Enrichment of Oral Microbiota in Early Cystic Precursors to Invasive Pancreatic Cancer

Rogier Aäron Gaiser, Asif Halimi, Hassan Alkharaan, Liyan Lu, Haleh Davanian, Katie Healy, Luisa W Hugerth, Zeeshan Ateeb, Roberto Valente, Carlos Férnandez Moro, Marco del Chiaro, Margaret Sällberg Chen

Supplemental Methods

Cyst fluid sample collection

Fresh resection specimens were received at the pathology laboratory within 20 minutes of surgical removal, under sterile conditions and on ice. A specialist pancreatic pathologist performed the macroscopic assessment to identify the cystic lesion and main pancreatic duct. Fluid from the latter was collected using a sterile syringe without needle. When the cystic lesion was readily identified in the intact specimen, its fluid was aspirated using a sterile syringe with needle. In those specimens in which the cystic lesion was not evident by the surface or the cyst content was too viscous for needle aspiration, the specimen was cut or the cyst wall incised and the content aspirated using syringe without needle. Aspirated fluid was placed in DNA-free, sterile collection tubes and stored frozen at -80 °C until further analysis. Extraction blanks (DNA-free water) was handled identical at the pathology laboratory to serve as environmental DNA contamination controls.

Histopathological diagnosis and cyst fluid classification

Resection tissue specimens were fixed in 4% formaldehyde and processed for routine histopathological diagnosis. The cystic lesion was classified by light microscopic examination of hematoxylin-eosin stained slides by specialized pancreatic pathologists as follows: intraductal papillary mucinous neoplasm (IPMN), mucinous cystic neoplasm (MCN), and serous cystadenoma (SCA). The grade of dysplasia was assessed using a 2-grade (high/low) scale, according to current international standard [1]. However, to make the IPMN classification more representative of the neoplastic epithelium that produces it, when the extent of high grade dysplasia was only focal (< 5 %), it was classified according to the predominant low grade. Specimens with IPMN and concomitant invasive carcinoma were classified as “Cancer” and considered as a separate class for further analyses.

Bacterial DNA isolation and absolute quantification of 16S rRNA gene by real-time PCR

To detect and quantify bacterial DNA in pancreatic cyst fluid, plasma and FFPE tissue, a real-time quantitative PCR assay targeting a highly conserved region of the bacterial 16S rRNA gene was used. Microbial DNA was purified from 200 μL pancreatic cyst fluid or plasma using the ZymoBIOMICS™ DNA Miniprep kit (Zymo Research, Irvine, USA) according to manufacturer's instructions for protein-rich biofluids, under sterile conditions in a biological flow cabinet and using 10% bleach-cleaned and UV-irradiated equipment. DNA was eluted in 100 μL nuclease-free H_2O. From FFPE tissue, DNA was purified using the AllPrep DNA/RNA FFPE Kit (Qiagen), which was eluted in 100 μL nuclease-free H_2O. DNA purity and concentration were assessed by spectrometric analysis using a NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, DE, USA) and fluorescence-based using the QuantiFluor® dsDNA System (Promega, Madison, USA) and a Qubit 3.0 (Thermo Fischer Scientific). Extraction blanks and nuclease-free H_2O were processed in parallel as controls to set the assay cut-off. Standard curves for the 16S rRNA gene were generated using 10-fold serial dilutions from purified genomic DNA of *E. coli* ATCC 25922 to allow 16S copy number quantification.

Quantitative PCR conditions were as follows: 95°C for 10 min, and 40 cycles of 95°C for 15 s and 58°C for 45 s on a 7500 Fast Real-Time PCR system (Applied Biosystems). 5 μL template DNA (from cyst fluid and plasma) or 25 ng total DNA (from FFPE tissue) was used in each reaction with 500 nM forward primer P891F (5'-TGGAGCATGTGGTTTAATTCGA-3') and reverse primer P1033R (5'-TGCGGGACTTAACCCAACA-3'), 200 nM probe (5'-FAM-CACGAGCTGACGACARCCATGCA-TAMRA-3') [2], or with *Fusobacteria nucleatum* primers 5'-AGGGTGAACGGCCACAAG-3', 5'-TCTCGGTCCATTGTCCAATATTCC-3' and probe 5'-FAM-ACACGGCCCTTACTCC-TAMRA-3' with 10 μL Fast Universal PCR Master Mix (Applied Biosystems) and nuclease free H_2O up to a total volume of 20 μL per reaction. Each reaction was run with technical duplicates and experiments were performed in duplicate. The lower limit of detection was set at the mean value + 3 SD of the controls (nuclease-free H_2O and extraction blanks); samples above the cut-off level were considered as bacterial DNA-positive samples.

PacBio full length 16S rRNA gene amplicon sequencing and microbiota analysis

Library preparation, quality control, sequencing, and raw data filtering were performed at GATC Biotech (Konstanz, Germany). Briefly, the V1-V8 region (1381 bp) of the 16S rRNA gene was amplified by PCR using primers 27F (5'-AGAGTTTGATCCTGGCTCAG-3') and BS-R1407 (5'-GACGGGCGGTGWGTRC-3'). DNA quality and integrity were checked using Agilent 2100 Bioanalyzer, Advanced Analytical Technologies Fragment Analyzer and Invitrogen Qubit® Fluorometer. Samples showing multiple amplicon peaks were discarded. The libraries that passed QC had a DNA concentration of > 0,6 ng/μL and fragment size range from 1489 to 1681 bp. These libraries were used as input for PacBio Single Molecule, Real-Time (SMRT) amplicon sequencing. Post-filtered reads of insert had a mean length of 1473 bp with 16 polymerase passes per insert on average, resulting in a mean read quality of 0.99695. Only reads of inserts with an accuracy greater than 0.99 were used for further analysis. Chimera sequences were filtered with UCHIME [3] using a full length, good quality, and non-chimeric 16S rRNA gene reference database. One FASTA file per sample was delivered by GATC. To filter out low quality sequences and decrease the amount of noise, all sequence files were concatenated, de-replicated and clustered to 99% identity, excluding singletons in this step. This generated 4277 centroid sequences, which were submitted to the online RDP classifier, v. 16 [4]. 589 sequences that were not classified as Bacteria with 100% bootstrap support were excluded. No centroids were classified as Archaea with >90% bootstrap support. Non-bacteria sequences (Chloroplast), which came overwhelmingly from a single cancer sample with low DNA yield, were excluded. The remaining sequences were kept and their taxonomic annotation was truncated to the phylum rank with >70% bootstrap support. To increase the granularity of the taxonomic assignment, sequences were also mapped to the HOMD 15.1 database with minimum 90% identity and maximum 10 best hits kept. This mapping was then parsed so that species-level assignments were accepted if they were in agreement for all hits with >97% identity, and genus-level assignments if they were in agreement for all hits with >95% identity. If no such hits were found, the RDP classification was kept. In total, 659 centroids could be mapped to HOMD. Finally, the FASTA files for each

sample were mapped back to the centroids and assigned to the best hit with >97% identity >97% of the query length, to quantify the relative abundance of centroids in each sample.

Statistical analyses

Statistically significant differences of 16S DNA copy number and IL-1β concentration between diagnose groups were assessed using Kruskal-Wallis test with Dunn's multiple comparison test. Differences between proportions of DNA-positive samples between groups were tested using the chi-square test. Log_{10}-transformed 16S DNA copy numbers were used as input for one-way ANOVA followed by a post test for linear trend by diagnose group. Spearman correlation was calculated between 16S DNA copy numbers and IL-1β concentrations per sample. All calculations were performed in GraphPad Prism 8.0.

Richness (Chao1), diversity (Shannon's entropy) and Inverted Simpson's were calculated at each taxonomic level using R packages Fossil 0.3.7 [5] and Vegan 2.4-5 [6], group-wise statistical comparisons were done using Welch's t-test with Benjamini-Hochberg FDR correction. Between-sample diversity was estimated using Bray-Curtis dissimilarity and plotted with package Pheatmap 1.0.8 [7]. PCoA calculation and plotting was performed with package Ape 5.2[8]. ANOSIM was calculated using the Vegan package, based on Bray-Curtis distances and using standard parameters. These calculations were performed in R v. 3.4.3 [9]. Linear Discriminant Analysis Effect Size (LEfSe) [10] was used to find differentially abundant taxa between three diagnosis groups (classes), with per-sample normalization to 1 million, an alpha cut-off value of 0.05 for the Kruskall-Wallis factorial test, the "one-against-all" multi-class analysis strategy, and a threshold for discriminative features at a logarithmic LDA score >3. Multivariate Analysis by Linear Models (MaAsLin, http://huttenhower.sph.harvard.edu/maaslin) was used to find correlations between clinical metadata and bacterial abundance data, with a q-value threshold of > 0.05, minimum feature abundance of 0.001, and minimum feature prevalence filter at 1%. Microbial co-occurrence analysis was done using only OTUs identified up to species level using the CoNet plugin [11] with the following five methods; Pearson,

Spearman, Mutual Information, Bray-Curtis, and Kullback-Leibler. For each method, the 300 top (positive correlation) and bottom (negative correlation) edges were calculated. The permutations output to compute p-values included randomisation, shuffling of rows, and renormalisation were stored in a file prior to method- and edge-specific permutation (100 iterations). This file was used to obtain bootstrap score (100 iterations) distributions to compute final p-values, merging all five method-specific p-values of each edge into one single p-value. The resulting network of statistically significant interactions with $p < 0.01$ and supported by >3 methods was visualized in Cytoscape 3.6.1. Generation of other graphs and statistical test were performed in GraphPad Prism 8.0.

Ethical considerations

This study follows the Helsinki convention and good clinical practice. This study was conducted at Karolinska University Hospital under permission of the Ethical Review Board Stockholm and Karolinska Biobank Board (Dnr 2015/1580-31/1). Written informed consent was obtained from all patients.

References

1 Basturk O, Hong SM, Wood LD, Adsay NV, Albores-Saavedra J, Biankin AV*, et al.* A Revised Classification System and Recommendations From the Baltimore Consensus Meeting for Neoplastic Precursor Lesions in the Pancreas. The American journal of surgical pathology 2015;**39**:1730-41.
2 Yang S, Lin S, Kelen GD, Quinn TC, Dick JD, Gaydos CA*, et al.* Quantitative multiprobe PCR assay for simultaneous detection and identification to species level of bacterial pathogens. Journal of clinical microbiology 2002;**40**:3449-54.
3 Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011;**27**:2194-200.
4 Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y*, et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic acids research 2014;**42**:D633-42.
5 Vavrek MJ. fossil: Palaeoecological and palaeogeographical analysis tools. Palaeontologia Electronica 2011;**14**.
6 Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB*, et al.* vegan: Community Ecology Package. 2017.
7 Kolde R. pheatmap: Pretty Heatmaps. 2015.
8 Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. 2018.
9 Team RDC. R: A Language and Environment for Statistical Computing. 2016.
10 Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS*, et al.* Metagenomic biomarker discovery and explanation. Genome biology 2011;**12**:R60.
11 Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape [version 2; referees: 2 approved]. F1000Research 2016;**5**.