

Cancer Genetic Markers of Susceptibility (CGEMS) Pancreatic Cancer Genome-Wide Association Scan PanScan I

The CGEMS data portal provides public access to summary results for approximately 558,000 SNPs genotyped at the NCI Core Genotyping Facility (using the Sentrix HumanHap550 and HumanHap550-Duo genotyping assays, Illumina, San Diego, CA) in 1,896 pancreatic cancer patients and 1,939 controls¹.

Access to a subset of the individual raw phenotype and genotype data analyzed in PanScan I will be possible for scientific research purposes only restricted to cancer related analyses. Registration by the individual investigator and the supporting institution will be required because of privacy concerns. The accessible data will include genotypes from the Genome-Wide Association Scan (GWAS) and a set of covariates, namely, disease phenotype (control, case), age (in 10 year categories: <51, 51-60, 61-70, 71-80, and >80), sex and study.

Study Population

Subjects were drawn from 12 cohort studies and one case-control study 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 Alpha-Tocopherol Beta-Carotene, Cancer Prevention Study (ATBC)², CLUE II³, the American Cancer Society Cancer Prevention Study-II (CPS-II)⁴, 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)⁵, Health Professionals Follow-up Study (HPFS)⁶, Nurses' Health Study (NHS)⁶, New York University Women's Health Study (NYUWHS)⁷, Physicians' Health Study I (PHS I)⁶, Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)⁸, Shanghai Men's and Women's Health Study (SMWHS)^{9,10}, Women's Health Initiative (WHI)¹¹, and the Women's Health Study (WHS)¹². 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 to each 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 (never, former, and

current cigarette smokers). To enhance power in PanScanI, 400 pancreatic adenocarcinoma cases and 400 controls were included from the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study¹³. The Molecular Epidemiology of Pancreatic Cancer study was initiated in 2000 and used an “ultra-rapid” case ascertainment system with > 95% of all patients from Minnesota, Iowa, and Wisconsin suspected with pancreatic cancer at the Mayo Clinic being approached. In patients diagnosed 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 coming in for a general medical exam in Community, General or Area Internal Medicine.

Sample handling

DNA samples were received from each cohort/study and visually inspected for adequate fluid in individual tubes. All studies provided blood derived DNA samples apart from CPSII, PLCO and SMWHS, where 131, 159 and 21 DNA samples derived from buccal specimens were included, respectively. Three measurements of quantification were performed according to the standard procedures at the Core Genotyping Facility of the National Cancer Institute. These include pico-green analysis, optical density spectrophotometry and real time PCR (<http://cgf.nci.nih.gov/dnaquant.cfm>). Samples were also analyzed with 15 short tandem repeats and the Amelogenin marker in the IdentifilerTM Assay (ABI, Foster City, CA). All samples advanced to genotype analysis completed no less than 13 of the 15 micro-satellite markers and were compatible with the gender of each participant.

After final review, a total of 4,063 DNA samples (including 311 buccal cell derived DNA samples) were selected for genotyping (representing 3,932 individuals). An additional 129 DNA samples were aliquoted and plated in duplicate for quality control purposes.

Genotype Quality control

Due to the multitude of studies of varying sample sizes in PanScan, the results of genotype clustering were compared to verify goodness of fit, to detect genotype discordances, and monitor potential cluster heterogeneity. 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
- 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.

Assessment of Call Rates

A total of 561,466 SNP genotype assays were attempted on the 4,063 DNA samples using the Human Hap500 Infinium Assay (Illumina, San Diego, CA). Samples that did not meet a 98% completion threshold after the second attempt were excluded from further analysis. A total of 2924 SNPs (0.52% overall) failed to provide accurate genotype results either due to a lack of calls or low call rates (<90%).

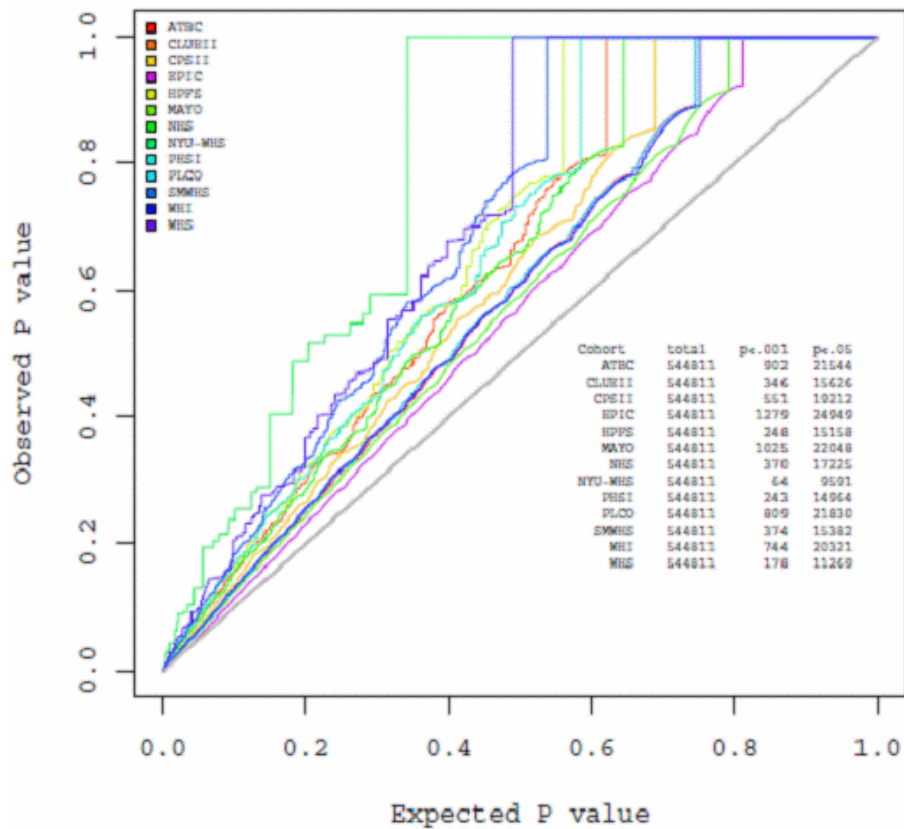
Assay concordance

The genotype concordance rate for SNP assays was evaluated using 139 pairs of known duplicated DNA assays (including 129 plated duplicate samples). These pairs of samples were separate aliquots from the same DNA preparation and all met quality control criteria required for the other samples, thereby, providing reliable data for comparison. An average discordance rate of 0.017% was observed (12,961 discordant genotypes out of 77,239,156 comparisons). No SNPs or samples were excluded from further analysis as a result of this analysis.

Hardy-Weinberg Proportions in control DNA

Deviation from Hardy-Weinberg proportions were tested using an exact test (see **Figure 1**)¹⁴. The analysis was conducted in all control samples with CEU ancestry >0.80 (by STRUCTURE) for each study. Significant deviations were observed for an average of 3.2% of all SNPs at the level of $p < 0.05$ and 0.1% at $p < 0.001$ per study. Supplementary Table 9 contains the numbers of SNPs that deviate from Hardy-Weinberg proportions for each cohort/study. No SNPs were excluded from analysis since the tests for association applied to such data are valid in the presence of departure from Hardy-Weinberg proportions (although with potentially reduced power when these deviations are due to systematic genotyping errors with equal effects among cases and controls).

Figure 1: Test for deviation from Hardy-Weinberg proportions.



Final sample selection for association analysis

Participants with valid genotypes were excluded from analysis based on: 1) Unanticipated inter-study duplicates (n=14); 2) Samples with completion rates lower than 98% (n=219 samples corresponding to 74 participants) were excluded; 3) Unexpected within study duplicate (n=1) and ineligible samples (n=8). No participants were excluded based on ethnicity. The final subject count for stage 1 association analysis is 1,896 cases and 1,939 controls (**Table 1**).

Table 1. Final participant counts for CGEMS Pancreatic Cancer GWAS

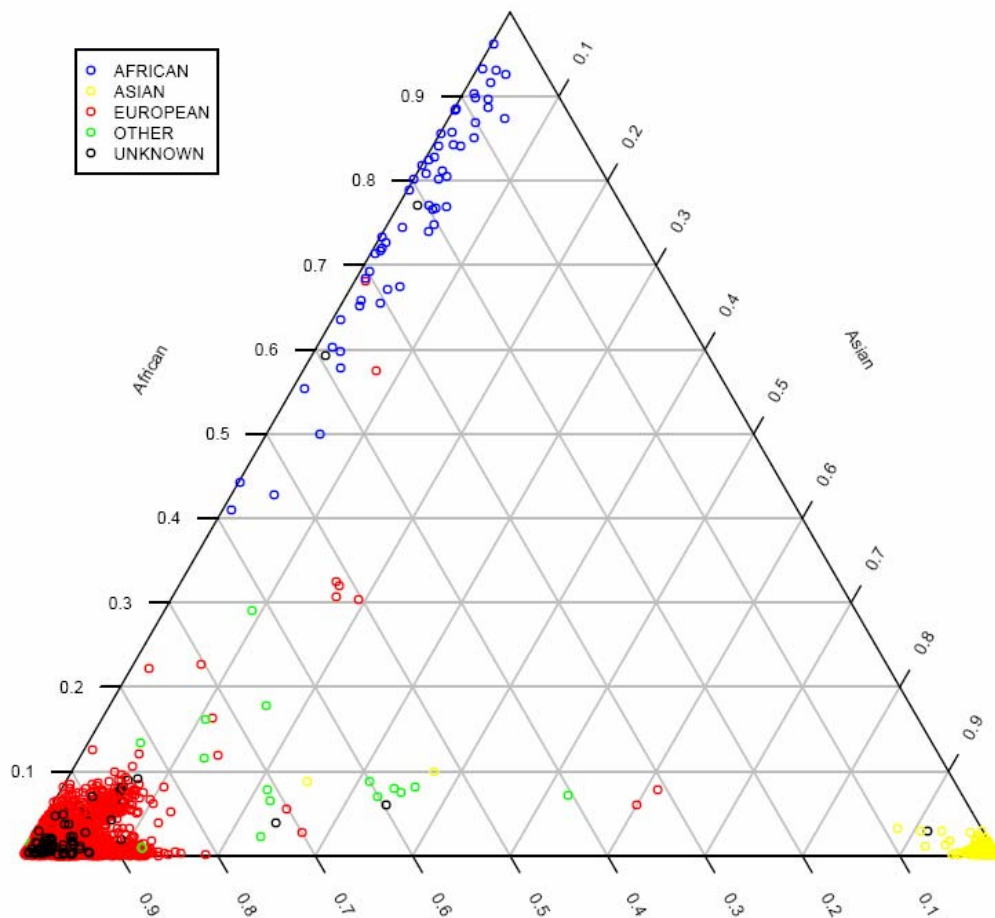
Cohort/Study	Cases	Controls
ATBC	194	208
CLUE II	68	71
CPS II	120	118
EPIC	421	436
HPFS	54	51
MAYO	368	345
NHS	82	84
NYU-WHS	13	13
PHS I	49	54
PLCO	199	220

SMWHS	58	65
WHI	245	242
WHS	25	32
Grand Total	1,896	1,939

Population structure

Similarity to three samples taken from different continental populations was estimated by using the STRUCTURE program¹⁵ by seeding the genotypes from the PanScan studies with those of the reference HapMap populations (based on SNPs in Build 22 for HapMap II with MAF>5% in any one of three HapMap populations)¹⁶. The number of clusters (the “k” parameter) was set to three and the CEU, YRI and JPT+CHB samples were each specified to a different cluster, these three panels are samples from populations of European, African and Asian origin respectively. The PanScan samples were left unspecified. A set of 9,405 SNPs with $r^2 < 0.004$ were selected for this analysis^{17,18,19}. 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% (**Figure 2**). African American ancestry was defined based on self-report and ranged in YRI similarity from 41% to 96%.

Figure 3. Plot of admixture defined by analysis with STRUCTURE.



To adjust for detectable differences in population substructure, a principal component analysis (PCA) of DNA samples in this study (excluding inferred sib and half-sib pairs) was performed with EIGENSTRAT²⁰. Five principal components were effective²¹ for distinguishing significant population groups and were included as quantitative covariates to correct for genetic admixture.

Assessment of Relationships

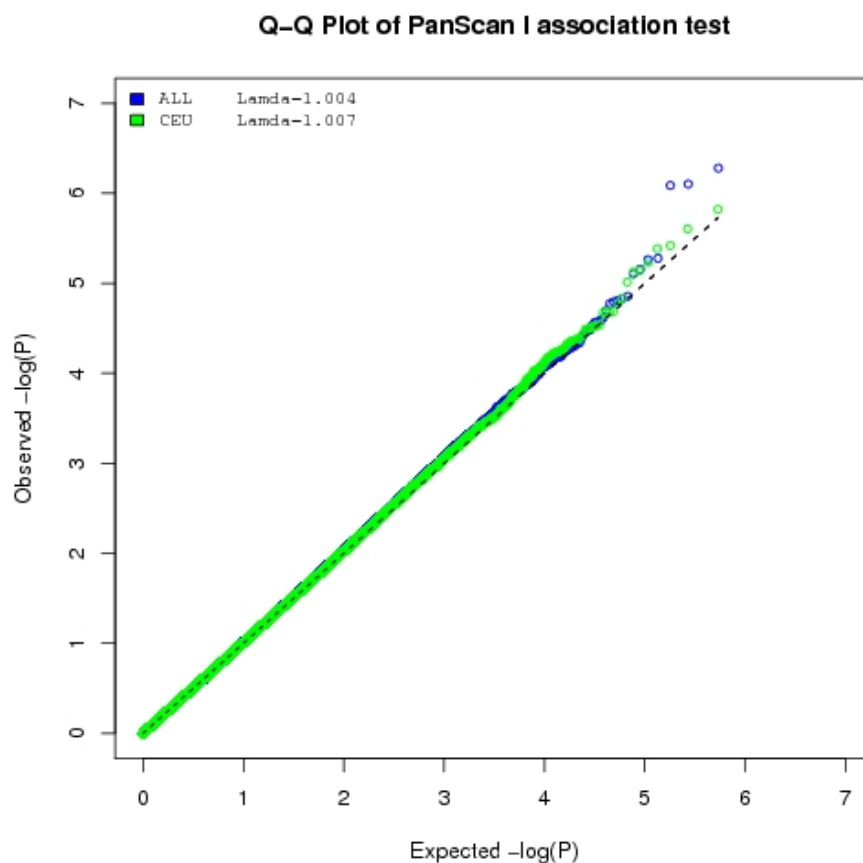
Genotype data for all SNPs on the Illumina HumanHap550 chip was used to identify 144 participants with 60-99% identity by state (IBS) as potential relatives. Two sets of SNPs with pairwise $r^2 < 0.004$ were selected separately for Asian (13,905 SNPs) and non-Asian studies (9,405 SNPs), respectively. These SNP sets were used to run PREST²² to identify 5 unexpected full-sib pairs and two unexpected half-sib pairs (7 cases and 7 controls), who were excluded from PCA but were included in the association analysis.

Association Analysis

The primary analysis of the initial PanScan GWAS study explores the association between single SNPs and pancreatic cancer susceptibility. All 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 allele²³. 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 coding²⁴. A score test was performed on all genetic parameters in each model to determine statistical significance, with one degree of freedom.

Results from the following two analytical procedures are provided: the first includes all 13 studies in PanScan I (ALL); the second includes all individuals in PanScan I with >80% admixture portion with the HapMap CEU population (CEU) according to STRUCTURE analysis.

Figure 4. Quantile-Quantile (QQ) plot of observed vs. expected P values in the GWAS.



Citation of data used:

Please cite the website for publications related to data available on this website

(<http://cgems.cancer.gov/>) and reference the full name of the study, Cancer Genetic Markers of Susceptibility.

References

1. Amundadottir, L. et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* **41**, 986-90 (2009).
2. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol* **4**, 1-10 (1994).
3. Gallicchio, L. et al. Single nucleotide polymorphisms in inflammation-related genes and mortality in a community-based cohort in Washington County, Maryland. *Am J Epidemiol* **167**, 807-13 (2008).
4. Calle, E.E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490-501 (2002).

5. Riboli, E. et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* **5**, 1113-24 (2002).
6. Wolpin, B.M. et al. Circulating insulin-like growth factor binding protein-1 and the risk of pancreatic cancer. *Cancer Res* **67**, 7923-8 (2007).
7. Zeleniuch-Jacquotte, A. et al. Postmenopausal levels of sex hormones and risk of breast carcinoma in situ: results of a prospective study. *Int J Cancer* **114**, 323-7 (2005).
8. Hayes, R.B. et al. Methods for etiologic and early marker investigations in the PLCO trial. *Mutat Res* **592**, 147-54 (2005).
9. Xu, W.H. et al. Joint effect of cigarette smoking and alcohol consumption on mortality. *Prev Med* **45**, 313-9 (2007).
10. Zheng, W. et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol* **162**, 1123-31 (2005).
11. Anderson, G.L. et al. Implementation of the Women's Health Initiative study design. *Ann Epidemiol* **13**, S5-17 (2003).
12. Rexrode, K.M., Lee, I.M., Cook, N.R., Hennekens, C.H. & Buring, J.E. Baseline characteristics of participants in the Women's Health Study. *J Womens Health Gen Based Med* **9**, 19-27 (2000).
13. McWilliams, R.R. et al. Polymorphisms in DNA repair genes, smoking, and pancreatic adenocarcinoma risk. *Cancer Res* **68**, 4928-35 (2008).
14. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887-93 (2005).
15. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).
16. Frazer, K.A. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).
17. Thomas, G. et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40**, 310-5 (2008).
18. Hunter, D.J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870-4 (2007).
19. Yu, K. et al. Population substructure and control selection in genome-wide association studies. *PLoS ONE* **3**, e2551 (2008).
20. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
21. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
22. Sun, L., Wilder, K. & McPeck, M.S. Enhanced pedigree error detection. *Hum Hered* **54**, 99-110 (2002).
23. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
24. Lettre, G., Lange, C. & Hirschhorn, J.N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol* **31**, 358-62 (2007).