U.PORTO

DOUTORAMENTO

BIOTECNOLOGIA MOLECULAR E CELULAR APLICADA ÀS CIÊNCIAS DA SAÚDE

Molecular-based approaches to characterize the microbiome: Exploring the gastric cancer scenario

Joana Pereira Marques

D
2019

U.PORTO

Joana Pereira Marques. Molecular-based approaches to characterize the microbiome: Exploring the gastric cancer scenario

Molecular-based approaches to characterize the microbiome: Exploring the gastric cancer scenario

Joana Pereira Marques

D.ICBAS **2019**

SEDE ADMINISTRATIVA

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR

FACULDADE DE FARMÁCIA

Joana Pereira Marques

Molecular-based approaches to characterize the microbiome: Exploring the gastric cancer scenario

Tese de Candidatura ao grau de Doutor em Biotecnologia Molecular e Celular Aplicada às Ciências da Saúde;
Programa Doutoral da Universidade do Porto (Instituto de Ciências Biomédicas Abel Salazar e Faculdade de Farmácia)

Orientador – Maria do Céu Fontes Herdeiro Figueiredo
Categoria – Professora Auxiliar
Afiliação – Faculdade de Medicina da Universidade do Porto (FMUP), Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup) e Instituto de Investigação e Inovação em Saúde, Universidade do Porto (i3S)

Co-orientador – Maria Raquel Campos Seruca
Categoria – Investigadora Principal e Professora Afiliada
Afiliação – Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup), Instituto de Investigação e Inovação em Saúde, Universidade do Porto (i3S) e Faculdade de Medicina da Universidade do Porto (FMUP)

Co-orientador – Leen-Jan van Doorn
Categoria – Chief Business Officer
Afiliação – DDL Diagnostic Laboratory, Rijswijk, The Netherlands

Ao avô Carlos…

The research presented in this PhD thesis was developed at:

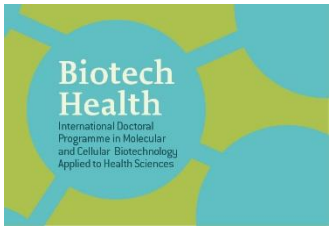
Epithelial Interactions and Cancer Group
i3S - Instituto de Investigação e Inovação em Saúde
Ipatimup - Instituto de Patologia e Imunologia Molecular
Universidade do Porto, Porto, Portugal

and

DDL Diagnostic Laboratory
Rijswijk, The Netherlands

FINANCIAL SUPPORT

This PhD thesis was financially supported through the PhD studentship from Fundação para a Ciência e a Tecnologia (FCT; PD/BD/114014/2015), and by the FCT PhD Programmes and by Programa Operacional Potencial Humano (POCH), specifically by the BiotechHealth Programme (Doctoral Programme on Cellular and Molecular Biotechnology Applied to Health Sciences - Reference PD/00016/2012). This research was financed in part by a Worldwide Cancer Research grant to CF and JCM (Reference 16-1352). i3S-Instituto de Investigação e Inovação em Saúde is funded by Fundo Europeu de Desenvolvimento Regional (FEDER) funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by Portuguese funds through Fundação para a Ciência e a Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Inovação (POCI-01-0145-FEDER-007274; POCI-01-0145-FEDER-032532).



Cofinanciado por:

LIST OF PUBLICATIONS

Ao abrigo do disposto do nº 2, alínea a) do artigo 31° do Decreto-Lei n.° 115/2013 de 7 de Agosto, fazem parte integrante desta tese de doutoramento os seguintes trabalhos já publicados ou em processo de revisão:

I. Ferreira RM, **Pereira-Marques J**, Pinto-Ribeiro I, Costa JL, Carneiro F, Machado JC, Figueiredo C. Gastric microbial community profilling reveals a dysbiotic cancer-associated microbiota. Gut. 2018; 67(2):226-236. [IF: 17.943]

II. **Pereira-Marques J**, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn LJ, Knetsch CS, Figueiredo C. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. Frontiers in Microbiology. 2019 [IF: 4.259*]

III. **Pereira-Marques J.***, Ferreira R.M.*, Pinto-Ribeiro I., Figueiredo C. *Helicobacter pylori* Infection, the Gastric Microbiome and Gastric Cancer. Advances in Experimental Medicine and Biology. Springer, New York, N. 2019 *Equal contribution. [IF: 2.126*]

IV. Molina-Castro S*, **Pereira-Marques J***, Figueiredo C, Machado JC, Varon C. Gastric cancer: Basic aspects. Helicobacter. 2017 Sep;22 Suppl 1. doi: 10.1111/hel.12412. *Equal contribution. [IF: 4.123]

ACKNOWLEDGMENTS

Crescimento. É a palavra que melhor define estes 4 anos de doutoramento. Foi de facto um período de grande crescimento pessoal e profissional. Uma nova cidade, novos colegas, novos mentores, novos amigos e novas experiências. Este percurso não teria sido possível sem a presença, disponibilidade, apoio e partilha de conhecimentos de várias pessoas, às quais quero deixar uma palavra de agradecimento, agora que esta fase está a terminar.

Em primeiro lugar gostaria de agradecer à minha orientadora, Céu Figueiredo, pela oportunidade que me deu de trabalhar no seu grupo, e de assim descobrir o fascinante mundo do microbioma gástrico. Muito obrigada por todo o apoio, incentivo, pelas discussões científicas e pela transmissão de conhecimentos ao longo destes 4 anos. É, para mim, um exemplo de persistência, dedicação e capacidade de trabalho.

À minha co-orientadora, Professora Raquel Seruca, agradeço a disponibilidade e a oportunidade dada em trabalhar no grupo Epithelial Interactions in Cancer (EPIC).

I would like to acknowledge Dr. Leen-Jan van Doorn, my co-supervisor, for accepting me at DDL Diagnostics Laboratory, in the Netherlands, to develop part of my PhD work. I am very thankful for the opportunity to work in an international environment, especially in a company. I have learned a lot during the 7 months I spent at DDL. Thank you for your availability, help, and discussions about the metagenomics project.

Ao Professor Mário Barbosa, Director do i3S, e Director do Programa Doutoral BiotechHealth, agradeço a oportunidade de fazer parte da segunda edição deste programa doutoral, que não só me permitiu levar a cabo este projecto de doutoramento, como contribuiu para o desenvolvimento de uma série de competências transversais que são uma mais valia no meu percurso profissional.

Ao Professor Sobrinho Simões pela dedicação demonstrada à instituição que é o Ipatimup.

Ao Instituto de Ciências Biomédicas Abel Salazar (ICBAS), ao Ipatimup e ao i3S por me terem recebido como aluna de doutoramento e à Fundação para a Ciência e Tecnologia pelo financiamento da minha bolsa de doutoramento.

Aos “Gastrobiotinhas”, Rui e Inês, o meu enorme e sentido obrigada. Não tenho palavras para descrever o que foram estes últimos anos que passávamos juntos no laboratório, de volta da “nossa” microbiota. Foi um longo percurso, com alguns altos e baixos, mas sempre repleto de novos e interessantes desafios. Foi convosco que tanto aprendi e partilhei o gosto por esta área.

Rui, muito obrigada por tanto me teres acompanhado e ajudado neste percurso. Por todo o apoio, disponibilidade e orientação ao longo destes anos. Obrigada também pela confiança que depositaste em mim, por todo o conhecimento partilhado, pelas ideias e discussões científicas, e por todas as tuas palavras de incentivo. Muito obrigada!
Inês, não me chegam as palavras para agradecer toda a amizade e companheirismo ao longo deste caminho. Obrigada por todo o teu apoio e disponibilidade desde que entrei no laboratório, e por toda a imensa partilha de ideias e experiências. Durante estes anos, fomos crescendo, lado a lado, não só cientificamente, mas também pessoalmente. Partilhámos muitos momentos, dentro e fora do laboratório. Recebeste-me sempre de braços abertos e estiveste sempre lá nas alturas em que mais precisei. Nunca esquecerei isso. Um gigante obrigada, vizinha de bancada e compincha das danças.

À Babi, a minha eterna companheira da biblioteca, agradeço o apoio incondicional. Já ouvia falar de ti e ainda não vivia no Porto. Obrigada pela tua amizade, por todo o teu carinho e compreensão, pelas imensas conversas e pela partilha de experiências. Estavas lá, quando sem muito conseguir dizer, mais precisei de um ombro amigo. Por tudo isto, um enorme obrigada.
P.S: Passados 4 anos, sou oficialmente uma Moura que adora o Porto.

À Rita, por todo o carinho, apoio e amizade. Muito obrigada pela alegria e boa energia que tanto transmites, e que tão bem te caracteriza. Por todas as boas gargalhadas que nos dias mais complicados fazem milagres!

Aos “Resistentes do Almoço”, Verónica, Ana, Joana, Rui e Marcelo obrigada pela vossa companhia, boa disposição e por todos os momentos divertidos durante aquela horinha do dia.

Agradeço também a todos os membros do grupo EPIC, em particular aos colegas do grupo *Helicobacter*, Marina, Miguel, Joana, Tânia, Vanessa e mais recentemente Cristina, pelo vosso apoio ao longo deste período.

Quero ainda agradecer ao Rob e à Mafalda, do serviço de genómica, por toda a simpatia, disponibilidade e ajuda.

Aos meus colegas da 2ª edição do BiotechHealth, obrigada pela partilha de momentos e pela entreajuda que existiu entre todos nas mais diversas etapas, ao longo destes anos. Agradeço também ao Doutor João Cortez, ao Daniel Vasconcelos e à Catarina Leite Pereira pela disponibilidade e ajuda nos vários assuntos relacionados com as unidades curriculares do BiotechHealth.

À Flavi, por estares nesta aventura comigo desde o primeiro dia. Muito obrigada por toda a tua amizade, apoio e carinho. Por todos os momentos que vivemos dentro e fora do

BiotechHealth e do Instituto. Pela partilha de tantas experiências, pelas longas conversas e pela entreajuda que sempre existiu. Por fazeres parecer que estávamos a viver na mesma cidade, quando ainda tinhamos alguns quilómetros a separar-nos. Pelas gargalhadas e momentos felizes, e por estares lá quando preciso. Que continuemos a partilhar muitos momentos destes daqui para a frente. Um gigante obrigada.

À Dani, por toda a amizade e carinho. Muito obrigada, por estares sempre disponível quando precisei. Por todo o teu apoio e ajuda ao longo destes anos. Pela partilha de experiências e pelos momentos divertidos que passámos. Por teres sempre uma palavra de encorajamento, especialmente nos últimos meses.

Ao Zé, agradeço o teu apoio, carinho e amizade. Muito obrigada, pelos “Bom dias” e pelos “Olás” no corredor que tanto faziam a diferença. Pelas conversas e pelos convívios. Pelas várias palavras de incentivo, apoio e força, especialmente nesta fase final.

I would also like to acknowledge everyone at DDL for welcoming me into such a nice work environment.
To Wilco, for all your support, guidance and helpful advices in the metagenomics project. Thank you, for all the scientific discussions but also for sharing your wisdom, and for allowing me to learn with your experience. Obrigada, for all the encouraging words!
To Anne, for all your help and partnership during my time at DDL. Thanks for our long discussions, for the opportunity to learn so much from your bioinformatics expertise. It was a really big challenge, but I have really enjoyed it. Obrigada, for trying to teach me a bit of Dutch every day and for our conversations during break time.
To Kim, for your precious time, availability and for helping with the lab stuff. Thanks for all your kindness and support. To Michiel, for all the support, ideas and helpful advices in metagenomics project, when your time was so valuable. To Jurgen, for your availability and for teaching me all the details about Ilumina library preparation and sequencing. Thanks for all the suggestions and for sharing your experience. To Eva, for sharing your experience and for your critical suggestions about the metagenomics project. To Ellen, for your support, kindness and for the discussions at the Microbiome group office. To Jaroslav for your kindness and for your funny mood at the office. To Dorota, for your kindness and for all the good and funny moments at DDL’s *borrel* and DDL’s games night. To Souvik, for your gentleness and for always have a funny joke to tell during break time. Thanks for you precious lessons about how to use R to build heatmaps, and of course, for being “my partner in crime” (security code man) in staying until late at DDL. To Annemiek, for all your help, support and friendship during my stay at DDL and in the Netherlands. Thanks for showing me Rotterdam and for the time we spent together.

To Frank and Martijn that so nicely opened the doors of their home for me, when I arrived in the Netherlands. Thank you for all your support, help and kindness. I was really lucky to have met you guys.

To Jolanda and Kees, for all your help and support when I was moving to the Netherlands. I am really thankful that we met back then and for the time we spent together.

A big thanks to all the friends I made at “Campo Alegre III” during the last 4 years. A special thanks to Atefeh, Paola, Vicky, Ruth, Marina, Rafa and Fernanda. For our dinners and lunches in the kitchen, for all the trips and good moments we spent together. We were like a family back then. Thank you so much for your friendship and support.

Por fim, aos amigos de sempre, que mesmo à distância, me acompanharam e apoiaram durante este percurso. Obrigada pela amizade incondicional e pelas constantes palavras de apoio, força e encorajamento. Cláudia, Catarina, Francisco, Buga, Joaninha, Mafalda, Sofia, Carlota, Rafa e David, obrigada por fazerem parte da minha vida.

Finalmente, um enorme obrigada ao meus pais e avós, por acreditarem sempre em mim e por apoiarem incondicionalmente todas as minhas decisões, mesmo que signifique ficarmos longe uns dos outros.

Por fim, um agradecimento muito especial à minha irmã. As palavras não chegam para descrever o que representas para mim e o quão bom é ter a tua amizade. Por toda a força e apoio incondicional ao longo desta aventura que foi o doutoramento. Muito obrigada, por seres o meu suporte, agora e sempre. Sou grata por te ter ao meu lado.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ACE	Abundance-based Coverage Estimator
ANOSIM	Analysis of similarity
AUC	Area under the curve
BLAST	Basic local alignment search tool
BMTagger	Best match tagger
bp	Base pair
cagA	cytotoxin-associated gene A
cDNA	Complementary DNA
COG	Cluster of Orthologous Groups
Ct	Threshold cycle
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxyribonucleic acid
emPCR	Emulsion polymerase chain reaction
EPIYA	Glutamic Acid-Proline-Isoleucine-Tyrosine-Alanine
FISH	Fluorescence *in situ* hybridization
Gb	Giga base pairs
GC	Guanine and Cistine
GI	Gastrointestinal
GIN	Gastrointestinal intraepithelial neoplasia
HMP	Human Microbiome Project
H. pylori	*Helicobacter pylori*
HUMAnN2	HMP Unified Metabolic Analysis Network
IL-1	Interleukin 1
IL-10	Interleukin 10
IL-1β	Interleukin 1 beta
INS-GAS	Insulin-gastrin
ISP	Ion sphere particles
ITS	Internal transcribed spacer
KEGG	Kyoto encyclopedia of genes and genomes
KO	Kyoto encyclopedia of genes and genomes orthology
LDA	Linear discriminant analysis
LefSe	Linear discriminant analysis effect size
MALDI-TOF	Matrix-Assisted Laser Desorption/ Ionization-Time of Flight
MetaHIT	Metagenomics of the Human Intestinal Tract

mRNA	Messenger RNA
MaAsLin	Multivariate association with linear models
MDI	Microbial dysbiosis index
MetaPhlAn2	Metagenomic phylogenetic analysis 2
NGS	Next-generation sequencing
NHNN	Non-*H. pylori* and non-NSAID
NIH	National Institute of Health
Non-NSAID	Non-steroidal anti-inflammatory drug
OTU	Operational taxonomic unit
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PCoA	Principle coordinate analysis
PERMANOVA	Permutational multivariate analysis of variance
PICRUSt	Phylogenetic investigation of communities by reconstruction of unobserved states
PPI	Proton pump inhibitor
QIIME	Quantitative insights into microbial ecology
qPCR	Quantitative polymerase chain reaction
RDP	Ribosomal database project
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
tRNA	Transfer RNA
ROC	Receiver-operating characteristic
T-RFLP	Terminal restriction fragment length polymorphism
TGGE	Temperature gradient gel electrophoresis
TNF-α	Tumor necrosis factor α
vacA	Vacuolating cytotoxin gene A
WMS	Whole metagenome sequencing

INTRODUCTION

1. The human microbiome

The human body is inhabited in its different niches by a vast collection of microbes, generally known as the microbiota. These microorganisms, their genetic information, as well as the information of the niche in which they interact, are usually referred to as the microbiome (Cho and Blaser 2012). Currently, the term microbiome is also used to refer to the microorganisms themselves, i.e. the microbiota (Knight *et al.* 2017). The number of microbial cells was commonly thought to outnumber the quantity of human cells by a 10-fold ratio, but recent assessments propose a 1:1 ratio as a better estimate (Sender, Fuchs, and Milo 2016). Bacteria constitute so far the best explored component of the microbiome. Progress in this research area had been hampered by the fact that only a very small fraction of the microbial species can be cultured *in vitro*. The advent of high-throughput sequencing technologies, together with the emergence of large international and interdisciplinary projects, have strongly contributed to expand our understanding of the microbiome structure and functions (Qin *et al.* 2010; Turnbaugh *et al.* 2007; Arnold, Roach, and Azcarate-Peril 2016).

1.1. The impact of the microbiome on human health and disease

It is currently accepted that the microbiome plays a major role in the maintenance of the normal physiology and health of the host, being involved in a wide variety of metabolic functions, like energy production and storage, fermentation and absorption of undigested carbohydrates, and participating in the normal maturation of the immune system (Gilbert *et al.* 2018; Lloyd-Price *et al.* 2017). Furthermore, the microbiome can also contribute for normal brain function, specifically in the regulation of digestive function and satiety, but also behaviour (Diaz Heijtz *et al.* 2011).

Disruption of the balance that exists between the microbiome and the host, called dysbiosis, may promote numerous diseases, including inflammatory bowel disease (Frank *et al.* 2007), obesity (Ley *et al.* 2006) and cancer (Kostic *et al.* 2013; Gilbert *et al.* 2018). For example, members of the gut microbiome such as *Fusobacterium nucleatum*, *Escherichia coli*, and *Bacteroides fragilis*, have been found enriched in colorectal cancer (Ahn *et al.* 2013; Bonnet *et al.* 2014; Goodwin *et al.* 2011). Nevertheless, and although the exact mechanisms linking microbial dysbiosis and cancer are still largely unknown, it can be anticipated that bacterial metabolites and toxins, as well as inflammation triggered by the microbiome contribute to the promotion of cancer.

1.2. Major human microbiome initiatives

The Human Microbiome Project (HMP), a major international initiative funded by the National Institute of Health (NIH), remains the largest high-scale effort to comprehensively characterize the human microbiome in order to understand its impact on health. The first phase of this project was focused on the description of microbial communities from healthy individuals across several body sites, including skin, nasal, oral, gastrointestinal, and urogenital areas, combining both amplicon and whole metagenome sequencing (Human Microbiome Project 2012a, 2012b). These studies have demonstrated that different body sites comprise significantly different microbial communities (Human Microbiome Project 2012a). Moreover, the composition of the normal microbiome varies between individuals and is influenced by local conditions inherent to the anatomic site, host genetics, diet, and antibiotic consumption (Gilbert *et al.* 2018; Lloyd-Price *et al.* 2017). Another similar initiative is the Metagenomics of the Human Intestinal Tract (MetaHIT), which aimed to characterize the human intestinal microbiota, in health and disease (Qin *et al.* 2010).

2. Approaches to study the microbiome

In order to better understand the human microbiome in terms of its composition and function, a variety of approaches can be employed **(Figure 1)**.

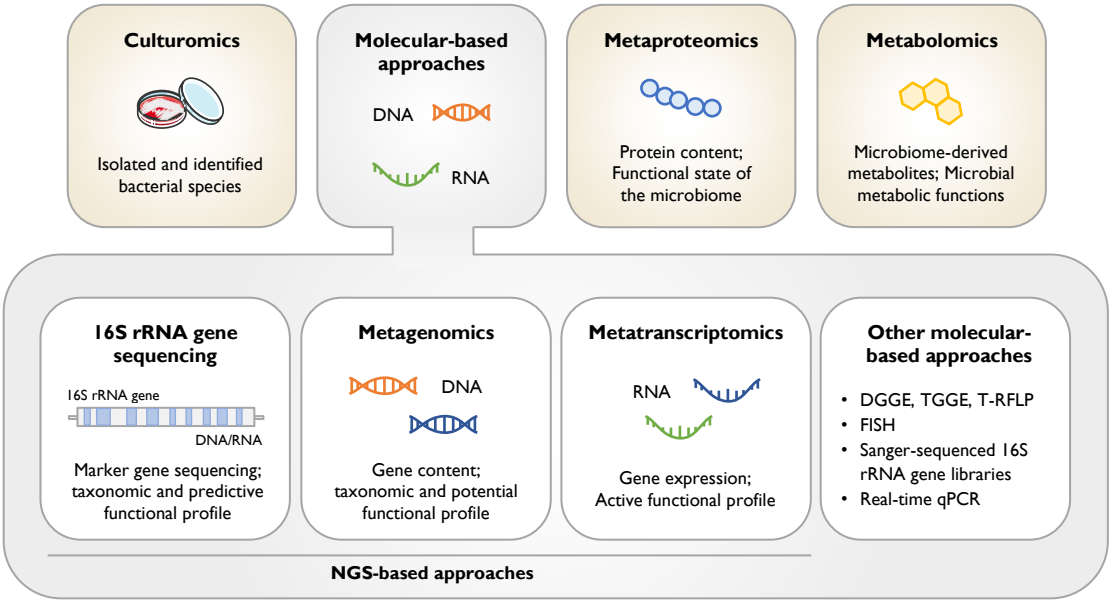


Figure 1. Overview of the techniques used to study the composition and/or function of the microbiome. NGS, Next-generation sequencing; DGGE or TGGE, Denaturing or temperature gradient gel electrophoresis; T-RFLP, Terminal restriction fragment length polymorphism; FISH, Fluorescence *in situ* hybridization; qPCR, quantitative polymerase chain reaction.

2.1. Culturomics

In the earlier microbiome studies, traditional culture-based methods were the gold-standard to analyze the composition of microbial communities. These techniques utilized differential selective media for isolation of specific bacterial species, followed by characterization of the cultured bacteria based on morphology and biochemical features (Moore and Holdeman 1974; Morgan and Huttenhower 2012; Savage 1977). Despite being a cost-effective approach that enabled selective detection of viable bacteria, conventional culture-based methods were laborious, time-consuming, and provided an incomplete view of the diversity of microbiota, since less than 30% of the gut microbiota have been cultured to date (Eckburg *et al.* 2005; Fraher, O'Toole, and Quigley 2012; Lagier *et al.* 2012; Seng *et al.* 2009).

Since the massive technological developments in molecular tools, culture techniques have been gradually replaced in microbiome studies (Human Microbiome Project 2012b; Lloyd-Price *et al.* 2017). However, more recently a renewed interest in culture methods has been demonstrated by researchers due to considerable advances, which enable the growth of previously uncultured bacteria (Ingham *et al.* 2007; Kaeberlein, Lewis, and Epstein 2002; Nichols *et al.* 2010). In this context, culturomics was developed to discover unknown bacteria inhabiting the human gut (Lagier *et al.* 2016).

Culturomics consists in a high-throughput culturing approach that uses several culture conditions in combination with matrix-assisted laser desorption/ ionization–time of flight (MALDI-TOF) mass spectrometry and 16S ribosomal RNA (rRNA) gene sequencing, to rapidly identify bacterial species (Lagier *et al.* 2018; Lagier *et al.* 2015). Even though culturomics is unable to directly give information about the function of the novel bacterial species isolated (Lagier *et al.* 2018), the application of this approach in human microbiome studies has considerably extended the collection of newly identified and isolated bacterial species associated with the human gut (Lagier *et al.* 2015).

2.2. Molecular-based approaches

To overcome the limitations of conventional culture-dependent methods, several molecular-based approaches were implemented in microbiome studies. Early molecular-based techniques typically focused on the analysis of the 16S rRNA gene, a standard phylogenetic marker (Pace 1997). These community profiling methods included DNA fingerprinting techniques, such as denaturing or temperature gradient gel electrophoresis (DGGE or TGGE) and terminal restriction fragment length polymorphism (T-RFLP), fluorescence *in situ* hybridization (FISH), DNA microarrays, real-time quantitative polymerase chain reaction (qPCR), and Sanger-sequencing of 16S rRNA gene libraries (Harmsen *et al.* 2000; Langendijk *et al.* 1995; Nadkarni *et al.* 2002; Osborn, Moore, and Timmis 2000; Paliy *et al.*

2009; Satokari *et al.* 2001; Zoetendal, Akkermans, and De Vos 1998; Suau *et al.* 1999). The majority of these approaches involve DNA extraction, followed by amplification of the 16S rRNA gene by polymerase chain reaction (PCR), except for FISH that can be performed directly in the sample, and DNA microarrays that can be either performed using 16S rRNA amplicons or total genomic DNA (Fraher, O'Toole, and Quigley 2012). Currently, most of these conventional methods have been replaced by next-generation sequencing (NGS) approaches, which enabled to move from collecting a few dozen sequences for each sample to start collecting a few hundred million. Therefore, the emergence of high-throughput DNA sequencing technologies has aided a deep profiling of the microbiome in terms of composition and function (Human Microbiome Project 2012a).

2.2.1. Real-time qPCR

Among the different molecular-based approaches available to examine the microbiome composition, qPCR enables identification and quantification of specific microbial taxa in complex microbial communities (Bartosch *et al.* 2004; Sokol *et al.* 2009). Typically, it involves DNA amplification using primers targeting a region of the 16S rRNA gene, and DNA quantification, which is obtained after each PCR cycle, using fluorescent DNA-binding dyes or probes that produce a fluorescent signal proportional to the amount of amplified product. Therefore, the amount of DNA present in a given sample can be measured using the standard curve method by plotting fluorescence against the number of PCR cycles using a logarithmic scale (Carey *et al.* 2007; Fraher, O'Toole, and Quigley 2012). Primers for qPCR can be designed either to target all bacterial taxa, allowing the quantification of the total bacterial load (Nadkarni *et al.* 2002) or to target specific bacterial taxa (Malinen *et al.* 2003; Mariat *et al.* 2009).

qPCR is a relatively fast, accurate, and sensitive molecular based-approach. However, this technique is unable to identify novel groups of bacteria since primers are designed to target a specific bacterial taxonomic group. In addition, it can be a laborious and technically difficult approach, namely during primer design (Fraher, O'Toole, and Quigley 2012; Sekirov *et al.* 2010). Several studies have shown alterations in the gut microbiome associated with disease, antibiotic treatment, and age by using qPCR alone (Mariat *et al.* 2009; Bartosch *et al.* 2004; Sokol *et al.* 2009). However, this molecular tool can also be utilized in combination with other approaches. For example, when combined with DGGE or DNA microarrays, which are semi-quantitative methods, qPCR offers more detailed information on the abundance of the members of the microbiota (Rinttila *et al.* 2004; Jalanka-Tuovinen *et al.* 2011). Moreover, in microbiome studies based on NGS approaches, qPCR has already

been used as a tool to validate the observed sequencing results (Kostic *et al.* 2012; Russo *et al.* 2017).

2.2.2. 16S rRNA gene sequencing

16S rRNA gene sequencing is one of the most commonly used NGS-based approaches to study the microbiome (Human Microbiome Jumpstart Reference Strains *et al.* 2010; Human Microbiome Project 2012a). This method uses PCR amplification and subsequent sequencing of a specific region of the 16S rRNA gene, in order to determine the taxonomic composition of the microbial communities within a given sample (Knight *et al.* 2018; Caporaso *et al.* 2011).

The 16S rRNA gene is the most typically used genetic marker in prokaryotes, because is phylogenetically informative and ubiquitous among *Bacteria* and *Archaea* (Pace 1997). This marker gene, which has a length of around 1500 bp, comprises nine hypervariable regions, each flanked by ten highly conserved regions suitable for PCR primer binding. The nine hypervariable regions have been used for taxonomic assignment, since they show sufficient sequence variability between different genera **(Figure 2)** (Baker, Smith, and Cowan 2003; Neefs *et al.* 1993; Knight *et al.* 2018).

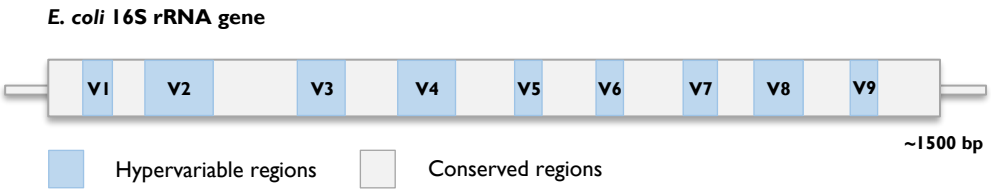


Figure 2. Schematic representation of the *E. coli* 16S rRNA gene, showing the nine hypervariable regions (V1-V9) (in blue) flanked by the ten conserved regions (in grey).

Usually one or multiple adjacent hypervariable regions of the 16S rRNA gene can be targeted, according to the sequencing technology and chemistry. However, there are no standards for the selection of variable region, which can lead to considerable misinterpretations of data and inconsistencies between microbiome studies (Claesson, Clooney, and O'Toole 2017; Claesson *et al.* 2010; Clooney *et al.* 2016). The choice of variable region to sequence is often based on its capacity to profile the highest number of taxa possible, based on the level of conservation of the adjacent nucleotide sequences and on the read length supported by the sequencing platform (Sarangi, Goel, and Aggarwal 2019). Besides selecting the region to be sequenced, it is also important to choose the PCR

primers taking into consideration their specificity (Allaband *et al.* 2019). Primers previously used in the literature, in studies with which the results will be compared can be selected. In case there are no published studies yet, selecting common primer sets, such as those designed for the HMP spanning the V1-V3 regions or V3-V5 regions (Human Microbiome Project 2012a), can also be a good alternative.

16S rRNA gene sequencing is a fast and cost-effective method to determine the composition of microbial communities (Knight *et al.* 2018; Thompson *et al.* 2017). It is suitable to an extensive variety of sample types and study designs, including samples that have high fraction of host DNA, such as tissue and low-biomass samples (Knight *et al.* 2018). Even though 16S rRNA gene sequencing is the most widely used NGS-based approach to study the microbiome, its application has several drawbacks. Normally this approach provides genus level taxonomic resolution for most bacterial taxa, but has a lower-resolution at the species level, unless a full-gene length approach is used (Petrosino *et al.* 2009). It is subject to PCR primer and amplification biases, since primers do not have the same exact affinity for all potential DNA sequences, binding differently to some taxa over others (Walker *et al.* 2015; Soergel *et al.* 2012). The selection of variable region (Claesson *et al.* 2010), the number of PCR cycles (Bonnet *et al.* 2002), and variations in the 16S rRNA gene copy number across bacterial species (Acinas *et al.* 2004) represent other sources of bias. Moreover, although 16S rRNA gene sequencing gives insights into the composition of the microbiome, information about its function cannot be directly obtained by this approach. The putative biological functions of the microbiota are inferred by associating the marker gene data with the microbial genome sequences using methods like phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) (Langille *et al.* 2013). Lastly, 16S rRNA gene sequencing is only applied to taxa from the *Bacteria* and *Archaea* domains. Therefore, for fungi and other single-cell eukaryotes the internal transcribed spacer (ITS) region and 18S ribosomal RNA gene are used as marker genes, respectively (Hoffmann *et al.* 2013; Nilsson *et al.* 2008).

16S rRNA gene sequencing data analysis

The 16S rRNA gene sequencing data analysis is a multistep process **(Figure 3)** that can be performed using several bioinformatic pipelines, including Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso *et al.* 2010), Mothur (Schloss *et al.* 2009), and UPARSE (Edgar 2013).

The first step consists in pre-processing the raw sequencing reads to reduce artifacts created during PCR amplification and sequencing. Therefore, high-quality data will be generated for the downstream processes (Jumpstart Consortium Human Microbiome

Project Data Generation Working 2012). After the overall quality of the data is assessed using software tools like FastQC (Andrews 2016), sequences from sequencing adapters and PCR primers are removed from the reads. In case paired-end sequencing has been performed, forward and reverse paired reads should be merged into one read, in order to increase the overall data quality and reads length. Then, length trimming and quality filtering are performed to remove low-quality bases from the datasets (Human Microbiome Project 2012b). Another important step in 16S rRNA gene sequencing data analysis is the detection and removal of chimeras, which are sequences created during PCR amplification from two or more different biological sequences, often due to incomplete template extension (Haas *et al.* 2011). These hybrid sequences when undetected may be incorrectly considered as new amplicons, and thus be misannotated as novel species (Edgar *et al.* 2011). Several computational tools have been utilized to detect and remove chimeric sequences, including ChimeraSlayer (Haas *et al.* 2011) and UCHIME2 (Edgar 2016). However, these methods might not be completely effective, since chimeras are usually difficult to distinguish from the parent sequences (Edgar *et al.* 2011). Once raw sequencing data is pre-processed, quality-filtered reads are clustered into operational taxonomic units (OTUs) based on DNA sequence similarity. Even though, other thresholds can be applied, the similarity threshold commonly used is 97%, due to an approximation of species-level resolution (Claesson, Clooney, and O'Toole 2017). There are different OTU clustering strategies, including *de novo*, closed-reference, and open-reference clustering. The most commonly applied method is *de novo* clustering, which groups reads by similarity without using a reference database (Westcott and Schloss 2015). When a reference database is available, the closed-reference approach can be used, clustering reads into OTUs according to their similarity to known sequences in the database. However, reads that do not align to the database are discarded. Open-reference clustering is a combination of both approaches, using first the closed-reference clustering and then, the de *novo* clustering of reads that do not match the reference database (Claesson, Clooney, and O'Toole 2017; Kopylova *et al.* 2016).

The taxonomic assignment of OTUs is then performed after clustering. A consensus sequence from each OTU is aligned to a reference ribosomal sequence database and then, the taxonomy of the closest match found in the database is assigned to the OTU. This step is carried out using a number of possible algorithms such as BLAST (Altschul *et al.* 1990), the Ribosomal Database Project (RDP) classifier (Cole *et al.* 2009) and UCLUST (Edgar 2010). Additionally, some commonly used reference databases are Greengenes (DeSantis *et al.* 2006), RDP (Maidak *et al.* 2001), and SILVA (Quast *et al.* 2013). Also, upon OTU clustering, computational methods like PICRUSt can predict the functional content of the metagenome using 16S rRNA sequencing data and a database of reference genomes (Langille *et al.* 2013).

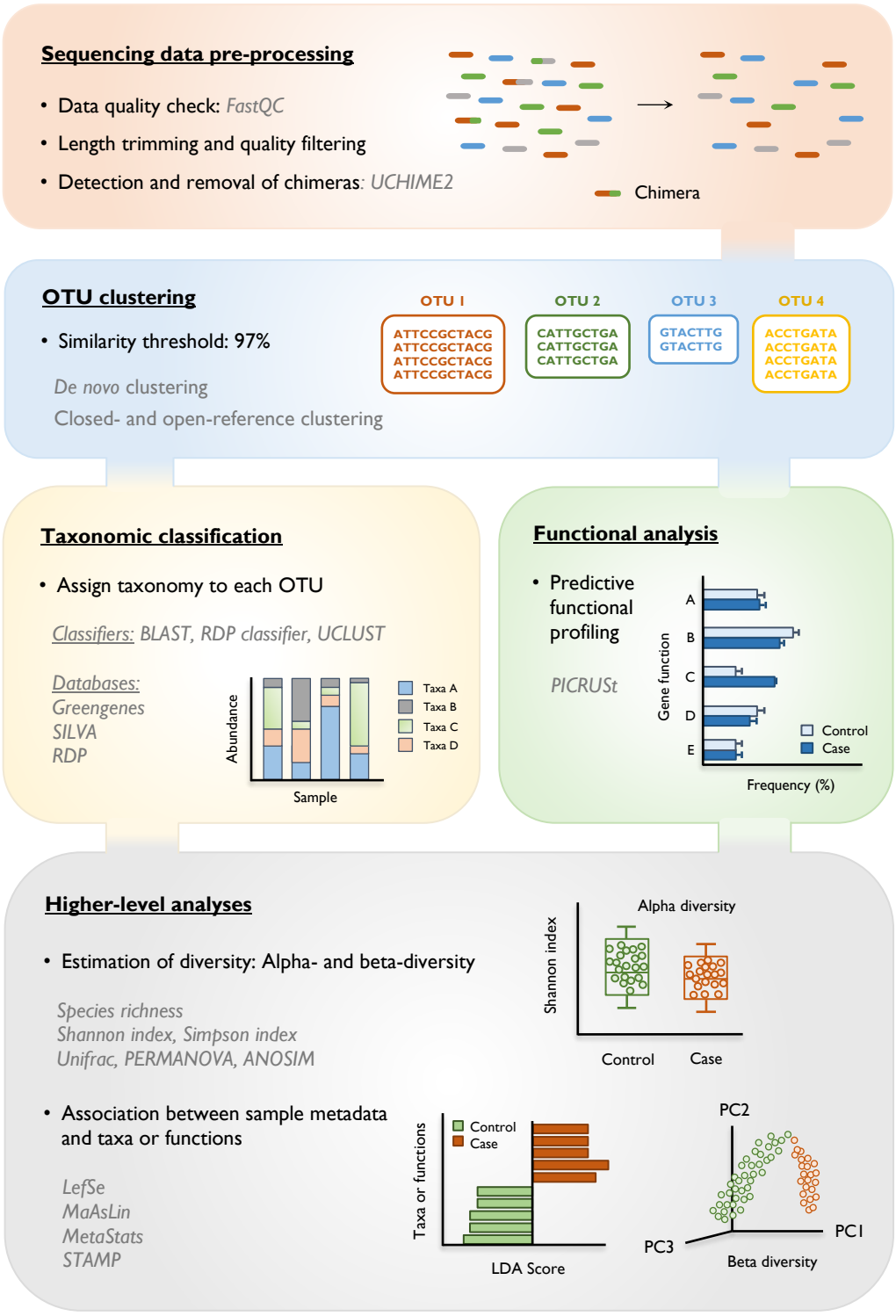


Figure 3. Workflow of 16S rRNA gene sequencing data analysis. During a pre-processing step, low-quality reads are filtered and chimeras removed. Then, highly similar sequences are clustered into OTUs, and taxonomy is assigned for each OTU using classifiers like RDP classifier and UCLUST, considering a minimum percentage of similarity to reference sequence databases, such as Greengenes, SILVA or RDP. Computational methods like PICRUSt can predict the gene function composition of the microbiome using 16S rRNA sequencing data and a database of reference genomes. The resulting matrix relating taxa or functions abundances to samples is used for further downstream analysis and data interpretation. Higher-level analyses focus on estimating diversity and on determining which taxa or functions significantly differ between study groups.

Predicted functional genes can be classified into clusters of orthologous groups (COG) that contains prokaryotic proteins of complete genomes or into Kyoto encyclopedia of genes and genomes orthology (KO) (Tatusov *et al.* 2000; Kanehisa and Goto 2000).

Finally, from the taxonomic classification and predictive functional profiling result matrices relating the relative abundance of taxa and gene functions to samples, which are further analyzed. These downstream analyses include estimation of taxa diversity and identification of differentially abundant taxa or functions between groups of interest (Knight *et al.* 2018). Classically, microbiome analyses have focused in assessing diversity using alpha- and beta-diversity measurements. Alpha-diversity quantifies the diversity within a sample. The simplest measure is species richness, which is the number of taxa identified (Chao 1 and Abundance-based Coverage Estimator (ACE)) (Chao 1984; Chao and Lee 1992). However, other measures also account for the taxa abundances (evenness) in a specific specimen (Shannon Index and Simpson Index) (Knight *et al.* 2018). Beta-diversity measures the diversity between samples by generating a distance matrix for all pairs of samples, using either phylogeny-dependent or -independent metrics. Unlike phylogeny-independent methods (Bray-Curtis distance), phylogeny-dependent methods take into consideration phylogenetic information to compare samples, the most common being UniFrac metrics (Knight *et al.* 2018; Lozupone, Hamady, and Knight 2006). While the unweighted UniFrac distance considers only the presence or absence of taxa across samples, the weighted UniFrac also accounts for relative abundance of taxa between samples (Lozupone *et al.* 2011). Beta-diversity data can be visualized by principal coordinate analysis (PCoA), a method that reduces distance matrices dimensionality into 2D or 3D representations. To assess significant sample clustering in beta-diversity analysis, the non-parametric permutation tests permutational multivariate analysis of variance (PERMANOVA) and analysis of similarity (ANOSIM) are used (Anderson and Walsh 2013).

Another commonly analysis applied to 16S rRNA gene sequencing data is the evaluation of relationships between metadata and variations in taxa or functions abundances, in order to identify microbial taxa or functions that explain differences between study groups. Tools such as linear discriminant analysis effect size (LefSe) (Segata *et al.* 2011), multivariate association with linear models (MaAsLin) (Morgan *et al.* 2012), MetaStats (White, Nagarajan, and Pop 2009), and STAMP (Parks and Beiko 2010) have been designed to carry out this type of analyses.

2.2.3. Metagenomics or whole metagenome sequencing

Whole metagenome sequencing (WMS) or metagenomics is another NGS-based approach currently performed to investigate the microbiome, specifically to characterize the

metagenome (Human Microbiome Jumpstart Reference Strains *et al.* 2010; Human Microbiome Project 2012a). WMS consists on untargeted DNA sequencing of fragments of all genomes within a sample, which produces high-complexity datasets with millions of short reads (Quince *et al.* 2017). Since all DNA present in the sample is captured, including bacterial, viral, and eukaryotic DNA, this method allows extensive characterization of the microbial communities living in a wide range of environments (e.g. soil, human-associated samples, among others), giving insights into both microbiome structure and potential function (Qin *et al.* 2010; Venter *et al.* 2004).

In comparison with 16S rRNA gene sequencing, WMS typically yields a more detailed taxonomic resolution, at the species or even strain-level (Truong *et al.* 2015; Truong *et al.* 2017), and allows assembly of whole microbial genomes (Mukherjee *et al.* 2017). It also enables the characterization of non-bacterial members of the microbiota, such as fungi and viruses, that are essentially lost with 16S rRNA gene sequencing (Knight *et al.* 2018). Moreover, this approach also reveals the functional potential of the microbiome by directly estimating the relative abundance of microbial functional gene families (Abubucker *et al.* 2012).

Still, this approach has been less implemented since it is more expensive than 16S rRNA gene profiling, it requires a greater depth of coverage, and the data analysis is more complex (Knight *et al.* 2012). Additionally, WMS is limited by the availability of annotated genomic sequences, since it relies on reference sequence databases that do not have complete genomes for many members of the microbiome (Quince *et al.* 2017).

A major technical challenge in whole metagenome analysis of human samples is the predominance of host DNA. Data from the HMP has revealed that the proportion of human DNA differs significantly by body site and sample type (Lloyd-Price *et al.* 2017; Human Microbiome Project 2012a). While stool samples comprise less than 10% of human DNA, samples such as saliva, throat, buccal mucosa, and vaginal swabs contain more than 90% of human-aligned reads (Human Microbiome Project 2012a; Lloyd-Price *et al.* 2017). The complexity of the latter type of samples, where only a limited fraction of the DNA represents the microbial content, requires a high quantity of sequences to obtain a reasonable coverage of the microbial genomes when using WMS. The influence of this limitation on the sensitivity of WMS to characterize the microbiome of host-derived complex samples remains to be explored. In fact, there are no recommendations on the proportion of host DNA a sample should have for a precise WMS analysis. Overcoming these issues is essential for future selection of appropriate sequencing depths that will guarantee the return of the maximum useful information, with a minimum cost possible.

WMS data analysis

As with 16S rRNA gene profiling analysis, a typical WMS bioinformatic pipeline of analysis starts with pre-processing of raw sequencing data **(Figure 4)**. First, the overall quality of the data is assessed (Andrews 2016). Then, quality-filtering that includes trimming low-quality bases and filtering reads based on length, is performed with tools like Trimmomatic (Bolger, Lohse, and Usadel 2014). Finally, host sequences decontamination is also done with tools like Best match tagger (BMTagger) (Rotmistrovsky and Agarwala 2018) and Bowtie2 (Langmead and Salzberg 2012), as host reads can lead to an incorrect taxonomy assignment, interfering with the downstream analysis of the microbiome data (Rotmistrovsky and Agarwala 2018).

After sequencing data pre-processing, a read-based or assembly-based approach can be used (Knight *et al.* 2018). Read-based profiling consists in assigning taxonomy without performing assembly, by comparing unassembled reads against publicly available reference databases. This approach identifies microbial species in a metagenome, and determines their relative abundances, generating a taxonomic profile. Read-based profiling improves the speed of data analysis, and can be performed efficiently with large and complex datasets. However, since it relies on external sequence data resources, it shows difficulties in profiling uncharacterized microorganisms (Knight *et al.* 2018; Quince *et al.* 2017). There are several read-based tools that can be utilized, including those that assign taxonomy to short DNA sequences of length k (the k-mers), such as Kraken (Wood and Salzberg 2014), but also marker gene methods that select specific genomic regions from reference genomes to perform taxonomy assignment (Metagenomic phylogenetic analysis 2 (MetaPhlAn2)) (Truong *et al.* 2015). The latter is a fast and precise assembly-free approach that can be used for large metagenomic datasets, in which performing assembly is unfeasible. In fact, MetaPhlAn2 was effective to profile the microbiome from multiple body sites for the HMP (Lloyd-Price *et al.* 2017).

Alternatively, WMS data can be analyzed by performing assembly of short reads into longer continuous sequences, the contigs. Therefore, this approach allows the construction of partial to full microbial genomes, possibly covering complete protein coding genes. However, in addition to be a computationally challenging and costly approach, there may not be enough coverage to assemble whole genomes (Knight *et al.* 2018; Quince *et al.* 2017). Metagenome assembly can be performed using either reference-based or *de novo*-based methods (Namiki *et al.* 2012; Luo *et al.* 2012). Reference-based assembly maps reads against available reference genomes in order to generate contigs. In contrast, metagenome *de novo* assembly, instead of relying on reference sequence resources to

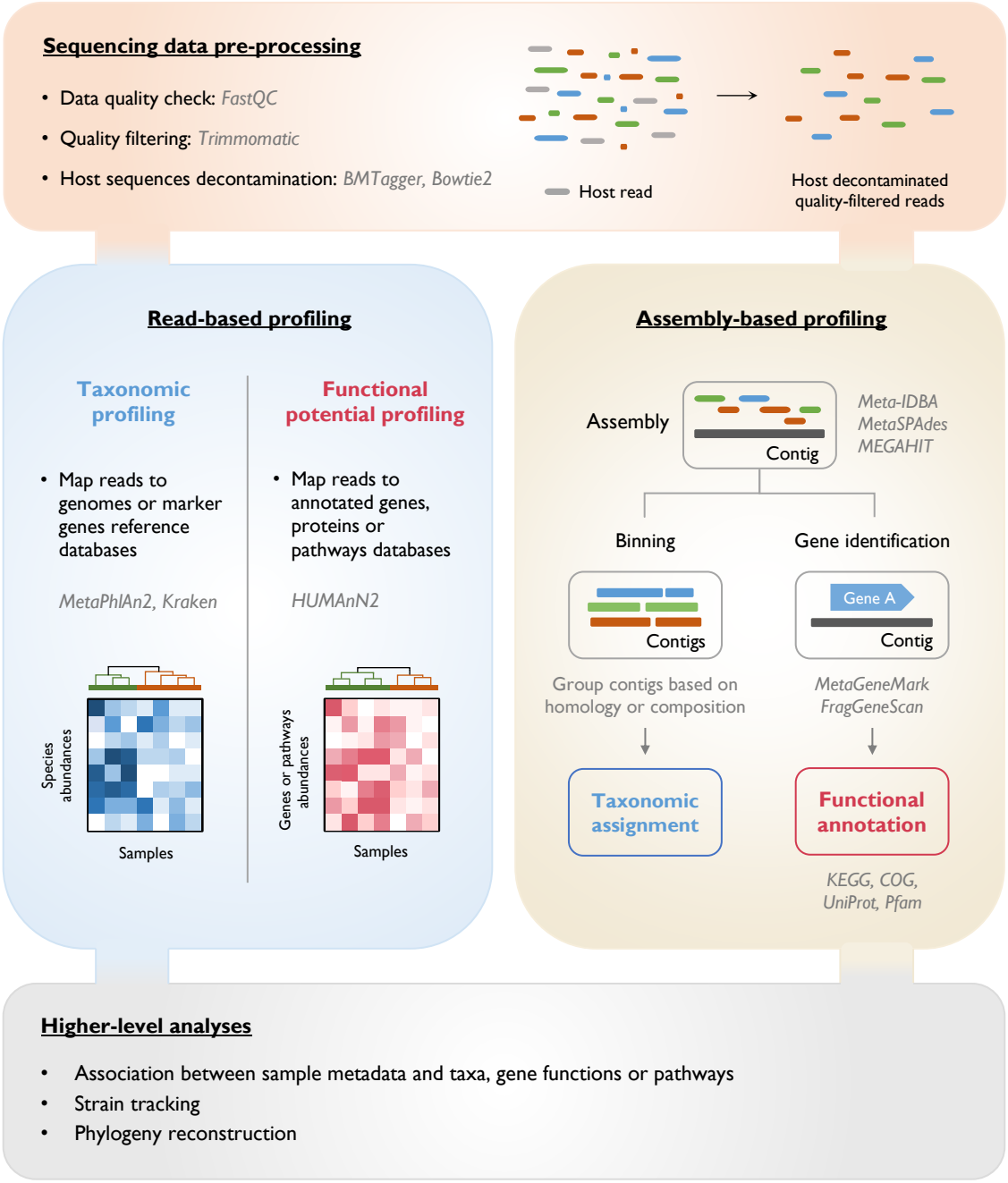


Figure 4. **Workflow of WMS data analysis**. During a pre-processing step, reads are trimmed based on quality and discarded based on length, and the host sequences are removed. The pre-processed data can be analyzed either using a read-based or assembly-based approach. In read-based profiling, unassembled reads are directly mapped to genomes or marker genes databases, and annotated genes, proteins or pathways databases, generating taxonomic and functional profiles, respectively. In assembly-based profiling, reads are assembled into contigs. Taxonomic assignment can be performed after binning of contigs. Identification of genes is carried out with tools, such as MetaGeneMark and FragGeneScan, followed by functional annotation using homology-based searches against databases like KEGG or COG. Processed microbiome data are subjected to higher-level analyses to disclose associations between sample metadata and taxa, gene functions or pathways.

assemble reads, applies graph theory algorithms, such as the de Bruijn graph (Pevzner, Tang, and Waterman 2001). Many metagenome-specific assemblers can be used, including MetaVelvet (Namiki *et al.* 2012), Meta-IDBA (Peng *et al.* 2011), MetaSPAdes (Bankevich *et al.* 2012), and MEGAHIT (Li *et al.* 2015). Upon metagenome assembly, the millions of assembled contigs generated are grouped according to their likely genomes of origin. This process is known as binning, and can either be performed based on similar sequence compositions or on sequence homology to known genes. This step is followed by taxonomy assignment (Claesson, Clooney, and O'Toole 2017; Quince *et al.* 2017).

Besides taxonomy assignment, the functional potential of the microbiome can be assessed by whole metagenome analysis, and both assembly-based and assembly-free approaches can be implemented. When functional analysis is carried out using assembly-free methods, reads are directly annotated without prior assembly through computational tools like the HUMAnN2 (Franzosa *et al.* 2018). In contrast, in an assembly-based approach, genes are identified within the assembled contigs using specific computational tools for gene finding, including MetaGeneMark (Zhu, Lomsadze, and Borodovsky 2010) and FragGeneScan (Rho, Tang, and Ye 2010). Once this identification step is performed, genes are functionally annotated by homology-based searches against various databases of enzymes, protein families and domains, such as Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa and Goto 2000), COG systems (Tatusov *et al.* 2000), UniProt (Consortium 2014), and Pfam (Finn *et al.* 2014). Regardless of whether assembly is performed or not, both approaches yield a functional profile in which gene families and pathways are identified, and their relative abundances quantified (Quince *et al.* 2017).

Finally, irrespectively of the profiling approach used, matrices of samples versus microbiome features, such as taxa, gene functions or pathways will be the outputs obtained from processing metagenomic data. Higher-level analyses focus not only in using these matrices to investigate significant associations between sample metadata and microbial features, but also in tracking and comparing strains across samples, and in reconstructing microbial phylogeny (Quince *et al.* 2017).

2.2.4. Metatranscriptomics

Metatranscriptomics is an emerging molecular-based approach that aims to characterize the active functional profile of the microbiome in a given sample, through analysis of the total transcribed RNA. Advances in high-throughput RNA sequencing have given the opportunity to investigate the genes actively expressed not only in the host but also in complex microbial communities (Knight *et al.* 2018; Bashiardes, Zilberman-Schapira, and Elinav 2016).

This approach is similar to WMS, however while metagenomics describes the potential microbial functions of the microbiome, metatranscriptomics reveals the actual functional activity of these microbial communities (Bashiardes, Zilberman-Schapira, and Elinav 2016; Claesson, Clooney, and O'Toole 2017; Gosalbes *et al.* 2011). Moreover, it enables discrimination of active live microbial communities, even though there is a bias towards microorganisms with a high rate of transcription (Knight *et al.* 2018).

Metatranscriptomics have several major challenges that can interfere with its large-scale application. RNA has a short half-life, being less stable than DNA, which requires careful sample collection and preservation to avoid compromising its integrity (Edri and Tuller 2014). The rRNA and transfer RNA (tRNA), which are highly abundant in total RNA samples, should be depleted before sequencing, because they extensively decrease coverage of messenger RNA (mRNA), interfering with the subsequent metatranscriptomics analysis (Peano *et al.* 2013; Sultan *et al.* 2014). Host RNA contamination is also another important issue to consider when performing metatranscriptomic studies, since it can be challenging to distinguish host from microbial RNA, especially in tissue specimens (Morgan and Huttenhower 2014). Finally, metatranscriptomic data should be analyzed simultaneously with WMS data from the same sample, in order to discriminate between different transcription levels and microbial abundances changes (Knight *et al.* 2018; Franzosa *et al.* 2014).

Typically, the first step in a metatranscriptomics experiment is the total RNA isolation from the sample. After depletion of rRNA, tRNA and host RNA, the enriched microbial mRNA is fractionated and the complementary DNA (cDNA) is synthetized. Then, the adapters are ligated to cDNA and libraries are subsequently, amplified and sequenced (Bashiardes, Zilberman-Schapira, and Elinav 2016; Giannoukos *et al.* 2012). Similar to whole metagenome analysis, metatranscriptomics data analysis starts with a pre-processing step of the raw reads, which involves filtering low-quality reads, and removal *in silico* of rRNA, tRNA, and host RNA reads still present (Morgan and Huttenhower 2014). In case reference genomes are available, the filtered mRNA sequences can then be directly mapped and annotated to specific databases like KEGG (Kanehisa and Goto 2000), COG or UniProt (Consortium 2014), using alignment tools, such as Bowtie2 (Langmead and Salzberg 2012) and BLAST (Altschul *et al.* 1990). Reads are then normalized for gene length and quantity in each sample, followed by calculation of the gene expression using software like CuffDiff (Trapnell *et al.* 2012), and by statistical analysis. However, when reference genomes are not available, reads are *de novo* assembled into transcripts, and the assembled transcripts are mapped against annotated protein databases with tools like Blast2GO (Conesa *et al.* 2005). Bioinformatic pipelines used in whole metagenome analysis, including HUMAnN2

(Franzosa *et al.* 2018) and MG-RAST (Keegan, Glass, and Meyer 2016), can also be used to perform functional assignment of metatranscriptomic data.

2.3. Metaproteomics

Metaproteomics aims to characterize the whole protein complement within a sample. It combines liquid-chromatography-based separation with mass spectrometry, to identify and quantify the wide range of proteins expressed (Wilmes and Bond 2006; Verberkmoes *et al.* 2009; Marchesi and Ravel 2015).

In contrast with metagenomics and metatranscriptomics that only allow the identification of genes or transcripts within the microbial community, metaproteomics directly monitors the microbial protein expression, thus capturing the actual functional state of the microbiome (Erickson *et al.* 2012; Verberkmoes *et al.* 2009). Also, by simultaneously measuring host and microbial proteins, this technology has the potential to unravel information about the dynamics of microbial communities, providing a better understanding of the host-microbiome interactions (Zhang *et al.* 2017; Zhang, Deeke, *et al.* 2018; Verberkmoes *et al.* 2009). In fact, in a recent ultradeep metaproteomic analysis of the gut microbiome, it was shown that the great majority of enriched human proteins in feces were functionally related to defense response against microorganisms (Zhang *et al.* 2017).

Despite the huge potential of metaproteomics to characterize the function of microbial communities, the application of this technology on microbiome studies is still at an early stage, since there are technical limitations to be addressed. Computational data analyses of metaproteomics is still challenging due to high complexity of datasets containing proteins at varying relative abundances from several different microbial species (Zhang *et al.* 2017), but also the limited availability of reference metaproteomic databases (Segal *et al.* 2019). Moreover, and despite the technical improvements in the performance of current mass spectrometers, they still have insufficient resolution to quantify certain low-abundance proteins (Muth *et al.* 2018; Zhang *et al.* 2017). Finally, novel bioinformatic tools for analysis of high complex metaproteomic data are necessary to obtain a precise and complete functional characterization of the microbiome (Muth, Renard, and Martens 2016).

2.4. Metabolomics and the microbiome

Metabolomics consists in the comprehensive description of non-protein small molecules, normally known as metabolites, in a given sample under specific environmental conditions (Allaband *et al.* 2019; Nicholson and Lindon 2008; Johnson *et al.* 2012). The use of chromatography separation methods (liquid- or gas-chromatography), coupled with

detection techniques (mass spectrometry or nuclear magnetic resonance spectrometry) enables the identification and characterization of the complete set of metabolites present in cells, tissues, and biofluids (urine, blood, saliva, etc.) (Johnson *et al.* 2012; Beckonert *et al.* 2007; Nicholson and Lindon 2008). These analytical techniques yield complex datasets containing mass spectral peaks, which are then compared with spectral databases in order to identify the metabolites present in a sample. Complex bioinformatics and statistical analysis are applied to process and interpret the metabolic profiling data (Johnson, Ivanisevic, and Siuzdak 2016).

The profiled metabolites can be generated by the host or the microbial communities, but can also be derived from exogenous sources, including those from the diet (Johnson *et al.* 2012). Therefore, by capturing all metabolites, metabolomics is another approach that can be used to study the function of the microbiome, providing a tool to understand the metabolic interactions between the microbiome and the host (Pedersen *et al.* 2016; Lamichhane *et al.* 2015). There are two different methodologies used in metabolomics analysis: untargeted and targeted approaches (Johnson, Ivanisevic, and Siuzdak 2016; Melnik *et al.* 2017). In particular, untargeted metabolomics aims to assess the widest variety of metabolites possible in a sample, allowing the discovery of novel compounds. In contrast, targeted metabolomics accurately measures a predetermined group of metabolites, using methods of analysis already optimized for the specific molecules. This approach provides a more sensitive and selective quantification than the untargeted metabolomics, but it is not suitable for the discovery of new compounds. For example, microbiome-derived metabolites that are commercially unavailable cannot be analyzed through this approach (Allaband *et al.* 2019; Johnson, Ivanisevic, and Siuzdak 2016).

Although metabolomics aims to cover the largest possible number of metabolites, it is not likely to capture the whole metabolome simultaneously, due to different chemical properties of metabolites in a sample, and to extraction methods (Johnson *et al.* 2012; Goodacre *et al.* 2004). Another important drawback is the difficulty to distinguish whether the identified metabolite derives from the host or the microbiome. Also, the identification of the microbial taxa that produced or altered a metabolite, constitutes a challenge in metabolomics of microbiome samples (Knight *et al.* 2018).

3. The gastric microbiome

3.1. *Helicobacter pylori* infection and gastric cancer

For many years, the human stomach was assumed to be sterile, given its high acidic pH, gastric peristalsis, and the presence of digestive enzymes, among other protective and

antimicrobial factors (Martinsen, Bergh, and Waldum 2005). With the discovery and isolation of *Helicobacter pylori* (Warren and Marshall 1983) this dogma was broken.

H. pylori is a bacteria species that colonizes the human gastric mucosa of about 45% of the population worldwide, with considerable variation according to the geographic region (Zamani *et al.* 2018). Since its initial isolation, *H. pylori* has been associated with chronic gastritis, peptic ulcer disease, gastric mucosa-associated lymphoid tissue lymphoma, and gastric cancer (Wotherspoon *et al.* 1991; Parsonnet *et al.* 1994; Parsonnet *et al.* 1991; Nomura *et al.* 1994)

H. pylori is considered as the major risk factor for the development of gastric cancer, being categorized as a class I carcinogen by the International Agency for Research on Cancer (IARC 1994). It has been estimated that at least 90% of all non-cardia gastric cancers worldwide are attributable to *H. pylori* (Plummer *et al.* 2015).

Gastric cancer is the fifth most incident and the third cause of cancer-related death worldwide (Ferlay *et al.* 2015). The incidence of gastric cancer shows wide geographic variation, with major geographic overlap with *H. pylori* prevalence, and in general countries with highest cancer incidence have high infection rates (Ferlay *et al*. 2015; Zamani *et al*. 2018). The magnitude of the risk of gastric cancer associated with *H. pylori* infection has now been estimated in different populations, and varies with the type of assay used to detect the infection, being about 3-fold if serology is used (Helicobacter and Cancer Collaborative 2001) and reaching over 20-fold when more sensitive assays are used (Gonzalez *et al.* 2011). As an additional piece of evidence that links *H. pylori* infection and gastric cancer, the eradication of the infection has an impact in reducing the incidence of this malignancy (Ford *et al.* 2015).

Gastric cancer is the result of a long and multistep process, which starts with *H. pylori*-associated chronic gastritis, followed by atrophic gastritis, intestinal metaplasia, dysplasia, and cancer (Correa 1992). Although the association between *H. pylori* and gastric cancer is extensively recognized, the majority of the infected patients do not develop this malignancy, which arguments in favour of the multifactorial nature of this disease. Host genetic susceptibility, namely polymorphisms in genes that are involved in the inflammatory response to *H. pylori* infection have been associated with the risk of gastric cancer. Among the best studied are those that encode interleukin (IL)-1β, IL-1 receptor antagonist, and tumour necrosis factor (TNF)-α pro-inflammatory cytokines and the anti-inflammatory IL-10. Genetic variation in the promoters or in non-coding regions of these genes are associated with increased risk for the development of gastric cancer (El-Omar *et al.* 2001; Machado *et al.* 2003; Persson *et al.* 2011). Remarkably, in genetically susceptible hosts, infection with more virulent *H. pylori* strains markedly enhances gastric cancer risk (Figueiredo *et al.* 2002).

Cigarette smoking, alcohol intake, and salt consumption are recognized environmental factors that influence the risk of gastric cancer (Praud *et al.* 2018; Rota *et al.* 2017; D'Elia *et al.* 2012). Adding to the influence of host and environmental factors in gastric cancer, the genetic diversity of *H. pylori*, and in particular variation in virulence genes associated with the pathogenicity of strains, also impact gastric cancer risk (Ferreira, Machado, and Figueiredo 2014). CagA is the best-documented *H. pylori* virulence factor influencing gastric cancer. CagA is encoded by a pathogenicity island that is present in about 60% to 70% of *H. pylori* strains worldwide. The same pathogenicity island also encodes a type IV secretion system, which functions as a molecular syringe and allows CagA to be delivered into the host cells (Backert, Tegtmeyer, and Fischer 2015). Once in the host cell cytoplasm, CagA can be phosphorylated by host kinases within EPIYA motifs. Both phosphorylated and non-phosphorylated CagA are capable of activating signalling pathways that influence host responses, including inflammation, proliferation, and cell polarity (Backert, Tegtmeyer, and Selbach 2010). CagA phosphorylation, however, appears to be important in gastric cancer development, as transgenic mice expressing wild-type CagA, but not phosphorylation-resistant CagA, develop gastric tumours (Ohnishi *et al.* 2008). Patients who are infected with *H. pylori cagA*-positive strains, and with strains with CagA harbouring higher number of phosphorylation motifs, are associated with increased risk for gastric premalignant lesions and for gastric cancer (Ferreira, Machado, and Figueiredo 2014). Additionally, CagA influences host disease progression, and infection with *H. pylori cagA*-positive strains increases the risk of progression of preneoplastic lesions (Plummer *et al.* 2015; Gonzalez *et al.* 2011). Variation in other *H. pylori* virulence factors, such as the VacA toxin, has also been associated with gastric precancerous lesions and cancer (Gonzalez *et al.* 2011; Ferreira, Machado, and Figueiredo 2014).

3.2. The gastric microbiota, is there more than *H. pylori*?

More recently, it has been shown that, in addition to *H. pylori*, the stomach harbours a complex bacterial community. Initial analyses of the bacteria present in the stomach relied on microbiological cultures. These have identified *Firmicutes* as the most common phylum, followed by *Proteobacteria*, *Bacteroidetes*, and *Actinobacteria*, and genera that were most commonly isolated included *Streptococcus*, *Lactobacillus*, *Bacteroides*, *Staphylococcus*, *Veillonella*, *Corynebacterium, Clostridium,* and *Neisseria* (Adamsson *et al.* 1999; Mowat *et al.* 2000; Stockbruegger 1985; Thorens *et al.* 1996; Zilberstein *et al.* 2007). This type of approach, however, yielded an incomplete and biased landscape of the gastric microbiota, since most of the bacteria are difficult to culture or are uncultivable. The development of culture-independent methods revealed that the human gastric ecosystem has a more

diverse and complex microbiota than initially anticipated (Andersson *et al.* 2008; Bik *et al.* 2006; Delgado *et al.* 2013; Li *et al.* 2009; Monstein *et al.* 2000; Schulz *et al.* 2018).

The bacterial community of the normal stomach has not been extensively characterized, probably due to difficulties in recruiting normal individuals for upper endoscopy. A 16S rRNA gene cloning and sequencing-based approach was undertaken to analyse the gastric microbial communities of five individuals with normal gastric mucosa and five patients with non-*H. pylori* and non-NSAID (non-steroidal anti-inflammatory drug) (NHNN) gastritis, all Chinese from Hong-Kong (Li *et al.* 2009). *Firmicutes* and *Proteobacteria* were the most represented phyla, and while in the normal stomach the *Proteobacteria* was the most abundant, in the NHNN gastritis the most abundant phylum was the *Firmicutes*. The five most common genera were *Streptococcus*, *Prevotella*, *Neisseria*, *Haemophilus*, and *Porphyromonas*; together, *Streptococcus* and *Prevotella* represented over 40% of all sequences.

Following studies exposed the diversity and the inter-individual variability of the gastric microbiota derived from the analysis of populations from distinct origins, but also from different sample types, and using various technical approaches. Overall, the most common gastric bacteria can be assigned to five major phyla – *Proteobacteria*, *Firmicutes, Bacteroidetes*, *Actinobacteria,* and *Fusobacteria*, and the two most prominent genera of the non-*H. pylori* infected stomach are *Streptococcus* and *Prevotella* (Andersson *et al.* 2008; Bik *et al.* 2006; Delgado *et al.* 2013; Li *et al.* 2009). A more recent study that included 20 Caucasians from the UK with a normal stomach, without evidence of *H. pylori* infection, concurred that the bacterial family *Prevotellaceae* was the most abundant (23%), followed by *Streptococcaceae* (10%). In fact, the microbiota of these stomachs had the highest levels of microbial diversity and bacterial richness in comparison with other groups of patients infected with *H. pylori* (Parsons *et al.* 2017).

According to the great majority of reports, when *H. pylori* is present, this bacterium is the most abundant microbial component, representing between 40% to over 95% of the gastric microbiota (Andersson *et al.* 2008; Bik *et al.* 2006; Klymiuk *et al.* 2017; Li *et al.* 2017; Parsons *et al.* 2017; Schulz *et al.* 2018) . In addition to finding *H. pylori* as the most abundant bacterium in the stomach of patients who test positive for *H. pylori*, it has been shown that the microbiota of *H. pylori*-positive subjects has lower diversity than that of *H. pylori*-negative subjects (Andersson *et al.* 2008; Bik *et al.* 2006; Schulz *et al.* 2018). A study that evaluated the gastric microbiota before and after *H. pylori* eradication treatment, showed that the eradication of *H. pylori* resulted in an increase in bacterial diversity (Li *et al.* 2017). The influence of *H. pylori* on the composition and dynamics of the gastric microbiota is still not fully understood. Difficulties may in part relate to the differences in methods to diagnose *H. pylori* infection and various studies using sequencing-based methods have demonstrated

that *H. pylori* could be detected at low levels in samples of subjects that were diagnosed as *H. pylori*-negative by conventional methods (histopathology, rapid urease test, serology, and PCR) (Bik *et al.* 2006; Delgado *et al.* 2013; Maldonado-Contreras *et al.* 2011; Thorell *et al.* 2017).

The majority of reports show no major alterations on the pattern of distribution of phyla between *H. pylori*-positive and *H. pylori*-negative patients (Bik *et al.* 2006; Maldonado-Contreras *et al.* 2011; Schulz *et al.* 2018). Using the PhyloChip microarray, Maldonado-Contreras *et al.* reported a similar representation of the four dominant phyla between *H. pylori*-infected and -uninfected rural Amerindians (Maldonado-Contreras *et al.* 2011). In regression analyses, authors were able to identify an association between *H. pylori* positivity and decreased relative abundance of *Actinobacteria*, *Bacteroidetes,* and *Firmicutes*. Experimental infections of the rhesus macaque model were used to assess the impact of *H. pylori* challenge upon the pre-existing gastric microbiota (Martin *et al.* 2013). Data showed that although *Helicobacter* became dominant in challenged animals, the removal of the *Helicobacter* reads from the libraries did not significantly alter the relative abundance of taxa between challenged and unchallenged animals. Nevertheless, the impact of *H. pylori* on relatively rare taxa was not determined. In contrast, in a mouse model of infection, challenge of animals with *H. pylori* significantly and consistently affected the abundance of several species, suggesting that *H. pylori* influences the gastric microbiota composition at lower taxonomic levels (Kienesberger *et al.* 2016).

It has been a matter of debate whether bacteria found in the stomach represent transient swallowed bacteria or active members of a resident microbiota colonizing the gastric mucosa. Comparisons of the microbial communities along different sites of the gastrointestinal (GI) tract have shown that the gastric microbiota is different from that at other sites. Although some proximity with the microbiota of the oral cavity and throat exists, the stomach microbial communities cluster together (Andersson *et al.* 2008; Delgado *et al.* 2013; Stearns *et al.* 2011). Recent data aiming to evaluate the metabolically active microbial communities in different regions of the GI tract found that the transient luminal microbiota present in gastric juice is closely related with that of saliva and of duodenal aspirates and significantly different from that of gastric biopsies, supporting the idea that the stomach has a local mucosa-associated microbiota (Schulz *et al.* 2018).

3.3. The gastric microbiota in gastric carcinogenesis

While *H. pylori* is recognized as being fundamental in gastric carcinogenesis, the role of non-*H. pylori* microbiota has not yet been established. The majority of the publications so far included low number of patients and/or had limitations in sensitivity and depth of

coverage, which in general did not allow producing statistically based conclusions. One of the first DNA-based descriptions of the gastric bacterial community in patients with gastric cancer, used T-RFLP in combination with 16S rRNA gene cloning and sequencing to characterize 10 patients with gastric cancer and five *H. pylori*-negative dyspeptics with normal gastric mucosa (Dicksved *et al.* 2009). A complex bacterial community dominated by different species of *Streptococcus*, *Lactobacillus*, *Veilonella* and *Prevotella*, and with low abundance of *H. pylori* was reported in the stomach of cancer patients.

A study of 15 patients from Mexico with non-atrophic gastritis, intestinal metaplasia, or gastric cancer, using the PhyloChip, showed a gastric microbiota profile separation between non-atrophic gastritis and gastric cancer based on the presence/absence of taxa. This analysis could neither separate non-atrophic gastritis and intestinal metaplasia, nor metaplasia and cancer (Aviles-Jimenez *et al.* 2014). Taxa with differences in abundance between non-atrophic gastritis and gastric cancer were identified, with significant decreases in the abundance of *Porphyromonas*, *Neisseria* and bacteria from the *TM7* phylum, and increases in the abundance of *Lactobacillus* and *Lachnospiraceae* observed in gastric cancer. Diversity, as measured by bacterial richness, was statistically significantly decreased from non-atrophic gastritis to gastric cancer. In contrast, a survey of the metabolic active bacteria of the stomach of 12 gastric cancer and 20 functional dyspepsia patients of Chinese ethnicity from Singapore and Malaysia, detected an increase in species richness and in phylogenetic diversity in cancer (Castano-Rodriguez *et al.* 2017). An earlier study of 10 chronic gastritis, 10 intestinal metaplasia and 11 gastric cancer patients from Korea, also suggested an increase in bacterial diversity from gastritis to cancer, but without supporting statistical analysis (Eun *et al.* 2014). Still, the majority of publications so far report a decrease in bacteria diversity and richness from non-atrophic gastritis to gastric cancer (Aviles-Jimenez *et al.* 2014; Coker *et al.* 2018; Li *et al.* 2017).

A complete gastric microbiota study in the gastric cancer field using 16S rRNA gene sequencing was published in the beginning of 2018 (Coker *et al.* 2018). Coker and colleagues studied the gastric mucosal microbiota in different histological stages of gastric carcinogenesis in 81 patients from Xi'an in China (Coker *et al.* 2018). The analysis of 21 superficial gastritis, 23 atrophic gastritis, 17 intestinal metaplasia, and 20 gastric cancer patients, demonstrated that the gastric microbiota of patients with intestinal metaplasia and with gastric cancer had significantly reduced microbial richness in comparison with that of superficial gastritis patients. Although no significant differences were found in microbiota profiles between superficial gastritis, atrophic gastritis and intestinal metaplasia, the microbiota of these stages were significantly different from that of the gastric cancer. The screen for differentially abundant taxa revealed 21 taxa enriched and 10 taxa depleted in gastric cancer in comparison with superficial gastritis, with increasing strengths of

interactions among them along the progression of disease. Among the cancer-enriched bacteria were members of the human oral microbiome *Peptostreptococcus*, *Streptococcus*, *Parvimonas*, *Slackia*, and *Dialister*, which were the most significant in network interaction analysis. These bacteria were able to distinguish gastric cancer from superficial gastritis in receiver-operating characteristic (ROC) analysis. The authors validated their results in a Chinese Inner Mongolian cohort of patients (Coker *et al.* 2018).

The role of the microbiota in the promotion of neoplasia is supported by data obtained in the insulin-gastrin (INS-GAS) transgenic mouse model. In comparison with germ-free INS-GAS mice, those harbouring a complex microbiota had higher levels of gastric inflammation, epithelial damage, oxyntic gland atrophy, hyperplasia, metaplasia, and dysplasia. When infected with *H. pylori*, INS-GAS mice that harboured a complex microbiota had more severe gastric lesions and an earlier development of gastrointestinal intraepithelial neoplasia (GIN) in comparison to *H. pylori*-infected germ-free INS-GAS mice (Lofgren *et al.* 2011). Furthermore, progression towards GIN occurred to a similar extent in *H. pylori*-infected INS-GAS mice with a complex microbiota and in *H. pylori*-infected INS-GAS mice colonized with a restricted microbiota consisting of only three species of commensal murine bacteria (*Clostridium* spp., *Lactobacillus murinus*, and *Bacteroides* spp.) (Lertpiriyapong *et al.* 2014). These results suggest that colonization of the stomach with commensal bacteria from other locations of the GI tract may promote *H. pylori*-associated gastric cancer. Altogether, these studies highlight that there is a shift in the composition of the stomach microbiome from gastritis to gastric cancer, with a likely reduction of bacterial diversity, and with increased microbial dysbiosis in the cancerous stomach.

3.4. Revisiting Correa's hypothesis of gastric carcinogenesis

In the multistep model of gastric carcinogenesis proposed by Pelayo Correa, persistent infection of the gastric mucosa with *H. pylori* initiates and perpetuates an inflammatory process that can progress to atrophic gastritis, intestinal metaplasia, dysplasia, and gastric cancer (Correa 1992). In this model, *H. pylori* infection plays an important role in the initial phases of the cascade. Indeed, *H. pylori* scarcely colonizes the severe atrophic stomach and may progressively disappear in gastric tissues at later steps of carcinogenesis (Correa 1992; Kuipers 1998). Analyses of the gastric microbiome have also described decreased relative abundance of *H. pylori* in gastric cancer (Dicksved *et al.* 2009; Eun *et al.* 2014; Hsieh *et al.* 2018), although this was not consistently observed or not reported (Coker *et al.* 2018; Yu *et al.* 2017).

The hypothesis of Correa contemplated that the loss of acid-secreting parietal cells in *H. pylori*-induced atrophic gastritis leads to higher gastric pH, and to proliferation in the

stomach of bacteria that are capable of reducing nitrate to nitrite, to form N-nitroso compounds that are mutagenic (Correa 1992; Correa *et al.* 1975). Actually, significant intragastric bacterial overgrowth has been demonstrated in patients on long-term acid suppression by the use of proton pump inhibitors (PPIs) or histamine-2 receptor antagonists (Stockbruegger 1985; Sanduleanu *et al.* 2001). A recent investigation of 24 dyspeptic Italian patients, showed that although PPI treatment did not have a major influence in the gastric microbiota composition, an increase in the relative abundance of *Firmicutes*, namely *Streptococcus* was reported (Paroni Sterbini *et al.* 2016). In accordance with these findings, in a study analysing the metabolically active gastric microbial communities of 19 patients from the UK receiving PPI therapy and 20 individuals with normal stomach, relatively few alterations in the gastric microbiota were detected, but *Streptococcus* was significantly enriched in PPI-treated patients (Parsons *et al.* 2017). An enrichment in *Streptococcaceae* in the gut microbiota of PPI users has also been reported in two large studies (Imhann *et al.* 2016; Jackson *et al.* 2016). The enrichment of upper GI tract commensals observed in the stomach and in the gut, may be related with the disruption of the highly acidic barrier of the stomach induced by the acid suppressive therapy.

Likewise, the increase of the pH of the stomach due to decreased acid production as a result of parietal cell loss in *H. pylori*-associated atrophy, may generate a niche that becomes suitable to the establishment of a different microbiome (Plottel and Blaser 2011). As we have recently proposed in a critical review of the literature (Pereira-Marques *et al.* 2019), this altered gastric microbiome where *H. pylori* is less abundant or absent, and where commensal bacteria from other locations of the GI tract may thrive, would act as continuous stimuli by maintaining the inflammatory process and/or inducing genotoxicity, thus promoting gastric carcinogenesis **(Figure 5)**. This would in part explain the lack of success of *H. pylori* eradication in preventing progression of preneoplastic lesions and gastric cancer in patients with atrophy or intestinal metaplasia at baseline (Wong *et al.* 2004; Mera *et al.* 2018).

The microbiome of the cancerous stomach is functionally different from that of the stomach without cancer (Coker *et al.* 2018). Although only a very limited number of studies have addressed this aspect, predictive functional analyses have revealed that gastric cancer patients have an enrichment of several microbial pathways, including those related with membrane transport, carbohydrate digestion and absorption, carbohydrate metabolism, xenobiotics biodegradation and metabolism, and lipid metabolism (Castano-Rodriguez *et al.* 2017; Coker *et al.* 2018; Tseng *et al.* 2016). Findings are, however, relatively divergent between studies and results should therefore be interpreted with caution.

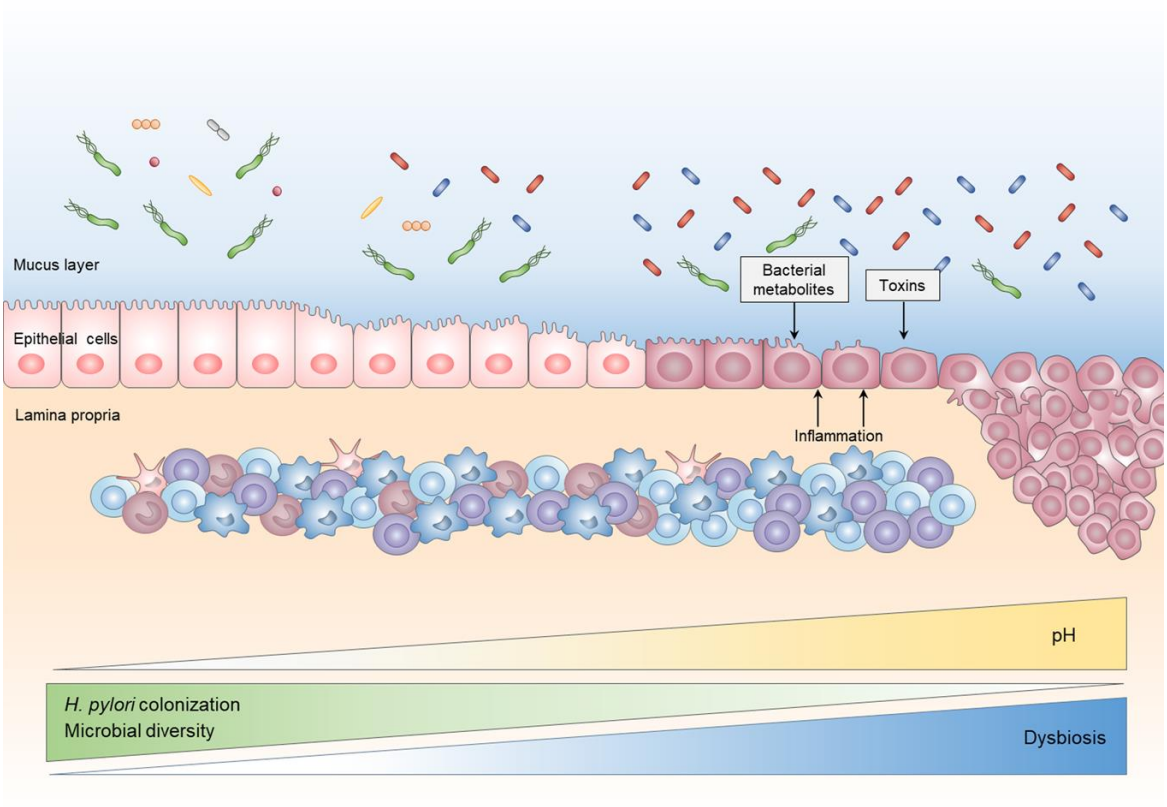


Figure 5. Model for microbial dysbiosis in gastric cancer development. *H. pylori* infection triggers and perpetuates an inflammatory response in the gastric mucosa that, in some of the infected individuals, leads to loss of acid secreting parietal cells with increase of the gastric pH. In this altered environment, *H. pylori* colonization decreases, and bacteria from other locations of the GI tract establish in the gastric niche, resulting in dysbiosis. This dysbiotic microbiome, characterized by reduced microbial diversity, may promote the development of gastric cancer by sustaining inflammation and/or inducing genotoxicity. Bacteria: green, *H. pylori*; orange, pink and grey, resident mucosa-associated microbiota; blue and red, dysbiotic microbiota; Inflammatory cells: dark blue, macrophages; pink, dendritic cells; dark pink, monocytes; light blue, CD4 T-lymphocytes; violet, CD8 T-lymphocytes. (From (Pereira-Marques *et al.* 2019) with permission of Springer Nature under licence number 4575310055215).

To revisit Correa's hypothesis that nitrate-reducing bacteria contribute to malignant transformation of the atrophic stomach by increasing the concentrations of nitrite and of N-nitroso compounds, the functional features of the microbiome involved in these reactions have been assessed (Ferreira *et al.* 2018). In comparison with chronic gastritis, the gastric cancer microbiome had an increased representation of nitrate reductase and of nitrite reductase functions, the enzymes that respectively reduce nitrate to nitrite and nitrite to nitric oxide (Ferreira *et al.* 2018). Interestingly, and in agreement with the abovementioned observations, are those of a follow-up study conducted in Taiwan to evaluate the effects of

subtotal gastrectomy as a treatment for early-stage gastric cancer. The alteration of the gastric environment by the surgery led to significant changes in the gastric microbial community, and nitrate reductase, nitrite reductase, and other functions related to nitrosation were enriched in the gastric microbiome before, but not after subtotal gastrectomy (Tseng *et al.* 2016; El-Omar *et al.* 2001). These data suggest that the gastric cancer microbiome has the potential to produce carcinogenic N-nitroso compounds. Additional features linked to the dysbiotic microbiome may be involved in the promotion of a carcinogenic environment in the stomach. Microbial metabolites and toxins, as well as inflammation by-products generated by the dysbiotic microbiome, may directly induce host cell damage or interfere with host signalling pathways that influence cell turnover and survival, thus increasing the risk for gastric malignant transformation **(Figure 5)**.

Outline and Aims

Despite the recent advances in the investigation of the human microbiome, research in the gastric microbiome is still limited. The majority of the publications had limitations in sensitivity and depth of coverage, and included low number of patients, which in general did not allow producing statistically based conclusions. Therefore, the general aim of this thesis was to establish solid molecular-based approaches, in order to characterize the gastric microbiome in the context of gastric cancer.

In the first part of this thesis, the aim was to use next-generation sequencing of the 16S rRNA gene to profile the gastric microbiota in chronic gastritis and gastric carcinoma patients. For that, the diversity and the composition of the gastric microbiota were investigated in 81 patients with chronic gastritis and in 54 patients with gastric carcinoma. Differences in microbial composition between the two patient groups were assessed using LEfSe. Associations between the most relevant taxa and clinical diagnosis were validated by real-time qPCR in DNA of the gastric biopsy specimens. An additional validation cohort of 38 gastric specimens from 15 patients with chronic gastritis and 23 patients with gastric carcinoma was used for confirmation of findings. Finally, microbial dysbiosis was compared between patient groups. These results are presented in **Part I** of the Results and in **Paper I**.

In the second part of the thesis, and since WMS could offer a higher degree of taxonomic and functional resolution by capturing the full genomic content, the aim was to establish a WMS strategy to characterize the gastric mucosa-associated microbiome. To accomplish this goal, the initial step was to set up a pipeline of analysis for WMS data for microbiome profiling, followed by optimization of sequencing data pre-processing. Validation of the optimized pipeline of analysis was achieved by evaluating the taxonomic profile of two mock DNA microbial communities. Further validation of the metagenomics workflow was performed using two additional published WMS datasets of both mock communities. These results are presented in **Part II.1** of the Results.

Since in WMS the amount of host DNA is a major challenge, and because there are no guidelines for generating an accurate analysis, **Part II.2** of the Results analyzed the sensitivity of WMS for microbiome taxonomic profiling, taking into account the wide range of sample complexities and sequencing depths. For that, synthetic samples with increasing levels of complexity were created by spiking DNA of a mock bacterial community, with DNA from a mouse-derived cell line. Additionally, datasets with variable or fixed sequencing depths were generated. The taxonomic profile of all samples and datasets was then evaluated with the WMS strategy optimized in **Part II.1**. These results generated **Paper II**.

Finally, **Part II.3** of the Results approaches the characterization of the gastric mucosa-associated microbiome using WMS, applied to metagenomes of human gastric carcinoma specimens (n = 2). Based on the results from **Part II.2**, showing that high levels of host DNA

decrease the sensitivity of WMS for microbiome profiling, the taxonomic characterization was determined in conditions in which DNA isolation of the specimens was performed without and with a human DNA depletion step.

The great majority of the data presented in the Results' section are compiled in **Papers I** and **II**. **Paper III** critically reviews the literature on the gastric microbiome, *H. pylori* infection, and gastric cancer, and was partially used in the Introduction section. An additional publication issued in the context of "The year in *Helicobacter pylori* 2017", which summarizes the findings of the literature of the previous year on the pathogenesis of *H. pylori*-associated gastric cancer, is also included (**Paper IV**).

MATERIALS AND METHODS

1. Materials

1.1. Gastric specimens

For the high-throughput profiling of the gastric bacterial communities by NGS of the 16S rRNA gene, 81 individuals with chronic gastritis and 54 with gastric carcinoma were included in the Portuguese discovery cohort. These were part of a case–control study aimed at investigating risk modifiers for gastric cancer (Figueiredo *et al.* 2002; Machado *et al.* 2003). Subjects with chronic gastritis (mean age 43.6±7.0 years; male-to-female ratio 39.5:1) were recruited during a screening programme for premalignant lesions of the gastric mucosa and underwent standard gastroscopy at Centro Hospitalar São João (CHSJ). Seventy patients had chronic superficial gastritis. The remaining eleven patients presented glandular atrophy with *foci* of intestinal metaplasia. Of these, 1 had mild corpus and moderate antral atrophy and the remaining 10 cases did not have corpus atrophy and had mild (n = 6), moderate (n = 2) or marked (n = 2) atrophy in the antrum (including incisura). Only individuals without evidence of past or present peptic ulcer disease were included. In addition, patients under PPI or antimicrobial treatments were excluded. Patients with gastric carcinoma (mean age 58.8 ± 13.2 years; male-to-female ratio 1.5:1) were diagnosed and underwent cancer resection at CHSJ. A validation cohort of an additional 38 gastric specimens from 15 patients with chronic gastritis and 23 patients with gastric carcinoma, diagnosed between 2014 and 2016, were retrieved from the tissue and tumour bank at CHSJ. For the whole metagenome analysis of the gastric microbial communities, gastric specimens from 2 patients with gastric carcinoma (2 males; one with 52 and the other with 79 years old), diagnosed in 2017, were retrieved from the tissue and tumour bank at CHSJ. All procedures were in accordance with the institutional ethical standards. Samples were delinked and unidentified from their donors.

1.2. Mock microbial communities

Genomic DNA from Microbial Mock Community B (HM-276D Even, High Concentration, v5.1H), and from Microbial Mock Community B (HM-277D Staggered, High Concentration v5.2H) was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project. These two mock microbial communities are composed of a combination of 20 bacterial genomic DNAs that differ in GC content (30% to 69%).

Microbial Mock Community B Even contains equimolar rRNA operon counts across all bacterial strains (10^6 copies per organism per µL). Microbial Mock Community B Staggered contains staggered rRNA operon counts differing by bacteria, ranging from 10^4 to 10^7 copies per organism per µL (as indicated by the manufacturer). The genomic GC content of each

species was obtained from the NCBI Genome Database. The number of 16S rRNA gene copies per genome, and the NCBI assembly accession number were obtained from the Ribosomal RNA Database Curated by the Schmidt Laboratory (Stoddard *et al.* 2015). To estimate the expected relative abundance of species, the theoretical number of genome copies per species was calculated as the ratio of input 16S rRNA copies to 16S rRNA copies per genome, and normalized by the sum of all theoretical genome copies of the species present in the mock **(Table 1 and 2)**.

Two published mock sequencing datasets, consisting of WMS data from the mock microbial communities Even (HM-276D) and Staggered (HM-277D), were retrieved from the Sequence Read Archive (PRJNA298489).

Table 1. Composition of the mock microbial community B: HM-276D, Even, High Concentration, v5.1H. Data was sorted from the highest to the lowest number of species genome copies.

Microbial species	NCBI assembly accession	GC content (%)	16S rRNA copies	16S rRNA copies per genome	No. species genome copies
Actinomyces odontolyticus ATCC 17982	GCF_000154225.1	65	1.000.000	2	500.000
Helicobacter pylori ATCC 700392	GCF_000307795.1	39	1.000.000	2	500.000
Rhodobacter sphaeroides ATCC 17023	GCF_003324715.1	69	1.000.000	3	333.333
Deinococcus radiodurans ATCC 13939	GCF_001638825.1	67	1.000.000	3	333.333
Propionibacterium acnes DSM16379	GCF_000008345.1	60	1.000.000	3	333.333
Pseudomonas aeruginosa ATCC 47085	GCF_000006765.1	67	1.000.000	4	250.000
Neisseria meningitidis ATCC BAA-335	GCF_000008805.1	52	1.000.000	4	250.000
Streptococcus pneumoniae ATCC BAA-334	GCF_000006885.1	40	1.000.000	4	250.000
Enterococcus faecalis ATCC 47077	GCF_000172575.2	38	1.000.000	4	250.000
Streptococcus mutans ATCC 700610	GCF_000007465.2	37	1.000.000	5	200.000
Staphylococcus aureus ATCC BAA-1717	GCF_000017085.1	33	1.000.000	5	200.000
Staphylococcus epidermidis ATCC 12228	GCF_000007645.1	32	1.000.000	5	200.000
Acinetobacter baumannii ATCC 17978	GCF_001593425.2	39	1.000.000	6	166.667
Listeria monocytogenes ATCC BAA-679	GCF_000196035.1	38	1.000.000	6	166.667
Lactobacillus gasseri ATCC 33323	GCF_000014425.1	35	1.000.000	6	166.667
Escherichia coli ATCC 700926	GCF_002843685.1	51	1.000.000	7	142.857
Bacteroides vulgatus ATCC 8482	GCF_000012825.1	42	1.000.000	7	142.857
Streptococcus agalactiae ATCC BAA-611	GCF_000007265.1	36	1.000.000	7	142.857
Bacillus cereus ATCC 10987	GCF_000008005.1	36	1.000.000	12	83.333
Clostridium beijerinckii ATCC 51743	GCF_000016965.1	30	1.000.000	14	71.429

Table 2. Composition of the mock microbial community B: HM-277D, Staggered, High Concentration, v5.2H. Data was sorted from the highest to the lowest number of species genome copies.

Microbial species	NCBI assembly accession	GC content (%)	16S rRNA copies	16S rRNA copies per genome	No. species genome copies
Rhodobacter sphaeroides ATCC 17023	GCF_003324715.1	69	10.000.000	3	3.333.333
Streptococcus mutans ATCC 700610	GCF_000007465.2	37	10.000.000	5	2.000.000
Staphylococcus epidermidis ATCC 12228	GCF_000007645.1	32	10.000.000	5	2.000.000
Escherichia coli ATCC 700926	GCF_002843685.1	51	10.000.000	7	1.428.571
Pseudomonas aeruginosa ATCC 47085	GCF_000006765.1	67	1.000.000	4	250.000
Staphylococcus aureus ATCC BAA-1717	GCF_000017085.1	33	1.000.000	5	200.000
Streptococcus agalactiae ATCC BAA-611	GCF_000007265.1	36	1.000.000	7	142.857
Bacillus cereus ATCC 10987	GCF_000008005.1	36	1.000.000	12	83.333
Clostridium beijerinckii ATCC 51743	GCF_000016965.1	30	1.000.000	14	71.429
Helicobacter pylori ATCC 700392	GCF_000307795.1	39	100.000	2	50.000
Propionibacterium acnes DSM16379	GCF_000008345.1	60	100.000	3	33.333
Neisseria meningitidis ATCC BAA-335	GCF_000008805.1	52	100.000	4	25.000
Acinetobacter baumannii ATCC 17978	GCF_001593425.2	39	100.000	6	16.667
Listeria monocytogenes ATCC BAA-679	GCF_000196035.1	38	100.000	6	16.667
Lactobacillus gasseri ATCC 33323	GCF_000014425.1	35	100.000	6	16.667
Actinomyces odontolyticus ATCC 17982	GCF_000154225.1	65	10.000	2	5.000
Deinococcus radiodurans ATCC 13939	GCF_001638825.1	67	10000	3	3.333
Streptococcus pneumoniae ATCC BAA-334	GCF_000006885.1	40	10.000	4	2.500
Enterococcus faecalis ATCC 47077	GCF_000172575.2	38	10.000	4	2.500
Bacteroides vulgatus ATCC 8482	GCF_000012825.1	42	10.000	7	1.429

2. Methods

2.1. DNA extraction

For 16S rRNA gene profiling using NGS, DNA was isolated from gastric biopsies or surgical specimens of non-neoplastic gastric mucosa adjacent to the tumour, as previously described (Figueiredo *et al.* 2002).

For whole metagenome analysis, DNA extraction of gastric carcinoma specimens was performed both with and without a host DNA depletion step. For that, each gastric specimen was divided in two fragments, and the DNA was isolated with two different DNA extraction kits, the Ultra-Deep Microbiome Prep (with host DNA depletion; Molzym) and the QIAmp DNA tissue kit (without host DNA depletion; Qiagen), according to the manufacturer's instructions, as briefly detailed below.

For the isolation with the Ultra-Deep Microbiome Prep kit (with host DNA depletion), tissue specimens were first incubated under denaturing conditions with proteinase K solution. Then, the partially digested tissue was treated with a chaotropic buffer for host cell lysis, followed by a DNAse treatment for degradation of the free-floating DNA, including the DNA released from host cells. Microbial cells were concentrated by centrifugation, and then supplemented with the reagent BugLysis and β-mercaptoethanol, to degrade the cell walls of bacteria and fungi. After microbial cells lysis with proteinase K solution, a bind-wash-elute spin column DNA purification was performed. The DNA was eluted in 100 µL Microbial-DNA free water (Qiagen).

For isolation using the QIAmp DNA tissue kit (without host DNA depletion), tissue specimens were homogenized with PBS 1X and Buffer ATL using the TissueRuptor. Then, samples were lysed under denaturing conditions with proteinase K solution. The resulting lysate was supplemented with RNAse A, Buffer AL, and ethanol (96-100%), before being transferred onto the QIAamp MinElute column. Finally, the DNA was captured on the silica-based membrane, washed successively, and eluted in 100 µL Microbial-DNA free water (Qiagen).

Genomic DNA from the MC-38 cell line, which is derived from C57BL/6 murine colon adenocarcinoma cells, was extracted with the QIAamp DNA Tissue kit as described above, but without the tissue homogenization step.

2.2. 16S rRNA gene sequencing

The 16S rRNA gene was amplified using primers U789F 5′-TAGATACCCTGGTAGTCC-3′ and U1053R 5′-CTGACGACAGCCATGC-3′ targeting the V5-V6 hypervariable regions and

sequenced in an Ion PGM Torrent platform following manufacturer's instructions. Primers were designed following recommendations reported by Andersson *et al*., and were extensively analysed using PrimerProspector (Andersson *et al.* 2008; Walters *et al.* 2011). The PCR reactions were performed in 25 µL containing 1X AmpliTaq Gold 360 Master Mix (Applied Biosystems, Foster City, CA) and 0.4 µM of forward and reverse primers. PCR was performed with 9 min of predenaturation at 95ºC, followed by 25 cycles of 30 seconds at 95ºC, 45 seconds at 52ºC, and 45 seconds at 72ºC. Final extension was performed for 10 minutes at 72ºC. Amplicons of approximately 280 bp were visualized and purified using the E-Gel SizeSelect Agarose Gels (Life Technologies, Foster City, CA), their concentration was determined with Qubit dsDNA HS Assay Kit (Life Technologies) and the respective size distribution with Qiaxcel DNA screening (Qiagen, Germany). Equal concentrations were used for library preparation to incorporate adaptor and barcode sequences. Next-generation sequencing was performed in the Ion PGM Torrent platform using Ion 316v2 Chips (Life, Technologies), following manufacturer's protocols. Briefly, 50 ng of amplified DNA was used to ligate the Ion Torrent A (containing the sequencing primer ligation site) and P1 adaptors (the site for Ion Sphere Particles-ISP ligation). Afterwards, the DNA molecules entered in a process of clonal amplification through an emulsion PCR (emPCR) using the OneTouch2 instrument (Ion Torrent, Life Technologies). During emPCR single DNA fragments become bound via the specific adaptor to single ISPs. This process leads to coating of each ISP with millions of copies of a single DNA fragment. To ensure only coated ISPs proceed for sequencing, an enrichment step was performed using the OneTouch ES instrument (Ion Torrent, Life Technologies). PCR negative controls containing Microbial DNA-free water (Qiagen), instead of DNA were processed as above mentioned. Data were deposited in Sequence Read Archive (PRJNA413125).

2.3. 16S rRNA gene sequencing data analysis

The performance of the UPARSE pipeline was evaluated and compared with that reported for samples of the HMP data set (Edgar 2013). The number of spurious OTUs and the number of biological meaningful OTUs were obtained according to similarity shared with sequences of the Greengenes Named Isolated database (DeSantis *et al.* 2006). OTUs were classified as "Named" (≥ 97% match from the reference database), "Chimeric" (<97% match to the reference database and chimeric with high confidence) and "Missing" (<97% match with low confidence or a biological sequence missing from the database).

Using the UPARSE pipeline (usearch_v7.0.1090_i86linux64), reads were filtered by imposing a maximum number of expected errors of 0.5 and a global trimming at 250 nucleotides (Edgar 2013). Reads were dereplicated and singletons were discarded. Filtered

reads were clustered into OTU assuming 97% similarity. Chimeric reads were reference removed using UCHIME (Edgar 2013). Each OTU was taxonomically assigned using UCLUST, considering a minimum percentage of similarity to a reference database (Greengenes Named Isolate database, release August 2013) match of 90% (Edgar 2010). Diversity analyses were performed using QIIME (V.1.9) (Caporaso *et al.* 2010). Alpha-diversity was determined by the Shannon index and with Good's estimator of coverage. Differences in alpha-diversity were assessed by the t-test controlled with 10^3 Monte Carlo permutations. Beta-diversity was assessed by unweighted and weighted UniFrac distance matrices and visualized by PCoA, controlled by 10^3 jackknife replicates (DeSantis *et al.* 2006). Sample clustering in beta-diversity analysis was tested using ANOSIM with 10^4 bootstrap replications (Anderson 2001). Comparisons between distance matrices were evaluated by the Mantel correlation controlled with 10^4 permutations.

2.4. Real-time qPCR

qPCR assays were performed using PowerUp SYBR Green Master Mix (Applied Biosystems) using different sets of primers. For quantification of each genus, two different assays were used, a universal assay (composed by universal primers targeting a conserved region of the 16S rRNA gene) and a specific assay (composed by genus-specific primers targeting the 16S rRNA gene) **(Table 3)**. The assays were designed to obtain the highest degree of specificity by comparison with sequences of Greengenes Named Isolate (release August 2013).

qPCR mixtures were prepared to a final volume of 10 µL, containing 1x PowerUp SYBR Green Master Mix, 1 µM of forward and reverse primers (Invitrogen, Foster City, CA), 2 µL of Microbial DNA-Free Water (Qiagen) and 1 µL of DNA. The qPCR was performed in a 7500 Fast Real-Time PCR System (Applied Biosystems) with the following conditions: 2 minutes at 50ºC, 10 minutes at 95ºC, followed by 40 cycles of denaturation at 95ºC for 15 seconds and annealing/extension at 60ºC for 1 minute. The amplification steps were followed by a melt dissociation step to check for nonspecific product formation. This step comprises an additional cycle of 95ºC for 15 seconds, 60ºC for 1 minute, 95ºC for 30 seconds and 60ºC for 15 seconds. In addition, the PCR product purity was also controlled by agarose gel electrophoresis. Two replicates were performed for each sample. To exclude any potential environmental contaminant in PCR reactions, blanks were prepared using Microbial DNA-Free Water (Qiagen) instead of DNA.

Table 3. Primers used in qPCR and, unless otherwise stated, designed for this study.

Primer	Sequence (5'-3')	Amplicon size (bp)	Target species
Helicobacter_F	GAAGATAATGACGGTATCTAAC	139	*Helicobacter* sp.
Helicobacter_R	ATTTCACACCTGACTGACTAT		
Neisseria_F	AACGATGTCAATTAGCTGTT	108	*Neisseria* sp.
Neisseria_R	CAATTCCTTTGAGTTTTAATC		
Achromo_F1	TCGGGCCTTGGTAGCG	77	*Achromobacter* sp.
Achromo_R1	TTCCTTTGAGTTTTAATCTT		
Phyllo_F	CTGCCTTTGATACTGGTAGT	202	*Phyllobacterium* sp.
Phyllo_R	CGGCTAGCTCTCATAGTTTA		
Clostr_F*	ATGCAAGTCGAGCGAKG	120	*Clostridium* sp.
Clostr_R*	TATGCGGTATTAATCTKCCTTT		
Rhodo_F	GGGTTCCTTCCACGGGAT	84	*Rhodococcus* sp.
Rhodo_R	CCTTTGAGTTTTAGCCTTG		
Lactob2_F	GAGGCAGCAGTAGGGAATCTTC	126	*Lactobacillus* sp.
Lactob2_R	GGCCAGTTACTACCTCTATCCTTCTTC		
Citro F1	GTAAAGTACTTTCAGCGAG	216	*Citrobacter* sp.
Citro R2	GTTTCGGATGCAGTTCCC		
Prevo_F	CACGGTAAACGATGGATGCC	113	*Prevotella* sp.
Prevo_R	CAATTCCTTTGAGTTTCACC		
Strepto_F	TGTCGTGAGATGTTGGGTTAAG	112	*Streptococcus* sp.
Strepto_R	CCACCTTCCTCCGGTTTATTAC		
340F§	TCCTACGGGAGGCAGCAGT	198	Universal
515R§	CGTATTACCGCGGCTGCTGGCAC		

*Rinttila T *et al.*, Journal of Applied Microbiology 2004 (Rinttila *et al.* 2004)

§Horz HP *et al.*, Journal of Clinical Microbiology 2005 (Horz *et al.* 2005)

To create standard curves, amplicons of each assay were cloned into pGEM-T easy vector system (Promega, Madison, WI). Dilution series of known plasmid concentrations were used to create a standard curve for each assay by plotting the log of each known concentration in the dilution series against the determined threshold cycle (Ct) value. From the standard curves, the reaction parameters (slope, y-intercept, correlation coefficient and efficiency) were obtained and the concentration of the target bacteria genus was extrapolated. The abundance of each genus was determined by the log10 ratio between

the DNA concentration determined for the specific assay and the DNA concentration determined for the universal assay.

2.5. Generation of synthetic samples

To create different synthetic complex samples with well-defined ratios of host to bacterial DNA, the mock microbial community DNA (HM-277D) was spiked with DNA from the mouse cell line (MC-38 cells). DNA concentrations of the mock microbial community and of the mouse cell line were measured using the NanoDrop 2000 UV spectrophotomer (ThermoFisher Scientific), and the exact volumes to be mixed in each condition were determined. Synthetic samples (SS) with increasing ratios of host to bacterial DNA were generated containing 10% (SS10), 90% (SS90), and 99% (SS99) host DNA. The mock microbial community sample (MS), which contains only microbial DNA, was used as control.

2.6. Library preparation and whole metagenome sequencing

Samples were first quantified and normalized to 0.2 ng/µL DNA material, using a Quant-It PicoGreen dsDNA assay (ThermoFisher Scientific), in order to use 1 ng input DNA for the library construction. Metagenomic library preparation was automated on the Hamilton Microlab STAR Liquid Handling Workstation, using a Nextera XT DNA library preparation kit (Illumina Inc., CA, USA), according to the manufacturer's protocol. Briefly, after normalized samples were fragmented and tagged by tagmentation, a limited-cycle PCR was performed to add the Index 1 (i7), Index 2 (i5) and full adapter sequences required for cluster generation. Amplification was followed by a cleanup step that purified the library DNA and removed small library fragments by using Agencourt AMPure XP beads (Beckman Coulter, Inc). The quality of the library was assessed using an Agilent Technology 2100 Bioanalyzer (Agilent Technologies, Wokingham, UK) and then, a bead-based normalization was performed using beads Nextera XT to ensure more equal library representation in the pooled library. Finally, the pooled library was sequenced as a paired-end 150-cycle run on the Illumina NextSeq 550 platform, at an expected sequencing depth of 5.5-6.5 Gb/sample.

2.7. Whole metagenome sequencing data analysis

The paired-end reads of each sample were concatenated into one single FASTQ file, which was used as input for our in-house pipeline of WMS sequencing data analysis **(Figure 6)**.

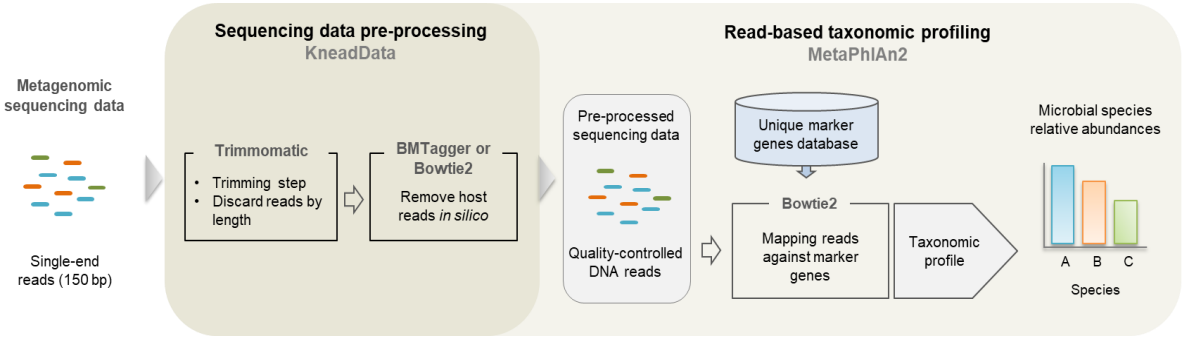


Figure 6. Schematic workflow of the pipeline of analysis for whole metagenome sequencing data. The generated metagenomes were pre-processed with KneadData. Initially, the reads of 150 bp length were trimmed based on quality and discarded based on a minimum length established, and the host reads were removed with Best Match Tagger (BMTagger) or Bowtie2 from the datasets. Next, the metagenomic samples were taxonomically profiled using MetaPhlAn2. This computational tool maps the quality-controlled WMS reads to a database of unique clade-specific marker genes, quantifying the relative abundances of each microbial clade in the samples with species-level resolution.

2.7.1. Sequencing data pre-processing

Sequencing data pre-processing was performed by KneadData (version 0.6.1), a computational tool designed to perform quality control on metagenomic sequencing data. KneadData integrates the tools FastQC (version 0.11.5) (Andrews 2016), Trimmomatic (version 0.33) (Bolger, Lohse, and Usadel 2014), and Bowtie2 (version 2.2) (Langmead and Salzberg 2012) or BMTagger (Rotmistrovsky and Agarwala 2018), to do quality check, quality filtering, and host sequences decontamination, respectively.

Quality filtering – Trimmomatic

First, reads were trimmed based on a sliding window trimming approach, cutting once the average base Phred quality score within a four-base sliding window dropped below 20, and then were discarded when the length of the read was below a specified minimum size (Minlen). Initially, the KneadData default parameters were used: Sliding window: 4:20 (4-base sliding window; average quality score 20) and Minlen: 50 (reads discarded when 50% of total input read length was trimmed). Then, optimized quality filtering parameters were

applied: Sliding window: 4:20 Minlen: 60 bp (reads discarded when the length was shorter than 60 bp).

Host sequences decontamination – Bowtie2 or BMTagger

After the quality filtering step, KneadData used Bowtie2 to identify and remove the mouse contaminant reads present in the datasets, by mapping the reads against the C57BL/6 reference genome (GCA_001632555.1 assembly). Bowtie2 was used with the default parameters (--very-sensitive end-to-end alignment). In order to remove human contaminant reads present in the datasets, KneadData used the BMTagger, a tool that identifies and separates the human from the non-human reads using the GRCh37 / hg19 assembly. The non-host quality-filtered reads were then used for the downstream analysis.

Quality check – FastQC

FastQC performed quality control checks on raw whole metagenome sequencing data but also on reads after sequencing data pre-processing, in order to assess the efficiency of the quality filtering and of host sequences decontamination steps in the generation of high-quality reads. FastQC was used with the default parameters (Andrews 2016).
The raw sequencing data was deposited in Sequence Read Archive (PRJNA521492).

2.7.2. Taxonomic profiling - MetaPhlAn2

The taxonomic profile was obtained using MetaPhlAn2 (version 2.7.5), an assembly-free taxonomic profiler (Segata *et al.* 2012; Truong *et al.* 2015). This computational tool mapped the quality-controlled reads to a database of unique clade-specific marker genes (read-based profiling) with high discriminatory power, estimating the relative abundances of each microbial clade in the samples with species-level resolution (Segata *et al.* 2012; Truong *et al.* 2015). Bowtie2, a fast DNA aligner, is used by MetaPhlAn2 to map the metagenomic reads against the unique clade-specific marker genes. Clade-specific markers constitute coding sequences that unambiguously identify specific microbial clades (at different taxonomic levels). MetaPhlAn2 relies on ~1 million unique clade-specific marker genes identified from ~17,000 reference genomes and > 7,000 unique species. Markers are now identified not only for *Bacteria* and *Archaea* (~13,500 bacterial and archaeal genomes), but also for Viruses (~3,500 viral genomes) and Eukaryotic microorganisms (*Fungi* and *Protozoa*; ~110 eukaryotic genome (Truong *et al.* 2015). The relative abundance of each clade is estimated with MetaPhlAn2 by: 1) normalizing the number of reads mapped to each marker in a clade by the markers length (estimating the coverage of each clade, clade-specific abundance); and 2) normalizing each clade-specific abundance by the sum of all

clades-specific abundances at the same taxonomic level (sum up to 100). When the abundance of a clade is larger than the sum of the abundances of the direct children clades, an unclassified sub-clade of the closest ancestor is added (i.e. when the genus abundance is larger than the sum of the abundances of its species, another unclassified species is added to that genus) (Segata *et al.* 2012; Truong *et al.* 2015).

2.8. Generation of datasets with reduced sequencing depths

The sample with the largest sequence dataset (SS90) comprising 50.8 million single-end reads and a high predominance of host DNA was used. Four datasets with reduced sequencing depths were generated by random subsampling paired-end reads using an in-house script. From the original SS90 dataset, we subsampled 50%, 25%, 10% and 5%, which correspond to 25.4, 12.7, 5.1 and 2.5 million single-end reads, respectively. For subsampling, the same seed was used in order to guarantee that the reads from the same pair were subsampled in the forward and reverse FASTQs. Then, a new set of paired FASTQ file for each random subset was created. The subsampling analysis was repeated five times for each depth.

2.9. Generation of simulated datasets with different host-microbial ratios

Simulated datasets with different host:microbial ratios were created by randomly selecting host and microbial reads from our previously sequenced datasets, and combining them in different proportions at a fixed sequencing depth of 10 million single-end reads. Using an in-house script, microbial single-end reads were randomly picked from the MS raw dataset, to assure that only microbial reads were selected. Host single-end reads were randomly picked from the mouse contaminant sequences removed by KneadData from the SS99 raw dataset, to guarantee sufficient sequences with host origin (the raw SS99 dataset contained 33.201.587 mouse single-end reads). Eighteen simulated datasets (SD) were generated, nine with progressive 10% increases in host reads (SD10 to SD90) and nine with progressive 1% increases in host reads (SD91 to SD99). For each simulated dataset, five replicates were randomly generated.

2.10. Statistical analyses

The normality of the data was evaluated using the Kolmogorov-Smirnov test. Correlations between variables were performed using Pearson's correlation (for normally distributed data). Statistical differences between groups was assessed by Student's t-test, Mann-Whitney test or by Kruskal-Wallis non-parametric test, followed by multiple comparisons

versus a control group using the Dunn's test. The differences were considered statistically significant with *P* values lower than 0.05. Statistical treatment was performed using the GraphPad Prism software (v. 6.01, GraphPad Software Inc., La Jolla, CA).

Taxonomic discovery analysis

Statistically significant differences in the relative abundance of taxa associated with groups of patients were performed using LEfSe (Segata *et al.* 2011). Only taxa with linear discriminant analysis (LDA) greater than 4 at a *P* value <0.05 were considered significantly enriched.

Microbial dysbiosis index (MDI)

The MDI was determined as the log transformation of the ratio between the total abundance of genera increased in gastric carcinoma and the total abundance of genera decreased in gastric carcinoma (Gevers *et al.* 2014). Unless otherwise stated, *Rhodococcus* spp., *Lactobacillus* spp., *Clostridium* spp., *Phyllobacterium* spp., *Achromobacter* spp. and *Citrobacter* spp. were included as increased in gastric carcinoma, and *Helicobacter* spp., *Neisseria* spp., *Prevotella* spp. and *Streptococcus* spp. were included as decreased in gastric carcinoma.

ROC analyses

ROC curves were constructed to evaluate the ability of the MDI to detect gastric carcinoma. The relative abundance was defined as the raw counts of genus-level taxa detected in at least one sample that were then normalized per sample by the total counts of all taxa in that sample so that the resulting relative abundances sum 100%. A mean ROC curve was reported including the 95% confident intervals. The best discrimination was determined by using the highest area under the curve at a significance value of $P \leq 0.05$.

RESULTS

PART I

Characterization of the gastric microbiota using next-generation sequencing of the 16S rRNA gene in chronic gastritis and gastric carcinoma patients

PART I. Characterization of the gastric microbiota using next-generation sequencing of the 16S rRNA gene in chronic gastritis and gastric carcinoma patients

Only a very small number of studies characterized the human gastric microbiota in health and disease. The limitations of these studies are the limitations in sensitivity and coverage compared with more recently developed techniques, and, most importantly, the inclusion of very limited numbers of subjects, making it difficult to generate statistically significant conclusions. In the context of gastric carcinogenesis, few studies have been conducted and no particular component of the microbiota has been identified as implicated in gastric carcinoma. Therefore, we performed high-throughput profiling of the gastric bacterial communities present in 135 gastric carcinoma cases and chronic gastritis controls, by NGS of the 16S rRNA gene. We used qPCR assays to validate the NGS results, and an additional cohort of patients to confirm our findings.

1.1. Quality control of 16S rRNA microbiota profiling

After sequencing and quality filtering, more than 10.8 million reads were obtained corresponding to a mean of 80.261 reads and 178 OTUs per sample **(Figure 7A)**. On average, patients with chronic gastritis had a significantly higher number of reads (86.957) than patients with cancer (67.954; $P < 0.05$). However, the number of OTUs was not significantly different between the two patient groups (186 and 169 OTUs, respectively; $P = 0.071$) **(Figure 7B)**. To control for the number of false OTUs and to measure the number of biologically meaningful OTUs, we classified them according to similarity shared with sequences of the Greengenes Named Isolated database **(Figure 7C and 7D)**. In our data set, the frequencies of misleading and valid OTUs were similar to those reported for the HMP data set processed with the UPARSE pipeline (Edgar 2013). In conclusion, our approach provides the most in-depth characterization of the gastric microbiota so far and generates robust and consistent data.

1.2. The gastric microbiota profile differs in chronic gastritis and gastric carcinoma

To evaluate alterations in the microbiota structure between patients with chronic gastritis and gastric carcinoma, we measured microbial alpha-diversity (i.e., within sample diversity) and beta-diversity (i.e., diversity between samples). By measuring alpha-diversity using the Shannon index, we found that patients with gastric carcinoma had significantly decreased microbial diversity in comparison with patients with chronic gastritis (**Figure 8A**, $P = 0.003$). To ensure good estimation of bacterial diversity, we measured the proportion of total bacterial species represented in samples of each patient group by the Good's estimator of

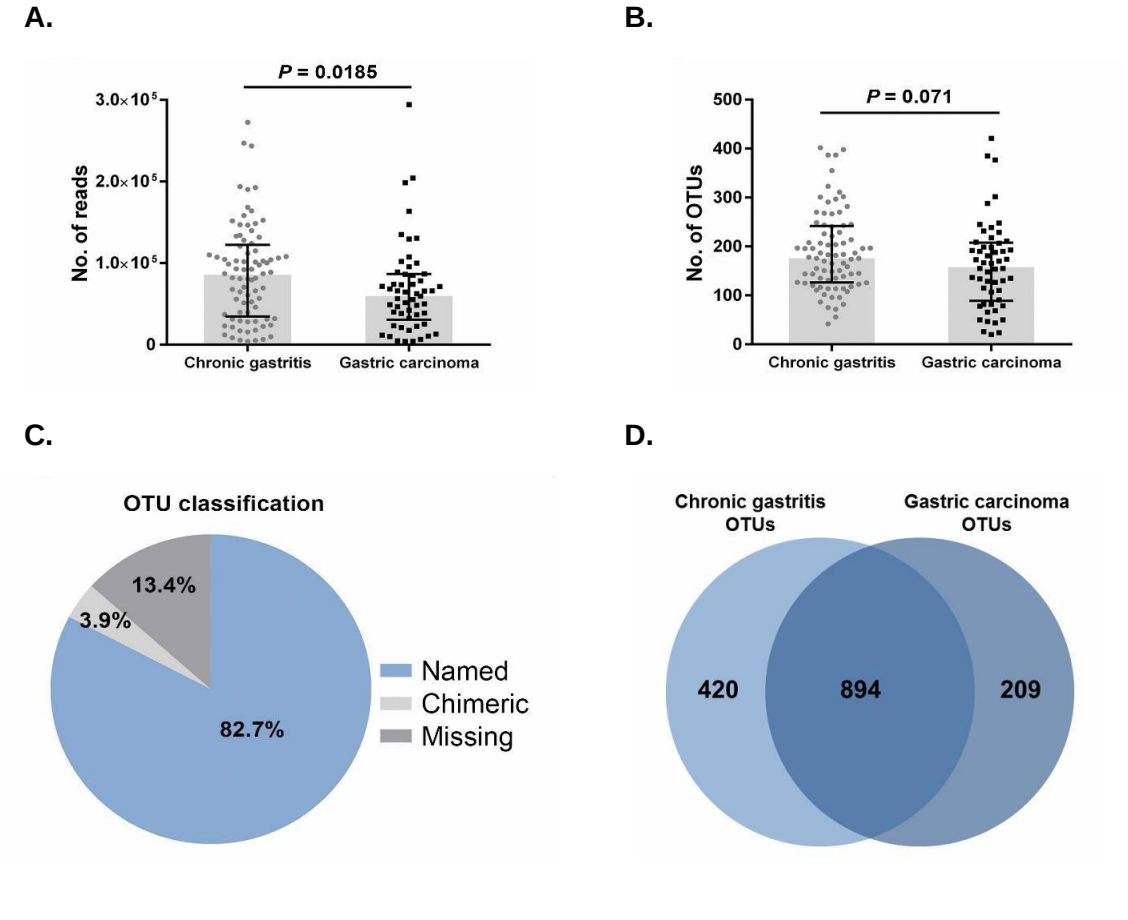


Figure 7. Performance of the UPARSE pipeline (White, Nagarajan, and Pop 2009) in 16S rRNA V5-V6 reads derived from 81 chronic gastritis and 54 gastric carcinoma patients. Number of reads **(A)** and OTUs **(B)** after quality filtering. Mann-Whitney test was used to compare chronic gastritis and gastric carcinoma patients in respect to the number of reads or OTUs. Data is shown as median with interquartile range. **(C)** Pie chart showing the average fraction of OTUs classified accordingly to similarity shared with sequences of the Greengenes Named Isolate database, as previously reported (White, Nagarajan, and Pop 2009): Named (≥ 97% match from the reference database), Chimeric (< 97% match to the reference database and chimeric with high confidence) and Missing (< 97% match with low confidence or a biological sequence missing from the database). **(D)** Venn diagram showing the number of OTUs exclusively identified in each group of patients and OTUs shared by the two groups of patients.

of coverage. Estimated coverage ranged from 0.94 to 0.98 in chronic gastritis and from 0.95 to 0.99 in gastric carcinoma (P = 0.225), suggesting that the 16S rRNA results from each (chronic gastritis and gastric carcinoma) library represent the majority of bacteria present in the gastric mucosa **(Figure 8B)**.

Beta-diversity was calculated using both unweighted (i.e., qualitative) and weighted (i.e., quantitative) UniFrac phylogenetic distance matrices, and visualised in PCoA plots. The total diversity captured by the top three principal coordinates was 29% and 47% for unweighted and weighted UniFrac, respectively. The microbiota composition of patients

with gastric carcinoma was significantly different from that of patients with chronic gastritis (ANOSIM R = 0.406, *P* = 0.0001; and R = 0.425, *P* = 0.0001, for unweighted and weighted distances, respectively; **Figure 8C and 8D**). These results show that there is a significant reduction in microbial diversity in gastric carcinoma. Furthermore, the fact that the weighted UniFrac captured more diversity than unweighted metrics suggests that alterations in the relative abundance of taxa are a major contributor for microbiota differences between gastritis and gastric carcinoma.

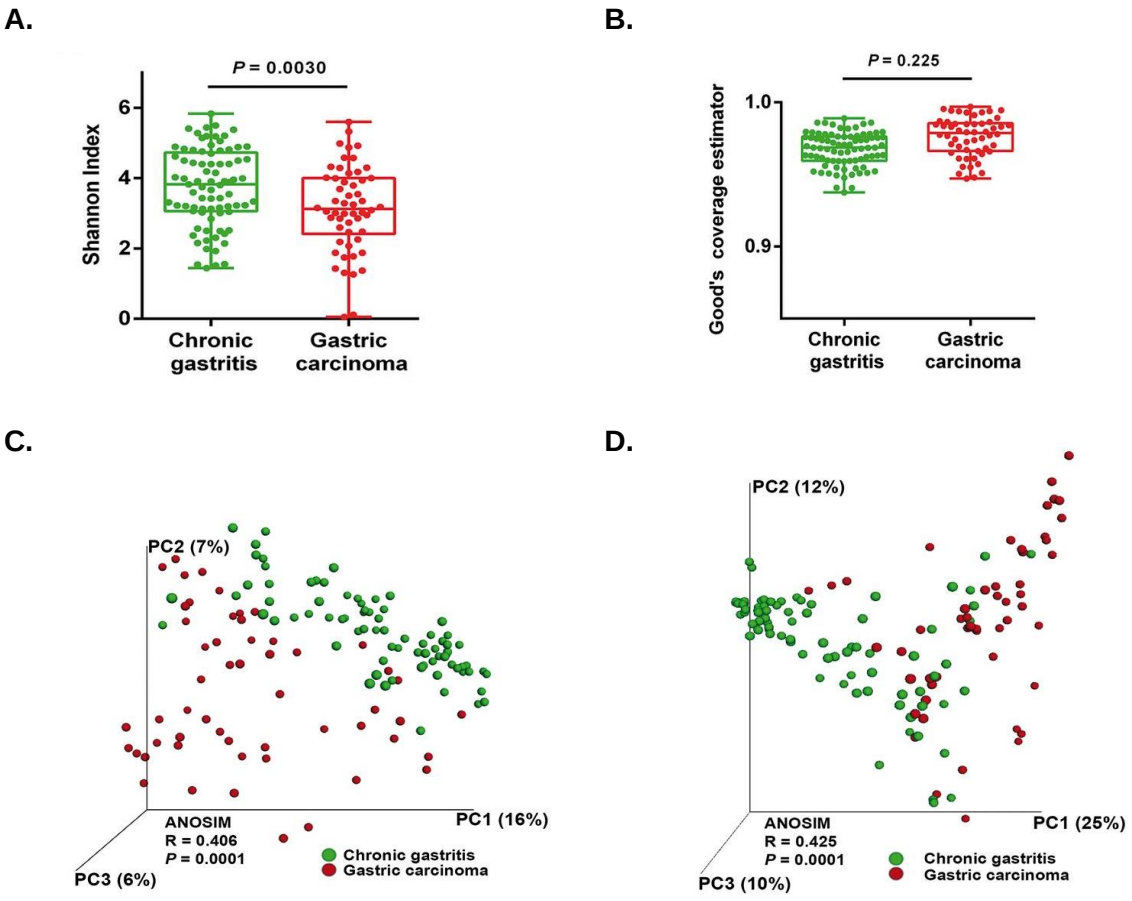


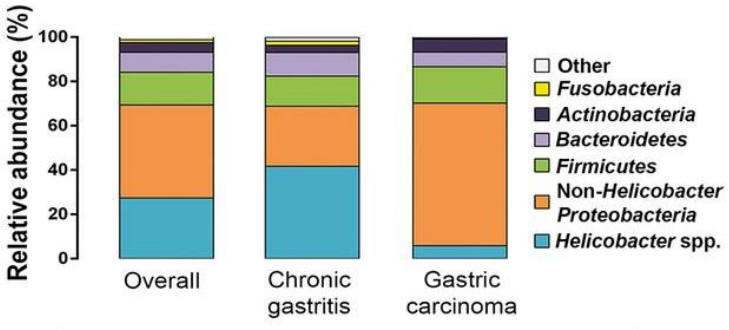
Figure 8. The gastric microbiota profile differs in chronic gastritis and gastric carcinoma. (A) Shannon index of diversity in patients with chronic gastritis and gastric carcinoma. **(B)** Good's estimator of coverage, measuring the proportion of total bacterial species represented in samples of each group of patients. Principal coordinate analysis (PCoA) plots of **(C)** unweighted and **(D)** weighted UniFrac distances in which samples were coloured by clinical outcome. The percentage of diversity captured by each coordinate is shown. ANOSIM, analysis of similarity.

Overall, the gastric microbiota was dominated by five phyla: *Proteobacteria* (69.3%), *Firmicutes* (14.7%), *Bacteroidetes* (9.0%), *Actinobacteria* (4.3%) and *Fusobacteria* (1.3%). Although these phyla were present in the two patient groups in the same order of relative

abundance, the gastric carcinoma microbiota had an over-representation of *Actinobacteria* ($P < 0.001$) and *Firmicutes* ($P = 0.040$), and lower abundance of *Bacteroidetes* ($P = 0.003$) and *Fusobacteria* ($P < 0.001$; **Figure 9**).

When reads assigned to *Proteobacteria* into *Helicobacter* spp. and non-*Helicobacter Proteobacteria* were separated, a significant reduction in the abundance of *Helicobacter* ($P < 0.001$) and an over-representation of non-*Helicobacter Proteobacteria* were detected in gastric carcinoma ($P < 0.001$; **Figure 9**).

These results show that for high taxonomic levels the stomach microbial communities differ in chronic gastritis and gastric carcinoma, suggesting that major changes also occur at lower taxonomic levels.



Taxa	Chronic gastritis (%)	Gastric carcinoma (%)	*P* value
Proteobacteria	68.8	70.2	NS
Helicobacter spp.	41.7	5.9	< 0.001
Non-Helicobacter Proteobacteria	27.1	64.3	< 0.001
Firmicutes	13.6	16.4	0.040
Bacteroidetes	10.6	6.6	0.003
Actinobacteria	3.3	5.9	< 0.001
Fusobacteria	1.8	0.5	< 0.001

Figure 9. Relative abundance of phyla in all subjects and in each group of patients.

1.3. Specific microbiota differentially abundant in chronic gastritis and gastric carcinoma

To identify the most relevant taxa responsible for the differences between clinical diagnoses, we conducted LEfSe analysis (Segata *et al.* 2011). This analysis identified 29 taxa, including 10 genera, which were differentially abundant in the two patient groups **(Figure 10)**. In gastric carcinoma, an enrichment in *Proteobacteria* taxa was observed, including the genera *Phyllobacterium* and *Achromobacter* and the families *Xanthomonadaceae* and *Enterobacteriaceae.* Although no specific genus could be identified within the *Xanthomonadaceae*, in the *Enterobacteriaceae*, the genus *Citrobacter* was identified as being significantly enriched in gastric carcinoma. Additionally, *Lactobacillus*, *Clostridium* and *Rhodococcus* were also significantly more abundant in

gastric carcinoma. *Helicobacter*, *Neisseria*, *Prevotella* and *Streptococcus* were most abundant in the microbiota of patients with chronic gastritis **(Figure 10)**.

A. **B.**

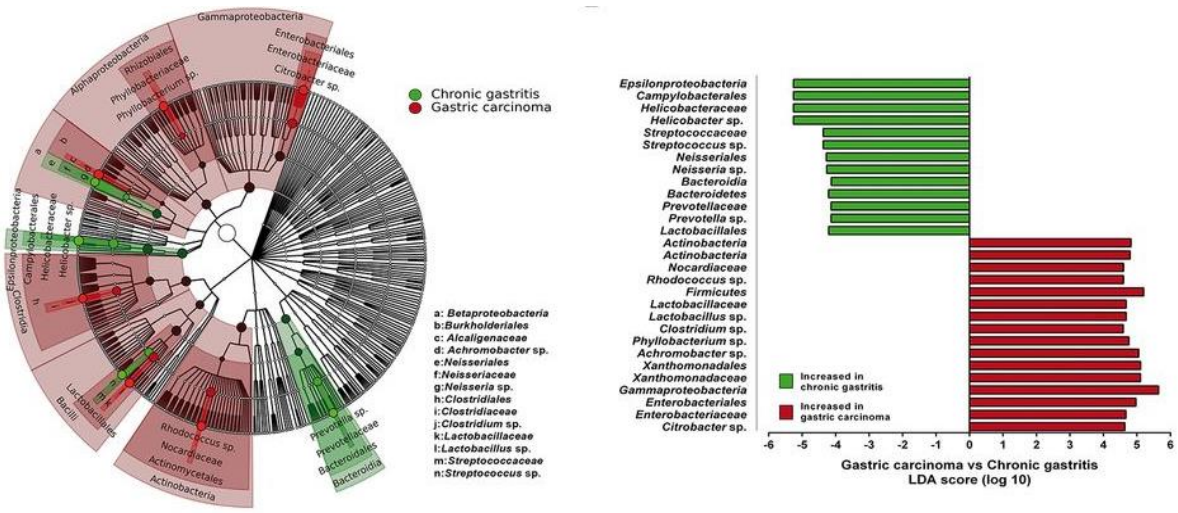


Figure 10. Microbial taxa associated with chronic gastritis and gastric carcinoma. (A) Cladogram representation of the gastric microbiota taxa associated with chronic gastritis and gastric carcinoma. **(B)** Association of specific microbiota taxa with the group of chronic gastritis and gastric carcinoma by linear discriminant analysis (LDA) effect size (LEfSe). Green indicates taxa enriched in chronic gastritis group and red indicates taxa enriched in gastric carcinoma group.

1.4. Validation of specific genera abundance in chronic gastritis and gastric carcinoma with qPCR

To demonstrate that our data were not biased by the microbiota profiling pipeline used, LEfSe results were validated by qPCR in the Portuguese discovery cohort. For that, two different sets of qPCR assays were used, a universal assay to quantify the total bacterial load, and a specific assay to quantify the 10 genera identified by LEfSe analysis.

The relative standard curve method was used to quantify the specific genera in patients with chronic gastritis and with gastric carcinoma. All the assays presented PCR efficiencies above 80% and correlation with targeted DNA concentration (correlation coefficient) above 98% **(Figure 11)**.

The qPCR confirmed significant decreases in the abundance of *Helicobacter* and *Neisseria*, and significant increases of *Achromobacter*, *Citrobacter*, *Phyllobacterium*, *Clostridium*, *Rhodococcus* and *Lactobacillus* in gastric carcinoma in comparison with chronic gastritis **(Figure 12A)**.

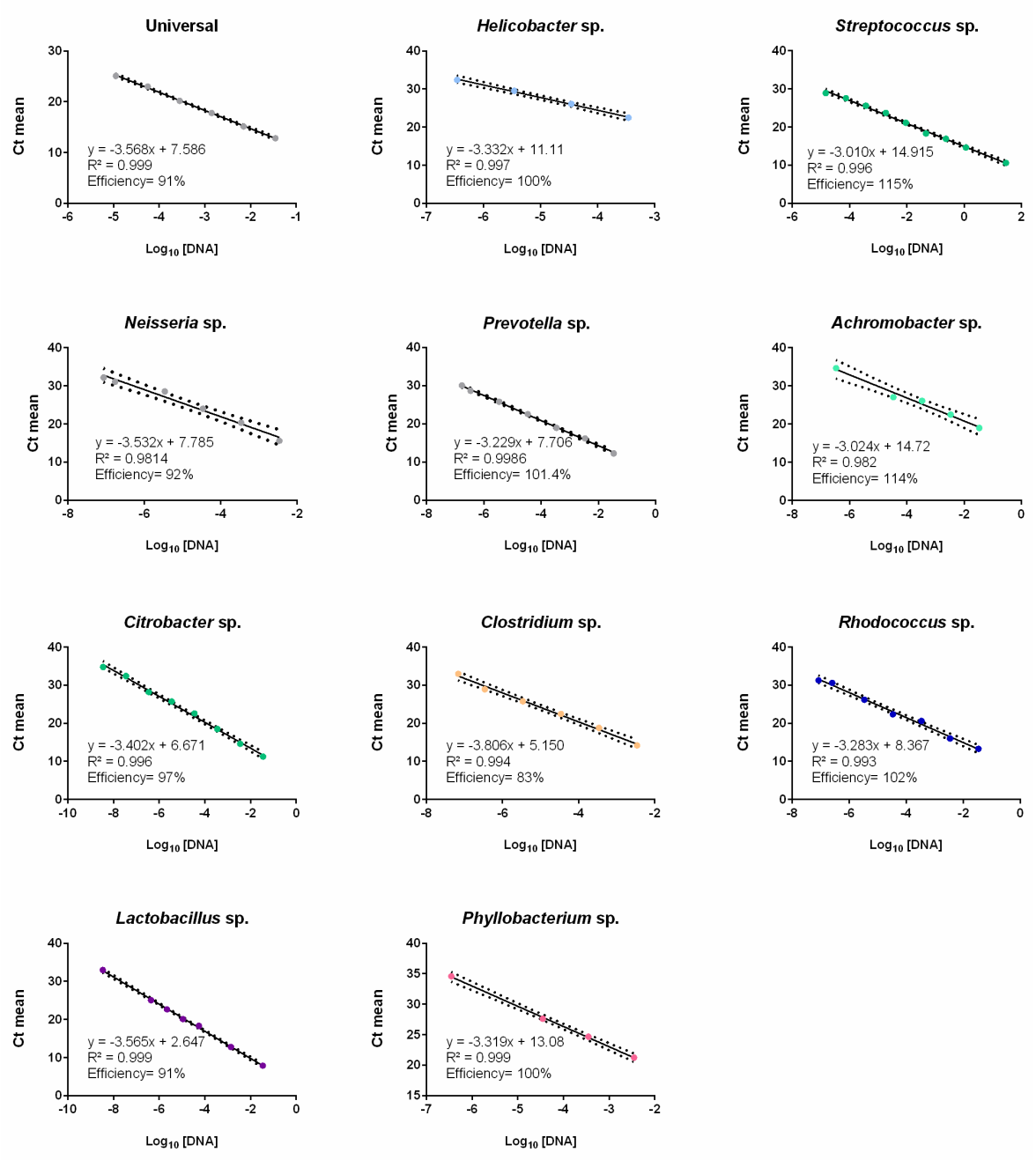


Figure 11. Standard curves for universal and genus-specific qPCR assays. The standard curves were created by cloning amplicons of each designed assay into pGEM-T easy vector system. The standard curves correlate the $\log_{10}$ of each known concentration with the threshold cycle (Ct) value. The DNA concentration of the target bacterial DNA was extrapolated using the standard curve equation (Ct mean = slope x $\log_{10}$[DNA]+ b). The correlation coefficient (R^2 value) was calculated, and the PCR efficiencies for each assay were determined as 10(-1/slope)-1. Dash lines represent the 95% confidence interval to contain the best-fit regression line.

To further confirm these findings, we have additionally used a validation cohort from Portugal, composed by 15 patients with chronic gastritis and 23 patients with gastric carcinoma. With the exception of *Prevotella* and *Streptococcus*, we were able to confirm the alterations in the abundance of the eight genera as identified by the original LEfSe analysis **(Figure 12B)**.

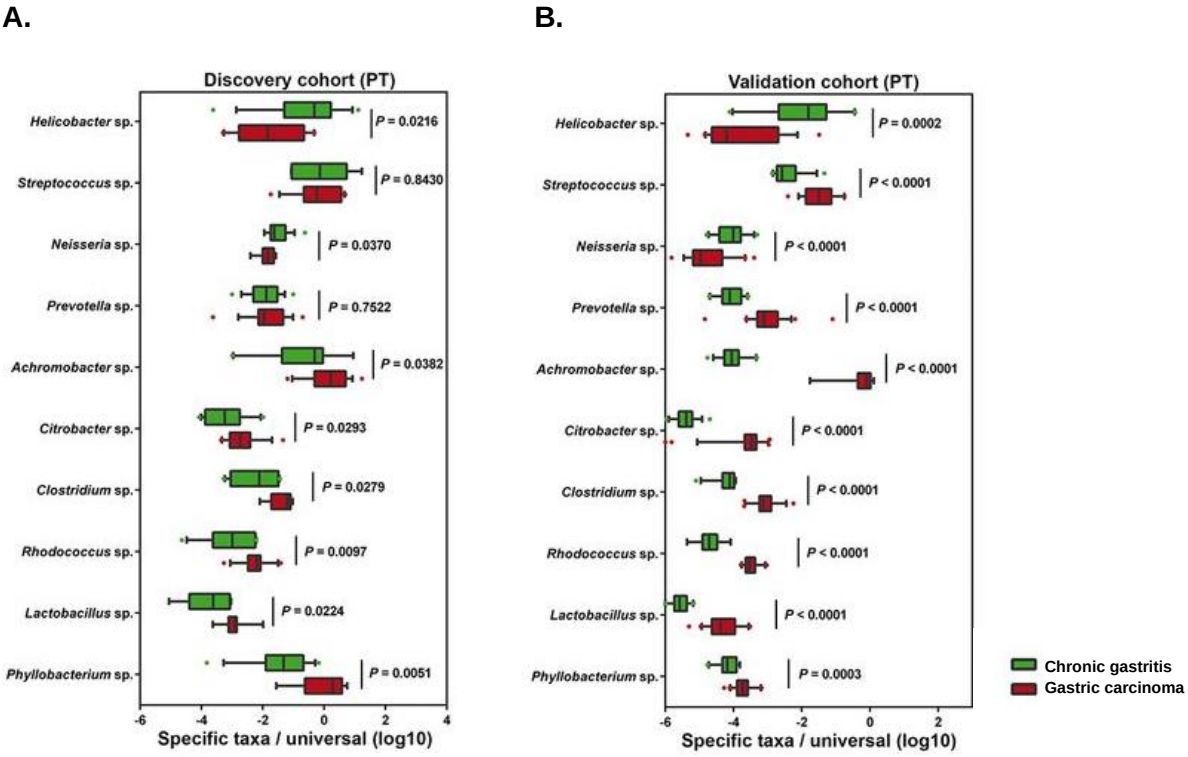


Figure 12. Validation of specific genera abundance in chronic gastritis and gastric carcinoma. Validation of LEfSe results by qPCR of the 10 genera differentially enriched in the discovery cohort **(A)** and in the Portuguese validation cohort **(B)**. Significance was obtained by Student's t-test.

1.5. Quantification of microbial dysbiosis in chronic gastritis and gastric carcinoma

We next combined the 10 most relevant taxa that characterized each group of patients and calculated the MDI (Gevers *et al.* 2014). The gastric microbiota of patients with gastric carcinoma had a higher MDI than that of patients with chronic gastritis both in the discovery cohort ($P < 0.0001$; **Figure 13A**) and in the Portuguese validation cohort as assessed using qPCR ($P < 0.0001$; **Figure 13B**). These results demonstrate that the gastric carcinoma microbiota has a high degree of dysbiosis.

We also evaluated whether the MDI could be used to discriminate between chronic gastritis and gastric carcinoma. In ROC analysis, the MDI showed excellent performance in

identifying gastric carcinoma, yielding an area under the curve (AUC) of 0.91 and 0.89 for the Portuguese discovery and validation cohorts, respectively **(Figure 13C and 13D)**.

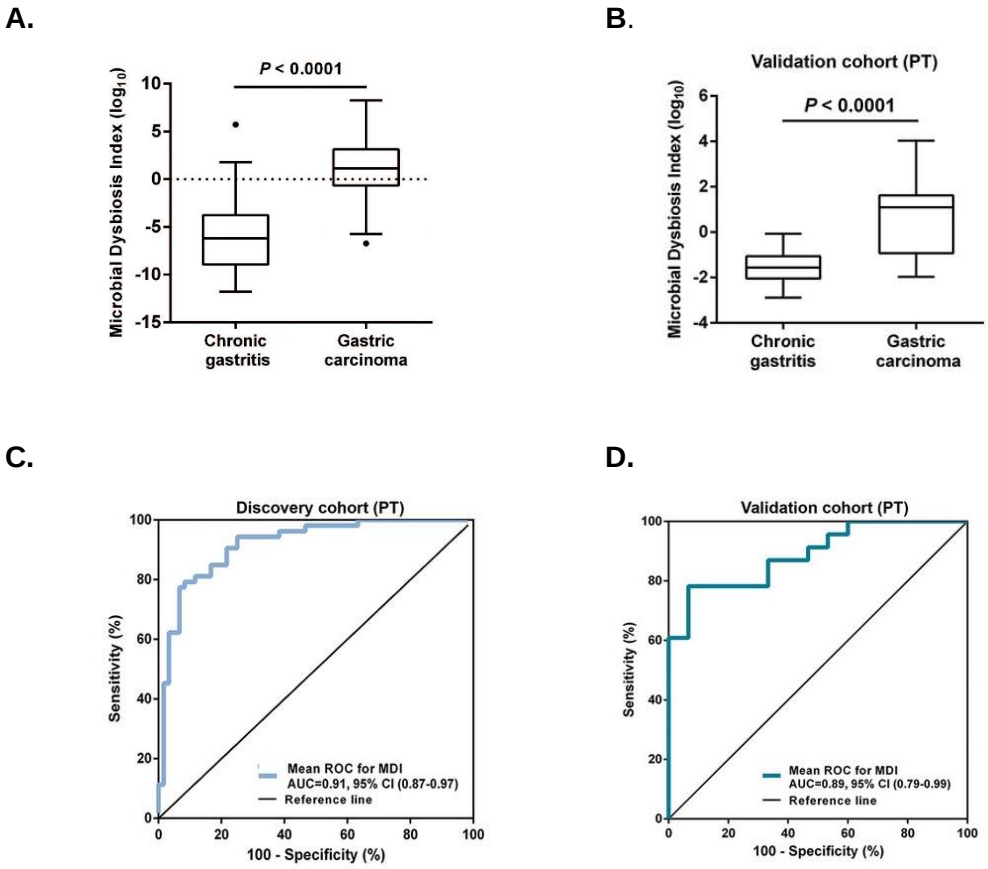


Figure 13. Quantification of microbial dysbiosis in chronic gastritis and gastric carcinoma. (A, B) Box plot showing the MDI in the discovery cohort **(A)** and in the Portuguese validation cohort **(B)**. Significance was obtained by Student's t-test. **(C, D)** ROC curves analysis to evaluate the discriminatory potential of MDI in gastric carcinoma detection in the discovery cohort **(C)** and in the Portuguese validation cohort **(D).** AUC, area under the curve; MDI, microbial dysbiosis index; ROC, receiver operating characteristic.

PART II

Establishment of a whole metagenome sequencing strategy to characterize the gastric mucosa-associated microbiome

PART II. Establishment of a whole metagenome sequencing strategy to characterize the gastric mucosa-associated microbiome

WMS has proved effective in the taxonomic and functional characterization of complex microbial communities residing in different human body sites (Lloyd-Price *et al.* 2017). Hence, we aimed to establish a WMS strategy to study the gastric mucosa-associated microbiome. For that, we started by setting up a pipeline of analysis for WMS data for microbiome profiling, using two synthetic DNA mock microbial communities with a well-defined composition. Since the level of host DNA remains a major challenge in WMS, next we evaluated the impact of sample complexity and sequencing depth on the taxonomic resolution of WMS. Finally, we applied this optimized method of analysis to metagenomes of human gastric carcinoma specimens.

1. Establishment of a pipeline of analysis for WMS data using mock communities

To establish, optimize, and validate a pipeline of analysis for WMS data for microbiome profiling, two mock DNA microbial communities (Mock Even and Staggered), used as reference samples in the HMP, were sequenced on the Illumina NextSeq 550.

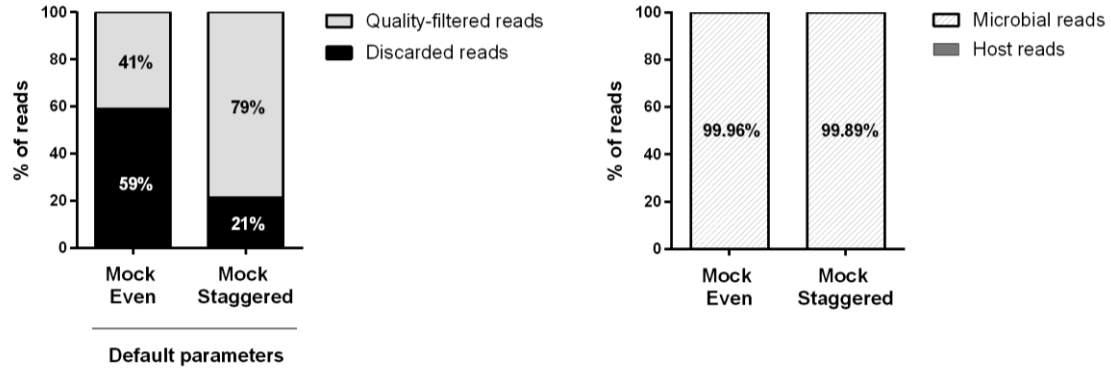
1.1. Optimization of sequencing data pre-processing

Initially, the metagenomes from the mock communities underwent a pre-processing step performed with KneadData.

The two datasets yielded a large number of raw single-end reads (more than 50 million). After sequencing data pre-processing (quality filtering and host sequences decontamination), the number of reads differed considerably between samples, being higher in mock Staggered (65.5 million reads) than in mock Even (26.8 million reads) **(Figure 14)**.

Of the total raw single-end reads, the proportion of discarded reads during quality filtering, with the default parameters, was higher in the mock Even (59%) than in the mock Staggered (21%) **(Figure 14A)**. Moreover, the percentage of reads removed during host sequences decontamination step was very low in both mock datasets (<0.15%), which is consistent with the absence of human DNA in these samples **(Figure 14B)**. Altogether, this data suggests that the lower number of pre-processed reads in the mock Even compared with the mock Staggered was due to high number of reads dropped during quality filtering rather than to host DNA sequences removed.

A. **B.**



Sample	Total No. of raw single-end reads	Total No. of quality-filtered reads	Total No. of quality-filtered and host decontaminated reads
Mock Even	65.398.644	26.805.292	26.797.216
Mock Staggered	83.356.856	65.562.313	65.493.026

Figure 14. Overview of the sequencing data pre-processing of mock communities metagenomes with the default quality filtering parameters (Sliding window of 4:20; minimum read length: 50% of total input read length). (A) The proportion of quality-filtered and discarded reads from the total raw single-end reads per mock community. **(B)** The percentage of microbial and host reads from the total quality-filtered reads per mock community.

To better understand the difference in the number of discarded reads during quality filtering across samples, the quality and minimum length of the reads were evaluated before and after sequencing data pre-processing. Both raw datasets showed an overall good quality, having 91% of the reads with average quality ≥ Q30, whereas the minimum length of the reads, after sequencing data pre-processing, was higher in the mock Even (105 bp) than in the mock Staggered (51 bp). These results indicate that the higher proportion of reads dropped during quality filtering in mock Even was associated with the minimum length of the reads parameter established by KneadData.

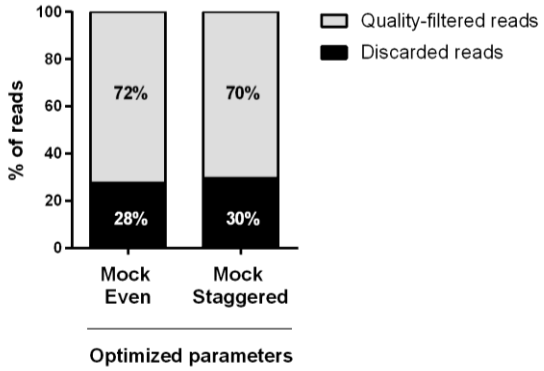
To increase the number of pre-processed reads in the mock Even, the quality filtering step was optimized by establishing a fixed minimum read length of 60 bp, instead of the default that removed the read when 50% of the total input read length was trimmed.

After sequencing data pre-processing with the optimized quality filtered parameters, the number of reads yielded was equivalent between the two mock communities **(Figure 15)**.

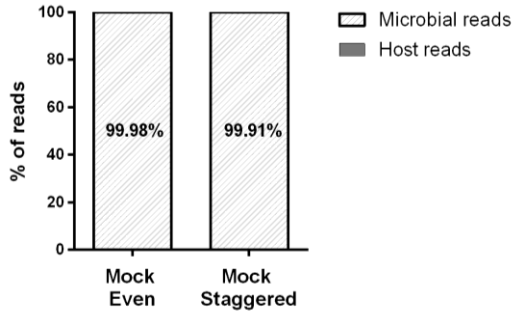
Of the total raw single-end reads, the proportion of discarded reads during quality filtering became similar between mock Even and Staggered (between 28% and 30%) **(Figure 15A)**,

and the proportion of reads removed as host DNA sequences from both datasets was still very low (<0.1%) **(Figure 15B)**. Moreover, the pre-processed datasets with the optimized quality filtering parameter showed an overall high quality, with 99% of the reads having an average quality ≥ Q30, and a minimum length of the reads high enough for the taxonomic analysis. Therefore, the optimized quality filtering parameters were used for further analyses.

A.



B.



Sample	Total No. of raw single-end reads	Total No. of quality-filtered reads	Total No. of quality-filtered and host decontaminated reads
Mock Even	65.398.644	47.348.084	47.337.044
Mock Staggered	83.356.856	58.699.734	58.644.547

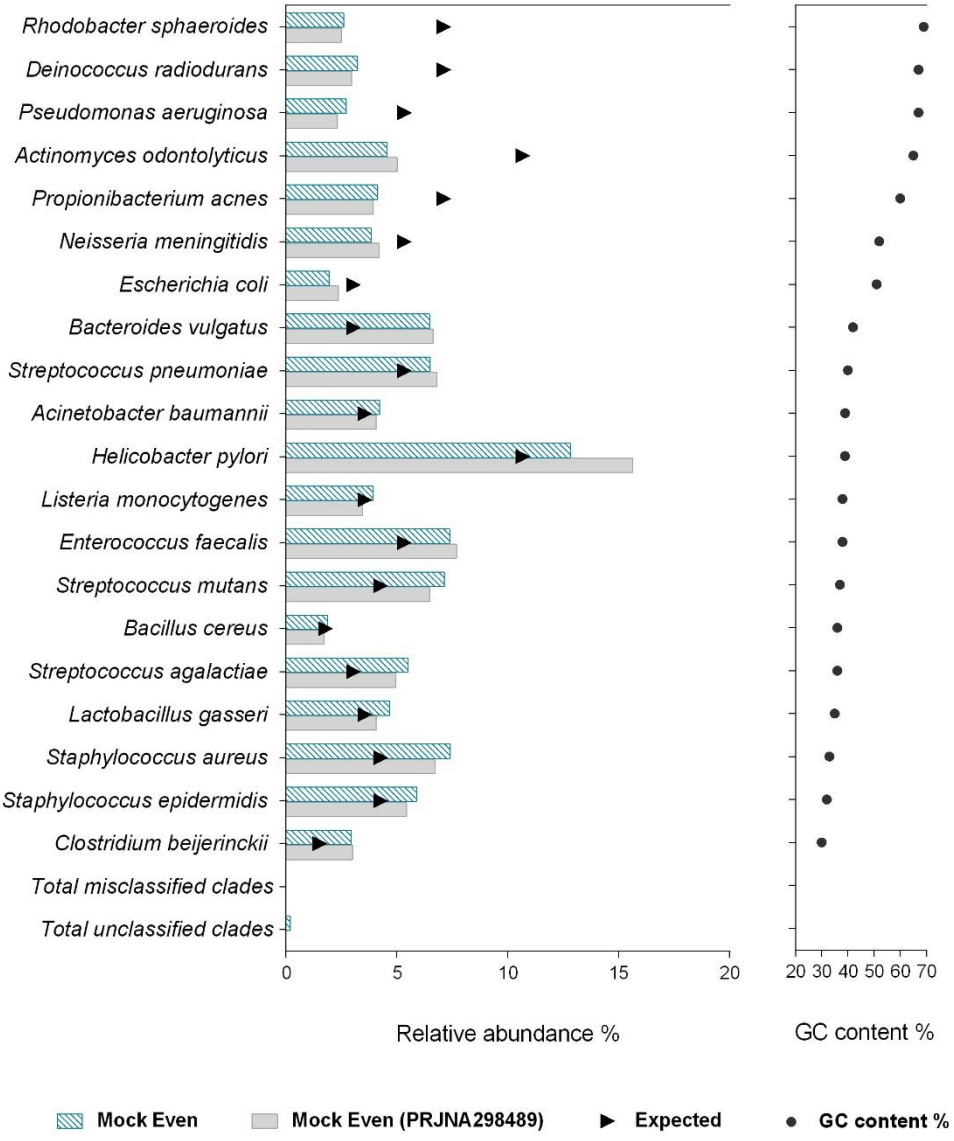
Figure 15. Overview of the sequencing data pre-processing of mock communities metagenomes with the optimized quality filtering parameters (Sliding window of 4:20; minimum read length: 60 bp). (A) The proportion of quality-filtered and discarded reads from the total raw single-end reads per mock community. **(B)** The percentage of microbial and host reads from the total quality-filtered reads per mock community

1.2. Reconstitution of the taxonomic profile of mock communities

After sequencing data pre-processing with the optimized quality filtering parameters, the taxonomic profile of the two mock communities was determined with MetaPhlAn2. The relative abundance of each taxa was then evaluated at species-level and represented in a bar plot **(Figure 16)**.

In the two mock communities, all 20 species of bacteria were successfully identified **(Figure 16A and 16B**; blue and orange bar) with a similar taxonomic profile compared to that of the expected, calculated based on the theoretical number of genome copies (**Figure 16A and 16B**; black triangle).

A.



B.

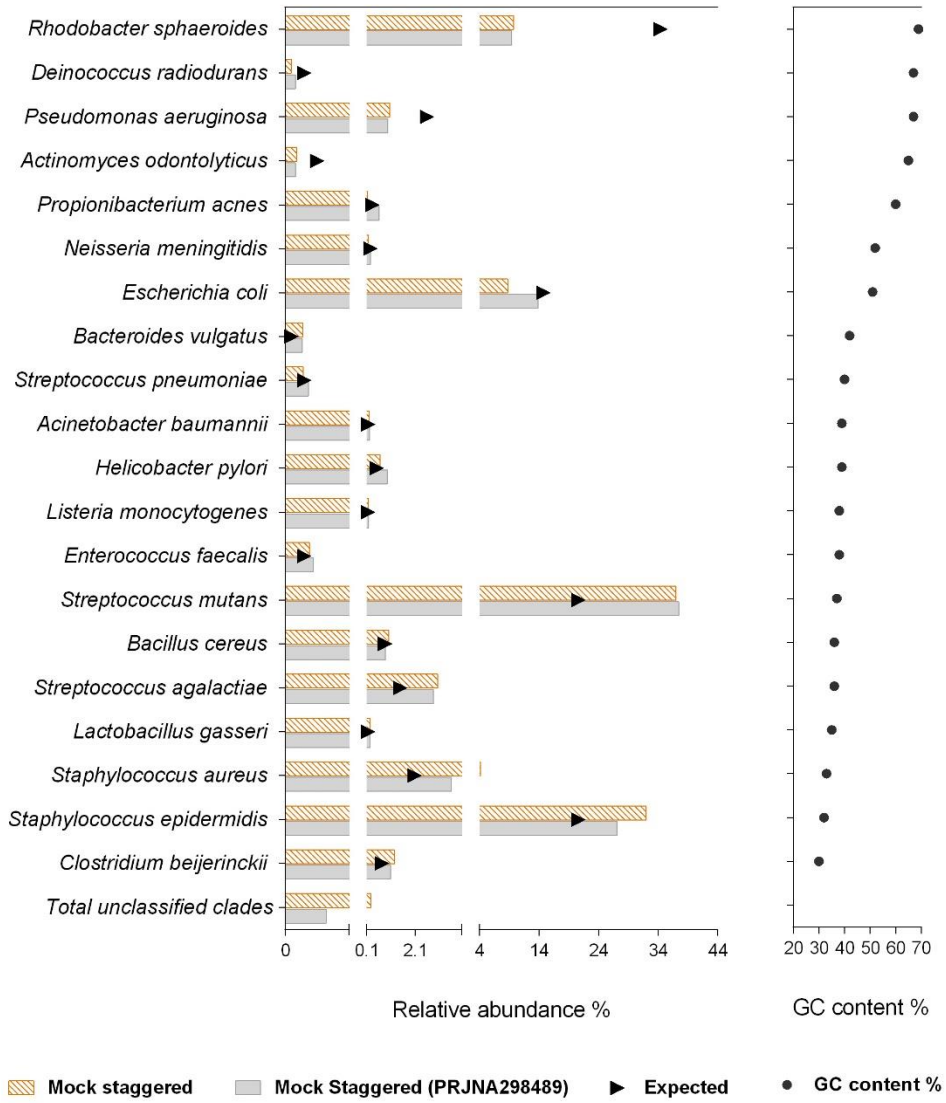


Figure 16. Reconstitution of the taxonomic profile of mock microbial communities from WMS reads. Taxonomic profile of the metagenomes of the mock Even **(A)** and mock Staggered **(B)** estimated with MetaPhlAn2, after the pre-processing step with the optimized quality filtering parameters, and expressed as relative abundance of species (blue and orange bars). The expected taxonomic results (black triangle) were estimated based on the theoretical number of species genome copies present in the mocks. The taxonomic profile of published WMS datasets (PRJNA298489) of both mocks (grey bar) was also determined using MetaPhlAn2. The genomic GC content (%) of the species included in the mock communities were obtained from the NCBI Genome Database (grey dot). Species were sorted from the highest to the lowest GC content.

With the aim of better evaluating differences in these profiles, we determined the ratio of observed to expected relative abundances, and considered underrepresentation or overrepresentation of species when there was a ≥ 2-fold change between the two conditions **(Table 4 and 5)**. In the two mock communities, the relative abundances of eight species

was over- or underestimated. In the species *Rhodobacter sphaeroides*, *Deinococcus radiodurans*, *Pseudomonas aeruginosa, Actinomyces odontolyticus* and *Propionibacterium acnes,* there was underestimation of the relative abundances, whereas in the species *Bacteroides vulgatus*, *Streptococcus agalactiae* and *Staphylococcus aureus,* there was overestimation of the relative abundances **(Table 4 and 5)**.

Table 4. Composition of the mock microbial community B Even (HM-276D, High concentration, v5.1H), showing the expected and the observed relative abundance of species after WMS. The ratio of observed to expected relative abundances of species is also shown. Orange shading indicates a ≥ 2-fold change.

Microbial species	GC content (%)	Expected relative abundance (%)	Observed relative abundance (%)	Ratio observed/ expected
Rhodobacter sphaeroides	69	7.12	2.60	0.36
Deinococcus radiodurans	67	7.12	3.22	0.45
Pseudomonas aeruginosa	67	5.34	2.71	0.51
Actinomyces odontolyticus	65	10.68	4.55	0.43
Propionibacterium acnes	60	7.12	4.14	0.58
Neisseria meningitidis	52	5.34	3.83	0.72
Escherichia coli	51	3.05	1.95	0.64
Bacteroides vulgatus	42	3.05	6.49	2.13
Streptococcus pneumoniae	40	5.34	6.51	1.22
Acinetobacter baumannii	39	3.56	4.23	1.19
Helicobacter pylori	39	10.68	12.83	1.20
Listeria monocytogenes	38	3.56	3.93	1.10
Enterococcus faecalis	38	5.34	7.39	1.38
Streptococcus mutans	37	4.27	7.14	1.67
Bacillus cereus	36	1.78	1.87	1.05
Streptococcus agalactiae	36	3.05	5.50	1.80
Lactobacillus gasseri	35	3.56	4.67	1.31
Staphylococcus aureus	33	4.27	7.42	1.74
Staphylococcus epidermidis	32	4.27	5.90	1.38
Clostridium beijerinckii	30	1.53	2.93	1.92
Total misclassified clades	NA	NA	0.007	NA
Total unclassified clades	NA	NA	0.19	NA

Table 5. Composition of the mock microbial community B Staggered (HM-277D, High concentration, v5.2H), showing the expected and the observed relative abundance of species after WMS. The ratio of observed to expected relative abundances of species is also shown. Orange shading indicates a ≥ 2-fold change.

Microbial species	GC content (%)	Expected relative abundance (%)	Observed relative abundance (%)	Ratio observed/ expected
Rhodobacter sphaeroides	69	34.4	9.78	0.28
Deinococcus radiodurans	67	0.03	0.01	0.28
Pseudomonas aeruginosa	67	2.58	1.06	0.41
Actinomyces odontolyticus	65	0.05	0.02	0.35
Propionibacterium acnes	60	0.34	0.15	0.45
Neisseria meningitidis	52	0.26	0.16	0.61
Escherichia coli	51	14.8	8.77	0.59
Bacteroides vulgatus	42	0.01	0.03	1.86
Streptococcus pneumoniae	40	0.03	0.03	1.08
Acinetobacter baumannii	39	0.17	0.20	1.16
Helicobacter pylori	39	0.52	0.65	1.27
Listeria monocytogenes	38	0.17	0.17	1.00
Enterococcus faecalis	38	0.03	0.04	1.47
Streptococcus mutans	37	20.7	37.0	1.79
Bacillus cereus	36	0.86	1.01	1.17
Streptococcus agalactiae	36	1.48	3.02	2.05
Lactobacillus gasseri	35	0.17	0.26	1.48
Staphylococcus aureus	33	2.07	4.16	2.02
Staphylococcus epidermidis	32	20.7	32.0	1.55
Clostridium beijerinckii	30	0.74	1.25	1.69
Total unclassified clades	NA	NA	0.28	NA

Rhodobacter sphaeroides, Deinococcus radiodurans, Pseudomonas aeruginosa, Actinomyces odontolyticus and Propionibacterium acnes compared with *Bacteroides vulgatus*, *Streptococcus agalactiae* and *Staphylococcus aureus* contain very different levels of GC content in their genomes (≥ 60% and ≤ 42%, respectively). Since it has been shown that the GC content introduces a bias during Illumina library preparation and sequencing (Jones *et al.* 2015), this potential bias was evaluated in the both mocks datasets. Pearson's correlation analysis showed that the GC content was negatively correlated with the ratio of observed to expected relative abundances in mock Even (**Figure 17A**; r = -0.8461, $P < 0.0001$) and in the mock Staggered (**Figure 17B**; r = -0.8811, $P < 0.0001$). Accordingly,

when a species had a higher genomic % GC content, the relative abundance observed after sequencing was lower than the expected. Therefore, this suggests that under- and overestimated relative abundances of species were due to this bias.

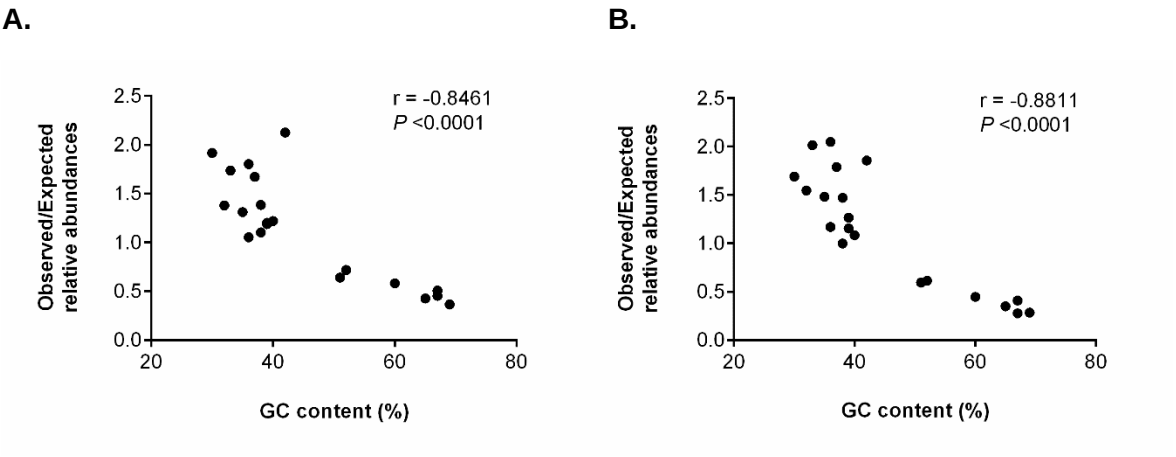


Figure 17. Pearson's correlation between genomic GC content and ratio of observed to expected relative abundances for the mock Even (A) and mock Staggered (B). Pearson's correlation coefficient (r) and *P*-value for the association are shown in the graphs.

In both mock datasets, unclassified clades were identified at low relative abundances (<0.3%), which likely represent bacterial species from the mocks that could not be assigned at the species level. Additionally, in mock Even, misclassified species with very low relative abundance (<0.01%) were detected.

To further validate our metagenomics workflow, we used published WMS data from both mock microbial communities (Jones *et al.* 2015), and evaluated the taxonomic profile with MetaPhlAn2 to compare with that of our metagenomics datasets (**Figure 16A and 16B**; grey bar). In both published mock WMS datasets, all 20 species of bacteria were detected with similar relative abundances to those of the mock Even and Staggered sequenced in our lab. Also unclassified and misclassified clades were identified in a very low relative abundance (<0.07%). Therefore, this data suggests that our sequencing protocol had no influence in the outcome results.

Overall, these results demonstrate that the taxonomic profile of both mock microbial communities was accurately reconstituted from WMS single end-reads using the optimized pipeline of analysis.

2. Evaluation of the impact of host DNA and sequencing depth on the taxonomic resolution of WMS for microbiome analysis

2.1. Generation of synthetic samples and pre-processing of sequencing data

To assess the influence of host DNA on the sensitivity of WMS for taxonomic profiling of the microbiome, three synthetic samples with distinct host:bacteria DNA ratios were generated to contain 10%, 90%, and 99% host DNA (SS10, SS90, and SS99, respectively). As control, the mock microbial community DNA sample (100% bacterial DNA; MS) was used **(Figure 18).** WMS was applied to all four samples, and the resulting metagenomes underwent a pre-processing step performed with KneadData.

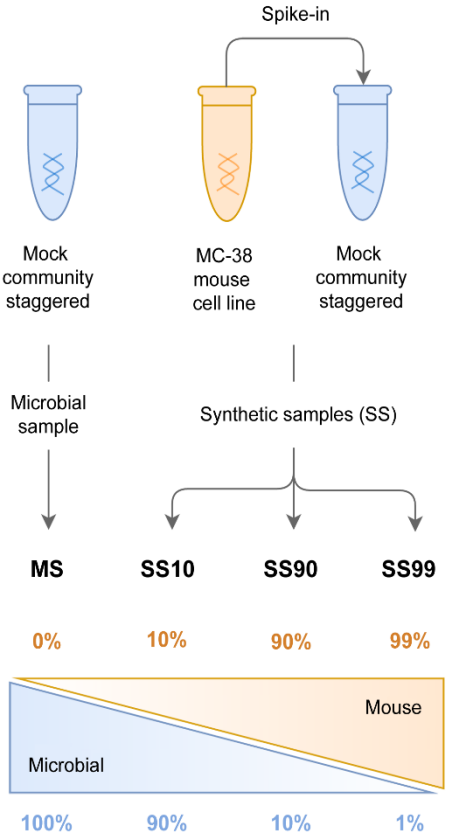
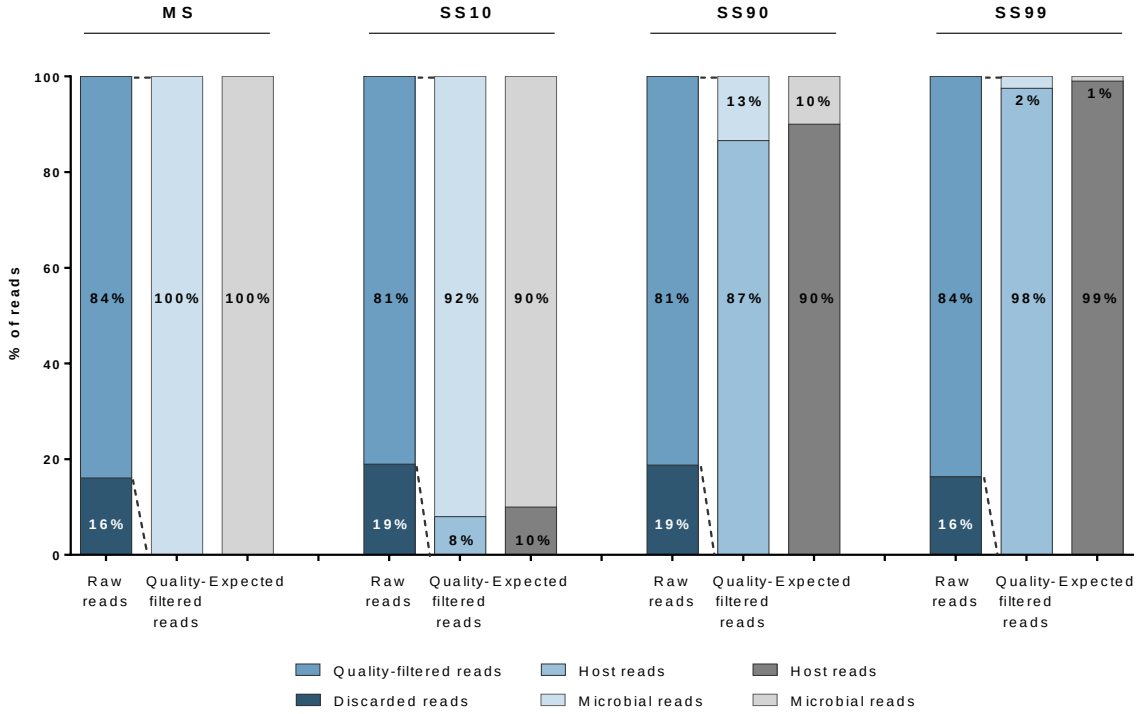


Figure 18. Schematic representation of the experimental design to generate synthetic samples. DNA samples from a mock microbial community staggered (HM-277D) were spiked with DNA from a mouse cell line (MC-38 cells), generating three synthetic samples containing 10%, 90%, and 99% host DNA (SS10, SS90, and SS99, respectively). DNA from the mock microbial community was used as control (MS).

The four datasets yielded a large number of raw single-end reads ranging from 35 to 51 million. After sequencing data pre-processing, the number of reads differed considerably between samples, being higher in MS (33.3 million reads) and SS10 (29.9 million reads) in comparison with SS90 (5.5 million reads) and SS99 (7.5 hundred thousand reads). The relative low number of pre-processed reads in SS90 and SS99 samples was due to high number of host DNA sequences removed rather than to reads dropped during quality filtering.

Of the total raw single-end reads, the proportion of discarded reads during quality filtering was similar between all samples (ranging from 16% to 19%) **(Figure 19A)**, which is consistent with the overall good quality of all raw datasets (between 80% and 90% of the reads with average quality ≥ Q30; **Figure 19B**). These results confirm that differences in the number of pre-processed reads across samples were associated with the host sequences decontamination step**.** Quality-filtered reads were comparable with the expected ratios of host to microbial DNA for each condition **(Figure 19A).** Overall, synthetic samples with the expected host to bacterial DNA ratios were successfully generated.

A.



B.

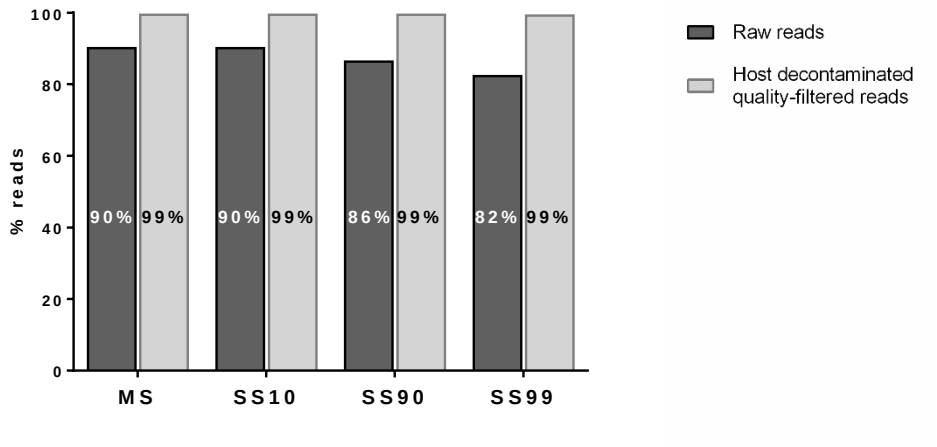


Figure 19. Overview of the sequencing data pre-processing from synthetic samples metagenomes. (A) The proportion of quality-filtered, discarded, microbial and host reads per sample. The *raw reads bar* represents the percentage of quality-filtered and discarded reads from the total raw single-end reads per sample. The *quality-filtered reads* bar constitutes the fraction of host and microbial reads from the total quality-filtered reads per sample. The *expected bar* consists in the theoretical percentages of host and microbial reads expected in each synthetic sample. **(B)** Percentage of reads with an average quality of ≥ Q30 in the MS, SS10, SS90 and SS99, in the raw and host decontaminated quality-filtered data.

2.2. Effect of the level of host DNA on the sensitivity of WMS for microbiome taxonomic profiling

After sequencing data pre-processing, the taxonomic profile of all samples was determined with MetaPhlAn2, with the aim to evaluate the effect of host DNA on the sensitivity of WMS for microbiome profiling. For that, bacteria species were grouped into the following categories, according to the number of 16S rRNA copies in the mock community: very low (10^4), low (10^5), high (10^6), and very high (10^7) abundant. The relative abundance of each taxa was then quantified at species-level and represented in a heat map **(Figure 20)**.

In the control MS, all 20 species of bacteria were successfully identified with a similar taxonomic profile compared to that of the expected, calculated based on the theoretical number of genome copies **(Figure 20)**.

The microbial profile of SS10 was comparable to the MS control, since all 20 bacterial species were detected with similar relative abundances to those of the MS **(Figure 20 and Table 6).** In SS90, however, there was a decrease in the ability to detect very low abundant species. Specifically, *Deinococcus radiodurans* could not be identified **(Figure 20)**, and the relative abundances of *Actinomyces odontolyticus*, *Enterococcus faecalis*, and *Bacteroides vulgatus* were underestimated **(Table 6).** The reduction in sensitivity was more striking in

SS99, where only two of the low and none of the very low abundant species were identified **(Figure 20).**

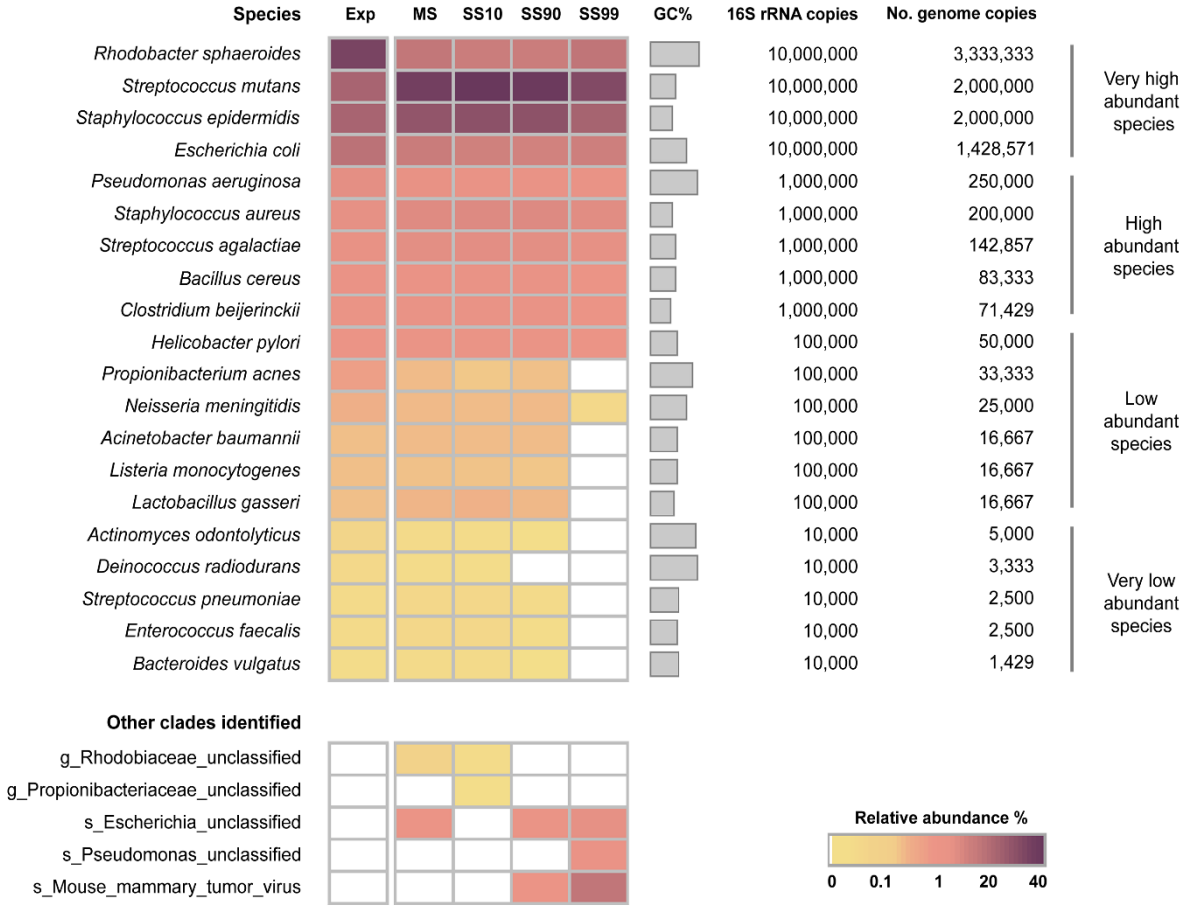


Figure 20. Effect of the levels of host DNA on the sensitivity of WMS for microbiome taxonomic profiling. Taxonomic profile of the metagenomes of the synthetic samples estimated with MetaPhlAn2, and expressed as relative abundance of species in a heat map. The expected (Exp) taxonomic results were estimated based on the theoretical number of species genome copies present in the mock. Species were sorted from the highest to the lowest expected relative abundances.

In all conditions, unclassified clades were identified at low relative abundances (<2%), which likely represent bacterial species from the mock microbial community that were identified only at the genus or family level. Also, in synthetic samples with the highest amount of host DNA (SS90 and SS99), a mouse mammary tumor virus was identified **(Figure 20)**. Since viruses are not included in the mock community, the virus was likely introduced in the generation of the synthetic samples by the spiking with DNA from the mouse cell line, which could have the virus integrated in its genome. Besides, viruses are not included in the database used by Bowtie2, and therefore viral sequences have not been filtered as host contaminant.

Results demonstrate that high ratios of host:bacterial DNA interfere with the sensitivity of WMS for taxonomic profiling. The increase in the proportion of host DNA leads to decreased sensitivity of WMS to detect very low and low abundant bacterial species.

Table 6. Ratio of relative abundances of species from synthetic samples to MS. Grey shading indicates a ≥ 2-fold change.

Microbial species	16S rRNA copies	MS	SS10/ MS	SS90/ MS	SS99/ MS
Rhodobacter sphaeroides	10.000.000	1	0.805	0.814	1.053
Streptococcus mutans	10.000.000	1	1.081	1.061	0.877
Staphylococcus epidermidis	10.000.000	1	1.069	1.053	0.776
Escherichia coli	10.000.000	1	0.816	0.774	0.886
Pseudomonas aeruginosa	1.000.000	1	0.742	0.798	0.714
Staphylococcus aureus	1.000.000	1	1.106	1.085	0.796
Streptococcus agalactiae	1.000.000	1	1.048	1.022	0.711
Bacillus cereus	1.000.000	1	1.021	1.237	0.574
Clostridium beijerinckii	1.000.000	1	1.064	1.070	0.589
Helicobacter pylori	100.000	1	0.911	0.907	0.561
Propionibacterium acnes	100.000	1	0.625	0.875	0
Neisseria meningitidis	100.000	1	0.952	1.011	0.169
Acinetobacter baumannii	100.000	1	0.967	0.962	0
Listeria monocytogenes	100.000	1	0.943	0.824	0
Lactobacillus gasseri	100.000	1	1.091	0.910	0
Actinomyces odontolyticus	10.000	1	0.812	0.435	0
Deinococcus radiodurans	10.000	1	0.626	0	0
Streptococcus pneumoniae	10.000	1	1.006	0.608	0
Enterococcus faecalis	10.000	1	1.065	0.310	0
Bacteroides vulgatus	10.000	1	0.935	0.181	0

2.3. Impact of sequencing depth on the sensitivity of WMS for microbiome taxonomic profiling

To assess the impact of sequencing depth on the sensitivity of WMS to detect bacterial species in samples with a high level of host DNA, reads from the SS90 metagenome were randomly subsampled, generating four datasets with reduced sequencing depths, corresponding to 50%, 25%, 10%, and 5% of the original dataset (SS90D50, SS90D25, SS90D10, and SS90D5, respectively). Their taxonomic profile was compared to that of the original SS90 dataset (SS90D100) **(Figure 21)**.

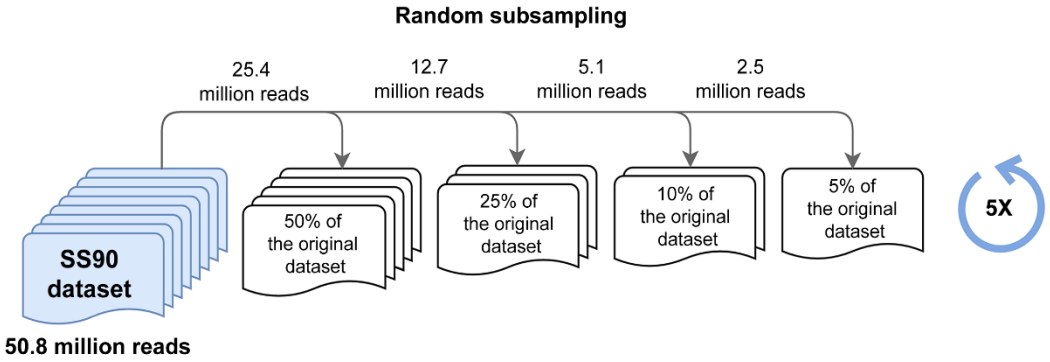


Figure 21. Schematic representation of the experimental design to generate random subsampling reads from the SS90 original dataset (90% host DNA). Random subsampling corresponding to 50%, 25%, 10%, and 5% of the reads from the original dataset (SS90D50, SS90D25, SS90D10, and SS90D5, respectively) were generated.

When the SS90 dataset was reduced to half of its original size (SS90D50), the number of very low abundant species that were not identified increased from one to three **(Figure 22)**, and the relative abundance of *E. faecalis* significantly decreased (P = 0.006; **Table 7 and 8)**. In SS90D25, none of the very low abundant species could be identified **(Figure 22)**. In comparison with the original dataset, no statistically significant differences were observed in the relative abundances of the remaining species **(Table 7 and 8)**. In SS90D10 and SS90D5, however, in addition to not identifying all very low abundant species, there were statistically significant decreases in the relative abundances of the low abundant species **(Figure 22 and Table 7 and 8)**.

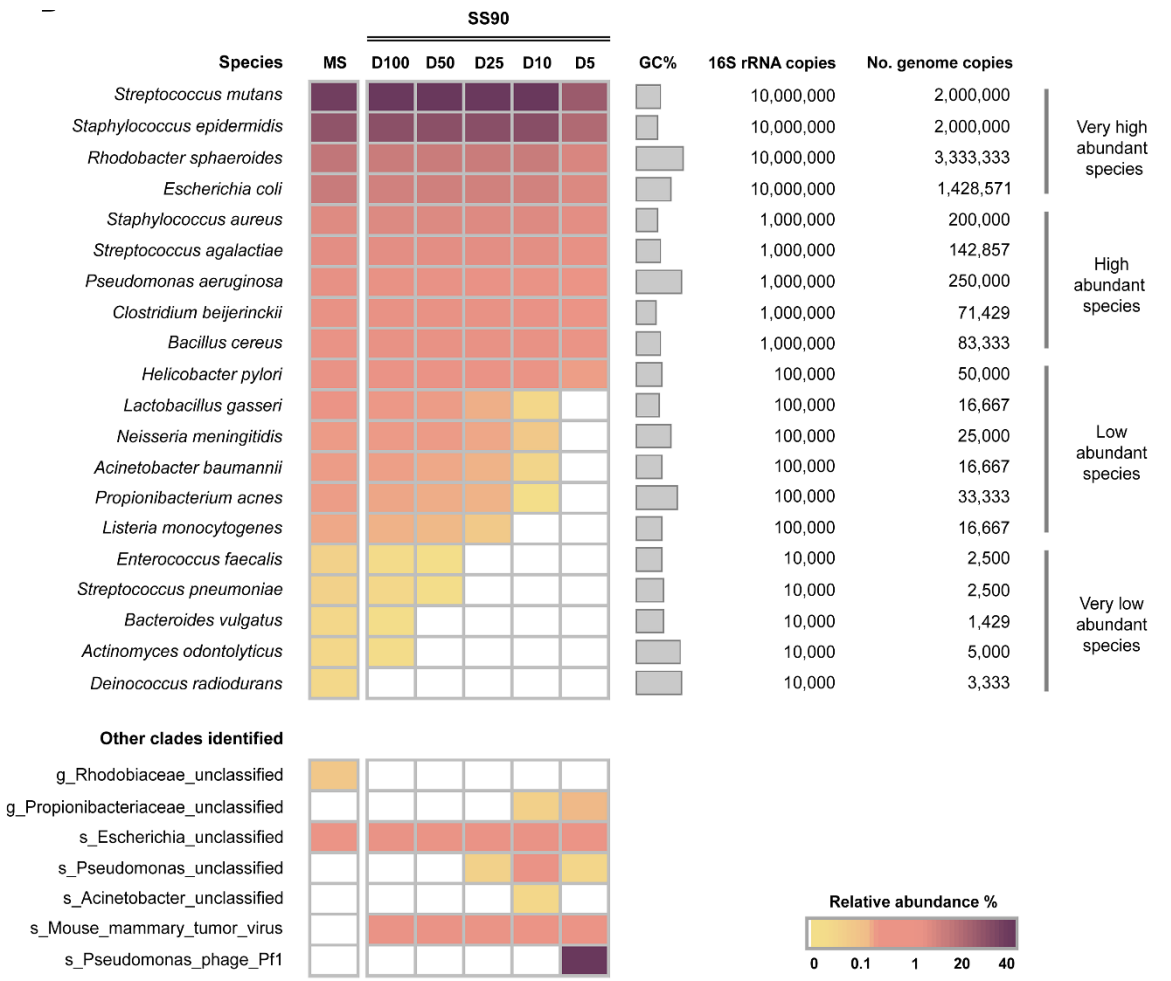


Figure 22. Impact of sequencing depth on the sensitivity of WMS for microbiome taxonomic profiling. Taxonomic profile of the generated datasets is represented as the average relative abundance from five independent experiments, and shown in a heat map. Data was sorted from the highest to the lowest relative abundances of species in the mock microbial community (MS).

The reduction of the dataset to 5% of its original size led to significantly lower relative abundances of the majority of high and very high abundant species **(Figure 22 and Table 7 and 8)**. In addition, a misclassified species (*Pseudomonas phage Pf1*) with a relative abundance of 40% was identified **(Figure 22)**. This likely constitutes an artifact originated by the reduction of the size of the dataset.

Overall, these results demonstrate that sequencing depth has a major impact on the sensitivity of WMS for taxonomic profiling of samples with 90% host DNA. When decreasing sequencing depth, the number of microbial species that are not detected increase, along with unclassified and misclassified clades.

Table 7. Ratio of relative abundances of species from each SS90 subset (SS90D50, SS90D25, SS90D10, SS90D5) to the SS90 original dataset (SS90D100). Random subsampling to generate each subset was performed in five independent experiments. Grey shading indicates a ≥ 2-fold change. nd: not detected in the SS90 original dataset.

Microbial Species	16S rRNA copies	SS90 D100	SS90D50/ SS90D100	SS90D25/ SS90D100	SS90D10/ SS90D100	SS90D5/ SS90D100
Streptococcus mutans	10.000.000	1	1.004	0.999	1.002	0.625
Staphylococcus epidermidis	10.000.000	1	1.004	1.023	1.037	0.621
Rhodobacter sphaeroides	10.000.000	1	0.995	0.995	1.019	0.585
Escherichia coli	10.000.000	1	1.029	1.009	0.961	0.622
Staphylococcus aureus	1.000.000	1	1.008	0.997	0.989	0.588
Streptococcus agalactiae	1.000.000	1	0.988	0.994	0.936	0.568
Pseudomonas aeruginosa	1.000.000	1	0.950	0.922	0.794	0.329
Clostridium beijerinckii	1.000.000	1	0.990	0.975	0.864	0.382
Bacillus cereus	1.000.000	1	1.000	0.950	0.971	0.465
Helicobacter pylori	100.000	1	0.957	0.964	0.776	0.257
Lactobacillus gasseri	100.000	1	0.949	0.677	0.105	0
Neisseria meningitidis	100.000	1	0.994	0.811	0.328	0
Acinetobacter baumannii	100.000	1	0.861	0.690	0.156	0
Propionibacterium acnes	100.000	1	0.871	0.772	0.011	0
Listeria monocytogenes	100.000	1	0.850	0.483	0	0
Enterococcus faecalis	10.000	1	0.090	0	0	0
Streptococcus pneumoniae	10.000	1	0.271	0	0	0
Bacteroides vulgatus	10.000	1	0	0	0	0
Actinomyces odontolyticus	10.000	1	0	0	0	0
Deinococcus radiodurans	10.000	1	nd	nd	nd	nd

Table 8. Statistical analysis (*P*-values) of the results presented in Table 7. The Kruskal-Wallis non-parametric test followed by multiple comparisons (SS90D50, SS90D25, SS90D10, or SS90D5) versus a control group (SS90D100) using Dunn's test was performed for each species. NA: not applicable, as it was not detected in the SS90 original dataset.

Microbial Species	16S rRNA copies	SS90D100 vs. SS90D50	SS90D100 vs. SS90D25	SS90D100 vs. SS90D10	SS90D100 vs. SS90D5
Streptococcus mutans	10.000.000	> 0.9999	> 0.9999	> 0.9999	0.0203
Staphylococcus epidermidis	10.000.000	> 0.9999	0.2548	0.0341	0.7829
Rhodobacter sphaeroides	10.000.000	> 0.9999	> 0.9999	> 0.9999	0.0102
Escherichia coli	10.000.000	0.9766	> 0.9999	> 0.9999	0.1242
Staphylococcus aureus	1.000.000	> 0.9999	> 0.9999	> 0.9999	0.0203
Streptococcus agalactiae	1.000.000	0.8729	> 0.9999	0.0382	0.0004
Pseudomonas aeruginosa	1.000.000	> 0.9999	> 0.9999	0.0231	0.0004
Clostridium beijerinckii	1.000.000	> 0.9999	0.6831	0.2230	0.0009
Bacillus cereus	1.000.000	> 0.9999	> 0.9999	> 0.9999	0.0048
Helicobacter pylori	100.000	> 0.9999	> 0.9999	0.0292	0.0006
Lactobacillus gasseri	100.000	> 0.9999	0.1874	0.0025	0.0004
Neisseria meningitidis	100.000	> 0.9999	0.7944	0.0242	0.0009
Acinetobacter baumannii	100.000	0.9033	0.2575	0.0024	0.0003
Propionibacterium acnes	100.000	> 0.9999	0.6463	0.0015	0.0015
Listeria monocytogenes	100.000	> 0.9999	0.1586	0.0006	0.0006
Enterococcus faecalis	10.000	0.0060	0.0007	0.0007	0.0007
Streptococcus pneumoniae	10.000	0.4097	0.0011	0.0011	0.0011
Bacteroides vulgatus	10.000	0.0005	0.0005	0.0005	0.0005
Actinomyces odontolyticus	10.000	0.0005	0.0005	0.0005	0.0005
Deinococcus radiodurans	10.000	NA	NA	NA	NA

2.4. Influence of the level of host DNA on the sensitivity of WMS for microbiome taxonomic profiling at a fixed sequencing depth

Having shown that high proportions of host DNA and reduced sequencing depths interfere with the sensitivity of WMS for microbiome profiling, the next aim was to investigate the influence of the level of host DNA on the sensitivity of the method at a fixed sequencing depth. For that, simulated datasets were generated with progressively greater proportions of host DNA (SD10 to SD99, ranging from 10% to 99% host reads), at the fixed sequencing depth of 10 million single-end reads with 150 bp length (1.5 Gb). This depth was chosen based on the recent guidelines for best practices for shotgun metagenomics, which suggest a minimum of 1 Gb sequencing depth per-sample (Quince *et al.* 2017). Simulated datasets were composed of microbial and mouse reads, randomly picked from our previously generated MS and SS99 raw datasets, respectively **(Figure 23)**. For each of the simulated datasets, 5 replicates were generated, and the taxonomic profile was estimated with MetaPhlAn2, after sequencing data pre-processing, and compared to that of MS.

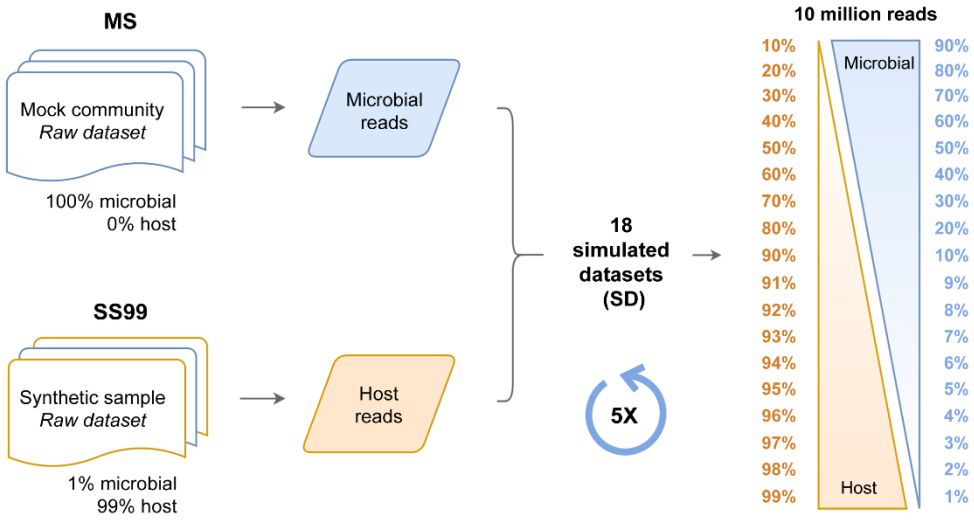


Figure 23. Schematic representation of the experimental design to generate simulated datasets with different host:microbial ratios. Microbial and host single-end reads were randomly selected from the mock microbial community (MS) and from the SS99 raw datasets, and were combined in different proportions, at a fixed sequencing depth of 10 million reads, to generate 18 simulated datasets (SD) with progressively greater complexity (host reads ranging from 10% to 99%).

In SD10 to SD60, all 20 species of bacteria were successfully detected, without significant differences in their relative abundances in comparison with the MS control **(Figure 24 and Table 9 and 10)**. In SD70 to SD90, there was a progressive reduction of the number of very low abundant species identified, none of them being detected in SD90 **(Figure 24 and Table 9 and 10)**, a result in line with the random subsampling analysis performed above **(Figure 24)**. In SD92 to SD99, in addition to not identifying all of the very low abundant species, there was a statistically significant decrease in the relative abundance of low abundant species **(Figure 24 and Table 9 and 10)**. In particular, when host reads represented 97% to 99% of the datasets, the low abundant species were mostly undetected, and the relative abundance of the majority of high abundant species significantly decreased. Additionally, the *Pseudomonas phage Pf1* was again identified with high relative abundances in these highly complex datasets **(Figure 24)**. In the most complex dataset (SD99) a *Staphylococcus* phage *StauST398-2* was also identified.

Overall, these results show that at the fixed depth of 10 million reads, which is currently used in metagenomic studies, high levels of host DNA interfere with an accurate reconstitution of the microbiome profile.

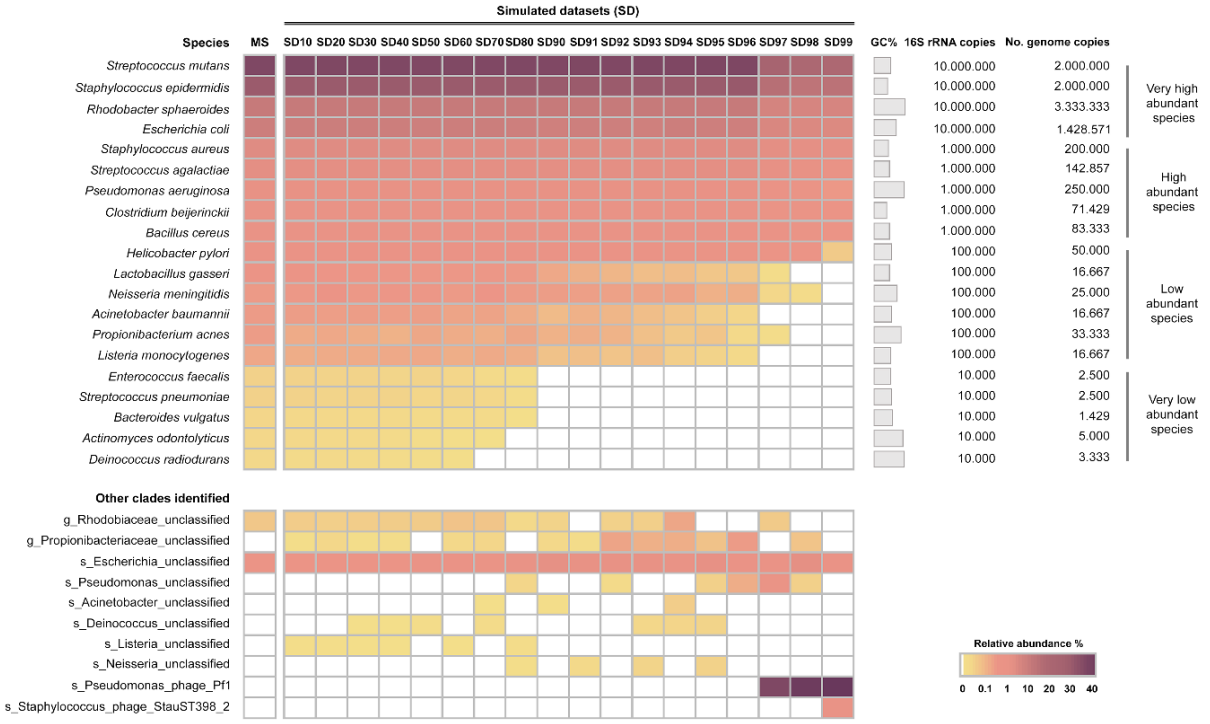


Figure 24. Influence of host DNA on the sensitivity of WMS for microbiome taxonomic profiling at a fixed sequencing depth. Taxonomic profile of the simulated datasets is represented as the average relative abundance from five independent simulations, and shown in a heat map. Data was sorted from the highest to the lowest relative abundances of species in the mock microbial community (MS).

Table 9. Ratio of relative abundances of species from each simulated dataset (SD) to MS dataset. Grey shading indicates a ≥ 2-fold change. Random subsampling to generate each simulated dataset was performed in five independent experiments.

Microbial Species	16S rRNA copies	MS	SD10 /MS	SD20 /MS	SD30 /MS	SD40 /MS	SD50 /MS	SD60 /MS	SD70 /MS	SD80 /MS	SD90 /MS	SD91 /MS	SD92 /MS	SD93 /MS	SD94 /MS	SD95 /MS	SD96 /MS	SD97 /MS	SD98 /MS	SD99 /MS
Streptococcus mutans	10.000.000	1	1.001	1.003	0.999	1.001	0.999	0.994	0.998	1.002	0.995	1.006	0.997	1.011	1.008	1.011	1.015	0.639	0.590	0.584
Staphylococcus epidermidis	10.000.000	1	1.001	1.005	1.003	1.005	1.005	1.006	1.002	1.005	1.015	1.002	1.003	1.027	1.013	1.005	1.038	0.659	0.572	0.603
Rhodobacter sphaeroides	10.000.000	1	1.002	0.993	0.996	1.000	1.001	1.001	0.998	0.994	0.988	1.000	1.005	0.985	1.001	0.983	1.005	0.638	0.570	0.572
Escherichia coli	10.000.000	1	1.003	1.014	0.997	0.987	0.974	0.987	1.008	0.960	0.913	0.946	0.965	0.911	0.982	0.959	0.940	0.597	0.520	0.499
Staphylococcus aureus	1.000.000	1	1.001	1.008	1.007	1.010	1.015	1.001	1.003	0.996	1.009	1.017	0.986	0.999	0.983	0.981	0.989	0.620	0.530	0.525
Streptococcus agalactiae	1.000.000	1	0.998	1.004	1.002	0.998	0.994	1.000	0.996	1.007	1.014	0.966	0.960	0.977	0.941	0.971	0.959	0.601	0.532	0.395
Pseudomonas aeruginosa	1.000.000	1	0.977	0.994	0.980	0.997	0.990	1.009	0.985	0.982	0.920	0.880	0.918	0.879	0.902	0.809	0.919	0.478	0.411	0.139
Clostridium beijerinckii	1.000.000	1	0.989	0.999	1.007	1.003	0.979	0.979	0.996	0.944	0.946	0.943	1.000	0.920	0.901	0.911	0.911	0.538	0.484	0.304
Bacillus cereus	1.000.000	1	1.039	0.977	1.037	0.978	1.036	1.023	0.991	1.003	0.903	0.979	1.065	0.997	1.011	1.015	0.851	0.698	0.568	0.435
Helicobacter pylori	100.000	1	0.992	0.982	0.990	0.990	0.955	0.963	0.939	0.923	0.909	0.953	0.967	0.851	0.806	0.827	0.798	0.435	0.301	0.072
Lactobacillus gasseri	100.000	1	0.941	0.939	0.919	0.924	0.910	0.903	0.906	0.890	0.698	0.588	0.509	0.413	0.424	0.305	0.295	0.033	0	0
Neisseria meningitidis	100.000	1	1.026	1.097	1.034	1.051	1.043	1.006	1.067	0.988	0.950	0.929	0.892	0.875	0.884	0.670	0.710	0.113	0.045	0
Acinetobacter baumannii	100.000	1	0.978	0.957	0.971	0.951	0.963	0.874	0.888	0.744	0.533	0.675	0.601	0.485	0.397	0.241	0.123	0	0	0
Propionibacterium acnes	100.000	1	0.780	0.830	0.749	0.674	0.843	0.760	0.712	0.858	0.676	0.736	0.659	0.460	0.320	0.381	0.078	0.034	0	0
Listeria monocytogenes	100.000	1	0.895	0.932	0.931	0.975	0.979	0.883	0.948	0.875	0.515	0.573	0.479	0.576	0.292	0.216	0.106	0	0	0
Enterococcus faecalis	10.000	1	0.858	0.820	0.837	0.705	0.820	0.588	0.325	0.210	0	0	0	0	0	0	0	0	0	0
Streptococcus pneumoniae	10.000	1	0.885	0.757	0.848	0.852	0.641	0.711	0.532	0.177	0	0	0	0	0	0	0	0	0	0
Bacteroides vulgatus	10.000	1	0.819	0.782	0.627	0.699	0.581	0.505	0.483	0.030	0	0	0	0	0	0	0	0	0	0
Actinomyces odontolyticus	10.000	1	0.778	0.655	0.685	0.538	0.319	0.157	0.015	0	0	0	0	0	0	0	0	0	0	0
Deinococcus radiodurans	10.000	1	0.704	0.831	0.662	0.523	0.428	0.212	0	0	0	0	0	0	0	0	0	0	0	0

Table 10. Statistical analysis (*P*-values) of the results presented in Table 9. The Kruskal-Wallis non-parametric test followed by multiple comparisons (SD10, SD20, SD30, SD40, SD50, SD60, SD70, SD80, SD90, SD91, SD92, SD93, SD94, SD95, SD96, SD97, SD98 or SD99) versus a control group (MS) using Dunn's test was performed for each species.

Microbial species	16S rRNA copies	MS vs. SD10	MS vs. SD20	MS vs. SD30	MS vs. SD40	MS vs. SD50	MS vs. SD60	MS vs. SD70	MS vs. SD80	MS vs. SD90	MS vs. SD91	MS vs. SD92	MS vs. SD93	MS vs. SD94	MS vs. SD95	MS vs. SD96	MS vs. SD97	MS vs. SD98	MS vs. SD99
Streptococcus mutans	10.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1720	0.2236	0.1963
Staphylococcus epidermidis	10.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1454	>0.999	>0.999	0.1608	>0.999	>0.999	>0.999
Rhodobacter sphaeroides	10.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.8504	0.1066	0.0994
Escherichia coli	10.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.0832	0.0195	0.0166
Staphylococcus aureus	1.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.6271	0.3169	0.3169
Streptococcus agalactiae	1.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.2796	>0.999	0.6096	0.0412	0.018	0.0053
Pseudomonas aeruginosa	1.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.7215	>0.999	>0.999	0.3926	>0.999	0.0747	>0.999	0.0085	0.0038	0.0007
Clostridium beijerinckii	1.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.8975	>0.999	0.5757	>0.999	0.0153	0.0105	0.0028
Bacillus cereus	1.000.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.3371	0.4981	0.1312
Helicobacter pylori	100.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.8736	>0.999	0.7417	>0.999	>0.999	0.9721	0.0462	0.2796	0.1899	0.0015	0.0005	0.0002
Lactobacillus gasseri	100.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.4485	0.0702	0.0371	0.0120	0.0154	0.0044	0.0039	0.0001	<0.0001	<0.0001
Neisseria meningitidis	100.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1467	0.1057	0.081
Acinetobacter baumannii	100.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.9099	0.1078	0.4353	0.1922	0.0841	0.0305	0.0078	0.0019	0.0002	0.0002	0.0002
Propionibacterium acnes	100.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.7519	0.0449	0.0154	0.0181	0.0004	0.0002	0.0001	0.0001
Listeria monocytogenes	100.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1784	0.3442	0.1697	0.7687	0.0247	0.0123	0.0035	0.0005	0.0005	0.0005
Enterococcus faecalis	10.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.5228	0.1474	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
Streptococcus pneumoniae	10.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.7679	0.1013	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
Bacteroides vulgatus	10.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.9671	0.0035	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
Actinomyces odontolyticus	10.000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.0036	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
Deinococcus radiodurans	10.000	>0.999	>0.999	>0.999	>0.999	>0.999	0.6434	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014

3. Analysis of WMS data from human gastric carcinoma specimens

To study the gastric mucosa-associated microbiome using WMS, we applied the previously optimized strategy to metagenomes of human gastric carcinoma specimens. Having shown that high levels of host DNA reduce WMS sensitivity for microbiome profiling, DNA isolation from gastric carcinoma specimens (n = 2) was performed with or without a human DNA depletion step. WMS was applied to all samples (n = 4), using the highest sequencing depth possible.

The four datasets yielded a large number of raw single-end reads ranging from 26 to 84 million. After quality filtering and host sequences decontamination, the number of reads was very low in all samples ranging from 3.7 hundred thousand to 1.3 million **(Table 11)**.

Table 11. Summary of the sequencing data pre-processing of gastric carcinoma metagenomes with the optimized quality filtering parameters.

Sample		Total number of raw single-end reads	Total number of quality-filtered reads	Total number of quality-filtered and host decontaminated reads
Host DNA depletion	Tumor #1	83.505.054	46.595.580	1.285.414
	Tumor #2	48.128.504	26.941.140	801.522
Without host DNA depletion	Tumor #1	46.217.930	35.331.215	633.504
	Tumor #2	26.734.786	20.709.707	368.891

Of the total raw single-end reads, the proportion of discarded reads during quality filtering was similar among gastric carcinoma specimens within the same DNA isolation method. However, it was different between the two DNA isolation approaches, being higher in the approach with host DNA depletion (44%) compared to that without (around 24%) **(Figure 25)**.

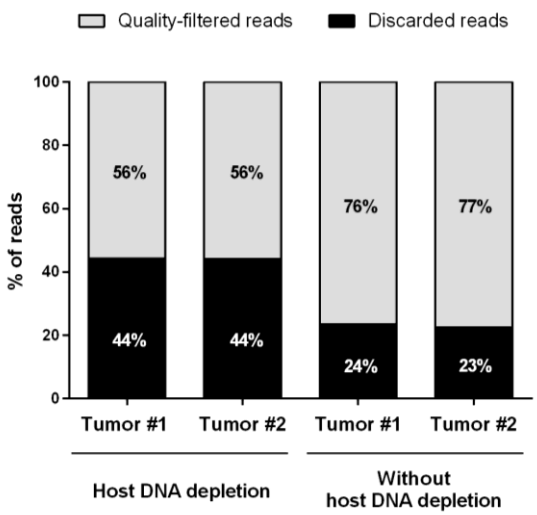


Figure 25. Overview of the quality filtering of gastric carcinoma metagenomes with the optimized parameters. The proportion of quality-filtered and discarded reads from the total raw single-end reads per gastric carcinoma specimen (with or without host DNA depletion approach).

To understand this discrepancy, the quality and length distribution of the raw reads were evaluated **(Table 12 and Figure 26)**. All raw datasets showed an overall good quality, having between 84% and 91% of the reads with average quality ≥ Q30, suggesting no association between the quality of the data and the variations in the fraction of reads dropped during quality filtering.

Table 12. Quality of the reads before and after sequencing data pre-processing of gastric carcinoma metagenomes with the optimized quality filtering parameters.

		Host DNA depletion		**Without host DNA depletion**	
		Tumor #1	**Tumor #2**	**Tumor #1**	**Tumor #2**
% reads with average quality ≥ Q30	Raw reads	91.1%	89.6%	87.1%	84.3%
	Quality-filtered and host decontaminated reads	98.3%	98.2%	97.9%	97.6%

In contrast, the read length distribution analysis showed differences between datasets from different DNA isolation approaches. There was a higher number of reads with ≤ 60 bp length, and a lower number of reads with ≥ 101 bp length in datasets from DNA isolation with host DNA depletion compared to those without **(Figure 26)**.

These results suggest that host DNA depletion led to fragmentation of the DNA from gastric carcinoma specimens, increasing the number of short reads, and thus the fraction of reads discarded based on length. In fact, there was a great fraction of reads with ≤ 60 bp (> 35%), which is the minimum used length considered by KneadData for quality filtering **(Figure 26)**.

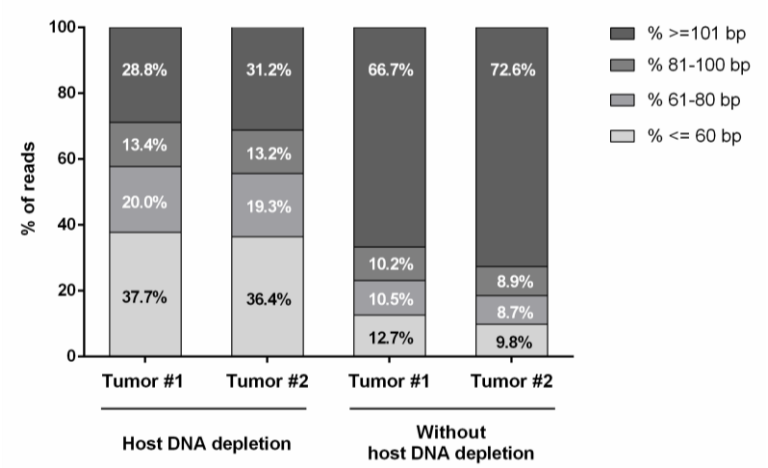


Figure 26. Raw reads length distribution analysis in gastric carcinoma metagenomes. Results are shown as the percent of raw reads with specific range sizes in each sample.

Of the total quality-filtered reads, the proportion of reads removed as host DNA sequences from gastric carcinoma metagenomes was very high ranging from 97% to 98%, even when host DNA depletion was performed **(Figure 27A)**. To confirm that quality filtering had no impact on the ratios of host to microbial reads obtained in the pre-processed metagenomes, the same ratio was determined by performing host sequences decontamination without the quality filtering step. Likewise, in all raw datasets the percentage of host reads was very high (around 96%), independently of the DNA isolation method used **(Figure 27B)**, demonstrating that the host DNA depletion step did not decrease the fraction of human DNA in gastric carcinoma samples, yielding a very low number of microbial reads**.**

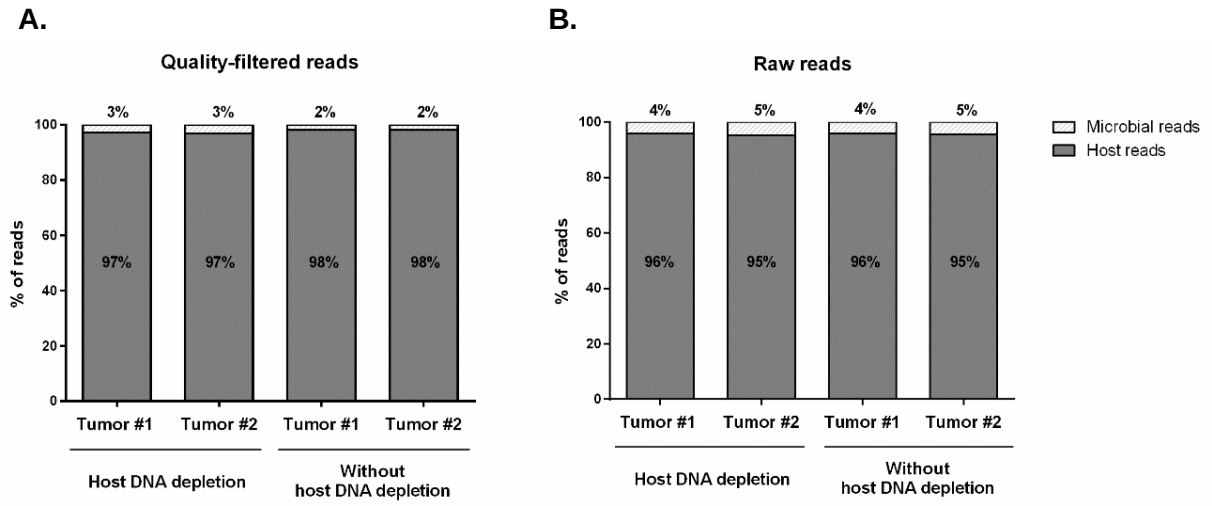


Figure 27. Microbial and host reads in the gastric carcinoma metagenomes. Percentage of microbial and host reads **(A)** from the total quality-filtered reads and **(B)** from the total raw single-end reads, per gastric carcinoma specimen.

After pre-processing the reads, the taxonomic profiles of the four samples of the two gastric carcinoma specimens was determined with MetaPhlAn2 **(Figure 28)**. Independently of the DNA isolation method, in two samples (Tumor #1, with host DNA depletion and Tumor #2, without host DNA depletion), all reads were assigned to an unknown genus of the *Propionibacteriaceae* family, while in the other two (Tumor #2, with host DNA depletion and Tumor #1, without host DNA depletion) all reads were assigned to unclassified clades **(Figure 28)**.

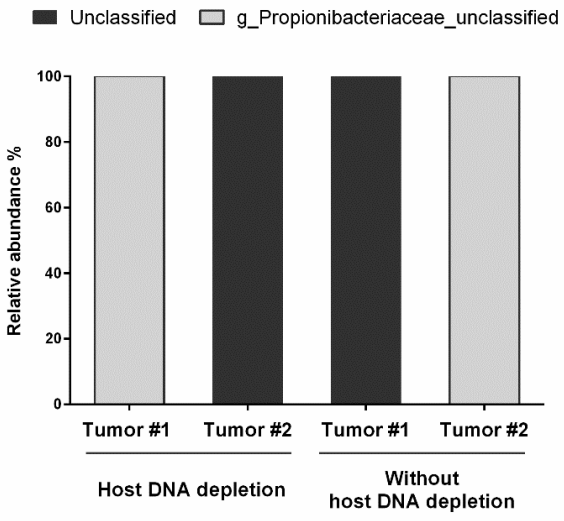


Figure 28. Taxonomic profile of gastric carcinoma specimens from WMS reads. Relative abundance of each taxa was determined with MetaPhlAn2 per gastric carcinoma specimen in the presence and absence of a host DNA depletion step during DNA isolation.

To confirm the low proportion of microbial reads in our WMS data from gastric carcinoma metagenomes, a specific qPCR assay was used to quantify the total bacterial load. Accordingly, in host DNA depleted samples, there was no detection of bacteria using the universal assay. However, in the absence of host DNA depletion, we were able to detect bacteria but at a very low DNA concentration **(Figure 29)**.

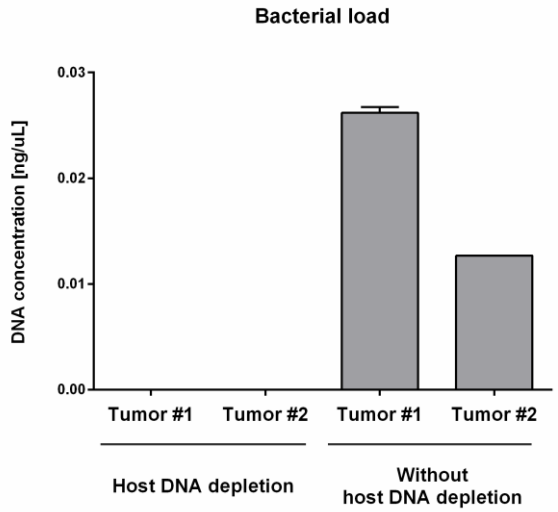


Figure 29. Quantification of the total bacterial load by quantitative PCR. Data are presented as the mean DNA concentration determined for the universal assay plus standard error.

These results confirm the low amount of microbial DNA present in all gastric carcinoma samples, independently of the DNA isolation approach used. Moreover, our data emphasize that the host DNA depletion approach promotes DNA fragmentation also impairing the detection of bacteria by WMS and qPCR. Nevertheless, in non-depleted host DNA samples there was not enough coverage to identify microbial species using WMS, because of the low number of bacteria present in the gastric carcinoma samples.

DISCUSSION

PART I. Characterization of the gastric microbiota using next-generation sequencing of the 16S rRNA gene in chronic gastritis and gastric carcinoma patients

In **Part I**, we successfully applied NGS of the 16S rRNA to characterize the gastric microbiota in the context of gastric carcinogenesis. In fact, our profiling of the gastric microbiota associated with chronic gastritis and gastric carcinoma constitutes the largest and most in-depth study to date. We have demonstrated that the gastric microbiota composition in patients with gastric carcinoma is significantly different from that of patients with chronic gastritis. Gastric carcinoma dysbiosis was consistent with a microbial community characterized by reduced microbial diversity, reduced *Helicobacter* abundance and over-representation of new bacterial genera. The major findings revealed in the Portuguese discovery cohort were confirmed in an additional validation cohort from Portugal.

In our study, the gastric microbial communities in gastritis and carcinoma were structurally different, with decreased alpha-diversity in carcinoma. Our findings are supported by previous data pointing to lower bacterial diversity among five patients with gastric cancer compared with five patients with non-atrophic gastritis (Aviles-Jimenez *et al.* 2014). Also supporting our data, another study identified significant decreases in microbial richness in intestinal metaplasia and in gastric carcinoma compared with superficial gastritis (Coker *et al.* 2018). Reduced microbial diversity has now been recognized as a feature of disease states, including inflammatory diseases and cancer (Ahn *et al.* 2013; Gevers *et al.* 2014; Lepage *et al.* 2011). For example, patients with colorectal cancer had decreased overall microbial community diversity in comparison to healthy controls (Ahn *et al.* 2013).

In terms of the composition of the gastric microbiota, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria,* and *Fusobacteria* were the five dominant phyla in the stomach, in accordance with previous descriptions (Andersson *et al.* 2008; Bik *et al.* 2006; Delgado *et al.* 2013). At the phylum level, we have identified differences between the two patient groups, with increased abundance of non-*Helicobacter Proteobacteria*, *Firmicutes* and *Actinobacteria* in cancer specimens. Importantly, by applying the LEfSe algorithm that was validated for high-dimensional microbiome data sets, we were able to determine the bacterial taxa that most likely explain differences between clinical diagnoses (Segata *et al.* 2011). Additionally, in this study, the major taxonomic differences that were detected after analyses of sequencing-generated and bioinformatics-treated data were further validated by real-time qPCR assays.

In chronic gastritis, and as expected, *Helicobacter* was detected as the most abundant genus. *Streptococcus*, *Prevotella* and *Neisseria* were also found significantly overabundant in this patient group, although *Streptococcus* and *Prevotella* could not be confirmed by

qPCR. Nevertheless, these genera have been identified earlier in *H. pylori*-positive and negative gastritis by 16S rDNA and rRNA sequencing, and by culture from gastric juice and gastric biopsies (Bik *et al.* 2006; Delgado *et al.* 2013; Thorens *et al.* 1996; Schulz *et al.* 2018). In fact, they are among the five most commonly found genera in the non-neoplastic stomach (Andersson *et al.* 2008; Bik *et al.* 2006; Li *et al.* 2009). *Streptococcus*, *Prevotella,* and *Neisseria* are commensals of the oral cavity and oesophagus and whether they constitute transient or active resident stomach microbes is not yet clarified. Interestingly, in a study that compared the gastric microbiota compositions in *H. pylori*-positive individuals from two populations with high and low gastric cancer risks in Colombia, *Neisseria* and *Streptococcus* were among the genera that occurred more abundantly in individuals from the low gastric cancer risk region (Yang *et al.* 2016).

In gastric carcinoma, there was a significant decrease in *Helicobacter* abundance, and several taxa were found to be significantly more abundant. These included *Citrobacter*, *Clostridium*, *Lactobacillus*, *Achromobacter,* and *Rhodococcus*, which reside in the intestinal mucosa as commensals but can be opportunistic pathogens (Kelly and LaMont 2008; Rajilic-Stojanovic and de Vos 2014). *Phyllobacterium*, which are environmental bacteria commonly found in plant roots, were too identified at higher abundance in gastric carcinoma (Jiao *et al.* 2015). In line with our results, in a study that combined T-RFLP with 16S rRNA gene cloning and sequencing, *Lactobacillus* was one of the dominating genera in 10 Swedish patients with gastric cancer (Dicksved *et al.* 2009). Additionally, the use of the microarray G3 PhyloChip to characterize the stomach microbiota of Mexican patients revealed a trend towards the increase of a *Lactobacillus* sp. from non-atrophic gastritis, to intestinal metaplasia, to gastric cancer (Aviles-Jimenez *et al.* 2014). Moreover, *Citrobacter*, *Clostridium,* and *Lactobacillus* have all been cultured from the gastric juice of achlorhydric patients, patients undergoing acid suppression therapy and patients with gastric cancer (Forsythe *et al.* 1988; Mowat *et al.* 2000; Sjostedt *et al.* 1985). Interestingly, infection with *Citrobacter rodentium* species increases epithelial cell proliferation and promotes colonic tumour formation in genetically susceptible mice as well as in chemically initiated colon carcinogenesis (Barthold and Jonas 1977; Newman *et al.* 2001). The integration of data from the most relevant genera that characterized each patient group allowed us to calculate the dysbiosis index that showed excellent capacity to discriminate between gastritis and gastric carcinoma.

Taken together our results and previously published data, we propose that colonization with bacteria other than *H. pylori*, namely gut commensals, contributes to alter the equilibrium between the ‘resident’ gastric microbiota and the host. This dysbiotic microbial community, by sustaining the gastric inflammatory process, and through its intrinsic genotoxic potential, may augment the risk for *H. pylori*-related gastric carcinoma development. In line with our

proposal, experimental evidence in the INS-GAS model showed that commensal intestinal bacteria play a role in the promotion of gastric cancer (Lertpiriyapong *et al.* 2014; Lofgren *et al.* 2011). Lertpiriyapong *et al.* showed that mice harboring a complex intestinal microbiota, and mice colonized with a restricted intestinal microbiota (that includes *Clostridium* and *Lactobacillus*), had an accelerated onset and progression of gastric cancer secondary to *H. pylori* infection. These mice also developed more severe gastric histopathology and higher expression levels of proinflammatory genes in comparison to germ-free mice (infected or not with *H. pylori*) (Lertpiriyapong *et al.* 2014).

Although our study is limited by its retrospective nature, and by the low number of patients with true premalignant lesions, our findings are consistent with a shift in the gastric microbial community structure along gastric carcinogenesis. In this sense, prospective follow-up studies of patients with premalignant lesions, successfully eradicated or not for *H. pylori* infection, would be crucial to ascertain the pathogenic effect of microbial dysbiosis in the progression to carcinoma. Additional studies aiming to address the effect of dysbiosis or of candidate bacterial species in an animal model of gastric carcinogenesis can also be considered, and in that regard, a humanised mouse model that better mimics the human immune response could be particularly informative. Ultimately, understanding the microbiota dynamics along gastric carcinogenesis may impact gastric carcinoma prevention and treatment strategies of patients with precancerous disease.

PART II. Establishment of a whole metagenome sequencing strategy to characterize the gastric mucosa-associated microbiome

WMS may offer a higher degree of taxonomic and functional resolution to the characterization of the microbiome, as it captures the full genomic content of a sample (Lloyd-Price *et al.* 2017). Studies using WMS to investigate the composition of the microbial communities present in human gastric mucosal tissues are still scarce, and therefore, the implementation of a WMS strategy to study the gastric-mucosa associated microbiome in the gastric carcinoma context was tackled in the **Part II** of the Results.

In **Part II.1** of the Results, we effectively established, optimized, and validated a pipeline to analyze WMS data generated by Illumina NextSeq 550 platform, using two mock DNA microbial communities from BEI resources. Mock microbial communities have been previously used not only for benchmarking novel technologies, but also to evaluate the performance of analytical methods, and to standardize protocols to ensure accuracy and consistency (Laursen, Dalgaard, and Bahl 2017; Jumpstart Consortium Human Microbiome Project Data Generation Working 2012; Human Microbiome Project 2012b). We selected

these synthetic communities (HM-276D and HM-277D) since their bacterial composition is well-defined, and they were generated by the HMP as reference standards. These mock samples were sequenced on the Illumina NextSeq 550, because this system provides a very high output (around 120 Gb per run), being well suited for whole metagenome analysis (Quince *et al.* 2017).

The pipeline of WMS data analysis implemented for microbiome profiling comprises two major steps: the sequencing data pre-processing and the taxonomic profiling, which were both performed using computational tools that are open source and are well established for whole metagenome analysis (Truong *et al.* 2015; Andrews 2016; Bolger, Lohse, and Usadel 2014; Rotmistrovsky and Agarwala 2018; Langmead and Salzberg 2012). The microbiome sequencing data was analyzed using a read-based profiling approach that maps the host decontaminated quality-filtered reads directly to unique marker genes without assembly (Quince *et al.* 2017). This computational approach provides a fast and accurate sequence data analysis to be applied to large WMS datasets. Moreover, it may allow profiling of low-abundance organisms that lack enough sequence coverage to allow assembly of their genome (Quince *et al.* 2017).

The sequencing data pre-processing performed with KneadData (Hu *et al.* 2018; Schirmer *et al.* 2018; Schirmer *et al.* 2016) was optimized, in order to obtain a similar proportion of quality-filtered reads across samples. In fact, after optimizing this parameter, around 30% of raw reads were discarded in both mock samples. Comparable proportions of low-quality reads from metagenomes of the mock community Even (HM-276D, BEI resources) were previously described, in a study using the same tool for quality filtering (Jones *et al.* 2015).

Validation of the metagenomics workflow is critical to assure a precise profiling of the taxonomic features of the microbiome in biological samples. Therefore, the successful validation of our optimized pipeline of WMS data analysis using the two mock communities, suggests high quality datasets. The taxonomic composition of these synthetic communities was investigated using MetaPhlAn2, an assembly-free taxonomic profiler that has proved effective in the human microbiome characterization, with a very low false positive rate (Lloyd-Price *et al.* 2017; Walsh *et al.* 2018; Asnicar *et al.* 2017; Truong *et al.* 2015). In fact, the use of this tool enabled an accurate reconstitution of the taxonomic profile of both mock communities from WMS reads, despite the over- and underrepresentation of some species due to GC bias in our Nextera XT libraries. This observation is in line with previous studies that showed a GC bias in libraries sequenced in different Illumina platforms (Jones *et al.* 2015; Jumpstart Consortium Human Microbiome Project Data Generation Working 2012; Laursen, Dalgaard, and Bahl 2017). Transposase insertion bias during tagmentation (Nextera libraries) (Goryshin *et al.* 1998), PCR amplification during library construction (Aird *et al.* 2011; Laursen, Dalgaard, and Bahl 2017), cluster amplification (Stein, Takasuka, and

Collings 2010) and sequencing (Aird *et al.* 2011) are known sources of GC bias in Illumina sequencing. Our data together with other published results indicate that the genomic GC content is an important aspect to consider when determining the abundances of microbial species using an Illumina system. To reduce this experimental bias related with Illumina sequencing, some strategies could be applied. For example, simulated datasets with a known taxonomic composition can be created to validate the pipeline of analysis, and PCR-free library methods (such as Kapa Hyper Prep PCR-free and TruSeq DNA PCR-free) can be used as an alternative to Illumina Nextera XT libraries that include a PCR amplification step (Jones *et al.* 2015; Siegwald *et al.* 2017).

According with data from the HMP, there is a high heterogeneity in the fraction of human DNA across different human sample types (Lloyd-Price *et al.* 2017). Moreover, samples with a high amount of host DNA, remains a main drawback in whole metagenome analysis, affecting the efficiency of microbiome profiling (Quince *et al.* 2017). Therefore, in **Part II.2** of the Results, we assessed the influence of sample complexity on the characterization of the taxonomic features of the microbiome using our optimized and validated pipeline of WMS data analysis.

Here, we showed that high proportions of host DNA, characterized by greater proportions of host DNA, reduce the sensitivity of WMS for microbiome profiling, in particular to detect very low and low abundant species of bacteria. It is plausible that high ratios of host:bacteria DNA reduce sequence coverage of the microbial genomes, hindering subsequent taxonomic analysis. This is consistent with previous studies addressing the issue of human DNA contamination on whole genome sequencing detection of the malaria parasite in clinical samples (Auburn *et al.* 2011; Oyola *et al.* 2013). Although these reports were not focused on microbiome characterization, they showed that low levels of human DNA (≤ 30%) in blood samples, resulted in higher average *Plasmodium* genome coverage (Auburn *et al.* 2011), whereas clinical samples containing > 80% human DNA, yielded a low number of reads assigned to *P. falciparum* genome (Oyola *et al.* 2013). In line with these observations, Hasan *et al.* found that by decreasing the human DNA background in a clinical sample, the sensitivity to detect microbial species was improved (Hasan *et al.* 2016).

We also showed that sequencing depth influences the sensitivity of microbiome profiling by WMS in samples that contain high levels of host DNA. The generation of SS90 datasets with reduced sequencing depths, resulted in gradually decreased capacity to accurately profile the microbiome. A reduction in sequencing depth from 51 million to 25 million reads already decreased WMS sensitivity, by preventing the identification of 60% of the species with very low abundance. In agreement with these findings, Jovel *et al.* showed that an increase in the size of the dataset leads to both an improvement of detection of microbial

species and a more consistent estimation of their relative abundances (Jovel *et al.* 2016). Moreover, in a metagenomic study of the faecal microbial community from beef cattle, the identification of new microbial taxa markedly improved with larger sequencing depths (Zaheer *et al.* 2018).

We also demonstrated that, besides preventing the identification of all species with very low abundance, a reduction in sequencing depth to 5 million reads additionally affected the relative abundance estimates of low abundant species. A depth as low as 2.5 million reads also resulted in major impairment in estimating the relative abundances of high and very high abundant species. In contrast with our findings, a recent study found no differences in the taxonomic profile of a mock community at divergent sequencing depths ranging from 0.1 to 7.5 single-end million reads (Walsh *et al.* 2018). These discrepancies may reflect the absence of host DNA in the mock sample analysed in that study, as compared to the high levels of host DNA in our study samples (90%). Taken together, our data and that of others, suggest that similar sequencing depths have distinct effects on the sensitivity of WMS for taxonomic profiling, depending on the sample. In fact, our analysis of simulated datasets with 10 million reads indicated that the reconstitution of the microbiome profile becomes more inaccurate as the amount of host DNA in a sample progressively increases. Interestingly, and based on this analysis, the outcomes of sequencing different types of host-derived samples, at a depth of 1.5 Gb per sample, can be extrapolated. For example, when sequencing a stool sample, the whole microbial community is expected to be accurately reconstituted, considering the low amount of host DNA in this type of sample (< 10%; (Human Microbiome Project 2012a). However, when sequencing samples like saliva, throat, buccal mucosa, and vaginal swabs (> 90% host DNA) (Human Microbiome Project 2012a), the detection of very low and low abundant species is expected to be impaired. This becomes more problematic in case of sequencing a tissue sample, as the detection of very low to high abundant species will be hampered, since this type of sample contains mostly human DNA (97% to 99% reads) and a low microbial biomass (Zhang *et al.* 2015).

To the best of our knowledge, this is the first in-depth analysis demonstrating that greater proportions of host DNA, together with low sequencing depths, reduce the sensitivity of WMS for microbiome profiling. Therefore, the results of this study can assist in the design of WMS experiments, by highlighting the importance of sample type and sequencing depth when characterizing the microbiome.

In **Part II.3** of the Results, we attempted to profile the gastric mucosa-associated microbiome using WMS. Our previous results suggest that when performing WMS on human mucosal samples, the taxonomic profiling of the microbiome will be impaired by the high level of sample complexity of these specimens. To overcome this limitation and

improve the detection power, we implemented strategies to decrease sample complexity and to increase sequencing depth. Therefore, to reduce sample complexity, we applied two different DNA isolation procedures, including or not a host DNA depletion step. Regarding sequencing depth, we sequenced these samples on an Illumina high-throughput platform (maximum output around 120 Gb) and we limited the level of multiplexing per run as much as possible, in order to maximize the retrieved sequences from low abundant species. Finally, we applied our optimized pipeline of analysis to the metagenomes from human gastric carcinoma specimens.

Our data indicated that the host DNA depletion approach used (Ultra-Deep Microbiome Prep kit) promoted DNA fragmentation. We hypothesize that the high number of short reads and thus, lower number of quality-filtered reads in host DNA depleted samples is likely a result of the excessive sequential lysis and digestion steps included in the experimental protocol. In fact, DNA fragmentation could promote DNA loss during library construction and impair the sequencing read length, reducing the quality of the microbiome analysis (Ferretti *et al.* 2017).

We also demonstrated that the Ultra-Deep Microbiome Prep kit was not effective in the human DNA depletion from gastric carcinoma samples, since it yielded a very low number of microbial reads. In all tissue specimens, independently of the DNA isolation approach used, there was a very high proportion of reads removed as host sequences (around 98%). These observations confirmed the previous reports showing that > 95% of the reads from human gastric biopsy libraries mapped to the human reference genome (Zhang *et al.* 2015; Zhang, Thakkar, *et al.* 2018). Additionally, a very low DNA concentration of bacteria was detected in non-depleted host DNA specimens using qPCR, confirming the low proportion of microbial reads found in WMS datasets from these samples. In host DNA depleted samples, we were not able to quantify the total bacterial load using the universal qPCR assay, since DNA fragmentation probably impaired detection of bacterial genera using this technique. Even though host DNA depletion techniques have already worked out for other human-derived samples such as saliva and blood (Marotz *et al.* 2018; Zhou and Pollard 2012), they still remain to be validated for mucosal specimens. To the best of our knowledge, this is the first study applying an approach of microbial DNA enrichment in human gastric specimens for WMS. The host DNA depletion method we selected has been previously applied to fluid samples (e.g. broncho-alveolar lavage) (Leo *et al.* 2017), and tissue samples, including hepatic surgical specimens (Lazarevic *et al.* 2018), breast biopsies (Costantini *et al.* 2018), and rabbit tissue (Arrazuria *et al.* 2016). The efficacy of this approach was variable across studies, and likely dependent on the intrinsic properties of each sample type, but also on the downstream analysis performed (either 16S rRNA

sequencing or WMS) (Arrazuria *et al.* 2016; Costantini *et al.* 2018; Lazarevic *et al.* 2018; Leo *et al.* 2017).

To date, there are only two studies using WMS to characterize the human mucosa-associated microbiome, and none of them were focused on the gastric carcinoma scenario (Khosravi Y 2017; Zhang *et al.* 2015). Here, the compositional analysis of gastric carcinoma specimens using MetaPhlAn2 demonstrated that there was not enough coverage for identification of microbial species using WMS, due to the low number of bacteria present in these samples and the high amount of host cells. Independently of the DNA isolation approach, all reads were assigned either to unclassified clades or to an unknown genus of the *Propionibacteriaceae* family. Since low microbial biomass samples have been described to be more susceptible to DNA contamination during processing steps (Eisenhofer *et al.* 2019) and *Propionibacterium* genus has been reported as a common contaminant taxon in blank controls (Salter *et al.* 2014), we cannot fully exclude the hypothesis of a potential contamination by these bacteria in our datasets. However, it is important to note that in other microbiome studies, the genus *Propionibacterium* has also been identified in the human stomach (Delgado *et al.* 2013; Zhang *et al.* 2015).

Zhang and colleagues performed microbiome profiling on 27 gastric biopsies using WMS, identifying a few bacterial species, such as *H. pylori*, *P. acnes, Staphylococcus epidermidis*, *Haemophilus parainfluenzae,* and *Lactobacillus gasseri* (Zhang *et al.* 2015). Additionally, in a pilot study using WMS to investigate the composition of the microbiome on gastric biopsies from 4 Asian patients, bacterial taxa including *S. aureus*, *H. pylori* and *Streptococcus* sp. were detected (Khosravi Y 2017). However, results from these studies should be interpreted with caution. In general, the short length of the reads (50 bp) together with the absent or poorly described quality filtering parameters, are highly suggestive of low-quality datasets. Moreover, while we applied a robust and validated approach to taxonomically profile our samples, which was based on alignments with unique clade-specific marker genes (MetaPhlAn2), both studies mentioned above performed a brute-force mapping of reads to bacterial genomes. This strategy can result in profiles with many false positives, since one read mapped to only one bacterial genome is sufficient to consider a species to be present in a sample. In contrast, MetaPhlAn2 demands a higher number of reads mapping to marker genes for calling the presence of that species, being a more accurate method. Finally, in Zhang *et al.* the majority of the gastric biopsies were from *H. pylori*-infected patients and only 2 samples were gastric tumour biopsies. In Khosravi *et al.* among gastric specimens from patients with different gastroduodenal diseases, only 1 was a gastric carcinoma. These differences in the type of samples used may explain the identification of *H. pylori* in these datasets, contrasting to our study that includes only gastric carcinoma specimens.

Interestingly, in a recent work conducting a WMS survey on the microbiome of gastric wash samples from 6 gastric carcinoma and 5 superficial gastritis patients, 13 bacterial taxa enriched and 31 taxa depleted were identified in gastric carcinoma patients using MetaPhlAn2 (Hu *et al.* 2018). One can speculate that the discrepancies between these observations and the data obtained in this thesis may result from differences in the type of sample collected, and its respective degree of complexity. Gastric wash samples have less host DNA compared with gastric mucosal samples that comprise mostly human cells. This likely explains the higher number of non-human quality-filtered reads for taxonomic analysis (7.8 million) in gastric wash samples than in gastric mucosal specimens from our study (< 1.3 million). Therefore, the higher number of reads available for mapping in the gastric wash samples likely improves the ability of MetaPhlAn2 to identify microbial species.

Overall, our current WMS strategy was not effective to characterize the gastric mucosa-associated microbiome, since the host DNA depletion approach chosen was unsuccessful in decreasing the degree of sample complexity of the mucosal specimens before sequencing. Therefore, there was an insufficient coverage to identify microbial species, which impaired the taxonomic profiling of the microbiome using our approach. Hence, our results together with published data emphasize the urgent need for effective host DNA depletion and/or microbial enrichment methods that enable a successful whole metagenome analysis of mucosal samples.

SUMMARY AND CONCLUSIONS

The microbiome plays an important role in human physiology and in the maintenance of health, but also has a major impact in the development of a wide range of diseases, including cancer. In gastric cancer, *Helicobacter pylori* is a major player, but recent microbiome studies have shown the existence of other bacteria in the stomach. Still, research in the gastric microbiome is limited. The majority of the publications had technical limitations in sensitivity and depth of coverage, and included low number of patients, which in general did not allow producing statistically based conclusions. Therefore, the general aim of this thesis was to establish solid molecular-based approaches, in order to characterize the gastric microbiome in the context of gastric cancer. In particular, we focused on 16S rRNA gene sequencing and on whole metagenome sequencing (WMS).

In the first part of the thesis, the gastric microbiota profile was evaluated in 54 patients with gastric carcinoma and 81 patients with chronic gastritis by next-generation sequencing of the 16S rRNA gene. The structure and composition of gastric microbial community in patients with gastric carcinoma was significantly different from that of the patients with chronic gastritis. The gastric carcinoma microbiota was characterized by reduced microbial diversity, by decreased abundance of *Helicobacter* and by the enrichment of other bacterial genera, mostly represented by intestinal commensals. The combination of these taxa into a microbial dysbiosis index revealed that dysbiosis has excellent capacity to discriminate between gastritis and gastric carcinoma. The major taxonomic differences detected between clinical diagnoses using 16S rRNA gene sequencing were validated by real-time qPCR. These findings were further confirmed in an additional validation cohort of Portuguese patients.

In the second part of the thesis, the aim was to establish a WMS strategy to characterize the gastric mucosa-associated microbiome. Initially, a pipeline of WMS data analysis was effectively established, optimized, and validated using two mock microbial communities. Then, the sensitivity of WMS for microbiome taxonomic profiling was analyzed. Results demonstrated, in the first in-depth analysis so far, that high levels of sample complexity, characterized by greater proportions of host DNA, together with low sequencing depths, reduce the sensitivity of WMS for microbiome profiling. Finally, the characterization of the gastric mucosa-associated cancer microbiome using WMS was approached. Results showed that, even after host DNA depletion, WMS was not effective to characterize the microbiome due to insufficient coverage to identify microbial species.

Altogether, the findings presented in this thesis show that 16S rRNA gene-based next-generation sequencing could be successfully applied to characterize the gastric microbiota. Indeed, this constitutes the largest and most in-depth study to date of

the microbiota in the context of gastric carcinogenesis. Furthermore, the results reported in this thesis may be a valuable tool to assist in the design of future WMS studies to functionally characterize the microbiome, by highlighting the importance of sample complexity and sequencing depth. The accurate characterization of the constituents and functions of the microbiome in health and in disease will be decisive in the translation of research findings into clinical practice. In particular, the profiling of the microbiome present in the stomach along the process of gastric carcinogenesis, may improve the efficacy of preventive and therapeutic strategies to reduce the incidence of gastric cancer.

SUMÁRIO E CONCLUSÕES

O microbioma desempenha um papel importante na fisiologia humana e na manutenção da saúde. Contudo, tem também um grande impacto no desenvolvimento de um vasto número de doenças, nas quais se inclui o cancro. No cancro gástrico, apesar de *Helicobacter pylori* desempenhar um dos papeis mais importantes, estudos recentes na área do microbioma humano demostraram a presença de outras bactérias no estômago. Ainda assim, a investigação no âmbito do microbioma gástrico é escassa, com grande parte dos estudos existentes a revelar limitações técnicas quanto à sensibilidade e profundidade da sequenciação, e a incluir um número reduzido de doentes, o que, de forma geral, resulta em conclusões sem significância estatística. Deste modo, o objetivo geral desta tese foi o de estabelecer métodos moleculares robustos para caracterizar o microbioma do estômago no contexto do cancro gástrico. Em particular, focamo-nos na sequenciação do gene 16S de RNA ribossomal e na sequenciação completa do metagenoma.

Na primeira parte desta tese foi avaliado o perfil da microbiota gástrica em 54 pacientes com carcinoma gástrico e em 81 pacientes com gastrite crónica, por sequenciação de última geração do gene 16S rRNA. Demonstrou-se que a estrutura e composição da comunidade microbiana gástrica em pacientes com carcinoma gástrico é significativamente diferente da que apresentam os pacientes com gastrite crónica. A microbiota do carcinoma gástrico revelou-se caracterizada pela redução da diversidade microbiana, pela diminuição da abundância de *Helicobacter* e pelo enriquecimento de outros géneros bacterianos, principalmente comensais intestinais. A combinação destes géneros num índice de disbiose microbiana revelou que a disbiose possui excelente capacidade de discriminar entre gastrite e carcinoma gástrico. As principais diferenças taxonómicas detetadas entre os diagnósticos clínicos utilizando a sequenciação do gene 16S rRNA foram validadas por PCR quantitativo em tempo real. Estes resultados foram confirmados numa coorte adicional de validação composta por pacientes portugueses.

Na segunda parte desta tese, pretendeu-se estabelecer uma estratégia de metagenómica para caracterizar o microbioma associado à mucosa gástrica. Inicialmente, uma pipeline de análise bioinformática de dados foi estabelecida, optimizada e validada com sucesso, utilizando duas comunidades microbianas sintéticas. De seguida, a sensibilidade da técnica de sequenciação do metagenoma para determinar o perfil taxonómico do microbioma foi analisada. Na primeira análise aprofundada até ao momento, os resultados demonstraram que altos níveis de complexidade da amostra, caracterizados por elevadas proporções de DNA do hospedeiro, juntamente com baixas profundidades de sequenciação, diminuem a sensibilidade da técnica de sequenciação do metagenoma para traçar o perfil do microbioma. Por último, foi abordada a caracterização do microbioma associado à mucosa

gástrica por metagenómica no contexto do cancro gástrico. Os resultados mostraram que, mesmo após a remoção do DNA do hospedeiro, a sequenciação do metagenoma não foi eficaz para traçar o perfil do microbioma devido à insuficiente cobertura de profundidade para identificar espécies microbianas.

Em suma, os resultados apresentados nesta tese evidenciam que a sequenciação de última geração baseada no gene 16S pode ser aplicada com sucesso na caracterização das comunidades microbianas gástricas. De facto, este trabalho constitui o maior e mais aprofundado estudo até à presente data, sobre a microbiota no âmbito da carcinogénese gástrica. Os resultados descritos nesta tese constituem uma ferramenta importante para o desenho experimental de futuros estudos de metagenómica para traçar o perfil funcional do microbioma, destacando a importância da complexidade da amostra e da profundidade da sequenciação. A caracterização rigorosa dos membros integrantes do microbioma e suas funções, no contexto de saúde e doença, tem um papel fundamental na translação dos dados de investigação para a prática clínica. Em particular, o conhecimento do perfil do microbioma presente no estômago ao longo do processo de carcinogénese gástrica pode vir a melhorar a eficácia de estratégias preventivas e terapêuticas dirigidas à redução da incidência de cancro gástrico.

REFERENCES

Abubucker, S., N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methe, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower. 2012. 'Metabolic reconstruction for metagenomic data and its application to the human microbiome', *PLoS Comput Biol*, 8: e1002358.

Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj, and M. F. Polz. 2004. 'Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons', *J Bacteriol*, 186: 2629-35.

Adamsson, I., C. E. Nord, P. Lundquist, S. Sjostedt, and C. Edlund. 1999. 'Comparative effects of omeprazole, amoxycillin plus metronidazole versus omeprazole, clarithromycin plus metronidazole on the oral, gastric and intestinal microflora in Helicobacter pylori-infected patients', *J Antimicrob Chemother*, 44: 629-40.

Ahn, J., R. Sinha, Z. Pei, C. Dominianni, J. Wu, J. Shi, J. J. Goedert, R. B. Hayes, and L. Yang. 2013. 'Human gut microbiome and risk for colorectal cancer', *J Natl Cancer Inst*, 105: 1907-11.

Aird, D., M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. 2011. 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries', *Genome Biol*, 12: R18.

Allaband, C., D. McDonald, Y. Vazquez-Baeza, J. J. Minich, A. Tripathi, D. A. Brenner, R. Loomba, L. Smarr, W. J. Sandborn, B. Schnabl, P. Dorrestein, A. Zarrinpar, and R. Knight. 2019. 'Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians', *Clin Gastroenterol Hepatol*, 17: 218-30.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. 'Basic local alignment search tool', *J Mol Biol*, 215: 403-10.

Anderson, Marti J. 2001. 'A new method for non-parametric multivariate analysis of variance', *Austral Ecology*, 26: 32-46.

Anderson, Marti J., and Daniel C. I. Walsh. 2013. 'PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?', *Ecological Monographs*, 83: 557-74.

Andersson, A. F., M. Lindberg, H. Jakobsson, F. Backhed, P. Nyren, and L. Engstrand. 2008. 'Comparative analysis of human gut microbiota by barcoded pyrosequencing', *PLoS One*, 3: e2836.

Andrews, S. 2016. 'FastQC A Quality Control tool for High Throughput Sequence Data', http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arnold, J. W., J. Roach, and M. A. Azcarate-Peril. 2016. 'Emerging Technologies for Gut Microbiome Research', *Trends Microbiol*, 24: 887-901.

Arrazuria, R., N. Elguezabal, R. A. Juste, H. Derakhshani, and E. Khafipour. 2016. 'Mycobacterium avium Subspecies paratuberculosis Infection Modifies Gut Microbiota under Different Dietary Conditions in a Rabbit Model', *Front Microbiol*, 7: 446.

Asnicar, F., S. Manara, M. Zolfo, D. T. Truong, M. Scholz, F. Armanini, P. Ferretti, V. Gorfer, A. Pedrotti, A. Tett, and N. Segata. 2017. 'Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling', *mSystems*, 2.

Auburn, S., S. Campino, T. G. Clark, A. A. Djimde, I. Zongo, R. Pinches, M. Manske, V. Mangano, D. Alcock, E. Anastasi, G. Maslen, B. Macinnis, K. Rockett, D. Modiano, C. I. Newbold, O. K. Doumbo, J. B. Ouedraogo, and D. P. Kwiatkowski. 2011. 'An effective method to purify Plasmodium falciparum DNA directly from clinical blood samples for whole genome high-throughput sequencing', *PLoS One*, 6: e22213.

Aviles-Jimenez, F., F. Vazquez-Jimenez, R. Medrano-Guzman, A. Mantilla, and J. Torres. 2014. 'Stomach microbiota composition varies between patients with non-atrophic gastritis and patients with intestinal type of gastric cancer', *Sci Rep*, 4: 4202.

Backert, S., N. Tegtmeyer, and W. Fischer. 2015. 'Composition, structure and function of the Helicobacter pylori cag pathogenicity island encoded type IV secretion system', *Future Microbiol*, 10: 955-65.

Backert, S., N. Tegtmeyer, and M. Selbach. 2010. 'The versatility of Helicobacter pylori CagA effector protein functions: The master key hypothesis', *Helicobacter*, 15: 163-76.

Baker, G. C., J. J. Smith, and D. A. Cowan. 2003. 'Review and re-analysis of domain-specific 16S primers', *J Microbiol Methods*, 55: 541-55.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. 2012. 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *J Comput Biol*, 19: 455-77.

Barthold, S. W., and A. M. Jonas. 1977. 'Morphogenesis of early 1, 2-dimethylhydrazine-induced lesions and latent period reduction of colon carcinogenesis in mice by a variant of Citrobacter freundii', *Cancer Res*, 37: 4352-60.

Bartosch, S., A. Fite, G. T. Macfarlane, and M. E. T. McMurdo. 2004. 'Characterization of bacterial communities in feces from healthy elderly volunteers and hospitalized elderly patients by using real-time PCR and effects of antibiotic treatment on the fecal microbiota', *Applied and Environmental Microbiology*, 70: 3575-81.

Bashiardes, S., G. Zilberman-Schapira, and E. Elinav. 2016. 'Use of Metatranscriptomics in Microbiome Research', *Bioinform Biol Insights*, 10: 19-25.

Beckonert, O., H. C. Keun, T. M. Ebbels, J. Bundy, E. Holmes, J. C. Lindon, and J. K. Nicholson. 2007. 'Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts', *Nat Protoc*, 2: 2692-703.

Bik, E. M., P. B. Eckburg, S. R. Gill, K. E. Nelson, E. A. Purdom, F. Francois, G. Perez-Perez, M. J. Blaser, and D. A. Relman. 2006. 'Molecular analysis of the bacterial microbiota in the human stomach', *Proc Natl Acad Sci U S A*, 103: 732-7.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30: 2114-20.

Bonnet, M., E. Buc, P. Sauvanet, C. Darcha, D. Dubois, B. Pereira, P. Dechelotte, R. Bonnet, D. Pezet, and A. Darfeuille-Michaud. 2014. 'Colonization of the human gut by E. coli and colorectal cancer risk', *Clin Cancer Res*, 20: 859-67.

Bonnet, R., A. Suau, J. Dore, G. R. Gibson, and M. D. Collins. 2002. 'Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs', *Int J Syst Evol Microbiol*, 52: 757-63.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. 'QIIME allows analysis of high-throughput community sequencing data', *Nat Methods*, 7: 335-6.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight. 2011. 'Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample', *Proc Natl Acad Sci U S A*, 108 Suppl 1: 4516-22.

Carey, C. M., J. L. Kirk, S. Ojha, and M. Kostrzynska. 2007. 'Current and future uses of real-time polymerase chain reaction and microarrays in the study of intestinal microbiota, and probiotic use and effectiveness', *Can J Microbiol*, 53: 537-50.

Castano-Rodriguez, N., K. L. Goh, K. M. Fock, H. M. Mitchell, and N. O. Kaakoush. 2017. 'Dysbiosis of the microbiome in gastric carcinogenesis', *Sci Rep*, 7: 15957.

Chao, Anne. 1984. 'Nonparametric Estimation of the Number of Classes in a Population', *Scandinavian Journal of Statistics*, 11: 265-70.

Chao, Anne, and Shen-Ming Lee. 1992. 'Estimating the Number of Classes via Sample Coverage', *Journal of the American Statistical Association*, 87: 210-17.

Cho, I., and M. J. Blaser. 2012. 'The human microbiome: at the interface of health and disease', *Nature Reviews Genetics*, 13: 260-70.

Claesson, M. J., A. G. Clooney, and P. W. O'Toole. 2017. 'A clinician's guide to microbiome analysis', *Nat Rev Gastroenterol Hepatol*, 14: 585-95.

Claesson, M. J., Q. Wang, O. O'Sullivan, R. Greene-Diniz, J. R. Cole, R. P. Ross, and P. W. O'Toole. 2010. 'Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions', *Nucleic Acids Res*, 38: e200.

Clooney, A. G., F. Fouhy, R. D. Sleator, O' Driscoll A, C. Stanton, P. D. Cotter, and M. J. Claesson. 2016. 'Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis', *PLoS One*, 11: e0148028.

Coker, O. O., Z. Dai, Y. Nie, G. Zhao, L. Cao, G. Nakatsu, W. K. Wu, S. H. Wong, Z. Chen, J. J. Y. Sung, and J. Yu. 2018. 'Mucosal microbiome dysbiosis in gastric carcinogenesis', *Gut*, 67: 1024-32.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. 2009. 'The Ribosomal Database Project: improved alignments and new tools for rRNA analysis', *Nucleic Acids Res*, 37: D141-5.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. 2005. 'Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research', *Bioinformatics*, 21: 3674-6.

Consortium, UniProt. 2014. 'Activities at the Universal Protein Resource (UniProt)', *Nucleic Acids Res*, 42: D191-8.

Correa, P. 1992. 'Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention', *Cancer Res*, 52: 6735-40.

Correa, P., W. Haenszel, C. Cuello, S. Tannenbaum, and M. Archer. 1975. 'A model for gastric cancer epidemiology', *Lancet*, 2: 58-60.

Costantini, L., S. Magno, D. Albanese, C. Donati, R. Molinari, A. Filippone, R. Masetti, and N. Merendino. 2018. 'Characterization of human breast tissue microbiota from core needle biopsies through the analysis of multi hypervariable 16S-rRNA gene regions', *Sci Rep*, 8: 16893.

D'Elia, L., G. Rossi, R. Ippolito, F. P. Cappuccio, and P. Strazzullo. 2012. 'Habitual salt intake and risk of gastric cancer: a meta-analysis of prospective studies', *Clin Nutr*, 31: 489-98.

Delgado, S., R. Cabrera-Rubio, A. Mira, A. Suarez, and B. Mayo. 2013. 'Microbiological survey of the human gastric ecosystem using culturing and pyrosequencing methods', *Microb Ecol*, 65: 763-72.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. 'Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB', *Appl Environ Microbiol*, 72: 5069-72.

Diaz Heijtz, R., S. Wang, F. Anuar, Y. Qian, B. Bjorkholm, A. Samuelsson, M. L. Hibberd, H. Forssberg, and S. Pettersson. 2011. 'Normal gut microbiota modulates brain development and behavior', *Proc Natl Acad Sci U S A*, 108: 3047-52.

Dicksved, J., M. Lindberg, M. Rosenquist, H. Enroth, J. K. Jansson, and L. Engstrand. 2009. 'Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls', *J Med Microbiol*, 58: 509-16.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. 2005. 'Diversity of the human intestinal microbial flora', *Science*, 308: 1635-8.

Edgar, R. C. 2010. 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, 26: 2460-1.

Edgar, R. C. 2013. 'UPARSE: highly accurate OTU sequences from microbial amplicon reads', *Nat Methods*, 10: 996-8.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. 2011. 'UCHIME improves sensitivity and speed of chimera detection', *Bioinformatics*, 27: 2194-200.

Edgar, Robert C. 2016. 'UCHIME2: improved chimera prediction for amplicon sequencing', *bioRxiv*: 074252.

Edri, S., and T. Tuller. 2014. 'Quantifying the effect of ribosomal density on mRNA stability', *PLoS One*, 9: e102308.

Eisenhofer, R., J. J. Minich, C. Marotz, A. Cooper, R. Knight, and L. S. Weyrich. 2019. 'Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations', *Trends Microbiol*, 27: 105-17.

El-Omar, E. M., M. Carrington, W. H. Chow, K. E. McColl, J. H. Bream, H. A. Young, J. Herrera, J. Lissowska, C. C. Yuan, N. Rothman, G. Lanyon, M. Martin, J. F. Fraumeni, Jr., and C. S. Rabkin. 2001. 'The role of interleukin-1 polymorphisms in the pathogenesis of gastric cancer', *Nature*, 412: 99.

Erickson, A. R., B. L. Cantarel, R. Lamendella, Y. Darzi, E. F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N. C. Verberkmoes, C. M. Fraser, R. L. Hettich, and J. K. Jansson. 2012. 'Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease', *PLoS One*, 7: e49138.

Eun, C. S., B. K. Kim, D. S. Han, S. Y. Kim, K. M. Kim, B. Y. Choi, K. S. Song, Y. S. Kim, and J. F. Kim. 2014. 'Differences in gastric mucosal microbiota profiling in patients with chronic gastritis, intestinal metaplasia, and gastric cancer using pyrosequencing methods', *Helicobacter*, 19: 407-16.

Ferlay, J., I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. 2015. 'Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012', *Int J Cancer*, 136: E359-86.

Ferreira, R. M., J. C. Machado, and C. Figueiredo. 2014. 'Clinical relevance of Helicobacter pylori vacA and cagA genotypes in gastric carcinoma', *Best Pract Res Clin Gastroenterol*, 28: 1003-15.

Ferreira, R. M., J. Pereira-Marques, I. Pinto-Ribeiro, J. L. Costa, F. Carneiro, J. C. Machado, and C. Figueiredo. 2018. 'Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota', *Gut*, 67: 226-36.

Ferretti, P., S. Farina, M. Cristofolini, G. Girolomoni, A. Tett, and N. Segata. 2017. 'Experimental metagenomics and ribosomal profiling of the human skin microbiome', *Exp Dermatol*, 26: 211-19.

Figueiredo, C., J. C. Machado, P. Pharoah, R. Seruca, S. Sousa, R. Carvalho, A. F. Capelinha, W. Quint, C. Caldas, L. J. van Doorn, F. Carneiro, and M. Sobrinho-Simoes. 2002. 'Helicobacter pylori and interleukin 1 genotyping: an opportunity to identify high-risk individuals for gastric carcinoma', *J Natl Cancer Inst*, 94: 1680-7.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta. 2014. 'Pfam: the protein families database', *Nucleic Acids Res*, 42: D222-30.

Ford, A. C., D. Forman, R. Hunt, Y. Yuan, and P. Moayyedi. 2015. 'Helicobacter pylori eradication for the prevention of gastric neoplasia', *Cochrane Database Syst Rev*: CD005583.

Forsythe, S. J., J. M. Dolby, A. D. Webster, and J. A. Cole. 1988. 'Nitrate- and nitrite-reducing bacteria in the achlorhydric stomach', *J Med Microbiol*, 25: 253-9.

Fraher, M. H., P. W. O'Toole, and E. M. Quigley. 2012. 'Techniques used to characterize the gut microbiota: a guide for the clinician', *Nat Rev Gastroenterol Hepatol*, 9: 312-22.

Frank, D. N., A. L. St Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace. 2007. 'Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases', *Proc Natl Acad Sci U S A*, 104: 13780-5.

Franzosa, E. A., L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, and C. Huttenhower. 2018. 'Species-level functional profiling of metagenomes and metatranscriptomes', *Nat Methods*, 15: 962-68.

Franzosa, E. A., X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, and C. Huttenhower. 2014. 'Relating the metatranscriptome and metagenome of the human gut', *Proc Natl Acad Sci U S A*, 111: E2329-38.

Gevers, D., S. Kugathasan, L. A. Denson, Y. Vazquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, X. C. Morgan, A. D. Kostic, C. Luo, A. Gonzalez, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, and R. J. Xavier. 2014. 'The treatment-naive microbiome in new-onset Crohn's disease', *Cell Host Microbe*, 15: 382-92.

Giannoukos, G., D. M. Ciulla, K. Huang, B. J. Haas, J. Izard, J. Z. Levin, J. Livny, A. M. Earl, D. Gevers, D. V. Ward, C. Nusbaum, B. W. Birren, and A. Gnirke. 2012. 'Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes', *Genome Biol*, 13: R23.

Gilbert, J. A., M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight. 2018. 'Current understanding of the human microbiome', *Nat Med*, 24: 392-400.

Gonzalez, C. A., C. Figueiredo, C. B. Lic, R. M. Ferreira, M. L. Pardo, J. M. Ruiz Liso, P. Alonso, N. Sala, G. Capella, and J. M. Sanz-Anquela. 2011. 'Helicobacter pylori cagA and vacA genotypes as predictors of progression of gastric preneoplastic lesions: a long-term follow-up in a high-risk area in Spain', *Am J Gastroenterol*, 106: 867-74.

Goodacre, R., S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. 2004. 'Metabolomics by numbers: acquiring and understanding global metabolite data', *Trends Biotechnol*, 22: 245-52.

Goodwin, A. C., C. E. Destefano Shields, S. Wu, D. L. Huso, X. Wu, T. R. Murray-Stewart, A. Hacker-Prietz, S. Rabizadeh, P. M. Woster, C. L. Sears, and R. A. Casero, Jr. 2011. 'Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis', *Proc Natl Acad Sci U S A*, 108: 15354-9.

Goryshin, I. Y., J. A. Miller, Y. V. Kil, V. A. Lanzov, and W. S. Reznikoff. 1998. 'Tn5/IS50 target recognition', *Proc Natl Acad Sci U S A*, 95: 10716-21.

Gosalbes, M. J., A. Durban, M. Pignatelli, J. J. Abellan, N. Jimenez-Hernandez, A. E. Perez-Cobas, A. Latorre, and A. Moya. 2011. 'Metatranscriptomic approach to analyze the functional human gut microbiota', *PLoS One*, 6: e17447.

Haas, B. J., D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, D. Ciulla, D. Tabbaa, S. K. Highlander, E. Sodergren, B. Methe, T. Z. DeSantis, J. F. Petrosino, R. Knight, and B. W. Birren. 2011. 'Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons', *Genome Res*, 21: 494-504.

Harmsen, H. J., G. R. Gibson, P. Elfferich, G. C. Raangs, A. C. Wildeboer-Veloo, A. Argaiz, M. B. Roberfroid, and G. W. Welling. 2000. 'Comparison of viable cell counts and fluorescence in situ hybridization using specific rRNA-based probes for the quantification of human fecal bacteria', *FEMS Microbiol Lett*, 183: 125-9.

Hasan, M. R., A. Rawat, P. Tang, P. V. Jithesh, E. Thomas, R. Tan, and P. Tilley. 2016. 'Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing', *J Clin Microbiol*, 54: 919-27.

Helicobacter, and Group Cancer Collaborative. 2001. 'Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts', *Gut*, 49: 347-53.

Hoffmann, C., S. Dollive, S. Grunberg, J. Chen, H. Li, G. D. Wu, J. D. Lewis, and F. D. Bushman. 2013. 'Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents', *PLoS One*, 8: e66019.

Horz, H. P., M. E. Vianna, B. P. Gomes, and G. Conrads. 2005. 'Evaluation of universal probes and primer sets for assessing total bacterial load in clinical samples: general

implications and practical use in endodontic antimicrobial therapy', *J Clin Microbiol*, 43: 5332-7.

Hsieh, Y. Y., S. Y. Tung, H. Y. Pan, C. W. Yen, H. W. Xu, Y. J. Lin, Y. F. Deng, W. T. Hsu, C. S. Wu, and C. Li. 2018. 'Increased Abundance of Clostridium and Fusobacterium in Gastric Microbiota of Patients with Gastric Cancer in Taiwan', *Sci Rep*, 8: 158.

Hu, Y. L., W. Pang, Y. Huang, Y. Zhang, and C. J. Zhang. 2018. 'The Gastric Microbiome Is Perturbed in Advanced Gastric Adenocarcinoma Identified Through Shotgun Metagenomics', *Front Cell Infect Microbiol*, 8: 433.

Human Microbiome Jumpstart Reference Strains, Consortium, K. E. Nelson, G. M. Weinstock, S. K. Highlander, K. C. Worley, H. H. Creasy, J. R. Wortman, D. B. Rusch, M. Mitreva, E. Sodergren, A. T. Chinwalla, M. Feldgarden, D. Gevers, B. J. Haas, R. Madupu, D. V. Ward, B. W. Birren, R. A. Gibbs, B. Methe, J. F. Petrosino, R. L. Strausberg, G. G. Sutton, O. R. White, R. K. Wilson, S. Durkin, M. G. Giglio, S. Gujja, C. Howarth, C. D. Kodira, N. Kyrpides, T. Mehta, D. M. Muzny, M. Pearson, K. Pepin, A. Pati, X. Qin, C. Yandava, Q. Zeng, L. Zhang, A. M. Berlin, L. Chen, T. A. Hepburn, J. Johnson, J. McCorrison, J. Miller, P. Minx, C. Nusbaum, C. Russ, S. M. Sykes, C. M. Tomlinson, S. Young, W. C. Warren, J. Badger, J. Crabtree, V. M. Markowitz, J. Orvis, A. Cree, S. Ferriera, L. L. Fulton, R. S. Fulton, M. Gillis, L. D. Hemphill, V. Joshi, C. Kovar, M. Torralba, K. A. Wetterstrand, A. Abouellleil, A. M. Wollam, C. J. Buhay, Y. Ding, S. Dugan, M. G. FitzGerald, M. Holder, J. Hostetler, S. W. Clifton, E. Allen-Vercoe, A. M. Earl, C. N. Farmer, K. Liolios, M. G. Surette, Q. Xu, C. Pohl, K. Wilczek-Boney, and D. Zhu. 2010. 'A catalog of reference genomes from the human microbiome', *Science*, 328: 994-9.

Human Microbiome Project, Consortium. 2012a. 'A framework for human microbiome research', *Nature*, 486: 215-21.

Human Microbiome Project, Consortium. 2012b. 'Structure, function and diversity of the healthy human microbiome', *Nature*, 486: 207-14.

IARC. 1994. 'Schistosomes, liver flukes and Helicobacter pylori. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Lyon, 7-14 June 1994', *IARC Monogr Eval Carcinog Risks Hum*, 61: 1-241.

Imhann, F., M. J. Bonder, A. Vich Vila, J. Fu, Z. Mujagic, L. Vork, E. F. Tigchelaar, S. A. Jankipersadsing, M. C. Cenit, H. J. Harmsen, G. Dijkstra, L. Franke, R. J. Xavier, D. Jonkers, C. Wijmenga, R. K. Weersma, and A. Zhernakova. 2016. 'Proton pump inhibitors affect the gut microbiome', *Gut*, 65: 740-8.

Ingham, C. J., A. Sprenkels, J. Bomer, D. Molenaar, A. van den Berg, J. E. van Hylckama Vlieg, and W. M. de Vos. 2007. 'The micro-Petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms', *Proc Natl Acad Sci U S A*, 104: 18217-22.

Jackson, M. A., J. K. Goodrich, M. E. Maxan, D. E. Freedberg, J. A. Abrams, A. C. Poole, J. L. Sutter, D. Welter, R. E. Ley, J. T. Bell, T. D. Spector, and C. J. Steves. 2016. 'Proton pump inhibitors alter the composition of the gut microbiota', *Gut*, 65: 749-56.

Jalanka-Tuovinen, J., A. Salonen, J. Nikkila, O. Immonen, R. Kekkonen, L. Lahti, A. Palva, and W. M. de Vos. 2011. 'Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms', *PLoS One*, 6: e23035.

Jiao, Y. S., H. Yan, Z. J. Ji, Y. H. Liu, X. H. Sui, X. X. Zhang, E. T. Wang, W. X. Chen, and W. F. Chen. 2015. 'Phyllobacterium sophorae sp. nov., a symbiotic bacterium isolated from root nodules of Sophora flavescens', *Int J Syst Evol Microbiol*, 65: 399-406.

Johnson, C. H., J. Ivanisevic, and G. Siuzdak. 2016. 'Metabolomics: beyond biomarkers and towards mechanisms', *Nat Rev Mol Cell Biol*, 17: 451-9.

Johnson, C. H., A. D. Patterson, J. R. Idle, and F. J. Gonzalez. 2012. 'Xenobiotic metabolomics: major impact on the metabolome', *Annu Rev Pharmacol Toxicol*, 52: 37-56.

Jones, M. B., S. K. Highlander, E. L. Anderson, W. Li, M. Dayrit, N. Klitgord, M. M. Fabani, V. Seguritan, J. Green, D. T. Pride, S. Yooseph, W. Biggs, K. E. Nelson, and J. C. Venter. 2015. 'Library preparation methodology can influence genomic and functional predictions in human microbiome research', *Proc Natl Acad Sci U S A*, 112: 14024-9.

Jovel, J., J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. L. Mason, K. L. Madsen, and G. K. Wong. 2016. 'Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics', *Front Microbiol*, 7: 459.

Jumpstart Consortium Human Microbiome Project Data Generation Working, Group. 2012. 'Evaluation of 16S rDNA-based community profiling for human microbiome research', *PLoS One*, 7: e39315.

Kaeberlein, T., K. Lewis, and S. S. Epstein. 2002. 'Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment', *Science*, 296: 1127-9.

Kanehisa, M., and S. Goto. 2000. 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic Acids Res*, 28: 27-30.

Keegan, K. P., E. M. Glass, and F. Meyer. 2016. 'MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function', *Methods Mol Biol*, 1399: 207-33.

Kelly, C. P., and J. T. LaMont. 2008. 'Clostridium difficile--more difficult than ever', *N Engl J Med*, 359: 1932-40.

Khosravi Y, Gan HM, Chia PJY, et al. . 2017. 'The Gastric Microbiome of Four Malaysian Gastroduodenal Disease Patients', *Arch Gene Genome Res*, 1: 1-9.

Kienesberger, S., L. M. Cox, A. Livanos, X. S. Zhang, J. Chung, G. I. Perez-Perez, G. Gorkiewicz, E. L. Zechner, and M. J. Blaser. 2016. 'Gastric Helicobacter pylori Infection Affects Local and Distant Microbial Populations and Host Responses', *Cell Rep*, 14: 1395-407.

Klymiuk, I., C. Bilgilier, A. Stadlmann, J. Thannesberger, M. T. Kastner, C. Hogenauer, A. Puspok, S. Biowski-Frotz, C. Schrutka-Kolbl, G. G. Thallinger, and C. Steininger. 2017. 'The Human Gastric Microbiome Is Predicated upon Infection with Helicobacter pylori', *Front Microbiol*, 8: 2508.

Knight, R., C. Callewaert, C. Marotz, E. R. Hyde, J. W. Debelius, D. McDonald, and M. L. Sogin. 2017. 'The Microbiome and Human Biology', *Annu Rev Genomics Hum Genet*, 18: 65-86.

Knight, R., J. Jansson, D. Field, N. Fierer, N. Desai, J. A. Fuhrman, P. Hugenholtz, D. van der Lelie, F. Meyer, R. Stevens, M. J. Bailey, J. I. Gordon, G. A. Kowalchuk, and J. A. Gilbert. 2012. 'Unlocking the potential of metagenomics through replicated experimental design', *Nat Biotechnol*, 30: 513-20.

Knight, R., A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L. I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein. 2018. 'Best practices for analysing microbiomes', *Nat Rev Microbiol*, 16: 410-22.

Kopylova, E., J. A. Navas-Molina, C. Mercier, Z. Z. Xu, F. Mahe, Y. He, H. W. Zhou, T. Rognes, J. G. Caporaso, and R. Knight. 2016. 'Open-Source Sequence Clustering Methods Improve the State Of the Art', *mSystems*, 1.

Kostic, A. D., E. Chun, L. Robertson, J. N. Glickman, C. A. Gallini, M. Michaud, T. E. Clancy, D. C. Chung, P. Lochhead, G. L. Hold, E. M. El-Omar, D. Brenner, C. S. Fuchs, M. Meyerson, and W. S. Garrett. 2013. 'Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment', *Cell Host Microbe*, 14: 207-15.

Kostic, A. D., D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero, J. Baselga, C. Liu, R. A. Shivdasani, S. Ogino, B. W. Birren, C. Huttenhower, W. S. Garrett, and M. Meyerson. 2012. 'Genomic analysis identifies association of Fusobacterium with colorectal carcinoma', *Genome Res*, 22: 292-8.

Kuipers, E. J. 1998. 'Review article: Relationship between Helicobacter pylori, atrophic gastritis and gastric cancer', *Aliment Pharmacol Ther*, 12 Suppl 1: 25-36.

Lagier, J. C., G. Dubourg, M. Million, F. Cadoret, M. Bilen, F. Fenollar, A. Levasseur, J. M. Rolain, P. E. Fournier, and D. Raoult. 2018. 'Culturing the human microbiota and culturomics', *Nat Rev Microbiol*: 540-50.

Lagier, J. C., P. Hugon, S. Khelaifia, P. E. Fournier, B. La Scola, and D. Raoult. 2015. 'The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota', *Clin Microbiol Rev*, 28: 237-64.

Lagier, J. C., S. Khelaifia, M. T. Alou, S. Ndongo, N. Dione, P. Hugon, A. Caputo, F. Cadoret, S. I. Traore, E. H. Seck, G. Dubourg, G. Durand, G. Mourembou, E. Guilhot, A. Togo, S. Bellali, D. Bachar, N. Cassir, F. Bittar, J. Delerce, M. Mailhe, D. Ricaboni, M. Bilen, N. P. Dangui Nieko, N. M. Dia Badiane, C. Valles, D. Mouelhi, K. Diop, M. Million, D. Musso, J. Abrahao, E. I. Azhar, F. Bibi, M. Yasir, A. Diallo, C. Sokhna, F. Djossou, V. Vitton, C. Robert, J. M. Rolain, B. La Scola, P. E. Fournier, A. Levasseur, and D. Raoult. 2016. 'Culture of previously uncultured members of the human gut microbiota by culturomics', *Nat Microbiol*, 1: 16203.

Lagier, J. C., M. Million, P. Hugon, F. Armougom, and D. Raoult. 2012. 'Human gut microbiota: repertoire and variations', *Front Cell Infect Microbiol*, 2: 136.

Lamichhane, S., C. C. Yde, M. S. Schmedes, H. M. Jensen, S. Meier, and H. C. Bertram. 2015. 'Strategy for Nuclear-Magnetic-Resonance-Based Metabolomics of Human Feces', *Anal Chem*, 87: 5930-7.

Langendijk, P. S., F. Schut, G. J. Jansen, G. C. Raangs, G. R. Kamphuis, M. H. Wilkinson, and G. W. Welling. 1995. 'Quantitative fluorescence in situ hybridization of Bifidobacterium spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples', *Appl Environ Microbiol*, 61: 3069-75.

Langille, M. G., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. 2013. 'Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences', *Nat Biotechnol*, 31: 814-21.

Langmead, B., and S. L. Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nat Methods*, 9: 357-9.

Laursen, M. F., M. D. Dalgaard, and M. I. Bahl. 2017. 'Genomic GC-Content Affects the Accuracy of 16S rRNA Gene Sequencing Based Microbial Profiling due to PCR Bias', *Frontiers in Microbiology*, 8: 1934.

Lazarevic, V., N. Gaia, M. Girard, S. Leo, A. Cherkaoui, G. Renzi, S. Emonet, S. Jamme, E. Ruppe, S. Vijgen, L. Rubbia-Brandt, C. Toso, and J. Schrenzel. 2018. 'When Bacterial Culture Fails, Metagenomics Can Help: A Case of Chronic Hepatic Brucelloma Assessed by Next-Generation Sequencing', *Front Microbiol*, 9: 1566.

Leo, S., N. Gaia, E. Ruppe, S. Emonet, M. Girard, V. Lazarevic, and J. Schrenzel. 2017. 'Detection of Bacterial Pathogens from Broncho-Alveolar Lavage by Next-Generation Sequencing', *Int J Mol Sci*, 18.

Lepage, P., R. Hasler, M. E. Spehlmann, A. Rehman, A. Zvirbliene, A. Begun, S. Ott, L. Kupcinskas, J. Dore, A. Raedler, and S. Schreiber. 2011. 'Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis', *Gastroenterology*, 141: 227-36.

Lertpiriyapong, K., M. T. Whary, S. Muthupalani, J. L. Lofgren, E. R. Gamazon, Y. Feng, Z. Ge, T. C. Wang, and J. G. Fox. 2014. 'Gastric colonisation with a restricted commensal microbiota replicates the promotion of neoplastic lesions by diverse intestinal microbiota in the Helicobacter pylori INS-GAS mouse model of gastric carcinogenesis', *Gut*, 63: 54-63.

Ley, R. E., P. J. Turnbaugh, S. Klein, and J. I. Gordon. 2006. 'Microbial ecology: human gut microbes associated with obesity', *Nature*, 444: 1022-3.

Li, D., C. M. Liu, R. Luo, K. Sadakane, and T. W. Lam. 2015. 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics*, 31: 1674-6.

Li, T. H., Y. Qin, P. C. Sham, K. S. Lau, K. M. Chu, and W. K. Leung. 2017. 'Alterations in Gastric Microbiota After H. Pylori Eradication and in Different Histological Stages of Gastric Carcinogenesis', *Sci Rep*, 7: 44935.

Li, X. X., G. L. Wong, K. F. To, V. W. Wong, L. H. Lai, D. K. Chow, J. Y. Lau, J. J. Sung, and C. Ding. 2009. 'Bacterial microbiota profiling in gastritis without Helicobacter pylori infection or non-steroidal anti-inflammatory drug use', *PLoS One*, 4: e7985.

Lloyd-Price, J., A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A. B. Hall, A. Brady, H. H. Creasy, C. McCracken, M. G. Giglio, D. McDonald, E. A. Franzosa, R. Knight, O. White, and C. Huttenhower. 2017. 'Strains, functions and dynamics in the expanded Human Microbiome Project', *Nature*, 550: 61-66.

Lofgren, J. L., M. T. Whary, Z. Ge, S. Muthupalani, N. S. Taylor, M. Mobley, A. Potter, A. Varro, D. Eibach, S. Suerbaum, T. C. Wang, and J. G. Fox. 2011. 'Lack of commensal flora in Helicobacter pylori-infected INS-GAS mice reduces gastritis and delays intraepithelial neoplasia', *Gastroenterology*, 140: 210-20.

Lozupone, C., M. Hamady, and R. Knight. 2006. 'UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context', *BMC Bioinformatics*, 7: 371.

Lozupone, C., M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight. 2011. 'UniFrac: an effective distance metric for microbial community comparison', *ISME J*, 5: 169-72.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, and J. Wang. 2012. 'SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler', *Gigascience*, 1: 18.

Machado, J. C., C. Figueiredo, P. Canedo, P. Pharoah, R. Carvalho, S. Nabais, C. Castro Alves, M. L. Campos, L. J. Van Doorn, C. Caldas, R. Seruca, F. Carneiro, and M. Sobrinho-Simoes. 2003. 'A proinflammatory genetic profile increases the risk for chronic atrophic gastritis and gastric carcinoma', *Gastroenterology*, 125: 364-71.

Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. 2001. 'The RDP-II (Ribosomal Database Project)', *Nucleic Acids Res*, 29: 173-4.

Maldonado-Contreras, A., K. C. Goldfarb, F. Godoy-Vitorino, U. Karaoz, M. Contreras, M. J. Blaser, E. L. Brodie, and M. G. Dominguez-Bello. 2011. 'Structure of the human gastric bacterial community in relation to Helicobacter pylori status', *ISME J*, 5: 574-9.

Malinen, E., A. Kassinen, T. Rinttila, and A. Palva. 2003. 'Comparison of real-time PCR with SYBR Green I or 5'-nuclease assays and dot-blot hybridization with rDNA-targeted oligonucleotide probes in quantification of selected faecal bacteria', *Microbiology*, 149: 269-77.

Marchesi, J. R., and J. Ravel. 2015. 'The vocabulary of microbiome research: a proposal', *Microbiome*, 3: 31.

Mariat, D., O. Firmesse, F. Levenez, V. Guimaraes, H. Sokol, J. Dore, G. Corthier, and J. P. Furet. 2009. 'The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age', *BMC Microbiol*, 9: 123.

Marotz, C. A., J. G. Sanders, C. Zuniga, L. S. Zaramela, R. Knight, and K. Zengler. 2018. 'Improving saliva shotgun metagenomics by chemical host DNA depletion', *Microbiome*, 6: 42.

Martin, M. E., S. Bhatnagar, M. D. George, B. J. Paster, D. R. Canfield, J. A. Eisen, and J. V. Solnick. 2013. 'The impact of Helicobacter pylori infection on the gastric microbiota of the rhesus macaque', *PLoS One*, 8: e76375.

Martinsen, T. C., K. Bergh, and H. L. Waldum. 2005. 'Gastric juice: a barrier against infectious diseases', *Basic Clin Pharmacol Toxicol*, 96: 94-102.

Melnik, A. V., R. R. da Silva, E. R. Hyde, A. A. Aksenov, F. Vargas, A. Bouslimani, I. Protsyuk, A. K. Jarmusch, A. Tripathi, T. Alexandrov, R. Knight, and P. C. Dorrestein. 2017. 'Coupling Targeted and Untargeted Mass Spectrometry for

Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples', *Anal Chem*, 89: 7549-59.

Mera, R. M., L. E. Bravo, M. C. Camargo, J. C. Bravo, A. G. Delgado, J. Romero-Gallo, M. C. Yepez, J. L. Realpe, B. G. Schneider, D. R. Morgan, R. M. Peek, Jr., P. Correa, K. T. Wilson, and M. B. Piazuelo. 2018. 'Dynamics of Helicobacter pylori infection as a determinant of progression of gastric precancerous lesions: 16-year follow-up of an eradication trial', *Gut*, 67: 1239-46.

Monstein, H. J., A. Tiveljung, C. H. Kraft, K. Borch, and J. Jonasson. 2000. 'Profiling of bacterial flora in gastric biopsies from patients with Helicobacter pylori-associated gastritis and histologically normal control individuals by temperature gradient gel electrophoresis and 16S rDNA sequence analysis', *J Med Microbiol*, 49: 817-22.

Moore, W. E., and L. V. Holdeman. 1974. 'Human fecal flora: the normal flora of 20 Japanese-Hawaiians', *Appl Microbiol*, 27: 961-79.

Morgan, X. C., and C. Huttenhower. 2012. 'Chapter 12: Human microbiome analysis', *PLoS Comput Biol*, 8: e1002808.

Morgan, X.C, and C. Huttenhower. 2014. 'Meta'omic analytic techniques for studying the intestinal microbiome', *Gastroenterology*, 146: 1437-48 e1.

Morgan, X. C., T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier, and C. Huttenhower. 2012. 'Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment', *Genome Biol*, 13: R79.

Mowat, C., C. Williams, D. Gillen, M. Hossack, D. Gilmour, A. Carswell, A. Wirz, T. Preston, and K. E. McColl. 2000. 'Omeprazole, Helicobacter pylori status, and alterations in the intragastric milieu facilitating bacterial N-nitrosation', *Gastroenterology*, 119: 339-47.

Mukherjee, S., R. Seshadri, N. J. Varghese, E. A. Eloe-Fadrosh, J. P. Meier-Kolthoff, M. Goker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H. P. Klenk, and N. C. Kyrpides. 2017. '1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life', *Nat Biotechnol*, 35: 676-83.

Muth, T., F. Kohrs, R. Heyer, D. Benndorf, E. Rapp, U. Reichl, L. Martens, and B. Y. Renard. 2018. 'MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go', *Anal Chem*, 90: 685-89.

Muth, T., B. Y. Renard, and L. Martens. 2016. 'Metaproteomic data analysis at a glance: advances in computational microbial community proteomics', *Expert Rev Proteomics*, 13: 757-69.

Nadkarni, M. A., F. E. Martin, N. A. Jacques, and N. Hunter. 2002. 'Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set', *Microbiology-Sgm*, 148: 257-66.

Namiki, T., T. Hachiya, H. Tanaka, and Y. Sakakibara. 2012. 'MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads', *Nucleic Acids Res*, 40: e155.

Neefs, J. M., Y. Van de Peer, P. De Rijk, S. Chapelle, and R. De Wachter. 1993. 'Compilation of small ribosomal subunit RNA structures', *Nucleic Acids Res*, 21: 3025-49.

Newman, J. V., T. Kosaka, B. J. Sheppard, J. G. Fox, and D. B. Schauer. 2001. 'Bacterial infection promotes colon tumorigenesis in Apc(Min/+) mice', *J Infect Dis*, 184: 227-30.

Nichols, D., N. Cahoon, E. M. Trakhtenberg, L. Pham, A. Mehta, A. Belanger, T. Kanigan, K. Lewis, and S. S. Epstein. 2010. 'Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species', *Appl Environ Microbiol*, 76: 2445-50.

Nicholson, J. K., and J. C. Lindon. 2008. 'Systems biology: Metabonomics', *Nature*, 455: 1054-6.

Nilsson, R. H., E. Kristiansson, M. Ryberg, N. Hallenberg, and K. H. Larsson. 2008. 'Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification', *Evol Bioinform Online*, 4: 193-201.

Nomura, A., G. N. Stemmermann, P. H. Chyou, G. I. Perez-Perez, and M. J. Blaser. 1994. 'Helicobacter pylori infection and the risk for duodenal and gastric ulceration', *Ann Intern Med*, 120: 977-81.

Ohnishi, N., H. Yuasa, S. Tanaka, H. Sawa, M. Miura, A. Matsui, H. Higashi, M. Musashi, K. Iwabuchi, M. Suzuki, G. Yamada, T. Azuma, and M. Hatakeyama. 2008. 'Transgenic expression of Helicobacter pylori CagA induces gastrointestinal and hematopoietic neoplasms in mouse', *Proc Natl Acad Sci U S A*, 105: 1003-8.

Osborn, A. M., E. R. Moore, and K. N. Timmis. 2000. 'An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics', *Environ Microbiol*, 2: 39-50.

Oyola, S. O., Y. Gu, M. Manske, T. D. Otto, J. O'Brien, D. Alcock, B. Macinnis, M. Berriman, C. I. Newbold, D. P. Kwiatkowski, H. P. Swerdlow, and M. A. Quail. 2013. 'Efficient depletion of host DNA contamination in malaria clinical sequencing', *J Clin Microbiol*, 51: 745-51.

Pace, N. R. 1997. 'A molecular view of microbial diversity and the biosphere', *Science*, 276: 734-40.

Paliy, O., H. Kenche, F. Abernathy, and S. Michail. 2009. 'High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray', *Appl Environ Microbiol*, 75: 3572-9.

Parks, D. H., and R. G. Beiko. 2010. 'Identifying biologically relevant differences between metagenomic communities', *Bioinformatics*, 26: 715-21.

Paroni Sterbini, F., A. Palladini, L. Masucci, C. V. Cannistraci, R. Pastorino, G. Ianiro, F. Bugli, C. Martini, W. Ricciardi, A. Gasbarrini, M. Sanguinetti, G. Cammarota, and B. Posteraro. 2016. 'Effects of Proton Pump Inhibitors on the Gastric Mucosa-Associated Microbiota in Dyspeptic Patients', *Appl Environ Microbiol*, 82: 6633-44.

Parsonnet, J., G. D. Friedman, D. P. Vandersteen, Y. Chang, J. H. Vogelman, N. Orentreich, and R. K. Sibley. 1991. 'Helicobacter pylori infection and the risk of gastric carcinoma', *N Engl J Med*, 325: 1127-31.

Parsonnet, J., S. Hansen, L. Rodriguez, A. B. Gelb, R. A. Warnke, E. Jellum, N. Orentreich, J. H. Vogelman, and G. D. Friedman. 1994. 'Helicobacter pylori infection and gastric lymphoma', *N Engl J Med*, 330: 1267-71.

Parsons, B. N., U. Z. Ijaz, R. D'Amore, M. D. Burkitt, R. Eccles, L. Lenzi, C. A. Duckworth, A. R. Moore, L. Tiszlavicz, A. Varro, N. Hall, and D. M. Pritchard. 2017. 'Comparison of the human gastric microbiota in hypochlorhydric states arising as a result of Helicobacter pylori-induced atrophic gastritis, autoimmune atrophic gastritis and proton pump inhibitor use', *PLoS Pathog*, 13: e1006653.

Peano, C., A. Pietrelli, C. Consolandi, E. Rossi, L. Petiti, L. Tagliabue, G. De Bellis, and P. Landini. 2013. 'An efficient rRNA removal method for RNA sequencing in GC-rich bacteria', *Microb Inform Exp*, 3: 1.

Pedersen, H. K., V. Gudmundsdottir, H. B. Nielsen, T. Hyotylainen, T. Nielsen, B. A. Jensen, K. Forslund, F. Hildebrand, E. Prifti, G. Falony, E. Le Chatelier, F. Levenez, J. Dore, I. Mattila, D. R. Plichta, P. Poho, L. I. Hellgren, M. Arumugam, S. Sunagawa, S. Vieira-Silva, T. Jorgensen, J. B. Holm, K. Trost, H. I. T. Consortium Meta, K. Kristiansen, S. Brix, J. Raes, J. Wang, T. Hansen, P. Bork, S. Brunak, M. Oresic, S. D. Ehrlich, and O. Pedersen. 2016. 'Human gut microbes impact host serum metabolome and insulin sensitivity', *Nature*, 535: 376-81.

Peng, Y., H. C. Leung, S. M. Yiu, and F. Y. Chin. 2011. 'Meta-IDBA: a de Novo assembler for metagenomic data', *Bioinformatics*, 27: i94-101.

Pereira-Marques, J., R. M. Ferreira, I. Pinto-Ribeiro, and C. Figueiredo. 2019. 'Helicobacter pylori Infection, the Gastric Microbiome and Gastric Cancer', *Adv Exp Med Biol*.

Persson, C., P. Canedo, J. C. Machado, E. M. El-Omar, and D. Forman. 2011. 'Polymorphisms in inflammatory response genes and their association with gastric cancer: A HuGE systematic review and meta-analyses', *Am J Epidemiol*, 173: 259-70.

Petrosino, J. F., S. Highlander, R. A. Luna, R. A. Gibbs, and J. Versalovic. 2009. 'Metagenomic pyrosequencing and microbial identification', *Clin Chem*, 55: 856-66.

Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. 'An Eulerian path approach to DNA fragment assembly', *Proc Natl Acad Sci U S A*, 98: 9748-53.

Plottel, C. S., and M. J. Blaser. 2011. 'Microbiome and malignancy', *Cell Host Microbe*, 10: 324-35.

Plummer, M., S. Franceschi, J. Vignat, D. Forman, and C. de Martel. 2015. 'Global burden of gastric cancer attributable to Helicobacter pylori', *Int J Cancer*, 136: 487-90.

Praud, D., M. Rota, C. Pelucchi, P. Bertuccio, T. Rosso, C. Galeone, Z. F. Zhang, K. Matsuo, H. Ito, J. Hu, K. C. Johnson, G. P. Yu, D. Palli, M. Ferraroni, J. Muscat, N. Lunet, B. Peleteiro, R. Malekzadeh, W. Ye, H. Song, D. Zaridze, D. Maximovitch, N. Aragones, G. Castano-Vinyals, J. Vioque, E. M. Navarrete-Munoz, M. Pakseresht, F. Pourfarzi, A. Wolk, N. Orsini, A. Bellavia, N. Hakansson, L. Mu, R. Pastorino, R. C. Kurtz, M. H. Derakhshan, A. Lagiou, P. Lagiou, P. Boffetta, S. Boccia, E. Negri, and C. La Vecchia. 2018. 'Cigarette smoking and gastric cancer in the Stomach Cancer Pooling (StoP) Project', *Eur J Cancer Prev*, 27: 124-33.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, H. I. T. Consortium Meta, P. Bork, S. D. Ehrlich, and J. Wang. 2010. 'A human gut microbial gene catalogue established by metagenomic sequencing', *Nature*, 464: 59-65.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner. 2013. 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools', *Nucleic Acids Res*, 41: D590-6.

Quince, C., A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. 2017. 'Shotgun metagenomics, from sampling to analysis', *Nat Biotechnol*, 35: 833-44.

Rajilic-Stojanovic, M., and W. M. de Vos. 2014. 'The first 1000 cultured species of the human gastrointestinal microbiota', *FEMS Microbiol Rev*, 38: 996-1047.

Rho, M., H. Tang, and Y. Ye. 2010. 'FragGeneScan: predicting genes in short and error-prone reads', *Nucleic Acids Res*, 38: e191.

Rinttila, T., A. Kassinen, E. Malinen, L. Krogius, and A. Palva. 2004. 'Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR', *J Appl Microbiol*, 97: 1166-77.

Rota, M., C. Pelucchi, P. Bertuccio, K. Matsuo, Z. F. Zhang, H. Ito, J. Hu, K. C. Johnson, D. Palli, M. Ferraroni, G. P. Yu, J. Muscat, N. Lunet, B. Peleteiro, W. Ye, H. Song, D. Zaridze, D. Maximovitch, M. Guevara, T. Fernandez-Villa, J. Vioque, E. M. Navarrete-Munoz, A. Wolk, N. Orsini, A. Bellavia, N. Hakansson, L. Mu, R. Persiani, R. C. Kurtz, A. Lagiou, P. Lagiou, C. Galeone, R. Bonzi, P. Boffetta, S. Boccia, E. Negri, and C. La Vecchia. 2017. 'Alcohol consumption and gastric cancer risk-A pooled analysis within the StoP project consortium', *Int J Cancer*, 141: 1950-62.

Rotmistrovsky, Kirill, and Richa Agarwala. 2018. *BMTagger: Best Match Tagger for removing human reads from metagenomics datasets*.

Russo, E., G. Bacci, C. Chiellini, C. Fagorzi, E. Niccolai, A. Taddei, F. Ricci, M. N. Ringressi, R. Borrelli, F. Melli, M. Miloeva, P. Bechi, A. Mengoni, R. Fani, and A. Amedei. 2017. 'Preliminary Comparison of Oral and Intestinal Human Microbiota in Patients with Colorectal Cancer: A Pilot Study', *Front Microbiol*, 8: 2699.

Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman, and A. W. Walker. 2014. 'Reagent and laboratory contamination can critically impact sequence-based microbiome analyses', *BMC Biol*, 12: 87.

Sanduleanu, S., D. Jonkers, A. De Bruine, W. Hameeteman, and R. W. Stockbrugger. 2001. 'Non-Helicobacter pylori bacterial flora during acid-suppressive therapy: differential findings in gastric juice and gastric mucosa', *Aliment Pharmacol Ther*, 15: 379-88.

Sarangi, A. N., A. Goel, and R. Aggarwal. 2019. 'Methods for Studying Gut Microbiota: A Primer for Physicians', *J Clin Exp Hepatol*, 9: 62-73.

Satokari, R. M., E. E. Vaughan, A. D. Akkermans, M. Saarela, and W. M. de Vos. 2001. 'Bifidobacterial diversity in human feces detected by genus-specific PCR and denaturing gradient gel electrophoresis', *Appl Environ Microbiol*, 67: 504-13.

Savage, D. C. 1977. 'Microbial ecology of the gastrointestinal tract', *Annu Rev Microbiol*, 31: 107-33.

Schirmer, M., E. A. Franzosa, J. Lloyd-Price, L. J. McIver, R. Schwager, T. W. Poon, A. N. Ananthakrishnan, E. Andrews, G. Barron, K. Lake, M. Prasad, J. Sauk, B. Stevens, R. G. Wilson, J. Braun, L. A. Denson, S. Kugathasan, D. P. B. McGovern, H. Vlamakis, R. J. Xavier, and C. Huttenhower. 2018. 'Dynamics of metatranscription in the inflammatory bowel disease gut microbiome', *Nat Microbiol*, 3: 337-46.

Schirmer, M., S. P. Smeekens, H. Vlamakis, M. Jaeger, M. Oosting, E. A. Franzosa, R. Ter Horst, T. Jansen, L. Jacobs, M. J. Bonder, A. Kurilshikov, J. Fu, L. A. B. Joosten, A. Zhernakova, C. Huttenhower, C. Wijmenga, M. G. Netea, and R. J. Xavier. 2016. 'Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity', *Cell*, 167: 1125-36 e8.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. 'Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities', *Appl Environ Microbiol*, 75: 7537-41.

Schulz, C., K. Schutte, N. Koch, R. Vilchez-Vargas, M. L. Wos-Oxley, A. P. A. Oxley, M. Vital, P. Malfertheiner, and D. H. Pieper. 2018. 'The active bacterial assemblages of the upper GI tract in individuals with and without Helicobacter infection', *Gut*, 67: 216-25.

Segal, J. P., B. H. Mullish, M. N. Quraishi, A. Acharjee, H. R. T. Williams, T. Iqbal, A. L. Hart, and J. R. Marchesi. 2019. 'The application of omics techniques to understand the role of the gut microbiota in inflammatory bowel disease', *Therap Adv Gastroenterol*, 12: 1756284818822250.

Segata, N., J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower. 2011. 'Metagenomic biomarker discovery and explanation', *Genome Biol*, 12: R60.

Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. 2012. 'Metagenomic microbial community profiling using unique clade-specific marker genes', *Nat Methods*, 9: 811-4.

Sekirov, I., S. L. Russell, L. C. Antunes, and B. B. Finlay. 2010. 'Gut microbiota in health and disease', *Physiol Rev*, 90: 859-904.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. 'Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans', *Cell*, 164: 337-40.

Seng, P., M. Drancourt, F. Gouriet, B. La Scola, P. E. Fournier, J. M. Rolain, and D. Raoult. 2009. 'Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry', *Clin Infect Dis*, 49: 543-51.

Siegwald, L., H. Touzet, Y. Lemoine, D. Hot, C. Audebert, and S. Caboche. 2017. 'Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics', *PLoS One*, 12: e0169563.

Sjostedt, S., A. Heimdahl, L. Kager, and C. E. Nord. 1985. 'Microbial colonization of the oropharynx, esophagus and stomach in patients with gastric diseases', *Eur J Clin Microbiol*, 4: 49-51.

Soergel, D. A., N. Dey, R. Knight, and S. E. Brenner. 2012. 'Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences', *ISME J*, 6: 1440-4.

Sokol, H., P. Seksik, J. P. Furet, O. Firmesse, I. Nion-Larmurier, L. Beaugerie, J. Cosnes, G. Corthier, P. Marteau, and J. Dore. 2009. 'Low counts of Faecalibacterium prausnitzii in colitis microbiota', *Inflamm Bowel Dis*, 15: 1183-9.

Stearns, J. C., M. D. Lynch, D. B. Senadheera, H. C. Tenenbaum, M. B. Goldberg, D. G. Cvitkovitch, K. Croitoru, G. Moreno-Hagelsieb, and J. D. Neufeld. 2011. 'Bacterial biogeography of the human digestive tract', *Sci Rep*, 1: 170.

Stein, A., T. E. Takasuka, and C. K. Collings. 2010. 'Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences?', *Nucleic Acids Res*, 38: 709-19.

Stockbruegger, R. W. 1985. 'Bacterial overgrowth as a consequence of reduced gastric acidity', *Scand J Gastroenterol Suppl*, 111: 7-16.

Stoddard, S. F., B. J. Smith, R. Hein, B. R. Roller, and T. M. Schmidt. 2015. 'rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development', *Nucleic Acids Res*, 43: D593-8.

Suau, A., R. Bonnet, M. Sutren, J. J. Godon, G. R. Gibson, M. D. Collins, and J. Dore. 1999. 'Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut', *Appl Environ Microbiol*, 65: 4799-807.

Sultan, M., V. Amstislavskiy, T. Risch, M. Schuette, S. Dokel, M. Ralser, D. Balzereit, H. Lehrach, and M. L. Yaspo. 2014. 'Influence of RNA extraction methods and library selection schemes on RNA-seq data', *BMC Genomics*, 15: 675.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. 'The COG database: a tool for genome-scale analysis of protein functions and evolution', *Nucleic Acids Res*, 28: 33-6.

Thompson, L. R., J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, J. A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vazquez-Baeza, A. Gonzalez, J. T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M. F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N. A. Bokulich, J. Lefler, C. J. Brislawn, G. Humphrey, S. M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J. A. Fuhrman, A. Clauset, R. L. Stevens, A. Shade, K. S. Pollard, K. D. Goodwin, J. K. Jansson, J. A. Gilbert, R. Knight, and Consortium Earth Microbiome Project. 2017. 'A communal catalogue reveals Earth's multiscale microbial diversity', *Nature*, 551: 457-63.

Thorell, K., J. Bengtsson-Palme, O. H. Liu, R. V. Palacios Gonzales, I. Nookaew, L. Rabeneck, L. Paszat, D. Y. Graham, J. Nielsen, S. B. Lundin, and A. Sjoling. 2017. 'In Vivo Analysis of the Viable Microbiota and Helicobacter pylori Transcriptome in Gastric Infection and Early Stages of Carcinogenesis', *Infect Immun*, 85.

Thorens, J., F. Froehlich, W. Schwizer, E. Saraga, J. Bille, K. Gyr, P. Duroux, M. Nicolet, B. Pignatelli, A. L. Blum, J. J. Gonvers, and M. Fried. 1996. 'Bacterial overgrowth during treatment with omeprazole compared with cimetidine: a prospective randomised double blind study', *Gut*, 39: 54-9.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. 2012. 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nat Protoc*, 7: 562-78.

Truong, D. T., E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata. 2015. 'MetaPhlAn2 for enhanced metagenomic taxonomic profiling', *Nat Methods*, 12: 902-3.

Truong, D. T., A. Tett, E. Pasolli, C. Huttenhower, and N. Segata. 2017. 'Microbial strain-level population structure and genetic diversity from metagenomes', *Genome Res*, 27: 626-38.

Tseng, C. H., J. T. Lin, H. J. Ho, Z. L. Lai, C. B. Wang, S. L. Tang, and C. Y. Wu. 2016. 'Gastric microbiota and predicted gene functions are altered after subtotal gastrectomy in patients with gastric cancer', *Sci Rep*, 6: 20701.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. 2007. 'The human microbiome project', *Nature*, 449: 804-10.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. 'Environmental genome shotgun sequencing of the Sargasso Sea', *Science*, 304: 66-74.

Verberkmoes, N. C., A. L. Russell, M. Shah, A. Godzik, M. Rosenquist, J. Halfvarson, M. G. Lefsrud, J. Apajalahti, C. Tysk, R. L. Hettich, and J. K. Jansson. 2009. 'Shotgun metaproteomics of the human distal gut microbiota', *ISME J*, 3: 179-89.

Walker, A. W., J. C. Martin, P. Scott, J. Parkhill, H. J. Flint, and K. P. Scott. 2015. '16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice', *Microbiome*, 3: 26.

Walsh, A. M., F. Crispie, O. O'Sullivan, L. Finnegan, M. J. Claesson, and P. D. Cotter. 2018. 'Species classifier choice is a key consideration when analysing low-complexity food microbiome data', *Microbiome*, 6: 50.

Walters, W. A., J. G. Caporaso, C. L. Lauber, D. Berg-Lyons, N. Fierer, and R. Knight. 2011. 'PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers', *Bioinformatics*, 27: 1159-61.

Warren, J. R., and B. Marshall. 1983. 'Unidentified curved bacilli on gastric epithelium in active chronic gastritis', *Lancet*, 1: 1273-5.

Westcott, S. L., and P. D. Schloss. 2015. 'De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units', *PeerJ*, 3: e1487.

White, J. R., N. Nagarajan, and M. Pop. 2009. 'Statistical methods for detecting differentially abundant features in clinical metagenomic samples', *PLoS Comput Biol*, 5: e1000352.

Wilmes, P., and P. L. Bond. 2006. 'Metaproteomics: studying functional gene expression in microbial ecosystems', *Trends Microbiol*, 14: 92-7.

Wong, B. C., S. K. Lam, W. M. Wong, J. S. Chen, T. T. Zheng, R. E. Feng, K. C. Lai, W. H. Hu, S. T. Yuen, S. Y. Leung, D. Y. Fong, J. Ho, C. K. Ching, J. S. Chen, and Group China Gastric Cancer Study. 2004. 'Helicobacter pylori eradication to prevent gastric cancer in a high-risk region of China: a randomized controlled trial', *JAMA*, 291: 187-94.

Wood, D. E., and S. L. Salzberg. 2014. 'Kraken: ultrafast metagenomic sequence classification using exact alignments', *Genome Biol*, 15: R46.

Wotherspoon, A. C., C. Ortiz-Hidalgo, M. R. Falzon, and P. G. Isaacson. 1991. 'Helicobacter pylori-associated gastritis and primary B-cell gastric lymphoma', *Lancet*, 338: 1175-6.

Yang, I., S. Woltemate, M. B. Piazuelo, L. E. Bravo, M. C. Yepez, J. Romero-Gallo, A. G. Delgado, K. T. Wilson, R. M. Peek, P. Correa, C. Josenhans, J. G. Fox, and S. Suerbaum. 2016. 'Different gastric microbiota compositions in two human populations with high and low gastric cancer risk in Colombia', *Sci Rep*, 6: 18594.

Yu, G., J. Torres, N. Hu, R. Medrano-Guzman, R. Herrera-Goepfert, M. S. Humphrys, L. Wang, C. Wang, T. Ding, J. Ravel, P. R. Taylor, C. C. Abnet, and A. M. Goldstein. 2017. 'Molecular Characterization of the Human Stomach Microbiota in Gastric Cancer Patients', *Front Cell Infect Microbiol*, 7: 302.

Zaheer, R., N. Noyes, R. Ortega Polo, S. R. Cook, E. Marinier, G. Van Domselaar, K. E. Belk, P. S. Morley, and T. A. McAllister. 2018. 'Impact of sequencing depth on the characterization of the microbiome and resistome', *Sci Rep*, 8: 5890.

Zamani, M., F. Ebrahimtabar, V. Zamani, W. H. Miller, R. Alizadeh-Navaei, J. Shokri-Shirvani, and M. H. Derakhshan. 2018. 'Systematic review with meta-analysis: the worldwide prevalence of Helicobacter pylori infection', *Aliment Pharmacol Ther*, 47: 868-76.

Zhang, C., K. Cleveland, F. Schnoll-Sussman, B. McClure, M. Bigg, P. Thakkar, N. Schultz, M. A. Shah, and D. Betel. 2015. 'Identification of low abundance microbiome in clinical samples using whole genome sequencing', *Genome Biol*, 16: 265.

Zhang, C., P. V. Thakkar, S. E. Powell, P. Sharma, S. Vennelaganti, D. Betel, and M. A. Shah. 2018. 'A Comparison of Homogenization vs. Enzymatic Lysis for Microbiome Profiling in Clinical Endoscopic Biopsy Tissue Samples', *Front Microbiol*, 9: 3246.

Zhang, X., W. Chen, Z. Ning, J. Mayne, D. Mack, A. Stintzi, R. Tian, and D. Figeys. 2017. 'Deep Metaproteomics Approach for the Study of Human Microbiomes', *Anal Chem*, 89: 9407-15.

Zhang, X., S. A. Deeke, Z. Ning, A. E. Starr, J. Butcher, J. Li, J. Mayne, K. Cheng, B. Liao, L. Li, R. Singleton, D. Mack, A. Stintzi, and D. Figeys. 2018. 'Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease', *Nat Commun*, 9: 2873.

Zhou, L., and A. J. Pollard. 2012. 'A novel method of selective removal of human DNA improves PCR sensitivity for detection of Salmonella Typhi in blood samples', *BMC Infect Dis*, 12: 164.

Zhu, W., A. Lomsadze, and M. Borodovsky. 2010. 'Ab initio gene identification in metagenomic sequences', *Nucleic Acids Res*, 38: e132.

Zilberstein, B., A. G. Quintanilha, M. A. Santos, D. Pajecki, E. G. Moura, P. R. Alves, F. Maluf Filho, J. A. de Souza, and J. Gama-Rodrigues. 2007. 'Digestive tract microbiota in healthy volunteers', *Clinics (Sao Paulo)*, 62: 47-54.

Zoetendal, E. G., A. D. Akkermans, and W. M. De Vos. 1998. 'Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria', *Appl Environ Microbiol*, 64: 3854-9.

APPENDIX

Paper I

Gastric microbial community profilling reveals a dysbiotic cancer-associated microbiota

Ferreira RM, **Pereira-Marques J**, Pinto-Ribeiro I, Costa JL, Carneiro F, Machado JC, Figueiredo C.

Gut. 2018; 67(2):226-236.

Gut Online First, published on November 4, 2017 as 10.1136/gutjnl-2017-314205

OPEN ACCESS

ORIGINAL ARTICLE

Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota

Rui M Ferreira,[1,2] Joana Pereira-Marques,[1,2,3] Ines Pinto-Ribeiro,[1,2,4] Jose L Costa,[1,2,4] Fatima Carneiro,[1,2,4,5] Jose C Machado,[1,2,4] Ceu Figueiredo[1,2,4]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2017-314205).

[1]i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
[2]Ipatimup – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal
[3]Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal
[4]Faculty of Medicine, University of Porto, Porto, Portugal
[5]Department of Pathology, Centro Hospitalar São João, Porto, Portugal

Correspondence to
Prof Ceu Figueiredo, Ipatimup - Institute of Molecular Pathology and Immunology of the University of Porto, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal; cfigueiredo@ipatimup.pt

Received 24 March 2017
Revised 25 October 2017
Accepted 26 October 2017

ABSTRACT

Objective Gastric carcinoma development is triggered by *Helicobacter pylori*. Chronic *H. pylori* infection leads to reduced acid secretion, which may allow the growth of a different gastric bacterial community. This change in the microbiome may increase aggression to the gastric mucosa and contribute to malignancy. Our aim was to evaluate the composition of the gastric microbiota in chronic gastritis and in gastric carcinoma.

Design The gastric microbiota was retrospectively investigated in 54 patients with gastric carcinoma and 81 patients with chronic gastritis by 16S rRNA gene profiling, using next-generation sequencing. Differences in microbial composition of the two patient groups were assessed using linear discriminant analysis effect size. Associations between the most relevant taxa and clinical diagnosis were validated by real-time quantitative PCR. Predictive functional profiling of microbial communities was obtained with PICRUSt.

Results The gastric carcinoma microbiota was characterised by reduced microbial diversity, by decreased abundance of *Helicobacter* and by the enrichment of other bacterial genera, mostly represented by intestinal commensals. The combination of these taxa into a microbial dysbiosis index revealed that dysbiosis has excellent capacity to discriminate between gastritis and gastric carcinoma. Analysis of the functional features of the microbiota was compatible with the presence of a nitrosating microbial community in carcinoma. The major observations were confirmed in validation cohorts from different geographic origins.

Conclusions Detailed analysis of the gastric microbiota revealed for the first time that patients with gastric carcinoma exhibit a dysbiotic microbial community with genotoxic potential, which is distinct from that of patients with chronic gastritis.

Significance of this study

What is already known on this subject?

- *Helicobacter pylori* infection increases the risk for gastric carcinoma by causing chronic inflammation and decreasing the number of acid-producing glands.
- Gastric acid reduction by acid-suppressive drugs results in bacterial overgrowth and high levels of gastric nitrite and *N*-nitrosamine.
- The context of a complex microbiota accelerates the onset and promotes neoplasia in the *H. pylori* insulin-gastrin mouse model of gastric cancer.

What are the new findings?

- The gastric microbiota profile of patients with carcinoma is significantly different from that of patients with chronic gastritis.
- The gastric carcinoma microbiota is dysbiotic and characterised by reduced microbial diversity, reduced *Helicobacter* abundance and over-representation of bacterial genera that include intestinal commensals.
- The microbial community found in gastric carcinoma has increased nitrosating functions consistent with increased genotoxic potential.

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

Our results provide a new interpretative frame for understanding the microbial dysbiosis associated with gastric carcinoma, and suggest that alterations in the gastric microbiota may need to be considered to maximise efficacy of preventive and therapeutic strategies tailored at reducing the incidence of gastric carcinoma.

INTRODUCTION

Gastric carcinoma is a major health problem worldwide, with an estimated 1 million new cases every year.[1] *Helicobacter pylori* infection plays a crucial role in the initial steps of carcinogenesis by causing enhanced inflammation and progres sive degradation of the architecture and function of the gastric epithelium.[2 3] From a certain point on, however, gastric carcinoma devel opment may be *H. pylori* independent, since colonisation decreases (and is eventually lost) in later steps of carcinogenesis.[4] Additionally, *H. pylori* eradication studies have shown that successful eradication does not completely prevent gastric carcinoma development.[5–7] These observations suggest that factors other than *H. pylori* contribute to persistent inflammation of the gastric mucosa and to gastric cancer development.

It has been proposed that changes that occur in the stomach as a result of chronic *H. pylori* infec tion leading to decreased acid secretion allow the successful establishment of a new microbiota that contributes to malignant transformation through maintenance of inflammation and conversion of nitrates into *N*-nitrosamines.[2 3] This is supported by earlier studies showing that reduction of gastric

To cite: Ferreira RM, Pereira-Marques J, Pinto-Ribeiro I, *et al*. *Gut* Published Online First: [*please include* Day Month Year]. doi:10.1136/gutjnl-2017-314205

acid by different types of drugs results in significant intragastric bacterial overgrowth, increased counts of nitrate-reducing bacteria and increased nitrite and *N*-nitrosamine levels.[8 9] This is further supported by studies in the hypergastrinaemic insulin-gastrin (INS-GAS) transgenic mouse model, which showed that *H. pylori*-induced gastric cancer is promoted by the pres ence of a complex gastric microbiota, as these animals develop more tumours than germ-free mice infected with *H. pylori* only.[10 11]

So far, only a very small number of studies characterised the human gastric microbiota in health and disease. Their major findings were that *H. pylori*-negative subjects contained a diverse microbiota in their stomach, whereas in *H. pylori*-infected patients the gastric mucosa was dominated by this species.[12–15] In the context of gastric carcinogenesis, few studies have been conducted and no particular component of the microbiota has been identified as implicated in gastric carcinoma.[15 16] The limitations of these studies are the limitations in sensitivity and coverage compared with more recently developed techniques, and, most importantly, the inclusion of very limited numbers of subjects, making it difficult to generate statistically significant conclusions. Therefore, we performed high-throughput profiling of the gastric bacterial communities present in 135 gastric carcinoma cases and chronic gastritis controls, by next-generation sequencing (NGS) of the 16S rRNA gene. We used validation cohorts from multiple geographic locations to confirm our findings.

MATERIALS AND METHODS

Patients

Eighty-one individuals with chronic gastritis and 54 with gastric carcinoma were included in the Portuguese discovery cohort (see online supplementary table S1). These were part of a case–control study aimed at investigating risk modifiers for gastric cancer.[17 18] Subjects with chronic gastritis (mean age 43.6±7.0 years; male-to-female ratio 39.5:1) were recruited during a screening programme for premalignant lesions of the gastric mucosa and underwent standard gastroscopy at Centro Hospitalar São João (CHSJ). Eleven patients presented glan dular atrophy with foci of intestinal metaplasia. Of these, 1 had mild corpus and moderate antral atrophy and the remaining 10 cases did not have corpus atrophy and had mild (n=6), moderate (n=2) or marked (n=2) atrophy in the antrum (including incisura). Only individuals without evidence of past or present peptic ulcer disease were included. In addition, patients under proton pump inhibitor or antimicrobial treatments were excluded. Patients with gastric carcinoma (mean age 58.8±13.2 years; male-to-female ratio 1.5:1) were diagnosed and underwent cancer resection at CHSJ. A validation cohort of an additional 38 gastric specimens from 15 patients with chronic gastritis and 23 patients with gastric carcinoma, diagnosed between 2014 and 2016, were retrieved from the tissue and tumour bank at CHSJ (see online supplementary table S1). All procedures were in accordance with the institutional ethical standards. Samples were delinked and unidenti fied from their donors.

Two additional validation series, consisting of NGS data of the 16S rRNA gene of 79 gastric carcinoma cases from a population from China and 53 gastric carcinoma cases from a population from Mexico, were retrieved from the Sequence Read Archive (BioProject PRJNA310127; see online supplementary table S2).[19]

16S rRNA gene sequencing

DNA was isolated from gastric biopsies or surgical specimens of non-neo plastic gastric mucosa adjacent to the tumour, as previously described.[17] The 16S rRNA gene was amplified using primers U789F 5′-TAGATACCCTGGTAGTCC-3′ and U1053R 5′-CTGACGACAGCCATGC-3′ targeting the V5-V6 hypervariable regions and sequenced in an Ion PGM Torrent platform following manufacturer's instructions. Primers were designed following recommendations reported by Andersson *et al*, and were extensively analysed using PrimerProspector (see online supplementary figure S1).[20 21]

Sequencing data analysis

The performance of the UPARSE pipeline was evaluated and compared with that reported for samples of the Human Microbiome Project (HMP) data set.[22] Using UPARSE (usearch _v7.0.1090_i86linux64), reads were filtered by imposing a maximum number of expected errors of 0.5 and a global trimming at 250 nucleotides.[22] Reads were dereplicated and singletons were discarded. Filtered reads were clustered into operational taxonomic units (OTU) assuming 97% similarity. Chimeric reads were reference removed using Uchime.[22] Each OTU was taxonomically assigned using Uclust considering a minimum percentage of similarity to a reference database (Greengenes Named Isolate database, release August 2013) match of 90%.[23] Diversity analyses were performed using QIIME (V.1.9).[24] Alpha diversity was determined by the Shannon index and with Good's estimator of coverage. Differences in alpha diversity were assessed by the t-test controlled with 10^3 Monte Carlo permutations. Beta diversity was assessed by unweighted and weighted UniFrac distance matrices and visualised by principal coordinate analysis (PCoA), controlled by 10^3 jackknife replicates.[25] Sample clustering in beta diversity analysis was tested using analysis of similarity (ANOSIM) with 10^4 bootstrap replications.[26] Comparisons between distance matrices were evaluated by the Mantel correlation controlled with 10^4 permutations.

Taxonomic discovery analysis

Statistically significant differences in the relative abundance of taxa associated with groups of patients were performed using linear discriminant analysis (LDA) effect size (LEfSe).[27] Only taxa with LDA greater than 4 at a P value <0.05 were considered significantly enriched.

Real-time quantitative PCR

Sequencing results were confirmed by quantitative PCR (qPCR) (see online supplementary table S3).

Functional metagenome predictions

For functional metagenome prediction, we captured OTU representative sequences from Greengenes database using the USEARCH global alignment command and discarding reads that did not hit the reference database. Reconstruction of the metagenome was performed using PICRUSt.[28] Accuracy of the predicted metagenomes was assessed by determining the nearest sequenced taxon index. Predicted functional genes were categorised into Clusters of Orthologous Groups (COG) and into Kyoto Encyclopedia of Genes and Genome (KEGG) orthology (KO), and compared across patient groups using STAMP.[29] Statistical differences in COG and KO frequencies were determined by White's non-parametric t-test with a Benjamini-Hochberg false discovery rate correction to adjust P values for multiple testing.[30]

For further details, see online supplementary information.

RESULTS

Quality control of 16S rRNA microbiota profiling

In the present study, we compared the gastric microbiota of patients with chronic gastritis with that of patients with gastric carcinoma by NGS of the 16S rRNA gene. After sequencing and quality filtering, more than 10.8 million reads were obtained corresponding to a mean of 80 261 reads and 178 OTUs per sample (see online supplementary figure S2). On average, patients with chronic gastritis had a significantly higher number of reads (86 957) than patients with cancer (67 954; $P<0.05$). However, the number of OTUs was not significantly different between the two patient groups (186 and 169 OTUs, respectively; $P=0.071$). To control for the number of false OTUs and to measure the number of biologically meaningful OTUs, we classified them according to similarity shared with sequences of the Greengenes Named Isolated database (see online supplementary figure S2). In our data set, the frequencies of misleading and valid OTUs were similar to those reported for the HMP data set processed with the UPARSE pipeline.[22]

To assure consistency between amplification and sequencing sets, 32 randomly selected samples were used to test reproducibility. The intraclass correlation coefficients showed good reproducibility for the assessment of the Shannon index, for the UniFrac distances and for the relative abundance of phyla (see online supplementary table S4). In conclusion, our approach provides the most in-depth characterisation of the gastric microbiota so far and generates robust and consistent data.

The gastric microbiota profile differs in chronic gastritis and gastric carcinoma

To evaluate alterations in the microbiota structure between patients with chronic gastritis and gastric carcinoma, we measured microbial alpha diversity (ie, within sample diversity) and beta diversity (ie, diversity between samples). By measuring alpha diversity using the Shannon index, we found that patients with gastric carcinoma had significantly decreased microbial diversity in comparison with patients with chronic gastritis (figure 1A, $P=0.003$; online supplementary figure S3). To ensure good estimation of bacterial diversity, we measured the proportion of total bacterial species represented in samples of each patient group by the Good's estimator of coverage. Estimated coverage ranged from 0.94 to 0.98 in chronic gastritis and from 0.95 to 0.99 in gastric carcinoma ($P=0.225$), suggesting that the 16S rRNA results from each (chronic gastritis and gastric carcinoma) library represent the majority of bacteria present in the gastric mucosa (figure 1B).

Beta diversity was calculated using both unweighted (ie, qualitative) and weighted (ie, quantitative) UniFrac phylogenetic distance matrices, and visualised in PCoA plots. The total

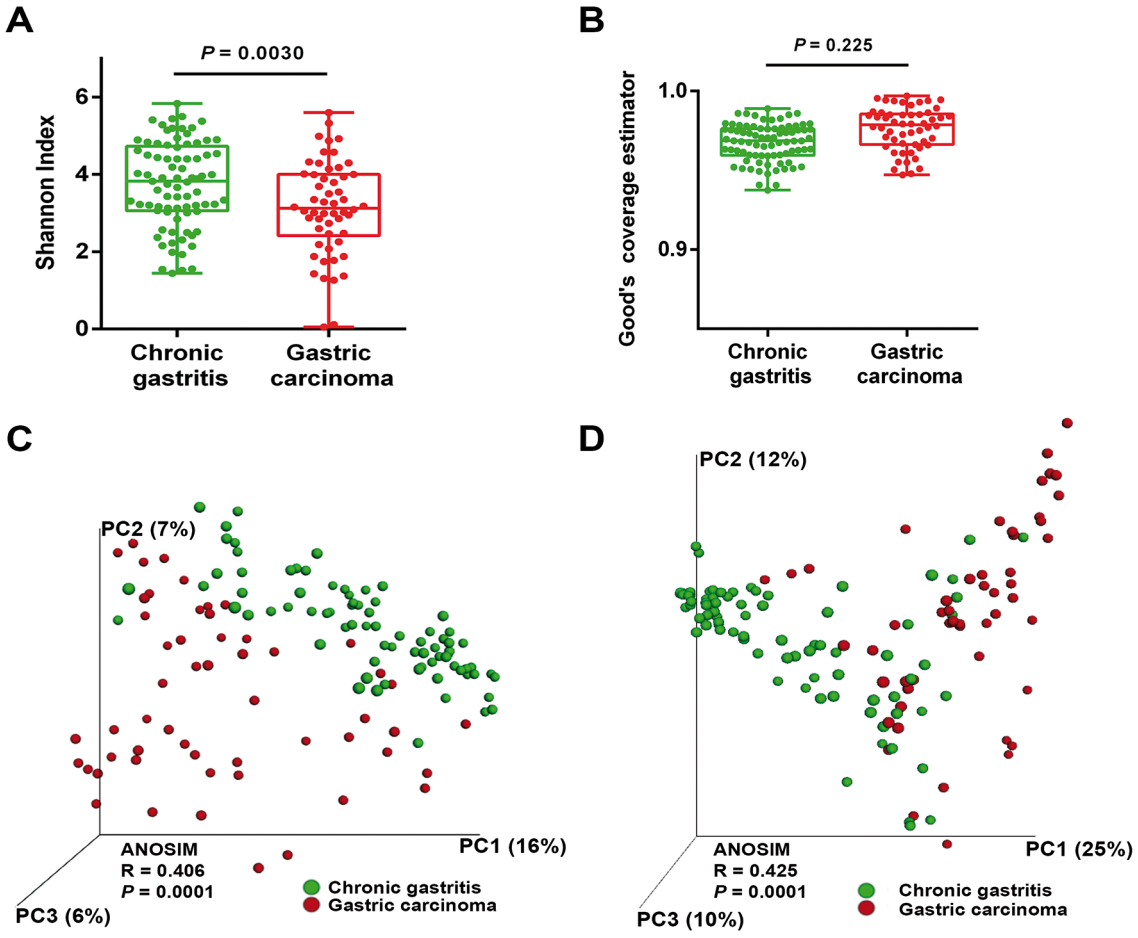


Figure 1 The gastric microbiota profile differs in chronic gastritis and gastric carcinoma. (A) Shannon index of diversity in patients with chronic gastritis and gastric carcinoma. (B) Good's estimator of coverage, measuring the proportion of total bacterial species represented in samples of each group of patients. Principal coordinate analysis (PCoA) plots of (C) unweighted and (D) weighted UniFrac distances in which samples were coloured by clinical outcome. The percentage of diversity captured by each coordinate is shown. ANOSIM, analysis of similarity.

diversity captured by the top three principal coordinates was 29% and 47% for unweighted and weighted UniFrac, respectively. The microbiota composition of patients with gastric carcinoma was significantly different from that of patients with chronic gastritis (ANOSIM R=0.406, *P*=0.0001; and R=0.425, *P*=0.0001, for unweighted and weighted distances, respectively; figure 1C,D).

Since age is an established risk factor for gastric carcinoma, and since patients with carcinoma were significantly older than patients with gastritis in our series (see online supplementary table S1), we next addressed whether the microbial profile was different between younger and older patients. Overall, increasing age could differentiate the microbiota profiles of the full sample set (see online supplementary figure S4A,B). However, when we performed age-matched comparisons of the microbiota in patients with chronic gastritis and carcinoma, we observed statistically significant differences in the unweighted and weighted UniFrac distances (see online supplementary figure S4C,D), reinforcing that the microbiota composition is different in the two clinical settings. Also in the age-matched comparisons, significantly decreased microbial alpha diversity was found in patients with gastric carcinoma (see online supplementary figure S4E, *P*=0.0096).

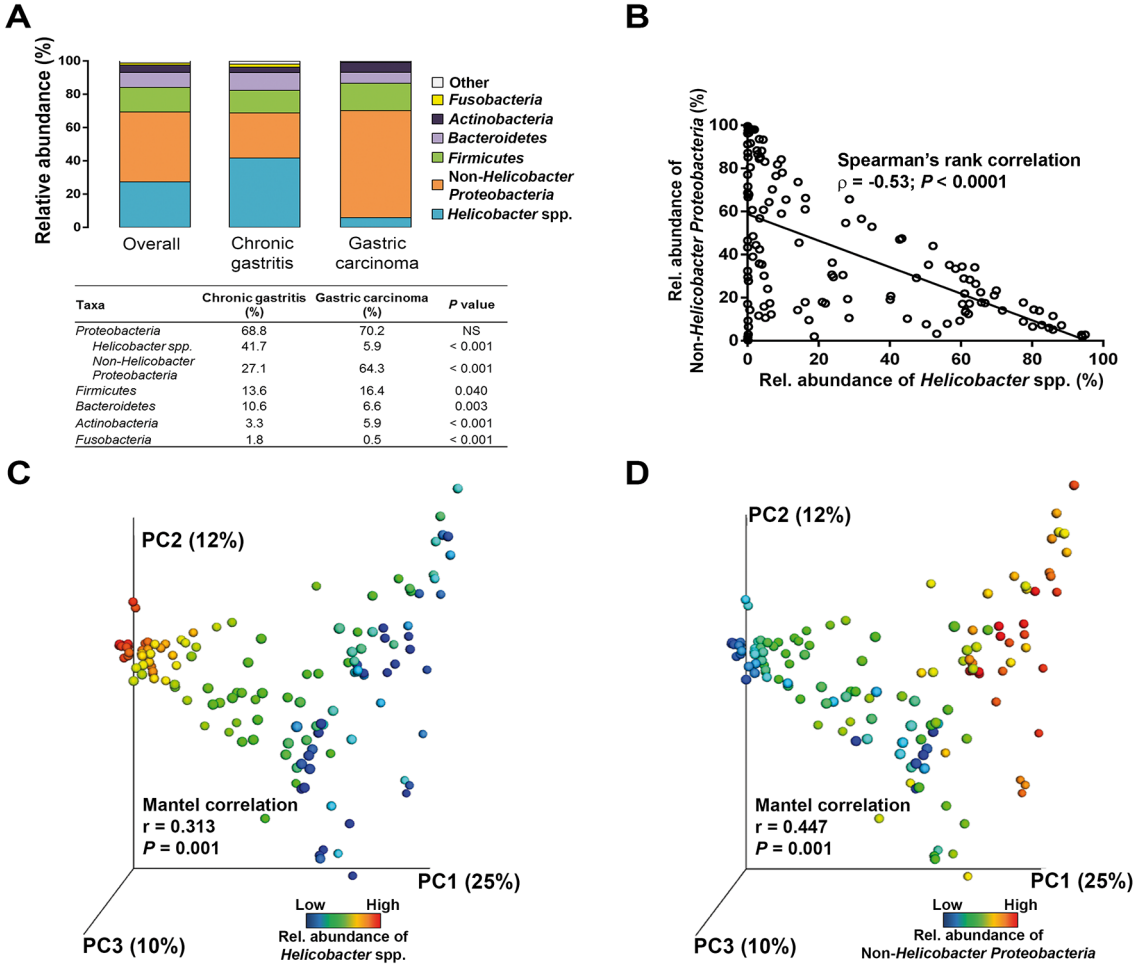
No statistically significant differences in the microbiota profiles of gastric carcinoma cases were observed for gender, histological type and tumour location (see online supplementary figure S5).

In patients with chronic gastritis, we could not detect differences in the alpha diversity and beta diversity between patients with non-atrophic gastritis (n=70) and patients with glandular atrophy (n=11; online supplementary figure S6), and therefore they were pooled together for the analyses.

These results show that there is a significant reduction in microbial diversity in gastric carcinoma. Furthermore, the fact that the weighted UniFrac captured more diversity than unweighted metrics suggests that alterations in the relative abundance of taxa are a major contributor for microbiota differences between gastritis and gastric carcinoma.

The influence of *H. pylori* in the microbiota composition of chronic gastritis and gastric carcinoma

Overall, the gastric microbiota was dominated by five phyla: *Proteobacteria* (69.3%), *Firmicutes* (14.7%), *Bacteroidetes* (9.0%), *Actinobacteria* (4.3%) and *Fusobacteria* (1.3%). Although these phyla were present in the two patient groups in the same order of relative abundance, the gastric carcinoma microbiota



Taxa	Chronic gastritis (%)	Gastric carcinoma (%)	*P* value
Proteobacteria	68.8	70.2	NS
Helicobacter spp.	41.7	5.9	< 0.001
Non-Helicobacter Proteobacteria	27.1	64.3	< 0.001
Firmicutes	13.6	16.4	0.040
Bacteroidetes	10.6	6.6	0.003
Actinobacteria	3.3	5.9	< 0.001
Fusobacteria	1.8	0.5	< 0.001

Figure 2 The influence of *Helicobacter pylori* in the microbiota composition of chronic gastritis and gastric carcinoma. (A) Relative abundance of phyla in all subjects and in each group of patients. (B) Spearman's rank correlation between relative abundance of *Helicobacter* spp. and non-*Helicobacter Proteobacteria* in all patients. Principal coordinate analysis (PCoA) plots of the weighted UniFrac distance matrix coloured by (C) increasing relative abundance of *Helicobacter* and of (D) non-*Helicobacter Proteobacteria*.

had an over-representation of *Actinobacteria* (P<0.001) and *Firmicutes* (P=0.040), and lower abundance of *Bacteroidetes* (*P*=0.003) and *Fusobacteria* (*P*<0.001; figure 2A).

When reads assigned to *Proteobacteria* into *Helicobacter* spp. and non-*Helicobacter Proteobacteria* were separated, a significant reduction in the abundance of *Helicobacter* (*P*<0.001) and an over-representation of non-*Helicobacter Proteobacteria* were detected in gastric carcinoma (*P*<0.001; figure 2A). Accordingly, a significant negative correlation was found between these taxa (r=−0.53, *P*<0.0001; figure 2B). In support of the above, the microbiota profile of the two patient groups could be distinguished by the abundance of *Helicobacter* (Mantel correlation, r=0.313, *P*=0.001; figure 2C) and by the abundance of non-*Helicobacter Proteobacteria* (Mantel correlation, r=0.447, *P*=0.001; figure 2D).

Regarding *Helicobacter* spp. in chronic gastritis, the mean relative abundance of this genus was 41.7%, but varied considerably between patients from 0.01% to 94.9% (figure 2A). The relative abundance of *Helicobacter* was inversely correlated with the abundance of non-*Helicobacter Proteobacteria* (r=−0.59, *P*<0.0001), *Firmicutes* (r=−0.49, *P*<0.0001), *Bacteroidetes* (r=−0.43, *P*<0.0001) and *Actinobacteria* (r=−0.54, *P*<0.0001; online supplementary table S5). In contrast, the great majority of patients with gastric carcinoma (80%) had a relative abundance of *Helicobacter* below 5%, including eight patients in which *Helicobacter* reads were not detected by NGS. The abundance of *Helicobacter* in gastric carcinoma was correlated with that of *Bacteroidetes* and *Fusobacteria* (see online supplementary table S5).

Overall, these results show that for high taxonomic levels the stomach microbial communities differ in chronic gastritis and gastric carcinoma, suggesting that major changes also occur at lower taxonomic levels. Additionally, our data validate that *Helicobacter* exists in the gastric carcinoma microbiota as a low abundant or absent genus.

Specific microbial taxa are associated with gastric carcinoma

To identify the most relevant taxa responsible for the differences between clinical diagnoses, we conducted LEfSe analysis.[27] This analysis identified 29 taxa, including 10 genera, which were differentially abundant in the two patient groups (figure 3A,B). In gastric carcinoma, an enrichment in *Proteobacteria* taxa was observed, including the genera *Phyllobacterium* and *Achromobacter* and the families *Xanthomonadaceae* and *Enterobacteriaceae*. Although no specific genus could be identified within the *Xanthomonadaceae*, in the *Enterobacteriaceae*, the genus *Citrobacter* was identified as being significantly enriched in gastric carcinoma. Additionally, *Lactobacillus*, *Clostridium* and *Rhodococcus* were also significantly more abundant in gastric carcinoma. *Helicobacter*, *Neisseria*, *Prevotella* and *Streptococcus* were most abundant in the microbiota of patients with chronic gastritis. Results of the LEfSe analysis in the age-matched subset closely recapitulated the bacteria taxa differentially abundant in the two patient groups (see online supplementary figure S4F).

To show that relationships among disease-associated taxa did not depend on differences observed in the abundance of *Helicobacter,* we conducted a reanalysis subtracting the *Helicobacter* reads from the data set. Considering the same parameters in the LEfSe analysis, we confirmed the enrichment of *Streptococcus, Prevotella* and *Neisseria* in chronic gastritis (see online supplementary figure S7). Additionally, we identified an enrichment in two *Proteobacteria* taxa, *Novosphingobium* and *Pasteurellales*, and in two *Bacteroidetes* families, *Chitinophagaceae* and *Saprospirae*. In gastric carcinoma, no additional taxa were detected after removing the *Helicobacter* reads from the data set.

To validate gastric carcinoma-enriched and depleted taxa, we used NGS data from an independent Chinese cohort of 79 gastric carcinoma cases. In this data set, and in agreement with the results obtained in the Portuguese discovery cohort, we could detect statistically significant enrichment in *Citrobacter*, *Rhodococcus*, *Lactobacillus* and *Phyllobacterium*, and depletion in *Helicobacter* and *Neisseria* (figure 3C). *Clostridium* reads were enriched in gastric carcinoma cases from the Chinese population, although not reaching statistical significance in the LEfSe analysis. *Achromobacter* reads were not detected in the Chinese validation cohort.

To demonstrate that our data were not biased by the microbiota profiling pipeline used, LEfSe results were validated by qPCR in the Portuguese discovery cohort using both genus-specific and universal primers. We confirmed significant decreases in the abundance of *Helicobacter* and *Neisseria*, and significant increases of *Achromobacter*, *Citrobacter*, *Phyllobacterium*, *Clostridium*, *Rhodococcus* and *Lactobacillus* in gastric carcinoma in comparison with chronic gastritis (figure 3D). We have additionally used a second validation cohort from Portugal, and with the exception of *Prevotella* and *Streptococcus*, we were able to confirm the alterations in the abundance of the eight genera as identified by the original LEfSe analysis (figure 3E).

Next, we compared gastric carcinoma cases and chronic gastritis control subjects for the prevalence of specific taxa. As shown in table 1, the six genera significantly enriched in gastric carcinoma and identified by LEfSe analysis were also significantly more prevalent in patients with gastric carcinoma than in patients with chronic gastritis. In logistic regression models with carriage of the genera *Phyllobacterium*, *Achromobacter, Citrobacter, Lactobacillus, Clostridium* or *Rhodococcus* as the independent variables, and gastric carcinoma as the dependent variable, the ORs for gastric carcinoma were 3.5 (95% CI 1.7 to 7), 20.5 (95% CI 7.4 to 59), 9.9 (95% CI 4.3 to 23), 6.3 (95% CI 2.9 to 14), 5.7 (95% CI 2.2 to 15) and 4.2 (95% CI 1.7 to 11), respectively. The associations remained significant after adjustment for age and sex.

Microbial dysbiosis is associated with gastric carcinoma

We next combined the 10 most relevant taxa that characterised each group of patients and calculated the microbial dysbiosis index (MDI).[31] The gastric microbiota of patients with gastric carcinoma had a higher MDI than that of patients with chronic gastritis both in the discovery cohort and in the validation cohorts (*P*<0.0001; figure 4A). Similar findings were observed in the age-matched subset of the discovery cohort (see online supplementary figure S4G). Likewise, significantly higher MDI was observed in the microbiota of patients with gastric carcinoma in comparison with that of patients with chronic gastritis, as assessed using qPCR in the Portuguese validation cohort (*P*<0.0001; figure 4B). The MDI showed an inverse correlation with the alpha diversity (r=−0.262, *P*=0.005; figure 4C) and a direct correlation with the beta diversity (r=0.208, *P*=0.001; figure 4D), resulting in a clear differentiation gradient among samples. These results demonstrate that the gastric carcinoma microbiota has a high degree of dysbiosis, consistent with reduced bacterial diversity.

We also evaluated whether the MDI could be used to discriminate between chronic gastritis and gastric carcinoma. In receiver operating characteristics (ROC) analysis, the MDI showed excellent performance in identifying gastric

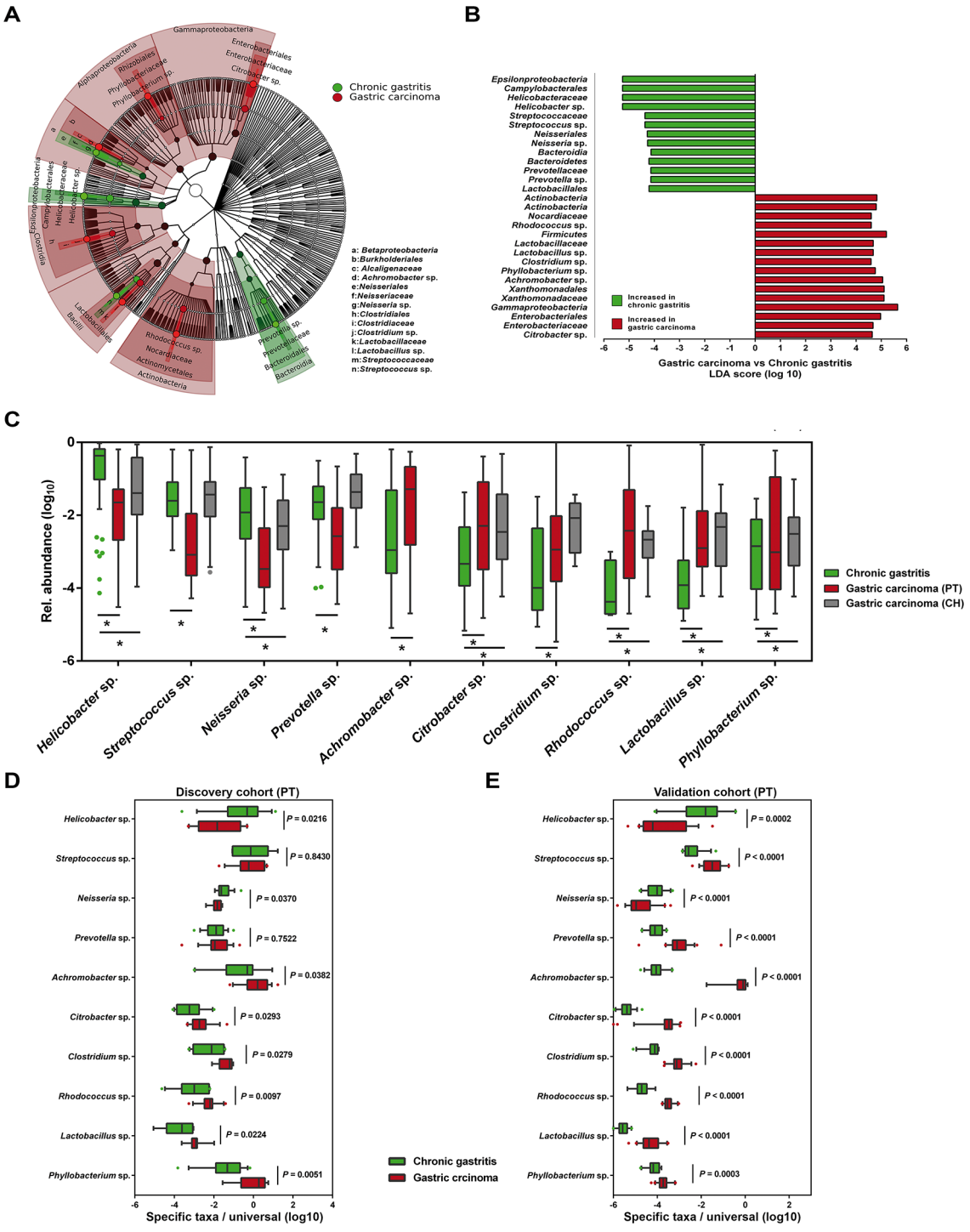


Figure 3 Microbial taxa associated with gastric carcinoma. (A) Cladogram representation of the gastric microbiota taxa associated with chronic gastritis and gastric carcinoma. (B) Association of specific microbiota taxa with the group of chronic gastritis and gastric carcinoma by linear discriminant analysis (LDA) effect size (LEfSe). Green indicates taxa enriched in chronic gastritis group and red indicates taxa enriched in gastric carcinoma group. (C) Relative abundance of the 10 genera differentially enriched in the two clinical settings across Portuguese discovery and Chinese validation cohorts. *Significance obtained by LEfSe analysis at $P<0.05$. (D,E) Validation of LEfSe results by quantitative PCR (qPCR) of the 10 genera differentially enriched in the discovery cohort (D) and in the Portuguese validation cohort (E). Significance was obtained by Student's t-test.

carcinoma, yielding an area under the curve (AUC) of 0.91 and 0.89 for the Portuguese discovery and validation cohorts, respectively (figure 4E,F). The MDI exhibited improved sensitivity and specificity to detect gastric carcinoma when compared with the use of single taxa (see online supplementary figure S8).

Since in the validation cohorts we could not confirm the differential abundance of *Prevotella* and *Streptococcus*, we recalculated the MDI excluding these genera. This analysis confirmed higher levels of dysbiosis in the gastric carcinoma microbiota in all cohorts, and similar AUCs in the ROC analysis (see online supplementary figure S9).

Table 1 Relative abundance and prevalence of selected microbial taxa in patients with chronic gastritis and gastric carcinoma in the discovery cohort

Taxa (phylum; class; order; family; genus)	Relative abundance (%)*			Prevalence (%)†			Univariate OR (95% CI)	Multivariate‡ OR (95% CI)
	Chronic gastritis	Gastric carcinoma	P§	Chronic gastritis	Gastric carcinoma	P¶		
Proteobacteria; Alphaproteobacteria; Rhizobiales; Phyllobacteriaceae; Phyllobacterium sp.	0.2	5.4	<0.0001	34.6	64.8	0.001	3.5 (1.7 to 7)	3.7 (1.2 to 11)
Proteobacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae; Achromobacter sp.	2.1	11.1	<0.0001	24.7	87.0	<0.0001	20.5 (7.4 to 59)	56.3 (9.5 to 333)
Proteobacteria; Betaproteobacteria; Neisseriales; Neisseriaceae; Neisseria sp.	3.9	0.4	<0.0001	92.6	75.9	<0.0001	0.25 (0.09 to 0.7)	0.27 (0.1 to 1)
Proteobacteria; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter sp.	41.7	5.9	<0.0001	100	85.2	<0.0001	0.0 (0.0 to 0.4)	NA
Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Citrobacter sp.	0.2	4.3	<0.0001	30.9	81.5	<0.0001	9.9 (4.3 to 23)	6.5 (2.1 to 20)
Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus sp.	0.025	4.7	<0.0001	19.7	61.1	<0.0001	6.3 (2.9 to 14)	3.1 (1.0 to 10)
Firmicutes; Bacilli; Lactobacillales; Streptococcaceae; Streptococcus sp.	7.7	2.9	<0.0001	97.5	92.6	0.217	NA	NA
Firmicutes; Clostridia; Clostridiales; Clostridiaceae; Clostridium sp.	0.046	3.7	<0.0001	8.6	35.2	<0.0001	5.7 (2.2 to 15)	7.8 (2.0 to 31)
Actinobacteria; Actinobacteria; Actinomycetales; Nocardiaceae; Rhodococcus sp.	0.002	3.3	<0.0001	9.9	31.5	0.003	4.2 (1.7 to 11)	3.8 (1.0 to 15)
Bacteroidetes; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella sp.	5.0	2.0	<0.0001	97.5	90.7	0.104	NA	NA

* Average relative abundance of specific taxa in subjects containing that taxa.
† Percentage of subjects carrying the specific taxa.
‡ Multivariate logistic regression analysis with non-carriers as reference adjusted for age and gender.
§ P values obtained through linear discriminatory analysis effect size (LEfSe).
¶ P values obtained by Fisher's exact test.
NA, not assessed.

The gastric carcinoma microbiota is characterised by nitrosating bacteria

To infer the metagenome functional content based on the microbial community profiles obtained from the 16S rRNA gene sequences we used PICRUSt.[28] Overall, the microbial communities present in patients with gastric carcinoma and chronic gastritis could be distinguished based on their functions (see online supplementary figure S10A). The predicted KEGG pathways significantly enriched in gastric carcinoma included membrane transport, carbohydrate metabolism, transcription, xenobiotics biodegradation and metabolism, cellular processes and signalling, metabolism, signal transduction, amino acid metabolism and lipid metabolism (see online supplementary table S6).

Because it has been hypothesised that nitrate-reducing bacterial species contribute to gastric malignant transformation by increasing intragastric concentrations of nitrite and *N*-nitroso compounds, we next compared chronic gastritis and gastric carcinoma regarding the microbial functional features involved in those metabolic reactions (see online supplementary table S7). The full reconstitution of the metagenomes showed that the functional composition of the total gastric carcinoma microbiota had increased nitrate reductase functions, which promote the reduction of nitrate to nitrite, and nitrite reductase functions, which promote the reduction of nitrite to nitric oxide, when compared with that of the chronic gastritis (figure 5A,B). Similar results were obtained when the 10 genera differentially abundant in the two patient groups were analysed (see online supplementary figure S10B,E). Collectively, these data provide evidence that a microbial community with genotoxic potential is present in gastric carcinoma.

DISCUSSION

We have profiled the gastric microbiota associated with chronic gastritis and gastric carcinoma in the largest and most in-depth study to date. We have demonstrated that the gastric microbiota composition in patients with gastric carcinoma is significantly different from that of patients with chronic gastritis. Gastric carcinoma dysbiosis was consistent with a microbial community with genotoxic potential, characterised by reduced microbial diversity, reduced *Helicobacter* abundance and over-representation of new bacterial genera. The major findings revealed in the Portuguese discovery cohort were confirmed in additional validation cohorts from multiple geographic locations.

In our study, the gastric microbial communities in gastritis and carcinoma were structurally different, with decreased alpha diversity in carcinoma. Our findings are supported by previous data pointing to lower bacterial diversity among five patients with gastric cancer compared with five patients with non-atrophic gastritis.[15] Also supporting our data, and while our paper was in revision, another paper was published in *Gut* that identified significant decreases in microbial richness in intestinal metaplasia and in gastric carcinoma compared with superficial gastritis.[32] Reduced microbial diversity has now been recognised as a feature of disease states, including inflammatory diseases and cancer.[31 33 34] For example, patients with colorectal cancer had decreased overall microbial community diversity in comparison to healthy controls.[34]

In terms of the composition of the gastric microbiota, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria* and *Fusobacteria* were the five dominant phyla in the stomach, in accordance with previous descriptions.[12 14 20] At the phylum level, we have already identified differences between the two patient groups,

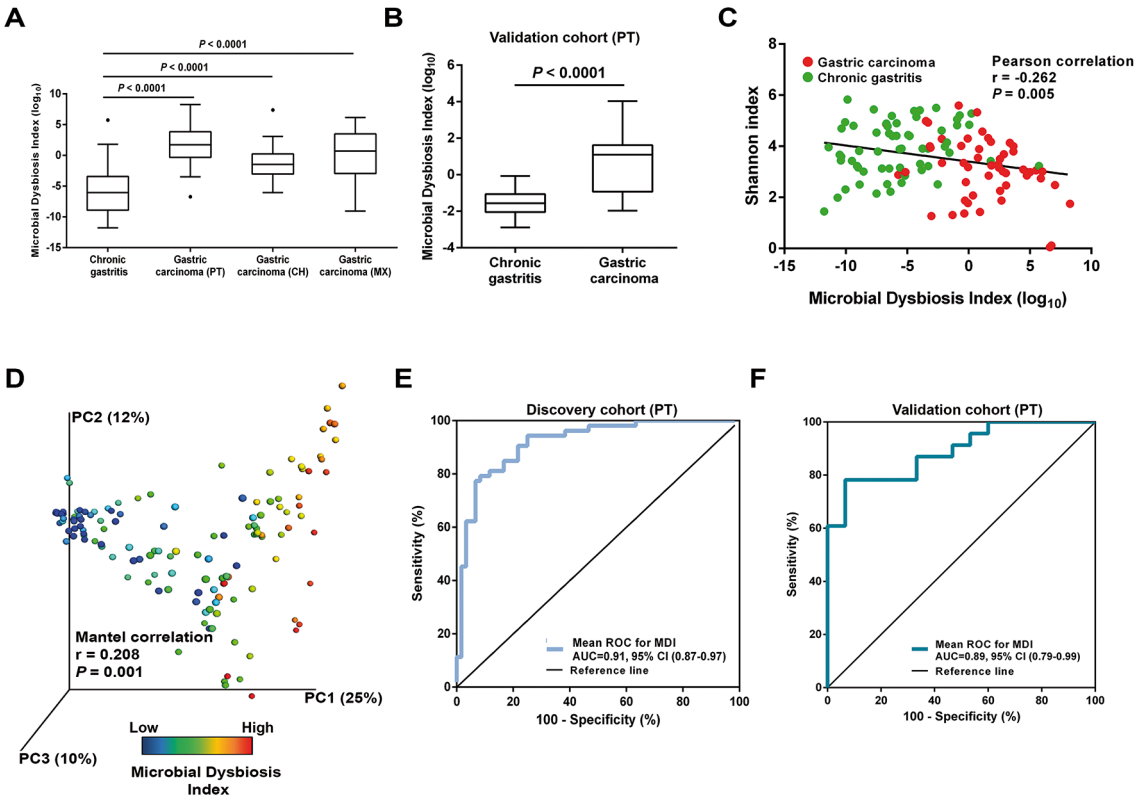


Figure 4 Microbial dysbiosis is associated with gastric carcinoma. (A) Box plot showing the MDI in the discovery cohort and in the Chinese and Mexican validation cohorts. Significance was obtained by one-way analysis of variance (ANOVA) corrected with Holm-Sidak test for multiple comparisons. (B) Box plot showing the MDI of the Portuguese validation cohort. Significance was obtained by Student's t-test. (C) Negative Pearson's correlation between MDI and Shannon index. (D) Principal coordinate analysis (PCoA) plot of the weighted UniFrac distance coloured by increasing MDI. The percentage of diversity captured by each coordinate is shown. Mantel correlations controlled with 10^4 permutations were used to compare distances. (E,F) ROC curves analysis to evaluate the discriminatory potential of MDI in gastric carcinoma detection in the discovery cohort (E) and in the Portuguese validation cohort (**F**). AUC, area under the curve; MDI, microbial dysbiosis index; ROC, receiver operating characteristic.

with increased abundance of non-*Helicobacter Proteobacteria*, *Firmicutes* and *Actinobacteria* in cancer specimens. Importantly, by applying the LEfSe algorithm that was validated for high-dimensional microbiome data sets, we were able to determine the bacterial taxa that most likely explain differences between clinical diagnoses.[27] Additionally, in this study, the major taxonomic differences that were detected after analyses of sequencing-generated and bioinformatics-treated data were further validated by real-time qPCR assays.

In chronic gastritis, and as expected, *Helicobacter* was detected as the most abundant genus. *Streptococcus, Prevotella* and *Neisseria* were also found significantly overabundant in this patient group, although *Streptococcus* and *Prevotella* could not be confirmed by qPCR. Nevertheless, these genera have been identified earlier in *H. pylori*-positive and negative gastritis by 16S rDNA and rRNA sequencing, and by culture from gastric juice and gastric biopsies.[12 14 35 36] In fact, they are among the five most commonly found genera in the non-neoplastic stomach.[12 20 37]*Streptococcus*, *Prevotella* and *Neisseria* are commensals of the oral cavity and oesophagus and whether they constitute transient or active resident stomach microbes is not yet clarified. Interestingly, in a study that compared the gastric microbiota compositions in *H. pylori*-positive individuals from two populations with high and low gastric cancer risks in Colombia, *Neisseria* and *Streptococcus* were among the genera that occurred more abundantly in individuals from the low gastric cancer risk region.[38]

In gastric carcinoma, there was a significant decrease in *Helicobacter* abundance, and several taxa were found to be significantly more abundant. These included *Citrobacter*, *Clostridium*, *Lactobacillus*, *Achromobacter* and *Rhodococcus*, which reside in the intestinal mucosa as commensals but can be opportunistic pathogens.[39 40] *Phyllobacterium*, which are environmental bacteria commonly found in plant roots, were too identified at higher abundance in gastric carcinoma.[41] All genera significantly overabundant in gastric carcinoma were also significantly more prevalent in gastric carcinoma cases than in chronic gastritis control patients, and these associations remained significant after adjustment for age and sex. In line with our results, in a study that combined terminal restriction fragment length polymorphism with 16S rRNA gene cloning and sequencing, *Lactobacillus* was one of the dominating genera in 10 Swedish patients with gastric cancer.[16] Additionally, the use of the microarray G3 PhyloChip to characterise the stomach microbiota of Mexican patients revealed a trend towards the increase of a *Lactobacillus* sp from non-atrophic gastritis, to intestinal metaplasia, to gastric cancer.[15] Moreover, *Citrobacter*, *Clostridium* and *Lactobacillus* have all been cultured from the gastric juice of achlorhydric patients, patients undergoing acid suppression therapy and patients with gastric cancer.[42–44] Interestingly, infection with *Citrobacter rodentium* species increases epithelial cell proliferation and promotes colonic tumour formation in genetically susceptible mice as well as in chemically initiated colon carcinogenesis.[45 46]

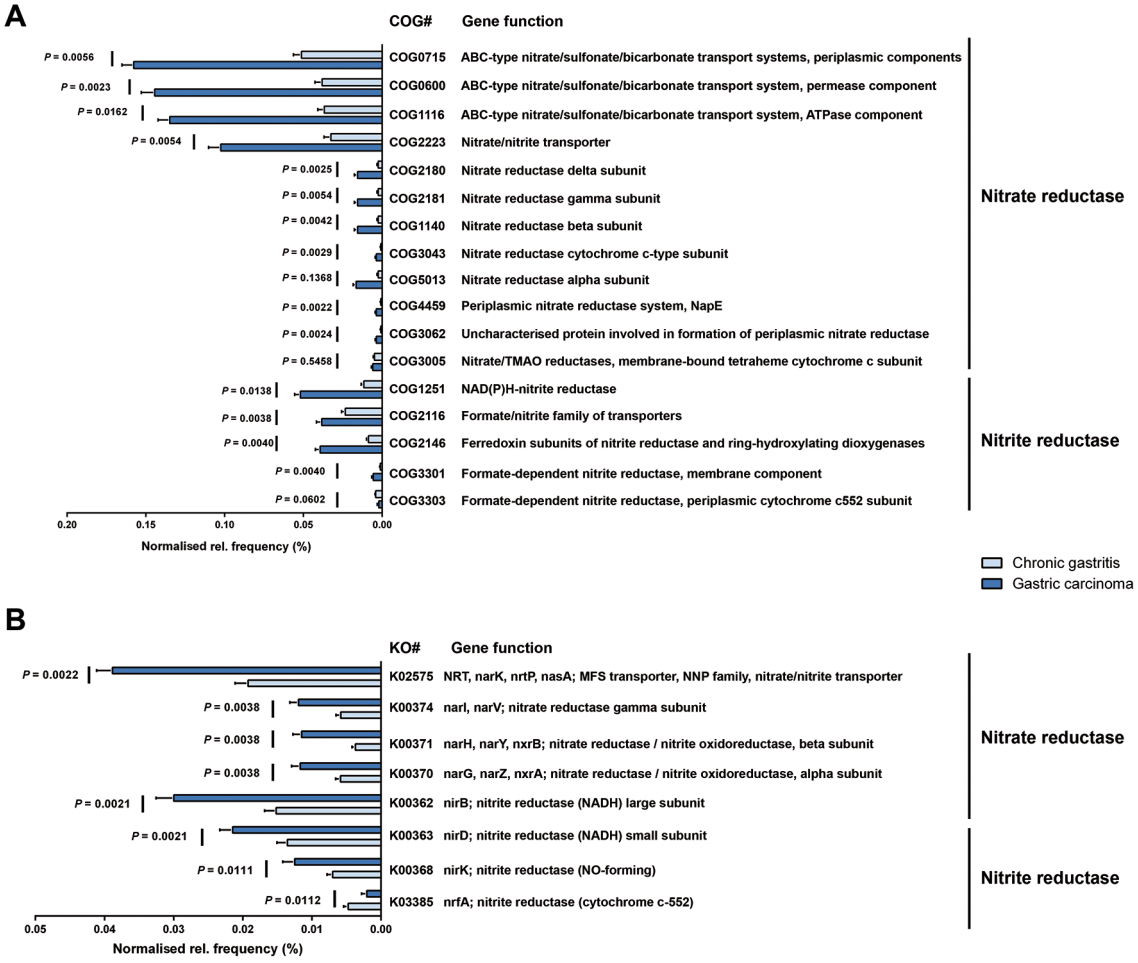


Figure 5 The gastric carcinoma microbiota is characterised by nitrosating bacteria. Functional classification of the predicted metagenome content of the microbiota of chronic gastritis and gastric carcinoma using (A) COG and (B) KO. The normalised relative frequency of nitrate reductase and nitrite reductase in patients with chronic gastritis and gastric carcinoma is shown. Significance was considered for adjusted $P<0.05$. COG, Clusters of Orthologous Groups; KO, Kyoto Encyclopedia of Genes and Genome (KEGG) orthology; NADH, nicotinamide adenine dinucleotide; NO, nitric oxide; TMAO, trimethylamine N-oxide.

The integration of data from the most relevant genera that characterised each patient group allowed us to calculate the dysbiosis index that showed excellent capacity to discriminate between gastritis and gastric carcinoma. Furthermore, the dysbiosis index had improved sensitivity and specificity to detect gastric carcinoma in comparison with the use of single genera, which suggests that changes in the microbial community rather than individual taxa contribute to gastric carcinoma development.

After having analysed the diversity and composition of the gastric microbiota and the microbial features associated with gastric carcinoma, we addressed the functional features of the microbiota. Specifically, we demonstrated that in comparison with chronic gastritis, the gastric carcinoma microbiota has increased nitrate reductase and nitrite reductase functions. This observation is compatible with the hypothesis that during carcinogenesis, changes in the stomach mucosa that lead to decreased acid secretion allow the growth of bacteria that are able to reduce nitrate to nitrite, a precursor of carcinogenic *N*-nitroso compounds.[2]

Taken together our results and previously published data, we propose that colonisation with bacteria other than *H. pylori*, namely gut commensals, contributes to alter the equilibrium between the 'resident' gastric microbiota and the host. This dysbiotic microbial community, by sustaining the gastric inflammatory process, and through its intrinsic genotoxic potential, may augment the risk for *H. pylori*-related gastric carcinoma development. In line with our proposal, experimental evidence in the INS-GAS model showed that commensal intestinal bacteria play a role in the promotion of gastric cancer.[12 13] Lertpiriyapong *et al* showed that mice harbouring a complex intestinal microbiota, and mice colonised with a restricted intestinal microbiota (that includes *Clostridium* and *Lactobacillus*), had an accelerated onset and progression of gastric cancer secondary to *H. pylori* infection. These mice also developed more severe gastric histopathology and higher expression levels of proinflammatory genes in comparison to germ-free mice (infected or not with *H. pylori*) and mice harbouring a complex or a restricted intestinal microbiota.[11]

Although our study is limited by its retrospective nature, and by the low number of patients with true premalignant lesions, our findings are consistent with a shift in the gastric microbial community structure along gastric carcinogenesis. In this sense, prospective follow-up studies of patients with premalignant lesions, successfully eradicated or not for *H. pylori* infection, would be crucial to ascertain the pathogenic effect of microbial dysbiosis in the progression to carcinoma. Additional studies to address the effect of dysbiosis or of candidate

bacterial species in an animal model of gastric carcinogenesis can also be considered, and in that regard, a humanised mouse model that better mimics the human immune response could be particularly informative. Ultimately, understanding the microbiota dynamics along gastric carcinogenesis may impact gastric carcinoma prevention and treatment strategies of patients with precancerous disease.

Contributors Study concept and design: CF, JCM. Data acquisition, analysis and interpretation: CF, JCM, RMF, JPM, IPR, JLC, FC. Drafting of the manuscript or revising it critically for important intellectual content: CF, JCM, RMF, JPM, IPR, JLC, FC. Obtained funding: CF, JCM.

Funding This research was supported by a Worldwide Cancer Research grant to CF and JCM (Reference 16-1352). RMF, JPM and IPR have fellowships from Fundação para a Ciência e a Tecnologia (FCT; SFRH/BPD/84084/2012, PD/BD/114014/2015 and SFRH/BD/110803/2015, respectively) through Programa Operacional Capital Humano (POCH) and the European Union. JPM's fellowship is in the framework of FCT's PhD Programme BiotechHealth (Ref PD/0016/2012). i3S-Instituto de Investigação e Inovação em Saúde is funded by Fundo Europeu de Desenvolvimento Regional (FEDER) funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by Portuguese funds through Fundação para a Ciência e a Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Inovação (POCI-01-0145-FEDER-007274).

Competing interests None declared.

Patient consent Detail has been removed from this case description/these case descriptions to ensure anonymity. The editors and reviewers have seen the detailed information available and are satisfied that the information backs up the case the authors are making.

Ethics approval Ethics Committee Centro Hospitalar São João.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Ferlay J, Soerjomataram I, Dikshit R, *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;136:E359–6.
2 Correa P. Human gastric carcinogenesis: a multistep and multifactorial process-first American cancer society award lecture on cancer epidemiology and prevention. *Cancer Res* 1992;52:6735–40.
3 Plottel CS, Blaser MJ. Microbiome and malignancy. *Cell Host Microbe* 2011;10:324–35.
4 El-Omar EM, Oien K, El-Nujumi A, *et al*. Helicobacter pylori infection and chronic gastric acid hyposecretion. *Gastroenterology* 1997;113:15–24.
5 Fukase K, Kato M, Kikuchi S, *et al*. Effect of eradication of Helicobacter pylori on incidence of metachronous gastric carcinoma after endoscopic resection of early gastric cancer: an open-label, randomised controlled trial. *Lancet* 2008;372:392–7.
6 Ma JL, Zhang L, Brown LM, *et al*. Fifteen-year effects of Helicobacter pylori, garlic, and vitamin treatments on gastric cancer incidence and mortality. *J Natl Cancer Inst* 2012;104:488–92.
7 Wong BC, Lam SK, Wong WM, *et al*. Helicobacter pylori eradication to prevent gastric cancer in a high-risk region of China: a randomized controlled trial. *JAMA* 2004;291:187–94.
8 Stockbruegger RW. Bacterial overgrowth as a consequence of reduced gastric acidity. *Scand J Gastroenterol Suppl* 1985;111:7–15.
9 Leach SA, Thompson M, Hill M. Bacterially catalysed N-nitrosation reactions and their relative importance in the human stomach. *Carcinogenesis* 1987;8:1907–12.
10 Lofgren JL, Whary MT, Ge Z, *et al*. Lack of commensal flora in helicobacter pylori-infected INS-GAS mice reduces gastritis and delays intraepithelial neoplasia. *Gastroenterol* 2011;140:210–20.
11 Lertpiriyapong K, Whary MT, Muthupalani S, *et al*. Gastric colonisation with a restricted commensal microbiota replicates the promotion of neoplastic lesions by diverse intestinal microbiota in the Helicobacter pylori INS-GAS mouse model of gastric carcinogenesis. *Gut* 2014;63:54–63.
12 Bik EM, Eckburg PB, Gill SR, *et al*. Molecular analysis of the bacterial microbiota in the human stomach. *Proc Natl Acad Sci U S A* 2006;103:732–7.
13 Maldonado-Contreras A, Goldfarb KC, Godoy-Vitorino F, *et al*. Structure of the human gastric bacterial community in relation to Helicobacter pylori status. *Isme J* 2011;5:574–9.
14 Delgado S, Cabrera-Rubio R, Mira A, *et al*. Microbiological survey of the human gastric ecosystem using culturing and pyrosequencing methods. *Microb Ecol* 2013;65:763–72.
15 Aviles-Jimenez F, Vazquez-Jimenez F, Medrano-Guzman R, *et al*. Stomach microbiota composition varies between patients with non-atrophic gastritis and patients with intestinal type of gastric cancer. *Sci Rep* 2014;4:4202.
16 Dicksved J, Lindberg M, Rosenquist M, *et al*. Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls. *J Med Microbiol* 2009;58:509–16.
17 Figueiredo C, Machado JC, Pharoah P, *et al*. Helicobacter pylori and interleukin 1 genotyping: an opportunity to identify high-risk individuals for gastric carcinoma. *J Natl Cancer Inst* 2002;94:1680–7.
18 Machado JC, Figueiredo C, Canedo P, *et al*. A proinflammatory genetic profile increases the risk for chronic atrophic gastritis and gastric carcinoma. *Gastroenterol* 2003;125:364–71.
19 Yu G, Torres J, Hu N, *et al*. Molecular characterization of the human stomach microbiota in gastric cancer patients. *Front Cell Infect Microbiol* 2017;7:302.
20 Andersson AF, Lindberg M, Jakobsson H, *et al*. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* 2008;3:e2836.
21 Walters WA, Caporaso JG, Lauber CL, *et al*. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* 2011;27:1159–61.
22 Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 2013;10:996–8.
23 Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;26:2460–1.
24 Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7:335–6.
25 DeSantis TZ, Hugenholtz P, Larsen N, *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;72:5069–72.
26 Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 2001;26:32–46.
27 Segata N, Izard J, Waldron L, *et al*. Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;12:R60.
28 Langille MG, Zaneveld J, Caporaso JG, *et al*. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;31:814–21.
29 Parks DH, Tyson GW, Hugenholtz P, *et al*. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 2014;30:3123–4.
30 White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009;5:e1000352.
31 Gevers D, Kugathasan S, Denson LA, *et al*. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;15:382–92.
32 Coker OO, Dai Z, Nie Y, *et al*. Mucosal microbiome dysbiosis in gastric carcinogenesis. *Gut* 2017.doi: 10.1136/gutjnl-2017-314281. [Epub ahead of print 1 Aug 2017].
33 Lepage P, Häsler R, Spehlmann ME, *et al*. Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterol* 2011;141:227–36.
34 Ahn J, Sinha R, Pei Z, *et al*. Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst* 2013;105:1907–11.
35 Thorens J, Froehlich F, Schwizer W, *et al*. Bacterial overgrowth during treatment with omeprazole compared with cimetidine: a prospective randomised double blind study. *Gut* 1996;39:54–9.
36 Schulz C, Schütte K, Koch N, *et al*. The active bacterial assemblages of the upper GI tract in individuals with and without Helicobacter infection. *Gut* 2016.doi: 10.1136/gutjnl-2016-312904. [Epub ahead of print 5 Dec 2016].
37 Li XX, Wong GL, To KF, *et al*. Bacterial microbiota profiling in gastritis without Helicobacter pylori infection or non-steroidal anti-inflammatory drug use. *PLoS One* 2009;4:e7985.
38 Yang I, Woltemate S, Piazuelo MB, *et al*. Different gastric microbiota compositions in two human populations with high and low gastric cancer risk in Colombia. *Sci Rep* 2016;6:18594.
39 Kelly CP, LaMont JT. Clostridium difficile--more difficult than ever. *N Engl J Med* 2008;359:1932–40.
40 Rajilić-Stojanović M, de Vos WM. The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol Rev* 2014;38:996–1047.
41 Jiao YS, Yan H, Ji ZJ, *et al*. Phyllobacterium sophorae sp. nov., a symbiotic bacterium isolated from root nodules of Sophora flavescens. *Int J Syst Evol Microbiol* 2015;65:399–406.
42 Forsythe SJ, Dolby JM, Webster AD, *et al*. Nitrate- and nitrite-reducing bacteria in the achlorhydric stomach. *J Med Microbiol* 1988;25:253–9.

43 Sjöstedt S, Heimdahl A, Kager L, *et al*. Microbial colonization of the oropharynx, esophagus and stomach in patients with gastric diseases. *Eur J Clin Microbiol* 1985;4:49–51.
44 Mowat C, Williams C, Gillen D, *et al*. Omeprazole, Helicobacter pylori status, and alterations in the intragastric milieu facilitating bacterial N-nitrosation. *Gastroenterol* 2000;119:339–47.
45 Barthold SW, Jonas AM. Morphogenesis of early 1, 2-dimethylhydrazine-induced lesions and latent period reduction of colon carcinogenesis in mice by a variant of Citrobacter freundii. *Cancer Res* 1977;37:4352–60.
46 Newman JV, Kosaka T, Sheppard BJ, *et al*. Bacterial infection promotes colon tumorigenesis in Apc(Min/+) mice. *J Infect Dis* 2001;184:227–30.

Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota

Rui M Ferreira, Joana Pereira-Marques, Ines Pinto-Ribeiro, Jose L Costa, Fatima Carneiro, Jose C Machado and Ceu Figueiredo

Gut published online November 4, 2017

Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota

Rui M. Ferreira[1,2], Joana Pereira-Marques[1,2,3], Ines Pinto-Ribeiro[1,2,4], Jose L. Costa[1,2,4], Fatima Carneiro[1,2,4,5], Jose C. Machado[1,2,4], Ceu Figueiredo[1,2,4*]

[1]i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal; [2]Ipatimup - Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal; [3]ICBAS - Institute of Biomedical Sciences Abel Salazar, University of Porto, Portugal; [4]Faculty of Medicine of the University of Porto, Porto, Portugal; and [5]Department of Pathology, Centro Hospitalar São João, Porto, Portugal

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY MATERIALS AND METHODS

Helicobacter pylori detection

DNA extracted from gastric tissue was used for *H. pylori* detection by amplification of the *vacA* and *cagA* genes, using biotin-labeled primers as previously described.[1] The detection of amplicons was performed by reverse hybridization onto a line probe assay (Labo Bio-Medical Products, Rijswijk, The Netherlands). DNA of *H. pylori* strains with known genotypes were used as templates for positive and negative controls.

16S rRNA gene sequencing

The PCR reactions were performed in 25µL containing 1X AmpliTaq Gold 360 Master Mix (Applied Biosystems, Foster City, CA) and 0.4µM of forward and reverse primers. PCR was performed with 9 min of predenaturation at 95ºC, followed by 25 cycles of 30 seconds at 95ºC, 45 seconds at 52ºC, and 45 seconds at 72ºC. Final extension was performed for 10 minutes at 72ºC. Amplicons of approximately 280 bp were visualized and purified using the E-Gel SizeSelect Agarose Gels (Life Technologies, Foster City, CA), their concentration was determined with Qubit dsDNA HS Assay Kit (Life Technologies) and the respective size distribution with Qiaxcel DNA screening (Qiagen, Germany). Equal concentrations were used for library preparation to incorporate adaptor and barcode sequences. Next-generation sequencing was performed in the Ion PGM Torrent platform using Ion 316v2 Chips (Life, Technologies), following manufacturer's protocols. Briefly, 50ng of amplified DNA was used to ligate the Ion Torrent A (containing the sequencing primer ligation site) and P1 adaptors (the site for Ion Sphere Particles-ISP ligation). Afterwards, the DNA molecules entered in a process of clonal amplification through an emulsion PCR (emPCR) using the OneTouch2 instrument (Ion Torrent, Life Technilogies). During emPCR single DNA fragments become bound via the specific adaptor to single ISPs. This process leads to coating of each ISP with millions of copies of a single DNA fragment. To ensure only coated ISPs proceed for sequencing, an enrichment step was performed using the OneTouch ES instrument (Ion Torrent, Life Technologies). PCR negative controls containing Microbial DNA-free water

(Qiagen), instead of DNA were processed as above mentioned. Data were deposited in Sequence Read Archive (PRJNA413125).

Target re-sequencing quality control

The number of spurious OTUs and the number of biological meaningful OTUs were obtained according to similarity shared with sequences of the Greengenes Named Isolated database (Supplementary Figure S2).[2] OTUs were classified as "Named" (≥ 97% match from the reference database), "Chimeric" (<97% match to the reference database and chimeric with high confidence) and "Missing" (<97% match with low confidence or a biological sequence missing from the database). Consistency in the values determined for alpha-diversity, beta-diversity and relative abundances of *Proteobacteria*, *Firmicutes*, *Bacteroidetes* and *Fusobacteria* across amplification and sequencing sets were evaluated by intraclass correlation coefficient in 32 samples derived from 12 chronic gastritis and 4 gastric carcinoma subjects (Supplementary Table S3).

NGS data analysis of the Chinese and Mexican validation cohorts

The gastric carcinoma raw paired-end reads of the 16S rRNA gene of the V3-V4 region were downloaded in FASTQ format from Sequence Read Archive (SRA) from the study SRP080738 (BioProject PRJNA310127). In total 80 samples from China and 54 samples from Mexico were downloaded. Reads were analysed using the UPARSE pipeline, as described for the discovery cohort. Chimeric reads were reference-removed using Uchime. Each OTU was taxonomically assigned using Uclust, considering a minimum of 90% of similarity to a reference database (Greengenes Named Isolate database, release August 2013). After quality filtering, OTU construction, chimera filtering, and removal of non-bacteria sequences, two samples (one from each cohort) were removed due to low number of reads (< 1000 sequences). The final set comprised 79 gastric cancer samples from China and 53 from Mexico (Supplementary Table S2). Noteworthy, the number of reads that could not be assigned to bacteria in these datasets were remarkably high (47.8% and 38.2% for China and Mexico, respectively) in comparison to those observed in the Portuguese discovery cohort (0.38%).

Real-time quantitative PCR (qPCR)

qPCR assays were performed using PowerUp SYBR Green Master Mix (Applied Biosystems) using different sets of primers. Two different assays were used for quantification, a universal assay (composed by universal primers targeting a conserved region of the 16S rRNA gene) and a specific assay (composed by genus-specific primers targeting the 16S rRNA gene) (Supplementary Table S2). The assays were designed to obtain the highest degree of specificity by comparison with sequences of Greengenes Named Isolate (release August 2013).

qPCR mixtures were prepared to a final volume of 10 µL, containing 1x PowerUp SYBR Green Master Mix, 1 µM of forward and reverse primers (Invitrogen, Foster City, CA), 2 µL of Microbial DNA-Free Water (Qiagen) and 1 µL of DNA. The qPCR was performed in a 7500 Fast Real-Time PCR System (Applied Biosystems) with the following conditions: 2 minutes at 50ºC, 10 minutes at 95ºC, followed by 40 cycles of denaturation at 95ºC for 15 seconds and annealing/extension at 60ºC for 1 minute. The amplification steps were followed by a melt dissociation step to check for nonspecific product formation. This step comprises an additional cycle of 95ºC for 15 seconds, 60ºC for 1 minute, 95ºC for 30 seconds and 60ºC for 15 seconds. In addition, the PCR product purity was also controlled by Agarose gel electrophoresis. Two replicates were performed for each sample. To exclude any potential environmental contaminant in PCR reactions, blanks were prepared using Microbial DNA-Free Water (Qiagen) instead of DNA.

The relative standard curve method was used to quantify the specific genera in patients with chronic gastritis and with gastric carcinoma. To create standard curves, amplicons of each designed assay were cloned into pGEM-T easy vector system (Promega, Madison, WI). Dilution series of known plasmid concentrations were used to create a standard curve for each assay by plotting the log of each known concentration in the dilution series against the determined C_T (threshold cycle) value.

From the standard curves, the reaction parameters (slope, *y*-intercept, correlation coefficient and efficiency) were obtained and the concentration of the target bacteria species was extrapolated. For each assay, the PCR efficiencies (determined as $10^{(-1/slope)} - 1$) obtained were: Universal – 91%; *Helicobacter* sp. – 100%; *Streptococcus* sp. – 115%; *Neisseria* sp. – 92%; *Prevotella* sp. – 101%; *Achromobacter* sp. – 114%; *Citrobacter* sp. – 97%; *Clostridium* sp. – 83%; *Rhodococcus*

sp. – 102%; *Lactobacillus* sp. – 91%; and *Phyllobacterium* sp. – 100%. The abundance of each genus was determined by the log10 ratio between the DNA concentration determined for the specific assay and the DNA concentration determined for the universal assay. The Student's t-test was used to compare the abundance of genus between chronic gastritis and gastric carcinoma cases.

Interpolation of linear and non-linear models and correlations

Interpolation of non-linear models was used to explain the relationship between two variables. These models were determined using the following equation (second order polynomial):

$y = B0 + B1x + B2x^2$

where *x* represents the relative abundance of *Helicobacter* spp. and *y* represents the Shannon index. B0, B1 and B2 represent the polynomial coefficients. Non-linear models were shown with the 95% confidence intervals lines. Deviation of data from non-linear models was evaluated by performing run tests, which apply a series of consecutive points that are above or below the interpolation curve and determine the probability that data diverge significantly (at a *P*- value ≤ 0.05).

The normality of the data was evaluated using the Kolmogorov-Smirnov test. Correlations between variables were performed using Pearson's correlation (for normally distributed data) or Spearman's rank correlation (for non-normally distributed data).

Logistic regression and receiver operating characteristic (ROC) analyses

The risk of gastric carcinoma was evaluated by multivariate logistic regression analysis, using clinical outcome as dependent variable and taxa prevalence as independent variable, adjusted for age and gender. The Hosmer-Lemeshow test was used to control goodness-of-fit of logistic regression models using 8 degrees of freedom and 10 steps. ROC curves were constructed to evaluate the ability of relative abundance of genera and of MDI to detect gastric carcinoma. The relative abundance was defined as the raw counts of genus-level taxa detected in at least one sample that were then normalized per sample by the total counts of all taxa in that sample so that the resulting relative abundances sum 100%. A mean ROC curve was reported including the 95%

confident intervals. The best discrimination was determined by using the highest area under the curve at a significance value of $P \leq 0.05$.

Microbial dysbiosis index (MDI)

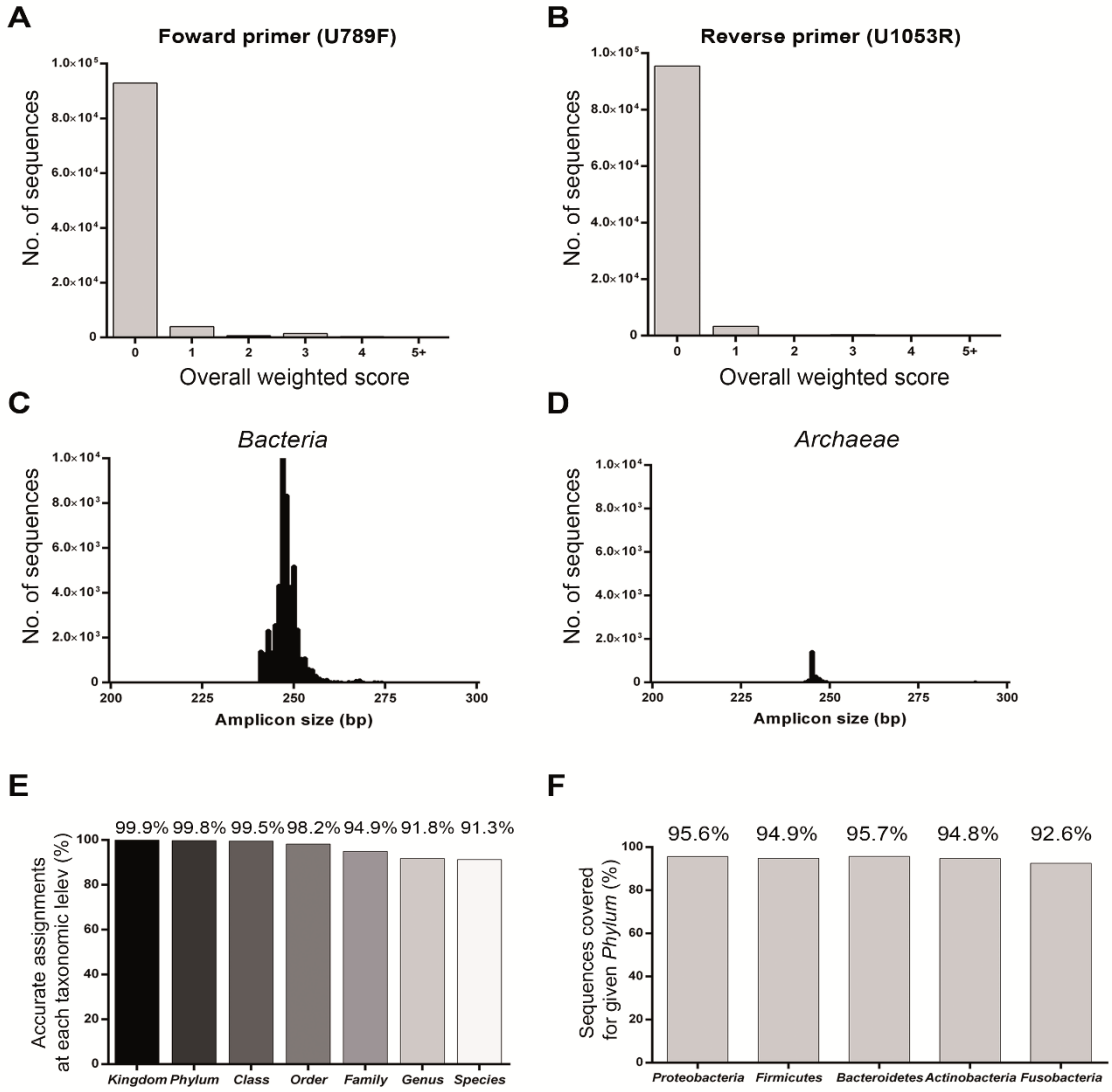
The MDI was determined as the log transformation of the ratio between the total abundance of genera increased in gastric carcinoma and the total abundance of genera decreased in gastric carcinoma.[3] Unless otherwise stated, *Rhodococcus* spp., *Lactobacillus* spp., *Clostridium* spp., *Phyllobacterium* spp., *Achromobacter* spp. and *Citrobacter* spp. were included as increased in gastric carcinoma, and *Helicobacter* spp., *Neisseria* spp., *Prevotella* spp. and *Streptococcus* spp. were included as decreased in gastric carcinoma.

Functional metagenome predictions

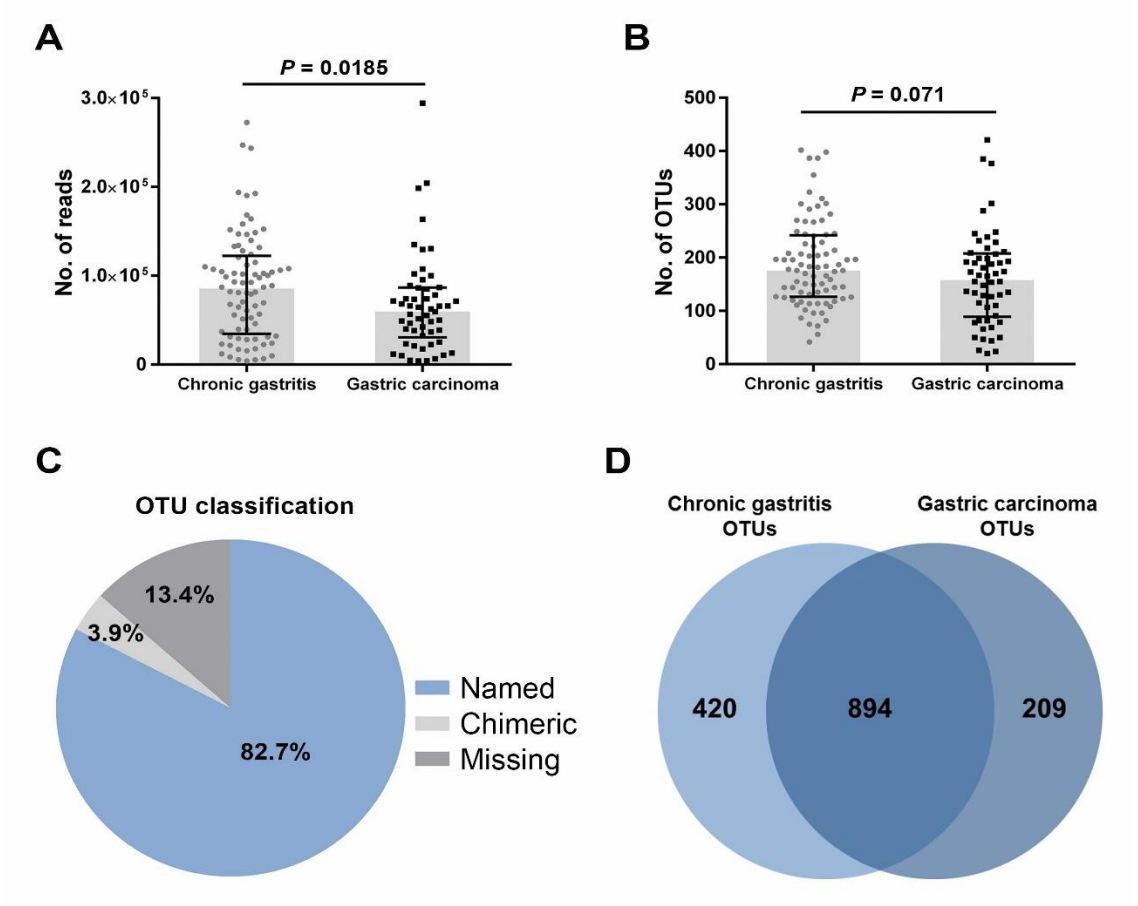
To predict the functional metagenome, we captured OTU representative sequences from Greengenes database using USEARCH global alignment command. Thus, only sequences with best hit to the corresponding taxonomy were kept, i.e. all OTUs with a hit of 97% similarity to the reference database were preserved. The reconstruction of the functional metagenome content was performed using PICRUSt.[4] This pipeline is a computational method to predict the gene function composition of a metagenome using 16S rRNA sequencing data and a database of reference genomes. The abundance of each OTU was corrected to reflect the true bacterial abundance by normalizing the 16S rRNA copy number for each OTU. The accuracy of the predicted metagenomes was determined by the NSTI (nearest sequenced taxon index). NSTI quantifies the availability of nearby representative genomes for each microbiome sample and is determined as the sum of phylogenetic distances for each microbe in the OTU table to its adjacent relative with a sequenced reference genome. Low NSTI values (closer to zero) indicate better accuracy in metagenome prediction.[4] Predicted functional genes were classified into clusters of orthologous groups (COG) that contains prokaryotic proteins of complete genomes or into Kyoto encyclopedia of genes and genomes orthology (KO) and submitted to a two-group comparison (gastritis vs carcinoma) using STAMP.[5] Differences in COG and KO relative frequencies were determined by White's non-parametric *t* test[6] with a Benjamini-Hochberg false discovery rate correction to adjust *P*-values for

multiple testing. The gene functions classified by KO were further categorized into KEGG Pathways using PICRUSt. The enrichment of predicted KEGG Pathways was assessed by White's non-parametric t test with a Benjamini-Hochberg false discovery rate correction and by LEfSe analysis. The functional contribution of taxa to selected predicted functions was evaluated with metagenome_contributions.py, a built-in function in PICRUSt.

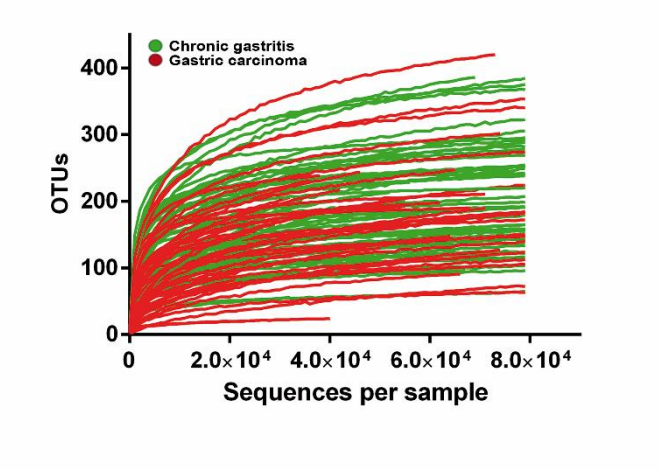
SUPPLEMENTARY FIGURES



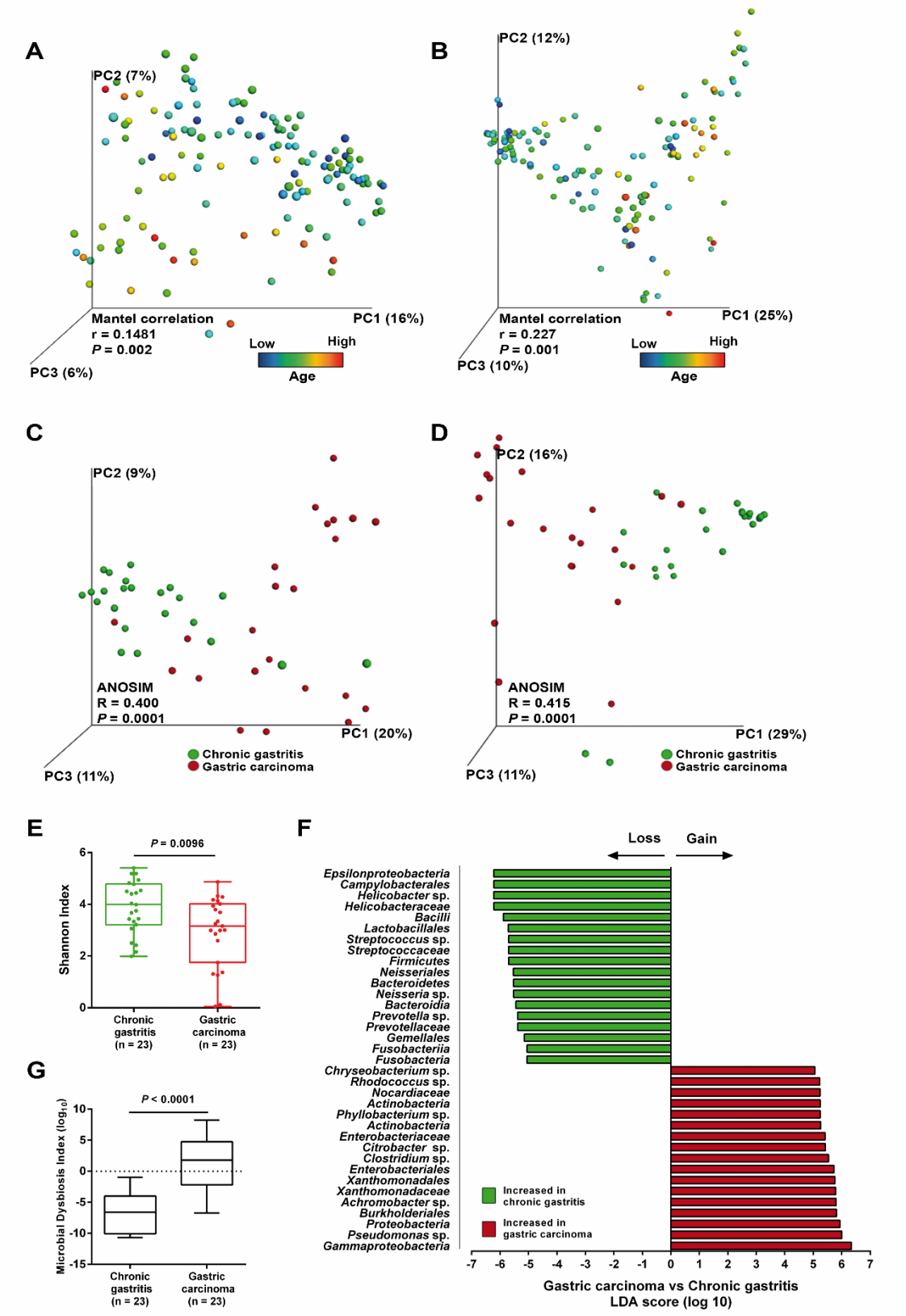
Supplementary Figure S1. Performance of the selected primers (U789F and U1053R) for 16S rRNA V5-V6 next-generation sequencing evaluated with Primer Analysis Pipeline implemented in PrimerProspector software package.[7] Number of sequences covered by the **(A)** forward and by the **(B)** reverse primer. The overall weighted score represents the performance of the primers by measuring penalties of primers to match reference sequences. Overall weighted score = non-3' mismatch × 0.40 + 3' mismatch ×1 + non-3' gaps × 1 + 3' gaps × 3. Simulated reads using the U789F and U1053R showed high number of **(C)** *Bacteria* sequences and low number of **(D)** *Archaea* sequences. Taxonomic coverage of simulated reads by **(E)** taxonomic level and by most relevant **(F)** *Phylum* represented in the gastric microbiota. The representative OUT sequences (clustered at 97% similarity) of the Greengenes Named Isolate database was used as reference for this analysis.



Supplementary Figure S2. Performance of the UPARSE pipeline[6] in 16S rRNA V5-V6 reads derived from 81 chronic gastritis and 54 gastric carcinoma patients. Number of reads **(A)** and OTUs **(B)** after quality filtering. Mann-Whitney test was used to compare chronic gastritis and gastric carcinoma patients in respect to the number of reads or OTUs. Data is shown as median with interquartile range. **(C)** Pie chart showing the average fraction of OTUs classified accordingly to similarity shared with sequences of the Greengenes Named Isolate database, as previously reported[6, 14]: Named (≥ 97% match from the reference database), Chimeric (< 97% match to the reference database and chimeric with high confidence) and Missing (< 97% match with low confidence or a biological sequence missing from the database). **(D)** Venn diagram showing the number of OTUs exclusively identified in each group of patients and OTUs shared by the two groups of patients.

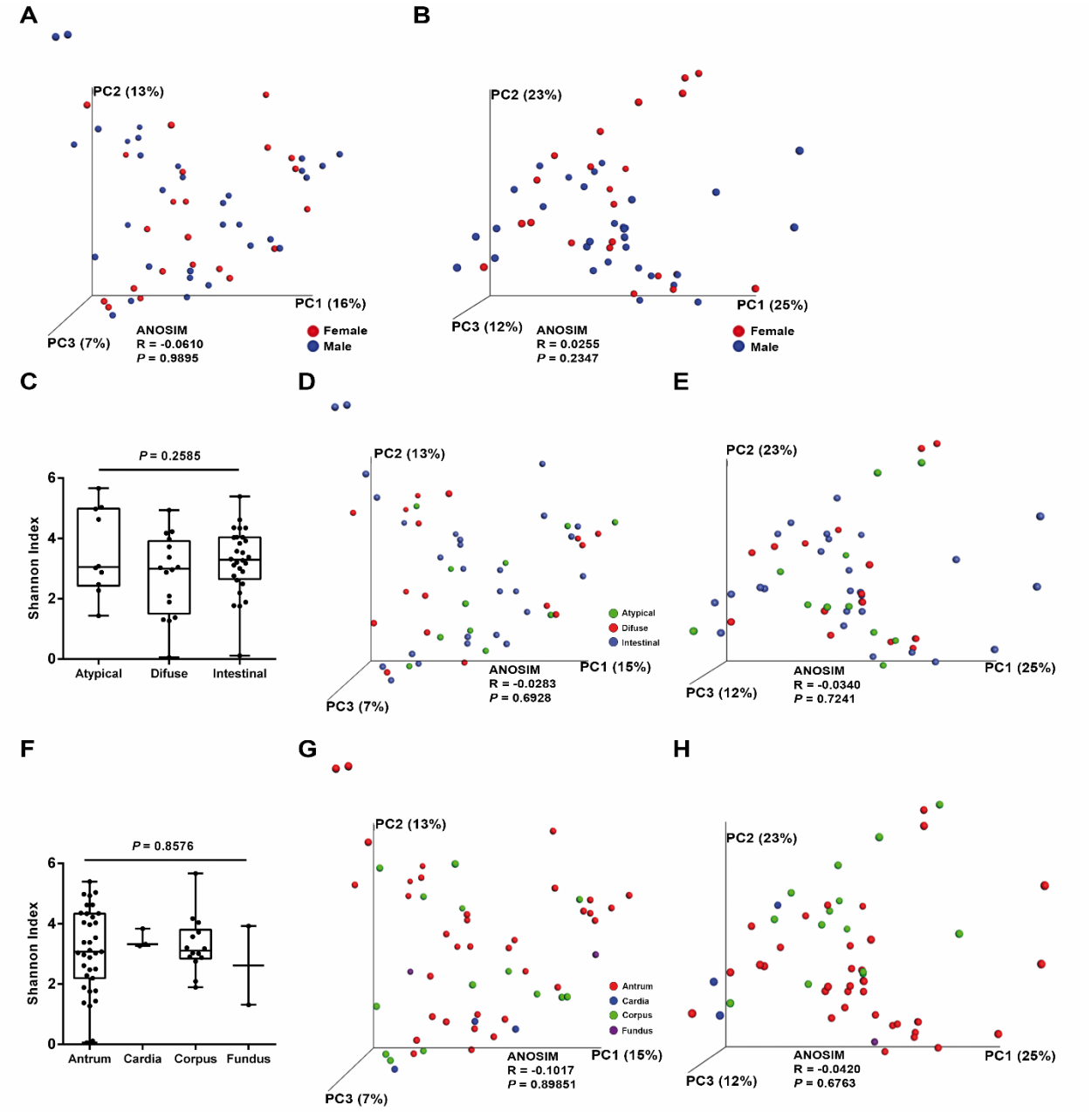


Supplementary Figure S3. Rarefaction curves of the number of OTUs versus the sequencing effort per sample. Rarefaction curves were estimated by bootstrapping of 20 random samples at 100 sequence increments to a maximum of 80,000 reads.

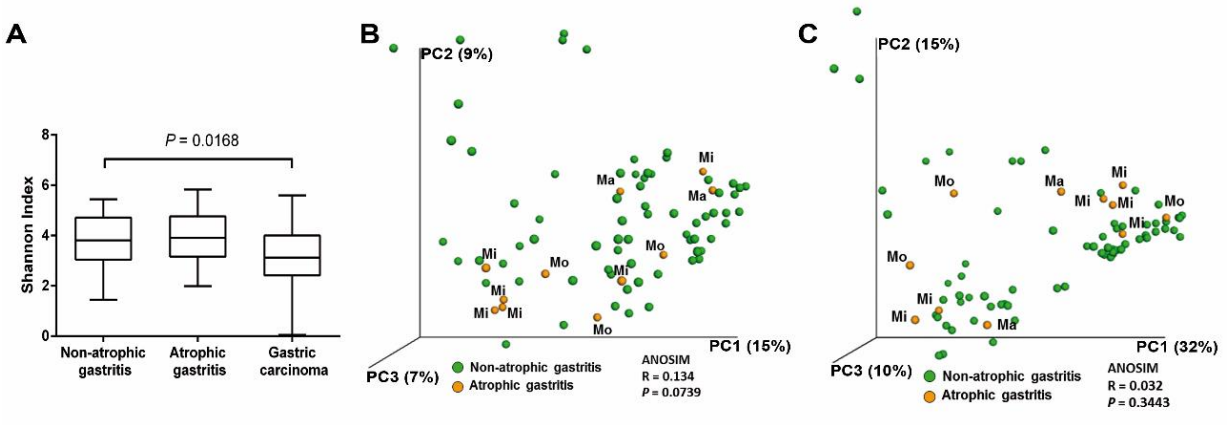


Supplementary Figure S4. PCoA plots of the unweighted **(A)** and weighted **(B)** UniFrac distances in the full sample set (discovery cohort) in which samples were coloured by increasing age. Mantel correlations controlled with 10^4 permutations were used to compare distances. PCoA plots of the unweighted **(C)** and weighted **(D)** UniFrac distances in age-matched chronic gastritis (n = 23) and

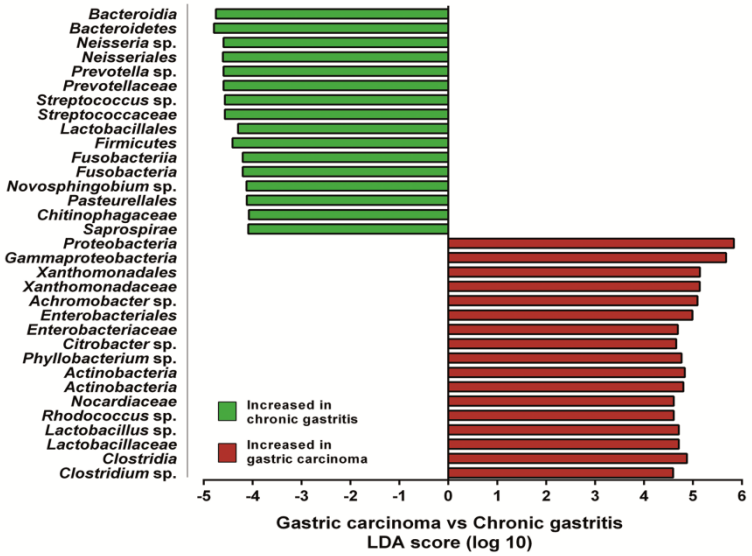
gastric carcinoma (n = 23) cases. Samples were coloured by diagnosis. Groups of patients were compared with t tests using 10^3 Monte Carlo permutations. **(E)** Shannon index of diversity among the age-matched chronic gastritis and gastric carcinoma cases. **(F)** LEfSe analysis showing the association of specific microbiota taxa with the clinical diagnosis. **(G)** Box plot showing the MDI in the age-matched chronic gastritis and gastric carcinoma cases.



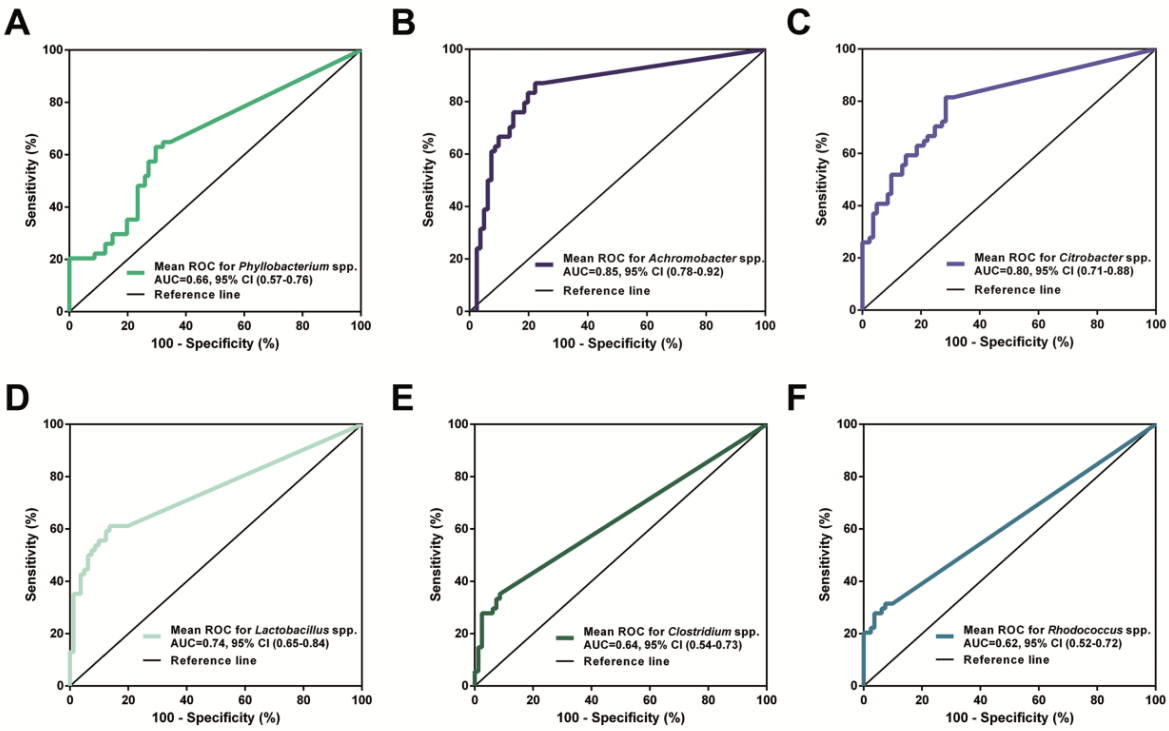
Supplementary Figure S5. (A-B) PCoA plots of the unweighted **(A)** and weighted UniFrac **(B)** distances in gastric carcinoma cases in the discovery cohort. Samples were coloured by gender. **(C)** Box plot showing the Shannon index across histological type of gastric carcinoma. **(D-E)** PCoA plots of the unweighted **(D)** and weighted **(E)** UniFrac distances in gastric carcinoma. Samples were coloured by histological type**. (F)** Box plot showing the Shannon index according to tumour location. **(G-H)** PCoA plots of the unweighted **(G)** and weighted **(H)** UniFrac distances in gastric carcinoma. Samples were coloured by tumour location.



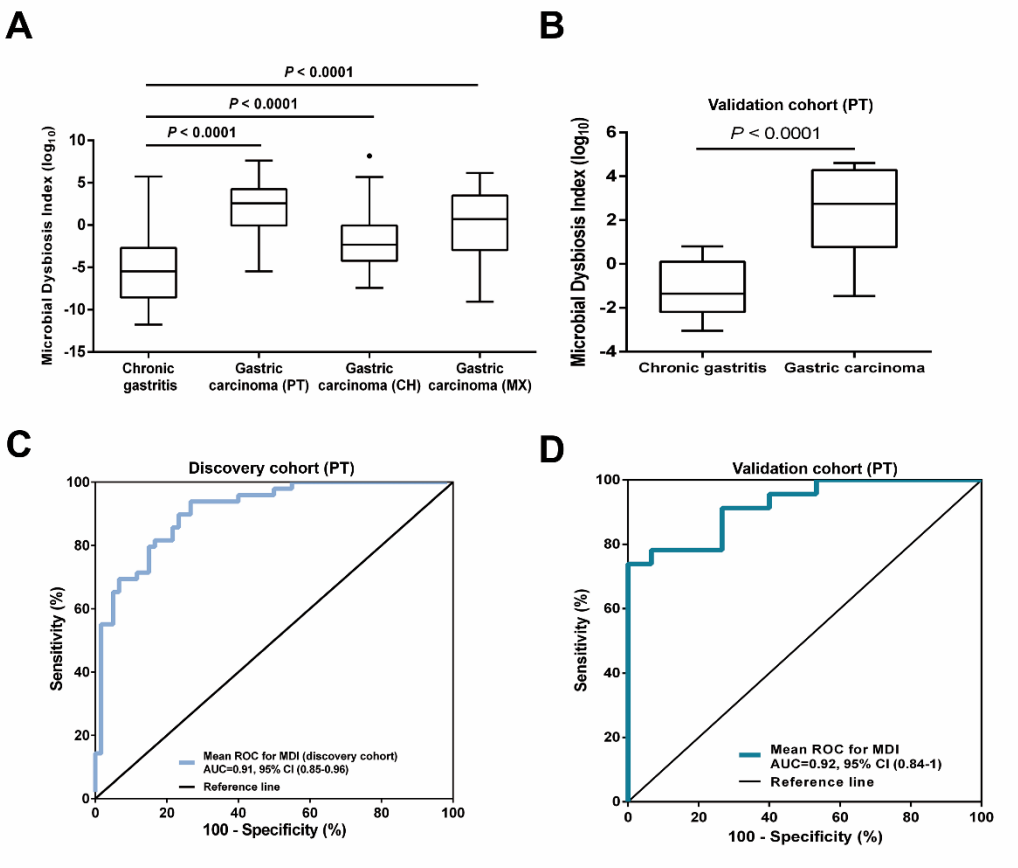
Supplementary Figure S6. (A) Boxplot showing the Shannon index of diversity in 70 non-atrophic gastritis patients, 11 patients with glandular atrophy and 54 gastric carcinoma patients. PCoA plots of the **(B)** unweighted and **(C)** weighted UniFrac distances in the group of chronic gastritis. Samples were colored by the presence of glandular atrophy. Mi, Mo, and Ma, stand for mild (n = 6), moderate (n = 3), and marked (n = 2) glandular atrophy, respectively. Gastric carcinoma samples were omitted in these plots. PC1, PC2 and PC3 represent the top three principal coordinates that captured the maximum diversity. The percentage of diversity captured by each coordinate is shown.



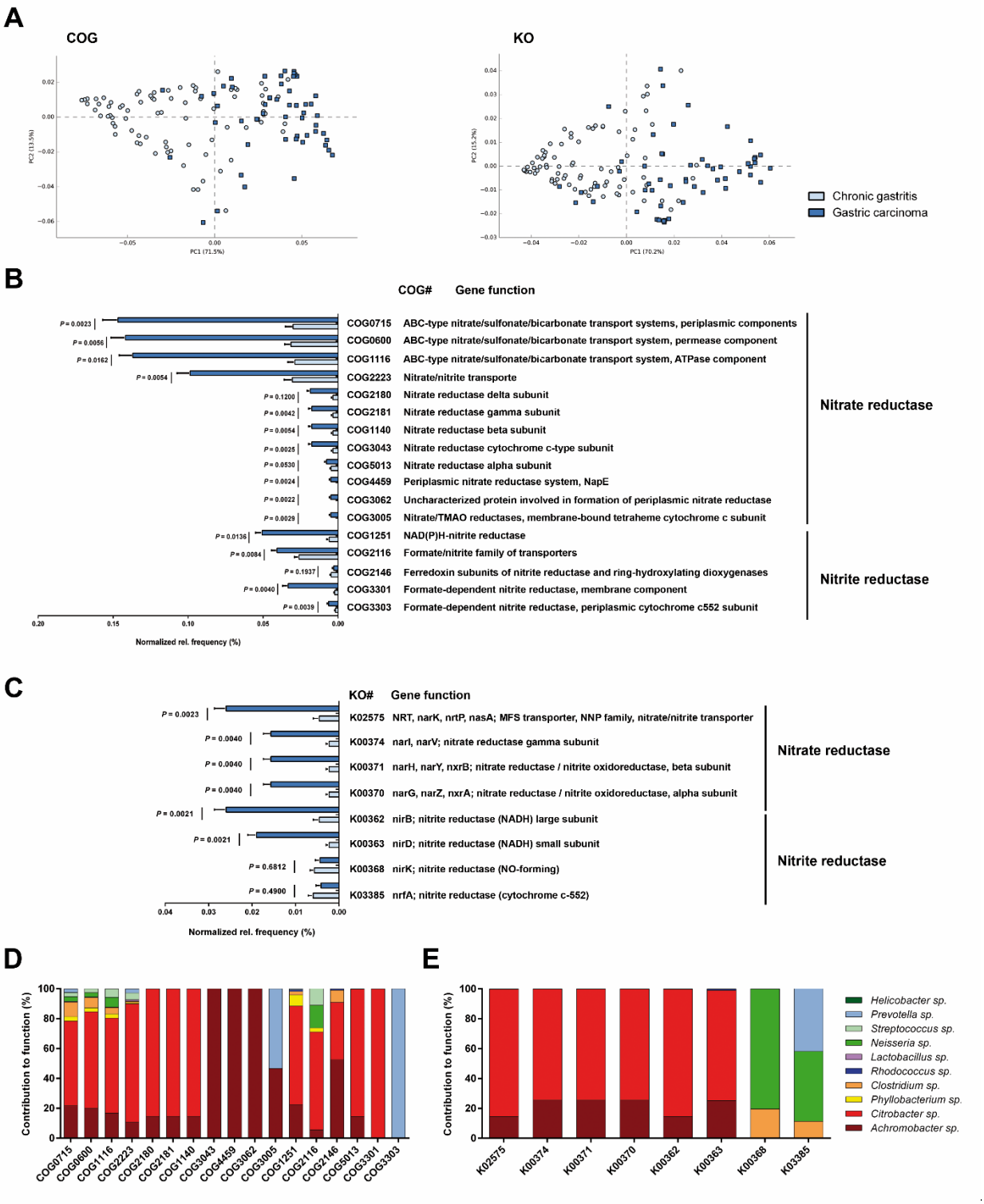
Supplementary Figure S7. LEfSe re-analysis, in which *Helicobacter* spp. reads were subtracted from the dataset. Green indicates taxa enriched in chronic gastritis group and red indicates taxa enriched in gastric carcinoma group.



Supplementary Figure S8. Receiver operating characteristic (ROC) curves analysis to evaluate the discriminatory potential of abundance of **(A)** *Phyllobacterium* spp., **(B)** *Achromobacter* spp., **(C)** *Citrobacter* spp., **(D)** *Lactobacillus* spp., **(E)** *Clostridium* spp. and **(F)** *Rhodococcus* spp. The discriminatory potential of a biomarker was determined by the highest area under the curve.



Supplementary Figure S9. Microbial dysbiosis index calculated excluding *Prevotella* sp. and *Streptococcus* sp. **(A)** Box plot showing the MDI in the discovery cohort and in the Chinese and Mexican validation cohorts. **(B)** Box plot showing the MDI the Portuguese validation cohort. Significance was obtained by Student's t-test. **(C-D)** ROC curve analysis to evaluate the discriminatory potential of MDI in gastric carcinoma detection in the discovery cohort **(C)** and in the Portuguese validation cohort **(D).**



Supplementary Figure S10. (A) Principal component analysis (PCA) plot comparing the metagenome predictions (for COG and KO using PICRUSt) of the microbiota of patients with chronic gastritis and gastric carcinoma. **(B-C)** Normalized relative frequency of nitrate reductase **(B)** and nitrite reductase **(C)** functions predicted with the 10 genera differentially abundant in the two patient groups. Significance was considered for adjusted $P < 0.05$. **(D-E)** Genera contribution to each nitrate and nitrate reductase functions (Supplementary Table S5).

SUPPLEMENTARY TABLES

Supplementary Table S1. Characteristics of the patients in the Portuguese discovery and validation cohorts.

Patient characteristics	Discovery cohort			Validation cohort		
	Chronic gastritis (n = 81)	Gastric carcinoma (n = 54)	*P* *	Chronic gastritis (n = 15)	Gastric carcinoma (n = 23)	*P* *
Age (mean ± S.D.)	43.6 ± 7.0	58.8 ± 13.2	< 0.001	58.1 ± 16	68.9 ± 13.5	0.032
Gender (n, %)						
Male	79 (97.5)	32 (59.3)	< 0.001	5 (33.3)	16 (69.6)	0.028
Female	2 (2.5)	22 (40.7)		10 (66.7)	7 (30.4)	
H. pylori status determined by PCR-LiPA (n, %)						
Positive	80 (98.8)	47 (87.0)	0.007	-	-	
Negative	1 (1.2)	7 (13.0)		-	-	
Histological type of gastric cancer (n, %)						
Intestinal	NA	27 (50.0)		NA	11 (47.8)	
Diffuse	NA	16 (29.6)		NA	0 (0)	
Indeterminate/mixed	NA	11 (20.4)		NA	12 (52.2)	
Tumor stage (n, %)						
T1	NA	4 (7.4)		NA	1 (4.3)	
T2	NA	3 (5.6)		NA	4 (17.4)	
T3	NA	47 (87.0)		NA	12 (52.2)	
T4	NA	0 (0)		NA	6 (26.1)	
Lymph node metastasis (n, %)						
Presence	NA	44 (81.5)		NA	16 (69.6)	
Absent	NA	10 (18.5)		NA	7 (30.4)	
Vascular invasion (n, %)						
Presence	NA	28 (51.9)		NA	16 (69.6)	
Absent	NA	26 (48.1)		NA	7 (30.4)	
Distant metastasis (n, %)						
Presence	NA	1 (1.9)		NA	0 (0)	
Absent	NA	53 (98.1)		NA	23 (100)	

NA, not applicable. **P*-value obtained by Fisher's exact test for categorical variables and by *t* test for continuous variables. Statistical tests were performed two-sided.

Supplementary Table S2. Characteristics of the patients and retrieval data in the Chinese and in the Mexican validation cohorts.

SampleID	Clinical_setting	Gender	Age	Local	SRA_Sample	SRA_Study	SRA_Project
Chi_GC_1	Gastric_carcinoma	Female	66	China	SRR3991143	SRP080738	PRJNA310127
Chi_GC_2	Gastric_carcinoma	Male	62	China	SRR3991145	SRP080738	PRJNA310127
Chi_GC_3	Gastric_carcinoma	Male	53	China	SRR3991147	SRP080738	PRJNA310127
Chi_GC_4	Gastric_carcinoma	Male	62	China	SRR3991149	SRP080738	PRJNA310127
Chi_GC_5	Gastric_carcinoma	Male	32	China	SRR3991151	SRP080738	PRJNA310127
Chi_GC_6	Gastric_carcinoma	Male	62	China	SRR3991153	SRP080738	PRJNA310127
Chi_GC_7	Gastric_carcinoma	Male	62	China	SRR3991155	SRP080738	PRJNA310127
Chi_GC_8	Gastric_carcinoma	Male	66	China	SRR3991157	SRP080738	PRJNA310127
Chi_GC_9	Gastric_carcinoma	Male	56	China	SRR3991159	SRP080738	PRJNA310127
Chi_GC_10	Gastric_carcinoma	Male	68	China	SRR3991161	SRP080738	PRJNA310127
Chi_GC_11	Gastric_carcinoma	Male	70	China	SRR3991163	SRP080738	PRJNA310127
Chi_GC_12	Gastric_carcinoma	Male	61	China	SRR3991165	SRP080738	PRJNA310127
Chi_GC_13	Gastric_carcinoma	Male	58	China	SRR3991167	SRP080738	PRJNA310127
Chi_GC_14	Gastric_carcinoma	Male	69	China	SRR3991169	SRP080738	PRJNA310127
Chi_GC_15	Gastric_carcinoma	Male	65	China	SRR3991171	SRP080738	PRJNA310127
Chi_GC_16	Gastric_carcinoma	Male	67	China	SRR3991173	SRP080738	PRJNA310127
Chi_GC_17	Gastric_carcinoma	Male	64	China	SRR3991175	SRP080738	PRJNA310127
Chi_GC_18	Gastric_carcinoma	Male	69	China	SRR3991177	SRP080738	PRJNA310127
Chi_GC_19	Gastric_carcinoma	Male	61	China	SRR3991179	SRP080738	PRJNA310127
Chi_GC_20	Gastric_carcinoma	Male	68	China	SRR3991181	SRP080738	PRJNA310127
Chi_GC_21	Gastric_carcinoma	Male	58	China	SRR3991183	SRP080738	PRJNA310127
Chi_GC_22	Gastric_carcinoma	Female	54	China	SRR3991185	SRP080738	PRJNA310127
Chi_GC_23	Gastric_carcinoma	Male	60	China	SRR3991187	SRP080738	PRJNA310127
Chi_GC_24	Gastric_carcinoma	Male	60	China	SRR3991189	SRP080738	PRJNA310127
Chi_GC_25	Gastric_carcinoma	Male	59	China	SRR3991191	SRP080738	PRJNA310127
Chi_GC_26	Gastric_carcinoma	Female	70	China	SRR3991193	SRP080738	PRJNA310127
Chi_GC_27	Gastric_carcinoma	Male	69	China	SRR3991195	SRP080738	PRJNA310127
Chi_GC_28	Gastric_carcinoma	Male	51	China	SRR3991197	SRP080738	PRJNA310127
Chi_GC_29	Gastric_carcinoma	Male	61	China	SRR3991199	SRP080738	PRJNA310127
Chi_GC_30	Gastric_carcinoma	Male	64	China	SRR3991201	SRP080738	PRJNA310127
Chi_GC_31	Gastric_carcinoma	Male	54	China	SRR3991203	SRP080738	PRJNA310127
Chi_GC_32	Gastric_carcinoma	Male	70	China	SRR3991205	SRP080738	PRJNA310127
Chi_GC_33	Gastric_carcinoma	Male	67	China	SRR3991207	SRP080738	PRJNA310127
Chi_GC_34	Gastric_carcinoma	Male	63	China	SRR3991209	SRP080738	PRJNA310127
Chi_GC_35	Gastric_carcinoma	Male	66	China	SRR3991211	SRP080738	PRJNA310127
Chi_GC_36	Gastric_carcinoma	Female	68	China	SRR3991213	SRP080738	PRJNA310127
Chi_GC_37	Gastric_carcinoma	Female	50	China	SRR3991215	SRP080738	PRJNA310127
Chi_GC_38	Gastric_carcinoma	Male	64	China	SRR3991217	SRP080738	PRJNA310127
Chi_GC_39	Gastric_carcinoma	Male	63	China	SRR3991219	SRP080738	PRJNA310127
Chi_GC_40	Gastric_carcinoma	Male	59	China	SRR3991221	SRP080738	PRJNA310127
Chi_GC_41	Gastric_carcinoma	Female	60	China	SRR3991223	SRP080738	PRJNA310127
Chi_GC_42	Gastric_carcinoma	Female	59	China	SRR3991225	SRP080738	PRJNA310127

Chi_GC_43	Gastric_carcinoma	Male	32	China	SRR3991227	SRP080738	PRJNA310127
Chi_GC_44	Gastric_carcinoma	Male	70	China	SRR3991229	SRP080738	PRJNA310127
Chi_GC_45	Gastric_carcinoma	Male	62	China	SRR3991231	SRP080738	PRJNA310127
Chi_GC_46	Gastric_carcinoma	Male	48	China	SRR3991232	SRP080738	PRJNA310127
Chi_GC_47	Gastric_carcinoma	Male	63	China	SRR3991234	SRP080738	PRJNA310127
Chi_GC_48	Gastric_carcinoma	Male	51	China	SRR3991236	SRP080738	PRJNA310127
Chi_GC_49	Gastric_carcinoma	Male	65	China	SRR3991237	SRP080738	PRJNA310127
Chi_GC_50	Gastric_carcinoma	Male	68	China	SRR3991239	SRP080738	PRJNA310127
Chi_GC_51	Gastric_carcinoma	Male	48	China	SRR3991241	SRP080738	PRJNA310127
Chi_GC_52	Gastric_carcinoma	Male	58	China	SRR3991243	SRP080738	PRJNA310127
Chi_GC_53	Gastric_carcinoma	Male	56	China	SRR3991245	SRP080738	PRJNA310127
Chi_GC_54	Gastric_carcinoma	Female	64	China	SRR3991247	SRP080738	PRJNA310127
Chi_GC_55	Gastric_carcinoma	Male	60	China	SRR3991249	SRP080738	PRJNA310127
Chi_GC_56	Gastric_carcinoma	Male	60	China	SRR3991251	SRP080738	PRJNA310127
Chi_GC_57	Gastric_carcinoma	Male	67	China	SRR3991253	SRP080738	PRJNA310127
Chi_GC_58	Gastric_carcinoma	Male	70	China	SRR3991255	SRP080738	PRJNA310127
Chi_GC_59	Gastric_carcinoma	Male	47	China	SRR3991257	SRP080738	PRJNA310127
Chi_GC_60	Gastric_carcinoma	Male	69	China	SRR3991259	SRP080738	PRJNA310127
Chi_GC_61	Gastric_carcinoma	Male	50	China	SRR3991261	SRP080738	PRJNA310127
Chi_GC_62	Gastric_carcinoma	Male	68	China	SRR3991263	SRP080738	PRJNA310127
Chi_GC_63	Gastric_carcinoma	Male	65	China	SRR3991265	SRP080738	PRJNA310127
Chi_GC_64	Gastric_carcinoma	Male	45	China	SRR3991267	SRP080738	PRJNA310127
Chi_GC_65	Gastric_carcinoma	Male	65	China	SRR3991269	SRP080738	PRJNA310127
Chi_GC_66	Gastric_carcinoma	Female	59	China	SRR3991271	SRP080738	PRJNA310127
Chi_GC_67	Gastric_carcinoma	Female	69	China	SRR3991273	SRP080738	PRJNA310127
Chi_GC_68	Gastric_carcinoma	Female	50	China	SRR3991274	SRP080738	PRJNA310127
Chi_GC_69	Gastric_carcinoma	Female	50	China	SRR3991276	SRP080738	PRJNA310127
Chi_GC_70	Gastric_carcinoma	Female	67	China	SRR3991278	SRP080738	PRJNA310127
Chi_GC_71	Gastric_carcinoma	Male	64	China	SRR3991280	SRP080738	PRJNA310127
Chi_GC_72	Gastric_carcinoma	Male	66	China	SRR3991282	SRP080738	PRJNA310127
Chi_GC_73	Gastric_carcinoma	Male	49	China	SRR3991284	SRP080738	PRJNA310127
Chi_GC_74	Gastric_carcinoma	Male	59	China	SRR3991286	SRP080738	PRJNA310127
Chi_GC_75	Gastric_carcinoma	Female	65	China	SRR3991288	SRP080738	PRJNA310127
Chi_GC_76	Gastric_carcinoma	Male	50	China	SRR3991290	SRP080738	PRJNA310127
Chi_GC_77	Gastric_carcinoma	Male	62	China	SRR3991292	SRP080738	PRJNA310127
Chi_GC_78	Gastric_carcinoma	Male	59	China	SRR3991294	SRP080738	PRJNA310127
Chi_GC_79	Gastric_carcinoma	Male	67	China	SRR3991296	SRP080738	PRJNA310127
Mex_GC_1	Gastric_carcinoma	Female	82	Mexico	SRR3991007	SRP080738	PRJNA310127
Mex_GC_2	Gastric_carcinoma	Male	72	Mexico	SRR3991010	SRP080738	PRJNA310127
Mex_GC_4	Gastric_carcinoma	Female	58	Mexico	SRR3991014	SRP080738	PRJNA310127
Mex_GC_5	Gastric_carcinoma	Female	38	Mexico	SRR3991016	SRP080738	PRJNA310127
Mex_GC_6	Gastric_carcinoma	Female	44	Mexico	SRR3991018	SRP080738	PRJNA310127
Mex_GC_7	Gastric_carcinoma	Female	70	Mexico	SRR3991020	SRP080738	PRJNA310127
Mex_GC_8	Gastric_carcinoma	Male	75	Mexico	SRR3991023	SRP080738	PRJNA310127
Mex_GC_9	Gastric_carcinoma	Female	66	Mexico	SRR3991025	SRP080738	PRJNA310127
Mex_GC_10	Gastric_carcinoma	Male	70	Mexico	SRR3991027	SRP080738	PRJNA310127
Mex_GC_11	Gastric_carcinoma	Female	56	Mexico	SRR3991028	SRP080738	PRJNA310127
Mex_GC_12	Gastric_carcinoma	Male	75	Mexico	SRR3991031	SRP080738	PRJNA310127
Mex_GC_13	Gastric_carcinoma	Female	56	Mexico	SRR3991033	SRP080738	PRJNA310127

Mex_GC_14	Gastric_carcinoma	Female	54	Mexico	SRR3991035	SRP080738	PRJNA310127
Mex_GC_15	Gastric_carcinoma	Male	69	Mexico	SRR3991038	SRP080738	PRJNA310127
Mex_GC_16	Gastric_carcinoma	Female	75	Mexico	SRR3991040	SRP080738	PRJNA310127
Mex_GC_17	Gastric_carcinoma	Female	42	Mexico	SRR3991042	SRP080738	PRJNA310127
Mex_GC_18	Gastric_carcinoma	Male	64	Mexico	SRR3991044	SRP080738	PRJNA310127
Mex_GC_19	Gastric_carcinoma	Female	52	Mexico	SRR3991046	SRP080738	PRJNA310127
Mex_GC_20	Gastric_carcinoma	Male	58	Mexico	SRR3991048	SRP080738	PRJNA310127
Mex_GC_21	Gastric_carcinoma	Female	82	Mexico	SRR3991051	SRP080738	PRJNA310127
Mex_GC_22	Gastric_carcinoma	Female	72	Mexico	SRR3991053	SRP080738	PRJNA310127
Mex_GC_23	Gastric_carcinoma	Male	57	Mexico	SRR3991056	SRP080738	PRJNA310127
Mex_GC_24	Gastric_carcinoma	Female	81	Mexico	SRR3991058	SRP080738	PRJNA310127
Mex_GC_25	Gastric_carcinoma	Male	62	Mexico	SRR3991061	SRP080738	PRJNA310127
Mex_GC_26	Gastric_carcinoma	Female	79	Mexico	SRR3991063	SRP080738	PRJNA310127
Mex_GC_27	Gastric_carcinoma	Male	38	Mexico	SRR3991066	SRP080738	PRJNA310127
Mex_GC_28	Gastric_carcinoma	Female	56	Mexico	SRR3991069	SRP080738	PRJNA310127
Mex_GC_29	Gastric_carcinoma	Female	72	Mexico	SRR3991072	SRP080738	PRJNA310127
Mex_GC_30	Gastric_carcinoma	Male	45	Mexico	SRR3991079	SRP080738	PRJNA310127
Mex_GC_31	Gastric_carcinoma	Male	69	Mexico	SRR3991082	SRP080738	PRJNA310127
Mex_GC_32	Gastric_carcinoma	Female	48	Mexico	SRR3991084	SRP080738	PRJNA310127
Mex_GC_33	Gastric_carcinoma	Male	60	Mexico	SRR3991088	SRP080738	PRJNA310127
Mex_GC_34	Gastric_carcinoma	Male	46	Mexico	SRR3991090	SRP080738	PRJNA310127
Mex_GC_35	Gastric_carcinoma	Male	56	Mexico	SRR3991092	SRP080738	PRJNA310127
Mex_GC_36	Gastric_carcinoma	Male	48	Mexico	SRR3991094	SRP080738	PRJNA310127
Mex_GC_37	Gastric_carcinoma	Male	80	Mexico	SRR3991096	SRP080738	PRJNA310127
Mex_GC_38	Gastric_carcinoma	Male	91	Mexico	SRR3991098	SRP080738	PRJNA310127
Mex_GC_39	Gastric_carcinoma	Female	56	Mexico	SRR3991100	SRP080738	PRJNA310127
Mex_GC_40	Gastric_carcinoma	Female	67	Mexico	SRR3991102	SRP080738	PRJNA310127
Mex_GC_41	Gastric_carcinoma	Male	61	Mexico	SRR3991104	SRP080738	PRJNA310127
Mex_GC_42	Gastric_carcinoma	Male	71	Mexico	SRR3991106	SRP080738	PRJNA310127
Mex_GC_43	Gastric_carcinoma	Female	77	Mexico	SRR3991110	SRP080738	PRJNA310127
Mex_GC_44	Gastric_carcinoma	Female	42	Mexico	SRR3991112	SRP080738	PRJNA310127
Mex_GC_45	Gastric_carcinoma	Female	67	Mexico	SRR3991114	SRP080738	PRJNA310127
Mex_GC_46	Gastric_carcinoma	Female	77	Mexico	SRR3991116	SRP080738	PRJNA310127
Mex_GC_47	Gastric_carcinoma	Male	37	Mexico	SRR3991118	SRP080738	PRJNA310127
Mex_GC_48	Gastric_carcinoma	Female	73	Mexico	SRR3991120	SRP080738	PRJNA310127
Mex_GC_49	Gastric_carcinoma	Female	67	Mexico	SRR3991123	SRP080738	PRJNA310127
Mex_GC_50	Gastric_carcinoma	Male	80	Mexico	SRR3991127	SRP080738	PRJNA310127
Mex_GC_51	Gastric_carcinoma	Male	76	Mexico	SRR3991129	SRP080738	PRJNA310127
Mex_GC_52	Gastric_carcinoma	Female	49	Mexico	SRR3991133	SRP080738	PRJNA310127
Mex_GC_53	Gastric_carcinoma	Male	67	Mexico	SRR3991135	SRP080738	PRJNA310127
Mex_GC_54	Gastric_carcinoma	Male	74	Mexico	SRR3991137	SRP080738	PRJNA310127

Supplementary Table S3. Primers used in qPCR and, unless otherwise stated, designed for this study.

Primer	Sequence (5'-3')	Amplicon size (bp)	Target species
Helicobacter_F	GAAGATAATGACGGTATCTAAC	139	*Helicobacter* sp.
Helicobacter_R	ATTTCACACCTGACTGACTAT		
Neisseria_F	AACGATGTCAATTAGCTGTT	108	*Neisseria* sp.
Neisseria_R	CAATTCCTTTGAGTTTTAATC		
Achromo_F1	TCGGGCCTTGGTAGCG	77	*Achromobacter* sp.
Achromo_R1	TTCCTTTGAGTTTTAATCTT		
Phyllo_F	CTGCCTTTGATACTGGTAGT	202	*Phyllobacterium* sp.
Phyllo_R	CGGCTAGCTCTCATAGTTTA		
Clostr_F*	ATGCAAGTCGAGCGAKG	120	*Clostridium* sp.
Clostr_R*	TATGCGGTATTAATCTKCCTTT		
Rhodo_F	GGGTTCCTTCCACGGGAT	84	*Rhodococcus* sp.
Rhodo_R	CCTTTGAGTTTTAGCCTTG		
Lactob2_F	GAGGCAGCAGTAGGGAATCTTC	126	*Lactobacillus* sp.
Lactob2_R	GGCCAGTTACTACCTCTATCCTTCTTC		
Citro F1	GTAAAGTACTTTCAGCGAG	216	*Citrobacter* sp.
Citro R2	GTTTCGGATGCAGTTCCC		
Prevo_F	CACGGTAAACGATGGATGCC	113	*Prevotella* sp.
Prevo_R	CAATTCCTTTGAGTTTCACC		
Strepto_F	TGTCGTGAGATGTTGGGTTAAG	112	*Streptococcus* sp.
Strepto_R	CCACCTTCCTCCGGTTTATTAC		
340F[§]	TCCTACGGGAGGCAGCAGT	198	Universal
515R[§]	CGTATTACCGCGGCTGCTGGCAC		

*Rinttila T *et al.*, Journal of Applied Microbiology 2004.[8]

[§]Horz HP *et al.*, Journal of Clinical Microbiology 2005.[9]

Supplementary Table S4. Quality control results of 32 gastric samples derived from 16 subjects of the discovery set distributed across amplification and sequencing sets*.

Microbiota measurement	Intraclass correlation coefficient[#] ICC (95% CI)	*P*[§]
Alpha diversity (Shannon index)	0.70 (-0.23 – 0.92)	< 0.0001
Beta diversity[&]		
Unweighted UniFrac	0.67 (-0.05 – 0.89)	0.003
Weighted UniFrac	0.91 (0.64 – 0.97)	< 0.0001
Relative abundance of major phyla		
Proteobacteria	0.69 (0.07 – 0.89)	0.004
Firmicutes	0.62 (-0.05 – 0.86)	0.034
Bacteroidetes	0.66 (0.07 – 0.88)	0.01
Fusobacteria	0.73 (0.27 – 0.91)	0.005

ICC, Intraclass correlation coefficient; CI, confidence intervals.

*The 16S rRNA reads were analyzed using the UPARSE pipeline applying the same quality filtering parameters for each set of sequencing (maximum expected errors of 0.50 and global trimming of 250nt).

[#]Intraclass correlation coefficients were computed assuming the two-way mixed average measures model and using the absolute agreement definition. Higher ICC and lower *P*-value indicate better reproducibility.

[§]*P*-values obtained by analysis of variance (ANOVA) for the proportion of variance estimated for between subjects.

[&]First principal coordinate of weighted UniFrac distance matrix.

Supplementary Table S5. Correlations between the relative abundance of *Helicobacter* spp. and other phyla.

Taxa	Chronic gastritis (ρ* and *P* - value)	Gastric carcinoma (ρ* and *P* - value)
Non-*Helicobacter Proteobacteria*	-0.59 (*P* < 0.0001)	-0.14 (*P* = 0.2640)
Firmicutes	-0.49 (*P* < 0.0001)	0.05 (*P* = 0.6889)
Bacteroidetes	-0.43 (*P* < 0.0001)	0.48 (*P* = 0.0003)
Actinobacteria	-0.54 (*P* < 0.0001)	-0.17 (*P* = 0.2405)
Fusobacteria	-0.15 (*P* = 0.1750)	0.29 (*P* = 0.0352)

*ρ, Spearman's rank coefficient.

Supplementary Table S6. Predicted KEGG Pathways differentially abundant between chronic gastritis and gastric carcinoma.

KEGG_Pathways	Chronic gastritis mean rel. freq. (%)	Gastric carcinoma mean rel. freq. (%)	*P*-values (corrected)*	LDA score	*P*-value (LEfSe)#	Enriched Category
Environmental Information Processing|Membrane Transport	10.92161	15.12360	0.00819	4.33586	1.44E-13	GC
Metabolism|Carbohydrate Metabolism	9.03294	9.83316	0.00178	3.61487	8.98E-05	GC
Genetic Information Processing|Transcription	1.88669	2.63088	0.00273	3.59340	4.64E-16	GC
Metabolism|Xenobiotics Biodegradation and Metabolism	2.50552	3.28664	0.00216	3.56750	1.54E-06	GC
Unclassified|Cellular Processes and Signaling	3.70073	4.38594	0.00186	3.53616	2.97E-09	GC
Unclassified|Poorly Characterized	4.57739	5.11194	0.00241	3.44255	3.58E-11	GC
Unclassified|Metabolism	2.19640	2.66576	0.00410	3.38182	6.97E-12	GC
Environmental Information Processing|Signal Transduction	1.97292	2.39848	0.00315	3.32954	4.67E-07	GC
Metabolism|Amino Acid Metabolism	9.43605	9.81096	0.04352	3.23132	1.19E-02	GC
Metabolism|Lipid Metabolism	3.03535	3.37183	0.00228	3.22117	1.87E-05	GC
Metabolism|Biosynthesis of Other Secondary Metabolites	0.61655	0.75324	0.02048	2.86853	3.31E-08	GC
Metabolism|Enzyme Families	1.79516	1.91412	0.00282	2.86758	3.75E-04	GC
Metabolism|Metabolism of Other Amino Acids	1.77550	1.87891	0.00410	2.77646	7.74E-05	GC
Organismal Systems|Excretory System	0.01130	0.02874	0.01365	2.67809	7.84E-11	GC
Genetic Information Processing|Replication and Repair	8.79776	6.76323	0.00195	4.00196	1.32E-12	CGa
Genetic Information Processing|Translation	5.77603	4.16490	0.00256	3.89954	2.36E-12	CGa
Metabolism|Energy Metabolism	6.03451	5.00229	0.00146	3.72885	1.63E-16	CGa
Unclassified|Genetic Information Processing	3.25102	2.34838	0.00171	3.65858	2.24E-15	CGa
Cellular Processes|Cell Motility	3.79669	2.98665	0.00455	3.64273	1.26E-03	CGa
Metabolism|Glycan Biosynthesis and Metabolism	2.80951	1.96937	0.00341	3.62993	4.66E-14	CGa
Metabolism|Nucleotide Metabolism	3.92975	3.13489	0.04096	3.59115	3.17E-11	CGa
Genetic Information Processing|Folding, Sorting and Degradation	2.82319	2.24511	0.00158	3.47981	1.33E-14	CGa
Metabolism|Metabolism of Cofactors and Vitamins	4.34925	3.85149	0.00293	3.41535	4.34E-12	CGa
Human Diseases|Neurodegenerative Diseases	0.52178	0.32498	0.00585	3.11480	1.09E-10	CGa
Organismal Systems|Circulatory System	0.10186	0.04917	0.00152	2.97205	5.55E-10	CGa
Cellular Processes|Cell Growth and Death	0.59027	0.42469	0.01024	2.96612	8.95E-13	CGa
Human Diseases|Metabolic Diseases	0.07488	0.07169	0.29446	2.91240	4.92E-02	CGa
Human Diseases|Cancers	0.17859	0.13235	0.00205	2.80085	2.31E-08	CGa

Environmental Information Processing|Signaling Molecules and Interaction	0.29797	0.20385	0.00372	2.76556	1.05E-11	CGa
Organismal Systems|Environmental Adaptation	0.17243	0.13526	0.00164	2.70978	3.76E-08	CGa
Human Diseases|Immune System Diseases	0.06094	0.04869	0.00683	2.69694	3.57E-08	CGa
Organismal Systems|Immune System	0.07854	0.05444	0.00512	2.63043	4.56E-07	CGa
Organismal Systems|Endocrine System	0.24091	0.25913	0.29240	-	-	-
Cellular Processes|Cell Communication	0.00000	0.00000	1.02500	-	-	-
Metabolism|Metabolism of Terpenoids and Polyketides	1.77672	1.77557	1.03658	-	-	-
Organismal Systems|Sensory System	0.00000	0.00000	1.00000	-	-	-
Human Diseases|Cardiovascular Diseases	0.00841	0.01131	0.15661	-	-	-
Human Diseases|Infectious Diseases	0.51979	0.49543	0.13653	-	-	-
Cellular Processes|Transport and Catabolism	0.24157	0.23500	0.57989	-	-	-
Organismal Systems|Nervous System	0.07261	0.08275	0.02510	-	-	-
Organismal Systems|Digestive System	0.03089	0.03518	0.18139	-	-	-

*White's non-parametric t-test corrected with Benjaminin-Hochberg FDR
[#]Wilcoxon test

Supplementary Table S7 (A-F). Predicted metagenome of the gastric microbiota using clusters of orthologous groups (COG) in PICRUSt.[4] Predicted COG and KO functions were analysed in STAMP[5] using the two-group comparison with White's non-parametric *t* test[6] and corrected for multiple test with Benjamini-Hochberg false discovery rate. The accuracy of the predicted metagenomes was assessed by determining the NSTI value for each microbiome sample. **Panel A**, contains the selected COG and KO functions predicted with the whole bacterial community; **Panel B**, contains the weighted NSTI values for each microbiome sample predicted (COG functions) with the whole bacterial community; **Panel C**, contains the weighted NSTI values for each microbiome sample predicted (KO functions) with the whole bacterial community; **Panel D**, contains the selected COG and KO functions predicted with 10 genera differentially abundant in the two patient groups (by LEfSe analysis); **Panel E**, contains the weighted NSTI values for each microbiome sample predicted (COG functions) with 10 genera differentially abundant in the two patient groups (by LEfSe analysis); **Panel F**, contains the weighted NSTI values for each microbiome sample predicted (KO functions) with 10 genera differentially abundant in the two patient groups (by LEfSe analysis).

Panel A.

Selected COG functions predicted with the whole bacterial community

#COG	Chronic_gastritis: mean rel. freq. (%)	Chronic_gastritis: std. dev. (%)	Gastric_carcinoma: mean rel. freq. (%)	Gastric_carcinoma: std. dev. (%)	p-values (corrected)	Difference between means	95.0% lower CI	95.0% upper CI
COG0600	0.0380803377962	0.0409023940591	0.144434615098	0.0623153699431	0.00232614809874	-0.106354277302	-0.125221671315	-0.0858149024638
COG0715	0.0511934133291	0.0458494571997	0.157705490347	0.0545866953999	0.00562539692974	-0.106512077018	-0.123411161547	-0.0869708850513
COG1116	0.0367227762667	0.0375718046599	0.134767467249	0.0547153422806	0.0162278399567	-0.0980446909824	-0.113905723223	-0.0821855767254
COG1140	0.0025440844066	0.00668464772303	0.0155970995959	0.0148806082199	0.00415556665557	-0.0130530151893	-0.017275905303	-0.00882205712479
COG1251	0.011611977352	0.0151913533865	0.0519315230767	0.0265861055036	0.0137960022686	-0.0403195457247	-0.0485243035945	-0.0319322400119
COG2116	0.0233649620796	0.0215994256599	0.0383015898969	0.0250015856269	0.00383590768206	-0.0149366278173	-0.0224787062647	-0.00737941715869
COG2146	0.00873193589374	0.0106158185296	0.0393208615594	0.0228568326087	0.00404667184042	-0.0305889256656	-0.0370139696261	-0.0238474856576
COG2180	0.0025440844066	0.00668464772303	0.0155970995959	0.0148806082199	0.00246255801811	-0.0130530151893	-0.0174246808685	-0.0088974980323
COG2181	0.0025440844066	0.00668464772303	0.0155970995959	0.0148806082199	0.00536082059038	-0.0130530151893	-0.0172555501534	-0.00875565238332
COG2223	0.0325291097158	0.0393469296998	0.102362968402	0.0569046880422	0.00538494126796	-0.0698338586861	-0.0868249683248	-0.0530108581266
COG3005	0.00488371320348	0.00661558452324	0.0058839829874	0.00724089219848	0.545794864476	-0.00100026978392	-0.00336821329968	0.00138511644863
COG3043	0.000775554391147	0.00136327433874	0.00373812723423	0.00526742345456	0.00291015974907	-0.00296257284309	-0.00449609407818	-0.001519306695
COG3062	0.000775554391147	0.00136327433874	0.00373812723423	0.00526742345456	0.00240080882007	-0.00296257284309	-0.00452292313383	-0.00156887470169
COG3301	0.00110220332143	0.00305540177519	0.00558613612147	0.00725633566029	0.00396292449273	-0.00448393280004	-0.00645199014736	-0.00259824027501
COG3303	0.00410815881234	0.00672591354877	0.00214585575317	0.0048409549419	0.0601975829892	0.00196230305917	0.0000265626740328	0.00375532751947
COG4459	0.000775554391147	0.00136327433874	0.00373812723423	0.00526742345456	0.00220101737343	-0.00296257284309	-0.00452548718722	-0.00152154912066
COG5013	0.00254435706316	0.00668457771702	0.016518260486	0.0143514761835	0.136777508206	-0.0139739034229	-0.0180142674333	-0.00970549825398

Selected KO functions predicted with the whole bacterial community

#KO	Chronic_gastritis: mean rel. freq. (%)	Chronic_gastritis: std. dev. (%)	Gastric_carcinoma: mean rel. freq. (%)	Gastric_carcinoma: std. dev. (%)	p-values (corrected)	Difference between means	95.0% lower CI	95.0% upper CI
K02575	0.019212388	0.017022566	0.038842644	0.016814508	0.002251173	-0.019630256	-0.025259151	-0.013779889
K00374	0.005812651	0.006131977	0.011933115	0.009218057	0.0037737	-0.006120464	-0.00876988	-0.003022912
K00371	0.003691426	0.004422298	0.011473263	0.009304943	0.003769578	-0.007781837	-0.01046599	-0.005165154
K00370	0.00585703	0.006274888	0.011669446	0.009237582	0.003767521	-0.005812416	-0.00887716	-0.003032478
K00362	0.015170144	0.014939699	0.029964508	0.018805713	0.002091545	-0.014794365	-0.020454654	-0.008977207
K00363	0.013528508	0.013678413	0.021447784	0.013696811	0.002090911	-0.007919277	-0.01298771	-0.003372292

K00368	0.006984151	0.007328579	0.012471795	0.012708158	0.011156408	-0.005487644	-0.009290866	-0.001638056
K03385	0.00473815	0.006007843	0.002046396	0.00567851	0.011201674	0.002691754	0.000679126	0.004591805

Panel B.

Weighted NSTI values for each microbiome sample predicted with the whole bacterial community: COG functions			
GC1	Gastric_carcinoma	Weighted NSTI	0.040489631851
GC2	Gastric_carcinoma	Weighted NSTI	0.0264613580275
GC3	Gastric_carcinoma	Weighted NSTI	0.0389247982662
GC4	Gastric_carcinoma	Weighted NSTI	0.0269396264775
GC5	Gastric_carcinoma	Weighted NSTI	0.0643272483964
GC6	Gastric_carcinoma	Weighted NSTI	0.034023649827
GC7	Gastric_carcinoma	Weighted NSTI	0.033400643341
GC8	Gastric_carcinoma	Weighted NSTI	0.0358000414837
GC9	Gastric_carcinoma	Weighted NSTI	0.093279905086
GC10	Gastric_carcinoma	Weighted NSTI	0.0936946331635
GC11	Gastric_carcinoma	Weighted NSTI	0.0366245943864
GC12	Gastric_carcinoma	Weighted NSTI	0.0287537167868
GC13	Gastric_carcinoma	Weighted NSTI	0.0382200425585
GC14	Gastric_carcinoma	Weighted NSTI	0.0362654260905
GC15	Gastric_carcinoma	Weighted NSTI	0.0472873255435
GC16	Gastric_carcinoma	Weighted NSTI	0.0307910984753
GC17	Gastric_carcinoma	Weighted NSTI	0.0383074646772
GC18	Gastric_carcinoma	Weighted NSTI	0.0147998010024
GC19	Gastric_carcinoma	Weighted NSTI	0.0250345494043
GC20	Gastric_carcinoma	Weighted NSTI	0.0196118356432
GC21	Gastric_carcinoma	Weighted NSTI	0.0554188523192
GC22	Gastric_carcinoma	Weighted NSTI	0.0185227799536
GC23	Gastric_carcinoma	Weighted NSTI	0.119390269582
GC24	Gastric_carcinoma	Weighted NSTI	0.0354898499502
GC25	Gastric_carcinoma	Weighted NSTI	0.0288768762561
GC26	Gastric_carcinoma	Weighted NSTI	0.0200459069109
GC27	Gastric_carcinoma	Weighted NSTI	0.0224176229194
GC28	Gastric_carcinoma	Weighted NSTI	0.0310864469154
GC29	Gastric_carcinoma	Weighted NSTI	0.0584871720662
GC30	Gastric_carcinoma	Weighted NSTI	0.0380997257276
GC31	Gastric_carcinoma	Weighted NSTI	0.0199768652024
GC32	Gastric_carcinoma	Weighted NSTI	0.0326920914222
GC33	Gastric_carcinoma	Weighted NSTI	0.0793141787269
GC34	Gastric_carcinoma	Weighted NSTI	0.0408301561178
GC35	Gastric_carcinoma	Weighted NSTI	0.174612125786
GC36	Gastric_carcinoma	Weighted NSTI	0.0367613957595
GC37	Gastric_carcinoma	Weighted NSTI	0.0346428244292
GC38	Gastric_carcinoma	Weighted NSTI	0.0271150607156
GC39	Gastric_carcinoma	Weighted NSTI	0.0344307993816
GC40	Gastric_carcinoma	Weighted NSTI	0.0666759514741
GC41	Gastric_carcinoma	Weighted NSTI	0.0335315244906
GC42	Gastric_carcinoma	Weighted NSTI	0.030448097136
GC43	Gastric_carcinoma	Weighted NSTI	0.0747453733925
GC44	Gastric_carcinoma	Weighted NSTI	0.0283546264849
GC45	Gastric_carcinoma	Weighted NSTI	0.0428279962603
GC46	Gastric_carcinoma	Weighted NSTI	0.0377074018348
GC47	Gastric_carcinoma	Weighted NSTI	0.0460050016118
GC48	Gastric_carcinoma	Weighted NSTI	0.0517707933249
GC49	Gastric_carcinoma	Weighted NSTI	0.0147181524782
GC50	Gastric_carcinoma	Weighted NSTI	0.0264651673989
GC51	Gastric_carcinoma	Weighted NSTI	0.0310639160311
GC52	Gastric_carcinoma	Weighted NSTI	0.0282792295803
GC53	Gastric_carcinoma	Weighted NSTI	0.0753222820128
GC54	Gastric_carcinoma	Weighted NSTI	0.0394247749183
Ga1	Chronic_gastritis	Weighted NSTI	0.0908111634672
Ga2	Chronic_gastritis	Weighted NSTI	0.0346916585769
Ga3	Chronic_gastritis	Weighted NSTI	0.117402392327
Ga4	Chronic_gastritis	Weighted NSTI	0.121956578849
Ga5	Chronic_gastritis	Weighted NSTI	0.0443341197835
Ga6	Chronic_gastritis	Weighted NSTI	0.0564023880138
Ga7	Chronic_gastritis	Weighted NSTI	0.057121902691
Ga8	Chronic_gastritis	Weighted NSTI	0.0283503088713
Ga9	Chronic_gastritis	Weighted NSTI	0.122618605685
Ga10	Chronic_gastritis	Weighted NSTI	0.137675241471
Ga11	Chronic_gastritis	Weighted NSTI	0.0313511145222
Ga12	Chronic_gastritis	Weighted NSTI	0.0359385545363
Ga13	Chronic_gastritis	Weighted NSTI	0.0267607348016
Ga14	Chronic_gastritis	Weighted NSTI	0.0399913286286
Ga15	Chronic_gastritis	Weighted NSTI	0.0334904226307
Ga16	Chronic_gastritis	Weighted NSTI	0.0555738114619
Ga17	Chronic_gastritis	Weighted NSTI	0.0590495787144

Ga18	Chronic_gastritis	Weighted NSTI	0.0310918907492
Ga19	Chronic_gastritis	Weighted NSTI	0.0381330828939
Ga20	Chronic_gastritis	Weighted NSTI	0.030050458735
Ga21	Chronic_gastritis	Weighted NSTI	0.0293018480457
Ga22	Chronic_gastritis	Weighted NSTI	0.0406953226685
Ga23	Chronic_gastritis	Weighted NSTI	0.0319070615753
Ga24	Chronic_gastritis	Weighted NSTI	0.0305824730562
Ga25	Chronic_gastritis	Weighted NSTI	0.0317728785598
Ga26	Chronic_gastritis	Weighted NSTI	0.0270866195057
Ga27	Chronic_gastritis	Weighted NSTI	0.0313910654489
Ga28	Chronic_gastritis	Weighted NSTI	0.0258875683418
Ga29	Chronic_gastritis	Weighted NSTI	0.0313598318794
Ga30	Chronic_gastritis	Weighted NSTI	0.0711050885032
Ga31	Chronic_gastritis	Weighted NSTI	0.0244993143869
Ga32	Chronic_gastritis	Weighted NSTI	0.024942786618
Ga33	Chronic_gastritis	Weighted NSTI	0.030719788708
Ga34	Chronic_gastritis	Weighted NSTI	0.0261586219593
Ga35	Chronic_gastritis	Weighted NSTI	0.0304169841173
Ga36	Chronic_gastritis	Weighted NSTI	0.0313967257136
Ga37	Chronic_gastritis	Weighted NSTI	0.0256214740952
Ga38	Chronic_gastritis	Weighted NSTI	0.0406025576612
Ga39	Chronic_gastritis	Weighted NSTI	0.0236312029641
Ga40	Chronic_gastritis	Weighted NSTI	0.0265096285674
Ga41	Chronic_gastritis	Weighted NSTI	0.0235599356195
Ga42	Chronic_gastritis	Weighted NSTI	0.0326502167956
Ga43	Chronic_gastritis	Weighted NSTI	0.0316268011843
Ga44	Chronic_gastritis	Weighted NSTI	0.0243789299692
Ga45	Chronic_gastritis	Weighted NSTI	0.0248214547776
Ga46	Chronic_gastritis	Weighted NSTI	0.0263703541274
Ga47	Chronic_gastritis	Weighted NSTI	0.0236490670651
Ga48	Chronic_gastritis	Weighted NSTI	0.0419295596547
Ga49	Chronic_gastritis	Weighted NSTI	0.0453555338905
Ga50	Chronic_gastritis	Weighted NSTI	0.0395148501086
Ga51	Chronic_gastritis	Weighted NSTI	0.0367415424425
Ga52	Chronic_gastritis	Weighted NSTI	0.0434132457438
Ga53	Chronic_gastritis	Weighted NSTI	0.0290510505889
Ga54	Chronic_gastritis	Weighted NSTI	0.0522665861212
Ga55	Chronic_gastritis	Weighted NSTI	0.0404586124022
Ga56	Chronic_gastritis	Weighted NSTI	0.0412643460406
Ga57	Chronic_gastritis	Weighted NSTI	0.0402077309941
Ga58	Chronic_gastritis	Weighted NSTI	0.0253999624615
Ga59	Chronic_gastritis	Weighted NSTI	0.0428094549349
Ga60	Chronic_gastritis	Weighted NSTI	0.0378768132663
Ga61	Chronic_gastritis	Weighted NSTI	0.0503554281182
Ga62	Chronic_gastritis	Weighted NSTI	0.0485291220322
Ga63	Chronic_gastritis	Weighted NSTI	0.0490305079779
Ga64	Chronic_gastritis	Weighted NSTI	0.0934052073583
Ga65	Chronic_gastritis	Weighted NSTI	0.0346860076556
Ga66	Chronic_gastritis	Weighted NSTI	0.0463552549761
Ga67	Chronic_gastritis	Weighted NSTI	0.0302419700077
Ga68	Chronic_gastritis	Weighted NSTI	0.0357172222252
Ga69	Chronic_gastritis	Weighted NSTI	0.0237149281501
Ga70	Chronic_gastritis	Weighted NSTI	0.0448817341401
Ga71	Chronic_gastritis	Weighted NSTI	0.0256301408615
Ga72	Chronic_gastritis	Weighted NSTI	0.0387136886567
Ga73	Chronic_gastritis	Weighted NSTI	0.038742864283
Ga74	Chronic_gastritis	Weighted NSTI	0.031060976654
Ga75	Chronic_gastritis	Weighted NSTI	0.0333810306831
Ga76	Chronic_gastritis	Weighted NSTI	0.0252564503647
Ga77	Chronic_gastritis	Weighted NSTI	0.0296099257201
Ga78	Chronic_gastritis	Weighted NSTI	0.0797687144484
Ga79	Chronic_gastritis	Weighted NSTI	0.055401172199
Ga80	Chronic_gastritis	Weighted NSTI	0.0282654611359
Ga81	Chronic_gastritis	Weighted NSTI	0.060914477945

Panel C.

Weighted NSTI values for each microbiome sample predicted with the whole bacterial community: KO functions			
#Sample	Clinical setting	Metric	Value
GC1	Gastric_carcinoma	Weighted NSTI	0.040489631851
GC2	Gastric_carcinoma	Weighted NSTI	0.0264613580275
GC3	Gastric_carcinoma	Weighted NSTI	0.0389247982662
GC4	Gastric_carcinoma	Weighted NSTI	0.0269396264775
GC5	Gastric_carcinoma	Weighted NSTI	0.0643272483964
GC6	Gastric_carcinoma	Weighted NSTI	0.034023649827
GC7	Gastric_carcinoma	Weighted NSTI	0.033400643341
GC8	Gastric_carcinoma	Weighted NSTI	0.0358000414837
GC9	Gastric_carcinoma	Weighted NSTI	0.093279905086

GC10	Gastric_carcinoma	Weighted NSTI	0.0936946331635
GC11	Gastric_carcinoma	Weighted NSTI	0.0366245943864
GC12	Gastric_carcinoma	Weighted NSTI	0.0287537167868
GC13	Gastric_carcinoma	Weighted NSTI	0.0382200425585
GC14	Gastric_carcinoma	Weighted NSTI	0.0362654260905
GC15	Gastric_carcinoma	Weighted NSTI	0.0472873255435
GC16	Gastric_carcinoma	Weighted NSTI	0.0307910984753
GC17	Gastric_carcinoma	Weighted NSTI	0.0383074646772
GC18	Gastric_carcinoma	Weighted NSTI	0.0147998010024
GC19	Gastric_carcinoma	Weighted NSTI	0.0250345494043
GC20	Gastric_carcinoma	Weighted NSTI	0.0196118356432
GC21	Gastric_carcinoma	Weighted NSTI	0.0554188523192
GC22	Gastric_carcinoma	Weighted NSTI	0.0185227799536
GC23	Gastric_carcinoma	Weighted NSTI	0.119390269582
GC24	Gastric_carcinoma	Weighted NSTI	0.0354898499502
GC25	Gastric_carcinoma	Weighted NSTI	0.0288768762561
GC26	Gastric_carcinoma	Weighted NSTI	0.0200459069109
GC27	Gastric_carcinoma	Weighted NSTI	0.0224176229194
GC28	Gastric_carcinoma	Weighted NSTI	0.0310864469154
GC29	Gastric_carcinoma	Weighted NSTI	0.0584871720662
GC30	Gastric_carcinoma	Weighted NSTI	0.0380997257276
GC31	Gastric_carcinoma	Weighted NSTI	0.0199768652024
GC32	Gastric_carcinoma	Weighted NSTI	0.0326920914222
GC33	Gastric_carcinoma	Weighted NSTI	0.0793141787269
GC34	Gastric_carcinoma	Weighted NSTI	0.0408301561178
GC35	Gastric_carcinoma	Weighted NSTI	0.174612125786
GC36	Gastric_carcinoma	Weighted NSTI	0.0367613957595
GC37	Gastric_carcinoma	Weighted NSTI	0.0346428244292
GC38	Gastric_carcinoma	Weighted NSTI	0.0271150607156
GC39	Gastric_carcinoma	Weighted NSTI	0.0344307993816
GC40	Gastric_carcinoma	Weighted NSTI	0.0666759514741
GC41	Gastric_carcinoma	Weighted NSTI	0.0335315244906
GC42	Gastric_carcinoma	Weighted NSTI	0.030448097136
GC43	Gastric_carcinoma	Weighted NSTI	0.0747453733925
GC44	Gastric_carcinoma	Weighted NSTI	0.0283546264849
GC45	Gastric_carcinoma	Weighted NSTI	0.0428279962603
GC46	Gastric_carcinoma	Weighted NSTI	0.0377074018348
GC47	Gastric_carcinoma	Weighted NSTI	0.0460050016118
GC48	Gastric_carcinoma	Weighted NSTI	0.0517707933249
GC49	Gastric_carcinoma	Weighted NSTI	0.0147181524782
GC50	Gastric_carcinoma	Weighted NSTI	0.0264651673989
GC51	Gastric_carcinoma	Weighted NSTI	0.0310639160311
GC52	Gastric_carcinoma	Weighted NSTI	0.0282792295803
GC53	Gastric_carcinoma	Weighted NSTI	0.0753222820128
GC54	Gastric_carcinoma	Weighted NSTI	0.0394247749183
Ga1	Chronic_gastritis	Weighted NSTI	0.0908111634672
Ga2	Chronic_gastritis	Weighted NSTI	0.0346916585769
Ga3	Chronic_gastritis	Weighted NSTI	0.117402392327
Ga4	Chronic_gastritis	Weighted NSTI	0.121956578849
Ga5	Chronic_gastritis	Weighted NSTI	0.0443341197835
Ga6	Chronic_gastritis	Weighted NSTI	0.0564023880138
Ga7	Chronic_gastritis	Weighted NSTI	0.057121902691
Ga8	Chronic_gastritis	Weighted NSTI	0.0283503088713
Ga9	Chronic_gastritis	Weighted NSTI	0.122618605685
Ga10	Chronic_gastritis	Weighted NSTI	0.137675241471
Ga11	Chronic_gastritis	Weighted NSTI	0.0313511145222
Ga12	Chronic_gastritis	Weighted NSTI	0.0359385545363
Ga13	Chronic_gastritis	Weighted NSTI	0.0267607348016
Ga14	Chronic_gastritis	Weighted NSTI	0.0399913286286
Ga15	Chronic_gastritis	Weighted NSTI	0.0334904226307
Ga16	Chronic_gastritis	Weighted NSTI	0.0555738114619
Ga17	Chronic_gastritis	Weighted NSTI	0.0590495787144
Ga18	Chronic_gastritis	Weighted NSTI	0.0310918907492
Ga19	Chronic_gastritis	Weighted NSTI	0.0381330828939
Ga20	Chronic_gastritis	Weighted NSTI	0.030050458735
Ga21	Chronic_gastritis	Weighted NSTI	0.0293018480457
Ga22	Chronic_gastritis	Weighted NSTI	0.0406953226685
Ga23	Chronic_gastritis	Weighted NSTI	0.0319070615753
Ga24	Chronic_gastritis	Weighted NSTI	0.0305824730562
Ga25	Chronic_gastritis	Weighted NSTI	0.0317728785598
Ga26	Chronic_gastritis	Weighted NSTI	0.0270866195057
Ga27	Chronic_gastritis	Weighted NSTI	0.0313910654489
Ga28	Chronic_gastritis	Weighted NSTI	0.0258875683418
Ga29	Chronic_gastritis	Weighted NSTI	0.0313598318794
Ga30	Chronic_gastritis	Weighted NSTI	0.0711050885032
Ga31	Chronic_gastritis	Weighted NSTI	0.0244993143869
Ga32	Chronic_gastritis	Weighted NSTI	0.024942786618
Ga33	Chronic_gastritis	Weighted NSTI	0.030719788708
Ga34	Chronic_gastritis	Weighted NSTI	0.0261586219593
Ga35	Chronic_gastritis	Weighted NSTI	0.0304169841173
Ga36	Chronic_gastritis	Weighted NSTI	0.0313967257136

Ga37	Chronic_gastritis	Weighted NSTI	0.0256214740952
Ga38	Chronic_gastritis	Weighted NSTI	0.0406025576612
Ga39	Chronic_gastritis	Weighted NSTI	0.0236312029641
Ga40	Chronic_gastritis	Weighted NSTI	0.0265096285674
Ga41	Chronic_gastritis	Weighted NSTI	0.0235599356195
Ga42	Chronic_gastritis	Weighted NSTI	0.0326502167956
Ga43	Chronic_gastritis	Weighted NSTI	0.0316268011843
Ga44	Chronic_gastritis	Weighted NSTI	0.0243789299692
Ga45	Chronic_gastritis	Weighted NSTI	0.0248214547776
Ga46	Chronic_gastritis	Weighted NSTI	0.0263703541274
Ga47	Chronic_gastritis	Weighted NSTI	0.0236490670651
Ga48	Chronic_gastritis	Weighted NSTI	0.0419295596547
Ga49	Chronic_gastritis	Weighted NSTI	0.0453555338905
Ga50	Chronic_gastritis	Weighted NSTI	0.0395148501086
Ga51	Chronic_gastritis	Weighted NSTI	0.0367415424425
Ga52	Chronic_gastritis	Weighted NSTI	0.0434132457438
Ga53	Chronic_gastritis	Weighted NSTI	0.0290510505889
Ga54	Chronic_gastritis	Weighted NSTI	0.0522665861212
Ga55	Chronic_gastritis	Weighted NSTI	0.0404586124022
Ga56	Chronic_gastritis	Weighted NSTI	0.0412643460406
Ga57	Chronic_gastritis	Weighted NSTI	0.0402077309941
Ga58	Chronic_gastritis	Weighted NSTI	0.0253999624615
Ga59	Chronic_gastritis	Weighted NSTI	0.0428094549349
Ga60	Chronic_gastritis	Weighted NSTI	0.0378768132663
Ga61	Chronic_gastritis	Weighted NSTI	0.0503554281182
Ga62	Chronic_gastritis	Weighted NSTI	0.0485291220322
Ga63	Chronic_gastritis	Weighted NSTI	0.0490305079779
Ga64	Chronic_gastritis	Weighted NSTI	0.0934052073583
Ga65	Chronic_gastritis	Weighted NSTI	0.0346860076556
Ga66	Chronic_gastritis	Weighted NSTI	0.0463552549761
Ga67	Chronic_gastritis	Weighted NSTI	0.0302419700077
Ga68	Chronic_gastritis	Weighted NSTI	0.0357172222252
Ga69	Chronic_gastritis	Weighted NSTI	0.0237149281501
Ga70	Chronic_gastritis	Weighted NSTI	0.0448817341401
Ga71	Chronic_gastritis	Weighted NSTI	0.0256301408615
Ga72	Chronic_gastritis	Weighted NSTI	0.0387136886567
Ga73	Chronic_gastritis	Weighted NSTI	0.038742864283
Ga74	Chronic_gastritis	Weighted NSTI	0.031060976654
Ga75	Chronic_gastritis	Weighted NSTI	0.0333810306831
Ga76	Chronic_gastritis	Weighted NSTI	0.0252564503647
Ga77	Chronic_gastritis	Weighted NSTI	0.0296099257201
Ga78	Chronic_gastritis	Weighted NSTI	0.0797687144484
Ga79	Chronic_gastritis	Weighted NSTI	0.055401172199
Ga80	Chronic_gastritis	Weighted NSTI	0.0282654611359
Ga81	Chronic_gastritis	Weighted NSTI	0.060914477945

Panel D.

#COG	Chronic_gastritis mean rel. freq. (%)	Chronic_gastritis std. dev. (%)	Gastric_carcinoma mean rel. freq. (%)	Gastric_carcinoma std. dev. (%)	p-values (corrected)	Difference between means	95.0% lower CI	95.0% upper CI
COG0600	0.0300418645333	0.0460843980869	0.146812141835	0.0737182608186	0.00231043088186	-0.116770277302	-0.138788727797	-0.0945946985164
COG0715	0.0314438736083	0.0378427117516	0.141752334688	0.0701383419265	0.00565196314901	-0.11030846108	-0.130485067169	-0.0895309924737
COG1116	0.0287626358989	0.0418888860315	0.136926410484	0.066415304397	0.0161185615731	-0.108163774585	-0.128837496949	-0.0887058740264
COG1140	0.00341973019684	0.00933044303473	0.0174785528014	0.0151631416482	0.00413046832374	-0.0140588226046	-0.0187097703159	-0.00942096154861
COG1251	0.00588295135724	0.0139110162012	0.0505727873507	0.0317275825981	0.0136387828696	-0.0446898359934	-0.0547299434599	-0.0351107891734
COG2116	0.0260441305425	0.0264772369147	0.0406912022371	0.0264159913953	0.00839272929034	-0.0146470716946	-0.0239204671983	-0.00526594122529
COG2146	0.00216100378478	0.00477317428429	0.0333160685549	0.0270121916546	0.00401949016559	-0.0311550647701	-0.0387130722049	-0.0241121414691
COG2180	0.00341973019684	0.00933044303473	0.0174785528014	0.0151631416482	0.00244995536705	-0.0140588226046	-0.0185814422875	-0.00920331912275
COG2181	0.00341973019684	0.00933044303473	0.0174785528014	0.0151631416482	0.00539100539101	-0.0140588226046	-0.0184239543674	-0.00958313672034
COG2223	0.0304687792229	0.04819766388	0.0987132679002	0.0636281509332	0.00541539908056	-0.0682444886773	-0.0886157069943	-0.0463657519206
COG3005	0.0045568813437	0.00692789916142	0.0076798560699	0.00846065593	0.052708949082	-0.00312297472628	-0.00583622152869	-0.000214930074913
COG3043	0.000004459489807	0.00001909505648	0.004824471628	0.00652515766495	0.00289782856369	-0.00482001213819	-0.00673429348284	-0.00305935378494
COG3062	0.000004459489807	0.00001909505648	0.004824471628	0.00652515766495	0.00238169790409	-0.00482001213819	-0.00659843293099	-0.00315031401884
COG3301	0.00170781910463	0.00466443438518	0.0063267980526	0.00731620351516	0.00393361773805	-0.00461897894799	-0.00686032626101	-0.00248275954591
COG3303	0.0045524218539	0.00692771582096	0.0028553844419	0.00594717363718	0.193659849192	0.00169703741191	-0.000513423867157	0.00393403320397
COG4459	0.000004459489807	0.00001909505648	0.004824471628	0.00652515766495	0.00218694051494	-0.00482001213819	-0.00662053015558	-0.00317972339968
COG5013	0.00342002704444	0.00933034241022	0.0184927461628	0.0144801404877	0.11968031968	-0.0150727191183	-0.0194776299944	-0.010939489594
#KO	Chronic_gastritis mean rel. freq. (%)	Chronic_gastritis std. dev. (%)	Gastric_carcinoma mean rel. freq. (%)	Gastric_carcinoma std. dev. (%)	p-values (corrected)	Difference between means	95.0% lower CI	95.0% upper CI
K02575	0.004509995	0.012372823	0.025966323	0.019687365	0.002297636	-0.021456328	-0.027262001	-0.01542714
K00374	0.002257451	0.006187414	0.01563650	0.012771472	0.004038676	-0.013379049	-0.017113515	-0.009667895
K00371	0.002257451	0.006187414	0.01563650	0.012771472	0.004036315	-0.013379049	-0.016970983	-0.009580145
K00370	0.002257451	0.006187414	0.01563650	0.012771472	0.004033956	-0.013379049	-0.016896677	-0.009447808
K00362	0.004510228	0.012372741	0.025965619	0.019687411	0.002109443	-0.021455391	-0.027549823	-0.015594188
K00363	0.002258658	0.006187010	0.018956808	0.014490694	0.002108799	-0.016698150	-0.020845732	-0.012818462
K00368	0.005625786	0.008819855	0.004356963	0.007688918	0.681206315	0.001268823	-0.001531841	0.004088364
K03385	0.005919455	0.010112995	0.004099644	0.008672612	0.489792805	0.001819811	-0.001240937	0.004759555

Panel E.

Weighted NSTI values for each microbiome sample predicted with 10 genera differentially abundant in the two patient groups (by LEfSe analysis): COG functions			
#Sample	Clinical setting	Metric	Value
GC1	Gastric_carcinoma	Weighted NSTI	0.00930809485847
GC2	Gastric_carcinoma	Weighted NSTI	0.0240044368429
GC3	Gastric_carcinoma	Weighted NSTI	0.0109010814095
GC4	Gastric_carcinoma	Weighted NSTI	0.0145213951541
GC5	Gastric_carcinoma	Weighted NSTI	0.023718870231
GC6	Gastric_carcinoma	Weighted NSTI	0.0293275862309
GC7	Gastric_carcinoma	Weighted NSTI	0.00497999715557
GC8	Gastric_carcinoma	Weighted NSTI	0.023024085928
GC9	Gastric_carcinoma	Weighted NSTI	0.0937727775039
GC10	Gastric_carcinoma	Weighted NSTI	0.0938399144357
GC11	Gastric_carcinoma	Weighted NSTI	0.0231372989755
GC12	Gastric_carcinoma	Weighted NSTI	0.0102467154304
GC13	Gastric_carcinoma	Weighted NSTI	0.0345476705214
GC14	Gastric_carcinoma	Weighted NSTI	0.00921013087128
GC15	Gastric_carcinoma	Weighted NSTI	0.0414846156894
GC16	Gastric_carcinoma	Weighted NSTI	0.0423956274665
GC17	Gastric_carcinoma	Weighted NSTI	0.0154562656542
GC18	Gastric_carcinoma	Weighted NSTI	0.0136105222154
GC19	Gastric_carcinoma	Weighted NSTI	0.0134520120571
GC20	Gastric_carcinoma	Weighted NSTI	0.0139521366017
GC21	Gastric_carcinoma	Weighted NSTI	0.0645339594707
GC22	Gastric_carcinoma	Weighted NSTI	0.0133382328625
GC23	Gastric_carcinoma	Weighted NSTI	0.0354155953905
GC24	Gastric_carcinoma	Weighted NSTI	0.0208627090301
GC25	Gastric_carcinoma	Weighted NSTI	0.0181078888419
GC26	Gastric_carcinoma	Weighted NSTI	0.0144427683349
GC27	Gastric_carcinoma	Weighted NSTI	0.0160166205041
GC28	Gastric_carcinoma	Weighted NSTI	0.0233542930185
GC29	Gastric_carcinoma	Weighted NSTI	0.025739929714
GC30	Gastric_carcinoma	Weighted NSTI	0.0113294205463
GC31	Gastric_carcinoma	Weighted NSTI	0.0148387168242
GC32	Gastric_carcinoma	Weighted NSTI	0.0260766868078
GC33	Gastric_carcinoma	Weighted NSTI	0.0804474660945
GC34	Gastric_carcinoma	Weighted NSTI	0.0157898849295
GC35	Gastric_carcinoma	Weighted NSTI	0.0367311418522
GC36	Gastric_carcinoma	Weighted NSTI	0.0141705107879
GC37	Gastric_carcinoma	Weighted NSTI	0.032360819261
GC38	Gastric_carcinoma	Weighted NSTI	0.011977852744
GC39	Gastric_carcinoma	Weighted NSTI	0.0181153823328
GC40	Gastric_carcinoma	Weighted NSTI	0.0262472709806
GC41	Gastric_carcinoma	Weighted NSTI	0.0164562967724
GC42	Gastric_carcinoma	Weighted NSTI	0.0154662402134
GC43	Gastric_carcinoma	Weighted NSTI	0.0381879853417
GC44	Gastric_carcinoma	Weighted NSTI	0.0160909987093
GC45	Gastric_carcinoma	Weighted NSTI	0.0218091017727
GC46	Gastric_carcinoma	Weighted NSTI	0.0277717197922
GC47	Gastric_carcinoma	Weighted NSTI	0.0368600029383
GC48	Gastric_carcinoma	Weighted NSTI	0.0162168230646
GC49	Gastric_carcinoma	Weighted NSTI	0.0152481813312
GC50	Gastric_carcinoma	Weighted NSTI	0.0220793569132
GC51	Gastric_carcinoma	Weighted NSTI	0.0233230434466
GC52	Gastric_carcinoma	Weighted NSTI	0.0201161357312
GC53	Gastric_carcinoma	Weighted NSTI	0.0202110531638
GC54	Gastric_carcinoma	Weighted NSTI	0.0414352900421
Ga1	Chronic_gastritis	Weighted NSTI	0.0308221809121
Ga2	Chronic_gastritis	Weighted NSTI	0.0190175746876
Ga3	Chronic_gastritis	Weighted NSTI	0.0222307471725
Ga4	Chronic_gastritis	Weighted NSTI	0.0209507992895
Ga5	Chronic_gastritis	Weighted NSTI	0.0319685746606
Ga6	Chronic_gastritis	Weighted NSTI	0.0153921588649
Ga7	Chronic_gastritis	Weighted NSTI	0.0217206904368
Ga8	Chronic_gastritis	Weighted NSTI	0.0232925028465
Ga9	Chronic_gastritis	Weighted NSTI	0.0247355306637
Ga10	Chronic_gastritis	Weighted NSTI	0.0213686407209
Ga11	Chronic_gastritis	Weighted NSTI	0.0233840454215
Ga12	Chronic_gastritis	Weighted NSTI	0.0229725174811
Ga13	Chronic_gastritis	Weighted NSTI	0.0236147972795
Ga14	Chronic_gastritis	Weighted NSTI	0.0188148415636
Ga15	Chronic_gastritis	Weighted NSTI	0.0239628754086
Ga16	Chronic_gastritis	Weighted NSTI	0.0258857082524
Ga17	Chronic_gastritis	Weighted NSTI	0.0266591487112
Ga18	Chronic_gastritis	Weighted NSTI	0.0234975410995
Ga19	Chronic_gastritis	Weighted NSTI	0.0275340254447
Ga20	Chronic_gastritis	Weighted NSTI	0.0265223213943
Ga21	Chronic_gastritis	Weighted NSTI	0.0237405518119
Ga22	Chronic_gastritis	Weighted NSTI	0.0246237461481
Ga23	Chronic_gastritis	Weighted NSTI	0.0261776323705
Ga24	Chronic_gastritis	Weighted NSTI	0.0240216954059
Ga25	Chronic_gastritis	Weighted NSTI	0.0216813543149
Ga26	Chronic_gastritis	Weighted NSTI	0.0245844643959
Ga27	Chronic_gastritis	Weighted NSTI	0.0236851326843
Ga28	Chronic_gastritis	Weighted NSTI	0.0235497859161
Ga29	Chronic_gastritis	Weighted NSTI	0.0276284179621
Ga30	Chronic_gastritis	Weighted NSTI	0.0324789141807
Ga31	Chronic_gastritis	Weighted NSTI	0.0235609158646
Ga32	Chronic_gastritis	Weighted NSTI	0.0234512561062
Ga33	Chronic_gastritis	Weighted NSTI	0.0258031842255
Ga34	Chronic_gastritis	Weighted NSTI	0.0239299087937
Ga35	Chronic_gastritis	Weighted NSTI	0.0276010614863
Ga36	Chronic_gastritis	Weighted NSTI	0.0250065251942
Ga37	Chronic_gastritis	Weighted NSTI	0.0233063614292
Ga38	Chronic_gastritis	Weighted NSTI	0.0321749195517

Ga39	Chronic_gastritis	Weighted NSTI	0.0233674455307
Ga40	Chronic_gastritis	Weighted NSTI	0.0239810959362
Ga41	Chronic_gastritis	Weighted NSTI	0.0231499448874
Ga42	Chronic_gastritis	Weighted NSTI	0.0232485466453
Ga43	Chronic_gastritis	Weighted NSTI	0.0265209393083
Ga44	Chronic_gastritis	Weighted NSTI	0.0235484717435
Ga45	Chronic_gastritis	Weighted NSTI	0.0236658097207
Ga46	Chronic_gastritis	Weighted NSTI	0.0234256131336
Ga47	Chronic_gastritis	Weighted NSTI	0.023129553295
Ga48	Chronic_gastritis	Weighted NSTI	0.0248635470616
Ga49	Chronic_gastritis	Weighted NSTI	0.0238738216243
Ga50	Chronic_gastritis	Weighted NSTI	0.0234692827327
Ga51	Chronic_gastritis	Weighted NSTI	0.0284928431634
Ga52	Chronic_gastritis	Weighted NSTI	0.0232210470076
Ga53	Chronic_gastritis	Weighted NSTI	0.0241237951842
Ga54	Chronic_gastritis	Weighted NSTI	0.0249728754939
Ga55	Chronic_gastritis	Weighted NSTI	0.0245525155737
Ga56	Chronic_gastritis	Weighted NSTI	0.0247968261668
Ga57	Chronic_gastritis	Weighted NSTI	0.0220259722126
Ga58	Chronic_gastritis	Weighted NSTI	0.0235104559057
Ga59	Chronic_gastritis	Weighted NSTI	0.0340624095354
Ga60	Chronic_gastritis	Weighted NSTI	0.0248388022169
Ga61	Chronic_gastritis	Weighted NSTI	0.0371165514265
Ga62	Chronic_gastritis	Weighted NSTI	0.0315506630709
Ga63	Chronic_gastritis	Weighted NSTI	0.0190374080374
Ga64	Chronic_gastritis	Weighted NSTI	0.0255534901151
Ga65	Chronic_gastritis	Weighted NSTI	0.0283541207743
Ga66	Chronic_gastritis	Weighted NSTI	0.0243016108554
Ga67	Chronic_gastritis	Weighted NSTI	0.0251248866666
Ga68	Chronic_gastritis	Weighted NSTI	0.0329422136093
Ga69	Chronic_gastritis	Weighted NSTI	0.0234692839506
Ga70	Chronic_gastritis	Weighted NSTI	0.031955855137
Ga71	Chronic_gastritis	Weighted NSTI	0.0231236853239
Ga72	Chronic_gastritis	Weighted NSTI	0.0316024332108
Ga73	Chronic_gastritis	Weighted NSTI	0.0290991240903
Ga74	Chronic_gastritis	Weighted NSTI	0.0228351266279
Ga75	Chronic_gastritis	Weighted NSTI	0.0244748764283
Ga76	Chronic_gastritis	Weighted NSTI	0.0233080136803
Ga77	Chronic_gastritis	Weighted NSTI	0.0232002651225
Ga78	Chronic_gastritis	Weighted NSTI	0.0236748666552
Ga79	Chronic_gastritis	Weighted NSTI	0.0408731774542
Ga80	Chronic_gastritis	Weighted NSTI	0.0233605857388
Ga81	Chronic_gastritis	Weighted NSTI	0.0286796217287

Panel F.

Weighted NSTI values for each microbiome sample predicted with 10 genera differentially abundant in the two patient groups (by LEfSe analysis): KO functions

#Sample	Clinical setting	Metric	Value
GC1	Gastric_carcinoma	Weighted NSTI	0.00930809485847
GC2	Gastric_carcinoma	Weighted NSTI	0.0240044368429
GC3	Gastric_carcinoma	Weighted NSTI	0.0109010814095
GC4	Gastric_carcinoma	Weighted NSTI	0.0145213951541
GC5	Gastric_carcinoma	Weighted NSTI	0.023718870231
GC6	Gastric_carcinoma	Weighted NSTI	0.0293275862309
GC7	Gastric_carcinoma	Weighted NSTI	0.00497999715557
GC8	Gastric_carcinoma	Weighted NSTI	0.023024085928
GC9	Gastric_carcinoma	Weighted NSTI	0.0937727775039
GC10	Gastric_carcinoma	Weighted NSTI	0.0938399144357
GC11	Gastric_carcinoma	Weighted NSTI	0.0231372989755
GC12	Gastric_carcinoma	Weighted NSTI	0.0102467154304
GC13	Gastric_carcinoma	Weighted NSTI	0.0345476705214
GC14	Gastric_carcinoma	Weighted NSTI	0.00921013087128
GC15	Gastric_carcinoma	Weighted NSTI	0.0414846156894
GC16	Gastric_carcinoma	Weighted NSTI	0.0423956274665
GC17	Gastric_carcinoma	Weighted NSTI	0.0154562656542
GC18	Gastric_carcinoma	Weighted NSTI	0.0136105222154
GC19	Gastric_carcinoma	Weighted NSTI	0.0134520120571
GC20	Gastric_carcinoma	Weighted NSTI	0.0139521366017
GC21	Gastric_carcinoma	Weighted NSTI	0.0645339594707
GC22	Gastric_carcinoma	Weighted NSTI	0.0133382328625
GC23	Gastric_carcinoma	Weighted NSTI	0.0354155953905
GC24	Gastric_carcinoma	Weighted NSTI	0.0208627090301
GC25	Gastric_carcinoma	Weighted NSTI	0.0181078888419
GC26	Gastric_carcinoma	Weighted NSTI	0.0144427683349
GC27	Gastric_carcinoma	Weighted NSTI	0.0160166205041
GC28	Gastric_carcinoma	Weighted NSTI	0.0233542930185
GC29	Gastric_carcinoma	Weighted NSTI	0.025739929714
GC30	Gastric_carcinoma	Weighted NSTI	0.0113294205463
GC31	Gastric_carcinoma	Weighted NSTI	0.0148387168242
GC32	Gastric_carcinoma	Weighted NSTI	0.0260766868078
GC33	Gastric_carcinoma	Weighted NSTI	0.0804474660945
GC34	Gastric_carcinoma	Weighted NSTI	0.0157898849295
GC35	Gastric_carcinoma	Weighted NSTI	0.0367311418522
GC36	Gastric_carcinoma	Weighted NSTI	0.0141705107879
GC37	Gastric_carcinoma	Weighted NSTI	0.032360819261

GC38	Gastric_carcinoma	Weighted NSTI	0.011977852744
GC39	Gastric_carcinoma	Weighted NSTI	0.0181153823328
GC40	Gastric_carcinoma	Weighted NSTI	0.0262472709806
GC41	Gastric_carcinoma	Weighted NSTI	0.0164562967724
GC42	Gastric_carcinoma	Weighted NSTI	0.0154662402134
GC43	Gastric_carcinoma	Weighted NSTI	0.0381879853417
GC44	Gastric_carcinoma	Weighted NSTI	0.0160909987093
GC45	Gastric_carcinoma	Weighted NSTI	0.0218091017727
GC46	Gastric_carcinoma	Weighted NSTI	0.0277717197922
GC47	Gastric_carcinoma	Weighted NSTI	0.0368600029383
GC48	Gastric_carcinoma	Weighted NSTI	0.0162168230646
GC49	Gastric_carcinoma	Weighted NSTI	0.0152481813312
GC50	Gastric_carcinoma	Weighted NSTI	0.0220793569132
GC51	Gastric_carcinoma	Weighted NSTI	0.0233230434466
GC52	Gastric_carcinoma	Weighted NSTI	0.0201161357312
GC53	Gastric_carcinoma	Weighted NSTI	0.0202110531638
GC54	Gastric_carcinoma	Weighted NSTI	0.0414352900421
Ga1	Chronic_gastritis	Weighted NSTI	0.0308221809121
Ga2	Chronic_gastritis	Weighted NSTI	0.0190175746876
Ga3	Chronic_gastritis	Weighted NSTI	0.0222307471725
Ga4	Chronic_gastritis	Weighted NSTI	0.0209507992895
Ga5	Chronic_gastritis	Weighted NSTI	0.0319685746606
Ga6	Chronic_gastritis	Weighted NSTI	0.0153921588649
Ga7	Chronic_gastritis	Weighted NSTI	0.0217206904368
Ga8	Chronic_gastritis	Weighted NSTI	0.0232925028465
Ga9	Chronic_gastritis	Weighted NSTI	0.0247355306637
Ga10	Chronic_gastritis	Weighted NSTI	0.0213686407209
Ga11	Chronic_gastritis	Weighted NSTI	0.0233840454215
Ga12	Chronic_gastritis	Weighted NSTI	0.0229725174811
Ga13	Chronic_gastritis	Weighted NSTI	0.0236147972795
Ga14	Chronic_gastritis	Weighted NSTI	0.0188148415636
Ga15	Chronic_gastritis	Weighted NSTI	0.0239628754086
Ga16	Chronic_gastritis	Weighted NSTI	0.0258857082524
Ga17	Chronic_gastritis	Weighted NSTI	0.0266591487112
Ga18	Chronic_gastritis	Weighted NSTI	0.0234975410995
Ga19	Chronic_gastritis	Weighted NSTI	0.0275340254447
Ga20	Chronic_gastritis	Weighted NSTI	0.0265223213943
Ga21	Chronic_gastritis	Weighted NSTI	0.0237405518119
Ga22	Chronic_gastritis	Weighted NSTI	0.0246237461481
Ga23	Chronic_gastritis	Weighted NSTI	0.0261776323705
Ga24	Chronic_gastritis	Weighted NSTI	0.0240216954059
Ga25	Chronic_gastritis	Weighted NSTI	0.0216813543149
Ga26	Chronic_gastritis	Weighted NSTI	0.0245844643959
Ga27	Chronic_gastritis	Weighted NSTI	0.0236851326843
Ga28	Chronic_gastritis	Weighted NSTI	0.0235497859161
Ga29	Chronic_gastritis	Weighted NSTI	0.0276284179621
Ga30	Chronic_gastritis	Weighted NSTI	0.0324789141807
Ga31	Chronic_gastritis	Weighted NSTI	0.0235609158646
Ga32	Chronic_gastritis	Weighted NSTI	0.0234512561062
Ga33	Chronic_gastritis	Weighted NSTI	0.0258031842255
Ga34	Chronic_gastritis	Weighted NSTI	0.0239299087937
Ga35	Chronic_gastritis	Weighted NSTI	0.0276010614863
Ga36	Chronic_gastritis	Weighted NSTI	0.0250065251942
Ga37	Chronic_gastritis	Weighted NSTI	0.0233063614292
Ga38	Chronic_gastritis	Weighted NSTI	0.0321749195517
Ga39	Chronic_gastritis	Weighted NSTI	0.0233674455307
Ga40	Chronic_gastritis	Weighted NSTI	0.0239810959362
Ga41	Chronic_gastritis	Weighted NSTI	0.0231499448874
Ga42	Chronic_gastritis	Weighted NSTI	0.0232485466453
Ga43	Chronic_gastritis	Weighted NSTI	0.0265209393083
Ga44	Chronic_gastritis	Weighted NSTI	0.0235484717435
Ga45	Chronic_gastritis	Weighted NSTI	0.0236658097207
Ga46	Chronic_gastritis	Weighted NSTI	0.0234256131336
Ga47	Chronic_gastritis	Weighted NSTI	0.023129553295
Ga48	Chronic_gastritis	Weighted NSTI	0.0248635470616
Ga49	Chronic_gastritis	Weighted NSTI	0.0238738216243
Ga50	Chronic_gastritis	Weighted NSTI	0.0234692827327
Ga51	Chronic_gastritis	Weighted NSTI	0.0284928431634
Ga52	Chronic_gastritis	Weighted NSTI	0.0232210470076
Ga53	Chronic_gastritis	Weighted NSTI	0.0241237951842
Ga54	Chronic_gastritis	Weighted NSTI	0.0249728754939
Ga55	Chronic_gastritis	Weighted NSTI	0.0245525155737
Ga56	Chronic_gastritis	Weighted NSTI	0.0247968261668
Ga57	Chronic_gastritis	Weighted NSTI	0.0220259722126
Ga58	Chronic_gastritis	Weighted NSTI	0.0235104559057
Ga59	Chronic_gastritis	Weighted NSTI	0.0340624095354
Ga60	Chronic_gastritis	Weighted NSTI	0.0248388022169
Ga61	Chronic_gastritis	Weighted NSTI	0.0371165514265
Ga62	Chronic_gastritis	Weighted NSTI	0.0315506630709
Ga63	Chronic_gastritis	Weighted NSTI	0.0190374080374
Ga64	Chronic_gastritis	Weighted NSTI	0.0255534901151

Ga65	Chronic_gastritis	Weighted NSTI	0.0283541207743
Ga66	Chronic_gastritis	Weighted NSTI	0.0243016108554
Ga67	Chronic_gastritis	Weighted NSTI	0.0251248866666
Ga68	Chronic_gastritis	Weighted NSTI	0.0329422136093
Ga69	Chronic_gastritis	Weighted NSTI	0.0234692839506
Ga70	Chronic_gastritis	Weighted NSTI	0.031955855137
Ga71	Chronic_gastritis	Weighted NSTI	0.0231236853239
Ga72	Chronic_gastritis	Weighted NSTI	0.0316024332108
Ga73	Chronic_gastritis	Weighted NSTI	0.0290991240903
Ga74	Chronic_gastritis	Weighted NSTI	0.0228351266279
Ga75	Chronic_gastritis	Weighted NSTI	0.0244748764283
Ga76	Chronic_gastritis	Weighted NSTI	0.0233080136803
Ga77	Chronic_gastritis	Weighted NSTI	0.0232002651225
Ga78	Chronic_gastritis	Weighted NSTI	0.0236748666552
Ga79	Chronic_gastritis	Weighted NSTI	0.0408731774542
Ga80	Chronic_gastritis	Weighted NSTI	0.0233605857388
Ga81	Chronic_gastritis	Weighted NSTI	0.0286796217287

SUPPLEMENTARY REFERENCES

1 van Doorn LJ, Figueiredo C, Rossau R, et al. Typing of Helicobacter pylori vacA gene and detection of cagA gene by PCR and reverse hybridization. *J Clin Microbiol* 1998;36:1271-6.
2 DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied & Environmental Microbiology* 2006;72:5069-72.
3 Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;15:382-92.
4 Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;31:814-21.
5 Parks DH, Tyson GW, Hugenholtz P, et al. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 2014;30:3123-4.
6 White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009;5:e1000352.
7 Walters WA, Caporaso JG, Lauber CL, et al. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* 2011;27:1159-61.
8 Rinttila T, Kassinen A, Malinen E, et al. Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. *J Appl Microbiol* 2004;97:1166-77.
9 Horz HP, Vianna ME, Gomes BP, et al. Evaluation of universal probes and primer sets for assessing total bacterial load in clinical samples: general implications and practical use in endodontic antimicrobial therapy. *J Clin Microbiol* 2005;43:5332-7.

Paper II

Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis

Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn LJ, Knetsch CS, Figueiredo C

Frontiers in Microbiology. 2019.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 12 June 2019
doi: 10.3389/fmicb.2019.01277

Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis

Joana Pereira-Marques[1,2,3], Anne Hout[4], Rui M. Ferreira[1,2], Michiel Weber[4], Ines Pinto-Ribeiro[1,2,5], Leen-Jan van Doorn[4], Cornelis Willem Knetsch[4] and Ceu Figueiredo[1,2,5]**

[1] i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal, [2] Ipatimup – Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal, [3] Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Porto, Portugal, [4] DDL Diagnostic Laboratory, Rijswijk, Netherlands, [5] Departmento de Patologia, Faculdade de Medicina da Universidade do Porto, Porto, Portugal

OPEN ACCESS

Edited by:
Angel Angelov,
Tübingen University Hospital,
Germany

Reviewed by:
Tatiana A. Vishnivetskaya,
The University of Tennessee,
Knoxville, United States
Jonathan Badger,
National Cancer Institute (NCI),
United States

***Correspondence:**
Cornelis Willem Knetsch
Wilco.Knetsch@ddl.nl
Ceu Figueiredo
cfigueiredo@ipatimup.pt

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: 14 February 2019
Accepted: 22 May 2019
Published: 12 June 2019

Citation:
Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn L-J, Knetsch CW and Figueiredo C (2019) Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. Front. Microbiol. 10:1277. doi: 10.3389/fmicb.2019.01277

The amount of host DNA poses a major challenge to metagenome analysis. However, there is no guidance on the levels of host DNA, nor on the depth of sequencing needed to acquire meaningful information from whole metagenome sequencing (WMS). Here, we evaluated the impact of a wide range of amounts of host DNA and sequencing depths on microbiome taxonomic profiling using WMS. Synthetic samples with increasing levels of host DNA were created by spiking DNA of a mock bacterial community, with DNA from a mouse-derived cell line. Taxonomic analysis revealed that increasing proportions of host DNA led to decreased sensitivity in detecting very low and low abundant species. Reduction of sequencing depth had major impact on the sensitivity of WMS for profiling samples with 90% host DNA, increasing the number of undetected species. Finally, analysis of simulated datasets with fixed depth of 10 million reads confirmed that microbiome profiling becomes more inaccurate as the level of host DNA increases in a sample. In conclusion, samples with high amounts of host DNA coupled with reduced sequencing depths, decrease WMS coverage for characterization of the microbiome. This study highlights the importance of carefully considering these aspects in the design of WMS experiments to maximize microbiome analyses.

Keywords: microbiome analysis, metagenomics, sample complexity, sequencing depth, mock community

INTRODUCTION

The collection of microorganisms present in a defined environment is known as the microbiota (Marchesi and Ravel, 2015). These microorganisms, which comprise bacteria, archaea, viruses, and microbial eukaryotes, along with their genetic information and specific characteristics of the niche they occupy are generally known as the microbiome (Marchesi and Ravel, 2015).

Targeted 16S rRNA gene and whole metagenome sequencing (WMS) are sequencing-based approaches that are currently used to explore the composition and functions of the microbiome (Human Microbiome Jumpstart Reference Strains Consortium, et al., 2010;

Human Microbiome Project Consortium, 2012a). WMS consists on untargeted DNA sequencing of fragments of all genomes within a sample, which produces high-complexity datasets with millions of short reads (Quince et al., 2017). Since all DNA present in the sample is captured, including bacterial, viral, and eukaryotic DNA, this method allows extensive characterization of the microbial communities living in a wide range of environments (e.g., soil, human-associated samples, among others) (Venter et al., 2004; Qin et al., 2010). In comparison with targeted 16S rRNA sequencing, WMS typically yields a more detailed taxonomic resolution, at the species or even strain-level. It also provides a more accurate insight into the functional composition of the microbiome (Qin et al., 2010; Abubucker et al., 2012; Truong et al., 2015; Truong et al., 2017). Still, this approach has been less implemented since it is more expensive than 16S rRNA profiling, it requires a greater depth of coverage, and the data analysis is more complex (Knight et al., 2012).

It is currently well recognized that the microbiome plays an important role in human physiology and in the maintenance of health, but also has a major impact in the development of a wide range of diseases, including obesity, inflammatory bowel disease, and cancer (Ley et al., 2006; Frank et al., 2007; Kostic et al., 2013; Gilbert et al., 2018). A major technical challenge in whole metagenome analysis of human samples is the predominance of host DNA. Data from the Human Microbiome Project (HMP) has revealed that the proportion of human DNA differs significantly by body site and sample type (Human Microbiome Project Consortium, 2012b; Lloyd-Price et al., 2017). While stool samples comprise less than 10% of human DNA, samples such as saliva, throat, buccal mucosa, and vaginal swabs contain more than 90% of human-aligned reads (Human Microbiome Project Consortium, 2012b; Lloyd-Price et al., 2017). The latter type of samples, where only a limited fraction of the DNA represents the microbial content, requires a high quantity of sequences to obtain a reasonable coverage of the microbial genomes when using WMS. Currently, very little is known about the impact of this technical limitation on the sensitivity of WMS to profile the microbiome of host-derived samples. In addition, there is no guidance for the reasonable amount of host DNA a sample should contain in order to generate an accurate WMS analysis. Overcoming these issues is crucial for future selection of appropriate sequencing depths that will guarantee the return of the maximum useful information, with a minimum cost possible. Therefore, this study aimed to evaluate the sensitivity of WMS for taxonomic profiling of microbiome samples, taking into account the wide range of host DNA in a sample and sequencing depths.

MATERIALS AND METHODS

Mock Microbial Community

Genomic DNA from Microbial Mock Community B (Staggered, High Concentration), v5.2H, for Whole Genome Shotgun Sequencing, HM-277D, was obtained through BEI Resources, NIAID, NIH as part of the HMP. This mock microbial community is composed of a combination of 20 bacterial genomic DNAs that differ in %GC content (30 to 69%), and contains staggered ribosomal RNA operon counts differing by bacteria, ranging from 10^4 to 10^7 copies per organism per μL (as indicated by the manufacturer). The genomic GC content of each species was obtained from the NCBI Genome Database. To estimate the expected relative abundance of species, the theoretical number of genome copies per species was calculated by the ratio of input 16S rRNA copies to 16S rRNA copies per genome, and normalized by the sum of all theoretical genome copies of the species present in the mock (sum up to 100). The detailed composition of the mock community, including %GC content, the number of 16S rRNA copies per genome, the number of 16S rRNA input copies, the number of species genome copies, and the expected relative abundance of species, is available in the **Supplementary Table S1**.

Mouse Cell Line and DNA Isolation

Total genomic DNA was extracted from the MC-38 cell line (a kind gift from Professor J. Machado, University of Porto), which is derived from C57BL/6 murine colon adenocarcinoma cells, with the QIAamp DNA Tissue kit (Qiagen, Germany), according to the manufacturer's instructions. DNA was eluted in 100 μl Microbial-DNA free water (Qiagen).

Generation of Synthetic Samples

To create different synthetic samples (SS) with well-defined ratios of host to bacterial DNA, the mock microbial community DNA was spiked with DNA from the mouse cell line. DNA concentrations of the mock microbial community and of the mouse cell line were measured using the NanoDrop 2000 UV spectrophotometer (Thermo Fisher Scientific), and the exact volumes to be mixed in each condition were determined. SS with increasing ratios of host to bacterial DNA were generated containing 10% (SS10), 90% (SS90), and 99% (SS99) host DNA. The mock microbial community sample (MS), which contains only microbial DNA, was used as control.

Library Preparation and Whole Metagenome Sequencing

Samples were first quantified and normalized to 0.2 ng/μl DNA material, using a Quant-It PicoGreen dsDNA assay (Thermo Fisher Scientific), in order to use 1 ng input DNA for the library construction. Metagenomic library preparation was automated on the Hamilton Microlab STAR Liquid Handling Workstation, using a Nextera XT DNA library preparation kit (Illumina Inc., CA, United States) according to the manufacturer's protocol. Briefly, after normalized samples were fragmented and tagged by tagmentation, a limited-cycle PCR was performed to add the Index 1 (i7), Index 2 (i5) and full adapter sequences required for cluster generation. Amplification was followed by a cleanup step that purified the library DNA and removed small library fragments by using Agencourt AMPure XP beads (Beckman Coulter, Inc.). The quality of the library was assessed using an Agilent Technology 2100 Bioanalyzer (Agilent Technologies, Wokingham, United Kingdom) and then, a bead-based normalization was performed using beads Nextera XT to

ensure more equal library representation in the pooled library. Finally, the pooled library was sequenced as a paired-end 150-cycle run on the Illumina NextSeq 550 platform, at an expected sequencing depth of 5.5 Gb/sample.

Sequencing Data Analysis

For each sample, the two FASTQ files with the forward and reverse paired-end reads were concatenated into one single FASTQ file, which was used as input for our in-house pipeline of WMS sequencing data analysis.

Sequencing Data Pre-processing

Sequencing data pre-processing was performed by KneadData (version 0.6.1), a computational tool designed to perform quality control on metagenomic sequencing data. KneadData integrates the tools FastQC (version 0.11.5) (Andrews, 2016), Trimmomatic (version 0.33) (Bolger et al., 2014), and Bowtie2 (version 2.2) (Langmead and Salzberg, 2012), to do quality check, quality filtering, and host sequences decontamination, respectively.

First, reads were trimmed based on a sliding window trimming approach, cutting once the average base Phred quality score within a four-base sliding window dropped below 20, and then were discarded when the length of the read was shorter than 60 bp. After the quality-filtering step, KneadData used Bowtie2 to identify and remove the mouse contaminant reads present in the datasets, by mapping the reads against the C57BL/6 reference genome (GCA_001632555.1 assembly). The non-mouse filtered reads were then used for the downstream analysis. Bowtie2 was used with the default parameters (–very-sensitive end-to-end alignment). FastQC, as a component of KneadData, performed quality control checks on raw whole metagenome sequencing data but also on reads after sequencing data pre-processing, in order to assess the efficiency of the quality filtering and of host sequences decontamination steps in the generation of high-quality reads. FastQC was used with the default parameters (Andrews, 2016).

Taxonomic Profiling – MetaPhlAn2

In our in-house pipeline of analysis, the host-filtered microbial reads were taxonomically profiled using MetaPhlAn2 (version 2.7.5), an assembly free taxonomic profiler (Segata et al., 2012; Truong et al., 2015). This computational tool mapped the quality-controlled shotgun reads to a database of unique clade-specific marker genes (read-based profiling) with high discriminatory power, estimating the relative abundances of each microbial clade in the samples with species-level resolution (Segata et al., 2012; Truong et al., 2015). Bowtie2, a fast DNA aligner, is used by MetaPhlAn2 to map the metagenomic reads against the unique clade-specific marker genes. Clade-specific markers constitute coding sequences that unambiguously identify specific microbial clades (at different taxonomic levels). MetaPhlAn2 relies on ∼1 million unique clade-specific marker genes identified from ∼17,000 reference genomes and >7,000 unique species. Markers are now identified not only for Bacteria and *Archaea* (∼13,500 bacterial and archaeal genomes), but also for Viruses (∼3,500 viral genomes) and Eukaryotic microorganisms (Fungi and Protozoa; ∼110 eukaryotic genome (Truong et al., 2015).

Generation of Datasets With Reduced Sequencing Depths

The sample with the largest sequence dataset (SS90) comprising 50.8 million single-end reads and a high predominance of host DNA was used. Four datasets with reduced sequencing depths were generated by random subsampling paired-end reads using an in-house script. From the original SS90 dataset, we subsampled 50, 25, 10, and 5%, which correspond to 25.4, 12.7, 5.1, and 2.5 million single-end reads, respectively. For subsampling, the same random seed was used in order to guarantee that the reads from the same pair were subsampled in the forward and reverse FASTQs. Then, it was created a new set of paired FASTQ file for each random subset. At each depth, the subsampling analysis was repeated five times.

Generation of Simulated Datasets of Microbiome Samples With Different Host-Microbial Ratios

Simulated datasets (SD) of microbiome samples with different host: microbial ratios were created by randomly selecting host and microbial reads from our previously sequenced datasets, and combining them in different proportions at a fixed sequencing depth of 10 million single-end reads, using an in-house script. Microbial single-end reads were randomly picked from the MS raw dataset, to assure that only microbial reads were selected. Host single-end reads were randomly picked from the mouse contaminant sequences removed by KneadData from the SS99 raw dataset, to guarantee sufficient sequences with host origin (the raw SS99 dataset contained 33.201.587 mouse single-end reads). Eighteen SD were generated, nine with progressive 10% increases in host reads (SD10 to SD90) and nine with progressive 1% increases in host reads (SD91 to SD99). For each simulated dataset, five replicates were randomly generated.

Statistical Analyses

Statistical treatment was performed using the GraphPad Prism software (v. 6.01, GraphPad Software Inc., La Jolla, CA, United States). Pearson's correlation was used to assess correlations between the species genomic %GC content and the ratio between observed and expected relative abundances in the MS control. A ≥2-fold difference was selected as arbitrary threshold to consider species as underestimated or overestimated in comparison with a reference condition. Differences between groups, when performing the random subsampling analysis and the simulated dataset analysis, were evaluated using the Kruskal–Wallis non-parametric test, followed by multiple comparisons versus a control group using the Dunn's test. The differences were considered statistically significant with P values lower than 0.05.

RESULTS

Generation of Synthetic Samples and Pre-processing of Sequencing Data

To assess the influence of host DNA on the sensitivity of WMS for taxonomic profiling of the microbiome, three synthetic samples with distinct host: bacteria DNA ratios were generated to contain 10, 90, and 99% host DNA (SS10, SS90, and SS99, respectively). As control, the mock microbial community DNA sample (100% bacterial DNA; MS) was used (**Figure 1A**).

The four datasets yielded a large number of raw single-end reads ranging from 35 to 51 million. After sequencing data pre-processing (quality filtering and host sequences decontamination), the number of reads differed considerably between samples, being higher in MS (33.3 million reads) and SS10 (29.9 million reads) in comparison with SS90 (5.5 million reads) and SS99 (7.5 hundred thousand reads). The relative low number of pre-processed reads in SS90 and SS99 samples was due to high number of host DNA sequences removed rather than to reads dropped during quality filtering (**Supplementary Table S2**).

Of the total raw single-end reads, the proportion of discarded reads during quality filtering was similar between all samples (ranging from 16 to 19%), confirming that differences in the number of pre-processed reads across samples were associated with the host sequences decontamination step (**Supplementary Figure S1A**). These results are consistent with the overall good quality of all raw datasets (between 80 and 90% of the reads with average quality ≥Q30). They also indicate that quality filtering was appropriate, resulting in datasets with reads of extremely high quality (99% of the reads had an average quality ≥Q30; **Supplementary Figures S1B,C**). Quality-filtered reads were comparable with the expected ratios of host to microbial DNA for each condition (**Supplementary Figure S1A**).

Overall, synthetic samples with the expected host to bacterial DNA ratios were successfully generated.

Effect of the Level of Host DNA on the Sensitivity of WMS for Microbiome Taxonomic Profiling

After sequencing data pre-processing, the taxonomic profile of all samples was determined with MetaPhlAn2, with the aim to evaluate the effect of host DNA on the sensitivity of WMS for microbiome profiling. For that, bacteria species were grouped into the following categories, according to the number of 16S rRNA copies in the mock community: very low (10^4), low (10^5), high (10^6), and very high (10^7) abundant. The relative abundance of each taxa was then quantified at species-level and represented in a heat map (**Figure 1B**).

In the control MS, all 20 species of bacteria were successfully identified with a similar taxonomic profile compared to that of the expected, calculated based on the theoretical number of genome copies (**Figure 1B**). In three species, there was over- or underestimation of the relative abundances due to the GC content bias introduced during Illumina sequencing (**Supplementary Table S1** and **Supplementary Figure S2**).

The microbial profile of SS10 was comparable to the MS control, since all 20 bacterial species were detected with similar relative abundances to those of the MS (**Figure 1B** and **Supplementary Table S3**). In SS90, however, there was a decrease in the ability to detect very low abundant species. Specifically, *Deinococcus radiodurans* could not be identified (**Figure 1B**), and the relative abundances of *Actinomyces odontolyticus*, *Enterococcus faecalis*, and *Bacteroides vulgatus* were underestimated (**Supplementary Table S3**). The reduction in sensitivity was more striking in SS99, where only two of the low abundant and none of the very low abundant species were identified (**Figure 1B**).

In all conditions, unclassified clades were identified at low relative abundances (<2%), which likely represent bacterial species from the mock microbial community that were identified only at the genus or family level. Also, in synthetic samples with the highest amount of host DNA (SS90 and SS99), a mouse mammary tumor virus was identified (**Figure 1B**). Since viruses are not included in the mock community, the virus was likely introduced in the generation of the synthetic samples by the spiking with DNA from the mouse cell line, which could have the virus integrated in its genome. Besides, viruses are not included in the database used by Bowtie2, and therefore viral sequences have not been filtered as host contaminant.

These results show that the taxonomic profile of the mock microbial community was accurately reconstituted. Results also demonstrate that high ratios of host: bacterial DNA interfere with the sensitivity of WMS for taxonomic profiling. The increase in the proportion of host DNA leads to decreased sensitivity of WMS to detect very low and low abundant bacterial species.

Impact of Sequencing Depth on the Sensitivity of WMS for Microbiome Taxonomic Profiling

To assess the impact of sequencing depth on the sensitivity of WMS to detect bacterial species in samples with a high level of host DNA, reads from the SS90 metagenome were randomly subsampled, generating four datasets with reduced sequencing depths, corresponding to 50, 25, 10, and 5% of the original dataset (SS90D50, SS90D25, SS90D10, and SS90D5, respectively). Their taxonomic profile was compared to that of the original SS90 dataset (SS90D100; **Figure 2A**).

When the SS90 dataset was reduced to half of its original size (SS90D50), the number of very low abundant species that were not identified increased from one to three (**Figure 2B**), and the relative abundance of *E. faecalis* significantly decreased ($P = 0.006$; **Supplementary Tables S4**, **S5**). In SS90D25, none of the very low abundant species could be identified (**Figure 2B**). In comparison with the original dataset, no statistically significant differences were observed in the relative abundances of the remaining species (**Supplementary Tables S4**, **S5**). In SS90D10 and SS90D5, however, in addition to not identifying all very low abundant species, there were statistically significant decreases in the relative abundances of the low abundant species (**Figure 2B** and **Supplementary Tables S4**, **S5**).

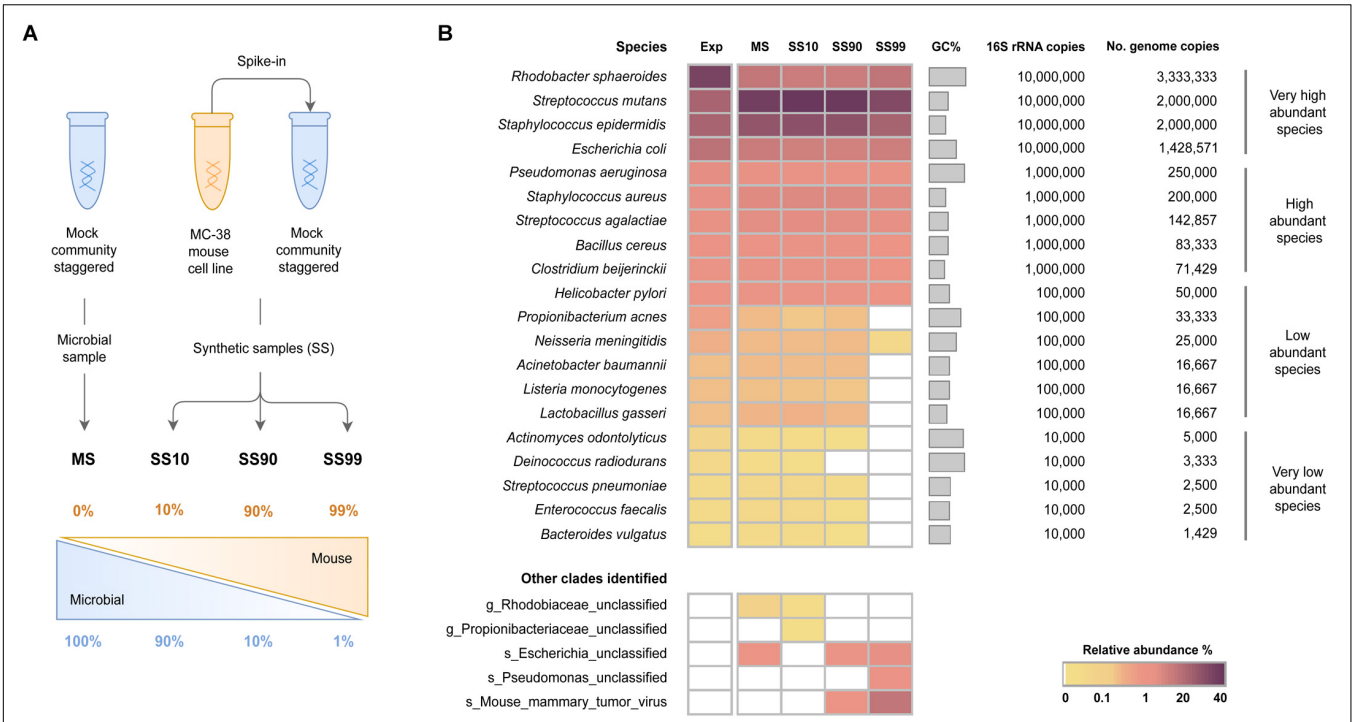


FIGURE 1 | Effect of the levels of host DNA on the sensitivity of WMS for microbiome taxonomic profiling. **(A)** Schematic representation of the experimental design to generate synthetic samples. DNA samples from a mock microbial community staggered (HM-277D) were spiked with DNA from a mouse cell line (MC-38 cells), generating three synthetic samples containing 10, 90, and 99% host DNA (SS10, SS90, and SS99, respectively). DNA from the mock microbial community was used as control (MS). **(B)** Taxonomic profile of the metagenomes of the synthetic samples determined with MetaPhlAn2, and expressed as relative abundance of species in a heat map. The expected (Exp) taxonomic results were estimated based on the theoretical number of species genome copies present in the mock. Species were sorted from the highest to the lowest expected relative abundances.

The reduction of the dataset to 5% of its original size led to significantly lower relative abundances of the majority of high and very high abundant species (**Figure 2B** and **Supplementary Tables S4**, **S5**). In addition, a misclassified species (*Pseudomonas* phage *Pf1*) with a relative abundance of 40% was identified (**Figure 2B**). This likely constitutes an artifact originated by the reduction of the size of the dataset.

Overall, these results demonstrate that sequencing depth has a major impact on the sensitivity of WMS for taxonomic profiling of samples with 90% host DNA. When decreasing sequencing depth, the number of microbial species that are not detected increase, along with unclassified and misclassified clades.

Influence of the Level of Host DNA on the Sensitivity of WMS for Microbiome Taxonomic Profiling at a Fixed Sequencing Depth

Having shown that high proportions of host DNA and reduced sequencing depths interfere with the sensitivity of WMS for microbiome profiling, the next aim was to investigate the influence of the level of host DNA on the sensitivity of the method at a fixed sequencing depth. For that, SD were generated with progressively greater proportions of host DNA (SD10 to SD99, ranging from 10 to 99% host reads), at the fixed sequencing depth of 10 million single-end reads with 150 bp length (1.5 Gb). This depth was chosen based on the recent guidelines for best practices for shotgun metagenomics, which suggest a minimum of 1 Gb sequencing depth per-sample (Quince et al., 2017). SD were composed of microbial and mouse reads, randomly picked from our previously generated MS and SS99 raw datasets, respectively (**Figure 3A**). For each of the SD, 5 replicates were generated, and the taxonomic profile was estimated with MetaPhlAn2, after sequencing data pre-processing, and compared to that of MS.

In SD10 to SD60, all 20 species of bacteria were successfully detected, without significant differences in their relative abundances in comparison with the MS control (**Figure 3B** and **Supplementary Tables S6**, **S7**). In SD70 to SD90, there was a progressive reduction of the number of very low abundant species identified, none of them being detected in SD90 (**Figure 3B** and **Supplementary Tables S6**, **S7**), a result in line with the random subsampling analysis performed above (**Figure 2B**). In SD92 to SD99, in addition to not identifying all of the very low abundant species, there was a statistically significant decrease in the relative abundance of low abundant species (**Figure 3B** and **Supplementary Tables S6**, **S7**). In particular, when host reads represented 97 to 99% of the datasets, the low abundant species were mostly undetected, and the relative abundance of the majority of high abundant species significantly

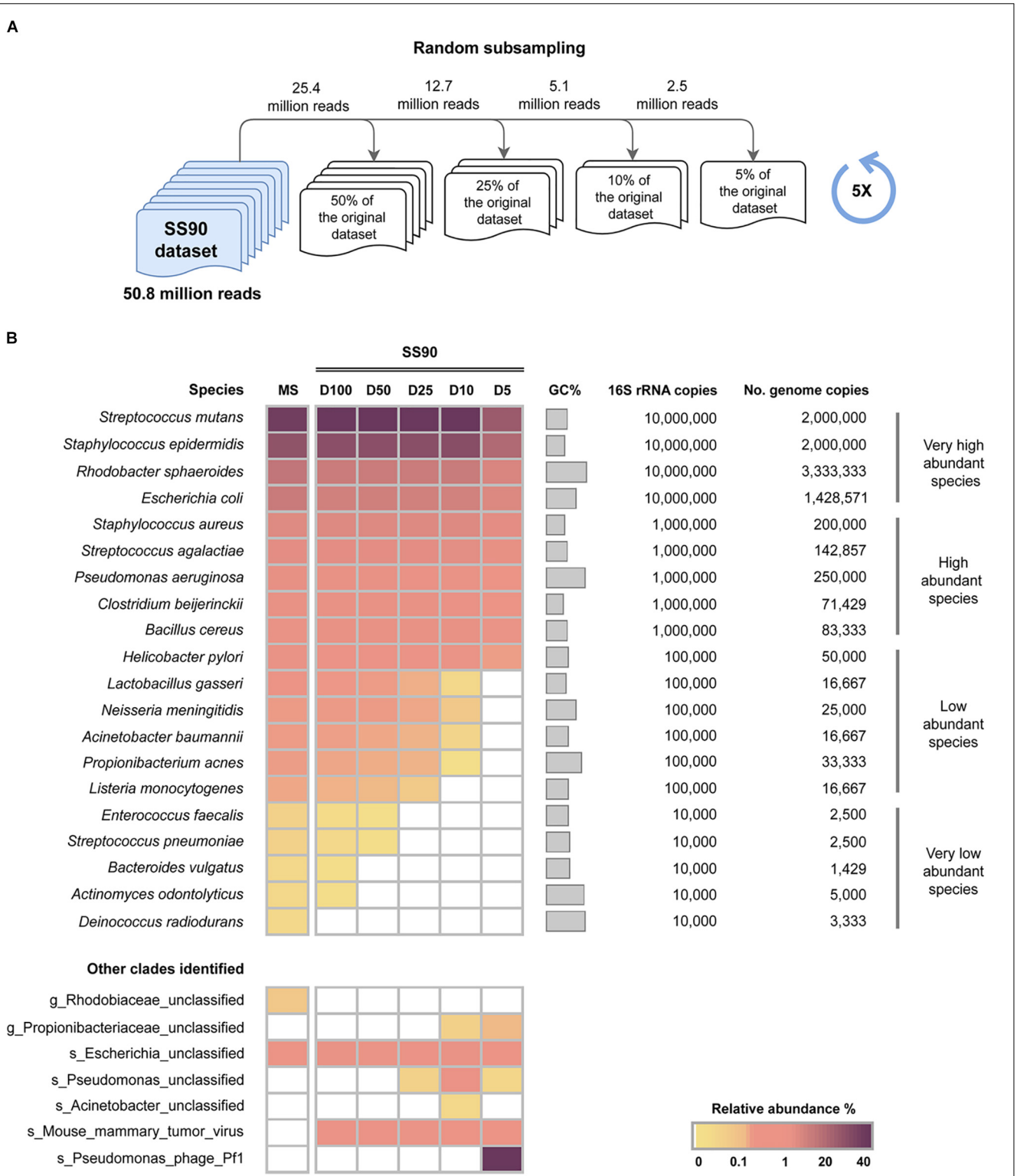


FIGURE 2 | Impact of sequencing depth on the sensitivity of WMS for microbiome taxonomic profiling. **(A)** Schematic representation of the experimental design to generate random subsampling reads from the SS90 original dataset (90% host DNA). Random subsampling corresponding to 50, 25, 10, and 5% of the reads from the original dataset (SS90D50, SS90D25, SS90D10, and SS90D5, respectively) were generated. **(B)** Taxonomic profile of the generated datasets are represented as the average relative abundance from five independent experiments, and shown in a heat map. Data was sorted from the highest to the lowest relative abundances of species in the mock microbial community (MS).

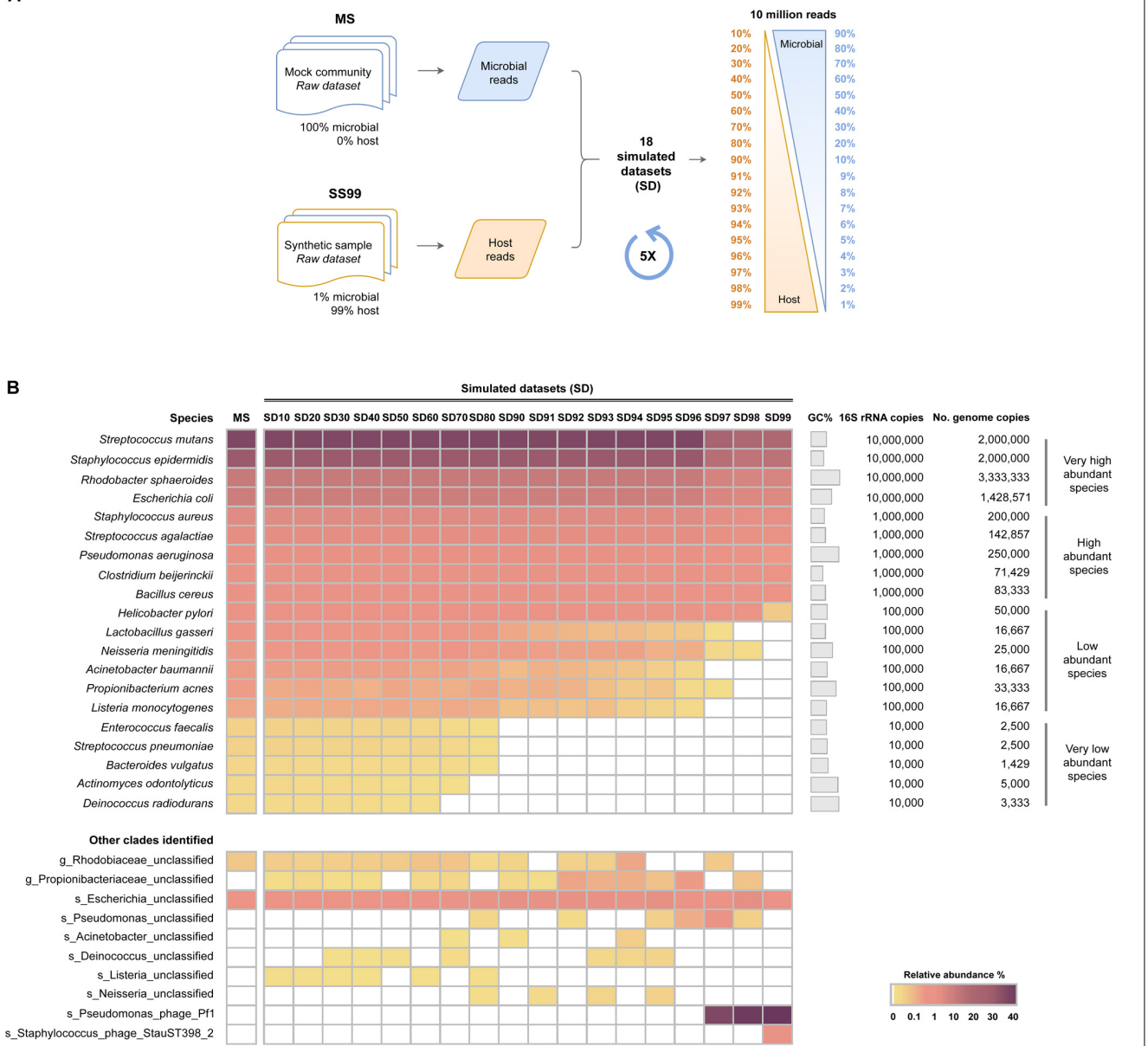


FIGURE 3 | Influence of host DNA on the sensitivity of WMS for microbiome taxonomic profiling at a fixed sequencing depth. **(A)** Schematic representation of the experimental design to generate simulated datasets (SD) with different host: microbial ratios. Microbial and host single-end reads were randomly selected from the mock microbial community (MS) and from the SS99 raw datasets, and were combined in different proportions, at a fixed sequencing depth of 10 million reads, to generate 18 simulated datasets (SD) with progressively higher host reads (ranging from 10 to 99%). **(B)** Taxonomic profile of the SD are represented as the average relative abundance from five independent simulations, and shown in a heat map. Data was sorted from the highest to the lowest relative abundances of species in the mock microbial community (MS).

decreased. Additionally, the *Pseudomonas* phage *Pf1* was again identified with high relative abundances in these highly complex datasets (**Figure 3B**). In the most complex dataset (SD99) a *Staphylococcus* phage *StauST398-2* was also identified.

Overall, these results show that at the fixed depth of 10 million reads, which is currently used in metagenomic studies, high levels of host DNA interfere with an accurate reconstitution of the microbiome profile.

DISCUSSION

Samples with a high amount of host DNA remains a major challenge in whole metagenome analysis, affecting the efficiency of microbiome profiling (Quince et al., 2017). Here, we show that high proportions of host DNA, reduce the sensitivity of WMS for microbiome profiling, in particular to detect very low and low abundant species of bacteria. It is plausible that high

ratios of host: bacteria DNA reduce sequence coverage of the microbial genomes, hindering subsequent taxonomic analysis. This is consistent with previous studies addressing the issue of human DNA contamination on WMS detection of the malaria parasite in clinical samples (Auburn et al., 2011; Oyola et al., 2013). Although these reports were not focused on microbiome characterization, they showed that low levels of human DNA ($\leq$30%) in blood samples, resulted in higher average *Plasmodium* genome coverage (Auburn et al., 2011), whereas clinical samples containing >80% human DNA, yielded a low number of reads assigned to *P. falciparum* genome (Oyola et al., 2013). In line with these observations, Hasan et al. (2016) found that by decreasing the human DNA background in a clinical sample, the sensitivity to detect microbial species was improved.

We also show that sequencing depth influences the sensitivity of microbiome profiling by WMS in samples that contain high levels of host DNA. The generation of SS90 datasets with reduced sequencing depths, resulted in gradually decreased capacity to accurately profile the microbiome. A reduction in sequencing depth from 51 million to 25 million reads already decreased WMS sensitivity, by preventing the identification of 60% of the species with very low abundance. In agreement with these findings, Jovel et al. (2016) showed that an increase in the size of the dataset leads to both an improvement of detection of microbial species and a more consistent estimation of their relative abundances. Moreover, in a metagenomic study of the fecal microbial community from beef cattle, the identification of new microbial taxa markedly improved with larger sequencing depths (Zaheer et al., 2018).

We also demonstrate that, besides preventing the identification of all species with very low abundance, a reduction in sequencing depth to five million reads additionally affected the relative abundance estimates of low abundant species. A depth as low as 2.5 million reads also resulted in major impairment in estimating the relative abundances of high and very high abundant species. In contrast with our findings, a recent study found no differences in the taxonomic profile of a mock community at divergent sequencing depths ranging from 0.1 to 7.5 single-end million reads (Walsh et al., 2018). These discrepancies may reflect the absence of host DNA in the mock sample analyzed in that study, as compared to the high levels of host DNA in our study samples (90%). Taken together, our data and that of others, suggest that similar sequencing depths have distinct effects on the sensitivity of WMS for taxonomic profiling, depending on the sample. In fact, our analysis of SD with 10 million reads indicated that the reconstitution of the microbiome profile becomes more inaccurate as the amount of host DNA in a sample progressively increases.

Interestingly, and based on this analysis, the outcomes of sequencing different types of host-derived samples, at a depth of 1.5 Gb per sample, can be extrapolated. For example, when sequencing a stool sample, the whole microbial community is expected to be accurately reconstituted, considering the low amount of host DNA in this type of sample [<10%; (Human Microbiome Project Consortium, 2012b)]. However, when sequencing samples like saliva, throat, buccal mucosa, and vaginal swabs (>90% host DNA), the detection of very low and low abundant species is expected to be impaired. This becomes more problematic in case of sequencing a tissue sample, as the detection of very low to high abundant species will be hampered, since this type of sample contains mostly human DNA (97 to 99% reads) and a low microbial biomass (Zhang et al., 2015). This also highlights another important aspect, which is the urgent need for effective host DNA depletion and/or microbial enrichment methods for whole metagenome analysis of tissue samples.

To the best of our knowledge, this is the first in-depth analysis demonstrating that greater proportions of host DNA, together with low sequencing depths, reduce the sensitivity of WMS for microbiome profiling. Therefore, the results of this study can assist in the design of WMS experiments, by highlighting the importance of sample type and sequencing depth when characterizing the microbiome.

DATA AVAILABILITY

The raw sequencing data has been deposited at the NCBI Sequence Read Archive (PRJNA521492).

AUTHOR CONTRIBUTIONS

RF, L-JvD, CK, and CF conceptualized and designed the study. JP-M, AH, IP-R, and MW acquired the data. JP-M, RF, CK, and CF performed the data analysis and interpretation. JP-M, RF, and CF drafted the manuscript. All authors revised the manuscript for important intellectual content.

FUNDING

This work was supported by European Regional Development Funds (ERDF) funds through the COMPETE 2020 – Operacional Programme for Competitiveness and Internationalization (POCI), Portugal 2020, and by FCT – Fundação para a Ciência e a Tecnologia (POCI-01-0145-FEDER-032532). JP-M and IP-R have fellowships from FCT (PD/BD/114014/2015 and SFRH/BD/110803/2015, respectively) through Programa Operacional Capital Humano (POCH) and the European Social Fund. JP-M's have fellowship from the framework of FCT's Ph.D. Program Biotech Health (Ref. PD/0016/2012).

ACKNOWLEDGMENTS

We thank Professor J. Machado, University of Porto, for the MC-38 cell line.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01277/full#supplementary-material

REFERENCES

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Andrews, S. (2016). *FastQC A Quality Control Tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed March 8, 2016).

Auburn, S., Campino, S., Clark, T. G., Djimde, A. A., Zongo, I., Pinches, R., et al. (2011). An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS One* 6:e22213. doi: 10.1371/journal.pone.0022213

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U.S.A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392–400. doi: 10.1038/nm.4517

Hasan, M. R., Rawat, A., Tang, P., Jithesh, P. V., Thomas, E., Tan, R., et al. (2016). Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. *J. Clin. Microbiol.* 54, 919–927. doi: 10.1128/JCM.03050-15

Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., et al. (2010). A catalog of reference genomes from the human microbiome. *Science* 328, 994–999. doi: 10.1126/science.1183605

Human Microbiome Project Consortium (2012a). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Human Microbiome Project Consortium (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'keefe, S., Mitchel, T., et al. (2016). Characterization of the gut microbiome using 16s or shotgun metagenomics. *Front. Microbiol.* 7:459. doi: 10.3389/fmicb.2016.00459

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* 30, 513–520. doi: 10.1038/nbt.2235

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013). *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14, 207–215. doi: 10.1016/j.chom.2013.07.007

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature* 550, 61–66. doi: 10.1038/nature23889

Marchesi, J. R., and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome* 3:31. doi: 10.1186/s40168-015-0094-5

Oyola, S. O., Gu, Y., Manske, M., Otto, T. D., O'brien, J., Alcock, D., et al. (2013). Efficient depletion of host DNA contamination in malaria clinical sequencing. *J. Clin. Microbiol.* 51, 745–751. doi: 10.1128/JCM.02507-12

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi: 10.1038/nbt.3935

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27, 626–638. doi: 10.1101/gr.216242.116

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74. doi: 10.1126/science.1093857

Walsh, A. M., Crispie, F., O'sullivan, O., Finnegan, L., Claesson, M. J., and Cotter, P. D. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome* 6:50. doi: 10.1186/s40168-018-0437-0

Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S. R., Marinier, E., Van Domselaar, G., et al. (2018). Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.* 8:5890. doi: 10.1038/s41598-018-24280-8

Zhang, C., Cleveland, K., Schnoll-Sussman, F., Mcclure, B., Bigg, M., Thakkar, P., et al. (2015). Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome Biol.* 16:265. doi: 10.1186/s13059-015-0821-z

Conflict of Interest Statement: AH, L-JvD, MW, and CK were employed by company DDL Diagnostic Laboratory, Rijswijk, Netherlands.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers

Supplementary Material

Impact of sample complexity and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis

Joana Pereira-Marques[1,2,3], Anne Hout[4], Rui M. Ferreira[1,2], Michiel Weber[4], Ines Pinto-Ribeiro[1,2,5], Leen-Jan van Doorn[4], Cornelis Willem Knetsch[4,*], Ceu Figueiredo[1,2,5,*]

[1]i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal;

[2]Ipatimup – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal;

[3]Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Porto, Portugal;

[4]DDL Diagnostic Laboratory, Rijswijk, The Netherlands; and

[5]Faculty of Medicine of the University of Porto, Porto, Portugal.

*** Correspondence:**
Cornelis Willem Knetsch: Wilco.Knetsch@ddl.nl

Ceu Figueiredo:cfigueiredo@ipatimup.pt

Supplementary Table 1. Composition of the mock microbial community (B: HM-277D, Staggered, High Concentration, v5.2H), showing the expected and the observed relative abundance of species after WMS. The ratio of observed to expected relative abundance of species is also shown. Grey shading indicates a ≥ 2-fold change**.**

Microbial species	NCBI assembly accession	16S rRNA copies	16S rRNA copies per genome [1]	No. genome copies	GC content [2] (%)	Expected relative abundance [3] (%)	Observed relative abundance (%)	Ratio observed/ expected
Rhodobacter sphaeroides ATCC 17023	GCF_003324715.1	10,000,000	3	3,333,333	69	34.43	12.79	0.37
Streptococcus mutans ATCC 700610	GCF_000007465.2	10,000,000	5	2,000,000	37	20.66	36.21	1.75
Staphylococcus epidermidis ATCC 12228	GCF_000007645.1	10,000,000	5	2,000,000	32	20.66	27.10	1.31
Escherichia coli ATCC 700926	GCF_002843685.1	10,000,000	7	1,428,571	51	14.75	10.81	0.73
Pseudomonas aeruginosa ATCC 47085	GCF_000006765.1	1,000,000	4	250,000	67	2.58	1.47	0.57
Staphylococcus aureus ATCC BAA-1717	GCF_000017085.1	1,000,000	5	200,000	33	2.07	4.27	2.07
Streptococcus agalactiae ATCC BAA-611	GCF_000007265.1	1,000,000	7	142,857	36	1.48	2.80	1.90
Bacillus cereus ATCC 10987	GCF_000008005.1	1,000,000	12	83,333	36	0.86	0.99	1.14
Clostridium beijerinckii ATCC 51743	GCF_000016965.1	1,000,000	14	71,429	30	0.74	1.14	1.55
Helicobacter pylori ATCC 700392	GCF_000307795.1	100,000	2	50,000	39	0.52	0.81	1.58
Propionibacterium acnes DSM16379	GCF_000008345.1	100,000	3	33,333	60	0.34	0.19	0.55
Neisseria meningitidis ATCC BAA-335	GCF_000008805.1	100,000	4	25,000	52	0.26	0.20	0.77
Acinetobacter baumannii ATCC 17978	GCF_001593425.2	100,000	6	16,667	39	0.17	0.19	1.13
Listeria monocytogenes ATCC BAA-679	GCF_000196035.1	100,000	6	16,667	38	0.17	0.16	0.93
Lactobacillus gasseri ATCC 33323	GCF_000014425.1	100,000	6	16,667	35	0.17	0.23	1.32
Actinomyces odontolyticus ATCC 17982	GCF_000154225.1	10,000	2	5,000	65	0.05	0.02	0.44
Deinococcus radiodurans ATCC 13939	GCF_001638825.1	10,000	3	3,333	67	0.03	0.02	0.56
Streptococcus pneumoniae ATCC BAA-334	GCF_000006885.1	10,000	4	2,500	40	0.03	0.04	1.55
Enterococcus faecalis ATCC 47077	GCF_000172575.2	10,000	4	2,500	38	0.03	0.04	1.57
Bacteroides vulgatus ATCC 8482	GCF_000012825.1	10,000	7	1,429	42	0.01	0.02	1.65

[1] Ribosomal RNA Database Curated by the Schmidt Laboratory (https://rrndb.umms.med.umich.edu/search/);
[2] NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/genomes/714);
[3] Expected relative abundance = No. genome copies of each species/ sum of genome copies of all species.

Supplementary Table 2. Summary of the sequencing data pre-processing of synthetic samples metagenomes.

Sample	Total number of raw single-end reads	Total number of quality-filtered reads	Total number of quality-filtered and host decontaminated reads
Microbial sample (MS)	39,682,202	33,309,964	33,309,243
Synthetic sample with 10% host DNA (SS10)	40,087,736	32,489,550	29,899,628
Synthetic sample with 90% host DNA (SS90)	50,846,240	41,297,894	5,535,588
Synthetic sample with 99% host DNA (SS99)	36,098,546	30,214,162	746,018

Supplementary Table 3. Ratio of relative abundances of species from synthetic samples to MS. Grey shading indicates a ≥ 2-fold change**.**

Microbial species	16S rRNA copies	MS	SS10/MS	SS90/MS	SS99/MS
Rhodobacter sphaeroides ATCC 17023	10,000,000	1	0.805	0.814	1.053
Streptococcus mutans ATCC 700610	10,000,000	1	1.081	1.061	0.877
Staphylococcus epidermidis ATCC 12228	10,000,000	1	1.069	1.053	0.776
Escherichia coli ATCC 700926	10,000,000	1	0.816	0.774	0.886
Pseudomonas aeruginosa ATCC 47085	1,000,000	1	0.742	0.798	0.714
Staphylococcus aureus ATCC BAA-1717	1,000,000	1	1.106	1.085	0.796
Streptococcus agalactiae ATCC BAA-611	1,000,000	1	1.048	1.022	0.711
Bacillus cereus ATCC 10987	1,000,000	1	1.021	1.237	0.574
Clostridium beijerinckii ATCC 51743	1,000,000	1	1.064	1.070	0.589
Helicobacter pylori ATCC 700392	100,000	1	0.911	0.907	0.561
Propionibacterium acnes DSM 16379	100,000	1	0.625	0.875	0
Neisseria meningitidis ATCC BAA-335	100,000	1	0.952	1.011	0.169
Acinetobacter baumannii ATCC 17978	100,000	1	0.967	0.962	0
Listeria monocytogenes ATCC BAA-679	100,000	1	0.943	0.824	0
Lactobacillus gasseri ATCC 33323	100,000	1	1.091	0.910	0
Actinomyces odontolyticus ATCC 17982	10,000	1	0.812	0.435	0
Deinococcus radiodurans ATCC 13939	10,000	1	0.626	0	0
Streptococcus pneumoniae ATCC BAA-334	10,000	1	1.006	0.608	0
Enterococcus faecalis ATCC 47077	10,000	1	1.065	0.310	0
Bacteroides vulgatus ATCC 8482	10,000	1	0.935	0.181	0

Supplementary Table 4. Ratio of relative abundances of species from each SS90 subset (SS90D50, SS90D25, SS90D10, SS90D5) to the SS90 original dataset (SS90D100). Random subsampling to generate each subset was performed in five independent experiments. Grey shading indicates a ≥ 2-fold change. nd: not detected in the SS90 original dataset.

Microbial Species	16S rRNA copies	SS90D100	SS90D50/ SS90D100	SS90D25/ SS90D100	SS90D10/ SS90D100	SS90D5/ SS90D100
Streptococcus mutans ATCC 700610	10,000,000	1	1.004	0.999	1.002	0.625
Staphylococcus epidermidis ATCC 12228	10,000,000	1	1.004	1.023	1.037	0.621
Rhodobacter sphaeroides ATCC 17023	10,000,000	1	0.995	0.995	1.019	0.585
Escherichia coli ATCC 700926	10,000,000	1	1.029	1.009	0.961	0.622
Staphylococcus aureus ATCC BAA-1717	1,000,000	1	1.008	0.997	0.989	0.588
Streptococcus agalactiae ATCC BAA-611	1,000,000	1	0.988	0.994	0.936	0.568
Pseudomonas aeruginosa ATCC 47085	1,000,000	1	0.950	0.922	0.794	0.329
Clostridium beijerinckii ATCC 51743	1,000,000	1	0.990	0.975	0.864	0.382
Bacillus cereus ATCC 10987	1,000,000	1	1.000	0.950	0.971	0.465
Helicobacter pylori ATCC 700392	100,000	1	0.957	0.964	0.776	0.257
Lactobacillus gasseri ATCC 33323	100,000	1	0.949	0.677	0.105	0
Neisseria meningitidis ATCC BAA-335	100,000	1	0.994	0.811	0.328	0
Acinetobacter baumannii ATCC 17978	100,000	1	0.861	0.690	0.156	0
Propionibacterium acnes DSM 16379	100,000	1	0.871	0.772	0.011	0
Listeria monocytogenes ATCC BAA-679	100,000	1	0.850	0.483	0	0
Enterococcus faecalis ATCC 47077	10,000	1	0.090	0	0	0
Streptococcus pneumoniae ATCC BAA-334	10,000	1	0.271	0	0	0
Bacteroides vulgatus ATCC 8482	10,000	1	0	0	0	0
Actinomyces odontolyticus ATCC 17982	10,000	1	0	0	0	0
Deinococcus radiodurans ATCC 13939	10,000	nd	nd	nd	nd	nd

Supplementary Table 5. Statistical analysis (*P*-values) of the results presented in Table S4. The Kruskal-Wallis non-parametric test followed by multiple comparisons (SS90D50, SS90D25, SS90D10, or SS90D5) versus a control group (SS90D100) using Dunn's test was performed for each species. NA: not applicable, as it was not detected in the SS90 original dataset.

Microbial species	16S rRNA copies	SS90D100 vs. SS90D50	SS90D100 vs. SS90D25	SS90D100 vs. SS90D10	SS90D100 vs. SS90D5
Streptococcus mutans ATCC 700610	10,000,000	> 0.9999	> 0.9999	> 0.9999	0.0203
Staphylococcus epidermidis ATCC 12228	10,000,000	> 0.9999	0.2548	0.0341	0.7829
Rhodobacter sphaeroides ATCC 17023	10,000,000	> 0.9999	> 0.9999	> 0.9999	0.0102
Escherichia coli ATCC 700926	10,000,000	0.9766	> 0.9999	> 0.9999	0.1242
Staphylococcus aureus ATCC BAA-1717	1,000,000	> 0.9999	> 0.9999	> 0.9999	0.0203
Streptococcus agalactiae ATCC BAA-611	1,000,000	0.8729	> 0.9999	0.0382	0.0004
Pseudomonas aeruginosa ATCC 47085	1,000,000	> 0.9999	> 0.9999	0.0231	0.0004
Clostridium beijerinckii ATCC 51743	1,000,000	> 0.9999	0.6831	0.223	0.0009
Bacillus cereus ATCC 10987	1,000,000	> 0.9999	> 0.9999	> 0.9999	0.0048
Helicobacter pylori ATCC 700392	100,000	> 0.9999	> 0.9999	0.0292	0.0006
Lactobacillus gasseri ATCC 33323	100,000	> 0.9999	0.1874	0.0025	0.0004
Neisseria meningitidis ATCC BAA-335	100,000	> 0.9999	0.7944	0.0242	0.0009
Acinetobacter baumannii ATCC 17978	100,000	0.9033	0.2575	0.0024	0.0003
Propionibacterium acnes DSM 16379	100,000	> 0.9999	0.6463	0.0015	0.0015
Listeria monocytogenes ATCC BAA-679	100,000	> 0.9999	0.1586	0.0006	0.0006
Enterococcus faecalis ATCC 47077	10,000	0.006	0.0007	0.0007	0.0007
Streptococcus pneumoniae ATCC BAA-334	10,000	0.4097	0.0011	0.0011	0.0011
Bacteroides vulgatus ATCC 8482	10,000	0.0005	0.0005	0.0005	0.0005
Actinomyces odontolyticus ATCC 17982	10,000	0.0005	0.0005	0.0005	0.0005
Deinococcus radiodurans ATCC 13939	10,000	NA	NA	NA	NA

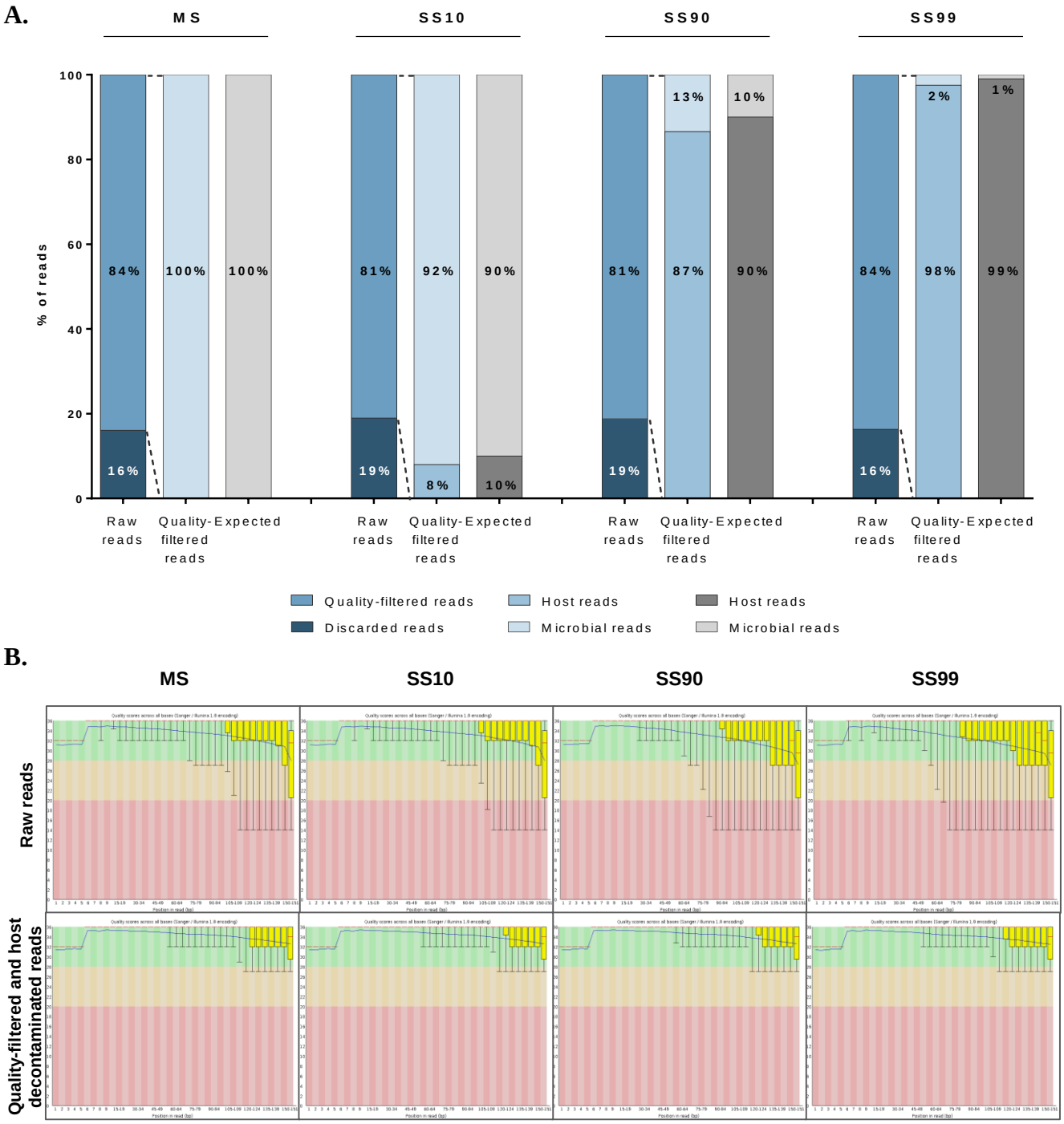
Supplementary Table 6. Ratio of relative abundances of species from each simulated dataset (SD) to MS dataset. Grey shading indicates a ≥ 2-fold change. Random subsampling to generate each simulated dataset was performed in five independent experiments.

Microbial Species	16S rRNA copies	MS	SD10 /MS	SD20 /MS	SD30 /MS	SD40 /MS	SD50 /MS	SD60 /MS	SD70 /MS	SD80 /MS	SD90 /MS	SD91 /MS	SD92 /MS	SD93 /MS	SD94 /MS	SD95 /MS	SD96 /MS	SD97 /MS	SD98 /MS	SD99 /MS
Streptococcus mutans	10,000,000	1	1.001	1.003	0.999	1.001	0.999	0.994	0.998	1.002	0.995	1.006	0.997	1.011	1.008	1.011	1.015	0.639	0.590	0.584
Staphylococcus epidermidis	10,000,000	1	1.001	1.005	1.003	1.005	1.005	1.006	1.002	1.005	1.015	1.002	1.003	1.027	1.013	1.005	1.038	0.659	0.572	0.603
Rhodobacter sphaeroides	10,000,000	1	1.002	0.993	0.996	1.000	1.001	1.001	0.998	0.994	0.988	1.000	1.005	0.985	1.001	0.983	1.005	0.638	0.570	0.572
Escherichia coli	10,000,000	1	1.003	1.014	0.997	0.987	0.974	0.987	1.008	0.960	0.913	0.946	0.965	0.911	0.982	0.959	0.940	0.597	0.520	0.499
Staphylococcus aureus	1,000,000	1	1.001	1.008	1.007	1.010	1.015	1.001	1.003	0.996	1.009	1.017	0.986	0.999	0.983	0.981	0.989	0.620	0.530	0.525
Streptococcus agalactiae	1,000,000	1	0.998	1.004	1.002	0.998	0.994	1.000	0.996	1.007	1.014	0.966	0.960	0.977	0.941	0.971	0.959	0.601	0.532	0.395
Pseudomonas aeruginosa	1,000,000	1	0.977	0.994	0.980	0.997	0.990	1.009	0.985	0.982	0.920	0.880	0.918	0.879	0.902	0.809	0.919	0.478	0.411	0.139
Clostridium beijerinckii	1,000,000	1	0.989	0.999	1.007	1.003	0.979	0.979	0.996	0.944	0.946	0.943	1.000	0.920	0.901	0.911	0.911	0.538	0.484	0.304
Bacillus cereus	1,000,000	1	1.039	0.977	1.037	0.978	1.036	1.023	0.991	1.003	0.903	0.979	1.065	0.997	1.011	1.015	0.851	0.698	0.568	0.435
Helicobacter pylori	100,000	1	0.992	0.982	0.990	0.990	0.955	0.963	0.939	0.923	0.909	0.953	0.967	0.851	0.806	0.827	0.798	0.435	0.301	0.072
Lactobacillus gasseri	100,000	1	0.941	0.939	0.919	0.924	0.910	0.903	0.906	0.890	0.698	0.588	0.509	0.413	0.424	0.305	0.295	0.033	0	0
Neisseria meningitidis	100,000	1	1.026	1.097	1.034	1.051	1.043	1.006	1.067	0.988	0.950	0.929	0.892	0.875	0.884	0.670	0.710	0.113	0.045	0
Acinetobacter baumannii	100,000	1	0.978	0.957	0.971	0.951	0.963	0.874	0.888	0.744	0.533	0.675	0.601	0.485	0.397	0.241	0.123	0	0	0
Propionibacterium acnes	100,000	1	0.780	0.830	0.749	0.674	0.843	0.760	0.712	0.858	0.676	0.736	0.659	0.460	0.320	0.381	0.078	0.034	0	0
Listeria monocytogenes	100,000	1	0.895	0.932	0.931	0.975	0.979	0.883	0.948	0.875	0.515	0.573	0.479	0.576	0.292	0.216	0.106	0	0	0
Enterococcus faecalis	10,000	1	0.858	0.820	0.837	0.705	0.820	0.588	0.325	0.210	0	0	0	0	0	0	0	0	0	0
Streptococcus pneumoniae	10,000	1	0.885	0.757	0.848	0.852	0.641	0.711	0.532	0.177	0	0	0	0	0	0	0	0	0	0
Bacteroides vulgatus	10,000	1	0.819	0.782	0.627	0.699	0.581	0.505	0.483	0.030	0	0	0	0	0	0	0	0	0	0
Actinomyces odontolyticus	10,000	1	0.778	0.655	0.685	0.538	0.319	0.157	0.015	0	0	0	0	0	0	0	0	0	0	0
Deinococcus radiodurans	10,000	1	0.704	0.831	0.662	0.523	0.428	0.212	0	0	0	0	0	0	0	0	0	0	0	0

Supplementary Table 7. Statistical analysis (*P*-values) of the results presented in Table S6. The Kruskal-Wallis non-parametric test followed by multiple comparisons (SD10, SD20, SD30, SD40, SD50, SD60, SD70, SD80, SD90, SD91, SD92, SD93, SD94, SD95, SD96, SD97, SD98 or SD99) versus a control group (MS) using Dunn's test was performed for each species.

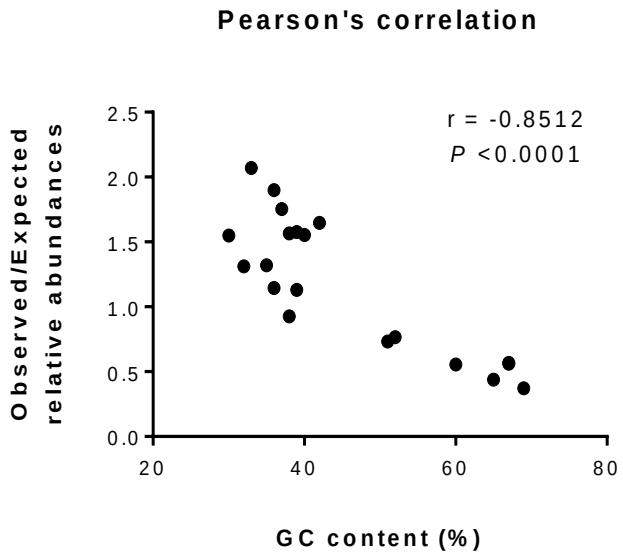
Microbial species	16S rRNA copies	MS vs. SD10	MS vs. SD20	MS vs. SD30	MS vs. SD40	MS vs. SD50	MS vs. SD60	MS vs. SD70	MS vs. SD80	MS vs. SD90	MS vs. SD91	MS vs. SD92	MS vs. SD93	MS vs. SD94	MS vs. SD95	MS vs. SD96	MS vs. SD97	MS vs. SD98	MS vs. SD99
Streptococcus mutans	10,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.172	0.2236	0.1963
Staphylococcus epidermidis	10,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1454	>0.999	>0.999	0.1608	>0.999	>0.999	>0.999
Rhodobacter sphaeroides	10,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.8504	0.1066	0.0994
Escherichia coli	10,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.0832	0.0195	0.0166
Staphylococcus aureus	1,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.6271	0.3169	0.3169
Streptococcus agalactiae	1,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.2796	>0.999	0.6096	0.0412	0.018	0.0053
Pseudomonas aeruginosa	1,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.7215	>0.999	>0.999	0.3926	>0.999	0.0747	>0.999	0.0085	0.0038	0.0007
Clostridium beijerinckii	1,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.8975	>0.999	0.5757	>0.999	0.0153	0.0105	0.0028
Bacillus cereus	1,000,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.3371	0.4981	0.1312
Helicobacter pylori	100,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.8736	>0.999	0.7417	>0.999	>0.999	0.9721	0.0462	0.2796	0.1899	0.0015	0.0005	0.0002
Lactobacillus gasseri	100,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.4485	0.0702	0.0371	0.012	0.0154	0.0044	0.0039	0.0001	<0.0001	<0.0001
Neisseria meningitidis	100,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1467	0.1057	0.081
Acinetobacter baumannii	100,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.9099	0.1078	0.4353	0.1922	0.0841	0.0305	0.0078	0.0019	0.0002	0.0002	0.0002
Propionibacterium acnes	100,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.7519	0.0449	0.0154	0.0181	0.0004	0.0002	0.0001	0.0001
Listeria monocytogenes	100,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.1784	0.3442	0.1697	0.7687	0.0247	0.0123	0.0035	0.0005	0.0005	0.0005
Enterococcus faecalis	10,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.5228	0.1474	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
Streptococcus pneumoniae	10,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.7679	0.1013	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
Bacteroides vulgatus	10,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.9671	0.0035	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
Actinomyces odontolyticus	10,000	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.0036	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
Deinococcus radiodurans	10,000	>0.999	>0.999	>0.999	>0.999	>0.999	0.6434	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014

Supplementary Figure 1.



Supplementary Figure 1. Overview of the sequencing data pre-processing from synthetic samples metagenomes. A. Summary view of the fraction of quality-filtered, discarded, microbial and host reads per sample. The *raw reads* bar represents the percentage of quality-filtered and discarded reads upon Trimmomatic quality filtering. The *quality-filtered reads* bar constitutes the fraction of host and microbial quality-filtered reads obtained after performing both quality-filtering and host sequences decontamination steps. The *expected* bar consists in the theoretical percentages of host and microbial reads expected in each synthetic sample. **B.** FastQC graphs showing per base sequence quality in the raw sequence data and in the host decontaminated quality-filtered data from the MS, SS10, SS90 and SS99. Quality scores values across all bases at each position are shown. **C.** Percentage of reads with an average quality of ≥Q30 in the MS, SS10, SS90 and SS99, in the raw and host decontaminated quality-filtered data.

Supplementary Figure 2.



Supplementary Figure 2. Correlation between genomic GC content and ratio of observed to expected relative abundances for the MS reference sample. Pearson's correlation coefficient (r) and *P*-value for the association are shown in the graph. Since it has been shown that the GC content introduces a bias during Illumina Nextera XT library preparation and sequencing (Jones *et al*. 2015), this potential bias was evaluated in the MS dataset. Pearson's correlation analysis showed that the GC content was negatively correlated with the ratio of observed to expected relative abundances (r = -0.8512, *P* <0.0001). This suggests that under and overestimated relative abundances were due to this bias.

Paper III

Helicobacter pylori Infection, the Gastric Microbiome and Gastric Cancer

Pereira-Marques J.*, Ferreira R.M.*, Pinto-Ribeiro I., Figueiredo C

Advances in Experimental Medicine and Biology. Springer, New York, N. 2019

*Equal contribution

Adv Exp Med Biol - Advances in Microbiology, Infectious Diseases and Public Health
https://doi.org/10.1007/5584_2019_366

Helicobacter pylori Infection, the Gastric Microbiome and Gastric Cancer

Joana Pereira-Marques, Rui M. Ferreira, Ines Pinto-Ribeiro, and Ceu Figueiredo

Abstract

After a long period during which the stomach was considered as an organ where microorganisms could not thrive, *Helicobacter pylori* was isolated *in vitro* from gastric biopsies, revolutionising the fields of Microbiology and Gastroenterology. Since then, and with the introduction of high-throughput sequencing technologies that allowed deep characterization of microbial communities, a growing body of knowledge has shown that the stomach contains a diverse microbial community, which is different from that of the oral cavity and of the intestine. Gastric cancer is a heterogeneous disease that is the end result of a cascade of events arising in a small fraction of patients colonized with *H. pylori.* In addition to *H. pylori* infection and to multiple host and environmental factors that influence disease development, alterations to the composition and function of the normal gastric microbiome, also known as dysbiosis, may also contribute to malignancy. Chronic inflammation of the mucosa in response to *H. pylori* may alter the gastric environment, paving the way to the growth of a dysbiotic gastric bacterial community. This dysbiotic microbiome may promote the development of gastric cancer by sustaining inflammation and/or inducing genotoxicity. This chapter summarizes what is known about the gastric microbiome in the context of *H. pylori*-associated gastric cancer, introducing the emerging dimension of the microbiome into the pathogenesis of this highly incident and deadly disease.

Keywords

Helicobacter pylori · Gastric microbiome · Gastric microbiota · Gastric cancer · Microbial dysbiosis

Author contributed equally with all other contributors. Joana Pereira-Marques and Rui M. Ferreira

J. Pereira-Marques
i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal

Ipatimup – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

ICBAS – Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal

R. M. Ferreira
i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal

Ipatimup – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

I. Pinto-Ribeiro and C. Figueiredo (✉)
i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal

Ipatimup – Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

Faculty of Medicine, University of Porto, Porto, Portugal
e-mail: cfigueiredo@ipatimup.pt

1 Introduction

The human body is inhabited in its different niches by a vast collection of microbes, generally known as the microbiota. These microorganisms, their genetic information, as well as the information of the niche in which they interact, are usually referred to as the microbiome (Cho and Blaser 2012). Currently, the term microbiome is also used to refer to the microorganisms themselves, i.e. the microbiota (Knight et al. 2017). The number of microbial cells was commonly thought to outnumber the quantity of human cells by a ten-fold ratio, but recent assessments propose a 1:1 ratio as a better estimate (Sender et al. 2016).

Bacteria constitute so far the best explored component of the microbiome. Progress in this research area had been hampered by the fact that only a very small fraction of the microbial species can be cultured *in vitro*. The advent of high-throughput sequencing technologies, together with the emergence of large international and interdisciplinary projects, have strongly contributed to expand our understanding of the microbiome structure and functions (Turnbaugh et al. 2007; Qin et al. 2010; Arnold et al. 2016).

It is currently accepted that the microbiome plays a major role in the maintenance of the normal physiology and health of the host, being involved in a wide variety of metabolic functions and participating in the normal maturation of the immune system (Gilbert et al. 2018). The composition of the normal microbiome varies between individuals and is influenced by local conditions inherent to the anatomic site, host genetics, diet, and antibiotic consumption (Lloyd-Price et al. 2017; Gilbert et al. 2018). Disruption of the balance that exists between the microbiome and the host, called dysbiosis, may promote numerous diseases, including cancer (Gilbert et al. 2018). For example, members of the gut microbiome such as *Fusobacterium nucleatum*, *Escherichia coli,* and *Bacteroides fragilis*, have been found enriched in colorectal cancer (Goodwin et al. 2011; Ahn et al. 2013; Bonnet et al. 2014). Nevertheless, and although the exact mechanisms linking microbial dysbiosis and cancer are still largely unknown, it can be anticipated that bacterial metabolites and toxins, as well as inflammation triggered by the microbiome contribute to the promotion of cancer. Here, we discuss in detail the current knowledge on the human gastric microbiome in the context of health and disease, and provide insights into the potential impact of microbial dysbiosis in the development of *H. pylori*-associated gastric cancer, by revisiting Correa's hypothesis of gastric carcinogenesis (Correa 1992).

2 Gastric Cancer

Gastric cancer is the fifth most incident cancer worldwide, with almost 1 million new cases per year (Ferlay et al. 2015). Gastric cancer is also the third cause of cancer-related death worldwide, with about 750,000 deaths estimated to having occurred in 2012. The incidence and mortality of gastric cancer show wide geographic variation, with East Asian countries registering the highest rates (Ferlay et al. 2015). Gastric cancer is a heterogeneous disease in what concerns morphology, genetics, and context. Histologically, gastric cancer heterogeneity is reflected by the diversity in classifications. The most commonly used histological classification systems are the one of the World Health Organization, comprising five main types – tubular, papillary, mucinous, poorly cohesive, and rare histological variants – and Lauren's, comprising two main types – diffuse and intestinal (Fenoglio-Preiser et al. 2010; Lauren 1965). Lauren's classification remains the most widely used and each cancer type has distinct epidemiologic and pathophysiological characteristics (Carneiro 1997; Spoto et al. 2018). Gastric cancer of the diffuse type occurs more frequently in females and at earlier ages, and is characterised by isolated or small groups of neoplastic cells that do not form glandular structures. In contrast, gastric cancer of the intestinal type is more prevalent at advanced ages, mainly in males, and is characterized by the presence of glandular structures and a higher to moderate degree of cell differentiation (Lauren 1965;

Carneiro 1997; Van Cutsem et al. 2016). The sequence of histological changes that culminate in intestinal type gastric cancer is better characterized than the one leading to diffuse type cancer, despite both types being associated with chronic gastritis as a consequence of *H. pylori* infection. Intestinal type gastric cancer is the result of a long, multifactorial and multistep process, which starts with *H. pylori* chronic gastritis, followed by atrophic gastritis, intestinal metaplasia, dysplasia, and cancer (Correa et al. 1975; Correa 1992).

Gastric cancer heterogeneity is also manifested at the molecular level (Ottini et al. 2006). Comprehensive analyses of gastric cancer tissues from large cohorts of patients recently emphasized the complexity of this disease and led to the proposal of different molecular classifications (Lei et al. 2013; Cancer Genome Atlas Research 2014; Cristescu et al. 2015). For example, the Cancer Genome Atlas research network classification proposed four main gastric cancer types (Cancer Genome Atlas Research 2014): chromosomally unstable tumours, which have marked aneuploidy, frequent mutations in *TP53*, amplification of receptor tyrosine kinases and *RAS*; microsatellite unstable tumours, which are characterised by *MLH1* promoter hypermethylation and a high mutational rate of genes including *TP53*, *KRAS*, *ARID1A*, *PIK3CA,* and *PTEN*; genomically stable tumours that have mutations in *CDH1*, encoding E-cadherin, *ARID1A* and *RHOA*; and Epstein-Barr virus-positive tumours, that show recurrent *PI3KCA* and *ARID1A*, but very rare *TP53* mutations, *CDKN2A* promoter hypermethylation, and amplification of *JAK2*, and of PD-L1- and PD-L2-encoding genes.

It is important to acknowledge that the great majority of gastric cancers occur in a sporadic setting, with about 10% of the cases having familial clustering, and 1–3% occurring in a hereditary setting (Oliveira et al. 2015; Van Cutsem et al. 2016). Hereditary diffuse gastric cancer (HDGC) is the most common and best-studied hereditary gastric cancer syndrome, where about 40% of the affected families have germline mutations in the *CDH1* gene, encoding the cell-cell adhesion protein E-cadherin (Oliveira et al. 2015). The other two syndromes, gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS) and familial intestinal gastric cancer (FIGC), are very rare, and while in the former germline point mutations in the *APC* promoter have been identified (Worthley et al. 2012; Li et al. 2016), in the latter no aetiological genetic alterations are known (Oliveira et al. 2015). The exact extent to which *H. pylori* infection and/or the microbiome of the stomach may contribute to the different molecular profiles and contexts of gastric cancer, however, remains largely unexplored.

3 *H. pylori* Infection and Gastric Cancer

H. pylori is considered as the major risk factor for the development of gastric cancer, being categorized as a class I carcinogen by the International Agency for Research on Cancer (IARC 1994). It has been estimated that at least 90% of all non-cardia gastric cancers worldwide are attributable to *H. pylori* (Plummer et al. 2015). The estimated worldwide prevalence of *H. pylori* is 44.3%, with considerable variation according to the geographic region (Zamani et al. 2018). There is a major geographic overlap between *H. pylori* prevalence and gastric cancer incidence, and in general countries with highest cancer incidence have high infection rates (Ferlay et al. 2015; Zamani et al. 2018). Since the initial collection of epidemiological and functional data that provided grounds for the classification of *H. pylori* as a class I carcinogen, numerous studies have been published demonstrating the causal relationship between chronic *H. pylori* infection and gastric cancer (IARC 2011). The magnitude of the risk of gastric cancer associated with *H. pylori* infection has now been estimated in different populations, and varies with the type of assay used to detect the infection, being about three-fold if serology is used (Helicobacter and Cancer Collaborative Group 2001) and reaching over 20-fold when more sensitive assays are used (Gonzalez et al. 2012). As an additional piece of evidence that links *H. pylori* infection and gastric cancer, the eradication of the infection has an

impact in reducing the incidence of this malignancy (Ford et al. 2015).

Although the association between *H. pylori* and gastric cancer is extensively recognized, the majority of the infected patients do not develop this malignancy, which arguments in favour of the multifactorial nature of this disease. Host genetic susceptibility, namely polymorphisms in genes that are involved in the inflammatory response to *H. pylori* infection have been associated with the risk of gastric cancer. Among the best studied are those that encode interleukin (IL)-1β, IL-1 receptor antagonist, tumour necrosis factor (TNF)-α pro-inflammatory cytokines and the anti-inflammatory IL-10. Genetic variation in the promoters or in non-coding regions of these genes are associated with increased risk for the development of gastric cancer (El-Omar et al. 2001; Machado et al. 2003; Persson et al. 2011). Remarkably, in genetically susceptible hosts, infection with more virulent *H. pylori* strains markedly enhances gastric cancer risk (Figueiredo et al. 2002).

Cigarette smoking, alcohol intake, and salt consumption are recognized environmental factors that influence the risk of gastric cancer. Indeed, ever and current smokers have higher risk to develop gastric cancer compared with never smokers, and among current smokers the risk increases with number of cigarettes per day (Praud et al. 2018). Heavy and very heavy alcohol drinkers have higher risks for developing gastric cancer in comparison with abstainers, and these associations are independent of the *H. pylori* infection status (Rota et al. 2017). Dietary salt intake is also associated with gastric cancer risk, the risk being gradually increased for higher consumption levels (D'Elia et al. 2012). Accordingly, in an animal model of infection, a diet with high salt content accelerated the development of gastric cancer, in particular in animals infected with *cagA*-positive *H. pylori* strains (Gaddy et al. 2013). On the other hand, the consumption of fruit and white vegetables, which are rich sources of vitamin C, are inversely associated with gastric cancer risk (Fang et al. 2015).

Adding to the influence of host and environmental factors in gastric cancer, the genetic diversity of *H. pylori*, and in particular variation in virulence genes associated with the pathogenicity of strains, also impact gastric cancer risk (Ferreira et al. 2014). CagA is the best-documented *H. pylori* virulence factor influencing gastric cancer. CagA is encoded by a pathogenicity island that is present in about 60–70% of *H. pylori* strains worldwide. The same pathogenicity island also encodes a type IV secretion system, which functions as a molecular syringe and allows CagA to be delivered into the host cells (Backert et al. 2015). Once in the host cell cytoplasm, CagA can be phosphorylated by host kinases within EPIYA motifs. Both phosphorylated and non-phosphorylated CagA are capable of activating signalling pathways that influence host responses, including inflammation, proliferation, and cell polarity (Backert et al. 2010). CagA phosphorylation, however, appears to be important in gastric cancer development, as transgenic mice expressing wild-type CagA, but not phosphorylation-resistant CagA, develop gastric tumours (Ohnishi et al. 2008). Patients who are infected with *H. pylori cagA*-positive strains, and with strains with CagA harbouring higher number of phosphorylation motifs, are associated with increased risk for gastric premalignant lesions and for gastric cancer (Ferreira et al. 2014). Additionally, CagA influences host disease progression, and infection with *H. pylori cagA*-positive strains increases the risk of progression of preneoplastic lesions (Plummer et al. 2007; Gonzalez et al. 2011). Variation in other *H. pylori* virulence factors, such as the VacA toxin, has also been associated with gastric precancerous lesions and cancer (Gonzalez et al. 2011; Ferreira et al. 2014). This and other virulence factors of *H. pylori* and their relationship with disease are discussed in Chap. 3 of this

volume. Additionally, the molecular mechanisms that underlie *H. pylori*-mediated malignant transformation are discussed in Chap. 8.

4 The Gastric Microbiota, Is There More Than *H. pylori*?

For many years, the human stomach was assumed to be sterile, given its high acidic pH, gastric peristalsis, and the presence of digestive enzymes, among other protective and antimicrobial factors (Martinsen et al. 2005). With the discovery and isolation of *H. pylori* (Warren and Marshall 1983) this dogma was broken, and more recently the idea that the stomach harbours a complex bacterial community became accepted. Initial analyses of the bacteria present in the stomach relied on microbiological cultures. These have identified *Firmicutes* as the most common phylum, followed by *Proteobacteria*, *Bacteroidetes*, and *Actinobacteria*, and genera that were most commonly isolated included *Streptococcus*, *Lactobacillus*, *Bacteroides*, *Staphylococcus*, *Veillonella*, *Corynebacterium, Clostridium,* and *Neisseria* (Stockbruegger 1985; Thorens et al. 1996; Adamsson et al. 1999; Mowat et al. 2000; Zilberstein et al. 2007). This type of approach, however, yielded an incomplete and biased landscape of the gastric microbiota, since most of the bacteria are difficult to culture or are uncultivable. The development of culture-independent methods revealed that the human gastric ecosystem has a more diverse and complex microbiota than initially anticipated (Monstein et al. 2000; Bik et al. 2006; Andersson et al. 2008; Li et al. 2009; Delgado et al. 2013; Schulz et al. 2018).

The bacterial community of the normal stomach has not been extensively characterised, probably due to difficulties in recruiting normal individuals for upper endoscopy. A 16S rRNA gene cloning and sequencing-based approach was undertaken to analyse the gastric microbial communities of five individuals with normal gastric mucosa and five patients with non-*H. pylori* and non-NSAID (non-steroidal anti-inflammatory drug) (NHNN) gastritis, all Chinese from Hong-Kong (Li et al. 2009). *Firmicutes* and *Proteobacteria* were the most represented phyla, and while in the normal stomach the *Proteobacteria* was the most abundant, in the NHNN gastritis the most abundant phylum was the *Firmicutes*. The five most common genera were *Streptococcus*, *Prevotella*, *Neisseria*, *Haemophilus*, and *Porphyromonas*; together, *Streptococcus* and *Prevotella* represented over 40% of all sequences.

Following studies exposed the diversity and the inter-individual variability of the gastric microbiota derived from the analysis of populations from distinct origins, but also from different sample types, and using various technical approaches. Overall, the most common gastric bacteria can be assigned to five major phyla – *Proteobacteria*, *Firmicutes, Bacteroidetes*, *Actinobacteria* and *Fusobacteria*, and the two most prominent genera of the non-*H. pylori* infected stomach are *Streptococcus* and *Prevotella* (Bik et al. 2006; Andersson et al. 2008; Li et al. 2009; Delgado et al. 2013). A more recent study that included 20 Caucasians from the UK with a normal stomach, without evidence of *H. pylori* infection, concurred that the bacterial family *Prevotellaceae* was the most abundant (23%), followed by *Streptococcaceae* (10%). In fact, the microbiota of these stomachs had the highest levels of microbial diversity and bacterial richness in comparison with other groups of patients infected with *H. pylori* (Parsons et al. 2017).

According to the great majority of reports, when *H. pylori* is present, this bacterium is the most abundant microbial component, representing between 40% to over 95% of the gastric microbiota (Bik et al. 2006; Andersson et al. 2008; Li et al. 2017; Klymiuk et al. 2017; Schulz et al. 2018; Ferreira et al. 2018; Parsons et al. 2017). In addition to finding *H. pylori* as the most abundant bacterium in the stomach of patients who test positive for *H. pylori*, it has been shown that the microbiota of *H. pylori*-positive subjects has lower diversity than that of *H. pylori*-negative subjects (Bik et al. 2006; Andersson et al. 2008; Schulz et al. 2018). Our analysis of the gastric microbiota of 81 chronic

gastritis cases from Portugal that were 99% *H. pylori*-positive, revealed that as *H. pylori* abundance increases, there is a significant decrease in diversity (data not shown). Accordingly, a study that evaluated the gastric microbiota before and after *H. pylori* eradication treatment showed that the eradication of *H. pylori* resulted in an increase in bacterial diversity (Li et al. 2017).

The influence of *H. pylori* on the composition and dynamics of the gastric microbiota is still not fully understood. Difficulties may in part relate to the differences in methods to diagnose *H. pylori* infection and various studies using sequencing-based methods have demonstrated that *H. pylori* could be detected at low levels in samples of subjects that were diagnosed as *H. pylori*-negative by conventional methods (histopathology, rapid urease test, serology, and PCR) (Bik et al. 2006; Maldonado-Contreras et al. 2011; Delgado et al. 2013; Thorell et al. 2017).

The majority of reports show no major alterations on the pattern of distribution of phyla between *H. pylori*-positive and *H. pylori*-negative patients (Bik et al. 2006; Maldonado-Contreras et al. 2011; Schulz et al. 2018). Using the PhyloChip microarray, Maldonado-Contreras *et al.* reported a similar representation of the four dominant phyla between *H. pylori*-infected and -uninfected rural Amerindians (Maldonado-Contreras et al. 2011). In regression analyses, authors were able to identify an association between *H. pylori* positivity and decreased relative abundance of *Actinobacteria*, *Bacteroidetes,* and *Firmicutes*. These results are sustained by our data on Portuguese patients with chronic gastritis, in which we found an inverse correlation between the relative abundance of *H. pylori* and non-*Helicobacter* α- and β-*Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes* (Ferreira et al. 2018).

Experimental infections of the rhesus macaque model were used to assess the impact of *H. pylori* challenge upon the pre-existing gastric microbiota (Martin et al. 2013). Data showed that although *Helicobacter* became dominant in challenged animals, the removal of the *Helicobacter* reads from the libraries did not significantly alter the relative abundance of taxa between challenged and unchallenged animals. Nevertheless, the impact of *H. pylori* on relatively rare taxa was not determined. In contrast, in a mouse model of infection, challenge of animals with *H. pylori* significantly and consistently affected the abundance of several species, suggesting that *H. pylori* influences the gastric microbiota composition at lower taxonomic levels (Kienesberger et al. 2016).

It has been a matter of debate whether bacteria found in the stomach represent transient swallowed bacteria or active members of a resident microbiota colonizing the gastric mucosa. Comparisons of the microbial communities along different sites of the gastrointestinal (GI) tract have shown that the gastric microbiota is different from that at other sites. Although some proximity with the microbiota of the oral cavity and throat exists, the stomach microbial communities cluster together (Andersson et al. 2008; Stearns et al. 2011; Delgado et al. 2013). Recent data aiming to evaluate the metabolically active microbial communities in different regions of the GI tract found that the transient luminal microbiota present in gastric juice is closely related with that of saliva and of duodenal aspirates and significantly different from that of gastric biopsies, supporting the idea that the stomach has a local mucosa-associated microbiota (Schulz et al. 2018).

5 The Gastric Microbiota in Gastric Carcinogenesis

While *H. pylori* is recognized as being fundamental in gastric carcinogenesis, the role of non-*H. pylori* microbiota has not yet been established. The majority of the publications so far included low number of patients and/or had limitations in sensitivity and depth of coverage, which in general did not allow producing statistically based conclusions. One of the first DNA-based descriptions of the gastric bacterial community in patients with gastric cancer, used terminal restriction fragment length polymorphism (T-RFLP) in combination with 16S rRNA

gene cloning and sequencing to characterize 10 patients with gastric cancer and five *H. pylori*-negative dyspeptics with normal gastric mucosa (Dicksved et al. 2009). A complex bacterial community dominated by different species of *Streptococcus*, *Lactobacillus*, *Veilonella* and *Prevotella*, and with low abundance of *H. pylori* was reported in the stomach of cancer patients.

A study of 15 patients from Mexico with non-atrophic gastritis, intestinal metaplasia, or gastric cancer, using the PhyloChip, showed a gastric microbiota profile separation between non-atrophic gastritis and gastric cancer based on the presence/absence of taxa. This analysis could neither separate non-atrophic gastritis and intestinal metaplasia, nor metaplasia and cancer (Aviles-Jimenez et al. 2014). Taxa with differences in abundance between non-atrophic gastritis and gastric cancer were identified, with significant decreases in the abundance of *Porphyromonas*, *Neisseria* and bacteria from the *TM7* phylum, and increases in the abundance of *Lactobacillus* and *Lachnospiraceae* observed in gastric cancer. Diversity, as measured by bacterial richness, was statistically significantly decreased from non-atrophic gastritis to gastric cancer. In contrast, a survey of the metabolic active bacteria of the stomach of 12 gastric cancer and 20 functional dyspepsia patients of Chinese ethnicity from Singapore and Malaysia, detected an increase in species richness and in phylogenetic diversity in cancer (Castano-Rodriguez et al. 2017). An earlier study of 10 chronic gastritis, 10 intestinal metaplasia and 11 gastric cancer patients from Korea, also suggested an increase in bacterial diversity from gastritis to cancer, but without supporting statistical analysis (Eun et al. 2014). Still, the majority of publications so far report a decrease in bacteria diversity and richness from non-atrophic gastritis to gastric cancer (Aviles-Jimenez et al. 2014; Li et al. 2017; Coker et al. 2018; Ferreira et al. 2018).

The two most complete gastric microbiota studies in the gastric cancer field using 16S rRNA gene sequencing were published in the beginning of 2018 (Coker et al. 2018; Ferreira et al. 2018). Coker and colleagues studied the gastric mucosal microbiota in different histological stages of gastric carcinogenesis in 81 patients from Xi'an in China (Coker et al. 2018). The analysis of 21 superficial gastritis, 23 atrophic gastritis, 17 intestinal metaplasia, and 20 gastric cancer patients, demonstrated that the gastric microbiota of patients with intestinal metaplasia and with gastric cancer had significantly reduced microbial richness in comparison with that of superficial gastritis patients. Although no significant differences were found in microbiota profiles between superficial gastritis, atrophic gastritis and intestinal metaplasia, the microbiota of these stages were significantly different from that of the gastric cancer. The screen for differentially abundant taxa revealed 21 taxa enriched and 10 taxa depleted in gastric cancer in comparison with superficial gastritis, with increasing strengths of interactions among them along the progression of disease. Among the cancer-enriched bacteria were members of the human oral microbiome *Peptostreptococcus*, *Streptococcus*, *Parvimonas*, *Slackia*, and *Dialister*, which were the most significant in network interaction analysis. These bacteria were able to distinguish gastric cancer from superficial gastritis in receiver-operating characteristic (ROC) analysis. The authors validated their results in a Chinese Inner Mongolian cohort of patients (Coker et al. 2018).

Our own studies analysing 135 Portuguese patients, showed significant differences in the structure as well as in the composition of the gastric microbial communities between chronic gastritis and gastric cancer patients (Ferreira et al. 2018). Overall, patients with cancer had significantly decreased gastric microbial diversity, as assessed by the Shannon index, in comparison with patients with chronic gastritis. The gastric microbiota profiles of the two patient groups could be separated based on both the presence/absence and the relative abundance of taxa. In our series, *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria* and *Fusobacteria* were identified as the five most abundant phyla in the stomach, in agreement with earlier descriptions (Bik et al. 2006; Aviles-Jimenez et al. 2014; Jo et al. 2016). Phyla ranked in the same relative abundance in the two patient

groups, with significantly increased abundance of non-*Helicobacter Proteobacteria*, *Actinobacteria* and *Firmicutes* and lower abundance of *Bacteroidetes* and *Fusobacteria* in the gastric cancer microbiota. While being the major genus in chronic gastritis with a mean relative abundance of 42% (varying from 0.01–95%), *Helicobacter* had a significant reduction in abundance in gastric cancer. In fact, and despite 87% of the gastric cancer patients were *H. pylori*-positive, the mean relative abundance of reads was just 6% (Ferreira et al. 2018). Actually, the gastric microbiota profiles of the two clinical settings could be distinguished based on *Helicobacter* abundance.

Overall, we have identified 29 microbial taxa, including 10 differentially abundant genera that best explain the differences between patient groups. Differential abundances in the great majority of these genera were further validated using quantitative polymerase chain reaction in the discovery cohort, and additionally confirmed in validation cohorts comprising patients from Portugal, China and Mexico (Ferreira et al. 2018). *Helicobacter*, *Neisseria*, *Prevotella*, and *Streptococcus* were enriched in the microbiota of chronic gastritis patients. *Streptococcus*, *Prevotella* and *Neisseria* are among the most abundant commensals of the oral cavity (Bik et al. 2010) and among the most frequently detected bacteria in the non-neoplastic stomach, having been cultured or identified in gastric juice and/or biopsies from *H. pylori*-positive and -negative gastritis and in the normal stomach (Thorens et al. 1996; Bik et al. 2006; Li et al. 2009; Delgado et al. 2013; Parsons et al. 2017; Schulz et al. 2018). Interestingly, in a comparison of the gastric microbiota of Colombian inhabitants from two regions with divergent gastric cancer risks, *Streptococcus* and *Neisseria* were identified only in individuals from the low risk, but not in those from the high risk gastric cancer region (Yang et al. 2016).

Genera that were enriched in the gastric cancer microbiota, and significantly more prevalent in patients with gastric cancer than in patients with chronic gastritis, were *Achromobacter*, *Citrobacter*, *Lactobacillus*, *Clostridium*, *Rhodococcus,* and *Phyllobacterium* (Ferreira et al. 2018). These bacteria comprise several intestinal residents that may become opportunistic pathogens (Kelly and LaMont 2008; Rajilic-Stojanovic and de Vos 2014), and indeed *Lactobacillus*, *Clostridium*, and *Citrobacter* have been detected in the gastric juice or gastric biopsies from patients taking acid suppressive drugs, and patients with intestinal metaplasia and gastric cancer (Sjostedt et al. 1985; Mowat et al. 2000; Dicksved et al. 2009; Aviles-Jimenez et al. 2014). In a recent study of nine gastritis and 11 gastric cancer patients from Taiwan, species of *Clostridium* and *Lactobacillus* were also found enriched in the gastric cancer microbiota (Hsieh et al. 2018).

Microbial dysbiosis was inversely correlated with the microbial diversity and was significantly higher in cancer than in gastritis, a finding that was validated in additional patient cohorts (Ferreira et al. 2018). Actually, microbial dysbiosis could distinguish gastric cancer and chronic gastritis patients in ROC analysis. Interestingly, microbial dysbiosis could discriminate gastric cancer better than individual genera, suggesting that alterations to the microbial community as a whole rather than particular bacteria contribute to the development of gastric cancer.

The role of the microbiota in the promotion of neoplasia is supported by data obtained in the insulin-gastrin (INS-GAS) transgenic mouse model. In comparison with germ-free INS-GAS mice, those harbouring a complex microbiota had higher levels of gastric inflammation, epithelial damage, oxyntic gland atrophy, hyperplasia, metaplasia, and dysplasia. When infected with *H. pylori*, INS-GAS mice that harboured a complex microbiota had more severe gastric lesions and an earlier development of gastrointestinal intraepithelial neoplasia (GIN) in comparison to *H. pylori*-infected germ-free INS-GAS mice (Lofgren et al. 2011). Furthermore, progression towards GIN occurred to a similar extent in *H. pylori*-infected INS-GAS mice with a complex microbiota and in *H. pylori*-infected INS-GAS mice colonized with a restricted microbiota consisting of only three species of commensal murine bacteria (*Clostridium* sp., *Lactobacillus*

murinus, and *Bacteroides* sp.) (Lertpiriyapong et al. 2014). These results suggest that colonization of the stomach with commensal bacteria from other locations of the GI tract may promote *H. pylori*-associated gastric cancer. Altogether, these studies highlight that there is a shift in the composition of the stomach microbiome from gastritis to gastric cancer, with a likely reduction of bacterial diversity, and with increased microbial dysbiosis in the cancerous stomach.

6 Revisiting Correa's Hypothesis of Gastric Carcinogenesis

In the multistep model of gastric carcinogenesis proposed by Pelayo Correa, persistent infection of the gastric mucosa with *H. pylori* initiates and perpetuates an inflammatory process that can progress to atrophic gastritis, intestinal metaplasia, dysplasia, and gastric cancer (Correa 1992). In this model, *H. pylori* infection plays an important role in the initial phases of the cascade. Indeed, *H. pylori* scarcely colonizes the severe atrophic stomach and may progressively disappear in gastric tissues at later steps of carcinogenesis (Correa 1992; Kuipers 1998). Analyses of the gastric microbiome have also described decreased relative abundance of *H. pylori* in gastric cancer (Dicksved et al. 2009; Eun et al. 2014; Ferreira et al. 2018; Hsieh et al. 2018), although this was not consistently observed or not reported (Yu et al. 2017; Coker et al. 2018).

The hypothesis of Correa contemplated that the loss of acid-secreting parietal cells in *H. pylori*-induced atrophic gastritis leads to higher gastric pH, and to proliferation in the stomach of bacteria that are capable of reducing nitrate to nitrite, to form N-nitroso compounds that are mutagenic (Correa et al. 1975; Correa 1992). Actually, significant intragastric bacterial overgrowth has been demonstrated in patients on long-term acid suppression by the use of proton pump inhibitors (PPIs) or histamine-2 receptor antagonists (Stockbruegger 1985; Sanduleanu et al. 2001). A recent investigation of 24 dyspeptic Italian patients, showed that although PPI treatment did not have a major influence in the gastric microbiota composition, an increase in the relative abundance of *Firmicutes*, namely *Streptococcus* was reported (Paroni Sterbini et al. 2016). In accordance with these findings, in a study analysing the metabolically active gastric microbial communities of 19 patients from the UK receiving PPI therapy and 20 individuals with normal stomach, relatively few alterations in the gastric microbiota were detected, but *Streptococcus* was significantly enriched in PPI-treated patients (Parsons et al. 2017). An enrichment in *Streptococcaceae* in the gut microbiota of PPI users has also been reported in two large studies (Imhann et al. 2016; Jackson et al. 2016). The enrichment of upper GI tract commensals observed in the stomach and in the gut, may be related with the disruption of the highly acidic barrier of the stomach induced by the acid suppressive therapy.

Likewise, the increase of the pH of the stomach due to decreased acid production as a result of parietal cell loss in *H. pylori*-associated atrophy, may generate a niche that becomes suitable to the establishment of a different microbiome (Plottel and Blaser 2011). One may speculate that this altered gastric microbiome, where *H. pylori* is less abundant or absent, and where commensal bacteria from other locations of the GI tract thrive, would act as continuous stimuli by maintaining the inflammatory process and/or inducing genotoxicity, thus promoting gastric carcinogenesis (Fig.1). This would in part explain the lack of success of *H. pylori* eradication in preventing progression of preneoplastic lesions and gastric cancer in patients with atrophy or intestinal metaplasia at baseline (Wong et al. 2004; Mera et al. 2018).

The microbiome of the cancerous stomach is functionally different from that of the stomach without cancer (Coker et al. 2018; Ferreira et al. 2018). Although only a very limited number of studies have addressed this aspect, predictive functional analyses have revealed that gastric cancer patients have an enrichment of several microbial pathways, including those related with membrane transport, carbohydrate digestion and absorption, carbohydrate metabolism, xenobiotics biodegradation and metabolism, and

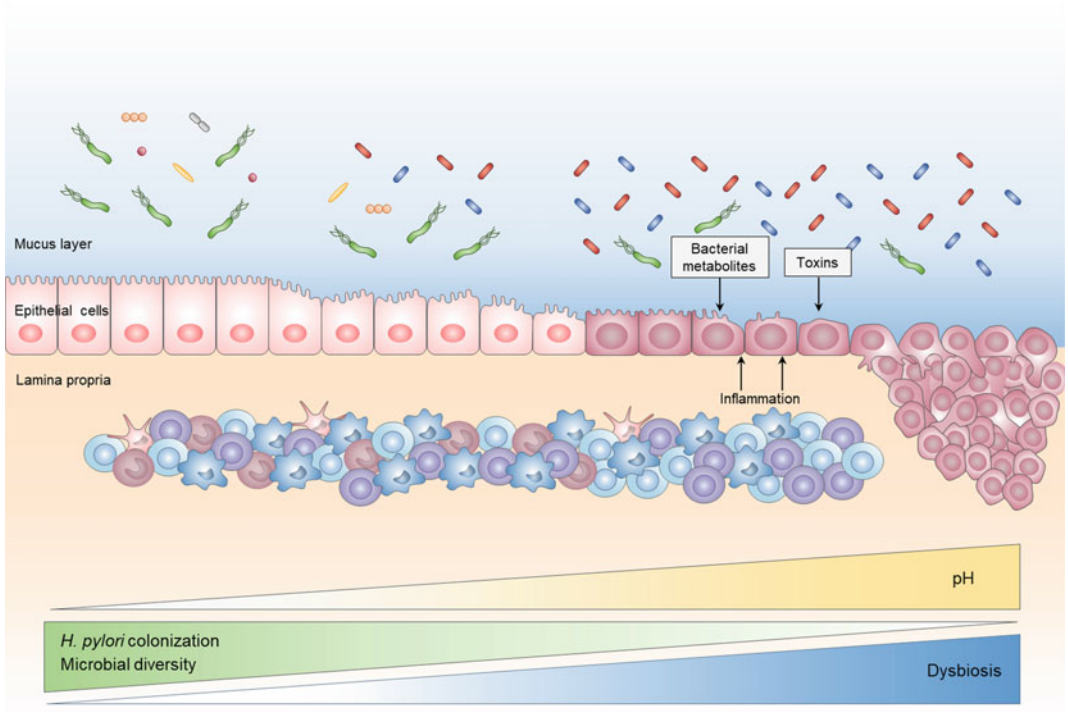


Fig. 1 Model for microbial dysbiosis in gastric cancer development. *H. pylori* infection triggers and perpetuates an inflammatory response in the gastric mucosa that, in some of the infected individuals, leads to loss of acid-secreting parietal cells with increase of the gastric pH. In this altered environment, *H. pylori* colonization decreases, and bacteria from other locations of the GI tract establish in the gastric niche, resulting in dysbiosis. This dysbiotic microbiome, characterized by reduced microbial diversity, may promote the development of gastric cancer by sustaining inflammation and/or inducing genotoxicity. Bacteria: green, *H. pylori;* orange, pink and grey, resident mucosa-associated microbiota; blue and red, dysbiotic microbiota; Inflammatory cells: dark blue, macrophages; pink, dendritic cells; dark pink, monocytes; light blue, CD4 T-lymphocytes; violet, CD8 T-lymphocytes

lipid metabolism (Tseng et al. 2016; Castano-Rodriguez et al. 2017; Coker et al. 2018; Ferreira et al. 2018). Findings are, however, relatively divergent between studies and results should therefore be interpreted with caution.

To revisit Correa's hypothesis that nitrate-reducing bacteria contribute to malignant transformation of the atrophic stomach by increasing the concentrations of nitrite and of N-nitroso compounds, we have assessed the functional features of the microbiome involved in these reactions (Ferreira et al. 2018). By fully reconstituting the metagenomes, based on the profiles obtained from the 16S rRNA gene sequences, we showed that in comparison with chronic gastritis, the gastric cancer microbiome had an increased representation of nitrate reductase and of nitrite reductase functions, the enzymes that respectively reduce nitrate to nitrite and nitrite to nitric oxide. The four genera *Citrobacter*, *Achromobacter*, *Clostridium* and *Phyllobacterium* were identified as the major contributors to these functions (Ferreira et al. 2018). Interestingly, and in agreement with our observations, are those of a follow-up study conducted in Taiwan to evaluate the effects of subtotal gastrectomy as a treatment for early-stage gastric cancer. The alteration of the gastric environment by the surgery led to significant changes in the gastric microbial community, and

nitrate reductase, nitrite reductase, and other functions related to nitrosation were enriched in the gastric microbiome before, but not after subtotal gastrectomy (Tseng et al. 2016). These data suggest that the gastric cancer microbiome has the potential to produce carcinogenic N-nitroso compounds. Additional features linked to the dysbiotic microbiome may be involved in the promotion of a carcinogenic environment in the stomach. Microbial metabolites and toxins, as well as inflammation by-products generated by the dysbiotic microbiome, may directly induce host cell damage or interfere with host signalling pathways that influence cell turnover and survival, thus increasing the risk for gastric malignant transformation (Fig.1).

7 Conclusions

Despite the recent advances in the investigation of the human gastric microbiome, research in this area remains limited. Although a number of papers about the microbiome of the stomach in the context of gastric carcinogenesis have been published, caution should be taken with the interpretation of the results of very distinct technical approaches. Additionally, differences in the geographic origin, genetic background, and environmental exposures of the populations should be taken into consideration.

While it is clear that the microbial community present in gastric cancer is distinct from that present in chronic gastritis, research conducted on the microbiome of the histological stages that precede gastric cancer is still lacking. Studies in large and clinically well-defined patient populations will be key to determine the role of microbial dysbiosis in progression to cancer. The shift from descriptive to functionally based studies that investigate the effects of specific taxa and/or bacterial derived-metabolites in the gastric mucosa, will allow gaining insights into the mechanisms that lead to dysbiosis-associated genotoxicity and inflammation. Uncovering these mechanisms will create the grounds for translating microbiome research into prevention, diagnosis, and treatment improvements to control and decrease gastric cancer burden.

Acknowledgements JPM, RMF and IPR have fellowships from Fundação para a Ciência e a Tecnologia (FCT; PD/BD/114014/2015, SFRH/BPD/84084/2012, and SFRH/BD/110803/2015, respectively) through Programa Operacional Capital Humano (POCH) and the European Union. JPM's fellowship is in the framework of FCT's PhD Programme BiotechHealth (Ref PD/0016/2012). i3S-Instituto de Investigação e Inovação em Saúde is funded by Fundo Europeu de Desenvolvimento Regional (FEDER) funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalisation (POCI), Portugal 2020, and by Portuguese funds through Fundação para a Ciência e a Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Inovação (POCI-01-0145-FEDER-007274).

References

Adamsson I, Nord CE, Lundquist P, Sjostedt S, Edlund C (1999) Comparative effects of omeprazole, amoxycillin plus metronidazole versus omeprazole, clarithromycin plus metronidazole on the oral, gastric and intestinal microflora in *Helicobacter pylori*-infected patients. J Antimicrob Chemother 44 (5):629–640

Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, Goedert JJ, Hayes RB, Yang L (2013) Human gut microbiome and risk for colorectal cancer. J Natl Cancer Inst 105 (24):1907–1911. https://doi.org/10.1093/jnci/djt300

Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One 3(7):e2836. https://doi.org/10.1371/journal.pone.0002836

Arnold JW, Roach J, Azcarate-Peril MA (2016) Emerging technologies for Gut microbiome research. Trends Microbiol 24(11):887–901. https://doi.org/10.1016/j.tim.2016.06.008

Aviles-Jimenez F, Vazquez-Jimenez F, Medrano-Guzman R, Mantilla A, Torres J (2014) Stomach microbiota composition varies between patients with non-atrophic gastritis and patients with intestinal type of gastric cancer. Sci Rep 4:4202. https://doi.org/10.1038/srep04202

Backert S, Tegtmeyer N, Selbach M (2010) The versatility of *Helicobacter pylori* CagA effector protein functions: the master key hypothesis. Helicobacter 15 (3):163–176. https://doi.org/10.1111/j.1523-5378.2010.00759.x

Backert S, Tegtmeyer N, Fischer W (2015) Composition, structure and function of the *Helicobacter pylori* cag pathogenicity island encoded type IV secretion system. Future Microbiol 10(6):955–965. https://doi.org/10.2217/fmb.15.32

Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, Francois F, Perez-Perez G, Blaser MJ, Relman DA (2006) Molecular analysis of the bacterial microbiota in the human stomach. Proc Natl Acad Sci U S A 103 (3):732–737. https://doi.org/10.1073/pnas.0506655103

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF, Nelson KE, Gill SR, Fraser-Liggett CM, Relman DA (2010) Bacterial diversity in the oral cavity of 10 healthy individuals. ISME J 4 (8):962–974. https://doi.org/10.1038/ismej.2010.30

Bonnet M, Buc E, Sauvanet P, Darcha C, Dubois D, Pereira B, Dechelotte P, Bonnet R, Pezet D, Darfeuille-Michaud A (2014) Colonization of the human gut by *E. coli* and colorectal cancer risk. Clin Cancer Res 20(4):859–867. https://doi.org/10.1158/1078-0432.CCR-13-1343

Cancer Genome Atlas Research (2014) Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513(7517):202–209. https://doi.org/10.1038/nature13480

Carneiro F (1997) Classification of gastric carcinomas. Curr Diagn Pathol 4(1):51–59. https://doi.org/10.1016/S0968-6053(97)80008-7

Castano-Rodriguez N, Goh KL, Fock KM, Mitchell HM, Kaakoush NO (2017) Dysbiosis of the microbiome in gastric carcinogenesis. Sci Rep 7(1):15957. https://doi.org/10.1038/s41598-017-16289-2

Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13 (4):260–270. https://doi.org/10.1038/nrg3182

Coker OO, Dai Z, Nie Y, Zhao G, Cao L, Nakatsu G, Wu WK, Wong SH, Chen Z, Sung JJY, Yu J (2018) Mucosal microbiome dysbiosis in gastric carcinogenesis. Gut 67(6):1024–1032. https://doi.org/10.1136/gutjnl-2017-314281

Correa P (1992) Human gastric carcinogenesis: a multistep and multifactorial process–First American Cancer Society Award Lecture on cancer epidemiology and prevention. Cancer Res 52(24):6735–6740

Correa P, Haenszel W, Cuello C, Tannenbaum S, Archer M (1975) A model for gastric cancer epidemiology. Lancet 2(7924):58–60

Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K, Ye XS, Do IG, Liu S, Gong L, Fu J, Jin JG, Choi MG, Sohn TS, Lee JH, Bae JM, Kim ST, Park SH, Sohn I, Jung SH, Tan P, Chen R, Hardwick J, Kang WK, Ayers M, Hongyue D, Reinhard C, Loboda A, Kim S, Aggarwal A (2015) Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 21(5):449–456. https://doi.org/10.1038/nm.3850

D'Elia L, Rossi G, Ippolito R, Cappuccio FP, Strazzullo P (2012) Habitual salt intake and risk of gastric cancer: a meta-analysis of prospective studies. Clin Nutr 31 (4):489–498. https://doi.org/10.1016/j.clnu.2012.01.003

Delgado S, Cabrera-Rubio R, Mira A, Suarez A, Mayo B (2013) Microbiological survey of the human gastric ecosystem using culturing and pyrosequencing methods. Microb Ecol 65(3):763–772. https://doi.org/10.1007/s00248-013-0192-5

Dicksved J, Lindberg M, Rosenquist M, Enroth H, Jansson JK, Engstrand L (2009) Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls. J Med Microbiol 58(Pt 4):509–516. https://doi.org/10.1099/jmm.0.007302-0

El-Omar EM, Carrington M, Chow WH, McColl KE, Bream JH, Young HA, Herrera J, Lissowska J, Yuan CC, Rothman N, Lanyon G, Martin M, Fraumeni JF Jr, Rabkin CS (2001) The role of interleukin-1 polymorphisms in the pathogenesis of gastric cancer. Nature 412(6842):99. https://doi.org/10.1038/35083631

Eun CS, Kim BK, Han DS, Kim SY, Kim KM, Choi BY, Song KS, Kim YS, Kim JF (2014) Differences in gastric mucosal microbiota profiling in patients with chronic gastritis, intestinal metaplasia, and gastric cancer using pyrosequencing methods. Helicobacter 19 (6):407–416. https://doi.org/10.1111/hel.12145

Fang X, Wei J, He X, An P, Wang H, Jiang L, Shao D, Liang H, Li Y, Wang F, Min J (2015) Landscape of dietary factors associated with risk of gastric cancer: a systematic review and dose-response meta-analysis of prospective cohort studies. Eur J Cancer 51 (18):2820–2832. https://doi.org/10.1016/j.ejca.2015.09.010

Fenoglio-Preiser C, Carneiro F, Correa P, Guilford P, Lambert R, Magraud F, Muñoz N, Powell SM, Ruggae M, Sasako M, Stolte M, Watanabe H (2010) Tumours of the stomach: gastric carcinoma. In: Bosman FT, Carneiro F, Hruban RH, Theise ND (eds) WHO classification of tumours of the digestive system, 4th edn. World Health Organization, Geneva, pp 37–67

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136(5):E359–E386. https://doi.org/10.1002/ijc.29210

Ferreira RM, Machado JC, Figueiredo C (2014) Clinical relevance of *Helicobacter pylori* vacA and cagA genotypes in gastric carcinoma. Best Pract Res Clin Gastroenterol 28(6):1003–1015. https://doi.org/10.1016/j.bpg.2014.09.004

Ferreira RM, Pereira-Marques J, Pinto-Ribeiro I, Costa JL, Carneiro F, Machado JC, Figueiredo C (2018) Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. Gut 67(2):226–236. https://doi.org/10.1136/gutjnl-2017-314205

Figueiredo C, Machado JC, Pharoah P, Seruca R, Sousa S, Carvalho R, Capelinha AF, Quint W, Caldas C, van Doorn LJ, Carneiro F, Sobrinho-Simoes M (2002) *Helicobacter pylori* and interleukin 1 genotyping: an

opportunity to identify high-risk individuals for gastric carcinoma. J Natl Cancer Inst 94(22):1680–1687

Ford AC, Forman D, Hunt R, Yuan Y, Moayyedi P (2015) *Helicobacter pylori* eradication for the prevention of gastric neoplasia. Cochrane Database Syst Rev (7): CD005583. https://doi.org/10.1002/14651858.CD005583.pub2

Gaddy JA, Radin JN, Loh JT, Zhang F, Washington MK, Peek RM Jr, Algood HM, Cover TL (2013) High dietary salt intake exacerbates *Helicobacter pylori*-induced gastric carcinogenesis. Infect Immun 81 (6):2258–2267. https://doi.org/10.1128/IAI.01271-12

Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R (2018) Current understanding of the human microbiome. Nat Med 24(4):392–400. https://doi.org/10.1038/nm.4517

Gonzalez CA, Figueiredo C, Lic CB, Ferreira RM, Pardo ML, Ruiz Liso JM, Alonso P, Sala N, Capella G, Sanz-Anquela JM (2011) *Helicobacter pylori* cagA and vacA genotypes as predictors of progression of gastric preneoplastic lesions: a long-term follow-up in a high-risk area in Spain. Am J Gastroenterol 106 (5):867–874. https://doi.org/10.1038/ajg.2011.1

Gonzalez CA, Megraud F, Buissonniere A, Lujan Barroso L, Agudo A, Duell EJ, Boutron-Ruault MC, Clavel-Chapelon F, Palli D, Krogh V, Mattiello A, Tumino R, Sacerdote C, Quiros JR, Sanchez-Cantalejo E, Navarro C, Barricarte A, Dorronsoro M, Khaw KT, Wareham N, Allen NE, Tsilidis KK, Bas Bueno-de-Mesquita H, Jeurnink SM, Numans ME, Peeters PH, Lagiou P, Valanou E, Trichopoulou A, Kaaks R, Lukanova-McGregor A, Bergman MM, Boeing H, Manjer J, Lindkvist B, Stenling R, Hallmans G, Mortensen LM, Overvad K, Olsen A, Tjonneland A, Bakken K, Dumeaux V, Lund E, Jenab M, Romieu I, Michaud D, Mouw T, Carneiro F, Fenge C, Riboli E (2012) *Helicobacter pylori* infection assessed by ELISA and by immunoblot and noncardia gastric cancer risk in a prospective study: the Eurgast-EPIC project. Ann Oncol 23 (5):1320–1324. https://doi.org/10.1093/annonc/mdr384

Goodwin AC, Destefano Shields CE, Wu S, Huso DL, Wu X, Murray-Stewart TR, Hacker-Prietz A, Rabizadeh S, Woster PM, Sears CL, Casero RA, Jr. (2011) Polyamine catabolism contributes to enterotoxigenic *Bacteroides fragilis*-induced colon tumorigenesis. Proc Natl Acad Sci U S A 108 (37):15354–15359. doi:https://doi.org/10.1073/pnas.1010203108

Helicobacter and Cancer Collaborative Group (2001) Gastric cancer and *Helicobacter pylori*: a combined analysis of 12 case control studies nested within prospective cohorts. Gut 49(3):347–353

Hsieh YY, Tung SY, Pan HY, Yen CW, Xu HW, Lin YJ, Deng YF, Hsu WT, Wu CS, Li C (2018) Increased abundance of *Clostridium* and *Fusobacterium* in gastric microbiota of patients with gastric cancer in Taiwan. Sci Rep 8(1):158. https://doi.org/10.1038/s41598-017-18596-0

IARC (1994) Schistosomes, liver flukes and *Helicobacter pylori*. IARC working group on the evaluation of carcinogenic risks to humans. Lyon, 7–14 June 1994. IARC Monogr Eval Carcinog Risks Hum 61:1–241

IARC (2011) IARC. Monographs on the evaluation of carcinogenic risks to humans, volume 100. A review of carcinogen—part B: biological agents. International Agency for Research on Cancer, Lyon

Imhann F, Bonder MJ, Vich Vila A, Fu J, Mujagic Z, Vork L, Tigchelaar EF, Jankipersadsing SA, Cenit MC, Harmsen HJ, Dijkstra G, Franke L, Xavier RJ, Jonkers D, Wijmenga C, Weersma RK, Zhernakova A (2016) Proton pump inhibitors affect the gut microbiome. Gut 65(5):740–748. https://doi.org/10.1136/gutjnl-2015-310376

Jackson MA, Goodrich JK, Maxan ME, Freedberg DE, Abrams JA, Poole AC, Sutter JL, Welter D, Ley RE, Bell JT, Spector TD, Steves CJ (2016) Proton pump inhibitors alter the composition of the gut microbiota. Gut 65(5):749–756. https://doi.org/10.1136/gutjnl-2015-310861

Jo HJ, Kim J, Kim N, Park JH, Nam RH, Seok YJ, Kim YR, Kim JS, Kim JM, Kim JM, Lee DH, Jung HC (2016) Analysis of gastric microbiota by pyrosequencing: minor role of bacteria other than *Helicobacter pylori* in the gastric carcinogenesis. Helicobacter 21(5):364–374. https://doi.org/10.1111/hel.12293

Kelly CP, LaMont JT (2008) *Clostridium difficile*–more difficult than ever. N Engl J Med 359(18):1932–1940. https://doi.org/10.1056/NEJMra0707500

Kienesberger S, Cox LM, Livanos A, Zhang XS, Chung J, Perez-Perez GI, Gorkiewicz G, Zechner EL, Blaser MJ (2016) Gastric *Helicobacter pylori* infection affects local and distant microbial populations and host responses. Cell Rep 14(6):1395–1407. https://doi.org/10.1016/j.celrep.2016.01.017

Klymiuk I, Bilgilier C, Stadlmann A, Thannesberger J, Kastner MT, Hogenauer C, Puspok A, Biowski-Frotz-S, Schrutka-Kolbl C, Thallinger GG, Steininger C (2017) The human gastric microbiome is predicated upon infection with *Helicobacter pylori*. Front Microbiol 8:2508. https://doi.org/10.3389/fmicb.2017.02508

Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, Sogin ML (2017) The microbiome and human biology. Annu Rev Genomics Hum Genet 18:65–86. https://doi.org/10.1146/annurev-genom-083115-022438

Kuipers EJ (1998) Review article: relationship between *Helicobacter pylori*, atrophic gastritis and gastric cancer. Aliment Pharmacol Ther 12(Suppl 1):25–36

Lauren P (1965) The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. Acta Pathol Microbiol Scand 64:31–49

Lei Z, Tan IB, Das K, Deng N, Zouridis H, Pattison S, Chua C, Feng Z, Guan YK, Ooi CH, Ivanova T, Zhang S, Lee M, Wu J, Ngo A, Manesh S, Tan E, Teh BT, So JB, Goh LK, Boussioutas A, Lim TK, Flotow H, Tan P, Rozen SG (2013) Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology 145(3):554–565. https://doi.org/10.1053/j.gastro.2013.05.010

Lertpiriyapong K, Whary MT, Muthupalani S, Lofgren JL, Gamazon ER, Feng Y, Ge Z, Wang TC, Fox JG (2014) Gastric colonisation with a restricted commensal microbiota replicates the promotion of neoplastic lesions by diverse intestinal microbiota in the *Helicobacter pylori* INS-GAS mouse model of gastric carcinogenesis. Gut 63(1):54–63. https://doi.org/10.1136/gutjnl-2013-305178

Li XX, Wong GL, To KF, Wong VW, Lai LH, Chow DK, Lau JY, Sung JJ, Ding C (2009) Bacterial microbiota profiling in gastritis without *Helicobacter pylori* infection or non-steroidal anti-inflammatory drug use. PLoS One 4(11):e7985. https://doi.org/10.1371/journal.pone.0007985

Li J, Woods SL, Healey S, Beesley J, Chen X, Lee JS, Sivakumaran H, Wayte N, Nones K, Waterfall JJ, Pearson J, Patch AM, Senz J, Ferreira MA, Kaurah P, Mackenzie R, Heravi-Moussavi A, Hansford S, Lannagan TRM, Spurdle AB, Simpson PT, da Silva L, Lakhani SR, Clouston AD, Bettington M, Grimpen F, Busuttil RA, Di Costanzo N, Boussioutas A, Jeanjean M, Chong G, Fabre A, Olschwang S, Faulkner GJ, Bellos E, Coin L, Rioux K, Bathe OF, Wen X, Martin HC, Neklason DW, Davis SR, Walker RL, Calzone KA, Avital I, Heller T, Koh C, Pineda M, Rudloff U, Quezado M, Pichurin PN, Hulick PJ, Weissman SM, Newlin A, Rubinstein WS, Sampson JE, Hamman K, Goldgar D, Poplawski N, Phillips K, Schofield L, Armstrong J, Kiraly-Borri C, Suthers GK, Huntsman DG, Foulkes WD, Carneiro F, Lindor NM, Edwards SL, French JD, Waddell N, Meltzer PS, Worthley DL, Schrader KA, Chenevix-Trench G (2016) Point mutations in exon 1B of APC reveal gastric adenocarcinoma and proximal polyposis of the stomach as a familial adenomatous polyposis variant. Am J Hum Genet 98(5):830–842. https://doi.org/10.1016/j.ajhg.2016.03.001

Li TH, Qin Y, Sham PC, Lau KS, Chu KM, Leung WK (2017) Alterations in gastric microbiota after *H. pylori* eradication and in different histological stages of gastric carcinogenesis. Sci Rep 7:44935. https://doi.org/10.1038/srep44935

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C (2017) Strains, functions and dynamics in the expanded human microbiome project. Nature 550(7674):61–66. https://doi.org/10.1038/nature23889

Lofgren JL, Whary MT, Ge Z, Muthupalani S, Taylor NS, Mobley M, Potter A, Varro A, Eibach D, Suerbaum S, Wang TC, Fox JG (2011) Lack of commensal flora in *Helicobacter pylori*-infected INS-GAS mice reduces gastritis and delays intraepithelial neoplasia. Gastroenterology 140(1):210–220. https://doi.org/10.1053/j.gastro.2010.09.048

Machado JC, Figueiredo C, Canedo P, Pharoah P, Carvalho R, Nabais S, Castro Alves C, Campos ML, Van Doorn LJ, Caldas C, Seruca R, Carneiro F, Sobrinho-Simoes M (2003) A proinflammatory genetic profile increases the risk for chronic atrophic gastritis and gastric carcinoma. Gastroenterology 125 (2):364–371

Maldonado-Contreras A, Goldfarb KC, Godoy-Vitorino F, Karaoz U, Contreras M, Blaser MJ, Brodie EL, Dominguez-Bello MG (2011) Structure of the human gastric bacterial community in relation to *Helicobacter pylori* status. ISME J 5(4):574–579. https://doi.org/10.1038/ismej.2010.149

Martin ME, Bhatnagar S, George MD, Paster BJ, Canfield DR, Eisen JA, Solnick JV (2013) The impact of *Helicobacter pylori* infection on the gastric microbiota of the rhesus macaque. PLoS One 8(10):e76375. https://doi.org/10.1371/journal.pone.0076375

Martinsen TC, Bergh K, Waldum HL (2005) Gastric juice: a barrier against infectious diseases. Basic Clin Pharmacol Toxicol 96(2):94–102. https://doi.org/10.1111/j.1742-7843.2005.pto960202.x

Mera RM, Bravo LE, Camargo MC, Bravo JC, Delgado AG, Romero-Gallo J, Yepez MC, Realpe JL, Schneider BG, Morgan DR, Peek RM Jr, Correa P, Wilson KT, Piazuelo MB (2018) Dynamics of *Helicobacter pylori* infection as a determinant of progression of gastric precancerous lesions: 16-year follow-up of an eradication trial. Gut 67(7):1239–1246. https://doi.org/10.1136/gutjnl-2016-311685

Monstein HJ, Tiveljung A, Kraft CH, Borch K, Jonasson J (2000) Profiling of bacterial flora in gastric biopsies from patients with *Helicobacter pylori*-associated gastritis and histologically normal control individuals by temperature gradient gel electrophoresis and 16S rDNA sequence analysis. J Med Microbiol 49 (9):817–822. https://doi.org/10.1099/0022-1317-49-9-817

Mowat C, Williams C, Gillen D, Hossack M, Gilmour D, Carswell A, Wirz A, Preston T, McColl KE (2000) Omeprazole, *Helicobacter pylori* status, and alterations in the intragastric milieu facilitating bacterial N-nitrosation. Gastroenterology 119(2):339–347

Ohnishi N, Yuasa H, Tanaka S, Sawa H, Miura M, Matsui A, Higashi H, Musashi M, Iwabuchi K, Suzuki M, Yamada G, Azuma T, Hatakeyama M (2008) Transgenic expression of *Helicobacter pylori* CagA induces gastrointestinal and hematopoietic neoplasms in mouse. Proc Natl Acad Sci U S A 105 (3):1003–1008. https://doi.org/10.1073/pnas.0711183105

Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F (2015) Familial gastric cancer: genetic susceptibility, pathology, and implications for management. Lancet Oncol 16(2):e60–e70. https://doi.org/10.1016/S1470-2045(14)71016-2

Ottini L, Falchetti M, Lupi R, Rizzolo P, Agnese V, Colucci G, Bazan V, Russo A (2006) Patterns of genomic instability in gastric cancer: clinical implications and perspectives. Ann Oncol 17(Suppl 7):vii97–vi102. https://doi.org/10.1093/annonc/mdl960

Paroni Sterbini F, Palladini A, Masucci L, Cannistraci CV, Pastorino R, Ianiro G, Bugli F, Martini C, Ricciardi W, Gasbarrini A, Sanguinetti M, Cammarota G, Posteraro B (2016) Effects of proton pump inhibitors on the gastric mucosa-associated microbiota in dyspeptic patients. Appl Environ Microbiol 82(22):6633–6644. https://doi.org/10.1128/AEM.01437-16

Parsons BN, Ijaz UZ, D'Amore R, Burkitt MD, Eccles R, Lenzi L, Duckworth CA, Moore AR, Tiszlavicz L, Varro A, Hall N, Pritchard DM (2017) Comparison of the human gastric microbiota in hypochlorhydric states arising as a result of *Helicobacter pylori*-induced atrophic gastritis, autoimmune atrophic gastritis and proton pump inhibitor use. PLoS Pathog 13(11): e1006653. https://doi.org/10.1371/journal.ppat.1006653

Persson C, Canedo P, Machado JC, El-Omar EM, Forman D (2011) Polymorphisms in inflammatory response genes and their association with gastric cancer: a HuGE systematic review and meta-analyses. Am J Epidemiol 173(3):259–270. https://doi.org/10.1093/aje/kwq370

Plottel CS, Blaser MJ (2011) Microbiome and malignancy. Cell Host Microbe 10(4):324–335. https://doi.org/10.1016/j.chom.2011.10.003

Plummer M, van Doorn LJ, Franceschi S, Kleter B, Canzian F, Vivas J, Lopez G, Colin D, Munoz N, Kato I (2007) *Helicobacter pylori* cytotoxin-associated genotype and gastric precancerous lesions. J Natl Cancer Inst 99(17):1328–1334. https://doi.org/10.1093/jnci/djm120

Plummer M, Franceschi S, Vignat J, Forman D, de Martel C (2015) Global burden of gastric cancer attributable to *Helicobacter pylori*. Int J Cancer 136(2):487–490. https://doi.org/10.1002/ijc.28999

Praud D, Rota M, Pelucchi C, Bertuccio P, Rosso T, Galeone C, Zhang ZF, Matsuo K, Ito H, Hu J, Johnson KC, Yu GP, Palli D, Ferraroni M, Muscat J, Lunet N, Peleteiro B, Malekzadeh R, Ye W, Song H, Zaridze D, Maximovitch D, Aragones N, Castano-Vinyals G, Vioque J, Navarrete-Munoz EM, Pakseresht M, Pourfarzi F, Wolk A, Orsini N, Bellavia A, Hakansson N, Mu L, Pastorino R, Kurtz RC, Derakhshan MH, Lagiou A, Lagiou P, Boffetta P, Boccia S, Negri E, La Vecchia C (2018) Cigarette smoking and gastric cancer in the Stomach Cancer Pooling (StoP) project. Eur J Cancer Prev 27 (2):124–133. https://doi.org/10.1097/CEJ.0000000000000290

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Meta HITC, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65. https://doi.org/10.1038/nature08821

Rajilic-Stojanovic M, de Vos WM (2014) The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 38(5):996–1047. https://doi.org/10.1111/1574-6976.12075

Rota M, Pelucchi C, Bertuccio P, Matsuo K, Zhang ZF, Ito H, Hu J, Johnson KC, Palli D, Ferraroni M, Yu GP, Muscat J, Lunet N, Peleteiro B, Ye W, Song H, Zaridze D, Maximovitch D, Guevara M, Fernandez-Villa T, Vioque J, Navarrete-Munoz EM, Wolk A, Orsini N, Bellavia A, Hakansson N, Mu L, Persiani R, Kurtz RC, Lagiou A, Lagiou P, Galeone C, Bonzi R, Boffetta P, Boccia S, Negri E, La Vecchia C (2017) Alcohol consumption and gastric cancer risk-a pooled analysis within the StoP project consortium. Int J Cancer 141(10):1950–1962. https://doi.org/10.1002/ijc.30891

Sanduleanu S, Jonkers D, De Bruine A, Hameeteman W, Stockbrugger RW (2001) Non-*Helicobacter pylori* bacterial flora during acid-suppressive therapy: differential findings in gastric juice and gastric mucosa. Aliment Pharmacol Ther 15(3):379–388

Schulz C, Schutte K, Koch N, Vilchez-Vargas R, Wos-Oxley ML, Oxley APA, Vital M, Malfertheiner P, Pieper DH (2018) The active bacterial assemblages of the upper GI tract in individuals with and without *Helicobacter* infection. Gut 67 (2):216–225. https://doi.org/10.1136/gutjnl-2016-312904

Sender R, Fuchs S, Milo R (2016) Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell 164(3):337–340. https://doi.org/10.1016/j.cell.2016.01.013

Sjostedt S, Heimdahl A, Kager L, Nord CE (1985) Microbial colonization of the oropharynx, esophagus and stomach in patients with gastric diseases. Eur J Clin Microbiol 4(1):49–51

Spoto CPE, Gullo I, Carneiro F, Montgomery EA, Brosens LAA (2018) Hereditary gastrointestinal carcinomas and their precursors: An algorithm for genetic testing. Semin Diagn Pathol 35(3):170–183. https://doi.org/10.1053/j.semdp.2018.01.004

Stearns JC, Lynch MD, Senadheera DB, Tenenbaum HC, Goldberg MB, Cvitkovitch DG, Croitoru K, Moreno-Hagelsieb G, Neufeld JD (2011) Bacterial

biogeography of the human digestive tract. Sci Rep 1:170. https://doi.org/10.1038/srep00170

Stockbruegger RW (1985) Bacterial overgrowth as a consequence of reduced gastric acidity. Scand J Gastroenterol Suppl 111:7–16

Thorell K, Bengtsson-Palme J, Liu OH, Palacios Gonzales RV, Nookaew I, Rabeneck L, Paszat L, Graham DY, Nielsen J, Lundin SB, Sjoling A (2017) In vivo analysis of the viable microbiota and *Helicobacter pylori* transcriptome in gastric infection and early stages of carcinogenesis. Infect Immun 85(10). https://doi.org/10.1128/IAI.00031-17

Thorens J, Froehlich F, Schwizer W, Saraga E, Bille J, Gyr K, Duroux P, Nicolet M, Pignatelli B, Blum AL, Gonvers JJ, Fried M (1996) Bacterial overgrowth during treatment with omeprazole compared with cimetidine: a prospective randomised double blind study. Gut 39(1):54–59

Tseng CH, Lin JT, Ho HJ, Lai ZL, Wang CB, Tang SL, Wu CY (2016) Gastric microbiota and predicted gene functions are altered after subtotal gastrectomy in patients with gastric cancer. Sci Rep 6:20701. https://doi.org/10.1038/srep20701

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. Nature 449(7164):804–810. https://doi.org/10.1038/nature06244

Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H (2016) Gastric cancer. Lancet 388 (10060):2654–2664. https://doi.org/10.1016/S0140-6736(16)30354-3

Warren JR, Marshall B (1983) Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet 1(8336):1273–1275

Wong BC, Lam SK, Wong WM, Chen JS, Zheng TT, Feng RE, Lai KC, Hu WH, Yuen ST, Leung SY, Fong DY, Ho J, Ching CK, Chen JS, China Gastric Cancer Study G (2004) *Helicobacter pylori* eradication to prevent gastric cancer in a high-risk region of China: a randomized controlled trial. JAMA 291(2):187–194. https://doi.org/10.1001/jama.291.2.187

Worthley DL, Phillips KD, Wayte N, Schrader KA, Healey S, Kaurah P, Shulkes A, Grimpen F, Clouston A, Moore D, Cullen D, Ormonde D, Mounkley D, Wen X, Lindor N, Carneiro F, Huntsman DG, Chenevix-Trench G, Suthers GK (2012) Gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS): a new autosomal dominant syndrome. Gut 61(5):774–779. https://doi.org/10.1136/gutjnl-2011-300348

Yang I, Woltemate S, Piazuelo MB, Bravo LE, Yepez MC, Romero-Gallo J, Delgado AG, Wilson KT, Peek RM, Correa P, Josenhans C, Fox JG, Suerbaum S (2016) Different gastric microbiota compositions in two human populations with high and low gastric cancer risk in Colombia. Sci Rep 6:18594. https://doi.org/10.1038/srep18594

Yu G, Torres J, Hu N, Medrano-Guzman R, Herrera-Goepfert R, Humphrys MS, Wang L, Wang C, Ding T, Ravel J, Taylor PR, Abnet CC, Goldstein AM (2017) Molecular characterization of the human stomach microbiota in gastric cancer patients. Front Cell Infect Microbiol 7:302. https://doi.org/10.3389/fcimb.2017.00302

Zamani M, Ebrahimtabar F, Zamani V, Miller WH, Alizadeh-Navaei R, Shokri-Shirvani J, Derakhshan MH (2018) Systematic review with meta-analysis: the worldwide prevalence of *Helicobacter pylori* infection. Aliment Pharmacol Ther 47(7):868–876. https://doi.org/10.1111/apt.14561

Zilberstein B, Quintanilha AG, Santos MA, Pajecki D, Moura EG, Alves PR, Maluf Filho F, de Souza JA, Gama-Rodrigues J (2007) Digestive tract microbiota in healthy volunteers. Clinics (Sao Paulo) 62(1):47–54. https://doi.org/10.1590/S1807-59322007000100008

Paper IV

Gastric cancer: Basic aspects.

Molina-Castro S*, **Pereira-Marques J***, Figueiredo C, Machado JC, Varon C

Helicobacter. 2017 Sep;22 Suppl 1

*Equal contribution.

DOI: 10.1111/hel.12412

REVIEW ARTICLE

WILEY Helicobacter

Gastric cancer: Basic aspects

Silvia Molina-Castro[1,2,*] | Joana Pereira-Marques[3,4,5,*] | Ceu Figueiredo[3,4,6] | Jose C. Machado[3,4,6] | Christine Varon[1]

[1]INSERM, UMR1053 Bordeaux Research in Translational Oncology, BaRITOn, University of Bordeaux, Bordeaux, France

[2]University of Costa Rica, San José, Costa Rica

[3]i3S - Instituto de Investigação e Inovação em Saúde (Institute of Research and Innovation in Health), University of Porto, Porto, Portugal

[4]Ipatimup - Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

[5]ICBAS - Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal

[6]Faculty of Medicine of the University of Porto, Porto, Portugal

*Molina-Castro and Pereira-Marques are contributed equally to the work.

Correspondence
Christine Varon, INSERM, UMR1053 Bordeaux Research in Translational Oncology, BaRITOn, University of Bordeaux, Bordeaux, France.
Email: christine.varon@u-bordeaux.fr

Abstract

Gastric cancer is one of the most incident and deadliest malignancies in the world. Gastric cancer is a heterogeneous disease and the end point of a long and multistep process, which results from the stepwise accumulation of numerous (epi)genetic alterations, leading to dysregulation of oncogenic and tumor suppressor pathways. Gastric cancer stem cells have emerged as fundamental players in cancer development and as contributors to gastric cancer heterogeneity. For this special issue, we will report last year's update on the gastric cancer molecular classification, and in particular address the gastric cancer groups who could benefit from immune checkpoint therapy. We will also review the latest advances on gastric cancer stem cells, their properties as gastric cancer markers and therapeutic targets, and associated signaling pathways. The understanding of the molecular basis underlying gastric cancer heterogeneity and of the role played by gastric cancer stem cells in cancer development and heterogeneity is of major significance, not only for identifying novel targets for cancer prevention and treatment, but also for clinical management and patient stratification for targeted therapies.

KEYWORDS

gastric cancer stem cells, immune checkpoint inhibition, marker, microRNAs, molecular classification

1 | MOLECULAR CLASSIFICATION OF GASTRIC CANCER

Gastric cancer (GC) is a complex heterogeneous disease. Recent comprehensive molecular analyzes have confirmed this heterogeneity and led to the emergence of several molecular classification schemes.[1-4] The molecular classification schemes and the (epi)genetics of GC have been recently reviewed.[5-7] While these classification systems are based on high-throughput genome-, transcriptome-, and proteome-wide approaches, Setia et al.[8] tested the validity of these classifications using in situ hybridization (ISH) and immunohistochemistry (IHC), which are widely available techniques in routine diagnostic practice. They have analyzed the expression of 14 different biomarkers, including EBV EBER, mismatch repair (MMR) proteins, p53, and E-cadherin, in a series of 146 GC. Using unsupervised hierarchical clustering analysis, they were able to identify five GC groups: EBV-positive (EBVaGC; 5%), MMR-deficient (16%), aberrant E-cadherin expression (21%), aberrant p53 expression (51%), and normal p53 expression (7%). EBVaCG had prominent lymphoid infiltrates, show a trend to better survival, and had a strong association with membranous expression of PD-L1. MMR-deficient GC had a predominant pattern of MLH1 and PMS2 loss, lower frequency of lymph node metastasis, and a trend to better survival. Aberrant E-cadherin GCs were, as expected, predominantly of the diffuse type, and comparable with the genomically stable group reported by the TCGA, the microsatellite stable (MSS)/EMT subtype of the Asian Cancer Research Group (ACRG), and the mesenchymal group reported by Lei et al.[2-4] The majority of the cases were tumors with aberrant p53 expression, which were comparable with the chromosomal instability group of the TCGA, the MSS/TP53- type of the ACRG, and the proliferative subtype of Lei et al.[2-4] Tumors in this group were predominantly of the intestinal type and had most frequently higher lymph node stage. The GC group with normal p53

expression were most frequently intestinal type and had increased MUC6 expression, and authors suggest that these features likely correspond to the metabolic subtype reported by Lei et al.[2]

Kim et al.[9] undertook a similar approach to analyze the expression of 10 molecular markers in a consecutive series of 438 advanced (metastatic) GCs from Korea. In addition to EBV EBER, MMR proteins, and p53, they have also analyzed the expression of receptor tyrosine kinases HER2, EGFR, and MET. Approximately 3.3% of the cases were EBVaGC, and 4.8% were MMR-deficient GCs. Overall RTK overexpression was found in 50% of the cases, being EGFR the most frequently overexpressed (40%), followed by HER2 (13.5%) and MET (12%), with over 10% GC simultaneously overexpressing two RTKs. Although the ISH/IHC results of Setia et al. and Kim et al. were, to some extent, consistent with previous comprehensive molecular studies, lower proportions of EBVaGC and MMR-deficient GC were detected in comparison with the TCGA and ACRG series.[3,4] Whether related to patient geographic origin and/or disease staging, these proposals need further validation to be universally relevant.

1.1 | Immune checkpoint inhibition in gastric cancer

The immune checkpoint programmed cell death (PD-1) is expressed by T cells, and their ligands PD-L1/2 are expressed by tumor cells or stromal immune cells. This pathway is involved in the regulation of T-cell activation and tolerance in response to antigenic stimulation. Tumor cells take advantage of this pathway to evade immune detection, by suppressing effector T-cell function. Blockade of this pathway with antibodies to PD-1 or its ligands has been recently explored with success in different types of cancer.[10-13]

In GC, the phase I trial of the anti-PD-1 pembrolizumab in a mostly pretreated population of patients with PD-L1-positive advanced cancer, showed a manageable toxicity profile and sustained antitumor responses in 22% of patients.[14] Interestingly, results pointed to a higher proportion of partial responses in patients with tumors with higher mononuclear inflammatory cell densities (score 3: 4/9, 44% vs score <2: 4/26, 15%). This was followed by a series of studies published over the last year addressing which GC groups could benefit from immune checkpoint therapy and which biomarkers are best suited to guide the use of PD-1/PD-L1 inhibitors.

Using a large cohort of 487 Japanese patients with GC, Kawazoe et al.[15] observed PD-L1 expression in tumor cells (23%) and in tumor infiltrating lymphocytes (TILs, 61%). PD-L1 expression was associated with older and male patients, and with poorly differentiated solid-type tumors, but not with patient prognosis. In EBVaGC (5% in this series), PD-L1 expression in tumor cells was more frequent and higher than that of EBV-negative GC (EBVnGC). PD-L1 expression in TILs was also more frequent in EBVaGC.

In agreement with the above, Derks et al.[16] observed, in a series of 81 GCs, that EBVaGCs (n=32) had PD-L1 IHC expression in tumor cells (50%) and in immune cells (94%), validating their first analysis of 12 EBVaGCs and 10 EBVnGCs from the TCGA study.

Similarly, Saito et al.[17] analyzed the TCGA database and showed that EBVaGC express higher levels of PD-L1 mRNA than the other TCGA GC subtypes. Further IHC analysis of a large series of 232 GCs from Japan showed that EBVaGC (42% of the series) had higher PD-L1 expression in tumor cells (34%) and in immune cells from the tumor stroma (45%) than non-EBVaGC. In EBVaGC, PD-L1 expression in tumor and immune cells was associated with the diffuse histologic type, tumor invasion depth, and poor patient prognosis.[17] Also corroborating the TCGA results,[3] 11% of EBVaGC had *PD-L1* gene amplification.

Kawazoe et al.[15] reported higher frequency and levels of PD-L1 expression, both in tumor cells and in TILs, in MMR-deficient GC than in MMR-proficient GC. Accordingly, Derks et al.[16] showed that among EBVnGCs, MSI GCs (n=15) had higher frequency of PD-L1 expression in tumor (33%) and immune cells (46%) than MSS GCs, which had no PD-L1 expression in tumor cells and PD-L1 expression in immune cells in only 35% of cases. Also consistent with these observations, are those of Ma et al., who reported significantly higher rates of PD-L1 expression in MSI GC and in EBVaGC in comparison with EBVnGC/MSS GC.[18] In their series of 44 GCs (7 EBVaGC, 16 MSI, and 21 EBVnGC/MSS), PD-L1 expression was not associated with tumor invasion depth invasion or nodal metastasis and was not an independent predictor of prognosis.

In a setting of 78 MSI-H/EBVnGC selected from different study cohorts, Cho et al.[19] found an overall PD-L1 expression of 62%, detected in tumor cells (9%) and in immune cells (60%). Compared to MSI-H GC without PD-L1 expression, immune cell PD-L1 expression was associated with intestinal-type GC, lower risk of lymph node metastasis, lower tumor stages, and independently associated with longer patient survival.

Tumor adaptive immune resistance is the mechanism of PD-L1 upregulation by tumor cells in response to interferon (IFN)-γ secreting $CD8^+$ T cells. In other cancer models, the efficacy of the anti-PD-1 treatment was associated with increased number of $CD8^+$ T cell and increased expression of PD-1/PD-L1, associated with an IFN-γ signature.[11,20,21] In this sense, several groups also addressed the density, type, and signatures of TILs in GC.

Li et al.[22] compared 44 GCs with massive lymphocyte infiltration (>80% TILs) with 93 GCs with relatively low TILs (<10%) and showed that higher density of TILs was significantly associated with increased PD-L1 expression levels.

Thompson et al.[23] showed PD-L1 expression in 12% GC tumor cells and 44% immune stromal cells, across all stages, in a sample comprising 29 GCs. They were the first to demonstrate that in comparison with PD-L1 negative cases, the density of $CD8^+$ T-cell infiltration was 16-fold higher within PD-L1-positive tumors and fourfold higher in the peri-tumoral/tumor-stroma interface regions. Increasing $CD8^+$ T-cell densities both within tumors and in the immune stroma were correlated with increasing levels of tumor and stromal PD-L1 expressions. Both PD-L1 expression and $CD8^+$ T-cell infiltration in the tumor and in the immune stroma were associated with worse patient survival. Kawazoe et al.[15] also confirmed strong association between expression of PD-L1 and high densities of $CD8^+$, $CD3^+$, and forkhead box P3 $(FOXP3)^+$ TILs.

Koh et al.[24] also addressed PD-L1 expression and $CD8^+$ TILs in 392 stage II/III GCs from Korea. The majority of EBVaGC (92%) and

MSI-H GC (67%) TILs were PD-L1^{+}/CD8^{High}, and analyzes of the TCGA and ACRG datasets validated that the EBVaGC and MSI-H CG in both were likely to be PD-L1^{+}/CD8^{High}.[3,4] PD-L1^{-}/CD8^{High} cases had the best overall survival, followed by PD-L1^{+}/CD8^{High}. CD8^{High} TILs were significantly associated overall survival in a multivariate analysis. Accordingly, Ma et al.[18] also demonstrated that PD-L1^{+} EBVaGC and PD-L1^{+} MSI GC had significantly more CD8^{+} T cells at the tumor invasive front than PD-L1^{+}EBVnGC/MSS GC.

Derks et al.[16] also investigated whether PD-L1 expression in EBVaGC was associated with a specific immune signature. Indeed, they found IFN-γ signaling genes CXCL9, CXCL10, CXCL11, IDO, and GZMB among the top 40 most enriched genes that discriminate EBVaGC. Additional analyzes of 214 individual GCs from the TCGA study were performed to evaluate the strength of the IFN-γ gene signature between different molecular GC subtypes. Interestingly, MSI GCs also had high IFN-γ response gene expression, unlike GS and CIN GCs.

Overall, these findings sustain the close relationship between PD-L1, CD8 status, and IFN-γ signatures and suggest that EBVaGCs and MSI are the GC groups with increased likelihood of benefiting from therapeutic PD-1/PD-L1 immune blockade.

Su et al.[25] used a genome editing approach based on CRISPR-Cas9 technology to disrupt PD-1 on primary T cells. PBMCs from healthy donors or EBVaGC patients were used to generate ΔPD-1 cytotoxic T cells (CTLs), upon stimulation with autologous dendritic cells loaded with an immunogenic EBV LMP2A epitope. ΔPD-1 LMP2A-specific CTLs showed enhanced immune response and cytotoxicity toward an EBV-positive GC cell line. The authors also demonstrated, in a xenograft mouse model of EBVaGCa, an antitumor effect of ΔPD-1 LMP2A-specific CTLs in combination with low dose radiotherapy. These results pave the way for the use of gene-editing strategies for cancer immunotherapy.

2 | GASTRIC CANCER STEM CELLS

2.1 | Unified model for tumor heterogeneity in gastric cancer

Cancer is now considered a stem cell disease. Tumor cells display significant functional and morphologic heterogeneity, which have been explained by two models: (1) the clonal evolution or stochastic model, in which only a small number of cells within a tumor can initiate and sustain tumor growth, but all the tumor cells are biologically homogeneous and could potentially be responsible for these processes, given that they undergo the stochastic events that govern tumor-initiating capacity; (2) the hierarchical or cancer stem cells (CSC) model, which is the most accepted and proposes that only a specific subset of tumor cells possesses the capacity to regenerate the tumor.[26] CSCs are the cells within a tumor with the capacity to initiate tumors and sustain their growth, thanks to their ability of self-renewal and asymmetric division, the potential to develop into any type of cell in the tumor, and proliferative capacity to drive the continued expansion of the tumor population.

Kreso and Dick proposed in 2009 a comprehensive model that integrates both models of heterogeneity, in which favorable mutations accumulate in the rare CSC population within an early tumor as it develops, generating subclones in the CSCs, some of which achieve higher capacity for self-renewal.[27] This model considers that tumor-initiating cells are not static entities but ones that can evolve, and this evolution process could be driven by genetic determinants, epigenetics, gene expression stochasticity, the CSC niche, and the tumor microenvironment.

Song et al.[28] recently proposed a unified model for GC, in which primitive GCSCs continually optimize their genome, improve the pattern of gene expression, and generate various subclones to adapt to the changing tumor microenvironment. Primitive GCSCs result from the accumulation of cancer-related mutations in stem cells derived from epithelium, mesenchyme, or dedifferentiated mature cells, which gradually acquire the properties of unlimited self-renewal and multilineage differentiation, creating the hierarchical structure of the tumor. GCSCs evolve into different subclones, and the dominant ones generate the respective hierarchical organization, while the less fit disappear. Multiple factors such as the tumor microenvironment, including the CSC niche, epigenetic alterations, and dysregulations of signaling pathways are proposed to have a role driving the evolution of the GCSCs. *Helicobacter pylori* infection undoubtedly influences several, if not all, of these aspects. For instance, Yong et al.,[29] confirmed the results of other studies showing that infection with *H. pylori* CagA^{+} strains induces an increase in the expression of CSC-markers (CD44, Lgr5), CSC properties in AGS and MKN45 gastric epithelial cells (GECs), and upregulates the stemness transcription factors Nanog and Oct4. The observed augmentation depends on *H. pylori*-activated Wnt/β-catenin signaling pathway. β-catenin was accumulated in the nucleus where it binds to the promoters of Nanog and Oct4. These results where further confirmed in an IHC evaluation that showed increased levels of Nanog and Oct4 in GC tissues of patients positive for *H. pylori* CagA^{+} strains in comparison with those infected with CagA^{-} strains.

Another important factor is the tumor microenvironment. Recently, the role of murine bone marrow-derived fibroblasts (mBMFs) was described as one of the microenvironment cues promoting CSC properties. Zhu et al.[30] showed that coculture of mBMFs with both murine and human GECs induced an EMT-like process, increased CD44 expression, formation of tumorspheres in vitro, and tumorigenesis in vivo in a xenograft model. These effects were due to the activation of mBMFs by cancer cells which stimulates the production of mIL-6 that activates the JAK2/STAT3 pathway resulting in the production of mHGF. Treatment with recombinant human IL-6 and HGF also induced tumorigenesis and CSC properties in MKN28 cells. HGF induced p-Met and p-STAT3 in human GEC lines and the consequent increase in the expression of TGF-β1, which in turn stimulate the IL-6/HGF production in mBMFs. In GC tissues evaluated by IHC, there was a positive correlation between IL-6 and TGF-β1 and between HGF and TGF- β1, which were also found in The Cancer Genome Atlas (TCGA) database. In an independent study, TGF-β1 signaling was identified as an inducer of the gastric cancer-associated fibroblasts (CAFs) ability to stimulate the invasive properties of GC cells.[31] CAFs isolated from patients with

GC increased their motility upon stimulation with TGF-β1 to a major extent in comparison with normal fibroblasts isolated from adjacent tissues. When GC cell lines were cocultured with TGF-β1-stimulated CAFs, they acquired increased invasive properties in vitro and in vivo.

Bone marrow-derived mesenchymal stem cells (BM-MSC) are recruited to primary tumor sites and contribute to the development of tumor niches for CSC and establish cytokine networks to promote tumor growth.[32,33] IL-17B has been described to be increased in GC tissues and in the serum of patients with GC. Lately, a priming effect of IL-17B in human umbilical cord and GC-derived MSC has been shown. IL-17B activated stemness-related pathways (STAT3, NF-κB, and Wnt/β-catenin) in MSC and their capacity to stimulate proliferation and migration in gastric epithelial cells through the production of soluble factors such as IL-6, IL-8, TGF-β, and CCL5.[34] *Helicobacter pylori* induces migration of BM-MSC toward infection site. The capacity of *H. pylori* to induce the expression of p53, bcl-2, and MMP2 in human adipocyte-derived MSC (hA-MSC) was recently described.[35] They also increase cancer cell resistance to chemotherapy. A study found that BM-MSCs enhance the anti-apoptotic abilities of $CD133^+$ CSCs in GC via the PI3K/Akt pathway, enhancing their resistance to cisplatin.[36] Cisplatin-resistance cell subpopulations were enriched in $CD133^+$ cells, which exhibited a higher IC50 than $CD133^-$. Using culture in Transwells, the authors showed that BM-MSCs increased the viability and the IC50 values of $CD133^+$ cells, while decreasing their apoptotic rate in response to cisplatin with an upregulation of Bcl-2 and a downregulation of Bax. P-AKT levels were increased as well, indicating an activation of the PI3K/Akt pathway. These properties were directly associated to CD133, as it exogenous overexpression enhanced the properties and knock down abrogated them. Furthermore, the inhibition of the PI3K-Akt pathway in $CD133^+$ cells abrogated their chemoresistance and anti-apoptotic capability. The results were confirmed in an in vivo model in which the injection of $CD133^+$ cells together with BM-MSC enhanced tumor growth and chemoresistance in comparison with the $CD133^+$ cells alone.

Nerves are another component of the SC niche which has been recently addressed in the context of gastric carcinogenesis. The Acetylcholine (Ach)-Nerve growth factor (NGF)-Muscarinic cholinergic receptor (M3R)-Yes-associated protein (YAP) axis has been identified to be central to GC biology.[37] It was showed that Dclk1 identifies the ACh producing cells in the gastrointestinal tract, including the stromal cholinergic nerves and the Tuft cells, both of which participate in the cholinergic signaling required for different stages of gastric carcinogenesis. ACh induces NGF production in GEC, which stimulates axonogenesis helping the innervation described in tumors and the proliferation of GEC (regulated by M3R). The activation of this axis was demonstrated to be sufficient for initiation of GC. Tff2-Cre; R26-NGF mice develop spontaneous metaplasia and dysplasia at 8 months, and by 18 months, gastric tumors with intramucosal adenocarcinoma, and accelerated its induction by N-methyl-N-nitrosourea (MNU)-treatment. Interestingly, M3R activation decreased YAP phosphorylation thereby increasing YAP-transcriptional activity, and the IHC evaluation of GC demonstrated correlation between YAP and NGF, which also correlated with more advanced stages.

2.2 | Markers

Markers for the accurate identification and isolation of GCSCs are vital for the CSC research field. Nguyen et al.[38] evaluated the expression of putative markers of GCSCs in primary tumors and patient-derived xenografts in immunodeficient mice, and tested the tumorigenic properties of the cells in vitro and in vivo. This study identified CD44 and ALDH as the most relevant and specific markers of chemoresistant GCSCs, along with CD133, EpCAM, and CD166. Another study by Chen et al.[39] identified by IHC Oct4, CD133, and EpCAM as markers of tumor progression. High CD133 expression was confirmed as marker of worst prognosis by qPCR. Lgr5 has been known as a marker of gastric stem cells, but it has been recently studied as a marker of GCSCs by IHC analysis of normal gastric mucosa, gastric dysplastic mucosa and early gastric tumors.[40] Furthermore, a specific isoform of ALDH, the ALDH-3A1, was identified as responsible for the Aldefluor activity of the cell lines MKN-45 and SGC-7901, and the presence of this isoform correlated with stem cell-properties in vitro and with dysplasia, lymph node metastasis and cancer stage in gastric tumors samples from patients.[41] Lgr5 expression was associated with the development of intestinal metaplasia in the body of the stomach. This work also showed one variant of CD44, the CD44v8-10 as a marker of GCSC. Experimentally, the tumor-initiating capacity $Lgr5^+$ cells were demonstrated upon deletion of Smad4 and PTEN, which promoted the invasive intestinal-type cancer in a murine model, while the same deletion in $Lgr5^-$ cells failed to do so.[42] Novel markers have also been identified, as is the case of the long noncoding RNA (lncRNA) ROR, which has been found highly expressed in $CD133^+$ GCSCs.[43] lncRNA ROR upregulation upregulated stemness factors such as OCT4, SOX2, and NANOG.

2.3 | Strategies for targeting gastric cancer stem cells

Given the importance of GCSCs in all the stages of GC, and the lack of specific, effective treatment for the disease, many research efforts have been put in developing strategies to target GCSCs. Recently, the inhibitory effect of a synthetic analog of genistein on the stemness properties, SC-markers (CD44, CD133, and ALDH) and EMT-like process in SGC-7901 GECs was described, and its effect was synergic with the inhibition of the transcription factor FoxM1, causing a decreased of the EMT transcription factor Twist1.[44] A Notch signaling pathway inhibitor, the gamma-secretase inhibitor IX (GSI) was also showed to have an inhibitory effect on cell proliferation, invasion and migration, and tumorsphere formation capacities of CD44+ GCSCs.[45]

Some drugs already used to treat other diseases have also been tested successfully for their inhibitory action on GCSCs. The proton-pump inhibitor pantoprazole reduced the expression of CSC-markers and SC-properties of GC cells, concomitantly increasing their susceptibility to 5-fluorouracil (5-FU).[46] All-trans-retinoic acid (tretinoin) exhibited the capacity of decreasing proliferation and tumorigenic capacity on GECs in vitro and in a mouse xenograft model, in association with a cell cycle arrest.[47] Sulfasalazine is an inhibitor of xCT (subunit

of the cysteine-glutamate transporter), which is used for treatment of Crohn's disease and has been described for having a suppressive activity on CD44v$^+$ cells in vitro and in vivo. Shitara et al.[48] recently published a study where sulfasalazine was administered orally to patients with GC to determine the effective dose. Half of the patients with CD44v$^+$ cells in their pretreatment biopsies reduced the level of this marker after treatment and the effective dose was determined at 8 g/d. The mechanisms of action of sulfasalazine are probably by decreasing the intracellular levels of glutathione and rendering the cell less resistant to oxidative stress. Glutathione was also reduced in the biopsies of two patients after treatment. The downside of this study is that it was performed in only eleven patients, but the results are promising.

2.4 | Role of microRNAs

While it is largely know that microRNAs play a role in gastric tumorigenesis and CSCs, new functions continued to be identified for specific microRNAs and the way they regulated cancer-related signaling pathways. Fan et al.[49] recently described an upregulation of miR-501-5p in GC cell lines and specimens from patients with GC, functioning as an activator of the Wnt/β-catenin signaling pathway and enhancing the CSC-phenotype. MiR-132 has been also found to be upregulated in Lgr5$^+$ GCSC-like cells in GC specimens, in correlation with chemoresistance and worse survival.[50] In vitro, miR-132 promoted resistance to cisplatin through the inhibition of SIRT, which results in enhanced expression of the transporter ABCG2, previously identified in multidrug resistance.[51] A microRNA expression profile of CD44$^+$ cells revealed a significant upregulation of miR-196a-5p in this cell population in comparison with the Cd44$^-$, and its suppression decreased the functional CSC properties of the CD44$^+$ cells by decreasing the expression of Smad4, which correlated with differentiation, metastatic properties and invasiveness of the tumors.[52] A regulation loop has been described between Bmi-1, miR-21, and miR-34a.[53] By upregulating miR-21, Bmi-1 enhances the CSC properties in GC, while miR-34a (also induced by Bmi-1), downregulated the CSC-features by inhibiting Bmi-1. Finally, the stress-associated hormone isoprotenerol was shown to induce CD44 expression and chemoresistance in GECs, but this effect was reversed by the overexpression of miR-373.[54] Isoprotenerol inhibited miR-373 by triggering the STAT3 pathway via β-2 adrenergic receptors. The authors also found that CD44 and miR-373 were inversely correlated in GC specimens and in the TCGA database. Interestingly, *H. pylori* infection inhibits miR-373 in AGS cells,[55] and this study unveils one of the mechanisms involved in the role of miR-373 in GCSC-emergence.

2.5 | Advances in signaling pathways

The Wnt/β-catenin signaling pathway role has been widely described in CSC-biology; however, new insights into its regulation have been studied recently. Zhou et al.[56] showed that the elevated expression of aquaporin 3 (AQP3) was associated with CD44 expression and correlated with more advanced clinical features in GC. In vitro and in vivo, the overexpression of AQP3 induced CSC properties in GECs, via the activation of the β-catenin signaling. Increased nuclear expression of cell cycle and apoptosis regulator 1 (CCAR1), a transcriptional co-activator of the Wnt/β-catenin pathway, has also been reported in correlation with a higher occurrence of GC, and its suppression in vitro lead to decreased CSC-related capacities in AGS and MKN28 cell lines. These results show that CCAR1 plays a role in gastric carcinogenesis and is a potential biomarker.[57] Furthermore, an upregulation of PMP22, a Wnt/β-catenin target gene, was found in GC cells resistant to cisplatin by Cai et al.,[58] and its inhibition decreased CSC properties both in vitro and in vivo. In this study, the treatment with a PMP22 inhibitor restores the sensibility to cisplatin treatment and PMP22 expression correlated with a worst clinical outcome in patients of GC. EMT-associated signaling pathways are also known to play a role in the emergence of GCSC. Hypoxia, specifically HIF-1α, has been recently described as an inductor of GCSC-like cells through the activation of Snail-1 (a major EMT-related transcription factor).[59]

ACKNOWLEDGEMENTS

JPM has a fellowship from Fundação para a Ciência e a Tecnologia (PD/BD/114014/2015) through Programa Operacional Capital Humano, in the framework of FCT's PhD Programme BiotechHealth (Ref. PD/0016/2012). SMC has a fellowship from the Universidad de Costa Rica and the Ministerio de Ciencia, Tecnología y Telecomunicaciones for the PhD Program in Health and Life Sciences (Université de Bordeaux).

CF and JCM are supported by NORTE-01-0145-FEDER-000029 (NORTE 2020) through ERDF funds.

DISCLOSURE OF INTERESTS

All coauthors declare no conflict of interest.

REFERENCES

1. Tan IB, Ivanova T, Lim KH, et al. Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. *Gastroenterology*. 2011;141:476-485, 85 e1-11.
2. Lei Z, Tan IB, Das K, et al. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology*. 2013;145:554-565.
3. Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*. 2014;513:202-209.
4. Cristescu R, Lee J, Nebozhyn M, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat Med*. 2015;21:449-456.
5. Figueiredo C, Constanza Camargo M, Leite M, Fuentes-Panana EM, Rabkin CS, Machado JC. Pathogenesis of gastric cancer: genetics and molecular classification. *Curr Top Microbiol Immunol*. 2017;400:277-304.
6. Strand MS, Lockhart AC, Fields RC. Genetics of gastric cancer. *Surg Clin North Am*. 2017;97:345-370.
7. Rocken C. Molecular classification of gastric cancer. *Expert Rev Mol Diagn*. 2017;17:293-301.

8. Setia N, Agoston AT, Han HS, et al. A protein and mRNA expression-based classification of gastric cancer. *Mod Pathol*. 2016;29:772-784.
9. Kim HS, Shin SJ, Beom SH, et al. Comprehensive expression profiles of gastric cancer molecular subtypes by immunohistochemistry: implications for individualized therapy. *Oncotarget*. 2016;7:44608-44620.
10. Ansell SM, Lesokhin AM, Borrello I, et al. PD-1 blockade with nivolumab in relapsed or refractory Hodgkin's lymphoma. *N Engl J Med*. 2015;372:311-319.
11. Herbst RS, Soria JC, Kowanetz M, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*. 2014;515:563-567.
12. Topalian SL, Sznol M, McDermott DF, et al. Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J Clin Oncol*. 2014;32:1020-1030.
13. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in tumors with mismatch-repair deficiency. *N Engl J Med*. 2015;372:2509-2520.
14. Muro K, Chung HC, Shankaran V, et al. Pembrolizumab for patients with PD-L1-positive advanced gastric cancer (KEYNOTE-012): a multicentre, open-label, phase 1b trial. *Lancet Oncol*. 2016;17:717-726.
15. Kawazoe A, Kuwata T, Kuboki Y, et al. Clinicopathological features of programmed death ligand 1 expression with tumor-infiltrating lymphocyte, mismatch repair, and Epstein-Barr virus status in a large cohort of gastric cancer patients. *Gastric Cancer*. 2017;20:407-415.
16. Derks S, Liao X, Chiaravalli AM, et al. Abundant PD-L1 expression in Epstein-Barr Virus-infected gastric cancers. *Oncotarget*. 2016;7:32925-32932.
17. Saito R, Abe H, Kunita A, et al. Overexpression and gene amplification of PD-L1 in cancer cells and PD-L1+ immune cells in Epstein-Barr virus-associated gastric cancer: the prognostic implications. *Mod Pathol*. 2017;30:427-439.
18. Ma C, Patel K, Singhi AD, et al. Programmed death-ligand 1 expression is common in gastric cancer associated with Epstein-Barr virus or microsatellite instability. *Am J Surg Pathol*. 2016;40:1496-1506.
19. Cho J, Lee J, Bang H, et al. Programmed cell death-ligand 1 expression predicts survival in patients with gastric carcinoma with microsatellite instability. *Oncotarget*. 2017;8:13320-13328.
20. Prat A, Navarro A, Pare L, et al. Immune-related gene expression profiling after PD-1 blockade in non-small cell lung carcinoma, head and neck squamous cell carcinoma and melanoma. *Cancer Res*. 2017;77:3540-3550.
21. Tumeh PC, Harview CL, Yearley JH, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*. 2014;515:568-571.
22. Li Z, Lai Y, Sun L, Zhang X, et al. PD-L1 expression is associated with massive lymphocyte infiltration and histology in gastric cancer. *Hum Pathol*. 2016;55:182-189.
23. Thompson ED, Zahurak M, Murphy A, et al. Patterns of PD-L1 expression and CD8 T cell infiltration in gastric adenocarcinomas and associated immune stroma. *Gut*. 2017;66:794-801.
24. Koh J, Ock CY, Kim JW, et al. Clinicopathologic implications of immune classification by PD-L1 expression and CD8-positive tumor-infiltrating lymphocytes in stage II and III gastric cancer patients. *Oncotarget*. 2017;8:26356-26367.
25. Su S, Zou Z, Chen F, et al. CRISPR-Cas9-mediated disruption of PD-1 on human T cells for adoptive cellular therapies of EBV positive gastric cancer. *Oncoimmunology*. 2017;6:e1249558.
26. O'Brien CA, Kreso A, Dick JE. Cancer stem cells in solid tumors: an overview. *Semin Radiat Oncol*. 2009;19:71-77.
27. Kreso A, Dick JE. Evolution of the cancer stem cell model. *Cell Stem Cell*. 2014;14:275-291.
28. Song Y, Wang Y, Tong C, et al. A unified model of the hierarchical and stochastic theories of gastric cancer. *Br J Cancer*. 2017;116:973-989.
29. Yong X, Tang B, Xiao YF, et al. *Helicobacter pylori* upregulates Nanog and Oct4 via Wnt/β-catenin signaling pathway to promote cancer stem cell-like properties in human gastric cancer. *Cancer Lett*. 2016;374:292-303.
30. Zhu L, Cheng X, Shi J, et al. Crosstalk between bone marrow-derived myofibroblasts and gastric cancer cells regulates cancer stemness and promotes tumorigenesis. *Oncogene*. 2016;35:5388-5399.
31. Ishimoto T, Miyake K, Nandi T, et al. Activation of transforming growth factor Beta 1 signaling in gastric cancer-associated fibroblasts increases their motility, via expression of Rhomboid 5 Homolog 2, and ability to induce invasiveness of gastric cancer cells. *Gastroenterology*. 2017;153:191-204.
32. Shaked Y, Voest EE. Bone marrow derived cells in tumor angiogenesis and growth: are they the good, the bad or the evil? *Biochim Biophys Acta*. 2009;1796:1-4.
33. Zhao Y, Feng F, Zhou YN. Stem cells in gastric cancer. *World J Gastroenterol*. 2015;21:112-123.
34. Bie Q, Zhang B, Sun C, Ji X, et al. IL-17B activated mesenchymal stem cells enhance proliferation and migration of gastric cancer cells. *Oncotarget*. 2017;8:18914-18923.
35. Moradi SL, Eslami G, Goudarzi H, et al. Role of *Helicobacter pylori* on cancer of human adipose-derived mesenchymal stem cells and metastasis of tumor cells-an in vitro study. *Tumour Biol*. 2016;37:3371-3378.
36. Ji N, Yu JW, Ni XC, et al. Bone marrow-derived mesenchymal stem cells increase drug resistance in CD133-expressing gastric cancer cells by regulating the PI3K/AKT pathway. *Tumour Biol*. 2016;37:14637-14651.
37. Hayakawa Y, Sakitani K, Konishi M, et al. Nerve growth factor promotes gastric tumorigenesis through aberrant cholinergic signaling. *Cancer Cell*. 2017;31:21-34.
38. Nguyen PH, Giraud J, Chambonnier L, et al. Characterization of biomarkers of tumorigenic and chemoresistant cancer stem cells in human gastric carcinoma. *Clin Cancer Res*. 2017;23:1586-1597.
39. Chen XL, Chen XZ, Wang YG, et al. Clinical significance of putative markers of cancer stem cells in gastric cancer: a retrospective cohort study. *Oncotarget*. 2016;7:62049-62069.
40. Choi YJ, Kim N, Lee HS, et al. Expression of leucine-rich repeat-containing G-protein coupled rreceptor 5 and CD44: potential implications for gastric cancer stem cell marker. *J Cancer Prev*. 2016;21:279-287.
41. Wu D, Mou YP, Chen K, et al. Aldehyde dehydrogenase 3A1 is robustly upregulated in gastric cancer stem-like cells and associated with tumorigenesis. *Int J Oncol*. 2016;49:611-622.
42. Li XB, Yang G, Zhu L, Tang YL, et al. Gastric Lgr5(+) stem cells are the cellular origin of invasive intestinal-type gastric cancer in mice. *Cell Res*. 2016;26:838-849.
43. Wang S, Liu E, Deng J, et al. Long noncoding RNA ROR regulates proliferation, invasion and stemness of gastric cancer stem cell. *Cell Reprogram*. 2016;18:319-326.
44. Cao X, Ren K, Song Z, et al. 7-Difluoromethoxyl-5,4′-di-n-octyl genistein inhibits the stem-like characteristics of gastric cancer stem-like cells and reverses the phenotype of epithelial-mesenchymal transition in gastric cancer cells. *Oncol Rep*. 2016;36:1157-1165.
45. Barat S, Chen X, Cuong Bui K, et al. Gamma-secretase inhibitor IX (GSI) impairs concomitant activation of notch and wnt-Beta-catenin pathways in CD44(+) gastric cancer stem cells. *Stem Cells Transl Med*. 2017;6:819-829.
46. Feng S, Zheng Z, Feng L, et al. Proton pump inhibitor pantoprazole inhibits the proliferation, self-renewal and chemoresistance of gastric cancer stem cells via the EMT/β-catenin pathways. *Oncol Rep*. 2016;36:3207-3214.
47. Nguyen PH, Giraud J, Staedel C, et al. All-trans retinoic acid targets gastric cancer stem cells and inhibits patient-derived gastric carcinoma tumor growth. *Oncogene*. 2016;35:5619-5628.
48. Shitara K, Doi T, Nagano O, et al. Dose-escalation study for the targeting of CD44v(+) cancer stem cells by sulfasalazine in

patients with advanced gastric cancer (EPOC1205). *Gastric Cancer*. 2017;20:341-349.
49. Fan D, Ren B, Yang X, Liu J, Zhang Z. Upregulation of miR-501-5p activates the wnt/β-catenin signaling pathway and enhances stem cell-like phenotype in gastric cancer. *J Exp Clin Cancer Res*. 2016;35:177.
50. Zhang L, Guo X, Zhang D, et al. Upregulated miR-132 in Lgr5(+) gastric cancer stem cell-like cells contributes to cisplatin-resistance via SIRT1/CREB/ABCG2 signaling pathway. *Mol Carcinog*. 2017;9999:1-13.
51. Mo W, Zhang JT. Human ABCG2: structure, function, and its role in multidrug resistance. *Int J Biochem Mol Biol*. 2012;3:1-27.
52. Pan Y, Shu X, Sun L, et al. miR-196a-5p modulates gastric cancer stem cell characteristics by targeting Smad4. *Int J Oncol*. 2017;50:1965-1976.
53. Wang X, Wang C, Zhang X, et al. Bmi-1 regulates stem cell-like properties of gastric cancer cells via modulating miRNAs. *J Hematol Oncol*. 2016;9:90.
54. Wei B, Sun X, Geng Z, et al. Isoproterenol regulates CD44 expression in gastric cancer cells through STAT3/MicroRNA373 cascade. *Biomaterials*. 2016;105:89-101.
55. Belair C, Baud J, Chabas S, et al. *Helicobacter pylori* interferes with an embryonic stem cell micro RNA cluster to block cell cycle progression. *Silence*. 2011;2:7.
56. Zhou Y, Wang Y, Wen J, et al. Aquaporin 3 promotes the stem-like properties of gastric cancer cells via Wnt/GSK-3β/β-catenin pathway. *Oncotarget*. 2016;7:16529-16541.
57. Chang TS, Wei KL, Lu CK, et al. Inhibition of CCAR1, a coactivator of β-catenin, suppresses the proliferation and migration of gastric cancer cells. *Int J Mol Sci*. 2017;18:460
58. Cai W, Chen G, Luo Q, et al. PMP22 regulates self-renewal and chemoresistance of gastric cancer cells. *Mol Cancer Ther*. 2017;16:1187-1198.
59. Yang SW, Zhang ZG, Hao YX, et al. HIF-1α induces the epithelial-mesenchymal transition in gastric cancer stem cells through the Snail pathway. *Oncotarget*. 2017;8:9535-9545.

How to cite this article: Molina-Castro S, Pereira-Marques J, Figueiredo C, Machado JC, Varon C. Gastric cancer: Basic aspects. *Helicobacter*. 2017;22(Suppl. 1):e12412. https://doi.org/10.1111/hel.12412