Supplementary Material

Human Oral Microbiome and Prospective Risk for Pancreatic Cancer: a Population-based Nested Case-Control Study

Xiaozhou Fan,[1] Alexander V. Alekseyenko,[2] Jing Wu,[1] Brandilyn A. Peters,[1] Eric J. Jacobs,[3] Susan M. Gapstur,[3] Mark P. Purdue,[4] Christian C. Abnet,[4] Rachael Stolzenberg-Solomon,[4] George Miller,[5,6,7] Jacques Ravel,[8] Richard B. Hayes,[1,7] Jiyoung Ahn,[1,7]

1. Department of Population Health, New York University School of Medicine, New York, NY, USA
2. Biomedical Informatics Center, Departments of Public Health Sciences and Oral Health Sciences, Program for Human Microbiome Research, Medical University of South Carolina, Charleston, SC, USA
3. Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA
4. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
5. Department of Surgery, New York University School of Medicine, New York, NY, USA
6. Department of Cell Biology, New York University School of Medicine, New York, NY, USA
7. NYU Perlmutter Cancer Center, New York, NY, USA
8. Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

ABBREVIATIONS

ACS-CPS II: American Cancer Society Cancer Prevention Study.
NCI-PLCO: National Cancer Institute Prostate, Lung, Colorectal, and Ovarian Cancer.
OTU: Operational Taxonomic Unit. Similar clusters of 16S small subunit rRNA genesare used as theory-agnostic approximations of microbial species.
HOMD: Human Oral Microbiome Database.
QC: Quality control. Specimens are used during method validation to detect, reduce, and correct deficiencies in the laboratory's internal analytical process.
PCoA: Principal coordinate analysis. Summaries and attempts to represent inter-object (dis)similarity in a low-dimensional space.
PERMANOVA: Permutational Mutivariate Analysis of Variance Using Distance Matrices. Analysis of variance using distance matrices — for partitioning distance matrices among sources of variation and fitting linear models to distance matrices; uses a permutation test with pseudo-F ratios.
LASSO: Least Absolute Shrinkage and Selection Operator. A regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.
BMI: Body mass index. An index calculated by dividing one's weight in kilograms by the square of one's height in meters.
OR: Odds Ratio. A measure of association between an exposure and an outcome.
CI: Confidence interval. A range expressing the degree of uncertainty associated with a sample statistic.

SUPPLEMENTAL METHODS

Study Population

The ACS CPS-II Nutrition Cohort is a subgroup of the larger CPS II Cohort, with a cohort size of more than 84,000 men and 97,000 women [1]. To assemble the CPS II Nutrition cohort, in 1992 and 1993 a baseline questionnaire was mailed to 516,671 CPS II men and women, age 50-74 years, who resided in 21 states with population-based state cancer registries. Incident cancer was captured primarily through self-report on periodic follow-up questionnaires and verified through medical records acquisition or through linkage with state cancer registries. Cancer deaths were ascertained via linkage with the National Death Index, usually followed by confirmation through linkage with state cancer with cancer registries. Pilot work indicated that self-report of cancer had high sensitivity, with range of 0.92 to 0.97, and high specificity (>0.99) [1].

The PLCO Cohort is a randomized controlled trail of screening tests for prostate, lung, colorectal and ovarian cancers [2]. Approximated 155,000 participants, aged 55-74 years, from 10 study centers across the U.S. were enrolled between 1993 and 2001. Exclusion criteria include a history of prostate, lung, colorectal or ovarian cancer; or a recent history of screening procedures for one of the PLCO cancers; or currently under treatment for any cancer except for basal or squamous cell skin cancer. Participants were individually randomized to the control arm (usual medical care) or intervention arm (receiving screening tests) in equal proportions. Incident cancer was captured through yearly questionnaires or through state registries, death certificates, physician reports, and reports from next to kin. Pancreatic cancer cases were confirmed by abstraction from medical records.

The current study employs a nested case-control design. In this design, cases are identified within a defined cohort, and for each case, one or more controls are selected who have not developed the disease by the time of disease occurrence in the case [3]. This design is ideal for molecular and genetic studies, where assaying the entire cohort is cost and time prohibitive; with the nested case-control design, temporality of exposure-disease is preserved, and valid effect estimates can be obtained, all with great cost efficiency. The selection of a control randomly from all cohort members at risk (excluding the index case) at the time of index case occurrence is known as incidence density sampling. Here, we have employed incidence density sampling without replacement, matching controls to cases on age (within 5 years), sex, race (white or other), and calendar year of oral wash sample collection. While traditional incidence density sampling selects controls with replacement, selection without replacement only produces slight bias [4], as 310 controls were selected from a cohort of >90,000, limiting the opportunity of resampling the same individual. Cases and controls were matched on several potential confounding factors in this analysis as matching may increase statistical efficiency when adjusting for these factors.

Statistical Analysis

Analysis of bacterial community composition

We used weighted (quantitative, accounts for relative abundance of each taxon) and unweighted (qualitative, accounts for presence/absence of taxon) UniFrac distances based on tables of OTU counts as the β-diversity (between-subject diversity) measures. UniFrac performs a pairwise comparison of all subjects in a dataset, defining the overall degree of phylogenetic similarity between any two subjects based on the degree of branch length their bacterial communities shared on a bacterial tree of life [5]. Principal coordinate analysis (PCoA) was used to reduce and represent patterns present in UniFrac distance matrices [6]. It attempts to explain most of the variance in the distance matrices by eigen-decomposition. The first and second rescaled eigenvector corresponding to the principal coordinates were used to display the distances between subjects of the dataset in a two dimensional space. Permutational Multivariate Analysis of Variance Using Distance Matrices (PERMANOVA) was conducted to test the association of microbial composition with pancreatic cancer status, allowing for the adjustment for covariates. These statistical significance tests are done using *F*-tests based on sequential sums of squares from permutations of the raw distance matrices [7].

Analysis of differential abundance of bacterial taxa

To limit false discovery rate, we used the L1 Penalized Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression classifier implemented in R “glmnet” package [8]. It generates a parsimonious classification model, which selects only a few features out of a potentially large set.

Our pipeline proceeds as follows:

1) “cv.glmnet” function was carried out to test for an optimal value of lambda (number of taxa to be used in the final model), in ten-fold cross-validated LASSO run on all data.
2) Since the folds were selected at random, we reduced this randomness by running this testing function 100 times and averaging the error curves to obtain the final optimal lambda.
3) We fit a generalized logistic model via penalized maximum likelihood using the “glmnet” function to generate a parsimonious taxa selection model.
4) We fit unconditional logistic regression using each selected taxon as a predictor of pancreatic cancer risk.

SUPPLEMENTAL RESULTS

Homogenous results from matched analysis

Unconditional logistic regression was used to compare carriage of the periodontal pathogens and relative abundance of LASSO-selected taxa in cases and controls. Instead of using the conditional logistic regression for matched cases and controls, the standard analysis can yield slightly better statistical precision [9]. For example, if two or more cases and their matched controls all have identical values for age, then combining them into a single stratum produces an estimator with lower variance and no less validity compared to the pair-matched analysis. However, we checked whether our results were consistent using conditional logistic regression. Results from conditional logistic regressions showed that carriage of *Porphyromonas gingivalis* was associated with a higher risk of pancreatic cancer (adjusted Odds Ratio [OR] for presence *vs.* absence=1.59 and 95% confidence interval [CI]= 1.13-2.26, *p*=0.0085), and a significant dose-response relationship (*P* trend=0.0055) for low carriers (below median relative abundance, OR 1.38 [0.88-2.19], *p*=0.16) and high carriers (above median relative abundance, OR 1.80 [1.16-2.79], *p*=0.0082) was also found. This dose-response relationship was confirmed by the paired Wilcoxon signed-rank test (*p*<0.0001). Conditional logistic regression also showed significant cancer risks related 4to carriage of *Aggregatibacter actinomycetemcomitans* (OR 2.02 [1.06-3.88], p=0.034), and lower abundance of *Fusobacteria* (OR 0.93 [0.88-0.98], *p*=0.011) and its genus *Leptotrichia* (OR 0.87 [0.79-0.95], *p*=0.0037).

Supplementary Table 1: Number of filtered sequence reads per sample*

	Total (N=732)	Cases (N=361)	Controls (N=371)
Cohort	Mean ± SD	Mean ± SD	Mean ± SD
CPS II	10,076 ± 2,258	10,126 ± 2,298	9,973 ± 2,188
PLCO	13,379 ± 2,263	13,425 ± 2,760	13,274 ± 1,722
Total	11,782 ± 2,799	11,871 ± 3,036	11,762 ± 2,550

* Poor-quality sequences were excluded using the default parameters of the QIIME script *split_libraries.py* (minimum average quality score=25, minimum/maximum sequence length=200/1000 base pairs, no ambiguous base calls, and no mismatches allowed in the primer sequence).

Supplementary Table 2a: Carriage of periodontal pathogens in pancreatic cancer cases and controls in the CPS II cohort

Periodontal pathogens	Cases (n=170) N (%)	Controls (n=170) N (%)	OR*	95% CI*	*p*-value*	*p* trend†
Porphyromonas gingivalis						
Non-carriers	116 (68.2)	134 (78.8)	1.00	-	-	
Any carriers	54 (31.8)	36 (21.2)	1.76	1.06-2.91	0.028	
Non-carriers	116 (68.2)	134 (78.8)	1.00	-	-	
Low carriers‡	26 (15.3)	18 (10.6)	1.63	0.84-3.19	0.15	
High carriers‡	28 (16.5)	18 (10.6)	1.89	0.97-3.65	0.060	
						0.031
Aggregatibacter actinomycetemcomitans						
Non-carriers	153 (90.0)	165 (97.1)	1.00	-	-	
Any carriers	17 (10.0)	5 (2.9)	3.90	1.37-11.07	0.011	
Tannerella forsythia						
Non-carriers	89 (52.4)	93 (54.7)	1.00	-	-	
Any carriers	81 (47.6)	77 (45.3)	1.08	0.70-1.66	0.74	
Prevotella intermedia						
Non-carriers	140 (82.4)	145 (85.3)	1.00	-	-	
Any carriers	30 (17.6)	25 (14.7)	1.27	0.70-2.30	0.44	

* Odds ratios, 95% confidence intervals, and *p*-values were calculated from logistic regression models after controlling for the covariates (age, race, sex, BMI, smoking status, alcohol consumption status, and history of diabetes).

† *p* trend was calculated by assigning values 0-2 to the non-carriers, low carriers, and high carriers of *Porphyromonas gingivalis*, respectively, and treating this as a continuous variable in the logistic regression model, after controlling for the covariates (age, race, sex, BMI, smoking status, alcohol consumption status, and history of diabetes).

‡ Cut-off point was based on the median relative abundance of *Porphyromonas gingivalis* in the control group of CPS II cohort (0.12%).

Supplementary Table 2b: Carriage of periodontal pathogens in pancreatic cancer cases and controls in the PLCO cohort

Periodontal pathogens	Cases (n=191) N (%)	Controls (n=201) N (%)	OR*	95% CI*	*p*-value*	*p* trend†
Porphyromonas gingivalis						
Non-carriers	117 (61.3)	141 (70.1)	1.00	-	-	
Any carriers	74 (38.7)	60 (29.9)	1.48	0.94-2.31	0.088	
Non-carriers	119 (61.3)	141 (70.2)	1.00	-	-	
Low carriers‡	33 (17.3)	30 (14.9)	1.31	0.73-2.35	0.36	
High carriers‡	41 (21.4)	30 (14.9)	1.64	0.94-2.88	0.082	
						0.070
Aggregatibacter actinomycetemcomitans						
Non-carriers	177 (92.7)	191 (95.0)	1.00	-	-	
Any carriers	14 (7.3)	10 (5.0)	1.52	0.64-3.57	0.34	
Tannerella forsythia						
Non-carriers	102 (53.4)	117 (58.2)	1.00	-	-	
Any carriers	89 (46.6)	84 (41.8)	1.28	0.84-1.93	0.25	
Prevotella intermedia						
Non-carriers	139 (72.7)	156 (77.6)	1.00	-	-	
Any carriers	52 (27.3)	45 (22.4)	1.47	0.89-2.41	0.13	

* Odds ratios, 95% confidence intervals, and *p*-values were calculated from logistic regression models after controlling for the covariates (age, race, sex, BMI, smoking status, alcohol consumption status, and history of diabetes).

† *p* trend was calculated by assigning values 0-2 to the non-carriers, low carriers, and high carriers of *Porphyromonas gingivalis*, respectively, and treating this as a continuous variable in the logistic regression model, after controlling for the covariates (age, race, sex, BMI, smoking status, alcohol consumption status, and history of diabetes).

‡ Cut-off point was based on the median relative abundance of *Porphyromonas gingivalis* in the control group of PLCO cohort (0.056%).

Supplementary Table 3a: Median relative abundance of selected oral microbial taxa* in pancreatic cancer cases and controls in the CPS II cohort

		Median relative abundance		OR†	95% CI†	*p*-value†	*p*-value‡
		Cases (n=361)	Controls (n=371)				
Phylum							
	Fusobacteria	3.20	3.47	0.93	0.87-0.99	0.019	0.018
	Bacteroidetes	7.73	7.70	1.00	0.97-1.04	0.76	0.95
Class							
	Fusobacteria	3.20	3.47	0.93	0.87-0.99	0.019	0.018
	SR1[C-1]	0.00	0.00	1.26	0.63-2.50	0.51	0.85
Order							
	Fusobacteriales	3.20	3.47	0.93	0.87-0.99	0.019	0.018
Family							
	Leptotrichiaceae	1.09	1.16	0.88	0.78-0.99	0.038	0.052
Genus							
	Leptotrichia	1.09	1.16	0.88	0.78-0.99	0.038	0.052
	Alloprevotella	0.22	0.21	1.52	1.07-2.16	0.021	0.036

* Taxa were selected from L1-penalized LASSO logistic models with the optimal value of lambda from 100 repeated 10-fold cross-validation on each taxonomic level in the combined CPS II and PLCO cohort in Table 3.

† Odds ratios, 95% confidence intervals, and *p*-values for 1% increase in relative abundance were calculated from logistic regression models, after controlling for the covariates (age, race, sex, BMI, smoking status, alcohol consumption status, and history of diabetes).

‡ Additionally controlled for the carriage of *Porphyromonas gingivalis* and *Aggregatibacter actinomycetemcomitans*.

Supplementary Table 3b: Median relative abundance of selected oral microbial taxa* in pancreatic cancer cases and controls in the PLCO cohort

		Median relative abundance		OR†	95% CI†	*p*-value†	*p*-value‡
		Cases (n=361)	Controls (n=371)				
Phylum							
	Fusobacteria	1.71	2.07	0.94	0.86-1.04	0.22	0.12
	Bacteroidetes	10.24	9.59	1.02	0.99-1.05	0.16	0.19
Class							
	Fusobacteria	1.71	2.07	0.94	0.86-1.04	0.22	0.12
	SR1[C-1]	0.021	0.022	1.16	0.96-1.40	0.13	0.17
Order							
	Fusobacteriales	1.71	2.07	0.94	0.86-1.04	0.22	0.12
Family							
	Leptotrichiaceae	0.67	0.90	0.85	0.73-0.99	0.032	0.021
Genus							
	Leptotrichia	0.67	0.90	0.85	0.73-0.99	0.032	0.021
	Alloprevotella	0.32	0.31	1.10	0.89-1.35	0.40	0.58

* Taxa were selected from L1-penalized LASSO logistic models with the optimal value of lambda from 100 repeated 10-fold cross-validation on each taxonomic level in the combined CPS II and PLCO cohort in Table 3.

† Odds ratios, 95% confidence intervals, and *p*-values for 1% increase in relative abundance were calculated from logistic regression models, after controlling for the covariates (age, race, sex, BMI, smoking status, alcohol consumption status, and history of diabetes).

‡ Additionally controlled for the carriage of *Porphyromonas gingivalis* and *Aggregatibacter actinomycetemcomitans*.

REFERENCES

1 Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML*, et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* 2002;**94**:2490-501.

2 Hayes RB, Reding D, Kopp W, Subar AF, Bhat N, Rothman N*, et al.* Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled clinical trials* 2000;**21**:349S-55S.

3 Ernster VL. Nested case-control studies. *Preventive medicine* 1994;**23**:587-90.

4 Richardson DB. An incidence density sampling program for nested case-control analyses. *Occupational and environmental medicine* 2004;**61**. http://oem.bmj.com/content/61/12/e59.long (accessed: 24 June 2004).

5 Lozupone C, Knight R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and environmental microbiology* 2005;**71**:8228-35.

6 Gower JC. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika* 1966;**53**:325.

7 Dixon P. VEGAN, a package of R functions for community ecology. *J Veg Sci* 2003;**14**:927-30.

8 Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *2010* 2010;**33**:22.

9 Pearce N. Analysis of matched case-control studies. *Bmj* 2016;**352**:i969.