

RESEARCH ARTICLE

Prediction model for pancreatic cancer risk in the general Japanese population

Masahiro Nakatochi¹, Yingsong Lin^{2*}, Hidemi Ito³, Kazuo Hara⁴, Fumie Kinoshita¹, Yumiko Kobayashi¹, Hiroshi Ishii⁵, Masato Ozaka⁶, Takashi Sasaki⁶, Naoki Sasahira⁶, Manabu Morimoto⁷, Satoshi Kobayashi⁷, Makoto Ueno⁷, Shinichi Ohkawa⁷, Naoto Egawa⁸, Sawako Kuruma⁹, Mitsuru Mori¹⁰, Haruhisa Nakao¹¹, Chaochen Wang², Takeshi Nishiyama¹², Takahisa Kawaguchi¹³, Meiko Takahashi¹³, Fumihiko Matsuda¹³, Shogo Kikuchi², Keitaro Matsuo¹⁴

1 Division of Data Science, Data Coordinating Center, Department of Advanced Medicine, Nagoya University Hospital, Nagoya, Japan, **2** Department of Public Health, Aichi Medical University School of Medicine, Nagakute, Japan, **3** Division of Cancer Information and Control, Aichi Cancer Center Research Institute, Nagoya, Japan, **4** Department of Gastroenterology, Aichi Cancer Center Hospital, Nagoya, Japan, **5** Clinical Research Center, National Hospital Organization Shikoku Cancer Center, Matsuyama, Japan, **6** Department of Hepato-biliary-pancreatic Medicine, The Cancer Institute Hospital of Japanese Foundation for Cancer Research, Tokyo, Japan, **7** Hepatobiliary and Pancreatic Medical Oncology Division, Kanagawa Cancer Center Hospital, Kanagawa, Japan, **8** Tokyo Metropolitan Hiroo Hospital, Tokyo, Japan, **9** Department of Internal Medicine, Tokyo Metropolitan Komagome Hospital, Tokyo, Japan, **10** Hokkaido Chitose College of Rehabilitation, Hokkaido, Japan, **11** Division of Hepatology and Pancreatology, Aichi Medical University School of Medicine, Nagakute, Japan, **12** Department of Public Health, Nagoya City University Graduate School of Medicine, Nagoya, Japan, **13** Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan, **14** Division of Cancer Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan

* linys@aichi-med-u.ac.jp

OPEN ACCESS

Citation: Nakatochi M, Lin Y, Ito H, Hara K, Kinoshita F, Kobayashi Y, et al. (2018) Prediction model for pancreatic cancer risk in the general Japanese population. PLoS ONE 13(9): e0203386. <https://doi.org/10.1371/journal.pone.0203386>

Editor: Amanda Ewart Toland, Ohio State University Wexner Medical Center, UNITED STATES

Received: April 10, 2018

Accepted: August 20, 2018

Published: September 7, 2018

Copyright: © 2018 Nakatochi et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by The Ministry of Health, Labor, and Welfare of Japan (<http://www.mhlw.go.jp/english/>): H21-11-1; The Ministry of Education, Culture, Sports, Science, and Technology of Japan (<http://www.mext.go.jp/en/>): 16H06277, 17K09095, 26253041, and 17015018. The funders had no role in study design, data

Abstract

Genome-wide association studies (GWASs) have identified many single nucleotide polymorphisms (SNPs) that are significantly associated with pancreatic cancer susceptibility. We sought to replicate the associations of 61 GWAS-identified SNPs at 42 loci with pancreatic cancer in Japanese and to develop a risk model for the identification of individuals at high risk for pancreatic cancer development in the general Japanese population. The model was based on data including directly determined or imputed SNP genotypes for 664 pancreatic cancer case and 664 age- and sex-matched control subjects. Stepwise logistic regression uncovered five GWAS-identified SNPs at five loci that also showed significant associations in our case-control cohort. These five SNPs were included in the risk model and also applied to calculation of the polygenic risk score (PRS). The area under the curve determined with the leave-one-out cross-validation method was 0.63 (95% confidence interval, 0.60–0.66) for the model that did or did not include cigarette smoking and family history of pancreatic cancer in addition to the five SNPs, respectively. Individuals in the lowest and highest quintiles for the PRS had odds ratios of 0.62 (0.42–0.91) and 1.98 (1.42–2.76), respectively, for pancreatic cancer development compared with those in the middle quintile. We have thus developed a risk model for pancreatic cancer that showed moderately good discriminatory ability with regard to differentiation of pancreatic cancer patients from control individuals. Our findings suggest the potential utility of a risk model that incorporates replicated GWAS-identified SNPs and established

collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

demographic or environmental factors for the identification of individuals at increased risk for pancreatic cancer development.

Introduction

Pancreatic cancer is a malignancy characterized by an elusive etiology, poor prognosis, and a lack of effective early detection tools. In Japan, pancreatic cancer represents the fourth leading cause of cancer deaths, with age-adjusted incidence and mortality rates having increased continuously over the past several decades[1]. The reason for this increasing pancreatic cancer burden is unclear, but it does not appear to be explained by risk factors, such as age and cigarette smoking, that have been established on the basis of epidemiologic studies[2].

The heritability of pancreatic cancer in Scandinavia has been estimated to be 36% by twin studies[3], suggestive of an important contribution of genetic variation to pancreatic cancer susceptibility. Focusing on common genetic variation represented by single nucleotide polymorphisms (SNPs), the National Cancer Institute Cohort Consortia (PanScan) have conducted four genome-wide association studies (GWASs) with populations of European ancestry and identified 18 risk loci that were robustly associated ($P < 5 \times 10^{-8}$) with pancreatic cancer risk[4]. GWASs in two populations (Japanese and Chinese) of East Asian ancestry identified additional risk loci that subsequently failed replication in other independent cohorts [10]. Whereas GWASs have improved our understanding of the role of common SNPs in pancreatic cancer, the identified SNPs confer relatively small increments in risk (1.1- to 1.5-fold) with regard to the development of this disease. Furthermore, it remains a challenge to identify low-frequency or rare variants and to further clarify their contribution to pancreatic cancer pathophysiology.

Detection of pancreatic cancer at an early stage in the general population is difficult. The relatively low prevalence of the condition, coupled with the lack of validated biomarkers or imaging modalities, has limited the feasibility of a population-based screening program. Such challenges can, in part, be addressed by the development of a risk prediction model that aims to identify a small subset of high-risk individuals by incorporating genetic variants as well as demographic and lifestyle risk factors[13]. As with risk prediction models for other cancer types, such as lung, breast, and colorectal cancer[14], the clinical utility of such risk models for pancreatic cancer is currently limited.

As far as we are aware, no risk model has been developed for the prediction of pancreatic cancer risk in the general population in Japan. As a first step toward the development of such a model, we examined SNPs identified by published GWASs for their associations with pancreatic cancer risk in an age- and sex-matched case-control data set of Japanese individuals. We then developed the model further by incorporating the GWAS-identified SNPs that showed significant associations in our case-control cohort as well as demographic and established risk factors. A validated risk model may help to identify individuals at increased risk for the development of pancreatic cancer, thereby raising awareness that can result in the adoption of risk-minimizing behavior or in early detection by follow-up examinations.

Materials and methods

Study subjects

To develop the risk model, we performed a genetic association study with 664 cases and 664 age- and sex-matched controls selected from two case-control data sets that included a total of

945 pancreatic cancer patients and 2109 control subjects. The first data set included 622 patients who were recruited for a multi-institutional case-control study of pancreatic cancer, the details of which have been described previously[19]. In this previous study, clinically or histologically (or both) diagnosed pancreatic cancer patients were recruited from January 2010 to July 2014 at five participating hospitals in central Japan, Kanto, and Hokkaido regions. Most of the patients were recruited by gastroenterologists and had tumors at stage 3 or stage 4 at the time of diagnosis. Approximately 33% of the patients underwent surgical resection. Questionnaire data on demographic and lifestyle factors and 7-ml blood samples were collected from the study participants. The second data set included 323 newly diagnosed pancreatic cancer patients as well as 2109 control subjects who were recruited to an epidemiologic research program at Aichi Cancer Center (HERPACC). All new outpatients at Aichi Cancer Center on their first visit were invited to participate in HERPACC. Those who agreed to participate filled out a self-administered questionnaire and provided a 7-ml blood sample. The data collected were entered into the HERPACC database and linked to the hospital cancer registry system periodically to confirm cancer diagnoses. The feasibility of using first-visit outpatients as control subjects within the framework of HERPACC has been addressed previously by comparing their epidemiologic features with those of randomly selected individuals from the general population[20]. In these two case-control data sets, the vast majority of pancreatic cancer cases had a histology of ductal adenocarcinoma, with a small proportion of endocrine tumors (1.7%) also being included. None of the control subjects had a diagnosis of cancer at the time of recruitment. For all case and control subjects in the present study, data on demographic and lifestyle factors, such as cigarette smoking and family history of pancreatic cancer in first-degree relatives, were extracted from questionnaire answers. Written informed consent was obtained from all study participants, and the study protocol was approved by the ethical board of Aichi Medical University, the institutional ethics committee of Aichi Cancer Center, the Human Genome and Gene Analysis Research Ethics Committee of Nagoya University, and the ethics committees of all participating hospitals.

For the present case-control genetic association study, cases diagnosed with endocrine tumors were excluded. Case and control subjects were matched according to sex and age (categorized in 5-year intervals). During the matching process, only individuals with available data for all variables, including age, sex, cigarette smoking status, and family history of pancreatic cancer, were selected. Totals of 664 cases and 664 control subjects were eligible for statistical analysis.

Genotyping and quality control

A total of 945 pancreatic cancer case subjects and 2109 control subjects were genotyped at the Center for Genomic Medicine, Kyoto University, with the use of a HumanCoreExome-12 v1.1 BeadChip array (Illumina, San Diego, CA, USA). Five samples with a genotype call rate of <0.98 were excluded. No samples showed a discrepancy between genetic and reported sex. The identity-by-descent method implemented in PLINK 1.9 software[21] detected 17 duplicate or closely related pairs of samples ($\pi\text{-hat} > 0.1875$), with one sample of each pair being excluded. Principal component analysis (PCA)[22] with the 1000 Genomes Project reference panel (phase 3)[23] detected seven subjects with estimated ancestries outside of the Japanese population. These seven samples were also excluded. Furthermore, PCA based on only our samples was performed to identify population outliers. On the basis of the first 10 principal components, nine population outliers were identified and were excluded from further analysis. Among the 542,585 SNPs that were genotyped with the array, we excluded nonautosomal SNPs as well as SNPs with a genotype call rate of <0.98 or a Hardy-Weinberg equilibrium

exact test *P* value of $<1 \times 10^{-6}$ in the control subjects, a minor allele frequency of <0.01 , or a departure from the allele frequency computed from the 1000 Genomes Project phase 3 EAS samples. Such quality control filtering resulted in the selection of 942 case subjects and 2074 control subjects as well as 248,185 SNPs.

Genotype imputation and postimputation processing

Genotype imputation was performed with SHAPEIT2[24] and Minimac3[25] software based on the 1000 Genomes Project cosmopolitan reference panel (phase 3)[23]. We searched GWAS Catalog[26] for published GWASs of pancreatic cancer and selected 77 candidate SNPs at 54 loci that had been characterized and found to be associated with pancreatic cancer (S1 Table). Of these 77 SNPs, 16 polymorphisms at 14 loci with an imputation quality score (r^2) of <0.8 (rs1747924, rs351365, rs4927850, rs35226131, rs6879627, rs73328514, rs6971499, rs10094872, rs1886449, rs7190458, rs7200646, rs4795218, rs77038344, rs11655237, rs7214041, rs6073450) were excluded. After imputation, we were finally left with 664 case subjects and 664 age- and sex-matched control subjects as well as 61 SNPs at 42 loci for statistical analysis.

Statistical analysis

To build a high-precision risk model, we applied a three-step approach. We initially performed a screening analysis with the 61 SNPs at 42 loci in which the cutoff *P* value was defined as <0.05 for logistic regression analysis. After this screening analysis and exclusion of SNPs in strong linkage disequilibrium with other polymorphisms, eight SNPs at seven loci that were significantly associated with pancreatic cancer remained for identification of SNPs that independently influence pancreatic cancer by logistic regression analysis with a stepwise forward selection procedure. Finally, we constructed two versions of a prediction model for pancreatic cancer: Model A included established risk factors and the five SNPs at five loci identified in the stepwise selection, and model B included only the SNPs.

Simple comparison of demographic and lifestyle risk factors between case and control groups was carried out with Fisher's exact test and Student's *t* test. In the screening step, the association between each SNP and pancreatic cancer was assessed with the use of logistic regression analysis. We used imputed genotype, which is the expected number of risk alleles for pancreatic cancer and is a continuous variable ranging from 0 to 2. We applied two types of analysis condition to assess the association of each SNP with pancreatic cancer: condition 1, in which no covariates were included, and condition 2, in which covariates comprised smoking status (nonsmoker = 0, ever-smoker = 1) and family history of pancreatic cancer (no = 0, yes = 1). In the subsequent step, multiple logistic regression analysis with a stepwise forward selection procedure was performed to identify SNPs that independently contribute to pancreatic cancer; the dependent variable was pancreatic cancer status (control = 0, case = 1) and independent variables included the imputed genotypes of each SNP. The significance level for inclusion in and exclusion from the model construction was $P < 0.05$. A version of the model including classical risk factors and SNPs identified by the stepwise selection was designated model A, whereas a version including only the five identified SNPs was designated model B. Receiver operating characteristic (ROC) analysis with the leave-one-out cross-validation (LOOCV) method was applied to evaluate model performance with the use of pROC of the R package[27]. Confidence intervals for area under the curve (AUC) values were assessed by 10,000-times bootstrap resampling.

We defined the polygenic risk score (PRS) for pancreatic cancer as the summation of the number of risk alleles multiplied by the corresponding natural logarithm of the odds ratio, In

(OR), in model B as follows:

$$\text{PRS}_{ij} = \sum_{i=1}^m \ln(OR_i) x_i$$

where m is the number of SNPs ($m = 5$ in this study), OR_i is the odds ratio for SNP i in model B, and x_i is the genotype coded as the number of risk alleles for SNP i . We calculated the PRS for each subject using this equation and then divided the study subjects into quintile groups (Q1 to Q5) with equal numbers of control subjects on the basis of the PRS. We compared the middle quintile group (Q3) with other groups (Q1, Q2, Q4, Q5) with the use of logistic regression analysis with adjustment for cigarette smoking and family history of pancreatic cancer.

Heritability analysis was performed with the use of GCTA software[28]. The analysis estimates the percentage of phenotypic variance explained by common SNPs. We assumed a prevalence of 0.000095 for pancreatic cancer in the Japanese population on the basis of data in the GLOBOCAN 2012 database[29]. To estimate the heritability, we used the data set comprising the 664 cases and 664 controls adopted for the association analysis as well as the 248,185 directly genotyped SNPs used for imputation.

A P value of <0.05 was considered statistically significant. All statistical analysis was performed with SAS software version 9.4 (SAS Institute, Cary, NC, USA) and the R project version 3.3 (www.r-project.org).

Results

We performed genotyping for 945 pancreatic cancer case subjects and 2109 control subjects with the use of a HumanCoreExome-12 v1.1 BeadChip array and imputed genotypes based on the 1000 Genomes Project cosmopolitan reference panel (phase 3). We picked up 77 candidate SNPs at 54 loci that had been characterized and found to be associated with pancreatic cancer in previous GWASs (S1 Table). Of these 77 SNPs, 16 SNPs at 14 loci were excluded from further analysis because of poor imputation quality. After postimputation processing, 664 case subjects and 664 age- and sex-matched control subjects as well as 61 SNPs at 42 loci remained for the subsequent analysis.

The demographic and lifestyle risk factors for the case and control data set selected for development of the risk model are shown in Table 1. The mean age was 60.8 years for the case subjects and 60.5 years for the controls ($P = 0.453$). Case subjects had a higher proportion of individuals with a family history of pancreatic cancer (6.0% versus 2.3%, $P < 0.001$) and ever-smokers (65.7% versus 53.0%, $P < 0.001$) compared with controls.

Table 1. Characteristics of case and control subjects.

| Characteristic | Controls ($n = 664$) | Cases ($n = 664$) | P value |
|--------------------------------------|------------------------|---------------------|-----------|
| Age (years) | 60.5 \pm 8.3 | 60.8 \pm 8.3 | 0.453 |
| Male | 440 (66.3%) | 440 (66.3%) | 1.000 |
| Height (cm) | 162.2 \pm 8.4 | 163.5 \pm 8.2 | 0.003 |
| Weight (kg) | 60.9 \pm 10.0 | 61.2 \pm 11.4 | 0.683 |
| Body mass index (kg/m ²) | 23.1 \pm 2.8 | 22.8 \pm 3.3 | 0.076 |
| Ever-smoker | 352 (53.0%) | 436 (65.7%) | <0.001 |
| Family history of pancreatic cancer | 15 (2.3%) | 40 (6.0%) | <0.001 |

Continuous data are means \pm s.d. Differences in continuous or noncontinuous variables between case and control groups were evaluated with Student's t test or Fisher's exact test, respectively.

<https://doi.org/10.1371/journal.pone.0203386.t001>

Table 2. Association of pancreatic cancer with 13 SNPs at seven loci in the case-control data set with a *P* value of <0.05.

| SNP | Locus | Position | Nearby genes | Alleles | Risk allele | RAF | | Condition 1 | | Condition 2 | |
|------------|---------|-----------|--|---------|-------------|-------|---------|--------------------------------|----------------|--------------------------------|----------------|
| | | | | | | Case | Control | OR (95% CI) | <i>P</i> value | OR (95% CI) | <i>P</i> value |
| rs13303010 | 1p36.33 | 894573 | <i>NOC2L</i> | G/A | G | 0.311 | 0.274 | 1.19 (1.01 _{IL} 1.41) | 0.039 | 1.20 (1.01 _{IL} 1.42) | 0.034 |
| rs12615966 | 2q12.1 | 105378957 | <i>LINC01114</i> , <i>LINC01158</i> | C/T | T | 0.113 | 0.089 | 1.34 (1.03 _{IL} 1.74) | 0.032 | 1.30 (1.00 _{IL} 1.70) | 0.053 |
| rs12478462 | 2q23.3 | 153654720 | <i>ARL6IP6</i> , <i>RPRM</i> | T/G | G | 0.687 | 0.648 | 1.19 (1.01 _{IL} 1.40) | 0.037 | 1.19 (1.01 _{IL} 1.40) | 0.039 |
| rs9854771 | 3q28 | 189508471 | <i>TP63</i> | G/A | G | 0.913 | 0.889 | 1.30 (1.00 _{IL} 1.68) | 0.046 | 1.27 (0.98 _{IL} 1.64) | 0.072 |
| rs687289 | 9q34.2 | 136137106 | <i>ABO</i> | G/A | A | 0.520 | 0.437 | 1.43 (1.22 _{IL} 1.67) | <0.001 | 1.42 (1.21 _{IL} 1.66) | <0.001 |
| rs657152 | 9q34.2 | 136139265 | <i>ABO</i> | C/A | A | 0.508 | 0.426 | 1.42 (1.21 _{IL} 1.66) | <0.001 | 1.41 (1.20 _{IL} 1.65) | <0.001 |
| rs505922 | 9q34.2 | 136149229 | <i>ABO</i> | T/C | C | 0.511 | 0.438 | 1.36 (1.16 _{IL} 1.60) | <0.001 | 1.36 (1.16 _{IL} 1.60) | <0.001 |
| rs630014 | 9q34.2 | 136149722 | <i>ABO</i> | A/G | G | 0.669 | 0.627 | 1.21 (1.03 _{IL} 1.43) | 0.019 | 1.21 (1.03 _{IL} 1.43) | 0.022 |
| rs9564966 | 13q22.1 | 73896221 | <i>KLF5</i> , <i>LINC00392</i> | A/G | A | 0.558 | 0.469 | 1.42 (1.22 _{IL} 1.65) | <0.001 | 1.40 (1.20 _{IL} 1.63) | <0.001 |
| rs9573163 | 13q22.1 | 73908846 | <i>KLF5</i> , <i>LINC00392</i> | C/G | C | 0.558 | 0.470 | 1.42 (1.22 _{IL} 1.65) | <0.001 | 1.40 (1.20 _{IL} 1.64) | <0.001 |
| rs4885093 | 13q22.1 | 73910026 | <i>KLF5</i> , <i>LINC00392</i> | G/A | G | 0.565 | 0.483 | 1.39 (1.19 _{IL} 1.62) | <0.001 | 1.37 (1.18 _{IL} 1.60) | <0.001 |
| rs9543325 | 13q22.1 | 73916628 | <i>KLF5</i> , <i>LINC00392</i> | C/T | C | 0.562 | 0.475 | 1.40 (1.21 _{IL} 1.63) | <0.001 | 1.39 (1.19 _{IL} 1.62) | <0.001 |
| rs16986825 | 22q12.1 | 29300306 | <i>ZNRF3</i> | C/T | T | 0.521 | 0.467 | 1.24 (1.07 _{IL} 1.44) | 0.005 | 1.22 (1.05 _{IL} 1.42) | 0.011 |

Risk alleles are defined in S1 Table with the exception of rs687289, the risk alleles for which are based on our data set. OR and *P* values in condition 2 were adjusted for cigarette smoking and family history of pancreatic cancer. OR values represent increased risk of pancreatic cancer per risk allele copy for each SNP. RAF, risk allele frequency.

<https://doi.org/10.1371/journal.pone.0203386.t002>

Thirteen SNPs at seven loci of the remaining 61 SNPs showed a significant association with pancreatic cancer in our case-control data set with a *P* value of <0.05 in condition 1 with no covariate or in condition 2 with the covariates of smoking status and family history of pancreatic cancer (Table 2, S2 Table). The OR for these 13 SNPs ranged from 1.19 to 1.43 for individuals with risk variants. One polymorphism of each SNP pair that exhibited strong linkage disequilibrium ($r^2 > 0.8$) was excluded, leaving eight SNPs at seven loci for stepwise logistic regression analysis. Five SNPs, cigarette smoking, and family history of pancreatic cancer remained significantly associated with pancreatic cancer risk in the stepwise logistic regression analysis and were therefore included in the development of two versions of the risk prediction model: Model A included cigarette smoking, family history of pancreatic cancer, and the five SNPs at five loci (Table 3), whereas model B included only the five SNPs (S3 Table). Akaike information criterion values for models A and B were 1764 and 1796, respectively. In model A, ever-smokers had a 1.5-fold increased risk compared with nonsmokers (OR = 1.58, with a 95% confidence interval [CI] of 1.26_{IL} 1.98). The OR for individuals with the effect alleles ranged from 1.20 to 1.43, after adjustment for cigarette smoking and family history of pancreatic cancer. The AUC values for the ROC curves derived from models A and B with the use of the LOOCV method were 0.63 (95% CI, 0.60_{IL} 0.66) and 0.61 (0.58_{IL} 0.64), respectively (Fig 1).

We calculated the polygenic risk score (PRS) for each study subject using model B and then divided the subjects into quintile groups (Q1 to Q5) with equal numbers of control individuals on the basis of the PRS (Fig 2a). The mean \pm s.d. values of the PRS for case and control subjects were 1.17 ± 0.42 and 1.01 ± 0.42 , respectively. The PRS was significantly associated with risk of pancreatic cancer (Fig 2b). Compared with subjects in the middle quintile of PRS values (Q3), the OR values were 0.62 (95% CI, 0.42_{IL} 0.90) and 0.83 (0.58_{IL} 1.20) for Q1 and Q2, respectively, and 1.98

Table 3. Demographic and lifestyle risk factors as well as the five SNPs included in risk model A.

| Factor or SNP | Locus | Position | Alleles | Risk allele | OR (95% CI) | <i>P</i> value |
|-------------------------------------|---------|-----------|---------|-------------|---------------------------------|----------------|
| Ever-smoker | | | | | 1.58 (1.26 _{1/2} 1.98) | <0.001 |
| Family history of pancreatic cancer | | | | | 2.61 (1.41 _{1/2} 4.86) | 0.002 |
| rs13303010 | 1p36.33 | 894573 | G/A | G | 1.20 (1.01 _{1/2} 1.42) | 0.039 |
| rs12615966 | 2q12.1 | 105378957 | C/T | T | 1.32 (1.01 _{1/2} 1.74) | 0.045 |
| rs657152 | 9q34.2 | 136139265 | C/A | A | 1.43 (1.22 _{1/2} 1.69) | <0.001 |
| rs9564966 | 13q22.1 | 73896221 | A/G | A | 1.42 (1.21 _{1/2} 1.66) | <0.001 |
| rs16986825 | 22q12.1 | 29300306 | C/T | T | 1.22 (1.04 _{1/2} 1.42) | 0.014 |

Risk alleles are defined in [S1 Table](#). OR represents increased risk of pancreatic cancer per risk allele copy for each SNP.

<https://doi.org/10.1371/journal.pone.0203386.t003>

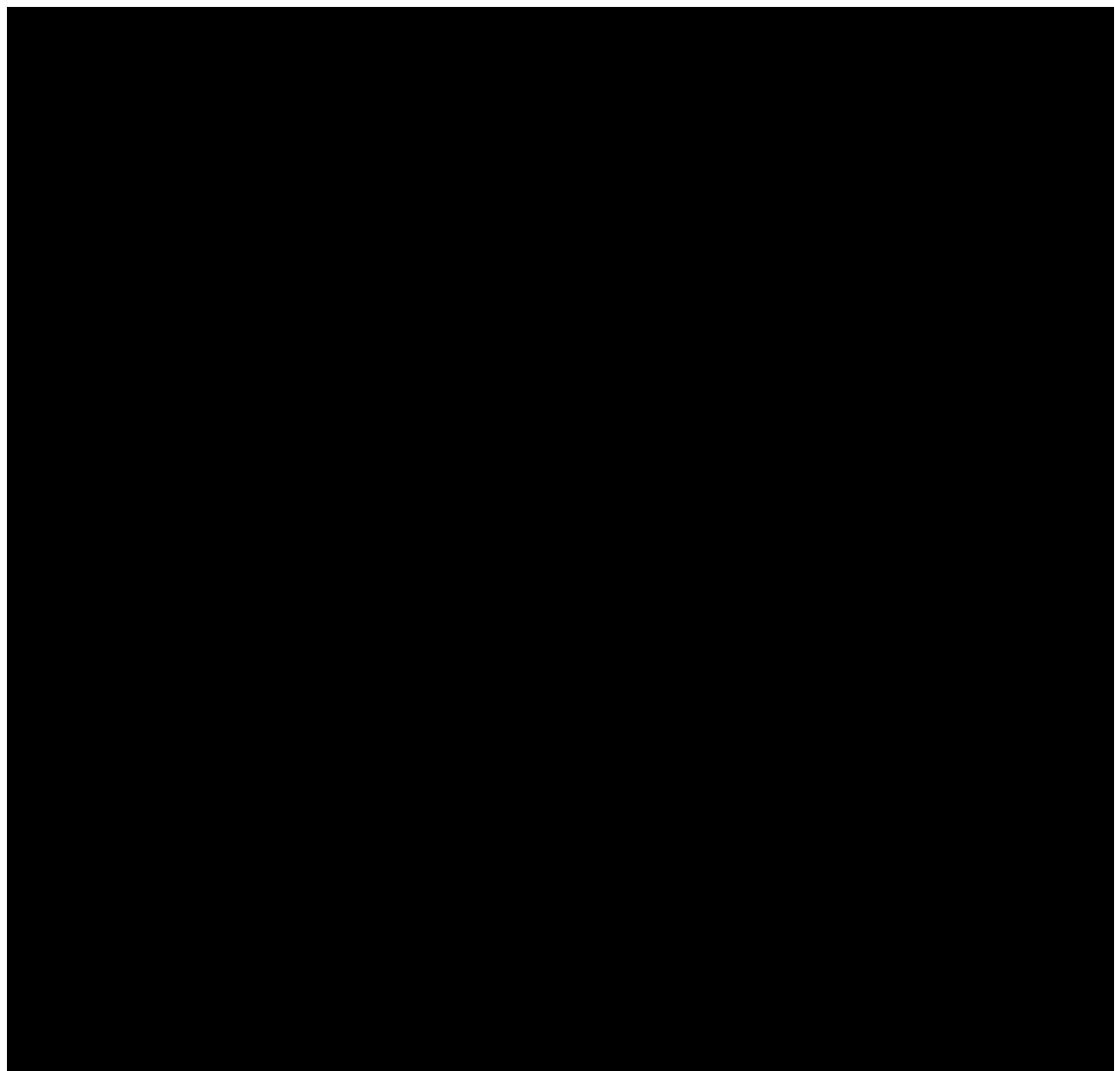


Fig 1. ROC curves for models A and B incorporating different variables according to the LOOCV method. Model A (blue line) incorporates classical risk factors and five GWAS-identified SNPs, whereas model B (red dashed line) includes only the five GWAS-identified SNPs. AUC values (95% CI) for models A and B are 0.63 (0.60_{1/2} 0.66) and 0.61 (0.58_{1/2} 0.64), respectively. The gray diagonal line corresponds to an AUC of 0.5 and no discrimination.

<https://doi.org/10.1371/journal.pone.0203386.g001>

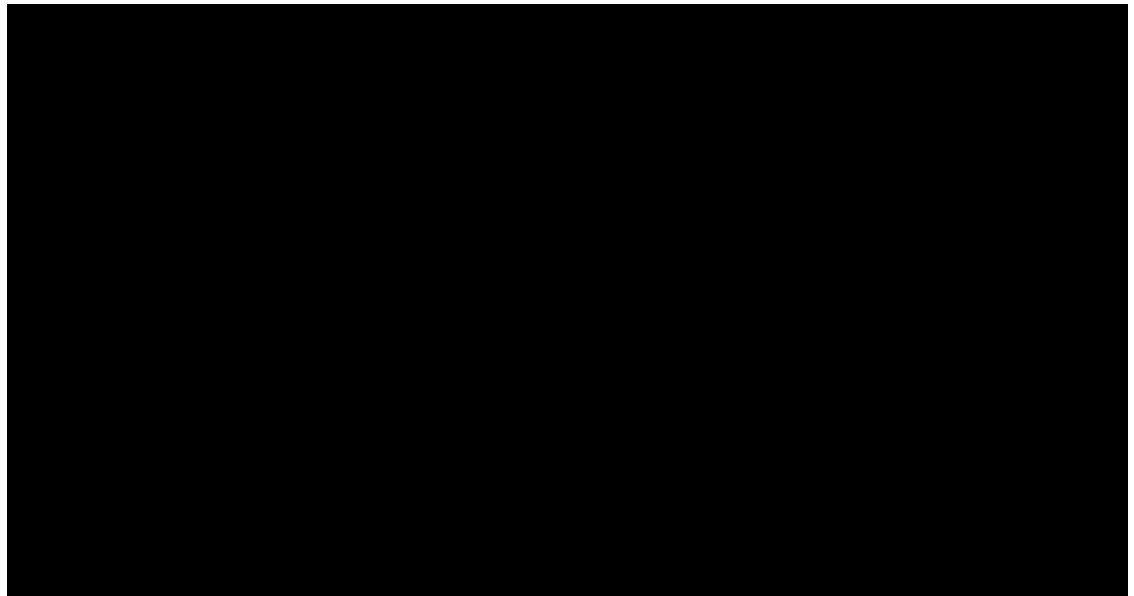


Fig 2. Percentage of subjects as well as the OR for pancreatic cancer according to PRS. (a) Distribution of the PRS in pancreatic cancer case and control subjects. (b) The OR for pancreatic cancer according to the quintiles of PRS. Vertical bars represent 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0). $P < 0.05$, $1.12 \leq 0.01$ versus Q3. The cutoffs for the quintiles of PRS in the control subjects were Q1 = 0.58, $0.58 < Q2 = 0.91$, $0.91 < Q3 = 1.12$, $1.12 < Q4 = 1.31$, and $Q5 > 1.31$. OR values were calculated by logistic regression analysis with adjustment for cigarette smoking and family history of pancreatic cancer.

<https://doi.org/10.1371/journal.pone.0203386.g002>

(1.42% \pm 2.76%) subjects in Q1, Q2, Q4, and Q5, respectively, after adjustment for cigarette smoking and family history of pancreatic cancer.

Finally, we estimated the heritability of pancreatic cancer due to common GWAS SNPs using only data for directly genotyped SNPs (248,185 SNPs for 664 cases and 664 age- and sex-matched controls). For a disease prevalence of 0.000095, we estimated that 16.1% (95% CI, 7.8% \pm 24.3%) of the total phenotypic variation in our data set was explained by common SNPs across the genome.

Discussion

Risk prediction models incorporating SNPs and environmental risk factors offer a means to identify a subset of individuals with increased cancer risk in the general population[30]. With the use of a case-control data set based on the Japanese population, we have now developed a risk model for pancreatic cancer and showed that it performed moderately well for discrimination of pancreatic cancer patients from individuals without the disease. Our findings suggest that a risk model that incorporates replicated GWAS-identified SNPs and established environmental factors is potentially useful for the identification of a subset of Japanese individuals with an increased risk for the development of pancreatic cancer.

Three risk models have been developed to date for the prediction of pancreatic cancer risk in general populations of different ethnicities[13,31,32]. To identify individuals at elevated risk for pancreatic cancer in a population of European ancestry, Klein et al. estimated the absolute risk of pancreatic cancer development with a risk model (based on three GWAS-identified SNPs, sex, age, ABO genotype, family history of pancreatic cancer, body mass index, cigarette smoking, and heavy alcohol intake) and incidence data from the SEER registries[13]. The AUC for the model was 0.61 (95% CI, 0.58 \pm 0.63) which demonstrated its superiority over a model that included only genetic or only nongenetic factors. However,

only a few individuals were estimated to have a 10-year absolute risk of >2% even if all genetic and nongenetic factors were present, indicating that the clinical utility of the model is low. By combining SEER data with a logistic model including risk factors (cigarette smoking, current use of proton pump inhibitors, recent diagnosis of diabetes mellitus and pancreatitis, Jewish ancestry, and ABO blood group other than O) that were identified from a population-based case-control study, Risch et al.[31] showed that 0.87% of controls with a combination of risk factors had an estimated 5-year absolute risk of >5%. It should be noted that these two models were developed for populations of European ancestry and that their performance in populations of Asian ancestry, including Japanese, awaits validation. With regard to East Asian populations, Yu et al. developed a risk prediction model to estimate individual risk of pancreatic cancer in the Korean population[32]. The model included biomarkers such as fasting blood glucose and urinary glucose levels as well as demographic and lifestyle risk factors, and it showed good discrimination ability with a validation set, with C-statistics of 0.81 (95% CI, 0.80 to 0.83) for men and 0.80 (0.79 to 0.82) for women. No genetic factors were included in this prediction model, however. The discrimination ability of our risk model is similar to that of the model of Klein et al.[13], and it would be similar to that of the model of Yu et al.[32] for the Korean population if we included matching factors such as age and sex. However, one key issue with all these risk models is the difficulty of their translation to clinical or public health practice. Further studies are thus needed to clarify the application of risk prediction models in different contexts, such as for population-wide use as a risk assessment tool or for screening for individuals with a high absolute risk of pancreatic cancer.

The PRS is independent of established risk factors and can provide risk stratification beyond family history[33]. Given that a polygenic component to pancreatic cancer risk was suggested by previous studies[34], we calculated the PRS using the five replicated GWAS-identified SNPs. Our results showed that this approach provided a good stratification of pancreatic cancer risk. Exploration of the polygenic contribution to pancreatic cancer risk beyond the known risk variants, however, will require studies with larger sample sizes and more sophisticated analytic approaches[35]. Although our sample size limited further evaluation of the PRS, risk prediction models for pancreatic cancer that incorporate the PRS are worth pursuing, given that fewer common genetic variants have been identified for this cancer than for other cancer types such as breast and colorectal cancer.

Risk models that incorporate GWAS-identified SNPs should be interpreted in the context of heritability that can be explained by common SNPs. In the present study, we estimated that 16.1% (95% CI, 7.8 to 24.3%) of the total phenotypic variation in our data set was explained by common SNPs across the genome. A prediction model based on all common SNPs across the genome would thus have a performance that corresponds to the heritability. Heritability of pancreatic cancer in individuals of European descent has been estimated on the basis of common SNPs across the genome. Childs et al. estimated that 16.4% (95% CI, 10.4 to 22.4%) and 13.1% (95% CI, 9.9 to 16.3%) of the total phenotypic variation in PanC4 and in the combined data set, respectively, was explained by common SNPs across the genome[9]. Our results are thus consistent with these previous findings. The heritability of pancreatic cancer might therefore be similar in populations of Japanese or European descent.

One strength of our study is that the risk model we constructed was based on case-control data for Japanese subjects. Our risk model represents the first attempt to use existing GWAS-identified SNPs to identify a subset of the general Japanese population at increased risk of developing pancreatic cancer. We were able to replicate 13 SNPs at seven loci out of 61 GWAS-identified SNPs at 42 loci in our case-control data set. Although the effect sizes for the associations of SNPs with pancreatic cancer in our study are small, they are consistent with the

results of previous GWASs for this cancer type. Furthermore, to address issues relating to overfitting, we used the LOOCV method to assess the performance of our prediction model. The AUC estimated with the LOOCV method was similar to those of previous models developed for pancreatic cancer, supporting the validity of our risk model based on the selected SNPs.

Our study also has several limitations. First, our GWAS was limited by the relatively small sample size and we did not validate our risk model in independent cohort samples. A clinically useful risk model needs to perform well with independent data sets and be generalizable to external populations. We will continue our efforts to find independent data sets with which to validate our risk model. Second, the clinical utility of our model as well as its potential contribution to a reduction in pancreatic cancer mortality in the general population are limited, although the discriminatory ability of the version of the model that included both demographic and lifestyle factors and replicated GWAS-identified SNPs was better than that of the version based only on SNPs. As shown previously[13], risk models that focus on rare cancer types such as pancreatic cancer may not offer clinically meaningful risk stratification because the percentage of individuals with a high absolute risk who warrant follow-up examination is small. Third, although we assessed all reported genome-wide significant SNPs documented for pancreatic cancer in GWASs, it is likely that other risk variants with borderline significance were not captured. Fourth, in addition to demographic factors and family history of pancreatic cancer, our study included only cigarette smoking as the most consistent risk factor for pancreatic cancer in Japanese. Further exploration and establishment of risk factors for pancreatic cancer in Japanese subjects may contribute to refinement of risk models.

In summary, we have developed a risk model for pancreatic cancer in Japanese individuals that showed a moderately good discriminatory ability with regard to differentiation of pancreatic cancer patients from control individuals. Further research is warranted to address the clinical utility of the model or its application to population-based screening. In particular, the goal of early detection can be pursued further by incorporation of established environmental risk factors, circulating biomarkers, the PRS, as well as other genomic data.

Supporting information

S1 Table. Information on the 77 SNPs at 54 loci extracted from published GWASs for pancreatic cancer.

(DOCX)

S2 Table. Association of pancreatic cancer with 61 SNPs at 42 loci extracted from published GWASs in our case-control cohort.

(XLSX)

S3 Table. The five SNPs at five loci included in the risk model.

(XLSX)

S1 Data. The data set used to construct the risk prediction model.

(XLSX)

Acknowledgments

We thank Mayuko Masuda, Kikuko Kaji, Kazue Ando, Etsuko Ohara, and Sumiyo Asakura for assistance with data collection.

Author Contributions

Conceptualization: Masahiro Nakatochi, Yingsong Lin, Chaochen Wang, Takeshi Nishiyama, Shogo Kikuchi, Keitaro Matsuo.

Formal analysis: Masahiro Nakatochi, Fumie Kinoshita, Yumiko Kobayashi.

Funding acquisition: Masahiro Nakatochi, Yingsong Lin, Shogo Kikuchi, Keitaro Matsuo.

Resources: Hidemi Ito, Kazuo Hara, Hiroshi Ishii, Masato Ozaka, Takashi Sasaki, Naoki Sasahira, Manabu Morimoto, Satoshi Kobayashi, Makoto Ueno, Shinichi Ohkawa, Naoto Egawa, Sawako Kuruma, Mitsuru Mori, Haruhisa Nakao, Takahisa Kawaguchi, Meiko Takahashi, Fumihiko Matsuda, Shogo Kikuchi, Keitaro Matsuo.

Supervision: Shogo Kikuchi, Keitaro Matsuo.

Writing original draft: Masahiro Nakatochi, Yingsong Lin.

Writing review & editing: Hidemi Ito, Kazuo Hara, Fumie Kinoshita, Yumiko Kobayashi, Hiroshi Ishii, Masato Ozaka, Takashi Sasaki, Naoki Sasahira, Manabu Morimoto, Satoshi Kobayashi, Makoto Ueno, Shinichi Ohkawa, Naoto Egawa, Sawako Kuruma, Mitsuru Mori, Haruhisa Nakao, Chaochen Wang, Takeshi Nishiyama, Takahisa Kawaguchi, Meiko Takahashi, Fumihiko Matsuda, Shogo Kikuchi, Keitaro Matsuo.

References

1. Lucas AL, Malvezzi M, Carioli G, Negri E, La Vecchia C, Boffetta P, et al. (2016) Global Trends in Pancreatic Cancer Mortality From 1980 Through 2013 and Predictions for 2017. *Clin Gastroenterol Hepatol* 14: 1452–1454. <https://doi.org/10.1016/j.cgh.2016.05.034> PMID: 27266982
2. Matsuo K, Ito H, Wakai K, Nagata C, Mizoue T, Tanaka K, et al. (2011) Cigarette smoking and pancreatic cancer risk: an evaluation based on a systematic review of epidemiologic evidence in the Japanese population. *Jpn J Clin Oncol* 41: 1292–1303. <https://doi.org/10.1093/jjco/hyr141> PMID: 21971423
3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. (2000) Environmental and heritable factors in the causation of cancer: analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 343: 78–85. <https://doi.org/10.1056/NEJM200007133430201> PMID: 10891514
4. Zhang M, Wang Z, Obazee O, Jia J, Childs EJ, Hoskins J, et al. (2016) Three new pancreatic cancer susceptibility signals identified on chromosomes 1q32.1, 5p15.33 and 8q24.21. *Oncotarget* 7: 66328–66343. <https://doi.org/10.18632/oncotarget.11041> PMID: 27579533
5. Klein AP, Wolpin BM, Risch HA, Stolzenberg-Solomon RZ, Mocci E, Zhang M, et al. (2018) Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nat Commun* 9: 556. <https://doi.org/10.1038/s41467-018-02942-5> PMID: 29422604
6. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* 41: 986–990. <https://doi.org/10.1038/ng.429> PMID: 19648918
7. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, et al. (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* 42: 224–227. <https://doi.org/10.1038/ng.522> PMID: 20101243
8. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, et al. (2014) Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat Genet* 46: 994–1000. <https://doi.org/10.1038/ng.3052> PMID: 25086665
9. Childs EJ, Mocci E, Campa D, Bracci PM, Gallinger S, Goggins M, et al. (2015) Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nat Genet* 47: 911–916. <https://doi.org/10.1038/ng.3341> PMID: 26098869
10. Campa D, Rizzato C, Bauer AS, Werner J, Capurso G, Costello E, et al. (2013) Lack of replication of seven pancreatic cancer susceptibility loci identified in two Asian populations. *Cancer Epidemiol Biomarkers Prev* 22: 320–328. <https://doi.org/10.1158/1055-9965.EPI-12-1182> PMID: 23250936

11. Wu C, Miao X, Huang L, Che X, Jiang G, Yu D, et al. (2012) Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat Genet* 44: 621–626. <https://doi.org/10.1038/ng.1182> PMID: 22686608
12. Low SK, Kuchiba A, Zembutsu H, Saito A, Takahashi A, Kubo M, et al. (2010) Genome-wide association study of pancreatic cancer in Japanese population. *PLoS One* 5: e11824. <https://doi.org/10.1371/journal.pone.0011824> PMID: 20686608
13. Klein AP, Lindstrom S, Mendelsohn JB, Stepilowski E, Arslan AA, Bueno-de-Mesquita HB, et al. (2013) An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. *PLoS One* 8: e72311. <https://doi.org/10.1371/journal.pone.0072311> PMID: 24058443
14. Abe M, Ito H, Oze I, Nomura M, Ogawa Y, Matsuo K (2017) The more from East-Asian, the better: risk prediction of colorectal cancer risk by GWAS-identified SNPs among Japanese. *J Cancer Res Clin Oncol* 143: 2481–2492. <https://doi.org/10.1007/s00432-017-2505-4> PMID: 28849422
15. Wen W, Shu XO, Guo X, Cai Q, Long J, Bolla MK, et al. (2016) Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. *Breast Cancer Res* 18: 124. <https://doi.org/10.1186/s13058-016-0786-1> PMID: 27931260
16. Gail MH (2008) Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst* 100: 1037–1041. <https://doi.org/10.1093/jnci/djn180> PMID: 18612136
17. Sueta A, Ito H, Kawase T, Hirose K, Hosono S, Yatabe Y, et al. (2012) A genetic risk predictor for breast cancer using a combination of low-penetrance polymorphisms in a Japanese population. *Breast Cancer Res Treat* 132: 711–722. <https://doi.org/10.1007/s10549-011-1904-5> PMID: 22160591
18. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. (2010) Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* 362: 986–993. <https://doi.org/10.1056/NEJMoa0907727> PMID: 20237344
19. Lin Y, Ueda J, Yagyu K, Ishii H, Ueno M, Egawa N, et al. (2013) Association between variations in the fat mass and obesity-associated gene and pancreatic cancer risk: a case-control study in Japan. *BMC Cancer* 13: 337. <https://doi.org/10.1186/1471-2407-13-337> PMID: 23835106
20. Inoue M, Tajima K, Hirose K, Hamajima N, Takezaki T, Kuroishi T, et al. (1997) Epidemiological features of first-visit outpatients in Japan: comparison with general population and variation by sex, age, and season. *J Clin Epidemiol* 50: 691–701. [https://doi.org/10.1016/S0950-2688\(97\)00171-1](https://doi.org/10.1016/S0950-2688(97)00171-1) PMID: 9048692
21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4: 7. <https://doi.org/10.1186/s13742-015-0047-8> PMID: 25722852
22. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2: e190. <https://doi.org/10.1371/journal.pgen.0020190> PMID: 17194218
23. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015) A global reference for human genetic variation. *Nature* 526: 68–74. <https://doi.org/10.1038/nature15393> PMID: 26432245
24. Delaneau O, Zagury JF, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10: 511–516. <https://doi.org/10.1038/nmeth.2307> PMID: 23269371
25. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. (2016) Next-generation genotype imputation service and methods. *Nat Genet* 48: 1284–1292. <https://doi.org/10.1038/ng.3656> PMID: 27571263
26. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42: D1001–D1006. <https://doi.org/10.1093/nar/gkt1229> PMID: 24316577
27. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12: 77. <https://doi.org/10.1186/1471-2105-12-77> PMID: 21414208
28. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 88: 294–303. <https://doi.org/10.1016/j.ajhg.2011.02.002> PMID: 21376301
29. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 136: E359–E386. <https://doi.org/10.1002/ijc.29210> PMID: 25220842
30. Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, et al. (2005) Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst* 97: 715–723. <https://doi.org/10.1093/jnci/dji128> PMID: 15900041

31. Risch HA, Yu H, Lu L, Kidd MS (2015) Detectable Symptomatology Preceding the Diagnosis of Pancreatic Cancer and Absolute Risk of Pancreatic Cancer Diagnosis. *Am J Epidemiol* 182: 261–268. <https://doi.org/10.1093/aje/kwv026> PMID: 26049860
32. Yu A, Woo SM, Joo J, Yang HR, Lee WJ, Park SJ, et al. (2016) Development and Validation of a Prediction Model to Estimate Individual Risk of Pancreatic Cancer. *PLoS One* 11: e0146473. <https://doi.org/10.1371/journal.pone.0146473> PMID: 26752291
33. Chatterjee N, Shi J, Garcia-Closas M (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 17: 392–404. <https://doi.org/10.1038/nrg.2016.27> PMID: 27140283
34. Lu Y, Ek WE, Whiteman D, Vaughan TL, Spurdle AB, Easton DF, et al. (2014) Most common 'sporadic' cancers have a significant germline genetic component. *Hum Mol Genet* 23: 6112–6118. <https://doi.org/10.1093/hmg/ddu312> PMID: 24943595
35. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P (2011) Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet Epidemiol* 35: 506–514. <https://doi.org/10.1002/gepi.20600> PMID: 21618606