Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

Ece Cevirgen, M.Sc.

born in: Malatya, Turkey

Oral examination: May 25th, 2020

Exploring the Potential of the Human Microbiome for Colorectal and Pancreatic Cancer Screening

Referees:

Dr. Athanasios Typas

Prof. Dr. Benedikt Brors

Exploring the Potential of the Human Microbiome for Colorectal and Pancreatic Cancer Screening

Ece Cevirgen, M.Sc.

Supervised by

Prof. Dr. Peer Bork

&

Prof. Dr. Magnus von Knebel Doeberitz

“Yani, öylesine ciddiye alacaksın ki yaşamayı,

yetmişinde bile, mesela, zeytin dikeceksin,

hem de öyle çocuklara falan kalır diye değil,

ölmekten korktuğun halde ölüme inanmadığın için,

yaşamak yanı ağır bastığından.”

Nazim Hikmet

Acknowledgements

Büyük ve küçük ailem, sevginiz ve güveniniz olmadan, bunlara asla cesaret edemezdim, iyi ki varsınız! Sizin parçanız olmaktan hep gurur duydum. Doktoraya başlamak, sizden uzak kalmak anlamına gelse de, bana "iyi ol" dediniz. Kaçırdığım her özel an ve sizi deli gibi özlemek dışında, burada geçirdiğim her saniyeye değdi. Hakan, her yorulduğumda ve şikayet ettiğimde, bana katlandın ve her şeyin tekrar iyi olacağını söylemeye devam ettin. Ali ve Abicik en zor zamanlarımda bile benimleydiniz ve bana her zaman kahkahaların değerini hatırlattınız… Bircan ve Ulduz, hiç kimseyle bu kadar ağlamadım, dostluğun anlamını hayatıma siz kattınız. Her düştüğümde "yaparsn" dediğinizi duydum ve ve tekrar kalktım. Şerefinize…

I owe a big thank you to my supervisor Peer Bork, not only for the knowledge I have gained during my PhD but also for life lessons. He handled me very well during my years at EMBL even though I did not have a filter. He believed in me, supported me and pushed me for better, even in these “another two weeks” moments. He taught me the importance of creating a happy group, seeing the positivity in human. Accepting mistakes wisely and learning from them were also strength I have learned during my PhD time. I still don’t know which kind of “bomb” am I, as he mentioned during the first interview, but I know that I will always remember my PhD years with a smile on my face thanks to him for all his effort.

I also want to thank my second supervisor Magnus von Knebel Doeberitz for his trust and support. My TAC members, Nassos Typas, Benedikt Brors deserves a big thank for their feedback during and in between TAC meetings. I want to thank Georg Zeller specially for his extra time and inspiring discussions, I have learnt a lot from him about statistics and felt like having a second group in EMBL. They all guided me though my PhD not only scientifically but also personally. Thank you, Michael Zimmermann and Robert Russell for agreeing to be

part of my thesis defense and for the nice conversations. Sebastian Schmidt, you were an amazing mentor and I felt lucky to have not only your guidance but also friendship during my PhD. I can hear “Take your time Ece, you are doing great, it will be alright”.

My PhD let me learn a lot of concepts and I have gained strength in science and I had chance to meet amazing people whom I would call “dost” not colleagues anymore. We shared happiness and laughed together, but more importantly we shared sadness, frustration and worries. Dear Alejandro, your Chilean spirit and positivity kept me alive, you always knew what I need. *Bacalhau* became a part of me thanks to you Renato and thank you for all the talks and walks. Oleks, your enthusiasm is so unique, you always made me remember why I do science, whenever I got frustrated by daily things and learn dark humor. Pamela, how would be my figures without you and my PhD without your cocktails… Lucas, you are the most optimistic person I have ever know, “Ain’t nobody got time for that”; thanks to you that I realized being a duck doesn’t hurt! Ivica, you are the best pizza/peace maker. Blond Polish Annas, you are both unique and great, we always end up with interesting discussions that let me to have a new view. Every morning coming to EMBL became fun more than work, thanks to you Thea, Simone, Anthony, Michael, Sanders, Steffi, Marja, Maria, Yan, Supriya, Wasiu, Askarbek, Kiley, Jakob, Chris, Suguru. Sara, you always made me smile and it was always special to spend time with you.

Diana, you will be there for me any time, I know I can cry hard and party hard with you. Keep rocking! Mariana, it was great to share several coffee breaks and parties with you. Judith, you were the first one helping me out to learn bioinformatics and coding; still on the way. Thank you for being so understanding and ambitious. First generation Zaugg group people including Pooja, Armando, Anna, Christian and monkey thank you for welcoming me at EMBL. You were the first reason to love Heidelberg and Germany.

GeneCore became my second lab, you were always supportive and fun. How can I forget all Friday coming together sessions, you always received me with joy and candies? Vladimir, you taught me how to make a family out of co-workers and you were a great role model. Rajna and Anja, it was great to work with you (even in S2 lab☺).

A big thanks to all of you again, without your contribution, guidance and friendship, this thesis would not exist.

Table of Contents

Summary

Microbiome research continues to move from purely descriptive approaches towards translation of findings into clinical applications. Microbiomics is a very active research field and the microbiome is now firmly established as an important factor in human health and disease, although mechanistic understanding currently remains limited. In this dissertation, I present two distinct studies which aim to (i) develop a microbiome-based screening approach for Colorectal Cancer (CRC), and (ii) investigate the microbiome of multiple body sites in Pancreatic Ductal Adenocarcinoma in order to describe a key microbiome signature that can be used for screening purposes.

First, I developed a qPCR-based method for the early detection of CRC named *Microbial Abundance-based Stool Test* (MAST), which, after optimization and selection, is based on the quantification of seven selected microbial species from fecal samples. MAST showed a high predictive power and accuracy (AUC = 0.88) in a first screening cohort (77 controls, 56 CRC) and was further validated on an independent cohort of 344 individuals (272 controls, 72 CRC). To assess the potential usage of MAST in clinical applications, I compared MAST with established CRC detection method gFOBT which targets occult blood presence in stool. The combination of gFOBT and MAST resulted in increased specificity in the validation cohort owing to the complementarity of the captured signature (i.e., microbiome species abundance vs. occult blood presence). In summary, my work establishes MAST as a promising tool towards the early detection of CRC, and as a significant step in taking microbiome-based diagnostics towards the clinic.

In my second project, I investigated the microbiome of multiple body sites of 57 Pancreatic Ductal Adenocarcinoma (PDAC) patients, 50 healthy controls and 29 chronic

pancreatitis patients by using 16S rRNA gene sequencing, metagenomic sequencing and fluorescent *in-situ* hybridization with the aim of establishing characteristic microbiome signatures of pancreatic cancer. While the overall microbiome community composition was not shifted between cases and controls in either saliva, stool or pancreatic tissue, the stool microbiome carried strong PDAC-specific signatures, distinguishing cancer patients from controls with high accuracy (AUC = 0.84 based on 22 microbiome species). The salivary microbiome, in contrast, showed only a weak signal overall. I validated the specificity of the fecal-based prediction model on external cohorts (3468 metagenomes) including healthy controls and patients affected by diseases such as type 1 or type 2 diabetes, liver cirrhosis, inflammatory bowel disease, and CRC. Moreover, pancreatic tumor and adjacent normal tissue showed markedly different microbiome profiles. Lastly, the presence of stool-related species/genera in both tumor and healthy tissue were validated by sequencing and in-situ hybridization methods.

In summary, I describe a novel screening test for CRC, and advance the understanding of the microbiome in Pancreatic Cancer across several body sides.

Zusammenfassung

Die Mikrobiomforschung bewegt sich weiterhin von rein deskriptiven Ansätzen in Richtung der Translation von Erkenntnissen in klinische Anwendungen. Die Mikrobiomik ist ein sehr aktives Forschungsgebiet: das Mikrobiom ist mittlerweile als wichtiger Faktor sowohl für die menschliche Gesundheit, als auch bei diversen Krankheiten etabliert, obwohl das mechanistische Verständnis vieler Zusammenhänge derzeit noch weitgehend begrenzt ist. In dieser Dissertation stelle ich zwei verschiedene Studien vor, die darauf abzielen, (i) einen mikrobiombasierten Screening-Ansatz für Darmkrebs (CRC) zu entwickeln und (ii) das Mikrobiom mehrerer Organe beim duktalen Pankreas-Adenokarzinom zu untersuchen, um eine wesentliche Mikrobiomsignatur zu beschreiben, die für Screeningzwecke verwendet werden kann.

Zunächst entwickelte ich eine qPCR-basierte Methode zur Früherkennung von CRC mit dem Namen Microbial Abundance-based Stool Test (MAST), die auf der Quantifizierung von sieben ausgewählten mikrobiellen Spezies aus Stuhlproben basiert. MAST zeigte in einer ersten Screening-Kohorte (77 Kontrollen, 56 CRC) eine hohe Vorhersagekraft und Genauigkeit (AUC = 0,88) und wurde in einer unabhängigen Kohorte von 344 Personen (272 Kontrollen, 72 CRC) weiter validiert. Um die potenzielle Verwendung von MAST in klinischen Anwendungen zu beurteilen, habe ich MAST mit der etablierten CRC-Nachweismethode gFOBT verglichen, die auf das Vorhandensein von okkultem Blut im Stuhl abzielt. Die Kombination von gFOBT und MAST führte aufgrund der Komplementarität der erfassten Signatur (d.h. das Messen der Häufigkeit von Mikrobiomspezies gegenüber okkultem Blut) zu einer erhöhten Spezifität in der Validierungskohorte. Zusammenfassend lässt sich sagen, dass meine Arbeit MAST als

vielversprechendes Instrument zur Früherkennung von CRC und als wichtigen Schritt für die mikrobiombasierte Diagnostik in der Klinik etabliert.

In meinem zweiten Projekt untersuchte ich das Mikrobiom mehrerer Körperstellen von 57 Patienten mit duktalem Pankreas- Adenokarzinom (PDAC), 50 gesunden Kontrollpersonen und 29 Patienten mit chronischer Pankreatitis mithilfe von 16S-rRNA-Gensequenzierung, metagenomischer Sequenzierung und fluoreszierender In-situ-Hybridisierung mit dem Ziel, charakteristische Mikrobiom-Signaturen von Bauchspeicheldrüsenkrebs zu etablieren. Während die Gesamtzusammensetzung des Mikrobioms weder in Speichel-, Stuhl- noch im Pankreasgewebe zwischen PDAC und Kontrollen verschoben war, wies das Stuhlmikrobiom starke PDAC-spezifische Signaturen auf, wodurch Krebspatienten von Kontrollen mit hoher Genauigkeit unterschieden werden konnten (AUC = 0,84 basierend auf 22 Mikrobiomspezies). Das Speichelmikrobiom zeigte dagegen insgesamt nur ein schwaches Signal. Ich habe die Spezifität des fäkalen Vorhersagemodells für externe Kohorten (> 3000 Metagenome) validiert, einschließlich gesunder Kontrollen und Patienten, die von Krankheiten wie Typ 1 oder Typ 2 Diabetes, Leberzirrhose, entzündlicher Darmerkrankung und CRC betroffen sind. Darüber hinaus zeigten der Pankreastumor und das angrenzende normale Gewebe deutlich unterschiedliche Mikrobiomprofile. Schließlich wurde das Vorhandensein von Stuhl-verwandten Arten / Gattungen sowohl im Tumor als auch im gesunden Gewebe durch Sequenzierungs- und In-situ-Hybridisierungsmethoden validiert.

Zusammenfassend beschreibe ich einen neuartigen Screening-Test für CRC und verbessere das Verständnis des Mikrobioms bei Bauchspeicheldrüsenkrebs über mehrere Körperseiten hinweg.

List of Figures

Abbreviations

A

Adenomatous Polyposis Coli: (APC)
Amplicon Sequence Variants: (ASVs)
Analysis of Variance: (ANOVA)
Area Under the Curve: (AUC)
Area Under the Receiver Operating Characteristics Curve: (AUROC)

B

Begleitende Evaluierung Innovativer Testverfahren Zur Darmkrebs-Früherkennung: (BLITZ)
Biodefense and Emerging Infections Research Resources Repository: (BEI Resources)

C

Carbohydrate Antigen: (CA)
Chronic Pancreatitis: (CP)
Colorectal Cancer: (CRC)
Control: CTR
Crohn's Disease: (CD)
Cycle Threshold: (Ct)

D

Darmkrebs: Chancen Der Verhütung Durch Screening: (DACHS)
Delta Ct: (Dct)

E

European Nucleotide Archive: (ENA)

F

False Discovery Rate: (FDR)
False Positive Rate: (Fpr)
Fecal Immunochemical Test: (FIT/iFOBT)
Fluorescence *in Situ* Hybridization: (FISH)
Food and Drug Administration: (FDA)

G

Gastrointestinal Tract: (GIT)
Genomic DNA: (gDNA)
German Cancer Research Center: (DKFZ)
Global Microbial Gene Catalog v1: (Gmgcv1)

Guaiac Fecal Occult Blood Test: (gFOBT)

I

Inflammatory Bowel Disease: (IBD)

L

Liquid Nitrogen: (LN2)
Liver Cirrhosis: (LC)

M

Mannose-Binding Lectin: (MBL)
Metagenomics of the Human Intestinal Tract (Metahit)
Microbial Abundance-Based Stool Test: (MAST)

O

Operational Taxonomic Units: (Otus)

P

Pancreatic Cancer: (PC)
Pancreatic Ductal Adenocarcinoma: (PDAC)
Permutational Multivariate Analysis of Variance: (PERMANOVA)

Q

Quantitative PCR: (qPCR)

R

Receiver Operating Characteristic: (ROC)

S

Short Chain Fatty Acids: (SCFA)
Single Nucleotide Variants: (SNVs)
Standardized Incidence Ratio: (SIR)

T

The Human Microbiome Project: (HMP)
The National Center for Biotechnology Information: (NCBI)
Type 1 Diabetes: (T1D)
Type 2 Diabetes: (T2D)

U

Ulcerative Colitis: (UC)

The Hidden World of Microbiome

1. Microbiome in human health and disease:

This chapter presents an overview of the current knowledge in the fast-paced field of microbiome and the relationship of it with cancer progression and detection. It is written by me as a result of literature search.

Recent technological advances have transformed classical microbiology towards more high throughput studies and allowed us to investigate more complex communities. *Microbiota* is a general term to summarize a community of organisms in a given region, such as an environment or a specific body site. The genomes and functional complements of all organisms in a microbiota are collectively referred to as *microbiome*.

The microbiome plays an essential role in maintaining human health, via multiple direct and indirect ways. The microbiome residing in the gastrointestinal tract (GIT), the organ with the highest bacterial load, is in particular known to be involved in nutrient adsorption, metabolite synthesis, detoxification and immune system modulation and regulation (Jandhyala et al. 2015; O'Hara & Shanahan 2006).

The gastrointestinal microbiome is dominated by bacteria, but also contains archaea, protists, viruses and fungi, creating a very complex and diverse community. Microbiome composition is affected by several factors that can be roughly classified in four main groups: host-intrinsic factors (host genetics, gender, immune system, BMI), host-extrinsic factors (diet, geography, transit time, medication, birth delivery mode), environmental (local environment and its interaction with host) and microbiome-intrinsic factors (age, disease status, confounding and stochastic effects) (Thomas S B Schmidt et al. 2018). These factors altogether create a complex system of variables, in which some are more interconnected than others, and make each individual unique. This inter-individual compositional variation in the gut may underlie differential responses to external stimuli, including for example some cancer therapies (Gharaibeh & Jobin 2018; Gopalakrishnan et al. 2018; Matson et al. 2018; Roberts et al. 2013; Routy et al. 2018; Wallace et al. 2015).

The gut microbiome is stable over time in terms of dominant phyla, which include Bacteroidetes, Firmicutes, Actinobacteria and Proteobacteria, provided that lifestyle and diet remains constant and no antibiotic is administered (Martinez-Guryn et al. 2019; Rodríguez et

al. 2015). However, despite the continuous efforts in the last decade, mostly via large-scale microbiome projects such as The Human Microbiome Project (HMP) and the Metagenomics of the Human Intestinal Tract (MetaHIT) (The MetaHIT Consortium & Ehrlich 2011; Turnbaugh et al. 2007), the precise definition of a healthy microbiome remains elusive. Since a myriad of factors influence the composition of the gut microbiome (see above), the high variability among individuals remains the biggest challenge to the identification of "health-specific" microbial signatures.

Several diseases including inflammatory bowel diseases (IBD), (Chong et al. 2019; Halfvarson et al. 2017), obesity (Maruvada et al. 2017; Wellen & Hotamisligil 2005), type 1 and type 2 diabetes mellitus (Jamshidi et al. 2019; Sharma & Tripathi 2019; Turnbaugh et al. 2009), several neurodegenerative disorders (Jangi et al. 2016; Shen & Ji 2019), cardiovascular diseases (Ahmadmehrabi & Tang 2017; Czesnikiewicz-Guzik & Müller 2018), rheumatoid arthritis (Konig 2020) and several malignancies (Schwabe & Jobin 2013; Scott et al. 2019; Sears & Garrett 2014) such as colorectal (Marchesi et al. 2011), liver (Schwabe & Greten 2020), stomach (Rajilic-Stojanovic et al. 2020), pancreas cancers (R. M. Thomas & Jobin 2020) have been associated with gut microbiome composition. Similarly, microbial communities in the human oral cavity have been linked with several diseases (Hall et al. 2017; Schwabe & Jobin 2013).

Cancer is a multifactorial disease affected by both genetic and environmental components. There is increasing evidence that commensal microbes may be a contributing factor to cancer development. Several studies found a strong association between microbial dysbiosis and cancer, suggesting that specific bacterial species can directly and indirectly modulate tumorigenesis (Bose & Mukherjee 2019; Clemente et al. 2012; Gentile & Weir 2018; Helmink et al. 2019; Kåhrström et al. 2016; Kamada et al. 2013; Lynch & Pedersen 2016; Sekirov et al. 2010; Shreiner et al. 2015), by producing toxins or tumorigenic molecules and by

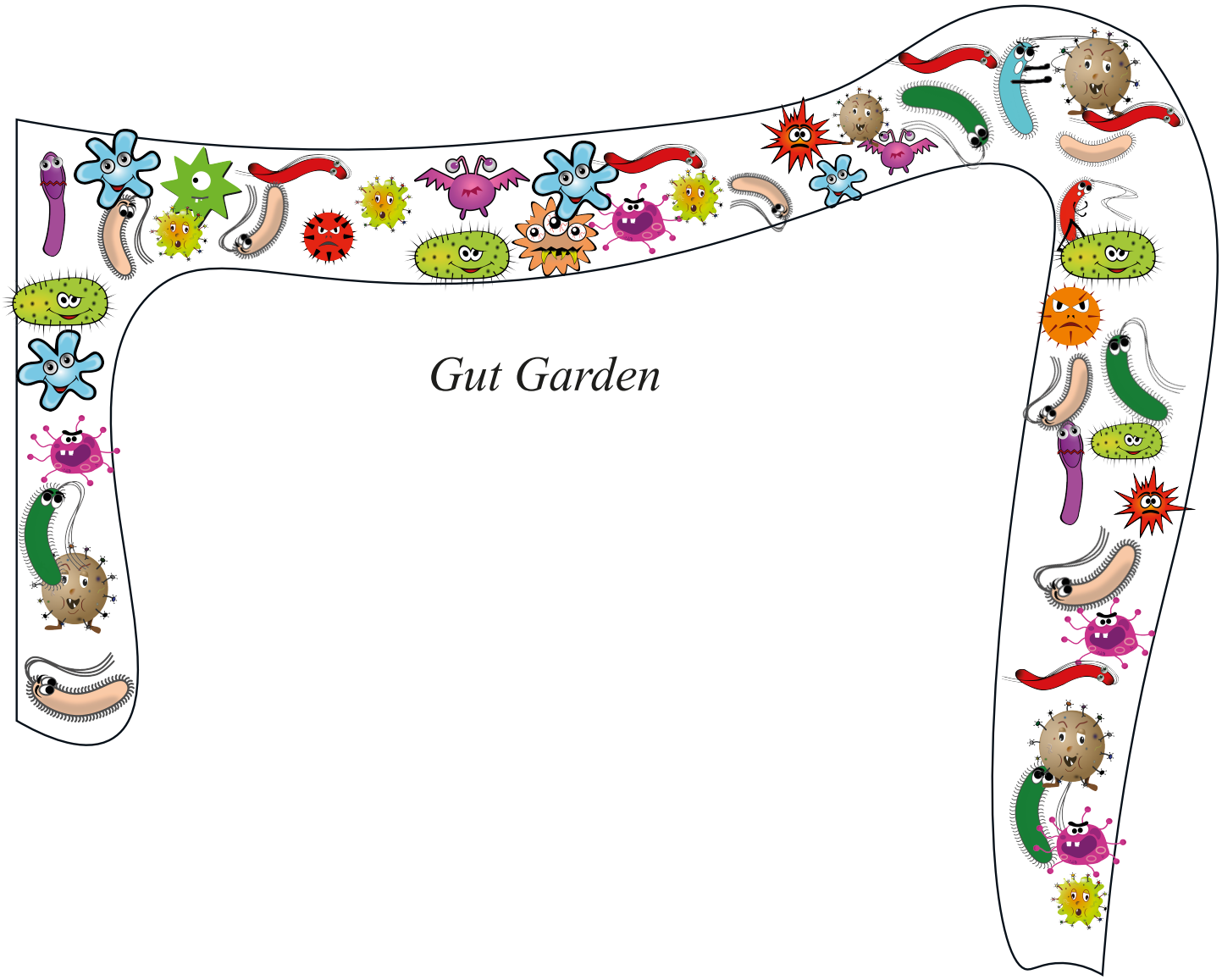
causing severe inflammation or immune suppression (Cani & Jordan 2018; Garrett 2019; Zitvogel et al. 2015, 2018). However the gut microbiome can also increase the host anti-tumor immunity via the production of short chain fatty acids (SCFA) like butyrate and lipopolysaccharides (LPS) (Zitvogel et al. 2017).

Most of the mentioned microbiome studies have reported associations rather than causation (Thomas S B Schmidt et al. 2018). This is due to the limitations of technical and biological knowledge. For instance, human studies are much more valuable than animal ones, given the considerable dissimilarities at the anatomical level (Nguyen et al. 2015). In human studies, detailed longitudinal or interventional designs are essential to establish causes and effects, but this kind of study design is costly and not widely adopted. Additional limiting factors include a lack of fine-grained functional annotations for most known species and limited taxonomic resolution (D'Argenio 2018). Moreover, focusing only on community-wide compositional changes may provide limited information, since microbial signatures can be limited to a small number of species or genes instead of the entire community (Zeller 2014).

Some species are notoriously linked with specific types of cancer; for example, *H. pylori* can cause gastric cancer by producing cytotoxins (Meng et al. 2015; Peek & Blaser 2002). Another organ routinely exposed to bacteria is the liver, with fundamental functions such as filtering blood, metabolizing drugs, detoxifying chemicals and also secreting bile acids to the gut. The liver is also directly linked with the gut via the so-called gut-liver axis. In some individuals, this connection may lead to inflammation and hepatotoxicity, important risk factors for carcinogenesis (Gupta et al. 2019; L.-X. Yu & Schwabe 2017). For instance, secondary bile acid conversion in liver sourced by some bacteria might cause DNA damage in liver cells leading to cancer (Yoshimoto et al. 2013). Colorectal cancer (CRC) has been associated with community-wide changes in gut microbiome composition, and for some species there is evidence of causal effects (Alhinai et al. 2019; Pleguezuelos-Manzano et al. 2020). Such

community-wide shifts (in this case in both in the oral cavity, pancreas and gut) have been described also in pancreatic cancer (PC) (R. M. Thomas & Jobin 2020)

During my PhD, I have investigated the association between the human microbiome, in particular oral cavity and gut, and different types of cancer and the identification of cancer-specific markers for early diagnostic purposes. In particular, my efforts have been towards the study of colorectal and pancreatic cancer. The next two chapters will be based on those projects leaded by me.


Gut Garden

2. The Fecal Microbiome as a Colorectal Screening Tool: *MAST: Microbial Abundance-based Stool Test*

This chapter discusses one of my main projects during the PhD and describes follow up analyses to understand to which extent bacterial species could be used for early detection and screening of colorectal cancer. Based on these analyses, I developed a qPCR-based test for early CRC detection, which we have termed *Microbial Abundance-based Stool Test (MAST)*. I have received support from collaborators, specifically Dr. Georg Zeller, Dr. Vladimir Benes, and Dr. Marja Driessen (all from EMBL). Most of the wet lab work and analyses reported here are unpublished and performed by me. A manuscript is currently under preparation based on this chapter.

I also contributed to a sub-project which resulted in to a paper, in collaboration with Dr. Georg Zeller and Jakob Wirbel and a review with Tilman Kühn and Solomon A Sowah (see below).

Wirbel, J., Pyl, P.T., **Kartal, E**. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nature Medicine 25, 679–689 (2019). https://doi.org/10.1038/s41591-019-0406-6

Solomon A Sowah, Lena Riedl, Antje Damms-Machado, Theron S Johnson, Ruth Schübel, Mirja Graf, **Ece Kartal**, Georg Zeller, Lukas Schwingshackl, Gabriele I Stangl, Rudolf Kaaks, Tilman Kühn, Effects of Weight-Loss Interventions on Short-Chain Fatty Acid Concentrations in Blood and Feces of Adults: A Systematic Review, *Advances in Nutrition*, Volume 10, Issue 4, July 2019, Pages 673–684, https://doi.org/10.1093/advances/nmy125

2.1. Introduction

Colorectal cancer (CRC) is the third most common cancer in westernized societies, with 2.5 million new cases and 1.2 million deaths expected worldwide by 2030 (https://gco.iarc.fr/tomorrow/). The genetic component is less than 20% (Carethers & Jung 2015). In addition to moderate genetic factors, such as mutations in the Adenomatous Polyposis Coli (APC) gene, an altered Wnt-signalling and DNA mismatch repair system (Loktionov 2020), there are other risk factors including inflammatory bowel diseases like CD or UC, obesity, alcohol consumption, tobacco usage, lack of physical activity and dietary intake (Moskal et al. 2016).

CRC progression is a multi-step progress which may take more than 10 years from normal epithelium to adenoma/carcinoma, so that early detection has a large impact on treatment opportunities and patient survival. Diagnosis at an early stage of the tumorigenesis results in a 5-years survival rate of >80%, however this decreases dramatically to < 20% as the disease advances (Siegel et al. 2017). Change in bowel habit, rectal bleeding, fatigue, weakness, weight loss, discomfort, pain or gas in bowel are some of CRC symptoms; however, they appear late during the disease's progression, emphasizing the importance of population screening.

Currently, colonoscopy is the "golden standard" with very high specificity, but the procedure is invasive, time-consuming, expensive and might cause distraction on gut microbiome and severe complications as well, limiting its wider usage (Rutter et al. 2014). Intense research on CRC biomarkers based on DNA, RNA, proteins, metabolites and the gut microbiome has accelerated recently (Coghlin & Murray 2015; Loktionov 2020). So far, guaiac fecal occult blood test (gFOBT), fecal immunochemical test (FIT/iFOBT) and multitarget stool test marketed as Cologuard are the most common and accepted non-invasive screening tests. gFOBT and iFOBT target heme of hemoglobin in stool which is an indicator of inflammation

in gut and these are not suitable to detect early lesions and their sensitivity (0.13-0.79 and 0.69-0.86, respectively) and specificity (0.86-0.98 and 0.92-0.95, respectively) are still low (Lee et al. 2014; Robertson et al. 2017). Cologuard detects abnormal DNA in stool and it is an enhanced version of iFOBT, however it is very expensive, requires sophisticated lab equipment and technical challenges still remains to be solved (Loktionov 2020). The prevalence of CRC, the benefits of early testing, and the limitations of current tests all highlight the importance of developing more sensitive and affordable screening tests for CRC.

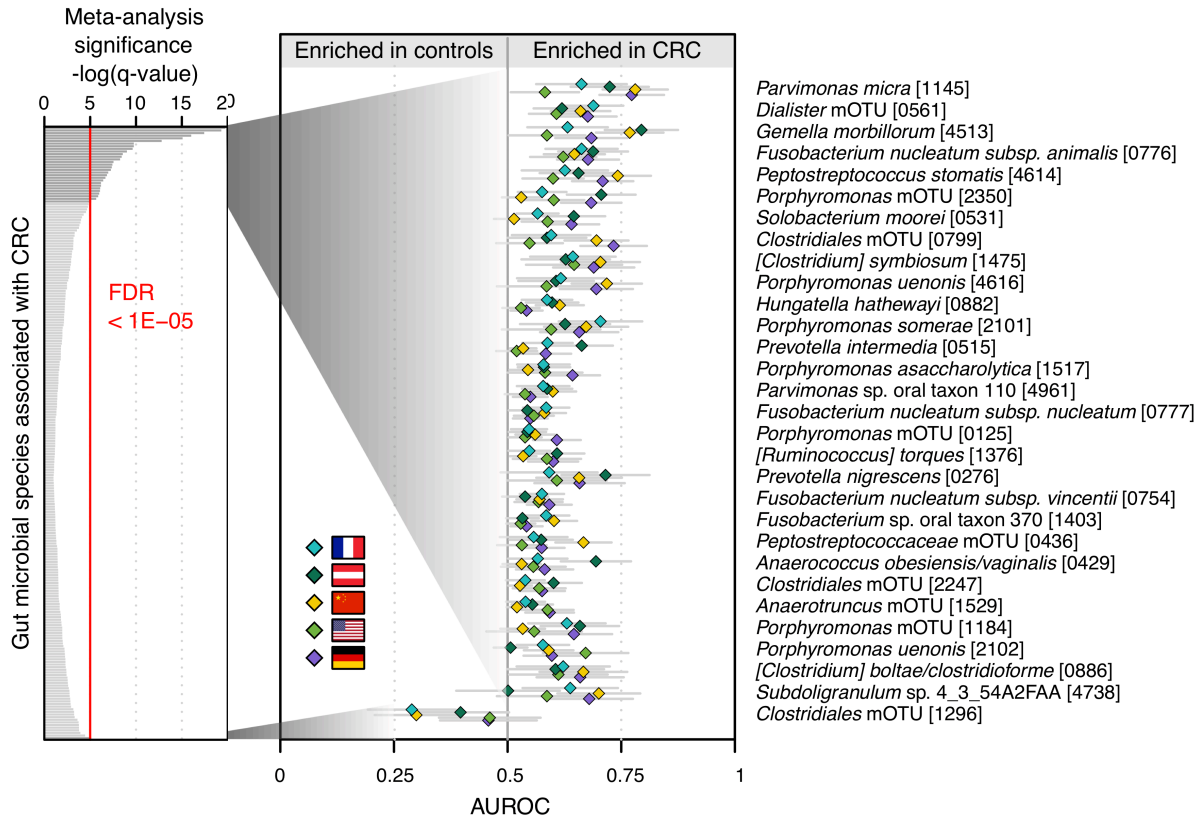


Figure 2.1 "Despite study differences, meta-analysis identifies a core set of gut microbes strongly associated with CRC.

a. The meta-analysis significance of gut microbial species derived from blocked Wilcoxon tests (n = 574 independent observations) is given by the bar height (FDR = 0.005) AT, Austria; CN, China; DE, Germany; FR, France; US, United States. **c.** For a core of highly significant species, association strength is quantified by the AUROC across individual studies, and the 95% confidence intervals are indicated by the gray lines. Family-level taxonomic information is color-coded above the species names." This figure is published in Wirbel et al., 2019 which I was also an author and it was produced by Jakob Wirbel.

Changes in the gut microbiome may create a microenvironment favorable for carcinogenesis (Gagnière et al. 2016; Kostic et al. 2013; Sheng et al. 2020; Tjalsma et al. 2012; Wang et al. 2012). Several studies confirmed the differential enrichment of bacteria between tumor and nontumor tissue (Allali et al. 2015; Flemer et al. 2017; Gao et al. 2017) and explored the CRC tumor virome and mycobiome (Coker et al. 2019; Nakatsu et al. 2018), however due to the challenges in targeting these biomes, their limited taxonomic resolution and their lower abundance relative to bacteria, research groups have mostly focused on role of bacteria and archaea.

Even though the fecal microbiome is only a proxy for gut microbiome, the possibility of non-invasive screening test has favored fecal-based research (Konstantinov et al. 2013; Rezasoltani et al. 2018; J. Yu et al. 2017). Numerous studies have compared fecal microbiome differences between CRC cases and controls. Though there was generally no overall community shift observed, significantly enriched or depleted species showed a link with CRC status (Ahn et al. 2013; Dai et al. 2018; Drewes et al. 2017; Feng et al. 2015; Marchesi et al. 2011; Mira-Pascual et al. 2015; Saito et al. 2019; N. Wu et al. 2013; Zeller et al. 2014) (Table 2.1, Figure 2.1). Fecal microbiome analysis showed increased levels of specific bacterial species in CRC cases, including *Fusobacterium nucleatum, Enterococcus faecalis, Escherichia coli, Bacterioides fragilis, Peptostreptococcus stomatitis* and *Parvimonas micra,* whereas other taxa were depleted, such as *Clostridiales, Faecalibacterium, Blautia,* and *Bifidobacterium* (Wong et al. 2017; J. Yu et al. 2017; Zeller et al. 2014; Zhang et al. 2019; Zou et al. 2018) (Table 2.1, Table 2.2).

Though these results are promising, the identified discriminative species varied by study. This variance is possibly due to cohort differences (e.g. due to geography, diet) as well as technical noise (e.g. sample quality, storage, extraction method, sequencing, choice of bioinformatics tools) (Paul I Costea et al. 2017; Duvallet et al. 2017; Villéger et al. 2018). To

diminish the effect of study-specific and background noise, a meta-analysis of publicly available CRC cohorts was performed and showed that there is a core set of CRC-associated microbes, independent of study effects, diet, geography or other factors, which have an impact on gut microbiome (Dai et al. 2018; Shah et al. 2018; A. M. Thomas et al. 2019; Wirbel et al. 2019).

The aim of this project in my PhD was to narrow down previously reported microbiome signals to target CRC by a non-invasive, cheap and fast method. It was mainly based on previous work (Zeller et al. 2014), which showed that a combination of 22 species gives a more accurate classification than any single taxon. This classifier supports the hypothesis that community composition is important to CRC development. Of these 22 species, four contribute disproportionately to the classifier, these being two *Fusobacterium species, Porphyromonas asaccharolytica* and *Peptostreptococcus stomatis,* which are patented under the EP2955232A1 number (https://patents.google.com/patent/EP2955232A1) (Zeller et al. 2014). Based on these findings, I have developed a quantitative PCR (qPCR) based non-invasive, rapid and low-cost CRC screening method, the Microbial Abundance-based Stool Test (MAST).

Table 2.1 Summary of CRC-related microbiome studies

Study	Country	Cohort size	Specimen	Method	Microbial change
Scanlan 2008	Belgium	20 CRC/ 20 Polyp/ 20 CTR	Stool	16s rRNA sequencing	↑ diversity of the *C. leptum* and *C. coccoides* subgroups
Sobhani 2012	France	60 CRC / 119 CTR	Stool	16s rRNA sequencing	↑ *Bacteroides, Prevotella*
Castellarin 2012	Canada	11 CRC with normal tissue	Tissue	RNA sequencing	↑ *Fusobacterium*
Kostic 2012	USA	95 CRC with normal tissue	Tissue	16s rRNA sequencing	↑ *Fusobacterium*
Sanapareddy 2012	USA	33 Adenoma / 38 CTR	Tissue	16s rRNA sequencing	↑ bacteria from 87 taxa
Chen 2012 Wang 2012	China	46 CRC/ 56 CTR	Tissue, stool, swab	16s rRNA sequencing	↑ *B. fragilis, Lactobacillales, Fusobacterium, Porphyromonas, Peptostreptococcus, Mogibacterium* ↓ diversity, *Bifidobacterium, Faecalibacterium*, Blautia butyrate-producing bacteria
Ahn 2013	USA	47 CRC/ 94 CTR	Stool	16s rRNA sequencing	↓ diversity, *Clostridia;* ↑ *Fusobacterium, Porphyromonas*
Dejea 2014	USA	23 CRC/ 2 Adenoma / 22 CTR	Tissue	16s rRNA sequencing	↑ *Fusobacterium*; difference by biofilm status
Zackular 2014	USA	30 CRC/ 30 Adenoma/ 30 CTR	Stool	16s rRNA sequencing	↑ *Bacteroides fragilis, Fusobacterium, Porphyromonas* ↓ butyrate-producing bacteria
Zeller 2014	France	91 CRC/ 42 Adenoma / 358 CTR	Stool	Metagenomic sequencing	↑ *Bacteroidetes, Fusobacteria and Proteobacteri;* ↓ *Actinobacteria and Firmicutes*
Burns 2015	USA	44 CRC	Tissue	16s rRNA sequencing	↑ diversity, *Fusobacterium* and *Providencia* ↑ virulence-related genes

Feng 2015	Canada	41 CRC/ 42 Adenoma/ 55 CTR	Stool	Metagenomic sequencing	↑ *B dorei, B vulgatus, E coli, Fusobacterium*; ↓ *Lactobacillus* and *Bifidobacterium*
Lu 2016	China	31 Adenoma / 20 CTR	Tissue	16s rRNA sequencing	↑ diversity, *Lactococcus* and *Pseudomonas* ↓ *Enterococcus*, *Bacillus* and *Solibacillus*
Vogtmann 2016	USA, France	52 CRC/ 52 CTR	Stool	Metagenomic sequencing	↑ *Fusobacterium, Porphyromonas, Clostridia*
Baxter 2016	USA, Canada	120 CRC/ 198 Adenoma / 172 no colonic lesions	Stool	16s rRNA sequencing	↑ *P. assaccharolytica, P. stomatis, P. micra, F. nucleatu*m ↓ *Lachnospiraceae*
Hale 2017	USA	233 Adenoma /547 CTR	Stool	16s rRNA sequencing	↑ *Bilophila, Desulfovibrio, Mogibacterium*; ↓ *Veillonella, Clostridia*
Yu 2017	Denmark, France, Austria	74 CRC / 54 CTR	Stool	Metagenomic sequencing	↑ *P. stomatis, F. nucleatum, P. micra, S. moorei*
Liang 2017	China	203 CRC / 236 CTR	Stool	16s rRNA sequencing	↑ *F. nucleatum, Clostridium hathewayi* ↓ *B. clarus*
Flemer 2018	Ireland	43 CRC / 37 CTR	Stool, tissue	16s rRNA sequencing	↓ *L. incertae sedis and Coprococcus*
Nakatsu 2018	Hong Kong, Austria, France, Germany	248 CRC/ 241 CTR	Stool	Shotgun sequencing	↑ *Orthobunyavirus*
Sheng 2019	China	67 CRC/ 30 CTR	Stool	16s rRNA sequencing	↑ *Prevotella, Collinsella, Peptostreptococcus* ↓ *Escherichia-Shigella*
Coker 2019	Hong Kong	184 CRC/ 197 Adenoma/ 204 CTR	Stool	Shotgun sequencing	↑ *Malasseziomycetes* ↓ *Saccharomycetes, Pneumocystidomycetes*
Yackida 2019	Japan	54 CRC/ 127 CTR	Stool	Shotgun sequencing	55 bacterial markers
Sarhadi 2020	Iran, Finland	83 CRC/ 60 CTR	Stool	16s rRNA sequencing	↑ *Prevotella, Clostridium, Ruminococcus* ↓ *Bacteroides*
Liu 2020	China	59 Polyp/ 54 Adenoma/ 51 CRC/ 42 CTR	Stool	16s rRNA sequencing	↓ *Bacteroidetes, Firmicutes*

Table 2.2 The role of intestinal bacteria in CRC.

Bacteria	Epidemiologic Evidence	Potential mechanisms	References
Associated with higher risk of CRC			
Fusobacterium nucleatum	↑ tumor tissue and fecal samples of patients lower infiltration by T cells higher risk of recurrence poorer patient survival	Tumor-permissive microenvironment Modulation of Ecadherin/β-catenin	(Kostic et al. 2013; Rubinstein et al. 2013)
Enterotoxigenic *Bacteroides fragilis* (ETBF)	↑ tumor tissue and fecal samples of patients associated with advanced cancer stage and proximal colon tumor.	DNA damage	(Goodwin et al. 2011; S. Wu et al. 2003)
pks+ *Escherichia coli*	↑ tumors than in normal flanking tissue ↑ late-stage tumors	Intestinal inflammation	(Pleguezuelos-Manzano et al. 2020)
Associated with lower risk of CRC			
SCFA-producing bacteria	↓ CRC patients than controls ↑ Native Americans with a low CRC incidence	Modulation of SCFAs	(Richard et al. 2018)

2.2. Results

We developed a microbial abundance-based stool test (MAST) for early detection of CRC; which targets selected species and strains based on previous knowledge (Zeller et al. 2014). We received fecal samples collected at the German Cancer Research Center (DKFZ) including 56 CRC patients from the ColoCare cohort and 77 gender- and age-matching healthy controls from the PRÄVENT cohort. Based on the ridge regression model we developed, we obtained an AUC of 0.87 when discriminating CRC patients from healthy controls.

To further validate our findings, we included another independent cohort collected at the same institute, composed of 37 CRC patients and 272 healthy individuals from the 'Begleitende Evaluierung innovativer Testverfahren zur Darmkrebs-Früherkennung' (BLITZ) study and an additional 35 CRC cases from the 'Darmkrebs: Chancen der Verhütung durch Screening' (DACHS) study (Brenner et al. 2014). Validation of the MAST qPCR model on this combined cohort (DE2) showed high accuracy as well. Moreover, we have tested the combination of gFOBT and MAST results in the DE2 cohort since gFOBT results were available only in this cohort. The combination increased accuracy and sensitivity, demonstrating that microbiome and gFOBT signals are partly complementary.

2.2.1. qPCR results per species

To quantify the relative abundances of selected species/strains to total DNA amount in a given samples, qPCR was performed in triplicates targeting genomic DNA (gDNA) for all selected species and 16S rRNA gene as proxy of total DNA. A universal primer for the 16S rRNA gene (Ritalahti et al. 2006), which was suitable for qPCR method was used to normalize cycle threshold (Ct) values. The SPUD assay was implemented as an inhibitor indicator. All primer and probe pairs shown in Table 2.3 had positive and negative controls in the same plate

to judge the reaction efficiency and to detect any potential contamination. We furthermore used a plate calibrator to detect pipetting errors caused by manual handling or the liquid handler. The detailed workflow is shown in Figure 2.2.

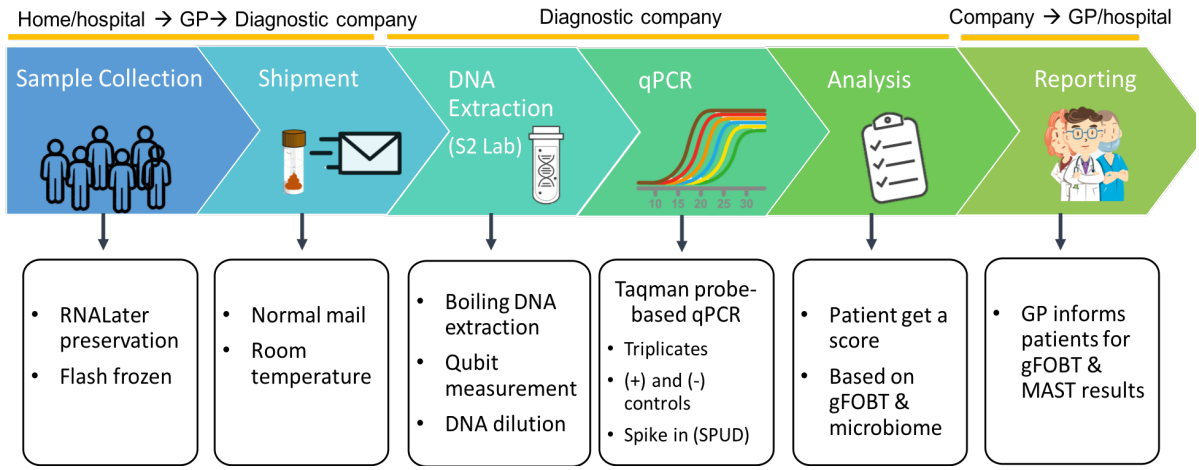


Figure 2.2 MAST workflow step by step.

Sample collection is first step of the process which can be done in either way. DNA extraction is done by boiling extraction method (see method section 2.4) as a next step after receiving the fecal material in the lab which requires S2 lab conditions. Normalized dCt values are used to perform machine learning approach and each patient receives a classification score based on their microbiome profile. This figure was produced by myself.

To assess the discriminative power of each species, we first calculated delta Ct (dCt) value based on the difference between a given species' Ct value and 16S rRNA gene. Receiver operating characteristic (ROC) analysis was based on dCt values of each species. We also had shotgun metagenomics data available for most of the individuals from DE1 cohort (Zeller et al. 2014; Wirbel et al. 2019). Some species were showed a very good area under the curve (AUC) score in agreement with both MAST and shotgun metagenomics sequencing, while species enriched in control samples did not have a high AUC for one or both methods (Figure 2.3). This figure also presents that qPCR approach is much more sensitive than shotgun metagenomics approach, especially detection of low abundant species/strains.

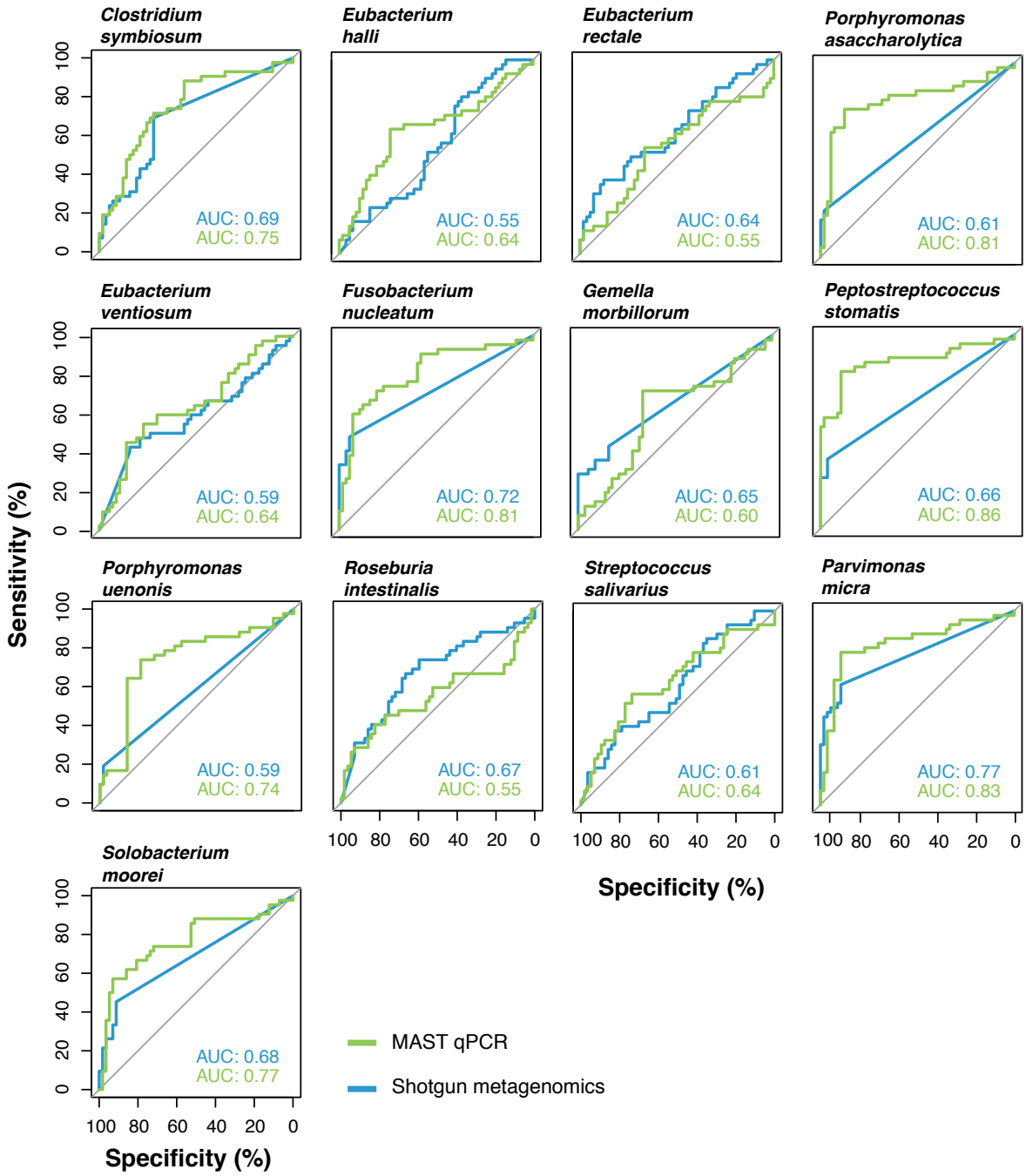


Figure 2.3 Comparison of qPCR and shotgun metagenomics sequencing showed that qPCR approach is more sensitive.

pROC package from R was used to produce ROC plots for each species. Green represents dCt value from MAST qPCR test, while blue shows relative abundance from shotgun metagenomics data for same samples. All shotgun metagenomics data was published in Zeller et al 2014 and Wirbel et al. 2019. This figure is part of the manuscript which is under preparation and all panels were produced by myself.

2.2.2. Combined model performs better than single species

To combine the power of different species information and see the total discriminative power, we applied a ridge regression model. Since our data had more observations (individuals) than predictors (targeted bacteria/strain), ridge regression was selected. The combined model differentiated CRC cases from controls with an AUC of 0.87 based on MAST of 13 species (Figure 2.5). The z-score normalized dCt values across all tested samples in DE1 cohort are shown as heatmap in Figure 2.4.

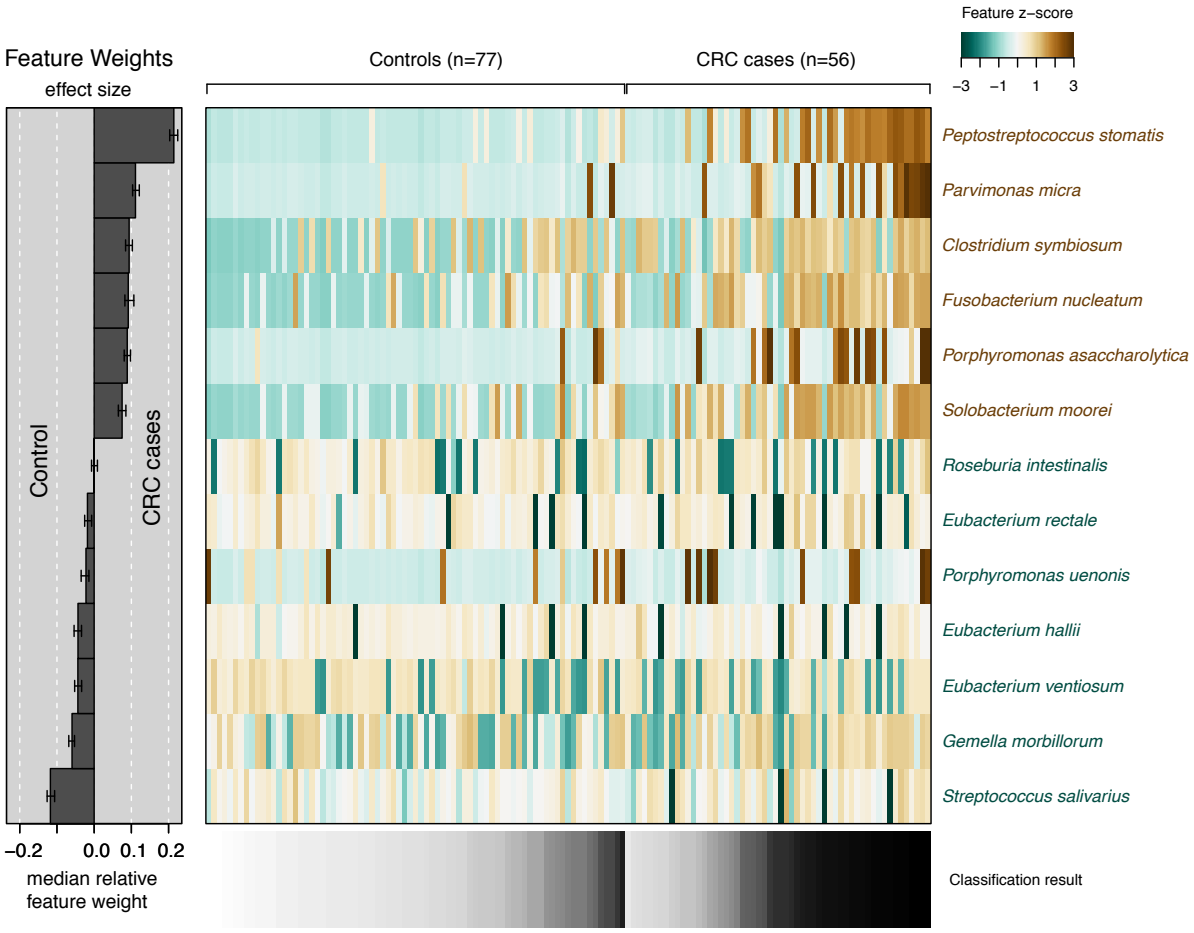


Figure 2.4 Heatmap showing the abundance of species among individuals across DE1 cohort.

z-score normalized dCt values of 13 species in DE1 cohort is shown for 56 CRC and 77 control samples. The left panel represents the main contribution of each selected feature to the overall model and below panel shows the classification score of model for each individual. This figure is part of the manuscript which is under preparation and it was produced by myself

We further decided to lower the number of potential species to decrease handling time and cost of the test in a clinical setup. We removed the species with lower weight contribution to model and we ended up with 7 species with the highest differentiation power, mostly CRC-enriched (Figure 2.5).

When we compared the current model with gFOBT theoretical sensitivity and specificity scores, our model performance was better (Figure 2.5). Regrettably, we did not have the information of gFOBT or iFOBT for our DE1 cohort.

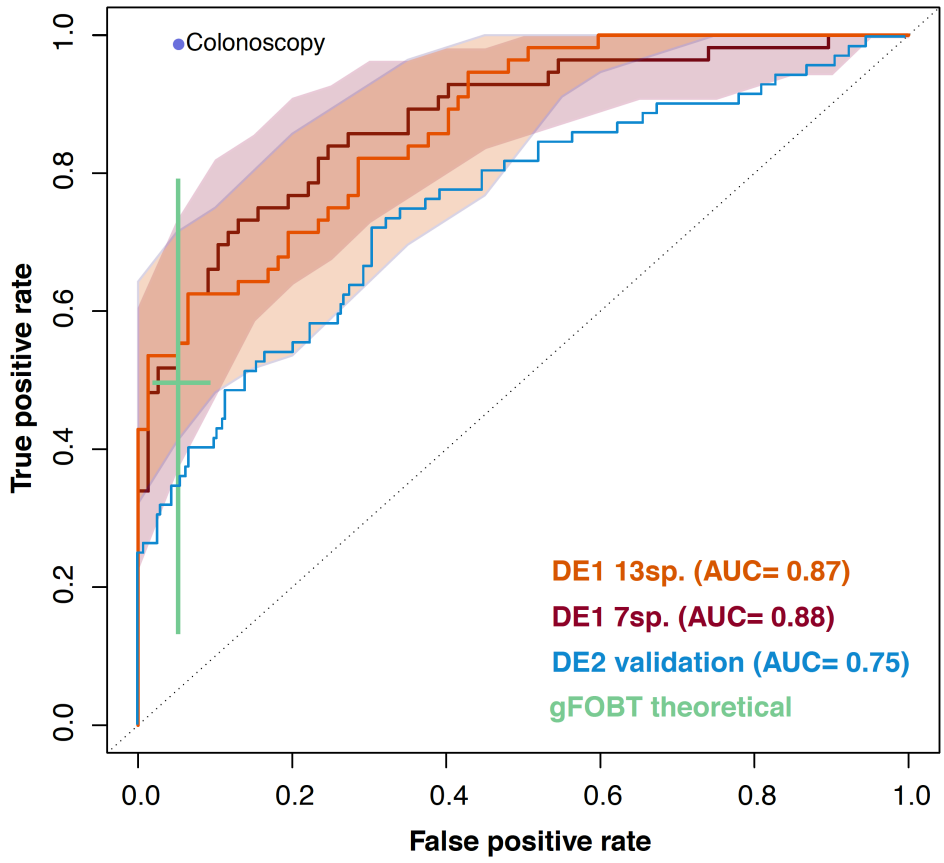


Figure 2.5 ROC curve based on MAST results of DE1 cohort.

Ridge regression model of DE1 cohort based on 13 species is represented by orange, while seven species model Ridge regression model of DE1 cohort based on 13 species is represented by orange, while seven species model is red. Models were trained on z-score normalized dCt values with 10 resampling followed by 10 cross validation. Validation of seven species model on DE2 cohort is displayed with blue line. The theoretical specificity and sensitivity of gFOBT test is shown with light green. 95% confidence intervals of true-positive rate are shaded accordingly. This figure is part of the manuscript which is under preparation and it was produced by myself.

2.2.3. Validation of MAST classifier using DE2 external cohort

To evaluate the accuracy of MAST model, we performed MAST qPCR in an independent cohort (DE2), which consists of 272 controls and 72 CRC cases. The same data generation and normalization procedures were applied. The AUC of 7 species-based MAST model performance on DE2 cohort was 0.75 (Figure 2.5). We also tested 13 species model and the performance of both models was very similar.

Since half of the CRC cases were from another study (BLITZ) in DE2 cohort, we further tested if those samples were different based on their profile, but effects were negligible (AUC=0.6), confirming that there is no difference between those 2 different study cohort (Figure S2. 1).

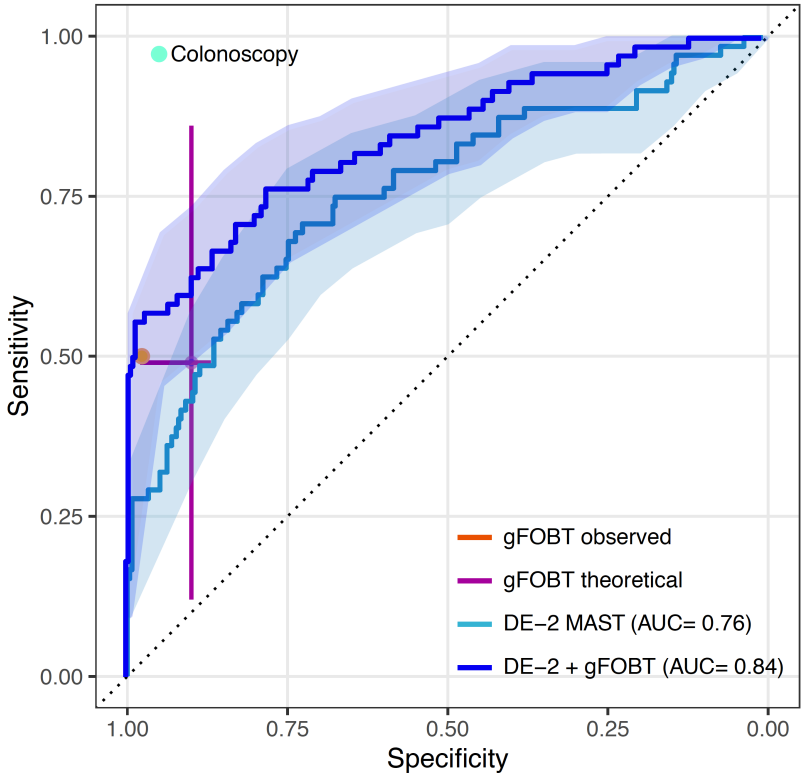


Figure 2.6 Combination of MAST and gFOBT results improves sensitivity in DE2 cohort.

Light blue line represents the model only based on normalized MAST results on DE2 cohort; while dark blue represents the combination of MAST and gFOBT for same cohort. Ridge regression model was trained on normalized dCt values with 10 resampling followed by 10 cross validation. This figure is part of the manuscript which is under preparation and it was produced by myself.

2.2.4. Testing complementary power of gFOBT and MAST results on DE2 study

gFOBT information was only available for the DE2 study. To determine the complementary effect of gFOBT with MAST, we first used only MAST data on DE2 cohort as predictor and then we added gFOBT results as predictor as well (Figure 2.6). Combined model achieved higher accuracy as measured by AUC of mean predictions score (0.84). This supports the previous observations about gFOBT and microbiome signals as capturing different signals and combination increases specificity (Zeller et al. 2014).

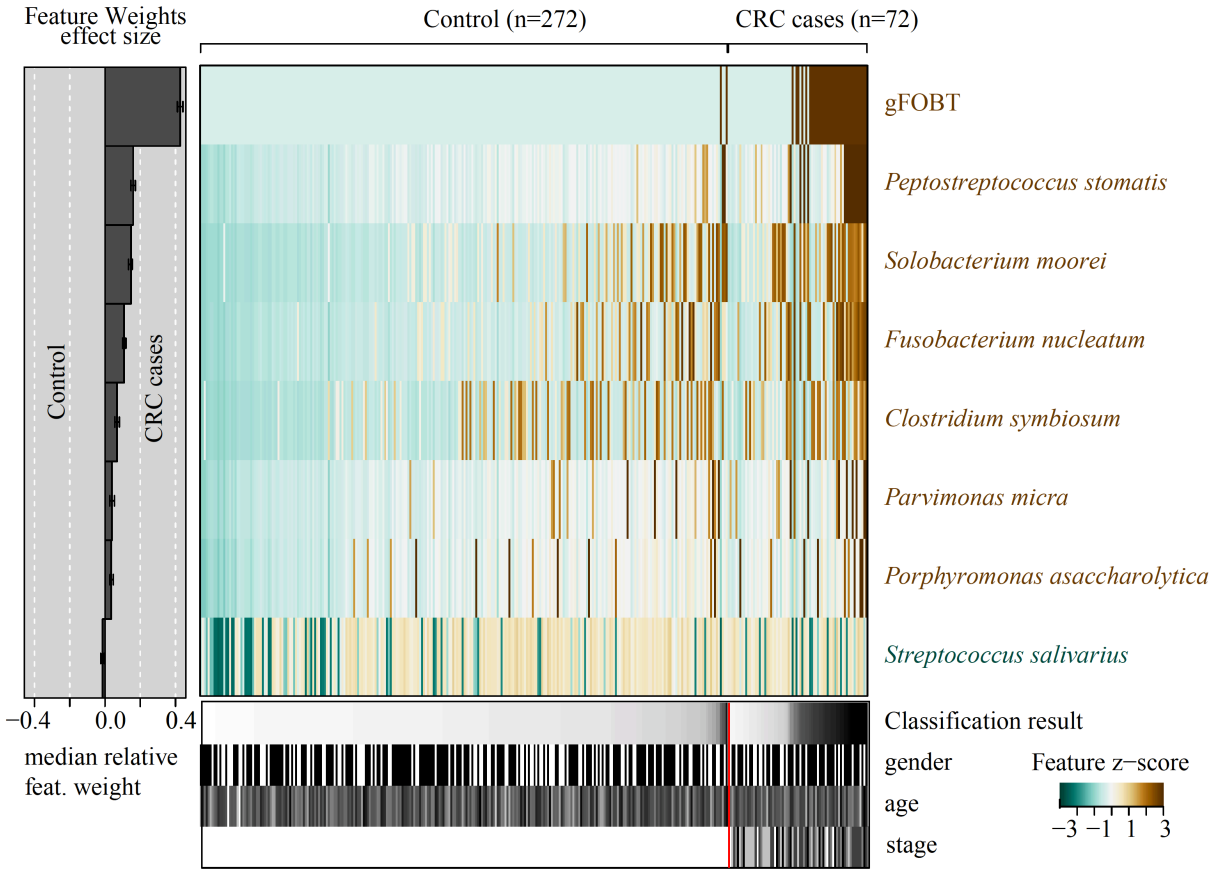


Figure 2.7 Heatmap showing results of MAST and gFOBT across DE2 cohort.

MAST results for selected seven species and gFOBT combination captured different signals which increased sensitivity. Feature weights are shown in left panel and heatmap represents the normalized dCT as abundance of seven species and gFOBT results. The below panel displays metavariables that we checked for potential confounder affect. This figure is part of the manuscript which is under preparation and it was produced by myself.

2.3. Discussion and Perspectives

Colorectal cancer (CRC) is a growing challenge in societies worldwide, and after lung cancer the second most deadliest cancer as measured by the total number of deaths per year worldwide (Keum & Giovannucci 2019). Thus, it is particularly important to improve early detection and strategies for monitoring, understanding and stabilizing disease progression.

Here we propose a microbiome-based screening technique called MAST that is low-cost, rapid and simple to process. It builds upon recent research that revealed a specific microbiome composition for CRC with some significantly abundant species when compared to healthy controls. For MAST, we initially selected 13 species/strains and used qPCR to target the species in the DE1 cohort (56 control and 77 CTR cases). The discriminative power of each species is shown in Figure 2.3. Reassuringly, species enriched in CRC mostly showed high accuracy to distinguish cancer cases from controls. A combination of those 13 species performed very promising in the differentiation of CRC cases from healthy controls with an AUC of 0.87.

Based on the results, we optimized the model and narrowed down the minimum number of species that still retain the signal from 13 to seven including *Clostridium symbiosum, Parvimonas micra, Peptostreptococcus stomatis, Porphyromonas asaccharolytica, Solobacterium moorei, Fusobacterium nucleatum* and *Streptococcus salivarius*. This allowed us to process more samples, decrease costs and handling time and in addition even have a better performance (AUC = 0.88). Next, we validated the accuracy of our seven species-based MAST model on an independent cohort DE2 (DE: 272 control vs. 72 CRC cases) with an AUC of 0.75 (Figure 2.5).

To determine the potential usage of MAST in clinics, in combination with gFOBT/iFOBT as shown in Zeller et al 2014, we wanted to compare and combine MAST

performance with gFOBT and iFOBT. gFOBT and iFOBT are simple and economical options while lacking sensitivity and specificity. Additionally, Cologuard performs better than gFOBT and iFOBT in CRC detection, yet its complexity, required laboratory equipment and high cost makes its usage limited in practice (Loktionov 2020). However, we did not yet have any of these test results available for the DE1 cohort and only got access to the gFOBT results for the DE2 cohort. When we built the models based on the DE2 cohort with and without gFOBT, we noted a big increase in specificity when the gFOBT results were included (AUC= 0.84, Figure 2.5). This strongly indicates that those two tests capture different types of signatures that complement each other (bacterial abundance and occult blood in stool, respectively) and the combination increases overall specificity of CRC detection.

Moreover, we hypothesize that including iFOBT results may even further increase AUC and also specificity more than either iFOBT or MAST alone. To test this in the future, we will conduct more external validation cohorts with gFOBT and/or iFOBT results to have a broader view of MAST applications and overall potential.

Potential improvements include further decreasing its current costs of just below 60 Euro by multiplexing primers to decrease labor costs and consumables or using a liquid handler to reduce overall handling time. Possible combination with gFOBT/iFOBT would also decrease the shipping cost and handling time.

Our study has a number of potential limitations. First, as with all studies, more and bigger external validation cohorts, preferably from different countries and continents to improve understanding of geographic signatures, are required to further evaluate and improve specificity and sensitivity of MAST. The used cohorts are from Germany only, and the possibility of geographic effect cannot be excluded. However, since meta-analyses of several available metagenomics data also demonstrated that CRC-related microbiome signal is present across all compared cohorts independent of geography, diet or technical differences (Dai et al.

2018; Shah et al. 2018; A. M. Thomas et al. 2019; Wirbel et al. 2019), this may not even be a limiting factor. Therefore, the developed early CRC screening method based on fecal microbiome may be used worldwide. Second, iFOBT results are yet missing in both cohorts as well as gFOBT results in the first cohort. Third, the in-silico model may also be further improved by adding or changing the species composition, generally more data to allow a better training of the data and estimation of parameters. In addition, since the current model is only based on species abundance, adding potential functional information such as the absence or presence of particular marker genes or their expression levels as well as merging taxonomic information may boost the model even further. Fourth, MAST currently targets only general CRC signatures but not individual subtypes. With more data, it may be possible to detect subtype-specific signatures such as those arising from MSI (CMS1), CIN (CMS2, CMS3, CMS4). However, the microbiome of the various subtypes has not been studied extensively either. In addition, stage-specific signatures may be detected with the availability of more data in order to build the model. Both extensions would significantly enhance the applicability and scope of the MAST test.

In summary, we aimed to develop an assay that can be used in clinics to predict early CRC cases and that has advantages over the current methods. Indeed, our results collectively suggest that MAST may be generally used in clinics in combination with gFOBT/iFOBT to increase detection accuracy of CRC.

▪

2.4. Methods

2.4.1. Overview of the samples and dataset included

Written informed consent was obtained from all CRC patients and healthy controls. The study was approved by the institutional review board (EMBL Bioethics Internal Advisory Board) and the ethics committee of the Medical Faculty at the University of Heidelberg. The study agrees with the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.

The first German (DE1) cohort consist of 56 CRC patients, recruited for ColoCare study (DKFZ, Heidelberg, (Böhm et al. 2017; Liesenfeld et al. 2015); 38 of which were published in (Zeller et al. 2014) with ENA accession ERP005534. All fecal samples were collected after colonoscopy. 77 gender- and age-matched participants of the PRÄVENT study from same institute were included as healthy controls; however, those individuals did not have colonoscopy examination, so the presence of undiagnosed carcinomas cannot be completely excluded. Considering the low prevalence of preclinical CRC in the general population, we still considered those individuals as healthy (Pox et al. 2012). 60 fecal metagenome of control samples were published with ENA accession (Wirbel et al. 2019)

We included a second cohort used as external validation recruited by same investigators (DKFZ, Heidelberg). 35 CRC patients were recruited from a population based study called Darmkrebs: Chancen der Verhütung durch Screening (DACHS) study (Brenner et al. 2014). 272 control and 37 CRC fecal samples were received from a screening cohort named Begleitende Evaluierung innovativer Testverfahren zur Darmkrebs-Früherkennung (BLITZ), which fecal samples were collected before colonoscopy (https://www.dkfz.de/en/imed/BLITZ.html).

2.4.2. DNA extraction procedure

DNA was extracted using 0.1 mg/20µl of stool by boiling extraction method, which has been also tested internally at EMBL. All the samples were thawed and 2ml saline solution (0,15M NaCl) was added after vortexing. After centrifugation at 20000 rpm for 2 min, liquid part was discarded and we added 500µl saline solution to bring pellet back. Biomixer with stirring rods was used to homogenize and then we used 70µM cell strainer to filter sample. 200µl solution was taken from filtered material and 800µl H_2O was added before vortexing. After mixing and dissolving pellet completely, samples were incubated at 99°C for 15 min to break all the cells. This extracted material was kept at -20°C until processing time. All samples were measured by Qubit fluorometer (Thermo Fisher) and diluted to 5ng/µl concentration to be able to use liquid handler for next steps. All prepared dilutions were measured again to be precise about concentration.

2.4.3. Species selection and primer design protocol

Species selection was based on a previous study (Zeller et al. 2014) which proposed 22-species metagenomic classifier for CRC prediction and also literature search for CRC related bacteria. In total, we selected 13 species as target species and we used *Bacteroides vulgatus* as internal control (Table 2.3). Primer design was aiming to target 10 marker genes and we used Primer3 software (http://bioinfo.ut.ee/primer3-0.4.0/) and the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/tools/primer-blast/) primer designing tools in combination and all designed primers were sequenced by sigma (https://www.sigmaaldrich.com/). Probes were also design by those tools or PrimerQuest tool (https://www.idtdna.com/pages/tools/primerquest) and were sequenced by IDT company. All

probes were 5′ nuclease probes with a fluorescent reporter dye and quencher FAM/ZEN/3' IBFQ. All the primer and probes were then diluted to 5NM/µl concentration.

Table 2.3 The list of targeted species, primer pairs and probe sequence.

Species	Size	Primer	Taqman probe
Bacteroides vulgatus	91	F-CAAGCTGAGAAAAGCAGCCAAA	AGTGGCAGTAGCCGGAGGGGTATCAGCCA
		R-GAATGAGTTACGAAGCCCGTTG	
Clostridium symbiosum	204	F-GTGAGATGATGTGCCAGGC	CACGCAAATATGATGCTGAAACCGGTG
		R-TACCGGTTGCTTCGTCGATT	
Eubacterium hallii	164	F-CGGCGTTGCTTCAAACTCTTC	CGCGCGCAGATGGAAGAAGCCTGCCG
		R-CCCAGCCATCACGAATACCTT	
Eubacterium rectale	180	F-CTCACACGCAGTACAGGAGG	AGCATGAGGAGCAGGACGGCTTCTTCTGGCA
		R-CCTTCCTCGCCAACTACAGG	
Eubacterium ventiosum	81	F- GTCGGACGCTAATTGCACA	AGATGCAATAGGTGTTACATATGGACCGG
		R-ACCAATAAGGCTCCAACTAAGC	
Fusobacterium nucleatum	101	F-AGTGCAGGTGATACATTTAGAG	TGCCCATTCTTCAAGTTGTTCAACTGC
		R-GATCAGCTCCTTCTCTTCCTTT	
Gemella morbillorum	101	F-CGTCGTAGTCTTCTTGTTGAGG	ACCGCAATTTCTTCACTACGTTTGTTGC
		R-TCAGTAGCATCTTCACGCTTAC	
Parvimonas micra	110	F-CAGAGCCAGCTTGGGAAA	AGACCCGGATGGCATCTTGAATGT
		R-ACCCGCATGAATGTCTATTGTA	
Peptostreptococcus stomatis	99	F-CAGAGGATGTCTGTGCATCTT	AGTCAAGACATCTACAACCGCGGC
		R-GGGTTATAGTCTTGTAGCCCTTATC	
Porphyromonas asaccharolytica	96	F-ACTGCTTCTTCCACCAAATAGA	ATACATCGACCGCAACGTGAGCT
		R-AAAGCGACTGAGCGAAGAA	
Porphyromonas uenonis	110	F-AGGCCACTGCACGTATCAAG	TCGACCTGACCCGTCCTGCGGTGCCT
		R-ATGACGAGAGCCTAGCGGA	
Roseburia intestinalis	132	F-TCTGTCTGCAGTCCCACTAT	ACTGACCAAACGAATCGCAGAATTAAAGGA
		R-TATCCTGCTGTACTGTTCCCT	
Solobacterium moorei	189	F-ATTAGAGGGTGTCAATGGCTCG	GCAGCGGACACATTCCGTGCTGGTGCG
		R-GGGTCACCGTTTTCTATTCCCT	
Streptococcus salivarius	166	F-AATACCCACAGTCTTCCGTTG	TGACGCCTTCCAAGAAGCTATGGA
		R-CGGTTTGATCATAATTGCCTGAAT	
16S rRNA	337	F-ATGGYTGTCGTCAGCT	CAACGAGCGCAACCC
		R-ACGGGCGGTGTGTAC	
SPUD spike-in	101	F-AACTTGGCTTTAATGGACCTCCA	GCACAAGCTATGGAACACCACGT
		R-ACATTCATCCTTACATGGCACCA	

A universal primer for the 16S rRNA gene is used to normalize Ct values (Ritalahti et al. 2006). The SPUD amplicon is spiked as an inhibitor indicator to all the test samples after dilution preparation (Nolan et al. 2006) (5- AACTTGGCTTTAATGGACCTCCAATT TTGAGTGTGCACAAGCTATGGAACACCACGTAAGACATAAAACGGCCACATATG GTGCCATGTAAGGATGAATGT-3). This artificial amplicon is part of potato genome which is lacking homology with any other known animal including human sequence. Presence of any inhibitor will increase quantification cycle (Cq) of SPUD assay, and absence of inhibitor will result by identical Cq.

2.4.4. Testing sensitivity and specificity of primers and probes

All primer pairs were tested against the main target species, sister species in same genus, human genome and *E.coli* to see potential mismatches or unspecific products Figure S2. 3. The primers which passed the specificity criteria were tested with probe and also sensitivity purposes. 1:10 DNA dilutions of main target were used to detect sensitivity of primers and probes.

2.4.5. Quantitative PCR

A DNA dilution of 5 ng/µl is used for processing stool samples; nuclease free water was used for all dilutions and reactions. All reactions were placed in triplicates. We used 384 well plates and QuantStudio 6 Flex Real time PCR SX was used to perform qPCR. The qPCR reaction was performed in 10µl volumes with 450nM primer and 300nM probe concentration. The one step cycling conditions consisted of an initial template denaturation of 95ºC for 10 min, followed by 40-cycles of denaturation at 95ºC for 15 sec, annealing, extension and reading fluorescence signal at 60ºC for 60 sec.

2.4.6. Quality control

We had several internal controls during the reaction. First of all, the primer pairs had a positive and negative control in the assay. SPUD, which is an artificial amplicon, was spiked in all samples during the reaction and used to see any potential inhibition in the given sample.

Table 2.4 Bacterial strains used as positive target to test primer and probes.

Species/strain	Reference	Tax id
Eubacterium halii VPI B4-27, type strain	DSMZ-3353	411469
Eubacterium rectale A1-86	DSMZ-17629	657318
Clostridium symbiosum WAL-14163	HM-309	742740
Clostridium symbiosum WAL-14673	HM-319	742741
Eubacterium ventriosum	DSMZ-3988	411463
F. nucleatum subsp. animalis D11	HM-75	997347
F. nucleatum subsp. animalis F0419	HM-758	997347
F. nucleatum subsp. nucleatum 1612A	DSMZ-15643	190304
F. nucleatum subsp. polymorphum	DSMZ-20482	393480
F. nucleatum subsp. vicentii	DSMZ-19508	209882
Gemella morbillorum M424	HM-240	29391
Parvimonas micra 3024A, VPI 5464	DSMZ-20468.	411465
Peptostreptococcus stomatis	DSMZ-17678	596315
Porphyromonas asaccharolytica	DSMZ-20707	879243
Porphyromonas uenonis	DSMZ-23387	281920/1122976
Roseburia intestinalis L1-82	DSMZ-14610	536231
Solobacterium moorei	DSMZ-22971	102148/1123263
Streptococcus salivarius-275	DSMZ-20560	1304

We furthermore used a plate calibrator to detect pipetting errors caused by manual handling or the liquid handler. We prepared calibrator reaction for each screening batch (DE1 and DE2) and only used that calibrator. Any seen difference would mean an error during the liquid handler process. Finally, we used *B. vulgatus* species which in present in all fecal

microbiome samples to monitor potential inhibitions as well. All primer and probe pairs have positive and negative controls in the same plate to judge the reaction efficiency and to detect contamination. All positive controls used in this assay is listed in Table 2.4 and they were ordered from Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures (https://www.dsmz.de/collection/catalogue) or Biodefense and Emerging Infections Research Resources Repository (BEI Resources, https://www.beiresources.org/) according to specified strain info.

2.4.7. Normalization of Ct value and model building

We used 16S rRNA gene between 1055-1392 to normalize mean Ct value. Delta Ct value was calculated based on the difference of 16s rRNA gene and the given primer Ct value Figure S2. 2. ROC curves per species was based on dCt values and pROC packages was used to perform ROC analysis.

After first normalization, we performed z-score normalization ($z = ((x - \mu) / \sigma)$; $x = Ct$; μ = mean; σ = standard deviation) to be able to use Ct values for machine learning approach. Z-score normalization is based on the standard deviations from their mean.

To assess the discriminate power of combined bacterial information, we have performed L2 regularization ridge linear regression model. Ridge regression modelling was chosen due to the characteristics of our dataset, which has a smaller number of predictors/features (targeted species) and more observations (individuals). Ridge regression includes all predictors, and does not perform predictor selection. Data were split into 80% training and 20% test set for 10 times repeat 10-fold cross validation ending up with 100 models in total. For each split, ridge linear regression was trained on the training set, and that model was then used to predict the test set. Models were then evaluated by checking the area under the Receiver Operating Characteristics curve (AUROC) based on the known CRC status in the metadata (Figure 2.5).

To test the gFOBT contribution to qPCR-based classifier, gFOBT status of all individuals was included to modelling and the same procedure described above was repeated (Figure 2.6).

"Biology enables, Culture forbids."

Yuval Noah Harari

3. The Potential of the Fecal Microbiome as a Novel Detection Approach for Pancreatic Cancer

This chapter discusses my second project during the PhD and details of how we identified fecal microbiome signature to detect Pancreatic Ductal Adenocarcinoma in collaboration with Dr. Nuria Malats and her team from CNIO, Spain. The text from following chapter has been mostly taken from a manuscript (close to submission) which has been originally written by myself. Parts of the text were also extended to include the analysis, method and literature details.

The sections including results from collaborators are explicitly stated and mostly involve Thomas Sebastian Schmidt, Georg Zeller and Jakob Wirbel from EMBL and Nuria Malats and her team from CNIO, Spain. Sections 3.2.5 and 3.2.6 are from Thomas Sebastian Schmidt and data of section 3.2.5 is provided by Sandra Rodriguez from Nuria Malats' team. All these findings are currently under preparation to be published as an original article, with the title:

"The Fecal Microbiome as a Novel Detection Approach for Pancreatic Ductal Adenocarcinoma"

3.1. Introduction

Pancreatic cancer (PC) is a disease on the rise worldwide, ranking seventh as leading cause of cancer-related deaths, with a higher incidence in developed countries (2.2-7.7 per 100,000 people) and a further increase predicted for the next decade (Rahib et al. 2014). PC mortality is almost as high as incidence rate and the 5-year survival rate is only 8% on average; however, survival rate for people with metastatic diagnosis is only 2% (Ferlay et al. 2018). Overall, the high mortality rate of PC is mostly due to the absence of characteristic symptoms, as early disease stages are usually asymptomatic or unspecific, and lack of routine screening methods (Kamisawa et al. 2016). Only a small subset of patients is eligible for surgical resection and most of the standard therapy options are not effective (Bracci 2017). The term 'pancreatic cancer' is an umbrella that encompasses several distinct types of disease origins, progressions and prognoses, categorized at both the anatomical (tumor site) and molecular level (patterns of mutations and chromosome rearrangements, (Collisson et al. 2019). Pancreatic ductal adenocarcinoma (PDAC), as a histological subtype, accounts for 85% of all pancreatic cancer cases (Hidalgo et al. 2015).

Population-scale screening programs could potentially alleviate the PDAC disease burden, but affordable, sensitive and non-invasive clinical tools for the early detection of PDAC are currently not yet available (Hasan et al. 2019). Because of the above, there is an urgent need to identify potential biomarkers for screening of high-risk groups, including familial pancreatic, ovarian, and breast cancers, hereditary pancreatitis, Peuts Jeghers Syndrome, and diabetes (Chang et al. 2014). Currently, the only Food and Drug Administration (FDA) approved biomarker for PDAC is based on carbohydrate antigen (CA) 19-9 protein (an antigen released in pancreas), which is mostly used for disease monitoring rather than diagnosis due to moderate sensitivity (72-86%) and specificity (68-80%) (Xing et al. 2018).

Table 3.1 Detailed list of current microbiome related studies in PDAC.

Study	Population	Cohort	Specimen	Method	Microbial change
Jandhyala 2017	Human	CP vs control	Feces	16S sequencing	↑ Firmicutes ↓ Bacteroidetes
Isaiah	Canine	EPI vs control	Feces	16S sequencing	↑ Lactobacillaceae and Streptococcaceae ↓ Lachnospiraceae and Ruminococcaceae
Hamada	Human	CP vs AIP	Feces	16S sequencing	↑ Bacteroides, Streptococcus and Clostridium spp.
Zhang	Human	AP vs control	Feces	16S sequencing	↑ Bacteroidetes and Proteobacteria ↓ Firmicutes and Actinobacteria
Beger	Human	AP	Pancreas	Culture	NA
Büchler	Human	AP	Pancreas	Culture	NA
Isenmann	Human	AP	Pancreas	Culture	NA
Thomas	Human	CP vs control	Pancreas	16S sequencing	NA
Wen	Mouse	DM	Feces	NA	NA
Kostic	Human	DM	Feces	16S sequencing	NA
de Goffau	Human	DM vs control	Feces	16S sequencing	↑ Bacteroides ↓ Bifidobacterium adolescentis and Bifidobacterium pseudocatenulatum
Endesfelder	Human	DM vs control	Feces	16S sequencing	NA
Mejía-León	Human	DM vs control	Feces	16S sequencing	↑ Bacteroides ↓ Prevotella, Megamonas and Acidaminococcus
Davis-Richardson	Human	DM vs control	Feces	16S sequencing	↑ Bacteroides dorei and Bacteroides vulgatus

Alkanani	Human	DM vs control	Feces	16S sequencing	↓ Lactobacillus and Staphylococcus
Vatanen	Human	DM vs control	Feces	16S sequencing	NA
Hu	Mouse	Pre-DM vs control	Feces	16S sequencing	↑ Gram-negative and Gram-positive ↓ Bacteroidetes and Erysipelotrichaceae
Torres	Human	PDAC (8) vs control (22)	Feces	16S sequencing	↑ Leptotrichia ↓ Porphyromonas Neisseria
Farrell	Human	PDAC (38), CP (27), control (38)	Oral	16S microarray	↓ Neisseria elongata and Streptococcus mitis
Michaud	Human	PDAC (405) vs control (410)	Plasma	Antibiotics to oral bacteria	↑ Porphyromonas gingivalis (ATTC 53978)
Ren	Human	PDAC vs control	Feces	16S sequencing	↑ Bacteroidetes ↓ Firmicutes and Proteobacteria
Pushalkar	Human	PDAC vs control	Feces	16S sequencing	↑ Proteobacteria, Synergistetes and Euryarchaeota
Thomas	Human	PDAC vs control	Pancreas	16S sequencing	NA
Riquelme	Human	PDAC LTS vs PDAC STS	Pancreas	16S sequencing	↑ Alpha diversity; ↑ Saccharopolyspora, Pseudoxanthomona, Streptomyces
Half	Human	PDAC (15) vs control (15)	Feces	16S mRNA	↑ Bacteroides and Verrucomicrobia ↓Firmicutes and Actinobacteria
Lin	Human	PDAC (13), CP (3) vs control (12)	Oral	16S mRNA	↑ Bacteroides ↓Corynebacterium and Aggregatibacter lower
Fan	Human	PDAC (361) vs control (371)	Oral	16S sequencing	↑ Porphyromonas gingivalis and Aggregatibacter actinomycetemcomitans
Michaud	Human	PDAC (405) vs control (416)	Plasma	Antibody	Porphyromonas gingivalis ATTC 53978
Olson	Human	PDAC (40) IPMN (39), control (58)	Oral	16S sequencing	↑ Firmicutes ↓ Proteobacteria

Like other cancers, pancreatic cancer also has only around 10% of hereditary contribution (Petersen 2016). Established non-hereditary risk factors include age, smoking, diabetes mellitus, chronic pancreatitis, periodontal diseases, obesity, alcohol consumption and dietary factors (Midha et al. 2016; Rawla et al. 2019). There have been several population studies showing increased risk for pancreatic cancer in pancreatitis patients with a standardized incidence ratio (SIR) of 19.0-26.7 (Malka et al. 2002). A Danish population-based study reported that people who were diagnosed with acute pancreatitis had a higher risk of developing pancreatic cancer 0.68% (95% CI 0.61-0.77) risk for 2-year and it increases to 0.85% (95% CI 0.76-0.94) at 5th year (Kirkegård et al. 2019).

Alterations in human microbiome composition including oral, fecal and organ-specific microbiota have also been associated with increased PC risk (Aykut et al. 2019; Del Castillo et al. 2019; Fan et al. 2018; Farrell et al. 2012; Half et al. 2019; Michaud et al. 2013a, 2013b; Pushalkar et al. 2018; Ren et al. 2019; Riquelme et al. 2019; Sethi et al. 2018; R. M. Thomas et al. 2018; Vogtmann et al. 2020). Some studies identified that poor oral microbiota and/or changes in bacterial composition are related with the increased risk of pancreatic cancer. For instance, periodontitis is a potential risk factor for pancreatic cancer which is an inflammation of oral cavity and it has an elevated risk of 1.5-2 (Maisonneuve et al. 2017). *Porphyromonas gingivalis* is major contributor of periodontitis and it has an increased risk for PDAC.

PDAC-associated dysbiosis of the salivary microbiome has been studied extensively. Global associations of the oral microbiome to PDAC have so far remained elusive; since previous studies are inconsistent and sometimes contradictory. One of the first studies was based on salivary microbiota of a small cohort, showing a significant decrease in the abundance of *Neisseria elongata and Streptococcus mitis,* through microarrays analysis. These findings were verified in an independent cohort via qPCR (28 pancreatic cancers, 27 chronic pancreatitis and 28 controls) (Farrell et al. 2012). Another study analyzed the blood serum antibody levels of 25 bacterial species in 405 cases and 416 controls and reported that an enrichment inn

Porphyromonas gingivalis is associated with an increased risk of PDAC (Michaud et al. 2013b). A study with a very small group of 8 cancer and 22 controls found high levels of *Leptotrichia* and low levels of genus *Porphyromonas* and *Neisseria* (Torres et al. 2015). A more recent study, in a prospective cohort, characterized the oral microbiota composition with 16S ribosomal sequencing of 361 carcinoma patients and 371 matched controls (Fan et al. 2018). In this study, *Porphyromonas gingivalis* and *Aggregatibacter actinomycetemcomitans* were linked with the high risk of pancreatic cancer while the phylum Fusobacteria and its genus *Leptotrichia* were associated with a decreased risk. Oral microbiome of 40 newly diagnosed cancer, 39 Intraductal papillary mucinous neoplasm (IPMN) and 58 normal individuals showed high level of Firmicutes in cancer patients and IPMN patients were similar to controls based on microbiome data (Olson et al. 2017).

Pancreas itself was believed to be a sterile organ, since pancreatic juice contains numerous proteases and is highly alkaline, which would prevent microbial growth (Maekawa et al. 2018). Yet a number of studies recently demonstrated the presence of microbes in the pancreas at healthy and diseased status via different techniques including 16S rRNA gene sequencing, qPCR, culturing and microscopy (Pushalkar et al. 2018; Sethi et al. 2018; R. M. Thomas et al. 2018). There is not yet a consensus to define a healthy or diseased pancreas microbiome; since each study reported slightly different results and could not confirm previous findings. This is mostly due to low-biomass of tissue samples, which makes contamination unavoidable and small cohort size that leads to statistical limitations. One study reported genus *Brevibacterium* and order Chlamydiales in diseased pancreas showed increased relative abundance (Pushalkar et al. 2018); while another study found an increase of *Acinetobacter* and *Pseudomonas* abundance in PDAC tissue compared with healthy specimen (R. M. Thomas et al. 2018). Even though the role of microbes in the pancreas remains unknown, it was also reported that, bacterial load had an effect not only on the survival time of PDAC but also the

response to some chemotherapy agents used in clinics (Sivan et al. 2015; Vétizou et al. 2015). Pancreas of long-term survivors harbors a different tumor microbiota with higher alpha diversity including high relative abundance of genus *Saccharopolyspora, Pseudoxanthomonas* and *Streptomyces* than short-term survivors (Riquelme et al. 2019). Presence of *Gammaproteobacteria* in PDAC tissue specimens showed a link with gemcitabine resistance (Geller et al. 2017).

Since several diseases have been associated with changes of the gut microbiome (Gharaibeh & Jobin 2018; Gopalakrishnan et al. 2018; Matson et al. 2018); the possible effect of gut microbiome on the progression PDAC has been recently investigated as well. A prospective study analyzed fecal microbiome of 85 PDAC patients before surgical resection and 57 healthy controls via 16S amplicon sequencing. Gut microbiome of PDAC patients was reported to be significantly different when compared with controls with lower alpha diversity measures, but not between different pancreatic sections (Ren et al. 2017). Another fecal microbiome study of 30 PDAC patients, 6 patients with pre-cancerous lesions, 13 healthy subjects and 16 with non-alcoholic fatty liver disease reported fourteen species to be discriminative between cancer and controls but only a small subset was shared with pre-cancerous individuals.

Moreover, another study reported that fungi can translocate from the gut to the pancreas, and genus *Malassezia* was enriched in diseased cancerous tissue followed by *Parastagonospora*, *Saccharomyces*, and *Septoriella* (Aykut et al. 2019). They proposed that ablation of mycobiome was slowing down the PDAC progression and repopulation with only *Malassezia species* promotes oncogenesis via Mannose-binding lectin (MBL)-C3 cascade activation. They also reported that the fungal composition of the gut and pancreas differs between healthy and diseased mice and anecdotally validated this in human patients as well.

All these studies demonstrate that the microbiome may have a role not only in carcinogenesis, but also on survival and probably response to PC therapy. Although mechanistic insight into putative gut-pancreas microbiome links remains limited, the possibility to use microbiome signature as a PDAC specific biomarker is promising. No effective screening methods for pancreatic cancer are currently available, so there is an urgent need to identify potential biomarkers. The implementation of effective early diagnosis tools will contribute to improve the treatment, prognosis, life expectancy and eventually prevention of PDAC. Even though we do not know yet the ultimate role of microbiome on PDAC, its potential as a biomarker needs to be further investigated.

Here, we conducted a study comparing oral, fecal and pancreas microbiome of newly diagnosed 57 PDAC patients, 29 with chronic pancreatitis (CP) and 50 healthy controls via 16S rRNA amplicon sequencing, shotgun metagenomics sequencing and fluorescence *in situ* hybridization (FISH) microscopy, to investigate the potential of microbiota as an early detection tool for PDAC.

3.2. Results

We observed several differentially abundant taxa between PDAC and controls in stool samples including several species from genus *Veillonella*, but not in saliva samples. We could not reproduce previous reports on individual oral taxa in our cohort, even when tested supervised. Modelling of fecal microbiome demonstrated a PDAC specific taxonomic microbial composition with AUC of 0.84. We applied the fecal-based prediction model to available external validation cohorts including healthy controls and some diseases as seen risk groups for PDAC like type 1 diabetes (T1D), type 2 diabetes (T2D), liver cirrhosis (LC), Crohn's Disease (CD), inflammatory bowel disease (IBD), UC, CRC to evaluate the specificity

of fecal signal. We also found different ASVs between healthy and diseased pancreas via 16S rRNA gene sequencing and further validation of those ASVs via microscopy proven that similar species which were present in stool samples were also present in different pancreatic tissue specimens.

3.2.1. Cohort details

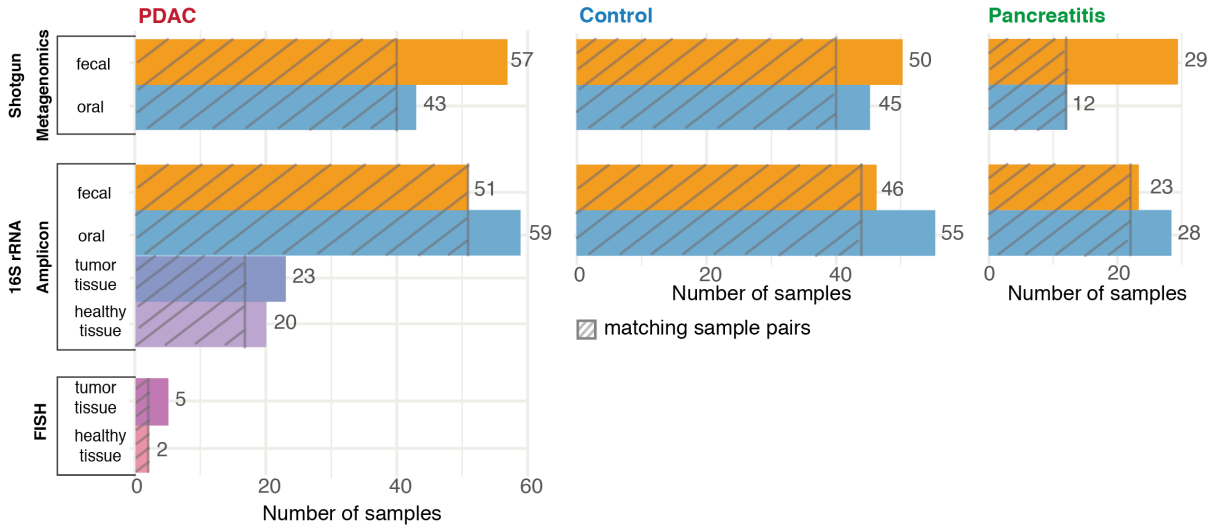


Figure 3.1 Cohort overview.

Strip bands on the bar plots represent the samples which have matching body sites. This figure is part of the manuscript which is under preparation and it was produced by myself.

We recruited 57 newly diagnosed, treatment-naïve PDAC patients, 29 CP patients and 50 age, gender-matched controls from Barcelona and Madrid, Spain (see Figure 3.1 for cohort overview, see Table 3.2). We obtained fecal shotgun metagenomes for all subjects, salivary metagenomes for a subset (45 PDAC, 12 CP and 43 controls), as well as matched 16S rRNA amplicon data from all available samples (see Methods). The analysis workflow is detailed in Figure 3.8.

Table 3.2 Demographic features of the PDAC cohort.

	Total N=136	PDAC N=57	Control N=50	CP N=29
Sex	87M + 49F	36M +21F	31M+19F	20M + 9F
Cholesterol	46	25	13	8
Diabetes	41	17	5	19
Age Mean	69.9	71.5	71.3	64.1
Smoking	79	28	26	25
Alcohol	105	40	39	26
Periodontitis	41	17	11	13
Jaundice	37	32	2	3
Obesity	36	16	13	7

3.2.2. Diversity measures and confounder analysis

Several environmental factors including smoking, alcohol consumption or diabetes have a direct effect on microbiome composition (Thomas S B Schmidt et al. 2018). Due to this knowledge, we tested potential confounders of microbial data in our cohort, in order to stratify the analyses accordingly. For a total of 22 demographic and clinical variables, we quantified marginal effects on microbiome community-level diversity (Figure S3. 1, Table S3. 1, Table S3. 2). Fecal and salivary microbiome species richness (as a proxy for alpha diversity) were not univariately associated with any tested variable including PDAC status after accounting for the most common PDAC risk factors and applying false discovery rate (FDR) correction (Figure 3.2a).

Microbiome community composition, in contrast, varied with diagnosed symptoms of jaundice (Permutational multivariate analysis of variance (PERMANOVA) on Bray-Curtis dissimilarities, R2=0.02, p=0.009) and age (R2=0.01, FDR-corrected p=0.03), diabetes (R2=0.01, p=0.04) in feces, and recruitment age (R2=0.02, p=0.04) and subject sex (R2=0.02, p=0.05) for saliva samples, with very low effect sizes (Figure 3.2b, Table S3. 2). After

stratification for those confounders, diagnosis was mildly but significantly associated with fecal (R2 = 0.02, p= 0.001), but not salivary (R2=0.01, p=0.5) (Figure S3. 2, Figure S3. 1) community composition.

3.2.3. Fecal metagenomic classifier identifies microbiome signatures of PDAC

Univariate tests showed nine differentially abundant taxa between cases and controls, even though the overall community composition shift was mild (Wilcoxon test on relative abundances, Benjamini-Hochberg-corrected $p < 0.05$; see Figure 3.2c). Most prominently, several *Veillonella atypica*, *Fusobacterium hwasookii nucleatum* and *Alloscardovia omnicolens* were enriched in feces of PDAC patients, whereas *Bacteroides coprocola, Romboutsia timonensis, Faecalibacterium prausnitzii* and *Bifidobacterium bifidum* were depleted. Those taxa were moderately predictive of PDAC state (Figure 3.3, Figure S3. 3).

Contrary to feces, we did not detect any species in the salivary microbiome with significantly differential abundance when correcting for multiple hypothesis tests, including *Porphyromonas gingivalis, Aggregatibacter actinomycetemcomitans* (Fan et al. 2018)*, Neisseria elongata* or *Streptococcus mitis* (Farrell et al. 2012) for which such associations were previously reported (Figure S3. 4). We hypothesized that this may be due to technical or geographical variation between datasets, yet we were unable compare our data to published sets, as raw and clinical data for these was not publicly available.

To combine individual signals into an accurate classifier, we next built multi-species metagenomic models, using L1-regularized LASSO regression with 10 resampling and 10-fold cross validation. The model discriminated PDAC cases and controls with AUROC of 0.85 ("model-1"; Figure 3.4).

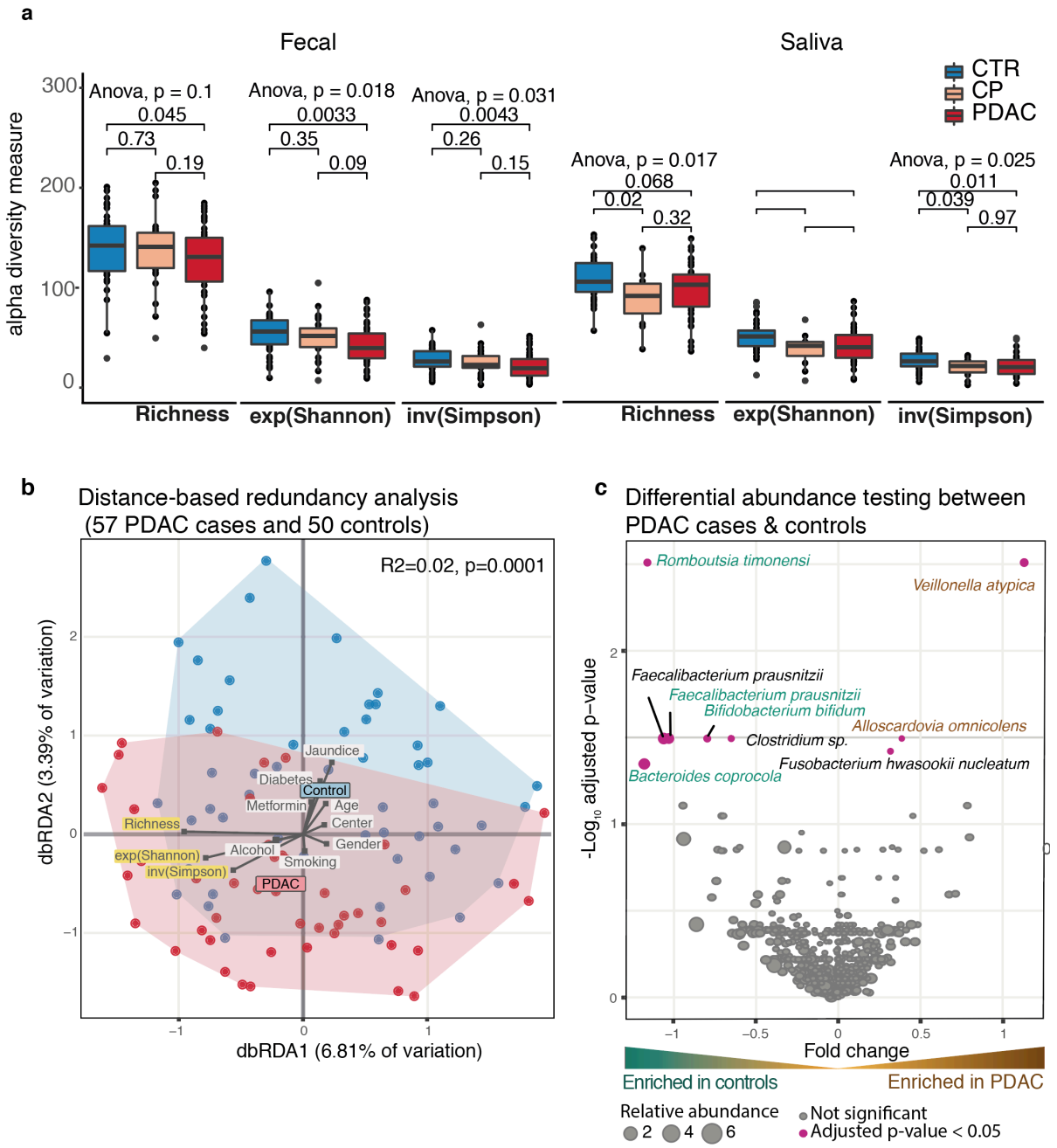


Figure 3.2 Community analysis of fecal microbiome data.

a. Different alpha diversity measures for both fecal and saliva metagenomics data. **b.** Bray-Curtis distance-based redundancy analysis (dbRDA) of PDAC and control fecal microbiome data. PDAC samples are shown as red colored circles and controls as blue. Richness, exponential Shannon (exp(Shannon)) and inverse Simpson (inv(Simpson)) diversity measures are also visualized with arrows similarly like tested metadata variables. The distance of meta-variable from the center represents the confounding effect size (see Methods). **c.** Wilcoxon test results of fecal microbiome data to test enriched taxa between PDAC and control cases (see Methods). Y-axis is FDR corrected p values and x-axis is log2 (fold change). Pink dots represent the significantly abundant species/strains in either group, while black dots show non-significant presence of species after FDR correction. Green and brown coded species are selected in metagenomic model-1 to predict PDAC. This figure is part of the manuscript which is under preparation. All panels were produced by myself.

The PDAC-related marker species in the model were *Methanobrevibacter smithii*, *Alloscardovia omnicolens*, *Veillonella atypica*, *Bacteroides finegoldii* and *Akkermansia muciniphila*; however, we note that by selection of machine learning model, LASSO regression procedure selects representatives for different species with highly similar abundance profiles. Despite that, models to detect CP patients from controls or PDAC cases had no predictive power and similarly, no robust PDAC signature was detected for the salivary microbiome in any group comparison (AUROC < 0.5, Figure S3. 4).

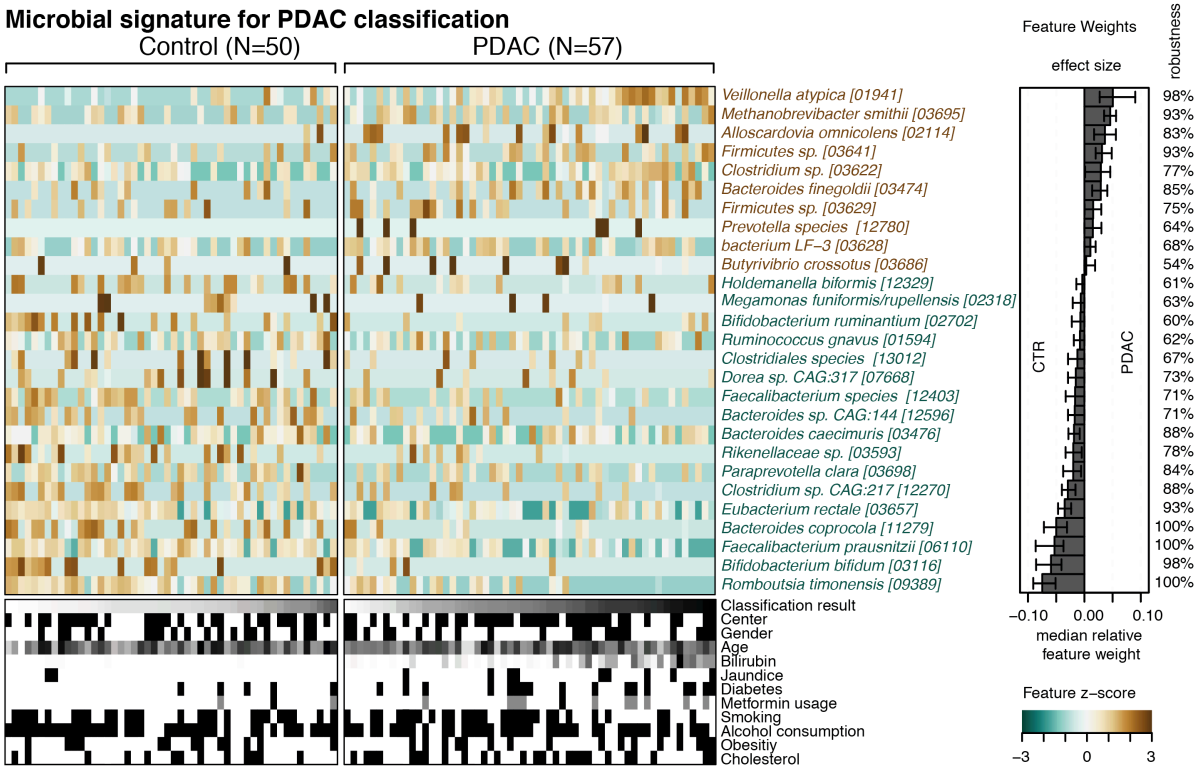


Figure 3.3 Metagenomics signature of PDAC-associated species.

Normalized abundance of 27 selected species in fecal microbiome across all samples shown as heatmap. The right panel represents the main contribution of each selected feature to the overall model-1 and robustness of each feature is presented in percentage. Classification scores from cross validation of each individual and condition for tested meta-variables are displayed at the bottom of the panel. This figure is part of the manuscript which is under preparation and it was produced by myself.

The initial gut microbiome-based classifier included species like *F. prausnitzii, B. coprocola, Bifidobacterium bifidum* or *Romboutsia timonensis* which were depleted in PDAC relative to controls (Figure 3.3). For some of these species, it was previously suggested that depletion is linked to intestinal inflammation in general, rather than to specific diseases (Duvallet et al. 2017) and also our experience from colorectal cancer shown similar patterns. We therefore re-trained a classifier preferentially that selected microbial features which were enriched in PDAC cases. The enrichment constrained model ("model-2") distinguished PDAC cases with an accuracy of AUROC=0.72, the difference with the unconstrained model-1 mostly due to a penalty on sensitivity (Figure 3.4).

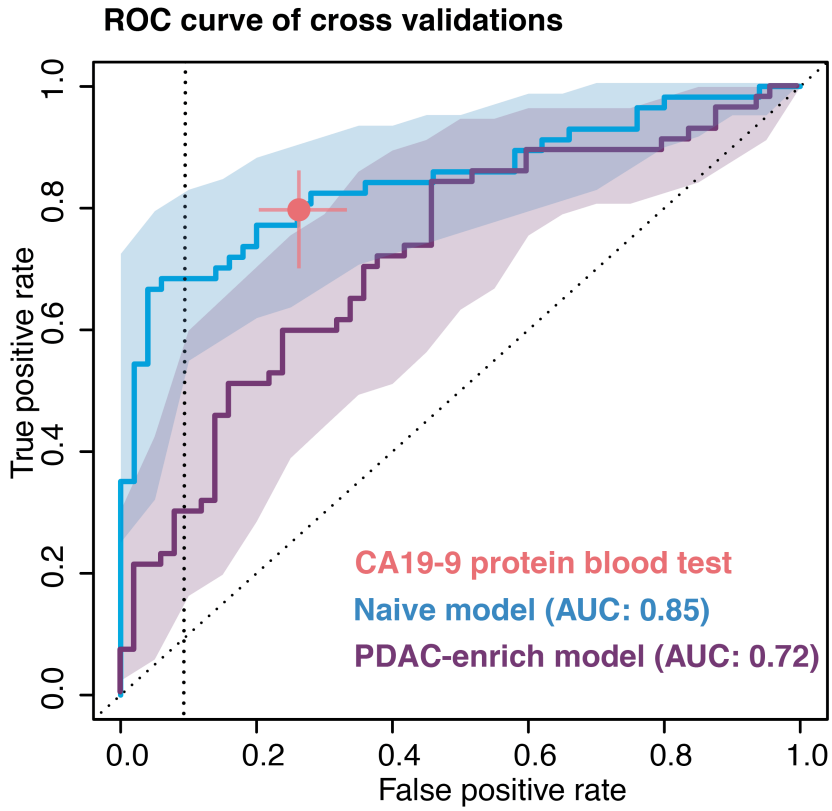


Figure 3.4 Internal cross validation.

The results are shown as receiver operating characteristic (ROC) curve with a 95% confidence interval shaded in grey. 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). The sensitivity and specificity of CA19-9 marker is shown in pink (Xing et al. 2018). This figure is part of the manuscript which is under preparation and it was produced by myself.

We also tested the prediction value of the 22 demographic and clinical variables describing our cohort, and none were selected as predictive features by the model and the microbiome signature was stronger than any other feature. The classifier captured a predictive gut microbiome signature of PDAC that is independent of other disease risk factors and potential confounders.

3.2.4. Specificity of metagenomic classifier

To confirm specificity to PDAC, we validated our classifiers against external cohorts in total 3,468 publicly available gut metagenomes from 15 studies across 7 countries, including subjects diagnosed by T1D or T2D, CRC, LD or IBD as well as healthy controls **(**Figure 3.5**,** Table S3. 3**)**. The unconstrained model-1 showed moderate specificity, predicting PDAC state for <20% of subjects in all controls, however Crohn's disease (37%) or liver diseases (43%) had higher false positive rate (fpr). We partly attributed this signal to technical variations, given that only Chinese IBD and LD patient cohorts showed high fpr. Furthermore, IBD has been associated with similar depletion signatures as observed in our model-1 in particular of *F. prausnitzii* (Cao et al. 2014), whereas liver diseases share some physiological characteristics with impaired pancreas function. In contrast, the enrichment-constrained model-2 was highly specific for PDAC, at 0-5% PDAC predictions in almost all external cohorts, at a maximum of 10-17% predictions among one of LD and IBD patient cohorts. In particular, the detected microbiome signatures were also robust against misclassification of type 2 diabetes patients – with relevance to potential screening applications, as these are a major PDAC risk group.

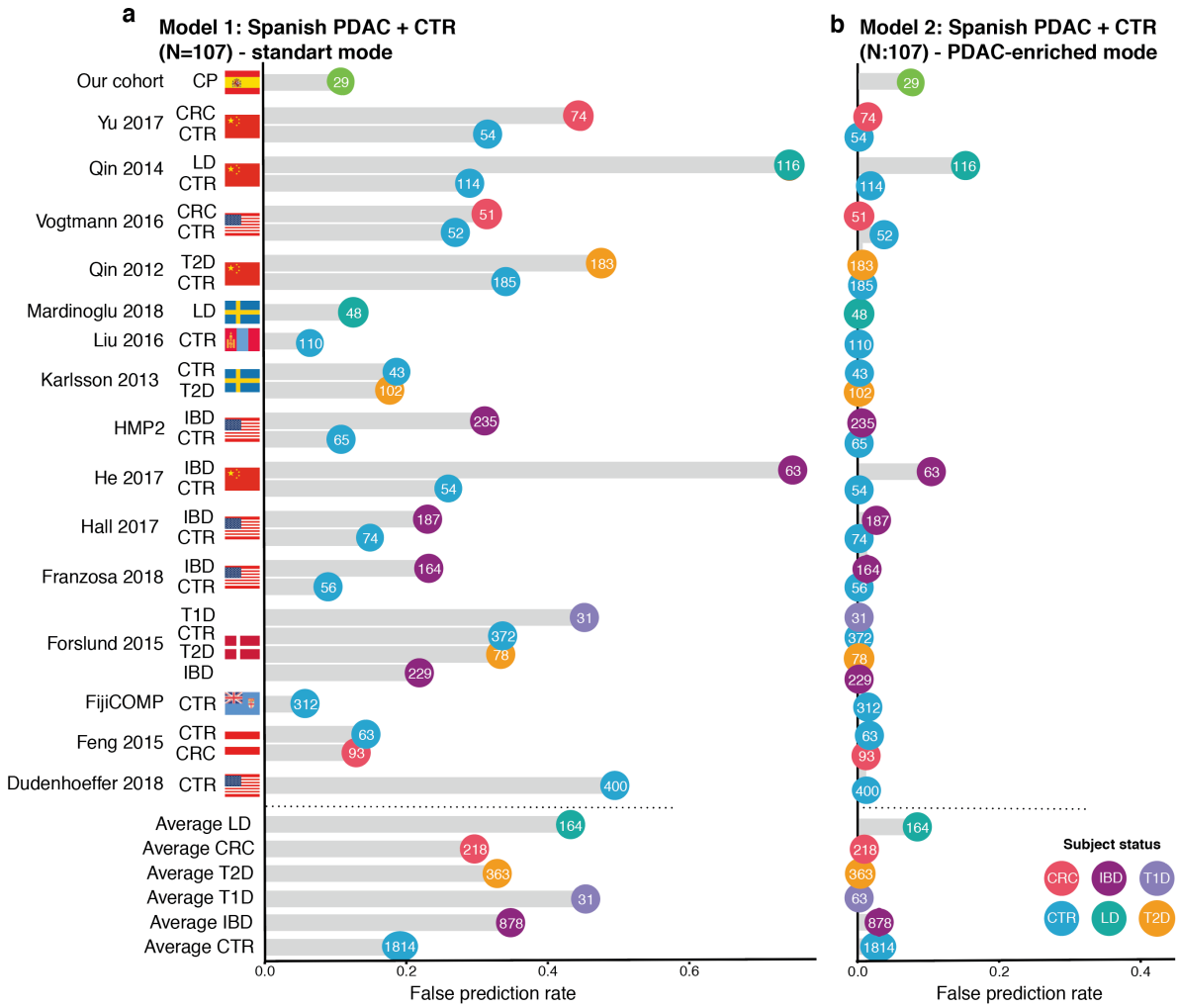


Figure 3.5 External validation of PDAC fecal microbiome models.

a. False positive rate (FPR) of metagenomic model-1 in 15 external test cohorts with 90% specificity cut-off (see Supplementary Table X for list of all studies included). Test cohorts were profiled and normalized in the same way as initial dataset (see Methods). Each study was stratified according to health status and model-1 was tested to predict in given group. FPR is displayed as a grey bar and the number of the cohort in color coded circles. **b.** False positive rate (FPR) of metagenomic model-2 in 15 external test cohorts with 90% specificity cut-off (see Table S3. 3 for list of all studies included). CTR: controls, CRC: colorectal cancer, CP: chronic pancreatitis, T1D: type-1 diabetes, T2D: type-2 diabetes, IBD: inflammatory bowel disease, LD: liver disease. This figure is part of the manuscript which is under preparation and both panels were produced by myself.

3.2.5. PDAC tumors harbour characteristic bacteria, with links to the oral and gut microbiome

Alterations in pancreatic secretion, as a consequence of tumor growth in the pancreatic duct, may plausibly underlie characteristic gut microbiome signatures, such as described above. This would imply that PDAC development indirectly causes microbiome shifts. Yet the pancreatic duct also provides a direct anatomical link between the gastrointestinal tract and the pancreas that may be traversed by bacteria (Pushalkar et al. 2018; R. M. Thomas et al. 2018) and fungi (Aykut et al. 2019) with a putative role in carcinogenesis (Riquelme et al. 2019).

Table 3.3 The tested genus and used FISH probes list.

Target species/genus	Sequence	Probe	Dye
Bifidobacterium (genus)	5'- GATAGGACGCGACCCCAT -3'	Bif228	Cy3
Veillonella (genus)	5'- AGACGCAATCCCCTCCTT -3'	Veil223	FITC
Akkermansia (species)	5'-CCTTGCGGTTGGCTTCAGAT-3'	MUC1437	FITC
Lactobacillus (genus)	5'- ACATGGAGTTCCACT -3'	Lact663	FITC
Bacteroides (genus)	5'- CCAATGTGGGGGACCTT -3'	Bac303	Cy3
Streptococcus (genus)	5' TTTAGCCGTCCCTTTCTGG -3'	Strc493	Cy3

We therefore expected that several microbial taxa with specific gut associations to PDAC should be detectable in pancreatic tumors. We taxonomically profiled biopsies of tumors (n=23) and adjacent healthy pancreatic tissue (n=20) in our cohort using 16S rRNA amplicon sequencing, applying strict filters against putative reagent contaminants as are frequently observed in low bacterial biomass samples (de Goffau et al. 2019; Salter et al. 2014) (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 fecal microbiome (Figure 3.6a, Figure S3. 5, Figure S3. 6).

Among these, *Lactobacillus sp.*, *Akkermansia muciniphila* and *Bacteroides sp.* were enriched in tumors relative to unaffected pancreatic tissue by relative abundance (Wilcoxon test, FDR-corrected $p<0.006$). In a subset of five tumor and two pancreatic tissue samples, we verified the presence/absence of *Lactobacillus sp.*, *Bifidobacterium sp., Veillonella sp., Bacteroides sp., Akkermansia sp.* and *Streptococcus sp.* using in-situ hybridization with genus-specific primers (Figure 3.6a-c & Table 3.3). *Akkermansia sp.*, though observed by amplicon sequencing in 26/30 subjects, were not detectable using FISH in six tested samples. This showed that depending on observations based on one method can be misleading due to the complications of low biomass samples.

To establish the provenance of tumor- and tissue-dwelling bacteria, we next traced exact 16S rRNA Amplicon Sequence Variants (ASVs) across salivary, fecal, tissue and tumor samples within subjects (Figure 3.7). *Veillonella sp*., characteristically enriched in stool among PDAC patients, were highly prevalent in both salivary (100% of subjects) and fecal (87.5%) samples across the entire cohort, while oral and fecal types also matched to tumor and tissue ASVs, implying that pancreas-dwelling *Veillonella* populations may be sourced from both the mouth and the gut. Interestingly, we found no intra-individual match in *Veillonella* ASVs between tumor and adjacent tissue samples, indicating that tumor-dwelling *Veillonella sp.* may be distinct from those in healthy tissue. In contrast, *Lactobacillus sp.*, *Bifidobacterium sp.* and *A. muciniphila* ASVs were matched between tumor and healthy tissue samples in several subjects. Pancreas-dwelling *A. muciniphila* populations moreover matched fecal types, suggesting that they may have originated from the gut. In *Lactobacillus sp.* and *Bifidobacterium*

sp., tissue and tumor types corresponded to both oral and fecal ASVs, but with divergent links, indicating that distinct pancreatic sub-populations may be linked to the mouth and the gut.

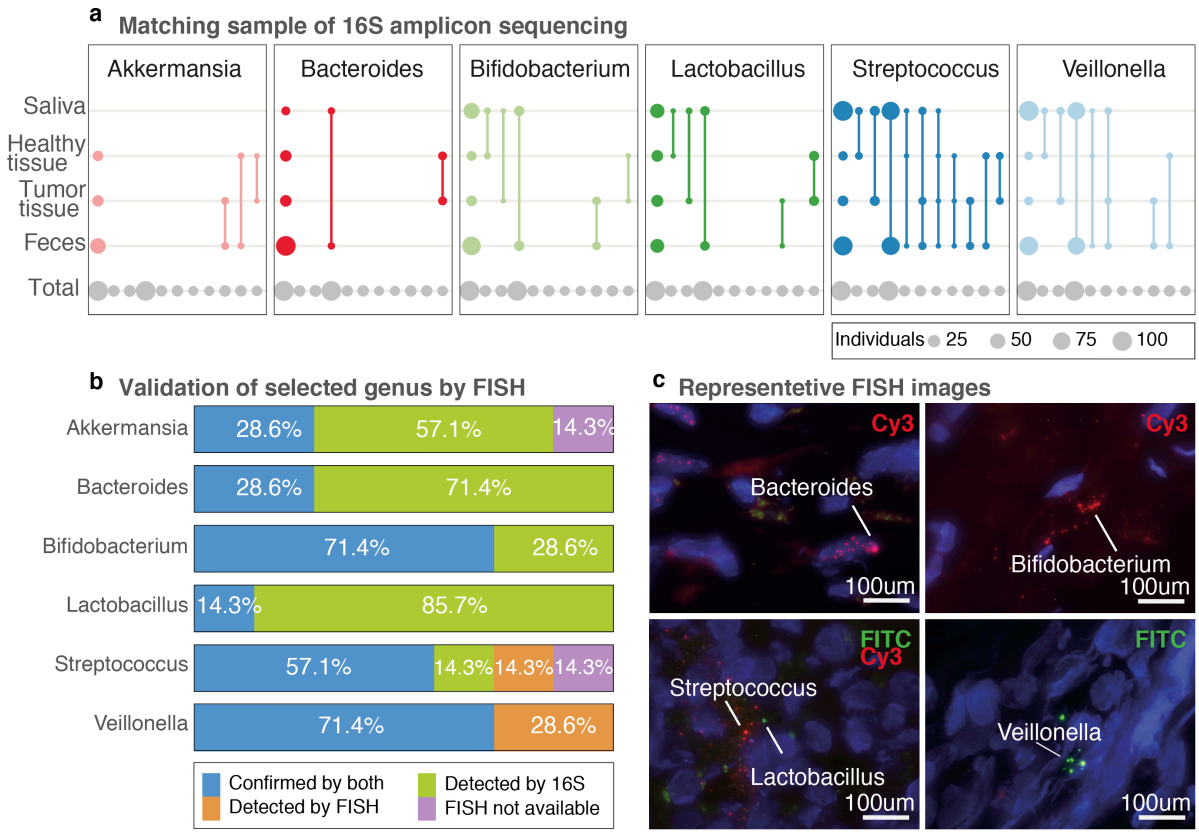


Figure 3.6 Presence of microbiome in different sections of pancreas with different conditions.

a. Presence of each genus in four different body sites including fecal, saliva, pancreatic tumor and healthy tissue samples. The number of the individuals is presented by circle size and matching pairs with connecting lines. The first column always presents the total number of cohort and the rest shows all combinations. **b.** Selected pancreatic tissue samples (five tumor and two non-tumor) to show bacterial presence with both 16S amplicon and FISH methods. The validation of bacterial presence with both 16S amplicon sequencing and in-situ hybridization is shown with blue color. The samples that showed bacterial presence via 16S; however, were not present according to FISH displayed in green. Bacterial presence validated only by FISH is shown by orange and samples that were not able for FISH validation due to lack of tissue material is shown by purple. The percentage is calculated by the number of samples divided by total number of tested samples (seven) and converted to percent. **c.** Representative microscopy images for *Bacteroides, Bifidobacterium, Lactobacillus, Streptococcus, Veillonella*. FITC fluorescent dye is used for green and Cy3 dye is used for red color labelling. This figure is part of the manuscript which is under preparation. Panel A was generated by Thomas Sebastian Schmidt, panel b was produced by myself and c was in collaboration with Nuria Malats from CNIO.

3.2.6. Several PDAC-associated species in the gut may be sourced from the oral cavity

Many microbial species traverse the gastrointestinal tract to form overlapping populations between the oral cavity and intestine, with increased levels of intra-individual strain transmission associated to diseases like colorectal cancer (T. S. Schmidt et al. 2019). Indeed, several prominent marker taxa with fecal enrichment signatures in PDAC are common oral commensals, such as *Veillonella sp.*, *Streptococcus sp.* or *Fusobacterium sp.*. We therefore hypothesized that intestinal populations of these PDAC-associated species were primarily of oral origin, with generally enhanced levels of autologous oral-intestinal strain exchange in PDAC patients.

We quantified oral-to-gut transmission based on the intra-individual overlap of microbial Single Nucleotide Variants (SNVs) for species prevalent in both the mouth and the gut, as a proxy for oral and intestinal strain populations (see Methods). We found that viewed across all subjects and species, PDAC was associated with increased levels of oral-intestinal strain overlap (Cohen's d = 0.33; Analysis of variance (ANOVA) $p<10^{-3}$ when stratifying for species-level effects and technical, demographic and clinical variables). This observation extended to individual PDAC-associated species, with enhanced levels of autologous transmission in several *Veillonellaceae sp.* (*V. dispar,* d=0.71; *V. atypica*, d=0.6; *V. parvula*, d=0.2; *Megasphaera micronuciformis*, d=2.47) and *Streptococcus sp.* (*S. salivarius*, d=0.51; *S. vestibularis*, d=0.49; *S. parasanguinis*, d=0.36). The situation was more nuanced among *Bifidobacteriaceae sp.*, with enhanced transmission in *B. longum* (d=2.16) and *A. omnicolens* (d=1.24), but less strain overlap in *B. dentium* (d=-0.89).

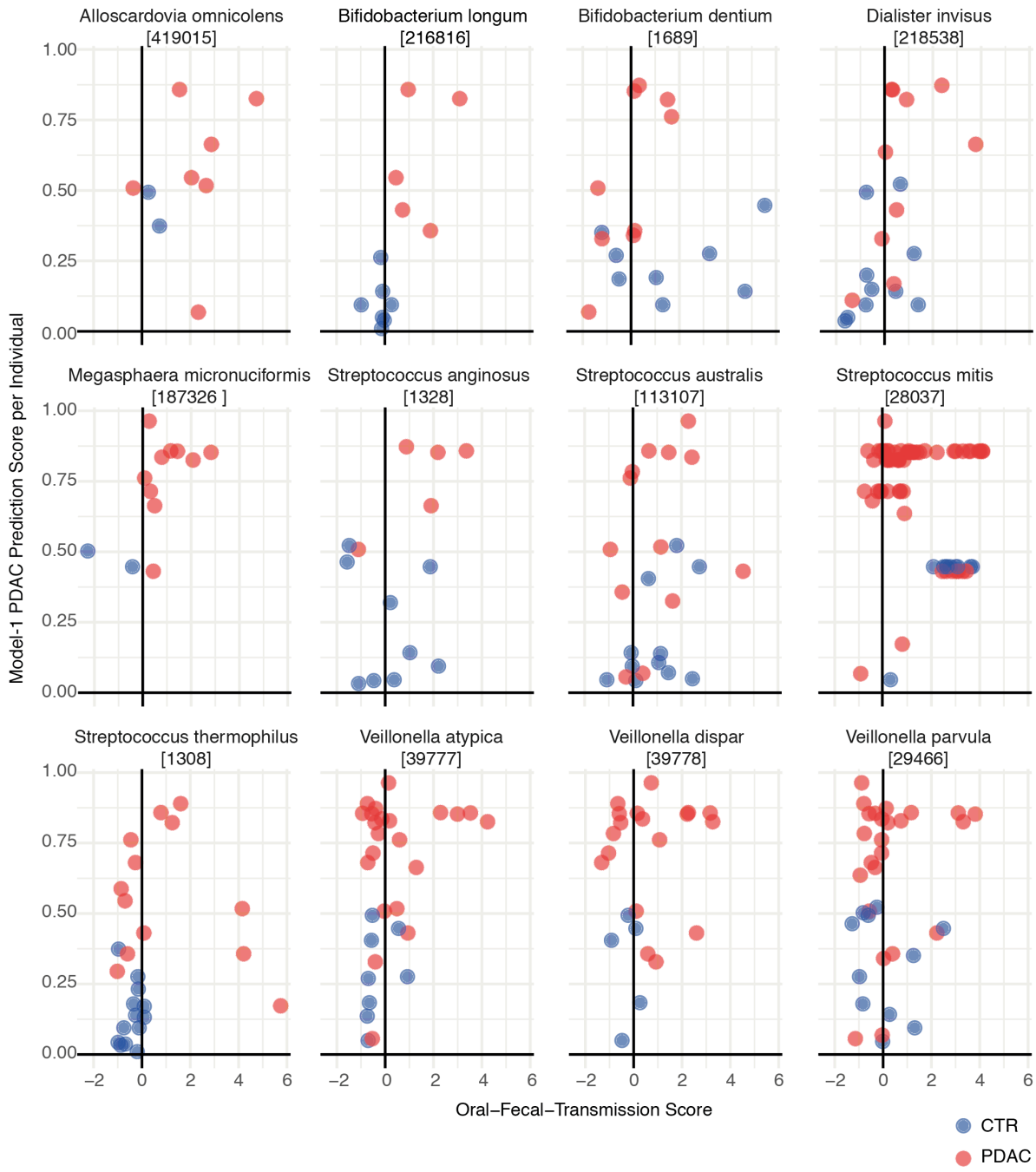


Figure 3.7 Oral-fecal transmission scores differ between PDAC cases and controls.

x-axis show the oral-fecal transmission score for each individual, while y-axis shows the prediction score of unconstrained model-1 per person. This figure is part of the manuscript which is under preparation and it is produced by Thomas Sebastian Schmidt.

However, due to limits in metagenomic coverage and species prevalence, our cohort size did not allow for sufficient statistical power to significantly discern these trends for individual species with confidence, in particular when stratifying for putative confounders and correcting for multiple hypothesis tests. Nevertheless, our data indicates that PDAC patients showed overall enhanced levels of oral-intestinal transmission, and that intestinal strain populations of PDAC signature species may be sourced autologously, from the oral cavity.

3.2.7. Functional profiling based on PDAC metagenomes

We also profiled KEGG modules to assess potential of them to distinguish PDAC from healthy controls. Feature selection was based on minimum redundancy maximum relevance (mRMR, http://home.penglab.com/proj/mRMR/) to choose only top 200 predictors, since the data by nature is very noisy with high number of features. We first performed univariate analysis between cases and controls and observed some differentially abundance functions (Figure S3. 7) and then build an L1-regularized lasso regression model with 10 resampling and 10 cross validation (Figure S3. 8). There are currently limitations to work with functional data, such as poor annotations of functions, noisy data which makes modelling very challenging due to high number of predictors than observations.

3.3. Discussion and Perspectives

The early detection of PDAC remains a formidable challenge, at the heart of ongoing efforts to mitigate the disease burden of pancreatic cancer. Currently, the only available biomarker for PDAC is CA19-9, with moderate sensitivity and specificity (Goonetilleke and Siriwardena 2007; Gui et al. 2014; Xing et al. 2018). There are several prediction studies including gene-carrier, absolute risk and also high-risk groups (Baecker et al. 2019; Hsieh et al. 2018; Muhammad et al. 2019; Nakatochi et al. 2018; Săftoiu et al. 2008). Pilot-scale studies have moreover explored urine and serum for suitable markers; however, most of these predictions are based on previous knowledge, not validated properly and none of them are adopted in clinics yet (Hasan et al. 2019).

Based on a cohort of newly diagnosed, treatment-naïve cases and clinically matched controls for which oral, fecal and tissue microbiomes were analyzed (Figure 3.1a), we developed a metagenomic classifier that robustly and accurately predicts PDAC, solely based on characteristic fecal microbial species (Figure 3.2). The PDAC signature captured by our multi-species model was orthogonal to clinical and demographic variables, including well-established PDAC risk factors. This suggests that in practice, the fecal microbiome may be used to screen for PDAC and may be complementary to other testable markers, with added accuracy in combined tests, as is the case e.g. for colorectal cancer (Zeller et al. 2014).

Enriched species in PDAC cases included several *Veillonella sp., Alloscardovia omnicolens, Methanobrevibacter smithii* (Figure 3.3). *Veillonella sp.* are common oral and gut commensals; however there are reports showing an enrichment in cystic fibrosis, several infections including meningitis, lung and oral carcinomas and even potential as a biomarker (Bhatti & Frank 2000; Nagy et al. 1998; Pustelny et al. 2015; Yan et al. 2015). The role of

Methanobrevibacter smithii, a prevalent methanogenic archaeon, in the human gut remains poorly understood. *M. smithii* has been associated with both health and disease status including anorexia and irritable bowel disease (Armougom et al. 2009; Chaudhary et al. 2018; Ghavami et al. 2018; Sogodogo et al. 2019).

The PDAC fecal microbiome signature stood out against high risk groups of different diseases such as T1D, T2D, LD, IBD, CRC and controls from 15 different studies and on more than 3,600 samples (Figure 3.5). We validated the specificity of our models and confirmed that the detected signature was PDAC specific, with low false positive rates among other diseases. Yet among PDAC risk groups, and specifically against type 2 diabetes patients, our model-1 and 2 were highly specific with 90% specificity. We note that between-studies, technical variation can often exceed biological differences in microbiome composition (Paul I Costea et al. 2017), and some false positive predictions might be due to technical variations as well. Meta-studies of multiple geographically and ethnically diverse PDAC cohorts may be required to further establish globally consistent PDAC microbiome signatures, as has been successfully shown for colon cancer (A. M. Thomas et al. 2019; Wirbel et al. 2019)

In the case of saliva and oral-microbiome markers, the existing controversy between oral microbiome studies and unavailable raw data makes the comparison of this studies challenging (Fan et al. 2018; Farrell et al. 2012; Michaud et al. 2013b; Olson et al. 2017; Torres et al. 2015). Several previously reported univariate PDAC associations including *P. gingivalis, A. actinomycetemcomitans*, *S. thermophilus* and Fusobacterium species of oral taxa were not confirmed in our cohort (Figure S3. 4). Indeed, we did not observe any salivary PDAC signature either for individual species or for multi-species models when stratifying for common demographic and clinical confounders. In fact, we found evidence for a strong, although indirect, association between PDAC and the oral microbiome, as levels of oral-intestinal

transmission were increased in PDAC, in particular for several PDAC signature taxa in the fecal microbiome.

While our data is strictly observational, there are strong indications that the observed characteristic fecal microbiome shifts are not merely a consequence of impaired pancreatic function and systemic effects thereof. Rather, several taxa could be traced between the gut and pancreas, with several univariate enrichment signals in tumors relative to adjacent healthy tissue. We confirmed previous observations (Pushalkar et al. 2018; Riquelme et al. 2019; R. M. Thomas et al. 2018) that the human pancreas harbors a microbiome, both by amplicon sequencing (accounting for putative contaminants to which low bacterial biomass samples are prone (Salter et al. 2014), and by performing *in situ* hybridization (FISH) for several genus (Figure 3.6). There is not yet a clear evidence for what represents a healthy or diseased pancreas microbiome; since each study reported slightly different results and could not confirm previous findings. This is mostly due to low-biomass of tissue samples, which makes contamination unavoidable and small cohort size that leads to statistical limitations. Getting conclusion based on one method for low biomass samples might be misleading due to technical issues, therefore, we confirmed presence of enriched genus in fecal samples (*Veillonella, Streptococcus, Akkermansia*) and also differentially abundant genus between healthy and tumor tissues (*Bacteroides, Lactobacillus, Bifidobacterium*) by FISH for validation.

For several taxa, we observed an intra-individual overlap of exact Amplicon Sequence Variants (ASVs) between oral, fecal and tissue samples, thus confirming links at the highest attainable taxonomic resolution for amplicon data. Given that fecal populations of several gut-enriched PDAC-associated taxa could thus be traced to pancreatic tumors, and in view of previous reports on microbe-mediated pancreatic carcinogenesis in murine models (Pushalkar et al. 2018; Riquelme et al. 2019; Sethi et al. 2018; R. M. Thomas et al. 2018), we believe that

the presented panel of PDAC-associated species in the gut may be relevant beyond merely diagnostic purposes.

The causal role of microbiome in PDAC is yet to be explored and with more prospective cohorts, we may improve our understanding of PDAC etiology and progression and the ulterior link with the gut microbiome. We should extend our knowledge not only on bacterial but also fungal and viral microbiome effect on PDAC progression. Based on the tumor, oral and fecal microbiome data, we could not make any causal links. However, the fecal microbiome provides promising knowledge for earlier detection of PDAC. The development of fecal microbiome-based, cost-effective assays for PDAC screening and progression opens new areas to explore. Yet, based on the demonstrated links to microbial populations in the tumor, oral and gut microbiome may also provide promising future entry points for therapeutic intervention.

3.4. Methods

3.4.1. Subject recruitment and sample collection

Subjects were recruited from two hospitals in Barcelona and Madrid in Spain by using same standard operating procedures (SOPs) for biological sample collection, processing and storage conditions. PDAC cases were identified early in their diagnostic process prior to any cancer treatment, while controls are age- and gender-matched hospitalized patients carefully selected with conditions from a list of selected pathologies known not to be associated with any PDAC risk factor.

Stool and saliva (mouthwash) samples were preserved in RNALater and immediately stored in at 4ºC, after ~12 hours, the sample was stored in an ultra-freezer at -20ºC, and 24 hours later at -80ºC until DNA extraction. Moreover, tumor and non-affected tissue samples were collected during the surgery for a subset of individuals and immediately flash-frozen onsite and preserved at -80ºC. All the samples were shipped with dry ice to keep the temperature stable. Information about lifestyle habits and medical history was surveyed with questionnaire and all the available information was stored through the web application EPIQuest developed by the Spanish National Institute of Bioinformatics (IBN-CNIO).

3.4.2. Sample processing

Fecal and salivary samples were thawed on ice, aliquoted and genomic DNA was extracted using the Qiagen Allprep PowerFecal DNA/RNA kit as per the manufacturer's instructions (Qiagen, Hilden, Germany). Genomic DNA from pancreatic tumoral and non-tumoral tissue samples was extracted using the Qiagen DNeasy blood and tissue kit in a protocol modified (Del Castillo et al. 2019): cells were lysed mechanically (with 5mm stainless steel

beads at 25Hz for 150s), followed by lysozyme treatment (20 mg/ml) and protease and RNAse digestion (56°C for 2h). All samples were randomly assigned to extraction batches. To account for potential bacterial contamination of extraction, PCR and sequencing kits. negative controls Each tissue DNA extraction batch included negative controls.

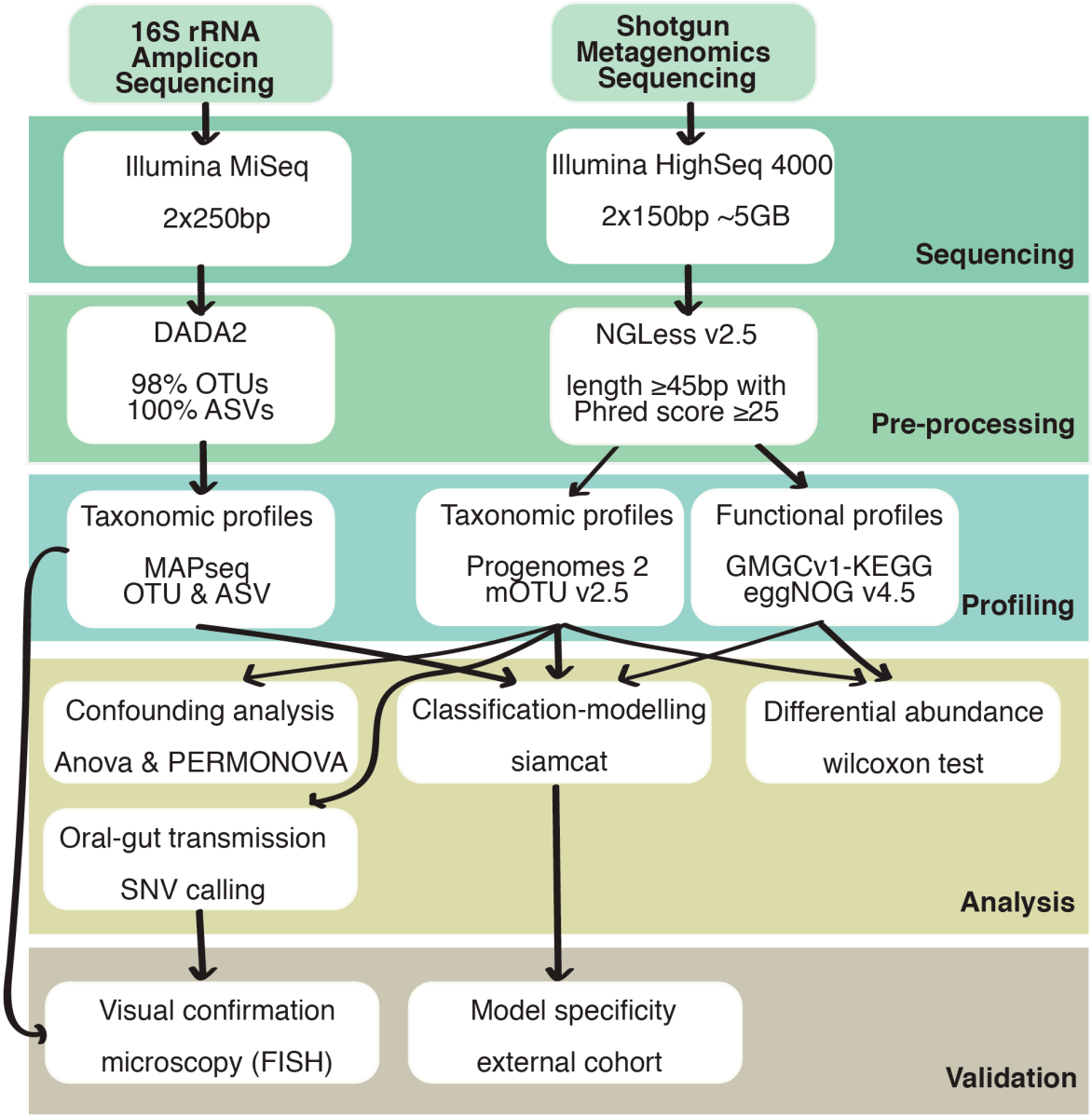


Figure 3.8 Analysis pipeline.

This pipeline shows quality controls we have done and the analysis we have performed. This figure is part of the manuscript which is under preparation and it was produced by myself.

3.4.3. 16S rRNA amplicon sequencing

Pancreatic tissue DNA was enriched for 16S rRNA in a pre-amplification PCR using primers 331F (5'-TCCTACGGGAGGCAGCAGT-3') (Nadkarni et al. 2002) and 979R (5'-GGTTCTKCGCGTTGCWTC-3') (Kramski et al. 2011). The cycling conditions consisted of an initial template denaturation of 98ºC for 2 min, followed by 30-cycles of denaturation at 98ºC for 10 sec, annealing at 65ºC for 20 sec, extension at 72ºC for 30 sec, and a final extension at 72ºC for 10 min. This was followed by a size-selective cleanup using SPRIselect magnetic beads (0.8 left-sized; Beckman Coulter, Brea, CA, USA). Fecal and salivary DNA were not pre-amplified.

Targeted amplification of the 16S rRNA V4 region (primer sequences F515 5'-GTGCCAGCMGCCGCGGTAA-3' and R806 5'-GGACTACHVGGGTWTCTAAT-3' (Caporaso et al. 2011), 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, TX, 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 2x250bp on an Illumina MiSeq (Illumina, San Diego, CA, USA) at the Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg.

3.4.4. 16S rRNA amplicon data processing

Raw reads were quality trimmed, denoised and filtered against chimeric PCR artefacts using DADA2 (Callahan et al. 2016). 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 (Matias Rodrigues et al. 2017). Reads that

did not confidently map to the reference were aligned to bacterial and archaeal secondary structure-aware SSU rRNA models using Infernal (Nawrocki & Eddy 2013) and clustered into OTUs with 98% average linkage using hpc-clust (Matias Rodrigues & von Mering 2014), as described previously (Thomas S B Schmidt et al. 2015). 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 filtered by removing samples retaining less than 500 reads and taxa observed in less than 5 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 in community composition. For pancreatic tissue and tumor samples, ASVs observed in negative control samples, as were reads mapping to known reagent kit contaminants (Salter et al. 2014). After these steps, we retained 298 samples 16S rRNA amplicon samples from 143 subjects for further analyses (130 salivary, 118 fecal, 25 of unaffected pancreatic tissue, 25 of tumor tissue with 17 matching PDAC tissue samples).

3.4.5. Shotgun metagenomic sequencing

Metagenomic libraries for 136 fecal 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, CA, USA) in 2x150bp paired-end setup at the Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg.

3.4.6. Metagenome data processing

Metagenomic data was processed using established workflows in NGless v2.5 (Coelho et al. 2019). Raw reads were quality trimmed (≥45bp at Phred score ≥25) and filtered against the human genome (version hg19, mapping at ≥90% identity across ≥45bp). The resulting filtered reads were mapped against representative genomes of 5,306 species-level genome clusters obtained from the proGenomes database v1 (Mende et al. 2020).

Taxonomic profiles were obtained using the mOTU profiler v2.5 (Milanese et al. 2019) and filtered to retain only species observed at a relative abundance ≥10^-5 in ≥5% of samples. Gene functional profiles were obtained from read counts by eggNOG v4.5 (Huerta-Cepas et al. 2016) annotations to orthologous groups and KEGG modules. Features with a relative abundance of ≥10^-6 in ≥15% of samples and a variance ≥10^-3 across all samples were retained for further analyses. We also used on minimum redundancy maximum relevance (mRMR, http://home.penglab.com/proj/mRMR/) approach to decrease the number of features.

3.4.7. 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 effective number of taxa with Hill coefficients of q=0 (i.e., 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 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 (Thomas Sebastian Benedikt Schmidt et al. 2017). 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 (Oksanen et al. 2010). To quantify potentially confounding univariate links between the abundance of individual taxa and subject clinical factors (see main text), we performed either ANOVA or non-parametric Kruskal-Wallis tests, depending on abundance distributions (Figure S3. 1, Table S3. 1, Table S3. 2)

3.4.8. Multivariable statistical modeling 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 cutoff: 0.001, prevalence cutoff: 0.05). Then, features were normalized by log10-transformation (to avoid infinite values from the logarithm, a pseudo-count of 1e-05 was added to all values) followed by standardization as z-scores. Data were randomly split into test and 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-regularized (LASSO) logistic regression model (Tibshirani 1996) using the implementation within the LiblineaR R package (Helleputte & Gramme 2017). 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).

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 introducing a preferential selection of features enriched in PDAC ('enrichment-constrained' model).

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

3.4.9. 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 (Table S3. 3 for the full list including accession numbers). We performed a literature search to identify publicly available datasets of fecal metagenomes in case-control or cohort studies for PDAC-related diseases. For a total set of 15 studies covering 3468 samples across 6 disease states, raw sequencing data was downloaded from the European Nucleotide Archive (ENA) and taxonomically profiled as described above.

The trained metagenomic classifiers for PDAC were then applied to each external dataset after a frozen normalization which utilizes the same set of features and normalization parameters (for example the mean of a feature for standardization) as in the normalization procedure from the pancreatic cancer dataset. Then, predictions were assessed for disease specificity in that high predictions for other diseases would indicate that the classifier relies on general features of dysbiosis in contrast to signals specific to pancreatic cancer. For this analysis, the cutoff for the predictions was set to a FPR of 0.1 in the pancreatic cancer dataset (Figure 3.5).

3.4.10. Sub-species and strain-level analyses

"Metagenomic reads were mapped against species representative genomes from the proGenomes v1 database (see above). Microbial single nucleotide variants (SNVs) were called from uniquely mapping reads using metaSNV (Paul Igor Costea et al. 2017) and within-species allele distances between samples were calculated as described previously. 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 (T. S. Schmidt et al. 2019). In short, the overlap between microbial SNVs in salivary and fecal samples within subjects was contrasted to a between-subject background to compute a quantitative oral-fecal transmission score and p value. Associations of species- and subject-specific transmission scores with clinical factors were tested for using ANOVA and *post hoc* tests, followed by Benjamini-Hochberg correction for multiple tests."

3.4.11. Fluorescence *in situ* Hybridization Microscopy

"FISH analyses were performed using probes specifically designed to a complementary 16S rRNA sequence unique to a particular taxon of bacteria (Figure 3.6, Table 3.3). Pancreatic tumor and normal pancreas samples were obtained from the pathologist immediately after their excision and they were directly frozen in liquid nitrogen (LN2). All the process lasted 30 minutes maximum. Sterile material was used to dissect the different samples. The minimum size of tissue for freezing was approximately 0.5 cm3 (0.5 x 0.5 x 0.5 cm). Samples were transferred from the temporary LN2 transport container to a locked –80°C freezer and kept until its analysis when they were transported with dry ice and put down in an OCT mould in LN2.

They were immediately cut with a cryostat to get 10 sections of 3-5 um each. All material was sterilized with ethanol after each sample handling.

Tissue sections of 5μm thickness were mounted on positively charged slides (SuperFrost, Thermo Scientific). Briefly, tissues were post fixed in freshly prepared 4% paraformaldehyde. After enhancement of the bacteria wall permeabilization by lysozyme treatment (10g/L Tris HCl 6.5M), the samples were hybridized for 1 hour at 45°C in the presence of the specific probe in a hybridizer machine (DAKO). Hybridization was done in 20 μl of hybridization buffer (20 nM Tris, pH 8.0. 0.9 M NaCl, 0.02% SDS, 30% formamide) added to 100 ng of the probe. Finally, the tissues were washed in washing solution (70% formamide, 10mM Tris pH7.2 and 01% BSA), dehydrated in a series of ethanol, air-dried and stained with 0.5 μg/ml DAPI/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 blinded analyzed to score for the number of FISH signals."

Appendix

A. Supplementary information for chapter 2

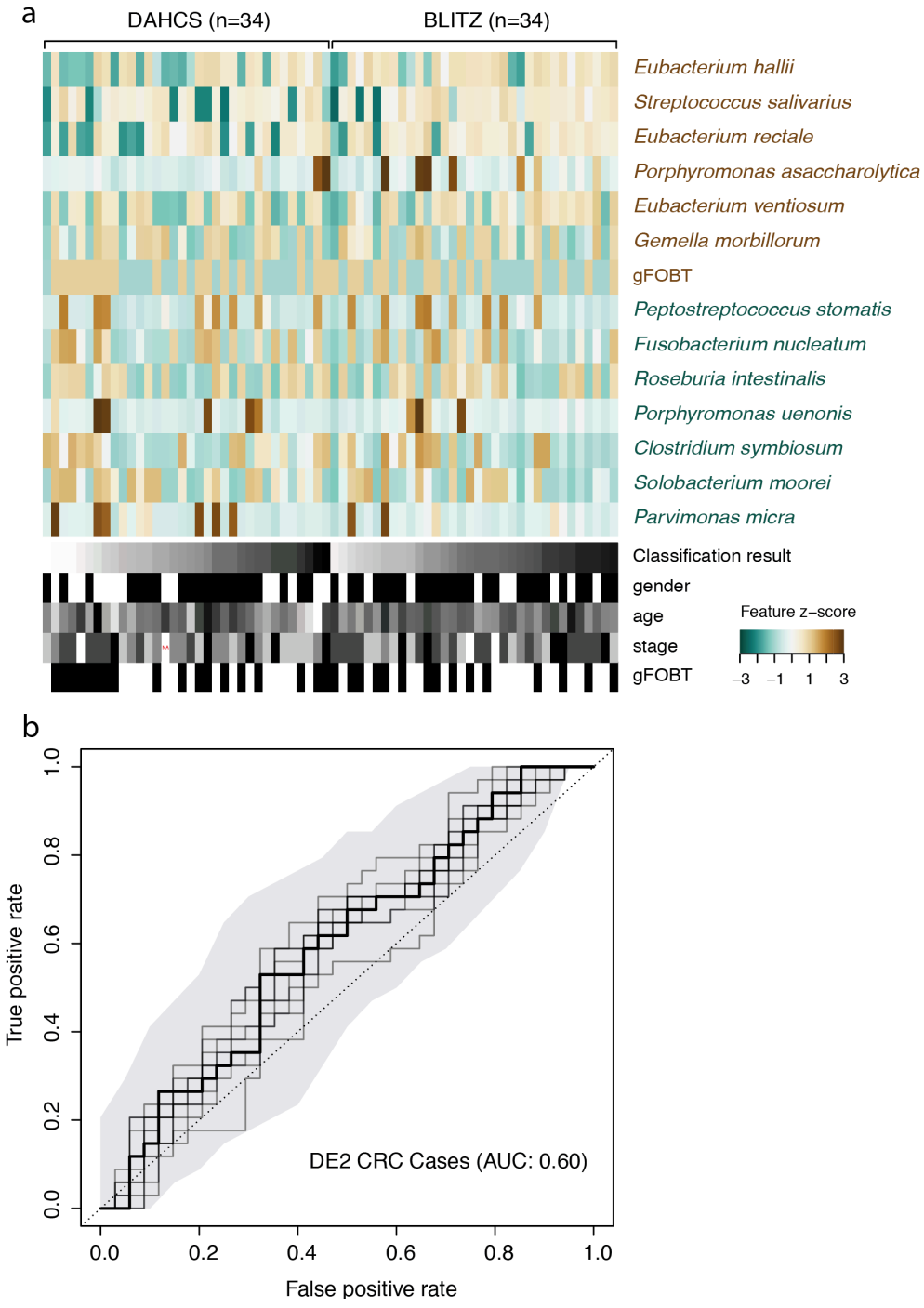


Figure S2. 1 Comparison of CRC cases from BLITZ and DACHS studies.

a. Same number of CRC cases were available from both studies. Heatmap shows the MAST and gFOBT results for both studies **b.** Ridge regression model with 10 resampling and 10 cross validation did not show a strong difference between two study groups based on MAST and gFOBT results. This figure is part of the manuscript which is under preparation and it was produced by myself.

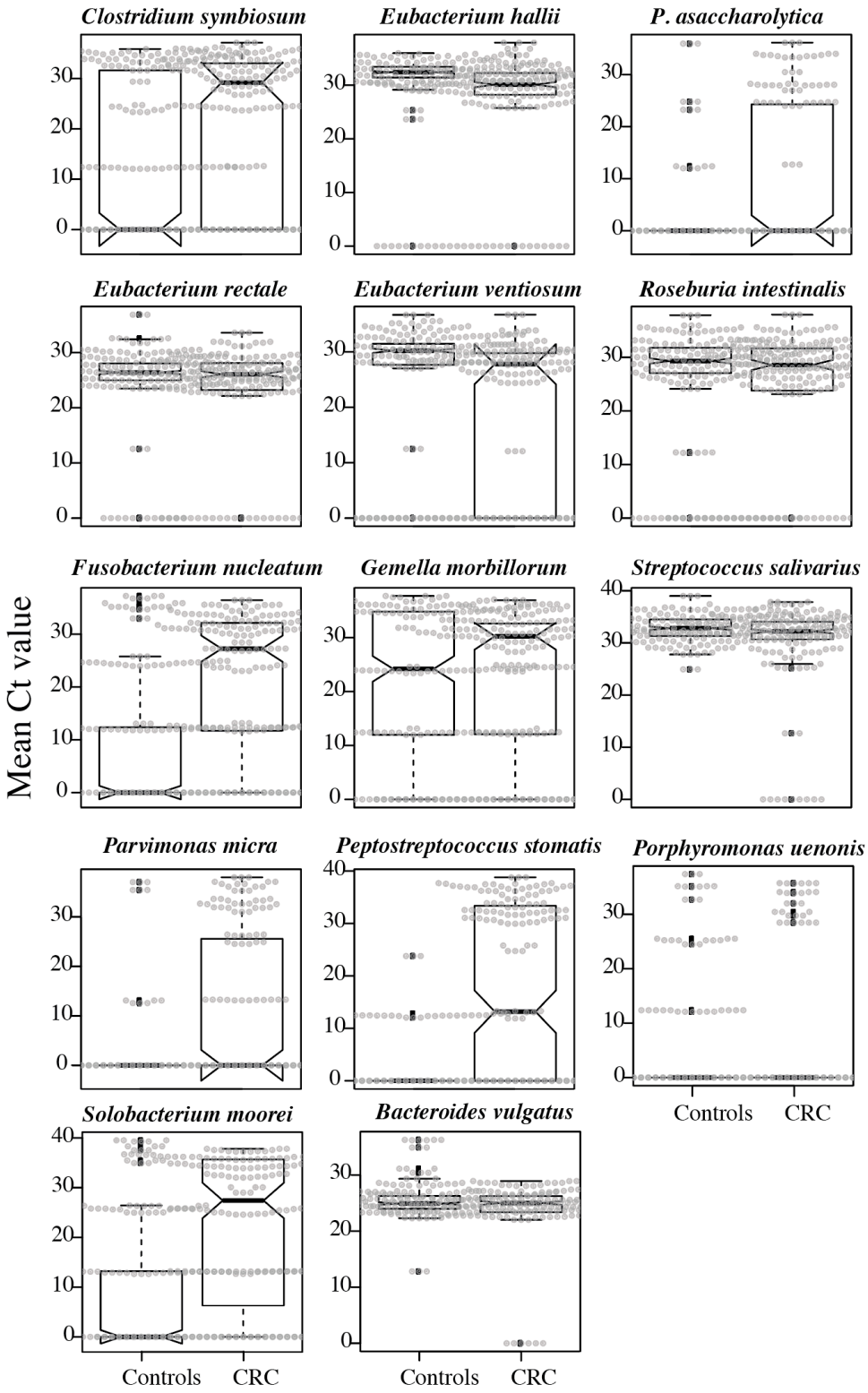


Figure S2. 2 Mean Ct values of DE1 cohort shown for each bacterium.

Mean Ct value was calculated based on the average of three reactions (triplicate) per samples. y-axis display the mean Ct value. This figure is part of the manuscript which is under preparation and it was produced by myself.

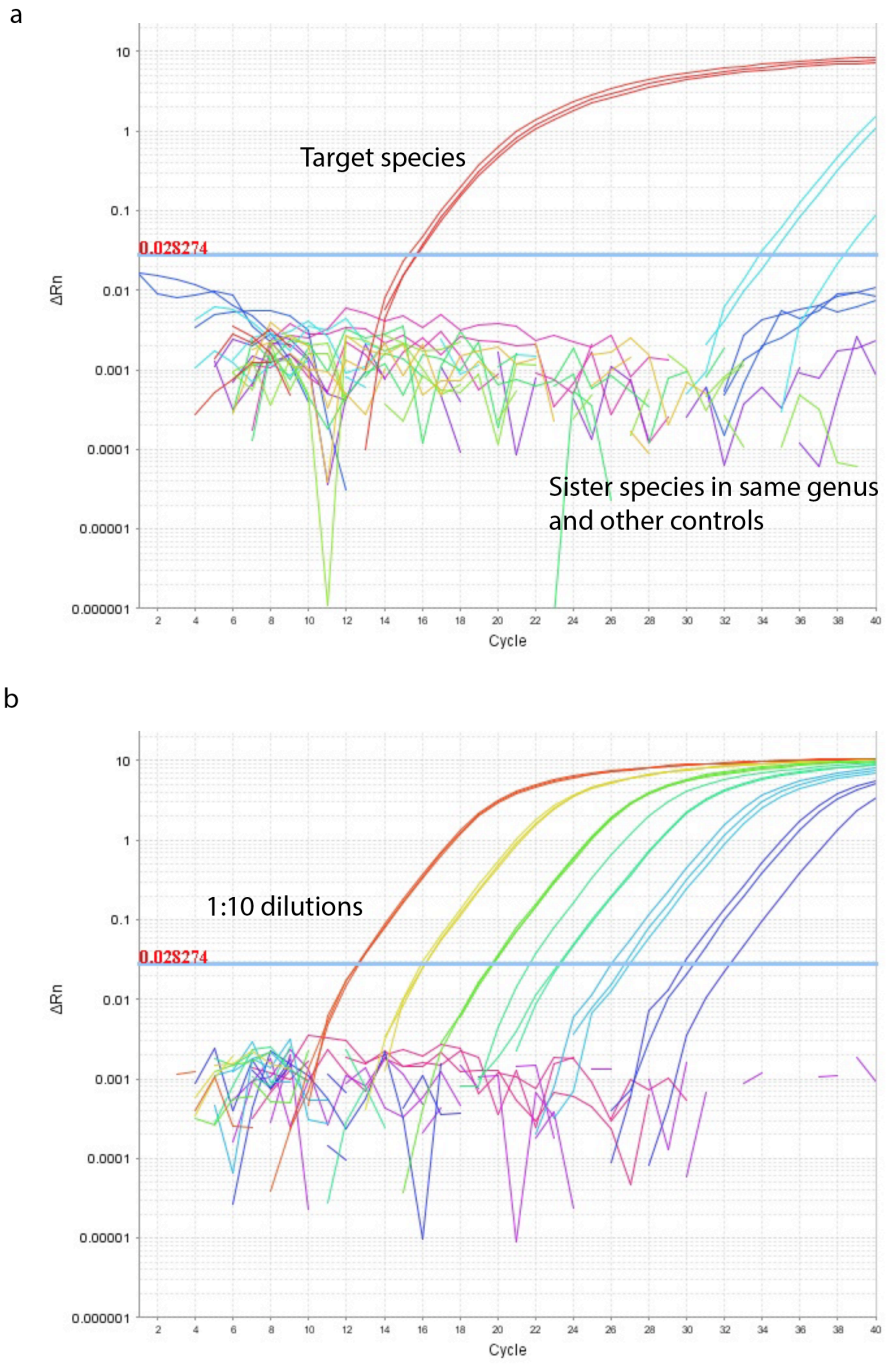


Figure S2. 3 Testing primer/ probe specificity and sensitivity.

a. Primer specificity was tested against sister species in same genus, *E.coli*, other available strains and also human genome. **b.** 1:10 DNA dilutions were used to test the detection limits for primer and probe. This figure is part of the manuscript which is under preparation and it was produced by myself.

B. Supplementary information for chapter 3

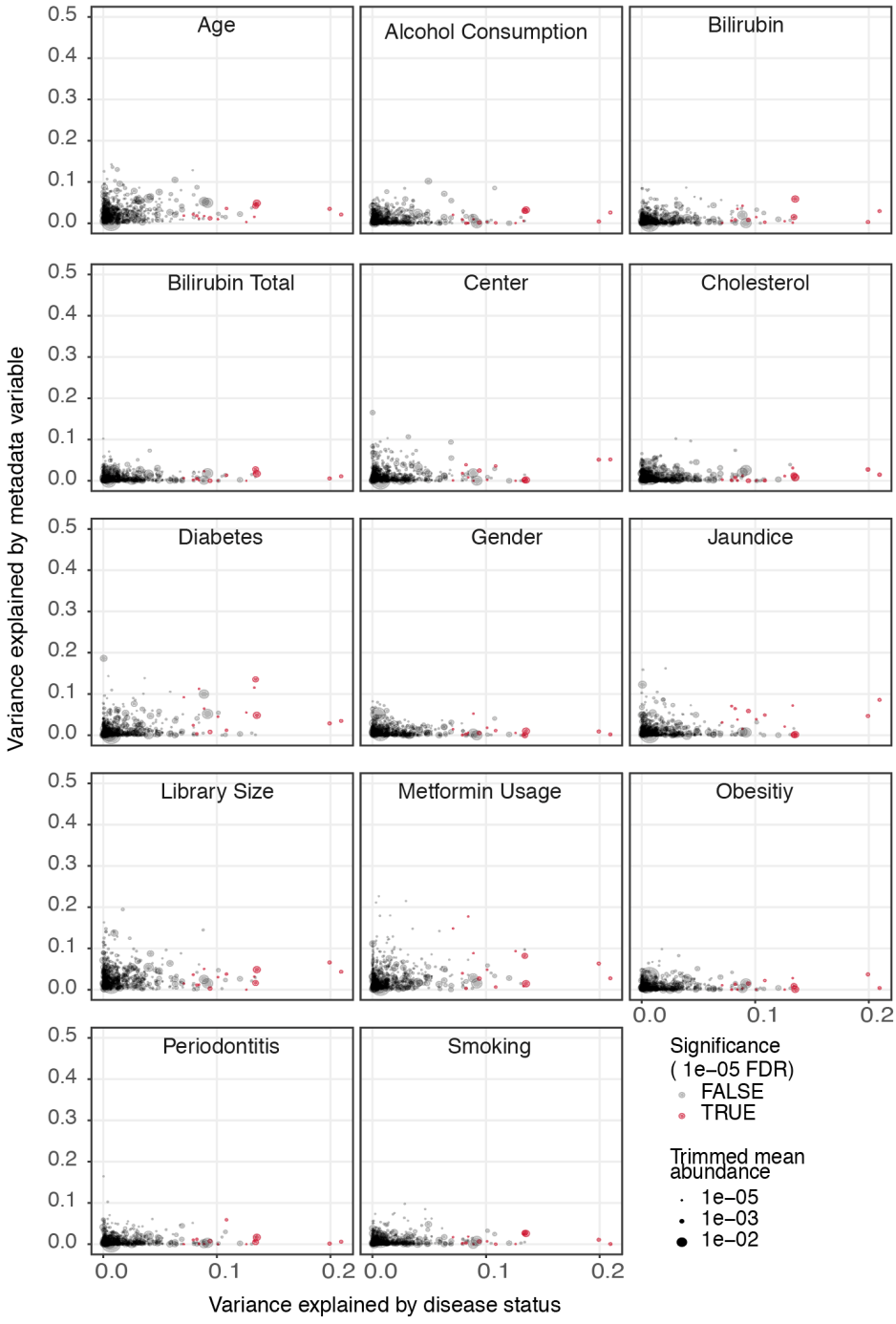


Figure S3. 1 Confounders analysis

Differentially abundant species are shown in red and the rest in in black. The size of the dot represents the mean abundance of species/strain. Disease status and tested variable was used as explanatory variable in linear model for feature abundance. This figure is part of the manuscript which is under preparation and it was produced by myself.

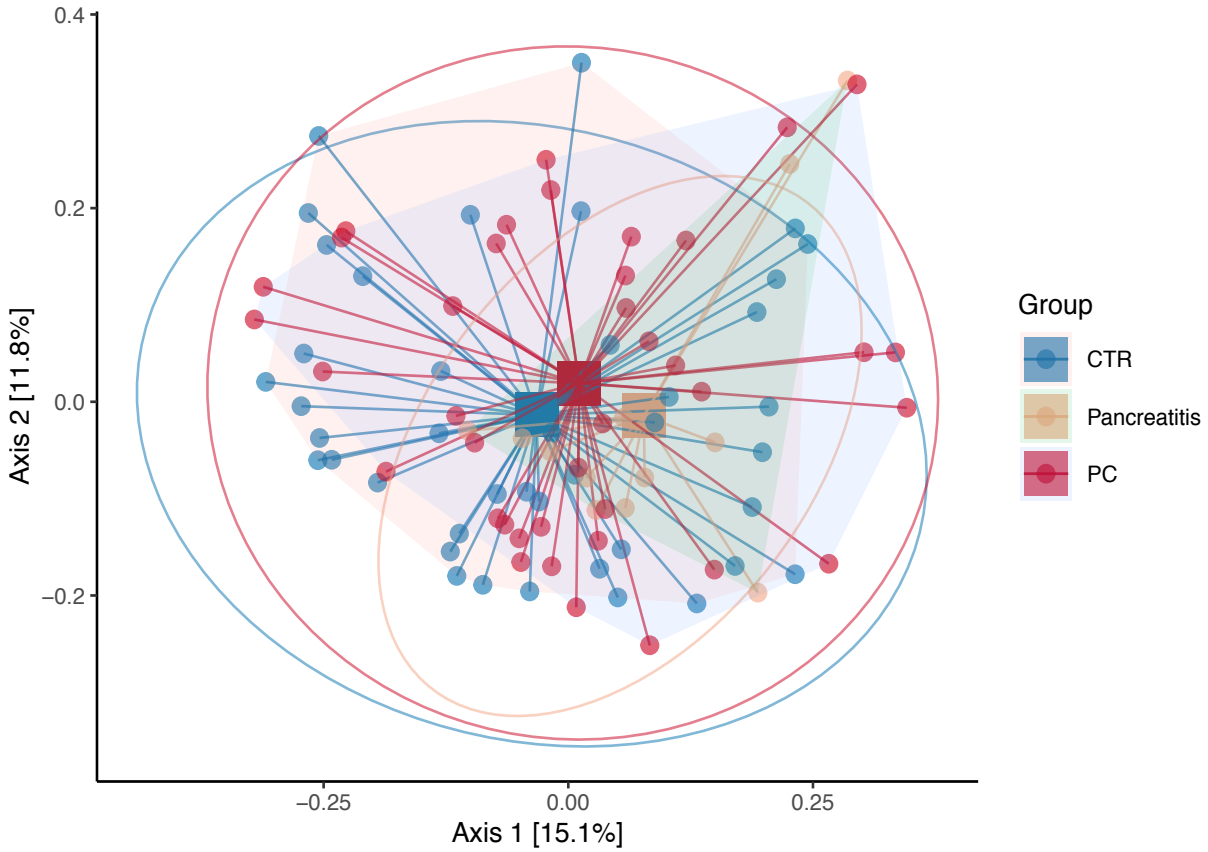


Figure S3. 2 PCoA plot based on Bray-Curtis dissimilarity of saliva microbiome.

PCoA plot based on Bray Curtis distance with the % explained by top two variance explained. PDAC cases: red, chronic pancreatitis: orange, controls: blue. This figure is part of the manuscript which is under preparation and it was produced by myself.

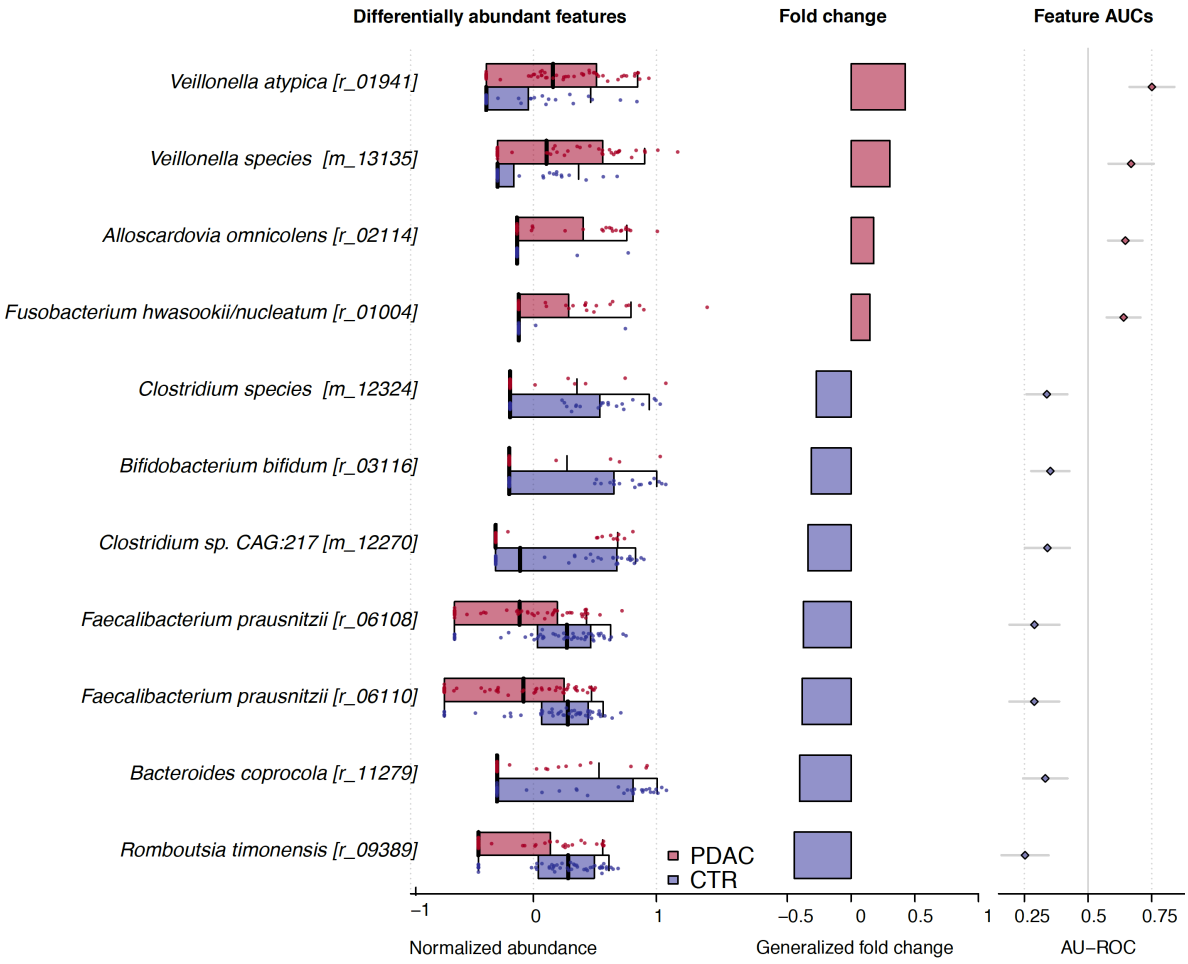


Figure S3. 3 Differential abundant species in fecal microbiome between PDAC cases and controls.

First panel shows the differentially abundant strains/species between cases (red) and controls (blue). Middle panel displays fold change of given strain and last panel presents the AUC of each feature to distinguish cases from controls. This figure is part of the manuscript which is under preparation and it was produced by myself.

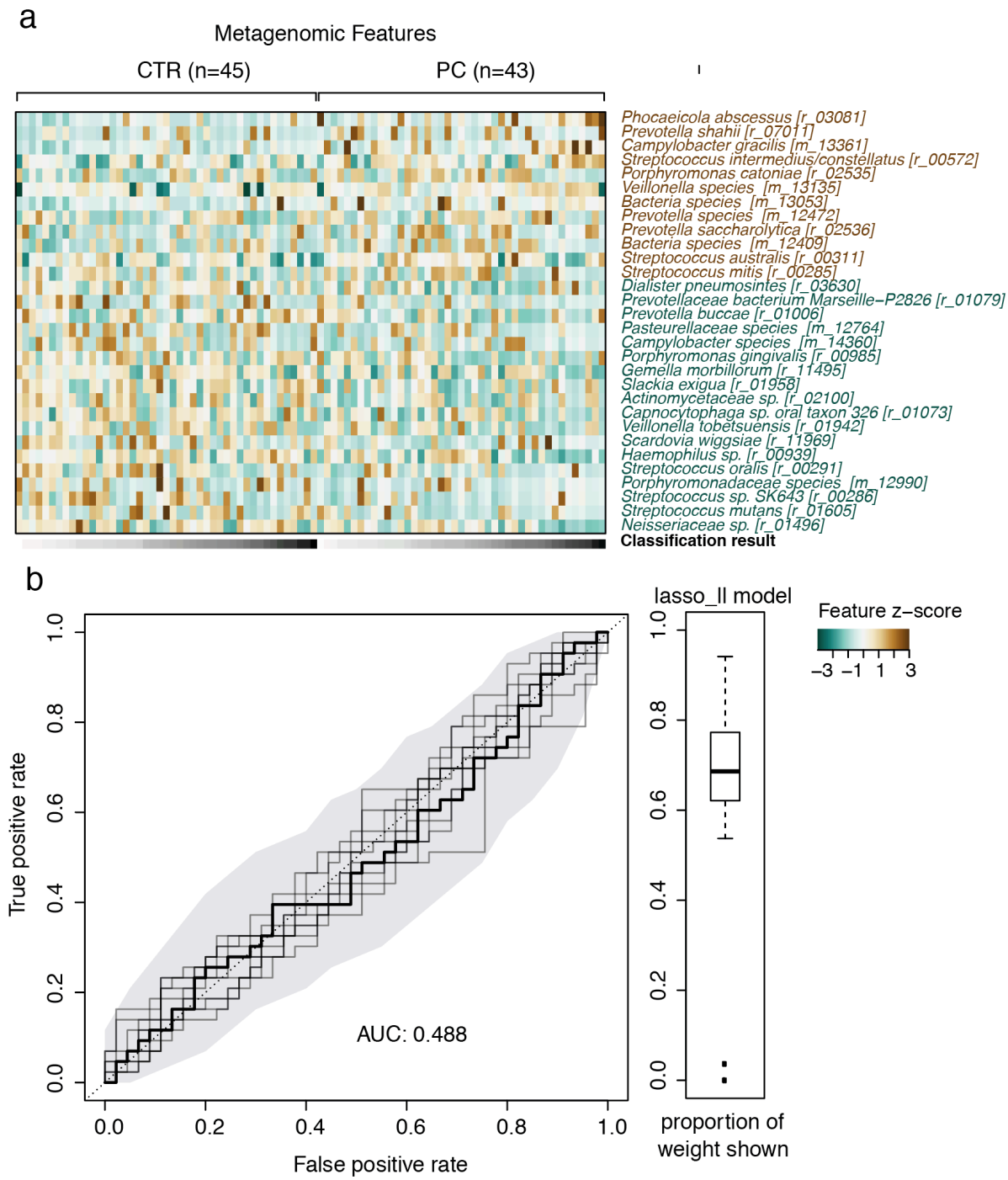


Figure S3. 4 Oral microbiome shows a weak power to distinguish cancer and control samples.

a. Heatmap represents the selected metagenomic features in the lasso_ll regression between cases and controls in saliva microbiome data. **b.** ROC curve based on 10 resampling and 10 cross validation (see 3.4 methods). All panels are part of the manuscript which is under preparation and they were produced by myself.

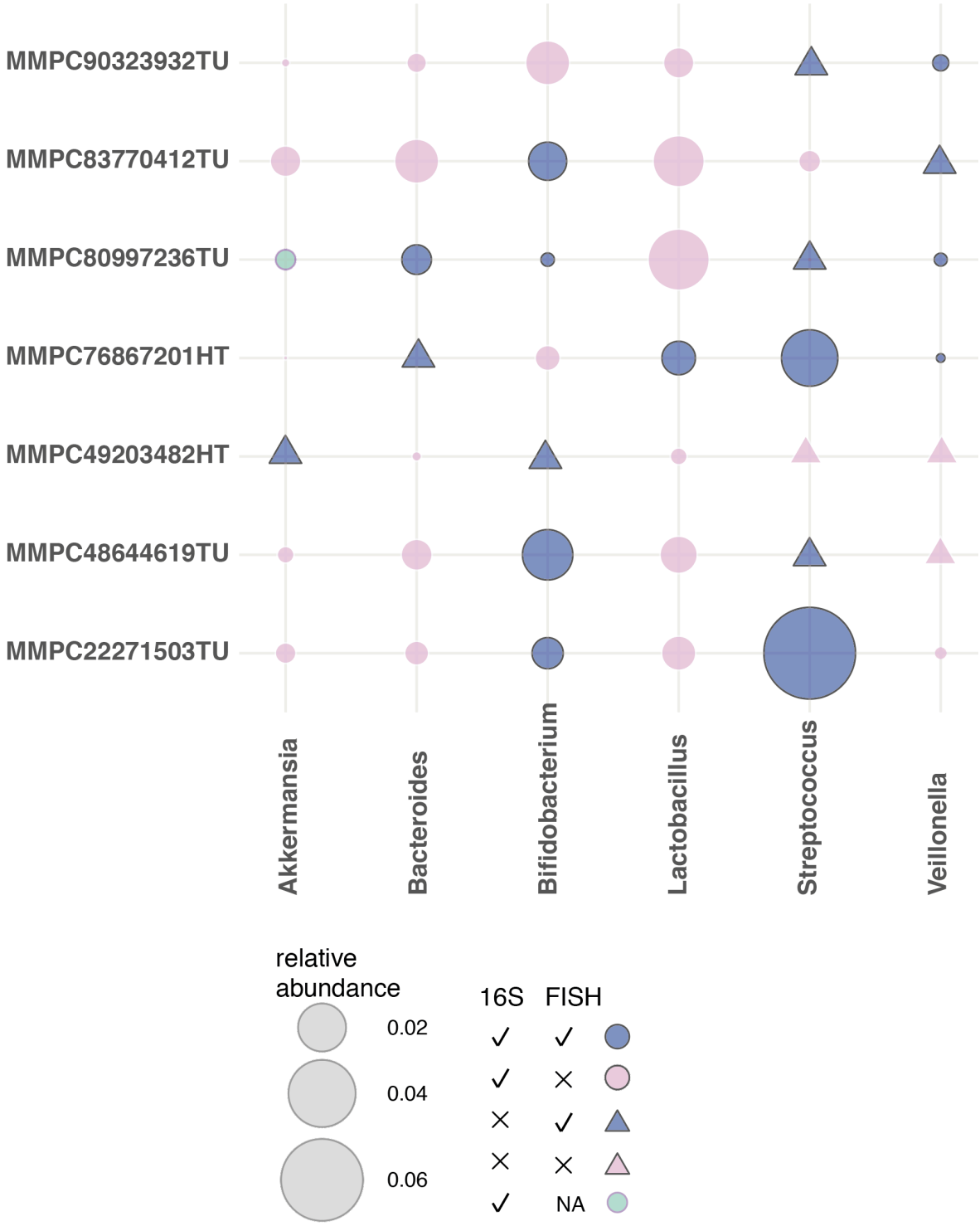


Figure S3. 5 Detailed information of tested samples via in-situ hybridization (FISH).

Rows displayed the tested sample IDs and columns shows the tested genus probes. The size of the dot represents relative abundance of genus in the given sample. Triangle shows that 16S was negative for given samples and color code show if FISH was positive or negative. One sample which displayed in green was not available for FISH testing. This figure is part of the manuscript which is under preparation and it was produced by myself.

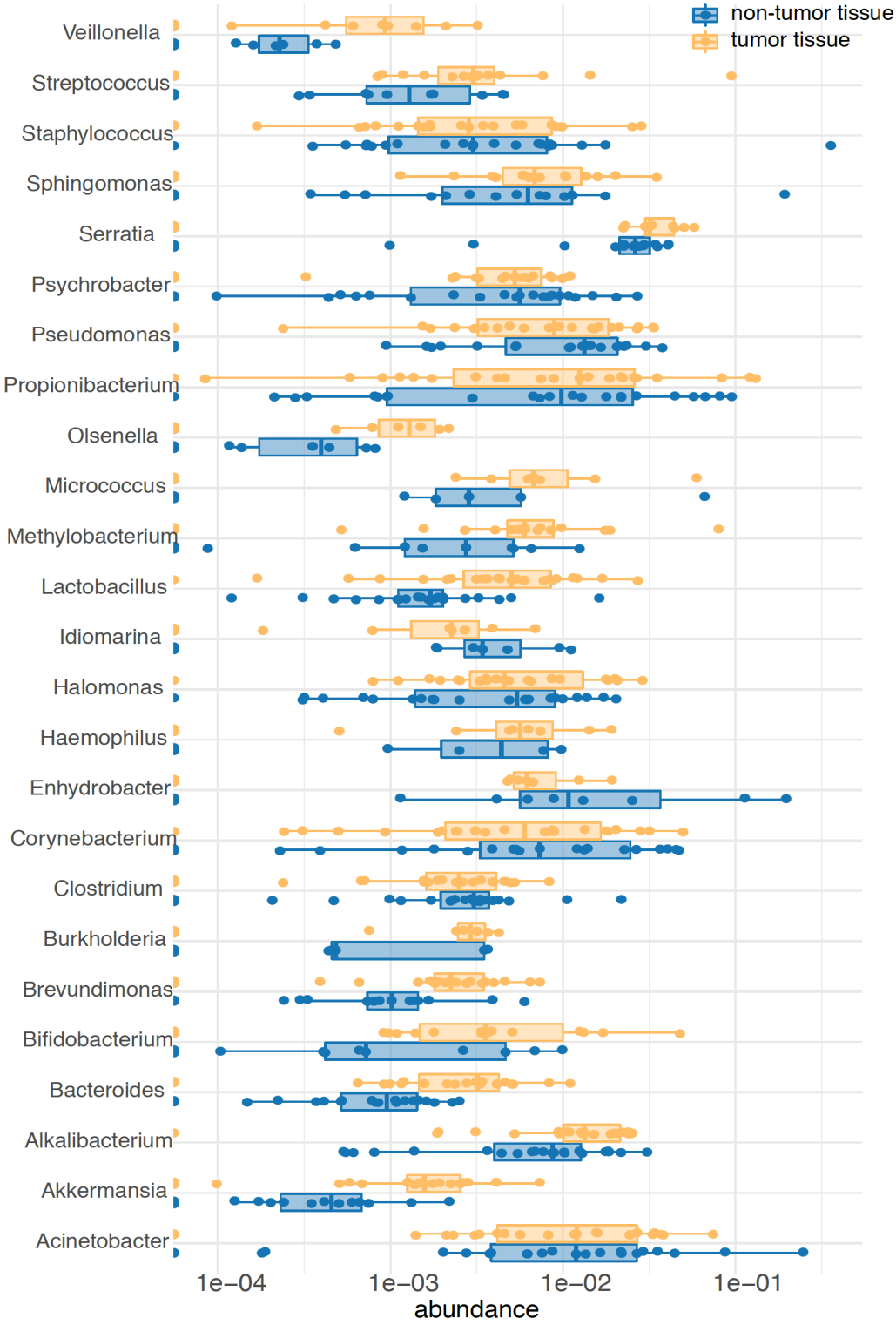


Figure S3. 6 Relative abundance of detected genus in pancreatic tumor and non-tumor tissue.

Relative abundance of detected genus in pancreatic tumor (orange) and non-tumor (blue) tissue is shown as bar plot. This figure is part of the manuscript which is under preparation and it was produced by Thomas Sebastian Schmidt.

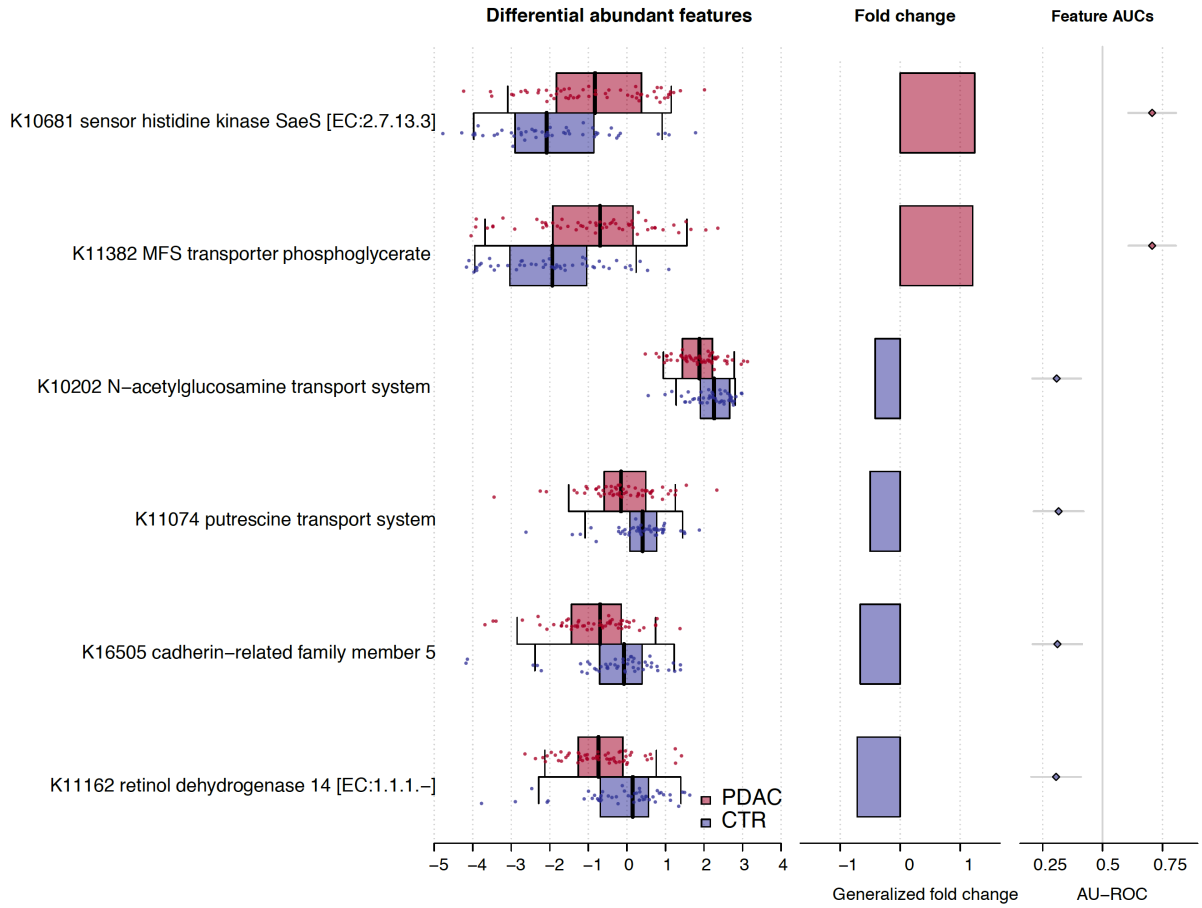


Figure S3. 7 Differential abundant KEGG modules between PDAC and controls.

First panel shows the differentially abundant KEGG modules between cases (red) and controls (blue). Middle panel displays fold change of given module and last panel presents the AUC of each feature to distinguish cases from controls. This figure is part of the manuscript which is under preparation and it was produced by myself.

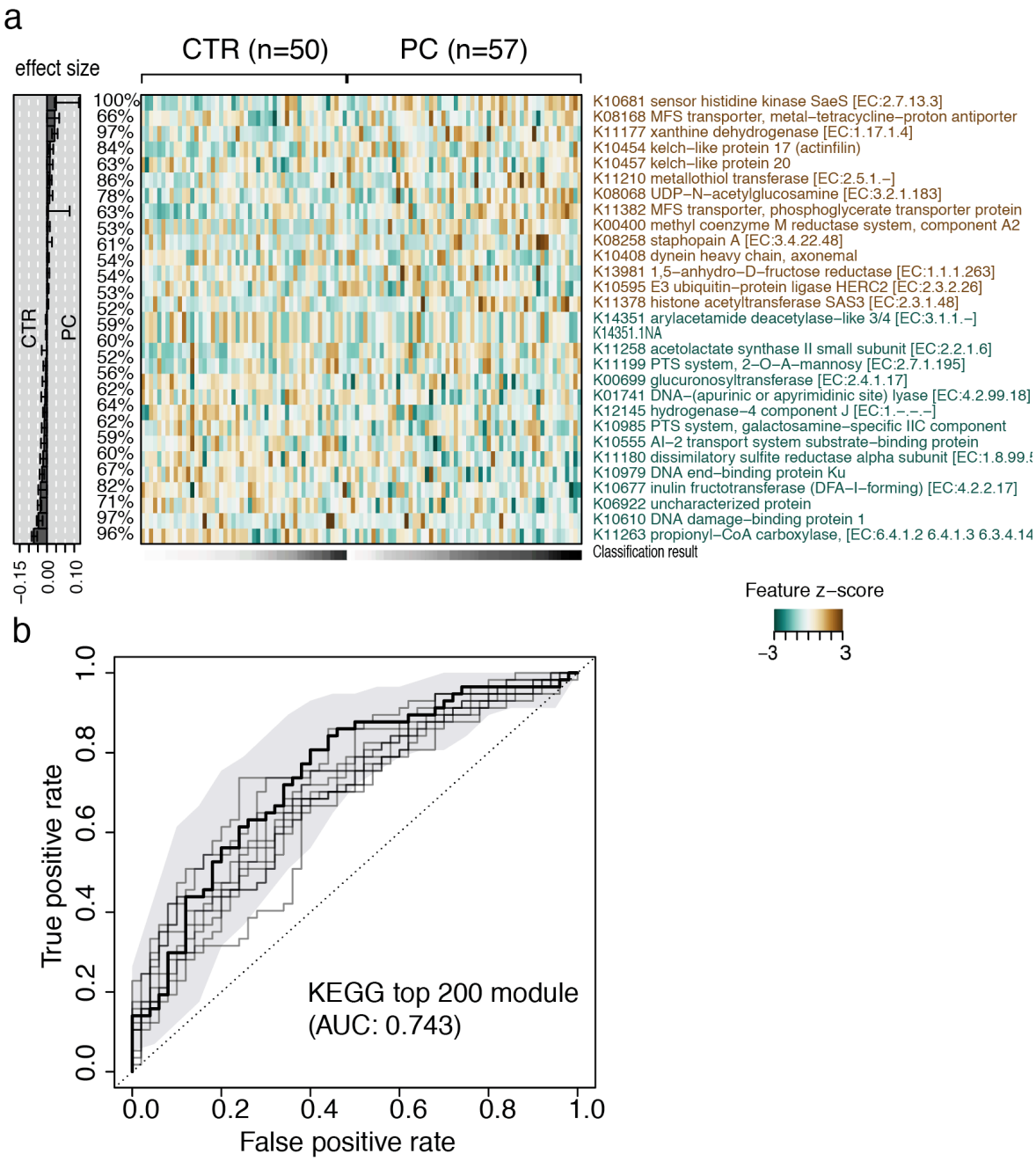


Figure S3. 8 Lasso_ll regression model based on top 200 KEGG modules.

a. Heatmap represents the selected KEGG modules in the lasso_ll regression model. **b.** ROC curve based on 10 resampling and 10 cross validation (see 3.4 methods). This figure is part of the manuscript which is under preparation and it was produced by myself.

Table S3. 1 Anova of richness diversity measure between PDAC cases and controls.

	Stool			Saliva		
Meta Variable	**Sum sq**	**p value**	**adj p**	**Sum sq**	**p value**	**p.adj**
Abmedication	7,40E+08	0.434882370	1.0	7,26E+07	0.726039062	1.0
Age	5,57E+10	0.348999552	1.0	1,43E+10	0.992133250	1.0
Alcohol Consumption	3,28E+09	0.105017752	1.0	4,22E+07	0.796587916	1.0
Acid Regurgitation	1,92E+09	0.215874045	1.0	2,05E+07	0.857472013	1.0
Diabetes	1,47E+07	0.914256854	1.0	4,78E+08	0.387311299	1.0
Heart Burn	7,20E+08	0.445550072	1.0	2,11E+09	0.059764110	1.0
High Blood Pressure	8,65E+06	0.934099012	1.0	4,70E+06	0.931832163	1.0
Rheumatoid Arthritis	7,36E+08	0.444113445	1.0	3,28E+07	0.821293457	1.0
Antibiotic	6,18E+07	0.825159237	1.0	3,55E+08	0.455050912	1.0
Aspirin/Paracetamol	1,36E+09	0.297843618	1.0	1,12E+09	0.181428572	1.0
Asthma	1,76E+09	0.236597169	1.0	4,57E+08	0.395506158	1.0
Center	4,90E+09	0.046953480	1.0	2,76E+06	0.947458361	1.0
Cholesterol	1,72E+09	0.246154381	1.0	2,93E+08	0.493014836	1.0
Cholesterol medication	6,33E+05	0.982226159	1.0	1,48E+08	0.663216095	1.0
Corticosteroids	2,24E+09	0.179684582	1.0	4,11E+08	0.406201881	1.0
Cpy1	7,16E+10	0.825078663	1.0	2,39E+10	0.800141666	1.0
Biliririrubin Direct	1,32E+11	0.543089087	1.0	5,08E+10	0.857634747	1.0
Bilirubin Lab	1,30E+11	0.455465155	1.0	4,85E+10	0.802784617	1.0
Fhpdac	2,40E+06	0.964420558	1.0	4,78E+08	0.383616587	1.0
Gender	1,04E+09	0.362627665	1.0	5,98E+08	0.330808453	1.0
Jaundice	9,45E+09	0.004406831	0.34373	4,30E+08	0.412200163	1.0
Jaundice Imputed	8,32E+09	0.009192609	0.70783	4,41E+08	0.403644362	1.0
Library Size	8,15E+09	0.086373435	1.0	1,55E+09	0.487294778	1.0
Metformin Usage	3,98E+09	0.205930855	1.0	3,12E+08	0.783170829	1.0
Obesity	8,37E+08	0.415414037	1.0	3,49E+08	0.450252695	1.0
Paracetamol	4,49E+09	0.058802839	1.0	2,03E+09	0.072918539	1.0
Periodontitis	1,73E+08	0.710712567	1.0	1,31E+09	0.136666489	1.0
Probiotic	1,11E+07	0.925379621	1.0	1,09E+07	0.894062783	1.0
Receding Gums	8,17E+08	0.415025012	1.0	2,69E+08	0.508708572	1.0
Salicylic Acid	1,85E+09	0.227061461	1.0	1,94E+08	0.582723500	1.0
Salicylic Acid	1,85E+09	0.227061461	1.0	1,94E+08	0.582723500	1.0
Smoking	9,62E+07	0.782721018	1.0	2,24E+09	0.059311754	1.0
Direct Bilirubin	2,75E+09	0.138132461	1.0	1,71E+09	0.102796407	1.0
Total Bilirubin	7,13E+08	0.440177205	1.0	2,87E+09	0.031822359	1.0

Table S3. 2 Adonis2 test of Bray-Curtis dissimilarity matrix between PDAC cases and controls of both stool and saliva microbiome.

	Stool		Saliva	
Meta Variable	**R^2 Bray-Curtis**	**p value**	**R^2 Bray-Curtis**	**p value**
Status	0.023120267	0.0001	0.010468749	0.6507
Center	0.011138008	0.2069	0.016751291	0.1285
Age	0.013765355	0.0320	0.018667727	0.0680
Gender	0.010546251	0.2884	0.017940759	0.0884
Jaundice Imputed	0.015980165	0.0087	0.007371362	0.9560
Diabetes	0.013610607	0.0396	0.008990599	0.8264
Obesity	0.008734297	0.6915	0.016455787	0.1449
Smoking	0.008943945	0.6466	0.015490596	0.1926
Alcohol Consumption	0.009730450	0.4545	0.017867998	0.0929
Periodontitis	0.009018863	0.6237	0.009449706	0.7709
Cholesterol	0.007848892	0.8750	0.010637989	0.6368
Metformin	0.008315828	0.7739	0.018288711	0.0898
Salicylic	0.009192529	0.5749	0.008768629	0.8462
Antibiotic	0.007858670	0.8685	0.006825246	0.9740
Aspirin/Paracetamol	0.012544261	0.0847	0.020375695	0.0432
Corticosteroids	0.010196054	0.3492	0.017140805	0.1277
Asthma	0.008577892	0.7244	0.014670641	0.2412
Acid Regurgitation	0.011797158	0.1335	0.006447749	0.9801
Rheumatoid Arthritis	0.008798014	0.6747	0.011646596	0.5123
Probiotic	0.009681840	0.4578	0.011088990	0.5759
Paracetamol	0.010198223	0.3573	0.009887242	0.7242
Heartburn	0.008810470	0.6760	0.013602585	0.3241
High Blood Pressure	0.007920655	0.8614	0.010496978	0.6491
Receding Gums	0.008765518	0.6873	0.007806693	0.9229
Fhpdac	0.010506425	0.2887	0.012944276	0.3752
Acid Med	0.010032433	0.3870	0.007517951	0.9390

Table S3. 3 Details of external validation cohorts.

Study	**Diagnosis**	**Cohort size**	**Accession number**
Dudenhoeffer 2018	CTR	400	PRJNA473126
FijiCOMP	CTR	312	PRJNA217052
Franzosa 2018	CTR	56	PRJNA400072
Franzosa 2018	IBD	164	PRJNA400072
Hall 2017	CTR	74	PRJNA385949
Hall 2017	IBD	187	PRJNA385949
Mardinoglu 2018	LD	48	PRJNA420817
Qin 2012	CTR	185	PRJNA422434
Qin 2012	T2D	183	PRJNA422434
Vogtmann 2016	CTR	52	PRJEB12449
Vogtmann 2016	CRC	51	PRJEB12449
Qin 2014	CTR	114	PRJEB6337
Qin 2014	LD	116	PRJEB6337
He 2017	CTR	54	PRJEB15371
He 2017	IBD	63	PRJEB15371
Karlsson 2013	CTR	43	PRJEB1786
Karlsson 2013	T2D	102	PRJEB1786
Feng 2015	CRC	93	PRJEB7774
Feng 2015	CTR	63	PRJEB7774
HMP2/Price-Lloyd 2019	IBD	235	PRJNA398089
HMP2/Price-Lloyd 2019	CTR	65	PRJNA398089
Liu 2016	CTR	110	PRJNA328899
Yu 2017	CRC	74	PRJEB10878
Yu 2017	CTR	54	PRJEB10878
Forslund 2015	CTR	372	PRJEB4336 PRJEB5224 PRJEB1220 PRJEB1786 PRJEB2054
Forslund 2015	T1D	31	
Forslund 2015	T2D	78	
Forslund 2015	IBD	229	

Bibliography

Ahmadmehrabi, S., & Tang, W. H. W. (2017). Gut microbiome and its role in cardiovascular diseases. *Current Opinion in Cardiology*, 32/6: 761–6. DOI: 10.1097/HCO.0000000000000445

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., et al. (2013). Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute*, 105/24: 1907–11. DOI: 10.1093/jnci/djt300

Alhinai, E. A., Walton, G. E., & Commane, D. M. (2019). The role of the gut microbiota in colorectal cancer causation. *International Journal of Molecular Sciences*, 20/21. DOI: 10.3390/ijms20215295

Allali, I., Delgado, S., Marron, P. I., Astudillo, A., Yeh, J. J., Ghazal, H., Amzazi, S., et al. (2015). Gut microbiome compositional and functional differences between tumor and non-tumor adjacent tissues from cohorts from the US and Spain. *Gut microbes*, 6/3: 161–72. DOI: 10.1080/19490976.2015.1039223

Armougom, F., Henry, M., Vialettes, B., Raccah, D., & Raoult, D. (2009). Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients. *Plos One*, 4/9: e7125. DOI: 10.1371/journal.pone.0007125

Aykut, B., Pushalkar, S., Chen, R., Li, Q., Abengozar, R., Kim, J. I., Shadaloey, S. A., et al. (2019). The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature*, 574/7777: 264–7. DOI: 10.1038/s41586-019-1608-2

Baecker, A., Kim, S., Risch, H. A., Nuckols, T. K., Wu, B. U., Hendifar, A. E., Pandol, S. J., et al. (2019). Do changes in health reveal the possibility of undiagnosed pancreatic cancer? Development of a risk-prediction model based on healthcare claims data. *Plos One*, 14/6: e0218580. DOI: 10.1371/journal.pone.0218580

Bhatti, M. A., & Frank, M. O. (2000). Veillonella parvula meningitis: case report and review of Veillonella infections. *Clinical Infectious Diseases*, 31/3: 839–40. DOI: 10.1086/314046

Böhm, J., Pianka, F., Stüttgen, N., Rho, J., Gigic, B., Zhang, Y., Habermann, N., et al. (2017). Discovery of novel plasma proteins as biomarkers for the development of incisional hernias after midline incision in patients with colorectal cancer: The ColoCare study. *Surgery*, 161/3: 808–17. DOI: 10.1016/j.surg.2016.08.025

Bose, M., & Mukherjee, P. (2019). Role of microbiome in modulating immune responses in cancer. *Mediators of Inflammation*, 2019: 4107917. DOI: 10.1155/2019/4107917

Bracci, P. M. (2017). Oral health and the oral microbiome in pancreatic cancer: an overview of epidemiological studies. *Cancer Journal*, 23/6: 310–4. DOI: 10.1097/PPO.0000000000000287

Brenner, H., Chang-Claude, J., Jansen, L., Knebel, P., Stock, C., & Hoffmeister, M. (2014). Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. *Gastroenterology*, 146/3: 709–17. DOI:

10.1053/j.gastro.2013.09.001

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13/7: 581–3. DOI: 10.1038/nmeth.3869

Cani, P. D., & Jordan, B. F. (2018). Gut microbiota-mediated inflammation in obesity: a link with gastrointestinal cancer. *Nature Reviews. Gastroenterology & Hepatology*, 15/11: 671–82. DOI: 10.1038/s41575-018-0025-6

Cao, Y., Shen, J., & Ran, Z. H. (2014). Association between Faecalibacterium prausnitzii Reduction and Inflammatory Bowel Disease: A Meta-Analysis and Systematic Review of the Literature. *Gastroenterology research and practice*, 2014: 872725. DOI: 10.1155/2014/872725

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl 1: 4516–22. DOI: 10.1073/pnas.1000080107

Carethers, J. M., & Jung, B. H. (2015). Genetics and genetic biomarkers in sporadic colorectal cancer. *Gastroenterology*, 149/5: 1177–1190.e3. DOI: 10.1053/j.gastro.2015.06.047

Chang, M.-C., Liang, P.-C., Jan, S., Yang, C.-Y., Tien, Y.-W., Wei, S.-C., Wong, J.-M., et al. (2014). Increase diagnostic accuracy in differentiating focal type autoimmune pancreatitis from pancreatic cancer with combined serum IgG4 and CA19-9 levels. *Pancreatology*, 14/5: 366–72. DOI: 10.1016/j.pan.2014.07.010

Chaudhary, P. P., Conway, P. L., & Schlundt, J. (2018). Methanogens in humans: potentially beneficial or harmful for health. *Applied Microbiology and Biotechnology*, 102/7: 3095–104. DOI: 10.1007/s00253-018-8871-2

Chong, P. P., Chin, V. K., Looi, C. Y., Wong, W. F., Madhavan, P., & Yong, V. C. (2019). The Microbiome and Irritable Bowel Syndrome - A Review on the Pathophysiology, Current Research and Future Therapy. *Frontiers in microbiology*, 10: 1136. DOI: 10.3389/fmicb.2019.01136

Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, 148/6: 1258–70. DOI: 10.1016/j.cell.2012.01.035

Coelho, L. P., Alves, R., Monteiro, P., Huerta-Cepas, J., Freitas, A. T., & Bork, P. (2019). NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome*, 7/1: 84. DOI: 10.1186/s40168-019-0684-8

Coghlin, C., & Murray, G. I. (2015). Biomarkers of colorectal cancer: recent advances and future challenges. *Proteomics. Clinical Applications*, 9/1–2: 64–71. DOI: 10.1002/prca.201400082

Coker, O. O., Nakatsu, G., Dai, R. Z., Wu, W. K. K., Wong, S. H., Ng, S. C., Chan, F. K. L., et al. (2019). Enteric fungal microbiota dysbiosis and ecological alterations in colorectal cancer. *Gut*, 68/4: 654–62. DOI: 10.1136/gutjnl-2018-317178

Collisson, E. A., Bailey, P., Chang, D. K., & Biankin, A. V. (2019). Molecular subtypes of pancreatic cancer. *Nature Reviews. Gastroenterology & Hepatology*, 16/4: 207–20. DOI: 10.1038/s41575-019-0109-y

Costea, Paul I, Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., Tramontano, M., et al. (2017). Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 35/11: 1069–76. DOI: 10.1038/nbt.3960

Costea, Paul Igor, Munch, R., Coelho, L. P., Paoli, L., Sunagawa, S., & Bork, P. (2017). metaSNV: A tool for metagenomic strain level analysis. *Plos One*, 12/7: e0182392. DOI: 10.1371/journal.pone.0182392

Czesnikiewicz-Guzik, M., & Müller, D. N. (2018). Scientists on the Spot: Salt, the microbiome, and cardiovascular diseases. *Cardiovascular Research*, 114/10: e72–3. DOI: 10.1093/cvr/cvy171

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K. K., Zhao, L., Chen, Z., Chan, F. K. L., et al. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome*, 6/1: 70. DOI: 10.1186/s40168-018-0451-2

D'Argenio, V. (2018). Human microbiome acquisition and bioinformatic challenges in metagenomic studies. *International Journal of Molecular Sciences*, 19/2. DOI: 10.3390/ijms19020383

Del Castillo, E., Meier, R., Chung, M., Koestler, D. C., Chen, T., Paster, B. J., Charpentier, K. P., et al. (2019). The Microbiomes of Pancreatic and Duodenum Tissue Overlap and Are Highly Subject Specific but Differ between Pancreatic Cancer and Noncancer Subjects. *Cancer Epidemiology, Biomarkers & Prevention*, 28/2: 370–83. DOI: 10.1158/1055-9965.EPI-18-0542

Drewes, J. L., White, J. R., Dejea, C. M., Fathi, P., Iyadorai, T., Vadivelu, J., Roslani, A. C., et al. (2017). High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *npj Biofilms and Microbiomes*, 3: 34. DOI: 10.1038/s41522-017-0040-3

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8/1: 1784. DOI: 10.1038/s41467-017-01973-8

Fan, X., Alekseyenko, A. V., Wu, J., Peters, B. A., Jacobs, E. J., Gapstur, S. M., Purdue, M. P., et al. (2018). Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut*, 67/1: 120–7. DOI: 10.1136/gutjnl-2016-312580

Farrell, J. J., Zhang, L., Zhou, H., Chia, D., Elashoff, D., Akin, D., Paster, B. J., et al. (2012). Variations of oral microbiota are associated with pancreatic diseases including pancreatic cancer. *Gut*, 61/4: 582–8. DOI: 10.1136/gutjnl-2011-300784

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature Communications*, 6: 6528. DOI: 10.1038/ncomms7528

Ferlay, J., Ervik, M., Lam, F., Colombet, M., & Mery, L. (2018). Global cancer observatory: cancer today. *for Research on Cancer*.

Flemer, B., Lynch, D. B., Brown, J. M. R., Jeffery, I. B., Ryan, F. J., Claesson, M. J., O'Riordain, M., et al. (2017). Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*, 66/4: 633–43. DOI: 10.1136/gutjnl-2015-309595

Gagnière, J., Raisch, J., Veziant, J., Barnich, N., Bonnet, R., Buc, E., Bringer, M.-A., et al. (2016). Gut microbiota imbalance and colorectal cancer. *World Journal of Gastroenterology*, 22/2: 501–18. DOI: 10.3748/wjg.v22.i2.501

Gao, R., Kong, C., Huang, L., Li, H., Qu, X., Liu, Z., Lan, P., et al. (2017). Mucosa-associated microbiota signature in colorectal cancer. *European Journal of Clinical Microbiology & Infectious Diseases*, 36/11: 2073–83. DOI: 10.1007/s10096-017-3026-4

Garrett, W. S. (2019). The gut microbiota and colon cancer. *Science*, 364/6446: 1133–5. DOI: 10.1126/science.aaw2367

Geller, L. T., Barzily-Rokni, M., Danino, T., Jonas, O. H., Shental, N., Nejman, D., Gavert, N., et al. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*, 357/6356: 1156–60. DOI: 10.1126/science.aah5043

Gentile, C. L., & Weir, T. L. (2018). The gut microbiota at the intersection of diet and human health. *Science*, 362/6416: 776–80. DOI: 10.1126/science.aau5812

Gharaibeh, R. Z., & Jobin, C. (2018). Microbiota and cancer immunotherapy: in search of microbial signals. *Gut*. DOI: 10.1136/gutjnl-2018-317220

Ghavami, S. B., Rostami, E., Sephay, A. A., Shahrokh, S., Balaii, H., Aghdaei, H. A., & Zali, M. R. (2018). Alterations of the human gut Methanobrevibacter smithii as a biomarker for inflammatory bowel diseases. *Microbial Pathogenesis*, 117: 285–9. DOI: 10.1016/j.micpath.2018.01.029

de Goffau, M. C., Lager, S., Sovio, U., Gaccioli, F., Cook, E., Peacock, S. J., Parkhill, J., et al. (2019). Human placenta has no microbiome but can contain potential pathogens. *Nature*, 572/7769: 329–34. DOI: 10.1038/s41586-019-1451-5

Goodwin, A. C., Destefano Shields, C. E., Wu, S., Huso, D. L., Wu, X., Murray-Stewart, T. R., Hacker-Prietz, A., et al. (2011). Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 108/37: 15354–9. DOI: 10.1073/pnas.1010203108

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., Prieto, P. A., et al. (2018). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science*, 359/6371: 97–103. DOI: 10.1126/science.aan4236

Gupta, H., Youn, G. S., Shin, M. J., & Suk, K. T. (2019). Role of gut microbiota in hepatocarcinogenesis. *Microorganisms*, 7/5. DOI: 10.3390/microorganisms7050121

Half, E., Keren, N., Reshef, L., Dorfman, T., Lachter, I., Kluger, Y., Reshef, N., et al. (2019). Fecal microbiome signatures of pancreatic cancer patients. *Scientific Reports*, 9/1: 16801. DOI: 10.1038/s41598-019-53041-4

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'Amato, M., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2: 17004. DOI: 10.1038/nmicrobiol.2017.4

Hall, A. B., Tolonen, A. C., & Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nature Reviews. Genetics*, 18/11: 690–9. DOI:

10.1038/nrg.2017.63

Hasan, S., Jacob, R., Manne, U., & Paluri, R. (2019). Advances in pancreatic cancer biomarkers. *Oncology reviews*, 13/1: 410. DOI: 10.4081/oncol.2019.410

Helleputte, T., & Gramme, P. (2017). LiblineaR: Linear predictive models based on the LIBLINEAR C/C++ Library. *R package version*.

Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. *Nature Medicine*, 25/3: 377–88. DOI: 10.1038/s41591-019-0377-7

Hidalgo, M., Cascinu, S., Kleeff, J., Labianca, R., Löhr, J.-M., Neoptolemos, J., Real, F. X., et al. (2015). Addressing the challenges of pancreatic cancer: future directions for improving outcomes. *Pancreatology*, 15/1: 8–18. DOI: 10.1016/j.pan.2014.10.001

Hsieh, M. H., Sun, L.-M., Lin, C.-L., Hsieh, M.-J., Hsu, C.-Y., & Kao, C.-H. (2018). Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer management and research*, 10: 6317–24. DOI: 10.2147/CMAR.S180791

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44/D1: D286-93. DOI: 10.1093/nar/gkv1248

Jamshidi, P., Hasanzadeh, S., Tahvildari, A., Farsi, Y., Arbabi, M., Mota, J. F., Sechi, L. A., et al. (2019). Is there any association between gut microbiota and type 1 diabetes? A systematic review. *Gut Pathogens*, 11: 49. DOI: 10.1186/s13099-019-0332-7

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., & Nageshwar Reddy, D. (2015). Role of the normal gut microbiota. *World Journal of Gastroenterology*, 21/29: 8787–803. DOI: 10.3748/wjg.v21.i29.8787

Jangi, S., Gandhi, R., Cox, L. M., Li, N., von Glehn, F., Yan, R., Patel, B., et al. (2016). Alterations of the human gut microbiome in multiple sclerosis. *Nature Communications*, 7: 12015. DOI: 10.1038/ncomms12015

Kåhrström, C. T., Pariente, N., & Weiss, U. (2016). Intestinal microbiota in health and disease. *Nature*, 535/7610: 47. DOI: 10.1038/535047a

Kamada, N., Seo, S.-U., Chen, G. Y., & Núñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. *Nature Reviews. Immunology*, 13/5: 321–35. DOI: 10.1038/nri3430

Kamisawa, T., Wood, L. D., Itoi, T., & Takaori, K. (2016). Pancreatic cancer. *The Lancet*, 388/10039: 73–85. DOI: 10.1016/S0140-6736(16)00141-0

Keum, N., & Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews. Gastroenterology & Hepatology*, 16/12: 713–32. DOI: 10.1038/s41575-019-0189-8

Kirkegård, J., Mortensen, F. V., Heide-Jørgensen, U., & Cronin-Fenton, D. (2019). Predictors of underlying pancreatic cancer in patients with acute pancreatitis: a Danish nationwide cohort study. *HPB: the Official Journal of the International Hepato Pancreato Biliary Association*. DOI: 10.1016/j.hpb.2019.08.013

Konig, M. F. (2020). The microbiome in autoimmune rheumatic disease. *Best Practice &*

Research. Clinical Rheumatology, 101473. DOI: 10.1016/j.berh.2019.101473

Konstantinov, S. R., Kuipers, E. J., & Peppelenbosch, M. P. (2013). Functional genomic analyses of the gut microbiota for CRC screening. *Nature Reviews. Gastroenterology & Hepatology*, 10/12: 741–5. DOI: 10.1038/nrgastro.2013.178

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., Clancy, T. E., et al. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe*, 14/2: 207–15. DOI: 10.1016/j.chom.2013.07.007

Kramski, M., Gaeguta, A. J., Lichtfuss, G. F., Rajasuriar, R., Crowe, S. M., French, M. A., Lewin, S. R., et al. (2011). Novel sensitive real-time PCR for quantification of bacterial 16S rRNA genes in plasma of HIV-infected patients as a marker for microbial translocation. *Journal of Clinical Microbiology*, 49/10: 3691–3. DOI: 10.1128/JCM.01018-11

Lee, J. K., Liles, E. G., Bent, S., Levin, T. R., & Corley, D. A. (2014). Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. *Annals of Internal Medicine*, 160/3: 171. DOI: 10.7326/M13-1484

Liesenfeld, D. B., Grapov, D., Fahrmann, J. F., Salou, M., Scherer, D., Toth, R., Habermann, N., et al. (2015). Metabolomics and transcriptomics identify pathway differences between visceral and subcutaneous adipose tissue in colorectal cancer patients: the ColoCare study. *The American Journal of Clinical Nutrition*, 102/2: 433–43. DOI: 10.3945/ajcn.114.103804

Loktionov, A. (2020). Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins? *World journal of gastrointestinal oncology*, 12/2: 124–48. DOI: 10.4251/wjgo.v12.i2.124

Lynch, S. V., & Pedersen, O. (2016). The human intestinal microbiome in health and disease. *The New England Journal of Medicine*, 375/24: 2369–79. DOI: 10.1056/NEJMra1600266

Maekawa, T., Fukaya, R., Takamatsu, S., Itoyama, S., Fukuoka, T., Yamada, M., Hata, T., et al. (2018). Possible involvement of Enterococcus infection in the pathogenesis of chronic pancreatitis and cancer. *Biochemical and Biophysical Research Communications*, 506/4: 962–9. DOI: 10.1016/j.bbrc.2018.10.169

Maisonneuve, P., Amar, S., & Lowenfels, A. B. (2017). Periodontal disease, edentulism, and pancreatic cancer: a meta-analysis. *Annals of Oncology*, 28/5: 985–95. DOI: 10.1093/annonc/mdx019

Malka, D., Hammel, P., Maire, F., Rufat, P., Madeira, I., Pessione, F., Lévy, P., et al. (2002). Risk of pancreatic adenocarcinoma in chronic pancreatitis. *Gut*, 51/6: 849–52. DOI: 10.1136/gut.51.6.849

Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H. M., Roelofs, R., Boleij, A., & Tjalsma, H. (2011). Towards the human colorectal cancer microbiome. *Plos One*, 6/5: e20447. DOI: 10.1371/journal.pone.0020447

Martinez-Guryn, K., Leone, V., & Chang, E. B. (2019). Regional diversity of the gastrointestinal microbiome. *Cell Host & Microbe*, 26/3: 314–24. DOI: 10.1016/j.chom.2019.08.011

Maruvada, P., Leone, V., Kaplan, L. M., & Chang, E. B. (2017). The Human Microbiome and Obesity: Moving beyond Associations. *Cell Host & Microbe*, 22/5: 589–99. DOI: 10.1016/j.chom.2017.10.005

Matias Rodrigues, J. F., & von Mering, C. (2014). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics*, 30/2: 287–8. DOI: 10.1093/bioinformatics/btt657

Matias Rodrigues, J. F., Schmidt, T. S. B., Tackmann, J., & von Mering, C. (2017). MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics*, 33/23: 3808–10. DOI: 10.1093/bioinformatics/btx517

Matson, V., Fessler, J., Bao, R., Chongsuwat, T., Zha, Y., Alegre, M.-L., Luke, J. J., et al. (2018). The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science*, 359/6371: 104–8. DOI: 10.1126/science.aao3290

Mende, D. R., Letunic, I., Maistrenko, O. M., Schmidt, T. S. B., Milanese, A., Paoli, L., Hernández-Plaza, A., et al. (2020). proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*, 48/D1: D621–5. DOI: 10.1093/nar/gkz1002

Meng, W., Bai, B., Sheng, L., Li, Y., Yue, P., Li, X., & Qiao, L. (2015). Role of Helicobacter pylori in gastric cancer: advances and controversies. *Discovery medicine*, 20/111: 285–93.

Michaud, D. S., Izard, J., Wilhelm-Benartzi, C. S., You, D.-H., Grote, V. A., Tjønneland, A., Dahm, C. C., et al. (2013a). Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. *Gut*, 62/12: 1764–70. DOI: 10.1136/gutjnl-2012-303006

——. (2013b). Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. *Gut*, 62/12: 1764–70. DOI: 10.1136/gutjnl-2012-303006

Midha, S., Chawla, S., & Garg, P. K. (2016). Modifiable and non-modifiable risk factors for pancreatic cancer: A review. *Cancer Letters*, 381/1: 269–77. DOI: 10.1016/j.canlet.2016.07.022

Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., Hingamp, P., et al. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications*, 10/1: 1014. DOI: 10.1038/s41467-019-08844-4

Mira-Pascual, L., Cabrera-Rubio, R., Ocon, S., Costales, P., Parra, A., Suarez, A., Moris, F., et al. (2015). Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers. *Journal of Gastroenterology*, 50/2: 167–79. DOI: 10.1007/s00535-014-0963-x

Moskal, A., Freisling, H., Byrnes, G., Assi, N., Fahey, M. T., Jenab, M., Ferrari, P., et al. (2016). Main nutrient patterns and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition study. *British Journal of Cancer*, 115/11: 1430–40. DOI: 10.1038/bjc.2016.334

Muhammad, W., Hart, G. R., Nartowt, B., Farrell, J. J., Johung, K., Liang, Y., & Deng, J. (2019). Pancreatic cancer prediction through an artificial neural network. *Frontiers in Artificial Intelligence*, 2. DOI: 10.3389/frai.2019.00002

Nadkarni, M. A., Martin, F. E., Jacques, N. A., & Hunter, N. (2002). Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology*, 148/Pt 1: 257–66. DOI: 10.1099/00221287-148-1-257

Nagy, K. N., Sonkodi, I., Szöke, I., Nagy, E., & Newman, H. N. (1998). The microflora associated with human oral carcinomas. *Oral Oncology*, 34/4: 304–8.

Nakatochi, M., Lin, Y., Ito, H., Hara, K., Kinoshita, F., Kobayashi, Y., Ishii, H., et al. (2018). Prediction model for pancreatic cancer risk in the general Japanese population. *Plos One*, 13/9: e0203386. DOI: 10.1371/journal.pone.0203386

Nakatsu, G., Zhou, H., Wu, W. K. K., Wong, S. H., Coker, O. O., Dai, Z., Li, X., et al. (2018). Alterations in enteric virome are associated with colorectal cancer and survival outcomes. *Gastroenterology*, 155/2: 529–541.e5. DOI: 10.1053/j.gastro.2018.04.018

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29/22: 2933–5. DOI: 10.1093/bioinformatics/btt509

Nguyen, T. L. A., Vieira-Silva, S., Liston, A., & Raes, J. (2015). How informative is the mouse for human gut microbiota research? *Disease Models & Mechanisms*, 8/1: 1–16. DOI: 10.1242/dmm.017400

Nolan, T., Hands, R. E., Ogunkolade, W., & Bustin, S. A. (2006). SPUD: a quantitative PCR assay for the detection of inhibitors in nucleic acid preparations. *Analytical Biochemistry*, 351/2: 308–10. DOI: 10.1016/j.ab.2006.01.051

O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Reports*, 7/7: 688–93. DOI: 10.1038/sj.embor.7400731

Oksanen, J., Blanchet, F. G., & Kindt, R. (2010). Vegan: community ecology package. R package version 1.17-4. */package= vegan*.

Olson, S. H., Satagopan, J., Xu, Y., Ling, L., Leong, S., Orlow, I., Saldia, A., et al. (2017). The oral microbiota in patients with pancreatic cancer, patients with IPMNs, and controls: a pilot study. *Cancer Causes & Control*, 28/9: 959–69. DOI: 10.1007/s10552-017-0933-8

Peek, R. M., & Blaser, M. J. (2002). Helicobacter pylori and gastrointestinal tract adenocarcinomas. *Nature Reviews. Cancer*, 2/1: 28–37. DOI: 10.1038/nrc703

Petersen, G. M. (2016). Familial pancreatic cancer. *Seminars in Oncology*, 43/5: 548–53. DOI: 10.1053/j.seminoncol.2016.09.002

Pleguezuelos-Manzano, C., Puschhof, J., Huber, A. R., van Hoeck, A., Wood, H. M., Nomburg, J., Gurjao, C., et al. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature*. DOI: 10.1038/s41586-020-2080-8

Pox, C. P., Altenhofen, L., Brenner, H., Theilmeier, A., Von Stillfried, D., & Schmiegel, W. (2012). Efficacy of a nationwide screening colonoscopy program for colorectal cancer. *Gastroenterology*, 142/7: 1460–7.e2. DOI: 10.1053/j.gastro.2012.03.022

Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C. P., Kurz, E., Mishra, A., Mohan, N., et al. (2018). The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer discovery*, 8/4: 403–16. DOI: 10.1158/2159-8290.CD-17-1134

Pustelny, C., Komor, U., Pawar, V., Lorenz, A., Bielecka, A., Moter, A., Gocht, B., et al. (2015). Contribution of Veillonella parvula to Pseudomonas aeruginosa-mediated

pathogenicity in a murine tumor model system. *Infection and Immunity*, 83/1: 417–29. DOI: 10.1128/IAI.02234-14

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., & Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Research*, 74/11: 2913–21. DOI: 10.1158/0008-5472.CAN-14-0155

Rajilic-Stojanovic, M., Figueiredo, C., Smet, A., Hansen, R., Kupcinskas, J., Rokkas, T., Andersen, L., et al. (2020). Systematic review: gastric microbiota in health and disease. *Alimentary Pharmacology & Therapeutics*. DOI: 10.1111/apt.15650

Rawla, P., Sunkara, T., & Gaduputi, V. (2019). Epidemiology of pancreatic cancer: global trends, etiology and risk factors. *World journal of oncology*, 10/1: 10–27. DOI: 10.14740/wjon1166

Ren, Z., Jiang, J., Xie, H., Li, A., Lu, H., Xu, S., Zhou, L., et al. (2017). Gut microbial profile analysis by MiSeq sequencing of pancreatic carcinoma patients in China. *Oncotarget*, 8/56: 95176–91. DOI: 10.18632/oncotarget.18820

Ren, Z., Li, A., Jiang, J., Zhou, L., Yu, Z., Lu, H., Xie, H., et al. (2019). Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. *Gut*, 68/6: 1014–23. DOI: 10.1136/gutjnl-2017-315084

Rezasoltani, S., Sharafkhah, M., Asadzadeh Aghdaei, H., Nazemalhosseini Mojarad, E., Dabiri, H., Akhavan Sepahi, A., Modarressi, M. H., et al. (2018). Applying simple linear combination, multiple logistic and factor analysis methods for candidate fecal bacteria as novel biomarkers for early detection of adenomatous polyps and colon cancer. *Journal of Microbiological Methods*, 155: 82–8. DOI: 10.1016/j.mimet.2018.11.007

Richard, M. L., Liguori, G., Lamas, B., Brandi, G., da Costa, G., Hoffmann, T. W., Pierluigi Di Simone, M., et al. (2018). Mucosa-associated microbiota dysbiosis in colitis associated cancer. *Gut microbes*, 9/2: 131–42. DOI: 10.1080/19490976.2017.1379637

Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., et al. (2019). Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell*, 178/4: 795–806.e12. DOI: 10.1016/j.cell.2019.07.008

Ritalahti, K. M., Amos, B. K., Sung, Y., Wu, Q., Koenigsberg, S. S., & Löffler, F. E. (2006). Quantitative PCR targeting 16S rRNA and reductive dehalogenase genes simultaneously monitors multiple Dehalococcoides strains. *Applied and Environmental Microbiology*, 72/4: 2765–74. DOI: 10.1128/AEM.72.4.2765-2774.2006

Roberts, A. B., Wallace, B. D., Venkatesh, M. K., Mani, S., & Redinbo, M. R. (2013). Molecular insights into microbial β-glucuronidase inhibition to abrogate CPT-11 toxicity. *Molecular Pharmacology*, 84/2: 208–17. DOI: 10.1124/mol.113.085852

Robertson, D. J., Lee, J. K., Boland, C. R., Dominitz, J. A., Giardiello, F. M., Johnson, D. A., Kaltenbach, T., et al. (2017). Recommendations on Fecal Immunochemical Testing to Screen for Colorectal Neoplasia: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology*, 152/5: 1217–1237.e3. DOI: 10.1053/j.gastro.2016.08.053

Rodríguez, J. M., Murphy, K., Stanton, C., Ross, R. P., Kober, O. I., Juge, N., Avershina, E., et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on

early life. *Microbial ecology in health and disease*, 26: 26050. DOI: 10.3402/mehd.v26.26050

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P. M., Alou, M. T., Daillère, R., Fluckiger, A., et al. (2018). Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science*, 359/6371: 91–7. DOI: 10.1126/science.aan3706

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., & Han, Y. W. (2013). Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host & Microbe*, 14/2: 195–206. DOI: 10.1016/j.chom.2013.07.012

Rutter, M. D., Nickerson, C., Rees, C. J., Patnick, J., & Blanks, R. G. (2014). Risk factors for adverse events related to polypectomy in the English Bowel Cancer Screening Programme. *Endoscopy*, 46/2: 90–7. DOI: 10.1055/s-0033-1344987

Săftoiu, A., Vilmann, P., Gorunescu, F., Gheonea, D. I., Gorunescu, M., Ciurea, T., Popescu, G. L., et al. (2008). Neural network analysis of dynamic sequences of EUS elastography used for the differential diagnosis of chronic pancreatitis and pancreatic cancer. *Gastrointestinal Endoscopy*, 68/6: 1086–94. DOI: 10.1016/j.gie.2008.04.031

Saito, K., Koido, S., Odamaki, T., Kajihara, M., Kato, K., Horiuchi, S., Adachi, S., et al. (2019). Metagenomic analyses of the gut microbiota associated with colorectal adenoma. *Plos One*, 14/2: e0212406. DOI: 10.1371/journal.pone.0212406

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12: 87. DOI: 10.1186/s12915-014-0087-z

Schmidt, T. S., Hayward, M. R., Coelho, L. P., Li, S. S., Costea, P. I., Voigt, A. Y., Wirbel, J., et al. (2019). Extensive transmission of microbes along the gastrointestinal tract. *eLife*, 8. DOI: 10.7554/eLife.42693

Schmidt, Thomas S B, Matias Rodrigues, J. F., & von Mering, C. (2015). Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environmental Microbiology*, 17/5: 1689–706. DOI: 10.1111/1462-2920.12610

Schmidt, Thomas S B, Raes, J., & Bork, P. (2018). The human gut microbiome: from association to modulation. *Cell*, 172/6: 1198–215. DOI: 10.1016/j.cell.2018.02.044

Schmidt, Thomas Sebastian Benedikt, Matias Rodrigues, J. F., & von Mering, C. (2017). A family of interaction-adjusted indices of community similarity. *The ISME Journal*, 11/3: 791–807. DOI: 10.1038/ismej.2016.139

Schwabe, R. F., & Greten, T. F. (2020). Gut microbiome in HCC - Mechanisms, diagnosis and therapy. *Journal of Hepatology*, 72/2: 230–8. DOI: 10.1016/j.jhep.2019.08.016

Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews. Cancer*, 13/11: 800–12. DOI: 10.1038/nrc3610

Scott, A. J., Alexander, J. L., Merrifield, C. A., Cunningham, D., Jobin, C., Brown, R., Alverdy, J., et al. (2019). International Cancer Microbiome Consortium consensus statement on the role of the human microbiome in carcinogenesis. *Gut*, 68/9: 1624–32. DOI: 10.1136/gutjnl-2019-318556

Sears, C. L., & Garrett, W. S. (2014). Microbes, microbiota, and colon cancer. *Cell Host & Microbe*, 15/3: 317–28. DOI: 10.1016/j.chom.2014.02.007

Sekirov, I., Russell, S. L., Antunes, L. C. M., & Finlay, B. B. (2010). Gut microbiota in health and disease. *Physiological Reviews*, 90/3: 859–904. DOI: 10.1152/physrev.00045.2009

Sethi, V., Kurtom, S., Tarique, M., Lavania, S., Malchiodi, Z., Hellmund, L., Zhang, L., et al. (2018). Gut microbiota promotes tumor growth in mice by modulating immune response. *Gastroenterology*, 155/1: 33–37.e6. DOI: 10.1053/j.gastro.2018.04.001

Shah, M. S., DeSantis, T. Z., Weinmaier, T., McMurdie, P. J., Cope, J. L., Altrichter, A., Yamal, J.-M., et al. (2018). Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut*, 67/5: 882–91. DOI: 10.1136/gutjnl-2016-313189

Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go? *The Journal of Nutritional Biochemistry*, 63: 101–8. DOI: 10.1016/j.jnutbio.2018.10.003

Shen, L., & Ji, H.-F. (2019). Associations between gut microbiota and alzheimer's disease: current evidences and future therapeutic and diagnostic perspectives. *Journal of Alzheimer's Disease*, 68/1: 25–31. DOI: 10.3233/JAD-181143

Sheng, Q.-S., He, K.-X., Li, J.-J., Zhong, Z.-F., Wang, F.-X., Pan, L.-L., & Lin, J.-J. (2020). Comparison of gut microbiome in human colorectal cancer in paired tumor and adjacent normal tissues. *OncoTargets and therapy*, 13: 635–46. DOI: 10.2147/OTT.S218004

Shreiner, A. B., Kao, J. Y., & Young, V. B. (2015). The gut microbiome in health and in disease. *Current Opinion in Gastroenterology*, 31/1: 69–75. DOI: 10.1097/MOG.0000000000000139

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., & Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67/3: 177–93. DOI: 10.3322/caac.21395

Sivan, A., Corrales, L., Hubert, N., Williams, J. B., Aquino-Michaels, K., Earley, Z. M., Benyamin, F. W., et al. (2015). Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science*, 350/6264: 1084–9. DOI: 10.1126/science.aac4255

Sogodogo, E., Doumbo, O., Aboudharam, G., Kouriba, B., Diawara, O., Koita, H., Togora, S., et al. (2019). First characterization of methanogens in oral cavity in Malian patients with oral cavity pathologies. *BMC oral health*, 19/1: 232. DOI: 10.1186/s12903-019-0929-8

The MetaHIT Consortium, & Ehrlich, S. D. (2011). Metahit: the european union project on metagenomics of the human intestinal tract. Nelson K. E. (ed.) *Metagenomics of the human body*, pp. 307–16. Springer New York: New York, NY. DOI: 10.1007/978-1-4419-7089-3_15

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, 25/4: 667–78. DOI: 10.1038/s41591-019-0405-7

Thomas, R. M., Gharaibeh, R. Z., Gauthier, J., Beveridge, M., Pope, J. L., Guijarro, M. V., Yu, Q., et al. (2018). Intestinal microbiota enhances pancreatic carcinogenesis in preclinical models. *Carcinogenesis*, 39/8: 1068–78. DOI: 10.1093/carcin/bgy073

Thomas, R. M., & Jobin, C. (2020). Microbiota in pancreatic health and disease: the next

frontier in microbiome research. *Nature Reviews. Gastroenterology & Hepatology*, 17/1: 53–64. DOI: 10.1038/s41575-019-0242-7

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58/1: 267–88. DOI: 10.1111/j.2517-6161.1996.tb02080.x

Tjalsma, H., Boleij, A., Marchesi, J. R., & Dutilh, B. E. (2012). A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nature Reviews. Microbiology*, 10/8: 575–82. DOI: 10.1038/nrmicro2819

Torres, P. J., Fletcher, E. M., Gibbons, S. M., Bouvet, M., Doran, K. S., & Kelley, S. T. (2015). Characterization of the salivary microbiome in patients with pancreatic cancer. *PeerJ*, 3: e1373. DOI: 10.7717/peerj.1373

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., et al. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457/7228: 480–4. DOI: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, 449/7164: 804–10. DOI: 10.1038/nature06244

Vétizou, M., Pitt, J. M., Daillère, R., Lepage, P., Waldschmitt, N., Flament, C., Rusakiewicz, S., et al. (2015). Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science*, 350/6264: 1079–84. DOI: 10.1126/science.aad1329

Villéger, R., Lopès, A., Veziant, J., Gagnière, J., Barnich, N., Billard, E., Boucher, D., et al. (2018). Microbial markers in colorectal cancer detection and/or prognosis. *World Journal of Gastroenterology*, 24/22: 2327–47. DOI: 10.3748/wjg.v24.i22.2327

Vogtmann, E., Han, Y., Caporaso, J. G., Bokulich, N., Mohamadkhani, A., Moayyedkazemi, A., Hua, X., et al. (2020). Oral microbial community composition is associated with pancreatic cancer: A case-control study in Iran. *Cancer medicine*, 9/2: 797–806. DOI: 10.1002/cam4.2660

Wallace, B. D., Roberts, A. B., Pollet, R. M., Ingle, J. D., Biernat, K. A., Pellock, S. J., Venkatesh, M. K., et al. (2015). Structure and Inhibition of Microbiome β-Glucuronidases Essential to the Alleviation of Cancer Drug Toxicity. *Chemistry & Biology*, 22/9: 1238–49. DOI: 10.1016/j.chembiol.2015.08.005

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., et al. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, 6/2: 320–9. DOI: 10.1038/ismej.2011.109

Wellen, K. E., & Hotamisligil, G. S. (2005). Inflammation, stress, and diabetes. *The Journal of Clinical Investigation*, 115/5: 1111–9. DOI: 10.1172/JCI25102

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25/4: 679–89. DOI: 10.1038/s41591-019-0406-6

Wong, S. H., Kwong, T. N. Y., Chow, T.-C., Luk, A. K. C., Dai, R. Z. W., Nakatsu, G., Lam, T. Y. T., et al. (2017). Quantitation of faecal Fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. *Gut*, 66/8: 1441–8.

DOI: 10.1136/gutjnl-2016-312766

Wu, N., Yang, X., Zhang, R., Li, J., Xiao, X., Hu, Y., Chen, Y., et al. (2013). Dysbiosis signature of fecal microbiota in colorectal cancer patients. *Microbial Ecology*, 66/2: 462–70. DOI: 10.1007/s00248-013-0245-9

Wu, S., Morin, P. J., Maouyo, D., & Sears, C. L. (2003). Bacteroides fragilis enterotoxin induces c-Myc expression and cellular proliferation. *Gastroenterology*, 124/2: 392–400. DOI: 10.1053/gast.2003.50047

Xing, H., Wang, J., Wang, Y., Tong, M., Hu, H., Huang, C., & Li, D. (2018). Diagnostic Value of CA 19-9 and Carcinoembryonic Antigen for Pancreatic Cancer: A Meta-Analysis. *Gastroenterology research and practice*, 2018: 8704751. DOI: 10.1155/2018/8704751

Yan, X., Yang, M., Liu, J., Gao, R., Hu, J., Li, J., Zhang, L., et al. (2015). Discovery and validation of potential bacterial biomarkers for lung cancer. *American journal of cancer research*, 5/10: 3111–22.

Yoshimoto, S., Loo, T. M., Atarashi, K., Kanda, H., Sato, S., Oyadomari, S., Iwakura, Y., et al. (2013). Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature*, 499/7456: 97–101. DOI: 10.1038/nature12347

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q. Y., Qin, Y., Tang, L., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66/1: 70–8. DOI: 10.1136/gutjnl-2015-309800

Yu, L.-X., & Schwabe, R. F. (2017). The gut microbiome and liver cancer: mechanisms and clinical translation. *Nature Reviews. Gastroenterology & Hepatology*, 14/9: 527–39. DOI: 10.1038/nrgastro.2017.72

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10: 766. DOI: 10.15252/msb.20145645

Zhang, X., Zhu, X., Cao, Y., Fang, J.-Y., Hong, J., & Chen, H. (2019). Fecal Fusobacterium nucleatum for the diagnosis of colorectal tumor: A systematic review and meta-analysis. *Cancer medicine*, 8/2: 480–91. DOI: 10.1002/cam4.1850

Zitvogel, L., Daillère, R., Roberti, M. P., Routy, B., & Kroemer, G. (2017). Anticancer effects of the microbiome and its products. *Nature Reviews. Microbiology*, 15/8: 465–78. DOI: 10.1038/nrmicro.2017.44

Zitvogel, L., Galluzzi, L., Viaud, S., Vétizou, M., Daillère, R., Merad, M., & Kroemer, G. (2015). Cancer and the gut microbiota: an unexpected link. *Science Translational Medicine*, 7/271: 271ps1. DOI: 10.1126/scitranslmed.3010473

Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G., & Gajewski, T. F. (2018). The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science*, 359/6382: 1366–70. DOI: 10.1126/science.aar6918

Zou, S., Fang, L., & Lee, M.-H. (2018). Dysbiosis of gut microbiota in promoting the development of colorectal cancer. *Gastroenterology report*, 6/1: 1–12. DOI: 10.1093/gastro/gox031