	63491605	63504266	17q23.3	100	2.35×10^{-10}	1.39×10^{-8}	1.78×10^{-10}	PanScan-PanC4

CHR: Chromosome. START: Start position for the gene window in the hg38 genome assembly. STOP: End position for the gene window in the hg38 genome assembly. Region: Chromosome bands. NSNP: Number of rare variants in the gene window. Burden, SKAT, STAAR-O: *p*-values from set-based tests.

5.4 DISCUSSION

Previous GWASs on PDAC risk have focused primarily on common variants, identifying numerous risk loci that individually contribute modestly to disease risk. This part of the thesis extends these findings by emphasizing the role of rare variants (defined as $0.01 \geq \text{MAF} \geq 1 \times 10^{-5}$), which can have substantial effects on disease susceptibility. This finding is consistent with the growing body of evidence suggesting that rare variants are crucial in understanding the genetic basis of complex diseases such as PDAC(Bomba et al., 2017; Momozawa & Mizukami, 2021).

Gene-based analyses focusing on rare variants have yielded significant findings in studies involving patients with a familial history of pancreatic cancer(Earl et al., 2020; Yu et al., 2022). However, these results differ from those of case-control studies involving sporadic PDAC patients. Studies have investigated the presence of rare germline variants, particularly pathogenic variants, in known pancreatic cancer-predisposing genes (such as *BRCA1/2*, *ATM*, *CDKN2A*, *PALB2*, and *MSH6*) among sporadic PDAC cases. Shindo and colleagues reported individual deleterious variants in 3.9% of 854 patients without familial pancreatic cancer history for those genes(Shindo et al., 2017). However, in terms of rare variant association studies, no pancreatic cancer-predisposing genes showed exome-wide significant associations (*p*-value $< 2.5 \times 10^{-6}$) in gene-based analyses in sporadic cases, to date.

This project differs from previous research by not limiting it to only pathogenic variants. Instead, I included all rare variants within gene regions, regardless of their predicted impact, and rare variants in other functional categories than pathogenicity. Additionally, while previous studies often focused on rare variants in known pancreatic cancer-predisposing genes, I applied gene-based analyses across the exome for all protein-coding and noncoding genes.

I applied the burden test to account for all rare variants in gene regions, and the STAAR-O test for rare variant sets categorized by function within the gene region. The burden test assumes that all variants are causal and contribute to the phenotype in the same direction, which is a limitation when there are variants with opposite effect directions(Neale et al., 2011). Nevertheless, the burden test normalizes the number of rare variants across genes by comparing a score between cases and controls. STAAR-O combines the *p*-values from the SKAT-O, burden, and ACAT-V tests using the Cauchy method, which enhances the power to detect associations in gene-based analyses when variants with both directions are present(S. Lee et al., 2014b). These combined approaches increase the statistical power to detect complex genetic associations, especially in the context of rare variants with heterogeneous effects, thereby improving the robustness and accuracy of gene-based association studies.

The characteristics of the two study populations appear to have influenced the results. While PanScan-PanC4's diverse ancestry panel captured a broader genetic spectrum, enhancing gene-level association detection, the UK Biobank's European-only (specifically, British-only for the vast majority of cases and controls) focus may have improved the power to detect rare variants specific to that population.

However, the high inflation scores observed in PanScan-PanC4 association analyses likely come from the inherent higher heterogeneity of the dataset, where cases and controls, although all of European ancestry, were collected from different countries and genotyped with different arrays. This heterogeneity, along with imputation using diverse ancestry panels and then merging of datasets, can introduce confounding due to population structure, leading to an overestimation of associations (**Supplementary Table 4** and **Supplementary Figure 1**). In this regard, more homogeneous datasets like the UK Biobank may be better suited for rare variant association analyses, as they reduce potential confounding. Importantly, the significant variants identified in the UK Biobank dataset had high info scores for those that were imputed, supporting their reliability.

Even if all technical challenges were resolved, such as genotyping all study participants with the same array or sequencing them using whole-genome sequencing and imputing to population-specific reference panels, rare variant analyses would still face an inherent biological limitation. Rare variants are often highly population-specific, meaning that variants strongly associated with disease risk in one country or region may simply not be present in other populations. This limits their utility for risk prediction and replication in diverse cohorts. The lack of high-quality whole-genome sequencing data across many global populations further complicates our ability to evaluate the degree of rare variant sharing. Thus, while homogeneous datasets can reduce technical variability, they also expose a paradox: rare variants, despite their potential strong effects, often lack generalizability beyond the populations in which they are discovered.

It is advantageous to identify potential genes using various gene-based methods. Further evaluating the clinical implications of these genes, such as analyzing their differential expression or impact on survival in PDAC patients, could significantly enhance the value of my findings. In the meta-analysis step, 15 genes - *PTPRT*, *MPZL1*, *WDR92*, *RBFOX1*, *PYROXD2*, *MED23*, *ST7L*, *OSGIN2*, *PARD3*, *RIPK2*, *GPC6*, *NUBP2*, *SCGB1D2*, *PNO1* and *SRGAP3* - in both gene-based analyses showed strong associations (**Figure 9a**). While the overlap of 15 genes between methods appears limited, it reflects the distinct aspects of genetic associations captured by each method. This emphasizes the need for diverse approaches to uncover the full spectrum of genetic contributors to PDAC risk. One of the limitations of this project is that the MAGMA tool only applies the fixed effect model for meta-analysis and provides no

information on the heterogeneity and effect size direction; therefore, I was unable to rerun the meta-analyses for these genes with a random effects model.

The *PARD3* gene reached study-wide significance in both meta-analyses approaches with around 6000 rare variants. *PARD3* encodes an adapter protein involved in asymmetrical cell division and cell polarization processes. It is involved in signaling via the transforming growth factor (TGF)-β receptor complex pathway. This pathway plays an important role in pancreatic cancer through its tumor-suppressive role in non-metastatic PDAC and its tumor-promoting role in metastatic PDAC(Hussain et al., 2021; Principe et al., 2021). In addition, genomic alterations in the PARD3 gene are reportedly associated with PDAC risk(Iorio et al., 2018) (Open Target Platform Database).

Mutations in the *PTPRT* gene have been observed in a variety of cancers, including colon, bladder, lung, and stomach cancers. Some of these *PTPRT* mutations are insertion, deletion and nonsense mutations that result in premature truncation of the protein (COSMIC database)(Z. Wang et al., 2004). In addition, two somatic *PTPRT* truncating mutations were reported in tumor tissues from two PDAC patients in the PACA-CA (Pancreatic Adenocarcinoma from ICGC Data Portal) cohort(J. Zhang et al., 2011). The reported mutations are chr20:42115303:G>A (rs763964405) and chr20:42771530:G>A, which cause a stop codon and therefore loss-of-function of the PTPRT protein. The gene encodes an important member of the JAK/STAT signaling regulators, a protein tyrosine phosphatase (PTP). PTPs are signaling molecules that regulate cellular processes, including cell growth, differentiation, the mitotic cycle, and oncogenic transformation.

The strong association of the *RIPK2* gene across multiple analyses (annotation-unweighted and annotation-weighted gene-based, gene-centric coding and gene-centric noncoding) is particularly noteworthy. RIPK2 functions as a key effector of the NOD1 and NOD2 signaling pathways in the inflammatory response and has been implicated in various cancers by promoting the disease progression(Jiao et al., 2023; Li et al., 2021; Yan et al., 2022; W. Zhang & Wang, 2022). Recently, Sang and colleagues reported that inhibition of RIPK2 leads to prolonged survival or complete regression of PDAC by sensitizing PDAC to anti-programmed cell death protein 1 (anti–PD–1) immunotherapy(Sang et al., 2024). My findings align with previous studies that highlighted the role of inflammation in PDAC pathogenesis.

The findings suggest that rare variations in the regulatory regions of these important proteins could affect their expression levels by changing their transcription factor binding affinity. Therefore, these rare variants might contribute to the disease progression. The same hypothesis is valid for the genes *RABGAP1L*, *NOSTRIN*, *ADH5*, *SEPTIN8* and *MX2*, because of several significant regulatory regions such as upstream, enhancer, and promoter regions (**Table 11**). In the UTR and enhancer

region of *SEPTIN8*, while rs190772633-C had a strong association p-value as 1.72×10^{-14} and a 1.77-fold increased PDAC risk in UK Biobank dataset.

NOSTRIN binds the endothelial nitric oxide (NO) synthase enzyme, which is responsible for NO production, and inhibits NO production. NO is a mediator of inflammatory response, neurotransmission and vascular homeostasis. Many studies have shown the role of NO in disease progression, such as type 2 diabetes, various cancers, and cardiovascular diseases(Bahadoran et al., 2019, p. 2; Sahebnasagh et al., 2022; Somasundaram et al., 2019). Moreover, one study suggested that the overexpression of *NOSTRIN* suppressed the migration and invasion of PDAC cells, and that lower *NOSTRIN* expression was associated with worse survival(J. Wang et al., 2016).

Additionally, I observed that one of the significant genes with rare variants in its promoter region, *CLU*, was highly expressed in pancreatic ductal cells (**Figure 12**). The *CLU* promoter (chr8:27,612,623-27,616,913) contains six non-genic regulatory elements (enhancers and silencers) and binding sites for 248 transcription factors, including *SMAD4*, *SP1*, *STAT3* and *KLF6*, making it a highly interactive regulatory hub (according to GeneCards).

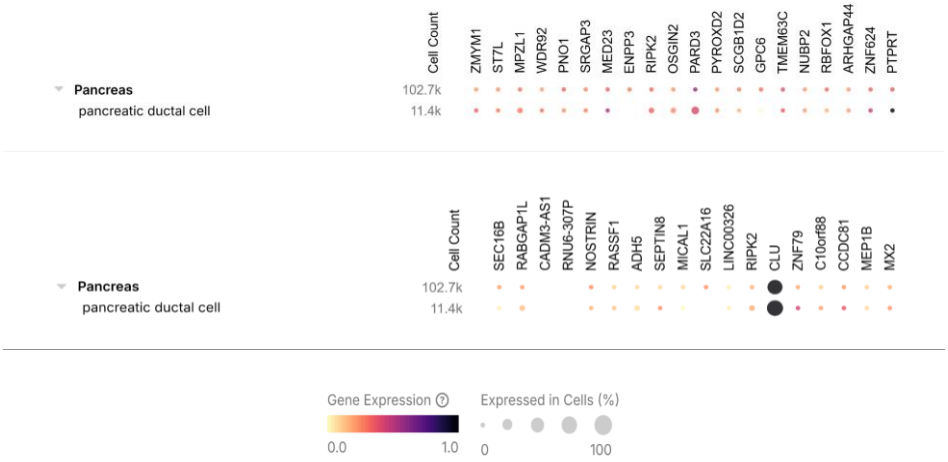


Figure 12. Expression of Associated Genes in Pancreatic Cells

The figure (adjusted by CZ CELLxGENE Discover data source). Dot plots show expression levels of significantly associated genes (from gene-based analysis) in pancreatic ductal cells and whole pancreas, based on Gene Expression - CZ CELLxGENE Discover data source. Dot color indicates average expression; size shows the percentage of cells expressing the gene. Two color scales are used: unscaled (range 0-8, comparable across plots) and scaled (0-1, not comparable across plots).

In the literature, *CLU*, a gene involved in the cholesterol pathway, has been proposed as a novel therapeutic target due to its overexpression in PDAC cells and its role in promoting cell proliferation. *CLU*, frequently upregulated in various cancers, acts as a

stress-response protein that promotes tumor survival and therapy resistance. Custirsen (OGX-011) is a second-generation antisense oligonucleotide designed to bind *CLU* mRNA, block its translation, and thereby prevent clusterin accumulation—restoring cancer cell sensitivity to treatment and overcoming a major mechanism of therapeutic resistance(Mitsufuji et al., 2022). One study reported immunohistochemically evidence of *CLU* overexpression in surgically resected PDAC tissues in approximately half of the cases examined(Amada et al., 2023).

Mitsufuji et al. demonstrated that inhibiting CLU induces cellular senescence in PDAC cells. Similarly, Tang et al. proposed that gemcitabine treatment increases chemoresistance through the upregulation of secretory clusterin (sCLU). They showed that inhibiting clusterin enhances the chemosensitivity of pancreatic cancer cells by blocking clusterin-mediated activation of pERK1/2. Their study concluded that knockdown of clusterin via OGX-011 transfection sensitizes PDAC cells to gemcitabine by preventing the gemcitabine-induced activation of the clusterin-pERK1/2 pathway(Tang et al., 2012). My findings suggest that rare variants in the promoter region of the *CLU* gene may contribute to its overexpression and thereby increase the risk of PDAC.

This project demonstrated the utility of imputation and large reference panels in uncovering rare variant associations and identified out several candidate genes associated with PDAC predisposition. Despite both datasets being of European ancestry, differences in rare variant allele frequencies across populations highlight the need for caution when interpreting results.

Additionally, the functional impact of genes remains to be elucidated. The diversity of sequencing/genotyping and imputation methods for rare variants add more complexity to rare variant studies, despite growing technological and methodological advances. The imputation of very rare variants can be imprecise, which may affect the accuracy of my results. Consequently, my findings should be interpreted with caution. Further studies, including functional assays and replication in independent cohorts, are necessary to validate my findings and understand the biological mechanisms underlying these associations.

In conclusion, my study highlights the significant contribution of rare variants to PDAC risk. While individual rare variants that were imputed with high confidence (INFO score > 0.99), such as rs12818275(C), rs12987101(A), and rs560926254(T), which showed strong association in score-test with PDAC risk, the genes *PARD3*, *RIPK2*, *PTPRT*, *NOSTRIN*, *SEPTIN8*, *CCDC81*, *CLU* and *SCGB1D2* were suggested as strong candidates by functional annotation-weighted gene-based analysis. The identification of associated genes provides new avenues for understanding the genetic basis of PDAC and potentially for developing targeted interventions. Future research should continue to leverage large-scale genomic data and advanced analytical methods to uncover the full spectrum of genetic factors involved in PDAC.

6 PROJECT 3: COGNITION GENES AS PDAC RISK FACTORS

6.1 INTRODUCTION

Pancreatic functions are regulated by the autonomic nervous system (ANS), particularly its sympathetic and parasympathetic branches, which adjust exocrine and endocrine secretions in response to physiological and environmental stimuli(T. B. and R. A. Travagli, 2016). Beyond its role in digestion, the ANS is a key component of the neuroendocrine system, facilitating communication between the brain and peripheral organs under conditions of psychological or physiological stress. Chronic activation of this axis has been implicated in immune modulation, inflammation, and tumor progression via catecholamine and glucocorticoid signaling(Antoni et al., 2006; Cole et al., 2015).

Although popular belief links negative thoughts and emotions to cancer risk, direct scientific evidence for causality is limited. However, persistent stress, anxiety, and depression are associated with cognitive dysfunction and alterations in brain structure and connectivity(McEwen & Gianaros, 2011). Cognition includes a range of mental processes such as attention, memory, decision-making, and problem-solving, and is supported by dynamic neural activity in regions such as the frontoparietal cortex, hippocampus, and amygdala. These areas are also central to top-down regulation of emotional and physiological states.

Genetic influences on cognition are well established, with twin and family-based studies estimating that 50-80% of the variance in cognitive ability is heritable(Davies

et al., 2011). Individual differences in cognitive processing and emotional regulation, shaped in part by this genetic variation, can affect systemic stress responses, potentially contributing to disease susceptibility.

Importantly, PDAC is one of the few malignancies in which perineural invasion (PNI) is not just common but clinically significant. PNI refers to the infiltration of cancer cells along nerve fibers and is associated with increased pain, early recurrence, and reduced survival(Ceyhan et al., 2008). It reflects a unique neural-tumor microenvironment in PDAC, characterized by intense nerve remodelling and reciprocal signaling between cancer cells and the peripheral nervous system. This neuroplasticity suggests that neural factors—potentially influenced by cognition-related genes and stress pathways—could modulate tumor behavior in a tissue-specific manner(Demir et al., 2010; Magnon, 2015).

In one study, 267 genes (258 of them autosomal) were identified as cognition-related and unique to modern humans(Zwir et al., 2022). These genes are mostly non-coding and are thought to contribute to higher-order traits such as abstract reasoning, creativity, and prosocial behavior. For clarity, I refer to this group as the cognitive gene network, defined by functional annotations, brain expression, and evolutionary signatures.

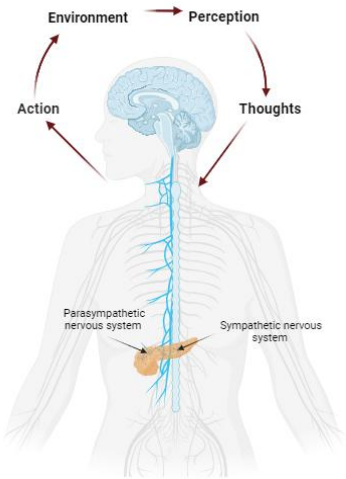


Figure 13. Hypothetical Model of Cognitive Regulation of Pancreatic Function via the Autonomic Nervous System

Given the ANS involvement in pancreatic physiology, the prevalence of perineural invasion in PDAC, and the known impact of cognition on stress reactivity, I hypothesize that genetic variations in cognition-related genes may influence

susceptibility to pancreatic cancer. This hypothesis is illustrated in **Figure 13**, which proposes that cognitive processes, shaped by perception, thoughts, and environmental stimuli, may influence pancreatic function via ANS pathways. To test this, I analyzed polymorphisms in 258 autosomal cognition-related genes. In addition, I considered the genes involved in related pathways and explored their potential interactions with cognitive gene variants in relation to pancreatic cancer risk.

I focused on genes involved in circadian rhythm regulation, serotonin signaling, and insomnia, as these pathways are mostly expressed in the brain and have well-established roles in cancer biology. Circadian disruption and sleep disturbances can promote inflammation, metabolic dysregulation, and impaired DNA repair, all contributing to tumor development(Huang et al., 2023; Shi et al., 2020). Serotonin signaling, similarly, affects both neural and peripheral systems and has been implicated in cancer progression through its effects on cell proliferation and immune modulation(Jiang et al., 2017; Lin et al., 2025). These pathways were therefore investigated for potential interactions with cognitive gene variants to better understand their combined influence on pancreatic cancer susceptibility.

6.2 METHODS

6.2.1 Compiling Gene List

I obtained the cognition gene list from the published study(Zwir et al., 2022), which identified 267 genes, including 258 located on autosomal chromosomes (**Supplementary Table 5**). Gene locations and annotations were confirmed using the following databases: AngioGenes (http://angiogenes.uni-frankfurt.de/gene), UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgLiftOver), and NCBI Gene (https://www.ncbi.nlm.nih.gov/gene/).

I extracted the gene lists related to serotonin, circadian rhythm, and insomnia from the KEGG and MSigDB databases under the respective terms "Circadian Rhythm," "Serotonin," and "Insomnia." These lists contained 202, 40, and 52 genes, respectively (**Supplementary Table 6**). Gene data were retrieved from the KEGG Pathway database (https://www.genome.jp/kegg/pathway.html), Reactome Pathway database (https://reactome.org/) and Molecular Signatures Database (MSigDB) (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp).

6.2.2 Genetic Data

I used the same post-QC genotype data from two cohorts as previously included in *Project 2: Rare Germline Variants in PDAC*: PanScan-PanC4 and the UK Biobank. This dataset comprised 25,275 individuals – 7,239 PDAC cases and 7,015 controls from PanScan-PanC4, and 1,021 PDAC cases and 10,000 controls from the UK Biobank.

The only difference from Project 2 was that I applied a more stringent imputation INFO score threshold (≥ 0.8 instead of ≥ 0.5) to reduce the total number of SNPs and enrich for rare variants with higher imputation quality. The final QC-passed genotype data were used in downstream analyses including logistic regression, gene-based testing, and epistasis analysis (**Figure 14**).

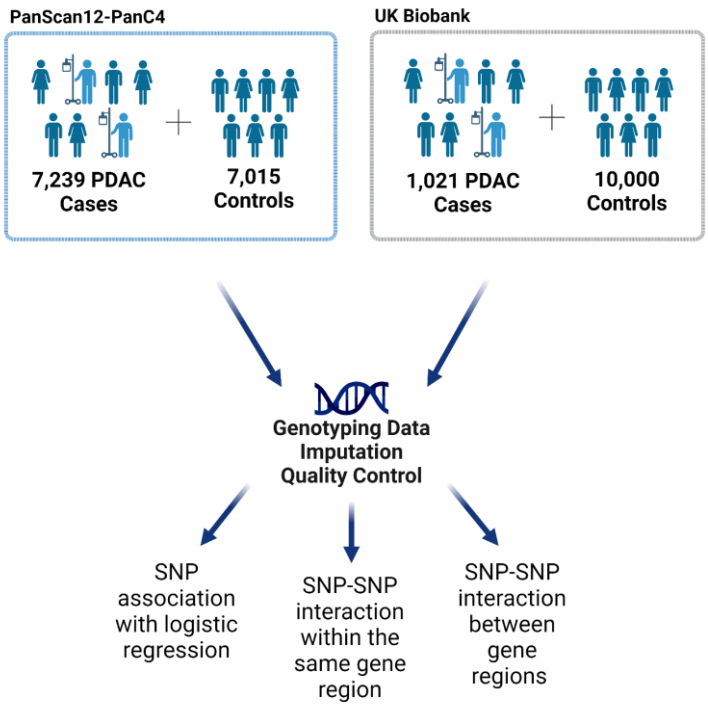


Figure 14. Overview of sample sources, genotype processing pipeline, and downstream analyses.

I defined gene regions as ±5 kb around gene boundaries. For each dataset, I extracted SNPs within these regions using PLINK version 2.0 with the ' *--extract* ' flag. I eventually generated 502 PLINK files, one for each gene region, instead of the expected 552, as 50 gene regions lacked SNPs in their regions in both datasets. I then

pruned variants in high LD by removing SNP pairs with $r^2 > 0.6$, also using PLINK version 2.0. It resulted in 3255 SNPs in PanScan-PanC4 and 3200 SNPs in UK Biobank datasets.

6.2.3 Logistic Regression

To identify associations between individual variants and PDAC, I ran logistic regression on each gene region after LD pruning. I adjusted for age, sex, genotyping array, and the first 10 PCs in both datasets separately. I performed this analysis in PLINK version 2.0 using the flags: ' *--logistic --covar --allow-no-sex* '

Additionally, I applied a meta-analysis between logistic regression results of each region from both dataset by using METAL software.

6.2.4 Gene-Based Analysis

I used the "ARTP2" R package(H. Zhang et al., 2016) to conduct gene-based association testing. For each gene region, I included both rare variants (MAF $> 1 \times 10^{-5}$) and common variants that passed LD pruning. ARTP2 combines *p*-values from individual SNPs using the adaptive rank truncated product (ARTP) method. This approach evaluates the joint association of multiple SNPs within a gene, increasing power to detect cumulative genetic effects.

6.2.5 Interaction (Epistasis) Analysis

SNP-SNP interactions within the same gene region

To explore intragenic interactions, I performed SNP-SNP epistasis tests within each gene region using PLINK. I used the following command for each region:

' *--bfile GENE1 --epistasis --epi1 0.001* '

The ' *--epi1* ' flag sets the *p*-value threshold for reporting significant SNP pairs.

SNP-SNP interactions between gene groups

To evaluate intergenic interactions, I first LD-pruned the full genotype data in both cohorts. I then created gene sets using the ' *--make-set* ' flag in PLINK version 1.9. I

grouped variants based on biological relevance into predefined sets: cognition network, circadian rhythm, serotonin, and insomnia-related genes.

I tested epistasis between these sets using the set-by-set epistasis function in PLINK version 1.9. Example command:
‘ *--epistasis set-by-set --set Cognition.set --epi 0.001* ‘

Additionally, I created all possible pairwise combinations of gene sets (e.g., Cognition × Serotonin, Cognition × Circadian Rhythm, Serotonin × Insomnia) by assigning group labels and generating combination sets via the ‘ *--make-set* ‘ flag. I then ran epistasis tests on these combined sets using:
‘ *--epistasis set-by-set --set Cognition×Serotonin.set --epi 0.001* ‘

6.2.6 Permutation Test

To evaluate the robustness of the epistasis results, I performed only one-time permutation test for each dataset due to computational constraints and the large number of tests performed. I randomly shuffled the case-control labels using the *sample()* function in R and repeated the epistasis analysis on the permuted data.

6.2.7 Firth Logistic Regression

I selected only SNP pairs that passed a Bonferroni-corrected significance threshold from epistasis results. Bonferroni-corrected significance thresholds were calculated as 0.05/number of interactions. I used genotype data in PLINK .raw format, where hard-called dosages were encoded as 0 (homozygous reference allele), 1 (heterozygous), or 2 (homozygous alternative allele). I re-coded the binary phenotype from PLINK convention (1 = control, 2 = case) to 0 and 1 respectively, to match logistic regression input. For each significant pair, I generated a 3×3 genotype combination table to examine the joint genotype distribution across cases and controls. I applied a sparsity filter to flag SNP pairs with fewer than five individuals in any genotype combination; although I retained these pairs for modeling, I marked them as potentially unstable.

To obtain more robust and bias-reduced effect size estimates, I applied Firth's penalized likelihood logistic regression using the "logistf" package in R. I defined the interaction term as a categorical variable representing the combined genotypes of both SNPs. I calculated odds ratios by exponentiating the model coefficients and derived *p*-values using penalized profile likelihood tests.

To identify case-exclusive SNP-SNP genotype combinations, I first generated combined genotype terms for each significant SNP pair by concatenating the individual genotypes (coded as 0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternate) from each SNP into a single interaction string (e.g., 0_2). I applied this procedure separately to the PanScan-PanC4 and UK Biobank datasets. I then compared the genotype combinations observed in cases versus controls and retained only those combinations that were present in at least one case but completely absent in all controls. For each case-exclusive combination, I constructed a 2×2 contingency table and performed a one-sided Fisher's exact test to evaluate whether the genotype combination was statistically enriched in cases by using R. I recorded the number of individuals carrying each combination and the corresponding *p*-values for all SNP pairs with case-exclusive genotype patterns.

6.2.8 Epistasis Test with Covariate Adjustment

I also performed epistasis tests within and between gene groups while adjusting for covariates, including age, sex, genotyping array, and the first 10 PCs, in both the PanScan-PanC4 and UK Biobank datasets. For this analysis, I used the CASSI software package. To reduce computational burden, I first extracted SNPs located within ±5 kilobases of the genes belonging to the cognition, circadian rhythm, serotonin, and insomnia-related gene groups. I then conducted pairwise SNP-SNP interaction tests within and across these predefined gene sets using CASSI's logistic regression framework with covariate adjustment.

6.3 RESULTS

I identified 3255 SNPs from PanScan-PanC4 and 3200 SNPs from UK Biobank across 502 gene regions. Only 950 SNPs overlapped between the two datasets. After applying logistic regression in both datasets, I performed a meta-analysis on these 950 mutual SNPs, which did not result in any genome-wide significant association. The most significant variant was rs2472632 (meta-analysis *p*-value = 2.01×10^{-6}) from the circadian rhythm gene group, which almost reached genome-wide significance in Project 1 (**Table 13**).

Table 13. Meta-analysis results of PanScan-PanC4 and UK Biobank datasets.

rsID	CHR:POS	PanScan-PanC4				UK Biobank				Meta-Analysis				
		EA	BETA	SE	*p*-value	EA	BETA	SE	*p*-value	EA	NEA	Z-score	*p*-value	Direction
rs2472632	11:27374328	A	0.1109	0.02446	5.81×10^{-6}	A	0.09862	0.04834	4.14×10^{-2}	A	C	4.752	2.01×10^{-6}	++
rs3847952	12:111308874	G	0.1007	0.02687	1.78×10^{-4}	G	0.07415	0.0541	1.71×10^{-1}	T	G	-3.721	1.98×10^{-4}	--
rs1337108	1:92753574	T	-0.1313	0.04059	1.22×10^{-3}	T	-0.1518	0.0831	6.78×10^{-2}	T	C	-3.636	2.77×10^{-4}	--
rs1344438	12:111298150	G	0.09886	0.02682	2.28×10^{-4}	G	0.06518	0.0542	2.29×10^{-1}	C	G	-3.565	3.64×10^{-4}	--
rs502396	18:659236	C	0.08554	0.02374	3.14×10^{-4}	C	0.0483	0.04686	3.03×10^{-1}	T	C	-3.387	7.07×10^{-4}	--
rs2972162	3:12383294	C	-0.04679	0.02363	4.77×10^{-2}	C	-0.1284	0.04719	6.52×10^{-3}	T	C	3.283	1.03×10^{-3}	++
rs12470503	2:162840541	C	0.08202	0.024	6.33×10^{-4}	C	0.04307	0.04808	3.70×10^{-1}	A	C	-3.16	1.58×10^{-3}	--
rs3898706	20:12921247	G	-0.17	0.04228	5.80×10^{-5}	G	-0.00222	0.08319	9.79×10^{-1}	A	G	3.042	2.35×10^{-3}	++
rs2595389	2:186669456	G	-0.06151	0.02361	9.20×10^{-3}	G	-0.07464	0.04678	1.11×10^{-1}	T	G	3.01	2.62×10^{-3}	++
rs1963187	10:30378681	A	0.1236	0.05017	1.37×10^{-2}	A	0.1438	0.09589	1.34×10^{-1}	A	T	2.841	4.49×10^{-3}	++
rs146893327	21:33338416	G	-0.07675	0.04715	1.04×10^{-1}	G	-0.2475	0.1025	1.58×10^{-2}	C	G	2.816	4.86×10^{-3}	++
rs6923988	6:128271989	C	-0.09839	0.0302	1.12×10^{-3}	C	-0.03279	0.05939	5.81×10^{-1}	A	C	2.811	4.94×10^{-3}	++

rsID: Reference SNP ID. CHR:POS: Chromosome and position in the hg38 genome assembly. EA: Effect allele. NEA: Non-effect allele. BETA: Regression coefficient. SE: Standard error of BETA. *p*-value: *p*-value of the association. Z-score: Z-statistic. Direction: Direction of the effect allele relative to the reference allele in both datasets; + indicates increased risk, – indicates decreased risk. SNPs are sorted in ascending order of the *p*-value from the meta-analysis.

A total of 502 genes were tested using the ARTP2 method in both PanScan-PanC4 and UK Biobank datasets. At a significance threshold of p-value < 0.05, 52 genes were significant in PanScan-PanC4 and 24 in the UK Biobank dataset. Only two genes showed significant associations in both datasets: *RP11-255M2.3* from the cognition gene group (p-value = 3.44×10^{-2} in PanScan-PanC4; p-value = 1.91×10^{-2} in UK Biobank) and *LGR4* from the circadian rhythm group (p-value = 1.45×10^{-4} in PanScan-PanC4; p-value = 2.76×10^{-2} in UK Biobank) (**Table 14**).

Table 14. The top gene-based association results using ARTP2 method in both datasets.

PanScan-PanC4				UK Biobank			
GENE	CHR	NSNP	p-value	GENE	CHR	NSNP	p-value
LGR4	11	51	1.45×10^{-4}	*RP11-536C5.2*	1	6	3.39×10^{-3}
AC002979.1	12	7	3.70×10^{-4}	*SIX3*	2	10	3.71×10^{-3}
MIR6760	12	7	5.35×10^{-4}	*HMGA1P5*	10	25	5.67×10^{-3}
CNR1	6	46	7.45×10^{-4}	*RP11-33E15.1*	8	5	6.48×10^{-3}
TYMS	18	33	8.55×10^{-4}	*CLIP2*	7	21	7.09×10^{-3}
RP11-647O20.1	14	7	8.60×10^{-4}	*NTRK3*	15	153	7.54×10^{-3}
HNF1B	17	115	1.21×10^{-3}	***RP11-255M2.3***	15	95	1.91×10^{-2}
RP11-255M2.3	15	251	3.44×10^{-2}	***LGR4***	11	29	2.76×10^{-2}

GENE: Gene name. CHR: Chromosome. NSNP: Number of SNPs analyzed. p-value: Gene-level p-value (ARTP2). Genes are sorted in ascending order of p-value. Bolded gene names indicate overlapping significant results in both datasets.

Significant SNP-SNP Interactions Identified Within Gene Regions Across Datasets

In total, I identified 63 significant SNP-SNP interactions in the PanScan-PanC4 dataset and 37 in the UK Biobank dataset that remained significant considering the Bonferroni-corrected significance threshold. These interactions were found in all gene groups:

- Cognition gene network: 19 interactions in PanScan-PanC4 and 9 in UK Biobank
- Circadian rhythm: 41 in PanScan-PanC4 and 18 in UK Biobank
- Serotonin: 3 in PanScan-PanC4 and 4 in UK Biobank
- Insomnia-related: 3 interactions (UK Biobank only)

Several SNP-SNP pairs showed strong and consistent signals across both datasets. In the cognition gene network, three gene regions, *RP11-436D23.1*, *RP11-244M2.1*, and

LINC00907, harbored interactions that were observed in both datasets (see **Table 15**). Notably, all associated SNP pairs showed infinite odds ratios (OR_INT = inf), suggesting complete genotype separation between cases and controls for these interaction combinations.

In the circadian rhythm gene group, SNP-SNP interactions within the genes *KCNH7*, *GSK3B*, *PPARGC1A*, *KCNMA1*, and *NTRK3* were significant in both PanScan-PanC4 and UK Biobank, suggesting robust cross-cohort evidence. Additionally, in the serotonin pathway, SNP-SNP pairs within *CHRM3* and *GRIN2A* showed consistent significance across both datasets, further supporting their potential role in modulating PDAC risk through interaction effects.

Table 15. Statistically significant SNP-SNP interactions within the same gene region from the cognition gene group that were found in both datasets.

SNP-SNP	CHR	Gene	Group	*p*-value		OR_INT	Study
				Interaction	Threshold		
rs571719082 rs117864995	6	*RP11-436D23.1*	Cognition	3.20×10^{-49}	2.27×10^{-8}	inf	PanScan-PanC4
rs566817611 rs4839955	6	*RP11-436D23.1*	Cognition	3.09×10^{-13}	5.05×10^{-8}	inf	UK Biobank
rs553048889 rs670850	18	*RP11-244M2.1*	Cognition	3.48×10^{-13}	6.72×10^{-8}	inf	PanScan-PanC4
rs780435835 rs117133219	18	*RP11-244M2.1*	Cognition	1.36×10^{-8}	1.49×10^{-7}	inf	UK Biobank
rs4392150 rs774580018	18	*LINC00907*	Cognition	5.94×10^{-48}	5.44×10^{-8}	inf	PanScan-PanC4
rs574957 rs568394635	18	*LINC00907*	Cognition	3.11×10^{-47}	1.56×10^{-7}	inf	UK Biobank

CHR: Chromosome. Gene: Gene containing the interacting SNPs. Group: Gene Set. *p*-value: Significance of the interaction. Interaction: SNP pair. Threshold: Bonferroni-corrected significance level (0.05 / number of tests). OR_INT: Odds ratio of interaction ("inf" indicates infinite value). Study: Source of the result.

Cross-Gene Region Interactions Observed Between Cognition, Circadian Rhythm, and Serotonin Gene Sets

I identified multiple significant SNP-SNP interactions between gene regions in both PanScan-PanC4 and UK Biobank datasets (**Table 16**). All reported interactions passed the Bonferroni-corrected significance threshold and showed infinite odds ratios (OR_INT = inf).

Table 16. Most significant SNP-SNP interactions between gene regions that were present in both datasets.

SNP-SNP	CHR	Gene	Group	*p*-value		OR_INT	Study
				Interaction	Threshold		
rs987942245 rs9509983	6 13	*PTPRK* *LINC00540*	Cognition Cognition	3.48×10^{-13}	2.92×10^{-9}	inf	PanScan-PanC4
rs370026517 rs9578450	6 13	*PTPRK* *LINC00540*	Cognition Cognition	3.43×10^{-13}	8.39×10^{-9}	inf	UK Biobank
rs960758960 rs3803479	6 15	*PTPRK* *RORA*	Cognition CircRhythm	3.48×10^{-13}	9.47×10^{-10}	inf	PanScan-PanC4
rs867410771 rs2553229	6 15	*PTPRK* *RORA*	Cognition CircRhythm	7.78×10^{-25}	2.80×10^{-9}	inf	UK Biobank
rs776774567 rs6494231	1 15	*RP4-663N10.1* *RORA*	Cognition CircRhythm	3.81×10^{-26}	9.47×10^{-10}	inf	PanScan-PanC4
rs759831036 rs7176798	1 15	*RP4-663N10.1* *RORA*	Cognition CircRhythm	5.94×10^{-48}	2.80×10^{-9}	inf	UK Biobank
rs4392150 rs891333831	18 15	*LINC00907* *RORA*	Cognition CircRhythm	5.94×10^{-48}	9.47×10^{-10}	inf	PanScan-PanC4
rs57192751 rs7176798	18 15	*LINC00907* *RORA*	Cognition CircRhythm	5.94×10^{-48}	2.80×10^{-9}	inf	UK Biobank
rs4392150 rs1015170124	18 7	*LINC00907* *AHR*	Cognition CircRhythm	5.94×10^{-48}	9.47×10^{-10}	inf	PanScan-PanC4
rs574957 rs999941481	18 7	*LINC00907* *AHR*	Cognition CircRhythm	1.36×10^{-47}	2.80×10^{-9}	inf	UK Biobank
rs4392150 rs926910292	18 4	*LINC00907* *MAPK10*	Cognition CircRhythm	5.94×10^{-48}	9.47×10^{-10}	inf	PanScan-PanC4
rs574957 rs541996588	18 4	*LINC00907* *MAPK10*	Cognition CircRhythm	1.36×10^{-47}	2.80×10^{-9}	inf	UK Biobank
rs1461710 rs973210063	18 6	*LINC00907* *BTBD9*	Cognition CircRhythm	3.48×10^{-13}	9.47×10^{-10}	inf	PanScan-PanC4
rs574957 rs181505995	18 6	*LINC00907* *BTBD9*	Cognition CircRhythm	1.36×10^{-47}	2.80×10^{-9}	inf	UK Biobank
rs75380201 rs763639347	18 7	*LINC00907* *KCND2*	Cognition CircRhythm	1.08×10^{-14}	9.47×10^{-10}	inf	PanScan-PanC4
rs574957 rs766445631	18 7	*LINC00907* *KCND2*	Cognition CircRhythm	1.36×10^{-47}	2.80×10^{-9}	inf	UK Biobank
rs75380201 rs141896823	18 16	*LINC00907* *HS3ST2*	Cognition CircRhythm	1.08×10^{-14}	9.47×10^{-10}	inf	PanScan-PanC4
rs76368980 rs176389	18 16	*LINC00907* *HS3ST2*	Cognition CircRhythm	3.99×10^{-25}	2.80×10^{-9}	inf	UK Biobank
rs141882392 rs12936117	18 17	*LINC00907* *HNF1B*	Cognition CircRhythm	3.48×10^{-13}	9.47×10^{-10}	inf	PanScan-PanC4
rs538218349 rs34064336	18 17	*LINC00907* *HNF1B*	Cognition CircRhythm	1.22×10^{-24}	2.80×10^{-9}	inf	UK Biobank
rs820665 rs541896749	1 3	*RP4-704D23.1* *NLGN1*	Cognition CircRhythm	2.98×10^{-10}	9.47×10^{-10}	inf	PanScan-PanC4
rs137973104 rs73882745	1 3	*RP4-704D23.1* *NLGN1*	Cognition CircRhythm	7.16×10^{-32}	2.80×10^{-9}	inf	UK Biobank
rs61941418 rs557776876	12 6	*RP11-407A16.4* *BTBD9*	Cognition CircRhythm	3.44×10^{-13}	9.47×10^{-10}	inf	PanScan-PanC4
rs150550111 rs4714165	12 6	*RP11-407A16.4* *BTBD9*	Cognition CircRhythm	1.83×10^{-24}	2.80×10^{-9}	inf	UK Biobank
rs1456832427	13	*RP11-427J23.1*	Cognition	3.48×10^{-13}	9.47×10^{-10}	inf	

SNP-SNP	CHR	Gene	Group	*p*-value		OR_INT	Study
				Interaction	Threshold		
rs62073558	17	*HNF1B*	CircRhythm				PanScan-PanC4
rs542776619 rs17626423	13 17	*RP11-427J23.1* *HNF1B*	Cognition CircRhythm	2.66×10^{-22}	2.80×10^{-9}	inf	UK Biobank
rs772612908 rs12691540	18 1	*LINC00907* *CHRM3*	Cognition Serotonin	3.48×10^{-13}	3.41×10^{-9}	inf	PanScan-PanC4
rs574957 rs752369551	18 1	*LINC00907* *CHRM3*	Cognition Serotonin	1.36×10^{-47}	1.17×10^{-8}	inf	UK Biobank
rs766048152 16:10101573_G	13 16	*RP11-427J23.1* *GRIN2A*	Cognition Serotonin	3.48×10^{-13}	3.41×10^{-9}	inf	PanScan-PanC4
rs777871168 rs1650420	13 16	*RP11-427J23.1* *GRIN2A*	Cognition Serotonin	2.12×10^{-47}	1.17×10^{-8}	inf	UK Biobank
rs544255966 rs10993750	18 9	*LINC00907* *SYK*	Cognition Serotonin	3.48×10^{-13}	3.41×10^{-9}	inf	PanScan-PanC4
rs73478549 rs290209	18 9	*LINC00907* *SYK*	Cognition Serotonin	5.28×10^{-25}	1.17×10^{-8}	inf	UK Biobank
rs967003743 rs10744895	6 12	*RP11-436D23.1* *NOS1*	Cognition Serotonin	3.48×10^{-13}	3.41×10^{-9}	inf	PanScan-PanC4
rs1167719419 rs7964845	6 12	*RP11-436D23.1* *NOS1*	Cognition Serotonin	7.59×10^{-25}	1.17×10^{-8}	inf	UK Biobank
rs7661844 rs549346754	4 15	*MAPK10* *RORA*	CircRhythm CircRhythm	1.76×10^{-183}	1.24×10^{-9}	inf	PanScan-PanC4
rs772559472 rs74735171	4 15	*MAPK10* *RORA*	CircRhythm CircRhythm	8.76×10^{-45}	3.76×10^{-9}	inf	UK Biobank
rs7766546 rs888794705	6 10	*HCRTR2* *KCNMA1*	CircRhythm CircRhythm	5.94×10^{-48}	1.24×10^{-9}	inf	PanScan-PanC4
rs545336125 rs17506197	6 10	*HCRTR2* *KCNMA1*	CircRhythm CircRhythm	2.00×10^{-22}	3.76×10^{-9}	inf	UK Biobank
rs7766546 rs886882195	6 11	*HCRTR2* *CAVIN3*	CircRhythm CircRhythm	5.94×10^{-48}	1.24×10^{-9}	inf	PanScan-PanC4
rs750226037 rs1051992	6 11	*HCRTR2* *CAVIN3*	CircRhythm CircRhythm	8.77×10^{-25}	3.76×10^{-9}	inf	UK Biobank
rs7766546 rs891333831	6 15	*HCRTR2* *RORA*	CircRhythm CircRhythm	5.94×10^{-48}	1.24×10^{-9}	inf	PanScan-PanC4
rs554119038 rs341402	6 15	*HCRTR2* *RORA*	CircRhythm CircRhythm	4.04×10^{-25}	3.76×10^{-9}	inf	UK Biobank
rs7766546 rs780240746	6 19	*HCRTR2* *CRTC1*	CircRhythm CircRhythm	5.94×10^{-48}	1.24×10^{-9}	inf	PanScan-PanC4
rs545336125 rs560165635	6 19	*HCRTR2* *CRTC1*	CircRhythm CircRhythm	5.55×10^{-24}	3.76×10^{-9}	inf	UK Biobank
rs777197736 rs6494231	9 15	*GNAQ* *RORA*	CircRhythm CircRhythm	1.12×10^{-24}	1.24×10^{-9}	inf	PanScan-PanC4
rs142620984 rs2278507	9 15	*GNAQ* *RORA*	CircRhythm CircRhythm	9.30×10^{-25}	3.76×10^{-9}	inf	UK Biobank
rs763860894 rs4467018	3 15	*NLGN1* *RORA*	CircRhythm CircRhythm	3.48×10^{-13}	1.24×10^{-9}	inf	PanScan-PanC4
rs556518321 rs74735171	3 15	*NLGN1* *RORA*	CircRhythm CircRhythm	8.76×10^{-45}	3.76×10^{-9}	inf	UK Biobank

CHR: Chromosome. Gene: Gene(s) containing the interacting SNPs. Group indicates the predefined gene set associated with the interacting SNPs: Cognition (cognition-related genes), CircRhythm (circadian rhythm-related genes), Serotonin (serotonergic signaling genes). *p*-value: Significance of the interaction. Interaction: SNP pair. Threshold: Bonferroni-corrected significance level (0.05/number of tests). OR_INT: Odds ratio of interaction ("inf" indicates infinite value). Study: Source of the result.

In the cognition gene set, the SNP-SNP interactions were observed between *PTPRK* and *LINC00540* in both datasets. Additional observed interactions included SNPs from *LINC00907* with *RORA*, *MAPK10*, *AHR*, *HCRTR2*, *HS3ST2*, *HNF1B*, *BTBD9*, *KCND2*, *GNAQ*, *NLGN1*, and *CRTC1* from the circadian rhythm group.

The gene *RORA* appeared in multiple significant interactions with cognition-related genes including *LINC00907*, *RP4-663N10.1*, *RP4-704D23.1*, *RP11-407A16.4*, and *RP11-427J23.1*. SNPs from *RP4-704D23.1* and *RP11-427J23.1* also showed SNP-SNP interactions with *MAPK10* and *HNF1B*.

In the serotonin gene set, the SNP-SNP interactions were detected between *CHRM3*, *GRIN2A*, *SYK*, and *NOS1* with cognition-related genes such as *LINC00907* and *RP11-427J23.1*. These SNP-SNP interactions between gene regions are shown in **Figure 15**.

Fewer and Non-Overlapping Significant Interactions Detected in Permutation Test

I performed a one-time permutation test for each dataset by randomly shuffling case-control labels and repeating the epistasis analyses. While some SNP-SNP interactions remained significant after permutation, the number of significant interactions within the same gene regions and between gene sets was notably lower compared to the original analysis. Moreover, the significant SNP pairs identified in the permuted data involved different variants and gene combinations, with no overlap with the main results.

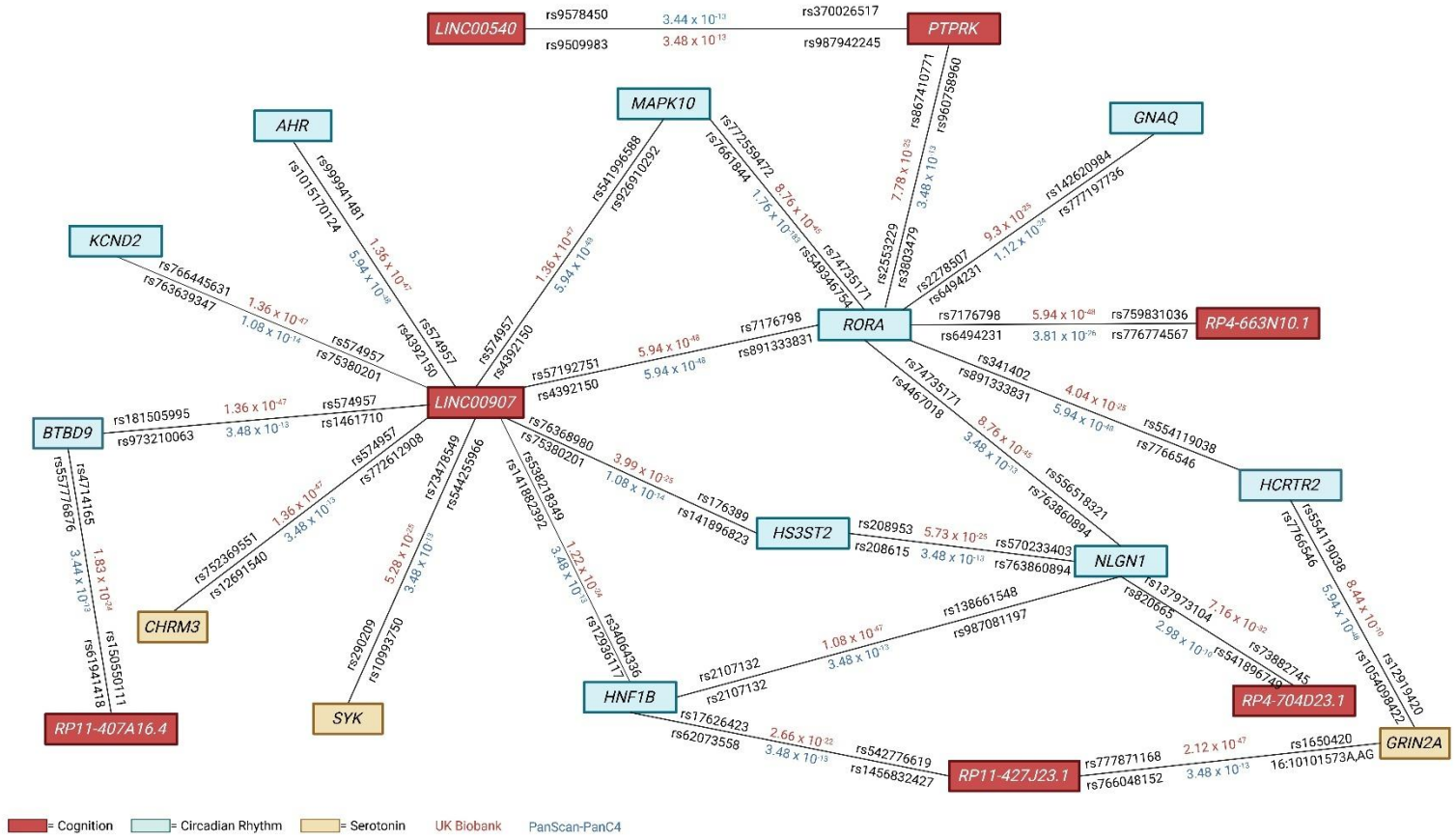


Figure 15. Network of mutual epistatic interactions identified in both PanScan-PanC4 and UK Biobank datasets.

Genes are colored by group: cognition (red), circadian rhythm (blue), and serotonin (yellow). Edges represent SNP-SNP interactions with corresponding *p*-values labeled in red (UK Biobank) and blue (PanScan-PanC4). Only SNP pairs showing significant interactions with an infinite odds ratio in both datasets are displayed.

Firth Logistic Regression and Case-Exclusive Genotype Combinations

I applied Firth logistic regression to the significant SNP-SNP interaction pairs identified in the initial PLINK epistasis analysis, separately for the PanScan-PanC4 and UK Biobank datasets. In both datasets, I observed that many models had rare genotype combinations, and some interaction terms were excluded due to lack of overlap between genotype groups.

I also identified genotype combinations that were present exclusively in cases but not in controls. In the PanScan-PanC4 dataset, I detected two such combinations: rs773576994_G (*RORB*) × rs565020865_C (*RORB*), both from the circadian rhythm gene group; and rs201351580_T (*SERPINE1*) × 7:101136489_G (*SERPINE1*), also from the circadian rhythm group. In the UK Biobank dataset, I found one case-exclusive combination: rs571814364_G (*THEMIS*) × rs76387204_G (*FARS2*), interaction between the cognition and insomnia gene groups (see **Table 17**).

Table 17. Case-exclusive SNP-SNP genotype combinations.

SNP_1	GENE_1	SNP_2	GENE_2	Genotype_Combo	Case_Count	Control_Count	Fisher_P	Study
rs773576994_G	*RORB*	rs565020865_C	*RORB*	0_2	1	0	0.508	PanScan-PanC4
rs201351580_T	*SERPINE1*	7:101136489_G	*SERPINE1*	0_2	1	0	0.508	PanScan-PanC4
rs571814364_G	*THEMIS*	rs76387204_G	*FARS2*	0_2	1	0	0.0926	UK Biobank

SNP_1 and SNP_2: Interacting variants. GENE_1 and GENE_2: Corresponding genes for the variants. Genotype_Combo: Case-exclusive genotype combination (0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternate). Case_Count and Control_Count: Number of individuals with the genotype combination in cases and controls, respectively. Fisher_P: One-sided Fisher's exact test *p*-value. Study: Indicates the dataset.

Covariate-adjusted epistasis analysis with CASSI

Using CASSI, I performed covariate-adjusted SNP-SNP epistasis tests within and between the cognition, circadian rhythm, serotonin, and insomnia gene groups in both cohorts. The lowest observed interaction *p*-value was in the order of 10^{-5} (the *p*-value of significancy thresholds shown in Table 15 and Table 16.). SNP pairs with a minor allele count (MAC) below 5 were automatically excluded by the software. No overlapping SNP-SNP interactions were detected between the CASSI results and those identified in the initial unadjusted epistasis analysis.

6.4 DISCUSSION

This study identified several statistically significant findings that advance our understanding of PDAC genetics. Notably, *RP11*-255M2.3 and *LGR4* emerged as significant in gene-based analyses across both PanScan-PanC4 and UK Biobank. This replication across two large, independent cohorts is particularly striking given the challenges of reproducing associations in complex trait genetics (J. P. Ioannidis et al., 2001). The biological relevance of *LGR4*, a key player in circadian rhythm regulation and Wnt signaling, underscores its potential role in pancreatic tumorigenesis, while *RP11-255M2.3*, a cognition-related non-coding RNA, suggests previously unrecognized regulatory mechanisms.

In the epistasis analysis, I detected 63 significant SNP-SNP interactions in PanScan-PanC4 and 37 in UK Biobank. Some gene regions showed strong and SNP-SNP interaction signals in both cohorts, including *RP11-436D23.1*, *RP11-244M2.1*, and *LINC00907* from the cognition gene group, as well as *KCNH7*, *GSK3B*, *PPARGC1A*, and *NTRK3* from the circadian rhythm group. The fact that these interactions replicated across two large and independent datasets underscores their potential biological relevance. Many of these signals passed strict Bonferroni correction, and the observation of infinite odds ratios, reflecting genotype combinations present exclusively in cases but absent in controls, suggests the possibility of strong, context-specific effects. While rare variant interactions are inherently challenging to interpret, the consistency of these patterns highlights potential neural and circadian regulatory hubs that warrant deeper investigation

One of the key observations in this study was the SNP-SNP interaction network connecting cognition-related genes with circadian rhythm and serotonin pathway genes across both cohorts. As shown in **Figure 15**, several interactions passed strict Bonferroni correction in both PanScan-PanC4 and UK Biobank, and consistently showed infinite odds ratios due to complete genotype separation between cases and controls. *LINC00907*, a long non-coding RNA from the cognition gene group, appeared as a central hub in the network, showing multiple interactions with circadian rhythm genes including *RORA*, *MAPK10*, *BTBD9*, *KCND2*, and *HNF1B*, as well as serotonin-related genes like *CHRM3* and *SYK*. Other cognition genes, such as *RP4-704D23.1* and *RP11-427J23.1*, also formed cross-group interactions with *MAPK10*, *HCRTR2*, and *GRIN2A*.

These SNP-SNP interaction signals suggest that certain cognition genes may act as points of convergence between neurogenetic and circadian or serotonergic regulatory mechanisms. While the functional consequences of these gene pairs are still unclear,

especially since many cognition genes are non-coding and poorly annotated, the consistent patterns observed across two large datasets support the idea that neural-related genetic networks may interact with stress and endocrine pathways to influence PDAC risk.

Still, there were several challenges. One major difficulty was that most cognition-related genes are non-coding RNAs. These genes often do not have well-characterized functions, and they are not covered well in common expression or regulatory databases like GTEx or ENCODE(ENCODE Project Consortium, 2004; Zwir et al., 2022). This makes it difficult to evaluate the biological impact of the significant variants or interpret their role in pancreatic biology.

None of the SNP–SNP interactions between common variants reached statistical significance after stringent correction. Many of the observed interactions involved rare–common variant pairs, which often led to very few individuals carrying certain genotype combinations and made the statistical models unstable. To reduce this bias, I applied Firth logistic regression(Heinze & Schemper, 2002), which confirmed that many models were affected by sparsity. Some interaction terms were dropped automatically due to separation, and others had wide confidence intervals, making the odds ratios difficult to interpret with confidence. These rare-common interactions are known to be hard to replicate in other datasets(Zuk et al., 2012).

Another limitation was technical heterogeneity. The PanScan-PanC4 dataset includes samples from over 12 studies with different genotyping arrays, which potentially causes batch effects. I adjusted for genotyping arrays and principal components to reduce this problem, but there may still be residual confounding. I also tried to run covariate-adjusted epistasis tests using the CASSI software, which supports logistic interaction modeling with covariates. However, CASSI was very slow to run, and it excluded SNP pairs with very low minor allele counts, including many of the rare-common interactions that were significant in the unadjusted model. In the end, no significant SNP-SNP interactions were between the CASSI and PLINK results. This outcome is not surprising, since epistasis detection is known to be very sensitive to filtering, statistical method, and allele frequencies(Cordell, 2009; Wei et al., 2014).

I also searched for case-exclusive genotype combinations, observed only in cases but never in controls. I found three such combinations (two in PanScan-PanC4 and one in UK Biobank), but each was observed in only one case. They did not reach significance in Fisher's test and did not overlap with top interaction signals. These combinations are likely due to chance and are hard to interpret due to low sample counts.

These findings provide one of the first systematic explorations of cognition-related genetic variation and gene–gene interactions in PDAC risk. The observed signals for *LGR4*, *RORA*, *RP11-255M2.3* and *LINC00907* point to a possible neuroendocrine axis linking cognitive, circadian, and serotonergic pathways to pancreatic tumorigenesis. While further validation and functional studies are needed, these results open new avenues for understanding the interplay between neural regulation and cancer biology. By integrating these genetic insights with multi-omics data and larger, diverse cohorts, future research could uncover novel biomarkers and therapeutic targets at the intersection of the nervous and endocrine systems in PDAC.

7 GENERAL DISCUSSION & CONCLUSION

In this thesis, I investigated the genetic susceptibility of PDAC by analyzing common and rare variants through a series of genome-wide and post-GWAS analyses. My focus was on utilizing available large-scale genotype datasets and applying secondary analysis techniques to extract novel insights that may not be immediately evident from traditional GWAS. Given the increasing recognition that many complex traits, including cancer, are influenced by non-coding and rare variants, I designed this work to move beyond conventional single-marker testing and toward a broader view of germline variation and disease risk.

I began by exploring TFBSs and enhancer regions as sources of functional non-coding variation. Among the tested variants, I identified rs2472632, located near *CCDC34*, as a candidate PDAC risk locus. This variant showed evidence of overlapping with enhancer activity and relevant TF motifs, including *LGR4*, which has known roles in Wnt/β-catenin signaling. Although the predictive annotations are based on in silico models, the convergence of enhancer markers, transcription factor binding predictions, and statistical association makes a compelling case for rs2472632 as a regulatory variant that could modulate gene expression in the pancreas. This project reinforced the idea that regulatory polymorphisms, though traditionally overlooked in GWAS, can meaningfully contribute to cancer susceptibility.

In the second part of the thesis, I applied rare variant association tests across the genome using both annotation-weighted and unweighted models. This approach led to the identification of several candidate genes, including *RIPK2*, *PTPRT*, *PARD3*, *CLU*, and *SEPTIN8*. These genes are either functionally relevant to pancreatic biology or involved in cancer-related pathways such as cell adhesion, immune signaling, or apoptosis. For instance, *CLU* exhibited both gene-level association and high expression in pancreatic ductal cells, aligning with its previously reported role in PDAC progression. The gene-based aggregation tests allowed me to capture the cumulative effects of multiple rare variants, including those that would not be individually

significant under single-marker GWAS thresholds. I also observed that annotation-weighted models slightly improved the signal-to-noise ratio, suggesting that incorporating biological priors into statistical testing can increase the interpretability and relevance of findings.

One of the most exploratory and hypothesis-driven aspects of this work was the investigation of cognition-associated genes as potential modulators of PDAC risk. Drawing from published genome-wide studies of cognitive traits and neuroendocrine regulation, I curated a gene list and tested its association with PDAC using multiple analytic strategies, including gene-based tests and epistasis models. While the rationale for this hypothesis is grounded in the bidirectional communication between the nervous system and pancreatic function, the genetic evidence I uncovered remains preliminary. Still, I found gene-level associations with *LGR4*, *RORA*, *RP11-255M2.3*, and *LINC00907*, along with a number of significant SNP-SNP interactions that were present in both the PanScan-PanC4 and UK Biobank datasets. These results suggest that the brain-pancreas axis could have a germline genetic component worth pursuing further in future studies.

Throughout the thesis, I remained acutely aware of the limitations inherent in working with imputed genotype data and merged datasets. One of the most significant technical limitations was the heterogeneity within the PanScan-PanC4 cohort, which is composed of multiple studies genotyped on different platforms. Despite adjusting for population structure and technical covariates, residual artifacts likely influenced the results. A viable solution, which I recommend for future studies, would be to re-impute the merged genotype data using a unified reference panel such as TOPMed. This could reduce batch effects and harmonize variant calling across sub-cohorts.

Another critical limitation is the difficulty of replicating rare variant associations. Rare variants are often population-specific and prone to imputation error, especially when imputation is based on reference panels not representative of the study population. I mitigated this by applying stringent quality control thresholds (e.g., INFO > 0.8), but acknowledge that some of the rare signals may still represent artifacts. Ultimately, replication in independent cohorts with whole genome sequencing data will be essential to validate the findings reported here. Moreover, rare-common interaction models often suffer from infrequent genotype distributions and complete separation between cases and controls in logistic regression. I addressed this using Firth correction, but the instability of these estimates still limits interpretation. More robust methods, including Bayesian or machine learning-based epistasis models, may offer a path forward.

Given the vast number of statistical tests I performed ranging from single-marker analyses to gene-based, interaction, and enrichment tests. I applied rigorous correction procedures to minimize false positives. Even so, some findings must be viewed as

hypothesis-generating until independently replicated. I designed my thesis around secondary analyses because generating new GWAS data is highly time-consuming and costly, whereas a wealth of high-quality primary datasets from large consortia and biobanks already exists but remains underexploited. By focusing on post-GWAS data mining, I was able to adopt a cost-effective strategy that extracts additional value from these resources, applying approaches such as fine-mapping, rare variant testing, and pathway-level analyses to uncover novel insights into PDAC genetics.

In conclusion, this thesis contributes to the growing body of evidence that PDAC risk is influenced not only by common GWAS loci but also by non-coding regulatory variants, rare germline mutations, and potentially by genes involved in cognitive and circadian rhythm pathways. By applying secondary analytical frameworks to existing datasets, I have identified novel candidate variants and pathways that warrant further functional and clinical investigation. These findings support a more complex and multi-layered model of PDAC susceptibility and open new directions for integrating genetic, neurological, and environmental factors in cancer risk assessment.

8 REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Aier, I., Semwal, R., Sharma, A., & Varadwaj, P. K. (2019). A systematic assessment of statistics, risk factors, and underlying features involved in pancreatic cancer. *Cancer Epidemiology*, *58*, 104–110. https://doi.org/10.1016/j.canep.2018.12.001

Amada, K., Hijiya, N., Ikarimoto, S., Yanagihara, K., Hanada, T., Hidano, S., Kurogi, S., Tsukamoto, Y., Nakada, C., Kinoshita, K., Hirashita, Y., Uchida, T., Shin, T., Yada, K., Hirashita, T., Kobayashi, T., Murakami, K., Inomata, M., Shirao, K., … Moriyama, M. (2023). Involvement of clusterin expression in the refractory response of pancreatic cancer cells to a MEK inhibitor. *Cancer Science*, *114*(5), 2189–2202. https://doi.org/10.1111/cas.15735

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R. Z., Fuchs, C. S., Petersen, G. M., Arslan, A. A., Bueno-de-Mesquita, H. B., Gross, M., Helzlsouer, K., Jacobs, E. J., LaCroix, A., Zheng, W., Albanes, D., Bamlet, W., Berg, C. D., Berrino, F., Bingham, S., Buring, J. E., Bracci, P. M., … Hoover, R. N. (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature Genetics*, *41*(9), 986–990. https://doi.org/10.1038/ng.429

Amundadottir, L. T. (2016). Pancreatic Cancer Genetics. *International Journal of Biological Sciences*, *12*(3), 314–325. https://doi.org/10.7150/ijbs.15001

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–1573. https://doi.org/10.1038/nprot.2010.116

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. https://doi.org/10.1038/nature12787

Antoni, M. H., Lutgendorf, S. K., Cole, S. W., Dhabhar, F. S., Sephton, S. E., McDonald, P. G., Stefanek, M., & Sood, A. K. (2006). The influence of bio-behavioural factors on tumour biology: Pathways and mechanisms. *Nature Reviews Cancer*, *6*(3), 240–248. https://doi.org/10.1038/nrc1820

Armitage, P. (1955). Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, *11*(3), 375–386. https://doi.org/10.2307/3001775

Bahadoran, Z., Mirmiran, P., Ghasemi, A., & Kashfi, K. (2019). Type 2 Diabetes and Cancer: The Nitric Oxide Connection. Crit Rev Oncog, 24(3), 235–242. https://doi.org/10.1615/CritRevOncog.2019031256

Band, G., & Marchini, J. (2018). *BGEN: A binary file format for imputed genotype and haplotype data* (p. 308296). bioRxiv. https://doi.org/10.1101/308296

Bartsch, D. K., Sina-Frey, M., Lang, S., Wild, A., Gerdes, B., Barth, P., Kress, R., Grützmann, R., Colombo-Benkmann, M., Ziegler, A., Hahn, S. A., Rothmund, M., & Rieder, H. (2002). CDKN2A Germline Mutations in Familial Pancreatic Cancer. *Annals of Surgery*, *236*(6), 730.

Basu, S., & Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, *35*(7), 606–619. https://doi.org/10.1002/gepi.20609

Boardman, L. A., Thibodeau, S. N., Schaid, D. J., Lindor, N. M., McDonnell, S. K., Burgart, L. J., Ahlquist, D. A., Podratz, K. C., Pittelkow, M., & Hartmann, L. C. (1998). Increased Risk for Cancer in Patients with the Peutz-Jeghers Syndrome. *Annals of Internal Medicine*, *128*(11), 896–899. https://doi.org/10.7326/0003-4819-128-11-199806010-00004

Bodmer, W., & Tomlinson, I. (2010). Rare genetic variants and the risk of cancer. *Current Opinion in Genetics & Development*, *20*(3), 262–267. https://doi.org/10.1016/j.gde.2010.04.016

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, *18*(1), 77. https://doi.org/10.1186/s13059-017-1212-4

Bosetti, C., Rosato, V., Li, D., Silverman, D., Petersen, G. M., Bracci, P. M., Neale, R. E., Muscat, J., Anderson, K., Gallinger, S., Olson, S. H., Miller, A. B., Bas Bueno-de-Mesquita, H., Scelo, G., Janout, V., Holcatova, I., Lagiou, P., Serraino, D., Lucenteforte, E., … Ghadirian, P. (2014). Diabetes, antidiabetic medications, and pancreatic cancer risk: An analysis from the International Pancreatic Cancer Case-Control Consortium. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, *25*(10), 2065–2072. https://doi.org/10.1093/annonc/mdu276

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., McVean, G., Leslie, S., Donnelly, P., & Marchini, J. (2017). *Genome-wide genetic data on ~500,000 UK Biobank participants* (p. 166298). bioRxiv. https://doi.org/10.1101/166298

Bykova, M., Hou, Y., Eng, C., & Cheng, F. (2022). Quantitative trait locus (xQTL) approaches identify risk genes and drug targets from human non-coding genomes. Human Molecular Genetics, 31(R1), R105–R113. https://doi.org/10.1093/hmg/ddac208

Campa, D., Gentiluomo, M., Stein, A., Aoki, M. N., Oliverius, M., Vodičková, L., Jamroziak, K., Theodoropoulos, G., Pasquali, C., Greenhalf, W., Arcidiacono, P. G., Uzunoglu, F., Pezzilli, R., Luchini, C., Puzzono, M., Loos, M., Giaccherini, M., Katzke, V., Mambrini, A., … Canzian, F. (2023). The PANcreatic Disease ReseArch (PANDoRA) consortium: Ten years' experience of association studies to understand the genetic architecture of pancreatic cancer. *Critical Reviews in Oncology/Hematology*, *186*, 104020. https://doi.org/10.1016/j.critrevonc.2023.104020

Canto, M. I., Almario, J. A., Schulick, R. D., Yeo, C. J., Klein, A., Blackford, A., Shin, E. J., Sanyal, A., Yenokyan, G., Lennon, A. M., Kamel, I. R., Fishman, E. K., Wolfgang, C., Weiss, M., Hruban, R. H., & Goggins, M. (2018). Risk of Neoplastic Progression in Individuals at High Risk for Pancreatic Cancer Undergoing Long-term Surveillance. *Gastroenterology*, *155*(3), 740-751.e2. https://doi.org/10.1053/j.gastro.2018.05.035

Cardon, L. R., & Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. *Trends in Genetics: TIG*, *19*(3), 135–140. https://doi.org/10.1016/S0168-9525(03)00022-2

Carrasquillo, M. M., McCallion, A. S., Puffenberger, E. G., Kashuk, C. S., Nouri, N., & Chakravarti, A. (2002). Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature Genetics*, *32*(2), 237–244. https://doi.org/10.1038/ng998

Carreras-Torres, R., Johansson, M., Haycock, P. C., Relton, C. L., Smith, G. D., Brennan, P., & Martin, R. M. (2018). Role of obesity in smoking behaviour: Mendelian randomisation study in UK Biobank. *BMJ*, *361*, k1767. https://doi.org/10.1136/bmj.k1767

Ceyhan, G. O., Demir, I. E., Altintas, B., Rauch, U., Thiel, G., Müller, M. W., Giese, N. A., Friess, H.,

& Schäfer, K.-H. (2008). Neural invasion in pancreatic cancer: A mutual tropism between neurons and cancer cells. *Biochemical and Biophysical Research Communications*, *374*(3), 442–447. https://doi.org/10.1016/j.bbrc.2008.07.035

Chaffee, K. G., Oberg, A. L., McWilliams, R. R., Majithia, N., Allen, B. A., Kidd, J., Singh, N., Hartman, A.-R., Wenstrup, R. J., & Petersen, G. M. (2018). Prevalence of germ-line mutations in cancer genes among pancreatic cancer patients with a positive family history. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *20*(1), 119–127. https://doi.org/10.1038/gim.2017.85

Chang, J., Tian, J., Zhu, Y., Zhong, R., Zhai, K., Li, J., Ke, J., Han, Q., Lou, J., Chen, W., Zhu, B., Shen, N., Zhang, Y., Gong, Y., Yang, Y., Zou, D., Peng, X., Zhang, Z., Zhang, X., … Miao, X. (2018). Exome-wide analysis identifies three low-frequency missense variants associated with pancreatic cancer risk in Chinese populations. *Nature Communications*, *9*(1), 3688. https://doi.org/10.1038/s41467-018-06136-x

Chen, F., Childs, E. J., Mocci, E., Bracci, P., Gallinger, S., Li, D., Neale, R. E., Olson, S. H., Scelo, G., Bamlet, W. R., Blackford, A. L., Borges, M., Brennan, P., Chaffee, K. G., Duggal, P., Hassan, M. J., Holly, E. A., Hung, R. J., Goggins, M. G., … Klein, A. P. (2019). Analysis of Heritability and Genetic Architecture of Pancreatic Cancer: A PanC4 Study. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *28*(7), 1238–1245. https://doi.org/10.1158/1055-9965.EPI-18-1235

Childs, E. J., Mocci, E., Campa, D., Bracci, P. M., Gallinger, S., Goggins, M., Li, D., Neale, R. E., Olson, S. H., Scelo, G., Amundadottir, L. T., Bamlet, W. R., Bijlsma, M. F., Blackford, A., Borges, M., Brennan, P., Brenner, H., Bueno-de-Mesquita, H. B., Canzian, F., … Klein, A. P. (2015). Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nature Genetics*, *47*(8), 911–916. https://doi.org/10.1038/ng.3341

Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2020). A guide to performing Polygenic Risk Score analyses. Nature Protocols, 15(9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, *6*(2), 121–133. https://doi.org/10.1038/nprot.2010.182

Clay, S., Alladina, J., Smith, N. P., Visness, C. M., Wood, R. A., O'Connor, G. T., Cohen, R. T., Khurana Hershey, G. K., Kercsmar, C. M., Gruchalla, R. S., Gill, M. A., Liu, A. H., Kim, H., Kattan, M., Bacharier, L. B., Rastogi, D., Rivera-Spoljaric, K., Robison, R. G., Gergen, P. J., … Dapas, M. (2024). Gene-based association study of rare variants in children of diverse ancestries implicates TNFRSF21 in the development of allergic asthma. *The Journal of Allergy and Clinical Immunology*, *153*(3), 809–820. https://doi.org/10.1016/j.jaci.2023.10.023

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., & Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, *37*(11), 1243–1246. https://doi.org/10.1038/ng1653

Cole, S. W., Nagaraja, A. S., Lutgendorf, S. K., Green, P. A., & Sood, A. K. (2015). Sympathetic nervous system regulation of the tumour microenvironment. *Nature Reviews Cancer*, *15*(9), 563–572. https://doi.org/10.1038/nrc3978

Consortium, Gte. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204. https://doi.org/10.1038/nature24277

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463–2468. https://doi.org/10.1093/hmg/11.20.2463

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews. Genetics*, *10*(6), 392–404. https://doi.org/10.1038/nrg2579

Corradi, C., Gentiluomo, M., Gajdán, L., Cavestro, G. M., Kreivenaite, E., Di Franco, G., Sperti, C., Giaccherini, M., Petrone, M. C., Tavano, F., Gioffreda, D., Morelli, L., Soucek, P., Andriulli, A., Izbicki, J. R., Napoli, N., Malecka-Panas, E., Hegyi, P., Neoptolemos, J. P., … Campa, D. (2021). Genome-wide scan of long noncoding RNA single nucleotide polymorphisms and

pancreatic cancer susceptibility. *International Journal of Cancer*, *148*(11), 2779–2788. https://doi.org/10.1002/ijc.33475

Corral, J. E., Mareth, K. F., Riegert-Johnson, D. L., Das, A., & Wallace, M. B. (2019). Diagnostic Yield From Screening Asymptomatic Individuals at High Risk for Pancreatic Cancer: A Meta-analysis of Cohort Studies. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, *17*(1), 41–53. https://doi.org/10.1016/j.cgh.2018.04.065

Cunha, D. M. G., Koike, M. K., Barbeiro, D. F., Barbeiro, H. V., Hamasaki, M. Y., Coelho Neto, G. T., Machado, M. C. C., & da Silva, F. P. (2014). Increased intestinal production of α-defensins in aged rats with acute pancreatic injury. *Experimental Gerontology*, *60*, 215–219. https://doi.org/10.1016/j.exger.2014.11.008

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, *19*, 73–96. https://doi.org/10.1146/annurev-genom-083117-021602

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M., McGhee, K., Lopez, L., Gow, A. J., Corley, J., Redmond, P., Fox, H. C., Haggarty, P., Whalley, L. J., McNeill, G., … Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*, *16*(10), 996–1005. https://doi.org/10.1038/mp.2011.85

De Marino, A., Mahmoud, A. A., Bose, M., Bircan, K. O., Terpolovsky, A., Bamunusinghe, V., Bohn, S., Khan, U., Novković, B., & Yazdi, P. G. (2022). A comparative analysis of current phasing and imputation software. *PloS One*, *17*(10), e0260177. https://doi.org/10.1371/journal.pone.0260177

Demir, I. E., Ceyhan, G. O., Liebl, F., D'Haese, J. G., Maak, M., & Friess, H. (2010). Neural Invasion in Pancreatic Cancer: The Past, Present and Future. *Cancers*, *2*(3), 1513–1527. https://doi.org/10.3390/cancers2031513

Derkach, A., Lawless, J. F., & Sun, L. (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology*, *37*(1), 110–121. https://doi.org/10.1002/gepi.21689

Devlin, B., & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, *29*(2), 311–322. https://doi.org/10.1006/geno.1995.9003

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics (Oxford, England)*, *30*(9), 1266–1272. https://doi.org/10.1093/bioinformatics/btu014

Earl, J., Galindo-Pumariño, C., Encinas, J., Barreto, E., Castillo, M. E., Pachón, V., Ferreiro, R., Rodríguez-Garrote, M., González-Martínez, S., Ramon Y Cajal, T., Diaz, L. R., Chirivella-Gonzalez, I., Rodriguez, M., de Castro, E. M., García-Seisdedos, D., Muñoz, G., Rosa, J. M. R., Marquez, M., Malats, N., & Carrato, A. (2020). A comprehensive analysis of candidate genes in familial pancreatic cancer families reveals a high frequency of potentially pathogenic germline variants. *EBioMedicine*, *53*, 102675. https://doi.org/10.1016/j.ebiom.2020.102675

Edwards, A. W. F. (2008). G. H. Hardy (1908) and Hardy-Weinberg equilibrium. *Genetics*, *179*(3), 1143–1150. https://doi.org/10.1534/genetics.104.92940

Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: Variant prioritization and Mendelian disease. *Nature Reviews. Genetics*, *18*(10), 599–612. https://doi.org/10.1038/nrg.2017.52

ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E.,

Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., … de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. https://doi.org/10.1038/nature05874

Evans, D. M., Marchini, J., Morris, A. P., & Cardon, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLoS Genetics*, *2*(9), e157. https://doi.org/10.1371/journal.pgen.0020157

Fan, G., Ye, D., Zhu, S., Xi, J., Guo, X., Qiao, J., Wu, Y., Jia, W., Wang, G., Fan, G., & Kang, J. (2017). RTL1 promotes melanoma proliferation by regulating Wnt/β-catenin signaling. *Oncotarget*, *8*(62), 106026–106037. https://doi.org/10.18632/oncotarget.22523

Fang, J., Jia, J., Makowski, M., Xu, M., Wang, Z., Zhang, T., Hoskins, J. W., Choi, J., Han, Y., Zhang, M., Thomas, J., Kovacs, M., Collins, I., Dzyadyk, M., Thompson, A., O'Neill, M., Das, S., Lan, Q., Koster, R., … Amundadottir, L. T. (2017). Functional characterization of a multi-cancer risk locus on chr5p15.33 reveals regulation of TERT by ZNF148. *Nature Communications*, *8*(1), 15034. https://doi.org/10.1038/ncomms15034

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

French, J. D., & Edwards, S. L. (2020). The Role of Noncoding Variants in Heritable Disease. *Trends in Genetics: TIG*, *36*(11), 880–891. https://doi.org/10.1016/j.tig.2020.07.004

Galeotti, A. A., Gentiluomo, M., Rizzato, C., Obazee, O., Neoptolemos, J. P., Pasquali, C., Nentwich, M., Cavestro, G. M., Pezzilli, R., Greenhalf, W., Holleczek, B., Schroeder, C., Schöttker, B., Ivanauskas, A., Ginocchi, L., Key, T. J., Hegyi, P., Archibugi, L., Darvasi, E., … Campa, D. (2021). Polygenic and multifactorial scores for pancreatic ductal adenocarcinoma risk prediction. Journal of Medical Genetics, 58(6), 369–377. https://doi.org/10.1136/jmedgenet-2020-106961

Gardner, T. B., Hessami, N., Smith, K. D., Ripple, G. H., Barth, R. J., Klibansky, D. A., Colacchio, T. A., Zaki, B., Tsapakos, M. J., Suriawinata, A. A., Putra, J., Tsongalis, G. J., Mody, K., Gordon, S. R., & Pipas, J. M. (2014). The effect of neoadjuvant chemoradiation on pancreatic cancer-associated diabetes mellitus. *Pancreas*, *43*(7), 1018–1021. https://doi.org/10.1097/MPA.0000000000000162

Geng, W., Liang, W., Fan, Y., Ye, Z., & Zhang, L. (2018). Overexpression of CCDC34 in colorectal cancer and its involvement in tumor growth, apoptosis and invasion. *Molecular Medicine Reports*, *17*(1), 465–473. https://doi.org/10.3892/mmr.2017.7860

Gentiluomo, M., Canzian, F., Nicolini, A., Gemignani, F., Landi, S., & Campa, D. (2022). Germline genetic variability in pancreatic cancer risk and prognosis. *Seminars in Cancer Biology*, *79*, 105–131. https://doi.org/10.1016/j.semcancer.2020.08.003

Gentiluomo, M., Lu, Y., Canzian, F., & Campa, D. (2019). Genetic variants in taste-related genes and risk of pancreatic cancer. *Mutagenesis*, *34*(5–6), 391–394. https://doi.org/10.1093/mutage/gez032

Gentiluomo, M., Peduzzi, G., Lu, Y., Campa, D., & Canzian, F. (2019). Genetic polymorphisms in inflammatory genes and pancreatic cancer risk: A two-phase study on more than 14 000 individuals. *Mutagenesis*, *34*(5–6), 395–401. https://doi.org/10.1093/mutage/gez040

Gentiluomo, M., Puchalt García, P., Galeotti, A. A., Talar-Wojnarowska, R., Tjaden, C., Tavano, F., Strobel, O., Kupcinskas, J., Neoptolemos, J., Hegyi, P., Costello, E., Pezzilli, R., Sperti, C., Lawlor, R. T., Capurso, G., Szentesi, A., Soucek, P., Vodicka, P., Lovecek, M., … Campa, D. (2019). Genetic variability of the ABCC2 gene and clinical outcomes in pancreatic cancer patients. *Carcinogenesis*, *40*(4), 544–550. https://doi.org/10.1093/carcin/bgz006

Ghadirian, P., Lynch, H. T., & Krewski, D. (2003). Epidemiology of pancreatic cancer: An overview. *Cancer Detection and Prevention*, *27*(2), 87–93. https://doi.org/10.1016/s0361-090x(03)00002-3

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383. https://doi.org/10.1371/journal.pgen.1004383

Glinka, A., Dolde, C., Kirsch, N., Huang, Y.-L., Kazanskaya, O., Ingelfinger, D., Boutros, M., Cruciat, C.-M., & Niehrs, C. (2011). LGR4 and LGR5 are R-spondin receptors mediating Wnt/β-catenin and Wnt/PCP signaling. *EMBO Reports*, *12*(10), 1055–1061. https://doi.org/10.1038/embor.2011.175

Goggins, M., Overbeek, K. A., Brand, R., Syngal, S., Chiaro, M. D., Bartsch, D. K., Bassi, C., Carrato, A., Farrell, J., Fishman, E. K., Fockens, P., Gress, T. M., Hooft, J. E. van, Hruban, R. H., Kastrinos, F., Klein, A., Lennon, A. M., Lucas, A., Park, W., … Bruno, M. (2020). Management of patients with increased risk for familial pancreatic cancer: Updated recommendations from the International Cancer of the Pancreas Screening (CAPS) Consortium. https://doi.org/10.1136/gutjnl-2019-319352

Goldstein, A. M., Struewing, J. P., Fraser, M. C., Smith, M. W., & Tucker, M. A. (2004). Prospective risk of cancer in CDKN2A germline mutation carriers. *Journal of Medical Genetics*, *41*(6), 421–424. https://doi.org/10.1136/jmg.2004.019349

Gong, Y., Qiu, W., Ning, X., Yang, X., Liu, L., Wang, Z., Lin, J., Li, X., & Guo, Y. (2015). CCDC34 is up-regulated in bladder cancer and regulates bladder cancer cell proliferation, apoptosis and migration. *Oncotarget*, *6*(28), 25856–25867. https://doi.org/10.18632/oncotarget.4624

Grimes, D. A., & Schulz, K. F. (2002). Cohort studies: Marching towards outcomes. *Lancet (London, England)*, *359*(9303), 341–345. https://doi.org/10.1016/S0140-6736(02)07500-1

Guerreiro, R. J., & Hardy, J. (2012). TOMM40 association with Alzheimer disease: Tales of APOE and linkage disequilibrium. *Archives of Neurology*, *69*(10), 1243–1244. https://doi.org/10.1001/archneurol.2012.1935

Han, F., & Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, *70*(1), 42–54. https://doi.org/10.1159/000288704

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Hart, P. A., Bellin, M. D., Andersen, D. K., Bradley, D., Cruz-Monserrate, Z., Forsmark, C. E., Goodarzi, M. O., Habtezion, A., Korc, M., Kudva, Y. C., Pandol, S. J., Yadav, D., & Chari, S. T. (2016). Type 3c (pancreatogenic) diabetes mellitus secondary to chronic pancreatitis and pancreatic cancer. The Lancet. Gastroenterology & Hepatology, 1(3), 226–237. https://doi.org/10.1016/S2468-1253(16)30106-6

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409–2419. https://doi.org/10.1002/sim.1047

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*(6), 226–231. https://doi.org/10.1007/BF01245622

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, *99*(6), 1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Hruban, R. H., Canto, M. I., Goggins, M., Schulick, R., & Klein, A. P. (2010). Update on familial pancreatic cancer. *Advances in Surgery*, *44*, 293–311. https://doi.org/10.1016/j.yasu.2010.05.011

Huang, C., Zhang, C., Cao, Y., Li, J., & Bi, F. (2023). Major roles of the circadian clock in cancer. Cancer Biology & Medicine, 20(1), 1–24. https://doi.org/10.20892/j.issn.2095-3941.2022.0474

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.-F., Gambaro, G., Richards, J. B., Durbin, R., Timpson, N. J., Marchini, J., & Soranzo, N. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, *6*(1), 8111. https://doi.org/10.1038/ncomms9111

Huang, J., Lok, V., Ngai, C. H., Zhang, L., Yuan, J., Lao, X. Q., Ng, K., Chong, C., Zheng, Z.-J., &

Wong, M. C. S. (2021). Worldwide Burden of, Risk Factors for, and Trends in Pancreatic Cancer. *Gastroenterology*, *160*(3), 744–754. https://doi.org/10.1053/j.gastro.2020.10.007

Hussain, S. M., Kansal, R. G., Alvarez, M. A., Hollingsworth, T. J., Elahi, A., Miranda-Carboni, G., Hendrick, L. E., Pingili, A. K., Albritton, L. M., Dickson, P. V., Deneve, J. L., Yakoub, D., Hayes, D. N., Kurosu, M., Shibata, D., Makowski, L., & Glazer, E. S. (2021). Role of TGF-β in pancreatic ductal adenocarcinoma progression and PD-L1 expression. Cellular Oncology (Dordrecht, Netherlands), 44(3), 673–687. https://doi.org/10.1007/s13402-021-00594-0

Hutchinson, A., Asimit, J., & Wallace, C. (2020). Fine-mapping genetic associations. *Human Molecular Genetics*, *29*(R1), R81. https://doi.org/10.1093/hmg/ddaa148

Igl, B.-W., König, I. R., & Ziegler, A. (2008). What Do We Mean by 'Replication' and 'Validation' in Genome-Wide Association Studies? *Human Heredity*, *67*(1), 66–68. https://doi.org/10.1159/000164400

Ilic, I., & Ilic, M. (2022). International patterns in incidence and mortality trends of pancreatic cancer in the last three decades: A joinpoint regression analysis. *World Journal of Gastroenterology*, *28*(32), 4698–4715. https://doi.org/10.3748/wjg.v28.i32.4698

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–58. https://doi.org/10.1038/nature09298

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, *29*(3), 306–309. https://doi.org/10.1038/ng749

Ioannidis, N. M., Davis, J. R., DeGorter, M. K., Larson, N. B., McDonnell, S. K., French, A. J., Battle, A. J., Hastie, T. J., Thibodeau, S. N., Montgomery, S. B., Bustamante, C. D., Sieh, W., & Whittemore, A. S. (2017). FIRE: Functional inference of genetic variants that regulate gene expression. *Bioinformatics (Oxford, England)*, *33*(24), 3895–3901. https://doi.org/10.1093/bioinformatics/btx534

Iorio, F., Garcia-Alonso, L., & Brammeld, J. S. (2018). Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. Sci Rep, 8(1). https://doi.org/10.1038/s41598-018-25076-6

Jiang, S.-H., Li, J., Dong, F.-Y., Yang, J.-Y., Liu, D.-J., Yang, X.-M., Wang, Y.-H., Yang, M.-W., Fu, X.-L., Zhang, X.-X., Li, Q., Pang, X.-F., Huo, Y.-M., Li, J., Zhang, J.-F., Lee, H.-Y., Lee, S.-J., Qin, W.-X., Gu, J.-R., … Zhang, Z.-G. (2017). Increased Serotonin Signaling Contributes to the Warburg Effect in Pancreatic Tumor Cells Under Metabolic Stress and Promotes Growth of Pancreatic Tumors in Mice. Gastroenterology, 153(1), 277-291.e19. https://doi.org/10.1053/j.gastro.2017.03.008

Jiao, J., Ruan, L., Cheng, C. S., Wang, F., Yang, P., & Chen, Z. (2023). Paired protein kinases PRKCI-RIPK2 promote pancreatic cancer growth and metastasis via enhancing NF-κB/JNK/ERK phosphorylation. Mol Med Camb Mass, 29(1). https://doi.org/10.1186/s10020-023-00648-z

Jin, C., & Bai, L. (2020). Pancreatic Cancer: Current Situation and Challenges. *Gastroenterology & Hepatology Letters*, *2*(1), 1–3. https://doi.org/10.18063/ghl.v2i1.243

Johnston, A. D., Simões-Pires, C. A., Thompson, T. V., Suzuki, M., & Greally, J. M. (2019). Functional genetic variants can mediate their regulatory effects through alteration of transcription factor binding. *Nature Communications*, *10*(1), 3472. https://doi.org/10.1038/s41467-019-11412-5

Jones, S., Hruban, R. H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D. W., Lin, J. C.-H., Palmisano, E., Brune, K., Jaffee, E. M., Iacobuzio-Donahue, C. A., Maitra, A., Parmigiani, G., Kern, S. E., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Eshleman, J. R., Goggins, M., & Klein, A. P. (2009). Exomic Sequencing Identifies PALB2 as a Pancreatic Cancer Susceptibility Gene. *Science*, *324*(5924), 217–217. https://doi.org/10.1126/science.1171202

Jung, S. W., Lee, J., & Cho, A. E. (2017). Elucidating the Bacterial Membrane Disruption Mechanism of Human α-Defensin 5: A Theoretical Study. *The Journal of Physical Chemistry B*, *121*(4), 741–748. https://doi.org/10.1021/acs.jpcb.6b11806

Jurgens, S. J., Choi, S. H., Morrill, V. N., Chaffin, M., Pirruccello, J. P., Halford, J. L., Weng, L.-C.,

Nauffal, V., Roselli, C., Hall, A. W., Oetjens, M. T., Lagerman, B., vanMaanen, D. P., Regeneron Genetics Center, Aragam, K. G., Lunetta, K. L., Haggerty, C. M., Lubitz, S. A., & Ellinor, P. T. (2022). Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nature Genetics*, *54*(3), 240–250. https://doi.org/10.1038/s41588-021-01011-w

Kang, Y. E., Kim, J.-M., Kim, K. S., Chang, J. Y., Jung, M., Lee, J., Yi, S., Kim, H. W., Kim, J. T., Lee, K., Choi, M. J., Kang, S. K., Lee, S. E., Yi, H.-S., Koo, B. S., & Shong, M. (2017). Upregulation of RSPO2-GPR48/LGR4 signaling in papillary thyroid carcinoma contributes to tumor progression. *Oncotarget*, *8*(70), 114980–114994. https://doi.org/10.18632/oncotarget.22692

Kastrinos, F., Mukherjee, B., Tayob, N., Wang, F., Sparr, J., Raymond, V. M., Bandipalliam, P., Stoffel, E. M., Gruber, S. B., & Syngal, S. (2009). Risk of Pancreatic Cancer in Families With Lynch Syndrome. *JAMA*, *302*(16), 1790–1795. https://doi.org/10.1001/jama.2009.1529

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Khoury, M. J., Beaty, T. H., Cohen, B. H., Khoury, M. J., Beaty, T. H., & Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., & Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, *10*(10), e1004722. https://doi.org/10.1371/journal.pgen.1004722

Kim, H. S., Gweon, T.-G., Park, S. H., Kim, T. H., Kim, C. W., & Chang, J. H. (2023). Incidence and risk of pancreatic cancer in patients with chronic pancreatitis: Defining the optimal subgroup for surveillance. Scientific Reports, 13(1), 106. https://doi.org/10.1038/s41598-022-26411-8

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, A. V., Neale, R. E., Tempero, M., Tuveson, D. A., Hruban, R. H., & Neoptolemos, J. P. (2016). Pancreatic cancer. *Nature Reviews. Disease Primers*, *2*, 16022. https://doi.org/10.1038/nrdp.2016.22

Klein, A. P. (2021). Pancreatic cancer epidemiology: Understanding the role of lifestyle and inherited risk factors. *Nature Reviews. Gastroenterology & Hepatology*, *18*(7), 493–502. https://doi.org/10.1038/s41575-021-00457-x

Klein, A. P., Wolpin, B. M., Risch, H. A., Stolzenberg-Solomon, R. Z., Mocci, E., Zhang, M., Canzian, F., Childs, E. J., Hoskins, J. W., Jermusyk, A., Zhong, J., Chen, F., Albanes, D., Andreotti, G., Arslan, A. A., Babic, A., Bamlet, W. R., Beane-Freeman, L., Berndt, S. I., … Amundadottir, L. T. (2018). Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nature Communications*, *9*(1), 556. https://doi.org/10.1038/s41467-018-02942-5

Kockum, I., Huang, J., & Stridh, P. (2023). Overview of Genotyping Technologies and Methods. *Current Protocols*, *3*(4), e727. https://doi.org/10.1002/cpz1.727

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90-97. https://doi.org/10.1093/nar/gkw377

Kumar, S., Ambrosini, G., & Bucher, P. (2017). SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Research*, *45*(D1), D139–D144. https://doi.org/10.1093/nar/gkw1064

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, *6*(1), 5–30. Scopus. https://doi.org/10.1017/S0266462300008916

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014a). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014b). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., & Lin, X. (2012). Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *American Journal of Human Genetics*, *91*(2), 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England)*, *13*(4), 762–775. https://doi.org/10.1093/biostatistics/kxs014

Lee, Y. H. (2015). Meta-Analysis of Genetic Association Studies. *Annals of Laboratory Medicine*, *35*(3), 283–287. https://doi.org/10.3343/alm.2015.35.3.283

Leeuw, C. A. de, Mooij, J. M., Heskes, T., & Posthuma, D. (2015a). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, *11*(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

Leeuw, C. A. de, Mooij, J. M., Heskes, T., & Posthuma, D. (2015b). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, *11*(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

Lewis, C. M., & Knight, J. (2012). Introduction to Genetic Association Studies. *Cold Spring Harbor Protocols*, *2012*(3), pdb.top068163. https://doi.org/10.1101/pdb.top068163

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, *83*(3), 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024

Li, B.-Q., Wang, L., Li, J., Zhou, L., Zhang, T.-P., Guo, J.-C., & Zhao, Y.-P. (2017). Surgeons' knowledge regarding the diagnosis and management of pancreatic cancer in China: A cross-sectional study. *BMC Health Services Research*, *17*(1), 395. https://doi.org/10.1186/s12913-017-2345-6

Li, D., Tang, L., Liu, B., Xu, S., Jin, M., & Bo, W. (2021). RIPK2 is an unfavorable prognosis marker and a potential therapeutic target in human kidney renal clear cell carcinoma. Aging, 13(7), 10450–10467. https://doi.org/10.18632/aging.202808

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233. https://doi.org/10.1093/genetics/165.4.2213

Li, S., & Tian, B. (2017). Acute pancreatitis in patients with pancreatic cancer: Timing of surgery and survival duration. *Medicine*, *96*(3), e5908. https://doi.org/10.1097/MD.0000000000005908

Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Aslibekyan, S., Ballantyne, C. M., Bielak, L. F., Blangero, J., Boerwinkle, E., Bowden, D. W., Broome, J. G., Conomos, M. P., Correa, A., Cupples, L. A., … Lin, X. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, *52*(9), 969–983. https://doi.org/10.1038/s41588-020-0676-4

Li, X., Xu, H., & Gao, P. (2018). ABO Blood Group and Diabetes Mellitus Influence the Risk for Pancreatic Cancer in a Population from China. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, *24*, 9392–9398. https://doi.org/10.12659/MSM.913769

Li, Z., Li, X., Zhou, H., Gaynor, S. M., Selvaraj, M. S., Arapoglou, T., Quick, C., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Auer, P. L., Bielak, L. F., Bis, J. C., Blackwell, T. W., Blangero, J., Boerwinkle, E., Bowden, D. W., … Lin, X. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, *19*(12), 1599–1611. https://doi.org/10.1038/s41592-022-01640-x

Lin, S., Tan, S., Peng, Y., Tulamaiti, A., Du, W., Ding, K., Chen, C., Wu, J., Li, H., Xu, W., Sun, J., Zhang, X.-L., Zhang, Z., & Zhao, X. (2025). Histone serotonylation promotes pancreatic cancer development via lipid metabolism remodeling. Nature Communications, 16(1), 5947. https://doi.org/10.1038/s41467-025-61197-z

Lin, Y., Nakatochi, M., Hosono, Y., Ito, H., Kamatani, Y., Inoko, A., Sakamoto, H., Kinoshita, F.,

Kobayashi, Y., Ishii, H., Ozaka, M., Sasaki, T., Matsuyama, M., Sasahira, N., Morimoto, M., Kobayashi, S., Fukushima, T., Ueno, M., Ohkawa, S., … Matsuo, K. (2020). Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. *Nature Communications*, *11*(1), 3175. https://doi.org/10.1038/s41467-020-16711-w

Liu, J., Jiang, W., Zhao, K., Wang, H., Zhou, T., Bai, W., Wang, X., Zhao, T., Huang, C., Gao, S., Qin, T., Yu, W., Yang, B., Li, X., Fu, D., Tan, W., Yang, S., Ren, H., & Hao, J. (2019). Tumoral EHF predicts the efficacy of anti-PD1 therapy in pancreatic ductal adenocarcinoma. *The Journal of Experimental Medicine*, *216*(3), 656–673. https://doi.org/10.1084/jem.20180749

Liu, L.-B., Huang, J., Zhong, J.-P., Ye, G.-L., Xue, L., Zhou, M.-H., Huang, G., & Li, S.-J. (2018). High Expression of CCDC34 Is Associated with Poor Survival in Cervical Cancer Patients. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, *24*, 8383–8390. https://doi.org/10.12659/MSM.913346

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *American Journal of Human Genetics*, *104*(3), 410–421. https://doi.org/10.1016/j.ajhg.2019.01.002

Löw, M., Stegmaier, C., Ziegler, H., Rothenbacher, D., Brenner, H., & ESTHER study. (2004). [Epidemiological investigations of the chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population (ESTHER study)]. *Deutsche Medizinische Wochenschrift (1946)*, *129*(49), 2643–2647. https://doi.org/10.1055/s-2004-836089

Lu, Y., Corradi, C., Gentiluomo, M., López de Maturana, E., Theodoropoulos, G. E., Roth, S., Maiello, E., Morelli, L., Archibugi, L., Izbicki, J. R., Sarlós, P., Kiudelis, V., Oliverius, M., Aoki, M. N., Vashist, Y., van Eijck, C. H. J., Gazouli, M., Talar-Wojnarowska, R., Mambrini, A., … Canzian, F. (2021). Association of Genetic Variants Affecting microRNAs and Pancreatic Cancer Risk. *Frontiers in Genetics*, *12*, 693933. https://doi.org/10.3389/fgene.2021.693933

Lu, Y., Ek, W. E., Whiteman, D., Vaughan, T. L., Spurdle, A. B., Easton, D. F., Pharoah, P. D., Thompson, D. J., Dunning, A. M., Hayward, N. K., Chenevix-Trench, G., Q-MEGA and AMFS Investigators, ANECS-SEARCH, UKOPS-SEARCH, BEACON Consortium, & Macgregor, S. (2014). Most common "sporadic" cancers have a significant germline genetic component. *Human Molecular Genetics*, *23*(22), 6112–6118. https://doi.org/10.1093/hmg/ddu312

Lu, Y., Gentiluomo, M., Lorenzo-Bermejo, J., Morelli, L., Obazee, O., Campa, D., & Canzian, F. (2020). Mendelian randomisation study of the effects of known and putative risk factors on pancreatic cancer. *Journal of Medical Genetics*, *57*(12), 820–828. https://doi.org/10.1136/jmedgenet-2019-106200

Lunetta, K. L. (2008). Genetic Association Studies. *Circulation*, *118*(1), 96–101. https://doi.org/10.1161/CIRCULATIONAHA.107.700401

Luo, W., Tao, J., Zheng, L., & Zhang, T. (2020). Current epidemiology of pancreatic cancer: Challenges and opportunities. *Chinese Journal of Cancer Research = Chung-Kuo Yen Cheng Yen Chiu*, *32*(6), 705–719. https://doi.org/10.21147/j.issn.1000-9604.2020.06.04

Lynch, S. M., Vrieling, A., Lubin, J. H., Kraft, P., Mendelsohn, J. B., Hartge, P., Canzian, F., Steplowski, E., Arslan, A. A., Gross, M., Helzlsouer, K., Jacobs, E. J., LaCroix, A., Petersen, G., Zheng, W., Albanes, D., Amundadottir, L., Bingham, S. A., Boffetta, P., … Stolzenberg-Solomon, R. Z. (2009). Cigarette smoking and pancreatic cancer: A pooled analysis from the pancreatic cancer cohort consortium. *American Journal of Epidemiology*, *170*(4), 403–413. https://doi.org/10.1093/aje/kwp134

Macgregor, S., & Khan, I. A. (2006). GAIA: An easy-to-use web-based application for interaction analysis of case-control data. *BMC Medical Genetics*, *7*, 34. https://doi.org/10.1186/1471-2350-7-34

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, *5*(2), e1000384. https://doi.org/10.1371/journal.pgen.1000384

Magnon, C. (2015). Role of the autonomic nervous system in tumorigenesis and metastasis. *Molecular & Cellular Oncology*, *2*(2), e975643. https://doi.org/10.4161/23723556.2014.975643

Mahawithitwong, P., Ohuchida, K., Ikenaga, N., Fujita, H., Zhao, M., Kozono, S., Shindo, K.,

Ohtsuka, T., Aishima, S., Mizumoto, K., & Tanaka, M. (2013). Kindlin-1 expression is involved in migration and invasion of pancreatic cancer. *International Journal of Oncology*, *42*(4), 1360–1366. https://doi.org/10.3892/ijo.2013.1838

Maisonneuve, P., & Lowenfels, A. B. (2015). Risk factors for pancreatic cancer: A summary review of meta-analytical studies. *International Journal of Epidemiology*, *44*(1), 186–198. https://doi.org/10.1093/ije/dyu240

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, J., Finch, K. L., Stevens, J. F., Livak, K. J., Slotterbeck, B. D., Slifer, S. H., Warren, L. L., Conneally, P. M., Schmechel, D. E., Purvis, I., Pericak-Vance, M. A., Roses, A. D., & Vance, J. M. (2000). SNPing away at complex diseases: Analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *American Journal of Human Genetics*, *67*(2), 383–394. https://doi.org/10.1086/303003

Masoudi, S., Momayez Sanat, Z., Mahmud Saleh, A., Nozari, N., Ghamarzad, N., & Pourshams, A. (2017). Menstrual and Reproductive Factors and Risk of Pancreatic Cancer in Women. *Middle East Journal of Digestive Diseases*, *9*(3), 146–149. https://doi.org/10.15171/mejdd.2017.65

Matsubayashi, H. (2011). Familial pancreatic cancer and hereditary syndromes: Screening strategy for high-risk individuals. *Journal of Gastroenterology*, *46*(11), 1249–1259. https://doi.org/10.1007/s00535-011-0457-z

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

Mazerbourg, S., Bouley, D. M., Sudo, S., Klein, C. A., Zhang, J. V., Kawamura, K., Goodrich, L. V., Rayburn, H., Tessier-Lavigne, M., & Hsueh, A. J. W. (2004). Leucine-rich repeat-containing, G protein-coupled receptor 4 null mice exhibit intrauterine growth retardation associated with embryonic and perinatal lethality. *Molecular Endocrinology (Baltimore, Md.)*, *18*(9), 2241–2254. https://doi.org/10.1210/me.2004-0133

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), 1097–1103. https://doi.org/10.1038/s41588-021-00870-7

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

McEwen, B. S., & Gianaros, P. J. (2011). Stress- and Allostasis-Induced Brain Plasticity. *Annual Review of Medicine*, *62*(Volume 62, 2011), 431–445. https://doi.org/10.1146/annurev-med-052209-100430

McMenamin, Ú. C., McCain, S., & Kunzmann, A. T. (2017). Do smoking and alcohol behaviours influence GI cancer survival? *Best Practice & Research. Clinical Gastroenterology*, *31*(5), 569–577. https://doi.org/10.1016/j.bpg.2017.09.015

Meyer, H. V. (2020). *meyer-lab-cshl/plinkQC: plinkQC 0.3.2 (v0.3.2). Zenodo.* https://doi.org/10.5281/zenodo.3934294

Midha, S., Chawla, S., & Garg, P. K. (2016). Modifiable and non-modifiable risk factors for pancreatic cancer: A review. *Cancer Letters*, *381*(1), 269–277. https://doi.org/10.1016/j.canlet.2016.07.022

Millstein, J., Siegmund, K. D., Conti, D. V., & Gauderman, W. J. (2005). Identifying susceptibility genes by using joint tests of association and linkage and accounting for epistasis. *BMC Genetics*, *6 Suppl 1*(Suppl 1), S147. https://doi.org/10.1186/1471-2156-6-S1-S147

Millstein, J., Zhang, B., Zhu, J., & Schadt, E. E. (2009). Disentangling molecular relationships with a

causal inference test. *BMC Genetics*, *10*, 23. https://doi.org/10.1186/1471-2156-10-23
Mitsufuji, S., Iwagami, Y., Kobayashi, S., Sasaki, K., Yamada, D., Tomimaru, Y., Akita, H., Asaoka, T., Noda, T., Gotoh, K., Takahashi, H., Tanemura, M., Doki, Y., & Eguchi, H. (2022). Inhibition of Clusterin Represses Proliferation by Inducing Cellular Senescence in Pancreatic Cancer. *Annals of Surgical Oncology*, *29*(8), 4937–4946. https://doi.org/10.1245/s10434-022-11668-0
Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics: EJHG*, *25*(7), 869–876. https://doi.org/10.1038/ejhg.2017.51
Modi, S., Kir, D., Banerjee, S., & Saluja, A. (2016). Control of Apoptosis in Treatment and Biology of Pancreatic Cancer. *Journal of Cellular Biochemistry*, *117*(2), 279–288. https://doi.org/10.1002/jcb.25284
Momozawa, Y., & Mizukami, K. (2021). Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics*, *66*(1), 11–23. https://doi.org/10.1038/s10038-020-00845-2
Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, *34*(2), 188–193. https://doi.org/10.1002/gepi.20450
Morton N E. (1982). *Outline of genetic epidemiology*. S. Karger.
Mudge, J. M., & Harrow, J. (2016). The state of play in higher eukaryote gene annotation. *Nature Reviews. Genetics*, *17*(12), 758–772. https://doi.org/10.1038/nrg.2016.119
Mueller, S. H., Lai, A. G., Valkovskaya, M., Michailidou, K., Bolla, M. K., Wang, Q., Dennis, J., Lush, M., Abu-Ful, Z., Ahearn, T. U., Andrulis, I. L., Anton-Culver, H., Antonenkova, N. N., Arndt, V., Aronson, K. J., Augustinsson, A., Baert, T., Freeman, L. E. B., Beckmann, M. W., … Kuchenbaecker, K. B. (2023). Aggregation tests identify new gene associations with breast cancer in populations with diverse ancestry. *Genome Medicine*, *15*(1), 7. https://doi.org/10.1186/s13073-022-01152-5
Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., … Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature*, *593*(7858), 238–243. https://doi.org/10.1038/s41586-021-03446-x
Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., & Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, *7*(3), e1001322. https://doi.org/10.1371/journal.pgen.1001322
Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888. https://doi.org/10.1371/journal.pgen.1000888
Noel, M., & Fiscella, K. (2019). Disparities in Pancreatic Cancer Treatment and Outcomes. *Health Equity*, *3*(1), 532–540. https://doi.org/10.1089/heq.2019.0057
Nussbaum, F. (2015). *Thompson and Thompson Genetics in Medicine E-Book: Thompson and Thompson Genetics in Medicine E-Book* (8th ed). Elsevier.
Onda, T., Uzawa, K., Nakashima, D., Saito, K., Iwadate, Y., Seki, N., Shibahara, T., & Tanzawa, H. (2007). Lin-7C/VELI3/MALS-3: An essential component in metastasis of human squamous cell carcinoma. *Cancer Research*, *67*(20), 9643–9648. https://doi.org/10.1158/0008-5472.CAN-07-1911
Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, *33*(6), 497–507. https://doi.org/10.1002/gepi.20402
Panoutsopoulou, K., & Wheeler, E. (2018). Key Concepts in Genetic Epidemiology. In E. Evangelou (Ed.), *Genetic Epidemiology: Methods and Protocols* (pp. 7–24). Springer. https://doi.org/10.1007/978-1-4939-7868-7_2
Park, W., Chawla, A., & O'Reilly, E. M. (2021). Pancreatic Cancer: A Review. *JAMA*, *326*(9), 851–862. https://doi.org/10.1001/jama.2021.13027
Patnala, R., Clements, J., & Batra, J. (2013). Candidate gene association studies: A comprehensive

guide to useful in silicotools. *BMC Genetics*, *14*(1), 39. https://doi.org/10.1186/1471-2156-14-39

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. Scopus. https://doi.org/10.1371/journal.pgen.0020190

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews. Genetics*, *14*(4), 288–295. https://doi.org/10.1038/nrg3458

Petersen, G. M., Amundadottir, L., Fuchs, C. S., Kraft, P., Stolzenberg-Solomon, R. Z., Jacobs, K. B., Arslan, A. A., Bueno-de-Mesquita, H. B., Gallinger, S., Gross, M., Helzlsouer, K., Holly, E. A., Jacobs, E. J., Klein, A. P., LaCroix, A., Li, D., Mandelson, M. T., Olson, S. H., Risch, H. A., … Chanock, S. J. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*, *42*(3), 224–228. https://doi.org/10.1038/ng.522

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, *94*(4), 559–573. https://doi.org/10.1016/j.ajhg.2014.03.004

Pierce, B. L., & Ahsan, H. (2011). Genome-Wide "Pleiotropy Scan" Identifies HNF1A Region as a Novel Pancreatic Cancer Susceptibility Locus. *Cancer Research*, *71*(13), 4352–4358. https://doi.org/10.1158/0008-5472.CAN-11-0124

Pistoni, L., Gentiluomo, M., Lu, Y., López de Maturana, E., Hlavac, V., Vanella, G., Darvasi, E., Milanetto, A. C., Oliverius, M., Vashist, Y., Di Leo, M., Mohelnikova-Duchonova, B., Talar-Wojnarowska, R., Gheorghe, C., Petrone, M. C., Strobel, O., Arcidiacono, P. G., Vodickova, L., Szentesi, A., … Campa, D. (2021). Associations between pancreatic expression quantitative traits and risk of pancreatic ductal adenocarcinoma. *Carcinogenesis*, *42*(8), 1037–1045. https://doi.org/10.1093/carcin/bgab057

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Principe, D. R., Timbers, K. E., Atia, L. G., Koch, R. M., & Rana, A. (2021). TGFβ Signaling in the Pancreatic Tumor Microenvironment. Cancers, 13(20). https://doi.org/10.3390/cancers13205086

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics*, *69*(1), 1–14. https://doi.org/10.1086/321275

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., & Willer, C. J. (2010). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics (Oxford, England)*, *26*(18), 2336–2337. https://doi.org/10.1093/bioinformatics/btq419

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Qi, W., Shao, F., & Huang, Q. (2017). Expression of Coiled-Coil Domain Containing 34 (CCDC34) and its Prognostic Significance in Pancreatic Adenocarcinoma. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, *23*, 6012–6018. https://doi.org/10.12659/msm.907951

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., & Matrisian, L. M. (2014). Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States. Cancer Research, 74(11), 2913–2921. https://doi.org/10.1158/0008-5472.CAN-14-0155

Raimondi, S., Lowenfels, A. B., Morselli-Labate, A. M., Maisonneuve, P., & Pezzilli, R. (2010). Pancreatic cancer in chronic pancreatitis; aetiology, incidence, and early detection. *Best Practice & Research. Clinical Gastroenterology*, *24*(3), 349–358. https://doi.org/10.1016/j.bpg.2010.02.007

Rawla, P., Sunkara, T., & Gaduputi, V. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World Journal of Oncology*, *10*(1), 10–27. https://doi.org/10.14740/wjon1166

Reid-Lombardo, K. M., & Petersen, G. M. (2010). Understanding genetic epidemiologic association

studies Part 1: Fundamentals. *Surgery*, *147*(4), 469–474. https://doi.org/10.1016/j.surg.2009.10.026

Rhee, S. Y., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews. Genetics*, *9*(7), 509–515. https://doi.org/10.1038/nrg2363

Riboli, E., Hunt, K. J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U. R., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., … Saracci, R. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): Study populations and data collection. *Public Health Nutrition*, *5*(6b), 1113–1124. https://doi.org/10.1079/PHN2002394

Risch, H. A., Lu, L., Wang, J., Zhang, W., Ni, Q., Gao, Y.-T., & Yu, H. (2013). ABO blood group and risk of pancreatic cancer: A study in Shanghai and meta-analysis. *American Journal of Epidemiology*, *177*(12), 1326–1337. https://doi.org/10.1093/aje/kws458

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Roberts, N. J., Jiao, Y., Yu, J., Kopelovich, L., Petersen, G. M., Bondy, M. L., Gallinger, S., Schwartz, A. G., Syngal, S., Cote, M. L., Axilbund, J., Schulick, R., Ali, S. Z., Eshleman, J. R., Velculescu, V. E., Goggins, M., Vogelstein, B., Papadopoulos, N., Hruban, R. H., … Klein, A. P. (2012). ATM Mutations in Patients with Hereditary Pancreatic Cancer. *Cancer Discovery*, *2*(1), 41–46. https://doi.org/10.1158/2159-8290.CD-11-0194

Roberts, N. J., Norris, A. L., Petersen, G. M., Bondy, M. L., Brand, R., Gallinger, S., Kurtz, R. C., Olson, S. H., Rustgi, A. K., Schwartz, A. G., Stoffel, E., Syngal, S., Zogopoulos, G., Ali, S. Z., Axilbund, J., Chaffee, K. G., Chen, Y.-C., Cote, M. L., Childs, E. J., … Klein, A. P. (2016). Whole Genome Sequencing Defines the Genetic Heterogeneity of Familial Pancreatic Cancer. *Cancer Discovery*, *6*(2), 166–175. https://doi.org/10.1158/2159-8290.CD-15-0402

Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genetics*, *16*(11), e1009049. https://doi.org/10.1371/journal.pgen.1009049

Russ, D., Williams, J. A., Cardoso, V. R., Bravo-Merodio, L., Pendleton, S. C., Aziz, F., Acharjee, A., & Gkoutos, G. V. (2022). Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models. PLOS ONE, 17(2), e0263390. https://doi.org/10.1371/journal.pone.0263390

Sahebnasagh, A., Saghafi, F., & Negintaji, S. (2022). Nitric Oxide and Immune Responses in Cancer: Searching for New Therapeutic Strategies. Curr Med Chem, 29(9), 1561–1595. https://doi.org/10.2174/0929867328666210707194543

Sang, W., Zhou, Y., & Chen, H. (2024). Receptor-interacting Protein Kinase 2 Is an Immunotherapy Target in Pancreatic Cancer. Cancer Discov, 14(2), 326–347. https://doi.org/10.1158/2159-8290.CD-23-0584

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews. Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Schijven, D., Soheili-Nezhad, S., Fisher, S. E., & Francks, C. (2024). Exome-wide analysis implicates rare protein-altering variants in human handedness. *Nature Communications*, *15*, 2632. https://doi.org/10.1038/s41467-024-46277-w

Selsted, M. E., & Ouellette, A. J. (2005). Mammalian defensins in the antimicrobial immune response. *Nature Immunology*, *6*(6), 551–557. https://doi.org/10.1038/ni1206

Shi, T., Min, M., Sun, C., Zhang, Y., Liang, M., & Sun, Y. (2020). Does insomnia predict a high risk of cancer? A systematic review and meta-analysis of cohort studies. Journal of Sleep Research, 29(1), e12876. https://doi.org/10.1111/jsr.12876

Shimizu, Y., Nakamura, K., Kikuchi, M., Ukawa, S., Nakamura, K., Okada, E., Imae, A., Nakagawa, T., Yamamura, R., Tamakoshi, A., & Ayabe, T. (2022). Lower human defensin 5 in elderly people compared to middle-aged is associated with differences in the intestinal microbiota

composition: The DOSANCO Health Study. *GeroScience*, *44*(2), 997–1009. https://doi.org/10.1007/s11357-021-00398-y

Shinawi, M., Sahoo, T., Maranda, B., Skinner, S. A., Skinner, C., Chinault, C., Zascavage, R., Peters, S. U., Patel, A., Stevenson, R. E., & Beaudet, A. L. (2011). 11p14.1 microdeletions associated with ADHD, autism, developmental delay, and obesity. *American Journal of Medical Genetics. Part A*, *155A*(6), 1272–1280. https://doi.org/10.1002/ajmg.a.33878

Shindo, K., Yu, J., Suenaga, M., Fesharakizadeh, S., Cho, C., Macgregor-Das, A., Siddiqui, A., Witmer, P. D., Tamura, K., Song, T. J., Navarro Almario, J. A., Brant, A., Borges, M., Ford, M., Barkley, T., He, J., Weiss, M. J., Wolfgang, C. L., Roberts, N. J., … Goggins, M. (2017). Deleterious Germline Mutations in Patients With Apparently Sporadic Pancreatic Adenocarcinoma. *Journal of Clinical Oncology*, *35*(30), 3382–3390. https://doi.org/10.1200/JCO.2017.72.3502

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians*, *71*(1), 7–33. https://doi.org/10.3322/caac.21654

Somasundaram, V., Basudhar, D., & Bharadwaj, G. (2019). Molecular Mechanisms of Nitric Oxide in Cancer Progression. Signal Transduction, and Metabolism. Antioxid Redox Signal, 30(8), 1124–1143. https://doi.org/10.1089/ars.2018.7527

Stolzenberg-Solomon, R. Z., Cross, A. J., Silverman, D. T., Schairer, C., Thompson, F. E., Kipnis, V., Subar, A. F., Hollenbeck, A., Schatzkin, A., & Sinha, R. (2007). Meat and meat-mutagen intake and pancreatic cancer risk in the NIH-AARP cohort. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *16*(12), 2664–2675. https://doi.org/10.1158/1055-9965.EPI-07-0378

Sun, J., Zheng, Y., & Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, *37*(4), 334–344. https://doi.org/10.1002/gepi.21717

Sun, W., Zhou, H., Cheng, M., Zhuang, S., & Qiu, Z. (2020). Association between Socioeconomic Status and One-Month Mortality after Surgery in 20 Primary Solid Tumors: A Pan-Cancer Analysis. *Journal of Cancer*, *11*(18), 5449–5455. https://doi.org/10.7150/jca.46088

Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Másson, G., Holm, H., Kong, A., Thorsteinsdottir, U., Sulem, P., Gudbjartsson, D. F., & Stefansson, K. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetics*, *48*(3), 314–317. https://doi.org/10.1038/ng.3507

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2019). *Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program* (p. 563866). bioRxiv. https://doi.org/10.1101/563866

Tang, Y., Liu, F., Zheng, C., Sun, S., & Jiang, Y. (2012). Knockdown of clusterin sensitizes pancreatic cancer cells to gemcitabine chemotherapy by ERK1/2 inactivation. *Journal of Experimental & Clinical Cancer Research : CR*, *31*(1), 73. https://doi.org/10.1186/1756-9966-31-73

Teare, M. D. (Ed.). (2011). *Genetic Epidemiology* (Vol. 713). Humana Press. https://doi.org/10.1007/978-1-60327-416-6

Thompson, E. A. (2013). Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*, *194*(2), 301–326. https://doi.org/10.1534/genetics.112.148825

Tobi, M., Kim, M., Weinstein, D. H., Rambus, M. A., Hatfield, J., Adsay, N. V., Levi, E., Evans, D., Lawson, M. J., & Fligiel, S. (2013). Prospective markers for early diagnosis and prognosis of sporadic pancreatic ductal adenocarcinoma. *Digestive Diseases and Sciences*, *58*(3), 744–750. https://doi.org/10.1007/s10620-012-2387-x

Travagli, R. A., Hermann, G. E., Browning, K. N., & Rogers, R. C. (2006). Brainstem circuits regulating gastric function. *Annual Review of Physiology*, *68*, 279–305. https://doi.org/10.1146/annurev.physiol.68.040504.094635

Travagli, T. B. and R. A. (2016). Neural Control of the Pancreas. *Pancreapedia: The Exocrine Pancreas Knowledge Base*. https://doi.org/10.3998/panc.2016.27

Tsai, H.-J., & Chang, J. S. (2019). Environmental Risk Factors of Pancreatic Cancer. *Journal of Clinical Medicine*, *8*(9), 1427. https://doi.org/10.3390/jcm8091427

Ueki, M., & Cordell, H. J. (2012). Improved statistics for genome-wide interaction analysis. *PLoS*

Genetics, *8*(4), e1002625. https://doi.org/10.1371/journal.pgen.1002625
Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9
Ünal, P., Lu, Y., Bueno-de-Mesquita, B., van Eijck, C. H. J., Talar-Wojnarowska, R., Szentesi, A., Gazouli, M., Kreivenaite, E., Tavano, F., Malecka-Wojciesko, E., Erőss, B., Oliverius, M., Bunduc, S., Nóbrega Aoki, M., Vodickova, L., Boggi, U., Giaccherini, M., Kondrackiene, J., Chammas, R., … Canzian, F. (2024). Polymorphisms in transcription factor binding sites and enhancer regions and pancreatic ductal adenocarcinoma risk. *Human Genomics*, *18*(1), 12. https://doi.org/10.1186/s40246-024-00576-x
Urbanowicz, R. J., Kiralis, J., Fisher, J. M., & Moore, J. H. (2012). Predicting the difficulty of pure, strict, epistatic models: Metrics for simulated model selection. *BioData Mining*, *5*(1), 15. https://doi.org/10.1186/1756-0381-5-15
US Preventive Services Task Force, Owens, D. K., Davidson, K. W., Krist, A. H., Barry, M. J., Cabana, M., Caughey, A. B., Curry, S. J., Doubeni, C. A., Epling, J. W., Kubik, M., Landefeld, C. S., Mangione, C. M., Pbert, L., Silverstein, M., Simon, M. A., Tseng, C.-W., & Wong, J. B. (2019). Screening for Pancreatic Cancer: US Preventive Services Task Force Reaffirmation Recommendation Statement. *JAMA*, *322*(5), 438–444. https://doi.org/10.1001/jama.2019.10232
van Andel, H., Ren, Z., Koopmans, I., Joosten, S. P. J., Kocemba, K. A., de Lau, W., Kersten, M. J., de Bruin, A. M., Guikema, J. E. J., Clevers, H., Spaargaren, M., & Pals, S. T. (2017). Aberrantly expressed LGR4 empowers Wnt signaling in multiple myeloma by hijacking osteoblast-derived R-spondins. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(2), 376–381. https://doi.org/10.1073/pnas.1618650114
Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03
Vigeland, M. D. (2020). Relatedness coefficients in pedigrees with inbred founders. *Journal of Mathematical Biology*, *81*(1), 185–207. https://doi.org/10.1007/s00285-020-01505-x
Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005
Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, T.-O. for P. M. C., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., Kooperberg, C., Roden, D., Darbar, D., Arnett, D. K., Regan, E. A., Boerwinkle, E., Rotter, J. I., Allison, M. A., … Visscher, P. M. (2019). *Recovery of trait heritability from whole genome sequence data* (p. 588020). bioRxiv. https://doi.org/10.1101/588020
Walsh, N., Zhang, H., Hyland, P. L., Yang, Q., Mocci, E., Zhang, M., Childs, E. J., Collins, I., Wang, Z., Arslan, A. A., Beane-Freeman, L., Bracci, P. M., Brennan, P., Canzian, F., Duell, E. J., Gallinger, S., Giles, G. G., Goggins, M., Goodman, G. E., … Stolzenberg-Solomon, R. Z. (2019). Agnostic Pathway/Gene Set Analysis of Genome-Wide Association Data Identifies Associations for Pancreatic Cancer. *JNCI: Journal of the National Cancer Institute*, *111*(6), 557–567. https://doi.org/10.1093/jnci/djy155
Wang, J., Yang, S., & He, P. (2016). Endothelial Nitric Oxide Synthase Traffic Inducer (NOSTRIN) is a Negative Regulator of Disease Aggressiveness in Pancreatic Cancer. Clin Cancer Res Off J Am Assoc Cancer Res, 22(24), 5992–6001. https://doi.org/10.1158/1078-0432.CCR-16-0511
Wang, L., Ai, M., Nie, M., Zhao, L., Deng, G., Hu, S., Han, Y., Zeng, W., Wang, Y., Yang, M., & Wang, S. (2020). EHF promotes colorectal carcinoma progression by activating TGF-β1 transcription and canonical TGF-β signaling. *Cancer Science*, *111*(7), 2310–2324. https://doi.org/10.1111/cas.14444
Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*, *55*, 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008
Wang, Z., Shen, D., & Parsons, D. W. (2004). Mutational Analysis of the Tyrosine Phosphatome in Colorectal Cancers. Science, 304(5674), 1164–1166. https://doi.org/10.1126/science.1096096

Wang, Z., Yin, P., Sun, Y., Na, L., Gao, J., Wang, W., & Zhao, C. (2020). LGR4 maintains HGSOC cell epithelial phenotype and stem-like traits. *Gynecologic Oncology*, *159*(3), 839–849. https://doi.org/10.1016/j.ygyno.2020.09.020

Weale, M. E. (2010). Quality Control for Genome-Wide Association Studies. In M. R. Barnes & G. Breen (Eds.), *Genetic Variation: Methods and Protocols* (pp. 341–372). Humana Press. https://doi.org/10.1007/978-1-60327-367-1_19

Wei, W.-H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews. Genetics*, *15*(11), 722–733. https://doi.org/10.1038/nrg3747

Weissenkampen, J. D., Jiang, Y., Eckert, S., Jiang, B., Li, B., & Liu, D. J. (2019). Methods for the analysis and interpretation for rare variants associated with complex traits. *Current Protocols in Human Genetics*, *101*(1), e83. https://doi.org/10.1002/cphg.83

Wellcome Trust Case Control Consortium, Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301. https://doi.org/10.1038/ng.2435

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190. https://doi.org/10.1093/bioinformatics/btq340

Wittke-Thompson, J. K., Pluzhnikov, A., & Cox, N. J. (2005). Rational Inferences about Departures from Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, *76*(6), 967–986. https://doi.org/10.1086/430507

Wolpin, B. M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G. M., Wang, Z., Arslan, A. A., Beane-Freeman, L., Bracci, P. M., Buring, J., Canzian, F., Duell, E. J., Gallinger, S., Giles, G. G., Goodman, G. E., Goodman, P. J., Jacobs, E. J., Kamineni, A., Klein, A. P., … Amundadottir, L. T. (2014). Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nature Genetics*, *46*(9), 994–1000. https://doi.org/10.1038/ng.3052

Wood, L. D., & Hruban, R. H. (2012). Pathology and molecular genetics of pancreatic neoplasms. *Cancer Journal (Sudbury, Mass.)*, *18*(6), 492–501. https://doi.org/10.1097/PPO.0b013e31827459b6

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., & Romano, L. A. (2003). The Evolution of Transcriptional Regulation in Eukaryotes. *Molecular Biology and Evolution*, *20*(9), 1377–1419. https://doi.org/10.1093/molbev/msg140

Wu, C., Miao, X., Huang, L., Che, X., Jiang, G., Yu, D., Yang, X., Cao, G., Hu, Z., Zhou, Y., Zuo, C., Wang, C., Zhang, X., Zhou, Y., Yu, X., Dai, W., Li, Z., Shen, H., Liu, L., … Lin, D. (2012). Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nature Genetics*, *44*(1), 62–66. https://doi.org/10.1038/ng.1020

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, *86*(6), 929–942. https://doi.org/10.1016/j.ajhg.2010.05.002

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Xiao, M., Wang, Y., & Gao, Y. (2013). Association between Helicobacter pylori infection and pancreatic cancer development: A meta-analysis. *PloS One*, *8*(9), e75559. https://doi.org/10.1371/journal.pone.0075559

Xu, M., Jung, X., Hines, O. J., Eibl, G., & Chen, Y. (2018). Obesity and Pancreatic Cancer: Overview of Epidemiology and Potential Prevention by Weight Loss. *Pancreas*, *47*(2), 158–162. https://doi.org/10.1097/MPA.0000000000000974

Yan, Y., Zhou, B., & Qian, C. (2022). Receptor-interacting protein kinase 2 (RIPK2) stabilizes c-Myc and is a therapeutic target in prostate cancer metastasis. Nat Commun, 13(1). https://doi.org/10.1038/s41467-022-28340-6

Yang, E., & Shen, J. (2021). The roles and functions of Paneth cells in Crohn's disease: A critical review. *Cell Proliferation*, *54*(1), e12958. https://doi.org/10.1111/cpr.12958

Yu, Y., Chang, K., Chen, J.-S., Bohlender, R. J., Fowler, J., Zhang, D., Huang, M., Chang, P., Li, Y.,

Wong, J., Wang, H., Gu, J., Wu, X., Schildkraut, J., Cannon-Albright, L., Ye, Y., Zhao, H., Hildebrandt, M. A. T., Permuth, J. B., … Huff, C. D. (2022). A whole-exome case-control association study to characterize the contribution of rare coding variation to pancreatic cancer risk. *Human Genetics and Genomics Advances*, *3*(1), 100078. https://doi.org/10.1016/j.xhgg.2021.100078

Yuan, C., Morales-Oyarvide, V., Babic, A., Clish, C. B., Kraft, P., Bao, Y., Qian, Z. R., Rubinson, D. A., Ng, K., Giovannucci, E. L., Ogino, S., Stampfer, M. J., Gaziano, J. M., Sesso, H. D., Cochrane, B. B., Manson, J. E., Fuchs, C. S., & Wolpin, B. M. (2017). Cigarette Smoking and Pancreatic Cancer Survival. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *35*(16), 1822–1828. https://doi.org/10.1200/JCO.2016.71.2026

Zaykin, D. V., & Zhivotovsky, L. A. (2005). Ranks of genuine associations in whole-genome scans. *Genetics*, *171*(2), 813–823. https://doi.org/10.1534/genetics.105.044206

Zhang, H., Wheeler, W., Hyland, P. L., Yang, Y., Shi, J., Chatterjee, N., & Yu, K. (2016). A Powerful Procedure for Pathway-Based Meta-analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations. *PLoS Genetics*, *12*(6), e1006122. https://doi.org/10.1371/journal.pgen.1006122

Zhang, J., Baran, J., & Cros, A. (2011). International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. Database J Biol Databases Curation, 2011:bar026. https://doi.org/10.1093/database/bar026

Zhang, J.-J., Zhu, Y., Xie, K.-L., Peng, Y.-P., Tao, J.-Q., Tang, J., Li, Z., Xu, Z.-K., Dai, C.-C., Qian, Z.-Y., Jiang, K.-R., Wu, J.-L., Gao, W.-T., Du, Q., & Miao, Y. (2014). Yin Yang-1 suppresses invasion and metastasis of pancreatic ductal adenocarcinoma by downregulating MMP10 in a MUC4/ErbB2/p38/MEF2C-dependent mechanism. *Molecular Cancer*, *13*, 130. https://doi.org/10.1186/1476-4598-13-130

Zhang, M., Wang, Z., Obazee, O., Jia, J., Childs, E. J., Hoskins, J., Figlioli, G., Mocci, E., Collins, I., Chung, C. C., Hautman, C., Arslan, A. A., Beane-Freeman, L., Bracci, P. M., Buring, J., Duell, E. J., Gallinger, S., Giles, G. G., Goodman, G. E., … Amundadottir, L. T. (2016). Three new pancreatic cancer susceptibility signals identified on chromosomes 1q32.1, 5p15.33 and 8q24.21. *Oncotarget*, *7*(41), 66328–66343. https://doi.org/10.18632/oncotarget.11041

Zhang, W., & Wang, Y. (2022). Activation of RIPK2-mediated NOD1 signaling promotes proliferation and invasion of ovarian cancer cells via NF-κB pathway. Histochem Cell Biol, 157(2), 173–182. https://doi.org/10.1007/s00418-021-02055-z

Zhang, Y. D., Hurson, A. N., Zhang, H., Choudhury, P. P., Easton, D. F., Milne, R. L., Simard, J., Hall, P., Michailidou, K., Dennis, J., Schmidt, M. K., Chang-Claude, J., Gharahkhani, P., Whiteman, D., Campbell, P. T., Hoffmeister, M., Jenkins, M., Peters, U., Hsu, L., … Garcia-Closas, M. (2020). Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications*, *11*(1), 3353. https://doi.org/10.1038/s41467-020-16483-3

Zhang, Y., Qi, G., Park, J.-H., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics*, *50*(9), 1318–1326. https://doi.org/10.1038/s41588-018-0193-x

Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., Laurie, C., & Levine, D. (2017). SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics (Oxford, England)*, *33*(15), 2251–2257. https://doi.org/10.1093/bioinformatics/btx145

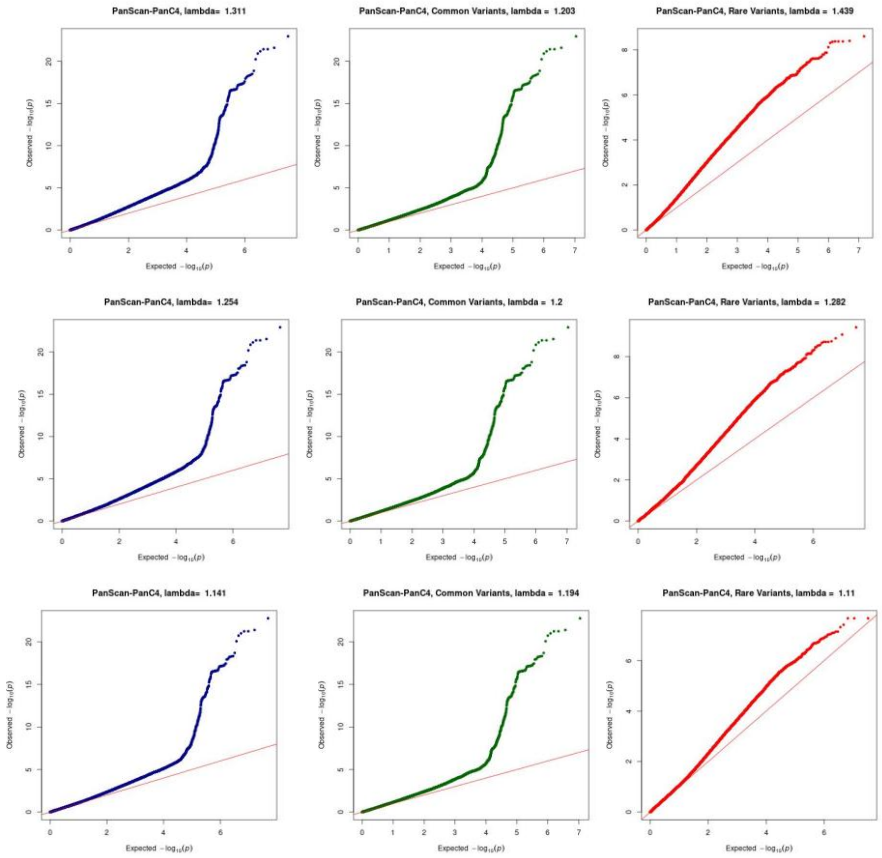
Zhou, H., Arapoglou, T., Li, X., Li, Z., Zheng, X., Moore, J., Asok, A., Kumar, S., Blue, E. E., Buyske, S., Cox, N., Felsenfeld, A., Gerstein, M., Kenny, E., Li, B., Matise, T., Philippakis, A., Rehm, H. L., Sofia, H. J., … Lin, X. (2023). FAVOR: Functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Research*, *51*(D1), D1300–D1311. https://doi.org/10.1093/nar/gkac966

Zhou, Q., & Melton, D. A. (2018). Pancreas regeneration. *Nature*, *557*(7705), 351. https://doi.org/10.1038/s41586-018-0088-0

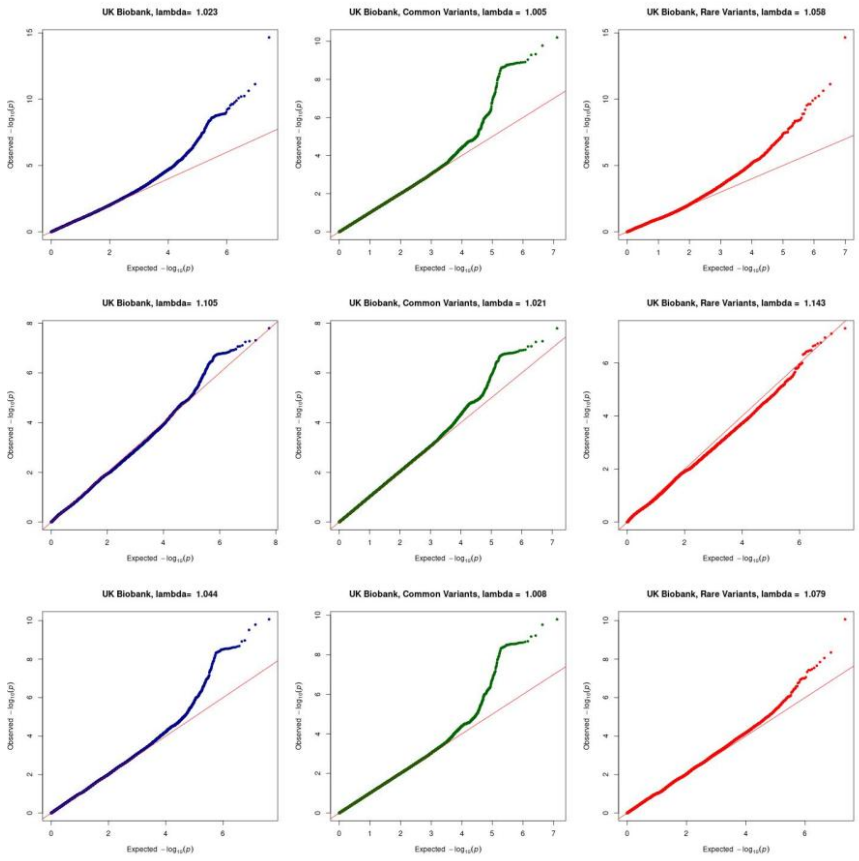
Zhou, T., Liu, J., Xie, Y., Yuan, S., Guo, Y., Bai, W., Zhao, K., Jiang, W., Wang, H., Wang, H., Zhao, T., Huang, C., Gao, S., Wang, X., Yang, S., & Hao, J. (2022). ESE3/EHF, a promising target of rosiglitazone, suppresses pancreatic cancer stemness by downregulating CXCR4. *Gut*,

71(2), 357–371. https://doi.org/10.1136/gutjnl-2020-321952
Zondervan, K. T., & Cardon, L. R. (2007). Designing candidate gene and genome-wide case-control association studies. *Nature Protocols*, *2*(10), 2492–2501. https://doi.org/10.1038/nprot.2007.366
Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(4), 1193–1198. https://doi.org/10.1073/pnas.1119675109
Zwir, I., Del-Val, C., Hintsanen, M., Cloninger, K. M., Romero-Zaliz, R., Mesa, A., Arnedo, J., Salas, R., Poblete, G. F., Raitoharju, E., Raitakari, O., Keltikangas-Järvinen, L., de Erausquin, G. A., Tattersall, I., Lehtimäki, T., & Cloninger, C. R. (2022). Evolution of genetic networks for human creativity. *Molecular Psychiatry*, *27*(1), 354–376. https://doi.org/10.1038/s41380-021-01097-y

9 SUPPLEMENTARY DATA



Supplementary Figure 1: Quantile-quantile (QQ) plots with lambda score for the Score Test, REGENIE, and PLINK Logistic Regression analyses performed on the PanScan-PanC4 dataset. The first row displays QQ plots for the Score Test results, the second row for REGENIE results, and the third row for Logistic Regression results. QQplots with dark blue points represent the results for all variants, dark green points represent results for common variants (MAF > 0.05), and red points represent results for rare variants with MAF between 0.00001 and 0.01.



Supplementary Figure 2: Quantile-quantile (QQ) plots with lambda score for the Score Test, REGENIE, and PLINK Logistic Regression analyses performed on the UK Biobank dataset. The first row displays QQ plots for the Score Test results, the second row for REGENIE results, and the third row for Logistic Regression results. QQplots with dark blue points represent the results for all variants, dark green points represent results for common variants (MAF > 0.05), and red points represent results for rare variants with MAF between 0.00001 and 0.01.

Supplemental Table 1: Genome-wide significant rare variant associations in the PanScan-PanC4 dataset by using score test. The first 7 columns correspond to chromosome (CHR), position in hg38 (POS), SNP ID (rsID),reference allele (REF), alternative allele (ALT), minor allele frequency (MAF) and sample size (N). INFO score indicates the imputation r2/information metric value of the variant. Observed minor allele counts in cases and controls are reported as C_Case and C_Control, respectively. For the Score Test, the table reports the p-value, test statistic (Score) with standard error (Score_se), and effect size (Est) with standart error (Est_se). REGENIE whole genome regression results include the regression coefficient (BETA), odds ratio (OR), and p-value metrics (LOG10P, P_VALUE). PLINK Logistic Regression results provide the odds ratio (OR), standard error of OR (SE), and p-value (P). The 'Consequence' column indicates the type of variant, while the 'GENE' column specifies the symbol of the gene harboring the variant.Groups of physically adjacent variants marked with same letter in the LD column show nearly identical association parameters, suggesting that they are in high linkage disequilibrium (r2>0.7).

										SCORE TEST						WHOLE GENOME REGRESSION					LOGISTIC REGRESSION				
CHR	POS	rsID	REF	ALT	MAF	N	INFO Score	C_Case	C_Control	pvalue	pvalue_log10	Score	Score_se	Est	Est_se	BETA	SE	OR	LOG10P	P_VALUE	OR	SE	P	Consequence	GENE	LD
1	70622967	rs17131364	A	G	0.0037	14254	0.654	84	22	4.00E-09	8.398	130.230	22.132	0.266	0.045	1.113	0.181	3.044	9.075	8.42E-10	3.173	0.232	6.28E-07	intergenic_variant	-	a
1	70623097	rs17131365	T	C	0.0045	14254	0.593	97	30	4.38E-09	8.359	141.851	24.168	0.243	0.041	1.008	0.166	2.741	8.891	1.28E-09	2.722	0.201	6.05E-07	intergenic_variant	-	a
1	71379769	rs356392	A	G	0.0045	14254	0.702	96	33	3.63E-08	7.440	133.873	24.305	0.227	0.041	0.935	0.165	2.546	7.813	1.54E-08	2.475	0.193	2.64E-06	intron_variant,non_coding_transcript_variant	ZRANB2-DT	
1	174968215	rs12072529	T	C	0.0075	14254	0.720	148	65	2.10E-08	7.677	179.147	31.973	0.175	0.031	0.716	0.126	2.046	7.930	1.17E-08	1.910	0.137	2.32E-06	intron_variant	RABGAP1L	b
1	174971448	rs16847652	C	T	0.0071	14254	0.734	141	61	3.18E-08	7.497	172.190	31.131	0.178	0.032	0.726	0.129	2.068	7.752	1.77E-08	1.939	0.141	2.85E-06	intron_variant	RABGAP1L	b
1	174974914	rs74128526	C	T	0.0071	14254	0.734	142	61	2.40E-08	7.620	174.063	31.190	0.179	0.032	0.732	0.129	2.079	7.891	1.29E-08	1.952	0.141	2.25E-06	intron_variant	RABGAP1L	b
1	174975627	rs10912862	C	T	0.0071	14254	0.733	141	60	2.19E-08	7.660	173.976	31.088	0.180	0.032	0.736	0.129	2.088	7.929	1.18E-08	1.966	0.142	2.07E-06	intron_variant	RABGAP1L	b
1	174983089	rs11806870	T	C	0.0068	14254	0.546	136	57	2.63E-08	7.580	170.184	30.585	0.182	0.033	0.744	0.131	2.105	7.846	1.43E-08	1.979	0.145	2.61E-06	intron_variant	RABGAP1L	b
1	174986433	rs4651250	G	A	0.0068	14254	0.547	136	57	2.63E-08	7.580	170.184	30.585	0.182	0.033	0.744	0.131	2.105	7.846	1.43E-08	1.979	0.145	2.61E-06	intron_variant	RABGAP1L	b
1	174990836	rs4650677	G	A	0.0071	14254	0.728	141	61	3.43E-08	7.465	171.892	31.151	0.177	0.032	0.724	0.129	2.063	7.716	1.92E-08	1.939	0.141	2.85E-06	downstream_gene_variant	RABGAP1L	b
1	175010590	rs10912866	C	A	0.0068	14254	0.735	136	58	3.64E-08	7.439	168.029	30.510	0.181	0.033	0.739	0.132	2.093	7.704	1.98E-08	1.962	0.145	3.26E-06	downstream_gene_variant	MRPS14	b
1	247457232	rs7526993	G	T	0.0052	14254	0.695	108	40	4.99E-08	7.302	139.440	25.577	0.213	0.039	0.869	0.157	2.384	7.512	3.07E-08	2.395	0.179	1.10E-06	intron_variant	OR2B11	
2	233828400	rs7606432	G	T	0.0030	14254	0.812	71	15	2.51E-09	8.601	116.722	19.581	0.304	0.051	1.285	0.205	3.613	9.427	3.74E-10	4.063	0.280	5.38E-07	intron_variant	MROH2A	
3	50495362	rs73835510	T	C	0.0045	14254	0.647	95	33	4.47E-08	7.349	128.909	23.562	0.232	0.042	0.958	0.170	2.605	7.722	1.90E-08	2.589	0.197	1.45E-06	intron_variant	CACNA2D2	
3	64816643	rs17071542	A	G	0.0039	14254	0.578	85	26	3.68E-08	7.434	123.571	22.446	0.245	0.045	1.019	0.179	2.769	7.896	1.27E-08	2.804	0.217	2.00E-06	intron_variant,non_coding_transcript_variant	ADAMTS9-AS2	c
3	64828991	rs17071578	G	A	0.0040	14254	0.579	87	26	1.62E-08	7.792	127.690	22.605	0.250	0.044	1.039	0.178	2.827	8.301	5.00E-09	2.873	0.217	1.12E-06	intron_variant,non_coding_transcript_variant	ADAMTS9-AS2	c
3	64829004	rs72892938	T	C	0.0040	14254	0.581	87	26	1.62E-08	7.792	127.690	22.605	0.250	0.044	1.039	0.178	2.827	8.301	5.00E-09	2.873	0.217	1.12E-06	intron_variant,non_coding_transcript_variant	ADAMTS9-AS2	c
3	64830028	rs28633489	T	C	0.0042	14254	0.643	93	28	7.61E-09	8.119	136.292	23.592	0.245	0.042	1.019	0.170	2.771	8.667	2.15E-09	2.804	0.208	6.88E-07	intron_variant,non_coding_transcript_variant	ADAMTS9-AS2	c
3	64830444	rs2002896	G	T	0.0040	14254	0.658	87	27	2.44E-08	7.613	125.604	22.520	0.248	0.044	1.028	0.178	2.796	8.085	8.22E-09	2.814	0.214	1.39E-06	intron_variant,non_coding_transcript_variant	ADAMTS9-AS2	c
3	64830496	rs1108243	G	A	0.0040	14254	0.663	87	27	2.44E-08	7.613	125.604	22.520	0.248	0.044	1.028	0.178	2.796	8.085	8.22E-09	2.814	0.214	1.39E-06	intron_variant,non_coding_transcript_variant	ADAMTS9-AS2	c
5	119083161	rs113283666	T	C	0.0043	14254	0.882	94	30	4.54E-08	7.343	131.828	24.108	0.227	0.041	0.939	0.167	2.556	7.757	1.75E-08	2.634	0.201	1.41E-06	intron_variant	DMXL1	
5	135082162	rs60411261	G	A	0.0074	14254	0.737	150	60	4.25E-09	8.371	187.524	31.924	0.184	0.031	0.755	0.126	2.127	8.714	1.93E-09	2.075	0.142	2.60E-07	intron_variant,non_coding_transcript_variant	PITX1-AS1	d
5	135082413	rs59392102	G	A	0.0074	14254	0.746	150	60	4.25E-09	8.371	187.524	31.924	0.184	0.031	0.755	0.126	2.127	8.714	1.93E-09	2.075	0.142	2.60E-07	intron_variant,non_coding_transcript_variant	PITX1-AS1	d
5	135082951	rs142178813	G	A	0.0074	14254	0.761	150	60	4.25E-09	8.371	187.524	31.924	0.184	0.031	0.755	0.126	2.127	8.714	1.93E-09	2.075	0.142	2.60E-07	intron_variant,non_coding_transcript_variant	PITX1-AS1	d
8	15358373	rs73667969	G	A	0.0015	14254	0.764	39	4	2.78E-08	7.556	72.387	13.031	0.426	0.077	1.831	0.308	6.243	8.565	2.72E-09	9.494	0.525	1.82E-05	intergenic_variant	-	e
8	15369294	rs139597717	G	A	0.0015	14254	0.859	39	4	3.05E-08	7.515	72.224	13.041	0.425	0.077	1.826	0.308	6.207	8.527	2.97E-09	9.494	0.525	1.82E-05	intergenic_variant	-	e
9	102245385	rs141914494	CAT	C	0.0051	14254	0.831	108	37	4.75E-08	7.323	148.889	27.267	0.200	0.037	0.828	0.147	2.289	7.725	1.89E-08	2.277	0.174	2.36E-06	intron_variant,non_coding_transcript_variant	-	
10	32487025	rs58980198	T	C	0.0039	14254	0.853	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant	CCDC7	f
10	32487031	rs56903806	G	A	0.0039	14254	0.853	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant	CCDC7	f
10	32491351	rs61515216	G	A	0.0039	14254	0.865	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant	CCDC7	f
10	32501489	rs112503862	T	G	0.0039	14254	0.866	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant	CCDC7	f
10	32509211	rs112394767	TATAGACCCATGAA	T	0.0038	14254	0.845	85	24	1.34E-08	7.873	126.579	22.281	0.255	0.045	1.064	0.180	2.897	8.455	3.51E-09	3.016	0.224	8.60E-07	intron_variant	CCDC7	f
10	32521097	rs28753068	A	G	0.0039	14254	0.848	85	26	4.85E-08	7.315	122.542	22.456	0.243	0.045	1.009	0.179	2.744	7.789	1.63E-08	2.804	0.217	2.00E-06	intron_variant	CCDC7	f
10	32549212	rs112676977	T	G	0.0038	14254	0.869	85	24	1.34E-08	7.873	126.579	22.281	0.255	0.045	1.064	0.180	2.897	8.455	3.51E-09	3.016	0.224	8.60E-07	intron_variant	CCDC7	f
10	32589354	rs57242880	C	T	0.0039	14254	0.866	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant,NMD_transcript_variant	CCDC7	f
10	32589958	rs79945449	G	T	0.0039	14254	0.868	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant,NMD_transcript_variant	CCDC7	f
10	32616391	rs58538625	C	A	0.0039	14254	0.869	85	25	2.46E-08	7.610	124.729	22.368	0.249	0.045	1.038	0.179	2.824	8.142	7.21E-09	2.906	0.220	1.31E-06	intron_variant,NMD_transcript_variant	CCDC7	f
12	126362718	rs16920552	G	C	0.0082	14254	0.839	160	73	4.45E-08	7.352	177.720	32.478	0.168	0.031	0.684	0.124	1.982	7.510	3.09E-08	1.949	0.135	7.34E-07	downstream_gene_variant	LINC02350	
13	83865310	rs11840233	T	C	0.0048	14254	0.646	101	37	4.69E-08	7.329	133.673	24.470	0.223	0.041	0.921	0.164	2.511	7.698	2.01E-08	2.451	0.187	1.66E-06	intergenic_variant	-	
13	87927557	rs113906642	C	A	0.0043	14254	0.646	92	30	3.49E-08	7.457	129.672	23.513	0.235	0.043	0.973	0.171	2.647	7.919	1.20E-08	2.639	0.203	1.71E-06	intergenic_variant	-	g
13	87940571	rs77635354	G	A	0.0043	14254	0.661	92	30	3.49E-08	7.457	129.672	23.513	0.235	0.043	0.973	0.171	2.647	7.919	1.20E-08	2.639	0.203	1.71E-06	intergenic_variant	-	g
13	110555518	rs9301462	G	A	0.0059	14254	0.526	125	44	4.91E-09	8.309	170.218	29.096	0.201	0.034	0.830	0.138	2.294	8.746	1.80E-09	2.258	0.161	4.56E-07	intron_variant	RAB20	
16	5515554	rs73512235	A	G	0.0021	14254	0.646	51	9	3.69E-08	7.433	87.381	15.873	0.347	0.063	1.470	0.253	4.349	8.216	6.08E-09	5.184	0.360	4.80E-06	intron_variant,non_coding_transcript_variant	RBFOX1	h
16	5582166	rs76387091	G	T	0.0026	14254	0.668	62	13	2.00E-08	7.699	103.511	18.444	0.304	0.054	1.293	0.218	3.643	8.546	2.85E-09	4.012	0.299	3.41E-06	intron_variant,non_coding_transcript_variant	RBFOX1	h
21	15874904	rs111271349	T	C	0.0065	14254	0.502	131	53	3.72E-08	7.430	159.214	28.929	0.190	0.035	0.777	0.139	2.174	7.662	2.18E-08	2.170	0.156	6.70E-07	intron_variant	USP25	k
21	15888955	rs9982588	T	C	0.0068	14254	0.550	137	56	2.44E-08	7.612	165.403	29.657	0.188	0.034	0.768	0.135	2.155	7.857	1.39E-08	2.147	0.152	4.75E-07	intergenic_variant	-	k

Supplemental Table 2: Genome-wide significant rare variant associations in the UK Biobank dataset by using score test. The first 7 columns correspond to chromosome (CHR), position in hg38 (POS), SNP ID (rsID),reference allele (REF), alternative allele (ALT), minor allele frequency (MAF) and sample size (N). INFO score indicates the imputation r2/information metric value of the variant. Observed minor allele counts in cases and controls are reported as C_Case and C_Control, respectively. For the Score Test, the table reports the p-value, test statistic (Score) with standard error (Score_se), and effect size (Est) with standard error (Est_se). REGENIE whole genome regression results include the regression coefficient (BETA), odds ratio (OR), and p-value metrics (LOG10P, P_VALUE). PLINK Logistic Regression results provide the odds ratio (OR), standard error of OR (SE), and p-value (P). The 'Consequence' column indicates the type of variant, while the 'GENE' column specifies the symbol of the gene harboring the variant.Groups of physically adjacent variants marked with same letter in the LD column show nearly identical association parameters, suggesting that they are in high linkage disequilibrium (r2>0.7).

										SCORE TEST						WHOLE GENOME REGRESSION					LOGISTIC REGRESSION				
CHR	POS	rsID	REF	ALT	MAF	N	INFO Score	C_Case	C_Control	pvalue	pvalue_log10	Score	Score_se	Est	Est_se	BETA	SE	OR	LOG10P	P_VALUE	OR	SE	P	Consequence	GENE	LD
1	55564788	rs182048854	T	G	0.0052	11021	0.963	28	86	1.77E-08	7.752	206.758	36.705	0.153	0.027	1.564	0.325	4.776	5.821	1.51E-06	3.253	0.220	8.41E-08	intron_variant,non_coding_transcript_variant	LINC01755	
1	101746319	rs545094345	T	C	0.0074	11021	0.836	36	125	5.96E-09	8.225	253.155	43.513	0.134	0.023	1.349	0.260	3.855	6.668	2.15E-07	2.892	0.192	3.24E-08	intron_variant,non_coding_transcript_variant	LINC01709	a
1	101756109	rs189431992	G	A	0.0061	11021	0.837	31	103	1.54E-08	7.813	224.910	39.757	0.142	0.025	1.287	0.281	3.620	5.337	4.60E-06	3.008	0.208	1.12E-07	intron_variant,non_coding_transcript_variant	LINC01709	a
1	101756136	rs182031558	A	G	0.0072	11021	0.837	35	123	1.37E-08	7.864	244.784	43.115	0.132	0.023	1.340	0.264	3.820	6.418	3.82E-07	2.857	0.195	6.79E-08	intron_variant,non_coding_transcript_variant	LINC01709	a
1	171679735	rs189475976	T	C	0.0010	11021	0.795	10	13	4.02E-08	7.396	90.862	16.551	0.332	0.060	2.412	0.623	11.160	3.968	1.08E-04	7.608	0.422	1.51E-06	downstream_gene_variant	RPL4P3	
1	238408532	rs577615576	C	A	0.0009	11021	0.797	9	11	4.49E-08	7.348	84.414	15.431	0.355	0.065	1.408	0.568	4.088	1.878	1.32E-02	8.043	0.451	3.73E-06	downstream_gene_variant	-	
2	168800230	rs554594944	T	G	0.0014	11021	0.869	13	18	2.37E-10	9.625	121.715	19.212	0.330	0.052	2.977	0.591	19.621	6.316	4.83E-07	7.153	0.365	7.30E-08	upstream_gene_variant	NOSTRIN	b
2	168802704	rs202204045	C	T	0.0014	11021	0.863	13	18	2.37E-10	9.626	121.720	19.212	0.330	0.052	3.038	0.597	20.862	6.439	3.64E-07	7.155	0.365	7.28E-08	intron_variant	NOSTRIN	b
2	168813599	rs560183362	A	G	0.0014	11021	0.858	13	18	2.92E-10	9.535	121.110	19.214	0.328	0.052	2.876	0.599	17.750	5.796	1.60E-06	7.144	0.365	7.45E-08	intron_variant	NOSTRIN	b
2	177689536	rs16865593	G	C	0.0010	11021	0.957	10	12	3.16E-08	7.500	88.977	16.084	0.344	0.062	2.913	0.698	18.405	4.517	3.04E-05	8.237	0.429	9.10E-07	intron_variant	PDE11A	c
2	177690009	rs1866214	C	T	0.0010	11021	0.959	10	12	3.19E-08	7.497	88.956	16.084	0.344	0.062	2.830	0.698	16.944	4.299	5.02E-05	8.229	0.429	9.21E-07	intron_variant	PDE11A	c
2	177696807	rs10172800	C	T	0.0010	11021	0.963	10	12	3.17E-08	7.499	88.972	16.084	0.344	0.062	2.869	0.688	17.624	4.516	3.05E-05	8.235	0.429	9.13E-07	intron_variant	PDE11A	c
2	177698909	rs10194370	T	C	0.0010	11021	0.959	10	12	3.19E-08	7.496	88.953	16.084	0.344	0.062	2.791	0.689	16.296	4.297	5.05E-05	8.228	0.429	9.22E-07	intron_variant	PDE11A	c
2	177702196	rs10192225	C	T	0.0010	11021	0.961	10	12	3.19E-08	7.496	88.951	16.084	0.344	0.062	2.736	0.682	15.421	4.214	6.11E-05	8.227	0.429	9.23E-07	intron_variant	PDE11A	c
2	177703238	rs73979531	C	T	0.0011	11021	0.965	11	13	4.43E-09	8.354	98.604	16.805	0.349	0.060	2.809	0.650	16.600	4.810	1.55E-05	8.370	0.411	2.35E-07	intron_variant	PDE11A	c
2	177706952	rs60501109	T	C	0.0011	11021	0.965	11	13	4.43E-09	8.354	98.604	16.805	0.349	0.060	2.791	0.648	16.304	4.780	1.66E-05	8.370	0.411	2.35E-07	intron_variant	PDE11A	c
2	177706985	rs16865698	G	C	0.0011	11021	0.966	11	13	4.43E-09	8.354	98.602	16.805	0.349	0.060	2.773	0.646	16.004	4.746	1.79E-05	8.369	0.411	2.35E-07	intron_variant	PDE11A	c
2	177707740	rs16865700	C	G	0.0011	11021	0.966	11	13	4.43E-09	8.354	98.604	16.805	0.349	0.060	2.792	0.648	16.318	4.782	1.65E-05	8.370	0.411	2.35E-07	intron_variant	PDE11A	c
2	177708092	rs16865706	C	T	0.0011	11021	0.957	11	13	4.42E-09	8.355	98.610	16.805	0.349	0.060	2.709	0.644	15.010	4.582	2.62E-05	8.372	0.411	2.34E-07	intron_variant	PDE11A	c
2	177708901	rs16865709	G	A	0.0012	11021	0.963	11	15	3.55E-08	7.450	96.441	17.497	0.315	0.057	2.628	0.632	13.841	4.491	3.23E-05	7.253	0.398	6.55E-07	intron_variant	PDE11A	c
2	177711216	rs6760512	C	G	0.0010	11021	0.965	11	12	1.26E-09	8.898	99.885	16.450	0.369	0.061	3.031	0.680	20.721	5.079	8.34E-06	9.060	0.419	1.42E-07	intron_variant	PDE11A	c
2	177712274	rs13429220	G	T	0.0012	11021	0.968	11	15	3.52E-08	7.454	96.469	17.497	0.315	0.057	2.670	0.628	14.439	4.673	2.12E-05	7.261	0.398	6.46E-07	intron_variant	PDE11A	c
2	177712914	rs6708025	C	A	0.0011	11021	0.966	11	14	1.27E-08	7.895	97.601	17.154	0.332	0.058	2.476	0.624	11.895	4.143	7.19E-05	7.770	0.404	3.94E-07	intron_variant	PDE11A	c
2	177718321	rs10210259	C	T	0.0014	11021	0.971	12	18	3.34E-08	7.477	103.815	18.798	0.294	0.053	2.689	0.603	14.710	5.081	8.29E-06	6.606	0.374	4.51E-07	intron_variant	PDE11A	c
2	234582093	rs533152804	T	A	0.0016	11021	0.707	13	22	1.21E-08	7.918	116.293	20.407	0.279	0.049	1.379	0.410	3.971	3.110	7.76E-04	5.845	0.351	5.05E-07	intergenic_variant	-	
3	61742389	rs150151986	C	A	0.0015	11021	0.970	13	21	7.01E-09	8.154	116.455	20.111	0.288	0.050	2.533	0.579	12.590	4.908	1.24E-05	6.127	0.354	3.15E-07	intron_variant	PTPRG	
3	72543943	rs574181877	C	A	0.0019	11021	0.880	15	27	2.53E-09	8.597	133.206	22.351	0.267	0.045	2.161	0.481	8.676	5.145	7.16E-06	5.508	0.324	1.36E-07	intron_variant,non_coding_transcript_variant	-	d
3	72578202	rs141393188	T	A	0.0021	11021	0.850	15	31	3.48E-08	7.458	128.988	23.387	0.236	0.043	1.955	0.480	7.062	4.339	4.59E-05	4.789	0.316	7.32E-07	downstream_gene_variant	RNU1-62P	d
4	45482110	rs191880803	G	C	0.0011	11021	0.852	11	13	6.08E-10	9.216	104.667	16.914	0.366	0.059	2.217	0.549	9.177	4.268	5.39E-05	8.366	0.411	2.36E-07	downstream_gene_variant	RNU6-931P	
5	164720055	rs572713612	A	G	0.0009	11021	0.659	9	11	4.82E-08	7.317	84.208	15.429	0.354	0.065	1.366	0.516	3.921	2.093	8.07E-03	8.080	0.451	3.55E-06	intron_variant,non_coding_transcript_variant	LINC03000	
6	19104435	rs559440781	A	AC	0.0015	11021	0.722	12	20	3.01E-08	7.521	108.162	19.522	0.284	0.051	1.050	0.361	2.858	2.439	3.64E-03	5.939	0.367	1.18E-06	intron_variant,non_coding_transcript_variant	-	
6	65076685	rs73765754	C	T	0.0010	11021	0.775	12	10	2.35E-11	10.629	107.149	16.034	0.417	0.062	2.962	0.637	19.328	5.480	3.32E-06	11.917	0.429	7.92E-09	intron_variant	EYS	
6	66294311	rs189013146	G	C	0.0011	11021	0.876	10	14	3.09E-08	7.510	93.610	16.908	0.327	0.059	2.908	0.654	18.316	5.056	8.79E-06	7.056	0.415	2.55E-06	intergenic_variant	-	
6	109460424	rs770823329	G	A	0.0010	11021	0.628	11	10	7.41E-12	11.130	108.386	15.824	0.433	0.063	2.855	0.560	17.372	6.471	3.38E-07	10.936	0.438	4.79E-08	downstream_gene_variant	ZBTB24	
7	21125677	rs536414907	T	C	0.0011	11021	0.889	10	14	3.67E-08	7.436	93.134	16.914	0.326	0.059	3.239	0.700	25.500	5.427	3.74E-06	7.069	0.415	2.50E-06	intron_variant,non_coding_transcript_variant	-	
8	9804061	rs182660759	G	C	0.0012	11021	0.859	11	16	1.56E-08	7.806	101.370	17.928	0.315	0.056	2.285	0.588	9.825	3.987	1.03E-04	6.789	0.393	1.10E-06	intron_variant,non_coding_transcript_variant	-	
8	29947735	rs193234554	G	T	0.0023	11021	0.819	16	35	3.47E-08	7.460	135.799	24.619	0.224	0.041	2.038	0.462	7.678	4.983	1.04E-05	4.526	0.304	6.58E-07	intron_variant,non_coding_transcript_variant	LINC02209	
8	95397133	rs568687252	C	T	0.0010	11021	0.916	10	13	1.37E-08	7.865	94.022	16.560	0.343	0.060	3.704	0.711	40.603	NA	NA	7.593	0.422	1.55E-06	intron_variant,non_coding_transcript_variant	CFAP418-AS1	
11	45036246	rs563490761	C	T	0.0012	11021	0.952	11	15	3.81E-09	8.419	103.708	17.601	0.335	0.057	3.188	0.681	24.248	5.546	2.85E-06	7.248	0.398	6.60E-07	intron_variant,non_coding_transcript_variant	-	e
11	45054101	rs559856794	G	A	0.0012	11021	0.964	11	15	3.82E-09	8.418	103.702	17.601	0.335	0.057	2.978	0.662	19.649	5.172	6.74E-06	7.247	0.398	6.62E-07	intron_variant,non_coding_transcript_variant	-	e
11	45078148	rs780853848	CTT	C	0.0011	11021	0.907	11	14	1.31E-09	8.882	104.697	17.260	0.351	0.058	2.633	0.629	13.915	4.541	2.87E-05	7.758	0.404	4.02E-07	intergenic_variant	-	e
12	63650391	rs760049940	AAAC	A	0.0011	11021	0.865	10	14	4.21E-08	7.376	92.707	16.912	0.324	0.059	2.868	0.682	17.598	4.578	2.64E-05	7.050	0.415	2.58E-06	intron_variant,NMD_transcript_variant	DPY19L2	
12	89651293	rs576889889	A	C	0.0010	11021	0.814	10	13	1.03E-08	7.988	94.810	16.557	0.346	0.060	3.168	0.660	23.752	5.793	1.61E-06	7.621	0.422	1.48E-06	intron_variant	ATP2B1	
12	90046386	rs369984058	T	C	0.0016	11021	0.949	17	18	2.23E-15	14.652	161.808	20.410	0.388	0.049	2.856	0.532	17.393	7.106	7.84E-08	9.386	0.340	4.41E-11	intron_variant,non_coding_transcript_variant	ATP2B1-AS1	f
12	90165370	rs188205290	C	T	0.0010	11021	0.946	10	13	9.81E-09	8.008	94.945	16.558	0.346	0.060	3.750	0.687	42.502	7.310	4.90E-08	7.635	0.422	1.45E-06	intergenic_variant	-	f
12	95666767	rs541888057	G	A	0.0023	11021	0.854	16	35	3.73E-08	7.428	135.516	24.625	0.223	0.041	2.240	0.467	9.397	5.791	1.62E-06	4.535	0.304	6.36E-07	intron_variant	NTN4	
15	61519250	rs139716382	C	T	0.0016	11021	0.766	14	20	8.34E-11	10.079	130.535	20.100	0.323	0.050	2.811	0.543	16.626	6.639	2.30E-07	6.925	0.350	3.23E-08	intron_variant,non_coding_transcript_variant	-	
15	87281896	rs188894299	T	C	0.0010	11021	0.883	11	12	1.28E-10	9.891	106.358	16.543	0.389	0.060	2.765	0.638	15.886	4.837	1.45E-05	9.058	0.419	1.42E-07	intergenic_variant	-	
16	4277823	rs554010513	C	T	0.0025	11021	0.853	17	38	1.83E-08	7.738	143.860	25.564	0.220	0.039	1.781	0.440	5.935	4.293	5.09E-05	4.435	0.294	3.93E-07	upstream_gene_variant	TFAP4	
16	48824988	rs149525318	G	A	0.0039	11021	0.938	24	61	3.50E-09	8.456	189.422	32.072	0.184	0.031	1.707	0.360	5.512	5.661	2.18E-06	3.770	0.241	3.71E-08	intergenic_variant	-	
16	48891059	rs138815923	T	A	0.0033	11021	0.892	23	49	5.79E-11	10.238	191.335	29.215	0.224	0.034	1.908	0.374	6.740	6.480	3.31E-07	4.685	0.255	1.38E-09	intergenic_variant	-	
16	71194272	rs562207241	G	A	0.0015	11021	0.789	13	21	6.62E-09	8.179	116.668	20.114	0.288	0.050	2.511	0.573	12.312	4.927	1.18E-05	6.121	0.354	3.20E-07	intron_variant	HYDIN	

Supplemental Table 3: The results from meta- analysis of Gene-based analysis and annotation-weighted gene-based analysis with MAGMA tool from PanScan-PanC4 and UK Biobank. GENE CODE: the gene codes assigned by MAGMA tool. CHR: chromosome. START,STOP: the gene region in hg38 included for the gene-based test. NSNP: The number of rare SNPs included in the gene-based test. For meta-analysis, this value represents the average number of SNPs across the two datasets, rather than the number of overlapping SNPs. N: Number of samples. ZTAT: Z score of the gene-based test. P: p-value of the gene-based test. DATASETS: Number of the datasets included in meta-analysis. Red colored p-values indicates the genes became significant in meta-analysis of gene-based analysis. Bonferroni-corrected significance threshold of 2.59 × 10-6. ORVs: Number of overlapped rare variants between PanScan-PanC4 and UK Biobank datasets.

					Gene-Based Analysis														Annotation-weighted Gene-based Analysis														
GENE NAME	GENE CODE	CHR	START	STOP	PanScan-PanC4				UK Biobank				Meta-Analysis					PanScan-PanC4				UK Biobank				Meta-Analysis					ORVs
					NSN	N	ZTA	P	NSN	N	ZTA	P	NSN	N	DATASI	ZTA	P	NSN	N	ZTA	P	NSN	N	ZTA	P	NSN	N	ZTA	DATASI	P	
DHDDS	79947	1	26427282	26473306	295	14254	4.6934	1.34E-06	289	11021	-0.015	5.06E-01	292	25275	2	3.5144	2.20E-04	258	14254	3.9698	3.60E-05	255	11021	-1.123	8.69E-01	257	25275	2.6434	2	4.10E-03	156
RPS6KA1	6195	1	26524758	26577030	344	14254	4.9929	2.97E-07	388	11021	-0.048	5.19E-01	366	25275	2	3.7175	1.01E-04	322	14254	4.6849	1.40E-06	355	11021	0.8675	1.93E-01	339	25275	3.991	2	3.29E-05	202
ZMYM1	79830	1	35054786	35117859	442	14254	4.1867	1.42E-05	609	11021	2.2356	1.27E-02	526	25275	2	4.6203	1.92E-06	404	14254	3.9508	3.89E-05	525	11021	2.2311	1.28E-02	465	25275	4.4954	2	3.47E-06	265
DAB1	1600	1	56995906	58255539	8687	14254	4.6344	1.79E-06	12046	11021	0.8433	2.00E-01	10367	25275	2	4.0372	2.71E-05	8111	14254	4.7723	9.11E-07	10768	11021	0.6999	2.42E-01	9440	25275	3.9898	2	3.31E-05	5198
ST7L	54879	1	112515804	112625838	807	14254	2.4723	6.71E-03	1091	11021	4.7459	1.04E-06	949	25275	2	4.9905	3.01E-07	736	14254	2.4009	8.18E-03	990	11021	4.5609	2.55E-06	863	25275	4.7695	2	9.23E-07	520
SPRR1A	6698	1	152979088	152987814	65	14254	5.1814	1.10E-07	82	11021	-1.486	9.31E-01	74	25275	2	2.9101	1.81E-03	58	14254	5.1621	1.22E-07	59	11021	-0.445	6.72E-01	59	25275	3.5151	2	2.20E-04	39
MPZL1	9019	1	167716950	167793919	539	14254	6.0108	9.23E-10	720	11021	1.2755	1.01E-01	630	25275	2	5.3562	4.25E-08	490	14254	5.9957	1.01E-09	616	11021	1.1151	1.32E-01	553	25275	5.1415	2	1.36E-07	321
PRRX1	5396	1	170657728	170741419	608	14254	3.6187	1.48E-04	766	11021	1.1593	1.23E-01	687	25275	2	3.483	2.48E-04	567	14254	4.7976	8.03E-07	710	11021	1.2439	1.07E-01	639	25275	4.4357	2	4.59E-06	360
WDR92	116143	2	68128149	68162560	299	14254	3.5604	1.85E-04	352	11021	2.9015	1.86E-03	326	25275	2	4.5897	2.22E-06	277	14254	4.0476	2.59E-05	336	11021	2.6055	4.59E-03	307	25275	4.9444	2	3.82E-07	189
PNO1	56902	2	68152873	68177959	232	14254	3.8278	6.47E-05	287	11021	2.4239	7.68E-03	260	25275	2	4.4751	3.82E-06	214	14254	4.1918	1.38E-05	275	11021	2.0762	1.89E-02	245	25275	4.5641	2	2.51E-06	130
GMCL1	64395	2	69824642	69883386	281	14254	4.9702	3.34E-07	484	11021	-0.672	7.49E-01	383	25275	2	3.2888	5.03E-04	252	14254	4.7879	8.43E-07	445	11021	-2.243	9.88E-01	349	25275	3.151	2	8.14E-04	170
DYSF	8291	2	71448119	71688763	1798	14254	5.918	1.63E-09	2397	11021	-0.307	6.20E-01	2098	25275	2	4.2418	1.11E-05	1678	14254	5.542	1.50E-08	2285	11021	-0.731	7.68E-01	1982	25275	3.9434	2	4.02E-05	1037
LRRTM4	80059	2	76745457	77527376	6642	14254	4.5362	2.86E-06	8933	11021	-0.955	8.30E-01	7788	25275	2	2.7759	2.75E-03	6033	14254	4.6117	2.00E-06	8307	11021	-1.435	9.24E-01	7170	25275	2.7848	2	2.68E-03	3988
LOC105373764	105373764	2	177697404	177730138	282	14254	2.3483	9.43E-03	285	11021	4.2903	8.92E-06	284	25275	2	4.5965	2.15E-06	267	14254	2.3046	1.06E-02	271	11021	4.0588	2.47E-05	269	25275	4.3778	2	5.99E-06	155
HIBCH	26275	2	190202634	190325045	785	14254	4.3308	7.43E-06	1254	11021	0.5183	3.02E-01	1020	25275	2	3.5945	1.62E-04	711	14254	4.5759	2.37E-06	1179	11021	0.2699	3.94E-01	945	25275	3.5919	2	1.64E-04	498
PARD3B	117583	2	204540793	205622162	7332	14254	4.6194	1.92E-06	11006	11021	0.7929	2.14E-01	9169	25275	2	3.9926	3.27E-05	6762	14254	4.4691	3.93E-06	10346	11021	0.7757	2.19E-01	8554	25275	3.8121	2	6.89E-05	4501
GRM7	2917	3	6856097	7743531	7428	14254	5.0076	2.76E-07	9876	11021	-0.105	5.42E-01	8652	25275	2	3.691	1.12E-04	6930	14254	4.9031	4.72E-07	9381	11021	-0.17	5.67E-01	8156	25275	3.5832	2	1.70E-04	4686
SRGAP3	9901	3	8978591	9367929	2748	14254	4.1842	1.43E-05	3836	11021	2.1787	1.47E-02	3292	25275	2	4.5809	2.32E-06	2586	14254	4.0463	2.60E-05	3663	11021	2.1702	1.50E-02	3125	25275	4.5589	2	2.57E-06	1675
SLC6A11	6538	3	10811200	10943471	860	14254	4.3249	7.63E-06	1216	11021	0.1199	4.52E-01	1038	25275	2	3.3271	4.39E-04	811	14254	5.105	1.65E-07	1170	11021	0.6142	2.70E-01	991	25275	4.152	2	1.65E-05	532
CLEC3B	7123	3	45021267	45038071	114	14254	4.6093	2.02E-06	108	11021	1.6183	5.28E-02	111	25275	2	4.5301	2.95E-06	100	14254	4.6414	1.73E-06	102	11021	1.4587	7.23E-02	101	25275	4.5405	2	2.81E-06	66
HTR1F	3355	3	87787767	87996856	1386	14254	4.6117	2.00E-06	1714	11021	-0.283	6.12E-01	1550	25275	2	3.2761	5.26E-04	1274	14254	4.4914	3.54E-06	1620	11021	-0.839	7.99E-01	1447	25275	3.1867	2	7.19E-04	744
TIMMDC1	51300	3	119493521	119526281	227	14254	4.7016	1.29E-06	302	11021	0.6841	2.47E-01	265	25275	2	3.9825	3.41E-05	194	14254	4.7851	8.55E-07	269	11021	0.8567	1.96E-01	232	25275	4.1024	2	2.04E-05	134
C4orf19	55286	4	37448454	37595510	1054	14254	4.5308	2.94E-06	1481	11021	-0.281	6.11E-01	1268	25275	2	3.2171	6.47E-04	959	14254	4.6034	2.08E-06	1379	11021	-0.475	6.83E-01	1169	25275	3.2664	2	5.45E-04	631
HERC3	8916	4	88587423	88711297	777	14254	4.2271	1.18E-05	982	11021	0.6763	2.49E-01	880	25275	2	3.621	1.47E-04	712	14254	4.6326	1.81E-06	941	11021	0.5313	2.98E-01	827	25275	3.7054	2	1.06E-04	506
TACR3	6870	4	103587468	103724816	1138	14254	4.6916	1.36E-06	1502	11021	-0.087	5.35E-01	1320	25275	2	3.4659	2.64E-04	1036	14254	5.0607	2.09E-07	1389	11021	0.1899	4.25E-01	1213	25275	3.957	2	3.79E-05	646
MARCH1	55016	4	163522298	164388989	6093	14254	5.2663	6.96E-08	8216	11021	0.5497	2.91E-01	7155	25275	2	4.3178	7.88E-06	5567	14254	5.2056	9.67E-08	7630	11021	0.332	3.70E-01	6599	25275	4.1513	2	1.65E-05	3540
FAT1	2195	4	186585783	186731915	1092	14254	4.6887	1.37E-06	1648	11021	-0.394	6.53E-01	1370	25275	2	3.2606	5.56E-04	1000	14254	3.9987	3.18E-05	1544	11021	-0.998	8.41E-01	1272	25275	2.9116	2	1.80E-03	689
CDH18	1016	5	19470930	20580873	6899	14254	5.3584	4.20E-08	10759	11021	-0.858	8.05E-01	8829	25275	2	3.4576	2.73E-04	6368	14254	5.3991	3.35E-08	10017	11021	-1.328	9.08E-01	8193	25275	3.514	2	2.21E-04	4593
SLC30A5	64924	5	69088949	69133072	266	14254	3.7827	7.76E-05	371	11021	-0.635	7.37E-01	319	25275	2	2.4211	7.74E-03	243	14254	4.7266	1.14E-06	350	11021	-0.934	8.25E-01	297	25275	3.4701	2	2.60E-04	175
FSTL4	23105	5	133192827	133847066	4649	14254	4.8861	5.14E-07	6573	11021	-0.177	5.70E-01	5611	25275	2	3.5522	1.91E-04	4368	14254	4.8423	6.42E-07	6286	11021	-0.297	6.17E-01	5327	25275	3.5373	2	2.02E-04	2887
HTR1E	3354	6	86932306	87018679	482	14254	4.7879	8.43E-07	722	11021	0.7841	2.16E-01	602	25275	2	4.1134	1.95E-05	447	14254	4.6148	1.97E-06	690	11021	1.37	8.53E-02	569	25275	4.3152	2	7.97E-06	304
MED23	9439	6	131571966	131633239	394	14254	3.7388	9.24E-05	546	11021	3.489	2.42E-04	470	25275	2	5.1117	1.60E-07	355	14254	3.5255	2.11E-04	500	11021	3.1665	7.72E-04	428	25275	4.8416	2	6.44E-07	237
ENPP3	5169	6	131632235	131749413	633	14254	3.9704	3.59E-05	990	11021	2.4113	7.95E-03	812	25275	2	4.5739	2.39E-06	587	14254	3.6786	1.17E-04	929	11021	2.1727	1.49E-02	758	25275	4.1865	2	1.42E-05	375
OPRM1	4988	6	154005496	154248867	1632	14254	4.6635	1.55E-06	2188	11021	-0.592	7.23E-01	1910	25275	2	3.1113	9.31E-04	1506	14254	4.7071	1.26E-06	2060	11021	-0.826	7.96E-01	1783	25275	3.1283	2	8.79E-04	1019
ACAT2	39	6	159756957	159781112	149	14254	4.7534	1.00E-06	220	11021	-0.277	6.09E-01	185	25275	2	3.3866	3.54E-04	136	14254	4.2439	1.10E-05	201	11021	-1.539	9.38E-01	169	25275	2.7898	2	2.64E-03	86
MYOM2	9172	8	2040043	2147456	1300	14254	4.6094	2.02E-06	1789	11021	0.8629	1.94E-01	1545	25275	2	4.0314	2.77E-05	1239	14254	4.6696	1.51E-06	1737	11021	1.1307	1.29E-01	1488	25275	4.3047	2	8.36E-06	810
CSMD1	64478	8	2933353	4999806	34253	14254	4.5621	2.53E-06	46496	11021	0.0745	4.70E-01	40375	25275	2	3.4752	2.55E-04	32523	14254	4.5588	2.57E-06	44545	11021	-0.534	7.03E-01	38534	25275	3.3212	2	4.48E-04	21269
EBF2	64641	8	25839730	26050124	1575	14254	4.7684	9.29E-07	2107	11021	-0.945	8.28E-01	1841	25275	2	2.9567	1.55E-03	1444	14254	4.5247	3.02E-06	1990	11021	-1.095	8.63E-01	1717	25275	2.9429	2	1.63E-03	919
CLU	1191	8	27594917	27620031	188	14254	4.5275	2.98E-06	246	11021	-0.167	5.66E-01	217	25275	2	3.2898	5.01E-04	174	14254	4.7205	1.18E-06	238	11021	-0.885	8.12E-01	206	25275	3.2239	2	6.32E-04	120
ADAM32	203102	8	39102531	39286917	1086	14254	5.1006	1.69E-07	1637	11021	-0.036	5.14E-01	1362	25275	2	3.8068	7.04E-05	998	14254	5.0483	2.23E-07	1495	11021	-0.887	8.13E-01	1247	25275	3.5582	2	1.87E-04	657
IDO2	169355	8	39929955	40018391	673	14254	4.8704	5.57E-07	951	11021	0.3353	3.69E-01	812	25275	2	3.879	5.25E-05	629	14254	4.7402	1.07E-06	888	11021	-0.494	6.89E-01	759	25275	3.5224	2	2.14E-04	450
ZMAT4	79698	8	40528590	40902824	2571	14254	4.4711	3.89E-06	3670	11021	0.1831	4.27E-01	3121	25275	2	3.4786	2.52E-04	2390	14254	4.6676	1.52E-06	3452	11021	-0.025	5.10E-01	2921	25275	3.5145	2	2.20E-04	1463
PXDNL	137902	8	51317575	51816029	3247	14254	4.698	1.31E-06	4688	11021	0.9309	1.76E-01	3968	25275	2	4.1428	1.72E-05	3030	14254	4.853	6.08E-07	4352	11021	0.7898	2.15E-01	3691	25275	4.1106	2	1.97E-05	1966
FAM110B	90362	8	57989509	58151718	1123	14254	4.4668	3.97E-06	1491	11021	-1.144	8.74E-01	1307	25275	2	2.599	4.67E-03	1042	14254	4.6576	1.60E-06	1430	11021	-1.194	8.84E-01	1236	25275	3.27	2	5.38E-04	696
DNAJC5B	85479	8	66010042	66102520	581	14254	4.451	4.27E-06	928	11021	0.3508	3.63E-01	755	25275	2	3.5742	1.76E-04	546	14254	4.5652	2.50E-06	889	11021	-0.383	6.49E-01	718	25275	3.7933	2	7.43E-05	352
C8orf34	116328	8	68325722	68821023	3500	14254	4.7097	1.24E-06	4984	11021	-1.063	8.56E-01	4242	25275	2	2.8348	2.29E-03	3237	14254	4.7007	1.30E-06	4711	11021	-1.6	9.45E-01	3974	25275	2.8287	2	2.34E-03	2079
RIPK2	8767	8	89752733	89793064	253	14254	5.2368	8.17E-08	330	11021	1.0405	1.49E-01	292	25275	2	4.6198	1.92E-06	235	14254	5.3559	4.26E-08	308	11021	0.525	3.00E-01	272	25275	4.625	2	1.87E-06	168
OSGIN2	734	8	89896868	89929868	215	14254	4.5962	2.15E-06	287	11021	1.8204	3.43E-02	251	25275	2	4.6537	1.63E-06	189	14254	4.0625	2.43E-05	261	11021	2.7451	3.02E-03	225	25275	4.7299	2	1.12E-06	128
NBN	4683	8	89931336	89989733	419	14254	4.7083	1.25E-06	535	11021	0.2197	4.13E-01	477	25275	2	3.6809	1.16E-04	374	14254	4.6006	2.11E-06	501	11021	-0.235	5.93E-01	438	25275	3.5951	2	1.62E-04	283
CALB1	793	8	90056608	90087881	161	14254	4.7475	1.03E-06	219	11021	0.0209	4.92E-01	190	25275	2	3.5791	1.72E-04	144	14254	4.5605	2.55E-06	199	11021	-1.684	9.54E-01	172	25275	2.9424	2	1.63E-03	105
LOC105375678	105375678	8	101209031	101323227	773	14254	3.6978	1.09E-04	1249	11021	2.8704	2.05E-03	1011	25275	2	4.6724	1.49E-06	715	14254	3.6662	1.23E-04	1183	11021	2.7605	2.89E-03	949	25275	4.5312	2	2.93E-06	500
EPC1	80314	10	32265716	32383798	1027	14254	4.7255	1.15E-06	1193	11021	-0.895	8.15E-01	1110	25275	2	2.958	1.55E-03	916	14254	4.7021	1.29E-06	1084	11021	-0.84	7.99E-01	1000	25275	3.0426	2	1.17E-03	1206
PARD3	56288	10	34107560	34820325	5287	14254	4.4666	3.97E-06	6663	11021	2.1853	1.44E-02	5975	25275	2	4.7973	8.04E-07	4833	14254	4.3073	8.26E-06	6246	11021	2.163	1.53E-02	5540	25275	4.6757	2	1.46E-06	6344
PYROXD2	84795	10	98381565	98420221	375	14254	3.4744	2.56E-04	480	11021	3.2117	6.60E-04	428	25275	2	4.73	1.12E-06	358	14254	3.4666	2.64E-04	449	11021	3.4565	2.74E-04	404	25275	4.8647	2	5.73E-07	494
ABLIM1	3983	10	114429110	114826643	2498	14254	4.4197	4.94E-06	3548	11021	1.1488	1.25E-01	3023	25275	2	4.0777	2.27E-05	2308	14254	4.5577	2.59E-06	3348	11021	1.3844	8.31E-02	2828	25275	4.3475	2	6.88E-06	2860
INSC	387755	11	15106430	15249208	951	14254	4.7653	9.43E-07	1239	11021	0.8781	1.90E-01	1095	25275	2	4.1584	1.60E-05	902	14254	4.3726	6.14E-06	1176	11021	1.1059	1.34E-01	1039	25275	3.9162	2	4.50E-05	527
SCGB1D2	10647	11	62236630	62246812	67	14254	3.2853	5.09E-04	113	11021	0.5437	2.93E-01	90	25275	2	2.8262	2.36E-03	64	14254	2.6389	4.16E-03	97	11021	3.8085	6.99E-05	81	25275	4.6033	2	2.08E-06	41
ATF7IP	55729	12	14360632	14504935	948	14254	4.7879	8.43E-07	1310	11021	-0.882	8.11E-01	1129	25275	2	3.0132	1.29E-03	853	14254	4.8746	5.45E-07	1205	11021	-1.958	9.75E-01	1029	25275	2.657	2	3.94E-03	526
GRIP1	23426	12	66345398	67074267	4818	14254	4.9375	3.96E-07	6574	11021	-0.363	6.42E-01	5696	25275	2	3.4681	2.62E-04	4419	14254	5.0831	1.86E-07	6087	11021	-0.857	8.04E-01	5253	25275	3.6456	2	1.33E-04	2719
GPC5	2262	13	91393619	92869237	9793	14254	5.7198	5.33E-09	14094	11021	-0.223	5.88E-01	11944	25275	2	4.1483	1.68E-05	8984	14254	5.6396	8.52E-09	13225	11021	-0.227	5.90E-01	11105	25275	4.1061	2	2.01E-05	5848
GPC6	10082	13	93221825	94410020	8119	14254	5.1862	1.07E-07	10966	11021	1.0242	1.53E-01	9543	25275	2	4.571	2.43E-06	7517	14254	5.0644	2.05E-07	10327	11021	1.2354	1.08E-01	8922	25275	4.6246	2	1.88E-06	4845
FGF14	2259	13	101717784	102406774	5109	14254	5.0599	2.10E-07	6800	11021	-0.454	6.75E-01	5955	25275	2	3.5003	2.32E-04	4610	14254	4.8556	6.00E-07	6370	11021	-1.245	8.93E-01	5490	25275	3.3813	2	3.61E-04	3102
TMEM63C	57156	14	77176759	77261495	529	14254	4.9269	4.18E-07	804	11021	1.5679	5.85E-02	667	25275	2	4.7353	1.09E-06	513	14254	4.6153	1.96E-06	722	11021	1.9938	2.31E-02	618	25275	4.5024	2	3.36E-06	337
EML5	161436	14	88610431	88797752	1094	14254	4.822	7.11E-07	1578	11021	-1.827	9.66E-01	1336	25275	2	2.4147	7.87E-03	987	14254	4.6045	2.07E-06	1401	11021	-1.576	9.43E-01	1194	25275	2.6998	2	3.47E-03	683
UNC13C	440279	15	53973203	54630608	4893	14254	4.9925	2.98E-07	6217	11021	0.5129	3.04E-01	5555	25275	2	4.0879	2.18E-05	4540	14254	4.9634	3.46E-07	5536	11021	-0.331	6.30E-01	5038	25275	3.9153	2	4.52E-05	2613
DYX1C1	161582	15	55415755	55513234	770	14254	4.6727	1.49E-06	1040	11021	-2.954	9.98E-01	905	25275	2	1.5586	5.95E-02	717	14254	4.5021	3.36E-06	937	11021	-1.315	9.06E-01	827	25275	2.9161	2	1.77E-03	448
SNX1	6642	15	64089076	64146234	344	14254	4.6124	1.99E-06	468	11021	-0.142	5.57E-01	406	25275	2	3.3698	3.76E-04	311	14254	4.9201	4.33E-07	404	11021	0.7213	2.35E-01	358	25275	4.2592	2	1.03E-05	216
ARIH1	25820	15	72469282	72588560	720	14254	4.5954	2.16E-06	980	11021	1.0766	1.41E-01	850	25275	2	4.1619	1.58E-05	643	14254	4.0669	2.38E-05	849	11021	0.9612	1.68E-01	746	25275	3.6541	2	1.29E-04	427
NTRK3	4916	15	87872999	88261796	3067	14254	4.577	2.36E-06	4316	11021	-1.517	9.35E-01	3692	25275	2	2.4352	7.44E-03	2882	14254	4.5126	3.20E-06	3968	11021	-1.276	8.99E-01	3425	25275	2.6799	2	3.68E-03	1811
PCSK6	5046	15	101301928	101494984	1727	14254	4.9692	3.36E-07	2385	11021	0.5547	2.90E-01	2056	25275	2	4.0981	2.08E-05	1620	14254	4.9503	3.71E-07	2137	11021	-0.327	6.28E-01	1879	25275	3.8918	2	4.98E-05	1104
CLCN7	1186	16	1442933	1480084	382	14254	3.6493	1.31E-04	559	11021	-0.591	7.23E-01	471	25275	2	2.3501	9.39E-03	359	14254	4.6661	1.53E-06	542	11021	-1.44	9.25E-01	451	25275	2.5439	2	5.48E-03	212
NUBP2	10101	16	1777923	1791191	164	14254	4.0487	2.58E-05	202	11021	3.0335	1.21E-03	183	25275	2	5.0436	2.28E-07	153	14254	3.7501	8.84E-05	198	11021	2.9102	1.81E-03	176	25275	4.6175	2	1.94E-06	101
RBFOX1	54715	16	5234782	7715340	31067	14254	5.5541	1.40E-08	42211	11021	0.8784	1.90E-01	36639	25275	2	4.751	1.01E-06	29502	14254	5.6435	8.33E-09	40406	11021	0.866	1.93E-01	34954	25275	4.8685	2	5.62E-07	18856
ARHGAP44	9912	17	12784512	12993643	1388	14254	3.9616	3.72E-05	2071	11021	2.5915	4.78E-03	1730	25275	2	4.6863	1.39E-06	1264	14254	3.8161	6.78E-05	1926	11021	2.3264	1.00E-02	1595	25275	4.4235	2	4.85E-06	845
ZNF624	57547	17	16597870	16660626	404	14254	4.2687	9.83E-06	430	11021	2.5374	5.58E-03	417	25275	2	4.8812	5.27E-07	365	14254	3.6552	1.29E-04	398	11021	2.5171	5.92E-03	382	25275	4.464	2	4.02E-06	207
MPRIP	23164	17	17037468	17194648	1159	14254	4.2966	8.67E-06	1419	11021	-1.195	8.84E-01	1289	25275	2	2.4374	7.40E-03	1069	14254	4.7969	8.06E-07	1334	11021	-1.851	9.68E-01	1202	25275	2.6826	2	3.65E-03	682
COPS3	8533	17	17244624	17286303	307	14254	4.6301	1.83E-06	386	11021	-0.141	5.56E-01	347	25275	2	3.3843	3.57E-04	281	14254	4.6807	1.43E-06	361	11021	-0.405	6.57E-01	321	25275	3.2571	2	5.63E-04	174
MSI2	124540	17	57251570	57702477	3189	14254	4.995	2.94E-07	4477	11021	-1.846	9.68E-01	3833	25275	2	2.532	5.67E-03	2988	14254	4.7522	1.01E-06	4266	11021	-1.115	8.68E-01	3627	25275	2.812	2	2.46E-03	1884
MEP1B	4225	18	32185024	32222404	243	14254	4.5884	2.23E-06	313	11021	-0.303	6.19E-01	278	25275	2	3.2459	5.86E-04	226	14254	4.4932	3.51E-06	300	11021	-0.993	8.40E-01	263	25275	3.0205	2	1.26E-03	163
PTPRT	11122	20	42070752	43194917	8065	14254	4.9523	3.67E-07	11552	11021	2.1527	1.57E-02	9809	25275	2	5.1405	1.37E-07	7532	14254	4.9211	4.30E-07	11021	11021	2.2166	1.33E-02	9277	25275	5.1945	2	1.03E-07	4961
TIAM1	7074	21	31116418	31564977	3121	14254	4.5027	3.36E-06	4327	11021	0.9183	1.79E-01	3724	25275	2	3.9878	3.33E-05	2854	14254	4.5822	2.30E-06	4043	11021	0.1006	4.60E-01	3449	25275	4.0469	2	2.59E-05	1797

Supplementary Table 4: Inflation Scores (Lambda) for Gene-Centric Analysis across PanScan-PanC4 and UK Biobank Cohorts. This table presents inflation scores calculated from the outputs of the "STAARpipeline" R package for different categories of gene-centric analysis.

Analysis Category	PanScan-PanC4	UK Biobank
Gene-Centric Coding Region Analysis	Inflation Score	Inflation Score
pLoF	0.854	1.031
pLoF_DS	0.989	0.992
missense	1.098	0.721
disruptive_missense	0.931	0.889
synonymous	1.291	0.781
Gene-Centric Noncoding Region Analysis	Inflation Score	Inflation Score
UTR	1.672	0.703
downstream	1.336	0.697
upstream	1.324	0.746
promoter_CAGE	1.205	0.806
promoter_DHS	1.768	0.693
enhancer_CAGE	1.367	0.757
enhancer_DHS	2.935	0.724

Supplementary Table 5: Cognition-related genes

Gene code	Gene name	Gene type	CHR	Start	End
ENSG00000264341	MIR4417	miRNA	1	5564071	5564143
ENSG00000234593	RP4-704D23.1	lincRNA	1	14338825	14419973
ENSG00000230424	RP1-43E13.2	antisense	1	19210501	19240704
ENSG00000176378	PFN1P10	processed_pseudogene	1	21459756	21460082
ENSG00000228237	EFCAB14-AS1	antisense	1	46674036	46692098
ENSG00000227883	ATP6V0E1P4	processed_pseudogene	1	47550196	47550753
ENSG00000234318	RP4-771M4.3	lincRNA	1	62896009	62901639
ENSG00000275678	RP4-547N15.3	sense_intronic	1	67121605	67123956
ENSG00000224493	RP11-87O11.1	processed_pseudogene	1	75521562	75522231
ENSG00000231999	FLJ27354	antisense	1	89581291	89632916
ENSG00000230735	RP11-413E1.4	sense_intronic	1	89629725	89676386
ENSG00000271949	RP11-302M6.4	protein_coding	1	89633140	89933250
ENSG00000231613	RP5-943J3.1	sense_intronic	1	89788914	89790492
ENSG00000229992	HMGB3P9	processed_pseudogene	1	92647048	92647630
ENSG00000234202	RP11-330C7.4	processed_pseudogene	1	92749175	92749330
ENSG00000223896	CCNJP2	processed_pseudogene	1	92755794	92756956
ENSG00000229052	RP11-386I23.1	transcribed_processed_pseudogene	1	92930696	92934098
ENSG00000235501	RP4-639F20.1	antisense	1	94927361	94963616
ENSG00000270911	RP11-272L13.4	processed_pseudogene	1	97855575	97856825
ENSG00000271810	RP11-426L16.10	protein_coding	1	112702614	112711433
ENSG00000231128	RP5-1073O3.2	antisense	1	113812379	113870159
ENSG00000228035	RP4-663N10.1	antisense	1	115283034	115368072
ENSG00000227242	NBPF13P	unprocessed_pseudogene	1	147019656	147056593
ENSG00000225279	RP11-536C5.2	antisense	1	160062461	160080769
ENSG00000213070	HMGB3P6	processed_pseudogene	1	164356767	164357364
ENSG00000227907	RP11-102C16.3	antisense	1	167052551	167058542
ENSG00000228687	RP3-419C19.3	unprocessed_pseudogene	1	192796533	192797205
ENSG00000226814	RP3-419C19.2	processed_pseudogene	1	192800571	192801141
ENSG00000226862	RP11-569A11.1	antisense	1	202604268	202605293
ENSG00000279333	RP11-75I2.3	TEC	1	210678315	210681932
ENSG00000232679	RP11-400N13.3	lincRNA	1	222010825	222064773
ENSG00000234601	CHRM3-AS1	antisense	1	239898016	239899872
ENSG00000227210	AC079145.4	antisense	2	19990209	20004795
ENSG00000228563	AL133247.2	antisense	2	31526942	31563464
ENSG00000232526	VDAC1P13	processed_pseudogene	2	42463139	42463997
ENSG00000236913	AC025750.5	processed_pseudogene	2	42469817	42470266
ENSG00000215263	AC025750.7	processed_pseudogene	2	42532766	42533442
ENSG00000226523	AC074375.1	processed_pseudogene	2	42680088	42680279
ENSG00000232202	AC098824.6	processed_pseudogene	2	42826322	42827617
ENSG00000230327	AC009234.1	processed_pseudogene	2	50588690	50589415
ENSG00000264525	MIR4778	miRNA	2	66358249	66358328

ENSG00000264051	ENSG00000264051		2	71125523	71125592
ENSG00000213301	HMGB3P11	processed_pseudogene	2	104771282	104771885
ENSG00000231731	AC010976.2	antisense	2	127455394	127514623
ENSG00000264684	MIR4773-1	miRNA	2	151368334	151368411
ENSG00000128692	EIF2S2P4	processed_pseudogene	2	170751805	170752788
ENSG00000239467	AC007405.6	lincRNA	2	170766878	170778729
ENSG00000235934	AC007405.8	antisense	2	170814686	170818037
ENSG00000237016	AC013410.1	processed_pseudogene	2	173575110	173576341
ENSG00000227227	AC017101.10	antisense	2	186641339	186695287
ENSG00000235337	AC114765.2	lincRNA	2	220127546	220212491
ENSG00000224514	LINC00620	antisense	3	13650696	13746638
ENSG00000221659	AC104441.1	miRNA	3	21138263	21138341
ENSG00000243715	CACNA2D3-AS1	antisense	3	54874605	54901255
ENSG00000242120	RP11-231I13.2	antisense	3	70194479	70312638
ENSG00000240241	RP11-314M24.1	lincRNA	3	78266940	78294731
ENSG00000270880	RP11-255E6.5	processed_pseudogene	3	113984037	113984655
ENSG00000241810	HMGN2P13	processed_pseudogene	3	153067278	153067551
ENSG00000242107	LINC01100	lincRNA	3	160016024	160031423
ENSG00000233441	CYP2AB1P	transcribed_unprocessed_ pseudogene	3	183895900	183910936
ENSG00000248343	RP11-556G22.2	antisense	4	21697450	21719026
ENSG00000249649	MRPS33P2	processed_pseudogene	4	38006784	38007105
ENSG00000249348	UGDH-AS1	antisense	4	39528019	39594707
ENSG00000221891	FAM218BP	processed_pseudogene	4	164929494	164929951
ENSG00000270960	RP11-366M4.18	processed_pseudogene	4	164933086	164933506
ENSG00000249557	HSPD1P15	processed_pseudogene	5	19233366	19234487
ENSG00000183666	GUSBP1	transcribed_unprocessed_ pseudogene	5	21341833	21589372
ENSG00000238335	ENSG00000238335		5	37801971	37802065
ENSG00000248587	GDNF-AS1	lincRNA	5	37811589	37953827
ENSG00000248489	CTD-2007H13.3	lincRNA	5	98929171	98995013
ENSG00000249619	HMGN1P13	processed_pseudogene	5	111572102	111572239
ENSG00000251099	CTD-2367A17.1	processed_pseudogene	5	111572236	111572398
ENSG00000248753	CTC-276P9.2	lincRNA	5	135120526	135139681
ENSG00000279739	CTB-105N12.2	TEC	5	168128856	168130422
ENSG00000253768	CTB-33O18.1	lincRNA	5	173562478	173573199
ENSG00000260239	RP11-274H24.1	lincRNA	6	4488798	4495769
ENSG00000272168	CASC15	lincRNA	6	21664185	22654455
ENSG00000278128	RP11-440J4.2	processed_pseudogene	6	23854139	23857646
ENSG00000218512	SPTLC1P2	processed_pseudogene	6	23856698	23856818
ENSG00000271162	RP1-57E3.1	processed_pseudogene	6	48952070	48952781
ENSG00000277043	EEF1A1P42	processed_pseudogene	6	49358185	49360420
ENSG00000221312	AL109922.1	miRNA	6	64974393	64974492
ENSG00000233237	LINC00472	lincRNA	6	71295173	71417436

ENSG00000269966	RP3-331H24.5	lincRNA	6	71343427	71420745
ENSG00000279289	RP3-331H24.6	TEC	6	71386852	71390829
ENSG00000261038	RP1-149C7.1	lincRNA	6	92387841	92388827
ENSG00000271860	RP11-436D23.1	lincRNA	6	97283303	98400869
ENSG00000216809	RP11-57K17.1	processed_pseudogene	6	118452469	118454992
ENSG00000169075	RP3-509L4.3	processed_pseudogene	6	118501430	118502906
ENSG00000275122	THEMIS	protein_coding	6	127708072	127918597
ENSG00000273993	PTPRK	protein_coding	6	127968779	128524810
ENSG00000276680	HYMAI	misc_RNA	6	144007227	144007464
ENSG00000248711	THUMPD3P1	unprocessed_pseudogene	7	40151613	40154151
ENSG00000232458	LINC01450	lincRNA	7	40964667	40979939
ENSG00000261467	RP11-731K22.1	lincRNA	7	73985992	73988767
ENSG00000231322	RPL13AP17	transcribed_processed_pseudogene	7	78347142	78359458
ENSG00000182965	NPM1P14	processed_pseudogene	7	112520488	112521670
ENSG00000234520	HRAT17	lincRNA	7	112953282	112995677
ENSG00000234826	AC003084.2	lincRNA	7	117998858	118004045
ENSG00000240790	AC073934.6	antisense	7	127644685	127652012
ENSG00000254007	RP11-281H11.1	lincRNA	8	5659679	5670140
ENSG00000254260	RP11-369E15.1	lincRNA	8	20941428	20950175
ENSG00000254033	RP11-431M3.2	lincRNA	8	34229294	34248946
ENSG00000270328	RP11-33E15.1	processed_pseudogene	8	56776283	56776874
ENSG00000254557	RP11-102F4.2	antisense	8	69713390	69719722
ENSG00000272254	RP11-531A24.7	antisense	8	73042910	73043346
ENSG00000253334	RP13-923O23.6	antisense	8	81791823	81815714
ENSG00000231942	HNRNPA1P36	processed_pseudogene	8	81807772	81808726
ENSG00000254689	RP11-354A14.1	lincRNA	8	81841952	81924348
ENSG00000253434	RP11-1101K5.1	lincRNA	8	111376639	111757710
ENSG00000253756	CARSP2	processed_pseudogene	8	114791653	114791929
ENSG00000253926	RP11-26E5.1	lincRNA	8	129415949	129445852
ENSG00000274287	SCRIB	protein_coding	8	143777927	143802386
ENSG00000276893	AL136231.1	miRNA	9	4719746	4719841
ENSG00000236924	RP11-390F4.6	lincRNA	9	6645956	6670635
ENSG00000229057	RPS3AP54	processed_pseudogene	9	6662424	6663792
ENSG00000231381	RNF2P1	processed_pseudogene	9	6668999	6670008
ENSG00000232946	RP11-390F4.2	processed_pseudogene	9	6675284	6675614
ENSG00000230920	RP11-527D15.1	lincRNA	9	9799423	9803784
ENSG00000276759	RP11-80I3.1	lincRNA	9	26066675	26118408
ENSG00000226562	CYP4F26P	lincRNA	9	33580695	33607874
ENSG00000275270	AL133391.1	miRNA	9	82705777	82705844
ENSG00000236115	RP11-62C3.6	antisense	9	92205356	92210158
ENSG00000241621	GOLGA2P6	unprocessed_pseudogene	10	30364410	30374386
ENSG00000220997	AL161651.1	miRNA	10	30365067	30365148

ENSG00000276347	MIR7162	miRNA	10	30368597	30368665
ENSG00000225482	RP11-542E6.3	processed_pseudogene	10	48292699	48293417
ENSG00000231906	RP11-541M12.3	processed_pseudogene	10	48546472	48546713
ENSG00000225299	RP11-344A5.1	antisense	10	67052609	67055028
ENSG00000226051	ZNF503-AS1	lincRNA	10	75243568	75373500
ENSG00000270087	RP11-399K21.11	lincRNA	10	75276376	75276646
ENSG00000233313	HMGA1P5	processed_pseudogene	10	75279726	75401246
ENSG00000108242	CYP2C18	protein_coding	10	94683729	94736190
ENSG00000276490	RP11-400G3.5	protein_coding	10	94688154	94853073
ENSG00000234640	RP11-264E18.1	lincRNA	10	128316282	128320214
ENSG00000229093	OR51AB1P	unprocessed_pseudogene	11	5291761	5292380
ENSG00000255202	RP4-541C22.5	antisense	11	33665220	33696701
ENSG00000254514	RP11-958J22.3	lincRNA	11	45582525	45583474
ENSG00000255217	OR5J7P	unprocessed_pseudogene	11	56165500	56165921
ENSG00000172489	OR5T3	protein_coding	11	56252254	56253222
ENSG00000272987	OR5AL1	polymorphic_pseudogene	11	56412696	56413680
ENSG00000254660	CTD-3051L14.14	processed_pseudogene	11	56509573	56510195
ENSG00000254603	OR5M6P	unprocessed_pseudogene	11	56512235	56513171
ENSG00000279295	RP11-683O4.1	TEC	11	79966805	79968708
ENSG00000254471	RP11-885L14.1	processed_pseudogene	11	79987513	79989630
ENSG00000279248	RP11-643G5.1	TEC	11	89533645	89534043
ENSG00000280385	AP000648.5	TEC	11	90193614	90198120
ENSG00000280367	RP11-121L10.2	TEC	11	90223153	90226538
ENSG00000254890	RP11-10A7.1	processed_pseudogene	11	109486968	109487424
ENSG00000262607	IQSEC3	protein_coding	12	70780	175481
ENSG00000256694	RP11-598F7.5	antisense	12	164664	166321
ENSG00000256540	RP11-598F7.6	lincRNA	12	166856	182399
ENSG00000262016	RP11-598F7.5	antisense	12	168679	170336
ENSG00000262193	RP11-598F7.6	antisense	12	170871	171722
ENSG00000247934	RP11-967K21.1	antisense	12	28163298	28190738
ENSG00000278733	RP11-425D17.1	antisense	12	28185625	28186190
ENSG00000229899	AC084290.2	processed_pseudogene	12	40286289	40286662
ENSG00000257528	KRT8P19	processed_pseudogene	12	61879318	61880772
ENSG00000268721	RP11-588H23.5	processed_pseudogene	12	70446460	70447585
ENSG00000258066	RP11-781A6.1	lincRNA	12	77775783	77783576
ENSG00000258846	EEF1A1P33	processed_pseudogene	12	96900697	96902698
ENSG00000221386	AC002979.1	miRNA	12	111302800	111302899
ENSG00000273700	MIR6760	miRNA	12	111304142	111304209
ENSG00000274697	MIR6761	miRNA	12	111799834	111799905
ENSG00000270018	RP3-462E2.5		12	111897336	111897906
ENSG00000229186	ADAM1A	unitary_pseudogene	12	111899263	111901391
ENSG00000213152	RPL7AP60	processed_pseudogene	12	112301870	112302663
ENSG00000257658	RP3-521E19.3	processed_pseudogene	12	112321759	112323089

ENSG00000275759	RP11-131L12.3	lincRNA	12	118428281	118428870
ENSG00000275409	RP11-131L12.4	lincRNA	12	118430147	118430699
ENSG00000247373	RP11-486O12.2	lincRNA	12	123575891	123585115
ENSG00000279953	RP11-338K17.5	TEC	12	123649068	123650485
ENSG00000278861	RP11-338K17.6	TEC	12	123655528	123656128
ENSG00000255839	RP11-338K17.8	lincRNA	12	123707602	123708386
ENSG00000279087	RP11-83B20.10	TEC	12	124637999	124638444
ENSG00000255944	RP11-407A16.4	lincRNA	12	126628172	126690322
ENSG00000256581	NLRP9P	unprocessed_pseudogene	12	129013336	129016663
ENSG00000276476	LINC00540	lincRNA	13	22040972	22276526
ENSG00000278305	RP11-326L9.1	lincRNA	13	31419989	31437999
ENSG00000275149	RP11-427J23.1	lincRNA	13	40079106	40273509
ENSG00000271140	PRR20FP	unprocessed_pseudogene	13	57173453	57173632
ENSG00000230009	MTCO2P3	processed_pseudogene	13	57204379	57204799
ENSG00000279657	RP11-521H3.3	TEC	13	60730176	60733600
ENSG00000258906	CTD-2552B11.3	processed_pseudogene	14	21184341	21184959
ENSG00000241984	RPL7AP2	processed_pseudogene	14	39156742	39157852
ENSG00000271424	RP11-647O20.1	processed_pseudogene	14	39370279	39370756
ENSG00000258487	RP11-99L13.1	lincRNA	14	44444747	44480626
ENSG00000258510	RP11-493G17.4	processed_pseudogene	14	77359423	77359916
ENSG00000258615	RP11-725G5.3	unprocessed_pseudogene	14	95199039	95199345
ENSG00000272291	ENSG00000272291		14	100875033	100875104
ENSG00000259448	RP11-16E12.1	lincRNA	15	31215622	31224445
ENSG00000259740	RP11-142J21.2	sense_intronic	15	47525169	47527756
ENSG00000259377	CTD-2308G16.1	sense_intronic	15	51752485	51756475
ENSG00000263155	MYZAP	protein_coding	15	57591904	57685364
ENSG00000259703	LINC00593	lincRNA	15	69835234	69843120
ENSG00000261229	MTHFS	lincRNA	15	79843547	79844304
ENSG00000271983	RP11-28H5.2	lincRNA	15	80693216	80693707
ENSG00000259561	RP11-300G22.2	sense_intronic	15	89704665	89705415
ENSG00000277654	RP11-255M2.3	lincRNA	15	94841059	95169864
ENSG00000260342	RP11-1035H13.3	protein_coding	16	18788063	18801519
ENSG00000260352	CTD-2288F12.1	antisense	16	18926863	18937043
ENSG00000265515	AC092287.1	miRNA	16	18930196	18930282
ENSG00000228480	AC130464.1	unprocessed_pseudogene	16	26354796	26358355
ENSG00000261190	C16orf97	processed_transcript	16	52005487	52078474
ENSG00000261436	RP11-354I13.2	lincRNA	16	60346478	60499364
ENSG00000260832	CTD-3253I12.1	antisense	16	82953230	82990298
ENSG00000232190	RP11-134D3.1	lincRNA	16	87057784	87070105
ENSG00000273388	RP11-401O9.4	antisense	17	10291820	10317926
ENSG00000263383	AC005549.1	miRNA	17	34174245	34174317
ENSG00000236770	AC079325.5	unprocessed_pseudogene	17	74549399	74552737
ENSG00000231027	AC079325.6	unprocessed_pseudogene	17	74560716	74567343

ENSG00000267051	RP11-128P17.2	lincRNA	18	11366913	11378409
ENSG00000277534	RP11-9E17.1	sense_intronic	18	26542971	26545791
ENSG00000266549	RP11-526I8.2	processed_pseudogene	18	26952201	26952558
ENSG00000267374	RP11-244M2.1	lincRNA	18	39206917	39800322
ENSG00000267586	LINC00907	lincRNA	18	42159283	42691422
ENSG00000267101	RP11-846C15.2	antisense	18	45104361	45181382
ENSG00000267098	RP11-325K19.1	lincRNA	18	60628963	60902726
ENSG00000267011	CTB-50L17.16	lincRNA	19	4457962	4471493
ENSG00000268623	CTB-52I2.3	processed_pseudogene	19	18040355	18041083
ENSG00000243455	RPS18P13	processed_pseudogene	19	18048637	18049090
ENSG00000280106	CTC-523E23.3	TEC	19	34949038	34951205
ENSG00000226510	UPK1A-AS1	antisense	19	35667948	35673291
ENSG00000267439	AC002398.11	antisense	19	35747057	35753415
ENSG00000267328	AC002398.12	antisense	19	35754566	35755490
ENSG00000267049	AC002398.13	antisense	19	35769144	35771028
ENSG00000225872	LINC01529	lincRNA	19	35785697	35797902
ENSG00000280023	LLNLR-276H7.1	TEC	19	36070239	36071070
ENSG00000267005	AC002984.2	processed_pseudogene	19	36173406	36173853
ENSG00000267470	ZNF571-AS1	antisense	19	37548914	37587348
ENSG00000268051	CTC-244M17.1	sense_intronic	19	37963853	37964790
ENSG00000228567	VN1R4	protein_coding	19	53266676	53267723
ENSG00000269758	CTD-2245F17.9	lincRNA	19	53267870	53270932
ENSG00000163098	BIRC8	protein_coding	19	53289601	53291426
ENSG00000276688	RP5-828H9.4	lincRNA	20	5407564	5407876
ENSG00000252819	ENSG00000252819		20	9066834	9066908
ENSG00000233048	RP5-1069C8.2	lincRNA	20	12865202	12952519
ENSG00000280240	AL049794.1	protein_coding	20	16551149	16553012
ENSG00000226229	RPLP0P1	processed_pseudogene	20	16586333	16586693
ENSG00000238034	RP5-1185K9.1	lincRNA	20	21947647	21970783
ENSG00000203266	RP11-560A15.3	lincRNA	20	57110640	57122745
ENSG00000223669	RP11-93B14.4	antisense	20	62732566	62735347
ENSG00000226501	USF1P1	processed_pseudogene	21	33334571	33335096
ENSG00000224790	AP000704.5	lincRNA	21	36966461	36975164
ENSG00000215734	MRPL20P1	processed_pseudogene	21	36994643	36995075
ENSG00000273212	AC000068.10	antisense	22	19456503	19456962
ENSG00000253874	IGLVIV-66-1	IG_V_pseudogene	22	22058433	22058892
ENSG00000254075	IGLVV-66	IG_V_pseudogene	22	22061077	22061538
ENSG00000254161	IGLVIV-65	IG_V_pseudogene	22	22070225	22070707
ENSG00000253242	IGLVIV-64	IG_V_pseudogene	22	22075366	22075666
ENSG00000253752	IGLVI-63	IG_V_pseudogene	22	22079770	22080263
ENSG00000253823	IGLV1-62	IG_V_pseudogene	22	22086561	22087110
ENSG00000275361	LL22NC03-88E1.18	unprocessed_pseudogene	22	22087924	22088085
ENSG00000279582	ENSG00000279582		22	35051326	35061427

ENSG00000243902	ELFN2	sense_overlapping	22	37339583	37427445
ENSG00000272694	RP1-63G5.8	antisense	22	37371684	37372858
ENSG00000227188	MGAT3-AS1	antisense	22	39475807	39476822
ENSG00000278881	FP325331.1	protein_coding	22	47115838	47117217

Supplementary Table 6: The gene lists of Circadian Rhythm, Serotonin and Insomnia related pathways

CIRCADIAN-RHYTHM RELATED GENES					
GENE CODE	**GENE NAME**	**GENE TYPE**	**CHR**	**START**	**END**
ENSG00000049246	PER3	PROTEIN_CODING	1	7784291	7845177
ENSG00000049247	UTS2	PROTEIN_CODING	1	7843083	7853512
ENSG00000117318	ID3	PROTEIN_CODING	1	23557926	23559501
ENSG00000121764	HCRTR1	PROTEIN_CODING	1	31617686	31632518
ENSG00000116478	HDAC1	PROTEIN_CODING	1	32292083	32333635
ENSG00000116560	SFPQ	PROTEIN_CODING	1	35176378	35193145
ENSG00000054118	THRAP3	PROTEIN_CODING	1	36224432	36305357
ENSG00000162409	PRKAA2	PROTEIN_CODING	1	56645314	56715335
ENSG00000177606	JUN	PROTEIN_CODING	1	58776845	58784048
ENSG00000143125	PROK1	PROTEIN_CODING	1	110451149	110457358
ENSG00000159208	CIART	PROTEIN_CODING	1	150282543	150287093
ENSG00000143437	ARNT	PROTEIN_CODING	1	150809713	150876708
ENSG00000143365	RORC	PROTEIN_CODING	1	151806071	151831845
ENSG00000160716	CHRNB2	PROTEIN_CODING	1	154567778	154580013
ENSG00000198400	NTRK1	PROTEIN_CODING	1	156815640	156881850
ENSG00000135829	DHX9	PROTEIN_CODING	1	182839347	182887982
ENSG00000163485	ADORA1	PROTEIN_CODING	1	203090654	203167405
ENSG00000117707	PROX1	PROTEIN_CODING	1	213983181	214041510
ENSG00000054277	OPN3	PROTEIN_CODING	1	241590102	241677376
ENSG00000153187	HNRNPU	PROTEIN_CODING	1	244840638	244864560
ENSG00000115738	ID2	PROTEIN_CODING	2	8678845	8684461
ENSG00000134318	ROCK2	PROTEIN_CODING	2	11179759	11348330
ENSG00000213639	PPP1CB	PROTEIN_CODING	2	28751640	28802940
ENSG00000138083	SIX3	PROTEIN_CODING	2	44941702	44946071
ENSG00000143947	RPS27A	PROTEIN_CODING	2	55231903	55235853
ENSG00000198380	GFPT1	PROTEIN_CODING	2	69319780	69387250
ENSG00000204640	NMS	PROTEIN_CODING	2	100470482	100483280
ENSG00000170485	NPAS2	PROTEIN_CODING	2	100820139	100996829
ENSG00000184611	KCNH7	PROTEIN_CODING	2	162371407	162838767
ENSG00000118260	CREB1	PROTEIN_CODING	2	207529737	207605988
ENSG00000132326	PER2	PROTEIN_CODING	2	238244044	238290102
ENSG00000134107	BHLHE40	PROTEIN_CODING	3	4979437	4985323
ENSG00000157017	GHRL	PROTEIN_CODING	3	10285666	10292947
ENSG00000132170	PPARG	PROTEIN_CODING	3	12287368	12434356
ENSG00000174738	NR1D2	PROTEIN_CODING	3	23945286	23980617
ENSG00000163421	PROK2	PROTEIN_CODING	3	71771655	71785206
ENSG00000151577	DRD3	PROTEIN_CODING	3	114127580	114199407
ENSG00000082701	GSK3B	PROTEIN_CODING	3	119821321	120094994

ENSG00000168875	SOX14	PROTEIN_CODING	3	137764315	137766334
ENSG00000169760	NLGN1	PROTEIN_CODING	3	173396284	174286644
ENSG00000181092	ADIPOQ	PROTEIN_CODING	3	186842704	186858463
ENSG00000109819	PPARGC1A	PROTEIN_CODING	4	23755041	23904089
ENSG00000134852	CLOCK	PROTEIN_CODING	4	55427903	55546909
ENSG00000138669	PRKG2	PROTEIN_CODING	4	81087370	81215222
ENSG00000138668	HNRNPD	PROTEIN_CODING	4	82352498	82374503
ENSG00000109339	MAPK10	PROTEIN_CODING	4	85990007	86594625
ENSG00000138823	MTTP	PROTEIN_CODING	4	99564081	99623997
ENSG00000151014	NOCT	PROTEIN_CODING	4	139015781	139045939
ENSG00000185149	NPY2R	PROTEIN_CODING	4	155208636	155217078
ENSG00000168412	MTNR1A	PROTEIN_CODING	4	186533655	186555567
ENSG00000066230	SLC9A3	PROTEIN_CODING	5	470456	524449
ENSG00000132356	PRKAA1	PROTEIN_CODING	5	40759389	40798374
ENSG00000164326	CARTPT	PROTEIN_CODING	5	71719275	71721048
ENSG00000152413	HOMER1	PROTEIN_CODING	5	79372636	79514134
ENSG00000113558	SKP1	PROTEIN_CODING	5	134148935	134176964
ENSG00000120738	EGR1	PROTEIN_CODING	5	138465479	138469303
ENSG00000171720	HDAC3	PROTEIN_CODING	5	141620876	141636849
ENSG00000072803	FBXW11	PROTEIN_CODING	5	171861549	172006873
ENSG00000184845	DRD1	PROTEIN_CODING	5	175440036	175444182
ENSG00000050748	MAPK9	PROTEIN_CODING	5	180233143	180292099
ENSG00000172201	ID4	PROTEIN_CODING	6	19837370	19842197
ENSG00000183826	BTBD9	PROTEIN_CODING	6	38168451	38640148
ENSG00000137252	HCRTR2	PROTEIN_CODING	6	55106460	55282617
ENSG00000196591	HDAC2	PROTEIN_CODING	6	113933028	114011308
ENSG00000106546	AHR	PROTEIN_CODING	7	16916359	17346152
ENSG00000136244	IL6	PROTEIN_CODING	7	22725884	22732002
ENSG00000164742	ADCY1	PROTEIN_CODING	7	45574140	45723116
ENSG00000132437	DDC	PROTEIN_CODING	7	50458436	50565405
ENSG00000189143	CLDN4	PROTEIN_CODING	7	73799542	73832690
ENSG00000106366	SERPINE1	PROTEIN_CODING	7	101127104	101139247
ENSG00000105835	NAMPT	PROTEIN_CODING	7	106248298	106285966
ENSG00000184408	KCND2	PROTEIN_CODING	7	120273175	120750337
ENSG00000106331	PAX4	PROTEIN_CODING	7	127610292	127618142
ENSG00000174697	LEP	PROTEIN_CODING	7	128241278	128257629
ENSG00000055130	CUL1	PROTEIN_CODING	7	148697914	148801110
ENSG00000106462	EZH2	PROTEIN_CODING	7	148807257	148884321
ENSG00000168484	SFTPC	PROTEIN_CODING	8	22156913	22164479
ENSG00000158941	CCAR2	PROTEIN_CODING	8	22604757	22620964
ENSG00000179388	EGR3	PROTEIN_CODING	8	22687659	22693480
ENSG00000147465	STAR	PROTEIN_CODING	8	38142700	38150992
ENSG00000253729	PRKDC	PROTEIN_CODING	8	47773111	47960178

ENSG00000147571	CRH	PROTEIN_CODING	8	66176376	66178464
ENSG00000140396	NCOA2	PROTEIN_CODING	8	70109782	70403808
ENSG00000155090	KLF10	PROTEIN_CODING	8	102648784	102655725
ENSG00000119138	KLF9	PROTEIN_CODING	9	70384604	70414657
ENSG00000198963	RORB	PROTEIN_CODING	9	74497335	74693177
ENSG00000156052	GNAQ	PROTEIN_CODING	9	77716097	78031811
ENSG00000165030	NFIL3	PROTEIN_CODING	9	91409045	91423832
ENSG00000107290	SETX	PROTEIN_CODING	9	132261356	132354986
ENSG00000107317	PTGDS	PROTEIN_CODING	9	136975092	136981742
ENSG00000152455	SUV39H2	PROTEIN_CODING	10	14878820	14904315
ENSG00000095794	CREM	PROTEIN_CODING	10	35126791	35212958
ENSG00000107643	MAPK8	PROTEIN_CODING	10	48306639	48439360
ENSG00000096717	SIRT1	PROTEIN_CODING	10	67884656	67918390
ENSG00000179774	ATOH7	PROTEIN_CODING	10	68230595	68232113
ENSG00000180644	PRF1	PROTEIN_CODING	10	70597348	70602759
ENSG00000156113	KCNMA1	PROTEIN_CODING	10	76869601	77638369
ENSG00000122375	OPN4	PROTEIN_CODING	10	86654518	86666460
ENSG00000171862	PTEN	PROTEIN_CODING	10	87862638	87971930
ENSG00000026103	FAS	PROTEIN_CODING	10	88953813	89029605
ENSG00000148680	HTR7	PROTEIN_CODING	10	90740823	90858039
ENSG00000166167	BTRC	PROTEIN_CODING	10	101354033	101557321
ENSG00000214285	NPS	PROTEIN_CODING	10	127549309	127553540
ENSG00000069696	DRD4	PROTEIN_CODING	11	637269	640706
ENSG00000180176	TH	PROTEIN_CODING	11	2163929	2171815
ENSG00000170955	CAVIN3	PROTEIN_CODING	11	6318946	6320532
ENSG00000133794	ARNTL	PROTEIN_CODING	11	13276652	13387266
ENSG00000129167	TPH1	PROTEIN_CODING	11	18017555	18046269
ENSG00000205213	LGR4	PROTEIN_CODING	11	27365961	27472790
ENSG00000121671	CRY2	PROTEIN_CODING	11	45847118	45883248
ENSG00000025434	NR1H3	PROTEIN_CODING	11	47248300	47269033
ENSG00000168539	CHRM1	PROTEIN_CODING	11	62908679	62921807
ENSG00000173933	RBM4	PROTEIN_CODING	11	66638667	66668374
ENSG00000173914	RBM4B	PROTEIN_CODING	11	66664998	66677887
ENSG00000172531	PPP1CA	PROTEIN_CODING	11	67398181	67421183
ENSG00000110090	CPT1A	PROTEIN_CODING	11	68754620	68844410
ENSG00000134640	MTNR1B	PROTEIN_CODING	11	92969651	92985066
ENSG00000149295	DRD2	PROTEIN_CODING	11	113409605	113475691
ENSG00000118058	KMT2A	PROTEIN_CODING	11	118436456	118526832
ENSG00000036672	USP2	PROTEIN_CODING	11	119355215	119381711
ENSG00000073614	KDM5A	PROTEIN_CODING	12	280057	389320
ENSG00000013583	HEBP1	PROTEIN_CODING	12	12974870	13000265
ENSG00000123095	BHLHE41	PROTEIN_CODING	12	26120030	26125037
ENSG00000029153	ARNTL2	PROTEIN_CODING	12	27332836	27425289

ENSG00000111602	TIMELESS	PROTEIN_CODING	12	56416363	56449426
ENSG00000135446	CDK4	PROTEIN_CODING	12	57747727	57756013
ENSG00000139287	TPH2	PROTEIN_CODING	12	71938845	72186618
ENSG00000008405	CRY1	PROTEIN_CODING	12	106991364	107093549
ENSG00000186298	PPP1CC	PROTEIN_CODING	12	110719680	110742939
ENSG00000150991	UBC	PROTEIN_CODING	12	124911604	124917368
ENSG00000121390	PSPC1	PROTEIN_CODING	13	19674752	19783019
ENSG00000005812	FBXL3	PROTEIN_CODING	13	76992598	77027195
ENSG00000057593	F7	PROTEIN_CODING	13	113105788	113120685
ENSG00000165819	METTL3	PROTEIN_CODING	14	21498133	21511342
ENSG00000100462	PRMT5	PROTEIN_CODING	14	22920525	22929408
ENSG00000136352	NKX2-1	PROTEIN_CODING	14	36516392	36521149
ENSG00000198894	CIPC	PROTEIN_CODING	14	77098126	77117287
ENSG00000182979	MTA1	PROTEIN_CODING	14	105419820	105470729
ENSG00000254585	MAGEL2	PROTEIN_CODING	15	23643549	23647867
ENSG00000114062	UBE3A	PROTEIN_CODING	15	25333728	25439051
ENSG00000069667	RORA	PROTEIN_CODING	15	60488284	61229302
ENSG00000140464	PML	PROTEIN_CODING	15	73994673	74047827
ENSG00000169375	SIN3A	PROTEIN_CODING	15	75369379	75455842
ENSG00000172379	ARNT2	PROTEIN_CODING	15	80404350	80597933
ENSG00000284059	MIR5572	MIRNA	15	80581103	80581239
ENSG00000140538	NTRK3	PROTEIN_CODING	15	87859751	88256768
ENSG00000122254	HS3ST2	PROTEIN_CODING	16	22814162	22916338
ENSG00000159723	AGRP	PROTEIN_CODING	16	67482571	67483547
ENSG00000140836	ZFHX3	PROTEIN_CODING	16	72782885	73891871
ENSG00000132382	MYBBP1A	PROTEIN_CODING	17	4538897	4555386
ENSG00000141510	TP53	PROTEIN_CODING	17	7661779	7687538
ENSG00000179094	PER1	PROTEIN_CODING	17	8140472	8156506
ENSG00000284117	MIR6883	MIRNA	17	8144994	8145071
ENSG00000141027	NCOR1	PROTEIN_CODING	17	16029157	16218185
ENSG00000108557	RAI1	PROTEIN_CODING	17	17681458	17811453
ENSG00000072310	SREBF1	PROTEIN_CODING	17	17810399	17837002
ENSG00000007171	NOS2	PROTEIN_CODING	17	27756766	27800529
ENSG00000108576	SLC6A4	PROTEIN_CODING	17	30194319	30236002
ENSG00000275410	HNF1B	PROTEIN_CODING	17	37686431	37745059
ENSG00000126368	NR1D1	PROTEIN_CODING	17	40092793	40100589
ENSG00000131747	TOP2A	PROTEIN_CODING	17	40388525	40417896
ENSG00000108784	NAGLU	PROTEIN_CODING	17	42536241	42544449
ENSG00000064300	NGFR	PROTEIN_CODING	17	49495293	49515008
ENSG00000108654	DDX5	PROTEIN_CODING	17	64498254	64508199
ENSG00000284564	MIR3064	MIRNA	17	64500774	64500839
ENSG00000284368	MIR5047	MIRNA	17	64501214	64501313
ENSG00000129673	AANAT	PROTEIN_CODING	17	76453351	76470117

ENSG00000141551	CSNK1D	PROTEIN_CODING	17	82239023	82273700
ENSG00000181408	UTS2R	PROTEIN_CODING	17	82371765	82377458
ENSG00000176890	TYMS	PROTEIN_CODING	18	657653	673578
ENSG00000141655	TNFRSF11A	PROTEIN_CODING	18	62325287	62391288
ENSG00000081913	PHLPP1	PROTEIN_CODING	18	62715541	62980433
ENSG00000088256	GNA11	PROTEIN_CODING	19	3094362	3123999
ENSG00000160113	NR2F6	PROTEIN_CODING	19	17231883	17245919
ENSG00000130522	JUND	PROTEIN_CODING	19	18279694	18281622
ENSG00000221983	UBA52	PROTEIN_CODING	19	18571730	18577550
ENSG00000105662	CRTC1	PROTEIN_CODING	19	18683678	18782333
ENSG00000104856	RELB	PROTEIN_CODING	19	45001449	45038198
ENSG00000105392	CRX	PROTEIN_CODING	19	47819779	47843330
ENSG00000105516	DBP	PROTEIN_CODING	19	48630030	48637379
ENSG00000169136	ATF5	PROTEIN_CODING	19	49928702	49933935
ENSG00000283842	MIR4751	MIRNA	19	49933064	49933137
ENSG00000126583	PRKCG	PROTEIN_CODING	19	53879190	53907652
ENSG00000101200	AVP	PROTEIN_CODING	20	3082556	3084724
ENSG00000101292	PROKR2	PROTEIN_CODING	20	5299218	5316954
ENSG00000101439	CST3	PROTEIN_CODING	20	23626706	23638473
ENSG00000101444	AHCY	PROTEIN_CODING	20	34280268	34311802
ENSG00000118702	GHRH	PROTEIN_CODING	20	37251086	37261819
ENSG00000198900	TOP1	PROTEIN_CODING	20	41028822	41124487
ENSG00000196839	ADA	PROTEIN_CODING	20	44619522	44652233
ENSG00000124089	MC3R	PROTEIN_CODING	20	56248732	56249815
ENSG00000125510	OPRL1	PROTEIN_CODING	20	64080082	64100643
ENSG00000180530	NRIP1	PROTEIN_CODING	21	14961235	15065936
ENSG00000157540	DYRK1A	PROTEIN_CODING	21	37365573	37526358
ENSG00000142178	SIK1	PROTEIN_CODING	21	43414483	43427131
ENSG00000128271	ADORA2A	PROTEIN_CODING	22	24417879	24442357
ENSG00000213923	CSNK1E	PROTEIN_CODING	22	38290691	38318084
ENSG00000283900	TPTEP2-CSNK1E	PROTEIN_CODING	22	38291918	38398522
ENSG00000128272	ATF4	PROTEIN_CODING	22	39519695	39522683
ENSG00000100393	EP300	PROTEIN_CODING	22	41092513	41092566
ENSG00000284015	MIR1281	MIRNA	22	41092592	41180077
ENSG00000186951	PPARA	PROTEIN_CODING	22	46150521	46243755

SEROTONIN-RELATED GENES

GENE CODE	GENE NAME	GENE TYPE	CHR	START	END
ENSG00000158748	HTR6	PROTEIN_CODING	1	19664875	19680966
ENSG00000179546	HTR1D	PROTEIN_CODING	1	23191895	23217502
ENSG00000184155	OR10J5	PROTEIN_CODING	1	159535078	159536007
ENSG00000158731	OR10J6P	UNPROCESSED_PSEUDOGENE	1	159598298	159599227

ENSG00000133019	CHRM3	PROTEIN_CODING	1	239386565	239915452
ENSG00000135914	HTR2B	PROTEIN_CODING	2	231108230	231125042
ENSG00000196639	HRH1	PROTEIN_CODING	3	11137093	11263557
ENSG00000179097	HTR1F	PROTEIN_CODING	3	87792706	87993839
ENSG00000186090	HTR3D	PROTEIN_CODING	3	184031544	184039369
ENSG00000178084	HTR3C	PROTEIN_CODING	3	184053047	184060673
ENSG00000186038	HTR3E	PROTEIN_CODING	3	184097064	184106995
ENSG00000145335	SNCA	PROTEIN_CODING	4	89700345	89838315
ENSG00000145545	SRD5A1	PROTEIN_CODING	5	6633427	6674386
ENSG00000178394	HTR1A	PROTEIN_CODING	5	63957874	63962507
ENSG00000164197	RNF180	PROTEIN_CODING	5	64165843	64372869
ENSG00000164270	HTR4	PROTEIN_CODING	5	148451032	148677235
ENSG00000113749	HRH2	PROTEIN_CODING	5	175658030	175710756
ENSG00000135312	HTR1B	PROTEIN_CODING	6	77460924	77463491
ENSG00000168830	HTR1E	PROTEIN_CODING	6	86937528	87016679
ENSG00000118432	CNR1	PROTEIN_CODING	6	88139864	88166347
ENSG00000175003	SLC22A1	PROTEIN_CODING	6	160121815	160158718
ENSG00000112499	SLC22A2	PROTEIN_CODING	6	160171061	160277638
ENSG00000146477	SLC22A3	PROTEIN_CODING	6	160348378	160452577
ENSG00000164638	SLC29A4	PROTEIN_CODING	7	5274369	5306912
ENSG00000181072	CHRM2	PROTEIN_CODING	7	136868652	137020255
ENSG00000157219	HTR5A	PROTEIN_CODING	7	155070324	155087392
ENSG00000036565	SLC18A1	PROTEIN_CODING	8	20144855	20183206
ENSG00000165025	SYK	PROTEIN_CODING	9	90801787	90898549
ENSG00000165646	SLC18A2	PROTEIN_CODING	10	117241093	117279430
ENSG00000180720	CHRM4	PROTEIN_CODING	11	46383789	46391776
ENSG00000181718	OR5T2	PROTEIN_CODING	11	56231282	56234255
ENSG00000172489	OR5T3	PROTEIN_CODING	11	56252254	56253222
ENSG00000149305	HTR3B	PROTEIN_CODING	11	113904796	113949079
ENSG00000166736	HTR3A	PROTEIN_CODING	11	113975075	113990313
ENSG00000181499	OR6T1	PROTEIN_CODING	11	123942867	123943838
ENSG00000123360	PDE1B	PROTEIN_CODING	12	54549601	54579239
ENSG00000089250	NOS1	PROTEIN_CODING	12	117208142	117452170
ENSG00000102468	HTR2A	PROTEIN_CODING	13	46831546	46897076
ENSG00000258806	OR11H7	POLYMORPHIC_PSEUDOGENE	14	20229402	20230346
ENSG00000176198	OR11H4	PROTEIN_CODING	14	20239286	20244349
ENSG00000184984	CHRM5	PROTEIN_CODING	15	33968497	34067458
ENSG00000183454	GRIN2A	PROTEIN_CODING	16	9753404	10182928
ENSG00000108405	P2RX1	PROTEIN_CODING	17	3896592	3916476
ENSG00000259207	ITGB3	PROTEIN_CODING	17	47253827	47313743
ENSG00000134489	HRH4	PROTEIN_CODING	18	24460637	24479961
ENSG00000171942	OR10H2	PROTEIN_CODING	19	15728024	15729052

ENSG00000171936	OR10H3	PROTEIN_CODING	19	15737985	15742343
ENSG00000172519	OR10H5	PROTEIN_CODING	19	15787661	15800357
ENSG00000186723	OR10H1	PROTEIN_CODING	19	15804549	15815664
ENSG00000176231	OR10H4	PROTEIN_CODING	19	15944299	15950350
ENSG00000104972	LILRB1	PROTEIN_CODING	19	54617158	54638022
ENSG00000101180	HRH3	PROTEIN_CODING	20	62214960	62220278

INSOMNIA-RELATED GENES

GENE CODE	GENE NAME	GENE TYPE	CHR	START	END
ENSG00000116288	PARK7	PROTEIN_CODING	1	7954291	7985505
ENSG00000184908	CLCNKB	PROTEIN_CODING	1	16040252	16057311
ENSG00000158828	PINK1	PROTEIN_CODING	1	20633458	20651511
ENSG00000116675	DNAJC6	PROTEIN_CODING	1	65248219	65415871
ENSG00000143641	GALNT2	PROTEIN_CODING	1	230057990	230282122
ENSG00000204843	DCTN1	PROTEIN_CODING	2	74361154	74392087
ENSG00000115317	HTRA2	PROTEIN_CODING	2	74529725	74533332
ENSG00000073734	ABCB11	PROTEIN_CODING	2	168915498	169031324
ENSG00000080819	CPOX	PROTEIN_CODING	3	98579446	98593648
ENSG00000197386	HTT	PROTEIN_CODING	4	3041363	3243957
ENSG00000154277	UCHL1	PROTEIN_CODING	4	41256413	41268455
ENSG00000145982	FARS2	PROTEIN_CODING	6	5261044	5829192
ENSG00000196126	HLA-DRB1	PROTEIN_CODING	6	32578769	32589848
ENSG00000179344	HLA-DQB1	PROTEIN_CODING	6	32659467	32668383
ENSG00000185345	PRKN	PROTEIN_CODING	6	161347417	162727775
ENSG00000009954	BAZ1B	PROTEIN_CODING	7	73440406	73522293
ENSG00000106638	TBL2	PROTEIN_CODING	7	73567537	73578791
ENSG00000049540	ELN	PROTEIN_CODING	7	74027789	74069907
ENSG00000106683	LIMK1	PROTEIN_CODING	7	74082933	74122525
ENSG00000049541	RFC2	PROTEIN_CODING	7	74231499	74254458
ENSG00000106665	CLIP2	PROTEIN_CODING	7	74289407	74405935
ENSG00000006704	GTF2IRD1	PROTEIN_CODING	7	74453790	74602605
ENSG00000263001	GTF2I	PROTEIN_CODING	7	74650231	74760692
ENSG00000005471	ABCB4	PROTEIN_CODING	7	87401696	87480435
ENSG00000004864	SLC25A13	PROTEIN_CODING	7	96120220	96322147
ENSG00000128567	PODXL	PROTEIN_CODING	7	131500262	131558217
ENSG00000138311	ZNF365	PROTEIN_CODING	10	62374192	62480288
ENSG00000133895	MEN1	PROTEIN_CODING	11	64803510	64811294
ENSG00000256269	HMBS	PROTEIN_CODING	11	119084866	119093834
ENSG00000059804	SLC2A3	PROTEIN_CODING	12	7919230	7936187
ENSG00000188906	LRRK2	PROTEIN_CODING	12	40196744	40369285
ENSG00000012504	NR1H4	PROTEIN_CODING	12	100473708	100564414
ENSG00000129003	VPS13C	PROTEIN_CODING	15	61852389	62060473
ENSG00000070915	SLC12A3	PROTEIN_CODING	16	56865207	56915850

ENSG00000258947	TUBB3	PROTEIN_CODING	16	89921392	89938761
ENSG00000161610	HCRT	PROTEIN_CODING	17	42184060	42185452
ENSG00000161653	NAGS	PROTEIN_CODING	17	44004622	44009068
ENSG00000081923	ATP8B1	PROTEIN_CODING	18	57646426	57803315
ENSG00000171867	PRNP	PROTEIN_CODING	20	4686350	4701590
ENSG00000159082	SYNJ1	PROTEIN_CODING	21	32628759	32728040