Detecting Pathway-Based Gene-Gene and Gene-Environment Interactions in Pancreatic Cancer

Eric J. Duell[1,2,3], **Paige M. Bracci**[5], **Jason H. Moore**[2,3,4], **Robert D. Burk**[6], **Karl T. Kelsey**[7], and **Elizabeth A. Holly**[5]

[1]International Agency for Research on Cancer, Lyon, France [2]Department of Community and Family Medicine, Lebanon, New Hampshire [3]Norris Cotton Cancer Center, Lebanon, New Hampshire [4]Department of Genetics, Computational Genetics Laboratory, Dartmouth Medical School, Lebanon, New Hampshire [5]Cancer Epidemiology Studies, Department of Epidemiology and Biostatistics, University of California, San Francisco, California [6]Departments of Pediatrics, Microbiology and Immunology, Epidemiology and Population Health, and Obstetrics, Gynecology and Women's Health, Albert Einstein College of Medicine, Bronx, New York [7]Departments of Community Health and Pathology and Laboratory Medicine, Center for Environmental Health and Technology, Brown University, Providence, Rhode Island

Abstract

Data mining and data reduction methods to detect interactions in epidemiologic data are being developed and tested. In these analyses, multifactor dimensionality reduction, focused interaction testing framework, and traditional logistic regression models were used to identify potential interactions with up to three factors. These techniques were used in a population-based case-control study of pancreatic cancer from the San Francisco Bay Area (308 cases, 964 controls). From 7 biochemical pathways, along with tobacco smoking, 26 polymorphisms in 20 genes were included in these analyses. Combinations of genetic markers and cigarette smoking were identified as potential risk factors for pancreatic cancer, including genes in base excision repair (OGG1), nucleotide excision repair (XPD, XPA, XPC), and double-strand break repair (XRCC3). XPD.751, XPD.312, and cigarette smoking were the best single-factor predictors of pancreatic cancer risk, whereas XRCC3.241*smoking and OGG1.326*XPC.PAT were the best two-factor predictors. There was some evidence for a three-factor combination of OGG1.326*XPD.751*smoking, but the covariate-adjusted relative-risk estimates lacked precision. Multifactor dimensionality reduction and focused interaction testing framework showed little concordance, whereas logistic regression allowed for covariate adjustment and model confirmation. Our data suggest that multiple common alleles from DNA repair pathways in combination with cigarette smoking may increase the risk for pancreatic cancer, and that multiple approaches to data screening and analysis are necessary to identify potentially new risk factor combinations.

Introduction

As the era of single-marker and single-nucleotide polymorphism association studies in epidemiology draws to a close, methods of data analysis are in development for main gene effects, gene-gene interactions, and gene-environment interactions among multiple markers. More traditional methods of analysis for higher-order interactions (2, 3, and more loci) such as logistic regression require more programming and computations, even in epidemiologic studies of modest size with 25 or more candidate gene markers. To address these issues, a number of data mining and data filtering software packages are now or soon will be freely available to researchers (1–6). One such method, multifactor dimensionality reduction, uses data reduction methods to detect higher-order interactions when the outcome variable is categorical.

Multifactor dimensionality reduction was developed as a nonparametric method. In case-control studies of complex disease, it does not require specification of a genetic model to detect gene-gene interactions without main gene effects (1). Multifactor dimensionality reduction software has existed since 2001 and has evolved over multiple versions. The core algorithm used to collapse high-dimension data and cross-validation consistency has remained unchanged, whereas user interface and the addition of graphical features characterize recent improvements. Newer versions of multifactor dimensionality reduction have incorporated data filtering methods such as Tuned ReliefF to assist in the analysis of interactions in genome-wide association studies (see Web site: www.epistasis.org; ref. 7).

Focused interaction testing framework software was developed to identify markers for gene-gene interaction and uses a parametric search algorithm in a pooled case-control group that reduces the number of tests done. The screening algorithm uses a goodness-of-fit m^2 statistic that compares observed to expected genotype frequencies in the pooled cases and controls (assuming no marginal or main gene effects). The main interaction testing framework then uses a likelihood ratio test to simultaneously test for higher-order multilocus effects while adjusting the threshold for significance by controlling false discovery rates. Thus, multifactor dimensionality reduction and focused interaction testing framework use fundamentally different strategies to detect interactions.

Pancreatic cancer is the fourth leading cause of cancer-related death in men and women in the United States (8). With the exception of cigarette smoking, few environmental risk factors for pancreatic cancer are known, and most cases (>90%) do not aggregate in families (9). Thus, it is reasonable to suggest that most pancreatic cancers are the result of complex interactions and cross-talk between alleles or gene products and environmental exposures such as tobacco smoke.

The main objectives of this article are to identify pathway-based higher-order interactions (gene-gene and gene-smoking) important in pancreatic cancer etiology and to compare the results of multifactor dimensionality reduction and focused interaction testing framework software packages with traditional multivariable logistic regression methods using genotype data from participants in a population-based case-control study of pancreatic adenocarcinoma. From 7 biological pathways, 26 polymorphisms in 20 genes were

evaluated in this study (Table 1). The following pathways and genes were included in the analyses: base excision repair (APE1, OGG1, XRCC1), nucleotide excision repair (XPA, XPC, XPD, ERCC1), double-strand break repair (XRCC3), carcinogen metabolism and oxidant stress (GSTM1, GSTT1, GSTP1, UGT1A7, SOD2), hormone metabolism (CYP1A1, CYP1B1, CCK), inflammation (TNF-A, RANTES, CCR5), and extracellular matrix (MMP3).

Materials and Methods

Study Population

A population-based case-control study of pancreatic cancer was conducted in six San Francisco Bay Area counties (Alameda, Contra Costa, Marin, San Francisco, San Mateo, and Santa Clara) between 1994 and 2005. Detailed study methods have been published (10–20). Briefly, cases were identified using rapid case ascertainment by the Northern California Cancer Center with a goal to identify patients in the study area within 1 month of diagnosis. Eligible cases were newly diagnosed between 1995 and 1999 with adenocarcinoma of the exocrine pancreas, were between 21 and 85 years of age, resided in 1 of the 6 Bay Area counties, were alive at the time of first attempted contact, and were able to complete an interview in English. A total of 532 eligible cases completed the interview for a 67% response rate (10–12). Patient diagnoses were confirmed by participants' physicians and by the Surveillance, Epidemiology, and End Results abstracts.

Control participants were identified within the six San Francisco Bay Area counties using random-digit dial and were frequency matched to cases in an approximately 3:1 ratio by sex and 5-year age group. Eligibility criteria were identical for case and control participants, except for pancreatic cancer status. Control recruitment for those older than 65 years was supplemented by random sampling of the Health Care Finance Administration (now the Centers for Medicare and Medicaid Services) lists for the 6 Bay Area counties. A total of 1,701 eligible control participants completed the study interview for a 67% response rate (10–12).

The analyses for this study are based on 308 cases and 964 controls who gave blood as part of the laboratory portion of the study. Detailed methods on case and control selection and the laboratory portion of the study have been published (12–14, 16). Eligible participants were those who had no portacath (a medical device surgically inserted under the skin that typically is used to deliver chemotherapy for cancer patients or for patients requiring long-term parenteral nutrition) in place and had no history of bleeding disorders. Blood was not requested from the out-of-area cases. It was not obtained from the remainder of the case participants for the following reasons: patient was too ill, had died, or refused; the blood draw was unsuccessful or insufficient; or the study had ended. An analysis comparing participants who provided blood with those who did not provide blood has been previously described (14). Among cases, age, sex, race, education, smoking status (never, former, current), and pack-years of smoking were not different between those who did and did not provide a blood specimen (all P values were greater than 0.05). Blood was not requested from out-of-area controls. It was not obtained from the remainder of the control participants for the following reasons: participant refused, had died, was lost to contact, or was too ill;

the blood draw was unsuccessful or insufficient; or the study had ended. Among controls, age, education, and pack-years of smoking were independent of venipuncture (all P values were greater than 0.05), whereas those who provided blood were more likely to be white, men, and ever smokers. Overall, case or control status was not related to providing blood (P = 0.60). The study interviewers obtained separate written informed consent from all participants before interview and venipuncture. Study methods and protocols were approved by the University of California Committee on Human Research.

Exposure and demographic information were obtained from participants during in-person interviews conducted by trained interviewers using structured questionnaires. No proxy interviews were conducted. Self-reported race was broadly defined as white or Caucasian, black or African American, or Asian. Of the participants, 5 cases and 15 controls did not fall into any of these 3 categories and were classified as "other race" for these analyses. Participants were defined as never smokers if they never had smoked more than 100 cigarettes in their lifetime and had not smoked cigars or pipes at least once per month for 6 months or more. Because there was a substantial number of participants who had never smoked and who reported a history of passive smoke exposure at home as an adult (women: 32 cases, 95 controls; men: 5 cases, 21 controls), these individuals were removed from the reference group of never smokers. In analyses of smoking status, passive smokers were combined with former active smokers and pipe or cigar smokers to form three groups (never, former or passive, and current). Smoking intensity (pack-years) was defined as the number of packs of cigarettes smoked per day multiplied by the number of years smoked. For gene-smoking interaction analyses using multifactor dimensionality reduction and focused interaction testing framework, pack-years were categorized to form three groups as follows: (a) never active or passive smokers; (b) former smokers, passive smokers, pipe or cigar smokers, or less than 41 pack-years; and (c) 41 or more pack-years. Age and sex were used to determine sampling probabilities and were therefore included in multivariable logistic models.

Genotyping

All genotyping was done on germ-line DNA (f50 ng) extracted from peripheral blood lymphocytes using the QIAmp DNA Blood Mini kit (Qiagen Inc.) according to the instruction of the manufacturer. PCR-RFLP analysis was used to genotype CYP1A1 m1 (T ! C, nucleotide 6235 in 3¶ flanking region), m2 (A ! G, nucleotide 4889), and m4 (C ! A) alleles. Genotypes for GSTM1 -null (homozygous gene deletion), GSTT1 -null, XPC -PAT+ [intron 9 poly(AT)], and CCR5 -D32 (32-bp deletion) were determined using PCR amplification and visualization on agarose gels. Detailed methods and results from our earlier analyses of polymorphisms in CYP1A1, GSTM1, and GSTT1 in this subset of the San Francisco Bay Area pancreatic cancer study have been published (13). CCR5-D32 was genotyped according to a gel-based PCR method and primers published by Martinson et al. (21). XPC -PAT+ was genotyped using primers and an optimized protocol from Khan et al. (22). XRCC1.194, XPA, ERCC1, XPD, and SOD2 variants were genotyped using validated Taqman assays (Applied Biosystems).

The CYP1B1 and UGT1A7 variants were detected via allele-specific oligonucleotide hybridization using similar methods to those developed to distinguish different human papillomavirus DNA genotypes (23). The CYP1B1 (Val432Leu) and CYP1B1 (Asn453Ser) variants were examined by amplifying genomic DNA using the primers listed in the footnotes to Table 1 (24). The biotinylated probe sequences used for the CYP1B1 hybridizations are listed in footnotes to Table 1. The UGT1A7.208 variant was examined by amplifying genomic DNA using primers listed in the footnotes to Table 1. The biotinylated probe sequences used for the UGT1A7*208 hybridizations also are listed in Table 1 footnotes. All PCRs were done in a final volume of 50 AL consisting of DNA (20 ng), deoxynucleotide triphosphate (0.2 mmol/L each; Invitrogen), $MgCl_2$ (2.5 mmol/L), 1 Amol/L each primer, AmplitaqGold polymerase (1.25 U; Perkin-Elmer), and 1× reaction buffer. Amplification was done with an initial denaturation at 94**j**C for 10 minutes, followed by 35 cycles of amplification at 94**j**C for 30 seconds, 55**j**C for 1 minute and 72**j**C for 1.5 minutes, and a final extension at 72**j**C for 5 minutes using a GeneAmp 9700 thermal cycler (Perkin-Elmer). The PCR products for each of the variants were denatured and blotted in duplicate onto Biodyne B membrane filters (pore size, 0.45 Am; Pall Biodyne) using a Robbins Hydra 96. The filters were treated with 3% hydrogen peroxide solution (Sigma Chemicals) at room temperature for 15 minutes then washed at 65jC for 30 minutes {wash solution consisted of 0.1× saline – sodium phosphate – EDTA [180 mmol/L NaCl, 10 mmol/L NaH_2PO_4, and 1 mmol/L EDTA (pH 7.4)] and 0.5% sodium dodecyl sulfate}. Afterward, the treatment the filters were hybridized with the biotinlabeled probes overnight (hybridization temperatures for the CYP1B1 and UGT1A7.208 variants were 57jC and 44**j**C, respectively), followed by two 1-hour washes at the hybridization temperature. Enhanced chemiluminescence reagent (Amersham) was used to detect hybridization, followed by exposure to autoradiography. All of the results were interpreted by two experienced investigators, and discrepancies were resolved by consensus.

APE1, OGG1, XRCC1.399, GSTP1, CCK, TNF-a, RANTES, XRCC3, and MMP3 (Table 1) were genotyped using the Masscode assay (BioServe Inc.; ref. 25). Methods and results for XRCC1.399, TNF-a, and RANTES have been published (14, 16). For participants in whom the mass spectrometry method failed to yield a conclusive genotype for TNF-a and RANTES, missing data were completed using PCR-RFLP assays according to Wilson et al. (26) and Hajeer et al. (27). A random sample of the data (3%) for TNF-a -308 and RANTES -403 were repeated using Masscode and PCR-RFLP and were found to agree for both genotyping methods. DNA samples that yielded "no calls" after three genotyping attempts were reported as missing.

Statistical Methods

Only markers with less than 7% missing data in cases or controls were included in these analyses. For multifactor dimensionality reduction and focused interaction testing framework analyses, genetic markers with missing genotype values were imputed to the most common genotype for that marker. Tests for Hardy-Weinberg equilibrium among all or white or Caucasian control participants were conducted by comparing observed with expected genotype frequencies using a m^2 test with 1 degree of freedom. Expected genotype frequencies were estimated from allele frequencies. Multifactor dimensionality reduction

and focused interaction testing framework analyses were run with and without variables for smoking status or pack-years.

Multifactor Dimensionality Reduction Analysis

Multifactor dimensionality reduction version 1.0.Orc1 was used, and best models were reported for interactions with up to three factors until the total cross-validation consistency was five or more. Multifactor dimensionality reduction uses cross-validation by dividing the data into a training set (e.g., 9/10 of the data) and a testing set (e.g., the remaining 1/10 of the data) to derive estimates of cross-validation consistency and testing accuracy. Multifactor dimensionality reduction models were considered statistically significant if the testing accuracy was greater than the cutoff based on a 1,000-fold permutation test. For permutation testing, the data were randomized 1,000 times by case or control status consistent with the null hypothesis of no association (testing accuracy, 0.5). The multifactor dimensionality reduction model-fitting procedure was run for each randomized data set to determine expected values for testing accuracy. Testing accuracies greater than the expected values based on the permutated data sets were considered statistically significant at the 0.05 level. This approach also accounts for multiple hypothesis testing. High- or low-risk summary graphics provided by multifactor dimensionality reduction were used to visualize potentially interacting genotypes and to build combined variables for testing in logistic regression models. Interaction dendrograms provided by multifactor dimensionality reduction were examined to assist in the visualization and interpretation of pot ential interactions (6). Connected red or orange lines indicate genetic markers that may interact synergistically, whereas blue or green lines indicate genetic markers that are redundant or do not interact. Shorter lines or leaves indicate stronger synergistic or redundant relations between variables.

Focused Interaction Testing Framework Analysis

Focused interaction testing framework software was downloaded in July 2006 using the Web site http://hydra.usc.edu/fitf (5). Data for interactions with up to three factors were evaluated. The overall a level of 0.05 was partitioned as follows: 0.01 in the first stage and 0.02 in the second and third stages. The m^2 subset (chi-square subset) statistical cutoff values were set to three and six for these analyses. All statistically significant interaction models were reported based on the false discovery rate P value. The model with the lowest false discovery rate P value was reported if no models were statistically significant.

Unconditional Logistic Regression

Unconditional multiple logistic regression with PROC LOGISTIC in SAS (version 9.1; SAS Institute) was used to compute covariate-adjusted odds ratios and 95% confidence intervals (95% CI) for genetic factors and pancreatic cancer risk. Interactions between variables were assessed in logistic regression models by forming a new variable with a common reference group (e.g., XRCC3.241 TT+TC+ never smoker) from two or three individual variables. For gene-gene interaction variables in logistic models, either previous information on variant function (if known) or interaction graph output from multifactor dimensionality reduction software was used to form the new combined variable (high versus low risk).

Results

Among white or Caucasian control participants, the following genetic markers were not in Hardy-Weinberg equilibrium: GSTP1.105 (P = 0.00007) and CYP1A1 .m2 (P = 0.003), and, thus, were removed from further analyses. Ignoring smoking among all participants, multifactor dimensionality reduction and focused interaction testing framework identified XPD as the single most important gene, with focused interaction testing framework favoring XPD.312 and multifactor dimensionality reduction favoring XPD.751. Among white or Caucasian participants, multifactor dimensionality reduction favored XRCC3.241, and focused interaction testing framework again favored XPD.312 as the best single-locus markers.

In two-loci models among all participants, multifactor dimensionality reduction favored a possible interaction between OGG1.326*XPC.PAT and focused interaction testing framework-identified XPD.312*XRCC3.241, although neither of these was considered statistically significant (Table 2). Focused interaction testing framework also identified a possible interaction between XPD.312 and XPA among whites or Caucasians that was not identified by multifactor dimensionality reduction. Thus, potential two-loci interactions OGG1.326*XPC.PAT and XPD.312*XPA were tested in subsequent logistic regression models.

Ignoring smoking among all participants, focused interaction testing framework identified XRCC1.399*XPC. PAT*XPD.312 and OGG1.326*XPD.156*XPD.751 and multifactor dimensionality reduction identified XRCC3. 241*XPC.PAT*CYP1B1.432 and OGG1.326*XPC.PAT*XPD. 156 as potential three-loci interactions (Table 2). In whites or Caucasians, multifactor dimensionality reduction identified XPD.156*XRCC3. 24 1*CYP1B1.453 and XPD.156*XRCC3.241*XPA and focused interaction testing framework identified XPD.312*ERCC1*XRCC3.241 as potential three-loci interactions. Cross-validation consistency was low (less than five) for all two- and three-loci interactions identified by multifactor dimensionality reduction that did not include smoking.

In general, similar results were obtained from models that evaluated tobacco smoking using smoking status (never, former, current) or smoking pack-years. In these models, multifactor dimensionality reduction and focused interaction testing framework identified smoking as the most important single risk factor for pancreatic cancer (Table 3). When smoking was included in multifactor dimensionality reduction and focused interaction testing framework models, multifactor dimensionality reduction tended to include smoking in all the best two- and three-factor combinations, whereas none of the best two- and three-factor combinations identified by focused interaction testing framework included smoking (Table 3). Multifactor dimensionality reduction results suggested that OGG1 and XPD and smoking interact (consistent among all participants and among whites or Caucasians), whereas focused interaction testing framework results suggested that OGG1 and XPD interact (among all participants, with or without smoking).

Next, logistic regression models were evaluated in all participants combined and restricted to whites or Caucasians for the best potential interactions (and single risk factors) identified

by multifactor dimensionality reduction and focused interaction testing framework software programs. Logistic regression models for XPD.312*XPA, OGG1. 326* XPC.PAT, XPD. 751, XRCC3.241, and smoking first were evaluated (Table 4) based on the high- or low-risk categories for XPD.312*XPA and OGG1.326*XPC.PAT provided by multifactor dimensionality reduction software (Fig. 1A and B). In these logistic models with smoking as a main effect, XPD.751 alone and XRCC.241 alone were statistically significant predictors of pancreatic cancer risk (Table 4). The odds ratios for all of the factors evaluated in these models were statistically significant and consistent between whites or Caucasians and all participants combined (Table 4).

To determine the specific combination of genotypes that may be associated with increased risk and to evaluate the potential interactions with smoking suggested by multifactor dimensionality reduction, combined odds ratios for XPD.312*XPA, OGG1.326*XPC.PAT, and XRCC3.241 *smoking were evaluated using logistic regression (Table 5). A two-loci interaction for OGG1. 326*XPC.PAT (combined variable with six categories) showed some heterogeneity in risk across categories (Table 5). Based on this logistic regression model, there did not seem to be an interaction between XPD.312 and XPA (Table 5). Among white or Caucasian participants, a likelihood ratio test comparing a model with XRCC3.241 and smoking as main effects versus a model with these factors combined resulted in a m^2 of 4.2 (P = 0.04).

Among all participants, the likelihood ratio test m^2 value testing for interaction between XRCC3.241 and smoking was 0.87 (P = 0.35). None of the likelihood ratio test tests for XPD.312*XPA and OGG1.326*XPC.PAT were statistically significant (all P > 0.05). We ran the final confirmatory logistic regression model in whites or Caucasians separately for men and for women (data not shown). In general, the direction and magnitude of odds ratios were similar between men and women. None of the likelihood ratio test tests for XRCC3. 241 *smoking by sex (men or women) or by race or ethnicity (in whites or Caucasians or in all participants) were statistically significant (all P > 0.2). In logistic models, we found some evidence for an interaction between smoking and XPD.751 in men and not in women (data not shown). The magnitude of the combined odds ratios (for XPD.751 and smoking) was not as strong as those observed for XRCC3.241 *smoking (data not shown).

Adjusted odds ratios for the combined effect of OGG1.326*XPD.751 genotypes also were evaluated because the multifactor dimensionality reduction analysis indicated a possible interaction between OGG1.326 and XPD.751. After adjustment for age, sex, XRCC3 *smoking, and XPD.312*XPA, there was no evidence of statistically significant associations or interactions for any combined category of OGG1.326*XPD.751 and pancreatic cancer (among whites or Caucasians, or among all participants; data not shown). Odds ratios adjusted for age, sex, XPD.312*XPA, and XRCC3 for the combined effect of OGG1.326*XPD.751, stratified by smoking status (current versus never or former), were evaluated because the multifactor dimensionality reduction analysis indicated that there may be an interaction between OGG1.326 and XPD.751 and smoking (Table 3). Combined odds ratios were evaluated for OGG1.326*XPD.751 in the following six categories: C/C*A/C +C/C (reference), C/C*A/A, C/G*A/C+C/C, C/G*A/A, G/G*A/C+C/C, and G/G*A/A. Among all current smokers, adjusted odds ratios for OGG1.326*XPD.751 were 3.0 (95%

CI, 1.1–8.4), 1.3 (95% CI, 0.47–3.5), 1.2 (95% CI, 0.25–4.2), 1.8 (95% CI, 0.25–13), and 0.91 (95% CI, 0.09–10). Among all former or never smokers, adjusted odds ratios for OGG1.326*XPD.751 were 1.3 (95% CI, 0.79–2.1), 0.85 (95% CI, 0.53–1.4), 0.68 (95% CI, 0.38–1.2), 1.1 (95% CI, 0.46–2.4), and 1.3 (95% CI, 0.55–3.0). Among white or Caucasian participants, adjusted odds ratios for the combined effect of OGG1.326*XPD.751 among current smokers were 3.2 (95% CI, 1.0–9.8), 1.2 (95% CI, 0.43–3.6), 2.1 (95% CI, 0.55– 8.4), 2.1 (95% CI, 0.28–16), and 1.5 (95% CI, 0.13–17). Among white or Caucasian never or former smokers, odds ratios for the combined effect of OGG1.326*XPD. 751 were 1.3 (95% CI, 0.77–2.1), 0.85 (95% CI, 0.52–1.4), 0.55 (95% CI, 0.28–1.1), 1.0 (95% CI, 0.39– 2.6), and 1.2 (95% CI, 0.44–3.5). There was little or no evidence in logistic regression models of three-loci interactions for XPD.312*XRCC3.241*ERCC1, OGG1*XPD. 156*XPD.751, and XRCC1.399*XPC.PAT*XPC.312 (data not shown).

A graphical depiction of the combined effect of XRCC3.241 and smoking as high- and low-risk groups and statistical interactions determined by multifactor dimensionality reduction are shown (Fig. 1C). As portrayed in the dendrogram (Fig. 2), OGG1 (OGG1.326) and XPC9 (XPC.PAT) may interact (connected red lines), whereas XPD.312 and XPD.751 are redundant or do not interact (blue lines).

Discussion

Multifactor dimensionality reduction and focused interaction testing framework software programs and logistic regression models were used to evaluate pathway-based gene-gene and gene-environment interactions in pancreatic adenocarcinoma using data from a population-based case-control study in the San Francisco Bay Area. Results showed that cigarette smoking, common variants in XPD (a nucleotide excision repair gene), and XRCC3 (a double-strand break repair gene) were the most important single-factor determinants of risk for pancreatic cancer in these data. For two-factor interactions, moderately increased risk estimates were observed for the combined effects of OGG1*XPC and XRCC3 *smoking. XPC is involved in damage detection that is specific for global genomic repair (nucleotide excision repair pathway), and OGG1 or 8-oxoguanine glycosylase is involved in removing mutagenic 7,8-dihydro-8-oxoguanine lesions as a consequence of free radical oxidation and is part of the base excision repair pathway. XRCC3 is involved in double-strand break repair (specifically, homologous recombination repair), often a result of radiation- and smoking-induced DNA damage. In these analyses, the strongest interaction observed was for XRCC3.241 *smoking. Smoking is one of the only established environmental risk factors for pancreatic cancer, but the precise mechanism of action in the pancreas is presently unknown. It is reasonable to assume that the effect of tobacco smoking on pancreatic tissues is a result of a complex combination of direct and indirect action of tobacco-associated carcinogens and metabolites (e.g., nitrosamines, oxygen free radicals) that are known to damage DNA (13, 28). Furthermore, there is mounting epidemiologic evidence that DNA repair polymorphisms in combination with heavy tobacco smoking increase the risk for pancreatic cancer (14, 29, 30). Interestingly, in these analyses, multifactor dimensionality reduction and focused interaction testing framework did not identify a previously reported interaction by our group (XRCC1 *smoking; ref. 14). Potential explanations for this include differences between the analytic

methods used to detect and rank interactions, and that the previously listed interactions gave stronger signals than XRCC1 *smoking (multifactor dimensionality reduction and focused interaction testing framework software programs give lower relative rankings to weaker interactions).

For three-factor combinations, there was some evidence for an interaction between XPD *OGG1 *smoking, but estimates based on logistic regression models lacked precision. In a recent analysis of a hospital-based case-control study from the MD Anderson Cancer Center, Jiao et al. (29) reported an interaction between XPD codon 312 variants and smoking in relation to risk of pancreatic cancer. Their analysis did not evaluate polymorphisms in OGG1. Overall, our results support the hypothesis that some common genetic variants in base excision repair, nucleotide excision repair, and double-strand break repair pathways define subgroups at higher risk for smoking-associated pancreatic cancer.

Although multifactor dimensionality reduction and focused interaction testing framework identified smoking as the best single-factor predictor of pancreatic cancer in our study, unlike multifactor dimensionality reduction, focused interaction testing framework did not identify any interactions involving smoking. Multifactor dimensionality reduction and focused interaction testing framework rarely agreed on the interaction factors and may reflect the different methods each program uses to identify interactions. Focused interaction testing frame-work screens markers based on a m^2 test that compares observed with expected frequencies in a pooled group of cases and controls, the same as assuming no marginal or main gene effects. In comparison, multifactor dimensionality reduction is less constrained than focused interaction testing framework and does not estimate model parameters or interaction terms and instead does cross-validation as part of the algorithm. With multifactor dimensionality reduction, any combination of genotypes (or environmental exposures) between two or more factors that results in an excess of cases compared with controls will be considered a potential interaction. Because the biology of many genes and of interactions between genes or gene products and environmental exposures are often unknown, the more "agnostic approach" of multifactor dimensionality reduction may be more appropriate for data mining. In contrast, focused interaction testing framework may have more power to detect interactions because of the prescreening procedure implemented in pooled cases and controls.

Because of the low magnitude of combined odds ratios from logistic models based on multifactor dimensionality reduction categories of high risk versus low risk, it is unclear if these categories represent risks due to true interactions (departures from additive or multiplicative effects of individual factors) or if they represent increased risks from multiple alleles or factors from distinct or overlapping pathways. There is considerable redundancy in DNA repair pathways. Because of this, the population-level effects of weakly or moderately interacting alleles and gene products are likely to be difficult to detect and interpret. Two markers (GSTP.105 and CYP1A1.m2) were not included in the analyses because they were not in Hardy-Weinberg equilibrium in Caucasian controls. Although the precise reasons for not satisfying Hardy-Weinberg equilibrium in our study are not known, they could include genotyping error (however, we repeated genotyping on 5% of the samples and found no differences), recent mutations that have not yet reached equilibrium in our population, or an

artifact of admixture due to subgroups that differ in allele frequency. It is important to note that our study had limited power to assess interactions; thus, our results require confirmation in larger epidemiologic studies and in studies from different populations.

Multifactor dimensionality reduction and focused interaction testing framework are tools that can provide analysts with guidelines for the evaluation of interactions, but neither method can compensate for a lack of precision in the data. Although traditional logistic regression is inadequate to analyze the large amount of data that are produced by genome-wide methods, it is useful for covariate adjustment and to describe relative risks for disease in association with various combinations of genetic and environmental factors. No single method is likely to identify all of the potential interacting alleles or genotypes in a data set. Based on our experience, it seems appropriate to recommend that researchers and analysts use more than one approach to screen for potential gene combinations or interactions. Our approach was to use multifactor dimensionality reduction and focused interaction testing framework as tools to identify potential interacting candidate alleles or genotypes for more efficient testing using traditional logistic regression techniques. Our result showing an increased association for pancreatic cancer with a "checkerboard-like" combination of XPC.PAT and OGG1.326 genotypes (Fig. 1A) may not have been observed using more traditional logistic regression methods such as interaction terms and $-2 \log -$ like likelihood ratio tests. It remains to be seen whether this pattern (XPC.PAT*OGG1.326) is observed in other study populations. It is becoming apparent that some interactions and allelic effects may be context dependent (31). Mutual collaborations and exchange of ideas among epidemiologists, computational biologists, mouse geneticists, and other basic scientists are necessary to further our understanding of these potentially important processes in complex human diseases.

Acknowledgments

References

1. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001; 69:138–47. [PubMed: 11404819]

2. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. 2004; 20:640–7. [PubMed: 15522460]

3. Chatterjee N, Kalaylioglu Z, Moslehi R, et al. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet. 2006; 79:1002–16. [PubMed: 17186459]

4. Goodman JE, Mechanic LE, Luke BT, et al. Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. Int J Cancer. 2006; 118:1790–7. [PubMed: 16217767]

5. Millstein J, Conti DV, Gilliland FD, et al. A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet. 2006; 78:15–27. [PubMed: 16385446]

6. Moore JH, Gilbert JC, Tsai CT, et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol. 2006; 241:252–61. [PubMed: 16457852]

7. Moore, JH.; White, BC. Tuning ReliefF for genome-wide genetic analysis. In: Rajapakse, JC., editor. Lecture notes in computer science. Springer; New York: 2007. p. 166-75.

8. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2007. CA Cancer J Clin. 2007; 57:43–66. [PubMed: 17237035]

9. Duell, EJ.; Bracci, PM.; Holly, EA. Environmental determinants of exocrine pancreatic cancer. In: Pour, PM., editor. Toxicology of the pancreas. CRC, Taylor & Francis; Boca Raton: 2005. p. 395-422.

10. Wang F, Gupta S, Holly EA. Diabetes mellitus and pancreatic cancer in a population-based case-control study in the San Francisco Bay Area, California. Cancer Epidemiol Biomarkers Prev. 2006; 15:1458–63. [PubMed: 16896032]

11. Eberle CA, Bracci PM, Holly EA. Anthropometric factors and pancreatic cancer in a population-based case-control study in the San Francisco Bay Area. Cancer Causes Control. 2005; 16:1235–44. [PubMed: 16215874]

12. Holly EA, Eberle CA, Bracci PM. Prior history of allergies and pancreatic cancer in the San Francisco Bay area. Am J Epidemiol. 2003; 158:432–41. [PubMed: 12936898]

13. Duell EJ, Holly EA, Bracci PM, et al. A population-based, case-control study of polymorphisms in carcinogen-metabolizing genes, smoking, and pancreatic adenocarcinoma risk. J Natl Cancer Inst. 2002; 94:297–306. [PubMed: 11854392]

14. Duell EJ, Holly EA, Bracci PM, et al. A population-based study of the Arg [399]Gln polymorphism in x-ray repair cross-complementing group 1 (XRCC1) and risk of pancreatic adenocarcinoma. Cancer Res. 2002; 62:4630–6. [PubMed: 12183419]

15. Duell EJ, Holly EA. Reproductive and menstrual risk factors for pancreatic cancer: a population-based study of San Francisco Bay Area women. Am J Epidemiol. 2005; 161:741–747. [PubMed: 15800266]

16. Duell EJ, Casella DP, Burk RD, et al. Inflammation, genetic polymorphisms in proinflammatory genes TNF-A, RANTES, CCR5, and risk of pancreatic adenocarcinoma. Cancer Epidemiol Biomarkers Prev. 2006; 15:726–31.

17. Hoppin JA, Tolbert PE, Holly EA, et al. Pancreatic cancer and serum organochlorine levels. Cancer Epidemiol Biomarkers Prev. 2000; 9:199–205. [PubMed: 10698482]

18. Slebos RJ, Hoppin JA, Tolbert PE, et al. K-ras and p53 in pancreatic cancer: association with medical history, histopathology, and environmental exposures in a population-based study. Cancer Epidemiol Biomarkers Prev. 2000; 9:1223–32. [PubMed: 11097231]

19. Holly EA, Chaliha I, Bracci PM, et al. Signs and symptoms of pancreatic cancer: a population-based case-control study in the San Francisco Bay area. Clin Gastroenterol Hepatol. 2004; 2:510–7. [PubMed: 15181621]

20. Chan JM, Wang F, Holly EA. Vegetable and fruit intake and pancreatic cancer in a population-based case-control study in the San Francisco Bay area. Cancer Epidemiol Biomarkers Prev. 2005; 14:2093–7. [PubMed: 16172215]

21. Martinson JJ, Chapman NH, Rees DC, et al. Global distribution of the CCR5 gene 32-basepair deletion. Nat Genet. 1997; 16:100–3. [PubMed: 9140404]

22. Khan SG, Metter EJ, Tarone RE, et al. A new xeroderma pigmentosum group C poly(AT) insertion/deletion polymorphism. Carcinogenesis. 2000; 21:1821–5. [PubMed: 11023539]

23. Jiang G, Qu W, Ruan H, et al. Elimination of false-positive signals in enhanced chemiluminescence (ECL) detection of amplified HPV DNA from clinical samples. Biotechniques. 1995; 19:566–8. [PubMed: 8777045]

24. Bailey LR, Roodi N, Dupont WD, et al. Association of cytochrome P450 1B1 (CYP1B1) polymorphism with steroid receptor status in breast cancer. Cancer Res. 1998; 58:5038–41. [PubMed: 9823305]

25. Kokoris M, Dix K, Moynihan K, et al. High-throughput SNP genotyping with the Masscode system. Mol Diagn. 2001; 5:329–340. [PubMed: 11172497]

26. Wilson AG, di Giovine FS, Blakemore AI, et al. Single base polymorphism in the human tumour necrosis factor alpha (TNF alpha) gene detectable by NcoI restriction of PCR product. Hum Mol Genet. 1992; 1:353. [PubMed: 1363876]

27. Hajeer AH, al Sharif F, Ollier WE. A polymorphism at position-403 in the human RANTES promoter. Eur J Immunogenet. 1999; 26:375–6. [PubMed: 10553505]

28. Schuller HM. Mechanisms of smoking-related lung and pancreatic adenocarcinoma development. Nat Rev Cancer. 2002; 2:455–63. [PubMed: 12189387]

29. Jiao L, Hassan MM, Bondy ML, et al. The XPD Asp312Asn and Lys751Gln polymorphisms, corresponding haplotype, and pancreatic cancer risk. Cancer Lett. 2007; 245:61–8. [PubMed: 16458430]

30. Jiao L, Bondy ML, Hassan MH, et al. Selected polymorphisms of DNA repair genes and risk of pancreatic cancer. Cancer Det Prev. 2006; 30:284–91.

31. To MD, Perez-Losada J, Mao JH, et al. A functional switch from lung cancer resistance to susceptibility at the Pas1 locus in Kras2LA2 mice. Nat Genet. 2006; 38:926–30. [PubMed: 16823377]
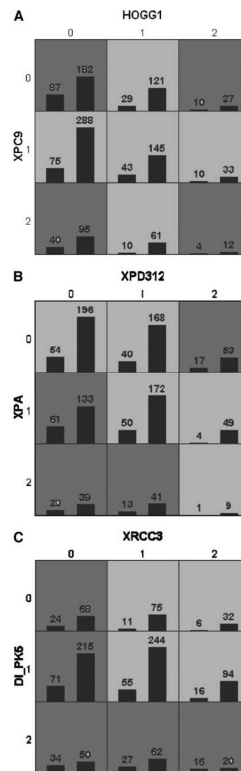
Figure 1.
Multifactor dimensionality reduction interaction diagrams. Darker-shaded cells show higher-risk combinations, whereas lighter-shaded cells show combinations not associated with elevated risk for pancreatic cancer. 0, homozygous for common allele; 1, heterozygous; 2, homozygous for minor allele. Case-control ratio was 1:3. Data from a population-based case-control study of pancreatic cancer in the San Francisco Bay Area, CA (1995–1999). A. XPC.PAT × OGG1.326 among all participants. B. XPD.312 × XPA among all participants. C. XRCC3.241 × smoking among whites or Caucasians.

Figure 2.
Multifactor dimensionality reduction interaction dendrogram (among all participants, without smoking). Shorter connections among nodes mean stronger synergistic (red and orange) or redundant (blue and green) interactions.

Table 1

Polymorphisms included in multifactor dimensionality reduction and focused interaction testing framework interaction analyses of pancreatic cancer in a population-based case-control study in the San Francisco Bay Area, CA (1995–1999)
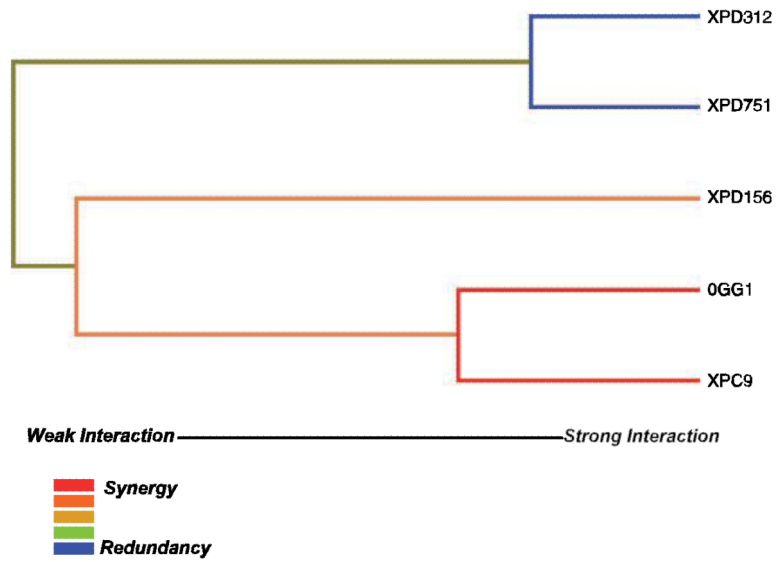
Pathway	Marker	Nucleotide position	Variation codon	Minor allele	All participants[*]		White or Caucasian participants[c]	
					Frequency		Frequency	
					Cases	Controls	Cases	Controls
Base excision repair								
APE1	rs3136820	2197	T/G (148)	G	0.43	0.44	0.44	0.44
hOGG1	rs1052133	8055	C/G (326)	G	0.21	0.24	0.19	0.23
XRCC1	rs25487	28152	G/A (399)	A	0.32	0.30	0.34	0.31
	rs1799782	26304	C/T (194)	T	0.08	0.08	0.08	0.06
Nucleotide excision repair								
XPA	rs1800975	−4 (5¶UTR)	G/A	A	0.34	0.32	0.35	0.31
XPC	AF076952	Intron 9	Poly(AT)	PAT+	0.38	0.42	0.39	0.43
ERCC1	rs3212986	8092	C/A	A	0.24	0.26	0.23	0.25
XPD	rs238406	22541	A/C (156)	C	0.46	0.43	0.50	0.45
	rs1799793	23591	G/A (312)	A	0.26	0.33	0.28	0.35
	rs13181	35931	A/C (751)	C	0.29	0.36	0.31	0.38
Double-strand break repair								
XRCC3	rs861539	18067	C/T (241)	T	0.31	0.37	0.32	0.39
Carcinogen metabolism								
GSTM1	—		Present/null	Null/null	0.52	0.53	0.53	0.53
GSTT1	—		Present/null	Null/null	0.20	0.19	0.17	0.17
GSTP1	rs1695	105	A/G	G	0.31	0.32	0.32	0.33
UGT1A7[b]	rs11692021	208	T/C	C	0.34	0.36	0.38	0.37
SOD2	—	47	C/T	T	0.48	0.48	0.48	0.50
Hormone metabolism								
CYP1A1	rs4646903	m1	T/C	C	0.14	0.14	0.12	0.13
	rs1048943	M2	A/G	G	0.06	0.07	0.05	0.06
	rs1799814	M4	C/A	A	0.03	0.04	0.04	0.05
CYP1B1[x]	rs1056836	432	C/G (432)	G	0.42	0.41	0.42	0.41
	rs1800440	453	A/G (453)	G	0.18	0.17	0.21	0.18
CCK	—	−45	C/T	T	0.14	0.12	0.12	0.12
Inflammation								
TNF-A	rs1800629	−308 (−487)	G/A	A	0.13	0.14	0.14	0.14
RANTES	rs2107538	−403 (−470)	G/A	A	0.19	0.20	0.17	0.18
CCR5	—		D32 (32-bp del)	Del	0.11	0.09	0.13	0.10
Extracellular matrix								
MMP3	—	−1171	5A/6A	A Ins	0.46	0.48	0.49	0.48

NOTE: Main gene effects (based on logistic regression models) for XRCC1.399, GSTM1, GSTT1, CYP1A1, TNF-a, RANTES, and CCR5 have been previously published 13, 14, 16. For GSTM1 and GSTT1, the frequency reported is for homozygous (null/null) deletion genotypes. UTR, untranslated region; del, deletion.

[*]308 cases and 964 controls.

[c]206 cases and 860 controls.

[b]UGT1A7 primers: 5¶-CATTGCGAAGTGCATTTTCT-3¶ and 5¶-GGGTTTGGAGAATTTCAGAGG-3¶. UGT1A7 biotinylated probe sequences: 208Trp-5¶-AGGAGAGAGTATTGAACCACA-3¶ and 208Arg-5¶-AGGAGAGAGTACTGAACCACA-3¶.

[x]CYP1B1 (Val432Leu) and CYP1B1 (Asn453Ser) primers: 5¶-GTGGTTTTTGTCAACCAGTGG-3¶ and 5¶-GCCCACTGAAAAAATCATCACTCT- GCTGGTCAGGTGC-3¶. CYP1B1 biotinylated probe sequences: 432Val, 5¶-TCATGACCCAGGAAGTGGCC-3¶; 432Leu, 5¶-TCATGACCCAGGA- AGTGGCC-3¶; 453Asn, 5¶-GGCCTCATCAGAAGGACCTG-3¶; and 453Ser, 5¶-GGCCTCATCAGAAGGACCTG-3¶.

Table 2

Gene interactions identified by multifactor dimensionality reduction and focused interaction testing framework in a population-based case-control study in the San Francisco Bay Area, CA (1995–1999)

	MDR	CVC	Testing accuracy[*]	FITF	FDR P
All participants[c]	Accuracy, z0.55 (P < 0.05)				
1 locus	XPD.751	5	0.55[*]	XPD.312	0.0015
2 loci	OGG1.326_XPC.PAT	4	0.54	XPD.312_XRCC3.241	0.0018
	OGG1.326_XPD.751	3	0.56[*]		
3 loci	XRCC3.241_XPC.PAT_CYP1B1.432	3	0.53	OGG1.326_XPD.156_XPD.751	0.00028[b]
	OGG1.326_XPC.PAT_XPD.156	2	0.54	XRCC1.399_XPC.PAT_XPD.312	0.00032[b]
White or Caucasian participants[x]	Accuracy, z0.56 (P < 0.05)				
1 locus	XRCC3.241	5	0.55	XPX.312	0.0042
2 loci	OGG1.326_XPD.751	3	0.55	XPD.312XPA	0.00040[b]
	OGG1.326_XPC.PAT	2	0.57[*]		
3 loci	XPD.156_XRCC3.241_CYP1B1.453	3	0.56[*]	XPD.312_ERCC1_XRCC3.241	0.00043
	XPD.156_XRCC3.241_XPA	2	0.55		

NOTE: For cross-validation consistencies less than five, the best multifactor dimensionality reduction models are shown until total cross-validation consistencies are five or more. Abbreviations: MDR, multifactor dimensionality reduction; CVC, cross-validation consistency; FITF, focused interaction testing framework; FDR, false discovery rate.

[*] Multifactor dimensionality reduction model is considered statistically significant (P < 0.05) if testing accuracy is greater than testing accuracy cutoff based on 1,000-fold permutation test for up to three-factor interactions.

[c] 308 cases and 964 controls.

[b] Focused interaction testing framework results reported for statistically significant or lowest false discovery rate P value. Statistically significant factors using focused interaction testing framework had false discovery rate P values less than the false discovery rate cutoff P value (all participants false discovery rate cutoff P values: one-factor P = 0.00037, two-factor P = 0.00019, three-factor P = 0.00034; white or Caucasian participants false discovery rate cutoff P values: one-factor P = 0.00037, two-factor P = 0.00075, three-factor P = 0.00028).

[x] 206 cases and 860 controls.

Table 3

Gene-smoking interactions identified by multifactor dimensionality reduction and focused interaction testing framework in a population-based case-control study in the San Francisco Bay Area, CA (1995–1999)

	MDR	CVC	Testing accuracy[*]	FITF	FDR P
All participants[c]	Accuracy z0.55 (P < 0.05)				
1 factor	Smoking	10	0.56[*]	Smoking[b]	0.000002[x]
2 factors	XPD.751_smoking	7	0.58[*]	XPD.312_XRCC3.241	0.0018
3 factors	OGG1.326_XPD.751 smoking	9	0.60[*]	OGG1.326_XPD.156_XPD.751	0.00028
White or Caucasian participants[k]	Accuracy z0.56 (P < 0.05)				
1 factor	Smoking[b]	10	0.57[*]	Smoking[b]	0.000014[x]
2 factors	XRCC3.241_smoking	7	0.58[*]	XPD.312_XPA	0.00040[b]
3 factors	OGG1.326XPD.751_smoking	6	0.60[*]	XPD.312_ERCC1_XRCC3.241	0.00043

NOTE: For cross-validation consistencies less than five, the best multifactor dimensionality reduction models are shown until total cross-validation consistencies are 5 or more.

[*] Multifactor dimensionality reduction model is considered statistically significant (P < 0.05) if testing accuracy is greater than testing accuracy cutoff based on 1,000-fold permutation test for up to three-factor interactions.

[c] 308 cases and 964 controls.

[b] Smoking defined as the following three categories: never active or passive smoker; former, passive, pipe, cigars, or less than 41 pack-years; or 41+ pack-years.

[x] Focused interaction testing framework results reported for statistically significant or lowest false discovery rate P value. Statistically significant factors using focused interaction testing framework had false discovery rate P values less than the false discovery rate cutoff P value (all participants false discovery rate cutoff P values: one-factor P = 0.00036, two-factor P = 0.00018, three-factor P = 0.00024; white or Caucasian participants false discovery rate cutoff P values: one-factor P = 0.00036, two-factor P = 0.00072, three-factors P = 0.00025).

[k] 206 cases and 860 controls.

Table 4

Logistic regression model of pancreatic cancer risk based on multifactor dimensionality reduction and focused interaction testing framework results in a population-based case-control study in the San Francisco Bay Area, CA (1995–1999)

Variable	All participants [308 cases, 964 controls; OR[*] (95% CI)]	White or Caucasian participants [260 cases, 860 controls; OR[c] (95% CI)]
XPD.312[*]XPA		
Low-risk genotypes (based on MDR)	1.0 (reference)	1.0 (reference)
High-risk genotypes (based on MDR)	1.6 (1.2–2.1)	1.8 (1.3–2.5)
hOGG1.326[*]XPC intron 9 poly(AT)		
Low-risk genotypes (based on MDR)	1.0 (reference)	1.0 (reference)
High-risk genotypes (based on MDR)	1.7 (1.2–2.2)	1.7 (1.2–2.3)
XPD.751		
A/A (Lys/Lys)	1.0 (reference)	1.0 (reference)
A/C (Lys/Gln)	0.80 (0.59–1.1)	0.80 (0.57–1.1)
C/C (Gln/Gln)	0.59 (0.37–0.94)	0.59 (0.36–0.96)
XRCC3.241		
C/C (Thr/Thr)	1.0 (reference)	1.0 (reference)
C/T (Thr/Met)	0.61 (0.44–0.83)	0.56 (0.40–0.78)
T/T (Met/Met)	0.69 (0.45–1.0)	0.63 (0.40–0.98)
Smoking		
Never active or passive	1.0 (reference)	1.0 (reference)
Former, passive, pipes, cigars	1.3 (0.88–1.9)	1.2 (0.75–1.8)
Current	2.8 (1.7–4.5)	2.8 (1.7–4.6)

Abbreviation: OR, odds ratio.

[*]Adjusted for all factors listed plus age, sex, and race or ethnicity.

[c]Adjusted for all factors listed plus age and sex.

Table 5

Confirmatory logistic regression model of pancreatic cancer risk in a population-based case-control study in the San Francisco Bay Area, CA (1995–1999)

Variable	All participants [308 cases, 964 controls; OR* (95% CI)]	White or Caucasian participants [260 cases, 860 controls; OR^c (95% CI)]
XPD.312*XPA		
G/A+A/A*G/G	1.0 (reference)	1.0 (reference)
G/G*G/G	0.86 (0.52–1.4)	0.84 (0.49–1.4)
G/A+A/A*G/A	0.73 (0.47–1.1)	0.78 (0.49–1.2)
G/G*G/A	1.2 (0.72–2.0)	1.3 (0.77–2.3)
G/A+A/A*A/A	0.94 (0.47–1.9)	1.1 (0.54–2.2)
G/G*A/A	1.5 (0.76–2.9)	1.8 (0.88–3.8)
hOGG1.326*XPC intron 9 poly(AT)		
C/G+G/G*PAT+/PAT+	1.0 (reference)	1.0 (reference)
C/C*PAT+/PAT+	2.3 (1.1–4.8)	2.7 (1.2–6.0)
C/G+G/G*PAT+/PAT−	1.6 (0.83–3.3)	1.8 (0.85–4.0)
C/C*PAT+/PAT−	1.4 (0.74–2.8)	1.7 (0.82–3.6)
C/G+G/G*PAT−/PAT−	1.3 (0.64–2.7)	1.8 (0.79–4.0)
C/C*PAT−/PAT−	2.5 (1.3–4.9)	2.9 (1.41–6.2)
XPD.751		
A/A (Lys/Lys)	1.0 (reference)	1.0 (reference)
A/C (Lys/Gln)	0.82 (0.55–1.2)	0.80 (0.53–1.2)
C/C (Gln/Gln)	0.65 (0.37–1.1)	0.67 (0.37–1.2)
XRCC3.241* smoking status		
T/T+T/C*never active or passive	1.0 (reference)	1.0 (reference)
C/C*never active or passive	2.0 (0.97–14.0)	2.4 (1.1–5.2)
T/T+T/C*former, passive, pipes, cigars	1.5 (0.85–2.7)	1.5 (0.81–2.7)
C/C*former, passive, pipes, cigars	2.3 (1.3–4.1)	2.1 (1.1–4.0)
T/T+T/C* current	3.1 (1.6–6.0)	2.8 (1.4–5.7)
C/C* current	5.2(2.6–11)	7.0 (3.3–15)

*
Adjusted for all factors listed plus age, sex, and race or ethnicity (34 cases and 48 controls had missing data for one or more variables).

^{c}Adjusted for all factors listed, plus age and sex (31 cases and 43 controls had missing data for one or more variables).