Nucleotide Excision Repair Pathway Polymorphisms and Pancreatic Cancer Risk: Evidence for role of *MMS19L*

Robert R. McWilliams,[1] William R. Bamlet,[2] Mariza de Andrade,[2] David N. Rider,[2] Julie M. Cunningham,[3] and Gloria M. Petersen[2]

Departments of [1]Oncology, [2]Health Sciences Research, and [3]Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota

Abstract

Background: **Nucleotide excision repair is a vital response to DNA damage, including damage from tobacco exposure. Single nucleotide polymorphisms (SNP) in the nucleotide excision repair pathway may encode alterations that affect DNA repair function and therefore influence the risk of pancreatic cancer development.**
Methods: **A clinic-based case-control study in non-Hispanic white persons compared 1,143 patients with pancreatic adenocarcinoma with 1,097 healthy controls. Twenty-seven genes directly and indirectly involved in the nucleotide excision repair pathway were identified and 236 tag-SNPs were selected from 26 of these (one had no SNPs identified). Association studies were done at the gene level by principal components analysis, whereas recursive partitioning analysis was utilized to identify potential gene-gene and gene-environment interactions within the pathway. At the individual SNP level, adjusted additive, dominant, and recessive models were investigated, and gene-environment interactions were also assessed.**
Results: **Gene level analyses showed an association of the *MMS19L* genotype (chromosome 10q24.1) with altered pancreatic cancer risk (P = 0.023). Haplotype analysis of *MMS19L* also showed a significant association (P = 0.0132). Analyses of seven individual SNPs in this gene showed both protective and risk associations for minor alleles, broadly distributed across patient subgroups defined by smoking status, sex, and age.**
Conclusion: **In a candidate pathway SNP association study analysis, common variation in a nucleotide excision repair gene, *MMS19L*, was associated with the risk of pancreatic cancer.** (Cancer Epidemiol Biomarkers Prev 2009;18(4):1295–302)

Introduction

DNA repair is a key mechanism in the function of human cells in response to DNA-damaging stimuli and consequent progression to cancer. It has also become an area of intense research in the study of genetic predisposition to pancreatic cancer, because mutations in genes involved with DNA repair, such as *BRCA1* and *BRCA2*, are known to increase the risk of pancreatic adenocarcinoma (1, 2). However, mutations in high-penetrance tumor suppressor genes explain only a small number (<5%) of cases of pancreatic cancer (3). In an effort to further characterize genetic risk of pancreatic cancer, the role of more common genetic variations (i.e. polymorphisms) has been increasingly studied.

Nucleotide excision repair represents a pathway involved in the detection and repair of DNA base damage such as pyrimidine dimers and bulky adducts, most notably those caused by environmental exposures such as UV light and chemical exposures (e.g., carcinogens; ref. 4). High-penetrance defects in this pathway in the *XPA*, *ERCC3/XPB*, *XPC*, *ERCC2/XPD*, *XPE*, and *ERCC5* genes have been implicated in the recessive clinical disorder xeroderma pigmentosum (5, 6), resulting in up to 1,000- to 2,000-fold increased risk for cutaneous malignancy as a result of UV damage in skin cells. Affected persons are also at increased risk of cancers of the brain and oral cavity at a young age (7). Cockayne syndrome (*ERCC8/CKN1/CSA*, *ERCC6/CSB*), an autosomal recessive severe developmental disorder with photosensitivity, is not known to confer increased cancer risk, although affected individuals often die in childhood of infectious causes, so lifelong cancer risk is unknown (8).

The nucleotide excision repair pathway consists of several primary steps that locate the damage, unwind the DNA duplex around the site, place incisions in the DNA upstream and downstream of the damage, and repair the gap (9, 10). Specifically, the protein XPC, bound to RAD23B, recognizes and binds to the damage. Next, several other proteins bind in a complex (RPA, XPA, GTF2H, MMS19L, and XPG) that unwinds the DNA helix, and the complex is then bound by ERCC1 and ERCC4/XPF which excise a 27- to 30-nucleotide fragment about the area of damage. DNA polymerases then repair the defect (4).

The importance of this pathway in carcinogenesis is suggested by prior associations of polymorphic variants with risk of certain cancers, especially tobacco-related cancers such as head/neck and lung cancer (11). Interactions between nucleotide excision repair polymorphisms and smoking have also been reported (12, 13). One potential mechanism for this is a reported

Received 11/20/08; revised 1/13/09; accepted 1/26/09; published OnlineFirst 3/24/09.
Grant support: National Cancer Institute CA K07 116303 (R. McWilliams.), P50 CA 102701 (G. Petersen).
Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).
Requests for reprints: Robert McWilliams, Gonda 10 South, 200 First St., SW, Rochester, MN 55905. Phone: 507-284-8432; Fax: 507-284-1803.
E-mail: Mcwilliams.robert@mayo.edu

doi:10.1158/1055-9965.EPI-08-1109

direct inhibition of nucleotide excision repair by tobacco smoke (14).

The effects of nucleotide excision repair gene polymorphisms and haplotypes have been shown to correlate with altered DNA repair capacity in some genes such as *ERCC1* and *ERCC2/XPD* (15), but conferred risk of pancreatic cancer by variation in the nucleotide excision repair pathway has not been definitively answered, with largely candidate SNP studies reported to date using relatively small sample sizes (16-19). Because candidate SNP studies inherently miss substantial variations in genes, we chose to do a systematic tag-SNP approach to the nucleotide excision repair pathway. The intent of such an approach is to use existing knowledge of linkage disequilibrium from HapMap (20) to comprehensively assess common variation in all identified genes in the pathway of interest. Using this approach, we did a case-control analysis utilizing the Mayo Clinic Biospecimen Resource for Pancreas Research.

Materials and Methods

Cases. This study was approved by the Mayo Clinic Institutional Review Board. Written, informed consent was obtained from each subject for participation in this study and provision of a blood sample. From October 2000 through March 2007, patients with pancreatic adenocarcinoma (ICD-O site codes C25.0-C25.3, C25.7, C25.9 and morphology codes 8140/3, 8140/6) were consecutively recruited to a registry (ultrarapid recruitment) during their visit to the Mayo Clinic (Rochester, Minnesota or Jacksonville, Florida). Ultrarapid recruitment is defined as recruitment at the time of clinic visit for the initial work-up for pancreatic cancer. Patients were identified by review of appointment calendars and pathology records, then approached by a study coordinator during a clinic visit or, if missed, contacted by mail. Of these, 71% consented to participate in the study. All records were reviewed and 1,949 were confirmed as pancreatic adenocarcinoma by a physician specialist (R.M.) in gastrointestinal medical oncology. Invasive intraductal papillary mucinous neoplasms, when identified by surgical pathology or clinical diagnosis, were excluded (n = 42). Eighty-seven percent of consenting participants provided blood samples for DNA analysis and 64% self-completed risk factor questionnaires specifically for pancreatic cancer. For those not completing questionnaires, data on clinical variables [smoking, body mass index (BMI), family history, race, ethnicity] were extracted from electronic and paper clinical records and death certificates by a single physician (R.M.). This data extraction step was assessed for intermethod reliability with 25 cases and 25 controls who completed questionnaires. For this study, 1,203 patients with pancreatic adenocarcinoma of all stages were initially included, representing 62% of all pancreatic adenocarcinoma patients identified at Mayo Clinic during this time period. Of these, 1,143 (95%) were non-Hispanic whites, so in order to prevent population stratification, analyses were limited to this demographic group. Ninety-six percent of cases had histologic confirmation of their diagnosis, with the remainder meeting the following criteria: having a pancreatic mass visualized on imaging and at least two of the following: elevated CA19-9, jaundice, weight loss, or abdominal pain. Upon enrollment, a risk factor and family history questionnaire was completed by the patient. Peripheral blood was collected for DNA analysis.

Controls. From May 2004 to February 2007, 1,511 control patients out of a total of 2,707 approached (56%) were recruited from the General Internal Medicine clinic at Mayo Clinic (Rochester) at the time of a general physical exam. Controls were attempted to be frequency-matched to cases on sex, residence [Olmsted County, Minnesota; three-state (MN, WI, IA); five-state area (MN, WI, IA, SD, ND); or outside of area], age at time of recruitment (in 5-year increments), and race/ethnicity. Controls with prior diagnoses of cancer except non-melanoma skin cancer were excluded. Upon enrollment, controls completed an equivalent risk factor and family history questionnaire to those administered to cases. Peripheral blood was collected for DNA analysis. For this study, 1,097 non-Hispanic white controls were randomly selected from those controls providing blood samples and completing questionnaires, using strata delineating age (in 5-year increments), sex, and location of residence to best approximate cases on a frequency-matching basis.

Study participants provided information about age at initiation and cessation of smoking and the number of packs smoked per day. If no smoking data were available from the self-completed questionnaire, smoking information was extracted from the participant's medical record (data were extracted for 24% of controls and 23% of cases). Smoking data were available for 99.7% of study participants. Total number of pack-years was calculated by multiplying the typical number of packs smoked daily with the number of years smoked. Pack-years were used as a measure of smoking exposure. Subjects were categorized as ''never smokers'' and ''ever smokers'' (>100 cigarettes in their lifetime). Ever smokers were further stratified by number of pack-years (≤20 pack-years, >20-40 pack-years, and >40 pack-years).

SNP Selection. Genes encoding proteins involved with the nucleotide excision repair pathway were selected from a review of the literature (21). In order to comprehensively assess common genetic variation in the genes selected, a linkage disequilibrium–based tag-SNP strategy was employed. To select linkage disequilibrium tag SNPs for the genes, genotype data from white populations were compiled from three different sources. Gene coordinates were calculated based on NCBI Build 36. For all but three genes, coordinates were calculated from the UCSC Genome Browser knownGene and knownToLocusLink tables. The coordinates for the other three genes were calculated from the gene2refseq file from the NCBI FTP site. One genome-wide genotyping project, Hapmap (http://www.hapmap.org) and two resequencing projects, SeattleSNPs (http://pga.mbt.washington.edu/) and NIEHS SNPs (http://egp.gs.washington.edu/), were utilized. We ran ldSelect software (Version 1.0; ref. 22) for SNP selection on each gene including 5 kb upstream/downstream using criteria of r^2 = 0.9 and minor allele frequency (maf) ≥0.05. We selected 3 tag SNPs for bins of size 30 or more, 2 tag SNPs for bins of size 10 or more, and 1 tag SNP otherwise. For genes with multiple sources, the optimal source of SNPs for each gene was chosen based on the most number of

linkage disequilibrium bins and most number of SNPs in each linkage disequilibrium bin. All known genes directly and indirectly involved in the nucleotide excision repair pathway were identified (N = 27), and 236 SNPs were selected. (No tag-SNPs were identified in *GTF2H2)*.

Genotyping. DNA samples were analyzed in the Mayo Clinic Genotyping Shared Resource using an Illumina Golden Gate Custom 768-plex OPA panel using the standard protocol. We selected SNPs with an Illumina design score of >0.4. BeadStudio II software was used to analyze the data and prepare reports. Cases and controls were intermixed on plates. Genotyping was successful for 1,189 cases and 1,126 controls, with a 99.7% average loci call rate. Locus success rate was 95.1% and sample success rate was 99.6%. Preset rules for dropping SNPs were poorly defined clusters, replicate or Mendelian errors, call rate < 90%, all samples heterozygous.

Quality Control. Positive and negative controls were run in parallel to ensure there was no contamination of the DNA. Other quality control measures included the addition of 56 CEPH family trios to the genotyping plates to test for non-Mendelian inheritance with 100% reproducibility and no Mendelian errors. Ten samples had low GenCall scores (<0.4; ref. 23) and were excluded from the analysis. All genotype clusters were manually inspected by a specialist scientist (JC); those with atypical clustering SNPs were flagged and excluded (n = 3 SNPs, 1 in ERCC5 and 2 in RPA3). Call rates were high for SNPs overall, at a 99.6% rate for samples and 95.1% for loci. Forty-seven pairs were used for duplicate concordance, with a 99.9% concordance rate. Twelve SNPs failed to amplify or were discarded due to poor quality and 91 samples had a call rate of 0.

Statistical Methods. Risk factor questionnaires were completed by 100% of controls and 71% of cases. For cases missing risk factor questionnaires, clinical data were extracted from available medical records as described above. To assess intermethod reliability between these two methods, we used the $\hat{E}$ coefficient to measure the inter-rater agreement (24).

Before analysis of disease-marker associations was done, we used χ^2 tests to determine whether the genotype distributions for each SNP showed Hardy-Weinberg equilibrium under Mendelian biallelic expectations.

For each polymorphism, we defined the major allele as the most common allele in controls, and the minor allele as the less common allele in controls. In order to examine the association between each SNP and disease we considered multiple unadjusted models [allelic, Cochran Armitage trend, genotypic (2df), additive, codominant, dominant, and recessive] among cases and controls using a combination of PLINK v0.99r (http://pngu.mgh.harvard.edu/purcell/plink/; ref. 25) and SAS (SAS software, version 9.1.2). Multivariable logistic analyses adjusted for age, sex, smoking status (ever/never), family history of pancreas cancer in a first-degree relative (yes/no), BMI, and personal history of diabetes (yes/no) were then done in the three different genetic models as well. (SAS software, version 9.1.2).

Table 1. Demographic and clinical characteristics of cases and controls

Variable	Cases (n = 1,143)	Controls (n = 1,097)	P*
Age at diagnosis (cases) or study entry (controls) in y (±SD)	65.5 (±10.7)	65.6 (±10.8)	0.79
Age <60 y	329 (29%)	297 (27%)	0.37
Male	668 (58%)	557 (51%)	<0.001
Non-Hispanic whites †	1,143 (100%)	1,097 (100%)	
Ever-smoker	682 (60%)	505 (46%)	
Smoking status			<0.001
Never-smoker ‡	455 (40%)	592 (54%)	
Former smoker	527 (47%)	458 (42%)	
Current smoker	148 (13%)	41 (4%)	
Missing	13	6	
Years smoked (±SD)	22.4 (±16.9)	18.2 (±14.0)	<0.001
Pack-years smoked (±SD)	17.0 (±23.0)	9.3 (±17.2)	<0.001
BMI in kg/m^2 (±SD)	27.8 (±5.5)	27.2 (±4.7)	0.010
Region			<0.001
MN, IA, or WI (Tristate)	579 (51%)	748 (68%)	
North or South Dakota	94 (8%)	40 (4%)	
Other USA	448 (39%)	308 (28%)	
Other country	22 (2%)	1 (0%)	
Diabetes mellitus (DM)			<0.001
No	801 (70%)	1,008 (92%)	
Yes	342 (30%)	89 (8%)	
DM (>2 y before pancreatic cancer dx)	224	0	
Pancreas cancer stage at enrollment			
Resectable	328 (29%)	0	
Locally advanced	379 (33%)	0	
Metastatic	430 (38%)	0	
NOS	6 (1%)	0	
Family history of pancreatic cancer (first-degree)	79 (7%)	43 (4%)	0.002

Abbreviation: NOS, not otherwise specified.
* P unadjusted.
† Only non-Hispanic whites included in the analysis.
‡ Defined as <100 cigarettes in lifetime.

A principal components analysis (26) approach was utilized in order to test for an overall association between disease and the multiple SNPs genotyped within each gene. The necessary number of principal components needed for each gene was determined using a 90% explained variance criteria. Once the necessary principal components were determined, univariate and multivariable logistic regression models were considered to assess the significance of each gene.

Haplotype-disease association was evaluated for each gene using Haplo.score (27), which accounted for ambiguous linkage phase. This method uses an expectation-maximization algorithm to infer haplotypes and accounts for ambiguity in haplotype assignment when comparing cases with controls and allows adjustment for nongenetic covariates, which are often critical when analyzing genetically complex phenotypes. The expectation-maximization method also provides global tests for association, as well as haplotype-specific tests, which give a meaningful advantage in attempting to understand the roles of different haplotypes. Haplotype odds ratios and 95% confidence intervals were calculated using Haplo.glm (28). Haplotype analyses were done using the Haplo.score and Haplo.glm functions included in HaploStats package version 1.2.1 in S-plus (Version 8.0.1).

Recursive partitioning (RPART) models (29), which implement binary trees to recursively partition the dataset into two subsets that are the most homogeneous with respect to the end point of interest (case/control status), were implemented to help identify potential interactions between SNPs (gene-gene) and environmental variables (gene-environment; ref. 30). These classification trees were built using all SNPs as well as the clinical variables used as adjusters in the multivariate analysis. After the first factor (and splitting point) has been chosen to maximize the homogeneity, each succeeding factor enters the tree conditional upon what has already entered and therefore represents an interaction (e.g. the second factor into the model would represent an interaction between the first factor and the second factor). Trees were grown using the standard defaults implemented by using standard functionality contained within the RPART library in S-plus (Version 8.0.1). The final trees were determined by pruning the tree to obtain a parsimonious model using cross-validation relative error rate and the 1-SE rule (29) as a guide to determine the best number of splits. The terminal nodes remaining after this pruning would define ''subgroups'' of interest whereas the splits resulting in those nodes would define potential interactions.

Results

Cases and controls (Table 1) were similar in age, but differed in BMI, sex (despite attempted frequency-matching), percent of ever-smokers, percent reporting a first-degree relative with pancreatic cancer, and diabetes (defined as diagnosed >2 years prior to cancer diagnosis for cases or participation for controls). When we validated medical record data to self-reported questionnaires, ê values for each variable for cases and controls, respectively, were: ever/never smoker (0.92, 0.75), pack-years (0.35, 0.64), family history of pancreatic cancer

Table 2. Results of principal components analysis of nucleotide excision repair genes and pancreatic cancer risk

Gene name	# SNPs	Chromosome location	Unadjusted P*	Adjusted P†	Principal components‡
ERCC1	4	19q13.32	0.2728	0.6705	3
ERCC2/XPD	10	19q13.32	0.3708	0.4229	5
ERCC3/XPB	7	2q14.3	0.4665	0.4154	4
ERCC4/XPF	4	16p13.12	0.7057	0.7866	3
ERCC5/XPG	14	13q33.1	0.5921	0.7214	5
ERCC6/CSB	12	10q11.23	0.4752	0.4524	4
ERCC8/CSA	11	5q12.1	0.2593	0.2623	4
XPA	7	9q22.33	0.1723	0.4038	5
XPC	9	3p25.1	0.3522	0.1697	5
RPA1	17	17p13.3	0.3186	0.3036	6
RPA2	7	1p35.3	0.2553	0.4069	3
RPA3	40	7p21.3	0.3151	0.3100	8
GTF2H1	7	11p15.1	0.2258	0.4933	3
GTF2H2	0	5q13.2	-	-	-
GTF2H3	3	12q24.31	0.7830	0.8007	2
GTF2H4	4	6p21.33	0.0515	0.0911	4
LIG1	14	19p13.2	0.2061	0.4489	3
RAD23A	1	19p13.13	0.4118	0.4864	1
RAD23B	20	9q31.2	0.9291	0.8288	6
CETN2	2	Xq28	0.9524	0.8358	2
CDK7	2	5q13.2	0.1080	0.1287	2
CCNH	2	5q14.3	0.3447	0.1898	2
MNAT1	15	14q23.1	0.1219	0.1316	5
XAB2 (HCNP)	6	19p13.2	0.7352	0.6284	3
DDB1	3	11q12.2	0.8722	0.9655	3
DDB2	8	11p11.2	0.4991	0.4876	5
MMS19L	7	10q24.1	0.0058	0.0230	3
Total					
27	236				

* Likelihood ratio test.
† Likelihood ratio test adjusted for age, sex, ever/never smoking, BMI, diabetes, first-degree family history of pancreatic cancer.
‡ Number of principal components needed to meet 90% explained variance criteria.

(1.0, 1.0), and race (1.0, 1.0). These results showed strong agreement between the two data sources.

The Principal Components Analysis approach was utilized to serve as an omnibus test for association between each candidate gene and disease. Adjusted and unadjusted principal components analyses were done for each gene in the nucleotide excision repair pathway to determine an overall gene level contribution to the risk of pancreatic cancer. *MMS19L* was the only gene that seemed to be significantly associated as shown in both unadjusted analyses (P = 0.0058) and after adjusting (0.0230) for age, sex, smoking status, BMI, diabetes, and family history of pancreatic cancer in a first-degree relative. Unadjusted and adjusted results for each of the genes are shown in Table 2. Based on our population, we determined that three independent principal components were sufficient to explain over 90% of the variability measured by the seven correlated SNPs of *MMS19L*. Unfortunately, this approach does not identify specific disease-causing variants and therefore additional analyses and/or follow-up studies would be necessary. Individual SNP level contributions to the eigenvectors and eigenvalue information for the first three principal components can be found in Supplemental Tables S1 and S2, respectively.

Logistic regression analyses at the single-SNP level for each gene were also done using additive, dominant, and recessive models model adjusted for age, sex, ever/never smoking, first-degree family history of pancreatic cancer, BMI, and diabetes. Overall odds ratios and subgroup analyses for *MMS19L* SNPs (total of 7) are shown in Table 3. Protective associations were observed in additive, dominant, and recessive models for minor alleles at rs872106 and rs2211243, whereas an increased risk was observed for rs2236575. The direction of risk effect for each SNP is largely consistent across demographic groups such as sex, location of residence, and smoking status, suggesting an effect independent of these factors (Table 4), although associations were more pronounced for females. Associations among smokers did not show a dose-dependent effect by pack-year categories, with risk changes more pronounced among ever than never smokers, although smokers with the least and the most pack-year exposure showed the highest effect and moderate pack-year smokers showed the least. No effect was detected in current smokers, but the numbers of current smokers in both cases and controls were small. The strongest associations for all three SNPs were seen among those former smokers who had quit at least 15 years prior to diagnosis/enrollment. The associations in the heaviest smokers were roughly comparable with those seen in nonsmokers.

Associations were identified among SNPs in several other genes, and are presented in the supplementary information (Supplemental Table S3).

Table 5 displays the results of the haplotype analysis for *MMS19L*. Of all possible combinations, seven haplotypes constituted 99% of haplotypes identified in controls, and were designated as A through F. We observed an overall effect on the risk of pancreatic cancer (global simulation P = 0.012). Two haplotypes (labeled in decreasing order of frequency in controls) were associated with statistically significant decreases in risk compared with the most common haplotype A (B, adjusted odds ratio, 0.85; 95% confidence interval, 0.73-0.99; and E,

Table 3. Association studies of SNPs in *MMS19L* for pancreatic cancer risk

Group ($N_{cases}/N_{controls}$)	*rs29001322*	*rs2236575*	*rs872106*	*rs3381*	*rs4917772*	*rs2211243*	*rs7915501*
	A>G	*A>T*	*G>C*	*G>A*	*A>G*	*A>G*	*A>C*
Genotype frequencies cases/controls	1,142/1,096	1,138/1,096	1,143/1,095	1,143/1,097	1,143/1,097	1,139/1,091	1,119/1,075
Major/major	675/673	340/384	628/550	1,058/994	611/588	342/269	719/683
Major/minor	414/379	559/521	433/436	85/99	458/441	543/532	361/346
Minor/minor	53/44	239/191	82/109	0/4	74/68	254/290	39/46
Hardy-Weinberg equilibrium	0.30	0.53	0.10	0.32	0.22	0.42	0.79
OR (95% CI) for pancreatic cancer risk							
Overall (codominant)* major/minor vs. major/major	1.093 (0.907-1.316)	1.184 (0.97-1.446)	0.914 (0.759-1.101)	0.749 (0.542-1.035)	0.98 (0.816-1.177)	0.792 (0.641-0.978)	0.921 (0.76-1.117)
Minor/minor vs. major/major	1.072 (0.691-1.661)	1.34 (1.039-1.727)	0.767 (0.555-1.059)	-	0.963 (0.666-1.393)	0.711 (0.557-0.908)	0.751 (0.473-1.194)
Overall (additive)*	1.099 (0.949-1.273)	1.19 (1.055-1.343)	0.841 (0.738-0.959)	0.719 (0.534-0.969)	1.008 (0.878-1.157)	0.829 (0.737-0.933)	0.941 (0.809-1.095)
Overall (dominant)*	1.121 (0.943-1.332)	1.287 (1.074-1.543)	0.836 (0.706-0.99)	0.74 (0.545-1.005)	1.009 (0.852-1.196)	0.769 (0.636-0.931)	0.952 (0.797-1.137)
Overall (recessive)*	1.107 (0.729-1.682)	1.223 (0.985-1.518)	0.702 (0.517-0.953)	-	1.013 (0.716-1.433)	0.785 (0.645-0.957)	0.808 (0.52-1.256)

Abbreviations: OR, odds ratio; 95% CI, 95% confidence interval.
*Adjusted for age, sex, ever/never smoking, diabetes, 1st degree relative with pancreatic cancer, and BMI.

Table 4. Pancreatic cancer risk analyses for associated *MMS19L* SNPs in selected subgroups

Group ($n_{cases}/n_{controls}$)*	*rs2236575*	*rs872106*	*rs2211243*
	A>T	*G>C*	A>G
Male (n = 668/557)	1.1 (0.94-1.287)	0.961 (0.805-1.147)	0.875 (0.748-1.024)
Female (n = 475/540)	1.301 (1.086-1.557)	0.712 (0.588-0.861)	0.787 (0.662-0.935)
Never smokers (n = 455/592)	1.156 (0.968-1.38)	0.852 (0.704-1.032)	0.88 (0.741-1.045)
Ever smokers (n = 682/505)	1.209 (1.029-1.421)	0.839 (0.703-1.002)	0.795 (0.677-0.933)
Pack-years <20 (n = 186/284)	1.202 (0.924-1.65)	0.725 (0.543-0.968)	0.744 (0.573-0.965)
Pack-years 20-40 (n = 149/119)	1.187 (0.853-1.51)	1.055 (0.727-1.531)	0.861 (0.622-1.191)
Pack-years >40 (n = 135/77)	1.239 (0.834-1.84)	0.805 (0.516-1.256)	0.723 (0.488, 1.07)
Current smokers† (n = 148/41)	0.679 (0.403-1.143)	1.175 (0.646-2.138)	1.026 (0.613-1.72)
Former smokers quitting 1-15 y prior to diagnosis/enrollment (n-447/134)	0.965 (0.718-1.298)	0.882 (0.642-1.212)	0.947 (0.708-1.265)
Former smokers quitting >15 y prior to diagnosis/enrollment (n = 228/365)	1.425 (1.131-1.797)	0.7 (0.535-0.916)	0.622 (0.492-0.787)
Age <60 y (n = 329/297)	1.064 (0.847-1.338)	0.968 (0.764-1.226)	0.891 (0.714, 1.111)
Age =60 y (n = 814/800)	1.245 (1.085-1.429)	0.785 (0.673-0.915)	0.805 (0.702-0.923)
Local (MN,WI,IA) (n = 579/748)	1.113 (0.953-1.299)	0.875 (0.738-1.036)	0.854 (0.733-0.996)
Nonlocal (n = 564/349)	1.263 (1.045-1.527)	0.799 (0.65-0.982)	0.825 (0.685-0.993)
BMI <30 kg/m^2(n = 769/858)	1.211 (1.055-1.389)	0.822 (0.707-0.955)	0.812 (0.71-0.929)
BMI >30 kg/m^2(n = 374/239)	1.162 (0.916-1.472)	0.858 (0.662-1.112)	0.862 (0.682-1.901)

* All analyses are unadjusted OR (95% CI) using an additive model. The ORs correspond to a unit increase in the number of variant alleles under the additive model.
† Defined as either current smoking at diagnosis/enrollment or quit within the preceding 1 y.

adjusted odds ratio, 0.65; 95% confidence interval, 0.47-0.90). Although these two haplotypes differed at rs872106, they carry the same alleles at rs2211243 and rs2236575.

Recursive partitioning analysis was also done as an exploratory method to assess SNP-SNP associations within the pathway and SNP-environment interactions, both overall and within the following subgroups: <age 60, ever/never smokers, BMI >30, self-reported diabetes (Y/N), and self-reported diabetes (Y/N) at least 2 years prior to either cancer diagnosis or enrollment as a control. After pruning the final trees using cross-validation error rate and the 1-SE rule, diabetes provided the only split in the overall analysis (342 of 1143 cases versus 89 of 1,097 controls). Ensuing splits that did not remain after pruning were ever/never smoking among nondiabetics, and then age < or ≥ 63.5 among smoking nondiabetics. No significant SNP-SNP or SNP-environment interactions were observed based on this analysis. In the subgroups, smoking (ever/never) provided a split among subjects <age 60 at cancer diagnosis or enrollment as a control (215 of 329 cases versus 120 of 297 controls) after pruning; among nondiabetics only, smoking (455 of 1008 cases versus 475 of 801 controls) and age < or ≥ 63.5 years among ever smokers (166 of 455 cases were < 63.5 versus 238 of 475 control ever smokers).

We previously reported an association of *ERCC4/XPF* SNP R415Q (rs1800067) showing an inverse association with pancreatic cancer, although this was attributed to chance given the low frequency of minor alleles (16). This prior study used a different control group and was of smaller sample size. The reported effect was not seen in this current study (adjusted odds ratio, 0.92; 95% confidence interval, 0.72-1.17).

Discussion

Nucleotide excision repair is a complex pathway integral to the repair of exogenous damage to DNA from a variety of sources. Small variations in this pathway may have an impact on DNA repair capacity, and, over time, could heighten the risk of malignancy. The effect of interactions of these variations among the

Table 5. Haplotype analysis of *MMS19L* and risk for pancreatic cancer

Haplotype	MMS19L SNP							*P*	Sim *P*	Haplotype frequencies			Unadjusted	Adjusted*
	1†	2	3	4	5	6	7							
	A>G	A>T	G>C	G>A	A>G	A>G	A>G			Overall	Controls	Cases	OR (95% CI)	OR (95% CI)
A(referent)	A	T	G	G	A	A	A	0.005	0.005	0.429	0.408	0.449	1.00	1.00
B	A	A	C	G	A	G	A	0.005	0.004	0.279	0.298	0.260	0.80 (0.69-0.92)	0.85 (0.73-0.99)
C	G	A	G	G	G	G	C	0.506	0.532	0.135	0.131	0.138	0.96 (0.80-1.15)	0.93 (0.77-1.13)
D	G	A	G	G	G	A	A	0.400	0.395	0.084	0.080	0.087	1.00 (0.79-1.25)	1.03 (0.82-1.31)
E	A	A	G	A	G	G	C	0.046	0.059	0.042	0.048	0.036	0.69 (0.51-0.93)	0.65 (0.47-0.90)
F	A	A	G	G	A	G	C	0.224	0.221	0.017	0.020	0.015	0.70 (0.44-1.10)	0.66 (0.41-1.08)
Rare	-	-	-	-	-	-	-			0.011	0.011	0.011	0.93 (0.49-1.77)	0.94 (0.47-1.88)

NOTE: Global Test [stat = 16.098, df = 6, P = 0.0132, simulated (S = 1000) P = 0.012].
*Adjusted for age, sex, ever/never smoking, diabetes, first-degree relative with pancreatic cancer, and BMI.
† SNP 1 is rs29001322, 2 is rs2236575, 3 is rs872106, 4 is rs3381, 5 is rs4917772, 6 is rs2211243, 7 is rs7915501.

many genes involved in nucleotide excision repair is largely unknown.

This study represents an analysis of common polymorphisms in the complete nucleotide excision repair pathway and associated genes with risk of pancreatic cancer. Due to the explosion of high-throughput technology in genetic analysis, large scale analyses are now possible for genetic epidemiology studies. Increasingly common among these are genome-wide association studies, which are agnostic, without the need for choosing candidate genes or pathways. These can be costly, and often are only done on a small subset of the sample in a staged approach, so only one question can be addressed (usually overall adjusted risk using an additive model) at the second stage. An alternative is the candidate pathway approach, which is based on prior suspicion of association, and this follows a classic hypothesis-testing approach. In these studies, tag-SNPs are chosen in every known gene in the pathway in an attempt to include most common sequence variations in the identified genes, through the assumption of linkage disequilibrium. Variations may have a direct effect on gene function, but more likely are linked to potential causal variants. This approach is limited by our knowledge of the genes involved in pathways and their interactions, and will miss less common variations (such as deleterious mutations).

In order to screen for overall gene effects, we did gene-level associations using a principal components analysis with each SNP of a gene included in the analysis, adjusted for important covariates.

Our study has implicated *MMS19L* (on 10q24.1), a human homolog of *MMS19*, a gene first noted in *Saccharomyces cerevisiae*, to be involved in nucleotide excision repair and RNA transcription, with separate domains required for each process (31, 32). *MMS19L* has not been well characterized in humans, but is believed to play a similar role in human nucleotide excision repair, with several regions highly conserved; and alternate splicing preserved across species (33). The protein binds to the GTF2H complex via ERCC2 and ERCC3, although its exact function is unclear (34). Analysis of *MMS19L* variants with cancer risk has only been reported in one study of lung cancer, with no alteration of outcome for one nonsynonymous SNP (35).

In addition to the gene level analyses by principal components analysis, we also did individual SNP, subgroup, haplotype, and interaction analyses within the pathway. As noted above, three SNPs in *MMS19L* seemed to associate with altered risk of pancreatic cancer. The association seemed to be strongest among women, ever smokers, former smokers quitting >15 years prior, and those with lower BMI. However, confidence intervals for these subgroups overlap with others, so these distinctions are considered exploratory.

In order to avoid missing possible associations of SNPs in genes not detected by the principal component approach, individual analyses were done for all SNPs in the pathway. Because many of these will be associated simply by chance, replication will be required to confirm our findings.

In the pathway interaction analyses undertaken using recursive partitioning (RPART), no significant associations were found, although we cannot rule out interactions. Pathway analysis is limited by many factors, including unknown biological function of variants, lack of ability to separate chance findings from true differences, and lack of consensus among the research community on how to best assess interactions. A potential limitation of RPART is that due to binary splitting, subgroups are created with rapidly diminishing numbers of cases and controls. Thus, it may not detect more complex associations due to a lack of power in the smaller groups. However, an advantage of RPART is that it is agnostic, and does not simply constitute a compilation of positive findings, many of which could be false positives.

Perhaps more important than our findings with *MMS19L*, there does not seem to be a large effect of nucleotide excision repair variation on pancreatic cancer risk overall. The low number of positive associations, when many are likely due to chance, suggests that perhaps this pathway is less important in pancreatic adenocarcinoma carcinogenesis. Replication of our findings, both positive and negative, in other study populations will be key to defining the role for polymorphisms this pathway in pancreatic cancer risk.

Limitations of this study include genotyping failure of 5% of our samples, which could affect power and results, but is unlikely to introduce a systematic bias. As this is a clinic-based case-control study, the choice of controls is always problematic, because no control group perfectly matches the patient population. Indeed, patients seen at a referral center are likely younger, healthier, and earlier stage than in the general population, and they must survive long enough to be seen. We attempted to minimize this with recruitment at the time of initial clinic appointment. In addition, using healthy patients seen in the General Internal Medicine Clinic as controls draws from a similar referral population at our institution, and the odds ratios seen for subjects from local and nonlocal locations of primary residence are consistent, at least for the *MMS19L* SNPs (Table 4). We also did not correct for multiple comparisons in our analyses, as we view these findings as exploratory and not conclusive. Methods such as the Bonferroni method can be overly conservative in genetic analyses due to linkage disequilibrium (36). The field has not yet reached a consensus on the correct adjustments needed, if any, aside from future replication, which we believe would represent the most important method of confirming our findings as not occurring by chance.

In conclusion, in a tag-SNP analysis of the nucleotide excision repair pathway and its associated genes, common variation in *MMS19L* is associated with altered risk of pancreatic cancer. Further studies to confirm the associations and identify the functional genetic variants in *MMS19L* responsible for the association are needed before these findings would be able to be included in risk modeling for pancreatic cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We appreciate the efforts of Martha Matsumoto (data analysis), Traci Hammer, Cindy Chan, and Jodie Cogswell (study coordinators, patient recruitment).

References

1. Thompson D, Easton DF, Breast Cancer Linkage Consortium. Cancer incidence in BRCA1 mutation carriers.[see comment]. J Natl Cancer Inst 2002;94:1358–65.
2. The Breast Cancer Linkage Consortium. Cancer risks in BRCA2 mutation carriers. J Natl Cancer Inst 1999;91:1310–6.
3. Petersen GM, Hruban RH. Familial pancreatic cancer: where are we in 2003?[comment]. J Natl Cancer Inst 2003;95:180–1.
4. Friedberg EC. How nucleotide excision repair protects against cancer. Nat Rev Cancer 2001;1:22–33.
5. Bootsma D, Hoeijmakers JH. The genetic basis of xeroderma pigmentosum. Ann Genet 1991;34:143–50.
6. Kraemer KH, Lee MM, Andrews AD, Lambert WC. The role of sunlight and DNA repair in melanoma and nonmelanoma skin cancer. The xeroderma pigmentosum paradigm. Arch Dermatol 1994;130:1018–21.
7. Kraemer KH, Lee MM, Scotto J. DNA repair protects against cutaneous and internal neoplasia: evidence from xeroderma pigmentosum. Carcinogenesis 1984;5:511–4.
8. de Boer J, Hoeijmakers JHJ. Nucleotide excision repair and human syndromes. Carcinogenesis 2000;21:453–60.
9. Millikan RC, Hummer A, Begg C, et al. Polymorphisms in nucleotide excision repair genes and risk of multiple primary melanoma: the Genes Environment and Melanoma Study. Carcinogenesis 2006;27:610–8.
10. Benhamou S, Sarasin A. Variability in nucleotide excision repair and cancer risk: a review. Mutat Res 2000;462:149–58.
11. Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. Cancer Epidemiol Biomarkers Prev 2002;11:1513–30.
12. Zhou W, Liu G, Miller DP, et al. Gene-environment interaction for the ERCC2 polymorphisms and cumulative cigarette smoking exposure in lung cancer. Cancer Res 2002;62:1377–81.
13. Hou SM, Falt S, Angelini S, et al. The XPD variant alleles are associated with increased aromatic DNA adduct level and lung cancer risk. Carcinogenesis 2002;23:599–603.
14. Mohankumar MN, Janani S, Prabhu BK, Kumar PR, Jeevanram RK. DNA damage and integrity of UV-induced DNA repair in lymphocytes of smokers analysed by the comet assay. Mutat Res 2002;520:179–87.
15. Zhao H, Wang LE, Li D, Chamberlain RM, Sturgis EM, Wei Q. Genotypes and haplotypes of ERCC1 and ERCC2/XPD genes predict levels of benzo[a]pyrene diol epoxide-induced DNA adducts in cultured primary lymphocytes from healthy individuals: a genotype-phenotype correlation analysis. Carcinogenesis 2008;29:1560–6.
16. McWilliams RR, Bamlet WR, Cunningham JM, et al. Polymorphisms in DNA repair genes, smoking, and pancreatic adenocarcinoma risk. Cancer Res 2008;68:4928–35.
17. Jiao L, Hassan MM, Bondy ML, Abbruzzese JL, Evans DB, Li D. The XPD Asp312Asn and Lys751Gln polymorphisms, corresponding haplotype, and pancreatic cancer risk. Cancer Lett 2007;245:61–8.
18. Wang LD, Lu XH, Miao XP. Polymorphisms of the DNA repair genes XRCC1 and XPC: relationship to pancreatic cancer risk. Wei Sheng Yan Jiu;35:534–6.
19. Duell EJ, Bracci PM, Moore JH, Burk RD, Kelsey KT, Holly EA. Detecting pathway-based gene-gene and gene-environment interactions in pancreatic cancer. Cancer Epidemiol Biomarkers Prev 2008;17:1470–9.
20. The International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–6.
21. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutat Res 2005;577:275–83.
22. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004;74:106–20.
23. Shen R, Fan JB, Campbell D, et al. High-throughput SNP genotyping on universal bead arrays. Mutat Res 2005;573:70–82.
24. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
25. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
26. Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 2007;31:383–95.
27. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–34.
28. Lake SL, Lyon H, Tantisira K, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 2003;55:56–65.
29. Therneau T, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. Technical Report Series, Section of Biostatistics, Mayo Clinic 1997. p. 61.
30. Rao DC. CAT scans, PET scans, and genomic scans. Genet Epidemiol 1998;15:1–18.
31. Lauder S, Bankmann M, Guzder SN, Sung P, Prakash L, Prakash S. Dual requirement for the yeast MMS19 gene in DNA repair and RNA polymerase II transcription. Mol Cell Biol 1996;16:6783–93.
32. Hatfield MD, Reis AM, Obeso D, et al. Identification of MMS19 domains with distinct functions in NER and transcription. DNA Repair (Amst) 2006;5:914–24.
33. Queimado L, Rao M, Schultz RA, et al. Cloning the human and mouse MMS19 genes and functional complementation of a yeast mms19 deletion mutant. Nucleic Acids Res 2001;29:1884–91.
34. Seroz T, Winkler GS, Auriol J, et al. Cloning of a human homolog of the yeast nucleotide excision repair gene MMS19 and interaction with transcription repair factor TFIIH via the XPB and XPD helicases. Nucleic Acids Res 2000;28:4506–13.
35. Michiels S, Danoy P, Dessen P, et al. Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers. Carcinogenesis 2007;28:1731–9.
36. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236–8.

Nucleotide Excision Repair Pathway Polymorphisms and Pancreatic Cancer Risk: Evidence for role of *MMS19L*

Robert R. McWilliams, William R. Bamlet, Mariza de Andrade, et al.

Cancer Epidemiol Biomarkers Prev 2009;18:1295-1302.