Màster Universitari
**Anàlisi de Dades Òmiques /
Omics Data Analysis**
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

Oral/Gut microbiome profiles in pancreatic cancer and their interactions with dietary patterns

by

Maria Ester Molina Montes

Supervisor: Núria Malats, Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Center - Centro Nacional de Investigaciones Oncológicas (CNIO)

Academic tutor: Malu Calle, Biosciences Department, Universitat de Vic

Biosciences Department

University of Vic – Central University of Catalonia

[10 September 2022]

Summary

Background and aims: Pancreatic cancer (PC) has a high case-fatality rate in Western countries, expected to rise in coming years if no immediate actions are taken. Several studies have pointed to an association between the human microbiome and PC risk. Diet is known to play a key role in the microbiota composition in terms of abundance and diversity, and could therefore interact in the microbiome-PC relationship. However, the association between diet and PC is complex. Within the PanGenEU and PanGen-MICROBIOME studies, we aimed to identify a gut and/or oral microbiome signature to detect PC, and to explore interactions with dietary factors.

Methods: The MICROBIOME study counts with metagenomic and 16srRNA data of the stool and salivary microbiota of over 50 PC cases and 50 matched controls. In addition, there is 16sRNA data on the salivary microbiota of more than 500 subjects (~250 cases and controls) from the PanGenEU study. After data processing, we compared alpha and beta-diversity measures among cases and controls regarding the oral and gut microbiome. Associations between taxa with case-control status were examined in univariate (Wilcoxon test) and multivariate analyses (via edgeR). Then, a microbiome-based classifier was explored (via LASSO regression) to discriminate between cases and controls. To explore interactions between dietary factors and bacterial taxa, we calculated several dietary scores (the diabetes risk reduction diet score DRRD and the relative Mediterranean Diet score rMED) and microbial risk scores based on the bacterial species conforming the signature. The latter comprised an abundance-based microbial risk score (MRS) and two alpha-diversity measures (richness score, rS, and Shannon score, SS). The association between dietary factors and these scores and with the PC risk was examined using logistic and linear regression models adjusted for potential confounders. Ridge Regression and LASSO as feature selection methods were used to identify foods, food groups and nutrients related to the microbial risk scores. Microbial species related with the dietary scores were also selected using this method. Tuning parameters were optimized on training and test sets. At the individual level, associations (e.g. Spearman correlations) between taxa of both the gut and oral microbiota with foods, food groups and nutrients were explored. Also, hierarchical clustering was applied to identify clusters of taxa and of foods, food groups and nutrients among PC cases and controls. P-values less than 0.05 were deemed significant; p-values were corrected for multiple testing (Benjamini-Hochberg, BH and applying FDR) in association analyses.

Results: A stool metagenomic signature of 27 microbial species that discriminated between PC and controls with an accuracy of up to 0.84 area under the curve (AUC). Its performance improved to up to 0.94 AUC when serum levels of CA-19.9, the current diagnostic marker of PC, were incorporated into the signature. The microbial risk signature MRS reflected better PC risk (OR $_{\text{per 1SD increase in MRS}}$=5.5, p= 3.01E-05) than those based on alpha-diversity measures (rS or SS). Regarding its relationship with dietary factors, the most meaningful variables were seafood, some alcoholic beverages, and some polyunsaturated fatty acids (PUFAs). None of the dietary scores, the DRRD and the rMED scores, were significantly associated with PC risk. However, some features in the gut microbiome appeared to be linked to these scores. Also, cluster analyses revealed the existence of gut microbial taxa and diet clusters, with groupings of these taxa with plant-based foods, seafood and PUFAs, mainly.

Conclusion: A distinctive gut microbial signature made up of 27 bacterial species allows to discriminate between PC cases and controls. Several dietary factors were related to this microbial signature in terms of a risk score. Future studies with larger sample size are warranted to confirm these findings.

Keywords: Pancreatic cancer, microbiome, dietary patterns, nutrients, biomarkers

Abbreviations:

PC: Pancreatic Cancer
T2D: Type 2 diabetes
DRRD: Diabetes risk reduction diet
rMED: relative Mediterranean Diet score
MRS: Microbial risk score
rS: richness score
SS: Shannon score
SFA: Saturated fatty acids
MFA: Monounsatureted fatty acids
PUFA: Polyunsatured fatty acids
E: Energy
OR: Odds ratios
CI: Confidence Intervals

1. Introduction

Pancreatic cancer (PC) has a high case-fatality rate in Western countries, expected to rise in coming years if no immediate actions are taken [1]. This situation is due to two major reasons: i) few risk factors of this disease have been established, which prevents the possibility of primary prevention; ii) the disease manifests itself at late stages owing to the lack of biomarkers for early diagnosis [2]. Specifically, well-established risk factors of PC include chronic pancreatitis (CP), diabetes type 2 (T2D) most notably if newly diagnosed, obesity, smoking and heavy alcohol intake, whereas there are is still scarce evidence regarding the role of dietary factors in PC aetiology [3].

Recently, some studies have also revealed that the gut microbiome is key in PC etiopathogenesis. Not only it shapes inflammation and immune function in PC disease, but it also seems to determine response to treatment [4]. Some studies have even suggested that certain microbial species of the gastrointestinal tract, detectable in human faeces by application of next generation sequencing technologies (metagenomics and 16S rRNA), could be involved in the development and progression of PC [5]. The majority of these studies relied on 16s rRNA gene sequencing data or qPCR analyses [6–8]. Bacterial species in tumoral tissue of the pancreas have been also evidenced by 16S rRNA sequencing [9], suggesting that pathogenic bacteria are able to migrate to this tissue in PC patients. It is also important to highlight that few studies have pointed to an association between the oral microbiome and PC risk, which might be driven by oral pathogens associated with periodontitis (e.g. *Porphyromonas gingivalis*) [10,11]. To the best of our knowledge, only two recent studies have explored gut microbiome profiles in the gut and oral microbiota using shotgun metagenomics sequencing data [12,13]; this technique provides much more sensitive information on bacterial species in a sample. In particular, a microbial signature with high accuracy (Area under the curve, AUC=0.94 when combined with the marker CA.19-9) to discriminate between PC and non-PC was identified among 50 PC cases and matched controls, and validated further within an external study sample [13]. Overall, these findings support that alterations in the gut microbiome composition could increase risk of PC, and could serve as potential early-diagnostic biomarkers for this disease.

On the other hand, diet is known to play a key role in the microbiota composition in terms of abundance and diversity [14–16], and could therefore interact in the microbiome-cancer relationship. This interaction can be very complex given that the human microbiome constitutes a community of bacterial species that symbiotically interacts together [17]. The complexity of this interaction is further compounded by the fact that dietary factors interact synergistically together, which makes dietary pattern analyses more appropriate [18]. In fact, instead of looking at individual nutrients or foods, dietary patterns derived from index scores (known as *a priori* patterns), factor or cluster analyses (known as *a posteriori* patterns), have emerged as complementary approaches in nutritional epidemiology to explore associations of overall diet on disease risk [18]. The effect of diet on the microbiome-cancer association, however, is an hitherto uninvestigated subject in PC [4]. This is possibly due to the lack of consistent associations between dietary factors with PC risk despite there are numerous studies on the likely impact of diet on this disease at an individual [19–22], and at an aggregated level [22–24]. In contrast, the interaction between diet and the human microbiome has been explored in other cancer types such as colorectal cancer. For instance, *Fusobacterium nucleatum* was depleted in tumor tissue and the intestinal microbiota in cancer patients who had a high adherence to a prudent dietary pattern (rich in whole grains and fiber) compared to those who followed a more westernized pattern (rich in red and processed meat, refined grains and desserts) [25]. A western-style diet was also stronger in tumors containing higher amounts of certain *Escherichia coli* strains [26]. Also, several bacteria were enriched in stool samples of colorectal cancer patients with higher consumption of sugars and sweets, eggs, oils and fats, amongst other foods [27].

Within the PanGenEU and MICROBIOME studies, we aimed to identify a gut/oral microbiome signature to detect PC, and to explore interactions with dietary factors. In particular, the aims of this study were: 1) To identify a stool and oral microbiota signature associated with PC using both 16s rRNA and metagenomics data; 2) To assess the relationship between dietary factors, individually and collectively, with the gut/oral microbiota signature and with PC; 3) To evaluate how dietary patterns cluster with gut/oral microbiotic profiles, also considering the aforementioned signatures, in PC cases and controls. The first aim has been addressed earlier within the MICROBIOME study and is part of the aforementioned study by Kartal, Schmidt and *Molina-Montes et al.* (co-first authors) [13]; the other two aims were addressed in the current study (Master thesis) and are based on data of both, the PanGenEU and MICROBIOME studies. Regarding the first aim, some non-published results derived from the PanGenEU study are also reported in this Master thesis.

2. Methods:

2.1. Study design: Case-control study.

2.2. Recruitment and data collection: Details on these issues within the MICROBIOME study have been described elsewhere [13]. In brief, there were 64 PC cases, 59 controls, and 29 CP patients enrolled in the MICROBIOME study with both saliva and stool samples (for 16sRNA and metagenome sequencing). Some non-eligible subjects were identified after reviewing the pathology reports. Therefore, 57 PC cases, 50 controls and 29 CP patients remained available for analyses. In addition, there were 515 subjects (317 PC cases and 198 controls) with saliva samples from the PanGenEU study. The latter PanGenEU subjects contributed only to the oral microbiome analyses via 16s rRNA sequencing. Study subjects were recruited from two Spanish hospitals for the MICROBIOME study (Hospital Ramón y Cajal in Madrid and Hospital Vall d'Hebron in Barcelona), whereas additional centers contributed with salivary samples to the PanGenEU study (Hospital del Mar, Hospital San Pau and Hospital de Elche in Spain, as well as Hospital TUM from Munich in Germany). All controls were matched to PC cases by age (± 10 years), sex and center, and all were admitted to hospitals for PC-non-related causes.

All participants provided information on demographic and medical history and lifestyle factors; these variables (smoking status, T2D, periodontitis, alcohol intake, height and weight as a proxy of body mass index -BMI, metformin use, dietary habits, etc.) were considered to control in the statistical analysis for their influence on the microbiome and PC disease. Within the MICROBIOME study, information on periodontal disease and on the use of antibiotics, and probiotic supplements was also collected. In the MICROBIOME study the information was collected using the same procedures and protocols used in the PanGenEU study to enable unified analyses of both studies. Only slight differences were adopted concerning the protocols used for the sampling of the saliva samples, which were modified to improve the sequencing efficiency in the MICROBIOME study, whereas those of the PanGenEU study were based on oral mouthwashes. In addition, clinical data was collected for PC cases, tumor samples were retrieved from same cases, and some clinical markers were measured (bilirubin and CA.19-9, the current tumor marker for PC).

The study counts with approval from independent Ethics Committees (e.g., CEI PI 26 2015- v7) and written informed consent were obtained from all study participants.

2.3. Sample processing and sequencing via 16S rRNA and shotgun metagenomics:

Oral and stool samples were collected from the participants of this study. Sample processing and sequencing, as well as the bioinformatics workflow, was performed by EMBL-Heidelberg. Data filtering and normalization was done jointly. Details on all processes are described elsewhere [13]. In short, at first, DNA was extracted from the RNALater-preserved samples using the Qiagen allprep powerfecal DNA/RNA kit. Secondly, 16S rRNA and shotgun metagenomics sequencing were performed:

- Targeted amplification of the 16S rRNA V4 region was carried out. Salivary samples with enough biomass were sequenced. Raw amplicon reads were denoised, filtered for read quality and chimeric reads, and matching paired reads were assembled using DADA2 [28]. The resulting Amplicon Sequence Variants (ASVs) were then clustered into open-reference Operational Taxonomic Units (OTUs) using MAPseq and other methods [29,30]. Thus, two taxa tables were generated for the data analyses: the open-OTU table with 2,081 OTUs, and the ASV table with 20,272 ASVs. To remove additional noise, we applied several other filtering steps to the 16s rRNA data including the removal of samples with less than 500 reads and taxa not present in at least 5 samples. The number of reads in the samples was normalized by rarefying to account for differences in sequencing depth across the runs. Across samples, OTU/ASV relative abundance was computed as the ratio of an OTU's/ASV´s absolute abundance to the total number of reads for that sample. Among the retained samples there were 18 duplicated saliva samples for quality controls. To check the performance of the sequencing we compared the microbiome composition between the saliva replicates against all other saliva samples using the Wilcoxon signed-rank test. Variation within the replicated saliva samples was smaller than variation between the duplicates and all other samples (i.e., p-value=2.109E-09 in ASV data). Thus, replicates were considered equivalent and reads of both replicates were pooled. The non-eligible cases were removed from the dataset too, leaving finally 573 samples with 3,393 ASV and 580 samples with 852 OTUs within the PanGenEU study (Supplementary Table 1). Within the MICROBIOME study, as described before

[13], 130 samples remained out of 142 salivary samples after quality control and filtering. Regarding stool samples sequenced via 16S rRNA, 118 samples remained out of 120.

- Metagenomic libraries for stool and salivary samples were prepared using the NEB Ultra II and SPRI HD kits, and then sequenced on an Illumina HiSeq 4000 platform (Illumina, San Diego, California, USA). For three salivary and one stool samples, technical replicates were merged after confirming their low within-sample variation. Stablished protocols were used for quality controls, data filtering and mapping. Taxonomic profiles were obtained using the mOTU profiler [31]. Stool shotgun metagenomes were obtained for all PC cases, controls and CP (57 PC cases, 50 controls and 29 CP patients), while salivary metagenomes for 43 PC cases, 45 controls and 12 CP patients.

For aim 1, all samples were considered, whereas for aims 2 and 3 only metagenomic and 16s rRNA data coming from the MICROBIOME study were considered. A summary table of the final study samples by aims, site and disease status is shown in Supplementary Table 2.

2.4. <u>Metadata processing and variables</u>:

As previously mentioned, PC cases and controls provided information about their lifestyle and environmental exposures and medical history in interviews conducted by trained staff, using the same structured questionnaire in all participating centers.

The data collected was recorded, curated and prepared for data analyses. Variables available for this study were: age, sex, center, pack-years of smoking in tertiles; BMI; obese vs non-obese; early-onset, late-onset T2D vs non-T2D; periodontitis vs non-periodontitis; family history of PC vs non-history, amongst others. As described elsewhere[13], random forest algorithm (n=100 trees) was used to impute the missing values of the metadata (missing rate: 3.1%) using *missForest* R package [32]. The resulting mean out-of-bag error was low (=0.12).

For the current dietary study (aims 2 and 3) we used information on diet, which was collected by means of food frequency questionnaire (FFQ). This tool was a semi-quantitative FFQ of 149-food items that was validated before [33], and adapted within the framework of the EPICURO study [34]. Within the MICROBIOME study, we added some further items on tea and coffee intake, on the use of vitamin and mineral supplements, and on the consumption of prebiotic and probiotic foods. All cases and controls were asked about their dietary intake two years before they entered the study. The questionnaire was structured by food groups, including some items for cooking methods of meat and fish. In the current study, we considered only the 82 food items accounting solely for food consumption regardless of cooking methods.

Firstly, the frequency per day was calculated using conversion factors for intakes on a weekly and monthly basis (for example, 3 per week equivalent to 3/7 per day). To convert frequencies of intake into grams per day we considered servings of intake according to the Spanish Food Pyramid Guidelines [35]. To estimate nutrient intakes (of macro and micronutrients, including vitamins, minerals and fatty acids), the Spanish Food Composition table BEDCA was used, which accounts for 46 nutrients of over 800 food items per 100 g of food [36]. For this dietary compilation process, a matrix of food items in g/day and the BEDCA database were multiplied by each other. In summary, the compilation process involved the following steps: frequency of intakes was transformed into frequencies/day, and then multiplied by servings/day (grams/day of each serving) to get grams/day of each food item. Then, intakes in grams were multiplied by the nutrient content per 100 g of each food. For food items not present in studies or for those combining various foods, average values were assigned for the compilation of these food items (for example: the mean values of nutrients contained in eggplant and zucchini were considered to adapt the food item).

Food groups were created to account for intakes in grams/day. In total, 38 food groups were generated (for instance, eggs, nuts, sauces, legumes, citrus fruits, fruits, leafy vegetables, vegetables, white fish, fatty fish, white meat, red meat, processed meat, ready-to eat dishes, sweetened beverages, artificially sweetened beverages, juices, etc.).

2.5. <u>Statistical data analysis</u>:

2.5.1. <u>Objective 1: Diversity measures and identification of the stool and oral microbiome signature</u>

Description of the data: Characteristics by PC cases and controls were compared by Chi-squared tests (for categorical variables) and Student´s t-test (for continuous variables) or nonparametric two-sample Wilcoxon (signed-rank) test for data far from normality distribution. To compare the relative abundance within the groups, genus-level summary abundance plots were generated for the 500 most abundant taxa.

Alpha-diversity: To determine microbial species diversity per group (PC cases, controls and pancreatitis), we calculated richness (number of species per sample) and two additional alpha-diversity measures: the Shannon and Simpson indexes based on Hill diversities [37,38]. Differences in these alpha diversity measures between groups were assessed by ANOVA and post hoc Tukey test to establish which specific groups differed in terms of diversity.

Community composition: To analyze trends in beta-diversity (community composition) between the groups of comparison, we calculated five different dissimilarity indexes: the Bray–Curtis index on non-transformed and square root transformed data, the abundance-weighted Jaccard-Chao index, and the unweighted and weighted TINA index [39]. Thereby we explored whether there are significant community-level compositional shifts between PC cases, controls and chronic pancreatitis patients both in the oral cavity and stool samples. To approach this analysis, the following steps were performed: 1) calculate pairwise sample compositional dissimilarities according to five beta-diversity indexes aforementioned, using custom codes developed at EMBL; 2) ordination analyses (Principal Coordinate Analysis, PCoA) for exploration and visualization, by projecting distance matrices to lower-dimensional Euclidean space; 3) test shifts in community composition for statistical significance by using permutational multivariate analysis of variance (PERMANOVA in adonis2 package in R) [40]. We accounted for potential confounders in PERMANOVA analyses (age, sex, center, smoking, as well as metformin use in stool samples) with 10,000 permutations. NMDs (non-metric multidimensional scaling) plots were also applied to explore variations in microbiome composition by potential confounders.

Per-taxa analyses: We first removed taxa with low overall abundance and prevalence (trimmed to retain ~ 200 taxa). Taxa that were differentially abundant between groups were detected using the non-parametric Wilcoxon test and the *edgeR* package in R to assess differential abundance between groups [41], followed by Benjamini-Hochberg (BH) multiple testing correction at an FDR-controlled p value cutoff of 0.05 [42]. Spearman correlations between gut and oral microbiota were also evaluated by characteristics of the study sample.

Signature, model evaluation and validation: Using the filtered data to retain sufficient overall abundance we developed the prediction model for PC. As described earlier [13], relative abundance data was normalized by log10 transformation and log-centred. Data were randomly split into test and training sets in a 10 times repeated 10- fold cross- validation, i.e., for each test fold, the remaining folds were used as training sets. Using the SIAMCAT R package (developed by EMBL-Heildelberg), we applied LASSO logistic regression models for feature selection [43]. The trained model was then used to predict the left-out test set and finally, all predictions were used to calculate the area under the curve (AUC) using the pROC R package. The obtained signature was further combined with other makers to test its predictive accuracy. Further details are available in Kartal, Schmidt, Molina-Montes, et al [13].

2.5.2. *Objective 2: Relationship between dietary factors with the signature and with PC risk*

Development of Dietary Scores (a priori patterns): Two dietary scores were calculated to evaluate their impact on PC risk and on the microbiome score, namely the Diabetes Risk Reduction Diet Score (DRRD) proposed by Kang et al [44], and the relative Mediterranean Diet score (rMED) proposed by Buckland et al [45]. Overall, quantile values were estimated among controls, and then applied to cases and controls for scoring. The scores were built on the overall PanGenEU and MICROBIOME study population (including 556 PC cases and 511 controls from Spain) to account for wider ranges of intakes.

To calculate the DRRD score, participants received a quintile value between 1 (intake consistent with the highest T2D risk) and 5 (for the lowest T2D risk) for the following five dietary factors: cereal fiber, nuts, coffee (caffeinated and decaffeinated), whole fruits, and ratio of polyunsaturated to saturated fat in ascending order; the scoring was reversed (1 for lowest and 5 for highest T2D risk) for other four dietary factors: glycemic index, trans fats, sugar-sweetened beverages, and red and processed meats. As in Turati et al, we modified the DRRDS by incorporating data on fruits and fruit juices in relation to diabetes. Also, sucrose intake was considered rather than glycemic index since this information was not available in our study. The DRRDS (range = 9-45) was the sum of the quintile values, whereby higher values relate to lower T2D risks.

The rMED score is an 18-point scale that incorporates nine selected components of the Mediterranean diet (MD). Each component was adjusted by energy density (g per1000 kcal per day), using the nutrient density model [46], and then divided into tertiles of intakes (except for olive oil). For the six components conforming the MD; fruits (including nuts), vegetables (excluding potatoes), legumes, fish (including seafood), olive oil and cereals (white and nonwhite), a score of 0-2 points was assigned to the first (0 points), second (1 point) and third (2 points) tertile of intake, respectively. For olive oil the scoring consisted of assigning 0 points to non-consumers, 1 point for participants below the median of intake and 2 points for levels of intake equal or above this median (10 g among controls). For the 2 components that do not conform with the MD, meat (including meat products) and dairy products, the scoring was reversed (first, second and third tertile: 2, 1 and 0 points, respectively). Because alcohol consumption has been potentially associated with PC [3], the alcohol component was removed from the rMED score. Thus, the range of the armed score contained eight components and the point scale ranks from 0 to, whereby 0 represents the lowest adherence to the MD pattern and the highest adherence.

Development of Microbial Risk Scores: Based on the signature previously identified in Kartal, Schmidt and Molina-Montes et al [13], a risk score was developed by applying several approaches. First, the 27 microbial species of the signature (Annex I) were combined in a microbial risk score (MRS) by summing up the relative counts of the positive microbial species, i.e., those with higher relative abundance among cases, while subtracting the negative ones (one minus the actual relative abundance); i.e. those with higher relative abundance among controls. Thus, higher scorings of the MRS were presumed to increase PC risk (Supplementary Figure 2). Microbial risk scores based on alpha-diversity measures on microbial signatures associated with a certain disease have been also proposed by some authors [47]. Therefore, we also calculated alpha-diversity measures on this sub-community of 27 microbial species. Specifically, richness and Shannon indexes were considered; the richness-Score (rS) and the Shannon-Score (SS), respectively. In this case, higher alpha-diversity scores were presumed to decrease PC risk. Thus, three microbial risk scores were used in association analyses with dietary factors: the MRS, the rS and the SS.

Association analyses: Linear regression models adjusted for age, sex, and center (Model 1), were used to test for associations between dietary factors (foods and nutrients) and alpha-diversity measures (outcome variable, on the log-scale for normalization). Likewise, associations were explored with regard to the MRS. Logistic regression models were also applied to test for associations between the dietary factors and the diet scores with PC risk (outcome variable), adjusting for potential confounders in multivariate models (Model 1 plus smoking status, T2D status family history of PC, energy intake and obesity). Results on the association analyses derived from logistic regression analyses are presented as Odds Rations (OR) and 95% Confidence Intervals (CI); those of the linear regression analyses are presented as beta-coefficients. Either two coefficients were calculated per 2-unit increment or per 1-SD (Standard Deviation) of the predictor variables. Potential effect modification by other variables were examined in stratified analyses and via the Wald test in models including interaction terms. P-values were corrected by BH. The study population comprised the overall PanGenEU and MICROBIOME study population (556 PC cases + 511 controls from Spain) for association analyses with PC risk, and the MICROBIOME study (50 PC cases + 50 controls) for association analyses with the microbial signatures.

Feature selection methods: In order to identify the most relevant dietary factors related to the MRS, as well as the most important microbial species (of the signature or the top 50 most prevalent ones) related to the dietary scores, several feature selection method (Ridge Regression, LASSO or Elastic Net regression-ENET) were examined (Supplementary Table 3). Optimal tuning parameters were calculated on a training and test set using a 10-fold cross-validation (and 5 repeats) procedure (R packages glment and mlbench). Overall, Ridge Regression for dietary variables (best tune for alpha 0 and lambda 1) and LASSO for taxa (best tune for alpha 1) were deemed more suitable (lowest root mean square error- RMSE, and highest explained variance - R^2). The selected variables were ranked according to variable importance.

2.5.3. Objective 3: Clustering of dietary factors with microbial species and the signature among PC cases and controls

Clustering of dietary factors and bacterial species: Unsupervised hierarchical clustering analysis was applied on the dietary factors (foods, food groups and nutrients) and the microbial species of the gut and oral microbiota using R package pheatmap. To reduce the input of microbial species, we restricted the analyses to the 50 most prevalent bacterial species. Also, the 27 microbial species of the previously identified gut signature were used in these analyses.

The distance matrix was defined by Euclidean and also by Manhattan distances, and Ward's method was used as linkage criteria to group the clusters. More specifically, based on the distance matrix, the clustering algorithm identified the closest observations (i.e., subjects with similar dietary and bacteria, in rows) and iteratively merged them within the same cluster until all clusters were merged together. Two or three clusters were retained, which were assumed to be the optimum number of clusters (2 to 5 clusters were tested) based on the silhouette method (R package Nbclust). These analyses were run overall to explore separations of groups of samples, and separately among cases and controls.

Correlations between dietary factors and microbial risk scores: Spearman correlation analyses were conducted to explore the strength of the association between dietary factors (foods and nutrients) and the microbial risk scores (the MRS, the rS and the SS) among cases and controls.

All statistics were conducted using software R-project (version 3.4 in the first study, and 4.2 in the MICROBIOME-dietary study) [48]. Overall, p-value of 0.05 were considered as statistically significant. In addition to some aforementioned R packages, we used specific packages for microbiome data analysis, such as Phyloseq [49], Microbiome [50] and Vegan [51]. The codes used for data analysis are available at: https://github.com/memmontes/FMP-Master-Omics

3. Results

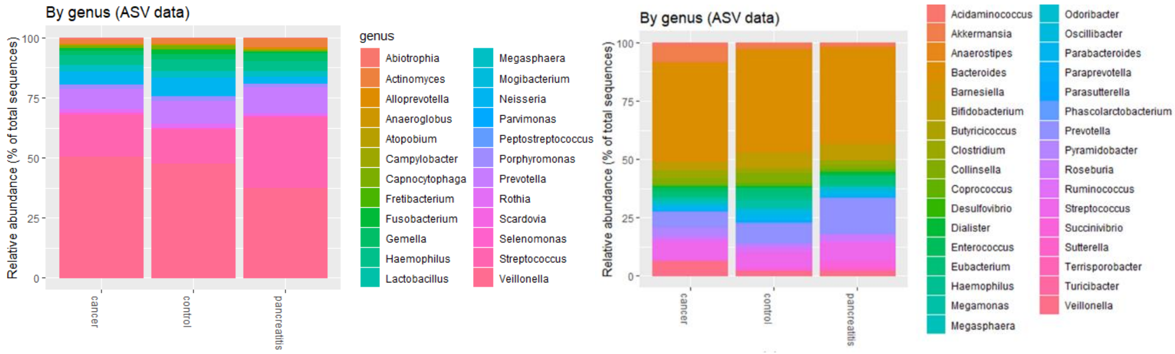
Results on aim 1 are shown in Reference 13 (Annex II). Results not reported in this scientific publication, mainly regarding the 16S rRNA data, are reported herein. Results on aims 2 and 3 are also non-published results.

Characteristics of the study populations are presented in Supplementary Table 4.

3.1. Descriptives on microbiome data and diversity measures

Figure 1 shows the prevalence of the most common bacterial species among PC cases and controls derived from 16S rRNA data. There were appreciable differences in the prevalence of several genera among groups. For instance, Veillonella spp. genera were enriched among PC patients in both saliva and stool samples. Similar findings were observed for the open-OTU and shotgun metagenomics data (data not shown).

Figure 1. Summed genus-level abundances of 500 most common taxa by disease status in saliva (left) and stool samples (right) at genus-level (ASV data). 16s rRNA data



Differences in alpha diversity measures in 16S rRNA data between groups were the following (Figure 2). In saliva, we found that PC cases had a significantly decreased microbial richness in comparison with controls (p-value=0.01 in ASV and p-value=0.05 in open-OTU). Significant differences were also noted for the other two alpha-diversity measures (e.g., p-value=0.01 and 0.02 in ASV data for Shannon and Simpson index, respectively). Similarly, CP patients had significantly reduced species richness, relative to controls (p-value=0.03 in ASV data only). In stool samples, there were statistically significant differences in diversity measures between PC and both controls and chronic pancreatitis patients (e.g., p-value=0.02 for richness in ASV data), though not so when comparing controls to CP patients. In shotgun data (Reference 13: Figure S2), the trends were in line with those reported for the 16S data, although in this case, diversity measures were significantly lower ($p<0.05$) in PC cases than in controls.

Figure 2. Alpha-diversity measures (richness, Shannon and Simpson) in PC cases, controls and CP patients in saliva (left) and stool (right) samples (ASV data). 16S rRNA data.

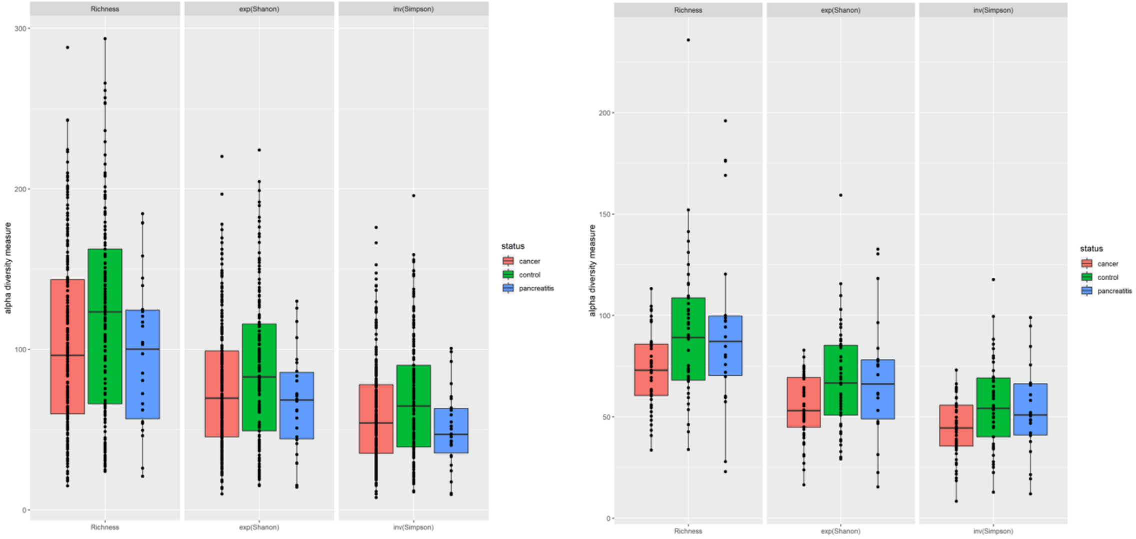
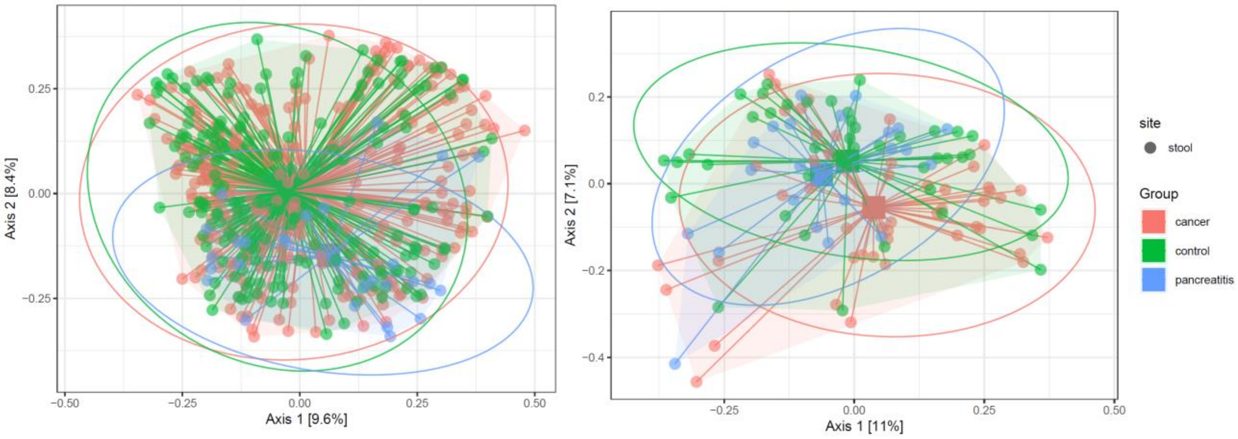


Figure 3 shows PCoA plots for Bray Curtis index, as the main beta-diversity measure derived from 16s rRNA data, for salivary and stool samples. Centroids for PC cases, controls and CP patients were far apart from each other in stool samples, pointing to different community compositions between the groups, while in saliva samples these centroids were closer, except for CP. Other beta-diversity measures showed similar results (Supplementary Figure 2). Regarding the metagenomic data, similar results were obtained (Reference 13: Figure 1B and Figure S3).

Figure 3. PCoA plots of saliva (left) and stool (right) samples for PDAC cases, controls and chronic pancreatitis patients. Bray Curtis index. 16s rRNA data.



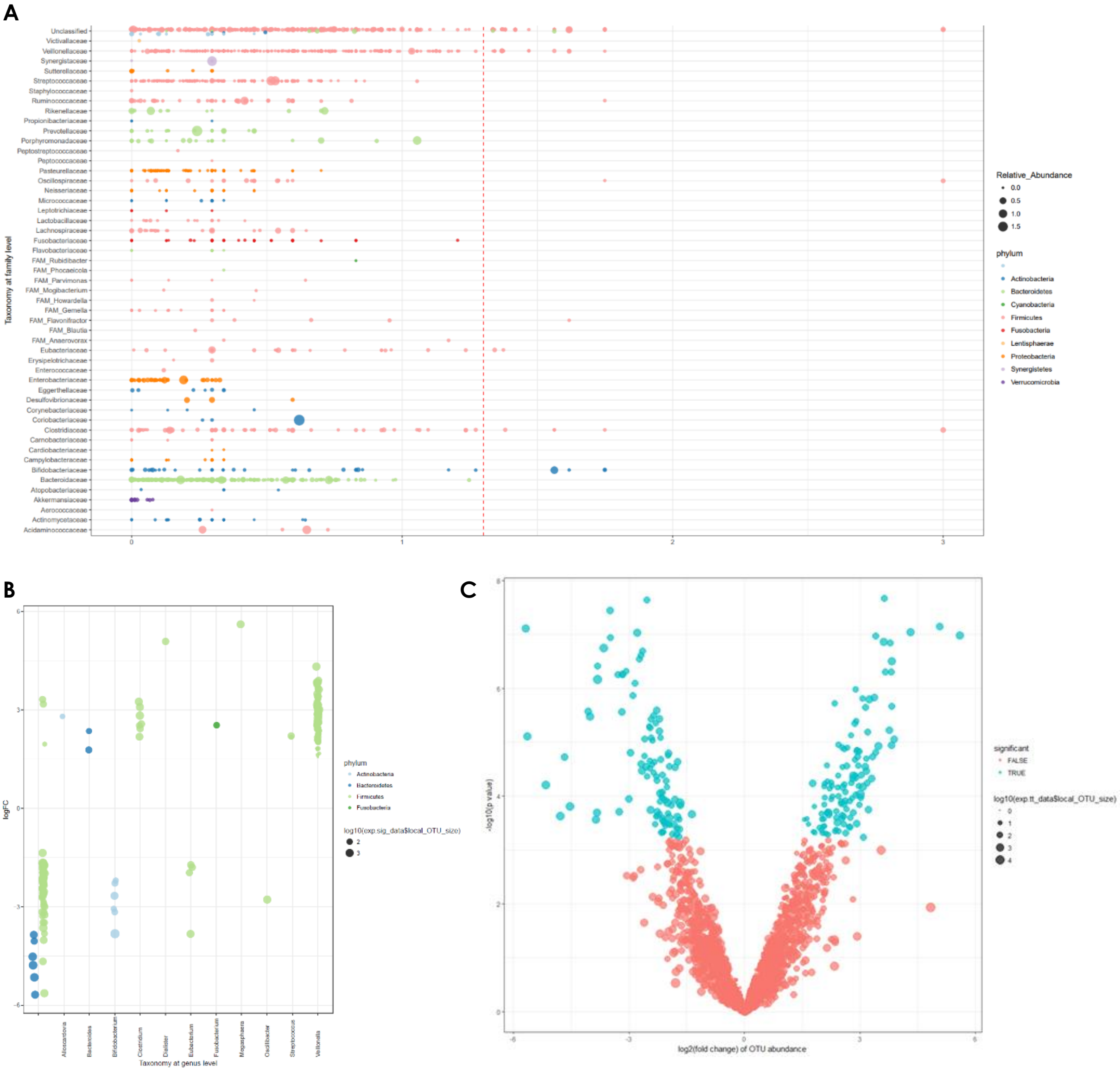
Beta-diversity indexes derived from 16S rRNA data were also used to test shifts in oral or gut microbiome composition according to disease status, controlling for potential confounders in PERMANOVA analyses (age, sex, saliva sampling method, and smoking, as well as metformin use in stool samples) with 10,000 permutations. In metagenomics data, confounders were diabetes and jaundice status in stool samples, and aspirin/paracetamol use in saliva samples (Reference 13). As shown in Supplementary Table 5 for 16S rRNA data, microbiome composition differed significantly between PC cases and controls in gut microbiome (Bray Curtis for ASV: p=0.0002) samples. Differences were also noted in oral samples between PC and CP patients (Bray Curtis for ASV: p=0.03). Interestingly, the statistically significant differences in community composition between CP and the controls were lost after controlling for confounders. Again, similar trends were observed for the open-OTU data (data not shown). Results for the metagenomic data are described in detail in Reference 13. In brief, disease status was significantly associated with community composition in stool (R2=0.02, p=0.001), but not in saliva (R2=0.01, p=0.5).

The potential confounding effect of variables was established with regard to alpha-diversity (Reference 13: Table S4) and beta-diversity measures (Reference 13: Table S5) in metagenomics and in 16S rRNA data (Suplementary Figure 3, and data not shown).

3.2. Per-taxa analyses and development of the microbial signature for PC

Several differentially abundant species (N=220) were found in stool (Figure 4A and 4B), and to a much lesser extent in saliva (data not shown). Both, 16s rRNA ASV and OPEN data retrieved similar results. As described in Reference 13 (Figure 1C), in shotgun data, species such as *Veillonella atypica* and *Fusobacterium nucleatum* were more abundant in stool samples of PC patients, whereas others (e.g., *Bifidobacterium bifidum*) were depleted. There was no significantly differential abundance in the oral microbiome (Reference 13, Figure S5).

Figure 4. Differentially abundant genus (per-taxa associations) in stool samples between PC cancer and control samples. 16S rRNA data



A) Wilcoxon test results of 16S rRNA stool on differentially abundant taxa between PC and control cases. X- axis is log10(FDR corrected p values). B) Edge R results, adjusted for age, sex and center. Y- axis is fold change, and dot size represents the relative abundance of a given species. C) Volcano plot of EdgeR results showing differentially abundant taxa between PC cases and controls. Blue dots represent significantly differentially abundant species in either group, while red dots show non-significant species after FDR correction. FDR, false discovery rate.

As described in Reference 13, the LASSO regression model (with 10-fold cross-validation) selected 27 faecal bacterial species as predictive features of PC, while any other metadata variable was selected by this model. This bacterial classifier of PC was therefore considered as independent of potential confounders. The AUC of this signature to predict PC reached 0.84. Some of these markers were enriched among PC cases (*Veillonella atypica*, *Fusobacterium nucleatum/hwasookii*, *Alloscardovia omnicolens*, etc) whereas others were less abundant (*Romboutsia timonensis*, *Faecalibacterium prausnitzii*, *Bacteroides coprocola*, *Bifidobacterium bifidum,* etc). In contrast, there was no robust signature derived from the salivary samples. As a consequence, the signature emerged from the shotgun metagenomics data only. The predictive accuracy did not improve after combining oral and gut microbiota samples. Importantly, it improved to a high extent (AUC=0.94) when the stool metagenomic signature was combined with the marker CA-19.9.

More details on the robustness of each feature (bacterial marker), its relative abundance among PC cases and controls and the overall predictive accuracy (AUC and ROC curves) are available in Reference 13, Figure 2, as well as in the supplementary material of this article (Figure S11). Results of the external validation study are also reported in this article (see Methods section in Reference 13).

3.3. Relationship between dietary factors, individually and collectively in diet scores, with PC risk

Table 1 shows results on the association analyses between dietary factors (nutrients and food groups) with PC risk. OR and 95%CI are given per 1 SD increase in intake. Results for foods on an individual basis are shown in Supplementary Table 6. In multivariate adjusted models for age, sex, center, T2D status, smoking in pack-years, family history of PC, and total energy intake in kcal, the most prominently associated dietary factors with PC risk were: Energy intake (in models without adjustment for this variable) (OR_{per1SD}=1.2), coffee intake (OR_{per1SD}=1.3) and consumption of canned fish (OR_{per1SD}=0.8). Other potential associations were lost after multiple testing correction. Energy intake influenced these associations the most (data not shown). In models without adjustment for energy intake, it was observed that energy intake, carbs, Vitamin E and coffee intake were positively associated with PC risk. Indeed, PC risk increased on overage by 10-30% (OR=1.1 to 1.3) per 1 SD increase in the intake of these dietary factors. Regarding other dietary factors, no other associations remained after multiple testing correction (p-values corrected by BH, p.bh>0.05).

Figure 5 shows results in the association between dietary factors when combined in diet scores with PC risk. OR and 95%CI are given per 2-units increase in adherence to these scores. Details on the distribution of these components in the study population are shown in Supplementary Table 7. While the DRRD score was expected to decrease the risk of developing PC given its potential inverse association with T2D, we observed a trend towards a positive association between adherence to this dietary score with PC risk in crude and multivariate adjusted models, even after removing recently diagnosed patients (<2 years) with T2D (Figure 5A). Interestingly, the association got non-significant when adjusting for T2D status in the models, which supports the potential confounding effect of this variable on this association. In addition, it was observed that the DRRD score increased the risk of PC significantly among T2D subjects ($OR_{per2units}$=1.25), but not so among non-T2D subjects (p>0.05). Thus, the DRRD does not have any potential prevention effect of PC risk if T2D is present. Moreover, interaction by T2D status tended to be statistically significant (p-value for interaction=0.05). These results reinforce that the DRRD score is likely to have an indirect impact on PC risk and on T2D, its main risk factor. There was no significant interaction by obesity (obesity and non-obesity) between the DRRD score and PC risk (p>0.05), as also shown in subgroup analyses. The same was true in subgroup analyses by sex (p for interaction by sex=0.09) despite a positive and significant association between adherence to the DRRD score and PC risk was apparent in women ($OR_{per2units}$=1.13; p<0.05), though not in men ($OR_{per2units}$=0.98; p>0.05). Adjustment for T2D status has almost no impact on these associations by subgroups. By adjusting for every component at once, it was seen that the DRRD score remained significantly associated with PC risk (Figure 5B).

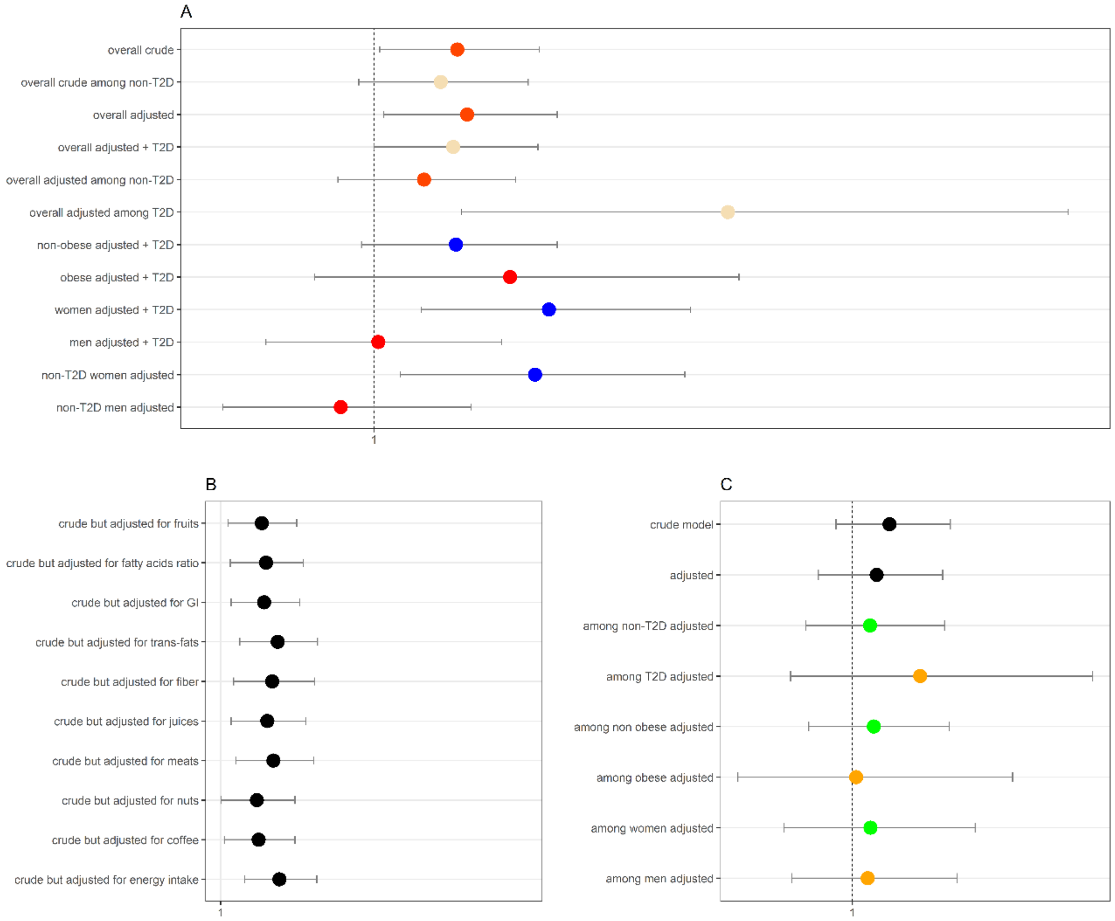
Table 1: Association between dietary factors (nutrients and food groups) and PC cancer risk within the PanGenEU study.

Nutrients	OR	LCI	HCI	P.value	p.bh	Food groups	OR	LCI	HCI	P.value	p.bh
kcal	1.21	1.06	1.38	**<0.001**	**<0.001**	**galldairy**	1.01	0.87	1.16	0.920	0.985
Lipids.g	0.95	0.72	1.25	0.700	0.770	**-sgmilkyogurt**	1	0.87	1.16	0.960	0.985
Proteins.g	0.71	0.56	0.91	**0.010**	0.220	**-sgcheesse**	0.99	0.86	1.14	0.880	0.978
Humidity.g	0.89	0.72	1.11	0.310	0.546	**-sgdairydessert**	1.05	0.91	1.2	0.510	0.756
Carbs.g	1.29	0.94	1.77	0.110	0.403	**gallmeat**	0.84	0.72	0.98	0.030	0.280
Sucrose.g	0.93	0.79	1.09	0.390	0.636	**-sgwhitemeat**	0.85	0.74	0.98	0.030	0.280
Fiber.g	1.01	0.81	1.25	0.960	0.982	**-sgredmeat**	0.91	0.78	1.05	0.180	0.525
Starch.g	1.06	0.9	1.26	0.490	0.653	**-sgorganmeat**	0.98	0.86	1.12	0.780	0.918
Sugar.g	1.1	0.93	1.31	0.250	0.458	**-sgcuredmeat**	1.03	0.9	1.18	0.650	0.839
Cholesterol.mg	0.86	0.73	1.01	0.060	0.264	**sgprocessedmeat**	0.91	0.79	1.05	0.210	0.525
VitA.ug	0.9	0.78	1.04	0.160	0.416	**sgcuredprocessed**	0.95	0.82	1.1	0.510	0.756
VitD.ug	0.9	0.78	1.04	0.170	0.416	**gallsea**	0.94	0.82	1.08	0.360	0.696
VitE.mg	1.12	0.89	1.4	0.330	0.558	**-sgfish**	0.97	0.85	1.11	0.700	0.848
VitB8.ug	1.04	0.89	1.2	0.640	0.741	**-sgothersea**	0.85	0.73	0.98	**0.020**	0.280
VitB9.ug	0.82	0.67	1	**0.050**	0.244	**gmeatseafood**	0.83	0.71	0.97	**0.020**	0.280
VitB3.mg	0.78	0.62	0.97	**0.030**	0.244	**gallreadydishes**	0.93	0.81	1.06	0.250	0.556
VitB5.mg	0.9	0.77	1.05	0.190	0.418	**gallvegetables**	0.9	0.78	1.05	0.190	0.525
VitB2.mg	0.89	0.73	1.08	0.230	0.458	**-sg1leafyveg**	0.88	0.77	1.01	0.060	0.280
VitB1.mg	0.8	0.62	1.02	0.080	0.320	**-sg1starchveg**	1	0.86	1.16	0.990	0.990
VitB12.ug	0.9	0.78	1.04	0.170	0.416	**-sg1sgfruitingveg**	0.96	0.83	1.1	0.550	0.772
VitB6.mg	0.91	0.78	1.05	0.180	0.417	**-sg1sggrainsveg**	0.93	0.81	1.06	0.290	0.611
VitC.mg	0.94	0.8	1.11	0.460	0.653	**-sg2redyellveg**	0.94	0.82	1.08	0.400	0.696
Calcium.mg	0.94	0.77	1.14	0.530	0.666	**-sg2greenveg**	0.91	0.79	1.05	0.200	0.525
Iron.mg	0.85	0.66	1.11	0.240	0.458	**-sg2whiteveg**	0.87	0.75	1	0.050	0.280
Potasium.mg	0.91	0.71	1.17	0.460	0.653	**glegumes**	0.97	0.85	1.12	0.680	0.848
Magnesium.mg	0.78	0.57	1.06	0.120	0.406	**gallfruits**	1.06	0.91	1.25	0.440	0.733
Sodium.mg	0.79	0.64	0.99	**0.040**	0.244	**golives**	1.06	0.93	1.22	0.400	0.696
Phosphorus.mg	0.75	0.57	0.99	**0.040**	0.244	**gnuts**	1.16	0.99	1.35	0.070	0.280
Copper.mg	0.99	0.86	1.13	0.850	0.890	**gallcereals**	1	0.86	1.17	0.950	0.985
Iodide.ug	0.89	0.74	1.07	0.230	0.458	**-sgbread**	1.03	0.89	1.2	0.650	0.839
Selenium.ug	0.81	0.67	1	**0.050**	0.244	**-sgricepasta**	0.99	0.86	1.13	0.830	0.949
Zinc.mg	0.73	0.56	0.95	**0.020**	0.244	**gallfats**	1.07	0.92	1.25	0.380	0.696
Linoleic.g	1.07	0.89	1.29	0.480	0.653	**gflour**	1.05	0.91	1.21	0.470	0.752
Linolenic.mg	1.07	0.89	1.29	0.480	0.653	**gchocolate**	1.12	0.98	1.29	0.110	0.367
Araquidonic.mg	0.84	0.72	0.98	**0.030**	0.244	**gsugars**	1.15	1	1.33	0.050	0.280
DHA.g	0.9	0.78	1.04	0.170	0.416	**gsauces**	1.09	0.94	1.27	0.250	0.556
EPA.g	0.9	0.78	1.04	0.160	0.416	**gsweetenedbev**	0.89	0.77	1.02	0.100	0.364
Estearic.g	1	0.83	1.21	0.990	0.990	**-sgsugarsweet**	0.88	0.76	1.01	0.070	0.280
Lauric.g	1.03	0.9	1.19	0.660	0.745	**-sgartificialsweet**	0.96	0.84	1.1	0.560	0.772
Miristic.g	0.95	0.8	1.11	0.510	0.660	**-sgjuices**	0.87	0.76	1	0.050	0.280
PUFA.g	0.97	0.79	1.19	0.750	0.805	**gcoffee**	1.26	1.09	1.45	**<0.001**	**<0.001**
SFA.g	0.9	0.71	1.15	0.420	0.653	**gdecoffee**	0.97	0.64	1.47	0.880	0.942
Trans.g	0.96	0.82	1.12	0.620	0.737	**gtea**	1.1	0.97	1.25	0.140	0.580
MFA.g	1.07	0.85	1.35	0.550	0.672						

Multivariate adjusted logistic regression models adjusted for age in years (continuous), sex (men, women), center (all five Spanish hospitals), diabetes status (no diabetes, diabetes: diagnosed less than 2 years, or since more than 2 years), pack-years of smoking (non-smokers, tertiles of pack-years), obese (no, yes: BMI>30 kg/m^2), and family history of PC (no, yes), as well as energy intake in Kcal. OR and 95% CI are derived from these models, and are related to PC risk per 1 SD increase in the intake of the dietary variable. P-values were corrected for multiple comparison testing by the Benjamini-Hochberg BH method (p.bh). For foods groups, the main group is indicated with "g" along with the corresponding subgroups "sg" before the food group´s name.

The apparent positive association between the DRRD score and PC risk seemed to diminish when removing the effect of nuts and coffee from the score. Among dietary factors, energy intake, again, had the biggest influence on this association. Regarding the rMED score (Figure 5C), we did not observe any significant association between this score with PC risk, neither in crude models, nor in multivariate adjusted models or by subgroups (all p-values>0.05).

Figure 5: Association between dietary scores (the DRRD and the rMED score) and PC cancer risk within the PanGenEU study.



Crude and multivariate adjusted logistic regression models as indicated. The latter were adjusted for age in years (continuous), sex (men, women), center (all five Spanish hospitals), T2D status (no T2D, T2D: diagnosed ≤ 2 years, or > 2 years), pack-years of smoking (non-smokers, tertiles of pack-years), obese (no, yes: BMI>30 kg/m^2), family history of PC (no, yes), and energy intake in Kcal. OR and 95% CI are derived from these models (dots and horizontal lines, respectively), and are related to PC risk per 2 units increase in adherence to the Diet score. A) Association between the DRRD score with PC risk overall and by subgroups. Recently diagnosed T2D was removed in all analyses. B) Association between the DRRD score with PC risk controlling for each component of the score at once. C) Association between the rMED score with PC risk overall and by subgroups

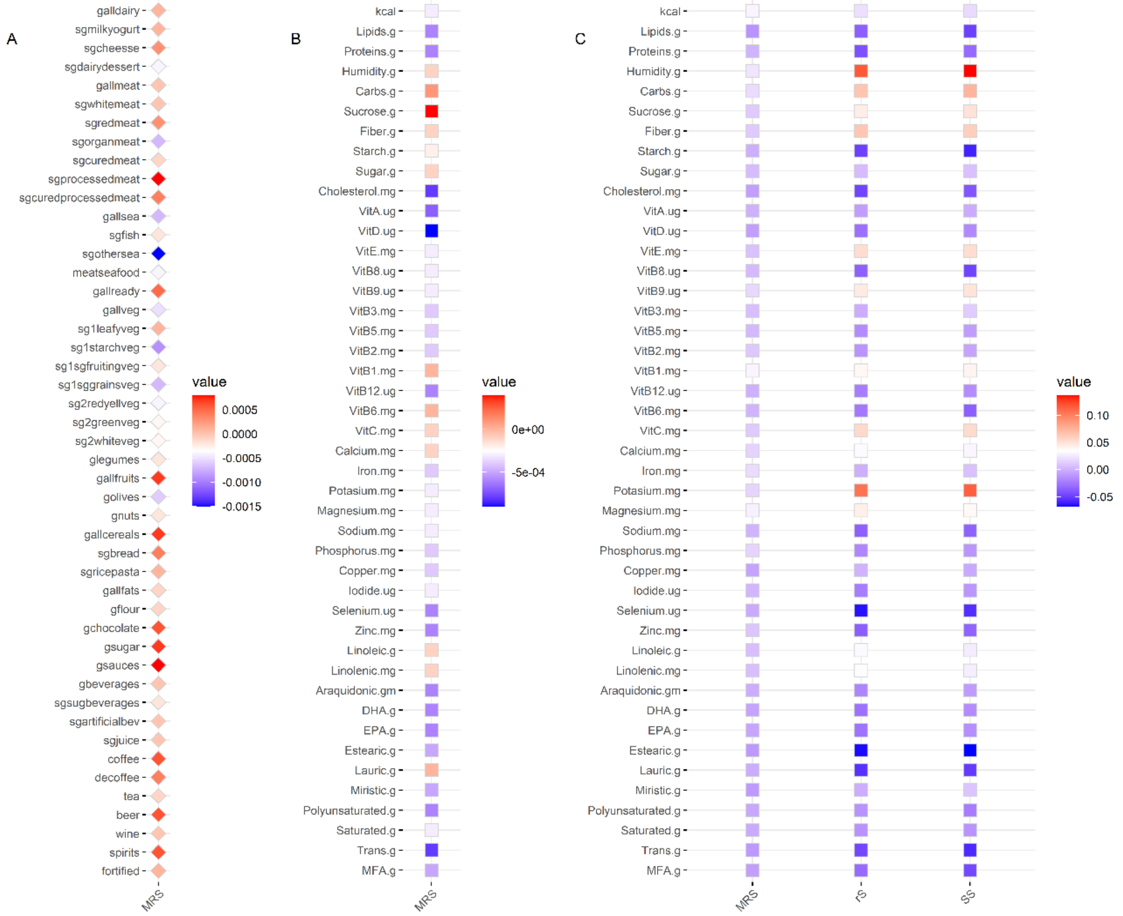
3.4. Relationship between dietary factors with the microbial signature (risk scores)

The risk scores based on the metagenomic signature of PC (Reference 13) was derived from the MICROBIOME study. The MRS, which distribution is shown in Supplementary Figure 1, was significantly associated with PC risk in age, sex and center-adjusted logistic regression models (OR_{per1SD}~5.5, p= 3.01E-05). As shown in Supplementary Table 8, multivariate adjusted linear regression models exploring associations between this score (log-transformed MRS) and the dietary factors (foods, food groups and nutrients) revealed that: high consumption of canned fish and of other seafoods such as squids, octopus and cuttlefish was inversely associated with the MRS (β_{per1SD}=-0.001; R2=0.24), and high consumption of hamburgers were positively associated with the MRS (β_{per1SD}=0.001; R2=0.25). Indeed, by food groups (Figure 6), the group of seafood emerged as inversely associated with the MRS. These associations were statistically significant at the nominal level; none of them held after multiple testing correction (p.bh>0.05). Other food groups tended to be associated with the MRS (starchy vegetables, and vegetables in general, processed foods and sugary foods), but did not reach statistical significance. There was no significant association between any nutrient and the MRS (Figure 6).

Regarding the risk signature in terms of alpha-diversity measures, the rS (richness score) and the SS (Shannon score), the following statistically significant results were observed (Supplementary Table 8 and Figure 6): positive associations with high intake of water and of potassium and some positive trends for overall fruits and few vegetables, and negative associations with fortified alcoholic beverages. P-values corrected by BH turned all non-significant. It is important to note that both scores were not associated with PC risk in age, sex and center-adjusted regression models. Nonetheless, an inverse non-significant association with PC risk was manifest (OR~0.8) given the link between a high taxonomic diversity and a healthy microbiota. Despite their weak association with PC risk, their association with dietary factors was more prominent if compared to the MRS (lower effect sizes) (Figure 6). Overall, further adjustment for use of probiotics had a negligible impact on the results, whereas adjustment for energy intake had a relatively high influence on the estimates.

Besides, there was a nearly significant association between DRRD score and the MRS in age, sex and center-adjusted linear regression models (per 1SD increase in MRS, the DRRDS increased on average by 0.04 points; p=0.07). This association weakened with other variables in the model including T2D. Associations with the rMED score were non-significant (β=-0.03 per 1SD increase in rMED, p=0.45). Both diet scores were not associated with the alpha-diversity scores. Also, there were no robust associations between dietary factors and alpha-diversity measures of the oral and gut microbiome, overall, within the PanGenEU study after multiple testing correction (data not shown).

Figure 6: Association between dietary factors and microbial risk scores (MRS, rS and SS) derived from stool samples and metagenomics data within the MICROBIOME study.



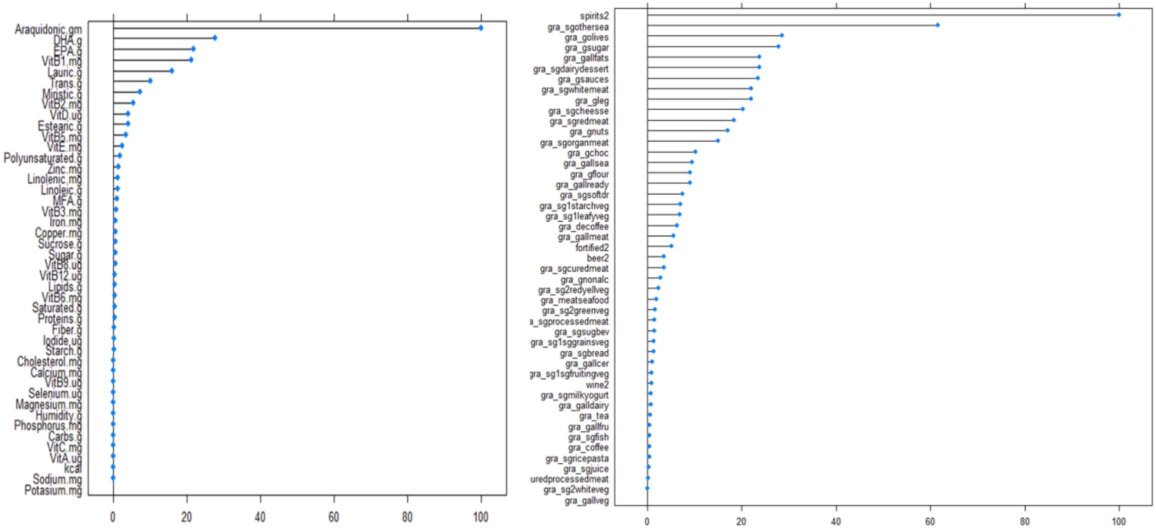
Multivariate adjusted linear regression models adjusted for age in years (continuous), sex (men, women), center (all five Spanish hospitals), T2D status (no T2D, T2D: diagnosed ≤ 2 years, or > 2 years)), pack-years of smoking (non-smokers, tertiles of pack-years), obese (no, yes: BMI>30 kg/m^2), family history of PC (no, yes), and energy intake in Kcal. β coefficients and 95% CI are derived from these models, per 1 SD increase in intake of the dietary variables. Microbial risk scores (outcome variable) were log-transformed to approximate a normal distribution. P-values were corrected for multiple comparisons by BH (p.bh). A and B) MRS associations for food groups and nutrients; C) Associations between all scores in a comparative manner. For foods groups, the main group is indicated as “g” along with the corresponding subgroups “sg” before the food group´s name.

3.5. *Selection of dietary factors (feature selection)*

Given the Correlation between dietary variables (Supplementary Figure 4), appropriate feature selection methods were applied. Figure 7 shows results on the features selected by Ridge Regression regarding the MRS. The most important features were araquidonic acid (AA), Eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) in the case of nutrients, all being polyunsatured fatty acids (PUFAs), seafood, and spirits (alcoholic beverages), olives, sugars and other less healthy foods. Somewhat similar results were obtained for the other microbial scores, the rS and SS; some saturated fatty acids were also selected including trans-fatty acids (negatively) and meristic acid (positively) (Supplementary Figure 5).

As for the taxa metagenomic features and relative to the diet scores (using LASSO as feature selection method), only *Clostridiales species* was selected with regard to DRRDS (negatively), whereas no taxa were selected for the rMED score (data not shown). When considering the most prevalent gut taxa (Supplementary Figure 6), it was seen that bacteria such as *Bacteroides caccae* (negatively) and *Prevotella sp CAG.279* (positively) were selected for DRRDS, while for rMED some selected bacteria were, again, *Prevotella sp CAG.279* and *Faecalibacterium.prausnitzii r_06110* (positively). Interestingly, in conventional linear regression analyses, these bacteria were also significantly associated with the dietary scores, even after multiple testing correction.

Figure 7: Feature importance of Nutrients (left) and food groups (right) selected by Ridge Regression models for the MRS.



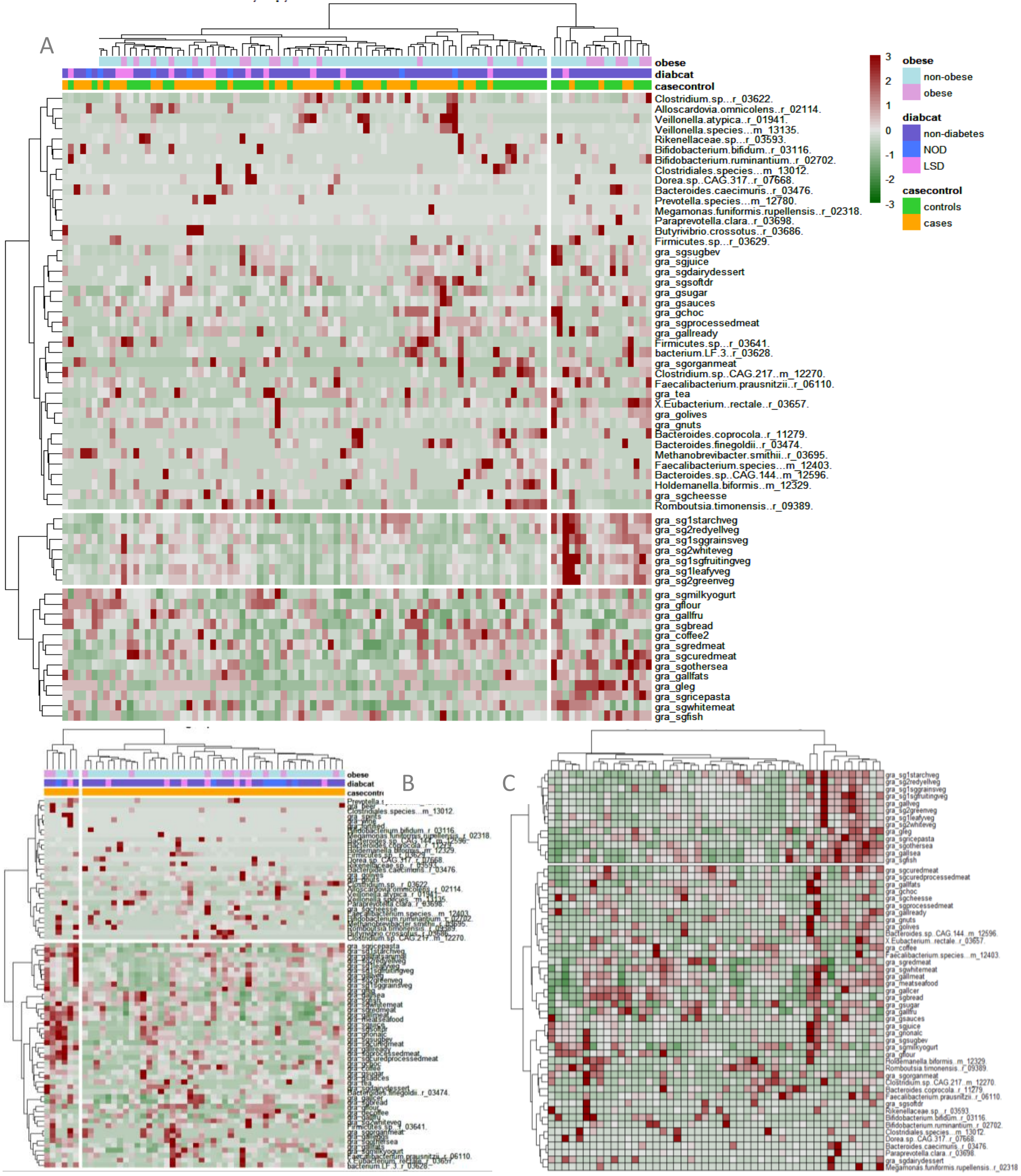
Feature importance plots, where variables were scaled relative to 100 according to scores. Highest coefficients were: spirits (1.29E-04), seafood (-7.86E-05), all fats (olive oil) (-1.06E-05), white meat (1.61E-05), olives (-4.56E-05), nuts (1.92E-05), sugar (2.07E-05), sauces (2.88E-05), araquidonic acid (-3.96E-03), DHA (-1.09E-03), EPA (-8.60E-04).

3.6. Clustering between dietary factors and taxa from the oral and gut microbiota

Cluster analyses retrieved two main dietary clusters among controls, a more Westernized dietary pattern (richer in consumption of less healthy foods including soft drinks, processed meat, ready dishes, etc.) and a prudent dietary pattern (Supplementary Figure 7). By combining dietary factors with microbial taxa, in general, two separate dietary clusters were observed among cases and controls. Figure 8 shows results on the clustering of dietary factors as food groups and the 27 microbial taxa of the microbial signature. The obtained clusters showed a group of overwhelmingly controls (right side) that featured a lower enrichment of taxa associated with PC disease (for example, *Alloscardovia omnicolens* and *Veillonella species)*, along with a higher enrichment of those associated with the controls in Reference 13 (for example, *Clostridium sp. CAG:217* and *Faecalibacterium prausnitzii r06110*). This group of subjects also featured a higher consumption (above the mean intake) of plant-based foods (all type of vegetables, cereals and legumes). Interestingly, seafood consumption was also relatively high in this group. There was also an unexpectedly low intake (below the mean) of fruits in this group. The other group of subjects (on the left) comprised a higher proportion of PC cases and showed to some extend opposite trends with regard to the intake of plant-based foods and the presence of the above mentioned taxa. By rows, the first cluster corresponded to all taxa and few food groups clustered within this group (mainly those that characterize a Westernized dietary pattern). No consistent clusters were obtained when examining foods individually (data not shown). Regarding nutrients, mixed results were obtained. For instance, DHA and other fatty acids clustered with taxa of the metagenomic signature, but not so when Manhattan distances were considered (Supplementary Figure 8).

Looking separately at PC cases and controls, it appeared that alcoholic beverages were more aggregated among PC with higher enrichment of pathogenic taxa. These patients had an overall high consumption pattern of all foods. Among the controls, no clear pattern was seen except the two common dietary patterns (Figure 8B and C). Similar findings were observed, with slight differences, when using other clustering distance methods (Supplementary Figure 9). Supplementary Figure 10 shows the same but relative to the most prevalent taxa in the gut microbiota. Less meaningful results for the oral microbiota (50 most prevalent taxa), were obtained for both metagenomics and 16s RNA data (data not shown).

Figure 8: Clusters of foods groups and the stool microbial signature taxa (27 species) overall and separately for PC cases and controls. Manhattan distance.



Hierarchical cluster obtained with Manhattan distances applied to taxa and food groups, after scaling all values y rows (value-mean/SD). For foods groups (all in grams of intake "gra"), the main group is indicated with "g" along with the corresponding subgroups "sg" before the food group´s name. A) Among PC cases and controls; B) Among PC cases; C) Among controls. For the latter, all taxa enriched among PC cases had to be removed to avoid zero values across rows and columns.

4. Discussion

This is the first study to elucidate both the within-sample diversity and individual components of the gut microbial community in association with dietary features within a Spanish study of PC patients.

This study has also unraveled a faecal metagenomic signature for PC diagnosis [13], and provides novel insights on the effect of the consumption of 82 food items, 38 food groups, 44 nutrients and two *a-priori* derived dietary patterns, on the overall microbial diversity of species richness in the salivary and faecal microbiota, and on the gut metagenomic signature.

Based on the above, a microbial risk signature was defined by considering summation of the relative abundance (MRS) or alpha-diversity measures (richness and Shannon, rS and SS, respectively). The latter approach was proposed earlier [47], whereas the first is founded on the principles of a generic risk score, where values are summed across samples. The current study shows that the MRS reflected better PC risk than those based on alpha-diversity measures. Indeed, positive association between the MRS and PC risk was expected, whereas negative associations with the others (a higher microbial diversity is supposed to be healthier) [47]. Importantly, any of these scores were associated with dietary factors (foods, food groups or nutrients) after multiple testing correction in linear models. Therefore, feature selection methods accounting for multicollinearity effects of dietary variables were used to score and select the best features. Ridge Regression was used, this being a common method to deal with dietary variables despite it shrinks coefficients towards zero [52]. Predictors of the MRS were some PUFAs (AA, EPA and DHA), which are mainly contained in seafood including canned fish, and some alcoholic beverages. The effects of omega-3 PUFAs on intestinal immunity and inflammation have been described in several studies [53,54]. These studies support that variability in omega-3 PUFA metabolism can be driven by cancer. Therefore, PC could influence omega-3 PUFAs-microbiome-immune system interactions.

Also, after multiple comparison adjustments, no significant associations were observed between diet consumption (foods, food groups or nutrients) and overall richness of the oral or gut microbiota from either 16S rRNA or metagenomics data. However, cluster analyses pointed to the existence of groups of subjects, and groups of taxa (of the gut microbial signature) and dietary factors. For instance, the group of controls tended to have a higher enrichment of beneficial taxa and a higher consumption of plant-based foods, as well as higher intake of PUFAs. To account for diet on a holistic way, two *a priori* dietary patterns were obtained, one for reducing T2D risk (the DRRD score) and another one to resemble a plant-based diet (the rMED score). Only the DRRD score was significantly associated with PC risk, although confounding and interaction for T2D was manifest. Therefore, both dietary scores did not appear to have a major impact on PC risk. In linear regression and by applying feature selection methods, some relevant gut taxa were selected: *Faecalibacterium prausnitzii,* which is part of the metagenomic signature in the case of rMED score, and some Prevotella sp in both dietary scores. Both are common in populations consuming a plant-rich diet [16]. In addition, *F. prausnitzli* has been linked to dietary fiber [16]. It is also important to note that among the controls, there were two dietary patterns present in the Spanish population: a more plant-based and a more Westernized dietary pattern [55]. Previous studies have shown the association of plant-based foods, which feature a high intake of bioactive compounds including fiber, with a healthy gut microbiota [15]. However, dietary fiber was not a key factor in this study.

The main limitation of this study is the sample size. Indeed, it is likely that significant results were not achieved due to this issue. Also, collection of dietary data is prone to measurement error [46]. Other biases that are likely relate to the collection of other variables and biological samples, despite the use of standard protocols and questionnaires. Therefore, the effects of these biases on the results cannot be rule out. These and others, such as residual confounding, may have driven some unexpected results (e.g., coffee). Regarding strengths, this study used high-level data (metagenomics) on two sites, gut and oral microbiome, and 16S rRNA to confirm the tendency and direction of the results. Both showed differences in the microbiome composition and richness between PC cases and controls [13]. However, the metagenomic data led to more focused results and allowed the identification of a gut microbial signature for PC diagnosis. Therefore, a major strength of this study is the use of this kind of data.

5. Conclusion

The gut microbiota hosts bacterial species that constitute a valid biomarker/signature for PC detection, thus with high potential for PC screening and monitoring. Indeed, microbial diversity differs between and within PC patients in both 16S rRNA and metagenomics data. In contrast, our results support that the oral microbiota is less implicated in this disease. Certain dietary factors, such as specific seafoods and alcoholic beverages, as well as nutrients involved in the modulating the inflammatory and immune response (e.g., PUFAs), could be related to some extent to this microbial signature. Diet as a whole, however, does not seem to impact PC risk, but could be key to retain some relevant taxa. Future studies with larger sample size are warranted to confirm these findings.

Acknowledgements: This research was supported by the Centro Nacional de Investigaciones Oncológicas (CNIO). I am thankful to all the people from the Genetic and Molecular Epidemiology Group GMEG (Lidia, Lola, Evangelina, Fran, Raquel, etc.), and I am especially grateful to Dr. Núria Malats, group leader of this group, for providing to me all the necessary facilities needed to carry out this Master thesis, and for her advice, revisions and inspiring discussions. As far as I am concerned, any mistakes are mine and should not tarnish the reputations of the GMEG thankful to the José Castillejo fellowship that allowed me a research stay at EMBL-Heidelberg, where I got familiar with data analysis of microbiome data. I am also thankful to members of the EMBL-Heidelberg group (Ece Kartal, Sebastina Schmidt and Peer Bork) for their continuous support.

References

1 Lepage C, Capocaccia R, Hackl M, Lemmens V, Molina E, Pierannunzio D, Sant M, Trama A FJE-5 WG. Survival in patients with primary liver cancer, gallbladder and extrahepatic biliary tract cancer and pancreatic cancer in Europe 1999-2007: Results of EUROCARE-5. *Eur J Cancer* 2015;**51**:2169–78. doi:10.1016/j.ejca.2015.07.034

2 Kleeff J, Korc M, Apte M, *et al.* Pancreatic cancer. *Nat Rev Dis Prim* 2016;**2**:16022.http://dx.doi.org/10.1038/nrdp.2016.22

3 Zanini S, Renzi S, Limongi AR, *et al.* A review of lifestyle and environment risk factors for pancreatic cancer. *Eur J Cancer* 2021;**145**:53–70. doi:10.1016/j.ejca.2020.11.040

4 Brandi G, Turroni S, McAllister F, *et al.* The human microbiomes in pancreatic cancer: Towards evidence-based manipulation strategies? *Int J Mol Sci* 2021;**22**:9914. doi:10.3390/ijms22189914

5 Yang Q, Zhang J, Zhu Y. Potential Roles of the Gut Microbiota in Pancreatic Carcinogenesis and Therapeutics. *Front Cell Infect Microbiol* 2022;**12**. doi:10.3389/fcimb.2022.872019

6 Ren Z, Jiang J, Xie H, *et al.* Gut microbial profile analysis by Miseq sequencing of pancreatic carcinoma patients in China. *Oncotarget* 2017;**8**:95176–91. doi:10.18632/oncotarget.18820

7 Riquelme E, Zhang Y, Zhang L, *et al.* Tumor Microbiome Diversity and Composition Influenc. *Cell* 2020;**178**:795–806. doi:10.1016/j.cell.2019.07.008.Tumor

8 Fan X, Alekseyenko A V, Wu J, *et al.* Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut* 2018;**57**:120–7. doi:10.1136/gutjnl-2016-312580

9 Pushalkar S, Hundeyin M, Daley D, *et al.* The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. *Cancer Discov* 2018;**8**:403–16. doi:10.1158/2159-8290.CD-17-1134

10 Michaud DS, Izard J, Wilhelm-benartzi CS, *et al.* Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. *Gut* 2013;**62**:1764–70.

11 Farrell JJ, Zhang L, Zhou H, *et al.* Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut* 2012;**64**:582–8. doi:10.1136/gutjnl-2011-300784

12 Nagata N, Nishijima S, Kojima Y, *et al.* Metagenomic Identification of Microbial Signatures Predicting Pancreatic Cancer From a Multinational Study. *Gastroenterology* 2022;**163**:222–38. doi:10.1053/j.gastro.2022.03.054

13 Kartal E, Schmidt TSB, Molina-Montes E, *et al.* A faecal microbiota signature with high specificity for pancreatic cancer. *Gut* 2022;**71**:1359–72. doi:10.1136/gutjnl-2021-324755

14 Shoaie S, Ghaffari P, Kovatcheva-Datchary P, *et al.* Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. *Cell Metab* 2015;**22**:320–31. doi:10.1016/j.cmet.2015.07.001

15 Flint HJ. The impact of nutrition on the human microbiome. *Nutr Rev* 2012;**70 Suppl 1**:S10-3. doi:10.1111/j.1753-4887.2012.00499.x

16 Campaniello D, Corbo MR, Sinigaglia M, *et al.* How Diet and Physical Activity Modulate Gut Microbiota: Evidence, and Perspectives. *Nutrients* 2022;**14**:2456. doi:10.3390/nu14122456

17 Cullin N, Azevedo Antunes C, Straussman R, *et al.* Microbiome and cancer. *Cancer Cell* 2021;**39**:1317–41. doi:10.1016/j.ccell.2021.08.006

18 Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol* 2002;**13**:3–9. doi:10.1097/00041433-200202000-00002

19 Molina-Montes E, Wark PA, Sánchez M-J, *et al.* Dietary intake of iron, heme-iron and magnesium and pancreatic cancer risk in the European prospective investigation into

cancer and nutrition cohort. *Int J Cancer* 2012;**131**:E1134-47. doi:10.1002/ijc.27547

20 Navarrete-Muñoz EM, Wark PA, Romaguera D, *et al.* Sweet-beverage consumption and risk of pancreatic cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Am J Clin Nutr* 2016;**104**:760–8. doi:10.3945/ajcn.116.130963

21 Rohrmann S, Linseisen J, Nöthlings U, *et al.* Meat and fish consumption and risk of pancreatic cancer: Results from the European Prospective Investigation into Cancer and Nutrition. *Int J Cancer* 2013;**132**:617–24. doi:10.1002/ijc.27637

22 Research WCRF/ AI for C, editor. *Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. 2018*. World Cancer Research Fund / American Institute for Cancer Research

23 Molina-Montes E, Sanchez M-J, Buckland G, *et al.* Mediterranean diet and risk of pancreatic cancer in the European Prospective Investigation into Cancer and Nutrition cohort. *Br J Cancer* 2017;**116**:811–20. doi:10.1038/bjc.2017.14

24 Cayssials V, Buckland G, Crous-Bou M, *et al.* Inflammatory potential of diet and pancreatic cancer risk in the EPIC study. *Eur J Nutr* 2022;**61**:2313–20. doi:10.1007/s00394-022-02809-y

25 Mehta RS, Nishihara R, Cao Y, *et al.* Association of Dietary Patterns With Risk of Colorectal Cancer Subtypes Classified by Fusobacterium Nucleatum in Tumor Tissue. *JAMA Oncol* 2017;**02215**:1–7. doi:10.1001/jamaoncol.2016.6374

26 McCormick BA, Inadomi JM. The Microbiome Modifies the Effect of Diet on Colorectal Cancer Incidence. *Gastroenterology* Published Online First: 5 September 2022. doi:10.1053/j.gastro.2022.07.066

27 Hoang T, Kim MJ, Park JW, *et al.* Nutrition-wide association study of microbiome diversity and composition in colorectal cancer patients. *BMC Cancer* 2022;**22**:1–19. doi:10.1186/s12885-022-09735-6

28 Callahan BJ, McMurdie PJ, Rosen MJ, *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3. doi:10.1038/nmeth.3869

29 João F Matias Rodrigues1, Thomas SB Schmidt1 JT and C von M. MAPseq: highly efficient k-mer search with confi- dence estimates, for rRNA sequence analysis. *Bioinformatics* 2017;**33**:3808–10. doi:10.1093/bioinformatics/xxxxx

30 Matias Rodrigues JF, Von Mering C. HPC-CLUST: Distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* 2014;**30**:287–8. doi:10.1093/bioinformatics/btt657

31 Milanese A, Mende DR, Paoli L, *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 2019;**10**:1014. doi:10.1038/s41467-019-08844-4

32 Stekhoven DJ, B??hlmann P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;**28**:112–8. doi:10.1093/bioinformatics/btr597

33 MARTIN-MORENO JM, BOYLE P, GORGOJO L, *et al.* Development and Validation of a Food Frequency Questionnaire in Spain. *Int J Epidemiol* 1993;**22**:512–9. doi:10.1093/ije/22.3.512

34 García-Closas R, García-Closas M, Kogevinas M, *et al.* Food, nutrient and heterocyclic amine intake and the risk of bladder cancer. *Eur J Cancer* 2016;**43**:1731–40. doi:10.1016/j.ejca.2007.05.007

35 (SENC =Grupo Colaborativo de la Sociedad Española de Nutrición Comunitaria, Aranceta Bartrina J, Arija Val V, *et al.* Dietary guidelines for the Spanish population (SENC, December 2016); the new graphic icon of healthy nutrition. Nutr. Hosp. 2016;**33**:1–48. doi:10.20960/nh.827

36 AESAN. BEDCA. Spanish Food Composition Database. 2007.https://www.bedca.net/bdpub/

37 Lozupone CA, Knight R. Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev* 2008;**32**:557–78. doi:10.1111/j.1574-6976.2008.00111.x

38 Chao A, Chiu C-H, Jost L. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annu Rev Ecol Evol Syst* 2014;**45**:297–324. doi:10.1146/annurev-ecolsys-120213-091540

39 Schmidt TSB, Matias Rodrigues JF, von Mering C. A family of interaction-adjusted indices of community similarity. *ISME J* 2017;**11**:791–807. doi:10.1038/ismej.2016.139

40 Anderson MJ. A new method for non parametric multivariate analysis of variance. *Austral Ecol* 2001;**26**:32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

41 Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;**26**:139–40. doi:10.1093/bioinformatics/btp616

42 Benjamini Y, Hochberg Y, Benjamini, Yoav HY. Benjamini and Y FDR.pdf. J. R. Stat. Soc. Ser. B. 1995;**57**:289–300. doi:10.2307/2346101

43 Wirbel J, Zych K, Essex M, *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol* 2021;**22**:1–27. doi:10.1186/s13059-021-02306-1

44 Kang JH, Peng C, Rhee JJ, *et al.* Prospective study of a diabetes risk reduction diet and the risk of breast cancer. *Am J Clin Nutr* 2020;**112**:1492–503. doi:10.1093/ajcn/nqaa268

45 Buckland G, Agudo A, Travier N, *et al.* Adherence to the Mediterranean diet reduces mortality in the Spanish cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Spain). *Br J Nutr* 2011;**106**:1581–91. doi:10.1017/S0007114511002078

46 Willett WC, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. *Am J Epidemiol* 1986;**124**:17–27. doi:10.1093/oxfordjournals.aje.a114366

47 Wang C, Segal LN, Hu J, *et al.* Microbial Risk Score for Capturing Microbial Characteristics, Integrating Multi-omics Data, and Predicting Disease Risk. *Microbiome* 2022;**10**:1–15. doi:10.1186/s40168-022-01310-2

48 Statistical R Core Team. R: A language and environment for statistical computing. 2020.https://www.r-project.org/

49 McMurdie PJ, Holmes S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 2013;**8**:e61217. doi:10.1371/journal.pone.0061217

50 Lahti, Leo; Sudarshan S. Tools for microbiome analysis in R. 2017.http://microbiome.github.com/microbiome

51 Jari, Oksanen; F Guillaume BRK. Vegan: community ecology package. 2019.https://github.com/vegandevs/vegan

52 Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970;**12**:55–67. doi:10.1080/00401706.1970.10488634

53 Fu Y, Wang Y, Gao H, *et al.* Associations among Dietary Omega-3 Polyunsaturated Fatty Acids, the Gut Microbiota, and Intestinal Immunity. *Mediators Inflamm* 2021;**2021**:8879227. doi:10.1155/2021/8879227

54 Bountziouka V, Bathrellou E, Constantinidis TC, *et al.* No Title. *J Appl Biobehav Res* 2010;**15**:31.

55 Pérez-Rodrigo MartaAU - Gil, ÁngelAU - González-Gross, MarcelaAU - Ortega, Rosa M.AU - Serra-Majem, LluisAU - Varela-Moreiras, GregorioAU - Aranceta-Bartrina, JavierTI - Lifestyle Patterns and Weight Status in Spanish Adults: The ANIBES Study C-G-C. Lifestyle Patterns and Weight Status in Spanish Adults: The ANIBES Study. Nutrients. 2017;**9**:606. doi:10.3390/nu9060606

Supplementary material

Supplementary Table 1: Filtering process applied to the 16s rRNA sequenced samples. 21

Supplementary Table 2: Samples used for data analyses by aims, site and disease status..............22

Supplementary Table 3: Parameter tuning of feature selection methods for food groups relative to the MRS and rS as an example. 23

Supplementary Table 4: Characteristics of the study population from the MICROBIOME and PanGenEU studies with valid microbiome and dietary data. More details provided in Reference 13 24

Supplementary Table 5: PERMANOVA analyses to test differences in community composition by case-control status (16S rRNA data) 28

Supplementary Table 6: Foods associated with PC cancer risk within the PanGenEU study (Spanish data sample). 29

Supplementary Table 7: Components and quantiles of the Diet Scores (DRRDS and rMED score) 31

Supplementary Table 8: Associations between foods on an individual basis with the microbial risk score MRS...33

Supplementary Figure 1: Distribution of the microbial risk score by disease status. 35

Supplementary Figure 2: PCoA plots of saliva (left) and stool (right) samples for PDAC cases, controls and chronic pancreatitis patients. Jaccard index. 16s rRNA data. 36

Supplementary Figure 3: Correlation between stool and saliva taxa at genus level among PC cases and controls and other subgroups (by diabetes, smoking and obesity status) 37

Supplementary Figure 4: Spearman correlation plots between dietary variables, foods and nutrients. 38

Supplementary Figure 6: Feature importance of the most prevalent taxa (N=50) for DRRD score selected by LASSO. 40

Supplementary Figure 7: Dietary clusters derived from intake for food items within the PanGenEU study and among 511 controls. Manhattan distances. 41

Supplementary Figure 8: Clusters of nutrients and the stool microbial signature taxa (27 species) overall for PC cases and controls. Euclidean (A) and Manhattan (B) distance. 42

Supplementary Figure 9: Clusters of foods groups and the stool microbial signature taxa (27 species) overall for PC cases and controls. Euclidean distance. 43

Supplementary Figure 10: Clusters of nutrients and the stool microbial signature taxa (top 50 most prevalent taxa) overall for PC cases and controls. Manhattan distance. 44

Supplementary Table 1: Filtering process applied to the 16s rRNA sequenced samples.

ASV table	**OTU Open table**
779 samples **20,272 ASVs**	**779 samples** **2,081 OTUs**
Remove samples with ≤ 500 reads (rowsum across samples)	Remove samples with ≤ 500 reads (rowsum across samples)
615 samples **20,272 ASVs**	**614 samples** **2,081 OTUs**
Remove taxa if not present in at least 5 samples	Remove taxa if not present in at least 5 samples
615 samples **3,393 ASVs**	**614 samples** **852 OTUs**
Keep samples which have ≥400 reads *across the retained OTUs*	Keep samples which have ≥400 reads *across the retained OTUs*
605 samples **3,393 ASVs**	**613 samples** **852 OTUs**
We retain 78.85% of samples, 16.44% of ASVs and 84.52% of total reads. The removed samples contained only 0.37% of total reads	We retain 78.69% of samples, 40.94% of OTUs and 98.35% of total reads. The removed samples contained only 0.37% of total reads
Remove duplicated samples (N=18 samples) after pooling (16 salivas)	Remove duplicated samples (N=19 samples) after pooling (17 salivas)
587 samples **3,393 ASVs**	**594 samples** **852 OTUs**
Remove non-eligible cases (N=14 samples)	Remove non-eligible cases (N=14 samples)
573 samples **3,393 ASVs**	**580 samples** **852 OTUs**

Supplementary Table 2: Samples used for data analyses by aims, site and disease status.

	16s rRNA	**Shotgun metagenomics**
Aim 1	Stool: 51 PC cases, 46 controls, 23 C Oral: 59 PC cases, 55 controls, 28 CP Oral: plus 191 PC cases and 132 controls	Stool: 57 PC cases, 50 controls, 29 CP Oral: 43 PC cases, 45 controls, 12 CP
Aim 2	Stool: 51 PC cases and 46 controls Oral: 59 PC cases and 55 controls Oral: plus 191 PC cases and 132 controls	Stool: 57 PC cases and 50 controls Oral: 43 PC cases and 45 controls Diet & Stool: 51 PC cases and 48 controls
Aim 3	Stool: 51 PC cases and 46 controls Oral: 59 PC cases and 55 controls Oral: plus 191 PC cases and 132 controls	Stool: 57 PC cases and 50 controls Oral: 43 PC cases and 45 controls

MICROBIOME study (Reference 13) with 58 PC cases and 57 controls that were sequenced. Numbers that remained for analyses are shown in the table

PanGenEU study with diet and lifestyle information: 556 PC cases and 511 controls. 498 PC cases and 454 controls were sequenced (16s rRNA data). 191 PC cases and 132 controls remained for analyses.

Supplementary Table 3: Parameter tuning of feature selection methods for food groups relative to the MRS and rS as an example.

MRS						
RMSE	Min.	1st	Median	Mean	3rd	Max.
LinearModel	0.08448413	0.13196141	0.15980575	0.17913991	0.19845087	0.6975618
Ridge	0.02670246	0.04968456	0.06075801	**0.07609767**	0.08934668	0.1972448
LASSO	0.03686794	0.05476516	0.0646346	0.08010137	0.09036256	0.2004859
ENET	0.03686794	0.05476516	0.0646346	0.08010137	0.09036256	0.2004859
Rsquared	Min.	1st	Median	Mean	3rd	Max.
LinearModel	2.11E-04	0.03070166	0.08116297	0.1535543	0.2459961	0.7955558
Ridge	5.03E-05	0.02561784	0.08759615	**0.1854319**	0.3004101	0.8633095
LASSO	1.11E-06	0.02213094	0.08214631	0.1691972	0.2512876	0.806744
ENET	1.11E-06	0.02213094	0.08214631	0.1691972	0.2512876	0.806744
rS						
RMSE	Min.	1st	Median	Mean	3rd	Max.
LinearModel	0.8307698	2.002407	2.300511	2.775733	3.14038	13.770613
Ridge	0.6203829	1.087438	1.248703	**1.843995**	1.439399	31.14373
LASSO	0.8161898	1.364145	1.663111	1.913209	2.243991	7.152979
ENET	0.8161898	1.364145	1.663111	1.913209	2.243991	7.152979
Rsquared	Min.	1st	Median	Mean	3rd	Max.
LinearModel	0.00056928	0.02808508	0.1612766	0.214404	0.3410273	0.8795427
Ridge	0.0003527	0.08172174	0.2502566	**0.3004052**	0.4804517	0.9161802
LASSO	0.00070851	0.03533278	0.1826459	0.256767	0.4588144	0.812833
ENET	0.00070851	0.03533278	0.1826459	0.256767	0.4588144	0.812833

Supplementary Table 4: Characteristics of the study population from the MICROBIOME and PanGenEU studies with valid microbiome and dietary data. More details provided in Reference 13

		Cases N=58		Controls N=57		p-value
center:						0.943
	2	21	36.20%	22	38.60%	
	8	37	63.80%	35	61.40%	
sex:						0.648
	0	22	37.90%	25	43.90%	
	1	36	62.10%	32	56.10%	
agec		71.9	10.5	72.2	12.2	0.896
smokingstatus:						0.825
	0	29	50.00%	28	49.10%	
	1	1	1.72%	3	5.26%	
	2	19	32.80%	19	33.30%	
	3	9	15.50%	7	12.30%	
cpy1		18.8	30.7	26.6	42.9	0.266
alcohol_status:						0.239
	0	20	34.50%	13	22.80%	
	1	38	65.50%	44	77.20%	
alldiab:						0.022
	0	41	70.70%	51	89.50%	
	1	17	29.30%	6	10.50%	
diabcat:						0.008
	0	41	70.70%	51	89.50%	
	1	7	12.10%	0	0.00%	
	2	10	17.20%	6	10.50%	
obese:						1
	0	42	72.40%	42	73.70%	
	1	16	27.60%	15	26.30%	
asthma:						1
	0	54	93.10%	53	93.00%	
	1	4	6.90%	4	7.02%	
nasal:						1
	0	49	84.50%	49	86.00%	
	1	9	15.50%	8	14.00%	
allacid:						0.421
	0	41	70.70%	45	78.90%	
	1	17	29.30%	12	21.10%	
allhburn:						0.536
	0	43	74.10%	46	80.70%	
	1	15	25.90%	11	19.30%	
allrheum:						0.166
	0	53	91.40%	46	80.70%	
	1	5	8.62%	11	19.30%	
allhbp:						0.307
	0	24	41.40%	30	52.60%	
	1	34	58.60%	27	47.40%	
cholesterol:						0.102
	0	30	51.70%	39	68.40%	

	1	28	48.30%	18	31.60%	
abmedication:						0.186
	0	35	60.30%	42	73.70%	
	1	23	39.70%	15	26.30%	
antibiotic:						0.404
	0	22	37.90%	27	47.40%	
	1	36	62.10%	30	52.60%	
asparmed:						0.912
	0	33	56.90%	34	59.60%	
	1	25	43.10%	23	40.40%	
salicylic.ever:						0.573
	0	51	87.90%	47	82.50%	
	1	7	12.10%	10	17.50%	
paracetamol.ever:						0.863
	0	44	75.90%	45	78.90%	
	1	14	24.10%	12	21.10%	
cholmedication:						0.09
	0	33	56.90%	42	73.70%	
	1	25	43.10%	15	26.30%	
cortmed:						0.743
	0	52	89.70%	53	93.00%	
	1	6	10.30%	4	7.02%	
diabdiet:						0.018
	0	41	70.70%	51	89.50%	
	1	13	22.40%	6	10.50%	
	2	4	6.90%	0	0.00%	
diabin:						0.032
	0	41	70.70%	51	89.50%	
	1	9	15.50%	2	3.51%	
	2	8	13.80%	4	7.02%	
diabmed:						0.04
	0	41	70.70%	51	89.50%	
	1	12	20.70%	4	7.02%	
	2	5	8.62%	2	3.51%	
metformin.ever:						0.043
	0	46	79.30%	53	93.00%	
	1	11	19.00%	3	5.26%	
	2	1	1.72%	1	1.75%	
probiot:						0.438
	0	56	96.60%	53	93.00%	
	1	2	3.45%	4	7.02%	
periodontitis:						0.433
	0	40	69.00%	44	77.20%	
	1	18	31.00%	13	22.80%	
recession:						0.131
	0	34	58.60%	42	73.70%	
	1	24	41.40%	15	26.30%	
FHPDAC:						0.717
	0	53	91.40%	54	94.70%	
	1	5	8.62%	3	5.26%	

		Cases		Controls		
		N=498		N=454		p-value
center:						<0.001
	0	44	8.84%	46	10.10%	
	1	29	5.82%	24	5.29%	
	2	57	11.40%	41	9.03%	
	3	50	10.00%	5	1.10%	
	6	15	3.01%	13	2.86%	
	7	54	10.80%	71	15.60%	
	8	108	21.70%	118	26.00%	
	9	141	28.30%	136	30.00%	
sex:						0.803
	0	222	44.60%	207	45.60%	
	1	276	55.40%	247	54.40%	
agec		64	12.3	64.2	12.8	0.834
smokingstatus:						0.001
	0	194	39.00%	209	46.00%	
	1	3	0.60%	12	2.64%	
	2	158	31.70%	140	30.80%	
	3	143	28.70%	93	20.50%	
alcohol_status:						0.529
	0	154	30.90%	150	33.00%	
	1	344	69.10%	304	67.00%	
alldiab:						<0.001
	0	355	71.30%	389	85.70%	
	1	143	28.70%	65	14.30%	
diabcat:						<0.001
	0	355	71.30%	389	85.70%	
	1	64	12.90%	16	3.52%	
	2	79	15.90%	49	10.80%	
obese:						0.929
	0	396	79.50%	359	79.10%	
	1	102	20.50%	95	20.90%	
panctype:						0.041
	0	477	95.80%	447	98.50%	
	1	18	3.61%	6	1.32%	
	2	3	0.60%	1	0.22%	
asthma:						0.044
	0	461	92.60%	402	88.50%	
	1	37	7.43%	52	11.50%	
nasal:						0.004
	0	428	85.90%	357	78.60%	
	1	70	14.10%	97	21.40%	
allacid:						0.013
	0	351	70.50%	353	77.80%	
	1	147	29.50%	101	22.20%	
allhburn:						0.003
	0	314	63.10%	328	72.20%	
	1	184	36.90%	126	27.80%	
allrheum:						0.367

	0	469	94.20%	420	92.50%	
	1	29	5.82%	34	7.49%	
allhbp:						0.256
	0	314	63.10%	269	59.30%	
	1	184	36.90%	185	40.70%	
cholesterol:						0.977
	0	321	64.50%	294	64.80%	
	1	177	35.50%	160	35.20%	
abmedication:						0.852
	0	377	75.70%	347	76.40%	
	1	121	24.30%	107	23.60%	
asparmed:						0.43
	0	387	77.70%	342	75.30%	
	1	111	22.30%	112	24.70%	
salicylic.ever:						0.722
	0	436	87.60%	393	86.60%	
	1	62	12.40%	61	13.40%	
paracetamol.ever:						0.054
	0	448	90.00%	389	85.70%	
	1	50	10.00%	65	14.30%	
cholmedication:						0.706
	0	376	75.50%	337	74.20%	
	1	122	24.50%	117	25.80%	
cortmed:						0.011
	0	482	96.80%	422	93.00%	
	1	16	3.21%	32	7.05%	
nsaidmed:						0.559
	0	428	85.90%	397	87.40%	
	1	70	14.10%	57	12.60%	
diabdiet:						<0.001
	0	355	71.30%	389	85.70%	
	1	104	20.90%	43	9.47%	
	2	39	7.83%	22	4.85%	
diabin:						<0.001
	0	355	71.30%	389	85.70%	
	1	72	14.50%	18	3.96%	
	2	71	14.30%	47	10.40%	
diabmed:						<0.001
	0	355	71.30%	389	85.70%	
	1	98	19.70%	50	11.00%	
	2	45	9.04%	15	3.30%	
periodontitis:						1
	0	420	84.30%	382	84.10%	
	1	78	15.70%	72	15.90%	
recession:						0.03
	0	339	68.10%	339	74.70%	
	1	159	31.90%	115	25.30%	
FHPDAC:						<0.001
	0	461	92.60%	444	97.80%	
	1	37	7.43%	10	2.20%	

Supplementary Table 5: PERMANOVA analyses to test differences in community composition by case-control status (16S rRNA data)

	Model 1		**Model 2**	
	R-squared	**p-values**	**R-squared**	**p-values**
BRAY_CURTIS				
Saliva samples				
global	0.0185	**0.0165**	0.0106	0.5368
cases (PC) vs controls (C)	0.0072	0.2763	0.0068	0.2531
cases (PC) vs pancreatitis (CP)	0.0193	**0.0282**	0.0055	0.9903
controls (C) vs pancreatitis (CP)	0.0246	**0.0038**	0.0117	0.2827
Stool samples				
global	0.0400	**0.0002**	0.0363	**0.0007**
cases (PC) vs controls (C)	0.0335	**0.0001**	0.0297	**0.0002**
cases (PC) vs pancreatitis (CP)	0.0244	0.1412	0.0245	0.1274
controls (C) vs pancreatitis (CP)	0.0206	0.3879	0.0181	0.5917

Model 1: crude
Model 2: adjusted for age (continuous), sex (female, male), center, smoking (never, former, current), metformin (no diabetes, yes metformin, no metformin) in stool samples

Supplementary Table 6: Foods associated with PC cancer risk within the PanGenEU study (Spanish data sample).

	OR	LCI	HCI	P.value	p.bh
fat milk	0.96	0.84	1.09	0.500	0.825
low fat milk	1.07	0.93	1.22	0.330	0.773
fat yogurt	0.95	0.83	1.08	0.430	0.825
low fat yogurt	0.99	0.87	1.13	0.900	0.942
fresh cheese	1	0.87	1.14	0.970	0.970
curated cheese	0.97	0.85	1.12	0.700	0.853
pudding	1.05	0.91	1.2	0.510	0.825
ice cream	0.93	0.81	1.06	0.280	0.773
eggs	0.93	0.81	1.07	0.320	0.773
chicken	0.88	0.77	1.01	0.060	0.445
beef	0.99	0.87	1.12	0.830	0.935
pork	0.87	0.75	1.02	0.080	0.445
lam	0.96	0.83	1.11	0.570	0.825
rabbit	0.91	0.79	1.05	0.190	0.650
liver	0.97	0.85	1.1	0.610	0.825
offals	1.01	0.89	1.15	0.860	0.942
bacon	0.99	0.86	1.14	0.900	0.942
white fish	0.98	0.86	1.11	0.710	0.854
blue fish	0.98	0.86	1.12	0.780	0.926
canned fish	0.81	0.71	0.94	**<0.001**	**<0.001**
salted fish	0.91	0.8	1.04	0.150	0.580
calamars	0.96	0.84	1.1	0.580	0.825
sausages	0.82	0.71	0.96	**0.010**	0.297
hamburguers	0.96	0.84	1.1	0.570	0.825
hot dogs	0.99	0.87	1.14	0.940	0.962
york ham	0.99	0.87	1.13	0.900	0.942
ham	1.07	0.93	1.22	0.360	0.801
salami	0.98	0.86	1.13	0.810	0.935
croquettes	0.96	0.84	1.1	0.560	0.825
fish fingers	0.97	0.85	1.1	0.600	0.825
ready dishes	0.97	0.85	1.1	0.610	0.825
spinach	0.85	0.74	0.98	**0.020**	0.445
coliflour	0.87	0.76	1	**0.050**	0.445
lettuce	0.94	0.82	1.07	0.330	0.773
tomato	0.97	0.84	1.11	0.630	0.825
onion	1.05	0.92	1.2	0.510	0.825
carrots	0.95	0.83	1.08	0.420	0.825
beans	0.95	0.83	1.08	0.410	0.825
peas	0.95	0.83	1.09	0.450	0.825
eggplant	0.97	0.85	1.11	0.620	0.825
peppers	1.07	0.94	1.23	0.290	0.773
artichokes	0.91	0.79	1.05	0.210	0.692
asparagus	0.86	0.75	0.98	**0.030**	0.445
maize	0.95	0.83	1.09	0.470	0.825
legumes	0.97	0.85	1.12	0.680	0.841
oranges	0.9	0.78	1.03	0.120	0.562
banana	0.87	0.76	1	**0.050**	0.445
apple	1.08	0.94	1.24	0.280	0.773
pear	1.05	0.91	1.21	0.540	0.825
peach	1.12	0.96	1.29	0.140	0.580
melon	1.01	0.88	1.16	0.840	0.935
grapes	1.15	0.99	1.34	0.070	0.445
prunes	1.07	0.94	1.23	0.310	0.773
kiwi	1.05	0.92	1.21	0.460	0.825
fruits almibar	0.99	0.87	1.13	0.940	0.962

olives	1.06	0.93	1.22	0.400	0.825
nuts	1.16	0.99	1.35	0.070	0.445
white bread	1.03	0.9	1.19	0.650	0.838
whole bread	1	0.87	1.14	0.970	0.970
breakfast cereals	1.07	0.94	1.23	0.300	0.773
french fries	0.97	0.84	1.11	0.630	0.825
boiled potatoes	1.11	0.96	1.28	0.150	0.580
rice	1.08	0.94	1.23	0.290	0.773
pasta	0.92	0.8	1.07	0.280	0.773
pizza	0.85	0.71	1.02	0.080	0.445
olive oil	1.04	0.91	1.2	0.540	0.825
other fats	1.01	0.89	1.15	0.830	0.935
cakes	1.11	0.97	1.28	0.120	0.562
croissants	0.94	0.82	1.07	0.340	0.776
chocolate	1.12	0.98	1.29	0.110	0.562
mermelade	1.16	1.01	1.33	**0.040**	0.445
sugar	1.03	0.9	1.18	0.680	0.841
mayonnaise	1.07	0.87	1.33	0.510	0.825
tomato sauce	1.11	0.95	1.3	0.190	0.650
ketchup	0.96	0.83	1.1	0.540	0.825
garlic	1.05	0.92	1.21	0.470	0.825
chips	0.91	0.8	1.04	0.170	0.630
sweet beverages	0.96	0.83	1.1	0.550	0.825
artificial bev	0.99	0.87	1.12	0.830	0.935
fresh orange juice	0.95	0.83	1.09	0.460	0.825
nectar	0.88	0.76	1	0.060	0.445
cereal beverage	0.97	0.85	1.11	0.670	0.841
coffee	1.26	1.09	1.45	**<0.001**	**<0.001**
decoffee	0.97	0.64	1.47	0.880	0.942
tea	1.1	0.97	1.25	0.140	0.580
beer	1.16	0.99	1.37	0.070	0.445
wine	1.04	0.9	1.19	0.600	0.825
sprits	1.16	0.99	1.36	0.070	0.445
fortified	0.94	0.83	1.06	0.310	0.773

Multivariate adjusted logistic regression models adjusted for age in years (continuous), sex (men, women), center (all five Spanish hospitals), diabetes status (no diabetes, diabetes: diagnosed less than 2 years, or since more than 2 years), pack-years of smoking (non-smokers, tertiles of pack-years), obese (no, yes: BMI>30 kg/m^2), and family history of PC (no, yes), as well as energy intake in Kcal. OR and 95% CI are derived from these models, and are related to PC risk per 1 SD increase in the intake of the dietary variable. P-values were corrected for multiple comparison testing by the Benjamini-Hochberg BH method (p.bh). For foods groups, the main group is indicated with "g" along with the corresponding subgroups "sg" before the food group´s name.

Supplementary Table 7: Components and quantiles of the Diet Scores (DRRDS and rMED score)

	Controls	N=511	Cases	N=560	p.overall
DRRDSCORE	26.7	4.67	27.5	4.62	0.008
SC_fiber	2.24	1.85	2.34	1.89	0.426
SC_nuts	3.14	1.74	3.36	1.76	0.04
SC_fruits	2.99	1.42	3.16	1.4	0.051
SC_coffee	2.51	1.62	2.7	1.67	0.05
SC_ratio	3	1.42	3.06	1.42	0.444
SC_gi	3	1.42	2.91	1.41	0.291
SC_trans	3	1.42	3	1.39	0.931
SC_meats	3	1.42	3.04	1.43	0.685
SC_juices	3.85	1.53	3.93	1.52	0.419
meats:					0.196
Q1	103	20.20%	132	23.60%	
Q2	102	20.00%	90	16.10%	
Q3	102	20.00%	106	18.90%	
Q4	102	20.00%	132	23.60%	
Q5	102	20.00%	100	17.90%	
fruits:					0.18
Q1	104	20.40%	87	15.50%	
Q2	101	19.80%	110	19.60%	
Q3	102	20.00%	123	22.00%	
Q4	102	20.00%	105	18.80%	
Q5	102	20.00%	135	24.10%	
coffee:					0.135
Non-consumers	245	47.90%	242	43.20%	
Q1	147	28.80%	159	28.40%	
Q2	119	23.30%	159	28.40%	
fiber:					0.465
Q1+Q2	352	68.90%	373	66.60%	
Q3	159	31.10%	187	33.40%	
gi:					0.787
Q1	103	20.20%	98	17.50%	
Q2	102	20.00%	116	20.70%	
Q3	102	20.00%	109	19.50%	
Q4	102	20.00%	113	20.20%	
Q5	102	20.00%	124	22.10%	
juices:					0.454
Non-consumers	304	59.50%	352	62.90%	
Q1	121	23.70%	116	20.70%	
Q2	86	16.80%	92	16.40%	
nuts:					0.038
Non-consumers	177	34.60%	175	31.20%	
Q1	121	23.70%	109	19.50%	
Q2	213	41.70%	276	49.30%	
ratio:					0.75
Q1	103	20.20%	101	18.00%	

Q2	102	20.00%	121	21.60%	
Q3	102	20.00%	101	18.00%	
Q4	102	20.00%	116	20.70%	
Q5	102	20.00%	121	21.60%	
trans:					0.778
Q1	103	20.20%	102	18.20%	
Q2	102	20.00%	127	22.70%	
Q3	102	20.00%	106	18.90%	
Q4	102	20.00%	117	20.90%	
Q5	102	20.00%	108	19.30%	
rMEDSCORE	7.86	2.13	8.04	2.21	0.18
medfru	274	164	289	181	0.14
medfruits	1	0.82	1.06	0.81	0.283
medvege	1.01	0.83	1.09	0.83	0.111
medleg	0.42	0.57	0.43	0.54	0.867
medcer	0.24	0.55	0.24	0.54	0.897
medfish	1.27	0.8	1.27	0.78	0.887
medoil	1.39	0.53	1.48	0.53	0.011
medmeat	1.34	0.8	1.31	0.8	0.493
meddai	1.18	0.77	1.16	0.76	0.621
fruits:					0.478
Q1	170	33.30%	167	29.80%	
Q2	170	33.30%	195	34.80%	
Q3	171	33.50%	198	35.40%	
vegetables:					0.28
Q1	173	33.90%	168	30.00%	
Q2	158	30.90%	171	30.50%	
Q3	180	35.20%	221	39.50%	
legumes:					0.259
Q1	314	61.40%	333	59.50%	
Q2	198	38.60%	227	40.50%	
cereals:					0.741
Q1	420	82.20%	455	81.20%	
Q2	91	17.80%	105	18.80%	
fish:					0.544
Q1	170	33.30%	116	20.70%	
Q2	170	33.30%	174	31.10%	
Q3	171	33.50%	270	48.20%	
oil:					0.038
Non-consumers	12	2.35%	9	1.61%	
Q1	286	56.00%	275	49.10%	
Q2	213	41.70%	276	49.30%	
meat:					0.757
Q1	279	54.60%	293	52.30%	
Q2	128	25.00%	147	26.20%	
Q3	104	20.40%	120	21.40%	

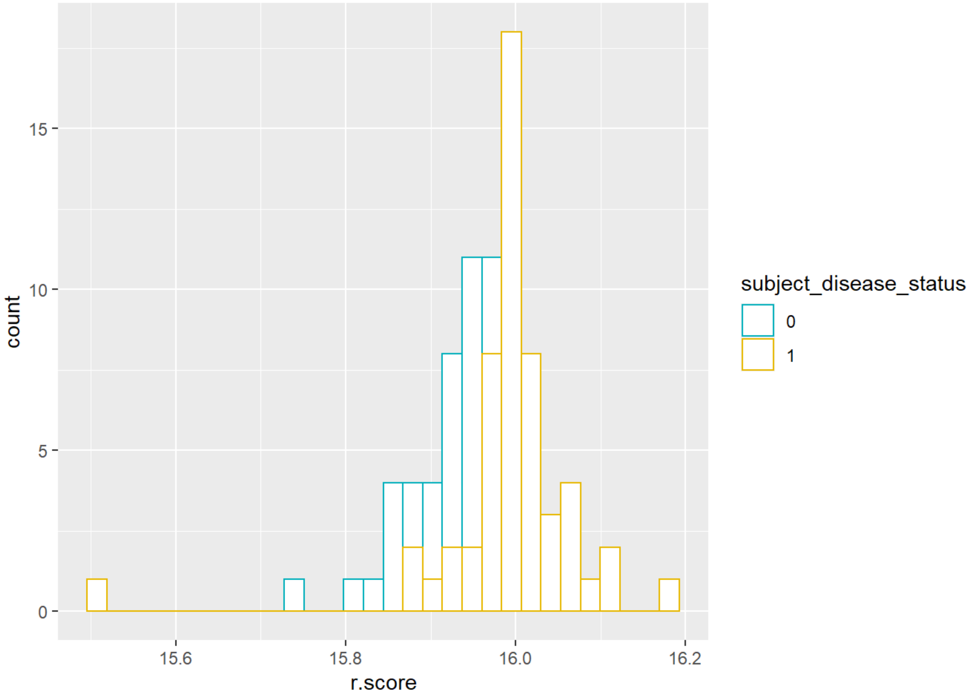
Supplementary Table 8: Associations between foods on an individual basis with the microbial risk score MRS.

	coeff	SE	p.value	CILow	CIHigh	R.squared	p.bh
fat milk	-0.0004	5.00E-04	0.471	-0.001	0.001	0.183	0.932
low fat milk	0.0005	5.00E-04	0.380	-0.001	0.002	0.186	0.932
fat yogurt	-0.0001	6.00E-04	0.922	-0.001	0.001	0.178	0.985
low fat yogurt	0.0002	5.00E-04	0.699	-0.001	0.001	0.180	0.972
fresh cheese	0.0002	5.00E-04	0.679	-0.001	0.001	0.180	0.965
curated cheese	0.0002	6.00E-04	0.727	-0.001	0.001	0.179	0.979
pudding	-0.0004	5.00E-04	0.398	-0.001	0.001	0.185	0.932
ice cream	0.0002	5.00E-04	0.645	-0.001	0.001	0.180	0.965
eggs	-0.0003	5.00E-04	0.516	-0.001	0.001	0.182	0.953
chicken	0.0004	5.00E-04	0.484	-0.001	0.001	0.183	0.937
beef	0.0001	5.00E-04	0.772	-0.001	0.001	0.179	0.985
pork	0.0003	6.00E-04	0.582	-0.001	0.001	0.181	0.965
lam	0	5.00E-04	0.985	-0.001	0.001	0.178	0.994
rabbit	-0.0004	5.00E-04	0.466	-0.001	0.001	0.183	0.932
liver	-0.0006	5.00E-04	0.223	-0.002	0.000	0.193	0.932
offals	-0.0006	5.00E-04	0.197	-0.002	0.000	0.194	0.932
bacon	0.0003	5.00E-04	0.610	-0.001	0.001	0.181	0.965
white fish	-0.0006	5.00E-04	0.287	-0.002	0.001	0.189	0.932
blue fish	0.0001	5.00E-04	0.785	-0.001	0.001	0.179	0.985
canned fish	-0.0012	6.00E-04	**0.032**	-0.002	0.000	0.222	0.713
salted fish	-0.0008	5.00E-04	0.112	-0.002	0.000	0.203	0.932
calamars	-0.0011	5.00E-04	**0.025**	-0.002	0.000	0.226	0.713
sausages	0.0001	7.00E-04	0.911	-0.001	0.002	0.178	0.985
hamburgers	0.0014	6.00E-04	**0.015**	0.000	0.003	0.234	0.713
hot dogs	0.0001	6.00E-04	0.867	-0.001	0.001	0.178	0.985
york ham	-0.0003	5.00E-04	0.537	-0.001	0.001	0.182	0.955
ham	0	5.00E-04	0.948	-0.001	0.001	0.178	0.993
salami	0.0002	5.00E-04	0.631	-0.001	0.001	0.180	0.965
croquettes	0.0003	1.00E-03	0.790	-0.002	0.002	0.179	0.985
fish fingers	-0.0001	5.00E-04	0.916	-0.001	0.001	0.178	0.985
readydishes	0.0004	5.00E-04	0.503	-0.001	0.001	0.183	0.952
spinach	-0.0001	5.00E-04	0.864	-0.001	0.001	0.178	0.985
coliflour	-0.0001	5.00E-04	0.861	-0.001	0.001	0.178	0.985
lettuce	0.0002	5.00E-04	0.631	-0.001	0.001	0.180	0.965
tomato	0.0001	5.00E-04	0.841	-0.001	0.001	0.179	0.985
onion	-0.0009	5.00E-04	0.112	-0.002	0.000	0.203	0.932
carrots	-0.0005	5.00E-04	0.282	-0.002	0.000	0.189	0.932
beans	-0.0006	5.00E-04	0.287	-0.002	0.001	0.189	0.932
peas	-0.0004	5.00E-04	0.394	-0.001	0.001	0.185	0.932
eggplant	-0.0002	5.00E-04	0.658	-0.001	0.001	0.180	0.965
peppers	-0.0005	5.00E-04	0.384	-0.001	0.001	0.186	0.932
artichokes	0	5.00E-04	0.981	-0.001	0.001	0.178	0.994
asparagus	-0.0006	5.00E-04	0.265	-0.002	0.000	0.190	0.932
maize	-0.0002	5.00E-04	0.683	-0.001	0.001	0.180	0.965
legumes	-0.0002	6.00E-04	0.724	-0.001	0.001	0.179	0.979
oranges	0.0006	5.00E-04	0.287	0.000	0.002	0.189	0.932
banana	-0.0007	5.00E-04	0.232	-0.002	0.000	0.192	0.932
apple	0.0011	5.00E-04	0.051	0.000	0.002	0.223	0.713
pear	-0.0004	6.00E-04	0.471	-0.001	0.001	0.183	0.932
peach	0.0004	5.00E-04	0.383	-0.001	0.001	0.186	0.932
melon	0.0003	5.00E-04	0.638	-0.001	0.001	0.180	0.965
grapes	0.0003	5.00E-04	0.622	-0.001	0.001	0.181	0.965
prunes	0.0002	5.00E-04	0.737	-0.001	0.001	0.179	0.979
kiwi	0	5.00E-04	0.972	-0.001	0.001	0.178	0.994
fruits almibar	-0.0004	5.00E-04	0.386	-0.001	0.001	0.185	0.932

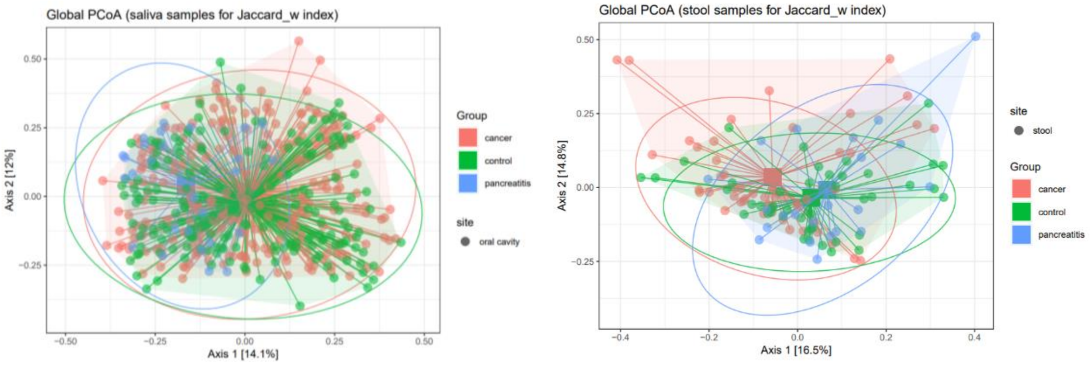
olives	-0.0006	5.00E-04	0.298	-0.002	0.001	0.189	0.932
nuts	-0.0002	5.00E-04	0.666	-0.001	0.001	0.180	0.965
white bread	0.0001	7.00E-04	0.929	-0.001	0.001	0.178	0.985
wholebread	0.0005	5.00E-04	0.324	0.000	0.002	0.188	0.932
breakfast cer	0.0007	5.00E-04	0.174	0.000	0.002	0.196	0.932
french fries	-0.0008	5.00E-04	0.139	-0.002	0.000	0.199	0.932
boiled potatoes	-0.0005	6.00E-04	0.414	-0.002	0.001	0.185	0.932
rice	0.0004	5.00E-04	0.419	-0.001	0.002	0.185	0.932
pasta	-0.0001	5.00E-04	0.850	-0.001	0.001	0.178	0.985
pizza	0.0006	6.00E-04	0.332	-0.001	0.002	0.187	0.932
olive oil	-0.0004	6.00E-04	0.567	-0.002	0.001	0.181	0.965
other fats	0.0007	5.00E-04	0.170	0.000	0.002	0.196	0.932
cakes	0.0001	6.00E-04	0.879	-0.001	0.001	0.178	0.985
croissants	-0.0003	5.00E-04	0.525	-0.001	0.001	0.182	0.953
chocolate	0.0006	6.00E-04	0.270	0.000	0.002	0.190	0.932
mermelade	0.0008	5.00E-04	0.093	0.000	0.002	0.205	0.932
sugar	0.0002	6.00E-04	0.677	-0.001	0.001	0.180	0.965
mayonnaise	0.0004	5.00E-04	0.380	-0.001	0.001	0.186	0.932
tomato sauce	0.0007	5.00E-04	0.133	0.000	0.002	0.200	0.932
ketchup	0.0004	5.00E-04	0.371	-0.001	0.001	0.186	0.932
garlic	0.0001	5.00E-04	0.893	-0.001	0.001	0.178	0.985
chips	0.0004	5.00E-04	0.411	-0.001	0.001	0.185	0.932
sweet bev	-0.0004	5.00E-04	0.448	-0.001	0.001	0.184	0.932
artificial bev	0.0007	5.00E-04	0.144	0.000	0.002	0.199	0.932
fresh juice	0.0006	6.00E-04	0.303	-0.001	0.002	0.188	0.932
nectar	-0.0001	6.00E-04	0.792	-0.001	0.001	0.179	0.985
cereal bev	0.0005	5.00E-04	0.294	0.000	0.002	0.189	0.932
coffee	0.0006	5.00E-04	0.302	-0.001	0.002	0.189	0.932
decoffee	0.0004	5.00E-04	0.447	-0.001	0.002	0.184	0.932
tea	-0.0001	5.00E-04	0.846	-0.001	0.001	0.178	0.985
beer	0.0006	5.00E-04	0.252	0.000	0.002	0.191	0.932
wine	0	5.00E-04	0.994	-0.001	0.001	0.178	0.994
sprits	0.0006	5.00E-04	0.208	0.000	0.002	0.194	0.932
fortified	0.0001	5.00E-04	0.857	-0.001	0.001	0.178	0.985

Multivariate adjusted linear regression models adjusted for age in years (continuous), sex (men, women), center (all five Spanish hospitals), diabetes status (no diabetes, diabetes: diagnosed less than 2 years, or since more than 2 years), pack-years of smoking (non-smokers, tertiles of pack-years), obese (no, yes: BMI>30 kg/m^2), and family history of PC (no, yes), as well as energy intake in Kcal. β coefficients and 95% CI are derived from these models, per 1 SD increase in the intake of the dietary variables. All microbial risk scores (outcome variable) were log-transformed to approximate a normal distribution. P-values were corrected for multiple comparison testing by the Benjamini-Hochberg BH method (p.bh).

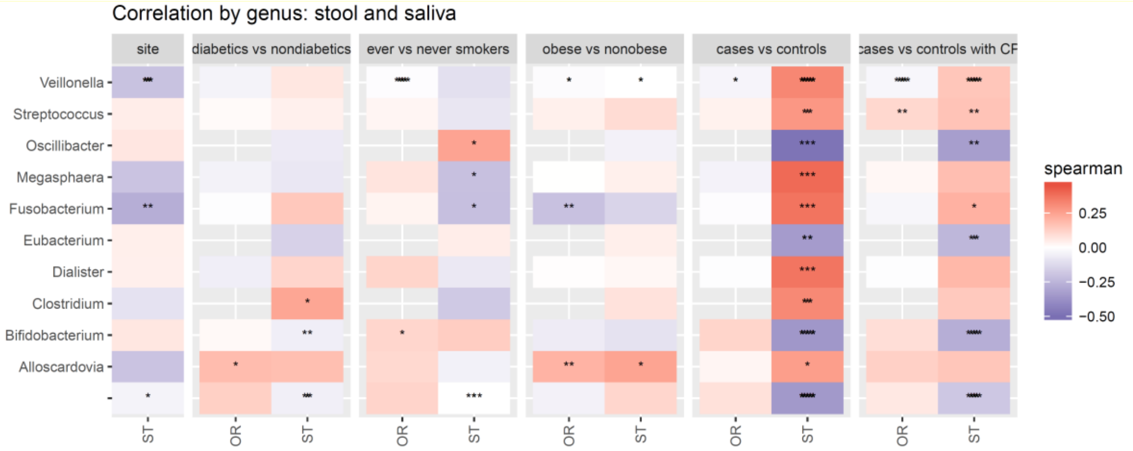
Supplementary Figure 1: Distribution of the microbial risk score by disease status.



Supplementary Figure 2: PCoA plots of saliva (left) and stool (right) samples for PDAC cases, controls and chronic pancreatitis patients. Jaccard index. 16s rRNA data.

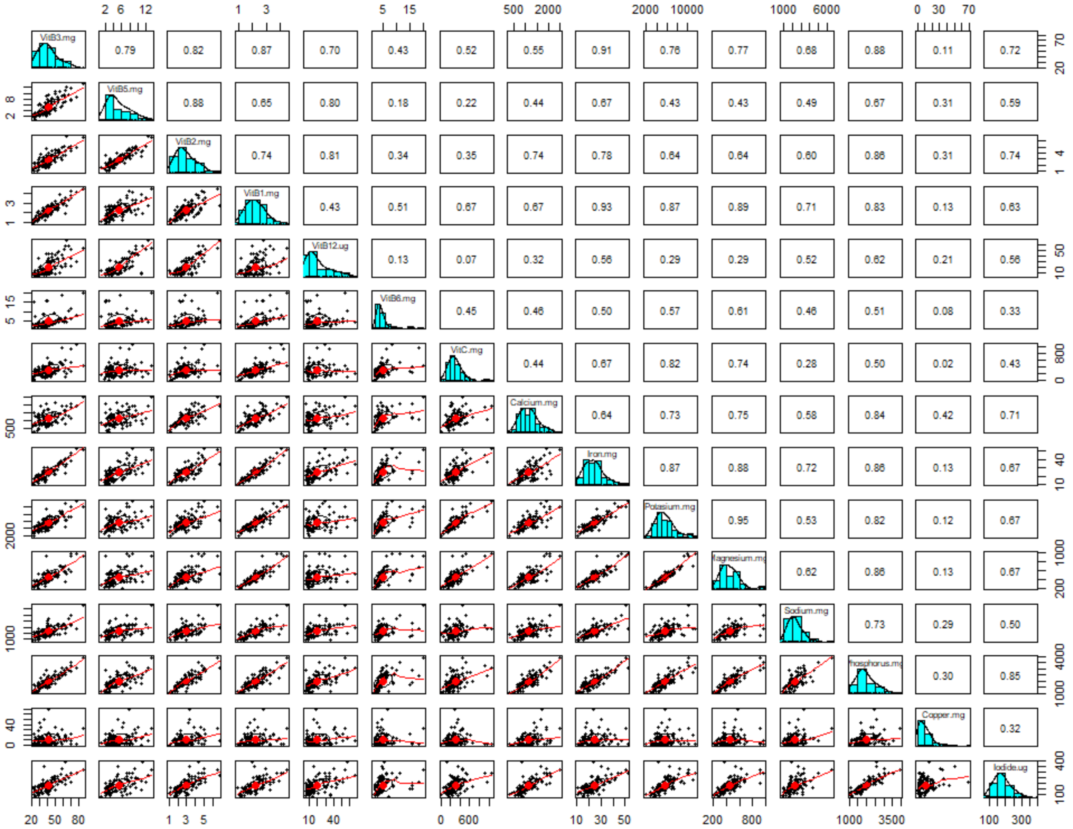


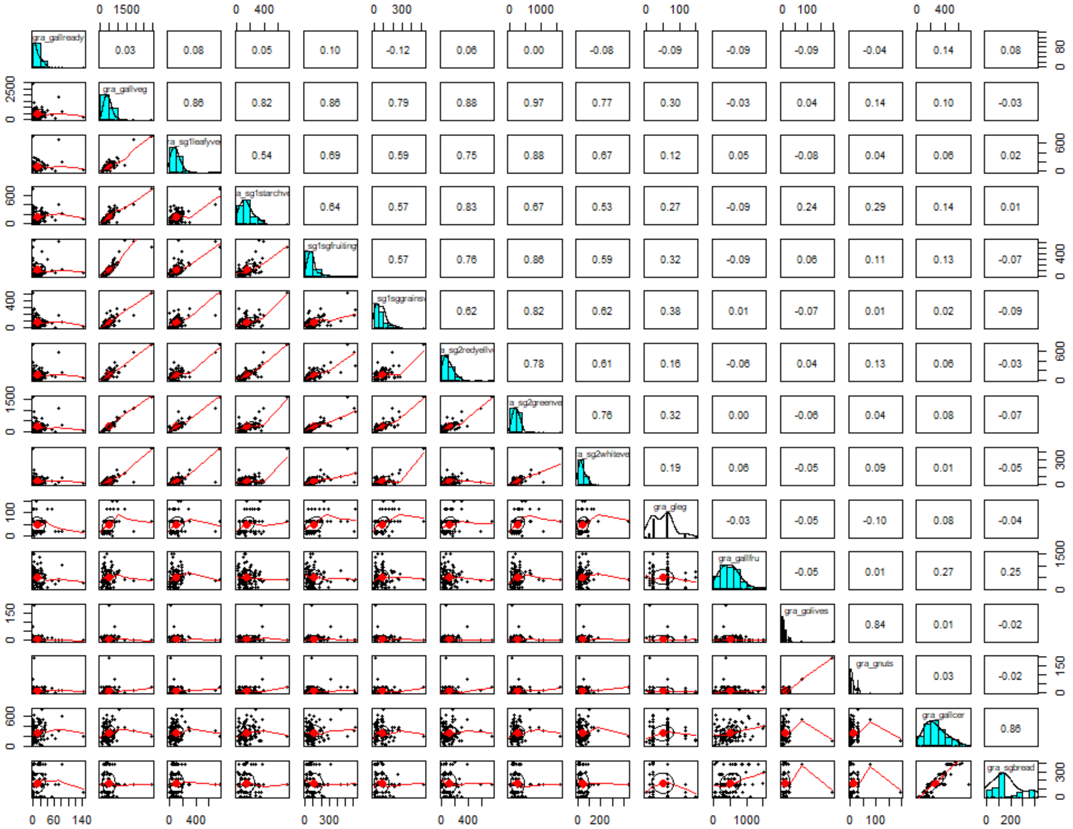
Supplementary Figure 3: Correlation between stool and saliva taxa at genus level among PC cases and controls and other subgroups (by diabetes, smoking and obesity status)



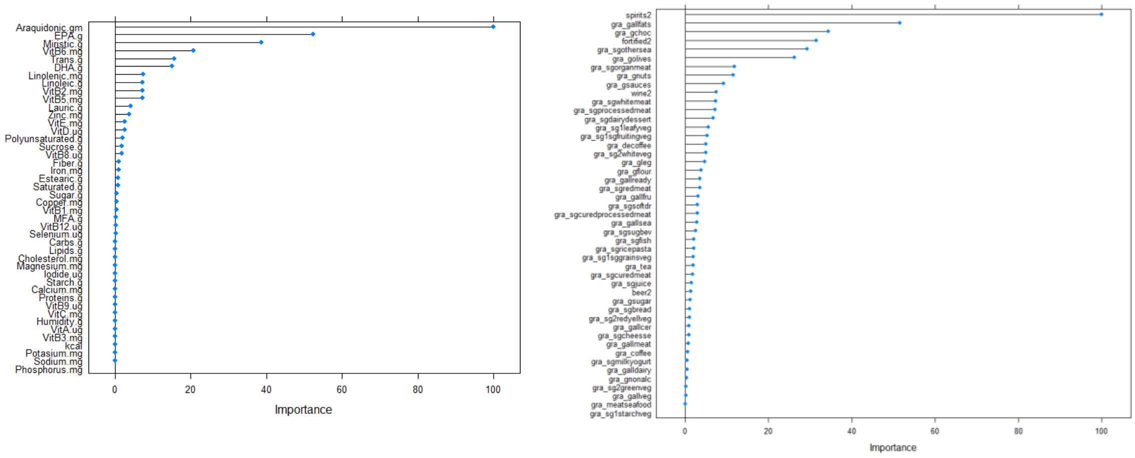
Plots derived from Spearman correlation analyses of bacterial taxa at genus level, whereby taxa present in both stool and oral samples were considered. Correlations between the taxa from both sites was analyzed by different conditions: diabetes status, smoking status, obesity, and disease status (with and without chronic pancreatitis). Significance levels: *p<0.05, **p<0.01, ***p<0.001

Supplementary Figure 4: Spearman correlation plots between dietary variables, foods and nutrients.



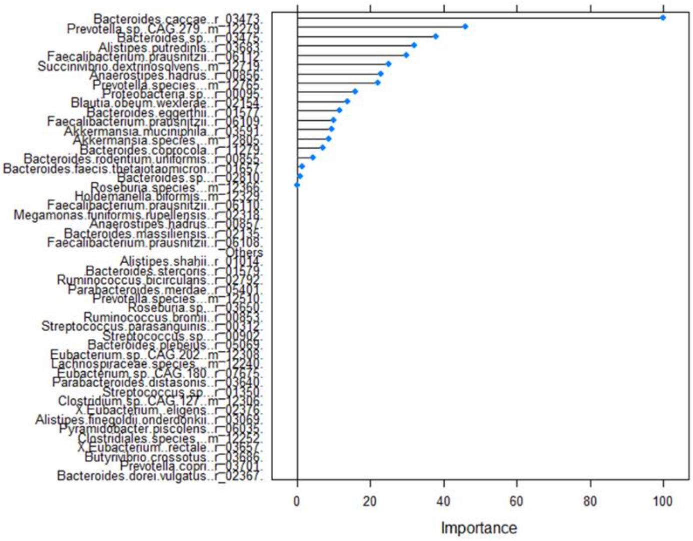


Supplementary Figure 5: Feature importance of Nutrients (left) and food groups (right) selected by Ridge Regression models for the richness score rS.



gra_galldairy	5.32E-05
gra_sgmilkyogurt	5.96E-05
gra_sgcheesse	1.07E-04
gra_sgdairydessert	-7.79E-04
gra_gallmeat	8.62E-05
gra_sgwhitemeat	8.48E-04
gra_sgredmeat	-4.06E-04
gra_sgorganmeat	-1.36E-03
gra_sgcuredmeat	2.10E-04
gra_sgprocessedmeat	-8.22E-04
gra_sgcuredprocessedmeat	-3.34E-04
gra_gallsea	3.20E-04
gra_sgfish	2.41E-04
gra_sgothersea	3.36E-03
gra_meatseafood	-6.71E-06
gra_gallready	4.16E-04
gra_gallveg	2.45E-05
gra_sg1leafyveg	6.45E-04
gra_sg1starchveg	3.88E-06
gra_sg1sgfruitingveg	-6.07E-04
gra_sg1sggrainsveg	2.26E-04
gra_sg2redyellveg	1.20E-04
gra_sg2greenveg	2.59E-05
gra_sg2whiteveg	5.70E-04
gra_gleg	5.42E-04
gra_gallfru	3.62E-04
gra_golives	-3.01E-03
gra_gnuts	1.33E-03
gra_gallcer	1.12E-04
gra_sgbread	1.21E-04
gra_sgricepasta	-2.40E-04
gra_gallfats	5.92E-03
gra_gflour	4.36E-04
gra_gchoc	3.95E-03
gra_gsugar	1.43E-04
gra_gsauces	1.06E-03
gra_gnonalc	4.42E-05
gra_sgsugbev	-3.00E-04
gra_sgsoftdr	3.45E-04
gra_sgjuice	-1.82E-04
gra_coffee	8.18E-05
gra_decoffee	5.80E-04
gra_tea	2.22E-04
beer2	-1.65E-04
wine2	-8.56E-04
spirits2	-1.15E-02
fortified2	-3.61E-03

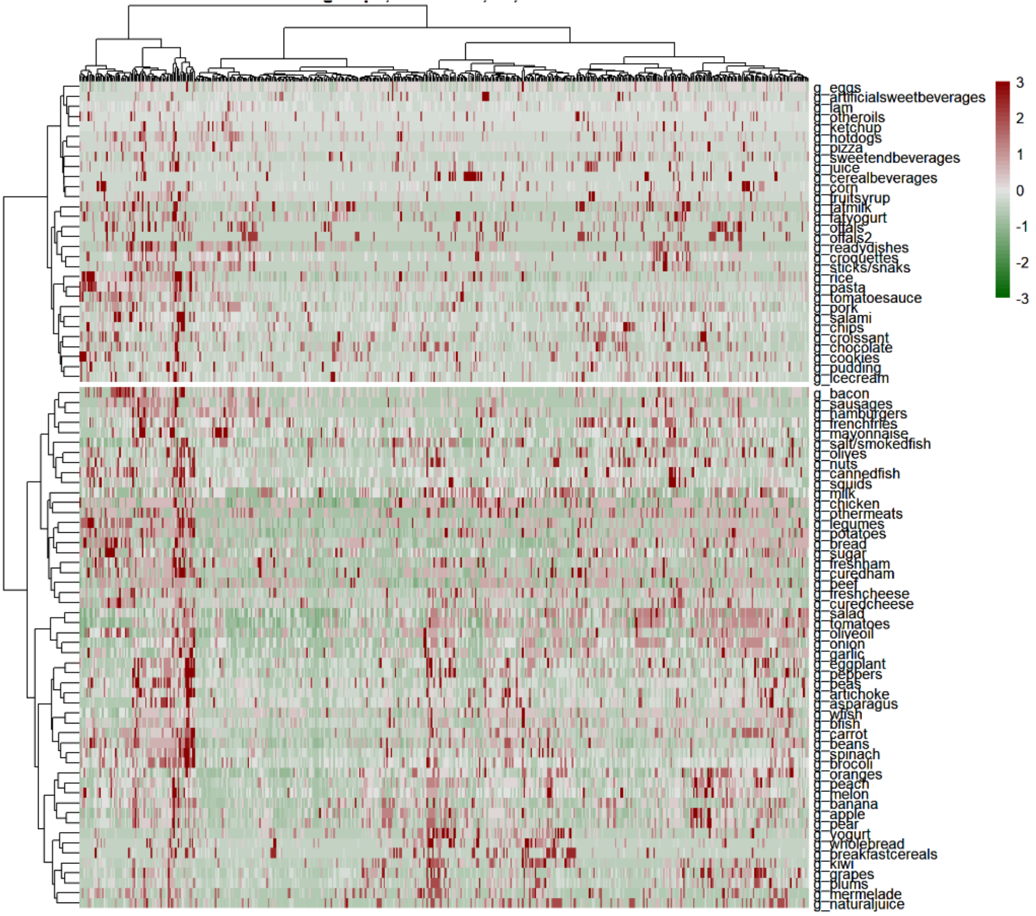
Supplementary Figure 6: Feature importance of the most prevalent taxa (N=50) for DRRD score selected by LASSO.



Coeficients:

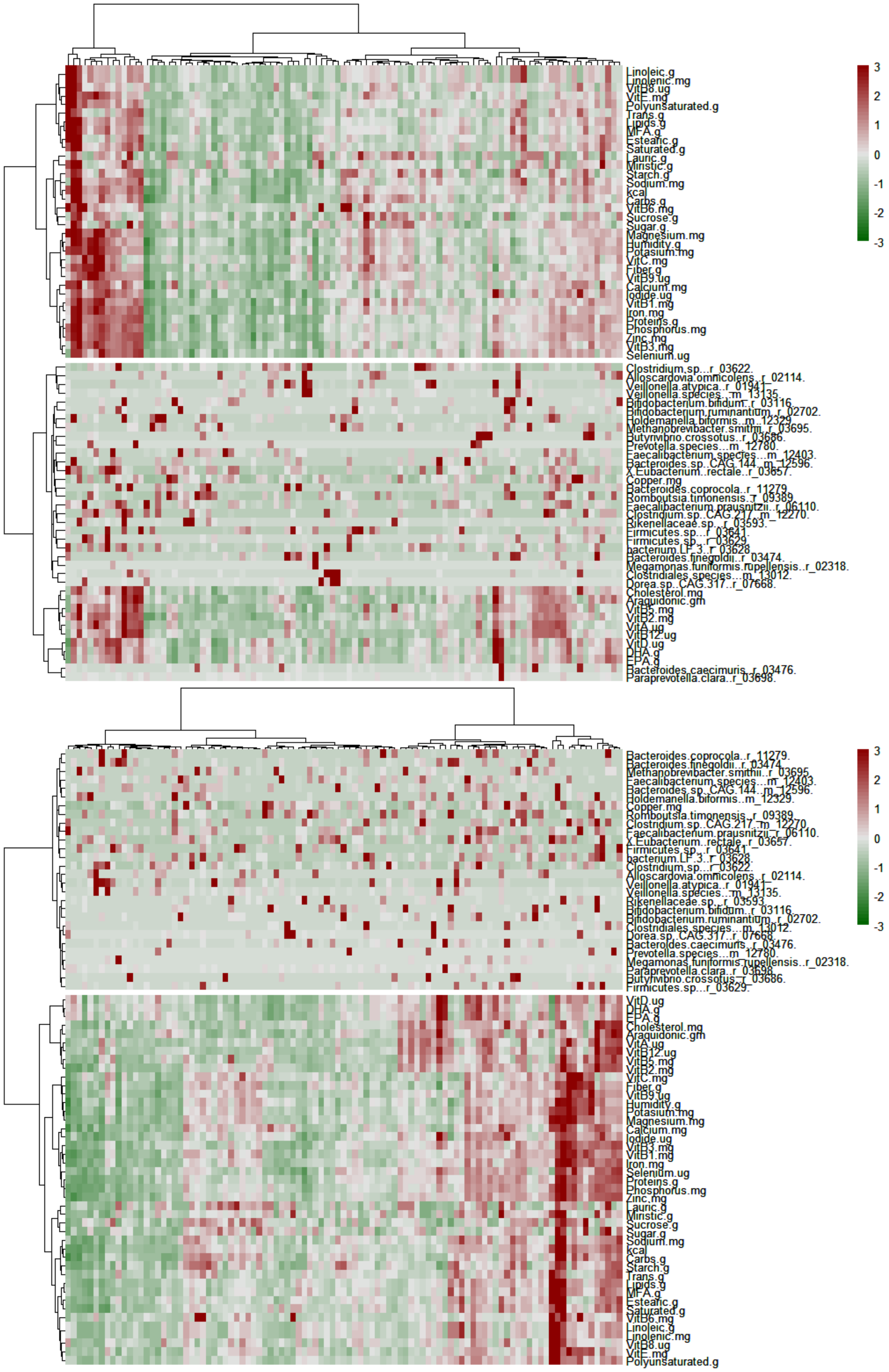
```
 (Intercept)        27.0455967
Bacteroides.rodentium.uniformis..r_00855.           3.8693444
Blautia.obeum.wexlerae..r_02154.           -12.2156586
Alistipes.putredinis..r_03683.        28.6387544
Proteobacteria.sp...r_00095.        -14.234156
Roseburia.species...m_12366.      -0.1244635
Akkermansia.muciniphila..r_03591.           8.4996434
Bacteroides.faecis.thetaiotaomicron..r_01657.      1.1985538
Faecalibacterium.prausnitzii..r_06109.      9.010479
Faecalibacterium.prausnitzii..r_06112.      26.6143472
Bacteroides.sp...r_03475.33.9770696
Bacteroides.coprocola..r_11279. -6.1956002
Bacteroides.eggerthii..r_01577.    -10.3307721
Bacteroides.caccae..r_03473.     -89.392068
Prevotella.species...m_12765.      -19.662022
Anaerostipes.hadrus..r_00856.     -20.3885194
Bacteroides.sp...r_02810.0.8033954
Akkermansia.species...m_12805. -7.7903256
Succinivibrio.dextrinosolvens..m_12719.    -22.4909647
Prevotella.sp..CAG.279..m_12279.           41.1459168
```

Supplementary Figure 7: Dietary clusters derived from intake for food items within the PanGenEU study and among 511 controls. Manhattan distances.



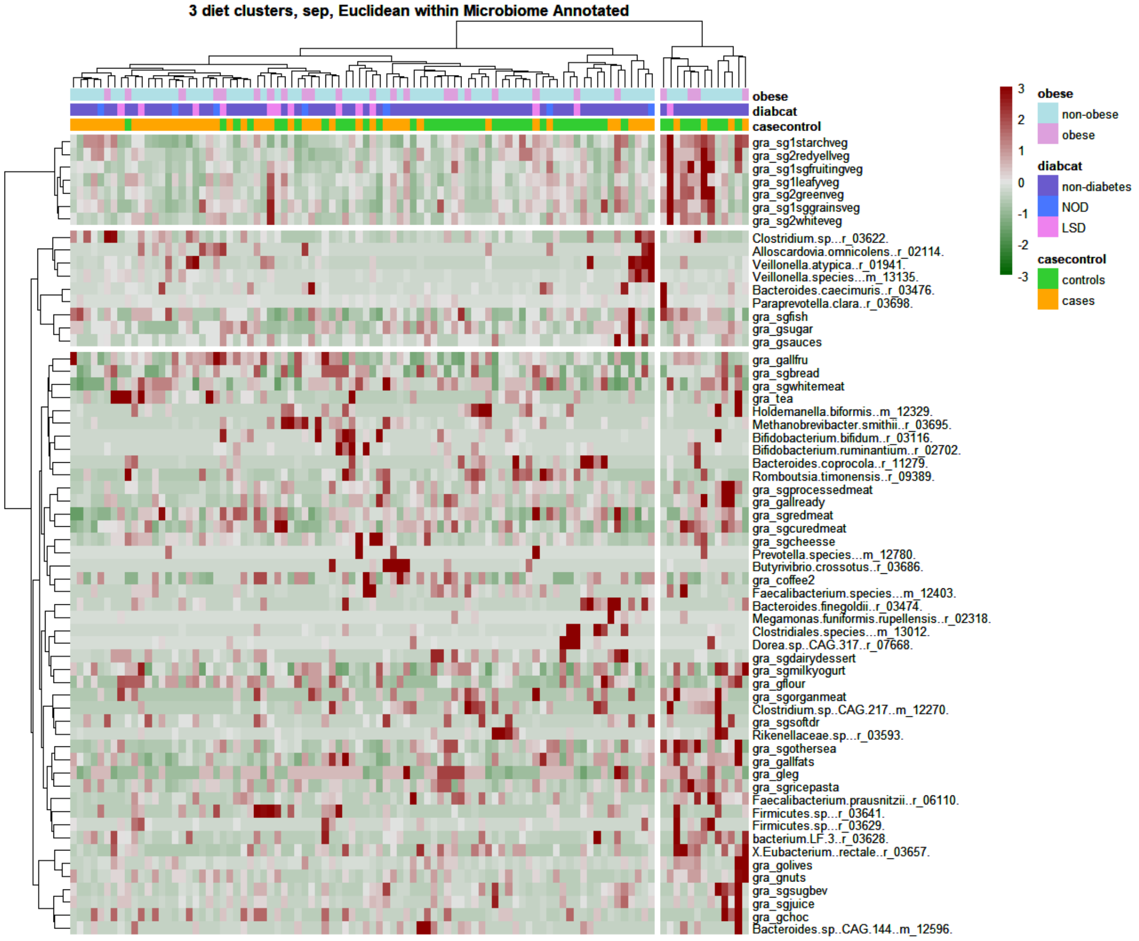
Hierarchical cluster obtained with Manhattan distances applied to taxa and foods in gram (g), after scaling all values (value-mean/SD).

Supplementary Figure 8: Clusters of nutrients and the stool microbial signature taxa (27 species) overall for PC cases and controls. Euclidean (A) and Manhattan (B) distance.



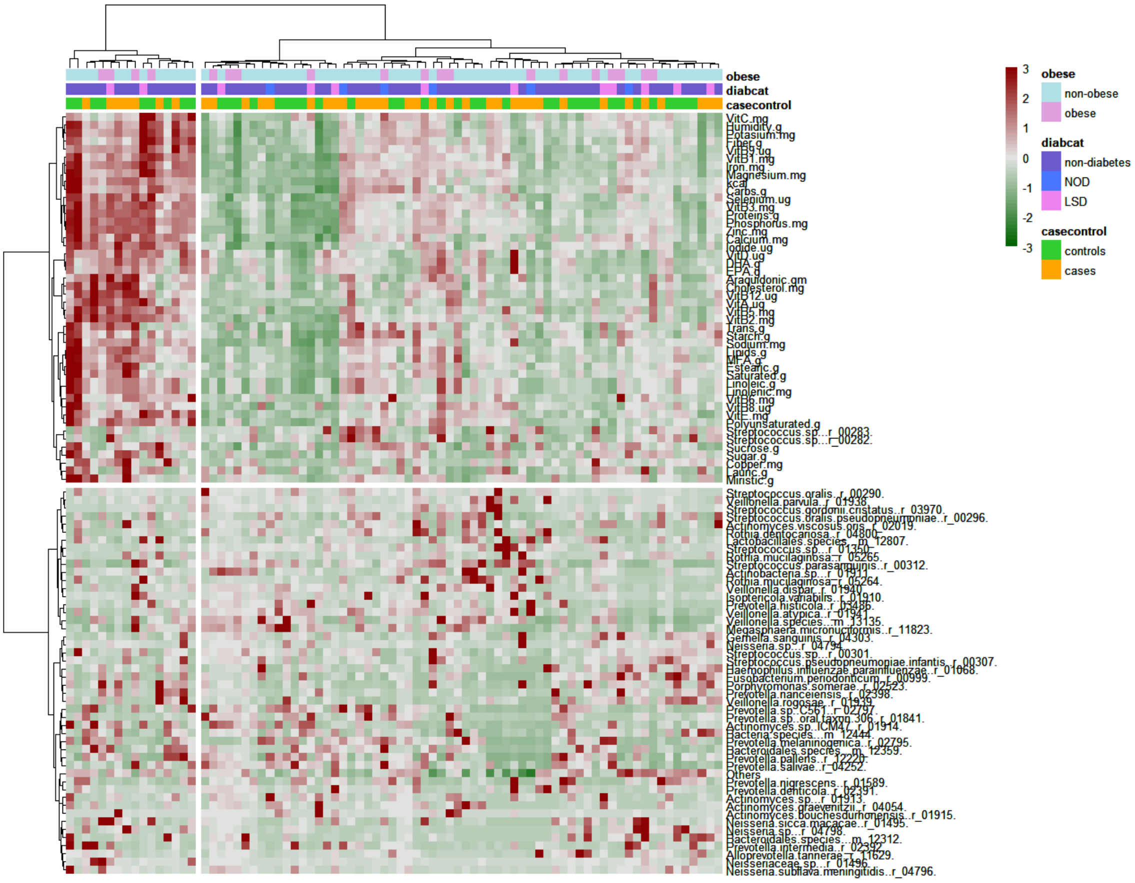
Hierarchical cluster obtained with Manhattan and Euclidean distances applied to taxa and nutrients, after scaling all values (value-mean/SD).

Supplementary Figure 9: Clusters of foods groups and the stool microbial signature taxa (27 species) overall for PC cases and controls. Euclidean distance.



Hierarchical cluster obtained with Euclidean distances applied to taxa and food groups, after scaling all values (value-mean/SD). For foods groups (all in grams of intake "gra"), the main group is indicated with "g" along with the corresponding subgroups "sg" before the food group´s name.

Supplementary Figure 10: Clusters of nutrients and the stool microbial signature taxa (top 50 most prevalent taxa) overall for PC cases and controls. Manhattan distance.



Hierarchical cluster obtained with Manhattan distances applied to taxa and food groups, after scaling all values (value-mean/SD).

Annex I

27 species of the faecal metagenimic signature (Reference 13: Kartal, Schmidt, Molina-Montes, et al. 2022):

More enriched among PC cases: "Methanobrevibacter smithii [r_03695]", "Veillonella atypica [r_01941]", "Firmicutes sp. [r_03641]", "Clostridium sp. [r_03622]", "Bacteroides finegoldii [r_03474]", "Firmicutes sp. [r_03629]", "bacterium LF-3 [r_03628]", "Alloscardovia omnicolens [r_02114]", "Prevotella species [m_12780]", "Veillonella species [m_13135]", "Butyrivibrio crossotus [r_03686]",

More enriched among controls: "Clostridiales species [m_13012]", "Megamonas funiformis/rupellensis [r_02318]", "Holdemanella biformis [m_12329]", "Dorea sp. CAG:317 [r_07668]", "Bifidobacterium ruminantium [r_02702]", "Bacteroides caecimuris [r_03476]", "Bacteroides sp. CAG:144 [m_12596]", "Faecalibacterium species [m_12403]", "Rikenellaceae sp. [r_03593]", "Paraprevotella clara [r_03698]", "Clostridium sp. CAG:217 [m_12270]", "[Eubacterium] rectale [r_03657]", "Bacteroides coprocola [r_11279]", "Faecalibacterium prausnitzii [r_06110]", "Bifidobacterium bifidum [r_03116]", "Romboutsia timonensis [r_09389]"

Annex II: Reference 13

A faecal microbiota signature with high specificity for pancreatic cancer.

Kartal E#, Schmidt TSB#, Molina-Montes E# (co-first authors), Rodríguez-Perales S, Wirbel J, Maistrenko OM, Akanni WA, Alashkar Alhamwe B, Alves RJ, Carrato A, Erasmus HP, Estudillo L, Finkelmeier F, Fullam A, Glazek AM, Gómez-Rubio P, Hercog R, Jung F, Kandels S, Kersting S, Langheinrich M, Márquez M, Molero X, Orakov A, Van Rossum T, Torres-Ruiz R, Telzerow A, Zych K; MAGIC Study investigators; PanGenEU Study investigators, Benes V, Zeller G, Trebicka J, Real FX, Malats N, Bork P.

Gut. 2022 Jul;71(7):1359-1372. doi: 10.1136/gutjnl-2021-324755. Epub 2022 Mar 8.

OPEN ACCESS

Original research

A faecal microbiota signature with high specificity for pancreatic cancer

Ece Kartal ,[1,2] Thomas S B Schmidt ,[1] Esther Molina-Montes ,[3,4] Sandra Rodríguez-Perales ,[4,5] Jakob Wirbel ,[1,2] Oleksandr M Maistrenko ,[1] Wasiu A Akanni ,[1] Bilal Alashkar Alhamwe ,[6] Renato J Alves ,[1] Alfredo Carrato ,[4,7,8] Hans-Peter Erasmus,[9] Lidia Estudillo ,[3,4] Fabian Finkelmeier,[9,10] Anthony Fullam ,[1] Anna M Glazek,[1] Paulina Gómez-Rubio,[3,4] Rajna Hercog,[11] Ferris Jung ,[11] Stefanie Kandels ,[1] Stephan Kersting ,[12,13] Melanie Langheinrich ,[13] Mirari Márquez,[3,4] Xavier Molero,[14,15,16] Askarbek Orakov ,[1] Thea Van Rossum ,[1] Raul Torres-Ruiz ,[4,5] Anja Telzerow ,[11] Konrad Zych ,[1] MAGIC Study investigators, PanGenEU Study investigators, Vladimir Benes ,[11] Georg Zeller ,[1] Jonel Trebicka ,[9,17] Francisco X Real ,[4,18,19] Nuria Malats ,[3,4] Peer Bork [1,20,21,22]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2021-324755).

For numbered affiliations see end of article.

Correspondence to
Dr Nuria Malats; nmalats@cnio.es
Dr Peer Bork, Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Baden, Germany; peer.bork@embl.org

EK, TSBS and EM-M contributed equally.

NM and PB are joint senior authors.

Received 5 April 2021
Accepted 5 December 2021

To cite: Kartal E, Schmidt TSB, Molina-Montes E, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2021-324755

ABSTRACT

Background Recent evidence suggests a role for the microbiome in pancreatic ductal adenocarcinoma (PDAC) aetiology and progression.

Objective To explore the faecal and salivary microbiota as potential diagnostic biomarkers.

Methods We applied shotgun metagenomic and 16S rRNA amplicon sequencing to samples from a Spanish case–control study (n=136), including 57 cases, 50 controls, and 29 patients with chronic pancreatitis in the discovery phase, and from a German case–control study (n=76), in the validation phase.

Results Faecal metagenomic classifiers performed much better than saliva-based classifiers and identified patients with PDAC with an accuracy of up to 0.84 area under the receiver operating characteristic curve (AUROC) based on a set of 27 microbial species, with consistent accuracy across early and late disease stages. Performance further improved to up to 0.94 AUROC when we combined our microbiome-based predictions with serum levels of carbohydrate antigen (CA) 19–9, the only current non-invasive, Food and Drug Administration approved, low specificity PDAC diagnostic biomarker. Furthermore, a microbiota-based classification model confined to PDAC-enriched species was highly disease-specific when validated against 25 publicly available metagenomic study populations for various health conditions (n=5792). Both microbiome-based models had a high prediction accuracy on a German validation population (n=76). Several faecal PDAC marker species were detectable in pancreatic tumour and non-tumour tissue using 16S rRNA sequencing and fluorescence in situ hybridisation.

Conclusion Taken together, our results indicate that non-invasive, robust and specific faecal microbiota-based screening for the early detection of PDAC is feasible.

Significance of this study

What is already known about this subject?

- Pancreatic ductal adenocarcinoma (PDAC) is on the rise worldwide, posing a high disease burden and mortality rate, yet accurate, non-invasive diagnostic options remain unavailable.
- Alterations in the oral, faecal and pancreatic microbiome composition have been associated with an increased risk of PDAC.

What are the new findings?

- Stool microbiota-based classifiers are described that predict PDAC with high accuracy and specificity, independent of disease stage, with potential as agents for non-invasive diagnostics.
- A faecal metagenomic classifier identified PDAC with an accuracy of 0.84 area under the receiver operating characteristic curve (AUROC) in a Spanish cohort, based on 27 species. The accuracy improved to up to 0.94 AUROC when combined with the less specific carbohydrate antigen (CA) 19–9 serum marker.
- The classifier was validated in an independent German PDAC cohort (0.83 AUROC), and PDAC disease specificity was confirmed against 25 publicly available metagenomic study populations with various health conditions (n=5792).
- The presence of marker taxa enriched in faecal samples (*Veillonella*, *Streptococcus*, *Akkermansia*) and also taxa with differential abundance in healthy and tumour pancreatic tissues (*Bacteroides*, *Lactobacillus*, *Bifidobacterium*) was validated by fluorescence *in situ* hybridisation.

Significance of this study

How might it impact on clinical practice in the foreseeable future?

- Faecal microbiome-based detection of PDAC may provide a non-invasive, cost-effective and robust approach to early PDAC diagnosis.
- The presented PDAC-specific microbiome signatures, including links between microbial populations across tissues, provide novel microbiome-related hypotheses regarding disease aetiology, prevention and possible therapeutic intervention.

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is the most common form of pancreatic cancer and a major cause of cancer-related deaths despite relatively low incidence rates.[1 2] The high lethality of PDAC is a consequence of both late diagnosis and limited therapeutic options[3]: symptoms are unspecific and often emerge only during late disease stages, at which point tumours can be either locally non-resectable or present as metastatic disease. At present, PDAC is diagnosed using imaging tests.[4] Sensitive and affordable tests for an early detection of PDAC could therefore improve outcome. PDAC markers have been explored in pancreatic tissue,[5] urine[6 7] and serum.[8 9] Yet to date, the sole Food and Drug Administration (FDA)-approved PDAC biomarker remains serum carbohydrate antigen (CA) 19-9. CA19-9 has limited disease specificity as levels can be elevated in several other concomitant conditions (eg, biliary obstruction) and is therefore mostly used as a marker for PDAC surveillance, rather than screening or diagnosis.[10–14]

PDAC has a complex aetiology, with established risk factors that include age, chronic pancreatitis, diabetes mellitus, obesity, asthma, blood group and lifestyle (eg, smoking and heavy alcohol consumption).[15 16] The role of these risk factors in PDAC aetiology may also be complemented—or sometimes indeed mediated—by alterations in the microbiome. For example, poor oral hygiene and periodontitis have been associated with an increased PDAC risk,[17] an observation that also extends to periodontitis- and caries-associated microbial species.[18 19] Shifts in these species are sometimes part of wider compositional changes in the oral microbiome[20 21] or have been explored as PDAC risk factors in their own right.[22] Similarly, microbial composition in the gut[23–25] and duodenum,[26 27] quantified via 16S rRNA amplicon sequencing, have previously been linked to PDAC risk.

The human pancreas harbours a microbiome that shares species with the mouth and the gut,[25 28–32] although its exact composition has remained elusive owing to the challenges associated with contamination control in low bacterial biomass samples.[33] In murine models, microbes originating from the intestine can contribute to carcinogenesis in the pancreatic duct,[25 30] suggesting a role for the microbiome in PDAC aetiology and progression that was recently extended to fungi.[34] Moreover, the pancreatic tumour microbiome may also be associated with disease progression and long-term survival in patients with PDAC.[31]

However, the translation of these advances into PDAC-specific microbiome signatures for clinical applications has so far remained largely unexplored. Here, we present the identification of robust, specific microbial PDAC signatures based on a metagenomic survey of a Spanish (ES) study population of 57 newly diagnosed and treatment-naïve patients with PDAC, 29 patients with chronic pancreatitis (CP), and 50 matched controls. We sampled saliva, faeces, pancreatic normal and tumour tissue and assessed microbial composition using whole-genome shotgun metagenomics, 16S rRNA amplicon sequencing, and fluorescence *in situ* hybridisation (FISH) assays. The best discrimination between patients with PDAC and non-PDAC subjects was achieved by statistical models based on a set of 27 faecal microbial species that could be quantified in a targeted manner in a diagnostic setting. The prediction accuracy of microbiome-based models was confirmed in an independent German (DE) PDAC validation population including 44 patients with PDAC and 32 controls and was further improved when combined with serum levels of CA19-9. We further validated the disease specificity of these models against existing data from 25 studies (n=5792) of nine diseases.[35–59] Several of the PDAC-enriched species were also detected in cancer tissue, with possible links to oral and intestinal populations, supporting their potential role in PDAC pathogenesis, as previously reported.[25 30 31 34]

METHODS

Subject recruitment and sample collection

A case–control design was applied. Subjects were prospectively recruited between 2016 and 2019 from the Hospital Ramón y Cajal in Madrid and Hospital Vall d'Hebron in Barcelona, Spain, using the same protocols for biological sample collection, processing and storage. Subjects with newly diagnosed PDAC (n=57), aged >18 years, were identified prior to any cancer treatment. Subjects in whom PDAC was suspected were recruited, and sampling was done before any treatment. Patients with chronic pancreatitis (CP, n=29) were recruited from the same hospitals. Controls matched for age, gender and hospital were selected from inpatients with a primary diagnosis for hospital admission not related to PDAC risk factors. Participants incapable of participating in the study owing to impairment of physical ability were excluded. Institutional review board ethical approval (CEI PI 26 2015-v7) and written informed consent were obtained from participating centres and study participants, respectively. Epidemiological and lifestyle data were collected by trained monitors during face-to-face interviews through a structured questionnaire. Clinical data, including stage of the diseases and follow-up data, were retrieved from hospital charts by the same monitors, likewise using structured questionnaires. Recorded jaundice status was additionally confirmed and extended by direct bilirubin measurements from blood samples in CNIO, Madrid. All data were entered, edited and managed using REDCap. Missing lifestyle and medication values in the metadata (missing overall in 3.1%) were imputed using a random forest-based algorithm for missing data imputation called missForest (n=100 trees).[60] The imputation accuracy was high according to the imputation error estimate (mean out-of-bag error=0.12). Serum CA19-9 levels were analysed by electrochemiluminescence immunoassay (ECLIA, Roche Diagnostics, Germany) following the manufacturer's instructions in the Institute of Laboratory Medicine and Pathobiochemistry, Marburg, Germany. Each sample was assayed in duplicate, with positive controls assayed in each plate (online supplemental table S1).

Stool and saliva (mouthwash) samples were preserved in RNALater and stored at 4°C immediately for 12 hours, then transferred to −20°C for another 24 hours, and then stored at −80°C until DNA extraction. Tumour and non-affected tissue samples were collected during surgery for a subset of individuals, immediately flash-frozen in liquid nitrogen after pathological assessment, and preserved at −80°C. All the samples were shipped on dry ice.

An independent validation population was recruited at the Department of Surgery, University Hospital of Erlangen (32 PDAC and 32 control samples) and Section for Translational Hepatology, Department of Internal Medicine I, Goethe University Clinic, Frankfurt (12 PDAC samples) using the same protocols for biological sample collection, processing and storage. Matched controls were selected from inpatients with a primary diagnosis for hospital admission not related to PDAC risk factors. The study was approved by the local ethics committees (SGI-3–2019, 451_18 B), and written informed consent from study participants was obtained. Clinical data, including disease stage and follow-up data, were retrieved from the clinical records of the hospital charts of the respective patients (online supplemental table S2). Serum CA19-9 levels were analysed by a routine immunoassay (Roche Diagnostics, Germany) following the manufacturer's instructions. Stool samples were preserved in OMNIgene-Gut OM-200 vials (Steinbrenner Laborsysteme GmbH, Germany) and stored at −80°C immediately until DNA extraction.

Sample processing

Faecal and salivary samples were thawed on ice, aliquoted, and genomic DNA was extracted using the Qiagen Allprep PowerFecal DNA/RNA kit according to the manufacturer's instructions (Qiagen, Hilden, Germany). Genomic DNA from pancreatic tumorous and non-tumoral tissue samples was extracted using the Qiagen DNeasy blood and tissue kit in a protocol modified from Del Castillo *et al*[26]: cells were lysed mechanically (with 5 mm stainless steel beads at 25 Hz for 150 s), followed by lysozyme treatment (20 mg/mL) and protease and RNAse digestion (56°C for 2 h). All samples were randomly assigned to extraction batches. To account for potential bacterial contamination of extraction, polymerase chain reaction (PCR) and sequencing kits, we included negative controls (extraction blanks) with each tissue DNA extraction batch (online supplemental figure 1).

16S rRNA amplicon sequencing

Pancreatic tissue DNA was enriched for 16S rRNA in a preamplification PCR using primers 331F (5'-TCCTACGGGAGGCAGCAGT-3')[61] and 979R (5'-GGTTCTKCGCGTTGCWTC-3').[62] The cycling conditions consisted of an initial template denaturation at 98°C for 2 min, followed by 30 cycles of denaturation at 98°C for 10 s, annealing at 65°C for 20 s, extension at 72°C for 30 s and a final extension at 72°C for 10 min. This was followed by a size-selective cleanup using SPRIselect magnetic beads (0.8 left-sized; Beckman Coulter, Brea, California, USA). Faecal and salivary DNA were not preamplified.

Targeted amplification of the 16S rRNA V4 region (primer sequences F515 5'-GTGCCAGCMGCCGCGGTAA-3' and R806 5'-GGACTACHVGGGTWTCTAAT-3'),[63] was performed using the KAPA HiFi HotStart PCR mix (Roche, Basel, Switzerland) in a two-step barcoded PCR protocol (NEXTflex 16S V4 Amplicon-Seq Kit; Bioo Scientific, Austin, Texas, USA) with minor modifications from the manufacturer's instructions. PCR products were pooled, purified using size-selective SPRIselect magnetic beads (0.8 left-sized) and then sequenced at 2×250 bp on an Illumina MiSeq (Illumina, San Diego, California, USA) at the Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg.

16S rRNA amplicon data processing

Raw reads were quality trimmed, denoised and filtered against chimeric PCR artefacts using DADA2.[64] The resulting exact amplicon sequence variants (ASVs) were taxonomically classified and mapped to a reference set of operational taxonomic units (OTUs) at 98% sequence similarity using MAPseq.[65] Reads that did not confidently map to the reference were aligned to bacterial and archaeal secondary structure-aware small subunit rRNA models using Infernal[66] and clustered into OTUs with 98% average linkage using HPC-CLUST,[67] as described previously.[68] As a result, we obtained taxa tables at two resolutions: 100% identical ASVs and 98% open-reference OTUs; unless otherwise indicated, analyses in the main text refer to OTUs.

Count tables were noise filtered by removing samples retaining less than 500 reads and taxa observed in fewer than five samples; this removed 2.5% of total reads from the dataset. For 18 salivary samples, technical replicates were merged after confirming that they strongly correlated with community composition. For pancreatic tissue and tumour samples, ASVs observed in negative control samples were removed, as were reads mapping to known reagent kit contaminants.[33] After these steps, we retained 308 16S rRNA amplicon samples from 143 subjects for further analyses (130 salivary, 118 faecal, 20 of unaffected pancreatic tissue, 23 of tumour tissue with 17 matching PDAC tissue samples).

Shotgun metagenomic sequencing

Metagenomic libraries for 212 faecal and 100 salivary samples were prepared using the NEB Ultra II and SPRI HD kits, depending on the concentration of starting material, with a targeted insert size of 350, and sequenced on an Illumina HiSeq 4000 platform (Illumina, San Diego, California, USA) in 2×150 bp paired-end setup to a target depth of 8 Gbp per sample at the Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg. Sequencing statistics for each sample are provided in the associated git repository (https://github.com/psecekartal/PDAC.git). For three salivary and one faecal samples, technical replicates were merged after confirming that they strongly correlated in community composition.

Metagenome data processing

Metagenomic data were processed using established workflows in NGLess v0.7.1.[69] Raw reads were quality trimmed (≥45 bp at Phred score ≥25) and filtered against the human genome (version hg19, mapping at ≥90% identity across ≥45 bp). The resulting filtered reads were mapped (≥97% identity across ≥45 bp) against the representative genomes of 5306 species-level genome clusters obtained from the proGenomes database v2.[70]

Taxonomic profiles were obtained using the mOTU profiler v2.5[71] and filtered to retain only species observed at a relative abundance $\geq 10^{-5}$ in ≥2% of samples. Gene functional profiles were obtained from mappings against a global microbioal gene catalogue (GMGCv1, Coelho *et al*[72], http://gmgc.embl.de/), by summarising read counts from eggNOG v4.5[73] annotations to orthologous groups and KEGG modules. Features with a relative abundance of $\geq 10^{-5}$ in ≥15% of samples were retained for further analyses.

Microbiome data statistical analyses

All data analyses were conducted in the R Statistical Computing framework v3.4 or higher.

Rarefied per-sample taxa diversity ('alpha diversity', averaged over 100 rarefaction iterations) was calculated as the effective number of taxa with Hill coefficients of q=0 (ie, taxa richness), q=1 (exponential of Shannon entropy) and q=2 (inverse Simpson index), and evenness measures as ratios thereof. Unless otherwise stated, results in the main text refer to taxa richness. Differences in alpha diversity were tested using analysis of variance (ANOVA) followed by post hoc tests and Benjamini-Hochberg correction, as specified in the main text.

Between-sample differences in community composition (‘beta diversity’) were quantified as Bray-Curtis dissimilarity on raw or square-root transformed counts, abundance-weighted Jaccard index, and abundance-weighted and unweighted TINA index, as described previously.[74] Trends between these indices were generally consistent, unless otherwise stated. Results are reported for Bray-Curtis dissimilarities on non-transformed data. Associations of community composition to microbiome-external factors were quantified using the ‘adonis2’ implementation of PERMANOVA and distance-based redundancy analysis in the R package vegan v2.5.[75] To quantify potentially confounding univariate links between the abundance of individual taxa and subject-specific variables (see main text), we performed either ANOVA or non-parametric Kruskal-Wallis tests, depending on abundance distributions (online supplemental figure 2-3 and online supplemental table S4-S5). Bilirubin levels were measured from blood samples, and jaundice status was confirmed by clinical records. Owing to missing jaundice status for several individuals, values used for further analysis were imputed from existing data (figure 1, online supplemental table S1-S3).

Multivariable statistical modelling and model evaluation

In order to train multivariable statistical models for the prediction of pancreatic cancer, we first removed taxa with low overall abundance and prevalence (abundance cut-off point: 0.001). Then, features were normalised by log10 transformation (to avoid infinite values from the logarithm, a pseudo-count of 1e-05 was added to all values) followed by standardisation as centred log-ratio (log.clr). Data were randomly split into test and

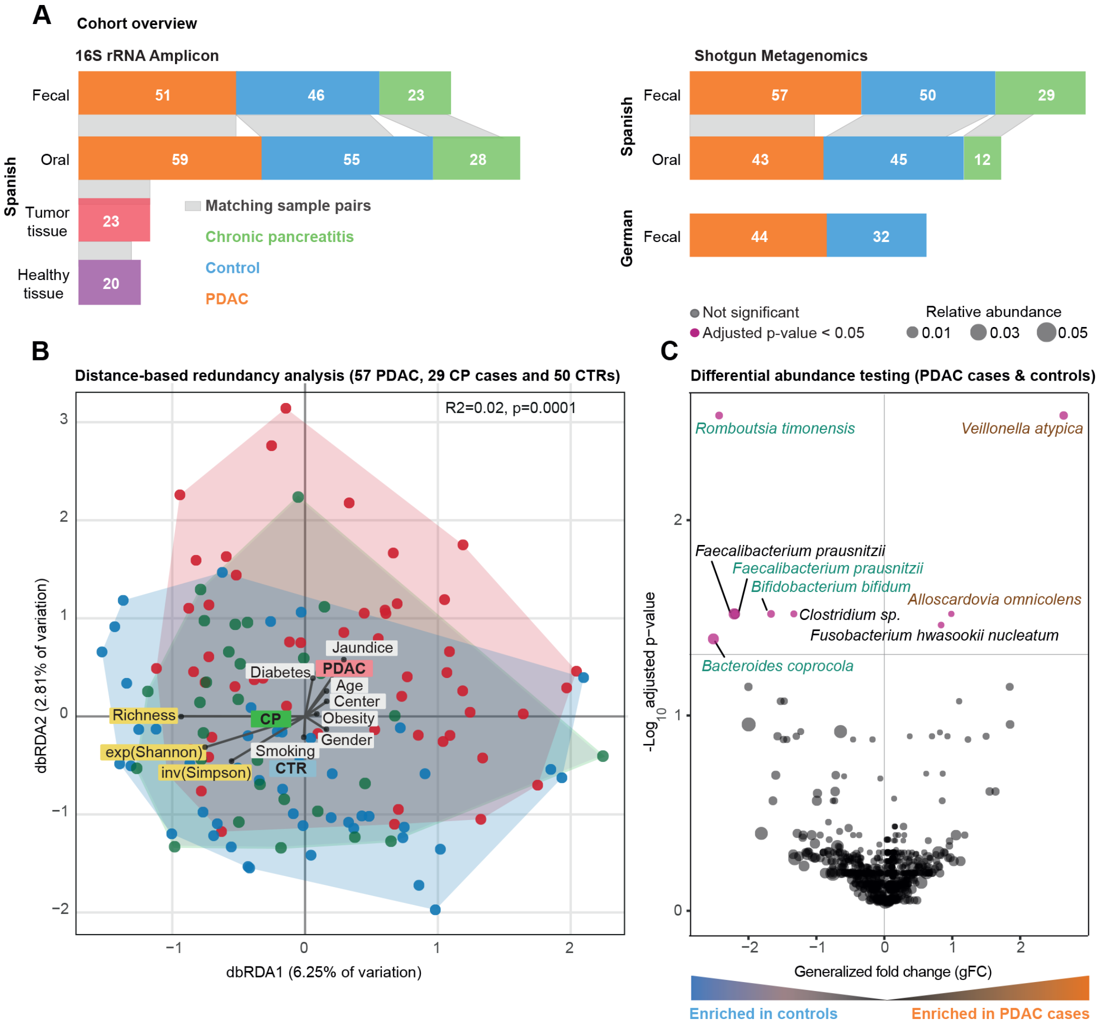


Figure 1 Community analysis of Spanish faecal microbiome data. (A) Study population overview. Grey bands between the bar plots indicate samples of matching body sites within individuals. (B) Bray-Curtis distance-based redundancy analysis (dbRDA) of pancreatic ductal adenocarcinoma (PDAC), chronic pancreatitis (CP) and control (CTR) faecal microbiome data in a Spanish (ES) cohort. PDAC samples are shown as red coloured circles, patients with CP as green and controls as blue. Richness, exponential Shannon (exp(Shannon)) and inverse Simpson (inv(Simpson)) diversity measures are also visualised with arrows similarly to tested metadata variables. The distance of the meta-variable from the centre represents the confounding effect size (see ‘Methods’). (C) Wilcoxon test results of ES faecal microbiome data to test enriched taxa between PDAC and control cases (see ‘Methods’). Y-axis is log10(FDR corrected p values), X-axis is generalised fold change, and dot size represents the relative abundance of a given species. Red dots represent significantly differentially abundant species in either group, while black dots show non-significant species after FDR correction. Green and brown-coloured species are selected in metagenomic model-1 as predictors of PDAC. FDR, false discovery rate.

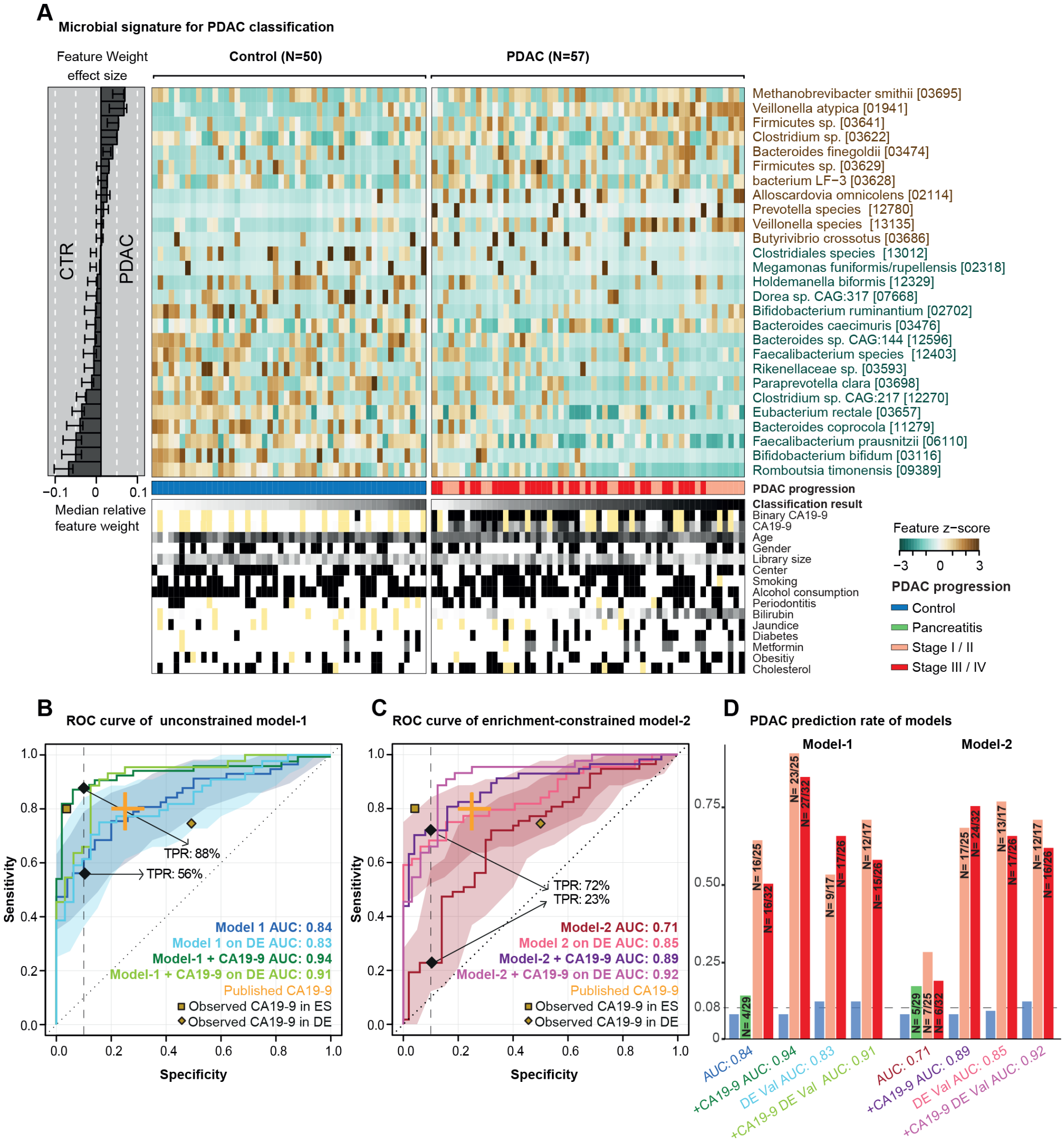


Figure 2 Predictive microbiome signatures of pancreatic ductal adenocarcinoma (PDAC). (A) Normalised abundance of 27 selected species in the faecal microbiome across samples shown as a heat map. The right panel represents the contribution of each selected feature to the overall model-1, and the robustness (the percentage of models in which the feature is included as predictor) of each feature is presented as percentage. Classification scores from cross-validation of each individual and condition for tested meta-variables are displayed at the bottom of the panel, yellow representing missing information. (B–D) Internal cross-validation results of unconstrained model-1 (without feature selection), enrichment-constrained model-2 (constrained to positive features) and combination of carbohydrate antigen (CA)19-9 (using a threshold of 37 µL/mL) with microbial features (see 'Methods') are shown as receiver operating characteristic (ROC) curve with 95% CI shaded in corresponding colour. True positive rates (TPRs) are given as a percentage at a 90% specificity cut-off. Validation of all models on an independent German (DE) PDAC test population (n=76) is represented as well. Published CA19-9 accuracy from a meta-study shown in orange. The yellow dots represent observed CA19-9 accuracies in our populations (data available for 33/50 controls (CTRs) and 44/57 patients with PDAC in the Spanish (ES) and for 8/32 CTRs and 44/44 patients with PDAC in the German (DE) population) (D) TPRs of all models at different PDAC progression stages and in addition, the false-positive rate for patients with chronic pancreatitis and controls at a 90% specificity cut-off are shown as bar plots. Stages I and II and stages III and IV are combined owing to the overall low sample size. The number of predicted cases compared with the total is also shown on the top of each bar. DE-Val, German validation population.

training sets in a 10 times repeated 10-fold cross-validation. For each test fold, the remaining folds were used as training data to train an L1-regularised (LASSO) logistic regression model[76] using the implementation within the LiblineaR R package v2.10.[77] The trained model was then used to predict the left-out test set and finally, all predictions were used to calculate the area under the receiver operating characteristics curve (AUROC) (figure 2).

In a second approach, features were filtered within the cross-validation (that is, for each training set) by first calculating the single-feature AUROC and then removing features with an AUROC <0.5, thereby selecting features enriched in PDAC ('enrichment-constrained' model).

In order to combine the predictions from the microbiome-based machine learning models with the CA19-9 marker, the coded CA19-9 marker (1 for positive, 0 for negative or not available) was added to the mean predictions from the repeated cross-validation runs, resulting in an OR combination. Alternatively, the AND combination was calculated by multiplying the predictions with the CA19-9 marker. ROC curves and AUROC values were calculated for both combinations using the pROC R package v1.15.[78] The 95% CI is shaded in corresponding colour and specified in figure legends for each ROC curve.

The trained ES metagenomic classifiers for PDAC were then applied to the DE dataset after applying a data normalisation routine, which selects the same set of features and uses the same normalisation parameters (for example, the mean of a feature for standardisation by using the frozen normalisation functionality in SIAMCAT) as in the normalisation procedure from the ES pancreatic cancer dataset. For this analysis, the cut-off point for the predictions was set to a false-positive rate of 10% among controls in the initial ES PDAC study population (figure 2).

All steps of data preprocessing (filtering and normalisation), model training, predictions and model evaluation were performed using the SIAMCAT R package v.1.5.0[79] (https://siamcat.embl.de/).

External validation of the metagenomic classifiers

To assess the disease specificity of the trained models, we obtained predictions for samples from other gut metagenomic datasets (online supplemental table S6) for the full list, including accession numbers). We performed a literature search to identify publicly available datasets of faecal metagenomes in case–control or cohort studies for relevant diseases. For a total set of 25 studies covering 5792 samples across nine disease states, raw sequencing data were downloaded from the European Nucleotide Archive and taxonomically profiled as described above.[35–59]

The trained metagenomic classifiers for PDAC were then applied to each external dataset after applying a data normalisation routine which selects the same set of features and uses the same normalisation parameters (for example, the mean of a feature for standardisation by using frozen normalisation functionality in SIAMCAT) as in the normalisation procedure from the pancreatic cancer dataset. Then, predictions were assessed for disease specificity because high prediction scores for samples from other disease samples would indicate that the classifier relies on general features of dysbiosis in contrast to signals specific to pancreatic cancer, which would not result in elevated false-positive rates on samples from other diseases. For this analysis, the cut-off point for the predictions was set at a false-positive rate of 10% among controls in the initial PDAC study population (figure 3). The effect of age, sex and sequencing depth of 25 populations on prediction score were tested by using the cor.test function (Spearman method) in the car R package v3.0–3.

Subspecies and strain-level analyses

Metagenomic reads were mapped against species-representative genomes from the proGenomes v1 database[80] (see above). Microbial single nucleotide variants were called from uniquely mapping reads using metaSNV,[81] and within-species allele distances between samples were calculated as described previously.[82] Associations between allele distance and PDAC disease state were quantified using PERMANOVA after stratifying for potential confounders (including sampled body site).

Oral-intestinal transmission of strains was quantified as described previously.[83] In short, the overlap between microbial single nucleotide variants in salivary and faecal samples within subjects was contrasted with a between-subject background to compute a quantitative oral-faecal transmission score and p value. Associations of species- and subject-specific transmission scores with clinical factors were tested using ANOVA and *post hoc* tests, followed by a Benjamini-Hochberg correction for multiple tests.

Fluorescence *in situ* hybridisation microscopy

FISH analyses were performed using probes specifically targeting the 16S rRNA sequence unique to a particular taxon of bacteria (figure 4). All probes were selected based on a literature search and the corresponding taxa are displayed in online supplemental table S7).

Pancreatic tumour and normal pancreas samples were obtained from the pathology department and immediately frozen in liquid nitrogen within less than 30 min of surgical excision. Sterile material was used to dissect the different samples. The minimum size of tissue for freezing was approximately 0.125 cm^3 (0.5×0.5×0.5 cm). Samples were transferred from the temporary liquid nitrogen transport container and kept in a locked freezer at –80°C. Before analysis they were transported on dry ice, moved to an optimal cutting temperature mould in liquid nitrogen and immediately cut on a cryotome to obtain 10 sections of 3–5 µm each. All material was sterilised with ethanol after each sample handling.

Tissue sections of 5 µm thickness were mounted on positively charged slides (SuperFrost, Thermo Scientific). Briefly, tissues were postfixed in freshly prepared 4% paraformaldehyde. After enhancement of the bacteria wall permeabilisation by lysozyme treatment (10 g/L Tris HCl 6.5M), samples were hybridised for 1 hour at 45°C in the presence of the specific probe in a hybridiser machine (DAKO). Hybridisation was done in 20 µL of hybridisation buffer (20 nM Tris, pH 8.0. 0.9 M NaCl, 0.02% sodium dodecyl sulfate, 30% formamide) added to 100 ng of the probe. Finally, the tissues were washed in washing solution (70% formamide, 10 mM Tris pH7.2 and 01% bovine serum albumin), dehydrated in a series of ethanol samples, air-dried and stained with 0.5 µg/mL DAPI (4',6,-diamidino-2-phenylindole)/antifade solution (Palex Medical). FISH images were captured using a Leica DM5500B microscope with a CCD camera (Photometrics SenSys) connected to a PC running the CytoVision software 7.2 image analysis system (Applied Imaging). Images were analysed blind and scored based on the intensity of the probe signal.

RESULTS

PDAC is associated with moderate shifts in microbiome composition when controlling for confounding factors in shotgun metagenomic data

We studied 57 newly diagnosed, treatment-naïve patients with PDAC, 29 patients with chronic pancreatitis (CP), and 50 controls matched for age, gender and hospital. Participants were prospectively recruited from two hospitals in Barcelona and Madrid, Spain, between 2016 and 2018, using the same standards (see subject characteristics in figure 1A and online supplemental table S1-S3 for the clinical data for each subject). We obtained faecal shotgun metagenomes for all subjects and salivary metagenomes for 45 patients with PDAC, 12 with CP, and 43 controls (see 'Methods'). The

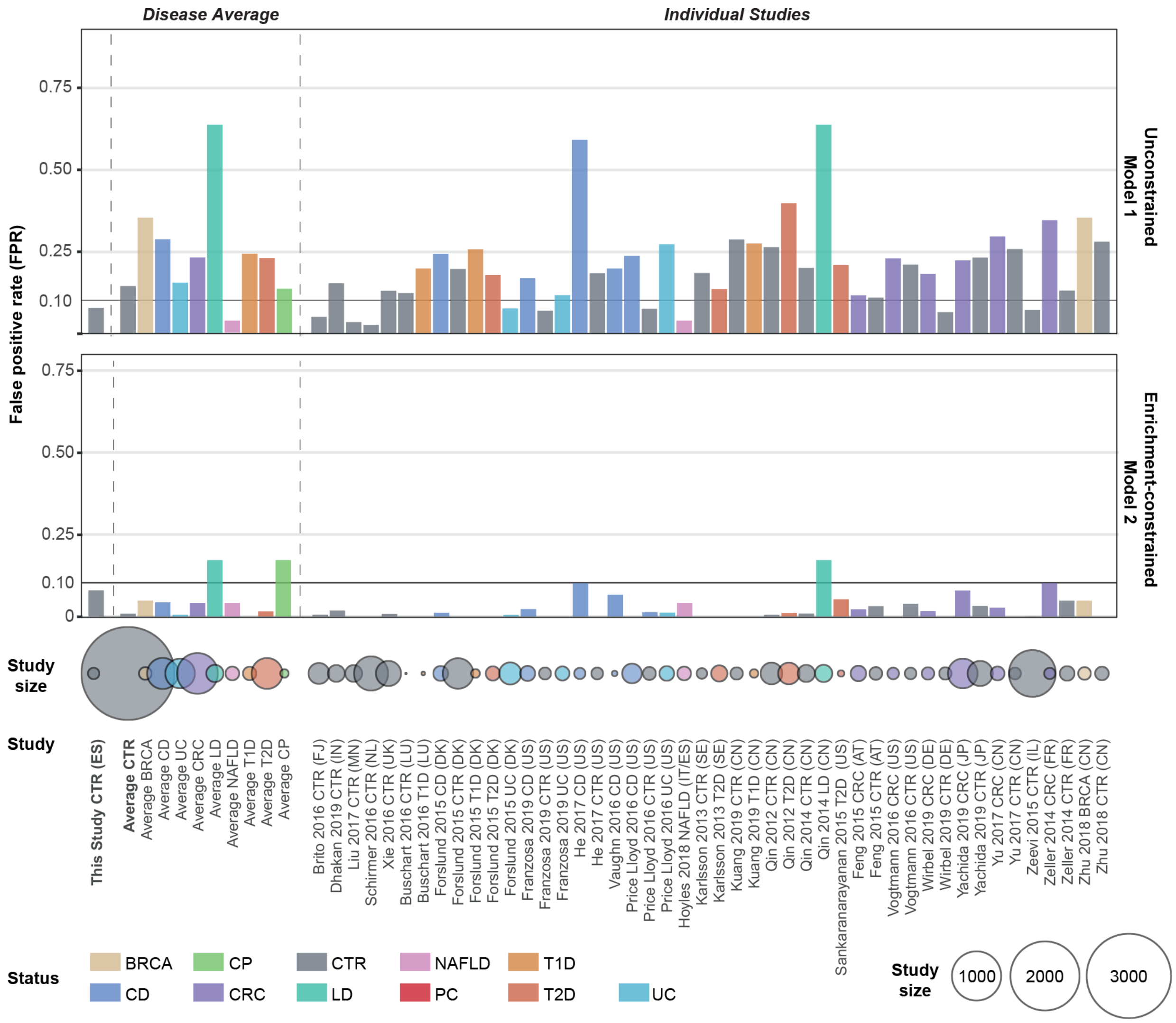


Figure 3 External validation of the disease specificity of pancreatic ductal adenocarcinoma (PDAC) faecal microbiome models. False positive rate (FPR) of metagenomic unconstrained model-1 and enrichment-constrained model-2 in 25 external test sets is shown as a bar plot (see online supplemental table S4 for a list of all studies included). Validation datasets were profiled and normalised in the same way as the initial dataset (see 'Methods'). Each study was stratified according to health status and models were tested to predict in the given group at a 90% specificity cut-off. A low FPR on metagenomes from patients with other disorders and healthy individuals indicates that the model is specific to PDAC. The number of subjects in each group is displayed as colour coded circles below. BRCA, breast cancer; CRC, colorectal cancer; CD, Crohn's disease; CP, chronic pancreatitis;, CTR, controls; LD, liver disease; NAFLD, non-alcoholic fatty liver disease; PC, pancreatic cancer; T1D, type 1 diabetes; T2D, type 2 diabetes; UC, ulcerative colitis; ES, Spanish; DE, German.

analysis workflow is detailed in online supplemental figure 1.

As several PDAC risk factors, such as tobacco smoking, alcohol consumption, obesity or diabetes, are themselves associated with microbiome composition[84], we first sought to establish potential confounders of microbiome signatures in our study population, in order to adjust analyses accordingly. For a total of 26 demographic and clinical variables, we quantified marginal effects on microbiome community-level diversity (online supplemental table S4). Faecal and salivary microbiome richness (as a proxy for alpha diversity) were not univariately associated with any tested variable, or with PDAC status, when accounting for the most common PDAC risk factors and applying a false discovery rate threshold of 0.05 (online supplemental figure 2, online supplemental table S4).

Microbiome community composition, in contrast, varied with age at diagnosis (PERMANOVA on between-sample Bray-Curtis dissimilarities, R2=0.01, Benjamini-Hochberg-corrected p=0.03), diabetes (R2=0.01, p=0.04) and jaundice status (R2=0.02, p=0.009) in faeces, and with aspirin/paracetamol use (R2=0.02, p=0.04) in saliva, albeit at very low effect sizes (online supplemental table S5). Even though cases and controls were matched for age and sex, we included these factors as strata for subsequent analyses. Under such adjustment, subject disease status was mildly but statistically significantly associated with community composition in faeces (R2=0.02, p=0.001), but not in saliva (R2=0.01, p=0.5) (figure 1B, online supplemental figure 3–4, online supplemental table S5). Indeed, the faecal microbiome composition of patients with PDAC differed from that of both controls (R2=0.02, p≤0.0001) and patients with CP

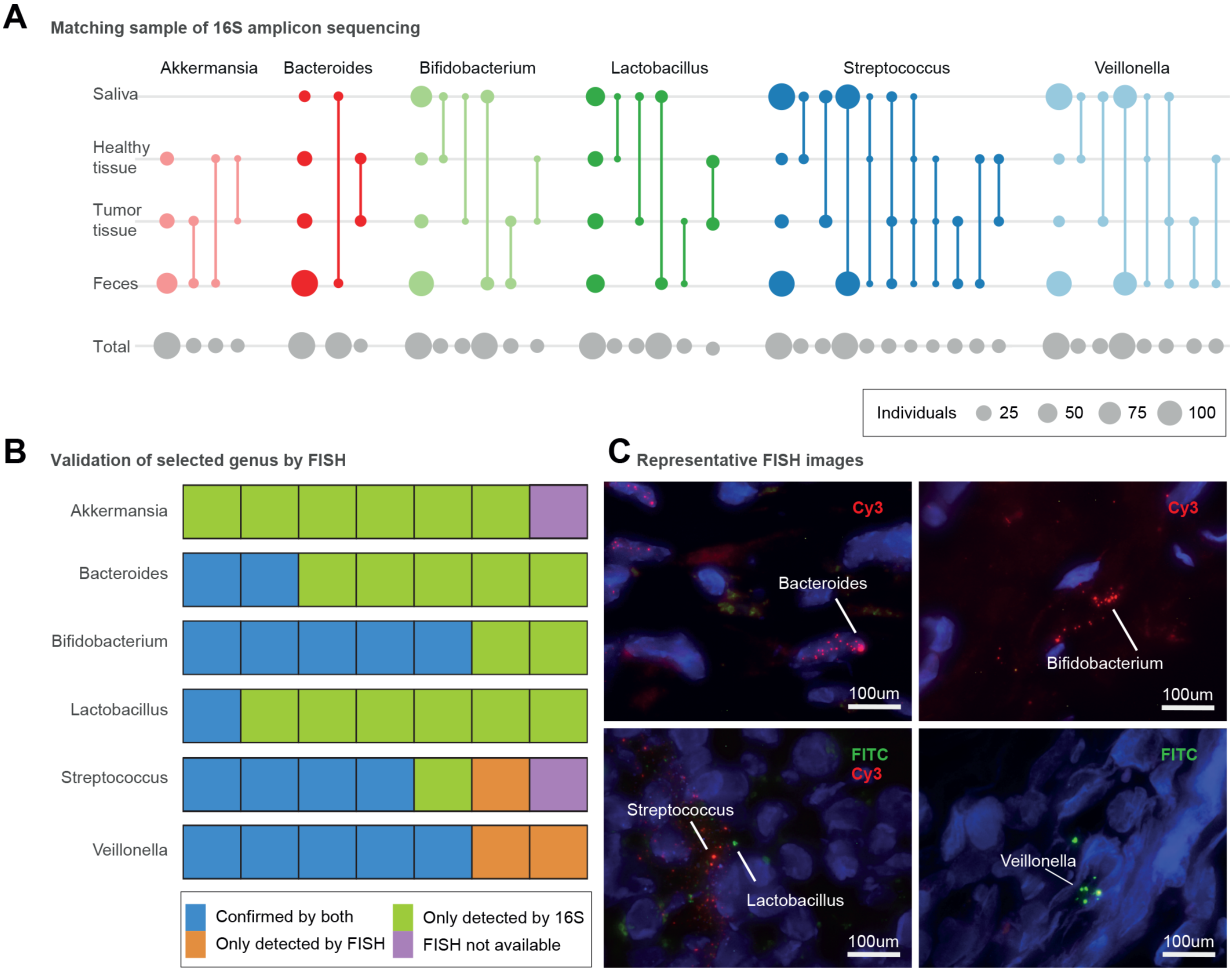


Figure 4 Presence of microbiomes in different sections of the pancreas with different conditions. (A) Presence of different genera in four different body sites including faecal, saliva, pancreatic tumour and healthy tissue samples, as inferred by 16S amplicon data. Circle size corresponds to the total number of subjects available for each comparison (grey, bottom row) or with intra-individually matched amplicon sequence variants (coloured); matched sample types are connected by lines. The first column shows the total number of samples per site in which the genus was detected. (B) Seven selected pancreatic tissue samples (five tumour and two non-tumour) to show bacterial presence/absence with both 16S amplicon and fluorescence *in situ* hybridisation (FISH) methods. Validation of bacterial presence with both 16S amplicon sequencing and FISH is shown in blue. Samples showing bacterial presence according to 16S only are displayed in green. Bacterial presence validated only by FISH is shown in orange, and samples not subjected to FISH validation owing to lack of tissue material are shown in purple. (C) Representative microscopy images for *Bacteroides* (intranuclear, tumour tissue), *Bifidobacterium* (extranuclear, tumour tissue), *Lactobacillus* (extranuclear, non-tumour tissue), *Streptococcus* (extranuclear, non-tumour tissue), *Veillonella* (extranuclear, tumour tissue). Fluorescein isothiocyanate (FITC) and Cy3 fluorescent dyes were used as indicated, and DAPI (4',6,-diamidino-2-phenylindole; blue) was used to label the nucleus.

(R2=0.02, p=0.003), although likewise at very small effect sizes.

High-accuracy metagenomic classifiers capture specific faecal microbiome signatures in patients with PDAC

Having established the presence of a gut microbiome signal for PDAC at the coarse level of overall community composition, we next identified nine species with disease-specific univariate associations (Wilcoxon test of relative abundances in PDAC cases vs controls, Benjamini-Hochberg-corrected $p<0.05$; see figure 1c). Most prominently, *Veillonella atypica*, *Fusobacterium nucleatum/hwasookii* and *Alloscardovia omnicolens* were enriched in faeces of patients with PDAC, whereas *Romboutsia timonensis, Faecalibacterium prausnitzii, Bacteroides coprocola* and *Bifidobacterium bifidum* species clusters were depleted. In contrast, we did not detect any species with significantly differential abundance in the salivary microbiome when correcting for multiple tests, including previously reported associations, such as *Porphyromonas gingivalis, Aggregatibacter actinomycetemcomitans,*[22] *Neisseria elongata* or *Streptococcus mitis*[18] (online supplemental figure 5).

Among the univariately associated faecal species, several were by themselves moderately predictive of PDAC state (online supplemental figure 5). To coalesce such individual signals into an overarching model, we next built multispecies metagenomic classifiers by fitting LASSO logistic regression models in 10-fold cross-validation (see 'Methods'). When applying no further constraints, the obtained model discriminated between patients with PDAC and controls with high accuracy in our study population ('model-1'; AUROC=0.84; Figure 2). The most prominent positive marker species in the model were *Methanobrevibacter smithii*, *Alloscardovia omnicolens*, *Veillonella atypica* and *Bacteroides finegoldii*. We note that by design, LASSO regression selects representative features among inter-correlated sets;

therefore, these species may be representatives of larger species sets with highly correlated abundances. None of the 26 demographic and epidemiological variables describing our study population were selected as predictive features by the model, and the microbiome signature was more informative than any other feature (see online supplemental figure 6 and 7). Further, none of these variables were individually associated with the microbial species represented in the model, ruling them out as potential confounders. This indicates that the classifier captured a diagnostic gut microbiome signature of PDAC that is probably independent of other disease risk factors and potential confounders.

An analogous model built to differentiate patients with CP from controls had no predictive power (AUROC=0.5; online supplemental figure 8), consistent with the observation that these groups were compositionally largely indistinguishable. Similarly, no robust PDAC signature was detected for the salivary microbiome (AUROC=0.48; online supplemental figure 9). However, a faecal model to distinguish patients with PDAC from those with CP performed better with an AUC of 0.75, but model robustness was limited by the low sample size in the group with CP (online supplemental figure 8). We further explored predictive associations at the higher resolution of functional microbiome profiles. Models based on the abundances of KEGG modules (online supplemental figure 10) achieved an accuracy of up to AUROC=0.74, but feature selection was likewise not robust across validation folds, as a consequence of fitting a high number of variables (modules) against a limited set of samples. We therefore pursued the species-based classifiers, as they provided stable models.

The initial gut microbiome-based classifier included several species depleted in PDAC relative to controls, such as *Faecalibacterium prausnitzii, Bacteroides coprocola, Bifidobacterium bifidum* or *Romboutsia timonensis* (figure 2B). For some of these species, it was previously suggested that depletion is linked to intestinal inflammation, in general, rather than to specific diseases.[85] We therefore retrained a classifier with the constraint that positively associated (enriched) microbial features were exclusively selected in each cross-validation fold. The resulting enrichment-constrained model (model-2) discerned patients with PDAC with an accuracy of AUROC=0.71. The difference with the unconstrained model, model-1, was mostly attributable to a penalty on sensitivity—that is, a decrease in confident detections of patients with PDAC, in line with expectations when training on sparse data.

Combination of metagenomic classifiers with antigen CA19-9 levels increases accuracy

Blood serum levels of the antigen CA19-9 are routinely used to monitor PDAC progress,[86 87] but have also been suggested as a potential marker for early diagnosis of PDAC, although with moderate reported sensitivity (0.80, 95%CI 0.72 to 0.86) and specificity (0.75, 95%CI 0.68 to 0.80).[12] CA19-9 serum levels were available for a subset of 77 individuals (33/50 controls and 44/57 patients with PDAC) in our Spanish population (online supplemental figure S11). Given that CA19-9 is directly secreted by tumours, we hypothesised that the readouts provided by CA19-9 serum levels and by our microbiome classifiers were complementary, and that their combination could improve the accuracy of PDAC prediction. Indeed, accounting for CA19-9 increased the accuracy of our unconstrained model-1 from AUROC=0.84 to 0.94, driven mostly by an increase in sensitivity (figure 2B). More strikingly, when we amended the enrichment-constrained model-2 with CA19-9 information, we observed a large increase in accuracy from AUC=0.71 to 0.89, likewise driven by a significant improvement in sensitivity, thereby essentially abolishing the performance penalty relative to model-1 (figure 2C, online supplemental figure S11). There was no significant bias towards higher CA19-9 levels in later disease stages in either the ES or DE populations (online supplemental figure S11).

Our Spanish study population included 25 patients with PDAC in early disease stages (T1, T2) and 32 subjects in later stages (T3, T4). Disease stage did not affect the performance of either microbiome-based model (figure 2D); in particular, recall was not biased towards later stages.

Performance of metagenome-based classifiers generalises to independent validation cohorts

To test whether the observed microbiome signatures generalise beyond our focal Spanish study population, we next challenged our models in two validation scenarios. First, we tested prediction accuracy in an independent study population of 44 patients with PDAC and 32 matched controls, recruited from two hospitals in Erlangen and Frankfurt am Main, Germany (see figure 1, Methods and online supplemental table S3), with the samples being processed identically to those of the Spanish population. On this DE validation population, both the unconstrained model-1 (figure 2B) and the enrichment-constrained model-2 (figure 2C) performed with comparable or indeed superior accuracies to the training population, both with and without complementation by CA19-9 levels, and with similar trends across disease stages (figure 2D).

Next, to confirm that our metagenomic classifiers captured PDAC-specific signatures, rather than unspecific, more general disease-associated variation, we further validated them against independent, external metagenomic datasets on various health conditions. In total, we classified 5792 publicly available gut metagenomes from 25 studies across 18 countries, including subjects with CP (this study), type 1 or type 2 diabetes, colorectal cancer, breast cancer, liver diseases, non-alcoholic fatty liver disease, including Crohn's disease and ulcerative colitis, as well as healthy controls (figure 3 and online supplemental table S6).

When tuned to 90% specificity (allowing for 10% false positive predictions) in our focal ES study population, the unconstrained model-1 showed a recall of 56% of patients with PDAC in the ES population and 48% in the DE validation population (with 6% false-positive rate), and up to 64% when complemented with information on CA19-9 levels (available for 8/32 controls and 43/44 patients with cases in the DE cohort). The disease specificity of model-1, however, was limited, with predictions of PDAC state for 15% of control subjects on average across all external datasets. Most of these false positive calls were observed in two Chinese populations of patients with Crohn's disease[48] or liver cirrhosis.[44] Crohn's disease has been associated with depletion signatures similar to those observed in our model (in particular of *F. prausnitzii*,[88]) whereas liver diseases share some physiological characteristics with impaired pancreas function. However, all other liver disease and Crohn's disease sets showed lower false detection rates, indicating that the effect was probably attributable, in part, to technical and demographic effects between studies. Indeed, we note that subjects in these two Chinese study populations were significantly younger than our populations (50±11 years for Qin_2014; 28.5±8 years for He_2017; 70±12y ears for our ES population). This age effect was systematic: across all validation sets, PDAC prediction scores were associated with subject age (ANOVA p=0.007; $\rho_{Spearman} = 0.16$), as well as with the sex of the subject ($p<10^{-6;}$) and sequencing depth (p=0.0008; $\rho_{Spearman} = 0.1$) (online supplemental figure S12, online supplemental table S6).

The enrichment-constrained model-2 showed lower detection rates in patients with PDAC in both populations, although recall was reinstated for CA19-9 combined models. Model-2 was highly specific for PDAC with, on average, just 0–5% PDAC predictions in almost all external populations, at a maximum of 17% predictions among the aforementioned[44] population with liver disease. In particular, the detected microbiome signatures were also robust against misclassification of patients with type 2 diabetes (<2% false-positive rate); this is relevant to potential screening applications, as these patients are a major PDAC risk group (figure 3).

PDAC harbours characteristic bacteria, consistent with oral and gut microbiome communities

Alterations in pancreatic secretion, as a consequence of tumour growth in the pancreatic duct, can affect digestive function and may thus plausibly underlie characteristic gut microbiome signatures, such as those described above. This would imply that PDAC progression can indirectly cause microbiome shifts (ie, reverse causation). In addition, the pancreatic duct directly communicates with the duodenum, providing an anatomical link for bacteria[25 30 89] and fungi[34] to colonise the pancreas and contribute to carcinogenesis.[31]

We therefore hypothesised that several gut microbial taxa associated with PDAC should be detectable in pancreatic tumours. We taxonomically profiled all faecal and salivary samples, as well as biopsies of tumours (n=23) and adjacent healthy pancreatic tissue (n=20) of patients with PDAC from our study population using 16S rRNA amplicon sequencing, applying strict filters to exclude putative reagent contaminants often seen in samples of low bacterial biomass[33 90] (see 'Methods'). We observed a surprisingly rich and diverse pancreas microbiome, with at least 13 bacterial genera present in ≥25% of samples, prominently including taxa with characteristic PDAC signatures in the faecal microbiome[91] (figure 4A, online supplemental figure 13). Among these, *Lactobacillus* spp, *Akkermansia muciniphila* and *Bacteroides* spp were enriched in tumours relative to non-tumour pancreatic tissue (Wilcoxon test, false discovery rate-corrected $p<0.006$).

In a subset of five tumour and two non-tumoral pancreatic tissue samples, we could further verify the prevalence of *Akkermansia* spp, *Lactobacillus* spp, *Bifidobacterium* spp, *Veillonella* spp, *Bacteroides* spp and *Streptococcus* spp using FISH assays with genus-specific primers (online supplemental figure 4, online supplemental table S7). Generally, amplicon and FISH data were concordant, though amplicon-based detection appeared more sensitive probably due to the amount of tissue analysed. Intriguingly, however, *Akkermansia* spp, although observed by amplicon sequencing in 26/30 subjects, were not detectable using FISH in any of the tested samples (figure 4B–C, online supplemental figure 14).

Links between oral, intestinal and pancreatic microbiomes

We next traced exact amplicon sequence variants (ASVs) across salivary, faecal, tumour and healthy tissue samples within subjects (figure 4A), at the highest taxonomic resolution attainable using 16S rRNA data. *Veillonella* spp, characteristically enriched in stool of patients with PDAC, were highly prevalent in both salivary (100% of subjects) and faecal (87.5%) samples across the entire study population, while oral and faecal types also matched tumour and non-tumour tissue ASVs. Interestingly, we found no intraindividual match in *Veillonella* ASVs between tumour and adjacent tissue samples, indicating that tumor-dwelling *Veillonella* spp may be distinct from those in healthy tissue. In addition, our data confirm previous reports that *Lactobacillus* spp[26] and *Bifidobacterium* spp[25] are present in both PDAC tumour and non-tumour tissue. For both genera, we found that tumour types corresponded to either oral or faecal ASVs, but not both, whereas no ASVs from healthy tissue were matched with faecal samples, indicating that distinct pancreatic subpopulations may be linked to the mouth and the gut.

Using paired salivary and faecal shotgun metagenomes, we further confirmed that strains of faecal PDAC-associated microbes may be sourced from the oral cavity (online supplemental results).

DISCUSSION

Early detection of PDAC remains a formidable challenge, at the heart of ongoing efforts to mitigate the burden of this cancer. Currently, the sole FDA-approved biomarker for PDAC is serum CA19-9, mostly used for disease monitoring rather than screening, due to inherent limits of sensitivity and specificity: CA19-9 levels can be elevated in several conditions unrelated to pancreatic cancer, while subjects lacking the Lewis-A antigen do not produce CA19-9 at all.[10–12] Small-scale studies have proposed PDAC markers based on pancreatic tissue,[5] urine[6 7] and blood serum[8 9] with limited applicability. Yet there are currently no screening tools for PDAC in the clinic—in particular, for early disease stages.

In a prospectively recruited study population of newly diagnosed, treatment-naïve patients and matched controls for whom oral, faecal and tissue microbiomes were analysed (figure 1A), we developed metagenomic classifiers that robustly and accurately predict PDAC solely based on characteristic faecal microbial species (figure 2). PDAC signatures captured by our multispecies models were orthogonal to well-established PDAC risk factors (figures 1B and 2A). This suggests that, in practice, the faecal microbiome may be used to screen for PDAC, complementary to other testable markers, with added diagnostic accuracy in combined tests, as has been proposed for colorectal cancer.[39] Indeed, a combination of our microbiome classifiers with CA19-9 data, available for a subset of our population, significantly enhanced the accuracy of PDAC detection (figure 2B–D).

Previous studies have explored links between PDAC and the oral[18–22 26 92 93] or faecal[23 24] microbiome at the limited taxonomic resolution of 16S rRNA sequencing, but provided conflicting reports regarding the association patterns of individual taxa, probably due to heterogeneous experimental and analytical approaches. The non-availability of raw sequence and patient-level clinical data for several PDAC datasets has made comparisons between studies challenging, and thus a consensus on PDAC-associated microbiome signatures has so far failed to emerge. Several previously reported univariate PDAC associations of oral taxa including *P. gingivalis, A. actinomycetemcomitans*, *S. thermophilus* and *Fusobacterium* spp were not confirmed in our study population (online supplemental figure 4); we generally did not observe any salivary PDAC signature either for individual species or for multispecies models.

We carefully checked our analyses for demographic, lifestyle, and clinical confounders, as these can show stronger microbiome associations than disease states.[84] We moreover validated our metagenomic classifiers against the independently sampled, yet consistently processed, DE population (figure 2B–D) and against external populations of various health states from 25 different studies (n=5792)[35–59] (figure 3). Both confounder control and external validation are essential when assessing the disease specificity of predictive models, in particular for diseases like PDAC with low incidence in the general population. This was confirmed in our analyses: among our two metagenomic classifiers, model-1 showed a high accuracy of AUROC=0.84 in our ES study population, driven by a high recall of patients with PDAC. However, model-1 showed only limited disease specificity in external validations, capturing non-specific species depletion signals discriminative between cases and controls in our population, but also shared by subjects with other diseases. These included generic inflammation signatures—for example, a depletion of *F. prausnitzii, E. rectale* or *B. bifidum*. Published metagenomic classifiers for various diseases, and in particular previously reported signatures for PDAC, share similar limitations: highly tuned accuracy on the focal population, but non-specific features shared with other diseases. This lack of specificity limits their translation into clinical practice. In contrast, our model-2, constrained to PDAC-enriched features, achieved only moderate accuracy within our populations (AUC=0.71 on ES, AUC=0.85 on DE) due to a penalty on sensitivity, but was highly PDAC-specific with very low false prediction rates in external populations, including known PDAC risk groups such as those with type 2 diabetes. In particular, PDAC-enriched features in both model-1 and model-2 showed little overlap with characteristic faecal microbiome features for other cancer types, such as colorectal cancer, indicating that a combination of our microbiome models with CA19-9 levels (highly sensitive, but not specific to PDAC) is promising. We note that the residual false positive rate among external populations may partly be due to technical heterogeneity, as all external populations were sampled and processed using independent protocols, and that univariate PDAC associations of individual species may be informative, but not disease-specific (Supplementary Discussion). The panel of PDAC-enriched species in model-2 thus shows potential for microbiome-based PDAC screening, given that a combination with complementary information on serum CA19-9 significantly increased accuracy (AUC=0.89 and 0.92).

Our models showed comparable performance across PDAC disease stages, with no bias towards later stages (figure 2B–D). This indicates that characteristic microbiome signatures emerge early during progression of the disease and that the faecal microbiome can serve for the early detection of PDAC.

Our data are strictly observational and cross-sectional. Nevertheless, there are strong indications that the identified faecal microbiome shifts are not merely a consequence of impaired pancreatic function or systemic effects thereof, although indirect effects cannot be ruled out. Several taxa could be traced between the gut and pancreas, with univariate enrichment in tumours relative to adjacent healthy tissue, indicating direct associations of PDAC with the gut microbiome. We confirmed previous observations[25 30 31 89 91] that the human pancreas harbours a microbiome, both by amplicon sequencing, and by FISH for the most comprehensive panel of taxa to date (figure 4). Pancreatic tissue and tumours contain only low bacterial biomass and are therefore prone to contamination in 16S rRNA amplicon data[33], whereas FISH testing requires specific hypotheses, so a comprehensive cataloguing of the healthy and diseased pancreatic microbiome composition is still emerging. In our study, we carefully filtered our dataset against known kit contaminants and confirmed the presence of various key genera using FISH assays. We moreover observed an intraindividual overlap of exact amplicon sequence variants between oral, faecal and tissue samples, confirming a shared presence across multiple sites for several species at the highest attainable taxonomic resolution for amplicon data.

Faecal populations of characteristic PDAC-associated taxa could thus be traced back to pancreatic tumours. Similarly, we observed significantly increased levels of oral-intestinal strain transmission in patients with PDAC, in particular of PDAC signature taxa, indicating that these may be sourced intraindividually, from the oral cavity (online supplemental results). These findings suggest that the oral, intestinal and pancreatic microbiomes may be intricately linked, and that multibody site study designs such as presented here will be necessary to disentangle their respective roles and interactions in PDAC aetiology.

In summary, the described faecal microbiome signatures enabled robust metagenomic classifiers for PDAC detection at high disease specificity, complementary to existing markers, and with potential towards cost-effective PDAC screening and monitoring. Furthermore, in view of previous reports on microbe-mediated pancreatic carcinogenesis in murine models and humans,[25 30 94] we believe that the presented panel of PDAC-associated bacterial species may be relevant beyond their use for diagnosis, providing promising future entry points for disease prevention and therapeutic intervention.

Author affiliations
[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[2]Collaboration for joint PhD degree, European Molecular Biology Laboratory and Heidelberg University, Heidelberg, Germany
[3]Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[4]Centro de Investigación Biomédica en Red de Oncología (CIBERONC), Madrid, Spain
[5]Molecular Cytogenetics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[6]Member of the German Center for Lung Research (DZL) and the Universities of Giessen and Marburg Lung School (UGMLC), Philipps University Marburg Faculty of Medicine, Marburg, Germany
[7]Medical Oncology Department of Oncology, Hospital Ramón y Cajal, Madrid, Spain
[8]University of Alcala de Henares, Alcala de Henares, Spain
[9]Translational Hepatology Department of Internal Medicine I, Goethe-Universitat Frankfurt am Main, Frankfurt am Main, Germany
[10]Frankfurt Cancer Institute, Goethe University Frankfurt, Frankfurt am Main, Hessen, Germany
[11]Genomic Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany
[12]Department of Surgery, Erlangen University Hospital, Erlangen, Germany
[13]Department of Surgery, University of Greifswald, Greifswald, Germany
[14]Hospital Universitari Vall d'Hebron, Institut de Recerca (VHIR), Barcelona, Spain
[15]Universitat Autònoma de Barcelona, Barcelona, Spain
[16]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Madrid, Spain
[17]EF Clif, European Foundation for the Study of Chronic Liver Failure, Barcelona, Spain
[18]Epithelial Carcinogenesis Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[19]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain
[20]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany
[21]Yonsei Frontier Lab (YFL), Yonsei University, Seoul, South Korea
[22]Max Delbrück Centre for Molecular Medicine, Berlin, Germany

Twitter Ece Kartal @ps_ecekartal, Thomas S B Schmidt @TSBSchm, Oleksandr M Maistrenko @o__maistrenko, Georg Zeller @ZellerGroup, Jonel Trebicka @JonelTrebicka, Nuria Malats @nmalats and Peer Bork @BorkLab

Acknowledgements We thank members of the Bork, Malats and Zeller groups for inspiring discussions and all contributions. Additionally, we thank the EMBL Genomics Core Facility for sequencing support.

Collaborators PanGenEU Study Investigators. Spanish National Cancer Research Centre (CNIO), Madrid, Spain: Núria Malats, Francisco X Real, Evangelina López de Maturana, Paulina Gómez-Rubio, Esther Molina-Montes, Lola Alonso, Mirari Márquez, Roger Milne, Ana Alfaro, Tania Lobato, Lidia Estudillo. Verona University, Italy: Rita Lawlor, Aldo Scarpa, Stefania Beghelli. National Cancer Registry Ireland, Cork, Ireland: Linda Sharp, Damian O'Driscoll. Hospital Madrid-Norte-Sanchinarro, Madrid, Spain: Manuel Hidalgo, Jesús Rodríguez Pascual. Hospital Ramon y Cajal, Madrid, Spain: Alfredo Carrato, Alejandra Caminoa, Carmen Guillén-Ponce, Mercedes Rodríguez-Garrote, Federico Longo-Muñoz, Reyes Ferreiro, Vanessa Pachón, M Ángeles Vaz. Hospital del Mar, Barcelona, Spain: Mar Iglesias, Lucas Ilzarbe, Cristina Álvarez-Urturi, Xavier Bessa, Felipe Bory, Lucía Márquez, Ignasi Poves, Fernando Burdío, Luis Grande, Javier Gimeno. Hospital Vall dHebron, Barcelona, Spain: Xavier Molero, Luisa Guarner, Joaquin Balcells, Mayte Salcedo. Technical University of Munich, Germany: Christoph Michalski, Irene Esposito, Jörg Kleeff, Bo Kong. Karolinska Institute, Stockholm, Sweden: Matthias Löhr, Jiaqui Huang, Caroline Verbeke, Weimin Ye, Jingru Yu. Hospital 12 de Octubre, Madrid, Spain: José Perea, Pablo Peláez. Hospital de la Santa Creu i Sant Pau, Barcelona, Spain: Antoni Farré, Josefina Mora, Marta Martín, Vicenç Artigas, Carlos Guarner, Francesc J Sancho, Mar Concepción, Teresa Ramón y Cajal. The Royal Liverpool University Hospital, UK: William Greenhalf, Eithne Costello. Queen's University Belfast, UK: Michael O'Rorke, Liam Murray, Marie Cantwell. Laboratorio de Genética Molecular, Hospital General Universitario de Elche, Spain: Víctor M Barberá, Javier Gallego. Instituto Universitario de Oncología del Principado de Asturias, Oviedo, Spain: Adonina Tardón, Luis Barneo. Hospital Clínico Universitario de Santiago de Compostela, Spain: Enrique Domínguez Muñoz, Antonio Lozano, Maria Luaces. Hospital Clínico Universitario de Salamanca, Spain: Luís Muñoz-Bellvís, J.M. Sayagués Manzano, M.L. Gutiérrez Troncoso, A. Orfao de Matos. University of Marburg, Department of Gastroenterology, Phillips University of Marburg, Germany: Thomas Gress, Malte Buchholz, Albrecht Neesse. Queen Mary University of London, UK: Tatjana Crnogorac-Jurcevic, Hemant M Kocher, Satyajit Bhattacharya, Ajit T Abraham, Darren Ennis, Thomas Dowe, Tomasz Radon. Scientific advisors of the PanGenEU Study: Debra T Silverman (NCI, USA) and Douglas Easton (U. of Cambridge, UK).MAGIC (MicrobiotA-focused German Interdisciplinary Collaboration) Study Investigators. Section for Translational Hepatology, Department of Internal Medicine I, Frankfurt Cancer Institute, Goethe University Frankfurt: Jonel Trebicka, Hans-Peter Erasmus, Fabian Finkelmeier, Robert Schierwagen, Wenyi Gu, Olaf Tyc, Frank Erhard Uschner, Stefan Zeuzem. Department of Surgery, University Greifswald: Stephan Kersting, Melanie Langheinrich. Department of Surgery, University Erlangen: Robert Grützmann, Georg F. Weber, Christian Pilarsky. Department of Internal Medicine, University Erlangen: Stefan Wirtz.

Contributors EK designed the study, conducted experimental work, acquired and analysed data, wrote the first manuscript draft and the revised manuscript.TSBS designed the study, acquired and analysed data, wrote the first manuscript draft and the revised manuscript.EM-M designed the study, contributed to patient recruitment and the collection of biomaterials and clinical data, acquired and analysed data, and wrote the first manuscript draft.SR-P contributed to patient recruitment and the collection of biomaterials and clinical data and conducted experimental work. JW, OMM, WAA, BAA, AC, HP-E, FF, PG-R, SKe, ML, MM, XM, RT-R, JT contributed to patient recruitment and the collection of biomaterials and clinical data. RJA, AF, AMG, KZ contributed to data analysis. LE contributed to patient recruitment and the collection of biomaterials and clinical data and conducted experimental work. RH, FJ, SKa, AT conducted experimental work and acquired data. AO, TvR contributed to data analysis. MSI, PSI contributed to patient recruitment. VB acquired data. GZ designed the study and contributed to data analysis. FXR designed the study and contributed to data analysis and wrote the first manuscript draft. NM conceived the study, designed the study, contributed to patient recruitment and the collection of biomaterials and clinical data and wrote the first manuscript draft. PB conceived of the study, designed the study, contributed to data analysis and wrote the first manuscript draft. All authors reviewed, edited and approved the final version of the manuscript.

Funding We acknowledge funding from EMBL, CNIO, World Cancer Research (#15–0391), the European Research Council (ERC-AdG-669830 MicrobioS), the BMBF-funded Heidelberg CenterCentre for Human Bioinformatics (HD-HuB) within the German Network for Bioinformatics Infrastructure (de.NBI #031A537B), Fondo de Investigaciones Sanitarias (FIS), Instituto de Salud Carlos III-FEDER, Spain (grant numbers PI15/01573, PI18/01347, FIS PI17/02303); Red Temática de Investigación Cooperativa en Cáncer, Spain (grant numbers RD12/0036/0034, RD12/0036/0050, RD12/0036/0073); III beca Carmen Delgado/Miguel Pérez-Mateo de AESPANC-ACANPAN; EU-6FP Integrated Project (#018771-MOLDIAG-PACA); EU-FP7-HEALTH (#259737-CANCERALIA). Funders had no involvement in the study design, patient enrolment, analysis, manuscript writing or reviewing.

Competing interests EK, TSBS, JW, OMM, EM-M, GZ, LE, SR-P, FXR, NM and PB have a pending patent application (application number: EP21382876.7) for early detection of pancreatic cancer based on microbial biomarkers. The other authors declare no conflicts of interest.

Patient consent for publication Not applicable.

Ethics approval Participants were prospectively recruited from the Hospital Ramón y Cajal in Madrid and Hospital Vall dHebron in Barcelona, Spain. Institutional review board ethical approval (CEI PI 26 2015-v7) and written informed consent was obtained from participating centres and study participants, respectively. An independent validation population was recruited at the Department of Surgery, University Hospital of Erlangen (32 PDAC and 32 control samples) and Section for Translational Hepatology, Department of Internal Medicine I, Goethe University Clinic Frankfurt (12 PDAC samples). The study was approved by the local ethics committees (SGI-3-2019, 451_18 B). Clinical data, including disease stage and follow-up data, were retrieved from the clinical records of the hospital charts of the respective patients.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data relevant to the study are included in the article or uploaded as supplementary information. The raw sequencing data for the samples are made available in the European Nucleotide Archive (ENA) under the study identifiers PRJEB38625 and PRJEB42013. Metadata for these samples are available as Supplementary Tables S1 and S2. Filtered taxonomic and functional profiles used as input for the statistical modelling pipeline are available in Supplementary Data S1 and S2. Analysis code and results available under https://github.com/psecekartal/PDAC.git.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Ece Kartal http://orcid.org/0000-0002-7720-455X
Thomas S B Schmidt http://orcid.org/0000-0001-8587-4177
Esther Molina-Montes http://orcid.org/0000-0002-0428-2426
Sandra Rodríguez-Perales http://orcid.org/0000-0001-7221-3636
Jakob Wirbel http://orcid.org/0000-0002-4073-3562
Oleksandr M Maistrenko http://orcid.org/0000-0003-1961-7548
Wasiu A Akanni http://orcid.org/0000-0002-2075-2387
Bilal Alashkar Alhamwe http://orcid.org/0000-0002-7120-0013
Renato J Alves http://orcid.org/0000-0002-7212-0234
Alfredo Carrato http://orcid.org/0000-0001-7749-8140
Lidia Estudillo http://orcid.org/0000-0003-3891-3713
Anthony Fullam http://orcid.org/0000-0002-0884-8124
Ferris Jung http://orcid.org/0000-0002-5534-7832
Stefanie Kandels http://orcid.org/0000-0002-4194-4927
Stephan Kersting http://orcid.org/0000-0002-2124-3103
Melanie Langheinrich http://orcid.org/0000-0002-0120-9135
Askarbek Orakov http://orcid.org/0000-0001-6823-5269
Thea Van Rossum http://orcid.org/0000-0002-3598-5001
Raul Torres-Ruiz http://orcid.org/0000-0001-9606-0398
Anja Telzerow http://orcid.org/0000-0001-9855-0809
Konrad Zych http://orcid.org/0000-0001-7426-0516
Vladimir Benes http://orcid.org/0000-0002-0352-2547
Georg Zeller http://orcid.org/0000-0003-1429-7485
Jonel Trebicka http://orcid.org/0000-0002-7028-3881
Francisco X Real http://orcid.org/0000-0001-9501-498X
Nuria Malats http://orcid.org/0000-0003-2538-3784
Peer Bork http://orcid.org/0000-0002-2627-833X

REFERENCES

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin* 2018;68:7–30. doi:10.3322/caac.21442

2 Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68:394–424. doi:10.3322/caac.21492

3 Kamisawa T, Wood LD, Itoi T, *et al*. Pancreatic cancer. *The Lancet* 2016;388:73–85. doi:10.1016/S0140-6736(16)00141-0

4 Park W, Chawla A, O'Reilly EM. Pancreatic cancer: a review. *JAMA* 2021;326. doi:10.1001/jama.2021.13027

5 Wang Y, Li Z, Zheng S, *et al*. Expression profile of long non-coding RNAs in pancreatic cancer and their clinical significance as biomarkers. *Oncotarget* 2015;6:35684–98. doi:10.18632/oncotarget.5533

6 Blyuss O, Zaikin A, Cherepanova V, *et al*. Development of PancRISK, a urine biomarker-based risk score for stratified screening of pancreatic cancer patients. *Br J Cancer* 2020;122:692–6. doi:10.1038/s41416-019-0694-0

7 Debernardi S, Massat NJ, Radon TP, *et al*. Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma. *Am J Cancer Res* 2015;5:3455–66.

8 Seifert AM, Reiche C, Heiduk M, *et al*. Detection of pancreatic ductal adenocarcinoma with galectin-9 serum levels. *Oncogene* 2020;39:3102–13. doi:10.1038/s41388-020-1186-7

9 Melo SA, Luecke LB, Kahlert C, *et al*. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* 2015;523:177–82. doi:10.1038/nature14581

10 Goonetilleke KS, Siriwardena AK. Systematic review of carbohydrate antigen (CA 19-9) as a biochemical marker in the diagnosis of pancreatic cancer. *Eur J Surg Oncol* 2007;33:266–70 http://www.sciencedirect.com/science/article/pii/S0748798306003763 doi:10.1016/j.ejso.2006.10.004

11 Gui J-C, Yan W-L, Liu X-D. CA19-9 and CA242 as tumor markers for the diagnosis of pancreatic cancer: a meta-analysis. *Clin Exp Med* 2014;14:225–33. doi:10.1007/s10238-013-0234-9

12 Xing H, Wang J, Wang Y, *et al*. Diagnostic value of CA 19-9 and carcinoembryonic antigen for pancreatic cancer: a meta-analysis. *Gastroenterol Res Pract* 2018;2018:1–9. doi:10.1155/2018/8704751

13 Hasan S, Jacob R, Manne U, *et al*. Advances in pancreatic cancer biomarkers. *Oncol Rev* 2019;13:410. doi:10.4081/oncol.2019.410

14 Qader G, Aali M, Smail SW, *et al*. Cardiac, hepatic and renal dysfunction and IL-18 polymorphism in breast, colorectal, and prostate cancer patients. *Asian Pac J Cancer Prev* 2021;22:131–7. doi:10.31557/APJCP.2021.22.1.131

15 Rawla P, Sunkara T, Gaduputi V. Epidemiology of pancreatic cancer: global trends, etiology and risk factors. *World J Oncol* 2019;10:10–27. doi:10.14740/wjon1166

16 Wood LD, Yurgelun MB, Goggins MG. Genetics of familial and sporadic pancreatic cancer. *Gastroenterology* 2019;156:2041–55. doi:10.1053/j.gastro.2018.12.039

17 Michaud DS, Lu J, Peacock-Villada AY, *et al*. Periodontal disease assessed using clinical dental measurements and cancer risk in the ARIC study. *J Natl Cancer Inst* 2018;110:843–54. doi:10.1093/jnci/djx278

18 Farrell JJ, Zhang L, Zhou H, *et al*. Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut* 2012;61:582–8. doi:10.1136/gutjnl-2011-300784

19 Michaud DS, Izard J, Wilhelm-Benartzi CS, *et al*. Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. *Gut* 2013;62:1764–70. doi:10.1136/gutjnl-2012-303006

20 Olson SH, Satagopan J, Xu Y, *et al*. The oral microbiota in patients with pancreatic cancer, patients with IPMNs, and controls: a pilot study. *Cancer Causes Control* 2017;28:959–69. doi:10.1007/s10552-017-0933-8

21 Lu H, Ren Z, Li A, *et al*. Tongue coating microbiome data distinguish patients with pancreatic head cancer from healthy controls. *J Oral Microbiol* 2019;11:1563409. doi:10.1080/20002297.2018.1563409

22 Fan X, Alekseyenko AV, Wu J, *et al*. Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut* 2018;67:120–7. doi:10.1136/gutjnl-2016-312580

23 Ren Z, Jiang J, Xie H, *et al*. Gut microbial profile analysis by MiSeq sequencing of pancreatic carcinoma patients in China. *Oncotarget* 2017;8:95176–91. doi:10.18632/oncotarget.18820

24 Half E, Keren N, Reshef L, *et al*. Fecal microbiome signatures of pancreatic cancer patients. *Sci Rep* 2019;9:16801. doi:10.1038/s41598-019-53041-4

25 Pushalkar S, Hundeyin M, Daley D, *et al*. The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discov* 2018;8:403–16. doi:10.1158/2159-8290.CD-17-1134

26 Del Castillo E, Meier R, Chung M, *et al*. The microbiomes of pancreatic and duodenum tissue overlap and are highly subject specific but differ between pancreatic cancer and noncancer subjects. *Cancer Epidemiol Biomarkers Prev* 2019;28:370–83. doi:10.1158/1055-9965.EPI-18-0542

27 Mei Q-X, Huang C-L, Luo S-Z, *et al*. Characterization of the duodenal bacterial microbiota in patients with pancreatic head cancer vs. healthy controls. *Pancreatology* 2018;18:438–45. doi:10.1016/j.pan.2018.03.005

28 Geller LT, Barzily-Rokni M, Danino T, *et al*. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* 2017;357:1156–60. doi:10.1126/science.aah5043

29 Mitsuhashi K, Nosho K, Sukawa Y, *et al*. Association of *Fusobacterium* species in pancreatic cancer tissues with molecular features and prognosis. *Oncotarget* 2015;6:7209–20. doi:10.18632/oncotarget.3109

30 Thomas RM, Gharaibeh RZ, Gauthier J, *et al*. Intestinal microbiota enhances pancreatic carcinogenesis in preclinical models. *Carcinogenesis* 2018;39:1068–78. doi:10.1093/carcin/bgy073

31 Riquelme E, Zhang Y, Zhang L, *et al*. Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell* 2019;178:795–806. doi:10.1016/j.cell.2019.07.008

32 Gaiser RA, Halimi A, Alkharaan H, *et al*. Enrichment of oral microbiota in early cystic precursors to invasive pancreatic cancer. *Gut* 2019;68:2186–94. doi:10.1136/gutjnl-2018-317458

33 Salter SJ, Cox MJ, Turek EM, *et al*. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014;12:87. doi:10.1186/s12915-014-0087-z

34 Aykut B, Pushalkar S, Chen R, *et al*. The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* 2019;574:264–7. doi:10.1038/s41586-019-1608-2

35 Heintz-Buschart A, May P, Laczny CC, *et al*. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* 2017;2:16180.

36 Dhakan DB, Maji A, Sharma AK, *et al*. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* 2019;8:giz004. doi:10.1093/gigascience/giz004

37 Feng Q, Liang S, Jia H, *et al*. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* 2015;6:6528. doi:10.1038/ncomms7528

38 Wirbel J, Pyl PT, Kartal E, *et al*. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 2019;25:679–89. doi:10.1038/s41591-019-0406-6

39 Zeller G, Tap J, Voigt AY, *et al*. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 2014;10:766. doi:10.15252/msb.20145645

40 Brito IL, Yilmaz S, Huang K, *et al*. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 2016;535:435–9. doi:10.1038/nature18927

41 Vaughn BP, Vatanen T, Allegretti JR, *et al*. Increased intestinal microbial diversity following fecal microbiota transplant for active Crohn's disease. *Inflamm Bowel Dis* 2016;22:2182–90. doi:10.1097/MIB.0000000000000893

42 Forslund K, Hildebrand F, Nielsen T, *et al*. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 2015;528:262–6. doi:10.1038/nature15766

43 Liu W, Zhang J, Wu C, *et al*. Unique features of ethnic Mongolian gut microbiome revealed by metagenomic analysis. *Sci Rep* 2016;6:34826. doi:10.1038/srep34826

44 Qin N, Yang F, Li A, *et al*. Alterations of the human gut microbiome in liver cirrhosis. *Nature* 2014;513:59–64. doi:10.1038/nature13568

45 Kuang Y-S, Lu J-H, Li S-H, *et al*. Connections between the human gut microbiome and gestational diabetes mellitus. *Gigascience* 2017;6:1–12.

46 Karlsson FH, Tremaroli V, Nookaew I, *et al*. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 2013;498:99–103. doi:10.1038/nature12198

47 Hoyles L, Fernández-Real J-M, Federici M, *et al*. Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat Med* 2018;24:1070–80. doi:10.1038/s41591-018-0061-3

48 Quing H, Gao Y, Jie Z, *et al*. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* 2017;6:gix050. doi:10.1093/gigascience/gix050

49 Franzosa EA, Sirota-Madi A, Avila-Pacheco J, *et al*. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* 2019;4:293–305. doi:10.1038/s41564-018-0306-4

50 Yu J, Feng Q, Wong SH, *et al*. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 2017;66:70–8. doi:10.1136/gutjnl-2015-309800

51 Zeevi D, Korem T, Zmora N, *et al*. Personalized nutrition by prediction of glycemic responses. *Cell* 2015;163:1079–94. doi:10.1016/j.cell.2015.11.001

52 Zhu J, Liao M, Yao Z, *et al*. Breast cancer in postmenopausal women is associated with an altered gut metagenome. *Microbiome* 2018;6:136. doi:10.1186/s40168-018-0515-3

53 Yachida S, Mizutani S, Shiroma H, *et al*. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* 2019;25:968–76. doi:10.1038/s41591-019-0458-7

54 Vogtmann E, Hua X, Zeller G, *et al*. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS One* 2016;11:e0155362. doi:10.1371/journal.pone.0155362

55 Sankaranarayanan K, Ozga AT, Warinner C, *et al*. Gut microbiome diversity among Cheyenne and Arapaho individuals from Western Oklahoma. *Curr Biol* 2015;25:3161–9. doi:10.1016/j.cub.2015.10.060

56 Qin J, Li Y, Cai Z, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55–60. doi:10.1038/nature11450

57 Lloyd-Price J, Mahurkar A, Rahnavard G, *et al*. Strains, functions and dynamics in the expanded human microbiome project. *Nature* 2017;550:61–6.

58 Schirmer M, Smeekens SP, Vlamakis H, *et al*. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* 2016;167:1125–36. doi:10.1016/j.cell.2016.10.020

59 Xie H, Guo R, Zhong H, *et al*. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst* 2016;3:572–84. doi:10.1016/j.cels.2016.10.004

60 Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;28:112–8. doi:10.1093/bioinformatics/btr597

61 Nadkarni MA, Martin FE, Jacques NA, *et al*. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* 2002;148:257–66. doi:10.1099/00221287-148-1-257

62 Kramski M, Gaeguta AJ, Lichtfuss GF, *et al*. Novel sensitive real-time PCR for quantification of bacterial 16S rRNA genes in plasma of HIV-infected patients as a marker for microbial translocation. *J Clin Microbiol* 2011;49:3691–3. doi:10.1128/JCM.01018-11

63 Caporaso JG, Lauber CL, Walters WA, *et al*. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 2011;108 Suppl 1:4516–22. doi:10.1073/pnas.1000080107

64 Callahan BJ, McMurdie PJ, Rosen MJ, *et al*. DADA2: high-resolution sample inference from illumina amplicon data. *Nat Methods* 2016;13:581–3. doi:10.1038/nmeth.3869

65 Matias Rodrigues JF, Schmidt TSB, Tackmann J, *et al*. MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics* 2017;33:3808–10. doi:10.1093/bioinformatics/btx517

66 Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29:2933–5. doi:10.1093/bioinformatics/btt509

67 Matias Rodrigues JF, von Mering C. HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* 2014;30:287–8. doi:10.1093/bioinformatics/btt657

68 Schmidt TSB, Matias Rodrigues JF, von Mering C. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol* 2015;17:1689–706. doi:10.1111/1462-2920.12610

69 Coelho LP, Alves R, Monteiro P, *et al*. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome* 2019;7:84.

70 Mende DR, Letunic I, Maistrenko OM, *et al*. proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res* 2020;48:D621-D625. doi:10.1093/nar/gkz1002

71 Milanese A, Mende DR, Paoli L, *et al*. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 2019;10. doi:10.1038/s41467-019-08844-4

72 Coelho LP, Alves R, Del Río Álvaro Rodríguez, del Río Á.R, *et al*. Towards the biogeography of prokaryotic genes. *Nature* 2021. doi:10.1038/s41586-021-04233-4. [Epub ahead of print: 15 Dec 2021].

73 Huerta-Cepas J, Szklarczyk D, Forslund K, *et al*. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 2016;44:D286–93. doi:10.1093/nar/gkv1248

74 Schmidt TSB, Matias Rodrigues JF, von Mering C. A family of interaction-adjusted indices of community similarity. *ISME J* 2017;11:791–807.

75 Oksanen J, Blanchet FG, Friendly M. Vegan: community ecology package, 2019. Available: https://CRAN.R-project.org/package=vegan

76 Tibshirani R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* 1996;58 http://doi.wiley.com/

77 Helleputte T. *LiblineaR: linear predictive models based on the LIBLINEAR C/C++ library, 2015. R package version*, 2015: 1–94.

78 Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77. doi:10.1186/1471-2105-12-77

79 Wirbel J, Zych K, Essex M, *et al*. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol* 2021;22:93.

80 Mende DR, Letunic I, Huerta-Cepas J, *et al*. proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res* 2017;45:D529–34. doi:10.1093/nar/gkw989

81 Costea PI, Munch R, Coelho LP, *et al*. metaSNV: a tool for metagenomic strain level analysis. *PLoS One* 2017;12:e0182392. doi:10.1371/journal.pone.0182392

82 Costea PI, Coelho LP, Sunagawa S, *et al*. Subspecies in the global human gut microbiome. *Mol Syst Biol* 2017;13:960. doi:10.15252/msb.20177589

83 Schmidt TS, Hayward MR, Coelho LP, *et al*. Extensive transmission of microbes along the gastrointestinal tract. *Elife* 2019;8. doi:10.7554/eLife.42693. [Epub ahead of print: 12 02 2019].

84 Schmidt TSB, Raes J, Bork P. The human gut microbiome: from association to modulation. *Cell* 2018;172:1198–215. doi:10.1016/j.cell.2018.02.044

85 Duvallet C, Gibbons SM, Gurry T, *et al*. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 2017;8:1784. doi:10.1038/s41467-017-01973-8

86 Azizian A, Rühlmann F, Krause T, *et al*. CA19-9 for detecting recurrence of pancreatic cancer. *Sci Rep* 2020;10:1332. doi:10.1038/s41598-020-57930-x

87 Winter JM, Yeo CJ, Brody JR. Diagnostic, prognostic, and predictive biomarkers in pancreatic cancer. *J Surg Oncol* 2013;107:15–22. doi:10.1002/jso.23192

88 Cao Y, Shen J, Ran ZH. Association between Faecalibacterium prausnitzii reduction and inflammatory bowel disease: a meta-analysis and systematic review of the literature. *Gastroenterol Res Pract* 2014;2014:1–7. doi:10.1155/2014/872725

89 Poore GD, Kopylova E, Zhu Q, *et al*. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 2020;579:567–74. doi:10.1038/s41586-020-2095-1

90 de Goffau MC, Lager S, Sovio U, *et al*. Human placenta has no microbiome but can contain potential pathogens. *Nature* 2019;572:329–34. doi:10.1038/s41586-019-1451-5

91 Nejman D, Livyatan I, Fuks G, *et al*. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* 2020;368:973–80. doi:10.1126/science.aay9189

92 Torres PJ, Fletcher EM, Gibbons SM, *et al*. Characterization of the salivary microbiome in patients with pancreatic cancer. *PeerJ* 2015;3:e1373. doi:10.7717/peerj.1373

93 Vogtmann E, Han Y, Caporaso JG, *et al*. Oral microbial community composition is associated with pancreatic cancer: a case-control study in Iran. *Cancer Med* 2020;9:797–806. doi:10.1002/cam4.2660

94 Sethi V, Kurtom S, Tarique M, *et al*. Gut microbiota promotes tumor growth in mice by modulating immune response. *Gastroenterology* 2018;155:33–7. doi:10.1053/j.gastro.2018.04.001

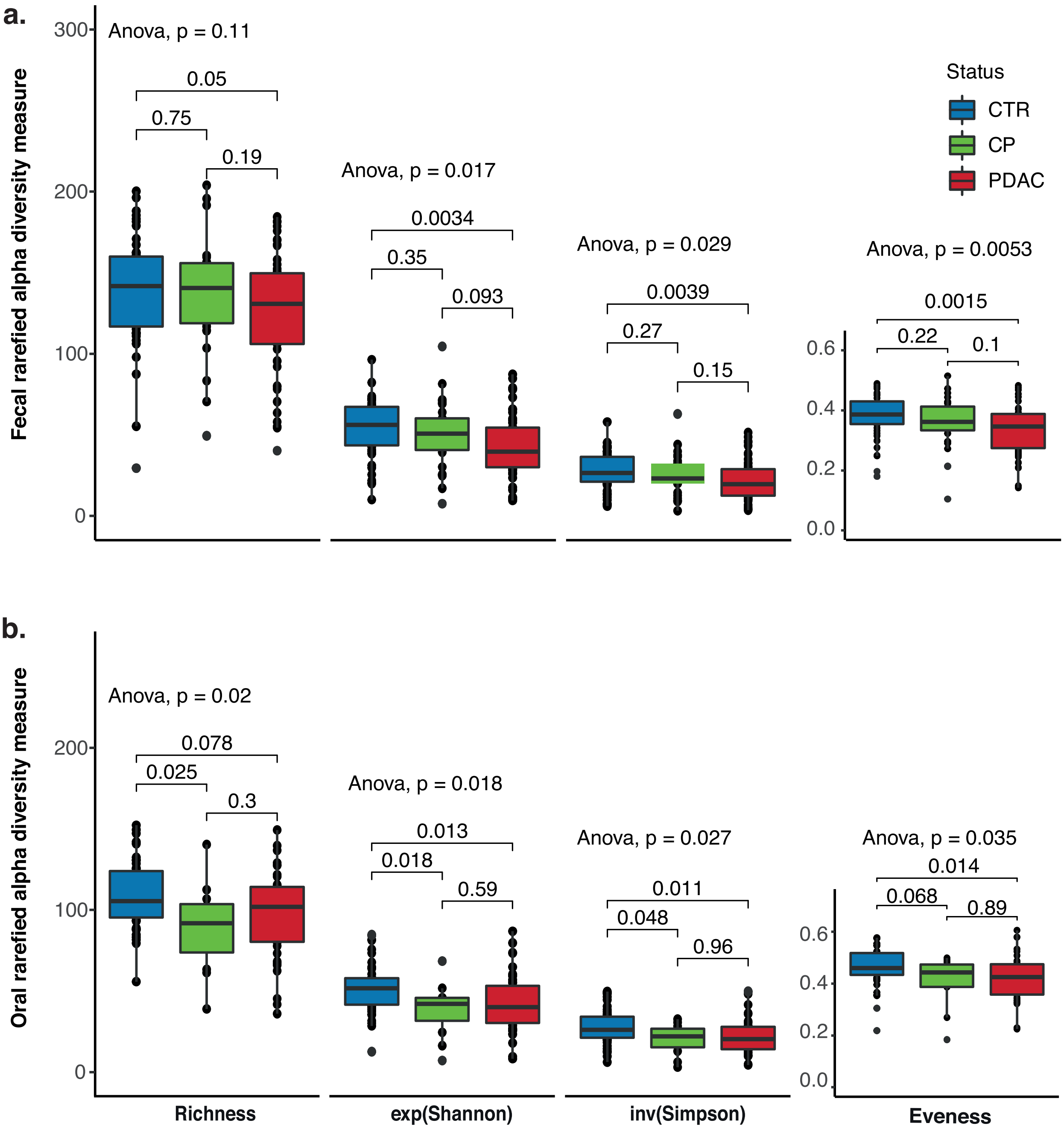


Figure S2. Alpha diversity measurements comparing PDAC and CP patients with controls.
Alpha diversity metrics for (a) fecal and (b) oral samples calculated as richness, exponential Shannon index (exp(Shannon)), inverse Simpson index (inv(Simpson)) and eveness. Colors denote groups, with blue for controls (CTR), green for chronic pancreatitis (CP) patients and red for PDAC cases. Pairwise comparisons were performed using Wilcoxon test and comparisons across all three groups were performed using ANOVA (see Methods).

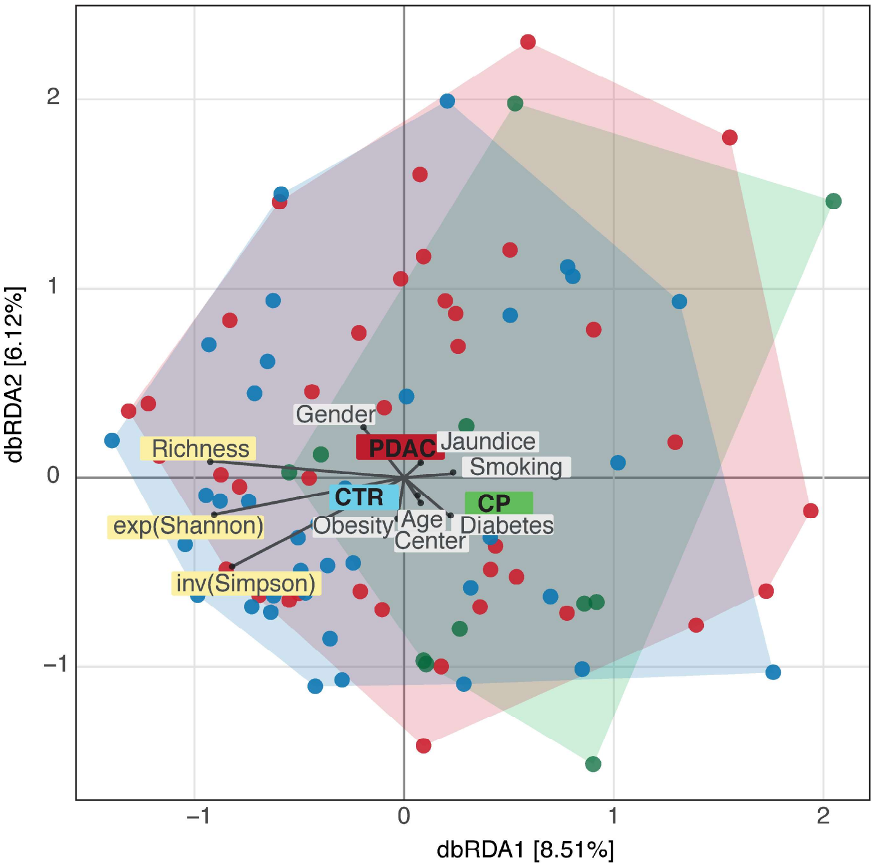


Figure S3. Distance-based redundancy analysis of saliva microbiome.

Bray-Curtis distance-based redundancy analysis (dbRDA) of PDAC, CP and control saliva microbiome data. PDAC samples are shown as red circles, CP patients as green and controls as blue. Association with metadata variables are shown as labeled lines. Richness, exponential Shannon (exp(Shannon)) and inverse Simpson (inv(Simpson)) diversity measures are also visualized with lines and were analysed similarly to metadata variables. The length of the metadata variable line represents the confounding effect size (see Methods).

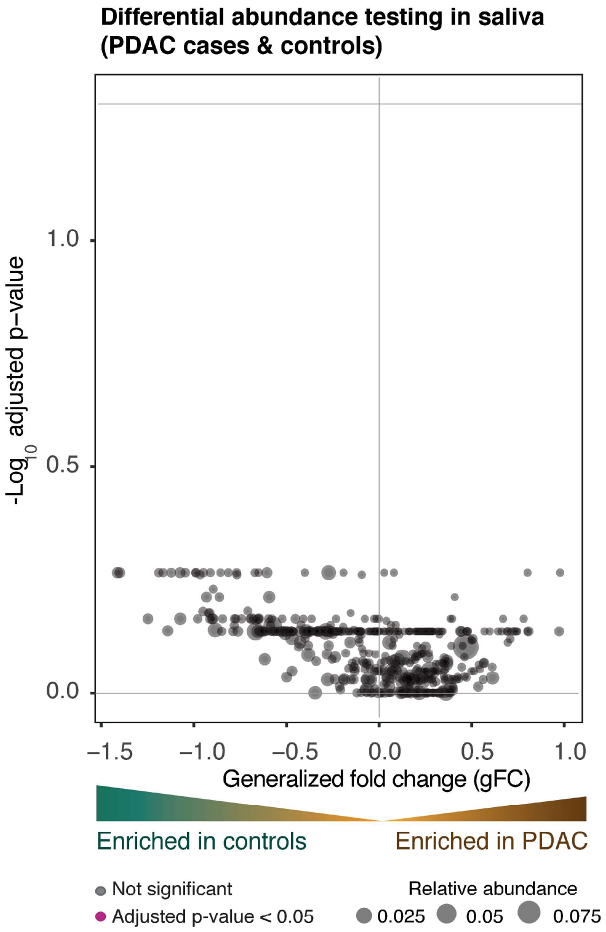


Figure S4. Differential abundance testing of saliva microbiome

Wilcoxon test results of saliva microbiome data to test for enrichment of taxa between PDAC cases and controls (see Methods). Y-axis is log10(FDR corrected p-values), x-axis is generalized fold change and dot size represents the relative abundance of given species and strains. Red dots represent significantly differentially abundant species/strains in either group, while black dots show non-significant species after FDR correction.

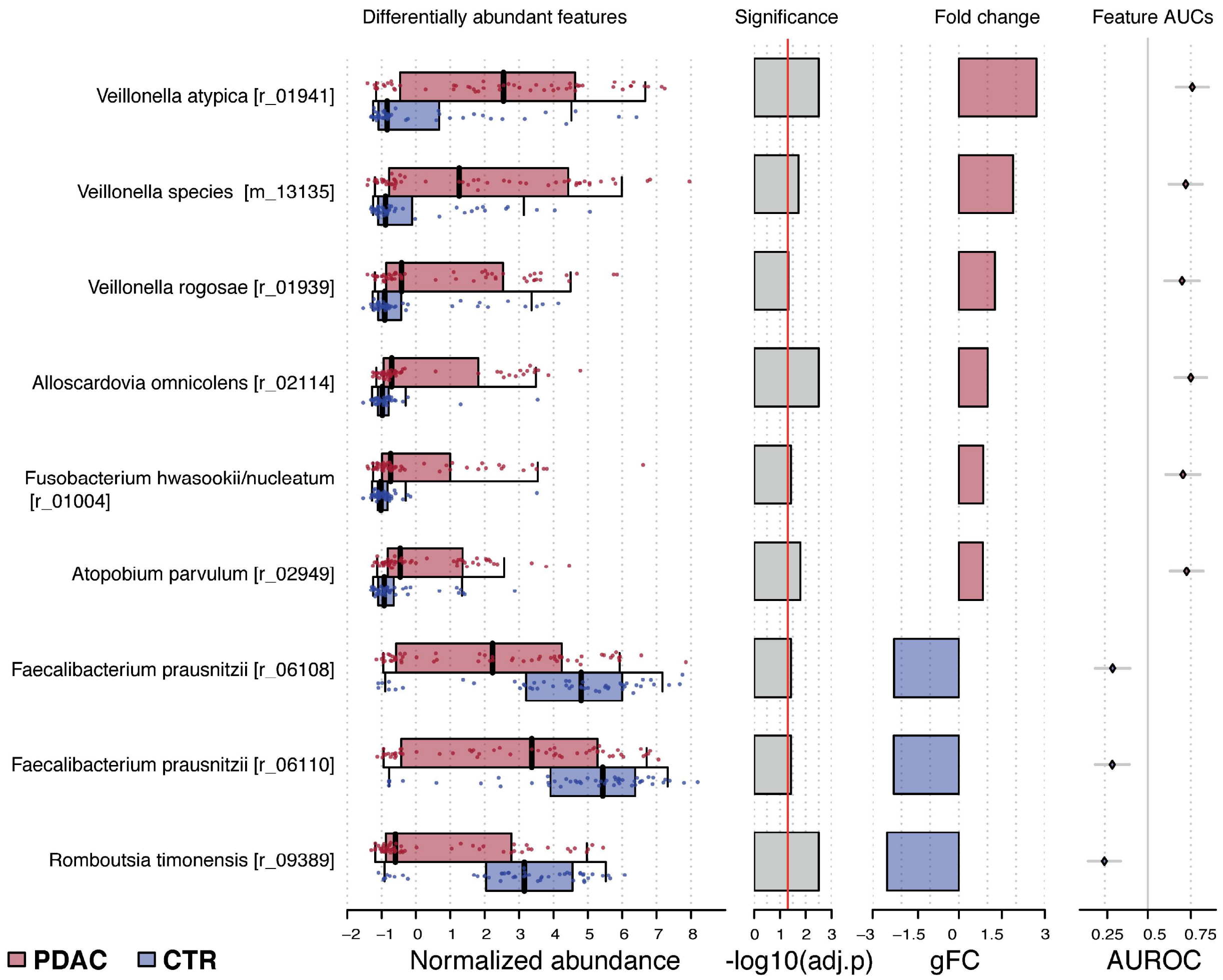


Figure S5. Differentially abundant species in fecal microbiome between PDAC cases and controls.
First column panel shows the differentially abundant species between PDAC cases (red) and controls (blue). Middle panels display the log10(FDR corrected p-values) and generalized fold change for each taxon and the last panel presents the AUC of each feature to distinguish cases from controls. gFC:Generalized fold change.

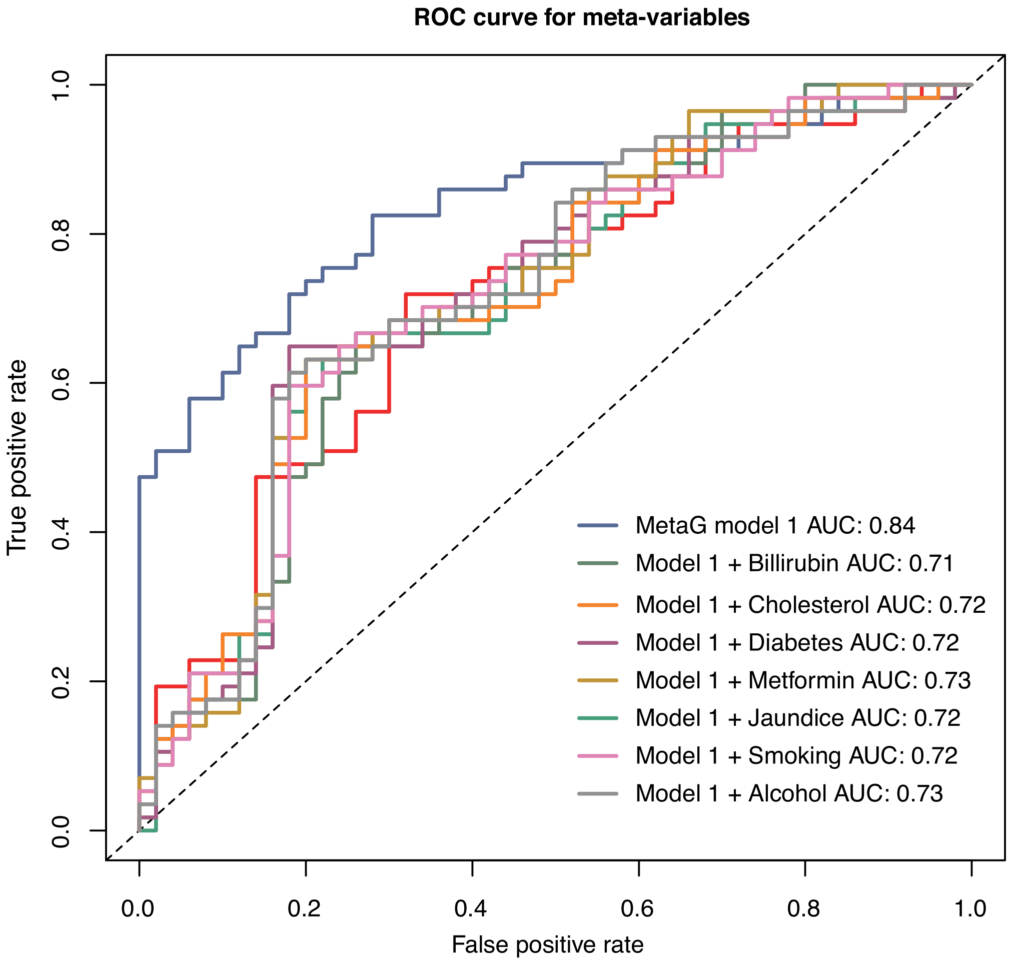


Figure S6. Contribution of confounding factors to the model.

The area under the ROC curve (AUROC) is used to show the performance of lasso_ll model based on fecal microbiome data of PDAC and control samples with 10 times resampling and 10 cross validation (see Methods). Each color corresponds to one specific model based on metagenomics features with an additional metadata variable. Shown metadata variables were added to the metagenomics features table with “add.meta.pred” function from “SIAMCAT” package v1.5.0.

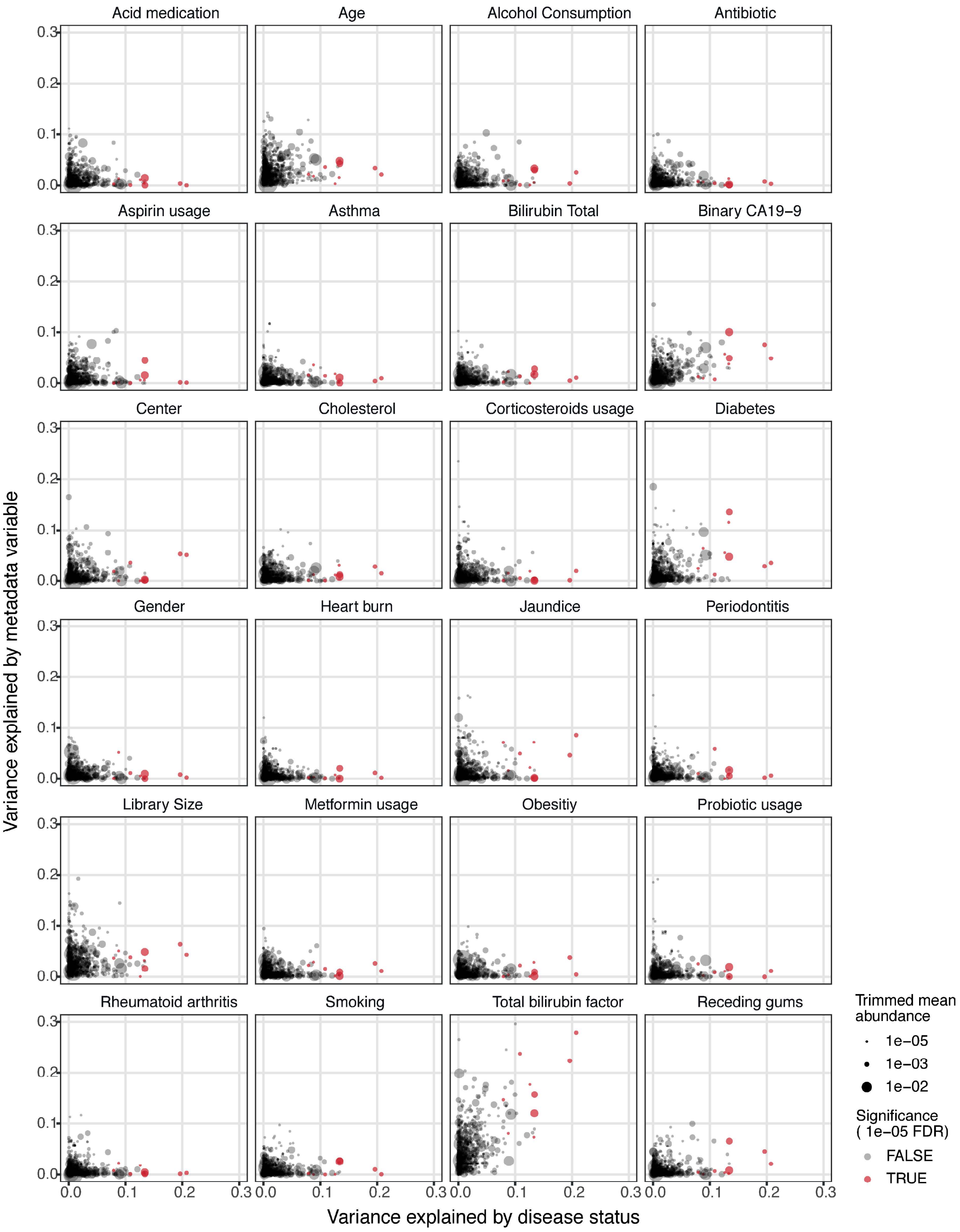


Figure S7. Potential confounder of single species associations by individual demographic and technical variables.

Variance explained by diagnosis is represented against confounding factors for single microbial species. Each circle is a strain or species and is colored red if it is differentially abundant between PDAC cases and controls. The size of each circle represents the mean abundance of that species or strain. Disease status and the tested variables were used as explanatory variables in the linear model for feature abundance.

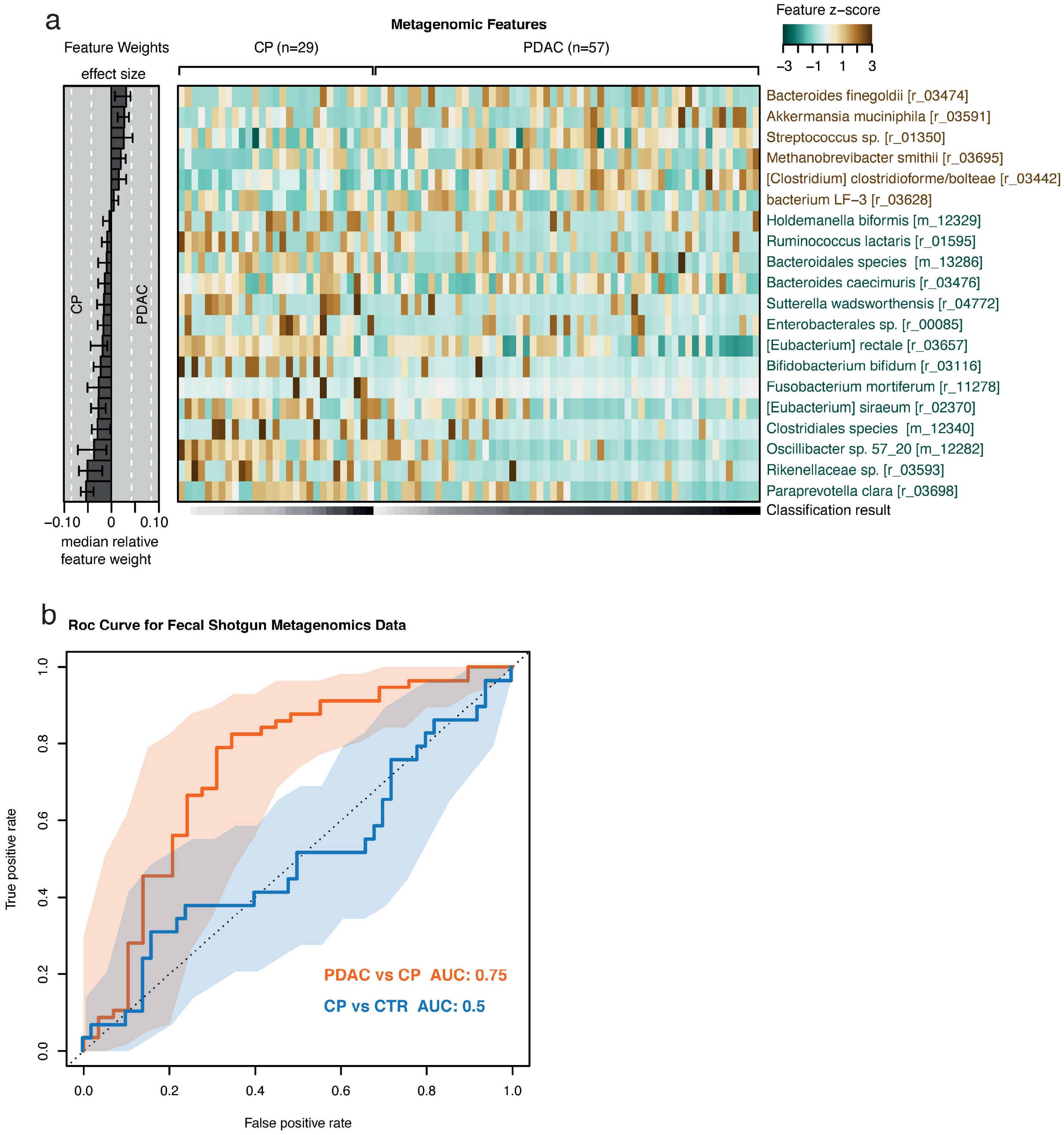


Figure S8. Fecal microbiome-based classifier distinguishes chronic pancreatitis cases from PDAC patients
(a) Heatmap representing the selected metagenomic features in the lasso_ll regression model between PDAC cases and chronic pancreatitis (CP) patients in the fecal microbiome data. (b) ROC curve based on 10 resamplings and 10-fold cross validation (see methods). The blue line represents the model for CP versus controls and the orange line for PDAC vs CP cases. Internal cross validation results are shown as receiver operating characteristic (ROC) curve with a 95% confidence interval shaded in corresponding color.

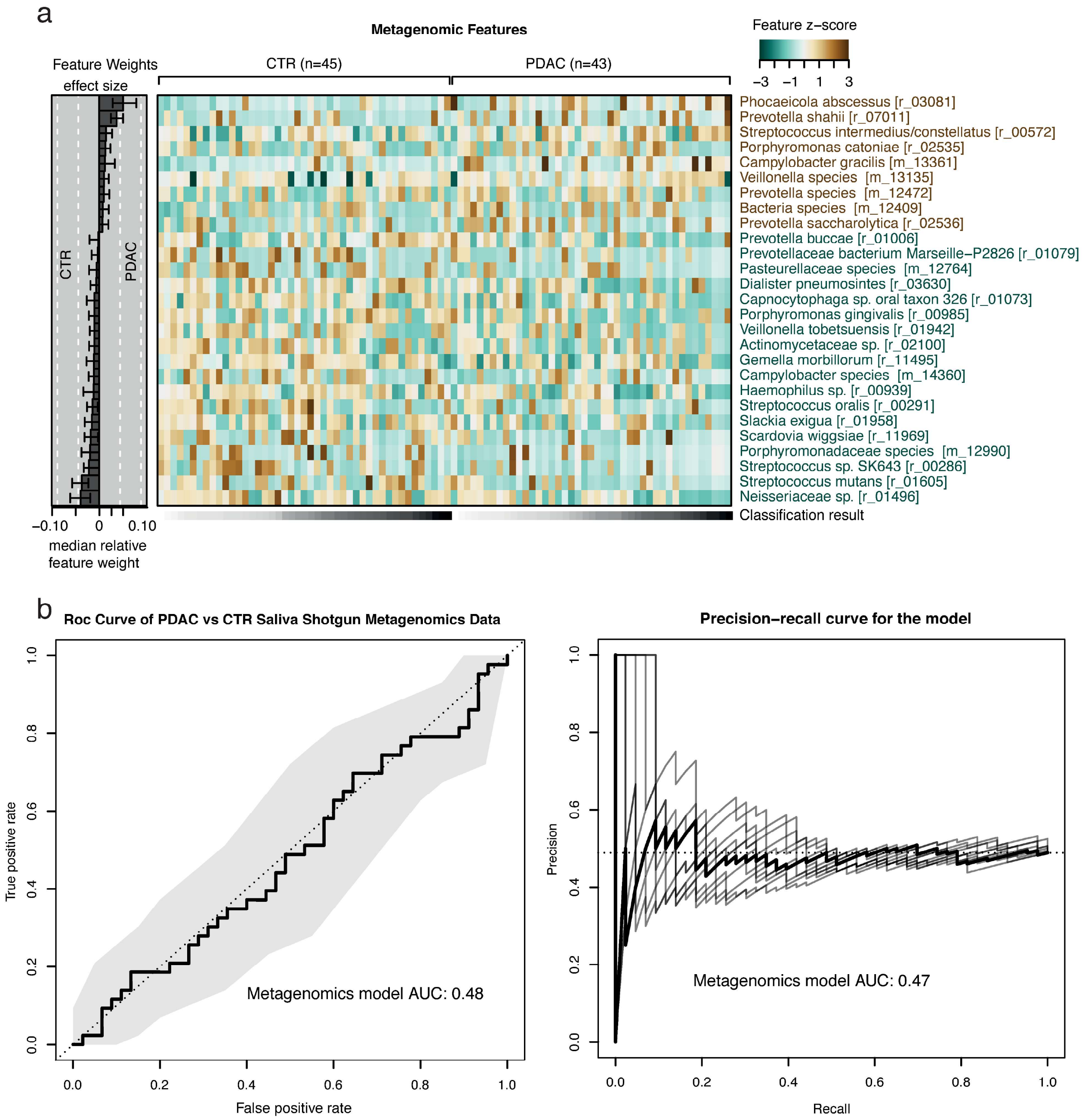


Figure S9. Oral microbiome does not distinguish PDAC samples from control samples.

(a) Heatmap representing the selected metagenomic features in the lasso_ll regression model between cases and controls in the saliva microbiome data. (b) ROC curve based on 10 resamplings and 10-fold cross validation (see methods) and precision recall curve. Internal cross validation results are shown as receiver operating characteristic (ROC) curve with a 95% confidence interval shaded in grey.

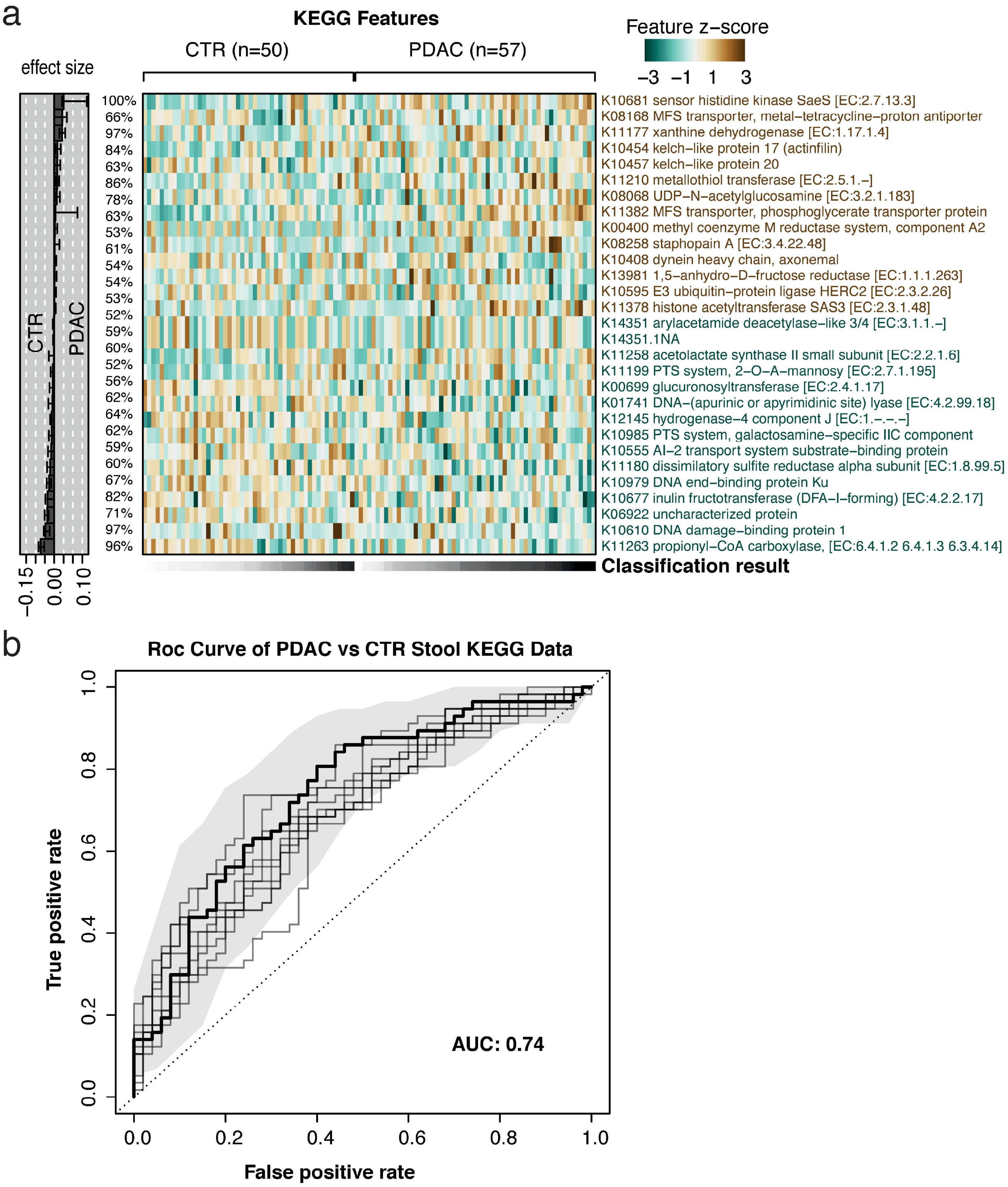


Figure S10. Lasso_ll regression model based on top 200 KEGG modules.

(a) Heatmap representing the selected KEGG modules in the lasso_ll regression model. (b) ROC curve based on 10 resamplings and 10-fold cross validation (see methods).

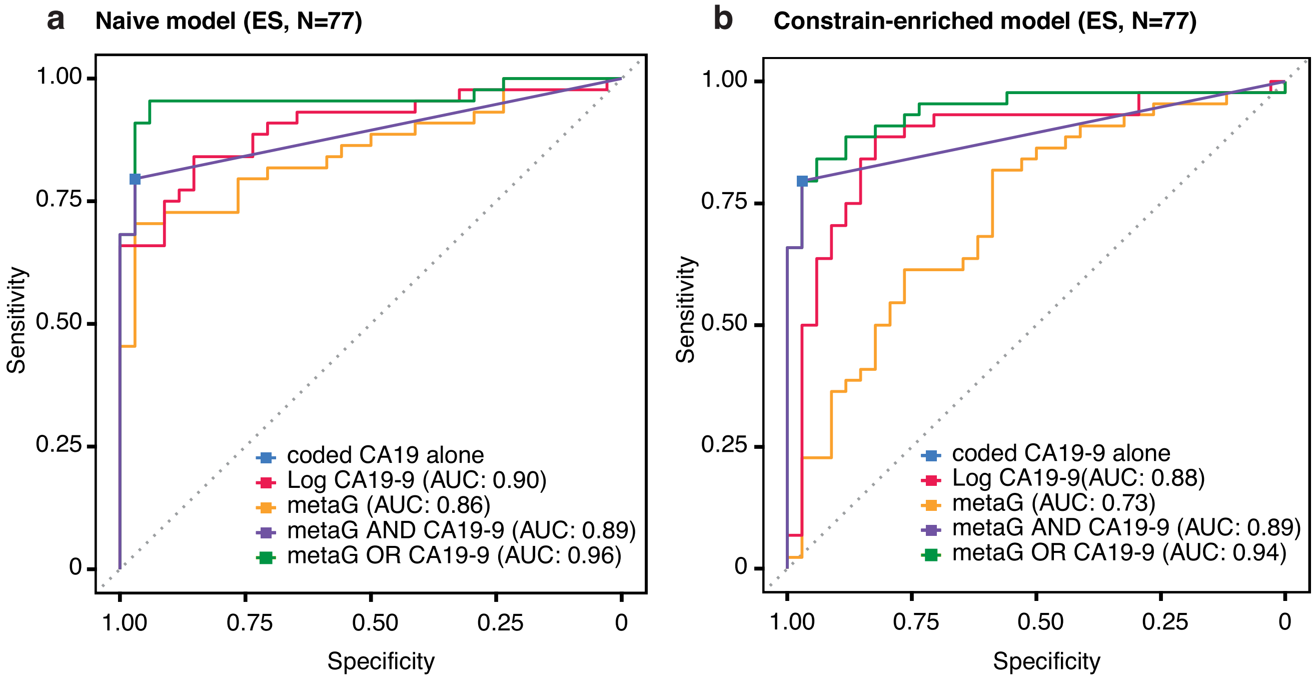


Figure S11. Combination of fecal microbiome data with CA19-9 results increase sensitivity.
77/107 (33/50 CTRs and 44/57 PDAC cases) individuals in Spanish (ES) whom CA19-9 data were available included in the modelling process explicitly. CA19-9 values were converted to binary values (>37ul/ml = 1 & <37ul/ml = 0) **(a)** ROC curve of full feature set. **(b)** ROC curve of enrichment-constrained models based on 77 individual fecal microbiomes. Coded CA19-9 is the binary version of data, which is represented by a blue dot. Log(CA19-9) is displayed with red, while “AND” and “OR” combinations are shown with purple and green respectively. 8/32 CTRs and 43/44 PDAC patients in the German (DE) cohort

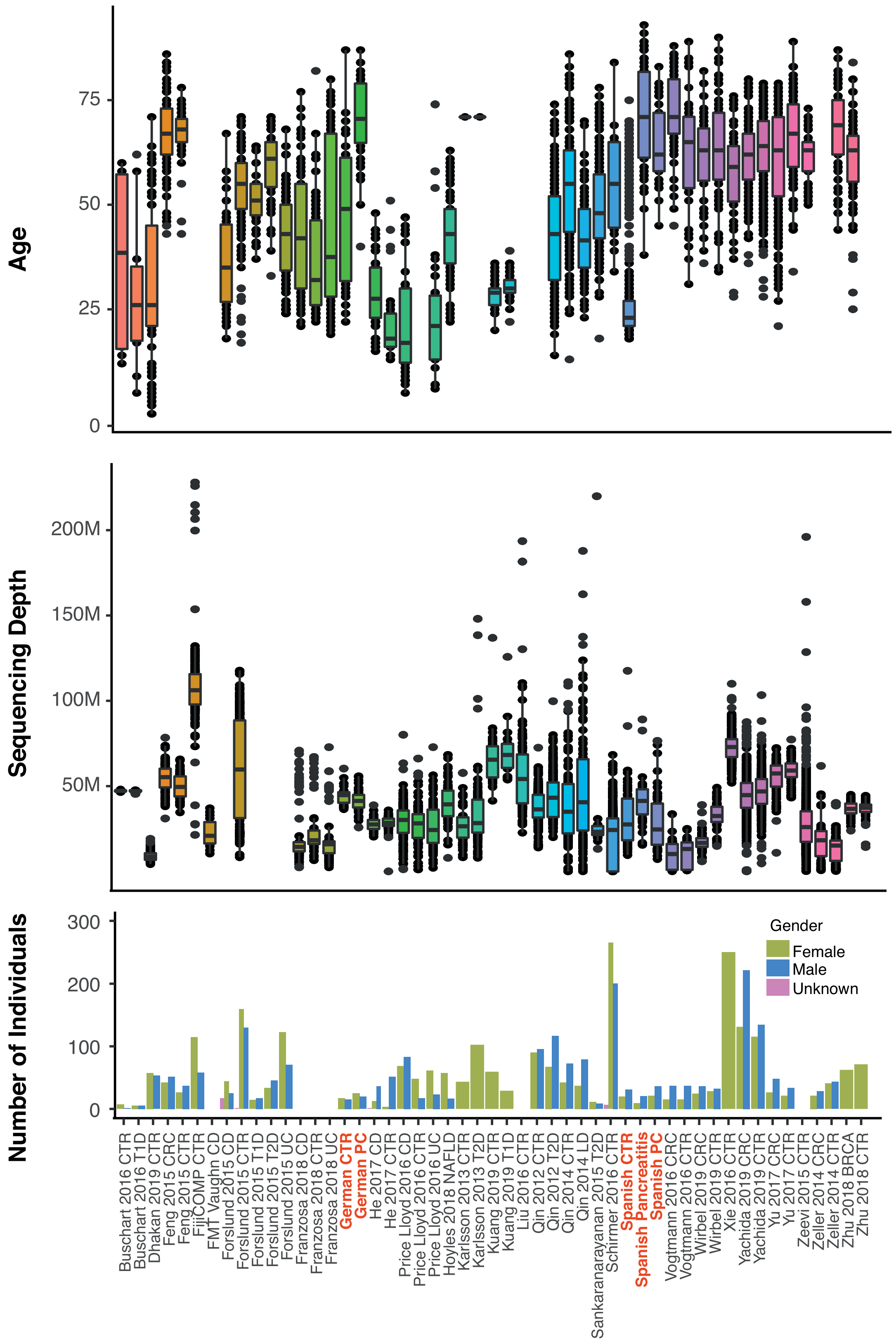


Figure S12. The overview of external validation cohorts.
(a) Age distribution is shown for all external datasets per group. X-axis shows all the studies and y-axis displays the age distribution. **(b)** Sequencing depth is represented across cohorts. **(c)** Gender information is displayed if available as bar plot. Green is used for females while blue is for males for available studies.
M:Million.

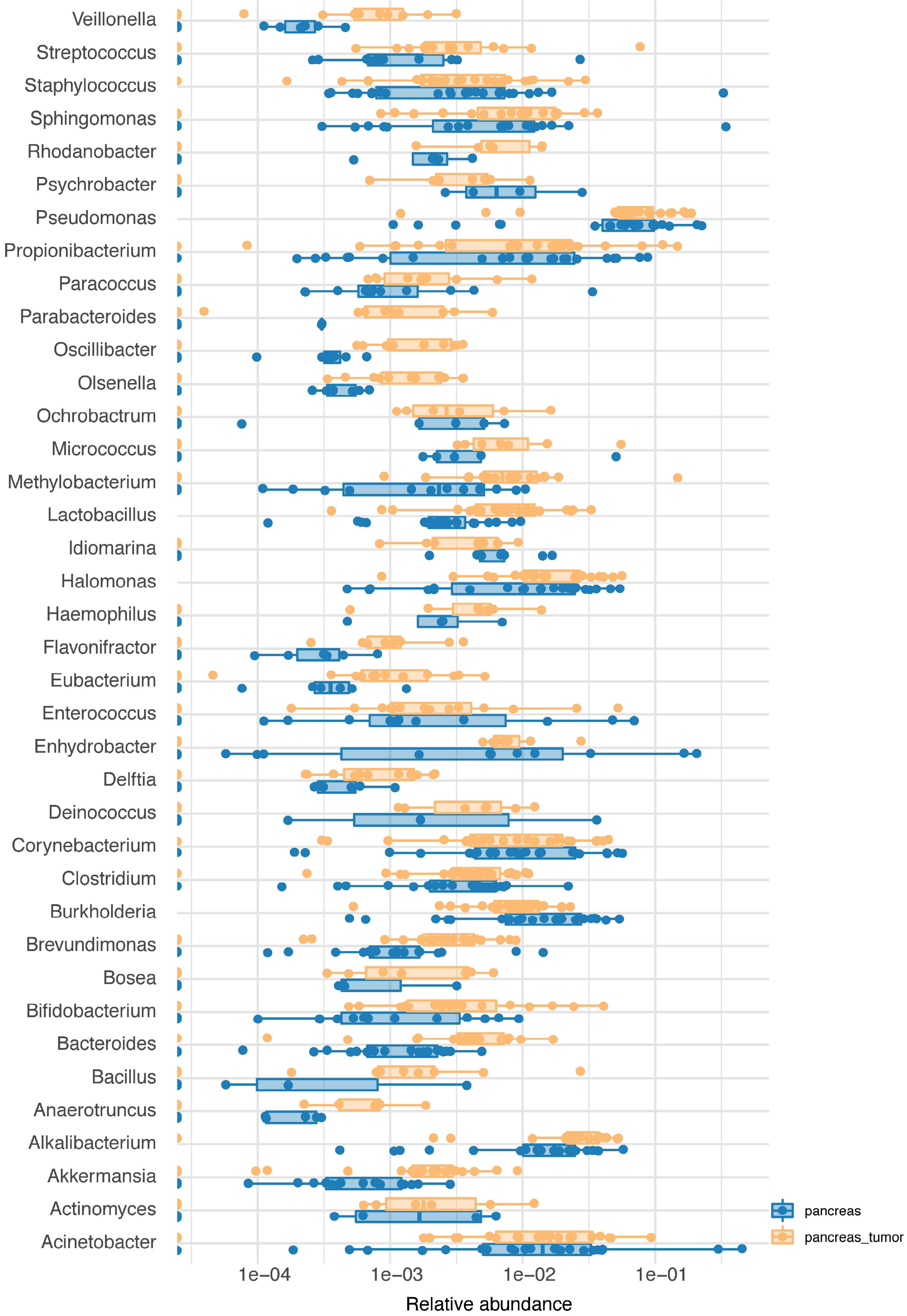


Figure S13. Relative abundance of genera in tumor and non-tumor pancreatic tissue.
Relative abundance of several genera is shown as bar plots. Orange is used to present the pancreatic tumor tissue, while blue is used for non-tumor tissue.

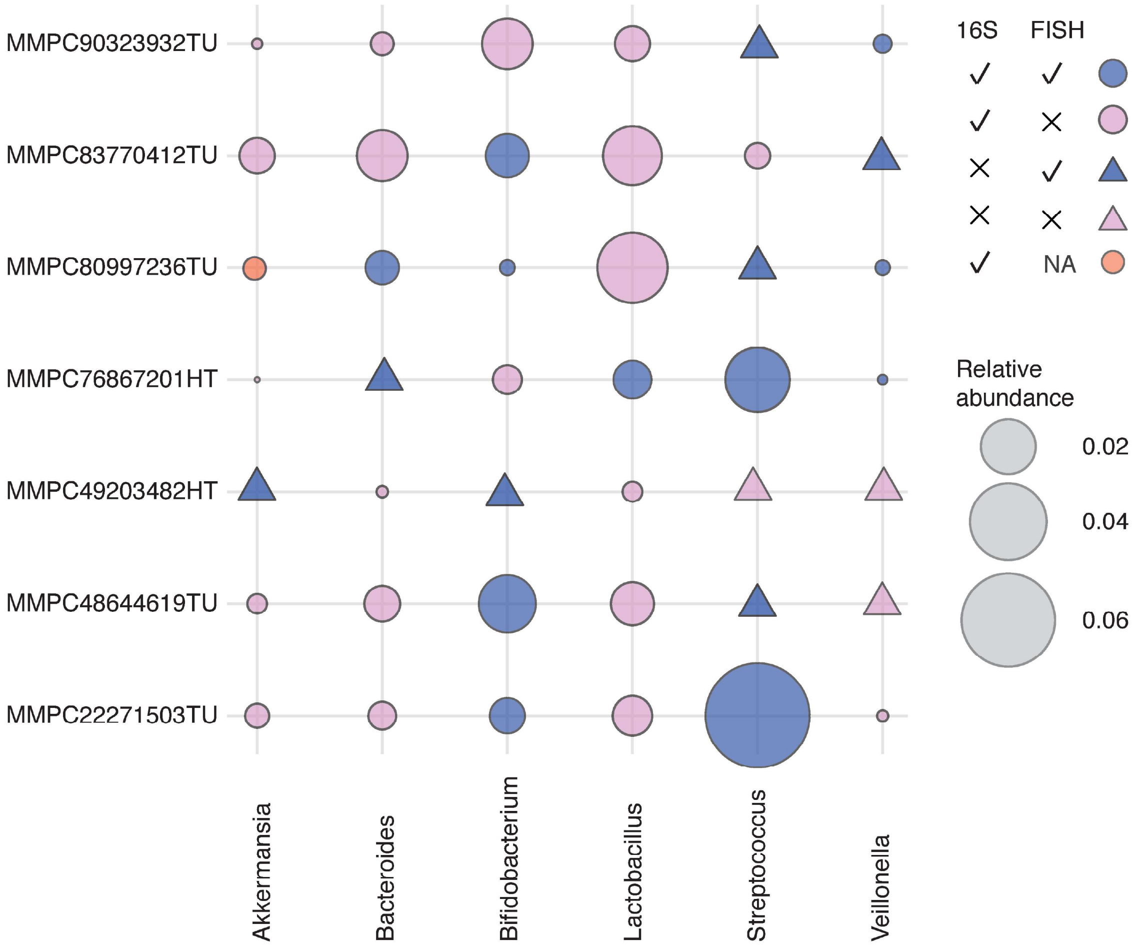


Figure S14. Detailed information of tested samples via in-situ hybridization (FISH).

Rows display the tested samples and columns show the tested genera. The size of the dot represents relative abundance of genus in the given sample. Triangles show that 16S was negative for given samples and color code displays if FISH was positive (blue) or negative (pink). One sample, displayed in orange, did not have enough tissue material for FISH testing.NA: Not available.

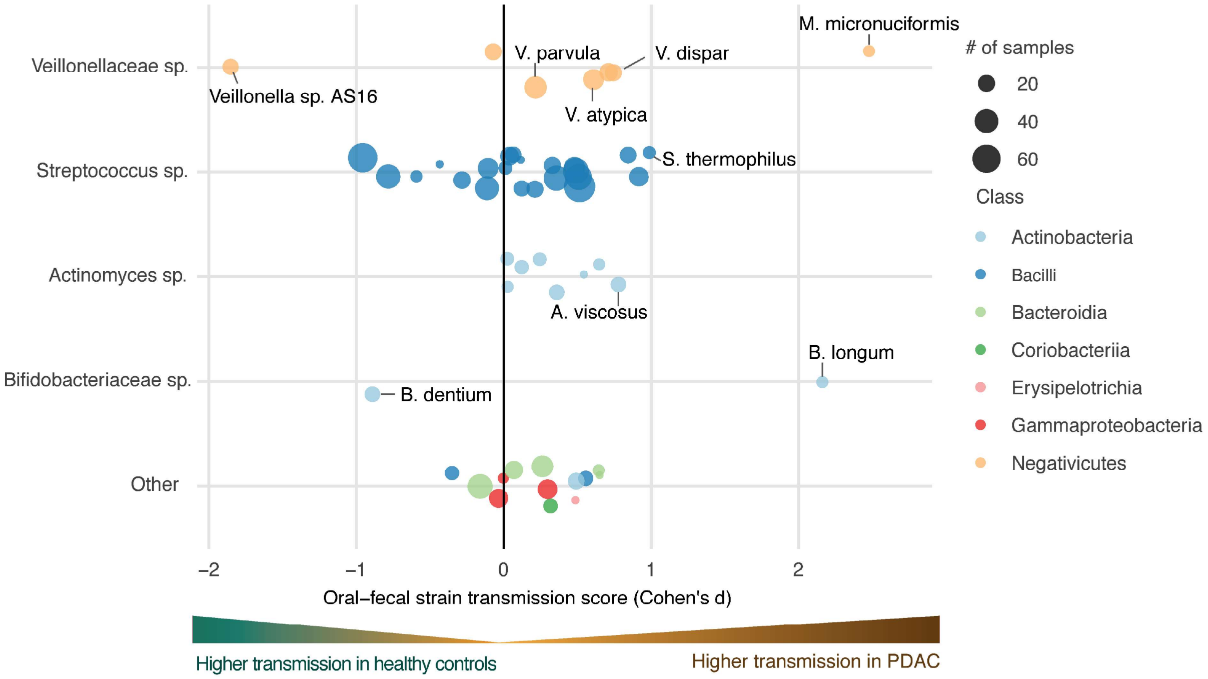


Figure S15. Oral-fecal transmission scores differ between PDAC cases and controls.
Oral-gut transmission scores (y-axis) of each species are displayed grouped by genus (x-axis). The number of subjects is represented by the size of the circle and the color represents the corresponding class group.