Online Methods

Study Subjects

Subjects in the first stage GWAS were drawn from 12 cohort studies and one case-control study (**Supplemental Table 1**) in the Pancreatic Cancer cohort consortium genome-wide association study (PanScan1) and are part of a larger international consortium, the National Cancer Institute sponsored Cohort Consortium. They include the American Cancer Society Cancer Prevention Study-II (CPS-II)[7], the Alpha-Tocopherol Beta-Carotene, Cancer Prevention Study (ATBC)[8], European Prospective Investigation into Cancer and Nutrition Study (EPIC- which is comprised of cohorts from Denmark, France, Germany, Great Britain, Greece, Italy, the Netherlands, Spain and Sweden)[9], CLUE II[10], Health Professionals Follow-up Study (HPFS)[11], Nurses' Health Study (NHS)[11], New York University Women's Health Study (NYUWHS)[12], Physicians' Health Study I (PHS I)[11], Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)[13], Shanghai Men's and Women's Health Study (SMWHS), Women's Health Initiative (WHI)[16], and the Women's Health Study (WHS)[17]. Each cohort that participated in PanScan, had a defined population from whom blood or buccal cells were collected prior to the diagnosis of pancreatic cancer. Incident primary pancreatic adenocarcinoma cases were identified by self report with subsequent medical record review, linkage with a cancer registry, or both. Cases were defined as primary adenocarcinoma of the exocrine pancreas (ICD-O-3 code C250-C259). Non-exocrine pancreatic tumors (histology type, 8150, 8151, 8153, 8155 and 8240) were excluded.

1,770 incident cases were identified among the cohorts as part of a nested case control study. An equal number of controls were selected within their respective cohort. Controls were alive and free of pancreatic cancer on the calendar date that their respective cases were diagnosed. One control was matched per case on calendar year of birth (+-5 years), sex, broad categories of race and ethnicity, as well as source of DNA (blood or buccal cell). The NHS, HPFS, WHS and PHS cohorts additionally matched on smoking status. 400 pancreatic adenocarcinoma cases and 400 controls were included from the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study[18]. The Molecular Epidemiology of Pancreatic Cancer study was initiated in 2000 and used an "ultra-rapid" case ascertainment system in which > 95% of patients from Minnesota, Iowa, and Wisconsin suspected with pancreatic cancer at the Mayo Clinic were approached. Among those with pancreatic cancer, 72% provided consent and a blood sample. Clinic controls were frequency matched to cases on age, race, gender, and area of residence from patients seeking general medical care.

Eight case-control studies from the PanC4 consortium participated in a replication of promising SNPs from the initial scan: University of Toronto[19], University of California San Francisco[20], Johns Hopkins University, MD Anderson Cancer Center[21], PACIFIC Study of Group Health and Northern California Kaiser Permanente, Memorial Sloan-Kettering Cancer Center[22], Yale University[23], and distinct cases and controls from the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study[18] (**Supplementary Table 2)**.

Each participating study obtained informed consent from study participants and approval from its Institutional Review Board (IRB) for this study. Each cohort study and the Mayo Clinic case control study has obtained IRB certification permitting data sharing in accordance with the NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS).

Genotyping and Quality Control

4,063 DNA samples (including 311 from buccal cells) were selected for genotyping (representing 3,932 individuals). 129 DNA samples were analyzed in duplicate.

Due to the multitude of studies of varying sample sizes in PanScan, results of genotype clustering were compared to verify goodness of fit, to detect genotype discordances, and monitor potential cluster heterogenity. The genotype models evaluated included: 1) Default cluster definitions provided by Illumina, 2) Clusters estimated from each study separately, 3) Clusters estimated from each study separately using samples with >98% completion rates, call the low completion samples using those cluster models, 4) Clusters estimated from all studies together using all samples, 5) Clusters estimated from all studies together using samples with >98% completion rates, then calling the low completion samples using those cluster model and 6) Clusters estimated from each study separately using samples with >98% completion rates, followed by grouping and re-clustering studies that showed similar cluster metrics. Genotypes for low completion samples were called using the corresponding cluster model. On the basis of completion rates and low discordance between known duplicate samples, the most rigorous clustering methods were 3), 5) and 6). Model 5 was chosen on the basis of parsimony.

561,466 SNP genotype assays were attempted on the 4,063 DNA samples using the Human Hap500 Infinium Assay (Illumina, San Diego, CA). Samples with less than 98%

completion after the second attempt were subsequently excluded. SNP assays with call rates <90% were excluded. An average discordance rate of 0.017% was observed for 139 pairs of duplicate DNA assays (including 129 plated duplicate samples).

Deviation from Hardy-Weinberg proportions were tested[34] in control samples (with CEU ancestry >0.80 by STRUCTURE) of each study (**Supplementary Figure 2**). No SNPs were excluded from analysis since the tests for association are valid in the presence of departure from Hardy-Weinberg proportions.

Participants with valid genotypes were excluded based on: 1) Unanticipated inter-study duplicates (n=14); 2) Completion rates lower than 98% (n=219 samples corresponding to 74 participants); 3) Unexpected within study duplicate (n=1) and ineligible samples (n=8). The final subject count for stage 1 association analysis is 1,896 cases and 1,939 controls (**Supplementary Table 4**).

Assessment of population structure was performed with STRUCTURE[35] by seeding the genotypes from the PanScan studies with the reference HapMap genotypes (based on Build 22 for HapMap II with MAF>5% in any of three HapMap populations)[36]. A set of 9,405 SNPs with $r^2<0.004$ were selected for this analysis[37,38,39]. A total of 59 participants (29 cases and 30 controls) were estimated to be of admixed origin with less than 80% similarity to CEU. No participants were excluded based on results from STRUCTURE but assigned the following categories for adjustment in the association analysis: European if CEU admixture portion was >80%, Asian if JPT/HCB admixture portion was >80% and other if admixture with no one continental group was greater than 80% (**Supplementary Figure 3**). African American ancestry was defined based on self-report and ranged similarity to YRI from 41% to 96%.

A principal component analysis (PCA) of DNA samples in this study (excluding inferred sib and half-sib pairs) was performed with EIGENSTRAT[40]. Five principal components were effective[41] for distinguishing significant population groups and were included as quantitative covariates to correct for genetic admixture.

Genotype data for the full scan was used to identify 144 participants with 60-99% identity by state (IBS) as potential relatives. Two sets of SNPs with pairwise $r^2<0.004$ were selected separately for Asian (13,905 SNPs) and non-Asian studies (9,405 SNPs), respectively and run on PREST[42] to identify 5 unexpected full-sib pairs and two unexpected half-sib pairs (7 cases and 7 controls), who were excluded from PCA but included in the association analysis.

TaqMan assays (ABI, Foster City, CA) were designed and optimized for 10 SNPs in the three notable regions as well as for a technical replica assay for rs505922 (rs687621) because this SNP could not be optimized (96% genotype concordance with HapMap samples) as per SNP500Cancer.

For the Fast-Track Replication Study, 5,845 samples were genotyped, including 180 duplicate DNA samples for quality control purposes. Genotyping was performed utilizing a multiplex integrated fluidic technology (Fluidigm Biomark, San Francisco, CA) and individual TaqMan assays (ABI, Foster City, CA). In the course of the follow-up replication genotyping, the opportunity to conduct a GWAS with the Illumina Infinium 610Quad was finalized. Since the same SNPs would be later genotyped, genotyping of only the top ten ranked SNPs was completed (second GWAS is ongoing). Consequently, genotyping of some of the samples for the Fast Track replication was performed with low DNA quantities (reserving sufficient DNA for the GWAS). Sample completion ranged from 28.90-99.40% per study and genotype completion rates per locus ranged from 57.7-99.8%. Overall genotype concordance between duplicate samples was 96.52% indicating the reliability of the current Fast Track replication results. Discordant genotypes between duplicates were set to missing. A small proportion (0.2%) of samples genotyped in Stage 2 were excluded as they were unanticipated inter-study or intra-study duplicates or had incomplete covariate data. The corresponding Infinium cluster plots for the 10 SNPs are shown in **Supplemental Figure 4**.

Association Analysis

All association analyses were conducted using logistic regression, adjusted for age (in ten-year categories), sex, study, arm (for WHI; intervention vs. observation), ancestry and five principal components of genetic structure. Each SNP genotype was coded as a count of minor alleles, with the exception of X-linked SNPs among males, which were coded as 2 if the participant carried the minor allele and 0 if he carried the major allele[25]. This log-linear odds model has near-optimal power across a wide range of alternative hypotheses, the main exception being rare recessive variants, for which we have limited power regardless of genotype coding[43]. A score test was performed on all genetic parameters in each model to determine statistical significance, with one degree of freedom.

We analyzed each study separately and conducted two analyses pooling multiple studies: the first included all cohorts (COHORTS); the second included all studies (ALL). We assessed heterogeneity in genetic effects across study using the Q and I^2 statistics[44].

We selected 10 SNPs from the three most notable regions from the GWAS (based on 2 or more SNPs per region ranked in the top 25 SNPs) for replication based on the results of the two pooled analyses. Association between pancreatic cancer and the replication SNPs was tested by fitting logistic regression models and testing the estimated genetic effects using the GLU software package. We analyzed each study separately in addition to pooling all eight studies. Models were adjusted for age in ten-year intervals, sex, self-reported race, and study. Genotypes were coded as counts of minor alleles (1-d.f. trend test). Combined single-SNP analyses pooled Stage 1 and Stage 2 data sets and adjusted for study, arm, age, sex, race and top five principal components of population stratification. In Stage 2 studies, principal components could not be calculated and were set to 0. **Supplementary Table 5** provides the results of the Stage 2 association analysis.

Data analysis and management used GLU (Genotyping Library and Utilities version 1.0), a suite of tools available as an open-source application for management, storage and analysis of GWAS data.

Data Access

The CGEMS data portal provides access to individual level data for 558,542 SNPs in 3,835 individuals to investigators from certified scientific institutions after approval of their submitted Data Access Request.

URLs:
CGEMS portal: http://cgems.cancer.gov/
CGF: http://cgf.nci.nih.gov/
EIGENSTRAT: http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm
GLU: http://code.google.com/p/glu-genetics/
SNP500Cancer: http://snp500cancer.nci.nih.gov/
STRUCTURE: http://pritch.bsd.uchicago.edu/structure.html
Tagzilla: http://tagzilla.nci.nih.gov/
Panc4: http://panc4.org