UNIVERSIDAD AUTONOMA DE MADRID

Programa de Doctorado en Epidemiología y Salud Pública

IDENTIFYING PANCREATIC CANCER RISK GENETIC VARIANTS

Evelina Mocci

Madrid 2018

Facultad de Medicina

Departamento de Medicina Preventiva, Salud Pública y Microbiología

IDENTIFYING PANCREATIC CANCER RISK GENETIC VARIANTS

Evelina Mocci

Director:
Dr. Alison Klein

Tutor:
Dr. Esther López García

Madrid, 2018

UNIVERSIDAD AUTONOMA
DE MADRID

Drs. Alison Klein and Fernando Rodríguez Artalejo, inform that the thesis entitled *"Identifying Pancreatic Cancer Risk Genetic Variants"* is an original work carried out by Miss Evelina Mocci under our guidance and supervision. This is an original work, rigorously carried out and is apt to be defended publicly in order to obtain the degree of Doctor on Epidemiology and Public Health.

For this to be recorded and to have the appropriate effects, this document is signed in Madrid, month 2018.

Alison Klein, PhD, MHS
Johns Hopkins University
School of Medicine
Director, GI SPORE
Professor of Oncology. Pathology and Epidemiology.

Fernando Rodríguez Artalejo, MD, PhD
Universidad Autónoma de Madrid
School of Medicine
Department of Preventive Medicine
Public Health and Microbiology.

Dedico questa tesi a tutta la mia Famiglia,

La mia dolcissima e grande mamma,
Al mio Angelo Sabrina,
My love Richard,
My baby Leia,
La mia adorata sorella Cristiana
E le mie bellissime nipoti
Benedetta & Angelica

ACKNOWLEDGMENTS

I want to thank the people who helped make this project possible.

Dr. Alison Klein, the scientific director of this thesis, for her valuable scientific contribution and her support.

Dr. Fernando Artalejo, the director of the PhD program for his availability and kindness.

Dr. Ghislane Scelo for kindly accepting to be my external thesis evaluator.

Drs. Goggins and Eshleman, great experts in the field of Pancreatic Cancer for kingly accepting to evaluating my thesis.

Milagros, David and Erica for their valid help in all administrative aspects of the PhD.

SUMMARY

Background: Pancreatic ductal adenocarcinoma (PDAC) is a highly deadly disease, with an incidence-mortality rates ratio close to one. Incidence rates vary widely across the world; more developed countries as Northern America and Western Europe show the highest number of PDAC cases versus the lower number registered in developing countries as South Central Asia and Middle Africa. While these statistics may be biased in developing countries due to the lack of appropriate health facilities, it is highly likely that these data reflect a different exposure to the main PDAC environmental risk factors.

About 20% of PDACs are attributable to cigarette smoking; long-standing diabetes, obesity, high alcohol consumption, and pancreatitis are other well-established risk factors. The main nonmodifiable risk factors for PDAC are age and family history of pancreatic cancer.

Approximately 5-10% of PDACs cluster within families. Part of these cases is attributable to high-penetrance germline mutations, mainly in DNA repair genes.

For the remaining 90% of PDACs, causes and genetic risk factors are under investigation. Genome-Wide Association Studies (GWASs) conducted so far in the Caucasian population discovered 23 independent common loci with a small-moderate (10-30%) effect on the disease. Despite the low impact on the risk of the disease, these findings have been of great importance to the knowledge of the molecular mechanisms involved in PDAC.

GWASs main limitation, in pancreatic cancer as in other complex diseases and traits, is the 'missing heritability'; it has been estimated that GWAS arrays explain for only about 13-16% of PDAC estimated heritability, meaning that most of the genetic factors associated with the disease are still unknown.

Geneticists are currently focusing on the study of rare and low-frequency genetic variants, particularly those located in the coding regions of the genome. Whole genome sequencing data from different studies demonstrate that the majority of human genome variants are rare. Rare variants are supposed to have a higher impact on the disease compared to common

ones; actually, deleterious variants accumulate in the population as extremely rare, because they undergo high negative selection pressure.

GWASs have limited statistical power to detect genome-wide significant association for rare variants; new analytical methods, generally addressed as aggregation tests, have been developed for this purpose. Aggregation tests measure the cumulative effect of rare variants located into a set (gene, region or pathway) on the phenotype.

This thesis presents two complementary studies conducted using the same dataset, the Pancreatic Cancer Case-Control Consortium (PanC4), including 4,164 PDAC cases and 3,792 controls recruited in 9 studies from North America, Central Europe and Australia.

Objectives:

Identifying novel genetic risk loci for PDAC

Methods: The first study, a two-stage GWAS, tests the association between PDAC and single common variants, by applying an unconditional logistic regression analysis adjusted for age and the top vectors of population ancestry under log-additive genetic model.

In stage 1, the association has been tested both on PanC4 genotyped data only and then through a combined analysis of genotyped and imputed data from the three datasets, PanC4, PanScan 1 and PanScan 2.

In stage 2, the top significant SNPs have been tested for replication in an independent population (PANDoRA) only and then in a combined analysis including PanC4, PanScan 1, PanScan 2 and PANDoRA data.

The second study, a whole exome-array analysis, focuses on rare and low frequency functional variants. We measured the association among the cumulative effect of these variants by gene and the disease using SKAT-O test.

Results: The two-stage GWAS identifies 3 novels common loci and replicated 1 genomic region previously described in the Han Chinese population. Furthermore it replicates 8 loci identified in previous GWASs.

The gene-based exome-array analysis does not show any exome-wide significant gene; however, it reports a novel attractive potential candidate gene with an excess of functional variants in cases compared to controls.

Conclusions: GWAS applied to common variants identifies novel loci with small –modest impact on the risk of PDAC. Larger sample size are needed to identifying exome-wide statistical association in functional variants with intermediate effect on the disease

Keywords: Pancreatic ductal adenocarcinoma, Genome-Wide Association Study, common genetic variants, rare and low-frequency genetic variants, aggregation tests

RESUMEN

Antecedentes El adenocarcinoma de páncreas (AP) es una enfermedad altamente mortal, con una tasa de incidencia-mortalidad cercana a uno. Las tasas de incidencia varían ampliamente en todo el mundo; los países más desarrollados como América del Norte y Europa occidental muestran el mayor número de casos de AP frente a la menor incidencia registrada en los países en desarrollo como el sur de Asia y África central.

Estos datos podrían estar sesgadas en los países en desarrollo debido a la falta de instalaciones de salud apropiadas para un correcto diagnostico; sin embargo es muy probable que estos datos reflejen una exposición diferente a los principales factores de riesgo ambientales del AP.

Alrededor del 20% de los AP son atribuibles al tabaquismo; la diabetes de larga data, la obesidad, el alto consumo de alcohol y la pancreatitis son otros factores de riesgo bien establecidos. Los principales factores de riesgo no modificables para el AP son la edad y los antecedentes familiares de cáncer de páncreas.

Aproximadamente el 5-10% de los AP se agrupan dentro de las familias. Parte de estos casos es atribuible a las mutaciones de la línea germinal, principalmente en los genes de reparación del ADN.

Para el restante 90% de los casos de AP, se sigue investigando sobre las causas y los factores de riesgo genéticos. Los estudios de Genoma-Wide Association (GWAS) realizado hasta ahora en la población Caucásica han descubierto 23 regiones genéticas independientes, con un efecto pequeño-moderado (10-30%) sobre la enfermedad. A pesar del bajo impacto sobre el riesgo de la enfermedad, estos hallazgos han sido de gran importancia para el conocimiento de los mecanismos moleculares implicados en el AP.

La principal limitación de GWAS, bien en el estudio de cáncer de páncreas como en otras enfermedades y rasgos complejos, es la "falta de heredabilidad"; de echo se ha estimado que las variantes incluida en el GWAS explican solo alrededor del 13-15% de la heredabilidad

estimada de AP, lo que significa que la mayoría de los factores genéticos asociados con la enfermedad aún se desconocen.

Actualmente, los genetistas se están enfocando en el estudio de variantes genéticas raras y de baja frecuencia, particularmente aquellas ubicadas en las regiones codificantes del genoma. De echo, los datos de secuenciación del genoma completo proveniente de diferentes estudios, demuestran que la mayoría de las variantes del genoma humano son raras. Además, se supone que las variantes raras tienen un mayor impacto en la enfermedades en comparación con las comunes; se ha demostrado que las variantes deletéreas se acumulan en la población como extremadamente raras como consecuencia de la fuerte presión negativa de selección

El análisis GWAS tiene bajo poder estadístico para identificar asociación entre enfermedad y variantes raras; nuevo métodos estadísticos, generalmente conocidos como test de agregación, han sido desarrollado para este propósito. El test de agregación calcula el efecto cumulativo de las variantes localizadas dentro de un conjunto (gen, región o pathway) en el fenotipo estudiado.

Esta tesis presenta dos estudios complementarios llevados a cabo en la misma base de datos que selecciona las muestra del consorcio de casos y controles de cáncer de páncreas (PanC4), y que incluye 4,164 casos de AP y 3,792 controles.

Objetivos:

El objetivo común es identificar los nuevos genes de riesgo para el AP.

Métodos: El primer estudio, un GWAS realizado en dos estadios, analiza la asociación entre AP y variantes comunes, mediante la aplicación de un análisis de regresión logística incondicional ajustado por edad y los principales vectores de ascendencia poblacional bajo el modelo genético log-aditivo.

En el estadio 1, la asociación entre variantes comunes y AP ha sido analizada en PanC4 solo y luego a través de un análisis combinado de las bases de datos, PanC4, PanScan 1 y PanScan 2.
En el estadio 2, las variantes con mas alta significatividad estadística han sido analizadas para la replicación en una población independiente (PANDoRA). Por ultimo, las mismas variantes han sido probado por asociación en un meta-analisis que incluye los datos de PanC4, PanScan 1, PanScan 2 y PANDoRA.
El segundo estudio, un análisis de asociación de genes en todo el genoma, se centra en variantes funcionales de frecuencia baja y rara. Medimos la asociación entre el efecto acumulativo de las variantes por gen y la enfermedad con SKAT-O test.

Resultados: el primer estudio identifica 3 regiones genéticas comunes y replica una región genómica previamente descrita en la población china Han.
Además, este estudio ha replicado ocho regiones genómica identificados en GWAS anteriores.
El análisis basado en el estudio de genes no identifica ningún gen significativo; sin embargo, destaca un nuevo gen candidato que presenta un exceso de las variantes funcionales en los casos en comparación con los controles.

Conclusiones: GWAS aplicado a variantes comunes identifica nuevas regiones genéticas que tienen un impacto pequeño-mediano sobre el riesgo de AP. Para identificar variantes funcionales con impacto intermedio en el riesgo de la enfermedad es necesario un tamaño de muestra mayor .
Palabras clave: adenocarcinoma ductal pancreático, estudio genómico de asociación, variantes genéticas comunes, variantes genéticas raras y de baja frecuencia, análisis de agregación

1. INTRODUCTION **1**

CHAPTER 1 EPIDEMIOLOGY OF PANCREATIC CANCER **1**

1.1 INCIDENCE, MORTALITY TRENDS, SURVIVAL PROGNOSIS 1
1.2 CIGARETTE SMOKING 2
1.3 DIABETES 3
1.4 BODY MASS INDEX 4
1.5 ALCOHOL 4
1.6 PANCREATITIS 5
1.7 DIETARY FACTORS 6
1.8 GASTROINTESTINAL MICROBIOME 6
1.9 ALLERGY 7
1.10 FAMILY HISTORY 8

CHAPTER 2 GENETIC LANDSCAPE OF PANCREATIC CANCER **9**

2.1 HEREDITARY PANCREATIC CANCER 9
2.2 COMMON VARIANTS ASSOCIATED WITH PANCREATIC CANCER 12
2.2.1 GWAS PANSCAN 1, 9Q24, *ABO* GENE 13
2.2.2 GWAS PANSCAN 2, 13Q22.*1 KLF5-KLF12*, 1Q32.1 *NR5A2*, 5P15.33 *TERT-CLPTM1L* 15
2.2.3 GWAS PANSCAN 3, 7Q32.2 *LINC-PINT*, 16Q23.Q *BCRA1*, 13Q12.2 *PDX1*, 22Q12.*1 ZNFR3* 18
2.2.4 GWAS PANC4, 17Q25.*1 LINC00673*, 7P13 *SUGT*, 3Q29 TP63, 2P13.3 *ETA1* 21
2.2.5 META-ANLYSIS PANSCAN1, PANSCAN2 AND PANSCAN 3:
NEW INDEPENDENT LOCI 1Q32.1, 5P15.33 AND 8Q24.1 21
2.2.6 META-ANLYSIS PANSCAN1, PANSCAN2, PANSCAN 3 AND PANC4
NEW LOCI 1P36.33, 7P12, 8Q21.11, 17Q12 AND 18Q21.32 22

CHAPTER 3 RARE VARIANTS **32**

3.1 DEFINITION, ORIGIN AND SOURCES 32

3.2 STATISTICAL METHODS FOR RARE VARIANTS 34
3.2.1 BURDEN TEST 36
3.2.2 VARIANCE COMPONENT 37
3.2.3 OMNIBUS 38

CHAPTER 4 GENOTYPE IMPUTATION **39**

2. HYPOTHESIS AND OBJECTIVES **42**

3. METHODS **45**

3.1 TWO-STAGE GWAS: COMMON VARIATION AT 2P13.3, 3Q29, 7P13 AND 17Q25.1 ARE ASSOCIATED WITH SUSCEPTIBILITY TO PDAC **45**

3.1.1 STUDY DESIGN 45

3.1.2 STAGE 1 47
3.1.2.1 STUDY POPULATION 47
PANC4 47
PANSCAN 1 AND PANSCAN 2 47
3.1.2.2 GENOTYPING AND QUALITY CONTROL 49
PANC4 49
PANSCAN 1 AND PANSCAN 2 49
3.1.2.3 ASSOCIATION ANALYSIS 52

3.1.3 STAGE 2 54
3.1.3.1 STUDY POPULATION 54
PANDORA REPLICATION STUDY 54
3.1.3.2 GENOTYPING AND QUALITY CONTROL 55
3.1.3.3 ASSOCIATION ANALYSIS 55

3.1.4 OTHER ANALYSES 55
3.1.4.1 HERITABILITY 55
3.1.4.2 HAPLOREG 56

3.2 EXOME ARRAY GENE-BASED ANALYSIS **57**

3.2.1 STUDY DESIGN 57
3.2.2 QUALITY CONTROL 59
3.2.3 GENOTYPE IMPUTATION 60
3.2.4 POST-IMPUTATION FILTERS 61
3.2.5 GENE ANNOTATION 61
3.2.6 POWER ANALYSIS 62
3.2.7 ASSOCIATION ANALYSIS 62

4. RESULTS **64**

4.1 TWO-STAGE GWAS: COMMON VARIATION AT 2P13.3, 3Q29, 7P13 AND 17Q25.1 ARE ASSOCIATED WITH SUSCEPTIBILITY TO PDAC **64**

4.1.1 PANC4 64
4.1.2 COMBINED STAGE 1 65
4.1.3 PANDORA REPLICATION 65
4.1.4 COMBINED STAGE 2 65
4.1.5 HERITABILITY ANALYSIS 74

4.2 EXOME ARRAY GENE-BASED ANALYSIS 75

5. DISCUSSION 78

5.1 TWO-STAGE GWAS: COMMON VARIATION AT 2P13.3, 3Q29, 7P13 AND 17Q25.1 ARE ASSOCIATED WITH SUSCEPTIBILITY TO PDAC 78

5.2 EXOME ARRAY GENE-BASED ANALYSIS 81

6. CONCLUSION 83

6. CONCLUSIONES 85

7. BIBLIOGRAPHY 87

8. ANNEXES 98

LIST OF FIGURES

PART 1 INTRODUCTION

CHAPTER 2
FIGURE 1. GENETIC LANDSCAPE OF PANCREATIC CANCER **31**

CHAPTER 4
FIGURE 1. GENOTYPE IMPUTATION SCHEME **41**

PART 3 METHODS

FIGURE 3.1.1 OVERVIEW OF STAGE 1 AND STAGE 2 OF THE TWO-STAGE GWAS **46**

FIGURE 3.1.2 PANC4 PRINCIPAL COMPONENTS ANALYSIS **53**

FIGURE 3.2.1 EXOME ARRAY ANALYSIS PIPELINE **58**

PART 4 RESULTS

FIGURE 4.1.1. PANC4 GWAS MANHATTAN PLOT **66**

FIGURE 4.1.2. LOCUS 17Q25.1 FROM PANC4 GWAS **67**

FIGURE 4.1.3 COMBINED STAGE 1 MANHATTAN PLOT **70**

FIGURE 4.1.4. LOCUS 3Q29 FROM COMBINED 1 GWAS **71**

FIGURE 4.1.5. LOCUS 2P13 FROM COMBINED STAGE 1 AND 2 **72**

FIGURE 4.1.6. LOCUS 7P13 FROM COMBINED STAGE 1 AND 2 **73**

FIGURE 4.2.1 EXOME-ARRAY ANALYSIS MANHATTAN PLOT **77**

LIST OF TABLES

PART 1 INTRODUCTION

CHAPTER 2
TABLE 1. HEREDITARY CANCER SYNDROMES ASSOCIATED WITH PDAC 11

TABLE 2. GWASS AND META-ANALYSIS DISCOVERS IN PDAC 25

PART 3 METHODS

TABLE 3.1.1 CHARACTERISTICS STUDIES SAMPLE 48

TABLE 3.1.2 PANC4 QUALITY CONTROL ANALYSIS TWO-STAGE GWAS 50

TABLE 3.1.3 IN PANSCAN 1 AND PANSCAN 2 QUALITY CONTROL ANALYSES 51

TABLE 3.1.4 PANDORA DATASET 54

TABLE 3.2.1 PANC4 QUALITY CONTROL ANALYSIS EXOME ARRAY STUDY 60

PART 4 RESULTS

TABLE 4.1.1. SIGNIFICANT AND HIGHLY SUGGESTIVE (P<1X10−6)
ASSOCIATION RESULTS FOR TWO-STAGE GWAS 68

TABLE 4.2.1. PANC4 SAMPLES CHARACTERISTICS 76

TABLE 4.2.2. SUGGESTIVE GENES FROM EXOME-ARRAY STUDY 76

LIST OF ABBREVIATIONS:

1000G	**1000GENOMES**
BMI	**BODY MASS INDEX**
CIDR	**CENTER FOR INHERITED DISEASE RESEARCH**
CI	**CONFIDENCE INTERVAL**
CMAF	**CUMULATIVE MINOR ALLELE FREQUENCY**
EPACT	**EFFICIENT AND PARALLELIZABLE ASSOCIATION CONTAINER TOOLBOX**
EGFR	**EPIDERMAL GROWTH FACTOR RECEPTOR**
EMT	**EPITHELIAL-TO-MESENCHYMAL TRANSITION**
FAMMM	**FAMILIAL ATYPICAL MULTIPLE MOLE AND MELANOMA**
FPC	**FAMILIAL PANCREATIC CANCER**
FDR	**FIRST-DEGREE RELATIVE**
GC	**GUANINE-CYTOSINE**
GWAS	**GENOME-WIDE ASSOCIATION STUDY**
HWE	**HARDY WEINBERG EQUILIBRIUM**
HBOC	**HEREDITARY BREAST AND OVARIAN CANCER**
HNPCC	**HEREDITARY NONPOLYPOSIS COLORECTAL CANCER**
HP	**HEREDITARY PANCREATITIS**
IBD	**IDENTICAL BY DESCENT**
IRB	**INSTITUTIONAL REVIEW BOARD**
IARC	**INTERNATIONAL AGENCY FOR RESEARCH ON CANCER**
LD	**LINKAGE DISEQUILIBRIUM**
LNCRNA	**LONG NON-CODING RNA**
LS	**LYNCH SYNDROME**
MODY	**MATURITY ONSET DIABETES OF THE YOUNG**

MAC	**MINIMUM MINOR ALLELE COUNT**
MAF	**MINOR ALLELE FREQUENCY**
MMR	**MISMATCH REPAIR**
MM	**MULTIPLE MYELOMA**
OR	**ODDS RATIO**
P	**P-VALUE**
PDAC	**PANCREATIC ADENOCARCINOMA**
PANC4	**PANCREATIC CANCER CASE-CONTROL CONSORTIUM**
PANSCAN 1	**PANCREATIC CANCER COHORT CONSORTIUM 1**
PANSCAN 2	**PANCREATIC CANCER COHORT CONSORTIUM 2**
PANSCAN 3	**PANCREATIC CANCER COHORT CONSORTIUM 3**
PANDORA	**PANCREATIC DISEASE RESEARCH**
PJS	**PEUTZ-JEGHER SYNDROME**
PCA	**PRINCIPAL COMPONENTS ANALYSIS**
RR	**RELATIVE RISK**
SKAT	**SEQUENCE KERNEL ASSOCIATION TEST**
SKAT-O	**SKAT OPTIMAL**
UCSF	**UNIVERSITY OF CALIFORNIA SAN FRANCISCO**
VC	**VARIANCE COMPONENT**
WGS	**WHOLE GENOME SEQUENCING**
WHO	**WORLD HEALTH ORGANIZATION**

1. INTRODUCTION

Chapter 1: Epidemiology of Pancreatic Cancer

1.1 Incidence, Mortality Trends, Survival Prognosis

In 2012, worldwide, there were approximately 338,000 individuals diagnosed with pancreatic cancer and approximately 331,000 individuals died from their disease making pancreatic cancer the seventh most common cause of cancer death [1]. Pancreatic cancer is strongly associated with increased age, with the majority of cases occurring after age 60. In the United States, from 2009–2013 the incidence of pancreatic cancer in Whites increased from less than 5 per 100,000 before age 45, to 30.0 per 100,000 in individuals aged 60–64, and 93.7 per 100,000 in individuals aged 80–84 [2]. Incidence is approximately equal in men and women. The disease burden is strongest in developed countries compared to developing countries [1]. This difference is likely driven, in large part, by differences in the age structure as well as access to medical care necessary for the diagnosis of pancreatic cancer [3].

In developed countries the overall incidence of pancreatic cancer is expected to continue to increase with the general aging of the population, particularly in high-income countries [4, 5]. Pancreatic cancer is projected to become the second leading cause of cancer death in the United States by 2030 [6]. However, other countries have seen a recent decrease in the incidence of pancreatic cancer that seems to reflect patterns in cigarette consumption.

As discussed later, cigarette smoking is a major risk factor for pancreatic cancer and never smoking or smoking cessation is strongly associated with a decrease in risk. In contrast, increased body mass index (BMI) and diabetes mellitus are both associated with a greater risk of pancreatic cancer and the increasing prevalence of these risk factors is projected to lead to a rise in incidence of pancreatic cancer. Pancreatic cancer is associated with an extremely poor prognosis with an estimated average 1-year relative survival rate of ~20%, and a 5-year rate of ~8% [4]. Survival rates have increased only slightly since the mid-1970s from 4–5% to around 8% in

the United States [2]. The low survival rates are mainly due to advanced stage at diagnosis with only ~20% of patients presenting with local disease [2]. Among patients who undergo surgical resection, the 5-year survival rate is ~15–25% [7]. Outcomes after surgical resection of the pancreas are highly dependent on the experience of the surgeon and the hospital; mortality rates are 70% lower among high-volume surgeons compared with low-volume surgeons, and hospitals with a high patient volume compared with low-volume hospitals [8].

1.2 Cigarette Smoking

Of modifiable risk factors, the relationship between active cigarette smoking and pancreatic cancer risk is well established. Approximately, 20% of all pancreatic cancers are attributable to cigarette smoking [9-11].

Numerous studies have explored the relationship between smoking and pancreatic cancer. A meta-analysis of 82 epidemiologic studies published between 1950 and 2007 [9, 11] reported a 1.74-fold (95% CI: 1.61–1.87) increased risk of pancreatic cancer among current smokers and a 1.2-fold (95% CI: 1.11–1.29) increased risk of pancreatic cancer among former smokers when compared with never smokers. Pooled analysis of individual-level data from the nested case-control studies within the Cohort Consortium (PanScan) [11] as well as analysis of data from 12 case-control studies in the Pancreatic Cancer Case-Control Consortium (PanC4) [10] showed that smokers have a 75–120% increased risk of pancreatic cancer compared with never smokers, and the risk persists for 10–20 years after smoking cessation [10, 11]. Risk also increased according to the number of cigarettes consumed per day; smokers of more than 35 cigarettes per day have a threefold (95%CI: 2.2–4.1) increased risk of pancreatic cancer compared with never smokers [10]. Quitting smoking is associated with a reduced pancreatic cancer risk with a decreased odds ratio in former smokers when compared with active smokers. Studies suggest that the risk in former smokers returns to that of never smokers 15–20 years after smoking cessation [10, 11].

1.3 Diabetes

The relationship between diabetes and pancreatic cancer is quite complex; many newly diagnosed pancreatic cancer patients report a recent onset of diabetes, and those with long-standing diabetes report a recent worsening of diabetes. Thus, it is generally considered that while longstanding diabetes is a risk factor for pancreatic cancer, diabetes can also result as a consequence of pancreatic cancer. There is considerable variability when estimating the prevalence of diabetes and/or glucose intolerance among newly diagnosed pancreatic cancer patients [12]. It has been estimated that up to 80% of newly diagnosed pancreatic cancer patients have glucose intolerance or diagnosed diabetes [13]. Studies that rely on patient or medical records of reported diabetes show lower prevalence estimates, including a large Mayo Clinic case-control study where 40% of patients reported diabetes [14]. Over 75% of pancreatic cancer patients who develop diabetes, do so within the 2 years preceding their pancreatic cancer diagnosis [15]. Thus, there is considerable interest in examining populations of newly diagnosed diabetics to determine whether this might enable earlier detection of pancreatic cancer. It has been shown that up to 1% of newly diagnosed diabetics develop pancreatic cancer within 3 years of their diabetes diagnosis [16]. While many pancreatic cancer patients develop diabetes as a consequence of their disease, there is considerable support from numerous population-based studies that long-standing diabetes (>3 yr) is associated with a modest increase in the risk of pancreatic cancer. Overall, the risk of pancreatic cancer in long-standing diabetes ranges from 1.5- to 2.4-fold [17-20]. However, as the duration of diabetes increases the association between diabetes and pancreatic cancer weakens, with some studies showing only modest or no increase in pancreatic cancer risk 15–20 years after diagnosis with diabetes [20, 21]; however, some studies still support an association with diabetes of 20 years or more [19]. In patients with new-onset diabetes who undergo surgical resection, diabetes often resolves after removal of the pancreatic cancer. In contrast, diabetes does not resolve in patients with long-standing diabetes after surgical removal of their cancer [13, 22].

1.4 Body Mass Index

In addition to diabetes, increased weight or BMI has consistently been associated with increased risk of pancreatic cancer. The World Health Organization (WHO) defines overweight individuals as those with a BMI of 25.0–29.9 kg/m2 and obese individuals as those with a BMI >30.0 kg/m2. Over the past 15 years many studies have demonstrated an increased risk of pancreatic cancer among obese individuals. In 2001, Michaud et al. reported a relative risk of pancreatic cancer of 1.72 (95% CI: 1.19–2.4) for individuals with a BMI >30 kg/m2 compared with individuals with a BMI <23 kg/m2 after controlling for the effect of age, smoking, and diabetes among participants of the Health Professionals Follow-Up Study and the Nurses' Health Study. Many subsequent studies have confirmed this finding; a pooled analysis of data from 13 prospective cohort studies reported an OR for pancreatic cancer of 1.33 (95% CI: 1.12–1.58) when comparing individuals in the lowest quartile of BMI with those in the highest quartile after controlling for the effects of age and smoking. Adjusting for the effect of diabetes attenuates this association slightly (OR = 1.21, 95% CI: 1.01–1.44) [23].

1.5 Alcohol

Numerous studies have examined the association between alcohol consumption and risk of pancreatic cancer. The results of these studies have been inconsistent, with some studies showing an association and others showing no relationship. One challenge to these studies is the strong relationship between smoking and heavy alcohol use, making it difficult to assess the independent association between alcohol use and pancreatic cancer risk. However, several recent large-scale studies that have pooled data across several studies, either using data from prospective cohort studies or retrospective case-control studies, have demonstrated that high levels of alcohol intake are associated with an increased risk of pancreatic cancer. These studies consistently report a ~20–45% increased risk of pancreatic cancer among heavy drinkers (defined as three drinks/day or ≥30 grams/ day of alcohol), compared with non- or occasional

drinkers [24-26]. In addition, in a pooled analysis of data from the Pancreatic Cancer Case-Control Consortium [27], the risk increases up to 60% among extremely heavy alcohol drinkers (≥9 drinks /day). Heavy alcohol consumption is associated with pancreatitis, an established risk factor for pancreatic cancer. Furthermore, acetaldehyde is an established carcinogen. Thus the association between alcohol and pancreatic cancer risk could be either via alcohol-induced pancreatitis or as a direct effect of acetaldehyde.

1.6 Pancreatitis

The relationship between pancreatitis and pancreatic cancer has been well established. Individuals with hereditary pancreatitis, a rare inherited condition, have a remarkably high lifetime risk of pancreatic cancer of 40% [28]. The risk is further increased by cigarette smoking [29]. Quantifying the association between pancreatitis and pancreatic cancer is challenging given the difficulties in diagnosis and differentiation between chronic and acute pancreatitis [30]. In addition, like diabetes, pancreatitis is both a risk factor and a manifestation of pancreatic cancer. The inflammation and damage of long-standing pancreatitis can lead to the development of pancreatic cancer. However, individuals with pancreatic cancer also experience pancreatitis as a consequence of their cancer. A recent large-scale study of 5,048 cases of ductal pancreatic adenocarcinoma and 10,947 controls from 10 case-control studies within the Pancreatic Cancer Case-Control Consortium examined the association between pancreatic cancer and pancreatitis. Overall, 6% of pancreatic cancer patients reported a history of pancreatitis compared to 1% of control individuals. The association between a recent diagnosis of pancreatitis (<1 yr) and pancreatic cancer was remarkably high (OR = 21.35, 95% CI: 12.03–37.86) [31]. In contrast, the association between a pancreatitis diagnosis of >2 years and pancreatic cancer was estimated to be (OR = 2.71, 95% CI: 1.96–3.74) [31]. The association between pancreatitis and pancreatic cancer persisted after controlling for other risk factors including smoking, alcohol consumption, BMI, and diabetes. Interestingly, there was evidence of

effect modification by age, with a stronger association between pancreatitis and pancreatic cancer in patients diagnosed before the age of 65 [31].

1.7 Dietary Factors

Given the generally late age of onset of pancreatic cancer and the complexity of lifetime dietary factors, identification of dietary factors that are consistently associated with pancreatic cancer risk has been remarkably challenging. Several studies have suggested a diet rich in fruit and vegetables may protect against pancreatic cancer with risk reduction in the order of 30–40%, when comparing the highest intake to the lowest intake of fruits and vegetables [32-34]. While a diet rich in fruit and vegetables may protect against pancreatic cancer, several studies have demonstrated an increased risk of pancreatic cancer among individuals who are frequent consumers of smoked or processed meats [35]. A meta-analysis including 6,643 pancreatic cancer cases from 11 prospective studies, reported that eating at least one serving of processed meat a day was associated with a 19% increased risk of pancreatic cancer [35].

1.8 Gastrointestinal Microbiome

In recent years, the importance of the microbiome in human health and disease has gained recognition. Several studies have shown that periodontal disease and tooth loss is associated with pancreatic cancer risk [36].

In 2007, a study among males participating in the Health Professionals Follow−up Study reported that individuals with a history of periodontal disease had a HR of pancreatic cancer of 1.54 (95% CI: 1.16–2.04) compared with those without such a history [37]. A recent study examined the association between specific oral pathogens and pancreatic cancer risk using prospective samples collected within the PLCO trial. This study found that individuals circulating antibodies to Porphyromonas gingivalis and Aggregatibacter actinomycetemcomitans had higher odds of pancreatic cancer (OR = 1.60, 95% CI: 1.15–2.20, and OR = 2.20, 95% CI: 1.16–4.18, respectively),

compared with noncarriers [38]. While some studies have shown an association between pancreatic cancer risk and Helicobacter pylori infection not all studies have shown a positive association. One possible explanation for these inconsistent results is that the relationship may vary between CagA-positive and CagA-negative infections; CagA-negative infection is positively associated with disease and CagA-positive infection potentially has a protective effect. A recent meta-analysis found an overall association of OR = 1.13, 95 % CI: 0.86–1.50 for H. pylori infection and pancreatic cancer risk. The association was OR = 0.78, 95% CI: 0.67– 0.91, and OR = 1.30, 95 % CI: 1.02–1.65 for CagA-positive and CagA-negative strains, respectively [39].

1.9 Allergy

Individuals with a history of allergies, including hay fever, allergic rhinitis, atopic dermatitis, and atopic asthma may have a lower risk of developing pancreatic cancer. A meta-analysis published in 2005 reported an overall association between allergies and pancreatic cancer risk of RR = 0.82, 95% CI: 0.68–0.99). A stronger protective effect was reported in atopic allergies (RR = 0.71, 95% CI: 0.64–0.80) and no association was reported for asthma or food allergies [40]. A recent pooled analysis of data from the Pancreatic Cancer Case-Control Consortium reported a protective effect of hay fever and animal allergies (OR = 0.74, 95% CI: 0.56, 0.96, and OR = 0.62, 95% CI: 0.41, 0.94, respectively), and no association with asthma [41]. In contrast, a recent case-control study from Spain reported a protective effect of both allergy and asthma (OR = 0.66, 95% CI: 0.52–0.83, and OR = 0.64, 95% CI: 0.47–0.88, respectively) [42].

1.10 Family History

One of the strongest risk factors for pancreatic cancer is having a family member with pancreatic cancer. The clustering of pancreatic cancer in families was first reported in the 1970s. Large-scale observational studies have consistently estimated an increased risk of pancreatic cancer among those with a family history of pancreatic pancreatic cancer [43-51]. A recent pooled analysis of data from one case-control and six cohort studies estimated the odds of pancreatic cancer to be 1.76 higher (95% CI: 1.19–2.61) among individuals who had at least one first-degree relative with pancreatic cancer compared with those with a family history of pancreatic cancer [51]. Risk is even higher in familial pancreatic cancer kindreds (defined as a having at least one pair of first-degree relatives with pancreatic cancer) with a 6.79-fold-increased risk of pancreatic cancer among first-degree relatives.

Mutations in the following genes have been associated with a markedly increased risk of pancreatic cancer: *BRCA2*, *BRCA1*, *PALB2*, *ATM*, *CDKN2A*, *STK11*, *PRSS1*, *MSH2*, *MLH1*, *MHS6*, and *PMS2* [52-58].

Chapter 2. Genetic Landscape of Pancreatic Cancer

2.1 Hereditary pancreatic cancer

Approximately 5-10% of pancreatic cancer (PDAC) cases cluster within families [59].

Familial Pancreatic Cancer (FPC) syndrome is defined as the presence of at least two first-degree relatives (FDR) affected with PDAC [59, 60].

Members of these families have up to a 32-fold increased risk of developing the disease, depending upon the number of family members affected [61].

Genetic analyses performed on FPC cases show a high percentage of high-penetrance germline mutations in DNA repair genes; the most frequent genes described in FPC cases are *ATM*, *BRCA2*, *CDKN2A*, and *PALB2* [62].

A smaller portion of PDAC cases reports a family history of cancer that is consistent with specific hereditary cancer syndromes [59].

One of these syndromes is called Hereditary Breast and Ovarian Cancer (HBOC). It is characterized by multiple cases of breast and ovarian cancer in the same family. It has been estimated that members of these families carrying a germline mutation in *BRCA2* gene have a 4-6 fold increased risk of developing PDAC compared to the general population [63, 64]. The risk of PDAC for *BRCA1* germline mutation carriers was highly variable depending on the study; it was estimated from null to ~ four-fold higher than the general population [63, 64].

In HBOC families BRCA negative, it was observed a mutation prevalence of 1.5% in *ATM* and 1.2% for *PALB2* genes, showing that these genes are associated with an increased risk of being affected with breast cancer other than pancreatic cancer [65].

The risk of developing PDAC is particularly high in Peutz-Jegher Syndrome (PJS), a rare (1/200 000 to 1/50 000 births) hereditary condition caused by high penetrant germline mutations in *STK11* gene [66].

Mutation carriers have a specific clinical phenotype as mucocutaneous pigmentation, gastrointestinal hamartomatous polyposis, and multi-systemic oncogenic predisposition. It has been estimated that *STK11* carriers have a 132-fold higher risk of developing PDAC compared to the general population [67]; furthermore these individuals are diagnosed very young, ~40 years old, compared to the average age at diagnosis of PDAC which is ~71 years old [59, 67].

Familial atypical multiple mole and melanoma (FAMMM) syndrome is a hereditary condition characterized by atypical nevi and multiple melanomas clustering within the same family side [68].

Germline mutations in *CDKN2A* gene explain ~40% of FAMMM cases [68]. Pancreatic cancer is the second most common cancer site after melanoma in FAMMM kindreds, being present in ~ 25% of the cases [69]. The literature shows that *CDKN2A* carriers in FAMMM kindreds have a 13-22 fold increased the risk of PDAC compared to the general population [70].

Hereditary pancreatitis (HP) is a rare genetic disorder characterized by the familial aggregation of acute recurrent or chronic pancreatitis. Germline mutations in serine protease 1 (*PRSS1*) gene explain up to 80% of HP cases [71, 72]. For these cases, the disease segregates in the family according to an autosomal dominant model of inheritance [71, 72]. Others genes have been associated with HP, serine protease inhibitor Kazal type 1 (*SPINK1*), chymotrypsin C (*CTRC*) and carboxypeptidase A1 (*CPA1*) [73-75].

Deleterious variants in SPINK1 have a low penetrance for pancreatitis, with homozygous carriers of deleterious variants having a significantly higher risk.

Recent studies highlighted that HP is a more complex disease, with multiple genetic variants and environmental factors involved in its pathogenesis [76, 77].

Member of families affected with HP have a remarkably increased risk of developing PDAC; several studies from different populations reported a 50-90 fold higher risk of PDAC compared to the general population. For members of families with HP syndrome, the risk of PDAC increases markedly after 50 years and is higher in smokers [28, 71, 78].

Lynch syndrome (LS), also known as hereditary nonpolyposis colorectal cancer (HNPCC), is characterized by the clustering of colorectal and endometrial cancers in the same family. However members of these kindreds are also susceptible to develop cancer in other sites. LS has been mainly associated with germline mutations in mismatch repair (MMR) genes as *MLH1*, *MSH2, MSH6*, *PMS2* and *EPCAM [79]*. A recent study estimated that members of families with LS syndrome have a 8.6-fold increase of being affected with PDAC compared to the general population [57].

Table 1 summarizes the hereditary PDAC cases concerning genetic cancer syndrome, responsible genes, and risk of PDAC.

Table 1. Hereditary cancer syndromes associated with PDAC

Genetic Syndrome	Gene mutated	Risk of PDAC estimated respect to the general population (fold increase)
FPC	*BRCA2, CDKN2A, PALB2, ATM*	2.3 one FDR 6.4 two FDR 32 three FDR
HBOC	*BRCA1* *BRCA2*	0-4 4-6
FAMMM	*CDKN2A*	13-22
PJS	*STK11*	132
HP	*PRSS1, SPINK1, CFTR, CPA1, CTRC*	50-90
LS	*MLH1, MSH2, MSH6, PMS2, EPCAM.*	8.6

Abbreviations: FPC: familial pancreatic cancer, FDR; first-degree relative, HBOC: hereditary breast and Ovarian cancer; FAMMM: Familial Atypical Multiple Mole Melanoma; PJS: Peutz Jeghers Syndrome; HP: Hereditary Pancreatitis; LS: Lynch Syndrome.

2.2 Common variants associated with pancreatic cancer

Common variants, defined here by convention as minor allele frequency (MAF) >1%, have been widely Investigated for association with complex diseases through genome-wide association studies (GWASs).

GWAS tests the hypothesis that common diseases are caused by common variants [80]; it applies an agnostic approach by looking for an association between genetic variants and disease all over the genome with no assumptions on biological or positional candidate loci, genes, and variants.

In the majority of GWASs, significant associations with the disease are found not directly with the functional variant, rather with other genotyped or imputed variants that are highly correlated with the causal variants. The genetic correlation between two variants, expressed as linkage disequilibrium (LD), is often measured as a squared correlation (r^2) and it strongly depends on allele frequency [81]. LD r^2 can be large only if the allele frequencies at the two genetic variants match [81]; it has been estimated that for $r^2 \geq .8$ and a locus with allele frequency .5, the other locus has allele frequency $0.5 \pm .06$ [81]. LD r^2 ranges from 0 (no correlation) to 1 (complete correlation).

To date, a large number of high-throughput SNPs arrays (200,000-2,000,000 SNPs) have been designed to be representative of LD genome landscape, and therefore they have been built to tag variants in the genome.

Other than LD and allele frequency, other factors that affect the statistical power of a GWAS to detect associations between disease and genetic variants are sample size and the proportion of phenotypic variance or effect size explained by a causal variant in the population. The smaller the size of the effect of the variant on the disease the larger is the number of samples needed to identify it.

GWASs has proved very useful; to date >10,000 genetic risk factors have been discovered for both complex traits and diseases [82], increasing the knowledge of their biology.

Because PDAC is a rare disease, with a worldwide age-standardized incidence rate of 4.2 affected per 100,000 individuals (GLOBOCAN 2012), the collection of a sample size appropriate to conduct GWAS required the creation of international data consortium.

2.2.1 GWAS Panscan 1, 9q24, ABO gene

The first GWAS on PDAC in Caucasian population was published in 2009 and was performed using the Pancreatic Cancer Cohort Consortium 1 (PanScan 1) dataset that included 1,896 cases and 1,939 controls, recruited from 12 prospective and 1 case-control studies [83].

For each SNP, cases and controls were compared for the count of minor alleles by fitting a logistic regression model, adjusted by study, age, sex and the top principal components of the population stratification analysis. Three genomic regions (9q34, 7q36 and 15q14) were selected because they were genome-wide significant or/and suggestive. These regions were then tested for replication in an independent population [83].

Combined analysis of discovery and replication datasets confirmed only the locus on chromosome 9 (9q34, $P = 5.37 \times 10^{8}$). The most significant SNP in this region, rs505922 (T/C), has been localized in the first intron of *the ABO* gene· [83].

ABO encodes a protein that catalyzes the transfer of carbohydrates to the H antigen, forming the antigenic structure of the ABO blood groups. Individuals with A, B, and AB alleles express glycosyltransferase activities that convert the H antigen into the A or B antigen, whereas individuals with O allele do not express any antigen because of a single nucleotide deletion that inactivates the glycosyltransferase activity [84].

Amundadottir L et al. observed that the major allele (T) of rs505922 was in complete linkage disequilibrium ($r^2=1$) with the allele O of *ABO* gene, meaning that individuals homozygous for allele T had O blood type, whereas individuals heterozygotes or homozygotes for the minor allele C had non-O blood type [83].

It has been demonstrated that the risk of PDAC for non-O blood type individuals, both C/T and C/C, was respectively 20% and 44% higher compared to individuals with O blood type (T/T) [83].

Another almost parallel study confirmed these findings; more than 100,000 individuals from two large independent cohorts, were followed-up for incident pancreatic cancer cases. In 9 years, 316 individuals from the cohort were diagnosed with PDAC [85]. This study highlighted that, although 45% of the total sample had O blood type, PDAC cases reported more frequently non-O blood group (A, B or AB). Specifically, the risk of PDAC for people whose blood group was A, AB and B were 1.32, 1.51, and 1.72 higher than people with O blood group, respectively [85].

Of note, PDAC is not first cancer to be found associated with *ABO* gene; actually, it was previously reported an increased risk of gastric cancer in individuals non-O blood type when compared to individuals with O blood type [86, 87].

These observations raised many hypotheses; one of these supposed that the association between ABO blood group and pancreatic cancer was the result of the infection by the bacterium *H pylori*. This hypothesis was supported by the fact that H *pylori* infection had been found associated with increased risk of both pancreatic and gastric cancers [88-90].

Another study showed that individuals tested positive to H *pylori* infection and belonging to non-O blood type had a statistically significant ~ 3-fold increased risk of PDAC compared to the individual tested negative for H *pylori* infection and whose blood type was O [91].

These observations led researchers to look for biological causes that justified the observed link between PDAC, *ABO* and H. *pylori*.

H. *pylori* colonize the gastric epithelium and activate an inflammatory response that causes an excess of acidity at the gastric level. In this context, the pancreas is enabled through the action of gastrointestinal hormones, and it responds by producing bicarbonate to neutralize the acidity. Because of the persistent state of gastric inflammation, the pancreas is chronically stimulated to produce bicarbonate, which causes hyperplasia accompanied by increase epithelial cell

activity, DNA synthesis and cell turnover; in this state, the pancreas becomes more susceptible to the action of carcinogens [92].

According to this theory, *the ABO* gene is important because both A and B antigens localized in the gastric epithelium have an essential role in the colonization process of H.*pylori* [92].

2.2.2 GWAS PanScan 2, 13q22.1 (*KLF5* and *KLF12*), 1q32.1 (*NR5A2*), 5p15.33 (*TERT-CLPTM1L*)

After one year the same group published a second GWAS realized by adding 1,955 cases and 1,995 controls to the PanScan 1 dataset used in the first GWAS. This additional dataset, named PanScan 2 collected samples from 8 case-control studies, and it was genotyped using approximately 620,000 SNPs [93].

The final sample, obtained by pooling the two datasets, included 3,851 cases and 3,934 controls genotyped with over 550,000 common markers [93].

The association was tested using a logistic regression model measuring a genotype trend effect on PDAC risk. The model was adjusted for age, sex, study and top principal components.

Three new genomic regions (13q22.1, 1q32.1 and 5p15.33) showed genome-wide significance P-values [93].

The locus on chromosome 13 (13q22.1), does not include any gene, however, is delimited by two genes, *KLF5* and *KLF12*, members of the Kruppel-like factors family. These genes are reported to have essential functions in the regulation of cell growth and transformation processes [94].

Because this region has been found associated with other cancer sites other than pancreas [95, 96], it has been hypothesized that it includes a critical gene involved in carcinogenesis.

A recent study performed fine mapping of the 13q22.1 locus, and it included other three neighboring genes, *PIBF1*, *DIS3*, and *BORA*, in addition to *KLF5* and *KLF12* [97].

Fine mapping identified eight additional genetic variants, highly correlated between them and significantly associated with an increased risk of PDAC [97]. These new variants were tested for association with expression levels of the five genes included in the region. All SNPs showed

significant association with *DIS3* expression levels; specifically for all eight variants the allele that determined an increased risk of PDAC was associated with a lower expression of *DIS3* gene [97].

The *DIS3* gene has an important role in both gene regulation and small RNA processing [98]. This gene has been found associated with other cancer sites as colorectal, melanoma, and multiple myeloma (MM); missense mutations in *DIS3* have been found in 11% of patients with MM [99].

The most significant SNP associated with PDAC at 1q32.1 locus is located in the first intron of *NR5A2* gene, a nuclear receptor subfamily 5 group A member 2 [93]. This gene is highly expressed in liver, exocrine pancreas, intestine, and ovary where it has fundamental roles in development, reverse cholesterol transport, bile-acid homeostasis and steroidogenesis processes [100, 101].

NR5A2 gene is essential during the early development of the pancreas and also for the organ adult homeostasis [102]; in fact, it has been proven that *NR5A2* promotes the regeneration of acinar cells after inflammation caused by chemically induced pancreatitis, and protects pancreas from *KRAS* driven pre-neoplastic changes.

These data have been supported from a recent crucial functional study on mice that showed as the loss of one Nr5a2 allele was able to produce a pro-inflammatory state, as demonstrated by the upregulation of inflammatory genes and the presence of chemokines and complement components in the pancreas [103].

The inflammatory state makes pancreas more susceptible to acquire mutations in the oncogene *KRAS;* genetic mutations in *KRAS* are found in precursor pancreatic lesions and are thought to be the first step in the carcinogenesis process [103].

Another study observed reduced expression levels of *NR5A2* gene in pancreatic cancer tissue compared to normal pancreas samples, adding evidence that *NR5A2* has a protective effect on pancreatic cancer [104]

The association on chromosome 5 lies in a small region, chr5p15.33 that includes two genes, *TERT* and *CLPTM1L*. Of note, this locus has been reported associated with other 10 cancer sites,

bladder, breast, lung, melanoma, non-melanoma skin, ovarian, prostate, testicular germ cell, chronic lymphocytic leukemia and glioma [105-112]; it has been observed that the same alleles can have risk-enhancing or protective effects on different cancer sites [108].

The gene *TERT* encodes a ribonucleoprotein polymerase that works in combination with an RNA template (*TERC*) to regulate telomere ends by addition of the telomere repeat TTAGGG at each cell division [113]. The activity of this enzymatic complex is essential for the life of the cell because the telomeres protect the chromosomes by a potential abnormality or other damage during cell division. Telomerase activity is at the higher level in germ cells and during early development, however, in most cells, telomeres become shorter as the number of cell divisions increases, till they reach a critical length that activates a signal for cellular senescence and apoptosis [113].

In the most of cancer cases, the activity of telomerase is upregulated, and cancer cells continue to divide; this pathway is crucial for initiating cancerogenesis and for tumor survival ([114, 115].

Recently, fine mapping of 5p15.33 locus identified a functional variant, rs36115365; increased risk of pancreatic and testicular cancers and decreased risk of lung and melanoma was observed in association with the minor allele of this SNP [116]. Proteomic analysis showed that rs36115365 regulates *TERT* expression by binding to a zinc finger protein (ZNF148); *ZNF148* knockout results in reduced expression of telomerase [116].

The neighboring gene cleft lip and palate associated transmembrane 1 like (*CLPTM1L)* encodes a protein that promotes growth and survival in pancreatic and lung cancer cells, respectively, and is overexpressed in other cancer sites [117].

This gene was found highly expressed in cisplatin-resistant ovarian tumor cell lines and is associated with cisplatin-induced apoptosis [117].

2.2.3 GWAS PanScan 3; 7q32.2 (LINC-PINT); 16q23.1 (*BCAR1);* 13q12.2 (*PDX1)*; 22q12.1(*ZNFR3)*

The third GWAS on PDAC was conducted in a new dataset, PanScan 3, which included new 1,582 cases and 5,203 controls of European ancestry recruited from 13 prospective cohort studies, 2 case series, and 1 case-control [118].

A meta-analysis including both PanScan 3 GWAS data and PanScan 1\2 pooled GWAS data identified 13 new genomic regions associated with PDAC; among them, five were replicated in an independent population [118].

A new independent signal was identified at 5p15.33; actually, in the previous GWAS (Petersen GM et al. 2010) the most significant SNP was located in the intron 13 of *CLPTM1L* gene, and it was associated with an increased risk of PDAC (rs401681, OR 1.19; 95%CI: 1.11-1.27, P= 3.66E-07) [93].

The new signal on 5p15.33 was found in a region of high LD that extended from the promoter region to exon 2 of *TERT* gene. The minor allele of the most significant SNP, located in exon 2 of *TERT*, was associated with a decreased risk of developing PDAC (rs2736098; OR 0.80; 95%CI: 0.76–0.85, P= 9.78 × 10×14) [118].

Another PDAC locus was identified at the long arm of chromosome 7 (7q32.2), intronic *to LINC-PINT,* a long intergenic non-protein coding RNA (lncRNA), p53 induced transcript [118].

Functional studies showed that lncRNAs genes are targets of the p53 tumor suppressor and as such, they are involved in a different process of tumorigenesis [119].

Because lncRNAs are detectable in the plasma, a recent study investigated the relation between plasma linc-pint levels and diagnosis of PDAC. It has been observed that patients with PDAC had lower plasma levels of linc-pint compared to healthy individuals [120].

The same study also compared plasma levels of linc-pint in different PDACs types, as pancreatic ductal adenocarcinomas, pancreatic cystic adenocarcinomas, pancreatic adenocarcinomas mixed with neuroendocrine carcinomas and rare pancreatic cancers; they found that low linc-pint plasma levels were associated with tumor recurrence and predicted poor prognosis. According to these data linc-pint plasma levels seem to be an interesting candidate non-

invasive biomarker for diagnosis and prognosis of PDAC; however further studies are needed to evaluate its specificity and sensitivity.

Increased risk of PDAC was observed in individuals carrying the minor allele of SNP rs7190458 (OR=1.46; 95%CI 1.30–1.65, P=1.13 × 10×10), located on chromosome 16 (16q23.1), in the last exon of Breast Cancer Anti-estrogen Resistance 1 (*BCAR1)* gene [118].

BCAR1 codes a member of the Crk-associated substrate (CAS) family of protein. It has been reported that the phosphorylation of this protein activates cell migration and enhances the invasive potential of carcinoma cells in vivo [121].

Over-expression of *BCAR1* has been observed in several cancer sites other than pancreas [121]. In pancreatic cancer, migration and metastasis related to BCAR1 protein are activated by epidermal growth factor receptor (EGFR) pathway. Once phosphorylated, BCAR1 forms a complex with Nck1 protein that promotes Ras-associated protein-1 (Rap1) signaling [122].

Another signal for PDAC was detected on chromosome 13 (13q12.2) near a SNP located 200bp upstream pancreatic and duodenal homeobox 1 (*PDX1*) gene [118].

Pdx1 is the first transcription factor expressed in the developing pancreas, and its role is of fundamental importance, as reported in a 1997 study regarding an infant born without the pancreas and carrying a homozygous deletion of a single nucleotide in the codon 63 of *PDX1* gene [123].

Other than pancreas genesis, Pdx1 is essential for adult β cells function; actually, it has been proved that the removal of Pdx1 resulted in hyperglycemia and increased secretion of glucagon [124].

The same study showed that adult β cells do not die from Pdx1 loss; instead, their phenotype change, assuming an α-like phenotype, characterized by the expression of glucagon or a loss of β-cell function with no expression of any pancreatic hormone [124].

Increased expression of Pdx1 was found in PDAC, suggesting that this transcription factor may be involved in tumorigenesis (Roy N et al. 2015). A recent study highlighted a double function of

PDX1 in pancreas tumorigenesis, depending on the stage of the process; actually, *PDX1* showed a protective role in the initial phase of oncogenic transformation by contrasting the change from acinar to ductal cells. However, *PDX1* acted as a promoter of cancer invasiveness by promoting epithelial-to-mesenchymal transition (EMT) and the metastatic processes [125].

An increased risk of PDAC was associated with the minor allele of a SNP located in the intron of the gene zinc and ring finger 3 (*ZNFR3*), at 22q12.1 [118].

ZNFR3 encodes an ubiquitin ligase and transmembrane protein [126]. Together with its functional homolog RNF43, these proteins have an essential role in the regulation of Wnt signaling pathway [126].

The normal functioning of these proteins, activated by their ligands, corresponds to down-regulation of the Wnt pathway [126].

In contrast, it has been demonstrated that mutations, which inactivated the *ZNFR3*/*RNF43* genes, corresponded to an increase in Wnt signaling pathway [126].

Of note, *RNF43* was identified as a tumor suppressor in cystic pancreatic, and it represents the first upstream Wnt pathway component mutated in cancers [127].

Wolphin et al. [118] also highlighted a region on 8q24.1, a genetic region frequently found amplified in several cancer sites cells [128].

The most significant SNP approached genome-wide significance level and was located close to plasmacytoma variant translocation 1 (*PVT1*) gene that encodes long non-coding RNA (lncRNA) [129].

PVT1 is located 57kb downstream *MYC oncogene*, and in some study, it has been described acting as an oncogene itself because its stabilizing role of MYC protein expression [130]; *PVT1* intragenic region includes several regulatory elements binding *MYC* [131].

A recent study demonstrates that *PVT1* promoter has an independent function than the rest of the gene; actually in contrast with the oncogene activity just described, *PVT1* promoter behaves

as tumor-suppressor as demonstrated by the fact that if it is silenced, the expression of *MYC* increases due to the loss of PVT1 lncRNA ([131].

A recent study aimed to identify the function of *PVT1* on pancreatic cancer did compare normal cells with PDAC cells and showed that *PVT1* plays an essential role in proliferation and migration processes; cells with *PVT1* inactivated had significantly reduced growth and migration abilities [132].

2.2.4 GWAS4 Panc4; 17q25.1 (*LINC00673*); 7p13 (*SUGCT*); 3q29 (*TP63*); 2p13.3 (*ETAA1*)

This study has been performed on the international Pancreatic Cancer Case-Control Consortium (PanC4), including 4,164 newly genotyped cases and 3,792 controls recruited from 9 studies from North America, Central Europe and Australia.

Since I have participated as the first co-author in the realization of this study, it will be reported in this thesis.

2.2.5 Meta-analysis PanScan 1, 2 and 3 GWASs: new independent loci at 1q32.1, 5p15.33 and 8q24.21

A follow-up analysis was performed using PanScan 1, 2 and 3 data. For all these datasets, SNPs were imputed using 1000 Genomes (1000G) as the reference panel. Association analysis was tested separately for each dataset, and then its results were combined in a fixed-effects meta-analysis including 5,107 pancreatic cancer cases and 8,845 controls. Promising signals were replicated in PANDoRA and PanC4 datasets.

Zhang et al. identified three independent signals in the previously reported loci, 1q32.1, 5p15.33 and 8q24.21 [104].

The new signal at 1q32.1 is located ~11 kb upstream of *NR5A2* gene [104], whose implication in pancreatic cancer has been extensively commented in paragraph 2.2 of this chapter.

The SNP at 8q24.21 is located ~28 kb upstream of *MYC* and ~850 kb upstream of the signal close to *PVT1* gene [104]. *MYC* and *PVT1* interaction have been commented in paragraph 2.3 of this chapter.

The new region on chr5p15.33 is located in *TERT* promoter (~200-500 bp upstream of the TSS) [104]. The role of *TERT* gene on PDAC has been described in chapter 2.2.2.

2.2.6 Meta-analysis PanScan 1, 2, 3 and PanC4 GWASs: 1p36.33, 7p12,8q21.11,17q12, and 18q21.32

Recently, five new regions have been reported in the most extensive study performed so far on pancreatic cancer, a meta-analysis of PanScan 1/2, PanScan 3, and PanC4 GWASs, including overall 9,040 PDAC cases and 12,496 controls of European ancestry [133]. The promising regions from the meta-analysis were then tested for replication in an independent population (PANDoRA) [133].

A new region was identified at the short arm of chromosome 1 (1p36.33); the most significant SNP of this region is located at the first intron of the novel inhibitor of histone acetyltransferase gene (*NOC2L*) [134]. Both *NOC2L* and the cytogenic band that harbor it has been recently described as differentially methylated in a study that compared the DNA methylation pattern between breast tumors and normal tissues [135]. Previous studies showed that *NOC2L* interacts directly with the tumor suppressor protein p53 and inhibits its function [134].

NOCL2 has also been reported as a negative regulator of TAp63 [136], the p53 homolog transcription factor essential for the development and differentiation of epithelial surfaces.

Both p53 and p63 have been described in pancreatic cancer; somatic mutation with loss of heterozygosity at the *p53* locus has been observed in ~ 75-90% of pancreatic carcinomas [137], whereas the minor allele of a common variant located in an intron of the p63 gene, was associated with a protective role of PDAC [138](**Table 2**).

Two related genes, the human hepatocyte nuclear factor 4 gamma (*HNF4G*) and Hepatocyte nuclear factor 1-beta (*HNF1B*) were associated with PDAC in this study [133]; both of them play an essential role in pancreas organogenesis and development [102, 139].

Pathway analysis of pancreatic cancer showed the strongest association (P-value 2.0 ×10-6) for the pathway including genes involved in pancreas development and cell differentiation processes compared to other pathways under study. Among the overall 22 genes included in pancreas development pathway, were both *HNF4G* and *HNF1B*, in addition to other genes before mentioned as *NR5A2*, *HNF1A*, and *PDX1* [139].

Both *HNF4G* and *HNF1B* also have essential functions in the regulation of glucose and fatty acid metabolism, and they have been found associated with early-onset autosomal-dominant type 2 diabetes [140, 141]. Furthermore, *HNF4G* was described as a susceptibility gene for hyperuricemia [142, 143] and body mass index [144].

Another gene found associated with pancreatic cancer in this study [133] is the Gastrin-releasing peptide (*GRP*) gene that is located on 18q21.32. Grp belongs to a family of peptides highly conserved among species and that have a high binding affinity with receptors located in the pancreas and gastrointestinal tract [145]. High expression of Grp receptor has been observed in gastrointestinal inflammatory diseases such as chronic pancreatitis [146]. Besides, Grp has been implicated in glucose homeostasis [147].

Of note, several studies support a strong association between acute pancreatitis and new-onset diabetes; it has been estimated that having acute pancreatitis increase >2 fold the risk of developing diabetes compared to the general population [147, 148].

GRP seems responsible for the hormone imbalance that leads to diabetes following inflammation of the pancreas [149].

A new genetic region for PDAC was identified at 7p12 locus [133]. The most significant SNP is intronic in *TNS3* gene, which encodes the third of the overall four proteins called tensins involved

in cell adhesion and migration processes. *TNS3* inhibits cell motility; its downregulation has been related to cancer cell metastatic behavior in human renal cell carcinoma [150].

Table 2 Summary of the regions associated with PDAC, identified through GWAS and meta-analysis studies.

Chr[a] SNP Position[b] Gene	Effect Allele (Minor)/ Reference Allele	Statistic	PanScan 1/2 3,535 cases 3,642 controls	PanScan 3 1,582 cases 5,203 controls	PANC4 3,933 cases 3,651 Controls	ALL GWAS[c] 9,040 cases 12,496 controls	PANDoRA 2,497 cases 4,611 controls	GWAS + PANDoRa[d] 11,537 cases 17,107 controls
1p36.33 rs13303010 894,573 *NOC2L*	G/A	maf[e] cases;controls	0.14; 0.13	0.12; 0.10	0.13; 0.11		0.14; 0.10	-
		info[f]	0.42	g	g		g	-
		OR (CI)[g]	1.15 (1.01 – 1.26)	1.22(1.09 – 1.33)	1.16(1.07 -1.24)	1.20 (1.12 -1.29)	1.45(1.33-1.57)	1.26 (1.19-1.35)
		p-value	$3.64x10^{-2}$	$1.48x10^{-3}$	$9.54x10^{-4}$	$7.30x10^{-7}$	$6.00x10^{-10}$	$8.36x10^{-14}$
		Heterogeneity p-value[h]				$6.49x10^{-1}$		$4.57x10^{-2}$
1q32.1 rs2816938 199,985,368 *NR5A2*	A/T	maf cases;controls	0.25; 0.22	0.25; 0.23	0.27; 0.23			
		info	1	1	1			
		OR (CI)	1.25 (1.17 - 1.33)	1.19 (1.08 -1.30)	1.19 (1.12 -1.27)	1.21 (1.17-1.26)		
		p-value	$1.81x10^{-8}$	$2.33x10^{-3}$	$2.80x10^{-6}$	$3.36x10^{-15}$		
		Heterogeneity p-value				$6.48x10^{-1}$		
1q32.1 rs3790844 200,007,432 *NR5A2*	G/A	maf cases;controls	0.20; 0.24	0.20; 0.24	0.20; 0.23			
		info	g	g	g			
		OR (CI)	0.77 (0.70 - 0.85)	0.86 (0.74-0.97)	0.83 (0.76 -0.91)	0.81 (0.76 - 0.86)		
		p-value	$2.16x10^{-10}$	$7.62x10^{-3}$	$6.87x10^{-6}$	$7.62x10^{-16}$		
		Heterogeneity p-value				$2.60x10^{-1}$		
2p13.3 rs2035565 67,619,656 *ETAA1*	C/T	maf cases;controls	0.30; 0.28	0.29; 0.28	0.30; 0.28			
		info	1	0.999	0.999			
		OR (CI)	1.10 (1.03 - 1.18)	1.09 (0.98 -1.19)	1.14 (1.07 -1.21)	1.12 (1.07 - 1.16)		
		p-value	$6.92x10^{-3}$	$1.24x10^{-1}$	$2.92x10^{-4}$	$2.56x10^{-6}$		
		Heterogeneity p-value				$7.21x10^{-1}$		

2p13.3 rs1486134 67,639,769 *ETAA1* (2236bp 3′)	G/T	maf cases;controls	0.30; 0.28	0.29; 0.28	0.30; 0.28		0.29; 0.27	
		info	g	g	g		g	
		OR (CI)	1.10 (1.03 - 1.17)	1.11 (1.00 -1.21)	1.14 (1.07 -1.21)	1.12 (1.07 - 1.16)	1.16 (1.06 - 1.27)	1.13 (1.09 - 1.17)
		p-value	$8.04x10^{-3}$	$5.90x10^{-2}$	$1.88x10^{-4}$	$9.80x10^{-7}$	$9.42x10^{-4}$	$4.61x10^{-9}$
		Heterogeneity p-value				$7.53x10^{-1}$	$3.32x10^{-2}$	$1.04x10^{-1}$
3q29 rs9854771 189,508,471 *TP63*	A/G	maf cases;controls	0.34; 0.37	0.34; 0.36	0.33; 0.36		0.34; 0.36	
		info	1	g	1			
		OR (CI)	0.87 (0.81 - 0.93)	0.96 (0.87 -1.06)	0.88 (0.82 -0.94)	0.89 (0.85 - 0.93)	0.93 (0.86 - 1.01)	0.90 (0.86 - 0.94)
		p-value	$7.98x10^{-5}$	$4.00x10^{-1}$	$1.28x10^{-4}$	$1.14x10^{-7}$	$1.01x10^{-1}$	$4.54x10^{-8}$
		Heterogeneity p-value				$2.44x10^{-1}$	$8.15x10^{-1}$	$6.50x10^{-1}$
5p15.33 rs2736098 1,294,086 *TERT*	T/C	maf cases;controls	0.26; 0.28	0.22; 0.27	0.24 0.27		0.22; 0.20	
		info	0.84	g	0.92		g	
		OR (CI)	0.85 (0.77 - 0.93)	0.78 (0.67-0.89)	0.83 (0.75 -0.91)	0.83 (0.78 - 0.88)	0.89 (0.79 - 0.99)	0.84 (0.79 - 0.88)
		p-value	$6.11x10^{-5}$	$1.28x10^{-5}$	$1.95x10^{-6}$	$5.80x10^{-14}$	$1.68x10^{-2}$	$6.86x10^{-15}$
		Heterogeneity p-value				$5.19x10^{-1}$		$4.18x10^{-1}$
5p15.33 rs35226131 1,295,373 *TERT, CLPTM1L*	T/C	maf cases;controls	0.02; 0.03	0.02; 0.03	0.02; 0.03			
		info	0.77	0.84	0.98			
		OR (CI)	0.61 (0.35 - 0.87)	0.66 (0.35 -0.97)	0.71 (0.51 -0.91)	0.67 (0.53 - 0.81)		
		p-value	$2.15x10^{-4}$	$9.05x10^{-3}$	$6.82x10^{-4}$	$2.19x10^{-8}$		
		Heterogeneity p-value				$6.95x10^{-1}$		
5p15.33 rs401681 1,322,087 *CLPTM1L*	T/C	maf cases;controls	0.49; 0.45	0.49; 0.45	0.49; 0.44			
		info	g	0.996	g			
		OR (CI)	1.19 (1.12 - 1.25)	1.20 (1.11 -1.30)	1.19 (1.13 -1.25)	1.19 (1.15 - 1.23)		
		p-value	$3.53x10^{-7}$	$1.27x10^{-4}$	$9.15x10^{-8}$	$9.32x10^{-17}$		
		Heterogeneity p-value				$9.73x10^{-1}$		

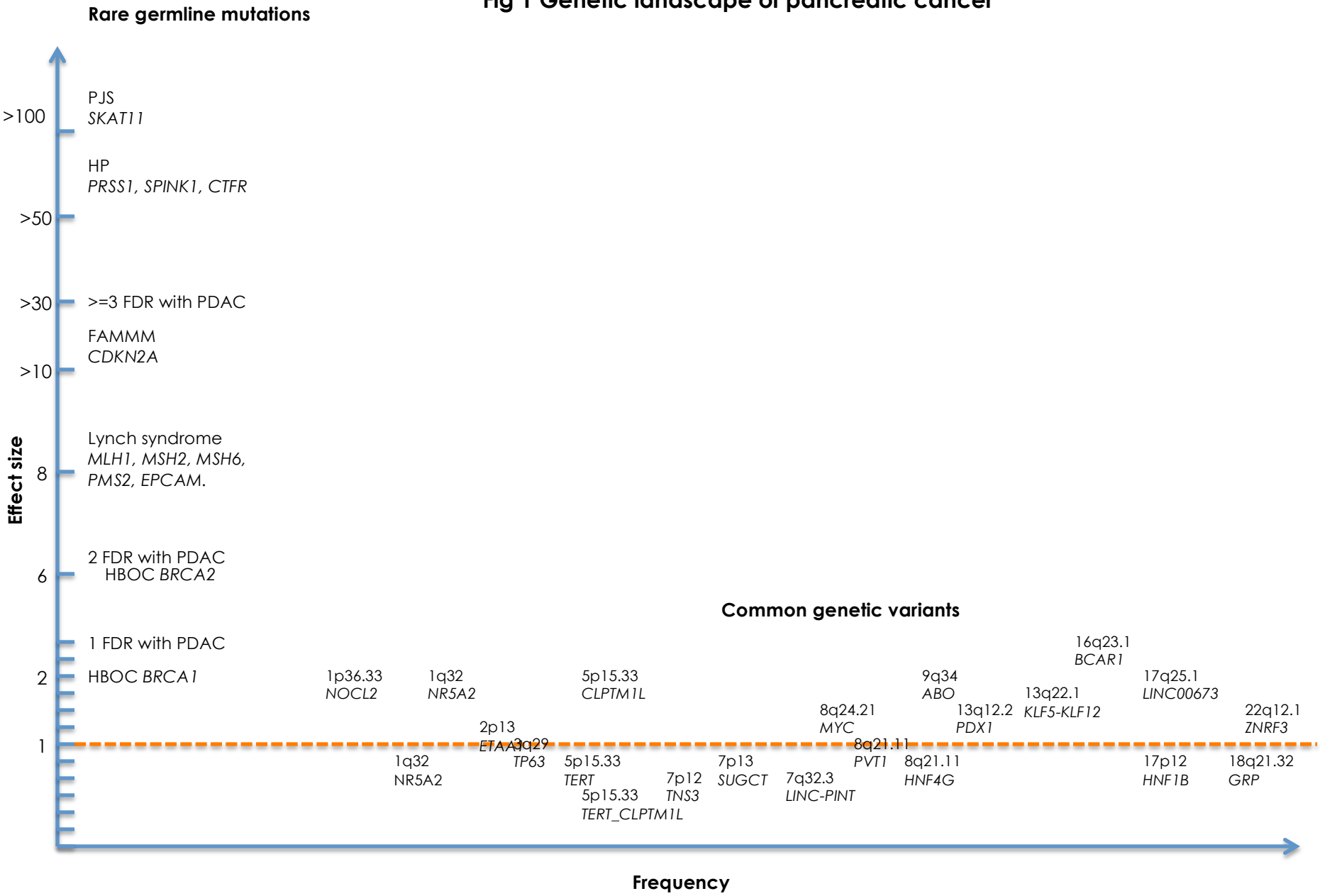
7p12 rs73,328,514 47488569 *TNS3*	T/A	maf cases;controls	0.09; 0.11	0.10; 0.12	0.10; 0.12		0.10;0.11	-
		info	0.93	0.97	0.97		g	-
		OR (CI)	0.80 (0.71-0.89)	0.88 (0.76 – 1.02)	0.82 (0.74-0.92)	0.83 (0.77-0.88)	0.94 (0.83-1.06)	0.85 (0.80 – 0.90)
		p-value	$8.38x10^{-5}$	$9.31x10^{-2}$	$3.61x10^{-4}$	**$4.35x10^{-8}$**	$3.08x10^{-1}$	$1.35x10^{-7}$
		Heterogeneity p-value				$5.98x10^{-1}$		$2.35x10^{-1}$
7p13 rs17688601 40,866,663 *SUGCT*	A/C	maf cases;controls	0.24; 0.27	0.25; 0.27	0.25; 0.27		0.25; 0.28	
		info	g	g	g		g	
		OR (CI)	0.85 (0.78 - 0.92)	0.92 (0.81 - 1.02)	0.88 (0.82 - 0.95)	0.88 (0.83 - 0.93)	0.91 (0.83 - 1)	0.88 (0.84 - 0.93)
		p-value	4.14x10-5	1.14x10-1	1.13x10-3	8.23x10-8	3.93x10-2	$1.11x10^{-8}$
		Heterogeneity p-value				5.63x10-1	7.25x10-2	$1.70x10^{-1}$
7q32.3 rs6971499 130,680,521 *LINC-PINT*	C/T	maf cases;controls	0.12;0.14	0.13; 0.16	0.13; 0.16		0.12; 0.15	
		info	0.95	g	g			
		OR (CI)	0.83 (0.73 - 0.92)	0.79 (0.66 -0.93)	0.82 (0.73 -0.91)	0.82 (0.76 - 0.88)	0.80 (0.67- 0.92)	0.81 (0.76 - 0.87)
		p-value	$1.52x10^{-4}$	$7.12x10^{-4}$	$2.32x10^{-5}$	$4.32x10^{-11}$	$3.82x10^{-4}$	$7.41x10^{-14}$
		Heterogeneity p-value				$8.79x10^{-1}$		$9.43x10^{-1}$
8q21.11 rs2941471 76,470,404 *HNF4G*	G/A	maf cases;controls	0.40; 0.43	0.41; 0.42	0.41; 0.43		0.40; 0.43	
		info	1.0	1.0	1.0		g	
		OR (CI)	0.87 (0.79 – 0.94)	0.91 (0.80-1.01)	0.89 (0.82-0.96)	0.89 (0.86-0.94)	0.86 (0.77-0.95)	0.89 (0.86-0.93)
		p-value	$2.39x10^{-4}$	$8.30x10^{-2}$	$2.19x10^{-3}$	$4.73x10^{-7}$	$2.42x10^{-3}$	$4.52x10^{-9}$
		Heterogeneity p-value				$7.73x10^{-1}$		$8.75x10^{-1}$
8q24.21 rs10094872 128,719,884 *MYC*	T/A	maf cases;controls	0.40; 0.36	0.39; 0.36	0.38; 0.36			
		info	0.94	0.96	0.97			
		OR (CI)	1.17 (1.10 - 1.24)	1.18 (1.08 -1.28)	1.11 (1.04 -1.18)	1.14 (1.10 - 1.19)		
		p-value	$1.28x10^{-5}$	$9.83x10^{-4}$	$3.25x10^{-3}$	$1.19x10^{-9}$		
		Heterogeneity p-value				$4.55x10^{-1}$		

8q24.21 rs1561927 129,568,078 *MIR1208*	C/T	maf cases;controls	0.25; 0.28	0.25; 0.27	0.24; 0.26		0.26; 0.28	
		info	g	g	g		g	
		OR (CI)	0.86 (0.79 - 0.94)	0.87 (0.76 -0.98)	0.92 (0.84 -0.99)	0.89 (0.84 - 0.93)	0.91 (0.81 - 0.99)	0.89 (0.85 - 0.93)
		p-value	$1.06x10^{-4}$	$1.06x10^{-2}$	$2.74x10^{-2}$	$6.18x10^{-7}$	$3.69x10^{-2}$	$7.09x10^{-8}$
		Heterogeneity p-value				$4.54x10^{-1}$		$6.25x10^{-1}$
9q34 rs505922 136,149,229 *ABO*	C/T	maf cases;controls	0.39; 0.35	0.41; 0.35	0.40; 0.35			
		info	1	1	g			
		OR (CI)	1.21 (1.14 - 1.28)	1.37 (1.27 -1.48)	1.28 (1.21 -1.34)	1.27 (1.22 - 1.31)		
		p-value	$4.78x10^{-8}$	$4.56x10^{-10}$	$1.00x10^{-12}$	$7.35x10^{-27}$		
		Heterogeneity p-value				$1.13x10^{-1}$		
13q12.2 rs9581943 28,493,997 *PDX1-AS1 - PDX1*	A/G	maf cases;controls	0.43;0.41	0.44; 0.40	0.43; 0.39		0.44; 0.41	
		info	1	g	g			
		OR (CI)	1.12 (1.06 - 1.19)	1.22 (1.13 -1.32)	1.17 (1.11 -1.24)	1.16 (1.12 - 1.21)	1.12 (103 - 1.20)	1.15 (1.12 - 1.19)
		p-value	$6.31x10^{-4}$	$3.10x10^{-5}$	$1.17x10^{-6}$	$1.21x10^{-12}$	$8.82x10^{-3}$	$5.12x10^{-14}$
		Heterogeneity p-value				$3.37x10^{-1}$		4.19x10-1
13q22.1 rs9543325 73,916,628 *KLF5 and KLF12*	C/T	maf cases;controls	0.44; 0.37	0.43; 0.38	0.43; 0.37			
		info	g	g	g			
		OR (CI)	1.26 (1.19 -1.33)	1.19 (1.09 -1.28)	1.24 (1.17 -1.30)	1.24 (1.19 - 1.28)		
		p-value	$2.87x10^{-11}$	$5.10x10^{-4}$	$1.91x10^{-10}$	$1.22x10^{-22}$		
		Heterogeneity p-value				$6.04x10^{-1}$		
16q23.1 rs7190458 75,263,661 *BCAR1*	A/G	maf cases;controls	0.06; 0.05	0.06; 0.04	0.06; 0.04		0.05; 0.04	
		info	0.74	g	g			
		OR (CI)	1.33 (1.16 - 1.50)	1.65 (1.43 -1.86)	1.27 (1.12 -1.41)	1.36 (1.26 -1.46)	1.34 (1.13 - 1.54)	1.36 (1.27 - 1.44)
		p-value	$9.38x10^{-4}$	$4.69x10^{-6}$	$1.39x10^{-3}$	$7.09x10^{-10}$	$5.07x10^{-3}$	$1.29x10^{-11}$
		Heterogeneity p-value				$1.27x10^{-1}$		$2.46x10^{-1}$

17q12 rs4795218 36,078,510 *HNF1B*	A/G	maf cases;controls	0.20 ; 0.23	0.22 ; 0.23	0.21 ; 0.23		0.21 ; 0.23	
		info	0.96	0.96	0.95		g	
		OR (CI)	0.87 (0.80 – 0.95)	0.88 (0.78 – 0.98)	0.88 (0.81-0.95)	0.88 (0.82 -0.93)	0.90 (0.82-0.98)	0.88 (0.84-0.92)
		p-value	$1.12x10^{-3}$	$2.29x10^{-2}$	$1.11x10^{-3}$	$2.73x10^{-7}$	$1.38x10^{-2}$	$1.32x10^{-8}$
		Heterogeneity p-value				$9.96x10^{-1}$		$9.78x10^{-1}$
17q25.1 rs11655237 70,400,166 *LINC00673*	T/C	maf cases;controls	0.13;0.11	0.13; 0.11	0.14; 0.11		0.13; 0.11	
		info	g	0.95	0.95			
		OR (CI)	1.17 (1.06 - 1.29)	1.26 (1.09 -1.47)	1.34 (1.21 -1.48)	1.25 (1.19 - 1.31)	1.24 (1.1 - 1.4)	1.25 (1.19 - 1.30)
		p-value	$2.17x10^{-3}$	$2.16x10^{-3}$	$1.05x10^{-8}$	$4.65x10^{-12}$	$6.40x10^{-4}$	$1.24x10^{-14}$
		Heterogeneity p-value				$1.55x10^{-1}$	$2.49x10^{-1}$	$2.39x10^{-1}$
17q25.1 rs7214041 70,401,476 *LINC00673*	T/C	maf cases;controls	0.13; 0.12	0.13; 0.11	0.14; 0.11		0.14; 0.12	
		info	0.96	g	g			
		OR (CI)	1.16 (1.05 - 1.28)	1.27 (1.10 -1.47)	1.32 (1.20 -1.46)	1.25 (1.18 - 1.31)	1.25 (1.11 - 1.41)	1.25 (1.19 - 1.30)
		p-value	$4.04x10^{-3}$	$1.39x10^{-3}$	$1.29x10^{-8}$	$6.58x10^{-12}$	$3.37x10^{-4}$	$9.49x10^{-15}$
		Heterogeneity p-value				$1.59x10^{-1}$	$3.69x10^{-1}$	$3.36x10^{-1}$
18q21.32 rs1517037 56,878,274 *GRP*	T/C	maf cases;controls	0.16; 0.19	0.17; 0.19	0.17; 0.18		0.17; 0.19	
		info	g	g	g			-
		OR (CI)	0.82 (0.75-0.89)	0.92(0.82 - 1.04)	0.90(0.83- 0.98)	0.87(0.82-0.93)	0.87(0.79-0.97)	0.86 (0.80-0.91)
		p-value	$7.56x10^{-6}$	$1.90X10^{-1}$	$1.64x10^{-2}$	$8.81x10^{-7}$	$1.17x10^{-2}$	$3.28x10^{-8}$
		Heterogeneity p-value				$1.87x10^{-1}$	$7.73x10^{-2}$	$1.03x10^{-1}$
22q12.1 rs16986825 29,300,306 *ZNRF3*	T/C	maf cases;controls	0.17; 0.15	0.18; 0.15	0.17; 0.15		0.20; 0.18	
		info	1	g	g			
		OR (CI)	1.16 (1.07 - 1.25)	1.22 (1.09 -1.35)	1.13 (1.04 -1.22)	1.16 (1.10 - 1.21)	1.14 (1.04 - 1.25)	1.15 (1.10 -1.20)
		p-value	$1.61x10^{-3}$	$2.02x10^{-3}$	$5.24x10^{-3}$	$2.93x10^{-7}$	$1.27x10^{-2}$	$1.21x10^{-8}$
		Heterogeneity p-value				$6.13x10^{-1}$		$7.97x10^{-1}$

a. Cytogenetic regions according to NCBI Human Genome Build 37
b. SNP position according to NCBI Human Genome Build 37
c. Results from the meta of PanScan 1 + PanScan 2, PanScan 3, and PanC4 genome-wide association analyses
d. Results from the meta of PanScan 1 + PanScan 2, PanScan 3, PanC4 and PANDoRA
e. Minor allele frequency
f. Quality of imputation metric. See online methods for more detail. If SNP is genotyped and not imputed, a 'g' is reported
g. Allelic Odds Ratio and corresponding 95% Confidence Interval
h. P---value from the test of heterogeneity of the Stage 1 studies (PanScan 1 & 2, PanScan 3 and PanC4)
and Stage2 (PanScan 1 & 2, PanScan 3, PanC4, and PANDoRA)

Fig 1 Genetic landscape of pancreatic cancer



Chapter 3. Rare variants

3.1 Definition, origin and sources

In general, a genetic variant is classified as rare when its minor allele frequency (MAF) is <0.01, low frequency when its MAF is between 0.01 and 0.05, and common when MAF is >0.05 [151].

At population level, the frequency of a genetic variant depends on its age and its effect on reproductive fitness, and it is regulated by natural selection [152, 153]. Population demographic events can affect the effect of natural selection on allele frequency [152, 153]. For example, selection has a much weaker effect on allele frequency in smaller populations derived by bottleneck events. In contrast, exploding growth, as the human population has experienced over the last ~5,000 years, resulting in an accumulation of extremely rare variants as results of higher selection pressure [152, 153].

In confirmation of this theory, whole-genome sequencing data collected from a large number of samples from different populations, show that a significant proportion of genetic variants (>1/3) are rare [154, 155].

It is reasonable to argue that a proportion of rare alleles derive from new mutations, particularly in genomic regions characterized by high mutation rate [155].

However, particularly for functional variants that are located in the coding part of the genome, it is hypothesized that their frequency in the population reflects the magnitude of their effect on the phenotype. A SNP with a substantial impact on a gene should be rare as a consequence of a stronger negative selection [154, 155].

Previous studies show that genes associated with common complex diseases, like cancer and cardiovascular diseases, have a notable excess of rare variants, particularly in patients with the related disease [152, 153]).

In the last ten years, there has been a significant advancement in sequencing technology, based on high-throughput parallel-sequencing approaches and known as next-generation sequencing. These new methods represent a great resource of rare and potentially functional

variants. However, the costs to apply this technique to the whole genome of a large number of samples, even if dropping quickly in time, is still very high and prohibitive for the most of research groups.

Low-depth whole genome sequencing (WGS) is a cheaper alternative than deep WGS and can be applied to a large number of samples as required in GWA studies. However, it has higher genotyping error rates, which reduces the statistical power [154] .

It has been estimated that the accuracy of common (MAF >5%) and low frequency (MAF 1-5%) polymorphisms identified through low-depth WGS can be improved by using large haplotypes reference panels as 1000 Genome Project; although it is still limited in the accurate detection of rare variants [156, 157].

Because WGS sequencing costs are dropping more quickly than genotyping arrays, literature is reporting the first GWASs performed with the new technology.

This new methodology has been recently applied with success in isolated populations [158, 159], where the frequency of rare functional variants may be higher as a consequence of bottleneck effect.

Exome-sequencing targets only the coding part of the genome, which represents only 1% - 2% of all genome with a high average depth (60x – 80x) [160]. Custom exome chips have been created for specific phenotype as Metabochip and Immunochip for metabolic and autoimmune disease, respectively [160]. Exome chips include both common and low frequency or rare variants located in target regions. Common variants usually include GWAS identified SNPs, and low frequency or rare SNPs are functional variants, mainly nonsynonymous, splicing, stop, gain or loss.

Illumina and Affymetrix have developed exome chips by applying exome-sequencing to 12,000 individuals mostly of European ancestry [161].

3.2 Statistical methods for studying rare variants

The application of single-variant association test, as GWAS, to rare variants (MAF <1%) has been proved useful only in rare cases when either the sample size or the variant effect size is very large.

Let us assume to design a case-control GWAS to study pancreatic cancer whose prevalence in the population is 1.6%, and we want to be able to detect a variant effect size of 1.4 with 80 % of statistical power and setting the type I error at 5×10^{-8}.

Which is the sample size we need?

We estimated that 6,500, 29,000, and 287,000 individuals are required to detect association with a variant whose MAF is 0.1, 0.01, and 0.001, respectively (http://zzz.bwh.harvard.edu/gpc/#cc_ins). This analysis demonstrates that the sample size increases exponentially with a ten units decrease of allele frequency.

Recently, new statistical methods have been created to overcome the reduced power issue associated with a single variant test applied to rare variants; overall they are known as aggregation tests because they measure the cumulative effects of multiple variants located into units of analysis; a unit can be a gene, a functional pathway including several genes or a locus identified through GWAS [162].

Among the high number of statistical methods developed for aggregation tests, regression-based methods are the most versatile as they allow to keep into account and adjust for the effect of covariates on the disease [162]. Logistic and linear regression tests are used for quantitative and binary phenotypes, respectively.

Let us suppose to apply a logistic regression model to test for association between a set of rare variants clustering into genes and a binary phenotype (disease/no disease), adjusting by age and gender.

$$\text{logit}\,(p_{disease}) = \beta_0 + \beta 1_{Age} + \beta 2_{Gender} + \beta 3 G_i \qquad \text{Equation 1}$$

For each individual in the analysis:

- logit $(p_{disease})$ = log $((p_{disease})\ (1- (p_{disease})^{-1})$ is the probability to be affected,
- β_0 is the intercept
- $\beta 1$ and $\beta 2$ are the regression coefficient of the covariates age and gender respectively.
- $\beta 3$ is the regression coefficient of the vector of genotypes $G_i = (G_{i1}, \ldots\ldots, G_{1m})$. The genotype vector is the allelic count of minor alleles (zero, one or two) for the m variants included in the set i.

According to the null hypothesis (H_0: $\beta 3 = 0$), there is no association between genes and disease and the resulting restricted model is:

$$\text{logit}\,(p_{disease}) = \beta_0 + \beta 1_{Age} + \beta 2_{Gender} \qquad \text{Equation 2}$$

H0 is tested with a likelihood ratio test that compares the likelihood of the full model (Equation 1) to that of the restricted model (Equation 2).

The likelihood ratio test statistic is $\chi 2$ distributed with *m* degrees of freedom, where *m* is the number of variants included in the genes.

Several analyses methods have been created to quantify the cumulative effect of rare genetic variants clustering into the same unit, and they mainly differ in the assumptions about the underlying relationship between the set of variants.

In general the score statistic for i genetic variants clustering into a set j is

$$Sj = \sum_{i=1}^{n} \text{Gij } (\text{yi} - \hat{\mu}\text{i}),$$

where μ is the estimated mean of the phenotype yi under the null hypothesis (H0: $\beta 3 = 0$).

The sign of Sj depends on the direction of the effect that the collapsed genetic variants have on the disease; Sj positive means increased disease risk and negative protective effect on the disease.

3.2.1 Burden Tests

Burden tests assume that all genetic variants included in the set are causal and have the same effect direction; violation of this assumption results in loss of power [162].

Burden tests collapse the information of multiple genetic variants into a single genetic score that is then used as a single SNP, to estimate the association with the disease. The most straightforward approach of burden tests considers all variants into the unit as having equal effect size and computes the genetic score by counting the number of minor alleles by unit; this test looks for an excess of the number of minor alleles in cases compared to controls [162].

In contrast, other burden tests compute the genetic score by incorporating additional information that determines a different weight for the variants included in the set. For example, variants might have different weight depending on sequencing or imputation quality scores, MAF and function ([163, 164].

Another class of burden tests is called data-adaptive because they estimate the weights of the variants from the data under study. Even if these methods result more robust compared to the classic burden test, they are often computationally intensive [162, 164].

3.2.2 Variance-Component Tests

Variance component (VC) tests have been developed to allow a different combination of genetic effects across the variants included in the set; actually, variants can have both a protective and at risk effect on the disease, and they can also vary by effect sizes [164].

Methods based on variant component test evaluate the distribution of statistics scores separately for each variant.

Among VC tests, Sequence kernel association test (SKAT) measures the genetic score as a weighted sum of squares of single-variant score statistics:

The weighting scheme that SKAT uses by default depends on MAF [165] according to the formula:

$$wj = Beta\ (MAFj, 1, 25) = 25(1 - MAF)$$

As reported in the above formula, SKAT up weights rare variants and down weights the more common ones.

Because SKAT collapses the squared of the statistic score test it is more powerful than Burden when the unit of analysis also includes non-causal variants or a mix of risk and protective variants; however it loses power when a significant percentage of variables into the set are causal [164].

.

3.2.3 Omnibus tests

Both methods above described, make strong assumptions on the genetic architecture of the disease, which is not possible to know a priori in the vast majority of the cases. To avoid the problem of making an assumption on the genetic model of the disease and reduce power, a more flexible class of methods have been developed, that combine burden and variance component tests.

The first of this kind is Fisher's method that combines SKAT and Burden p-values and then evaluates the significance of the test through permutation [166].

Another combined approach is a modified version of SKAT, called SKAT Optimal (SKAT-O), created to increase statistical power. SKAT-O test is a linear combination of SKAT and burden tests [162].

It includes a parameter Rho (ρ) that measures the pairwise correlation between the effects of the genetic variants involved in the same set. The parameter ρ is measured by a grid of values that range from 0 to 1 and uses it like a weight in the linear combination. The value 0 corresponds to the scenario of no correlation between variants, and in this scenario, SKAT test has the maximum power; on the other hand, the value 1 corresponds to the case where all variants in the set are casual, and Burden test has the most potent power.

SKAT-O tests all combination of genetic effect and directions among the variants and then selects the rho value that corresponds to the smaller p-value [162].

Chapter 4. Genotype Imputation

Genotype Imputation is a relatively new methodology increasingly used in the field of genetics as a source of new in silico genotypes at no cost.

This technique uses public databases that include phased sequencing data from a large number of individuals recruited from different populations. The most commonly used public datasets for genotype imputation are HapMap [167] and 1000 Genome [168].

The most recent phase of 1000 Genome (phase 3) includes 2,504 individuals sampled from 26 populations in Africa, East, and South Asia, Europe and America [168]. All individuals have been sequenced using both whole-genome sequencing and targeted exome sequencing for a total of 88 million variant sites [168]. Genetic data have been phased using a multi-stage approach that includes both the use of trios data and bioinformatics tools [168].

Genotype imputation is frequently applied to increase the power of Genome-Wide Association studies (GWASs) by increasing the number of SNPs analyzed, to allow meta-analysis of GWASs performed using different genotyping arrays and to fine mapping target genetic regions.

Several tools based on different algorithms have been developed to perform genotype imputation. The common mechanism for each algorithm is to identify shared regions among the haplotypes of sample study and those of the reference panel. These shared regions are identical by descent (IBD) segments of chromosomes that pass along generations from common ancestors [169].

Members of the same family share extended IBD regions because they have a close common ancestor; but for unrelated individuals, it is expected to see much shorter IBD regions because their common ancestor is far more distant in generational time ([154, 169].

Li et al estimated that unrelated European samples share from 100 to 200 Kb IBD regions [154].

The main steps of genotype imputation process are described in detail in **Figure 1**.

Figure 1A highlights the difference in genetic variants density between the study sample and reference panel.

Figure 1B represents the alignment of shared IBD regions between the sample and reference. During this step, it is possible that more than one haplotype from the reference is a potential match for the sample. In fact, this step depends on the degree of uncertainty of the imputed genotype. Figure 1C shows the new inferred 'in silico' genotypes in sample study.

All imputation programs give an estimate of imputation quality, often indicated as r squared or info score, whose value (range 0 – 1) measures the correlation between the imputed and the correct genotype [154].

Several factors affect the quality of imputation; in general, the larger the size of the study and reference samples, both concerning the number of individuals and marker density, the better the quality of the genotype imputation [170].

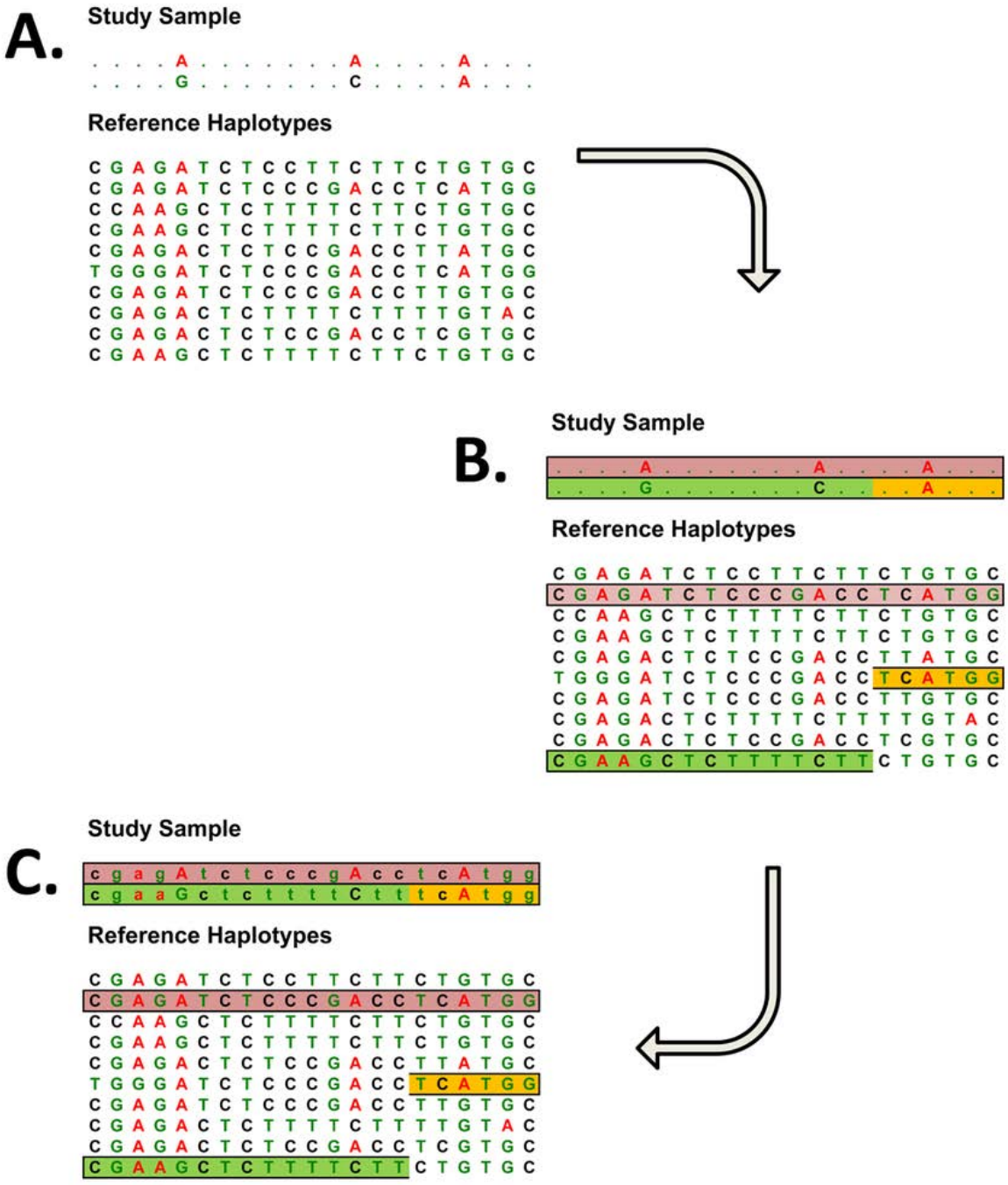
Among the factors that negatively affect imputation quality are low variant heterozygosity, low allele frequency, high sequence similarity to other genomic regions, and high GC content [170].

Another factor influencing the quality of imputation is the recombination rate [170].

It has been observed that genomic regions including genes involved in hematological traits and immune system diseases were characterized by low imputation quality.

This observation may be justified by the fact that these genes, whose function is to regulate the organism response to the attack by pathogens, are subjected to positive evolutionary pressure and the genomic regions that harbor them are characterized by high recombination rates [170].

Figure 1 Genotype imputation scheme



The picture has originally published in Annual review of genomics and human genetics 10(1):387-406 article. They authorized its use in this thesis

2. Hypothesis and Objectives

State of the art

GWASs conducted to date have highlighted the highly polygenic nature of pancreatic cancer, where many common loci have a small to moderate (10-30%) effect on the disease, by increasing or decreasing its risk.

Fine mapping and functional studies applied to genomic regions identified through GWASs have shown that despite the small-moderate effect on PDAC at the population level, common variants tag regions harboring loci involved in molecular processes of fundamental importance for the proper functioning of the pancreas.

GWAS will continue to be applied to the study of pancreatic cancer as to other complex diseases and traits since this approach will continue to identify new regions as the number of samples and genomic coverage increase.

However, GWASs main limitation, in pancreatic cancer as in other complex diseases and traits, is the 'missing heritability'. Indeed it has been estimated that GWASs findings, to date, explain only a small proportion of the disease-estimated heritability. This means that the majority of loci associated with complex diseases/ traits have yet to be identified.

In this context, the attention of geneticists shifted towards the study of rare variants for identifying genetic risk factors that explain the missing heritability.

This new phase in the field of genetics has been made possible by the recent advent of next-generation sequencing methodology.

Whole genome sequencing of a large number of samples allowed to establish that the majority of variants in the genome are low-frequency and rare (minor allele frequency <0.5%)[168].

According to the evolutionary theory, deleterious variants accumulate in the population as extremely rare, because they undergo high negative selection pressure [152].

Previous studies show that genes associated with complex diseases have an excess of rare variants in individuals affected with the disease compared to individuals not affected [162].

GWAS has limited statistical power to detect genome-wide significant association for rare variants; new analytical methods, generally addressed as aggregation tests, have been developed for this purpose. Aggregation tests measure the cumulative effect of rare variants physically located into a set on the phenotype.

Hypotheses and Objectives

In this thesis, I present two studies that have a common primary objective of identifying new genetic risk factors associated with pancreatic cancer but have been designed to test different hypotheses.

The first study is a two-stage GWAS that tests the association between common variants and PDAC risk.

This study tests the following hypotheses:

- We expect that common variants (minor allele frequency >5%) with a direct effect on PDAC or that are highly correlated with functional variants, will be found more commonly in affected individuals than in healthy individuals.
- We expect that the analysis of a large new dataset as PanC4 will find new genomic regions associated with the disease.
- We expect that by combining existing pancreatic cancer datasets with the new dataset PanC4, we will reach a sample size that allows detecting genomic regions that were not possible to identify in the single datasets.
- Some of the genomic regions found associated with PDAC might be due to the effect of confounding factors.

The second study is a case-control gene-based analysis that tests the cumulative effect of non-synonymous rare and low frequency variants clustering by gene on the risk of pancreatic cancer

- We expect that rare variants have a higher impact on the risk of pancreatic cancer compared to common variants and that they may explain a higher percentage of pancreatic cancer phenotypic variation due to genetic factors.
- We expect that genes associated with the disease will have an excess of non-synonymous variants in cases compared to controls and that each variant will contribute to the phenotype.

3. METHODS

3.1 Two-stage GWAS: *Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 are associated with susceptibility to pancreatic cancer.*

3.1.1 Study Design

We conducted a two-stage GWAS of pancreatic cancer (Fig. 1). First, genome-wide genotyping of 8,052 subjects from nine studies within the Pancreatic Cancer Case-Control Consortium (PanC4) was conducted using the HumanOmniExpressExome-8v1 array. After quality control, 7,956 individuals (4,164 cases and 3,792 controls) and 654,470 SNPs were analyzed for association with PDAC using unconditional logistic regression adjusted by age and principal components eigenvectors.

We then conducted a genome-wide meta-analysis of the PanC4 data with data from PanScan 1 [83] and PanScan 2 [93](Combined Stage 1, Fig. 1). After quality control, we analyzed 528,179 SNPs and 3,746 individuals (1,856 cases and 1,890 controls) from PanScan 1 and 557,555 SNPs and 3,300 individuals (1,618 cases and 1,682 controls) from PanScan 2. Since the genotyping platforms differed across studies, missing genotypes were imputed using IMPUTE v2 [171], with 1000 Genomes [172] (release Dec 2013) and HapMap3 [173](release #2,2009) as reference panels.

For PanScan 1 and PanScan 2, we conducted association analysis using unconditional logistic regression including age and principal components eigenvectors as covariates. Data from PanC4, PanScan 1, and PanScan 2 were combined (7,638 cases and 7,364 controls and 866,891 SNPs) and analyzed using a fixed-effects model.

We next conducted a Stage 2 analysis in an independent set of 2,497 cases and 4,611 controls from the PANDoRA consortium [174]. After quality control, 2,287 cases and 4,205 controls from the PANDoRA study were analyzed.

Twenty-five SNPs with p-values below 10×5 in either PanC4 or the Combined Stage 1 analyses

were tested for replication in PANDoRA. Lastly we conducted a combined analysis of the Stage 1 and 2 data for the 25 SNPs. **Figure 3.1.1** summaries the study main steps.

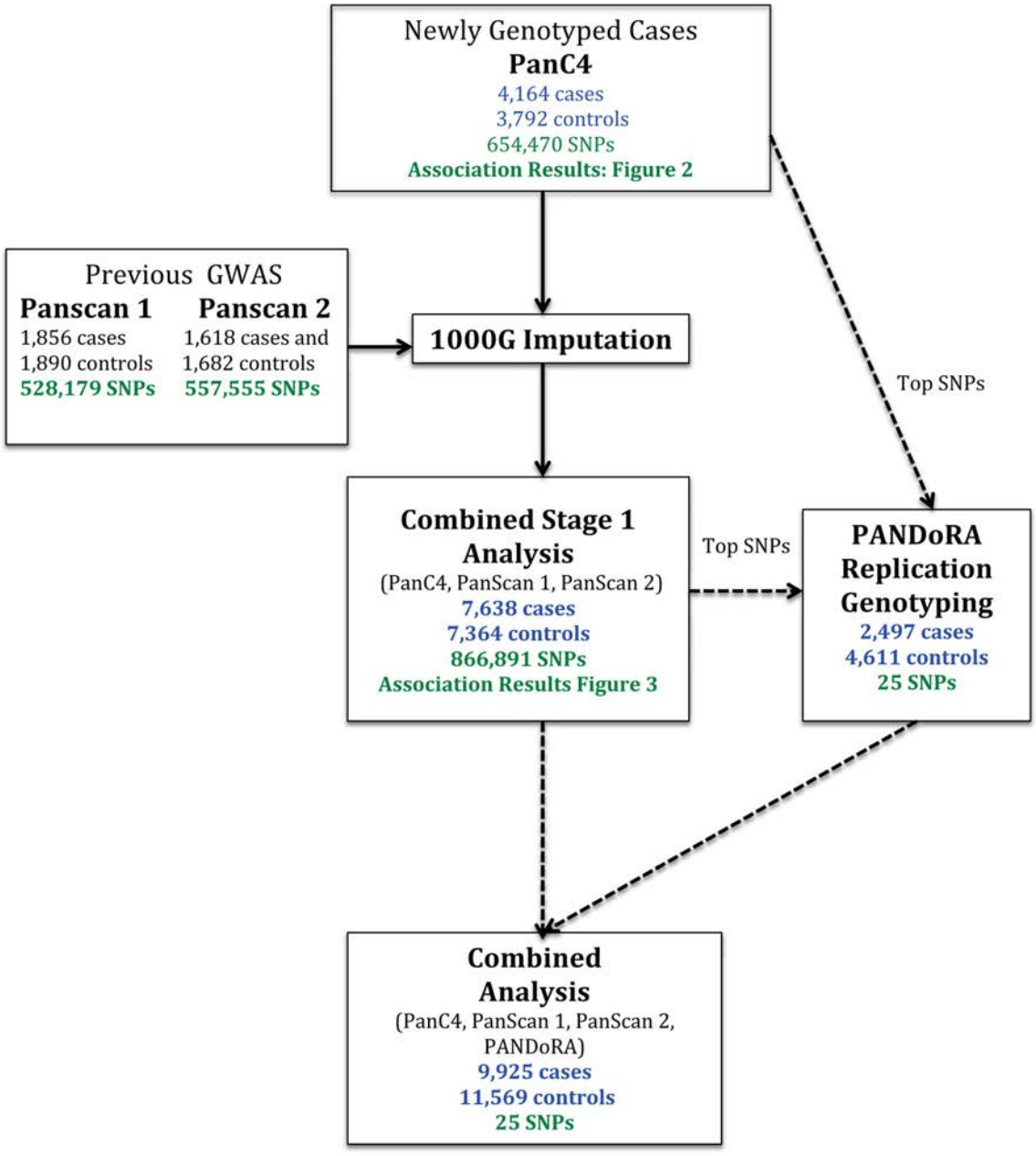


Figure 3.1.1 Overview of Stage 1 and Stage 2 analyses

3.1.2 STAGE 1

3.1.2.1 Study Population

PanC4

In total, 8,052 individuals were selected for genotyping from studies participating in the Pancreatic Cancer Case-Control Consortium (PanC4). Participating sites included: The Central Europe study coordinated by the International Agency for Research on Cancer (IARC/Central Europe)[175], Johns Hopkins Hospital [61, 176] Mayo Clinic [177], MD Anderson Cancer Center [178], Memorial Sloane-Kettering Cancer Center [53] [179], University of Toronto [180] , Queensland [181], University of California San Francisco (UCSF) [182], and Yale University [57] [183] (**Table 3.1.1**). Cases were defined as individuals with adenocarcinoma of the pancreas. DNA samples from these individuals from PanC4, 180 study duplicates, 176 HapMap control samples, and 26 replicates from the previous pancreatic cancer GWAS PanScan 2 [93],

The Johns Hopkins Institutional Review Board (IRB) approved the overall study. Each individual study obtained IRB approval from their parent institution.

PanScan 1 and PanScan 2

PanScan 1 and PanScan 2 data were obtained from dbGAP [184, 185] (dbGaP study accession: phs000206.v4.p3). Data from all participating sites apart from Group Health (which required a separate data sharing agreement) were included in the analysis.

Table 3.1.1 Characteristics study samples

Study	Cases	Controls	Accrual Years	Source of Cases	Source of Controls	Control Matching	Age at diagnosis Cases (SD)[a]	Age interview Controls (SD)	Male (%)	Cauc (%)[b]
IARC	448	456	2006-2010	Academic hospitals	General practitioners	Age Sex Region	63.84 (11.16)	61.88 (11.88)	57	100
Johns Hopkins Hospital	315	81	2007-2011	Clinic	Spouse in law	None	64.29 (11.53)	63.45 (14.73)	51	93
Mayo Clinic	1104	1027	2000-2010	Clinic	Primary Care patients	Age Sex Race Residence	65.92 (11.09)	63.34 (10.58)	56	93
MD Anderson	616	509	1997-2007	Hospital	Friends and spouses of non PC patients	Age Sex Race	62.6 (9.69)	59.2 (10.6)	59	100
Memorial Sloan Kettering	317	139	2000-2008	Clinic	Patients Spouses and visitors	None	64.03 (10.39)	61.68 (10.98)	64	87
Toronto	402	401	2003-2012	Population based cancer registry	Family medicine Clinic database	Age Sex Ethnicity	64.92 (11.03)	62.95 (11.73)	50	85
UCSF	253	248	2006-2010	Two UCSF Clinics	Three UCSF Clinics	Age Sex	62.52 (10.35)	60.4 (10.96)	54	82
Yale	156	366	2005-2009	Population- based hospitals and cancer registry	Enhancer RDD	Age Sex	67.02 (10.43)	65.15 (10.6)	59	93
Total[c]	**4170**	**3831**					**64.76**	**63.09**	**58**	**95**

a- SD- Standard Deviation, b- Percent self-identifying as Caucasian, c- Table excludes the 31 failed samples and 20 samples with unresolved issues

3.1.2.2 Genotyping and Quality Control

PanC4

Samples were genotyped on the IlluminaHumanOmniExpressExome-8v1 array at the Johns Hopkins Center for Inherited Disease Research (CIDR). Genotypes were called using GenomeStudio version 2011.1, Genotyping Module 1.9.4 and GenTrain version 1.0.

Genotyping results were inspected for quality by assessing the missing call rate, allelic imbalance, heterozygosity, discordance in reported versus genotyped gender, relatedness, ancestry and chromosomal anomalies. Unexpected relatedness between pairs of samples was assessed using the method of moments [186] implemented in *SNPRelate* [187]. The median genotype call rate was 99.9%, with all individuals having a call rate greater than 98%. After removing individuals with excessive allele sharing, duplicates and subjects with incomplete information on age, 7,956 subjects (4,164 cases and 3,792 controls) were available for statistical analyses (**Table 3.1.2**). SNPs with the following characteristics were excluded from statistical analyses: positional duplicates, more than two discordant calls in study duplicates, technical failures or missing call rate greater than 2%, more than one Mendelian error in HapMap control trios, Hardy-Weinberg equilibrium p-value<10−6, sex difference in allele frequency greater than 0.2 for autosomes/XY in samples of European ancestry, and minor allele frequencies (MAF) less than 0.005. Overall 654,470 SNPs passed the quality control filters applied; the median missing call rate was 0.024% and 98% of SNPs had a missing call rate less than 1% (Table 2).

PanScan1 and PanScan 2

PanScan 1 [83] and PanScan 2 [93] studies used the Illumina HumanHap550 and Illumina Human 610-Quad chips respectively. Quality control was performed as described above for PanC4. Forty-five unexpected duplicates between PanScan 1, PanScan 2, and PanC4 were identified and removed from analyses of the PanScan datasets. After data cleaning, 528,179 SNPs and

3,746 individuals (1,856 cases, 1,890 controls) remained in PanScan 1, and 557,555 SNPs and 3,300 individuals (1,618 cases and 1,682 controls) remained in PanScan 2 (See **Table 3.1.3**).

Table 3.1.2 Quality control steps applied to both Samples and SNPs in PanC4 dataset

Filters Applied	**N**
Samples	
Total Samples Genotyped	8,052
Failed Samples	31
Unresolved Identity Issues	20
Relatedness Issues	45
Total Samples Analyzed	**7,956**
SNPs	
Total SNPs Genotyped	951,117
MAF ≤ 0.005	244,744
Technical failures or missing call rate >2%	22,865
HWE p-value <10-6	9,477
Positional duplicates	18,699
>1 Mendelian error in HapMap controls trios	755
>2 discordant calls in study duplicates	102
Sex difference in allele frequency > 0.2	5
Total SNPs Analyzed	**654,470**

Table 3.1.3 Quality control steps applied to both Samples and SNPs in PanScan 1 and PanScan 2 datasets

Filters Applied	PanScan 1 N	PanScan2 N
Samples		
Total Samples Genotyped	3,937	3,484
Failed Samples	3	38
Unresolved Identity Issues	27	9
Replicated subjects	152	137
Relatedness Issues	9	0
Total Samples Analyzed	**3,746**	**3,300**
SNPs		
Total SNPs Genotyped	561,466	620,901
MAF ≤ 0.005	13,701	43,531
Technical failures or missing call rate >2%	12,443	16,131
HWE p-value <10-6	7,143	3,682
Missingness varying by phenotype (p –value < 10-5	0	2
Total SNPs Analyzed	**528,179**	**557,555**

3.1.2.3 Association Analysis

To investigate population structure, principal components analysis (PCA) was conducted separately for PanC4 (**Fig. 3.1.2**), PanScan 1 and PanScan 2 using *SNPRelate* [187] .

Genotype imputation was performed separately for PanScan 1, PanScan 2 and PanC4 using IMPUTE v2 [171]. Since PanScan 1 and PanScan 2 SNPs were originally mapped using an older genome assembly (NCBI build 36), we converted their genome position to genome assembly NCBI build 37 by using LIFTOVER.

Markers not identified in the build 37 assembly were removed. To decrease computational time, we pre-phased genotypes to produce best-guess haplotypes using SHAPEIT v2 software [188]. Both 1000 Genomes [172] Phase I-integrated haplotypes (release Dec 2013) and HapMap3 [173](release #2,2009) were used as reference panels during imputation.

After imputation, SNPs with quality scores < 0.3 were excluded from all subsequent analysis. Only SNPs directly genotyped in either PanC4, PanScan 1, or PanScan 2 and passing quality control filters were retained for analysis. This resulted in 866,891 SNPs in the Combined Stage 1 analysis. The expected genotype counts were then analyzed using the frequentist test option of SNPTEST [189]. Decade of age and eigenvectors from PCA were included as covariates. The number of eigenvectors to include was chosen based on inspection of the screen plot and p-values from association between eigenvectors and pancreatic cancer status.

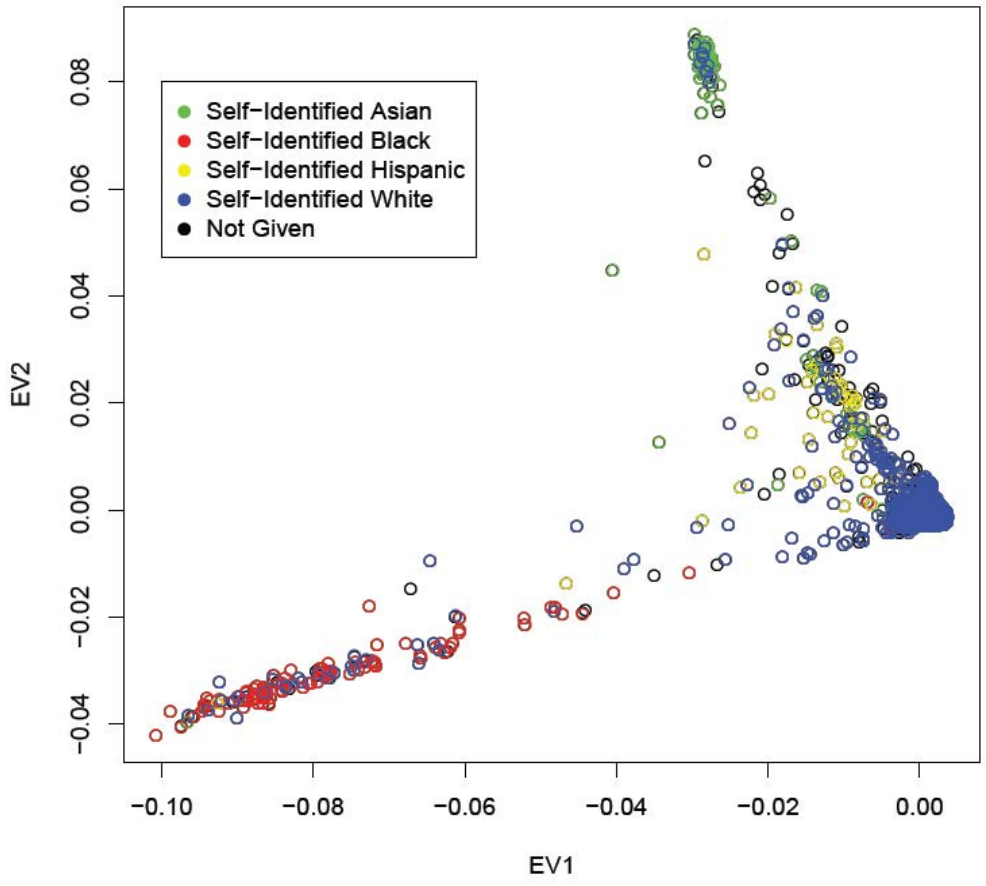
QQ plots (**Supplementary Fig.1.1 and Fig. 1.2**) indicated appropriate control of type-1 errors, with λ values of 1.025 for PanC4, 0.998 for PanScan 1, and 1.017 for PanScan 2.

The results from each study were then combined using a fixed-effects inverse standard error approach implemented by METAL [190] ('Combined Stage 1').

Test statistics for PanC4 and PanScan 2 were adjusted to account for small amounts of population stratification using METAL's genomic control option. Our sample size gives us over 80% power to detect an odds ratio of 1.2 for SNPs with a minor allele frequency greater than 0.20. To examine whether our association results were confounded by population stratification, we

conducted a secondary analysis, restricting our samples to those of European ancestry based on a PCA analysis performed with PanC4 and Hapmap3 samples. The loci identified through association testing did not change, and their odds ratios and p-values did not vary significantly (results not shown).

Figure 3.1.2 Plot of the first two eigenvectors from principal components analysis of PanC4



3.1.3 STAGE 2

3.1.3.1 Study Population

PANDoRA Replication Study

PANcreatic Disease ReseArch (PANDoRA) [174]consortium includes case-control studies from different European countries. In this study we analyzed 2,497 PDACs and 4,611 controls from six European countries: Czech Republic, Germany, Greece, Italy, Lithuania, and Poland (**Table 3.1.4**)

Table 3.1.4 PANDoRA dataset

	Pre Quality Control Filters		Post Quality Control Filters	
	Cases	**Controls**	**Cases**	**Controls**
By Country				
Germany	1166	1800	1071	1729
Italy	983	1702	914	1448
Czech Republic	60	542	57	531
Lithuania	58	192	57	174
Poland	106	207	90	173
Greece	124	168	98	150
By Gender				
Men	1392	2415	1307	2265
Women	1066	2126	980	1940
Missing (%)	1.56	1.52	0	0
By Age				
Mean (SD)	59.81 (10.35)	53.53 (11.21)	59.87 (10.39)	53.56 (11.19)
Missing (%)	2.36	6.96	0	0
Total	**2497**	**4611**	**2287**	**4205**

3.1.3.2 Genotyping and Quality Control

Twenty-five SNPs from 23 independent regions identified as showing evidence of association (P<1.10−5) in either the PanC4 analysis or the Combined Stage 1 analysis, were genotyped in samples from PANDoRA [174] with TaqMan technology.

Among all samples, 8% were duplicated and overall concordance was >99%. Samples missing more than 2 SNPs (~15%) or missing covariate information were excluded from analyses. In total, 2,287 cases and 4,205 controls from the PANDoRA study remained after quality control. Two SNPs, rs16867971 for Greece and rs10850078 for Lithuania, showed evidence of departure from HWE in controls (P<0.001). The SNP violating HWE was not analyzed for that country.

3.1.3.3 Association Analysis

The 25 SNPs chosen for inclusion in Stage 2 were tested for association, separately for each country of PANDoRA dataset. Logistic regression models with additive effects of each allele were fit, as implemented in PLINK [186]. Then we conducted a fixed-effects meta-analysis of each country ('PANDoRA') and lastly we performed meta-analysis of the 25 SNPs using PanC4, PanScan 1, PanScan 2, and PANDoRA datasets (Combined Stage 1 and 2 analysis).

3.1.4 Other Analyses

3.1.4.1 Heritability Analysis

Heritability analysis was performed using GCTA software. This analysis estimates the percentage of phenotypic variance explicated by common SNPs. We assumed a prevalence of 0.0149 (risk to age 90 in the US Caucasian population; SEER data collected in 2009–2011). We excluded individuals not clustering with HapMap [173] CEU (CEPH- Utah residents with ancestry from northern and western Europe) samples in PCA analysis as well as individuals with estimated relationships > 0.05 or missing genotype rate >0.01. SNPs with missing rate>0.05, MAF <0.01 and HWE p-value<5.10−4 were also excluded. We estimated the overall heritability in the PanC4

study using SNP data, as well as the heritability attributed to the 12 regions with significant evidence of association in the Caucasian population plus the 6 suggestive regions identified.

3.1.4.2 HaploReg

HaploReg is a tool used for exploring functional annotations of non-coding variants. For each variant and region identified in this study, we used HaploReg to gain insight into functional annotations including chromatin state (promoters and enhancers), conserved regions, variant effect on regulatory motifs and protein binding sites. Regions were defined by SNPs with r2>0.8 to the associated SNP.

3.2 *Exome-Array Gene-based Analysis*

3.2.1 Study Design

The analysis has been performed in PanC4 datasets.

Quality control steps were identical for samples, whereas for the SNPs, we changed the filtering criteria by MAF and we removed only monomorphic SNPs. After run all quality control steps we kept 4,164 PDACs and 3,792 controls genotyped with 821,150 SNPs for association analysis.

More than 81 million new variants were imputed using 1000 Genomes (phase 3) as reference panel. To maintain only good quality imputed SNPs, particularly for rare variants, we applied different quality score filters according to the variants MAF, keeping ~ 11 million variants for the next analysis step.

Imputed genotypes were converted into dosage format and all variants were annotated for gene location and function.

After annotation we selected only the non-synonymous variants that included also splice acceptor or donor sites, 2 bases only, and start or stop-altering variants. Approximately 138,000 variants were non-synonymous, and only 15% of them had been imputed.

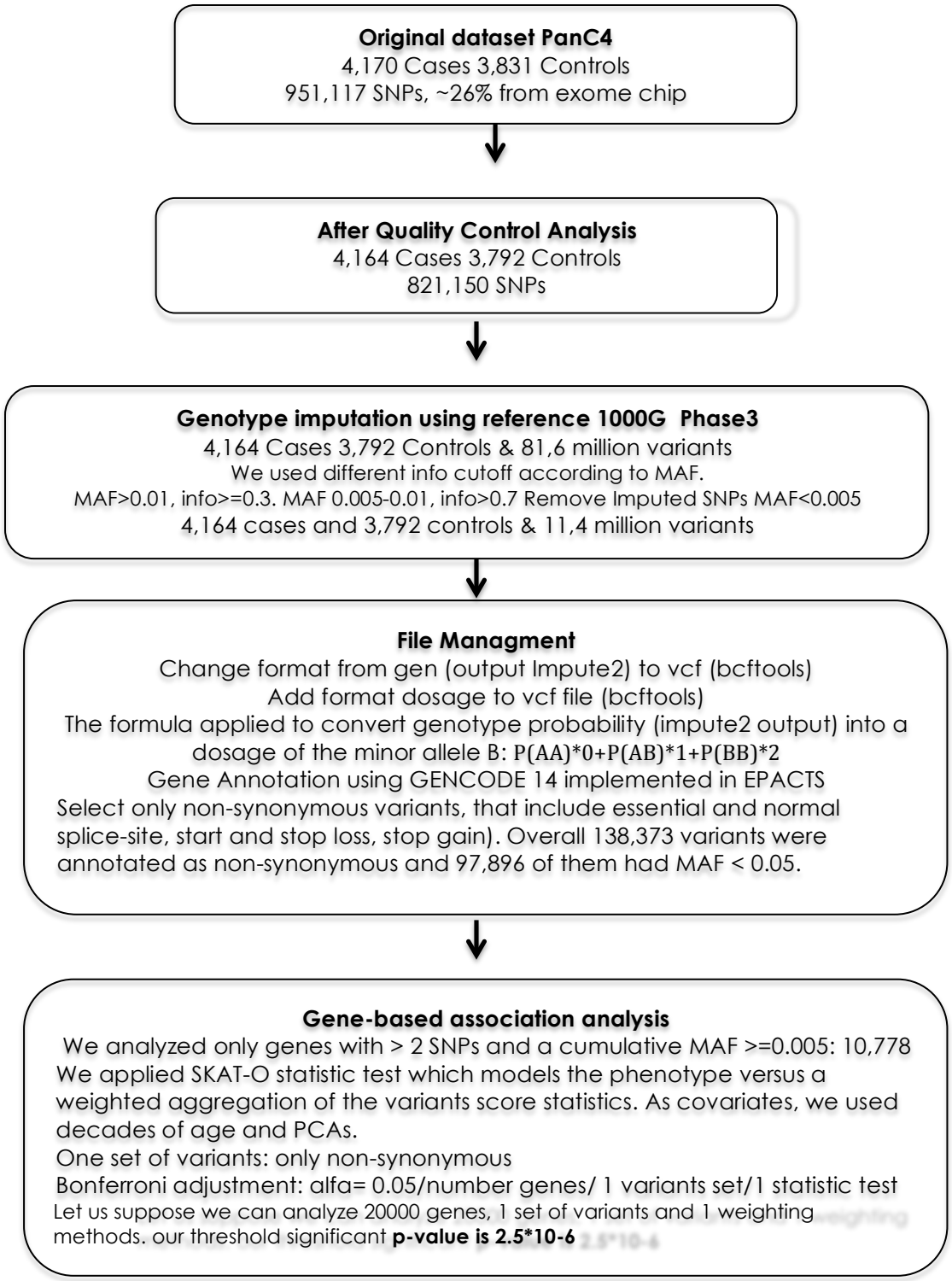
Among non-synonymous variants we filtered out those whose MAF was higher than 0.05, keeping ~ 98,000 variants in the final analysis.

We then clustered the selected variants into genes, and analyzed only genes that included a minimum of two variants and/or whose cumulative MAF was higher than 0.005. Our final analysis included ~ 98,000 clustering into 10,778 genes.

Gene-based association was tested using SKAT-O and only genes whose association p-value was $< 4.6 \times 10^{-6}$ were considered exome wide significant.

Figure 3.2.1 summarizes the main steps of this study.

Figure 3.2.1 Study Pipeline



3.2.2 Quality Control

Before imputation our dataset included 4,170 cases and 3,831 controls genotyped with ~951,000 autosomal SNPs (**Fig.3.2.1**).

In the quality control analysis, SNPs were removed if monomorphic, positional duplicates, missing call rate ≥ 2%, Hardy Weinberg Equilibrium (HWE) P-value ≤10^{-6}, MAF difference by sex >0.2, >2 discordant calls duplicate samples or > 1 Mendelian error in HapMap trios samples.

A detailed description of sample quality control analyses have been reported in Childs EJ et al. paper; among 8,052 individuals genotyped, 31 failed, and 21 had identity issues [138].

We assessed cryptic relatedness by conducting IBD analysis on a subset of common (MAF > 5%) and uncorrelated SNPs (r2 < 0.1), using SNPRelate [187] (R package). Overall 45 individuals were removed because of relatedness issues [138].

Individuals were also checked for differences between self-reported and X-chromosome genotype determined gender, chromosome anomalies, and ancestry group outliers.

To identify population structure including group outliers we performed principal component analysis (PCA), using SNPRelate [187] (R package). As in our previously published report (Childs et al. 2015), the top 7 eigenvectors were included as covariates in the regression models to controls for population substructure **(Supplementary Table 2.1).**

Approximately 130,000 SNPs and 96 samples were removed after quality control analysis, leaving 4,164 cases, 3,792 controls and 821,150 SNPs for the next analysis step (**Fig. 3.2.1, Table 3.2.1**).

Table 3.2.1 SNPs quality control steps

SNPs	
Total SNPs Genotyped	951,117
Not Autosomal	25,681
Monomorphic	57,582
Missing call rate ≥ 2%	19,465
HWE[a] P-value < 10-6	9,417
Positional Duplicates	17,305
> 1 Mendelian Error	424
>2 discordant calls in duplicates	93
MAF[b] difference by sex>0.2	0
Total SNPs analyzed	**821,150**

a = HWE: Hardy Weinberg Equilibrium; b= MAF: Minor Allele Frequency

3.2.3 Genotype imputation

To increase marker density and gene coverage, we applied genotype imputation using 1000 Genomes (Phase 3) as reference panel [168]. Before imputation, we checked that genotypes had the same strand alignment as that of the reference panel. When study and reference alleles did not match, we flipped study alleles. SNPs with A/T or G/C alleles and minor allele frequency close to 0.45 were removed because it was not possible to establish the correct alignment. After strand alignment, our data were pre-phased using Shape-IT [188]. Genotype imputation was performed using IMPUTE2 software [189].

For each chromosome, we created 4Mb sliding windows, separately for p and q chromosome arms.

3.2.4 Post-imputation filters

After imputation, our dataset included more than 81,6 million variants, 99.5% of them imputed. We observed that MAF median and mean values were 0 and 0.022 respectively and the average quality info score was 0.43 (**Supplementary Table 2.2**). Prior to annotation and generation of the .vcf files, we removed low-quality variants by applying a different quality score cutoff according to MAF; info >= 0.3 for variants with MAF > 0.01 and info > 0.7 for SNPs with MAF between 0.5% and 1%; rare and poorly imputed variants were removed, leaving approximately only 14% of imputed variants, with a mean MAF of 0.16 and a mean info-score of 0.94. The details of the imputation results are shown in **Supplementary Table 2.2.**

A final .vcf file including the minor allele dosage for imputed variants was generated using call format (vcf) file that was created using BCFTOOL tool.

https://samtools.github.io/bcftools/bcftools.html.

3.2.5 Gene annotation

Selected variants were annotated using GENECODE 14 [191] in EPACTS v3.2.3 (Efficient and Parallelizable Association Container Toolbox) http://genome.sph.umich.edu/wiki/ EPACTS.

In the present study, we selected 138,373 variants predicted to be nonsynonymous; 85% of them were from the exome-chip of Illumina HumanOmniExpressExome-8v1 platform, and the rest have been imputed. Overall non-synonymous variants were distributed in 19,213 genes. Since the GENCODE 14 gene reference set predicted 20,078 protein-coding genes, 865 genes were not included in our analysis (**Supplementary Table 2.3**).

3.2.6 Power analysis

We estimated the sample size we needed to detect an odds ratio (OR) from 2 to 3.5 with 80% of statistical power in SKAT, assuming disease prevalence of 0.015 (lifetime risk of pancreatic cancer years in the US Caucasian population; https://seer.cancer.gov) and a significance level of 0.05/10,778 or 4.6×10^{-6}.

3.2.7 Association analysis

After annotation, only non-synonymous SNPs were retained for analysis. We also included altered splice acceptor or donor sites, 2 bases only, and start or stop-altering variants.

The selected variants were collapsed into genes, and only genes including at least two variants and a minimum cumulative MAF of 0.005 were included in the final analysis.

A gene-based analysis was performed using Optimal Sequence Kernel Association Test (SKAT-O) [162] implemented in EPACTS (v3.2.3; http://genome.sph.umich.edu/wiki/ EPACTS; date last accessed ??). We applied the default software parameters; SNPs with MAF between 1e-6 and 0.05 and a minimum minor allele count (MAC) of 1 were included in the analysis.

SKAT-O uses a log-additive genetic model and a beta distribution weight proposed by Wu et al. [192]. The weight of the variants in the gene is inversely proportional to their MAF. As covariates, we included the first seven eigenvectors from principal component analysis and decade of age.

We tested for over-dispersion of gene-based analysis P-values by generating QQ-plot and estimating lambda (λ) as inflation factor value. Lambda was computed by converting SKAT-O p-values to chi-square statistics first and then by dividing their median value to the expected median value (0.456).

SKAT-O gene-based association P-values indicated appropriate control of type-1 errors, with $\lambda = 1.02$. QQ plot is showed in **Supplementary Figure 2.1**.

To correct for multiple testing, we used a Bonferroni-corrected threshold of 0.05/number of genes analyzed. Because we examined 10,778, P= 4.6×10^{-6} was the study cutoff to identify genes significantly associated with pancreatic cancer.

4. RESULTS

4.1 Two-stage GWAS

4.1.1 PanC4

Analysis of 7,956 newly genotyped PanC4 individuals identified a novel locus at 17q25.1 (*LINC00673*, rs7214041, OR=1.38, 95%CI:1.26–1.51, P=1.95.10×10) significantly associated with pancreatic cancer risk (**Figure 4.1.1, Figure 4.1.2 and Table 4.1.1**, column 'PanC4').

We observed 11 SNPs on chromosome 9q31.3 (**Supplementary Fig. 1.3e**) in moderate to high LD (r2 values between 0.6 and 1) with p-values from 7.00.10×8 to 2.73.10×6, including rs10991043 (OR=1.19, 95%CI:1.12–1.26, P=7.00.10×8) nearby the *SMC2* (structural maintenance of chromosome 2) gene.

In addition we replicate regions that had previously been reported to be associated with pancreatic cancer in the Caucasian population (**Supplementary Table 1.1**). These include: 9q34.2 [83] (*ABO*, rs505922, OR=1.27, 95%CI:1.19–1.35, P=1.72.10×13), 13q22.1 [93] (*KLF5,* rs9543325, OR=1.24, 95%CI:1.16–1.32, P=2.26.10×10), 5p15.33 [93] (*CLPTM1*, rs401681, OR=1.2, 95% CI:1.13–1.28, P=2.7.10×8), 13q12.2 [118] (*PDX1*, rs9581943, OR=1.17, 95%CI:1.10–1.24, P=1.94.10×7), 1q32.1 [93] (*NR5A2*, rs3790844, OR=0.83, 95%CI:0.77–0.90, P=3.05.10×6), 7q32.3 [118] (*LINC-PINT*, rs6971499, OR=0.81, 95%CI:0.74–0.88, P=7.1.10×6), 5p15.33 [118] (*TERT*, rs2736098, OR=0.85, 95%CI: 0.78–0.93, P=2.31.10×5), 16q23.1 [118] (*BCAR1*, rs7190458, OR=1.4, 95%CI=1.22–1.60, P=1.01.10×4), and 22q12.1 [118] (*ZNRF3*, rs16986825, OR=1.14, 95% CI= 1.04–1.24, P= 2.72.10×3). In contrast, other than 2p13.3 (*ETAA1*, rs2035565, OR=1.15, 95%CI=1.07–1.25, P=2.69.10×4) (**Supplementary Table 1.1**) we observed no evidence of association (P>0.05) for SNPs previously reported to be associated (P<1.10×6) with pancreatic cancer in Asian populations [[193, 194].

While all ethnic groups were included in our analyses, over 92% of our study population reported Caucasian ancestry. We obtained similar results when analysis was limited to individuals reporting European ancestry. Because of limited sample sizes we did not conduct independent analysis of other ethnic groups (results not shown).

4.1.2 Combined Stage 1

The Combined Stage 1 Analysis (**Table 4.1.1**, column 'Combined Stage 1') yielded a second novel region of association at 3q29 (*TP63*, rs9854771, OR=0.87, 95%CI:0.83–0.92, P=4.08.10×8) (**Figure 4.1.3, Figure 4.1.4 and Table 4.1.1**, column 'Combined Stage 1').

A second SNP on 17q25.1 (rs11655237, OR=1.27, 95%CI:1.19–1.36, P=6.74.10×12), which is in high LD (r2=0.95) with rs7214041, also gave significant evidence of association in these combined data.

4.1.3 PANDoRA Replication

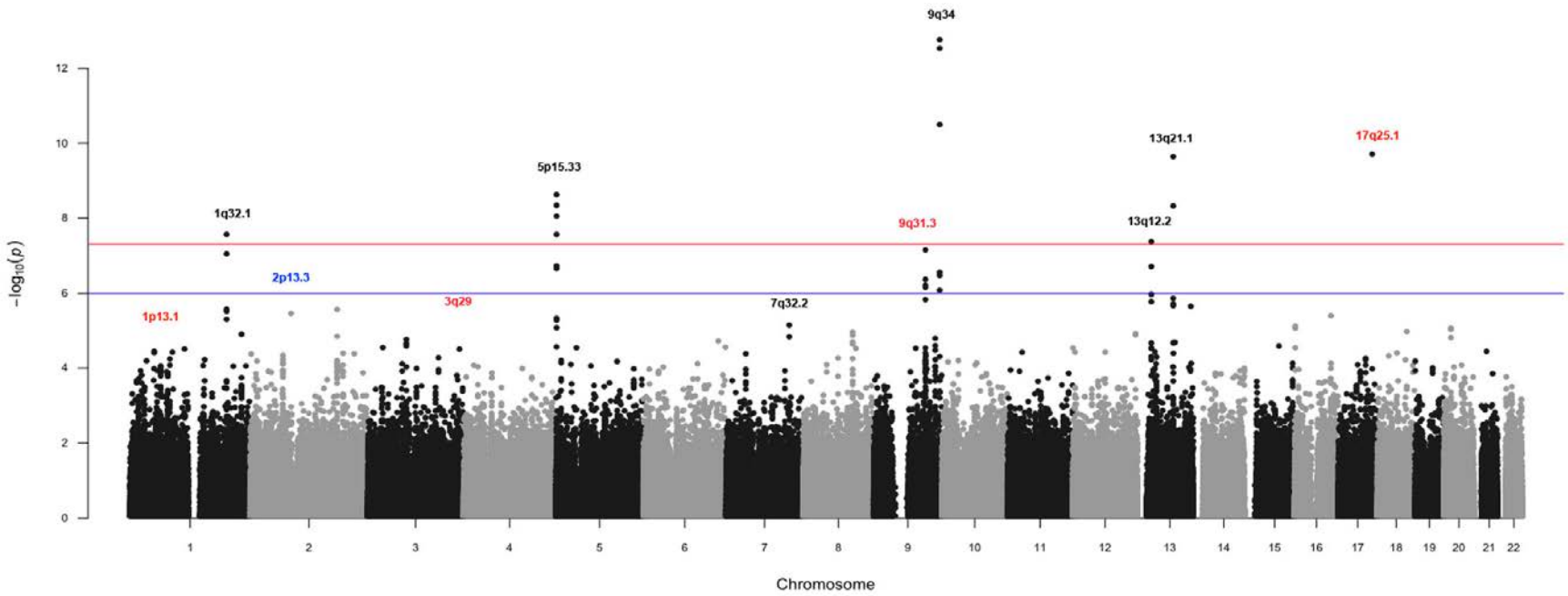
We observed independent evidence of association (p-value<0.05) at 17q25.1, 13q12.2, 2p13.3, 3q29 18q21.2, 20q13.11 in the PANDoRA study (**Supplementary Table 1.2, column 'PANDoRA'**).

4.1.4 Combined analysis of the Stage 1 and 2

Combined analysis of the Stage 1 and 2 data for the 25 SNPs (**Table 4.1.1** and **Supplementary Table 1.2, column 'Combined Stage 1&2'**) revealed two additional significantly associated loci: 2p13.3(*ETAA1*, rs1486134, OR=1.14, 95%CI:1.09–1.19, P=3.36.10×9) **(Figure Table 4.1.1 and Figure 4.1.5)** 7p13(*SUGCT*, rs17688601, OR=0.88, 95%CI:0.84–0.92, P=1.41.10×8) (**Table 4.1.1 and Figure 4.1.6**).

Promising signals (**Supplementary Table 1.3** and **Supplementary Figure 1.3**) arose at 18q21.2 (*GRP*, rs1517037, OR=0.87, 95%CI:0.83–0.92, P=3.17.10×7), 12q24.31(*HNF1A*, rs7310409, OR= 1.11, 95%CI:1.06–1.15, P=6.34.10×7), 1p13.1(*WNT2B*, rs351365, OR=0.89, 95%CI:0.85–0.93, P=7.39.10×7), and 20q13.11 (rs6073450, OR=1.11, 95%CI:1.06–1.15, P=9.21.10×7).

Figure 4.1.1. Manhattan plot of PanC4 association analysis



Loci previously associated with pancreatic cancer in Caucasians are shown in black, 2p13.3 in blue and novel loci in red.

Figure 4.1.2. Regional association and linkage disequilibrium (LD) plots of the novel genome wide significant locus 17q25.1 from PanC4 GWAS.

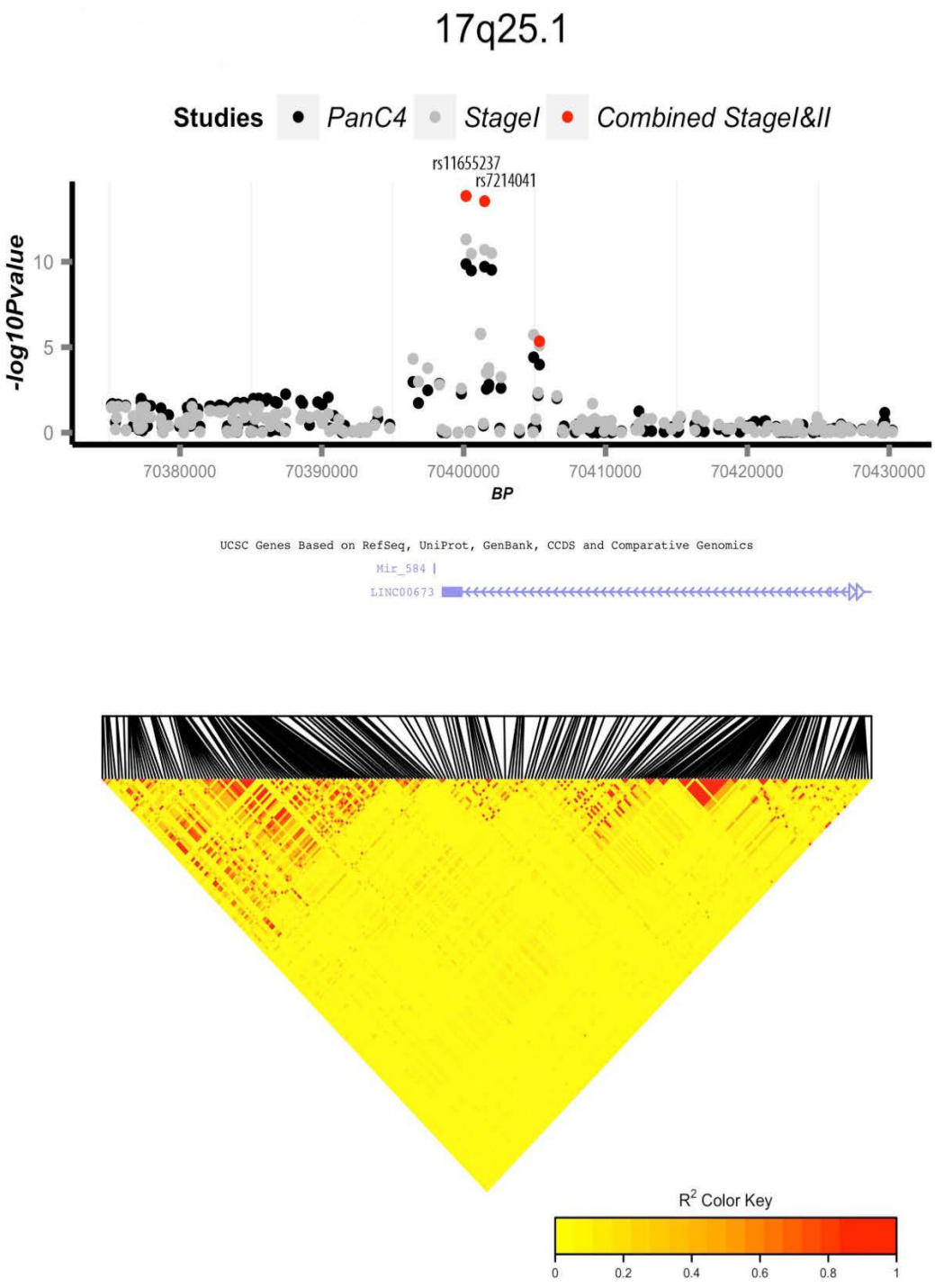


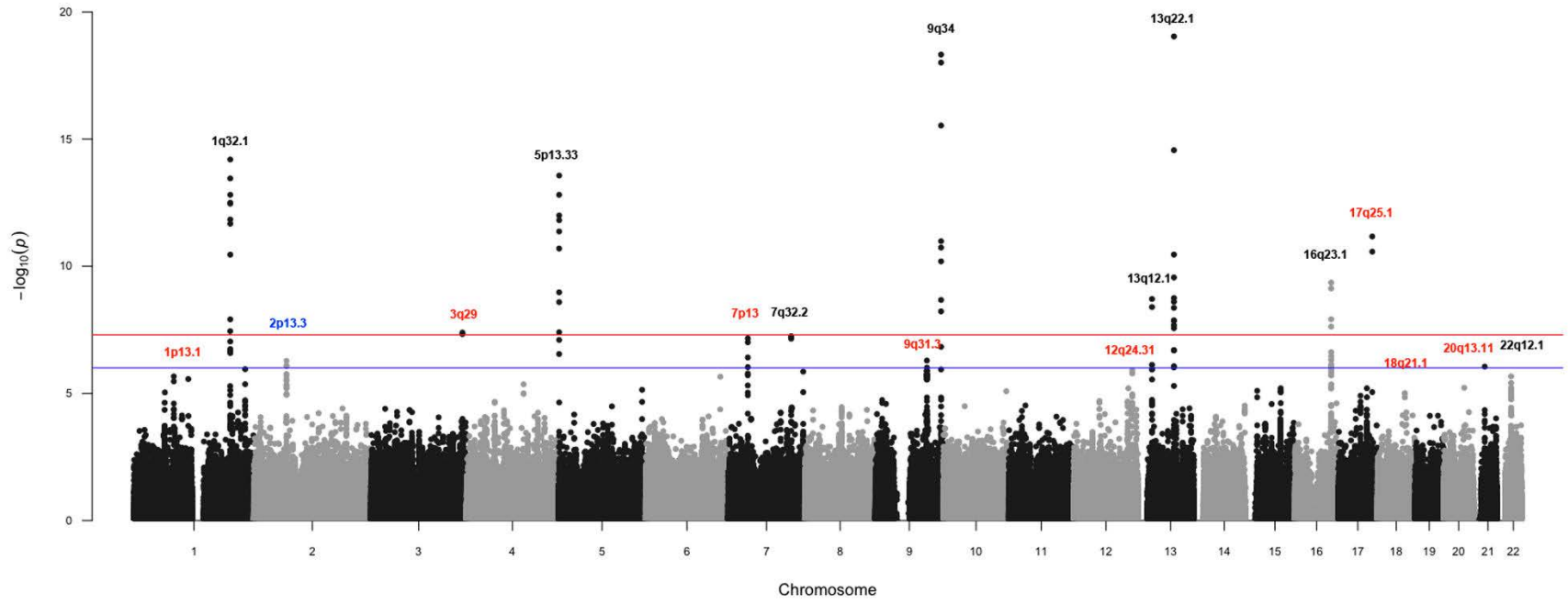
Table 4.1.1. Significant and highly suggestive ($P<1x10^{-6}$) association results for PDAC in both Stage 1 and Stage 2

			Stage 1				Stage 2	
Chr[a] SNP Position[b] Gene	**Effect Allele (Minor)/ Reference Allele**	**Statistic**	**PanC4** 4,164 cases 3,792 controls	**PanScan 1** 1,856 cases, 1,890 controls	**PanScan 2** 1,618 cases 1,682 controls	**Combined Stage 1[c]** 7,638 cases 7,364 controls	**PANDoRA** 2,497 cases 4,611 controls	**Combined Stage 1&2[d]** 9,925 cases 11,569 controls
Loci reaching genome–wide significance for association ($P<5x10^{-8}$)								
17q25.1[h] rs11655237 70,400,166 *LINC00673*	T/C	maf[e] cases;controls	0.146; 0.110	0.139; 0.129	0.149; 0.116		0.135; 0.114	
		info[f]	0.963	g	g			
		OR (CI)[g]	1.38 (1.26 – 1.52)	1.09 (0.96 – 1.25)	1.34 (1.16 – 1.55)	1.27 (1.19 – 1.36)	1.24 (1.10 – 1.40)	1.26 (1.19 – 1.34)
		p–value	$1.38 X 10^{-10}$	$1.95 X 10^{-1}$	$2.95 X 10^{-4}$	$6.74 X 10^{-12}$	$6.40 X 10^{-4}$	$1.42 X 10^{-14}$
17q25.1[h] rs7214041 70,401,476 *LINC00673*	T/C	maf cases;controls	0.148; 0.112	0.140; 0.133	0.150; 0.117		0.139; 0.117	
		info	g	0.966	0.96			
		OR (CI)	1.38 (1.26 – 1.51)	1.07 (0.93 – 1.22)	1.33 (1.15 – 1.53)	1.26 (1.18 – 1.35)	1.25 (1.11 – 1.41)	1.26 (1.19 – 1.34)
		p–value	$1.95 X 10^{-10}$	$3.36 X 10^{-1}$	$3.69 X 10^{-4}$	$2.67 X 10^{-11}$	$3.37 X 10^{-4}$	$2.88 X 10^{-14}$
2p13.3 rs1486134 67,639,769 *ETAA1* (2236bp 3')	G/T	maf cases;controls	0.302; 0.275	0.305; 0.292	0.305; 0.276		0.292; 0.273	
		info	g	g	g			
		OR (CI)	1.14 (1.06 – 1.22)	1.06 (0.96 – 1.18)	1.15 (1.03 – 1.28)	1.13 (1.08 – 1.19)	1.16 (1.06 – 1.27)	1.14 (1.09 – 1.19)
		p–value	$5.96 X 10^{-5}$	$1.57 X 10^{-1}$	$5.18 X 10^{-3}$	$8.35 X 10^{-7}$	$9.42 X 10^{-4}$	$3.36 X 10^{-9}$
7p13 rs17688601 40,866,663 *SUGCT*	A/C	maf cases;controls	0.241; 0.263	0.218; 0.254	0.237; 0.268		0.254; 0.277	
		info	g	g	g			
		OR (CI)	0.89 (0.83 – 0.96)	0.82 (0.73 – 0.91)	0.85 (0.76 – 0.94)	0.87 (0.82 – 0.91)	0.91 (0.83 – 1.00)	0.88 (0.84 – 0.92)
		p–value	$1.98 X 10^{-3}$	$1.66 X 10^{-4}$	$8.72 X 10^{-3}$	$9.77 X 10^{-8}$	$3.93 X 10^{-2}$	$1.41 X 10^{-8}$

3q29 rs9854771 189,508,471 *TP63*	A/G	maf cases;controls	0.328; 0.362	0.336; 0.366	0.325; 0.356		0.341; 0.356	
		info	g	0.998	0.998			
		OR (CI)	0.86 (0.81 – 0.92)	0.88 (0.80 – 0.90)	0.87 (0.79 – 0.97)	0.87 (0.83 – 0.92)	0.93 (0.86 – 1.01)	0.89 (0.85 – 0.93)
		p–value	3.10×10^{-5}	7.94×10^{-3}	1.55×10^{-2}	4.08×10^{-8}	1.01×10^{-1}	2.35×10^{-8}

[a] Cytogenetic regions according to NCBI Human Genome Build 37 and NCBI's Map Viewer
[b] SNP position according to NCBI Human Genome Build 37
[c] Results from the Combined Stage 1 meta–analysis of PanC4, PanScan 1, and PanScan 2
[d] Results from the Combined Stage 1 and 2 meta–analysis of PanC4, PanScan 1, PanScan 2, and PANDoRA
[e] MAF– minor allele frequency
[f] Quality of imputation metric. See online methods for more detail. If snp is genotyped and not imputed, a 'g' is reported
[g] Allelic Odds Ratio and corresponding 95% Confidence Interval
[h] $r^2>0.95$

Figure 4.1.3 Manhattan plot of Combined Stage 1 association analysis.



Loci previously associated with pancreatic cancer in Caucasians are shown in black, 2p13.3 in blue and novel loci in red.

Figure 4.1.4. Regional association and linkage disequilibrium (LD) plots of the novel genome wide significant locus 3q29 from Combined 1 GWAS.

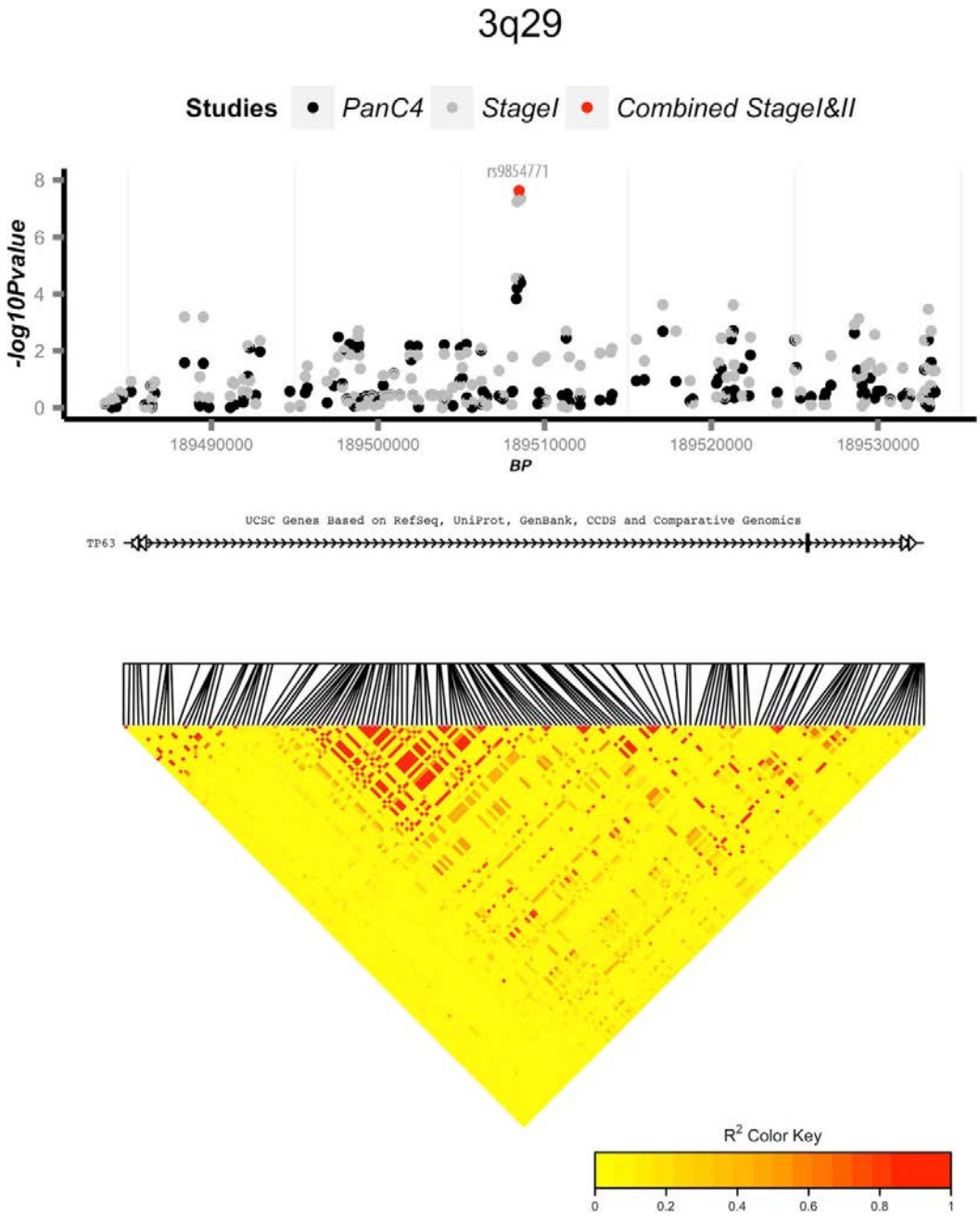


Figure 4.1.5. Regional association and linkage disequilibrium (LD) plots of the novel genome-wide significant locus 2p13 from Combined Stage 1 and 2.

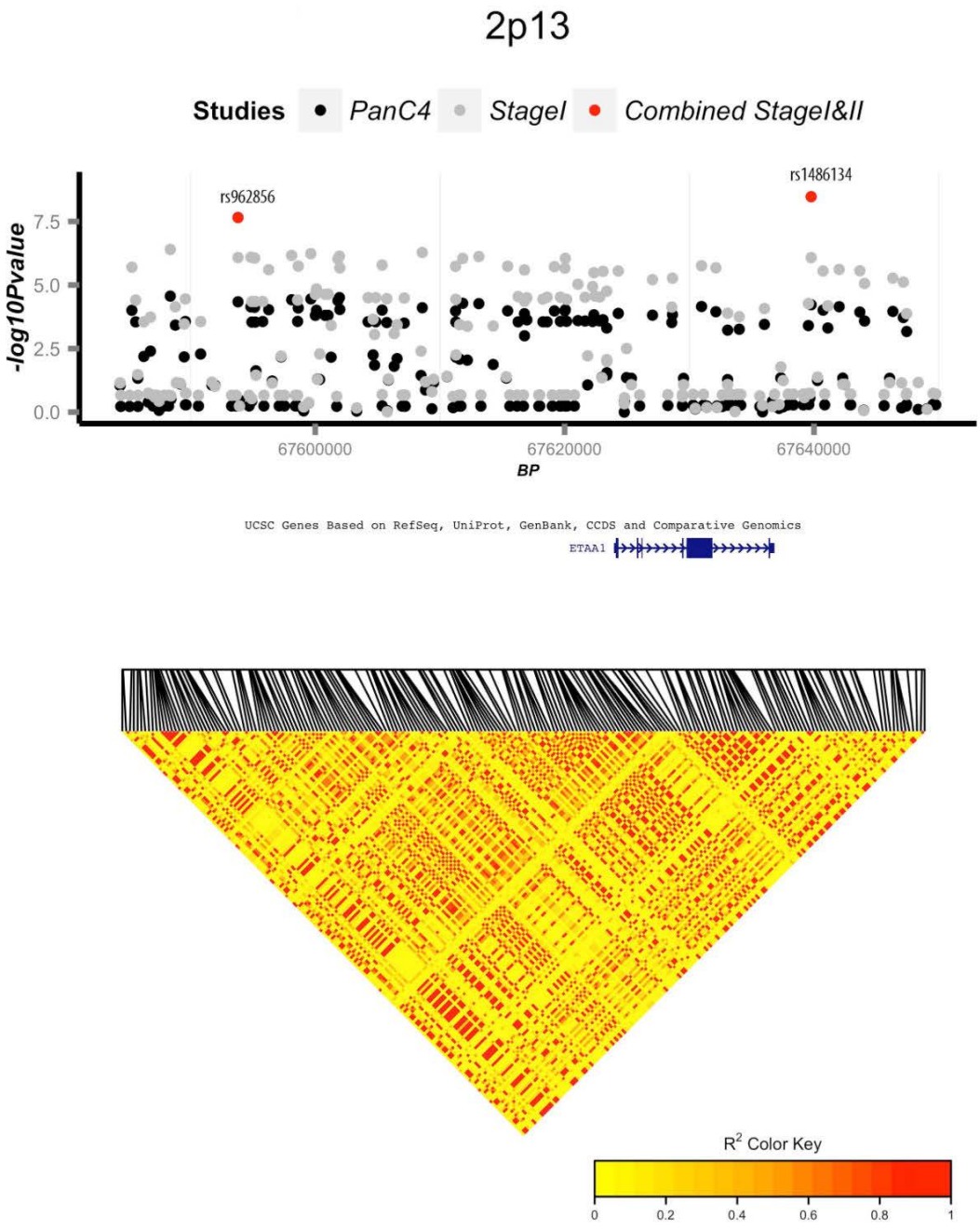
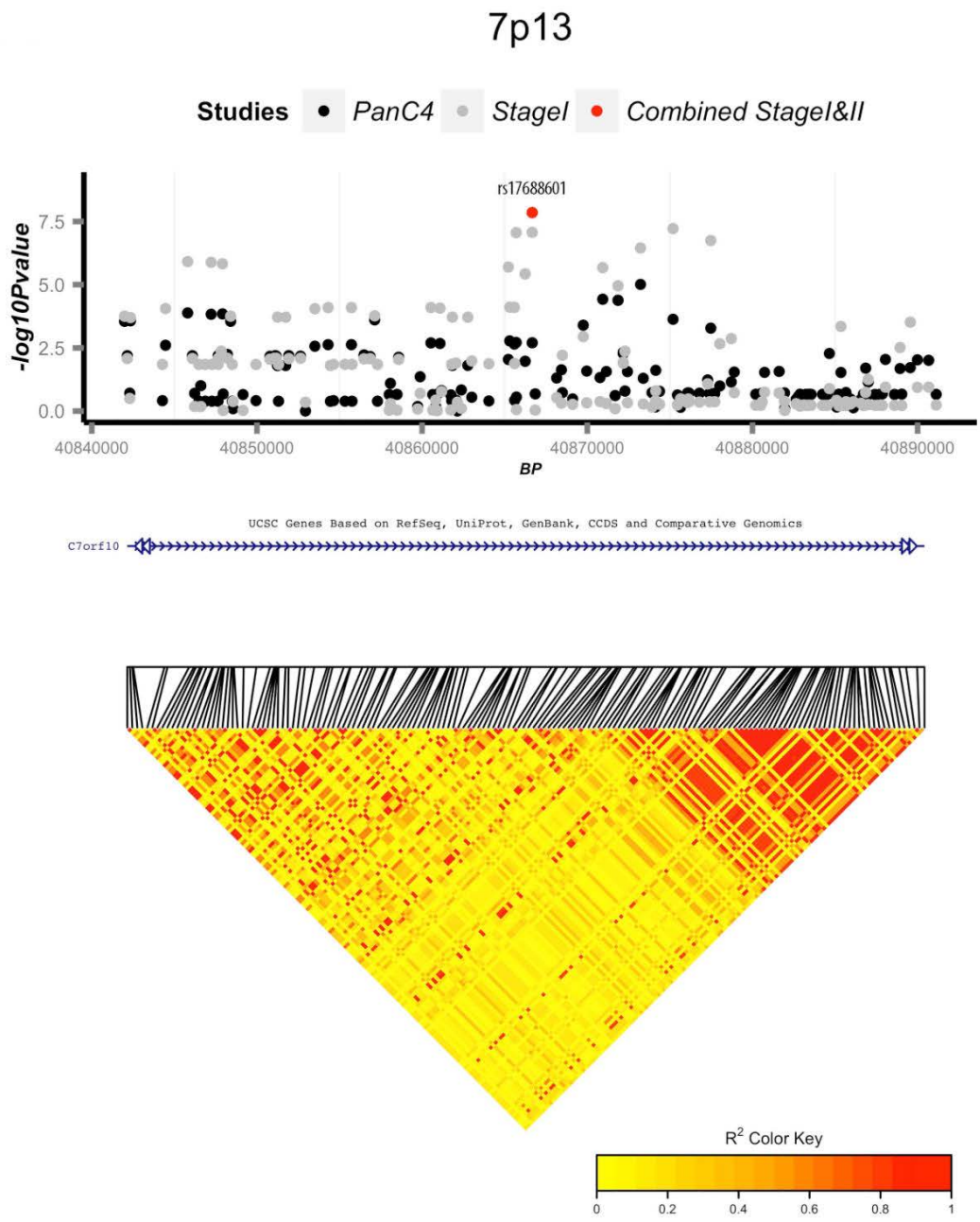


Figure 4.1.6. Regional association and linkage disequilibrium (LD) plots of the novel genome-wide significant locus 7p13 from Combined Stage 1 and 2.



4.1.5 Heritability Analysis

We estimated the heritability of pancreatic cancer due to common GWAS SNPs using data from PanC4 samples of Caucasian ancestry using only directly genotyped SNPs (3,828 cases, 3,551 controls and 620,357 SNPs) as well as the combined stage 1 dataset (7,032 cases 6,866 controls 268,681 SNPs). Using a disease prevalence of 0.0149, reflecting the lifetime risk of pancreatic cancer, we estimated that 16.4% (95%CI: 10.4%–22.4%) in PanC4 and 13.1% (95%CI 9.9%–16.3%) for the combined dataset of the total phenotypic variation was explained by genome-wide common SNPs. The established associated regions (loci in Table 1 and Supplemental Table 3), accounted for 3.0% (95%CI: 2.0%-3.9%) and 2.1%(95%CI 1.7%-3.1%) of the total phenotypic variation in the Panc4 population and the combined dataset, respectively.

4.2 *Exome-Array Gene-based Analysis*

Table 4.2.1 shows sample distribution by gender and decades of age, specifically age at diagnosis for cases and age at interview for controls. Overall cases and controls matched by number in age and gender strata.

Only genes including at least 2 non-synonymous variants and/or those with a minimum cMAF of 0.005 were kept in the final dataset that included 97,896 variants grouped into 10,778 genes; 44% of the genes were excluded in the filtering process.

Filtered genes included on average 9.1 variants (range 2 - 461). Among all variants, approximately 15% of them were singletons, and they were found just in one sample, whether case or control.

In our analysis, no gene reached an exome-wide significance P-value (P=4.6×10-6); however, we selected two genes as suggestive findings as their P-value were <1×10-4 (**Table 4.2.2, Figure 4.2.1**).

The most significant gene was *PDE12* or phosphodiesterase 12, located on chromosome 3 (3p14.3). Four missense variants have been selected in this gene; however one of them was excluded because its MAF was higher than 0.05. Accordingly, the resulting association test for *PDE12* derives from the cumulative effect of three rare variants, one of them a singleton, with complete concordant effect (rho =1) (**Table 4.2.2**).

The present analysis identified another interesting gene, *RAD52,* which encodes a protein with an important role in DNA double-strand damage repair through homologous recombination. Seven rare non-synonymous variants, 5 genotyped and 2 imputed, contributed with the same effect direction to the final gene signal.

Table 4.2.1 Sample characteristics

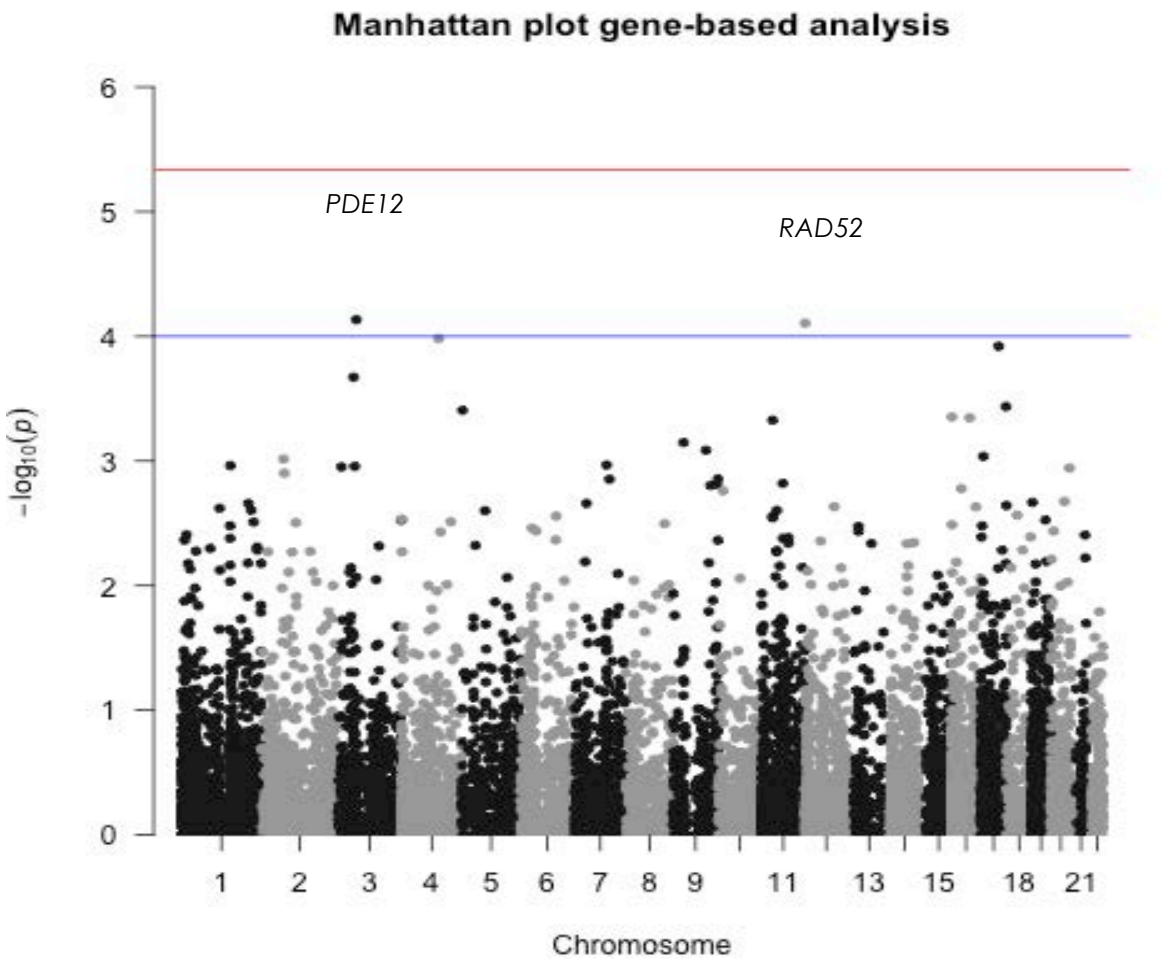
	Cases	Controls
Number	4164	3792
By Gender		
Men	2396	2106
Women	1768	1686
Age decades		
20	7	11
30	54	88
40	329	320
50	905	950
60	1475	1266
70	1054	888
80	324	263
90	16	6

Table 4.2.2 Top genes from gene-based association analysis and contributing SNPs

	Gene-based analysis									
Region	**Gene**	**All[a]**	**Pvalue[b]**	**Rho[c]**	**cMAF[d]**	**rsid[e]**	**alA[f]**	**alB[g]**	**Info**	**MAF**
3p14.3	*PDE12*	3	7.35E-05	1	0.017	exm325824	T	C	1	0.0115
						exm325832	C	A	1	0.0058
						exm2059005	C	G	1	0.0000
12p13.3	*RAD52*	7	7.83E-05	1	0.059	rs4987208	A	C	0.98	0.0192
						exm973603	G	T	1	0.0113
						exm973613	G	A	1	0.0001
						exm973619	G	C	1	0.0011
						rs7487683	C	T	0.68	0.0265
						exm973637	C	T	1	0.0002
						exm973655	C	T	1	0.0003

a. All: All variant clustering into the gene;
b. P-value: Gene-based association analysis P-value
c. Rho: correlation of the effect of the variants into the set.
d. cMAF: cumulative MAF.
e. rsid= SNP name using rs nomenclature
f. alA=Allele A
g. alB=Allele B
h. Info: imputation quality score
i. MAF: minor allele frequency

Figure 4.2.1 Results analysis



Red line exome wide significant –log10(Pvalue=4.6 × 10-6), Blue line exome wide suggestive –log10(Pvalue=1 × 10-4)

5. DISCUSSION

5.1 Two-Stage GWAS: Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 are associated with susceptibility to PDAC

The present study is a genome-wide association study on 9,925 pancreatic cancer cases and 11,569 controls, including 4,164 newly genotyped cases and 3,792 controls in 9 studies from North America, Central Europe and Australia.

We identified and replicated a novel region for association on 17q25.1 (**Fig. 4.1.2**). Two highly correlated variants (rs11655237 and rs7214041, r2=0.95) were associated with pancreatic cancer risk. Variant rs7214041 is to LINC00673 (long inter-genic non-protein coding RNA 673). rs11655237, a non-coding transcript variant, shows significant DNase hypersensitivity in multiple cancer cell lines and binds transcription factors including *P300*, *FOXA1*, *FOXA2*, and the DNA repair protein *RAD21* according to HaploReg v2 [195].

HaploRegV2 also indicated rs7214041 alters regulatory motifs for HNF1 [195].

Interestingly, we also found suggestive evidence of an association with rs7310409 located at the *HNF1A* locus (12q24.31, **Supplementary Fig. 1.3a** and **Supplementary Table 1.3**).

A recent study of the pancreatic cancer transcriptome suggests *HNF1A* may act as a tumor suppressor in pancreatic cancers [196]. Variation in *HNF1A* has been associated with risk of Type 2 diabetes [197, 198], a well-established risk factor for pancreatic cancer [16, 17, 21], and maturity onset diabetes of the young (MODY) [199]. Furthermore, variants in *HNF1A* (in particular rs7310409) and *HNF4A* were identified as risk factors for pancreatic cancer in pathway-based and candidate-SNP-based analyses of the PanScan data [139, 200].

We also identified significant association for two variants in high LD (rs9854771 and rs1515496, r2=0.99) located in an intron of *TP63* on 3q29 (**Fig. 4.1.4**). p63 is a p53 homologue implicated in tumorigenesis and metastasis [201] by playing a role in cell-cycle arrest and apoptosis. Overexpression of p63 can mimic p53 activation in certain experimental models [202].

Interestingly, different isoforms of p63 have opposing effects; TAp63 has tumor suppressive effects while DNp63 has oncogenic effects [203]. Danilov and colleagues suggested DNp63α was the predominant isoform in pancreatic cancer cell lines and promoted pancreatic cancer growth, motility and invasion [204]. Previous GWAS studies of lung cancer and bladder cancer have demonstrated significant evidence of association for SNPs in *TP63* [105, 205-208]. HaploReg query of this region showed that both are predicted to be conserved elements via GERP, suggesting functional roles.

Our analysis revealed genome-wide significance in a region on 2p13.3 (rs1486134). A pancreatic cancer GWAS in Han Chinese subjects [12] found suggestive evidence for another SNP on 2p13.3 (rs2035565) (**Supplementary Table 1.1**). High LD is present throughout this region (**Fig. 4.1.5**), including strong LD between rs1486134 and rs2035565 in European and Asian populations based on 1000 Genomes [15] samples (r2=0.91 and r2=0.90 respectively). This region includes the gene *ETAA1* (Ewing tumor-associated antigen 1), alias *ETAA16*, that may function as a tumor-specific cell surface antigen in the Ewing's family of tumors [209].

We observed significant association on 7p13 for rs17688601, located in an intron of the *SUGCT* (succinyl-CoA:glutarate-CoA transferase) gene (alias c7orf10) (**Fig. 4.1.6**).

This variant is predicted in HaploREGV2 to alter binding of *HNF1-4* and other DNA binding proteins [195]. The SUGCT protein is involved in glutarate metabolism and mutations in this gene are associated with glutaric aciduria [210]. While there is evidence of altered tricarboxylic acid cycle metabolism in pancreatic cancer [211], the role of this gene in pancreatic cancer risk is unclear.

Combined Stage 1 and Stage 2 identified suggestive evidence of association ($P<1.10\times6$) in four regions: 12q24.31(*HNF1A*) (**Supplementary Fig. 1.3a**), 18q21.2(*GRP*) (**Supplementary Fig. 1.3b**), 1p13.1(5' of WNT2B) (**Supplementary Fig. 1.3c**), and 20q13.11 (**Supplementary Fig. 1.3d**). GRP (gastrin releasing peptide) production has been associated with pancreatic tumor growth in vitro [212]. WNT signaling plays an important role in pancreas development. *WNT2B* (Wingless-

Type MMTV Integration Site Family, Member 2B) is overexpressed in pancreatic cancer and has been associated with decreased survival [213]. The 20q13.11 variant is located ~20kb of the *HNF4A* (MODY) gene, mutations of which are associated with early-onset diabetes [214].

In the PanC4 study we observed a new region on 9q31,3 near *SMC2* gene, that have not been replicated in the other datasets and combined analyses. *SMC2* plays an important role in DNA repair in humans. While there was no evidence of association in the other study populations examined, the strong signal across multiple SNPs in PanC4 suggest that this region merits further investigation.

Further functional characterization of these associated regions is needed, including examining if these SNPs are functional through eQTL. Performing eQTL analysis of pancreatic tissues is challenging. Normal pancreatic tissue is primarily comprised of acinar cells (>90%), but pancreatic ductal adenocarcinoma has a ductal phenotype, and the appropriate normal tissue to analyze is debatable because the cell of origin of pancreatic ductal adenocarcinomas is debated. eQTL analysis of pancreatic tumor tissue is also problematic because the tumor tissue of a pancreatic ductal adenocarcinomas contains a variable mixture of cell types including fibroblasts, multiple types of immune cells, non-neoplastic pancreatic cells and cancer cells, with cancer cells representing only a minority of the total cell population. Furthermore, gene expression analysis of normal pancreatic tissue is often limited by the RNA degradation associated with high level RNAase expression in pancreatic acinar cells. An ideal study of pancreatic eQTLs for pancreatic cells would take into account these challenges.

Smoking is a well-established risk factor for pancreatic cancer [[9, 215-217]. For all nine SNPs identified in **Table 4.1.1** and **Supplementary Table 1.3**, we conducted an analysis stratified by smoking status (ever smoker vs. never smoker) in PanC4 samples. No qualitative differences in effect size between current smokers and never smokers were observed (results not shown). Furthermore, when we included an interaction term in the model, this term was not significant at the 0.05 level.

5.2 Exome Array Gene-Based Analysis

The present study is the larger case-control analysis that focuses on the effect of rare and low-frequency functional variants on the risk of pancreatic cancer.

We estimated that our study had 80% power to detect an exome-wide significant association (P= 4.6 × 10-6) with a gene only when OR was 5; we observed that the statistical power dropped to 16% when the OR was 2.

None of the genes analyzed reached exome-wide significance.

The gene that provided the strongest evidence of association was *PDE12*. This gene is highly expressed in all tissues, and it encodes for a mitochondrial protein that removes poly(A) tails from mitochondrial mRNAs [218]. High expression of *PDE12* has been associated with severe inhibition of mitochondrial ATP production, thereby limiting oxidative metabolism [218]. *PDE12* has also been described as a negative regulator of the innate immune response; its inactivation has been associated with enhanced cellular resistance to viral pathogens [219].

Similar evidence of association was found for variation in *RAD52*, a gene that encodes a protein with an important role in DNA double-strand break repair process. Actually, in yeast *RAD52* interacts directly with *RAD51*, the protein that catalyzes DNA damage repair through homologous recombination [220].

In humans, *BRCA2* plays *RAD52* role and interacts directly with *RAD51,* supported in its function by other 2 crucial components of the homologous recombination pathway, *PALB2* (partner and localizer of *BRCA2*) and *BRCA1* [221].

It has been observed that *RAD52* has an essential role in an alternative DNA repair pathway that is *RAD51*-mediated; actually it has been reported that cells deficient in *BRCA2* [222]*, PALB2 or BRCA1* genes showed impact on cell growth only if also *RAD52* was depleted, demonstrating that Rad52 role is essential for cell survival when one of the components of the homologous recombination pathway is inactivated [223].

The association of pancreatic cancer with *RAD52*, described for the first time in this study, is

particularly interesting because mutations in *BRCA2, PALB2*, and *BRCA1* genes are well known genetic risk factors for pancreatic cancer, both for hereditary [54, 224, 225] and sporadic cases (Shindo K et al. 2017, Blair AB et al. 2018).

Furthermore, overexpression of *RAD51* has been observed in approximately 66% of pancreatic cancer cases [226].

In our analysis, the association with *RAD52* gene derives from the cumulative effect of seven rare non-synonymous SNPs (cMAF 0.06) with the same effect direction on the disease (rho=1) (**Table 4.2.2**).

6. CONCLUSIONS

Pancreatic cancer is a leading cause of cancer mortality in developed countries. Unlike other cancers, the incidence of pancreatic cancer has increased in recent years and by 2030 PDAC may emerge as the second leading cause of cancer death after lung.

Major modifiable risk factors include cigarette smoking, diabetes, obesity, alcohol intake.

Nonmodifiable risk factors include age and family history of pancreatic cancer.

The present thesis gives a detailed description of the current state of knowledge of genetic risk factors for PDAC.

Understanding the genetic causes of PDAC may be crucial both for developing new non-invasive diagnostic methods and more effective targeted treatments.

GWASs conducted so far, highlighted the highly polygenic nature of PDAC, where numerous polymorphisms give a small contribution to disease predisposition and the importance of common variation in pancreatic cancer risk

The two-stage GWAS reported in this thesis showed novel loci associated with PDAC and replicated the findings from previous studies. While it is of interest that many of these highly associated variants are located in the introns of genes, these associations could be due to more distant genomic effects. Follow-up studies, including functional studies, are needed to fully understand how these variants (either directly or indirectly) impact risk of pancreatic cancer.

In the two-stage GWAS, we estimated that common SNPs from the GWAS array explained only ~13-16% of PDAC heritability, meaning that > 80% of the genetic variants associated with PDAC have yet to be identified.

A decade of GWAS studies have led to similar conclusions for many complex diseases and traits.

In this context, and given the recent advances in next-generation sequencing technology, the attention of the geneticists is moving towards rare variants (MAF <1%).

Unlike the previous method described, GWAS using common SNPs that is hypothesis-free, the gene-based analysis gives different weight to the variants according to their frequency, as rare variants are supposed to have a stronger effect on the disease compared to common ones.

The second study reported in this thesis describes a case-control exome array gene-based association analysis. The study calculates the cumulative effect of rare and low frequency functional variants on the risk of pancreatic cancer.

Although our gene-based analysis did not find any whole genome significance gene, it nevertheless revealed a very strong functional candidate gene, *RAD52*.

This gene is involved in homologous recombination by interacting with *RAD51* in human cancer cells deficient in *BRCA1, PALB2,* or *BRCA2* genes.

Ultimately this thesis reports two complementary approaches to identify new genetic risk factors for PDAC. While previous GWAS, including ours, demonstrate the importance of common genetic variants, mainly as a great tool to shed light on the molecular mechanisms leading to PDAC, the focus of the research is shifting to functional variants and the development of new analysis methods for these variants.

6. CONCLUSIONES

El adenocarcinoma de páncreas (AP) es la principal causa de mortalidad por cáncer en los países mas desarrollados. A diferencia de otros cánceres, la incidencia de cáncer de páncreas ha aumentado en los últimos años y para 2030 el AP podría emerger como la segunda causa principal de muerte por cáncer después el cáncer del pulmón.

Los principales factores de riesgo modificables incluyen el tabaquismo, la diabetes, la obesidad y el consumo de alcohol. Los factores de riesgo no modificables incluyen la edad y los antecedentes familiares de cáncer de páncreas.

La presente tesis ofrece una descripción detallada del estado actual del conocimiento de los factores de riesgo genéticos para el AP.

Comprender las causas genéticas de AP podría ser crucial tanto para el desarrollo de nuevos métodos de diagnóstico no invasivos como para tratamientos dirigidos más efectivos. Los estudios de asociación en todo el genoma, llevados a cabo hasta hoy, resaltaron la natura altamente poligénica de AP, donde numerosos polimorfismos dan una pequeña contribución al desarrollo de la enfermedad y sobre la importancia de la variación genéticas común en el riesgo de cáncer de páncreas.

El estudio de asociación en todo el genoma en dos etapas, muestra regiones en el genoma que están reportada por primera vez en el AP y además replica los hallazgos de estudios anteriores.

Aunque es muy interesante el echo que muchas de estas variantes altamente asociadas con la enfermedad se encuentran en los intrones de genes, estas asociaciones podrían deberse a efectos genómicos más distantes.

Por lo tanto se necesitan estudios de seguimiento, que incluyan estudios funcionales, para comprender completamente cómo estas variantes (si directamente o indirectamente) afectan el riesgo de cáncer de páncreas

En el estudio de asociación en todo el genoma en dos etapas, se ha estimado que las variantes comunes de la plataforma de genotipado, explican solo un pequeño porcentaje (~ 13-16%) de

la herencia de la enfermedad; lo que significa que la mayoría (~ 80%) de las variantes genéticas asociadas con el cáncer de páncreas aún no se han identificado.

Una década de estudios de asociación genética en todo el genoma destaca un escenario similar para otras enfermedades y rasgos complejos.

Por esta razón y también gracias a la innovación de la tecnología de secuenciación del genoma, en los últimos años la atención de los genetistas se ha movido al estudio de variantes genéticas raras (frecuencia alelo menor <1%).

El segundo estudio reportado en esta tesis describe un análisis de asociación de casos y controles donde la unidad de estudio es el gen. El estudio calcula el efecto cumulativo de variantes funcionales raras y de baja frecuencia sobre el riesgo de cáncer de páncreas. A diferencia del método anterior descrito, estudio de asociación con variantes comunes, que no hace suposiciones sobre la genética de la enfermedad, el segundo estudio le da un peso diferente a las variables de acuerdo a su frecuencia en la población; de hecho el método utilizado se basa en la hipótesis que las variantes raras tienen un efecto más fuerte sobre la enfermedad en comparación con las comunes.

Aunque nuestro estudio no encontró ningún gen significativo en todo el genoma, sin embargo, reveló un nuevo fuerte candidato funcional para el cáncer de páncreas, *RAD52*. Este gen participa en la recombinación homóloga interactuando con *RAD51* en células cancerosas humanas deficientes en genes *BRCA1*, *PALB2* o *BRCA2*.

En última instancia, esta tesis informa dos enfoques complementarios para identificar nuevos factores de riesgo genéticos para el cáncer de páncreas. Mientras que el estudio de asociación en todo el genoma demuestra la importancia de variantes genéticas comunes, principalmente como una gran herramienta para descubrir los mecanismos moleculares que conducen al cáncer de páncreas, la atención de la investigación genética se está enfocando en el estudio de las variantes funcionales y en el desarrollo de nuevos métodos de análisis para estos variantes.

7. BIBLIOGRAPHY

1. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* (1097-0215 (Electronic)).
2. SEER*) SEER*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2016 Sub (1973-2014) <Katrina/Rita Population Adjustment>* 2016.
3. Lucas, A.L., et al., *Global Trends in Pancreatic Cancer Mortality From 1980 Through 2013 and Predictions for 2017.* (1542-7714 (Electronic)).
4. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2016.* (1542-4863 (Electronic)).
5. Malvezzi, M., et al., *European cancer mortality predictions for the year 2016 with focus on leukaemias.* (1569-8041 (Electronic)).
6. Rahib, L., et al., *Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States.* (1538-7445 (Electronic)).
7. He, J., et al., *2564 resected periampullary adenocarcinomas at a single institution: trends over three decades.* (1477-2574 (Electronic)).
8. Nathan, H., et al., *The volume-outcomes effect in hepato-pancreato-biliary surgery: hospital versus surgeon contributions and specificity of the relationship.* (1879-1190 (Electronic)).
9. Iodice, S., et al., *Tobacco and the risk of pancreatic cancer: a review and meta-analysis.* (1435-2451 (Electronic)).
10. Bosetti, C., et al., *Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4).* (1569-8041 (Electronic)).
11. Lynch, S.M., et al., *Cigarette smoking and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium.* (1476-6256 (Electronic)).
12. Muniraj, T. and S.T. Chari, *Diabetes and pancreatic cancer.* (1121-421X (Print)).
13. Permert, J., et al., *Pancreatic cancer is associated with impaired glucose metabolism.* (1102-4151 (Print)).
14. Chari, S.T., et al., *Pancreatic cancer-associated diabetes mellitus: prevalence and temporal association with diagnosis of cancer.* (1528-0012 (Electronic)).
15. Pannala, R., et al., *Prevalence and clinical profile of pancreatic cancer-associated diabetes mellitus.* (1528-0012 (Electronic)).
16. Chari, S.T., et al., *Probability of pancreatic cancer following diabetes: a population-based study.* (0016-5085 (Print)).
17. Everhart, J. and D. Wright, *Diabetes mellitus as a risk factor for pancreatic cancer. A meta-analysis.* (0098-7484 (Print)).
18. Huxley, R., et al., *Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies.* (0007-0920 (Print)).
19. Bosetti, C., et al., *Diabetes, antidiabetic medications, and pancreatic cancer risk: an analysis from the International Pancreatic Cancer Case-Control Consortium.* (1569-8041 (Electronic)).

20. Elena, J.W., et al., *Diabetes and risk of pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium.* (1573-7225 (Electronic)).
21. Li, D., et al., *Diabetes and risk of pancreatic cancer: a pooled analysis of three large case-control studies.* (1573-7225 (Electronic)).
22. Permert, J., et al., *Improved glucose metabolism after subtotal pancreatectomy for pancreatic cancer.* (0007-1323 (Print)).
23. Arslan, A.A., et al., *Anthropometric measures, body mass index, and pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan).* (1538-3679 (Electronic)).
24. Genkinger, J.M., et al., *Alcohol intake and pancreatic cancer risk: a pooled analysis of fourteen cohort studies.* (1055-9965 (Print)).
25. Tramacere, I., et al., *Alcohol drinking and pancreatic cancer risk: a meta-analysis of the dose-risk relation.* (1097-0215 (Electronic)).
26. Jiao, L., et al., *Alcohol use and risk of pancreatic cancer: the NIH-AARP Diet and Health Study.* (1476-6256 (Electronic)).
27. Lucenteforte, E., et al., *Alcohol consumption and pancreatic cancer: a pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4).* (1569-8041 (Electronic)).
28. Lowenfels, A.B., et al., *Hereditary pancreatitis and the risk of pancreatic cancer. International Hereditary Pancreatitis Study Group.* J Natl Cancer Inst, 1997. **89**(6): p. 442-6.
29. Lowenfels Ab Fau - Maisonneuve, P., et al., *Cigarette smoking as a risk factor for pancreatic cancer in patients with hereditary pancreatitis.* (0098-7484 (Print)).
30. Talamini, G., et al., *Chronic pancreatitis: relationship to acute pancreatitis and pancreatic cancer.* (1590-8577 (Electronic)).
31. Duell, E.J., et al., *Pancreatitis and pancreatic cancer risk: a pooled analysis in the International Pancreatic Cancer Case-Control Consortium (PanC4).* (1569-8041 (Electronic)).
32. Koushik, A., et al., *Intake of fruits and vegetables and risk of pancreatic cancer in a pooled analysis of 14 cohort studies.* (1476-6256 (Electronic)).
33. Bae, J.M., G. Lee Ej Fau - Guyatt, and G. Guyatt, *Citrus fruit intake and pancreatic cancer risk: a quantitative systematic review.* (1536-4828 (Electronic)).
34. Paluszkiewicz, P., et al., *Main dietary compounds and pancreatic cancer risk. The quantitative analysis of case-control and cohort studies.* (1877-783X (Electronic)).
35. Larsson, S.C. and A. Wolk, *Red and processed meat consumption and risk of pancreatic cancer: meta-analysis of prospective studies.* (1532-1827 (Electronic)).
36. Stolzenberg-Solomon, R.Z., et al., *Tooth loss, pancreatic cancer, and Helicobacter pylori.* (0002-9165 (Print)).
37. Michaud, D.S., et al., *Periodontal disease, tooth loss, and cancer risk in male health professionals: a prospective cohort study.* (1474-5488 (Electronic)).
38. Fan, X., et al., *Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study.* (1468-3288 (Electronic)).
39. Schulte, A., et al., *Association between Helicobacter pylori and pancreatic cancer risk: a meta-analysis.* (1573-7225 (Electronic)).
40. Gandini, S., et al., *Allergies and the risk of pancreatic cancer: a meta-analysis with review of epidemiology and biological mechanisms.* (1055-9965 (Print)).

41. Olson, S.H., et al., *Allergies and risk of pancreatic cancer: a pooled analysis from the Pancreatic Cancer Case-Control Consortium.* (1476-6256 (Electronic)).
42. Gomez-Rubio, P., et al., *Reduced risk of pancreatic cancer associated with asthma and nasal allergies.* (1468-3288 (Electronic)).
43. Falk, R.T., et al., *Life-style risk factors for pancreatic cancer in Louisiana: a case-control study.* (0002-9262 (Print)).
44. Friedman, G.D. and S.K. van den Eeden, *Risk factors for pancreatic cancer: an exploratory study.* (0300-5771 (Print)).
45. Fernandez, E., et al., *Family history and the risk of liver, gallbladder, and pancreatic cancer.* (1055-9965 (Print)).
46. Ghadirian, P., et al., *Reported family aggregation of pancreatic cancer within a population-based case-control study in the Francophone community in Montreal, Canada.* (0169-4197 (Print)).
47. Coughlin, S.S., et al., *Predictors of pancreatic cancer mortality among a large cohort of United States adults.* (0957-5243 (Print)).
48. Silverman, D.T., *Risk factors for pancreatic cancer: a case-control study based on direct interviews.* (0270-3211 (Print)).
49. Price, T.F., M.G. Payne Rl Fau - Oberleitner, and M.G. Oberleitner, *Familial pancreatic cancer in south Louisiana.* (0162-220X (Print)).
50. Schenk, M., et al., *Familial risk of pancreatic cancer.* (0027-8874 (Print)).
51. Jacobs, E.J., et al., *Family history of cancer and risk of pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan).* (1097-0215 (Electronic)).
52. Goggins, M., et al., *Germline BRCA2 gene mutations in patients with apparently sporadic pancreatic carcinomas.* Cancer Res, 1996. **56**(23): p. 5360-4.
53. Al-Sukhni, W., et al., *Germline BRCA1 mutations predispose to pancreatic adenocarcinoma.* (1432-1203 (Electronic)).
54. Jones, S., et al., *Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene.* (1095-9203 (Electronic)).
55. Roberts, N.J., et al., *ATM mutations in patients with hereditary pancreatic cancer.* Cancer Discov, 2012. **2**(1): p. 41-6.
56. Goldstein, A.M., et al., *Increased risk of pancreatic cancer in melanoma-prone kindreds with p16INK4 mutations.* (0028-4793 (Print)).
57. Kastrinos, F., et al., *Risk of pancreatic cancer in families with Lynch syndrome.* (1538-3598 (Electronic)).
58. Whitcomb, D.C., et al., *Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene.* (1061-4036 (Print)).
59. Brand, R.E., et al., *Advances in counselling and surveillance of patients at risk for pancreatic cancer.* (0017-5749 (Print)).
60. Welinsky, S. and A.L. Lucas, *Familial Pancreatic Cancer and the Future of Directed Screening.* (2005-1212 (Electronic)).
61. Klein, A.P., et al., *Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds.* (0008-5472 (Print)).
62. Roberts, N.J., et al., *Whole Genome Sequencing Defines the Genetic Heterogeneity of Familial Pancreatic Cancer.* (2159-8290 (Electronic)).

63. Moran, A., et al., *Risk of cancer other than breast or ovarian in individuals with BRCA1 and BRCA2 mutations.* (1573-7292 (Electronic)).
64. Mocci, E., et al., *Risk of pancreatic cancer in breast cancer families from the breast cancer family registry.* Cancer Epidemiol Biomarkers Prev, 2013. **22**(5): p. 803-11.
65. Hauke, J.A.-O.h.o.o., et al., *Gene panel testing of 5589 BRCA1/2-negative index patients with breast cancer in a routine diagnostic setting: results of the German Consortium for Hereditary Breast and Ovarian Cancer.* (2045-7634 (Electronic)).
66. McGarrity Tj Fau - Amos, C.I., M.J. Amos Ci Fau - Baker, and M.J. Baker, *Peutz-Jeghers Syndrome BTI - GeneReviews((R)).*
67. Beggs, A.D., et al., *Peutz-Jeghers syndrome: a systematic review and recommendations for management.* (1468-3288 (Electronic)).
68. Eckerle Mize D Fau - Bishop, M., et al., *Familial Atypical Multiple Mole Melanoma Syndrome BTI - Cancer Syndromes.*
69. Bergman, W., et al., *Systemic cancer and the FAMMM syndrome.* (0007-0920 (Print)).
70. Vasen, H.F., et al., *Risk of developing pancreatic cancer in families with familial atypical multiple mole melanoma associated with a specific 19 deletion of p16 (p16-Leiden).* (0020-7136 (Print)).
71. Rebours, V., et al., *The natural history of hereditary pancreatitis: a national series.* (1468-3288 (Electronic)).
72. Joergensen, M.T., et al., *Genetic, epidemiological, and clinical aspects of hereditary pancreatitis: a population-based cohort study in Denmark.* (1572-0241 (Electronic)).
73. LaRusch, J., et al., *Whole exome sequencing identifies multiple, complex etiologies in an idiopathic hereditary pancreatitis kindred.* (1590-8577 (Electronic)).
74. Rosendahl, J., et al., *Chymotrypsin C (CTRC) variants that diminish activity or secretion are associated with chronic pancreatitis.* (1546-1718 (Electronic)).
75. Witt, H., et al., *Variants in CPA1 are strongly associated with early onset chronic pancreatitis.* (1546-1718 (Electronic)).
76. Cohn, J.A., P.S. Mitchell Rm Fau - Jowell, and P.S. Jowell, *The impact of cystic fibrosis and PSTI/SPINK1 gene mutations on susceptibility to chronic pancreatitis.* (0272-2712 (Print)).
77. Schneider, A., et al., *Combined bicarbonate conductance-impairing variants in CFTR and SPINK1 variants are associated with chronic pancreatitis in patients without cystic fibrosis.* (1528-0012 (Electronic)).
78. Howes, N., et al., *Clinical and genetic characteristics of hereditary pancreatitis in Europe.* (1542-3565 (Print)).
79. Ligtenberg, M.J., et al., *Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1.* (1546-1718 (Electronic)).
80. Lander, E.S., *The new genomics: global views of biology.* (0036-8075 (Print)).
81. Wray, N.R., *Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies.* (1832-4274 (Print)).
82. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* (1537-6605 (Electronic)).
83. Amundadottir L Fau - Kraft, P., et al., *Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer.* (1546-1718 (Electronic)).

84. Yamamoto, F., et al., *An integrative evolution theory of histo-blood group ABO and related genes.* (2045-2322 (Electronic)).
85. Wolpin, B.M., et al., *ABO blood group and the risk of pancreatic cancer.* (1460-2105 (Electronic)).
86. CA, C., *Blood groups and disease.* In: Steinberg AG, editor. Progress in medical genetics, Vol. 1. New York, NY: Grune and Stratton; 1961. pp. 81–119., 1961.
87. Edgren, G., et al., *Risk of gastric cancer and peptic ulcers in relation to ABO blood type: a cohort study.* (1476-6256 (Electronic)).
88. Polk, D.B. and R.M. Peek, Jr., *Helicobacter pylori: gastric cancer and beyond.* (1474-1768 (Electronic)).
89. Raderer, M., et al., *Association between Helicobacter pylori infection and pancreatic cancer.* (0030-2414 (Print)).
90. Stolzenberg-Solomon, R.Z., et al., *Helicobacter pylori seropositivity as a risk factor for pancreatic cancer.* (0027-8874 (Print)).
91. Risch, H.A., et al., *ABO blood group, Helicobacter pylori seropositivity, and risk of pancreatic cancer: a case-control study.* (1460-2105 (Electronic)).
92. Risch, H.A., *Pancreatic cancer: Helicobacter pylori colonization, N-nitrosamine exposures, and ABO blood group.* (1098-2744 (Electronic)).
93. Petersen, G.M., et al., *A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33.* (1546-1718 (Electronic)).
94. Dong, J.T. and C. Chen, *Essential role of KLF5 transcription factor in cell proliferation and differentiation and its implications for human diseases.* (1420-9071 (Electronic)).
95. Chen, M.M., et al., *GWAS meta-analysis of 16 852 women identifies new susceptibility locus for endometrial cancer.* (1460-2083 (Electronic)).
96. Couch, F.J., et al., *Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer.* (2041-1723 (Electronic)).
97. Hoskins, J.W., et al., *Functional characterization of a chr13q22.1 pancreatic cancer risk locus reveals long-range interaction and allele-specific effects on DIS3 expression.* (1460-2083 (Electronic)).
98. Robinson, S.R., et al., *The 3′ to 5′ Exoribonuclease DIS3: From Structure and Mechanisms to Biological Functions and Role in Human Disease.* (2218-273X (Electronic)).
99. Weissbach, S., et al., *The molecular spectrum and clinical impact of DIS3 mutations in multiple myeloma.* (1365-2141 (Electronic)).
100. Fayard, E., K. Auwerx J Fau - Schoonjans, and K. Schoonjans, *LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis.* (0962-8924 (Print)).
101. Martin, M., et al., *Transcription factors in pancreatic development. Animal models.* (1421-7082 (Print)).
102. Naqvi, A.A.T., G.M. Hasan, and M.I. Hassan, *Investigating the role of transcription factors of pancreas development in pancreatic cancer.* (1424-3911 (Electronic)).
103. Cobo, I., et al., *Transcriptional regulation by NR5A2 links differentiation and inflammation in the pancreas.* (1476-4687 (Electronic)).
104. Zhang, M., et al., *Three new pancreatic cancer susceptibility signals identified on chromosomes 1q32.1, 5p15.33 and 8q24.21.* (1949-2553 (Electronic)).

105. Rothman, N., et al., *A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci.* (1546-1718 (Electronic)).
106. Bojesen, S.E., et al., *Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer.* (1546-1718 (Electronic)).
107. McKay, J.D., et al., *Lung cancer susceptibility locus at 5p15.33.* (1546-1718 (Electronic)).
108. Rafnar, T., et al., *Sequence variants at the TERT-CLPTM1L locus associate with many cancer types.* (1546-1718 (Electronic)).
109. Stacey, S.N., et al., *New common variants affecting susceptibility to basal cell carcinoma.* (1546-1718 (Electronic)).
110. Turnbull, C., et al., *Genome-wide association study identifies five new breast cancer susceptibility loci.* (1546-1718 (Electronic)).
111. Berndt, S.I., et al., *Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia.* (1546-1718 (Electronic)).
112. Zhao, Y., et al., *Fine-mapping of a region of chromosome 5p15.33 (TERT-CLPTM1L) suggests a novel locus in TERT and a CLPTM1L haplotype are associated with glioma susceptibility in a Chinese population.* (1097-0215 (Electronic)).
113. Cesare, A.J. and R.R. Reddel, *Alternative lengthening of telomeres: models, mechanisms and implications.* (1471-0064 (Electronic)).
114. Cheung, A.L. and W. Deng, *Telomere dysfunction, genome instability and cancer.* (1093-9946 (Print)).
115. Jafri, M.A., et al., *Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies.* (1756-994X (Electronic)).
116. Fang, J., et al., *Functional characterization of a multi-cancer risk locus on chr5p15.33 reveals regulation of TERT by ZNF148.* (2041-1723 (Electronic)).
117. Tang, J., et al., *CLPTM1L gene rs402710 (C > T) and rs401681 (C > T) polymorphisms associate with decreased cancer risk: a meta-analysis.* (1949-2553 (Electronic)).
118. Wolpin, B.M., et al., *Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer.* (1546-1718 (Electronic)).
119. Sanchez, Y., et al., *Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature.* (2041-1723 (Electronic)).
120. Li, L., et al., *Plasma and tumor levels of Linc-pint are diagnostic and prognostic biomarkers for pancreatic cancer.* (1949-2553 (Electronic)).
121. Klemke, R.L., et al., *CAS/Crk coupling serves as a "molecular switch" for induction of cell migration.* (0021-9525 (Print)).
122. Huang, M., et al., *EGFR-dependent pancreatic carcinoma cell metastasis through Rap1 activation.* (1476-5594 (Electronic)).
123. Stoffers, D.A., et al., *Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence.* (1061-4036 (Print)).
124. Gannon, M., et al., *pdx-1 function is specifically required in embryonic beta cells to generate appropriate numbers of endocrine cell types and maintain glucose homeostasis.* (1095-564X (Electronic)).
125. Roy, N., et al., *PDX1 dynamically regulates pancreatic ductal adenocarcinoma initiation and maintenance.* (1549-5477 (Electronic)).

126. Zebisch, M. and E.Y. Jones, *ZNRF3/RNF43--A direct linkage of extracellular recognition and E3 ligase activity to modulate cell surface signalling.* (1873-1732 (Electronic)).
127. Wu, J., et al., *Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways.* (1091-6490 (Electronic)).
128. Huppi, K., et al., *The 8q24 gene desert: an oasis of non-coding transcriptional activity.* (1664-8021 (Electronic)).
129. Cui, M., et al., *Long non-coding RNA PVT1 and cancer.* (1090-2104 (Electronic)).
130. Tseng, Y.Y., et al., *PVT1 dependence in cancer with MYC copy-number increase.* (1476-4687 (Electronic)).
131. Marchese, F.P. and M. Huarte, *A "Counter-Enhancer" in Tumor Suppression.* (1097-4172 (Electronic)).
132. Zhou, D.D., et al., *Long non-coding RNA PVT1: Emerging biomarker in digestive system cancer. LID - 10.1111/cpr.12398 [doi].* (1365-2184 (Electronic)).
133. Klein, A.P., et al., *Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer.* (2041-1723 (Electronic)).
134. Hublitz, P., et al., *NIR is a novel INHAT repressor that modulates the transcriptional activity of p53.* (0890-9369 (Print)).
135. Titus, A.J.A.-O.h.o.o., et al., *Deconvolution of DNA methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes.* (2045-2322 (Electronic)).
136. Heyne, K., et al., *NIR, an inhibitor of histone acetyltransferases, regulates transcription factor TAp63 and is controlled by the cell cycle.* (1362-4962 (Electronic)).
137. Hruban, R.H., et al., *Genetics of pancreatic cancer. From genes to families.* Surg Oncol Clin N Am, 1998. **7**(1): p. 1-23.
138. Childs, E.J., et al., *Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer.* (1546-1718 (Electronic)).
139. Li, D., et al., *Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer.* (1460-2180 (Electronic)).
140. Plengvidhya, N., et al., *Hepatocyte nuclear factor-4gamma: cDNA sequence, gene organization, and mutation screening in early-onset autosomal-dominant type 2 diabetes.* (0012-1797 (Print)).
141. Johnson, S.R.A.-O.h.o.o., et al., *Whole-exome sequencing for mutation detection in pediatric disorders of insulin secretion: Maturity onset diabetes of the young and congenital hyperinsulinism.* (1399-5448 (Electronic)).
142. Chen, B.D., et al., *TT genotype of rs2941484 in the human HNF4G gene is associated with hyperuricemia in Chinese Han men.* (1949-2553 (Electronic)).
143. Kottgen, A., et al., *Genome-wide association analyses identify 18 new loci associated with serum urate concentrations.* (1546-1718 (Electronic)).
144. Speliotes, E.K., et al., *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* (1546-1718 (Electronic)).
145. Ischia, J., et al., *Gastrin-releasing peptide: different forms, different functions.* (0951-6433 (Print)).

146. Petronilho, F., et al., *Gastrin-releasing peptide as a molecular target for inflammatory diseases: an update.* (2212-4055 (Electronic)).
147. Pendharkar, S.A., et al., *Gastrin-Releasing Peptide and Glucose Metabolism Following Pancreatitis.* (1918-2805 (Print)).
148. Petrov, M.S., *Diabetes of the exocrine pancreas: American Diabetes Association-compliant lexicon.* (1424-3911 (Electronic)).
149. Ewald, N., et al., *Prevalence of diabetes mellitus secondary to pancreatic diseases (type 3c).* (1520-7560 (Electronic)).
150. Carter, J.A., et al., *CpG dinucleotide-specific hypermethylation of the TNS3 gene promoter in human renal cell carcinoma.* (1559-2308 (Electronic)).
151. Kent, J.W., Jr., *Rare variants, common markers: synthetic association and beyond.* (1098-2272 (Electronic)).
152. Maher, M.C., et al., *Population genetics of rare variants and complex diseases.* (1423-0062 (Electronic)).
153. Lohmueller, K.E., et al., *Proportionally more deleterious genetic variation in European than in African populations.* (1476-4687 (Electronic)).
154. Li, Y., et al., *Low-coverage sequencing: implications for design of complex trait association studies.* (1549-5469 (Electronic)).
155. Gorlov, I.P., et al., *Evolutionary evidence of the effect of rare variants on disease etiology.* (1399-0004 (Electronic)).
156. Pasaniuc, B., et al., *Extremely low-coverage sequencing and imputation increases power for genome-wide association studies.* (1546-1718 (Electronic)).
157. Southam, L.A.-O.h.o.o., et al., *Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits.* (2041-1723 (Electronic)).
158. Gilly, A., et al., *Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation.* (1460-2083 (Electronic)).
159. Hatzikotoulas K Fau - Gilly, A., E. Gilly A Fau - Zeggini, and E. Zeggini, *Using population isolates in genetic association studies.* (2041-2657 (Electronic)).
160. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* (1471-0064 (Electronic)).
161. Do, R., G.R. Kathiresan S Fau - Abecasis, and G.R. Abecasis, *Exome sequencing and complex disease: practical aspects of rare variant association studies.* (1460-2083 (Electronic)).
162. Lee, S., X. Wu Mc Fau - Lin, and X. Lin, *Optimal tests for rare variant effects in sequencing association studies.* (1468-4357 (Electronic)).
163. Asimit, J.L., et al., *ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data.* (1423-0062 (Electronic)).
164. Nicolae, D.L., *Association Tests for Rare Variants.* (1545-293X (Electronic)).
165. Drichel, D., et al., *Rare variant testing of imputed data: an analysis pipeline typified.* (1423-0062 (Electronic)).
166. Derkach, A., L. Lawless Jf Fau - Sun, and L. Sun, *Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests.* (1098-2272 (Electronic)).

167. Frazer Ka Fau - Ballinger, D.G., et al., *A second generation human haplotype map of over 3.1 million SNPs.* (1476-4687 (Electronic)).
168. Auton A Fau - Brooks, L.D., et al., *A global reference for human genetic variation.* (1476-4687 (Electronic)).
169. Su, S.Y., et al., *Detection of identity by descent using next-generation whole genome sequencing data.* (1471-2105 (Electronic)).
170. Liu Q Fau - Cirulli, E.T., et al., *Systematic assessment of imputation performance using the 1000 Genomes reference panels.* (1477-4054 (Electronic)).
171. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* (1546-1718 (Electronic)).
172. Abecasis Gr Fau - Auton, A., et al., *An integrated map of genetic variation from 1,092 human genomes.* (1476-4687 (Electronic)).
173. Altshuler Dm Fau - Gibbs, R.A., et al., *Integrating common and rare genetic variation in diverse human populations.* (1476-4687 (Electronic)).
174. Campa, D., et al., *Genetic susceptibility to pancreatic cancer and its functional characterisation: the PANcreatic Disease ReseArch (PANDoRA) consortium.* (1878-3562 (Electronic)).
175. Urayama, K.Y., et al., *Body mass index and body size in early adulthood and risk of pancreatic cancer in a central European multicenter case-control study.* (1097-0215 (Electronic)).
176. Brune, K.A., et al., *Importance of age of onset in pancreatic cancer kindreds.* J Natl Cancer Inst, 2010. **102**(2): p. 119-26.
177. McWilliams, R.R., et al., *Polymorphisms in DNA repair genes, smoking, and pancreatic adenocarcinoma risk.* (1538-7445 (Electronic)).
178. Hassan, M.M., et al., *Risk factors for pancreatic cancer: case-control study.* (0002-9270 (Print)).
179. Olson, S.H., et al., *Allergies, variants in IL-4 and IL-4R alpha genes, and risk of pancreatic cancer.* (1525-1500 (Electronic)).
180. Eppel, A., S. Cotterchio M Fau - Gallinger, and S. Gallinger, *Allergies are associated with reduced pancreas cancer risk: A population-based case-control study in Ontario, Canada.* (1097-0215 (Electronic)).
181. Tran, B., et al., *Association between ultraviolet radiation, skin sun sensitivity and risk of pancreatic cancer.* (1877-783X (Electronic)).
182. Duell, E.J., et al., *Detecting pathway-based gene-gene and gene-environment interactions in pancreatic cancer.* (1055-9965 (Print)).
183. Risch, H.A., *Etiology of pancreatic cancer, with a hypothesis concerning the role of N-nitroso compounds and excess gastric acidity.* (1460-2105 (Electronic)).
184. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes.* (1546-1718 (Electronic)).
185. Tryka, K.A., et al., *NCBI's Database of Genotypes and Phenotypes: dbGaP.* (1362-4962 (Electronic)).
186. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* (0002-9297 (Print)).
187. Zheng, X., et al., *A high-performance computing toolset for relatedness and principal component analysis of SNP data.* (1367-4811 (Electronic)).

188. Delaneau O Fau - Zagury, J.-F., J. Zagury Jf Fau - Marchini, and J. Marchini, *Improved whole-chromosome phasing for disease and population genetic studies.* (1548-7105 (Electronic)).
189. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* (1471-0064 (Electronic)).
190. Willer, C.J., G.R. Li Y Fau - Abecasis, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* (1367-4811 (Electronic)).
191. Wright, J.A.-O., et al., *Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow.* (2041-1723 (Electronic)).
192. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* (1537-6605 (Electronic)).
193. Wu, C., et al., *Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations.* (1546-1718 (Electronic)).
194. Low, S.K., et al., *Genome-wide association study of pancreatic cancer in Japanese population.* (1932-6203 (Electronic)).
195. Ward, L.D. and M. Kellis, *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.* (1362-4962 (Electronic)).
196. Hoskins, J.W., et al., *Transcriptome analysis of pancreatic cancer reveals a tumor suppressor function for HNF1A.* (1460-2180 (Electronic)).
197. Voight, B.F., et al., *Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.* (1546-1718 (Electronic)).
198. Hegele, R.A., et al., *The hepatic nuclear factor-1alpha G319S variant is associated with early-onset type 2 diabetes in Canadian Oji-Cree.* (0021-972X (Print)).
199. Yamagata, K., et al., *Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3).* (0028-0836 (Print)).
200. Pierce, B.L. and H. Ahsan, *Genome-wide "pleiotropy scan" identifies HNF1A region as a novel pancreatic cancer susceptibility locus.* (1538-7445 (Electronic)).
201. Bergholz, J. and Z.X. Xiao, *Role of p63 in Development, Tumorigenesis and Cancer Progression.* (1875-2284 (Electronic)).
202. Flores, E.R., et al., *Tumor predisposition in mice mutant for p63 and p73: evidence for broader tumor suppressor functions for the p53 family.* (1535-6108 (Print)).
203. Melino, G., *p63 is a suppressor of tumorigenesis and metastasis interacting with mutant p53.* (1476-5403 (Electronic)).
204. Danilov, A.V., et al., *DeltaNp63alpha-mediated induction of epidermal growth factor receptor promotes pancreatic cancer cell growth and chemoresistance.* (1932-6203 (Electronic)).
205. Figueroa, J.D., et al., *Genome-wide association study identifies multiple loci associated with bladder cancer risk.* (1460-2083 (Electronic)).
206. Shiraishi, K., et al., *A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population.* (1546-1718 (Electronic)).
207. Miki, D., et al., *Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations.* (1546-1718 (Electronic)).
208. Lan, Q., et al., *Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia.* (1546-1718 (Electronic)).

209. Borowski, A., et al., *Structure and function of ETAA16: a novel cell surface antigen in Ewing's tumours.* (0340-7004 (Print)).
210. Marlaire, S., M. Van Schaftingen E Fau - Veiga-da-Cunha, and M. Veiga-da-Cunha, *C7orf10 encodes succinate-hydroxymethylglutarate CoA-transferase, the enzyme that converts glutarate to glutaryl-CoA.* (1573-2665 (Electronic)).
211. Son, J., et al., *Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway.* (1476-4687 (Electronic)).
212. Avis, I., et al., *Effect of gastrin-releasing peptide on the pancreatic tumor cell line (Capan).* (0899-1987 (Print)).
213. Jiang, H., et al., *Expression of Gli1 and Wnt2B correlates with progression and clinical outcome of pancreatic cancer.* (1936-2625 (Electronic)).
214. Yamagata, K., et al., *Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1).* (0028-0836 (Print)).
215. Fuchs, C.S., et al., *A prospective study of cigarette smoking and the risk of pancreatic cancer.* (0003-9926 (Print)).
216. Jang, J.H., et al., *Genetic variants in carcinogen-metabolizing enzymes, cigarette smoking and pancreatic cancer risk.* (1460-2180 (Electronic)).
217. Talamini, R., et al., *Tobacco smoking, alcohol consumption and pancreatic cancer risk: a case-control study in Italy.* (1879-0852 (Electronic)).
218. Rorbach, J., M. Nicholls Tj Fau - Minczuk, and M. Minczuk, *PDE12 removes mitochondrial RNA poly(A) tails and controls translation in human mitochondria.* (1362-4962 (Electronic)).
219. Wood, E.R., et al., *The Role of Phosphodiesterase 12 (PDE12) as a Negative Regulator of the Innate Immune Response and the Discovery of Antiviral Inhibitors.* (1083-351X (Electronic)).
220. Shinohara, A. and T. Ogawa, *Stimulation by Rad52 of yeast Rad51-mediated recombination.* (0028-0836 (Print)).
221. Moynahan, M.E. and M. Jasin, *Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis.* (1471-0080 (Electronic)).
222. Feng, Z., et al., *Rad52 inactivation is synthetically lethal with BRCA2 deficiency.* (1091-6490 (Electronic)).
223. Lok, B.H., et al., *RAD52 inactivation is synthetically lethal with deficiencies in BRCA1 and PALB2 in addition to BRCA2 through RAD51-mediated homologous recombination.* (1476-5594 (Electronic)).
224. Hu, C., et al., *Association Between Inherited Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer.* (1538-3598 (Electronic)).
225. Takeuchi, S., et al., *Mutations in BRCA1, BRCA2, and PALB2, and a panel of 50 cancer-associated genes in pancreatic ductal adenocarcinoma.* (2045-2322 (Electronic)).
226. Nagathihalli, N.S. and G. Nagaraju, *RAD51 as a potential biomarker and therapeutic target for pancreatic cancer.* (0006-3002 (Print)).
227. Cramer-Morales, K., et al., *Personalized synthetic lethality induced by targeting RAD52 in leukemias identified by gene mutation and expression profile.* (1528-0020 (Electronic)).

8. ANNEXES

ANNEX 1

Includes Supplementary Tables and Figures of the Two-Stage GWAS

ANNEX 2

Includes Supplementary Tables and Figures of the Exome-array gene-based analysis

ANNEX 1

Supplementary Table 1.1. Association results for ten loci previously identified in pancreatic cancer GWASs

Chr	Gene[a]	SNP	Position[b]	Study[c]	Reported OR (CI)[d]	Reported P---value[e]	PanC4 OR (CI)[f]	PanC4 P---value[g]
9q34	*ABO*	rs505922	136149229	PanScan 1	1.2 (1.12 --- 1.28)	5.37×10^{-8}	1.27 (1.19 --- 1.35)	1.72×10^{-13}
13q22.1	*KLF5 AND KLF12*	rs9543325	73916628	PanScan 2	1.26 (1.18 --- 1.35)	3.27×10^{-11}	1.24 (1.16 --- 1.32)	2.26×10^{-10}
1q32.1	*NR5A2*	rs3790844	200007432	PanScan 2	0.77 (0.71 --- 0.84)	2.45×10^{-10}	0.83 (0.77 --- 0.90)	3.05×10^{-6}
5p15.33	*CLPTM1L*	rs401681	1322087	PanScan 2	1.19 (1.11 --- 1.27)	3.66×10^{-7}	1.20 (1.13 ---1.28)	2.70×10^{-8}
5p15.33	*TERT*	rs2736098	1294086	PanScan 3	0.80 (0.76 --- 0.85)	9.78×10^{-14}	0.85 (0.78 --- 0.93)	2.31×10^{-5}
7q32.3	*LINC---PINT*	rs6971499	130680521	PanScan 3	0.79 (0.74 --- 0.84)	2.98×10^{-12}	0.81 (0.74 --- 0.88)	7.10×10^{-6}
16q23.1	*BCAR1*	rs7190458	75263661	PanScan 3	1.46 (1.3 --- 1.65)	1.13×10^{-10}	1.40 (1.22 --- 1.60)	1.01×10^{-4}
13q12.2	*PDX1*	rs9581943	28493997	PanScan 3	1.15 (1.1 --- 1.2)	2.35×10^{-9}	1.17 (1.10 --- 1.24)	1.94×10^{-7}
22q12.1	*ZNRF3*	rs16986825	29300306	PanScan 3	1.18 (1.12 --- 1.25)	1.18×10^{-8}	1.14 (1.04 --- 1.24)	2.72×10^{-3}
8q24.21	*MIR1208*	rs1561927	129568078	PanScan 3	0.87 (0.83 --- 0.92)	1.3×10^{-7}	0.92 (0.85--- 0.99)	2.20×10^{-2}
2p13.3	*ETAA1*	rs2035565	67,619,656	China	1.33 (1.19 --- 1.49)	5.46×10^{-7}	1.15 (1.07 --- 1.25)	2.69×10^{-4}

a. RefSeq Gene symbol of the closest gene to the listed SNP. For SNPs not intragenic to the listed gene, the gene is listed in grey.
b. Position of the SNP according to NCBI Human Genome Build 37
c. GWAS where the SNP was first found to be associated with Pancreatic Cancer
d. Odds ratio and Confidence Interval for the SNP in publication listed in the Study column
e. P---value listed in the original study
f. OR and CI from a test for association of this SNP with pancreatic cancer in the PanC4 Study
g. P---value from a test for association of this SNP with pancreatic cancer in the PanC4 Study

Supplementary Table 1.2. Association results for 25 SNPs in Stages 1 and 2

			Stage 1					Stage 2			
Chr[a] SNP Position[b] Gene	Effect Allele (Minor)/ Reference Allele	Statistic	PanC4	PanScan 1	PanScan 2	All Stage 1[c]	Heterogeneity[d]	PANDoRA	Heterogeneity[e]	Combined Stage 1&2[f]	Heterogeneity[g]
17q25.1		maf[h] cases;controls	0.146;0.11	0.139;0.129	0.149;0.116			0.135; 0.114			
rs11655237	T/C	info[i]	0.963	g	g						
70400166		OR (CI)[j]	1.38 (1.26 -- 1.52)	1.09 (0.96 -- 1.25)	1.34 (1.16 -- 1.55)	1.27 (1.19 -- 1.36)		1.24 (1.1 -- 1.4)		1.26 (1.19 -- 1.34)	
LINC00673		p---value	1.38E---10	1.95E---01	2.95E---04	6.74E---12	2.29E---02	6.40E---04	2.49E---01	1.42E---14	7.34E---02
17q25.1		maf cases;controls	0.148;0.112	0.140;0.133	0.150;0.117			0.139; 0.117			
rs7214041	T/C	info	g	0.966	0.96						
70401476		OR (CI)	1.38 (1.26 -- 1.51)	1.07 (0.93 -- 1.22)	1.33 (1.15 -- 1.53)	1.26 (1.18 -- 1.35)		1.25 (1.11 -- 1.41)		1.26 (1.19 -- 1.34)	
LINC00673		p---value	1.95E---10	3.36E---01	3.69E---04	2.67E---11	1.46E---02	3.37E---04	3.69E---01	2.88E---14	8.52E---02
13q12.2		maf cases;controls	0.440;0.399	0.428;0.416	0.437;0.405			0.450; 0.424			
rs9554197	T/C	info	g	g	g						
28476978		OR (CI)	1.18 (1.11 -- 1.26)	1.05 (0.96 -- 1.15)	1.14 (1.03 -- 1.26)	1.15 (1.1 -- 1.21)		1.11 (1.03 -- 1.2)		1.14 (1.1 -- 1.19)	
PDX1---AS1		p---value	4.22E---08	3.20E---01	1.05E---03	1.93E---09	6.25E---02	1.03E---02	8.35E---01	8.89E---11	4.12E---01
2p13.3		maf cases;controls	0.302;0.275	0.305;0.292	0.305;0.276			0.292; 0.273			
rs1486134	G/T	info	g	g	g						
67639769		OR (CI)	1.14 (1.06 -- 1.22)	1.06 (0.96 -- 1.18)	1.15 (1.03 -- 1.28)	1.13 (1.08 -- 1.19)		1.16 (1.06 -- 1.27)		1.14 (1.09 -- 1.19)	
ETAA1 (2236 bp on 3')		p---value	5.96E---05	1.57E---01	5.18E---03	8.35E---07	4.54E---01	9.42E---04	3.32E---02	3.36E---09	8.32E---02
2p13.3		maf cases;controls	0.407;0.379	0.401;0.391	0.406;0.371			0.386; 0.31			
rs962856	C/T	info	g	g	g						
67593803		OR (CI)	1.12 (1.06 ---1.20)	1.04 (0.95 -- 1.14)	1.16 (1.05 -- 1.28)	1.12 (1.07 -- 1.18)		1.12 (1.03 -- 1.22)		1.12 (1.08 -- 1.17)	
		p---value	4.58E---05	3.40E---01	1.37E---03	8.30E---07	1.82E---01	8.05E---03	1.87E---01	2.21E---08	2.08E---01
7p13		maf cases;controls	0.241;0.263	0.218;0.254	0.237;0.268			0.254; 0.277			
rs17688601	A/C	info	g	g	g						
40866663		OR (CI)	0.89 (0.83 -- 0.96)	0.82 (0.73 -- 0.91)	0.85 (0.76 -- 0.94)	0.87 (0.82 -- 0.91)		0.91 (0.83 -- 1)		0.88 (0.84 -- 0.92)	
SUGCT		p---value	1.98E---03	1.66E---04	8.72E---03	9.77E---08	3.99E---01	3.93E---02	7.25E---02	1.41E---08	1.23E---01
3q29		maf cases;controls	0.328;0.362	0.336;0.366	0.325;0.356			0.341; 0.356			
rs9854771	A/G	info	g	0.998	0.998						
189508471		OR (CI)	0.86 (0.81 -- 0.92)	0.88 (0.80 -- 0.97)	0.87 (0.79 -- 0.97)	0.87 (0.83 -- 0.92)		0.93 (0.86 -- 1.01)		0.89 (0.85 -- 0.93)	
TP63		p---value	3.10E---05	7.94E---03	1.55E---02	4.08E---08	9.71E---01	1.01E---01	8.15E---01	2.35E---08	8.61E---01
18q21.2		maf cases;controls	0.172;0.182	0.170;0.195	0.164;0.189			0.168;0.187			
rs1517037	T/C	info	g	g	g						
56878274		OR (CI)	0.93 (0.86 -- 1.01)	0.85 (0.75 -- 0.95)	0.84 (0.74 -- 0.95)	0.87 (0.82 -- 0.93)		0.87 (0.79 -- 0.97)		0.87 (0.83 -- 0.92)	
GRP (9126 bp on 5')		p---value	3.39E---02	3.82E---03	1.82E---03	9.93E---06	2.56E---01	1.17E---02	7.73E---02	3.17E---07	1.24E---01
12q24.31		maf cases;controls	0.407;0.386	0.423;0.375	0.421;0.392			0.426;0.415			
rs7310409	A/G	info	g	g	g						
121424861		OR (CI)	1.09 (1.02 -- 1.16)	1.22 (1.11 -- 1.34)	1.13 (1.02 -- 1.24)	1.12 (1.07 -- 1.18)		1.07 (0.98 -- 1.16)		1.11 (1.06 -- 1.15)	
HNF1A		p---value	1.80E---02	5.35E---05	1.76E---02	1.24E---06	1.45E---01	1.26E---01	9.95E---01	6.34E---07	7.08E---01
1p13.1		maf cases;controls	0.228;0.257	0.24;0.254	0.239;0.258			0.229;0.248			
rs351365	T/C	info	g	0.891	0.889						
113046395		OR (CI)	0.85 (0.79 -- 0.92)	0.93 (0.84 -- 1.03)	0.91 (0.81 -- 1.01)	0.88 (0.83 -- 0.93)		0.92 (0.83 -- 1.01)		0.89 (0.85 -- 0.93)	
WNT2B		p---value	3.08E---05	1.62E---01	4.45E---02	2.72E---06	5.21E---01	8.02E---02	3.97E---01	7.39E---07	5.33E---01
20q13.11		maf cases;controls	0.415;0.381	0.429;0.411	0.411;0.384			0.413;0.403			
rs6073450	A/G	info	g	g	g						
43086648		OR (CI)	1.15 (1.08 -- 1.23)	1.08 (0.98 -- 1.18)	1.12 (1.02 -- 1.24)	1.12 (1.06 -- 1.17)		1.09 (1 -- 1.18)		1.11 (1.06 -- 1.15)	
		p---value	1.82E---04	1.20E---01	2.62E---02	6.01E---06	6.79E---01	4.92E---02	9.03E---01	9.21E---07	9.52E---01
17q25.1		maf cases;controls	0.312;0.282	0.292;0.281	0.317;0.29			0.289;0.272			
rs11652288	G/A	info	g	0.892	0.802						
70405340		OR (CI)	1.15 (1.08 -- 1.23)	1.05 (0.95 -- 1.16)	1.14 (1.02 -- 1.26)	1.13 (1.07 -- 1.18)		1.07 (0.98 -- 1.17)		1.11 (1.06 -- 1.16)	
LINC00673		p---value	1.04E---04	2.42E---01	2.56E---02	8.90E---06	4.98E---01	1.52E---01	5.59E---01	4.53E---06	6.11E---01
1p31.3		maf cases;controls	0.222;0.233	0.207;0.233	0.220;0.243			0.230;0.237			
rs1747924	C/A	info[g]	g	g	g						
64538961		OR (CI)[h]	0.94 (0.87 -- 1.01)	0.86 (0.77 -- 0.96)	0.88 (0.78 -- 0.98)	0.88 (0.84 -- 0.93)		0.94 (0.85 -- 1.03)		0.9 (0.86 -- 0.94)	
ROR1		p---value	2.38E---02	4.06E---03	3.13E---03	9.16E---06	3.33E---01	1.94E---01	1.94E---01	6.74E---06	2.12E---01
21q11.2		maf cases;controls	0.494;0.464	0.234;0.227	0.247;0.209			0.253;0.252			
rs1822660	A/G	info	0.619	g	g						
22658894		OR (CI)	1.13 (1.06 ---1.20)	1.04 (0.93 ---1.16)	1.24 (1.11 -- 1.40)	1.15 (1.09 -- 1.22)		1.02 (0.93 -- 1.11)		1.11 (1.06 -- 1.17)	
NCAM2		p---value	2.19E---05	5.61E---01	6.53E---04	8.86E---07	6.72E---02	7.41E---01	9.29E---02	1.08E---05	1.01E---02
4q26		maf cases;controls	0.111;0.098	0.117;0.099	0.117;0.088			0.117;0.116			
rs1380376	G/A	info	0.999	g	g						
118636270		OR (CI)	1.15 (1.03 -- 1.27)	1.20 (1.04 -- 1.39)	1.38 (1.17 -- 1.62)	1.19 (1.11 -- 1.28)		1.05 (0.93 -- 1.2)		1.15 (1.08 -- 1.23)	
		p---value	2.35E---02	9.52E---03	3.64E---04	4.38E---06	1.91E---01	4.04E---01	1.46E---01	1.22E---05	7.74E---02
9q31.3		maf cases;controls	0.407;0.366	0.391;0.371	0.390;0.388			0.365;0.364			
rs10991043	C/T	info	g	g	g						
106797388		OR (CI)	1.19 (1.12 -- 1.27)	1.09 (0.99 -- 1.2)	1 (0.91 -- 1.11)	1.13 (1.08 -- 1.18)		1 (0.92 -- 1.08)		1.1 (1.05 -- 1.14)	
		p---value	7.00E---08	5.18E---02	7.09E---01	5.10E---07	2.56E---02	9.19E---01	3.02E---01	1.35E---05	1.05E---02
15q22.32		maf cases;controls	0.376;0.354	0.383;0.357	0.400;0.356			0.352;0.341			
rs8024986	C/T	info	g	0.999	0.999						
73024868		OR (CI)	1.10 (1.03 -- 1.17)	1.12 (1.02 -- 1.23)	1.21 (1.09 -- 1.33)	1.12 (1.06 -- 1.17)		1.04 (0.95 -- 1.13)		1.1 (1.05 -- 1.14)	
BBS4		p---value	1.25E---02	9.68E---03	2.30E---03	6.30E---06	1.00E+00	4.13E---01	2.28E---01	1.37E---05	2.11E---01
6q25.2		maf cases;controls	0.300;0.268	0.288;0.261	0.290;0.285			0.286;0.288			
rs7762516	C/T	info	g	0.776	0.744						
154806742		OR (CI)	1.17 (1.09 -- 1.25)	1.15 (1.04 -- 1.27)	1.02 (0.92 -- 1.14)	1.14 (1.08 -- 1.2)		1.02 (0.94 -- 1.12)		1.11 (1.06 -- 1.16)	
CNKSR3		p---value	1.90E---05	3.17E---03	8.81E---01	2.21E---06	1.04E---01	6.29E---01	2.11E---01	1.82E---05	4.34E---02
1p22.3		maf cases;controls	0.448;0.481	0.472;0.493	0.449;0.474			0.450;0.452			
rs318405	G/A	info	g	g	g						
83014374		OR (CI)	0.88 (0.82 -- 0.93)	0.92 (0.84 -- 1)	0.91 (0.82 -- 1)	0.9 (0.86 -- 0.94)		0.99 (0.91 -- 1.07)		0.92 (0.88 -- 0.95)	
		p---value	5.69E---05	5.69E---02	6.22E---02	2.15E---06	6.88E---01	8.10E---01	2.66E---02	2.19E---05	2.13E---02
12q24.21		maf cases;controls	0.436;0.408	0.439;0.414	0.452;0.42			0.422;0.408			
rs10850078	G/A	info	g	g	g						
113209519		OR (CI)	1.12 (1.05 -- 1.2)	1.11 (1.01 -- 1.21)	1.14 (1.03 -- 1.25)	1.11 (1.06 -- 1.17)		1.02 (0.94 -- 1.11)		1.09 (1.05 -- 1.13)	
		p---value	1.56E---03	1.63E---02	2.85E---02	6.29E---06	9.80E---01	6.26E---01	1.79E---01	2.98E---05	2.11E---01
1q42.2		maf cases;controls	0.484;0.465	0.491;0.465	0.480;0.465			0.469; 0.455			
rs1317764	G/T	info	g	g	g						
230877400		OR (CI)	1.08 (1.01 -- 1.15)	1.11 (1.01 -- 1.21)	1.06 (0.96 -- 1.17)	1.12 (1.07 -- 1.17)		0.98 (0.9 -- 1.06)		1.09 (1.04 -- 1.13)	
		p---value	1.79E---04	2.37E---02	2.88E---02	1.12E---06	9.62E---01	6.53E---01	6.43E---01	5.72E---05	1.87E---01
5q35.2		maf cases;controls	0.208;0.224	0.199;0.213	0.204;0.24			0.210;0.211			
rs6894235	A/G	info	g	g	g						
173285545		OR (CI)	0.91 (0.84 -- 0.98)	0.92 (0.82 -- 1.03)	0.81 (0.72 -- 0.91)	0.88 (0.83 -- 0.93)		1.01 (0.91 -- 1.11)		0.91 (0.87 -- 0.95)	
		p---value	8.44E---03	1.27E---01	1.29E---04	7.23E---06	1.64E---01	9.20E---01	5.39E---02	1.10E---04	1.14E---02
10q26.3		maf cases;controls	0.101;0.085	0.1;0.088	0.114;0.091			0.082;0.086			
rs7086546	C/T	info	g	g	g						
130347966		OR (CI)	1.21 (1.09 -- 1.35)	1.16 (0.99 -- 1.36)	1.3 (1.11 -- 1.52)	1.2 (1.11 -- 1.29)		0.88 (0.76 -- 1.02)		1.12 (1.04 -- 1.2)	
		p---value	2.37E---03	5.63E---02	5.82E---03	8.23E---06	7.82E---01	9.31E---02	6.14E---01	1.84E---03	2.94E---02
8q22.3		maf cases;controls	0.431;0.391	0.383;0.397	0.420;0.420			0.408;0.403			
rs16867971	C/T	info	g	0.99	0.99						
102609456		OR (CI)	1.18 (1.11 -- 1.26)	0.94 (0.86 -- 1.03)	1 (0.91 -- 1.1)	1.05 (1.01 -- 1.1)		1.06 (0.98 -- 1.16)		1.06 (1.01 -- 1.1)	
GRHL2		p---value	6.85E---06	2.13E---01	4.12E---01	2.64E---02	2.52E---04	1.47E---01	7.04E---01	8.15E---03	8.92E---03
2p11.2		maf cases;controls	0.440;0.478	0.471;0.445	0.474;0.457			0.465;0.459			
rs6706539	A/G	info	g	0.773	0.758						
84381680		OR (CI)	0.86 (0.80 -- 0.91)	1.11 (1.01 -- 1.22)	1.07 (0.97 -- 1.18)	0.96 (0.92 -- 1.01)		1.06 (0.98 -- 1.15)		0.99 (0.95 -- 1.03)	
		p---value	3.44E---06	8.63E---03	4.71E---02	1.29E---01	3.05E---07	1.52E---01	1.68E---01	5.69E---01	1.46E---06

a Cytogenetic regions according to NCBI Human Genome Build 37
b SNP position according to NCBI Human Genome Build 37
c Results from the meta---analysis of Stage 1: PanC4, PanScan 1, and PanScan 2
d P---value from test of heterogeneity of the Stage 1 studies (PanC4, PanScan 1, and PanScan 2)
e P---value from test of heterogeneity of the PANDoRA countries
f Results from the meta---analysis of Stage1 and Stage 2: PanC4, PanScan 1, PanScan 2, and PANDoRA
g P---value from test of heterogeneity of the Stage 1 and Stage 2 studies (PanC4, PanScan 1, PanScan 2, and PANDoRA)
h Minor allele frequency

i Quality of imputation metric. See online methods for more detail. If snp is genotyped and not imputed, a 'g' is reported
j Allelic Odds Ratio and corresponding 95% Confidence Interval

Supplementary Table 1.3. Association results for loci highly suggestive ($P<1x10^{-6}$) for pancreatic cancer

			Stage 1				Stage 2	
Chr[a] SNP Position[b] Gene	**Effect Allele (Minor)/ Reference Allele**	**Statistic**	**PanC4** 4,164 cases 3,792 controls	**PanScan 1** 1,856 cases, 1,890 controls	**PanScan 2** 1,618 cases and 1,682 controls	**Combined Stage 1[c]** 7,638 cases 7,364 controls	**PANDoRA** 2497 cases 4611 controls	**Combined Stage 1&2[d]** 9,925 cases 11,569 controls
18q21.2 rs1517037 56,878,274 *GRP* (9126bp 5')	T/C	maf cases;controls	0.172; 0.182	0.170; 0.195	0.164; 0.189		0.168; 0.187	
		info	g	g	g			
		OR (CI)	0.93 (0.86 --- 1.01)	0.85 (0.75 --- 0.95)	0.84 (0.74 --- 0.95)	0.87 (0.82 --- 0.93)	0.87 (0.79 --- 0.97)	0.87 (0.83 --- 0.92)
		p---value	$3.39 X 10^{---2}$	$3.82 X 10^{---3}$	$1.82 X 10^{---3}$	$9.93 X 10^{---6}$	$1.17 X 10^{---2}$	$3.17 X 10^{---7}$
12q24.31 rs7310409 121,424,861 *HNF1A*	A/G	maf cases;controls	0.407; 0.386	0.423; 0.375	0.421; 0.392		0.426; 0.415	
		info	g	g	g			
		OR (CI)	1.09 (1.02 --- 1.16)	1.22 (1.11 --- 1.34)	1.13 (1.02 --- 1.24)	1.12 (1.07 --- 1.18)	1.07 (0.98 --- 1.16)	1.11 (1.06 --- 1.15)
		p---value	$1.80 X 10^{---2}$	$5.35 X 10^{---5}$	$1.76 X 10^{---2}$	$1.24 X 10^{---6}$	$1.26 X 10^{---1}$	$6.34 X 10^{---7}$
1p13.1 rs351365 113,046,395 *WNT2B*	T/C	maf cases;controls	0.228; 0.257	0.240; 0.254	0.239; 0.258		0.229; 0.248	
		info	g	0.891	0.889			
		OR (CI)	0.85 (0.79 --- 0.92)	0.93 (0.84 --- 1.03)	0.91 (0.81 --- 1.01)	0.88 (0.83 --- 0.93)	0.92 (0.83 --- 1.01)	0.89 (0.85 --- 0.93)
		p---value	$3.08 X 10^{---5}$	$1.62 X 10^{---1}$	$4.45 X 10^{---2}$	$2.72 X 10^{---6}$	$8.02 X 10^{---2}$	$7.39 X 10^{---7}$
20q13.11 rs6073450 43,086,648	A/G	maf cases;controls	0.415; 0.381	0.429; 0.411	0.411; 0.384		0.413; 0.403	
		info	g	g	g			
		OR (CI)	1.15 (1.08 --- 1.23)	1.08 (0.98 --- 1.18)	1.12 (1.02 --- 1.24)	1.12 (1.06 --- 1.17)	1.09 (1.00 --- 1.18)	1.11 (1.06 --- 1.15)
		p---value	$1.82 X 10^{---4}$	$1.20 X 10^{---1}$	$2.62 X 10^{---2}$	$6.01 X 10^{---6}$	$4.92 X 10^{---2}$	$9.21 X 10^{---7}$

9q31.3 rs6073450 106797388	C/T	maf cases;controls	0.407; 0.366	0.391; 0.371	0.390; 0.388		0.365; 0.364	
		info	g	g	g			
		OR (CI)	1.19 (1.12 --- 1.27)	1.09 (0.99 --- 1.2)	1.00 (0.91 --- 1.11)	1.13 (1.08 --- 1.18)	1.00 (0.92 --- 1.08)	1.10 (1.05 --- 1.14)
		p---value	$7.00 \times 10^{---8}$	$5.18 \times 10^{---2}$	$7.09 \times 10^{---1}$	$5.10 \times 10^{---7}$	$9.19 \times 10^{---1}$	$1.35 \times 10^{---5}$

[a] Cytogenetic regions according to NCBI Human Genome Build 37 and NCBI's Map Viewer
[b] SNP position according to NCBI Human Genome Build 37
[c] Results from the Combined Stage 1 meta---analysis of PanC4, PanScan 1, and PanScan 2
[d] Results from the Combined Stage 1 and 2 meta---analysis of PanC4, PanScan 1, PanScan 2, and PANDoRA
[e] MAF--- minor allele frequency
[f] Quality of imputation metric. See online methods for more detail. If snp is genotyped and not imputed, a 'g' is reported
[g] Allelic Odds Ratio and corresponding 95% Confidence Interval
[h] $R^2>0.9$

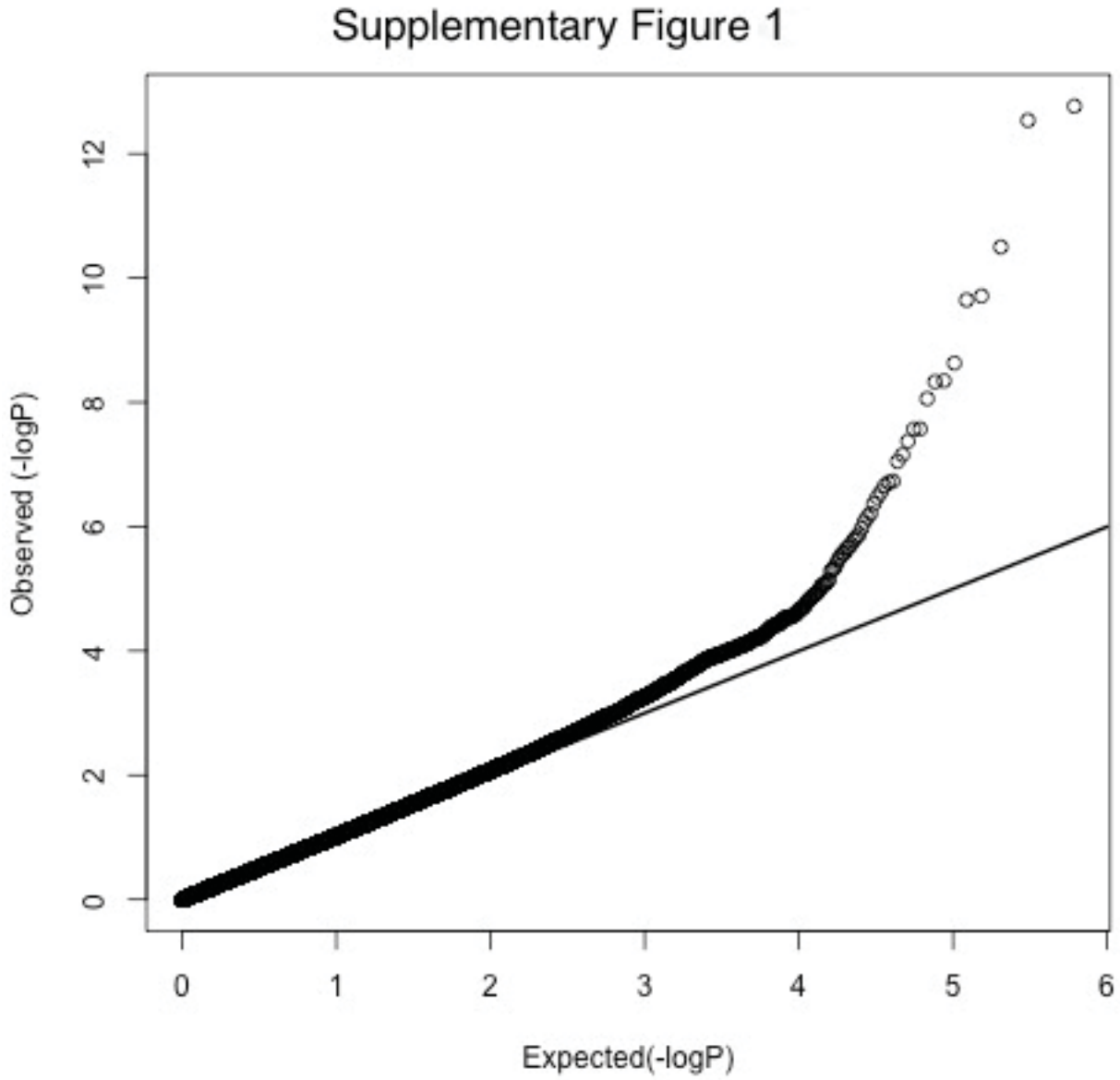
Supplementary Figure 1.1. Quantile-Quantile (Q-Q) plot of association results fro m PanC4 analysis

Supplementary Figure 1.2. Quantile-Quantile (Q-Q) plot of association results fro m the Combined Stage 1 analysis

Supplementary Figure 1.3. Regional association and linkage disequilibrium (LD) plots for five suggestive loci: (a) 12q24.31 (b) 18q21.2, (c) 1p13.1, (d) 20q13.11, and (e) 9q31.3. Association p-values are shown for three analyses: PanC4 only (black circles), Combined Stage 1 (PanC4, PanScan 1, and PanScan 2) (grey circles), and Combined Stage 1 and 2 (PanC4, PanScan 1, PanScan 2, and PANDoRA) (red circles). LD plots are based on 1000 Genomes European samples

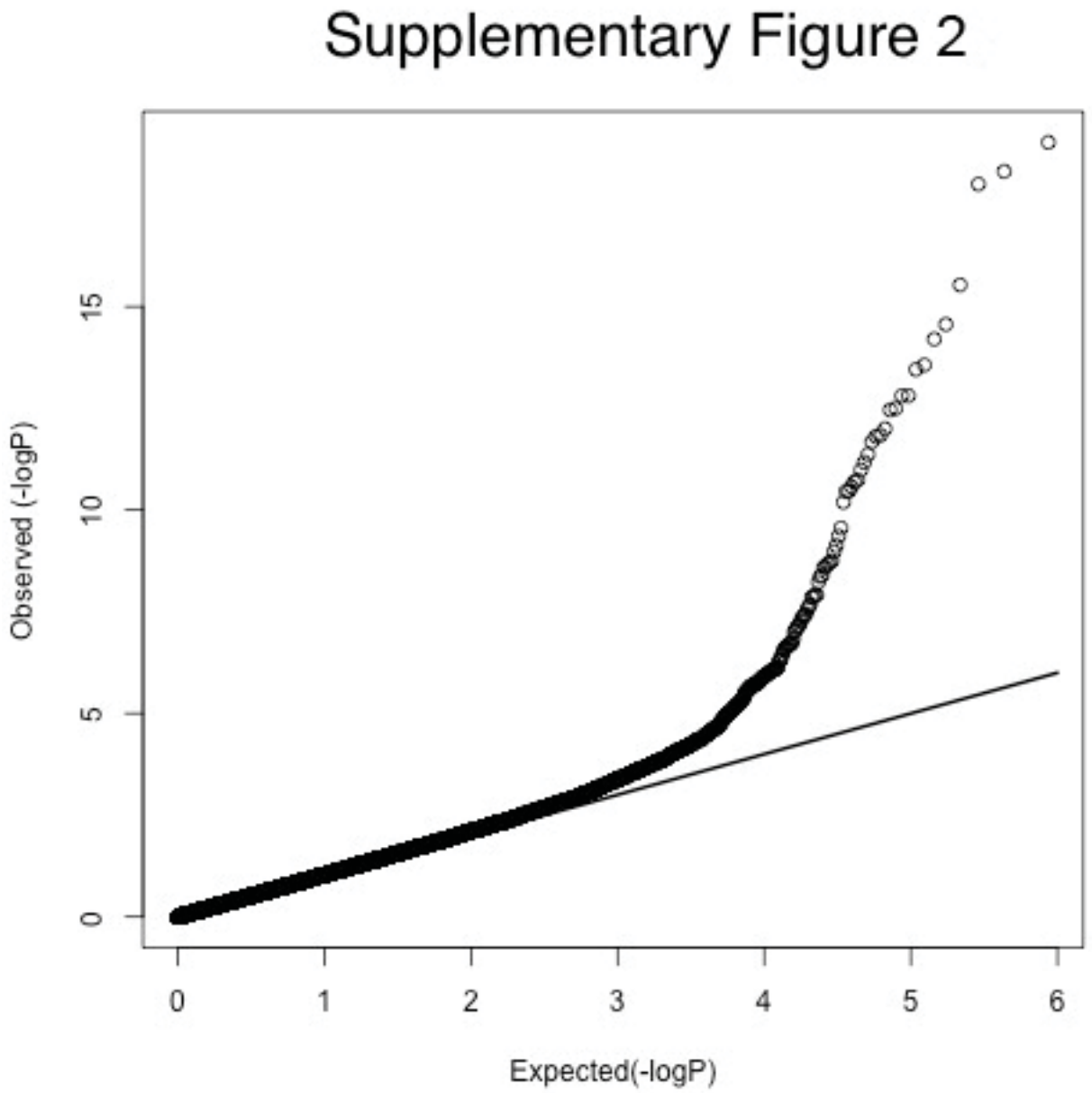
Supplementary Figure 1.4a-4i. Forest plots for the nine top associations reported

Supplementary Figure 1.1. Quantile-Quantile (Q-Q) plot of association results fro m PanC4 analysis



Nature Genetics: doi:10.1038/ng.3341

Supplementary Figure 1.2. Quantile-Quantile (Q-Q) plot of association results fro m the Combined Stage 1 analysis

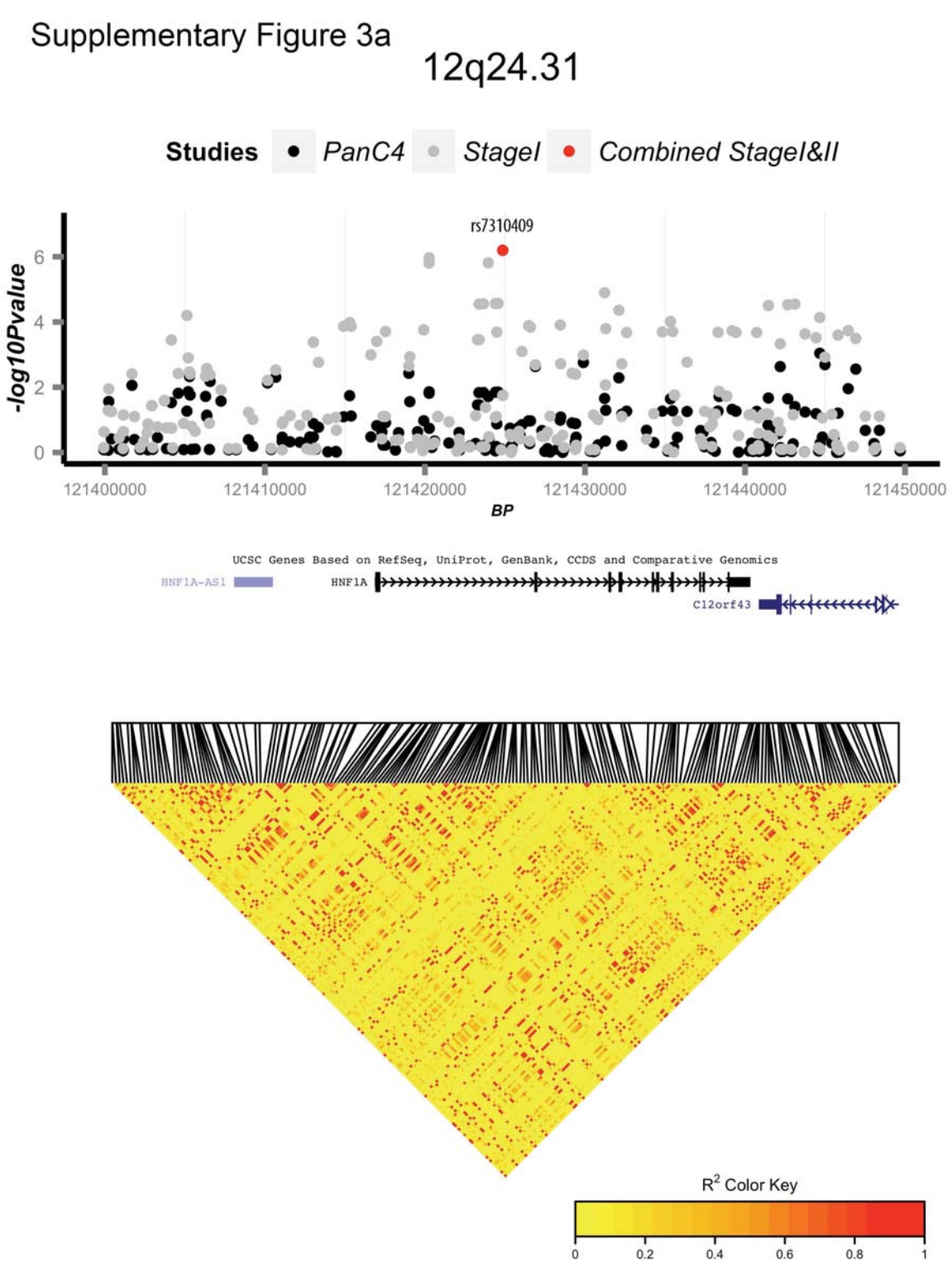


Nature Genetics: doi:10.1038/ng.3341

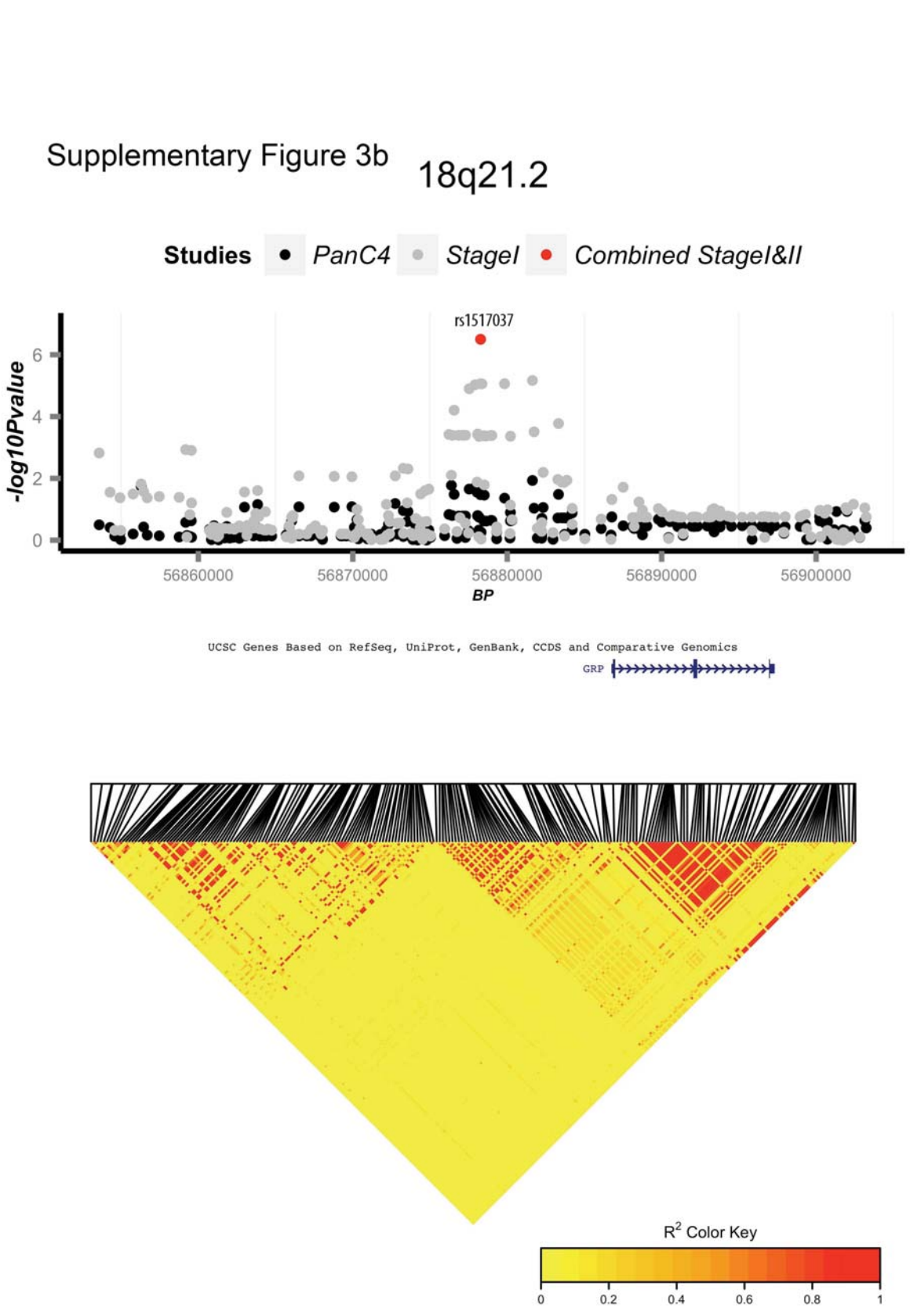
Supplementary Figure 1.3. Regional association and linkage disequilibrium (LD) plots for five suggestive loci: (a) 12q24.31 (b) 18q21.2, (c) 1p13.1, (d) 20q13.11, and (e) 9q31.3. Association p-values are shown for three analyses: PanC4 only (black circles), Combined Stage 1 (PanC4, PanScan 1, and PanScan 2) (grey circles), and Combined Stage 1 and 2 (PanC4, PanScan 1, PanScan 2, and PANDoRA) (red circles). LD plots are based on 1000 Genomes European samples

Nature Genetics: doi:10.1038/ng.3341

Supplementary Figure 3a

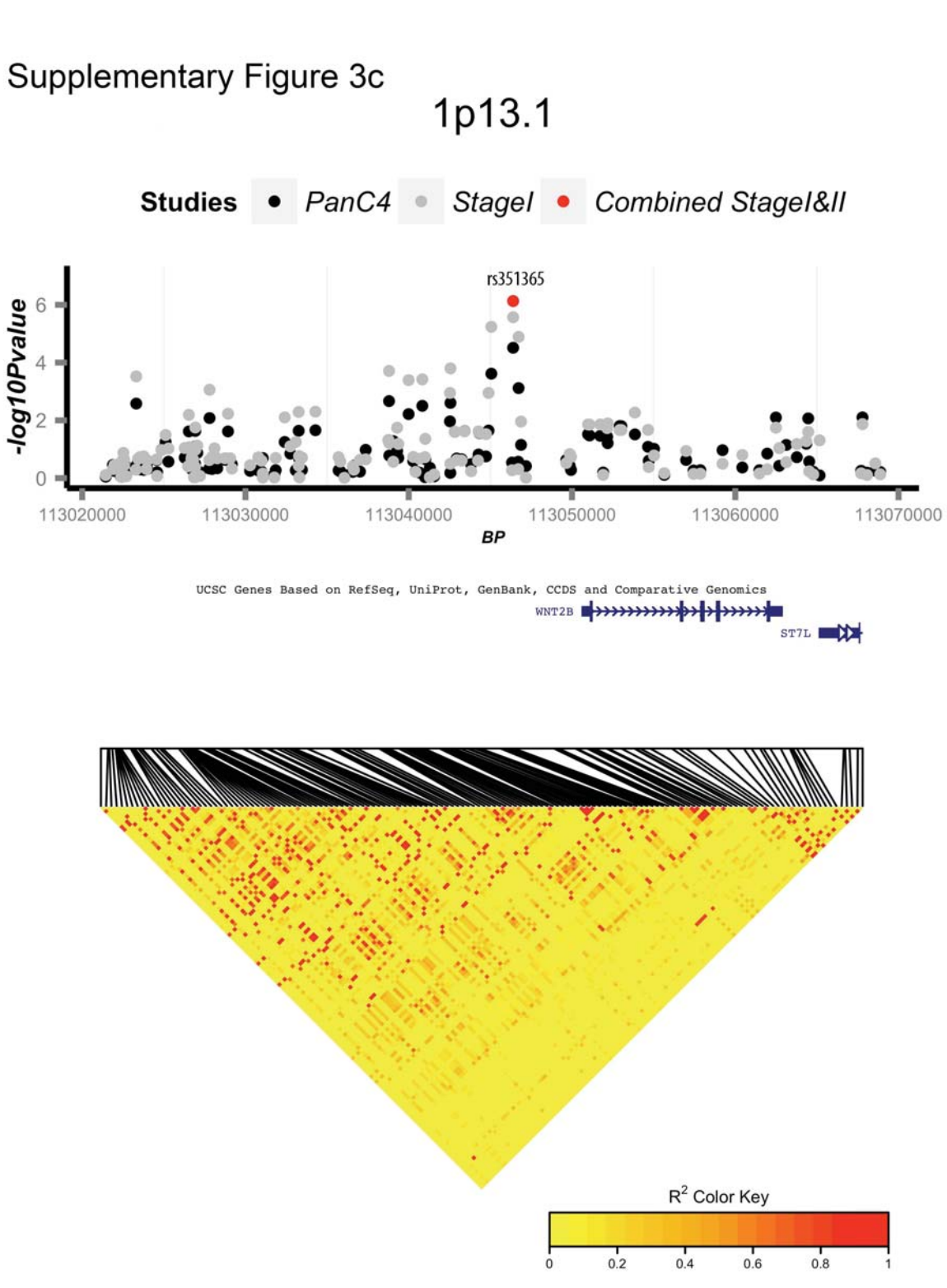


Supplementary Figure 3b

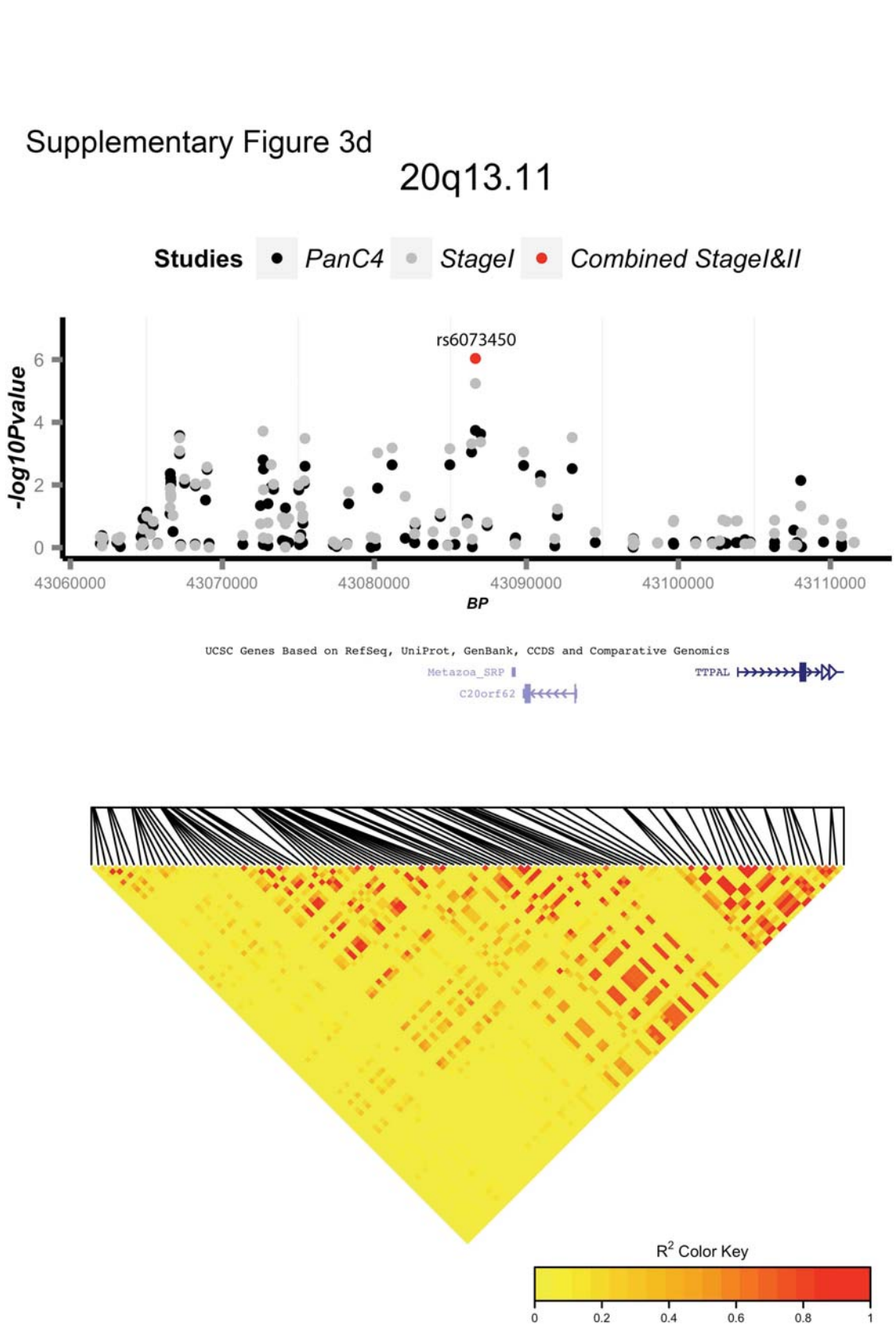


Nature Genetics: doi:10.1038/ng.3341

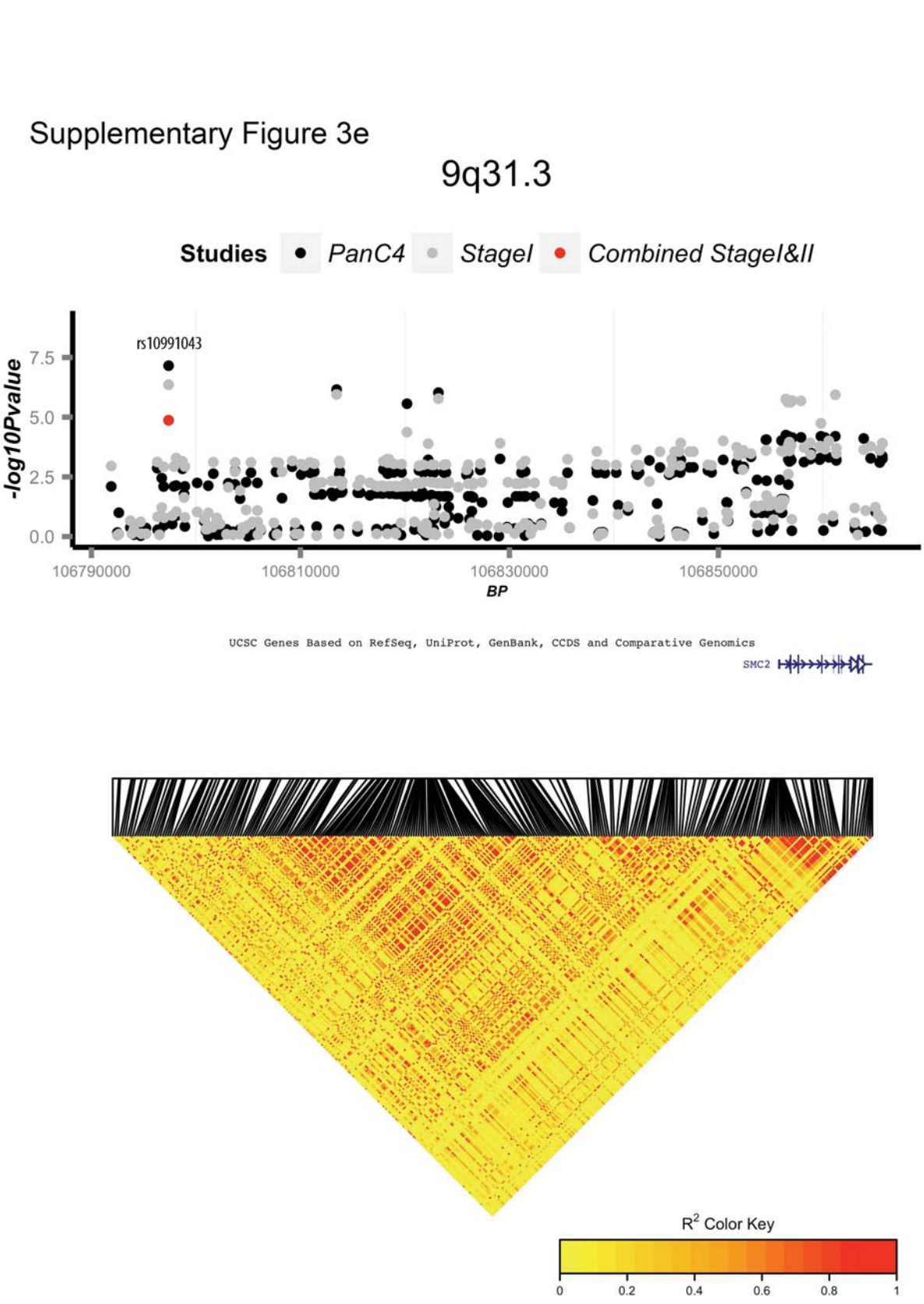
Supplementary Figure 3c



Supplementary Figure 3d

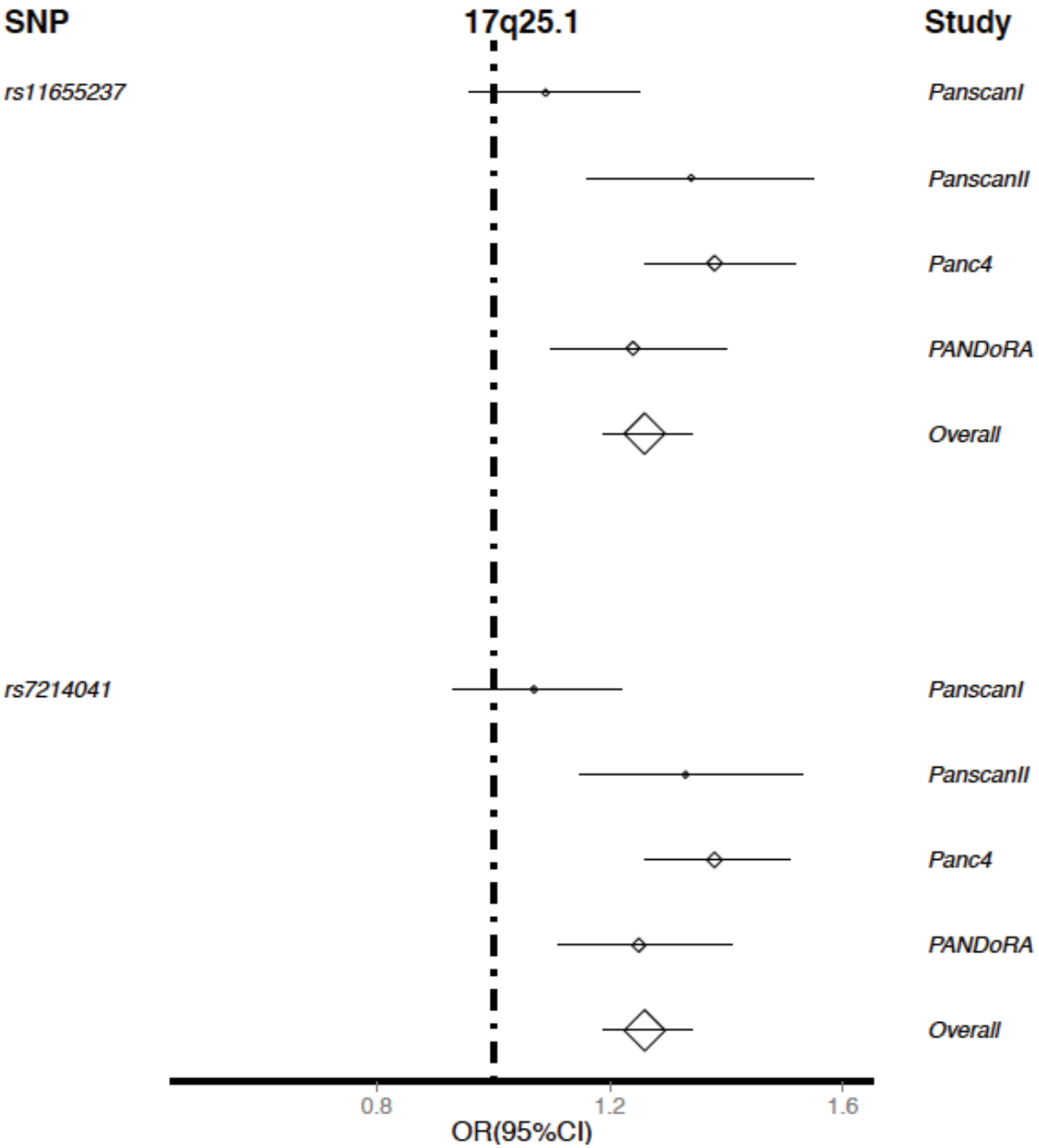


Supplementary Figure 3e

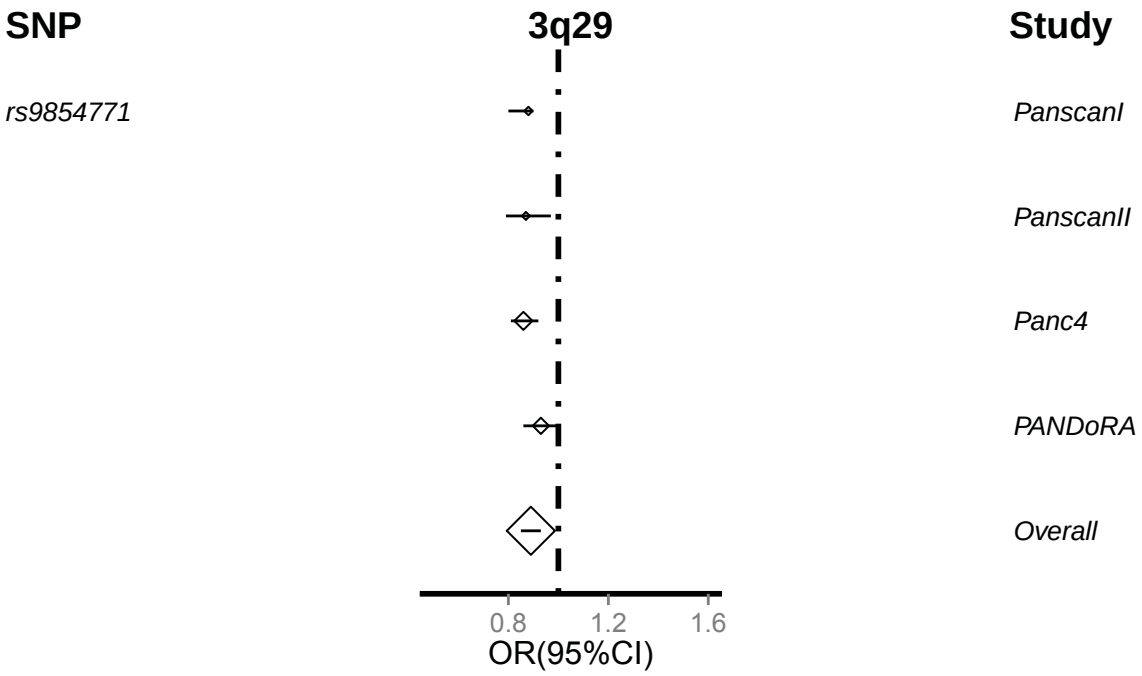


Supplementary Figure 4a-4i. Forest plots for the nine top associations reported

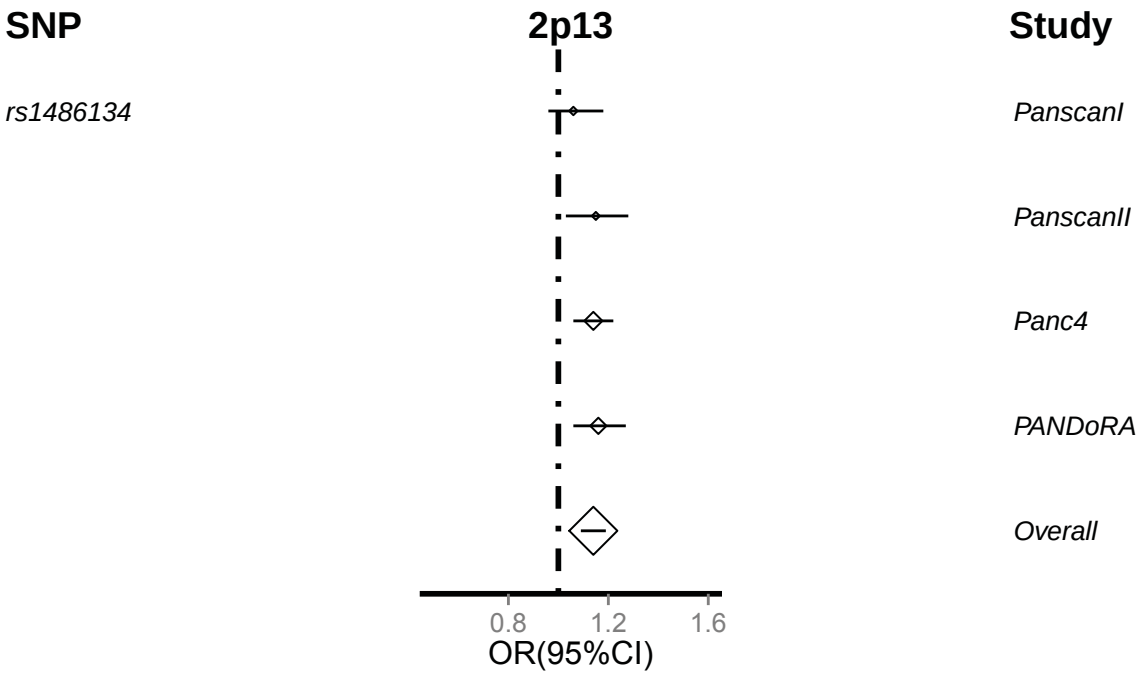
Supplementary Figure 4a



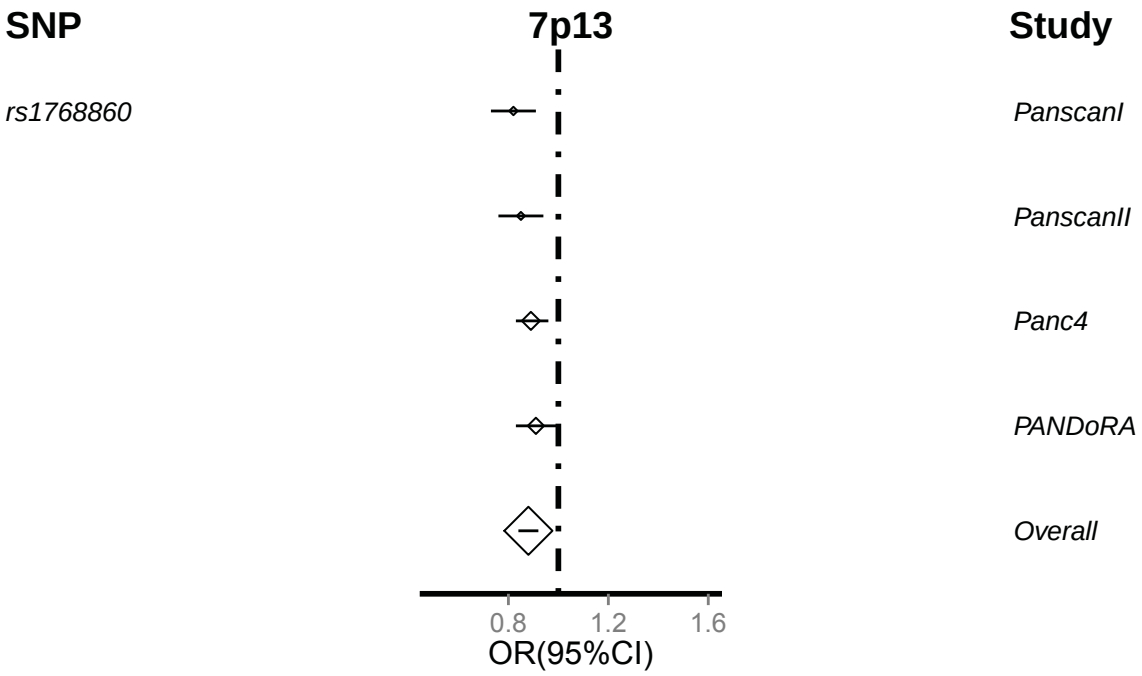
Supplementary Figure 4b



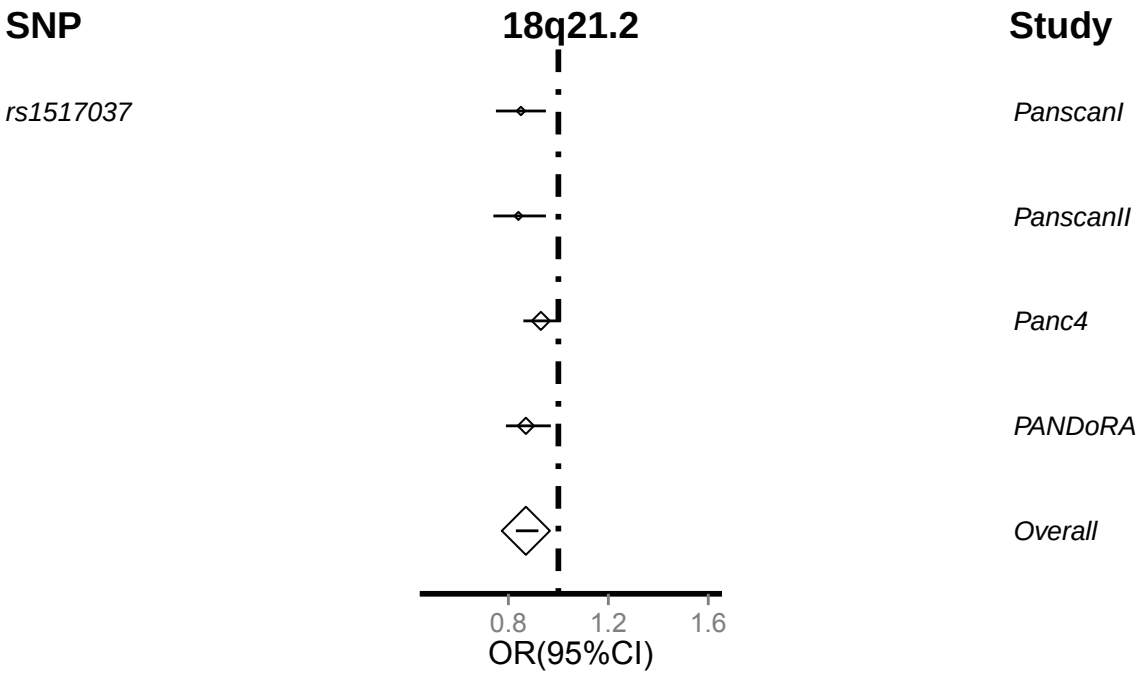
Supplementary Figure 4c



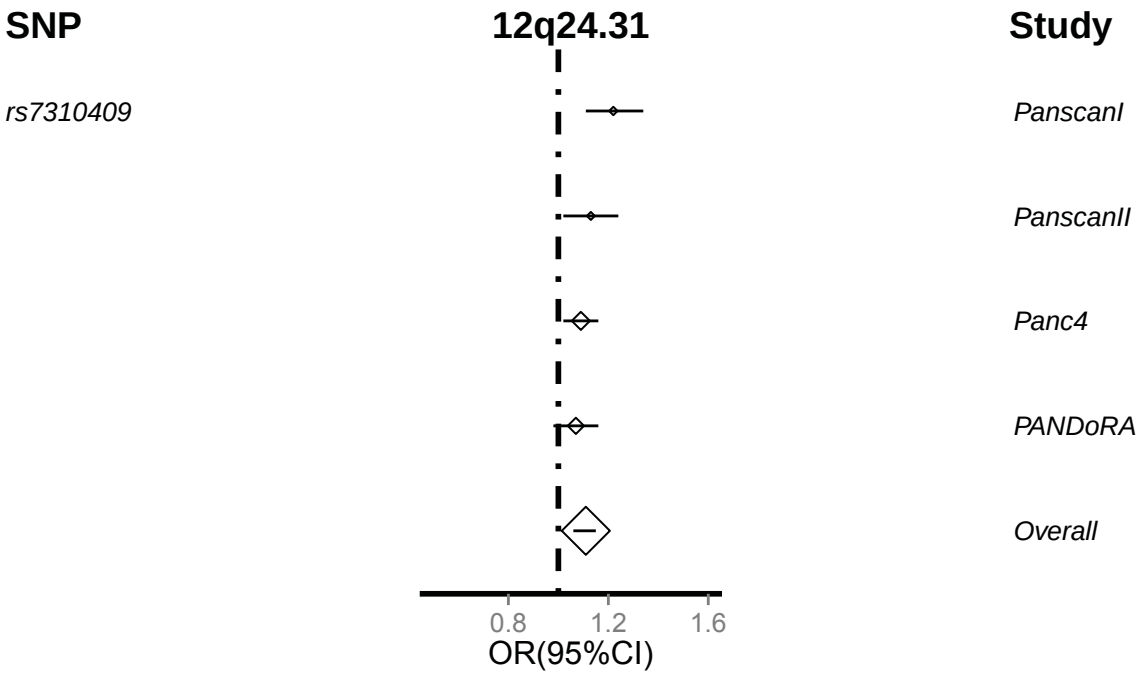
Supplementary Figure 4d



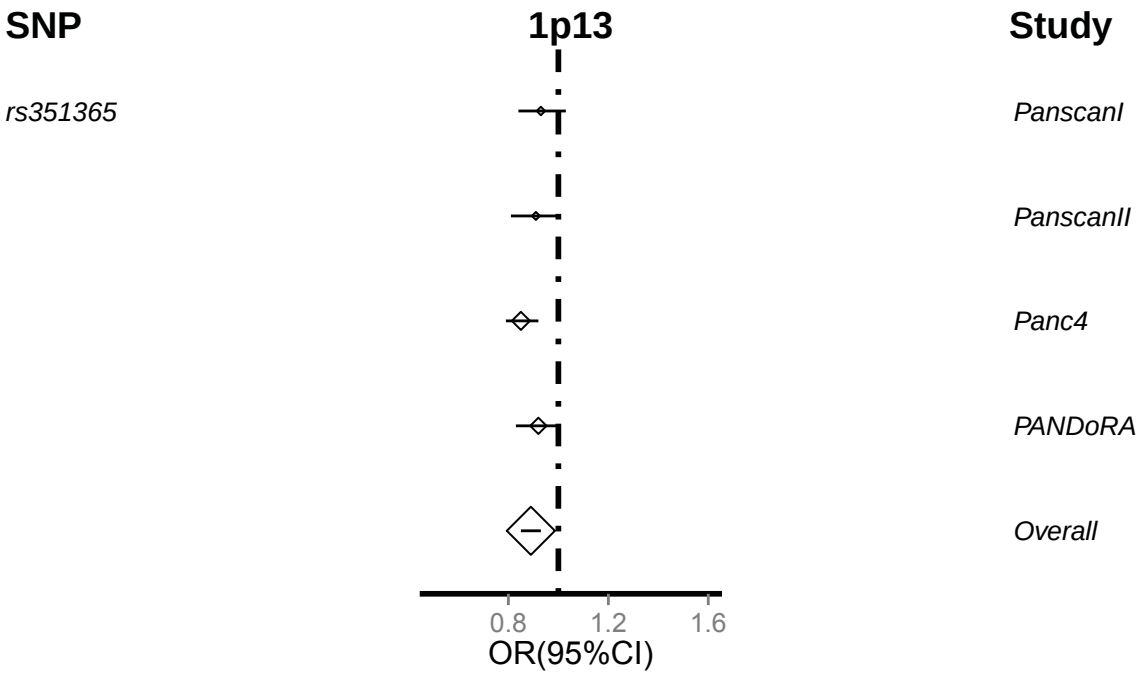
Supplementary Figure 4e



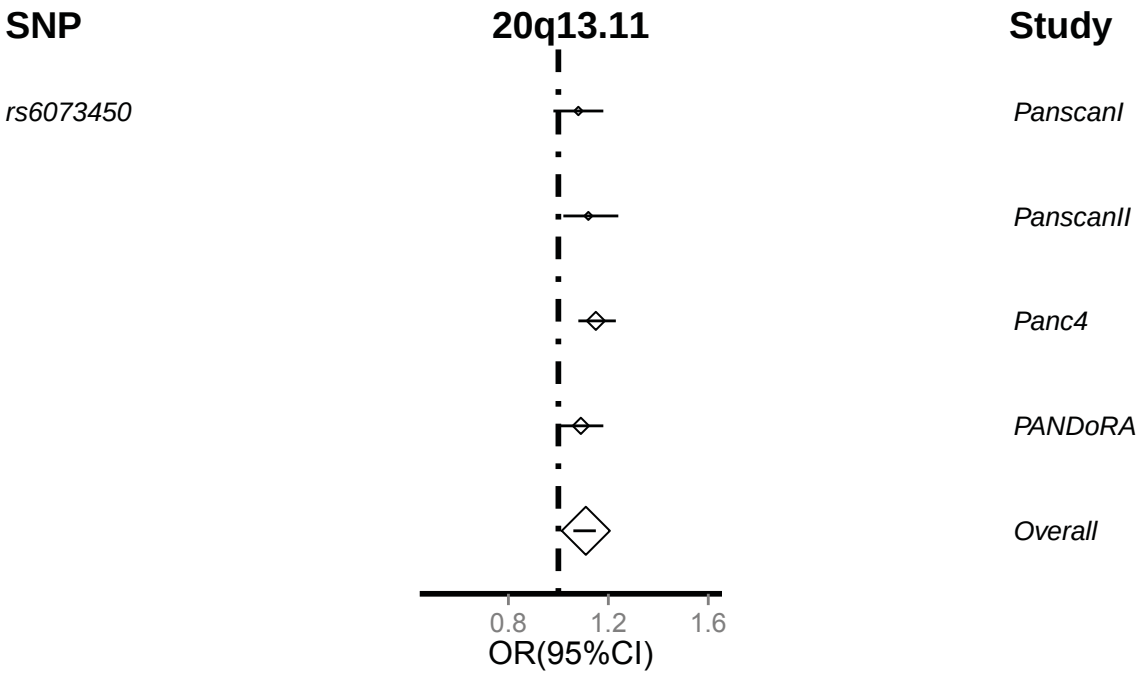
Supplementary Figure 4f



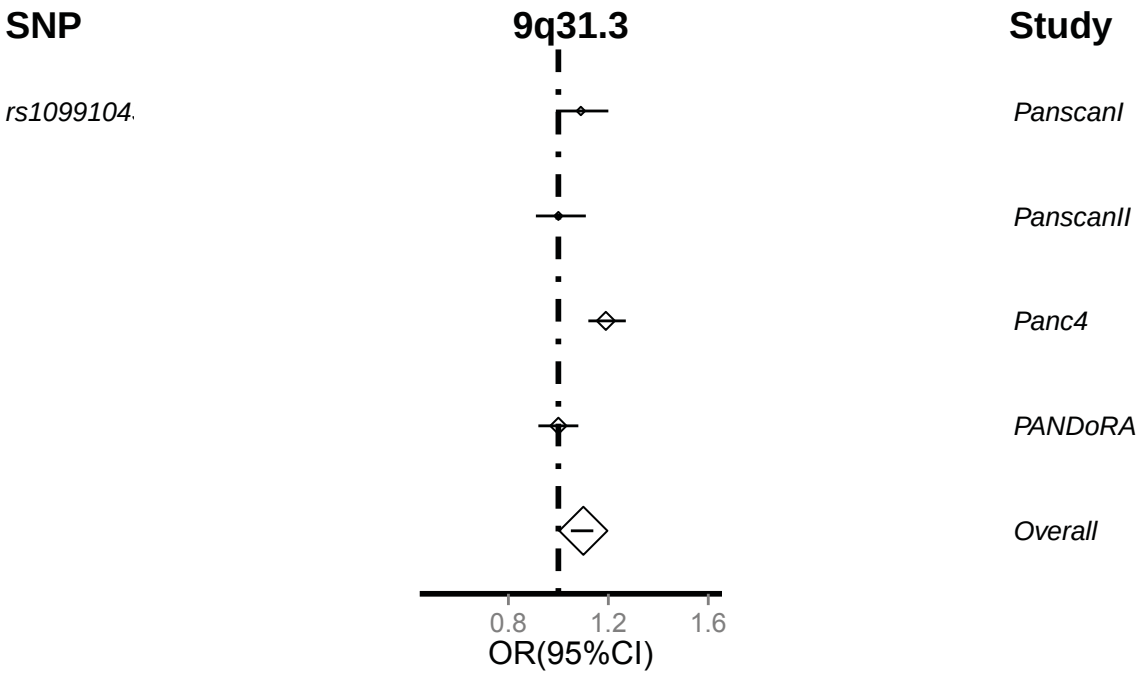
Supplementary Figure 4g



Supplementary Figure 4h



Supplementary Figure 4i



ANNEX 2

Supplementary Table 2.1. Regression P-value of each Eigenvector on cases-control study

N	Eigenvector	P-value	Significant
1	85.906	5.44E-06	***
2	41.118	1.36E-01	
3	17.130	1.86E-05	***
4	6.747	2.21E-01	
5	5.796	3.55E-01	
6	4.264	4.32E-03	**
7	3.605	3.11E-02	*
8	2.610	4.71E-01	
9	2.352	1.75E-01	
10	2.181	1.23E-01	

*** P-value < 5.00E-05; ** P-value < 5.00E-03; * P-value< 5.00E-02

Supplementary Table 2.2. Imputed genotypes MAF and INFO quality score

	All imputed				Filtered			
	INFO		MAF		INFO		MAF	
CHR	Median	Mean	Median	Mean	Median	Mean	Median	Mean
1	0.375	0.427	0	0.021	0.989	0.937	0.093	0.152
2	0.367	0.424	0	0.021	0.989	0.941	0.095	0.152
3	0.378	0.431	0	0.022	0.989	0.942	0.1	0.155
4	0.387	0.434	0	0.023	0.988	0.939	0.101	0.156
5	0.382	0.432	0	0.021	0.989	0.944	0.099	0.154
6	0.407	0.445	0	0.024	0.991	0.948	0.103	0.157
7	0.385	0.431	0	0.022	0.987	0.935	0.1	0.157
8	0.374	0.429	0	0.021	0.989	0.942	0.094	0.154
9	0.374	0.426	0	0.021	0.987	0.937	0.095	0.151
10	0.394	0.437	0	0.023	0.99	0.943	0.095	0.153
11	0.377	0.431	0	0.022	0.99	0.944	0.098	0.156
12	0.389	0.435	0	0.022	0.989	0.942	0.095	0.153
13	0.4	0.440	0	0.023	0.991	0.946	0.099	0.156
14	0.384	0.431	0	0.022	0.988	0.932	0.094	0.152
15	0.376	0.425	0	0.021	0.985	0.931	0.094	0.154
16	0.358	0.412	0	0.020	0.978	0.925	0.093	0.152
17	0.38	0.421	0	0.021	0.979	0.923	0.103	0.158
18	0.393	0.433	0	0.022	0.988	0.938	0.097	0.155
19	0.405	0.428	0	0.023	0.971	0.917	0.105	0.155
20	0.392	0.430	0	0.022	0.987	0.937	0.098	0.155
21	0.394	0.429	0	0.024	0.985	0.932	0.106	0.161
22	0.395	0.428	0	0.023	0.977	0.908	0.097	0.156
Average	**0.385**	**0.430**	**0**	**0.022**	**0.986**	**0.936**	**0.098**	**0.155**

Supplementary Table 2.3. Distribution of genes and NSV by chromosome

	Ensembl release 87	All NSV and Genes (Gencode14)				Only filtered genes (≥ 2 NSV; CMAF≥0.005)		
Chr	**Nº genes**	**Nº genes**	**Nº all NSV**	**Nº filtered NSV**	**Nº Singletons**	**Nº genes**	**Nº NSV**	**Nº Singletons**
1	2058	2012	14715	12237	1870	1180	10643	1576
2	1309	1210	9151	7729	1186	676	6736	985
3	1078	1067	8217	6965	1090	622	5940	890
4	752	800	5869	4836	803	435	4190	667
5	876	892	6315	5299	801	491	4575	670
6	1048	1020	8479	6664	963	623	5917	819
7	989	941	6486	5274	777	498	4462	632
8	677	696	4899	4110	648	351	3527	539
9	786	788	5822	4928	754	448	4264	627
10	733	770	5597	4560	750	413	3904	598
11	1298	1298	9404	7631	1188	746	6523	954
12	1034	1048	7006	5873	955	579	4968	767
13	327	322	2139	1794	263	167	1539	217
14	830	753	4822	3841	559	379	3227	444
15	613	628	4829	4024	638	340	3496	550
16	873	887	6246	5174	809	509	4550	683
17	1197	1140	7757	6310	994	614	5314	818
18	270	304	2131	1741	311	156	1493	259
19	1472	1378	9865	7704	1186	859	6660	954
20	544	531	3596	3021	492	298	2521	383
21	234	241	1709	1357	223	131	1160	174
22	488	487	3319	2614	409	263	2287	356
Total	19,486	19,213	138,373	113,686	17,669	10,778	97897,896	14562 14,562

Abbreviations: Chr=chromosome; **Nº**= Number; NSV=non-synonymous variants; CMAF=cumulative minimum allele frequency

Supplementary table 2.4. Other interesting genes with a P-value ≤ 5 × 10-4.

Region	BEGIN	END	Gene	Function	IDFc[a]	All[b]	Pass[c]	Sing[d]	PVALUE	RHO	CMAF
4q26	115769497	115998158	*NDST4*	*candidate tumor suppressor gene*	0.017	9	9	0	1.04E-04	0	0.008
17q22	56435080	56492800	*RNF43*	*tumor suppressor,* ***PC***	0.016	14	8	0	1.20E-04	0	1.321
3p21.31	49039984	49043292	*P4HTM*	hypoxia-inducible transcription factors	0.016	5	5	2	2.13E-04	1	0.007
17q25	79093270	79104992	*AATK*	induced during apoptosis	0.116	13	10	1	3.66E-04	0.1	1.081
5p15.3	1254594	1294166	*TERT*	*maintains telomere ends,* ***PC***	0.102	4	4	1	3.92E-04	0	0.054
16p13.3	1306346	1308333	*TPSD1*	*inflammatory disorders*	0.030	11	6	0	4.44E-04	0	1.528
16q12.2	56533694	56548584	*BBS2*	*Bardet-Biedl syndrome.*	0.018	8	7	2	4.49E-04	1	0.19
11p13	34129779	34167728	*NAT10*	*histone acetylation, tRNA acetylation, the biosynthesis of 18S rRNA,*	0.109	16	16	2	4.70E-04	0	0.063

Supplementary Figure 2.1. Q-Q plot