

# **Insights Into the Human Gut Microbiome and its Link with Obesity and Cardiometabolic Diseases**

## **Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät  
der Eberhard Karls Universität Tübingen  
zur Erlangung des Grades eines  
Doktors der Naturwissenschaften  
(Dr. rer. nat.)

vorgelegt von  
Juan Jacobo de la Cuesta Zuluaga  
aus Medellín, Kolumbien

Tübingen

2022



Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der  
Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	30.05.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Ruth Ley, Ph.D.
2. Berichterstatter:	Prof. Dr. Daniel Huson
3. Berichterstatter:	Prof. Jack Gilbert, Ph.D.



# Index

<b>Index</b>	<b>5</b>
<b>Acknowledgements</b>	<b>7</b>
<b>Summary</b>	<b>11</b>
<b>Zusammenfassung</b>	<b>13</b>
<b>Introduction</b>	<b>17</b>
The human gut microbiome	17
The interplay between the gut microbiome, obesity and cardiometabolic conditions	18
Confounders and the importance of well-characterized cohorts	21
Culture-independent methods to study the gut microbiome	22
Internationalizing microbiome research	25
Open data as a key element for studying the microbiome	26
Outline	27
<b>Publications</b>	<b>31</b>
<b>Declaration of contributions in collaborative works</b>	<b>33</b>
<b>Journal reuse policies</b>	<b>35</b>
<b>Chapter I</b>	<b>37</b>
Abstract	37
Contributions	38
<b>Chapter II</b>	<b>39</b>
Abstract	39
Contributions	40
<b>Chapter III</b>	<b>41</b>
Abstract	41
Contributions	41
<b>Chapter IV</b>	<b>43</b>
Abstract	43
Contributions	44
<b>Chapter V</b>	<b>45</b>
Abstract	45

Contributions	46
<b>Discussion and outlook</b>	<b>47</b>
Contribution of this thesis	47
Pitfalls and shortcomings	48
Cross-sectional, computational and correlational studies	48
Impact of medications and other confounders on the gut microbiome	49
The limitations of databases and methods that rely on them	50
Metagenomics vs Metatranscriptomics or Metaproteomics	51
Promises and future challenges of microbiome science	53
<b>Bibliography</b>	<b>55</b>
<b>Appendices</b>	<b>67</b>
Appendix I: Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults	69
Appendix II: Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut	83
Appendix III: Struo - a pipeline for building custom databases for common metagenome profilers	101
Appendix IV: Obesity is the main driver of functional alterations of the gut microbiome in cardiometabolic disease	105
Appendix V: Gut metagenomes and assembled microbial genomes from human adults from urban cohorts from Colombia, South America	153

# Acknowledgements

It was sometime in the middle of my first year, while I was dragging myself through my first incursion into the valley of shit<sup>1</sup> that this section of my thesis came to my mind. I am not really sure why, but somehow, it soothed me. It may have been the effect of visualizing success, after all, if you are reading this is because my thesis is finished. It may have also been a way of reminding myself how many great, brilliant, caring and supporting people surround me. Or it could just have been the musings of an anxious mind.

I often found myself lost in those thoughts, daydreaming while I walked from home to the institute. A couple of years into my PhD the COVID-19 pandemic put a stop to the daydreamy commute for quite some time, but it made me realize how fortunate I am to be where I have been and do what I have done. I am aware, now more than ever, of the luck, the support and the privilege that I have had so that I can think all these commuting, and later lockdown, thoughts. Some of which I finally get to put into text.

I would like to thank:

## *My mentors and supervisors*

Ruth Ley, my PhD supervisor, for allowing me to be in the best place I could be, for putting me to the test, for pushing me to grow as a scientist. Daniel Huson and Juan S. Escobar, members of my thesis advisory committee, for the fruitful discussions, their feedback and encouragement. Nick Youngblut, for creating and maintaining the bioinformatic structure of the lab, for his feedback, and the productive work relationship we developed, without which my work would have been way harder than it was. Miryan Sánchez, Gabriel Saavedra, and all my past mentors, who channeled my curiosity and my love of learning and that influenced me in one way or another to be where I am right now.

### *My labmates and other people around the institute*

*Frau Dr.* Claudia Mirretta Barone, Guillermo Luque, Andrea Borbón, Albane Ruaud, Leo Moreno, Sarah Keim and Tanja Schön for all the laughs, jokes, bantering, pranks weird conversations, venting sessions, vending machine trips, coffee breaks, lunches, cooking weekends, beer hours and any other time spent together. Tony Walters, Daphne Welter, Zach Henseler and Jess Sutter, for all the Sundays of chili, table-top RPGs and discussions of weird internet stuff; when I grow up I wanna be as generous as you are. Adrian Contreras, Sergio Latorre, Thanvi Srikant, Rebecca Schwab, Christa Lanz, George Deffner and everyone with whom I ever enjoyed a beer during B.E.E.R hour, a coffee on a random break or just a small chat on any of the halls of the institute. Karin Klein and Ursula Schach for all the support that keeps the department and the lab up and running. Hagay Enav, Taichi Suzuki, Xiaoying Liu and Liam Fitzstevens, for all the feedback and the many pleasant conversations about science or science-adjacent stuff.

### *My collaborators*

Vidarium Research Center and all my friends there, for the many years of fruitful scientific collaboration and tireless efforts that were vital to the success of this venture and my growth as a scientist. All my coauthors across continents and institutions, for all your help and support with my work; it was also my pleasure to help you with yours whenever I could.

### *My Friends and family*

María Lucía Zuluaga, Alberto de la Cuesta and Luisa de la Cuesta: Mom, Dad and Sis, for whom my love is like the universe: infinite yet continues to expand. Fabián Mejía and Juan Pablo Lopera, true friends across the ocean, always willing to hear me and to share the growing pains of becoming a researcher. John Mario Gutierrez, whose morning calls, weird internet memes and selection of tropical music became a staple of my day to day. Juan Sebastian Villa, porci, ademán.



*People who don't know me*

All the students, postdocs and researchers who developed the software, methods and tutorials I used along my PhD. Gus, Zeta and Charly, whose melodies and lyrics have been the soundtrack of my life for years, yet somehow, their meaning always fit my current life stage. The brothers Green, Mike Duncan, Dan Carlin, Roman Mars and the many other people whose podcasts kept me company on my way to the institute, while I cooked, cleaned my apartment, exercised, tried to fall asleep or just looked for a way of driving away the silence of a universe indifferent to all human emotion.

*Last, but certainly not least*

Aleja Duque, who I love and admire. Whose presence has been invaluable through my academic career, navigating adulthood and surviving pandemics. Who, with patience, has endured my silliness and bad jokes (nah, they're all really good). Who kept me company while I was in some really dark places during my PhD. With whom I have grown from the larva of a biologist to a fully grown Dr. Bug.

Parafraseando al eterno Gustavo Cerati: *no solo no hubiera sido nada sin ustedes, sino con toda la gente que estuvo a mi alrededor desde el comienzo; algunos siguen hasta hoy. ¡Gracias... totales!*

*PS* - I also want to anti-thank (?) COVID-19, my anxiety and all the people who were examples of that which I do not want to become. Goddamn you all to hell!

*PPS* - in case I find myself reading this many years in the future: Future Jacobo, I wish I knew as much as you know. I did the best I could, please be kind to us both. The same applies to other people, they are doing the best they can. Please be kind to them.

1: Mewburn I. The Valley Of Shit. The Thesis Whisperer. 2012 [cited 2021 Sep 1]. Available from: <https://thesiswhisperer.com/2012/05/08/the-valley-of-shit/>



# Summary

The gut microbiome is a complex microbial community that inhabits the gastrointestinal tract comprising archaea, bacteria, viruses and fungi. This community lies at the interface between our environment and our cells. As such, it plays an important role in multiple nutritional, physiological and immune processes, including the synthesis of vitamins and other compounds, the energy harvest from food, and the tight regulation of innate and adaptive immunity. The gut microbiome is implicated in the pathophysiology of obesity, type 2 diabetes and cardiovascular disease. This is of particular relevance in the context of the epidemiologic and dietary transition that characterizes westernization, a process in which low- and middle-income countries shift towards increased consumption of processed foods and reduced physical activity with a concomitant increase in non-communicable diseases. This thesis contributes to our understanding of the role of the gut microbiome in cardiometabolic disease and obesity

In chapter I, I studied the gut microbiome of adults from multiple populations to describe the association between the host's age and sex and the gut microbial diversity using 16S rRNA gene sequencing. I showed that the microbiome diversity increased with age until 40 years of age, and that young, but not middle-aged adult women had higher gut microbial diversity than men. These observations were robust to the use of antibiotics or the cardiometabolic health of the subjects. However, the pattern was not universal since it was not observed in all studied populations.

In chapter II, I described the diversity, ecological distribution and genomic characteristics of the archaeal order *Methanomassiliicoccales*, which have potential as microbiome-based therapeutics. I carried out genomic and phylogenetic comparisons and confirmed that the *Methanomassiliicoccales* order forms two large phylogenetic clades. Based on abundance across host-associated and environmental metagenomes, I showed that the clades largely differ in environment preference and genomic potential.

Chapter III introduces a modular pipeline that aids with the retrieval of microbial genomes from public databases, which are then used to create custom databases for several

metagenome profiler programs. Here I carried out benchmark analyses on synthetic and real datasets and showed that the use of custom databases result in an increase of mappability of sequencing reads. Databases created using this pipeline were used in other chapters of this dissertation.

For chapter IV, I evaluated the functional potential of the gut microbiome of a cohort of Colombian adults to detect variation in the microbiome associated with obesity or cardiometabolic health. I used shotgun gut metagenomes from the Colombian cohort to test the reproducibility of a set of functional characteristics previously reported to be associated with cardiometabolic conditions in other populations. Using host metadata, I classified subjects according to their obesity and cardiometabolic status, and identified which microbiome functions were uniquely associated with each condition. I found that obesity drives associations of the microbiome with cardiometabolic disease when both are present.

Chapter V describes the retrieval of genomes from the Colombian gut microbiome using the metagenome sequence data I collected in the previous chapter. I evaluated the quality of the genome assemblies, performed the taxonomic classification, established their taxonomic novelty compared to what is currently reported, and annotated functional and genomic characteristics.

All in all, the works presented in this thesis advance our knowledge of the role of the gut microbiome in obesity and cardiometabolic disease. I expect this will help guide future studies that use metagenomics to look into the associations and mechanisms of the microbiome with these non-communicable conditions.

# Zusammenfassung

Das Darmmikrobiom ist eine komplexe Gemeinschaft von Mikroorganismen, welche im Gastrointestinaltrakt angesiedelt ist und sich aus Archaeen, Bakterien, Viren und Pilzen zusammensetzt. Diese Gemeinschaft befindet sich direkt am Übergang von unserer Außenwelt zu unseren Zellen und übernimmt wichtige Rollen bei ernährungsbezogenen, physiologischen und immunologischen Prozessen. Zum Beispiel, bei der Synthese von Vitaminen und anderen Verbindungen, bei der Energiegewinnung aus der Nahrung und bei der Regulation der angeborenen und erworbenen Immunität. Darüber hinaus wird das Darmmikrobiom mit pathophysiologischen Prozessen wie Fettleibigkeit, Typ-2-Diabetes und Herz-Kreislauf-Erkrankungen in Verbindung gebracht. Dies ist von besonderer Bedeutung im Kontext des epidemiologischen und ernährungsbedingten Wandels, den die Westernisierung mit sich bringt, ein Prozess, bei dem für Länder mit niedrigem und mittlerem Einkommen ein erhöhter Konsum von verarbeiteten Lebensmitteln und weniger körperliche Aktivität festgestellt werden kann, was mit einem Anstieg der Inzidenz nicht übertragbarer Krankheiten einhergeht. Diese Arbeit trägt zu einem besseren Verständnis der Rolle des Darmmikrobioms bei kardiometabolischen Erkrankungen und Fettleibigkeit bei.

In Kapitel I untersuchte ich das Darmmikrobiom erwachsener Probanden verschiedener Populationen auf einen Zusammenhang zwischen Alter, Geschlecht und der mikrobiellen Vielfalt im Darm mittels 16S rRNA-Gen-Sequenzierung. Ich zeigte eine mit dem Alter zunehmende Diversität des Mikrobioms bis zu einem Lebensalter von 40 Jahren, als auch eine höhere Diversität des Darmmikrobioms bei jungen erwachsenen Frauen, jedoch nicht mittleren Alters, im Gegensatz zu Männern. Diese Erkenntnisse waren unabhängig von der Antibiotikaeinnahme oder der kardiometabolischen Gesundheit der Probanden. Allerdings war dieses Muster nicht universell, da es nicht in allen untersuchten Populationen nachgewiesen wurde.

In Kapitel II habe ich die Vielfalt, die ökologische Verteilung und die genomischen Merkmale von Archaeen der Ordnung *Methanomassiliicoccales* beschrieben, welche das Potential haben, als mikrobiombasierte Therapeutika eingesetzt zu werden. Hierzu habe ich

genomische und phylogenetische Vergleiche durchgeführt und so bestätigte, dass die Ordnung der *Methanomassiliicoccales* zwei große phylogenetische Gruppen beschreibt. Basierend auf ihrer Abundanz in Metagenomen von Proben die dem Wirt und der Umwelt entnommen wurden, konnte ich zeigen, dass sich die beiden Gruppen hinsichtlich ihrer Präferenzen für ihre Umgebungen und in ihrem genomischen Potenzial stark unterscheiden.

In Kapitel III habe ich einen anpassungsfähigen Verarbeitungsablauf vorgestellt, der beim Abrufen mikrobieller Genome aus öffentlichen Datenbanken hilft, welche dann zur Erstellung benutzerdefinierter Datenbanken für verschiedene Metagenomprofiler-Programme verwendet werden können. Im Rahmen einer Benchmarkanalyse mit synthetischen und realen Datensätzen zeigte ich, dass die Verwendung benutzerdefinierter Datenbanken zu einer verbesserten Zuordnungsfähigkeit der Sequenzierungsdaten führt. Datenbanken, welche Mittels dieser Pipeline erstellt wurden, wurden in anderen Kapiteln dieser Dissertation verwendet.

In Kapitel IV habe ich das funktionelle Potenzial des Darmmikrobioms eines Kohorts kolumbianischer Erwachsener analysiert, um Variationen im Mikrobiom aufzudecken, die mit Fettleibigkeit oder kardiometabolischer Gesundheit assoziiert sind. Ich habe Shotgun-Darm-Metagenome des Kolumbien-Kohorts verwendet, um die Reproduzierbarkeit einer Reihe von funktionellen Merkmalen zu zeigen, von denen zuvor berichtet worden war, dass diese mit kardiometabolischen Erkrankungen in anderen Populationen assoziiert sind. Anhand von Metadaten die für den Wirt erhoben wurden, habe ich die Probanden aufgrund ihrer Fettleibigkeit und ihres kardiometabolischen Status klassifiziert und so ermittelt, welche funktionellen Eigenschaften der Mikrobiome eindeutig mit dem jeweiligen Zustand assoziiert waren. Ich fand heraus, dass Fettleibigkeit zur Assoziation des Mikrobioms mit kardiometabolischen Erkrankungen führt, wenn beide Erkrankungen vorhanden sind.

Kapitel V beschreibt die Erstellung von Genomen aus den Metagenom-Sequenzdaten des Darmmikrobioms der kolumbianischen Kohorte, welche im vorangegangenen Kapitel beschrieben wurden. Ich beurteilte die Qualität der erstellten Genome, vollzog die

taxonomische Klassifizierung, ermittelte ihre taxonomische Neuheit im Vergleich zu den derzeit vorliegenden Informationen, und annotierte ihre funktionellen sowie genomischen Merkmale.

Zusammengefasst, erweitern die vorgestellten Ergebnisse dieser Dissertation unser Wissen über die Rolle des Darmmikrobioms bei Fettleibigkeit und kardiometabolischen Erkrankungen. Ich gehe davon aus, dass die hier erlangten Erkenntnisse für künftige Studien hilfreich sein werden, bei denen Metagenom-Analysen genutzt werden, um die Assoziationen und Mechanismen des Mikrobioms mit diesen nicht übertragbaren Krankheiten zu untersuchen.





# Introduction

## The human gut microbiome

The gut microbiome is the collection of bacteria, archaea, viruses and fungi that inhabit the human gastrointestinal tract, together with their theater of activity (Marchesi and Ravel, 2015). This theater includes the nucleic acids, proteins, lipids and other metabolites produced by the microbes, plus compounds produced by the host (Berg *et al.*, 2020). The study of the microorganisms within the human gut has become a flourishing area of evolutionary, ecological, nutritional, and medical research (Knight *et al.*, 2017; McCarville *et al.*, 2020; Suzuki and Ley, 2020). The interest in this field has both benefited from and led to advances in sequencing of nucleic acids (Integrative HMP (iHMP) Research Network Consortium, 2019); algorithms for the manipulation of biological sequences (Bağcı, Patz and Huson, 2021) or the analysis of interaction networks (Knight *et al.*, 2018; Beghini *et al.*, 2021); novel statistical frameworks (Quinn *et al.*, 2018; Morton *et al.*, 2019), machine learning approaches (Topçuoğlu *et al.*, 2020); and the isolation of as-of-recently uncultured microorganisms (Forster *et al.*, 2019; Zou *et al.*, 2019). At a more fundamental level, it has changed the understanding of our relationship with microbes, insofar as we constitute a metaorganism, whose functioning is intimately linked to the microbial communities that colonize our bodies (Theis *et al.*, 2016).

The gut microbiome can be understood using two complementary conceptual frameworks: one ecological and the other physiological. From an ecological perspective, the microbiome is an ecosystem subject to changes in flows of nutrients (David *et al.*, 2014), environmental stresses (Janzon *et al.*, 2019) and other evolutionary factors (Ley, Peterson and Gordon, 2006), where its members establish ecological relations with each other and with their environment (Banerjee, Schlaeppli and van der Heijden, 2018). The ecological relations with their environment, the human gut, is what distinguishes a host-associated microbiota from other microbial communities and bridges the ecological and the physiological

frameworks. The environment itself is a living organism which establishes relations with the microbes, favoring ecological niches for commensal species (Suzuki and Ley, 2020) while avoiding colonization by pathogens (Litvak and Bäumler, 2019), and utilizing compounds produced by the microbes. Indeed, in the human gut, it is estimated that the microbial density can reach up to  $10^{11}$  microbes  $\text{gram}^{-1}$ , making it the main source of microbial metabolites in the body (Sender, Fuchs and Milo, 2016). These metabolites include, but are not limited to, short-chain fatty acids, sphingolipids, methylamines, in addition to antigens such as flagellin or lipopolysaccharide, which modulate host health. Because of this, the microbiome can be thought of as a microbial organ, that is, a collection of cells that resides as a structural unit within the body of the host and that have a common function (Byndloss and Bäumler, 2018).

## The interplay between the gut microbiome, obesity and cardiometabolic conditions

Obesity is a non-communicable disease that has been increasing at alarming rates across the world. It is estimated that 2 billion people are overweight (body mass index [BMI]  $\geq 25 \text{ kg m}^{-2}$ ), one third of them being obese (BMI  $\geq 30 \text{ kg m}^{-2}$ ) (Seidell and Halberstadt, 2015). Obesity is considered a risk factor for other major non-communicable conditions, including type 2 diabetes, cardiovascular disease, coronary heart disease, stroke and multiple cancers (Grover *et al.*, 2015). Such conditions are, in turn, associated with a decrease in quality of life and a reduction in life expectancy (Nyberg *et al.*, 2018). The rising incidence of obesity and associated disorders is not limited to high-income countries; on the contrary, it is also a major issue in low- and middle-income countries (Dinsa *et al.*, 2012).

The gut microbiome is implicated in the pathophysiology of obesity and the associated conditions type 2 diabetes and cardiovascular disease. The role of the microbial community is multifaceted; at the most basic level, the overall diversity of the microbiome (Le Chatelier *et al.*, 2013; Walters, Xu and Knight, 2014), the abundance of multiple taxa (Duvall *et al.*, 2017) and the functions they perform (Armour *et al.*, 2019) differs between

diseased and healthy individuals. Moreover, the microbiome can influence the host physiology under such conditions in several ways. First, the gut microbiome is closely connected to the host diet. Individuals with diets rich in ultra processed foods, fat and sugars have a microbiota enriched in bile-tolerant and putrefactive microorganisms (García-Vega *et al.*, 2020). In contrast, subjects whose diets are rich in plant polysaccharides harbor a microbiome with the ability to degrade dietary fiber and produce short-chain fatty acids (SCFA). Fiber consumption has been associated with leanness in cross-sectional (Hadrévi, Søggaard and Christensen, 2017) and intervention studies (Buscemi *et al.*, 2018), and in murine models the addition of fiber prevented the onset of metabolic syndrome induced by a high-fat diet (Zou *et al.*, 2018). Relatedly, methanogenic *Archaea* can increase the energetic efficiency of primary fermenters by reducing partial pressures of H<sub>2</sub> through methanogenesis, which results in an increased production of SCFA (Horz and Conrads, 2010). SCFAs have a positive impact on host health (Morrison and Preston, 2016): butyrate is the main energy source for colonocytes, and evidence from *in vitro* assays indicate that it influences the maintenance of the gut barrier by preserving luminal anaerobiosis and promoting the assembly of tight junction proteins (Kelly *et al.*, 2015). Acetate, propionate and butyrate regulate the homeostasis of glucose and lipids in the liver (den Besten *et al.*, 2015), and circulating acetate is negatively correlated with plasma insulin levels (Layden *et al.*, 2012).

Conversely, the microbiome impacts the progression of obesity and cardiometabolic conditions. A large component of the progression of obesity, diabetes and cardiovascular disease is linked to various inflammatory processes, both at the systemic level as well as on particular host tissues. Recent evidence underscores the importance of intestinal inflammation in the development of obesity (Cox, West and Cripps, 2015). Low-fiber, high-fat diets (Martinez-Medina *et al.*, 2014; O’Keefe *et al.*, 2015; Statovci *et al.*, 2017) or the consumption of antibiotics (Palleja *et al.*, 2018), to name a couple, are pro-inflammatory challenges that favor changes in the metabolism of the gut epithelium from  $\beta$ -oxidation of butyrate towards anaerobic glycolysis. This causes the gut to lose its hypoxic status (Litvak, Byndloss and Bäumler, 2018) and results in environmental conditions favorable to facultative

anaerobes (Shin, Whon and Bae, 2015; Zeng, Inohara and Nuñez, 2017). The expansion of anaerobes from the family *Enterobacteriaceae* (phylum *Proteobacteria*) is considered a signature of gut epithelial dysfunction (Litvak *et al.*, 2017). Perturbations in the integrity of the epithelial barrier result in translocation of microbial antigens, such as lipopolysaccharide (LPS), which promote low-grade inflammation, exacerbating epithelial dysfunction (Mohammad and Thiernemann, 2020). In turn, increased circulating LPS levels promote a rise in pro-inflammatory cytokines, inducing a state of low-grade systemic inflammation that is linked to glucose intolerance and insulin resistance (Ding and Lund, 2011).

A further way in which the microbiome can promote detrimental cardiovascular health outcomes is the synthesis of methylamines (Zeisel and Warrier, 2017). The utilization of dietary compounds such as carnitine or choline by various microorganisms results in the synthesis of trimethylamine (TMA). This compound is absorbed by the host and carried to the liver by the portal vein, where it is oxidized to trimethylamine N-oxide (TMAO). Multiple epidemiological and experimental studies indicate that circulating levels of TMAO are directly related to cardiovascular disease (Brown and Hazen, 2018), and it has been suggested that TMAO inhibits cholesterol transport and promotes its accumulation in macrophages, resulting in the formation of atherosclerotic plaques (Geng *et al.*, 2018).

The associations and mechanisms described above are a non-exhaustive list of the ways in which the microbiome is involved in obesity and cardiometabolic disease. Yet they serve to illustrate an important characteristic of the microbial community: unlike other human organs, it can potentially be targeted rapidly and with relative ease by interventions utilizing pharmacological, nutritional or probiotic elements, or a combination thereof (Zimmermann *et al.*, 2021). The challenge is to robustly determine associations between the host phenotypes and microbiome features. This requires access to well-characterized and data-rich cohorts, as I describe below.

## Confounders and the importance of well-characterized cohorts

Conditions with highly overlapping phenotypes, such as obesity, cardiovascular disease or diabetes, require a thorough characterization of the subjects. This is not, however, the only reason why good host data is required for the successful study of the microbiome in a human cohort. The gut microbiome is a highly plastic ecosystem, and its composition and structure are influenced by a myriad of host and environmental factors such as diet, medication usage, physical activity, host genetics, among others. Many of the aforementioned host parameters are known risk factors for multiple diseases, therefore, it is not uncommon for these factors to confound the variable or condition studied.

The ultimate goal of cross-sectional studies is to obtain a small group of species or functions with a robust association to the host's phenotype. These sets of features can serve as the source of hypotheses to be evaluated in intervention or mechanistic studies. However, the identification of causal relationships between specific microbes or functions with disease is hindered by the low concordance between studies (Duvall *et al.*, 2017; Armour *et al.*, 2019). While it cannot be discarded that some links between microbial features and host physiology might be population-specific or contingent on particular configurations of the microbiota, the inclusion of known confounders can lessen the risk of obtaining false positives in cross-sectional population studies (Ghosh *et al.*, 2020; Vujkovic-Cvijin *et al.*, 2020). In other words, accounting for common confounders might facilitate the comparison between studies by reducing biases introduced by the confounding variables (Vujkovic-Cvijin *et al.*, 2020).

Previous studies have pointed out pervasive confounding variables in cohort studies of the gut microbiome (Vujkovic-Cvijin *et al.*, 2020). Some of these factors are associated with the structure and composition of the microbiome and represent risk factors of non-communicable diseases; namely, age (Biagi *et al.*, 2010; Odamaki *et al.*, 2016; Ghosh *et al.*, 2020) and sex (Markle *et al.*, 2013; Wallis *et al.*, 2017; Sinha *et al.*, 2019) of the host. Other confounders appear precisely because they are used to treat conditions such as hypertension or dyslipidemia, and are known to have a direct effect over gut microbes,

although not all individuals with such conditions are exposed to them. Such is the case of medications (Maier *et al.*, 2018; Vich Vila *et al.*, 2020) including antibiotics (Forslund *et al.*, 2013), lipid-lowering medication (Kummen *et al.*, 2020), antidiabetics (Forslund *et al.*, 2015; de la Cuesta-Zuluaga *et al.*, 2017) or proton pump inhibitors (Jackson *et al.*, 2016). Finally, geographical origin is linked to the composition of the microbiome even at the regional level, as has been shown in several populations (He *et al.*, 2018), including the Colombian cohort I studied in the present thesis (de la Cuesta-Zuluaga *et al.*, 2018).

Therefore, there is the need to include microbiome-associated confounding host variables in studies that look into the links between host health and physiological, dietary or epidemiological factors. This necessitates the use of well-phenotyped populations, where the collected host data allows consideration of the influence of confounding host variables over the composition of the microbiome while performing statistical analyses. Moreover, as I will briefly discuss in the ‘*Open data as a key element for studying the microbiome*’ section of this introduction, sharing this information in a way that guarantees its accessibility, interoperability and reusability is key for the advancement of microbiome research as a collective endeavor (Ryan *et al.*, 2021).

## Culture-independent methods to study the gut microbiome

The insights that can be obtained about the role of the microbiome in health and disease are contingent on the approaches used to study the microbial community. Culture-based methods are key to understanding the links between microbes and host (Maier *et al.*, 2018), the interactions between members of the microbiota (Ruaud *et al.*, 2020) and the mechanisms by which microbial compounds affect host systems (Johnson *et al.*, 2020). However, culture-based methods fall short when characterizing the complete microbial community since not all gut microbes are as-of-yet culturable, they are low-throughput, and do not provide an overview of all members of the community and their relative abundances (Almeida *et al.*, 2021). Sequencing-based methods are culture-independent and can overcome

some of these pitfalls, since they can potentially detect a large fraction of the community with an even-increasing throughput and an ever-decreasing cost (Youngblut and Ley, 2021).

The most widely used culture-independent method to survey microbiomes is the sequencing of a single marker gene, such as the 16S rRNA gene (Goodrich *et al.*, 2014). After sample collection, total DNA extracted from the microbial community is used as starting material. Particular phylogenetically and taxonomically informative regions of the selected marker are amplified by PCR and then sequenced. Amplicon reads are quality-filtered and clustered into operational taxonomic units (OTUs) based on sequence identity or denoised into sequence variants (SVs) based on sequencing error profiles (Bolyen *et al.*, 2019). Representative sequences from SVs or OTUs are then taxonomically annotated by matching them against reference databases (de la Cuesta-Zuluaga and Escobar, 2016). In addition, these markers can be used to infer phylogenies that encompass all detected members of the community, which in turn allow assessing intra- and inter-sample diversity (Lozupone *et al.*, 2011). The end result of marker-gene workflows are tables enumerating the abundance of the members of the microbial community across the sequenced samples. However, marker-gene-based methods can only provide information about the presence and abundance of microbes in the microbiota, no inference about the functions they potentially or actually perform can be directly obtained (Aßhauer *et al.*, 2015).

Alternatively, community DNA can also be used to perform shotgun metagenome sequencing (hereafter metagenomics), where the genetic material is randomly sequenced without the use of marker-specific primers. This method has gained popularity in recent years thanks to the declining costs of sequencing and computational resources (Hillmann *et al.*, 2018). An advantage of metagenomics over marker-gene sequencing is the greater amount of information it provides, even at relatively shallow coverages (Nayfach *et al.*, 2015; Hillmann *et al.*, 2018). This includes information regarding the metabolic potential of the microbial community (Franzosa *et al.*, 2018), estimation of species-level abundance (Lu *et al.*, 2017) and the ability of retrieving genomes or gene catalogs (Almeida *et al.*, 2021). Moreover, metagenomics allows to obtain functional and phylogenetic diversity measures from whole

genomes and high-quality multi-locus phylogenies, which help to resolve associations between the microbiome and host phenotypes, as we recently showed (Youngblut, de la Cuesta-Zuluaga and Ley, 2022).

To obtain the taxonomic and functional profile of the microbiome, shotgun metagenome reads are mapped against databases of genomes or genes and their abundance is calculated (Franzosa *et al.*, 2018; Wood, Lu and Langmead, 2019). In any database-dependent method or step, be it marker gene or metagenomic sequencing, the content of the databases will heavily influence the final result. Indeed, genomes from many novel microbial taxa have been recovered in recent years thanks to advances in culturing methods (Forster *et al.*, 2019; Zou *et al.*, 2019) and metagenome sequencing and assembly (Pasolli *et al.*, 2019), which has led to the creation of large collections of Bacteria and Archaea genomes (Parks *et al.*, 2018) and comprehensive databases that leverage them (de la Cuesta-Zuluaga, Ley and Youngblut, 2020; Youngblut and Ley, 2021), thus reducing the risk of oversights by including the most up-to-date microbial data. Similar to marker gene sequencing, the end result of metagenomic profiling workflows are tables of abundances of taxa, genes or metabolic pathways across all samples.

In addition to functional profiling, metagenomics allows the assembly of whole genomes. Metagenome-assembled genomes (MAGs) are most commonly generated by assembling shotgun reads into contigs on a per-sample basis; the contigs are then grouped into bins according to similarities in k-mer frequency or patterns of sequence coverage across multiple samples (Sieber *et al.*, 2018). Once binned, the quality of the MAGs is assessed in terms of completeness and contamination (also called redundancy) according to the presence of clade-specific single-copy genes (Parks *et al.*, 2015). Quality-filtered MAGs can then be subjected to taxonomic classification, phylogenetic inference and gene calling (Parks *et al.*, 2017). As with any large collection of genomes, MAGs can be used to assess the diversity of a microbiome (Youngblut *et al.*, 2020), to compare diversification patterns of microbes with their hosts (Suzuki *et al.*, 2021), to expand databases for metagenome profiling (de la Cuesta-Zuluaga, Ley and Youngblut, 2020; Youngblut and Ley, 2021), to perform



comparative genomics analyses of specific taxa (Tett *et al.*, 2019; De Filippis, Pasoli and Ercolini, 2020), among others. All this with the advantage of not requiring the microorganisms in pure culture in order to have information about the genetic potential they encode (Almeida *et al.*, 2019).

## Internationalizing microbiome research

As mentioned above, the gut microbiome has broad relevance to the health of the host. Given the wide variation in the composition of the microbiome between and within populations, the identification of generalizable associations and mechanistic links between the microbial community and human health requires the study of diverse human populations. Yet, there is a very strong bias in the representation of human populations in repositories of genomic data (Abdill, Adamowicz and Blekhman, 2022). A recent survey of human microbiome data found that publicly available samples are dominated by highly developed countries: The United States contributed 40.2% of available 16S rRNA amplicon sequencing or shotgun metagenome samples while representing only 4.3% of the global population. Likewise, China and European countries also contribute to the bulk of the available samples, while countries from the global south, including Southeast Asia, Africa and Latin America are underrepresented (Abdill, Adamowicz and Blekhman, 2022).

There are still unanswered questions in the field of microbiome science which would benefit from studying a wide range of populations, including but not limited to: which characteristics of the gut microbiome are specific to certain populations and which are universal? Do associations of the microbiome with human health described in subjects from high-income countries extend to lower- and middle-income countries? How does the distribution of potentially beneficial microbes vary between countries? (Porrás and Brito, 2019). In the absence of studies in these populations, it is not possible to determine how generalizable the associations between microbiota and host health are.

For this reason, there have been calls for initiatives that identify and include populations with socioeconomic and environmental factors outside of high-income

countries, so that a universal understanding of the human microbiome and its effect on host health can be achieved. Otherwise, the benefits of microbiome research may be extended to only a fraction of the world's population (Porrás and Brito, 2019). As I will describe in the *Outline* section of this introduction, the works included in this thesis address questions regarding the diversity of the gut microbiome, its association with human health, and the methodological challenges of studying this microbial community, while incorporating data from multiple populations.

## Open data as a key element for studying the microbiome

The phrase '*standing on the shoulders of giants*' has become commonplace but that does not make it any less true. One of the essential elements of science is the use of existing information to solve new questions. It is now possible, even expected, for data generated as part of a research project to be made public in a transparent, reproducible and reusable manner (Wilkinson *et al.*, 2016), especially if such data come from research funded with public money.

Secondary analyses, in which researchers combine and reanalyze public data in a manner not planned by the original authors, are valuable for the study of microbiomes (Pasolli *et al.*, 2016; Duvallet *et al.*, 2017; Armour *et al.*, 2019; Ruaud *et al.*, 2020; Youngblut *et al.*, 2020). This has been facilitated by centralized repositories of raw high-throughput DNA sequences, from which microbial and host data can be retrieved and processed. Results that are robust to reanalysis are more credible, and studies that produce new knowledge from underutilized data make the practice of science more efficient, particularly in cases where the generation of new data is not possible (Rajesh *et al.*, 2021) (even though this may upset some people (Longo and Drazen, 2016)).

Despite this, about one-fifth of microbiome studies do not make public the data they generated (Eckert *et al.*, 2020). The works included in the present thesis benefited extensively from a large and diverse amount of genomic and metagenomic data that can be accessed through various databases. In turn, I have strived to guarantee that the new genome and

metagenome data I produced as part of my research projects, in addition to the code to analyze them, are made public in databases and code repositories.

## Outline

In chapter I, I present a multi-population study where I described the relationship of age and sex to gut bacterial diversity in young and middle-aged adults from four geographic regions: the United States, the United Kingdom, Colombia, and China. I observed that microbial diversity increased with age yet plateaued at about 40 years, and that young, but not middle-aged, adult women had higher gut microbial diversity than men. Microbial diversity was not associated with cardiometabolic health and medication consumption; the observed patterns remained after adjusting for cardiometabolic parameters in the Colombian cohort and antibiotic usage in the US and UK cohorts. The aforementioned association of age and sex with microbial diversity was not evidenced in the Chinese cohort, therefore, its universality remains an open question.

In chapter II, I carried out a study on the diversity, ecological distribution and genomic characteristics of the archaeal order *Methanomassiliicoccales*, lesser-known members of the human gut microbiota. Microorganisms from this order use methylated amines, including trimethylamine (TMA), for methane production. TMA is a compound known to induce atherosclerosis, which makes these taxa potential targets for microbiome-based interventions. I characterized a *Methanomassiliicoccales* MAG retrieved from samples of the TwinsUK cohort and used it, together with publicly available genomes, to perform phylogenetic analyses and genomic comparisons. I confirmed that the *Methanomassiliicoccales* order forms two large phylogenetic clades. Using publicly available metagenomes from environmental and non-human animal guts, I showed that these clades differ in their environmental preference, with some exceptions. Host-enriched taxa tended to have smaller genomes and possessed genes related to bile resistance and aromatic amino acid precursors. Using publicly available human gut metagenomes, I showed that these taxa were

absent from multiple populations, yet when present, they were correlated with bacteria known to produce TMA.

The characterization of microbial communities, such as the one presented in chapter II, relies heavily on database-dependent methods. The profiling of a metagenome sample requires, thus, that relevant genomes are included in the databases used. Such was not the case when my work on *Methanomassiliicoccales* started. In chapter III, I present Struo, a modular pipeline to assist the creation of custom databases of genes and genomes for commonly used metagenome profilers. The custom databases created with Struo provide a substantial increase in mappability of reads in synthetic and real metagenomic datasets. I employed databases created with this pipeline to obtain taxonomic and human gut metagenome samples in the work presented in chapters II and IV.

Chapter IV presents my efforts to assess the functional potential of the gut microbiome of a sample of community-dwelling Colombian adults. This with the aim of determining variation uniquely and robustly associated with obesity or cardiometabolic health by incorporating phenotypic data that help disentangle said conditions. I selected functional features linked to obesity, cardiovascular disease or type 2 diabetes from published studies in diverse populations, and tested their replication in the Colombian cohort. I performed shotgun metagenome sequencing from stool DNA to assess the gut microbiome of these subjects. Members of this cohort were very well characterized in terms of biochemical, anthropometric, medication and dietary data, which I sought to include in my analyses to reduce the effect of possible confounders. Moreover, these data allowed me to classify subjects according to their obesity and cardiometabolic status, and to determine which functions were associated with one condition while accounting for the other. Overall, I found that obesity drives the microbiome associations with cardiometabolic disease when both conditions are present.

An advantage of shotgun metagenome sequencing over 16S rRNA gene sequencing is that it allows to obtain functional data and the assembly of genomes of the members of the community, in addition to providing insights into the taxonomic profile of a microbiome.

The assembly of novel microbial genomes is of particular relevance for the study of the gut microbiome of human populations from low- and middle-income countries and other understudied populations. Thus, in chapter V I present the retrieval of metagenome-assembled genomes from the gut microbiome of Colombians using the metagenome sequence data I produced in chapter IV. I assessed the assembly quality, performed the taxonomic classification of this set of genomes, determined their taxonomic novelty and annotated functional and genomic features. In each of the chapters of the present work I relied on publicly available data; I consider it my duty to also contribute to the scientific community the data that I generated as part of my work. Therefore, I present this set of MAGs as a data descriptor.



# Publications

## Peer reviewed papers included in this dissertation

- de la Cuesta-Zuluaga J, Kelley ST, Chen Y, Escobar JS, Mueller NT, Ley RE, et al. **Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults.** mSystems. 2019 Jul;4(4). <http://dx.doi.org/10.1128/mSystems.00261-19>
- de la Cuesta-Zuluaga J, Ley RE, Youngblut ND. **Struo: a pipeline for building custom databases for common metagenome profilers.** Bioinformatics. 2020 Apr 1;36(7):2314–5. <http://dx.doi.org/10.1093/bioinformatics/btz899>
- de la Cuesta-Zuluaga J, Spector TD, Youngblut ND, Ley RE. **Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut.** mSystems. 2021 Feb 9;6(1). <http://dx.doi.org/10.1128/mSystems.00939-20>

## Advanced manuscripts included in this dissertation

- de la Cuesta-Zuluaga J, Youngblut ND, Escobar JS, Ley RE. **Obesity is the main driver of functional alterations of the gut microbiome in cardiometabolic disease.** In preparation.
- de la Cuesta-Zuluaga J, Youngblut ND, Escobar JS, Ley RE. **Gut metagenomes and assembled microbial genomes from human adults from urban cohorts from Colombia, South America.** In preparation.

## Peer reviewed papers, preprints, and advanced manuscripts not included in this dissertation

- Ruaud A, Esquivel-Elizondo S, de la Cuesta-Zuluaga J, Waters JL, Angenent LT, Youngblut ND, et al. **Syntrophy via Interspecies H<sub>2</sub> Transfer between**

- Christensenella* and *Methanobrevibacter* Underlies Their Global Cooccurrence in the Human Gut.** mBio. 2020 Feb 4;11(1). <http://dx.doi.org/10.1128/mBio.03235-19>
- Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, Walzer C, et al. **Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity.** mSystems. 2020 Nov 3;5(6). <http://dx.doi.org/10.1128/mSystems.01045-20>
  - Liu X, Sutter JL, de la Cuesta-Zuluaga J, Waters JL, Youngblut ND, Ley RE. **Reclassification of *Catabacter hongkongensis* as *Christensenella hongkongensis* comb. nov. based on whole genome analysis.** Int J Syst Evol Microbiol . 2021 Apr;71(4). <http://dx.doi.org/10.1099/ijsem.0.004774>
  - Youngblut ND, de la Cuesta-Zuluaga J, Ley RE. **Incorporating genome-based phylogeny and functional similarity into diversity assessments helps to resolve a global collection of human gut metagenomes.** Environmental microbiology. 2022. <https://doi.org/10.1111/1462-2920.15910>
  - Sutter JS, Walters WA, de la Cuesta-Zuluaga J, Welter DK, Liu X, Youngblut ND, Ley RE. **Mapping function onto phylogeny in rhizosphere *Pseudomonas*.** In preparation.
  - Clasen S, Bell M, Borbón A, Henseler Z, de la Cuesta-Zuluaga J, Lee D, et al. **Silent recognition of flagellins by Toll-like receptor 5 allows immune evasion by human gut commensal bacteria.** In preparation.
  - Salazar-Jaramillo L, Chica LA, de la Cuesta-Zuluaga J, Cadavid M, Velásquez-Mejía EP, Ley RE, Reyes A, Escobar JS. **Genetic diversity within *Clostridia* assessed by alternative genetic markers and its association with human phenotypes.** In preparation.



# Declaration of contributions in collaborative works

EBERHARD KARLS  
UNIVERSITÄT  
TÜBINGEN



Mathematisch-  
Naturwissenschaftliche  
Fakultät

Erklärung nach § 5 Abs. 2 Nr. 8 der Promotionsordnung der Math.-Nat.  
Fakultät

**-Anteil an gemeinschaftlichen Veröffentlichungen-  
Nur bei kumulativer Dissertation erforderlich!**

**Declaration according to § 5 Abs. 2 No. 8 of the PhD regulations of the  
Faculty of Science  
-Collaborative Publications-  
For Cumulative Theses Only!**

Last Name, First Name: de la Cuesta Zuluaga, Juan Jacobo.

## List of Publications

1. de la Cuesta-Zuluaga J, Kelley ST, Chen Y, Escobar JS, Mueller NT, Ley RE, et al. **Age- and Sex Dependent Patterns of Gut Microbial Diversity in Human Adults.** mSystems. 2019 Jul;4(4). <http://dx.doi.org/10.1128/mSystems.00261-19>
2. de la Cuesta-Zuluaga J, Ley RE, Youngblut ND. **Struo: a pipeline for building custom databases for common metagenome profilers.** Bioinformatics. 2020 Apr 1;36(7):2314–5. <http://dx.doi.org/10.1093/bioinformatics/btz899>
3. de la Cuesta-Zuluaga J, Spector TD, Youngblut ND, Ley RE. **Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut.** mSystems. 2021 Feb 9;6(1). <http://dx.doi.org/10.1128/mSystems.00939-20>
4. de la Cuesta-Zuluaga J, Youngblut ND, Escobar JS, Ley RE. **Obesity is the main driver of functional alterations of the gut microbiome in cardiometabolic disease.** In preparation.
5. de la Cuesta-Zuluaga J, Youngblut ND, Escobar JS, Ley RE. **Gut metagenomes and assembled microbial genomes from human adults from urban cohorts from Colombia, South America.** In preparation.



Nr.	Accepted publication yes/no	List of authors	Position of candidate in list of authors	Scientific ideas by the candidate (%)	Data generation by the candidate (%)	Analysis and Interpretation by the candidate (%)	Paper writing done by the candidate (%)
1	Yes	de la Cuesta-Zuluaga J, Kelley ST, Chen Y, Escobar JS, Mueller NT, Ley RE	First	50 %	50 %	66 %	66 %
2	Yes	de la Cuesta-Zuluaga J, Ley RE, Youngblut ND	First	80 %	75 %	90 %	90 %
3	Yes	de la Cuesta-Zuluaga J, Spector TD, Youngblut ND	First	80 %	90 %	90 %	90 %
4	No	de la Cuesta-Zuluaga J, Youngblut ND, Escobar JS, Ley RE	First	85 %	90 %	90 %	95 %
5	No	de la Cuesta-Zuluaga J, Youngblut ND, Escobar JS, Ley RE	First	90 %	90 %	90 %	95 %

I confirm that the above-stated is correct.

08.02.2022 , \_\_\_\_\_

Date, Signature of the ca

I/We certify that the above-stated is correct.

\_\_\_\_\_  
Date, Signature of the doctoral committee or at least of one of the supervisors

# Journal reuse policies

The following are the reuse and distribution policies of the journals in which the articles included in this thesis were published.

## mSystems - American Society for Microbiology

- Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults
- Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut

"(...) authors retain copyright. To allow for maximum dissemination of research, articles in mSystems are published under the Creative Commons CC BY license, which allows unrestricted reuse of the material with proper attribution. Users have the right to read, download, copy, distribute, print, search, or link to the full texts of these articles."

American Society for Microbiology. mSystems FAQs. Cited 2021 Oct 1. <https://journals.asm.org/journal/msystems/faq>

## Bioinformatics - Oxford University Press

- Struo: a pipeline for building custom databases for common metagenome profilers

"After publication you may reuse the following portions of your content without obtaining formal permission for the activities expressly listed below (...) inclusion within your thesis or dissertation"

Oxford University Press. Author Reuse and Self-Archiving. Cited 2021 Oct 1. <https://global.oup.com/academic/rights/permissions/autperm/?cc=de&lang=en&>



# Chapter I

## Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults

The content of this chapter has been published as:

**de la Cuesta-Zuluaga J**, Kelley ST, Chen Y, Escobar JS, Mueller NT, Ley RE, et al.

mSystems. 2019 Jul;4(4).

Available from: <http://dx.doi.org/10.1128/mSystems.00261-19>

See *appendix I*

### Abstract

Gut microbial diversity changes throughout the human life span and is known to be associated with host sex. We investigated the association of age, sex, and gut bacterial alpha diversity in three large cohorts of adults from four geographical regions: subjects from the United States and United Kingdom in the American Gut Project (AGP) citizen-science initiative and two independent cohorts of Colombians and Chinese. In three of the four cohorts, we observed a strong positive association between age and alpha diversity in young adults that plateaued after age 40 years. We also found sex-dependent differences that were more pronounced in younger adults than in middle-aged adults, with women having higher alpha diversity than men. In contrast to the other three cohorts, no association of alpha diversity with age or sex was observed in the Chinese cohort. The association of alpha diversity with age and sex remained after adjusting for cardiometabolic parameters in the Colombian cohort and antibiotic usage in the AGP cohort. We further attempted to predict the microbiota age in individuals using a machine-learning approach for the men and women in each cohort. Consistent with our alpha-diversity-based findings, U.S. and U.K. women had a significantly higher predicted microbiota age than men, with a reduced difference being seen

above age 40 years. This difference was not observed in the Colombian cohort and was observed only in middle-aged Chinese adults. Together, our results provide new insights into the influence of age and sex on the biodiversity of the human gut microbiota during adulthood while highlighting similarities and differences across diverse cohorts.

## Contributions

Project conception and outline: JdlCZ, STK, VGT, JSE, NTM. Project implementation and coordination: JdlCZ, STK, DM. 16S rRNA amplicon data curation: JdlCZ, STK, YC, DM. Statistical analyses: JdlCZ, STK, YC. Statistical advice: JSE, NTM, DM. Machine learning: SH, ADS. Supervision, discussion of analysis and interpretation: REL, RK, VGT. Manuscript preparation: JdlC, STK. Manuscript review: REL, JSE, NTM, DM, RK, VGT. Comments: all authors.

## Chapter II

### Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut

The content of this chapter has been published as:

**de la Cuesta-Zuluaga J**, Spector TD, Youngblut ND, Ley RE.

mSystems. 2021 Feb 9;6(1).

Available from: <http://dx.doi.org/10.1128/mSystems.00939-20>

See *appendix II*

#### Abstract

Archaea of the order *Methanomassiliicoccales* use methylated amines such as trimethylamine as the substrates for methanogenesis. They form two large phylogenetic clades and reside in diverse environments, from soil to the human gut. Two genera, one from each clade, inhabit the human gut: *Methanomassiliicoccus*, which has one cultured representative, and “*Candidatus* Methanomethylophilus,” which has none. Questions remain regarding their distribution across biomes and human populations, their association with other taxa in the gut, and whether host genetics correlate with their abundance. To gain insight into the *Methanomassiliicoccales* clade, particularly its human-associated members, we performed a genomic comparison of 72 *Methanomassiliicoccales* genomes and assessed their presence in metagenomes derived from the human gut (n = 4,472, representing 22 populations), nonhuman animal gut (n = 145), and nonhost environments (n = 160). Our analyses showed that all taxa are generalists; they were detected in animal gut and environmental samples. We confirmed two large clades, one enriched in the gut and the other enriched in the environment, with notable exceptions. Genomic adaptations to the gut include genome reduction and genes involved in the shikimate pathway and bile resistance.

Genomic adaptations differed by clade, not habitat preference, indicating convergent evolution between the clades. In the human gut, the relative abundance of *Methanomassiliicoccales* spp. correlated with trimethylamine-producing bacteria and was unrelated to host genotype. Our results shed light on the microbial ecology of this group and may help guide *Methanomassiliicoccales*-based strategies for trimethylamine mitigation in cardiovascular disease.

## Contributions

Project conception and outline: REL, NDY, TDS. Metagenome assembly: JdlCZ, NDY. Comparative genomics, phylogenetic analysis and metagenome profiling: JdlCZ. Public data retrieval: JdlCZ, NDY. Bioinformatics and statistics support: NDY. Supervision, discussion of analysis and interpretation: REL, NDY. Manuscript preparation: JdlC. Manuscript review: REL, NDY. Comments: all authors.



## Chapter III

### Struo: a pipeline for building custom databases for common metagenome profilers

The content of this chapter has been published as:

**de la Cuesta-Zuluaga J**, Ley RE, Youngblut ND.

Bioinformatics. 2020 Apr 1;36(7):2314–5.

Available from: <http://dx.doi.org/10.1093/bioinformatics/btz899>

*See appendix III*

#### Abstract

Taxonomic and functional information from microbial communities can be efficiently obtained by metagenome profiling, which requires databases of genes and genomes to which sequence reads are mapped. However, the databases that accompany metagenome profilers are not updated at a pace that matches the increase in available microbial genomes, and unifying database content across metagenome profiling tools can be cumbersome. To address this, we developed Struo, a modular pipeline that automatizes the acquisition of genomes from public repositories and the construction of custom databases for multiple metagenome profilers. The use of custom databases that broadly represent the known microbial diversity by incorporating novel genomes results in a substantial increase in mappability of reads in synthetic and real metagenome datasets.

#### Contributions

Project conception and outline: JdlC, NDY. Assessment of genome collections: JdlC. Pipeline benchmark and statistical analysis: JdlC. Implementation of snakemake workflow:

NDY. Manuscript preparation: JdlC. Manuscript review: REL, NDY. Comments: all authors.

# Chapter IV

## Obesity is the main driver of functional alterations of the gut microbiome in cardiometabolic disease

The content of this chapter is yet to be published.

**de la Cuesta-Zuluaga J**, Youngblut ND, Escobar JS, Ley RE.

*In preparation.*

*See appendix IV*

### Abstract

The discovery of distinct links between obesity (OB) and the cardiometabolic health status (CHS) with the gut microbiome is hindered by the overlap between these conditions. Moreover, differences in study design and covariates used encumber the comparison of study outcomes. Here, we describe features of gut microbiome function associated independently with OB or CHS in a cohort of adults; and test for the replication of associations previously reported for microbiome and OB/CHS. We enrolled 459 deeply-phenotyped Colombians from whom we obtained 408 gut metagenomes. We measured three OB indices and classified individuals according to their CHS using blood biochemistry and anthropometric data. We evaluated the association of 136 KEGG modules and 2 653 orthologs previously linked with OB, cardiovascular disease or diabetes. Medication use, city, sex and age were included as covariates. We found that metagenome sequence diversity negatively correlated with OB; subjects with CHS had lower diversity than healthy subjects with similar OB levels. OB explained a higher proportion of variance for sequence diversity and functional beta-diversity. Similarly, more modules and orthologs were uniquely associated with OB than with CHS or shared by both conditions. The microbiome potential of diseased individuals in both conditions showed a decreased fermentative ability and an increased response to oxygen.

Disease-linked features were mainly contributed by members of Proteobacteria. Our results suggest that OB drives the microbiome associations with CHS when both are present

## Contributions

Project conception and outline: JdlCZ, REL, JSE. Sample processing and metagenome sequencing: JdlCZ. Sequence data processing, host data processing, metagenome profiling, diversity indices calculation, literature review, selection of features to include in analysis, statistical analyses: JdlCZ. Bioinformatics support: NDY. Supervision, discussion of analysis and interpretation: REL, JSE, NDY. Manuscript preparation: JdlC.

## Chapter V

### Gut metagenomes and assembled microbial genomes from human adults from urban cohorts from Colombia, South America

The content of this chapter is yet to be published.

**de la Cuesta-Zuluaga J**, Youngblut ND, Escobar JS, Ley RE.

*In preparation.*

See *appendix V*

#### Abstract

The human gut microbiome is an important mediator of multiple physiological processes. The identification of generalizable associations and mechanistic links between this microbial community and human health requires the study of diverse human populations. Yet the microbiomes of subjects from low- and middle-income countries are understudied. Here, we present a set of shotgun gut metagenomes of 459 deeply-phenotyped male and female adults (18-62 years old) living in geographically distinct urban areas of Colombia (South America), studied in the context of westernization and the epidemiological transition. We assembled these metagenomes and retrieved 2 266 medium- and high-quality metagenome-assembled genomes (MAGs), which we annotated, classified taxonomically, and compared to large collections of microbial genomes. The metagenomes, MAGs, and accompanying host data presented here will benefit initiatives looking into the human microbiome's diversity and its role in westernization, nutrition, obesity and cardiometabolic disease.

## Contributions

Project conception and outline: JdlCZ, REL, JSE. Sample processing and metagenome sequencing: JdlCZ. Sequence data processing, metagenome assembly, genome quality assessment and taxonomic characterization: JdlCZ. Bioinformatics support: NDY. Supervision, discussion of analysis and interpretation: REL, JSE, NDY. Manuscript preparation: JdlC. Manuscript review: REL, JSE, NDY.

## Discussion and outlook

The present dissertation comprises work of various kinds that revolve around a central theme, namely, the use of culture-independent methods and computational approaches to study the human gut microbiome in the context of obesity and cardiometabolic diseases. In this section I will not delve into the details of any particular work since they are discussed in their corresponding manuscript. Instead, I will discuss their contribution and shortcomings as a whole.

### Contribution of this thesis

The collection of works I present here can be grouped into three broad categories that are closely intertwined, and which encompass the different tasks a computational biologist can perform.

First, obtaining biological knowledge about the studied phenomena. The works presented in chapters I, II and IV aimed to link the composition and diversity of the gut microbiome, or the genomic potential of certain of its members, to host phenotypes. In particular, those related to obesity and cardiometabolic health. As such, I described patterns of microbiota diversity in multiple populations and assessed whether these patterns were associated with host health; characterized microorganisms that are understudied but have the potential for establishing interventions to tackle cardiovascular disease; and assessed the generality of associations between the functional potential of the intestinal microbiome and non-communicable diseases in an understudied population.

Second, the generation of new data from microbial communities, and the assembly and characterization of novel microbial genomes. To perform the analyses reported in chapters II and IV, I had to generate new shotgun metagenome sequences. I then used these sequence reads to functionally and taxonomically profile the microbiome of the study participants. Moreover, these sequences also served as the basis for the assembly of microbial genomes reported in chapters II and V. These data enrich our knowledge of microbial

diversity, and importantly, are made available to the scientific community for others to use to answer questions that go beyond the objectives of this dissertation.

Third, the development and benchmarking of tools, which is closely linked to the generation of new data. Just as the data I generated are publicly available, I made use of data contributed by other researchers to develop databases that facilitate the study of microbiomes. In turn, these databases were instrumental in maximizing the microbial diversity I was able to detect in the analysis of taxonomic and functional profiles, thus completing the circle.

I expect the results I present in my thesis to help guide and inform future studies, be them cohort studies that take into account the covariates they measure and how they are included in the analyses; studies that use metagenomic methods to assess microbial communities incorporating custom databases to maximize the microorganisms detected; intervention studies that explore novel mechanisms to treat cardiometabolic disease and obesity; or *in vitro* studies that investigate how microbial metabolism may influence host gut homeostasis.

## Pitfalls and shortcomings

As with any scientific endeavor, the works presented in this thesis are not without limitations. I will briefly address some of them.

### Cross-sectional, computational and correlational studies

The manuscript presented in chapter II was desk rejected the first time it was submitted to a peer-reviewed journal. The only comment from the editor read ‘*This informatics study is correlational*’. This sentence did not sit very well with me at first, as it seemed aggressively obvious and an insufficient argument to reject a manuscript. I still think it is self-evident, however, I now consider it an assertion worth discussing.

The findings from chapters I, II and IV are derived from cross-sectional data from a single (chapter IV) or multiple human populations (I and II). Cross-sectional data provides a



snapshot of the microbiome of a subject at a single point in time; they do not account for the temporal variability of the microbial community (Vandeputte *et al.*, 2021). The gut microbiome is a complex and dynamic system that can vary over time as a response to shifts in diet, consumption of medications or disease progression (Johnson *et al.*, 2019). The temporal variation on biomarkers of diversity, species, and potential function presented in these chapters could be assessed by future studies focusing on the response to stimuli of interest, whether dietary, pharmacological or lifestyle related. Moreover, as the editor correctly pointed out, the kind of data only allowed me to report associations between the diversity, microbial and functional features measured and host phenotypes, not to provide causal inference.

Indeed, the computational and correlational nature of these works makes them hypothesis-proposing rather than hypothesis-testing. This is not a liability, rather it is a strength. Data exploration is the process of finding correlations and patterns that can later be tested for causality (Yanai and Lercher, 2020), and while there are some who treat such approaches with contempt, it is undeniable that computational thinking and methods are central to current biology (Markowitz, 2017). The aim of this sort of studies is to provide a selection of taxa, functions or indices that are strongly correlated with the phenotype evaluated, in other words, to put forward hypotheses. This set can, in turn, inform the design of hypothesis-testing mechanistic or intervention studies that elucidate how the microbiome is affecting or being affected by the host.

## Impact of medications and other confounders on the gut microbiome

As mentioned in the introduction and discussed in the relevant sections of chapters I, II and IV, the exclusion of subjects with various conditions such as cancer, neurological diseases, gastrointestinal diseases, or who consumed antibiotics from the analyses allowed me to rule out that these diseases were responsible for the observed associations. Likewise, the inclusion of covariates such as age, sex, and geographic origin, in addition to the consumption of medications for hypertension, diabetes or dyslipidemia allowed me to reduce the potential for confounding (Forslund *et al.*, 2021). Nevertheless, the growth of commensal

microorganisms in the gut microbiome is inhibited by drugs with human targets of all therapeutic classes, as demonstrated by *in vitro* studies (Maier *et al.*, 2018). Since my analyses only consider antibiotics, and medications with direct relevance to obesity and cardiometabolic disease without considering their dosage, I am unable to rule out residual confounding from other medications and variables not measured or not included in the analyses (Forslund *et al.*, 2021).

I sought to be judicious with the inclusion of covariates that could influence the outcome of my statistical analyses. The number and nature of the variables I included depended on the quality of information available: from the information-rich Colombian cohort to the relatively poor public data I used to characterize the distribution of *Methanomassiliicoccales* in multiple populations. I certainly did the best I could with what I had at my disposal. I expect that as we gain knowledge about the microbiome, better controlled studies will be performed, so that the noise introduced by confounding variables is accounted for, and the associations of the microbiome with the conditions studied are robustly elucidated.

## The limitations of databases and methods that rely on them

Metagenomics involves sequencing the total DNA of a microbial community; sequencing reads are then mapped against databases of genes and genomes (Beghini *et al.*, 2021). This allows to simultaneously investigate the taxonomic composition of the community and the metabolic potential encoded by the microbes present (de la Cuesta-Zuluaga, Ley and Youngblut, 2020). Similarly, the taxonomic assignment of 16S rRNA gene amplicons also requires the use of databases against which to contrast the sequences obtained (de la Cuesta-Zuluaga and Escobar, 2016).

Database dependency influences the ability of different algorithms to identify and annotate genes or taxa present in a microbial community, and flaws in the databases will certainly lead to flaws in the assessment of the microbiome (Nasko *et al.*, 2018). The issue is exacerbated when undescribed microorganisms dominate the community under study. This

pitfall is currently unavoidable: at some point all analyses require adding a label to the unit of study, be it a taxonomy to an OTU or a MAG, or a protein name to a predicted CDS. A database of some sort is thus required to match a name or category with a unit of study.

Fortunately, great strides have been recently made regarding the creation, maintenance and expansion of public databases of microbial genomes obtained by culture or metagenomic assembly. The proGenomes database (Mende *et al.*, 2020) and the genome taxonomy database (GTDB) (Parks *et al.*, 2022) did not exist or were in an incipient state when I started my PhD. Yet they now comprise hundreds of thousands of bacterial and archaeal genomes which correspond to tens of thousands of microbial species. These databases are continuously expanding. For example, the latest release of the GTDB (release 202, as of the writing of this discussion) covers 258 406 genomes belonging to 47 894 species clusters, as defined by their standardized taxonomic ranking (Parks *et al.*, 2018). These values represent an increase of 33 % in the number of genomes and 50 % in species clusters compared to the previous release (release 95) (Australian Centre for Ecogenomics, 2021). Moreover, the tools that leverage these databases are also under continuous development: *Struo*, the pipeline for the creation of custom databases for taxonomic profilers I presented in chapter III, continued its development under the responsibility of other researchers in the laboratory (Youngblut and Ley, 2021). Therefore, while the issue of database dependence will affect microbiome studies for the foreseeable future, its pervasiveness can be alleviated by the broad and systematic inclusion of novel genomes from the ever expanding public databases.

## Metagenomics vs Metatranscriptomics or Metaproteomics

The presence of a gene in the genome of a given taxon does not necessarily imply that said gene is expressed, therefore, metagenomics can only inform us about the potential of the microbial community, but not about how much of that potential is fulfilled. To obtain insights about the activity of the microbial community and how it changes in response to a given stimulus or condition, the use of metatranscriptomics or metaproteomics is necessary. These approaches directly measure the transcripts, proteins and metabolites actually

expressed by the members of the microbiome (Shakya, Lo and Chain, 2019). In this section I refer to the functional information encoded by the metagenome as the ‘potential profile’. Conversely, I refer to the information derived from metatranscriptomics or metaproteomics as the ‘active profile’.

The active profile of the microbiome will vary in a context-dependent manner. Factors like host genetics and metabolism, immune status or dietary patterns will impact the activity landscape of the gut microbiome (Tanca *et al.*, 2017). In other words, it only provides a snapshot of the activity of the community at a specific point in time. Different stimuli can temporarily influence the activity of the microbial community without altering its structure (Maurice, Haiser and Turnbaugh, 2013; Franzosa *et al.*, 2014). Changes in the composition of the potential and taxonomic profile require perturbations of greater duration or intensity.

The information carried by the active and potential profiles differ, since by definition, the expressed transcripts and proteins correspond to a fraction of what is encoded by the genomes of the members of the community (Tanca *et al.*, 2017). Certain features can be missed because they were not expressed at the moment of sampling. Moreover, the abundance of genes in a metagenome is only moderately correlated with their mRNA expression (Franzosa *et al.*, 2014) and weakly correlated with the protein levels (Tanca *et al.*, 2017). It has been suggested that the positive correlation between the potential and active profiles, that is, that the abundance of a feature in the metagenome is a determinant of its corresponding expression, indicates that most genes across the majority of microbial genomes are transcribed at similar, relatively fixed rates (Franzosa *et al.*, 2014).

This pitfall is not exclusive to metagenomics; there are also discrepancies between the abundance of microbial taxa and their metabolic activity. In particular, there is a large difference in the abundance of taxa from the phyla *Bacteroidetes* and *Firmicutes* and their respective activity as measured by metaproteomics (Tanca *et al.*, 2017). Thus, this caveat should also be extended to the taxonomic analysis of microbial communities using 16S rRNA gene sequencing or shotgun metagenomics.

Its shortcomings notwithstanding, metagenomics is a useful approach since it provides a broad overview of the functions the microbiota can perform. The abundance measurements, in turn, provide reasonable estimations of the functions that are actually being performed. Moreover, metagenome sequencing allows us to perform ecological and evolutionary assessments of the microbial community by the use of strain-level taxonomic profiles and metagenome-assembled genomes. This, however, is beyond the scope of the present discussion.

## Promises and future challenges of microbiome science

Microbiome science has bloomed over the last two decades. Its link with human development, health, and evolution, together with the potential of manipulating this microbial community are some of the reasons why it has attracted attention from basic, applied, and translational science (Clavel *et al.*, 2022).

Nevertheless, a dose of skepticism is needed to ward off hype and its detrimental effects; this is the most important challenge the field currently faces (Hanage, 2014). The first step to avoid this, and to which this dissertation contributes, is to employ study designs and statistical frameworks in cross-sectional studies that narrow down the set of microbial features to be evaluated in subsequent studies (Vujkovic-Cvijin *et al.*, 2020).

The next challenges stem from the aforementioned robust set of microbial features linked with the host phenotype. The logical step after the identification of microbial associations is to elucidate the molecular mechanisms by which the microbes and the microbial-derived metabolites regulate host physiology; this will provide with insights into the genetic, biochemical, ecological and evolutionary dynamics at play between hosts and their microbes (Suzuki and Ley, 2020; Zimmermann *et al.*, 2021). In turn, deeper knowledge of the mechanisms by which host and microbes interact will contribute to the identification of targets for intervention and thus, the development of tools to be used in clinical and nutritional settings. Such tools could include tests for the diagnosis or monitoring of disease (Schlaberg, 2020), or treatments targeting specific conditions (Sorbara and Pamer, 2022). The

implementation of therapeutic tools will also require overcoming diverse regulatory obstacles (Cordailat-Simmons, Rouanet and Pot, 2020). For this, the establishment of tractable methods that allow systematic and controlled tests are required (Mirzayi *et al.*, 2021), so that the relevant quality, efficacy and safety standards are met (Cordailat-Simmons, Rouanet and Pot, 2020)

The present dissertation illustrates how culture-independent methods can be used to study microbial communities and specific taxa within them. However, the use of isolated microbes will be key to overcome some of the aforementioned issues, and is itself a challenge to be faced. The development of high-throughput methods for the cultivation of microbes (Forster *et al.*, 2019; Zou *et al.*, 2019), together with the establishment of molecular and bioinformatic workflows that allow their taxonomic classification and functional characterization will be key (Meyer *et al.*, 2021). Likewise, the deployment of the required laboratory and computational infrastructure that enables the storage, distribution and analysis of these materials and data will be crucial (Stephens *et al.*, 2015).

The aforementioned challenges are mostly of technical nature; one must also consider those related to openness in data sharing and equity in the study of overlooked populations mentioned in the introduction of this dissertation.

Many challenges lie ahead, however, there are even more interesting research avenues and novel findings waiting to be discovered. The combination of high-throughput sequence-based methods with improved culture techniques and novel computational approaches will certainly lead to insights about the functioning of the microbial community, its adaptation to living in a host, the metabolic processes it carries out and how it is associated with multiple human phenotypes in different populations. If the aforementioned hurdles are overcome, it is not unreasonable to foresee a scenario where microbiome-based applications are used on a routine basis in clinical and nutritional settings.

There has never been a better time in history to be a microbiologist, a computational biologist or a combination thereof. I am personally excited for what we will learn about our relationship with microbes in the years to come.

# Bibliography

Abdill, R.J., Adamowicz, E.M. and Blekhman, R. (2022) 'Public human microbiome data are dominated by highly developed countries', *PLoS biology*, 20(2), p. e3001536. doi:10.1371/journal.pbio.3001536.

Almeida, A. *et al.* (2019) 'A new genomic blueprint of the human gut microbiota', *Nature*, 568(7753), pp. 499–504. doi:10.1038/s41586-019-0965-1.

Almeida, A. *et al.* (2021) 'A unified catalog of 204,938 reference genomes from the human gut microbiome', *Nature biotechnology*, 39(1), pp. 105–114. doi:10.1038/s41587-020-0603-3.

Armour, C.R. *et al.* (2019) 'A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome', *mSystems*, 4(4). doi:10.1128/mSystems.00332-18.

Aßhauer, K.P. *et al.* (2015) 'Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data: Fig. 1.', *Bioinformatics*, 31(17), pp. 2882–2884. doi:10.1093/bioinformatics/btv287.

Australian Centre for Ecogenomics (2021) *GTDB Release 202 Statistics, Genome Taxonomy Database*. Available at: <https://gtdb.ecogenomic.org/stats/r202> (Accessed: 21 December 2021).

Bağcı, C., Patz, S. and Huson, D.H. (2021) 'DIAMOND+MEGAN: Fast and Easy Taxonomic and Functional Analysis of Short and Long Microbiome Sequences', *Current protocols*, 1(3), p. e59. doi:10.1002/cpz1.59.

Banerjee, S., Schlaeppli, K. and van der Heijden, M.G.A. (2018) 'Keystone taxa as drivers of microbiome structure and functioning', *Nature reviews. Microbiology*, 16(9), pp. 567–576. doi:10.1038/s41579-018-0024-1.

Beghini, F. *et al.* (2021) 'Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3', *eLife*, 10. doi:10.7554/eLife.65088.

Berg, G. *et al.* (2020) 'Microbiome definition re-visited: old concepts and new challenges', *Microbiome*, 8(1), p. 103. doi:10.1186/s40168-020-00875-0.

den Besten, G. *et al.* (2015) 'Short-Chain Fatty Acids Protect Against High-Fat Diet-Induced

Obesity via a PPAR $\gamma$ -Dependent Switch From Lipogenesis to Fat Oxidation', *Diabetes*, 64(7), pp. 2398–2408. doi:10.2337/db14-1213.

Biagi, E. *et al.* (2010) 'Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians', *PloS one*, 5(5), p. e10667. doi:10.1371/journal.pone.0010667.

Bolyen, E. *et al.* (2019) 'Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2', *Nature biotechnology*, 37(8), pp. 852–857. doi:10.1038/s41587-019-0209-9.

Brown, J.M. and Hazen, S.L. (2018) 'Microbial modulation of cardiovascular disease', *Nature reviews. Microbiology*, 16(3), pp. 171–181. doi:10.1038/nrmicro.2017.149.

Buscemi, J. *et al.* (2018) 'Associations between fiber intake and Body Mass Index (BMI) among African-American women participating in a randomized weight loss and maintenance trial', *Eating behaviors*, 29, pp. 48–53. doi:10.1016/j.eatbeh.2018.02.005.

Byndloss, M.X. and Bäumler, A.J. (2018) 'The germ-organ theory of non-communicable diseases', *Nature reviews. Microbiology*, 16(2), pp. 103–110. doi:10.1038/nrmicro.2017.158.

Clavel, T. *et al.* (2022) 'Next steps after 15 stimulating years of human gut microbiome research', *Microbial biotechnology*, 15(1), pp. 164–175. doi:10.1111/1751-7915.13970.

Cordaillat-Simmons, M., Rouanet, A. and Pot, B. (2020) 'Live biotherapeutic products: the importance of a defined regulatory framework', *Experimental & molecular medicine*, 52(9), pp. 1397–1406. doi:10.1038/s12276-020-0437-6.

Cox, A.J., West, N.P. and Cripps, A.W. (2015) 'Obesity, inflammation, and the gut microbiota', *The lancet. Diabetes & endocrinology*, 3(3), pp. 207–215. doi:10.1016/S2213-8587(14)70134-2.

de la Cuesta-Zuluaga, J. *et al.* (2017) 'Metformin Is Associated With Higher Relative Abundance of Mucin-Degrading Akkermansia muciniphila and Several Short-Chain Fatty Acid – Producing Microbiota in the Gut', *Diabetes care*, 40(1), pp. 54–62. doi:10.2337/dc16-1324.

de la Cuesta-Zuluaga, J. *et al.* (2018) 'Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization', *Scientific reports*, 8(1), p. 11356. doi:10.1038/s41598-018-29687-x.

de la Cuesta-Zuluaga, J. and Escobar, J.S. (2016) 'Considerations For Optimizing



Microbiome Analysis Using a Marker Gene', *Frontiers in Nutrition*, 3(26). doi:10.3389/fnut.2016.00026.

de la Cuesta-Zuluaga, J., Ley, R.E. and Youngblut, N.D. (2020) 'Struo: a pipeline for building custom databases for common metagenome profilers', *Bioinformatics*, 36(7), pp. 2314–2315. doi:10.1093/bioinformatics/btz899.

David, L.A. *et al.* (2014) 'Diet rapidly and reproducibly alters the human gut microbiome', *Nature*, 505(7484), pp. 559–563. doi:10.1038/nature12820.

De Filippis, F., Pasolli, E. and Ercolini, D. (2020) 'Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease', *Current biology: CB*, 30(24), pp. 4932–4943.e4. doi:10.1016/j.cub.2020.09.063.

Ding, S. and Lund, P.K. (2011) 'Role of intestinal inflammation as an early event in obesity and insulin resistance', *Current opinion in clinical nutrition and metabolic care*, 14(4), pp. 328–333. doi:10.1097/MCO.0b013e3283478727.

Dinsa, G.D. *et al.* (2012) 'Obesity and socioeconomic status in developing countries: a systematic review', *Obesity reviews: an official journal of the International Association for the Study of Obesity*, 13(11), pp. 1067–1079. doi:10.1111/j.1467-789X.2012.01017.x.

Duvallet, C. *et al.* (2017) 'Meta-analysis of gut microbiome studies identifies disease-specific and shared responses', *Nature communications*, 8(1), p. 1784. doi:10.1038/s41467-017-01973-8.

Eckert, E.M. *et al.* (2020) 'Every fifth published metagenome is not available to science', *PLoS biology*, 18(4), p. e3000698. doi:10.1371/journal.pbio.3000698.

Forslund, K. *et al.* (2013) 'Country-specific antibiotic use practices impact the human gut resistome', *Genome research*, 23(7), pp. 1163–1169. doi:10.1101/gr.155465.113.

Forslund, K. *et al.* (2015) 'Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota', *Nature*, 528(7581), pp. 262–266. doi:10.1038/nature15766.

Forslund, S.K. *et al.* (2021) 'Combinatorial, additive and dose-dependent drug-microbiome associations', *Nature*, 600(7889), pp. 500–505. doi:10.1038/s41586-021-04177-9.

Forster, S.C. *et al.* (2019) 'A human gut bacterial genome and culture collection for improved metagenomic analyses', *Nature biotechnology*, 37(2), pp. 186–192. doi:10.1038/s41587-018-0009-7.

- Franzosa, E.A. *et al.* (2014) ‘Relating the metatranscriptome and metagenome of the human gut’, *Proceedings of the National Academy of Sciences*, 111(22), pp. E2329–E2338. doi:10.1073/pnas.1319284111.
- Franzosa, E.A. *et al.* (2018) ‘Species-level functional profiling of metagenomes and metatranscriptomes’, *Nature methods*, 15(11), pp. 962–968. doi:10.1038/s41592-018-0176-y.
- García-Vega, Á.S. *et al.* (2020) ‘Diet Quality, Food Groups and Nutrients Associated with the Gut Microbiota in a Nonwestern Population’, *Nutrients*, 12(10). doi:10.3390/nu12102938.
- Geng, J. *et al.* (2018) ‘Trimethylamine N-oxide promotes atherosclerosis via CD36-dependent MAPK/JNK pathway’, *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 97, pp. 941–947. doi:10.1016/j.biopha.2017.11.016.
- Ghosh, T.S. *et al.* (2020) ‘Adjusting for age improves identification of gut microbiome alterations in multiple diseases’, *eLife*, 9. doi:10.7554/eLife.50240.
- Goodrich, J.K. *et al.* (2014) ‘Conducting a microbiome study’, *Cell*, 158(2), pp. 250–262. doi:10.1016/j.cell.2014.06.037.
- Grover, S.A. *et al.* (2015) ‘Years of life lost and healthy life-years lost from diabetes and cardiovascular disease in overweight and obese people: a modelling study’, *The lancet. Diabetes & endocrinology*, 3(2), pp. 114–122. doi:10.1016/S2213-8587(14)70229-3.
- Hadrévi, J., Søgaard, K. and Christensen, J.R. (2017) ‘Dietary Fiber Intake among Normal-Weight and Overweight Female Health Care Workers: An Exploratory Nested Case-Control Study within FINALE-Health’, *Journal of nutrition and metabolism*, 2017, p. 1096015. doi:10.1155/2017/1096015.
- Hanage, W.P. (2014) ‘Microbiology: Microbiome science needs a healthy dose of scepticism’, *Nature*, 512(7514), pp. 247–248. doi:10.1038/512247a.
- He, Y. *et al.* (2018) ‘Regional variation limits applications of healthy gut microbiome reference ranges and disease models’, *Nature medicine*, 24(10), pp. 1532–1535. doi:10.1038/s41591-018-0164-x.
- Hillmann, B. *et al.* (2018) ‘Evaluating the Information Content of Shallow Shotgun Metagenomics’, *mSystems*, 3(6). doi:10.1128/mSystems.00069-18.
- Horz, H.-P. and Conrads, G. (2010) ‘The discussion goes on: What is the role of

- Euryarchaeota in humans?', *Archaea*, 2010, p. 967271. doi:10.1155/2010/967271.
- Integrative HMP (iHMP) Research Network Consortium (2019) 'The Integrative Human Microbiome Project', *Nature*, 569(7758), pp. 641–648. doi:10.1038/s41586-019-1238-8.
- Jackson, M.A. *et al.* (2016) 'Proton pump inhibitors alter the composition of the gut microbiota', *Gut*, 65(5), pp. 749–756. doi:10.1136/gutjnl-2015-310861.
- Janzon, A. *et al.* (2019) 'Interactions between the Gut Microbiome and Mucosal Immunoglobulins A, M, and G in the Developing Infant Gut', *mSystems*, 4(6). doi:10.1128/mSystems.00612-19.
- Johnson, A.J. *et al.* (2019) 'Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans', *Cell host & microbe*, 25(6), pp. 789–802.e5. doi:10.1016/j.chom.2019.05.005.
- Johnson, E.L. *et al.* (2020) 'Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels', *Nature communications*, 11(1), p. 2471. doi:10.1038/s41467-020-16274-w.
- Kelly, C.J. *et al.* (2015) 'Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and Intestinal Epithelial HIF Augments Tissue Barrier Function', *Cell host & microbe*, 17(5), pp. 662–671. doi:10.1016/j.chom.2015.03.005.
- Knight, R. *et al.* (2017) 'The Microbiome and Human Biology', *Annual review of genomics and human genetics*, 18, pp. 65–86. doi:10.1146/annurev-genom-083115-022438.
- Knight, R. *et al.* (2018) 'Best practices for analysing microbiomes', *Nature reviews. Microbiology*, 16(7), pp. 410–422. doi:10.1038/s41579-018-0029-9.
- Kummen, M. *et al.* (2020) 'Rosuvastatin alters the genetic composition of the human gut microbiome', *Scientific reports*, 10(1), p. 5397. doi:10.1038/s41598-020-62261-y.
- Layden, B.T. *et al.* (2012) 'Negative association of acetate with visceral adipose tissue and insulin levels', *Diabetes, metabolic syndrome and obesity: targets and therapy*, 5, pp. 49–55. doi:10.2147/DMSO.S29244.
- Le Chatelier, E. *et al.* (2013) 'Richness of human gut microbiome correlates with metabolic markers', *Nature*, 500(7464), pp. 541–546. doi:10.1038/nature12506.
- Ley, R.E., Peterson, D.A. and Gordon, J.I. (2006) 'Ecological and evolutionary forces shaping microbial diversity in the human intestine', *Cell*, 124(4), pp. 837–848.

doi:10.1016/j.cell.2006.02.017.

Litvak, Y. *et al.* (2017) 'Dysbiotic Proteobacteria expansion: a microbial signature of epithelial dysfunction', *Current opinion in microbiology*, 39, pp. 1–6. doi:10.1016/j.mib.2017.07.003.

Litvak, Y. and Bäumler, A.J. (2019) 'Microbiota-Nourishing Immunity: A Guide to Understanding Our Microbial Self', *Immunity*, 51(2), pp. 214–224. doi:10.1016/j.immuni.2019.08.003.

Litvak, Y., Byndloss, M.X. and Bäumler, A.J. (2018) 'Colonocyte metabolism shapes the gut microbiota', *Science*, 362(6418). doi:10.1126/science.aat9076.

Longo, D.L. and Drazen, J.M. (2016) 'Data Sharing', *The New England journal of medicine*, pp. 276–277. doi:10.1056/NEJMe1516564.

Lozupone, C. *et al.* (2011) 'UniFrac: an effective distance metric for microbial community comparison', *The ISME journal*, 5(2), pp. 169–172. doi:10.1038/ismej.2010.133.

Lu, J. *et al.* (2017) 'Bracken: estimating species abundance in metagenomics data', *PeerJ Computer Science*, p. e104. doi:10.7717/peerj-cs.104.

Maier, L. *et al.* (2018) 'Extensive impact of non-antibiotic drugs on human gut bacteria', *Nature*, 555(7698), pp. 623–628. doi:10.1038/nature25979.

Marchesi, J.R. and Ravel, J. (2015) 'The vocabulary of microbiome research: a proposal', *Microbiome*, 3, p. 31. doi:10.1186/s40168-015-0094-5.

Markle, J.G.M. *et al.* (2013) 'Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity', *Science*, 339(6123), pp. 1084–1088. doi:10.1126/science.1233521.

Markowitz, F. (2017) 'All biology is computational biology', *PLoS biology*, 15(3), p. e2002050. doi:10.1371/journal.pbio.2002050.

Martinez-Medina, M. *et al.* (2014) 'Western diet induces dysbiosis with increased E coli in CEABAC10 mice, alters host barrier function favouring AIEC colonisation', *Gut*, 63(1), pp. 116–124. doi:10.1136/gutjnl-2012-304119.

Maurice, C.F., Haiser, H.J. and Turnbaugh, P.J. (2013) 'Xenobiotics shape the physiology and gene expression of the active human gut microbiome', *Cell*, 152(1-2), pp. 39–50. doi:10.1016/j.cell.2012.10.052.

- McCarville, J.L. *et al.* (2020) ‘Microbiota Metabolites in Health and Disease’, *Annual review of immunology*, 38, pp. 147–170. doi:10.1146/annurev-immunol-071219-125715.
- Mende, D.R. *et al.* (2020) ‘proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes’, *Nucleic acids research*, 48(D1), pp. D621–D625. doi:10.1093/nar/gkz1002.
- Meyer, F. *et al.* (2021) ‘Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit’, *Nature protocols*, 16(4), pp. 1785–1801. doi:10.1038/s41596-020-00480-3.
- Mirzayi, C. *et al.* (2021) ‘Reporting guidelines for human microbiome research: the STORMS checklist’, *Nature medicine*, 27(11), pp. 1885–1892. doi:10.1038/s41591-021-01552-x.
- Mohammad, S. and Thiemermann, C. (2020) ‘Role of Metabolic Endotoxemia in Systemic Inflammation and Potential Interventions’, *Frontiers in immunology*, 11, p. 594150. doi:10.3389/fimmu.2020.594150.
- Morrison, D.J. and Preston, T. (2016) ‘Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism’, *Gut microbes*, 7(3), pp. 189–200. doi:10.1080/19490976.2015.1134082.
- Morton, J.T. *et al.* (2019) ‘Establishing microbial composition measurement standards with reference frames’, *Nature communications*, 10(1), p. 2719. doi:10.1038/s41467-019-10656-5.
- Nasko, D.J. *et al.* (2018) ‘RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification’, *Genome biology*, 19(1), p. 165. doi:10.1186/s13059-018-1554-6.
- Nayfach, S. *et al.* (2015) ‘Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes’, *PLoS computational biology*, 11(11), p. e1004573. doi:10.1371/journal.pcbi.1004573.
- Nyberg, S.T. *et al.* (2018) ‘Obesity and loss of disease-free years owing to major non-communicable diseases: a multicohort study’, *The Lancet. Public health*, 3(10), pp. e490–e497. doi:10.1016/S2468-2667(18)30139-7.
- Odamaki, T. *et al.* (2016) ‘Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study’, *BMC microbiology*, 16, p. 90. doi:10.1186/s12866-016-0708-5.

- O’Keefe, S.J.D. *et al.* (2015) ‘Fat, fibre and cancer risk in African Americans and rural Africans’, *Nature communications*, 6, p. 6342. doi:10.1038/ncomms7342.
- Palleja, A. *et al.* (2018) ‘Recovery of gut microbiota of healthy adults following antibiotic exposure’, *Nature microbiology*, 3(11), pp. 1255–1265. doi:10.1038/s41564-018-0257-9.
- Parks, D.H. *et al.* (2015) ‘CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes’, *Genome research*, 25(7), pp. 1043–1055. doi:10.1101/gr.186072.114.
- Parks, D.H. *et al.* (2017) ‘Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life’, *Nature microbiology*, 2(11), pp. 1533–1542. doi:10.1038/s41564-017-0012-7.
- Parks, D.H. *et al.* (2018) ‘A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life’, *Nature biotechnology*, 36(10), pp. 996–1004. doi:10.1038/nbt.4229.
- Parks, D.H. *et al.* (2022) ‘GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy’, *Nucleic acids research*, 50(D1), pp. D785–D794. doi:10.1093/nar/gkab776.
- Pasolli, E. *et al.* (2016) ‘Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights’, *PLoS computational biology*, 12(7), p. e1004977. doi:10.1371/journal.pcbi.1004977.
- Pasolli, E. *et al.* (2019) ‘Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle’, *Cell*, 176(3), pp. 649–662.e20. doi:10.1016/j.cell.2019.01.001.
- Porrás, A.M. and Brito, I.L. (2019) ‘The internationalization of human microbiome research’, *Current opinion in microbiology*, 50, pp. 50–55. doi:10.1016/j.mib.2019.09.012.
- Quinn, T.P. *et al.* (2018) ‘Understanding sequencing data as compositions: an outlook and review’, *Bioinformatics*, 34(16), pp. 2870–2878. doi:10.1093/bioinformatics/bty175.
- Rajesh, A. *et al.* (2021) ‘Improving the completeness of public metadata accompanying omics studies’, *Genome biology*, 22(1), p. 106. doi:10.1186/s13059-021-02332-z.
- Ruaud, A. *et al.* (2020) ‘Syntrophy via Interspecies H<sub>2</sub> Transfer between *Christensenella* and *Methanobrevibacter* Underlies Their Global Cooccurrence in the Human Gut’, *mBio*, 11(1).

doi:10.1128/mBio.03235-19.

Ryan, M.J. *et al.* (2021) 'Towards a unified data infrastructure to support European and global microbiome research: a call to action', *Environmental microbiology*, 23(1), pp. 372–375. doi:10.1111/1462-2920.15323.

Schlaberg, R. (2020) 'Microbiome Diagnostics', *Clinical chemistry*, 66(1), pp. 68–76. doi:10.1373/clinchem.2019.303248.

Seidell, J.C. and Halberstadt, J. (2015) 'The global burden of obesity and the challenges of prevention', *Annals of nutrition & metabolism*, 66 Suppl 2, pp. 7–12. doi:10.1159/000375143.

Sender, R., Fuchs, S. and Milo, R. (2016) 'Revised Estimates for the Number of Human and Bacteria Cells in the Body', *PLoS biology*, 14(8), p. e1002533. doi:10.1371/journal.pbio.1002533.

Shakya, M., Lo, C.-C. and Chain, P.S.G. (2019) 'Advances and Challenges in Metatranscriptomic Analysis', *Frontiers in genetics*, 10, p. 904. doi:10.3389/fgene.2019.00904.

Shin, N.-R., Whon, T.W. and Bae, J.-W. (2015) 'Proteobacteria: microbial signature of dysbiosis in gut microbiota', *Trends in biotechnology*, 33(9), pp. 496–503. doi:10.1016/j.tibtech.2015.06.011.

Sieber, C.M.K. *et al.* (2018) 'Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy', *Nature microbiology*, 3(7), pp. 836–843. doi:10.1038/s41564-018-0171-1.

Sinha, T. *et al.* (2019) 'Analysis of 1135 gut metagenomes identifies sex-specific resistome profiles', *Gut microbes*, 10(3), pp. 358–366. doi:10.1080/19490976.2018.1528822.

Sorbara, M.T. and Pamer, E.G. (2022) 'Microbiome-based therapeutics', *Nature reviews. Microbiology* [Preprint]. doi:10.1038/s41579-021-00667-9.

Statovci, D. *et al.* (2017) 'The Impact of Western Diet and Nutrients on the Microbiota and Immune Response at Mucosal Interfaces', *Frontiers in immunology*, 8, p. 838. doi:10.3389/fimmu.2017.00838.

Stephens, Z.D. *et al.* (2015) 'Big Data: Astronomical or Genomical?', *PLoS biology*, 13(7), p. e1002195. doi:10.1371/journal.pbio.1002195.

- Suzuki, T.A. *et al.* (2021) ‘Codiversification of gut microbiota with humans’, *bioRxiv*. doi:10.1101/2021.10.12.462973.
- Suzuki, T.A. and Ley, R.E. (2020) ‘The role of the microbiota in human genetic adaptation’, *Science*, 370(6521). doi:10.1126/science.aaz6827.
- Tanca, A. *et al.* (2017) ‘Potential and active functions in the gut microbiota of a healthy human cohort’, *Microbiome*, 5(1), p. 79. doi:10.1186/s40168-017-0293-3.
- Tett, A. *et al.* (2019) ‘The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations’, *Cell host & microbe*, 26(5), pp. 666–679.e7. doi:10.1016/j.chom.2019.08.018.
- Theis, K.R. *et al.* (2016) ‘Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes’, *mSystems*, 1(2). doi:10.1128/mSystems.00028-16.
- Topçuoğlu, B.D. *et al.* (2020) ‘A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems’, *mBio*, 11(3). doi:10.1128/mBio.00434-20.
- Vandeputte, D. *et al.* (2021) ‘Temporal variability in quantitative human gut microbiome profiles and implications for clinical research’, *Nature communications*, 12(1), p. 6740. doi:10.1038/s41467-021-27098-7.
- Vich Vila, A. *et al.* (2020) ‘Impact of commonly used drugs on the composition and metabolic function of the gut microbiota’, *Nature communications*, 11(1), p. 362. doi:10.1038/s41467-019-14177-z.
- Vujkovic-Cvijin, I. *et al.* (2020) ‘Host variables confound gut microbiota studies of human disease’, *Nature*, 587(7834), pp. 448–454. doi:10.1038/s41586-020-2881-9.
- Wallis, A. *et al.* (2017) ‘Support for the microgenderome invites enquiry into sex differences’, *Gut microbes*, 8(1), pp. 46–52. doi:10.1080/19490976.2016.1256524.
- Walters, W.A., Xu, Z. and Knight, R. (2014) ‘Meta-analyses of human gut microbes associated with obesity and IBD.’, *FEBS letters*, 588(22), pp. 4223–4233. doi:10.1016/j.febslet.2014.09.039.
- Wilkinson, M.D. *et al.* (2016) ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific data*, 3, p. 160018. doi:10.1038/sdata.2016.18.
- Wood, D.E., Lu, J. and Langmead, B. (2019) ‘Improved metagenomic analysis with Kraken



2', *Genome biology*, 20(1), p. 257. doi:10.1186/s13059-019-1891-0.

Yanai, I. and Lercher, M. (2020) 'A hypothesis is a liability', *Genome biology*, 21(1), p. 231. doi:10.1186/s13059-020-02133-w.

Youngblut, N.D. *et al.* (2020) 'Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity', *mSystems*, 5(6). doi:10.1128/mSystems.01045-20.

Youngblut, N.D., de la Cuesta-Zuluaga, J. and Ley, R.E. (2022) 'Incorporating genome-based phylogeny and functional similarity into diversity assessments helps to resolve a global collection of human gut metagenomes', *Environmental microbiology* [Preprint]. doi:10.1111/1462-2920.15910.

Youngblut, N.D. and Ley, R.E. (2021) 'Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets', *Cold Spring Harbor Laboratory*. doi:10.1101/2021.02.10.430604.

Zeisel, S.H. and Warriar, M. (2017) 'Trimethylamine N-Oxide, the Microbiome, and Heart and Kidney Disease', *Annual review of nutrition*, 37, pp. 157–181. doi:10.1146/annurev-nutr-071816-064732.

Zeng, M.Y., Inohara, N. and Nuñez, G. (2017) 'Mechanisms of inflammation-driven bacterial dysbiosis in the gut', *Mucosal immunology*, 10(1), pp. 18–26. doi:10.1038/mi.2016.75.

Zimmermann, M. *et al.* (2021) 'Towards a mechanistic understanding of reciprocal drug-microbiome interactions', *Molecular systems biology*, 17(3), p. e10116. doi:10.15252/msb.202010116.

Zou, J. *et al.* (2018) 'Fiber-Mediated Nourishment of Gut Microbiota Protects against Diet-Induced Obesity by Restoring IL-22-Mediated Colonic Health', *Cell host & microbe*, 23(1), pp. 41–53.e4. doi:10.1016/j.chom.2017.11.003.

Zou, Y. *et al.* (2019) '1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses', *Nature biotechnology*, 37(2), pp. 179–185. doi:10.1038/s41587-018-0008-8.



# Appendices



Appendix I: Age- and Sex-Dependent Patterns of Gut Microbial  
Diversity in Human Adults

# Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults

Jacobo de la Cuesta-Zuluaga,<sup>a</sup> Scott T. Kelley,<sup>b</sup> Yingfeng Chen,<sup>b</sup> Juan S. Escobar,<sup>c</sup> Noel T. Mueller,<sup>d,e</sup> Ruth E. Ley,<sup>a</sup> Daniel McDonald,<sup>f</sup> Shi Huang,<sup>f</sup> Austin D. Swafford,<sup>g</sup> Rob Knight,<sup>f,g,h,i</sup> Varykina G. Thackray<sup>j</sup>

<sup>a</sup>Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany

<sup>b</sup>Department of Biology, San Diego State University, San Diego, California, USA

<sup>c</sup>Vidarium—Nutrition, Health and Wellness Research Center, Grupo Empresarial Nutresa, Medellin, Colombia

<sup>d</sup>Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

<sup>e</sup>Welch Center for Epidemiology, Prevention and Clinical Research, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA

<sup>f</sup>Department of Pediatrics, University of California, San Diego, La Jolla, California, USA

<sup>g</sup>Center for Microbiome Innovation, University of California, San Diego, La Jolla, California, USA

<sup>h</sup>Department of Computer Science, University of California, San Diego, La Jolla, California, USA

<sup>i</sup>Department of Bioengineering, University of California, San Diego, La Jolla, California, USA

<sup>j</sup>Department of Obstetrics, Gynecology and Reproductive Sciences, University of California, San Diego, La Jolla, California, USA

**ABSTRACT** Gut microbial diversity changes throughout the human life span and is known to be associated with host sex. We investigated the association of age, sex, and gut bacterial alpha diversity in three large cohorts of adults from four geographical regions: subjects from the United States and United Kingdom in the American Gut Project (AGP) citizen-science initiative and two independent cohorts of Colombians and Chinese. In three of the four cohorts, we observed a strong positive association between age and alpha diversity in young adults that plateaued after age 40 years. We also found sex-dependent differences that were more pronounced in younger adults than in middle-aged adults, with women having higher alpha diversity than men. In contrast to the other three cohorts, no association of alpha diversity with age or sex was observed in the Chinese cohort. The association of alpha diversity with age and sex remained after adjusting for cardiometabolic parameters in the Colombian cohort and antibiotic usage in the AGP cohort. We further attempted to predict the microbiota age in individuals using a machine-learning approach for the men and women in each cohort. Consistent with our alpha-diversity-based findings, U.S. and U.K. women had a significantly higher predicted microbiota age than men, with a reduced difference being seen above age 40 years. This difference was not observed in the Colombian cohort and was observed only in middle-aged Chinese adults. Together, our results provide new insights into the influence of age and sex on the biodiversity of the human gut microbiota during adulthood while highlighting similarities and differences across diverse cohorts.

**IMPORTANCE** Microorganisms in the human gut play a role in health and disease, and in adults higher gut biodiversity has been linked to better health. Since gut microorganisms may be pivotal in the development of microbial therapies, understanding the factors that shape gut biodiversity is of utmost interest. We performed large-scale analyses of the relationship of age and sex to gut bacterial diversity in adult cohorts from four geographic regions: the United States, the United Kingdom, Colombia, and China. In the U.S., U.K., and Colombian cohorts, bacterial biodiversity correlated positively with age in young adults but plateaued at about age 40 years, with no positive association being found in middle-aged adults. Young, but not middle-aged, adult women had higher gut bacterial diversity than men, a pattern confirmed via supervised machine learning. Interestingly, in the Chinese cohort, min-


**Citation** de la Cuesta-Zuluaga J, Kelley ST, Chen Y, Escobar JS, Mueller NT, Ley RE, McDonald D, Huang S, Swafford AD, Knight R, Thackray VG. 2019. Age- and sex-dependent patterns of gut microbial diversity in human adults. *mSystems* 4:e00261-19. <https://doi.org/10.1128/mSystems.00261-19>.

**Editor** Sergio Baranzini, University of California, San Francisco

**Copyright** © 2019 de la Cuesta-Zuluaga et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Varykina G. Thackray, [vthackray@ucsd.edu](mailto:vthackray@ucsd.edu).

J.D.L.C.-Z. and S.T.K. contributed equally to this article.

 Large human gut microbiome cohort study reveals patterns of age- and sex-dependent alpha diversity in young adults, with women having higher alpha diversity than men. This positive association of age and gut biodiversity plateaus by age 40.

**Received** 24 April 2019

**Accepted** 30 April 2019

**Published** 14 May 2019

imal associations were observed between gut biodiversity and age or sex. Our results highlight the patterns of adult gut biodiversity and provide a framework for future research.

**KEYWORDS** 16S rRNA amplicon, age, diversity, microbiome, sex

The human gut microbiota is a highly diverse ecosystem that is extremely variable among individuals (1). This microbial community may play a key role in human health and disease (2). Since the gut microbiota may be pivotal to the development of microbial therapies, understanding the factors that shape overall gut microbiota biodiversity over the different human life stages is of utmost interest.

There is increasing evidence suggesting that host genes, gene expression patterns, environmental exposures (including medication and diet), and lifestyle factors play an important role in delimiting the boundaries of microbial diversity in the gut (3, 4). While a detailed longitudinal study of the interplay of each of these factors would be scientifically, logistically, and financially challenging, the chronological age of the host may be conceived of as a proxy variable that represents the accumulation of these effects for a given individual. Several studies have reported a positive correlation between age and gut microbiota alpha diversity from birth to adulthood (5–8). Likewise, it has been shown that alpha diversity is maintained in old age, until comorbidities contribute to its decline (9). Another intriguing host-associated pattern identified in humans and rodents is the link between the gut microbiota and biological sex. Several studies have reported that women have higher microbial diversity than men and that sex differences in microbial composition emerge after puberty (8, 10–13). These differences may contribute to the sexual dimorphism of autoimmune (12, 14, 15) and neuroimmune (16, 17) diseases. Therefore, it is key to consider the impact of age and sex differences in different human populations to adequately discriminate changes and variations in the microbiome of individuals.

To better understand how the age and sex of the host relate to the diversity of the gut microbiota during adulthood, we explored the association of these factors using data from individuals in three cross-sectional studies from four geographical origins, including the citizen-science American Gut Project (AGP), comprised of individuals from the United States and the United Kingdom (4); a cohort of individuals from China (18); and a study of community-dwelling adults from Colombia (19).

## RESULTS

The basic characteristics of the individuals from the four cohorts, stratified by sex and age group, are summarized in Table 1. We defined adults as individuals between 20 and 69 years of age and divided the age groups by the middle point of this range (i.e., 45 years); subjects above 70 years of age were excluded from the analysis.

To assess changes in alpha diversity with age during adulthood, we fit a simple linear regression model and a regression model with linear splines, in which the model is fit as two consecutive segments (20 to 45 years and 46 to 69 years; see Fig. S1 in the supplemental material). We then evaluated the goodness of fit of each model using the Akaike information criterion (AIC), which indicated that changes in alpha diversity are better explained by distinguishing between young adults (20 to 45 years of age) and middle-aged adults (46 to 69 years of age). In the U.S., U.K., and Colombian cohorts, we observed a positive but nonlinear association between alpha-diversity measures and age in both women and men. Loess curves fit independently by sex showed an inflection point after 40 years of age in each of these cohorts (Fig. 1A to C). In contrast, we did not observe such a pattern in the Chinese cohort, in which alpha diversity displayed a slight decrease with age (Fig. 1D).

Next, for each population, we fit linear regression models to examine associations of microbial diversity, age, and sex in each age group separately. In both the U.S. and U.K. cohorts, we observed a positive relationship between microbial richness and age for both sexes in young adults (adjusted *P* value [*P*-adjust], <0.001 for the U.S. cohort and

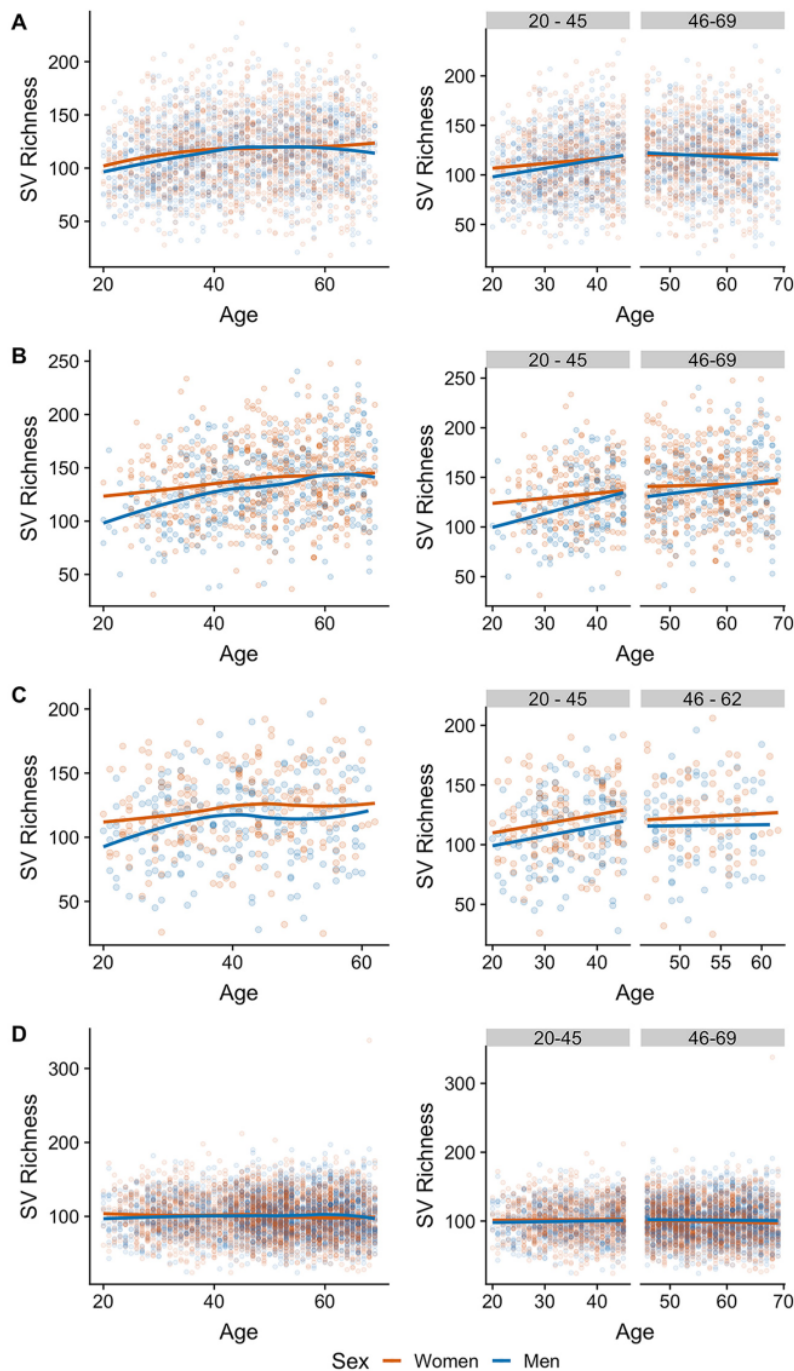
**TABLE 1** General characteristics of the participants of the included cohorts<sup>a</sup>

Cohort and characteristic	Young adults (ages 20–45 yr)		Middle-aged adults (ages 46–69 yr) <sup>b</sup>	
	Women	Men	Women	Men
<b>AGP, U.S.</b>				
No. of subjects	627	644	734	583
Age (yr)	34.60 (6.79)	33.76 (6.61)	56.13 (6.37)	57.13 (6.41)
SV richness	113.80 (33.04)	110.95 (31.51)	120.40 (0.89)	119.0 (0.78)
Shannon index	4.87 (0.83)	4.83 (0.80)	4.98 (0.89)	5.01 (0.78)
<b>AGP, U.S. antibiotic consumers</b>				
No. of subjects	136	83	147	91
Age (yr)	33.0 (7.75)	58.07 (6.56)	34.71 (6.87)	58.35 (6.52)
SV richness	100.38 (27.61)	97.80 (31.85)	107.07 (30.06)	108.71 (30.11)
Shannon index	4.64 (0.74)	4.70 (0.83)	4.72 (0.80)	4.83 (0.85)
<b>AGP, U.K.</b>				
No. of subjects	195	173	344	224
Age (yr)	35.90 (6.02)	36.40 (6.32)	56.45 (6.68)	57.75 (6.85)
SV richness	132.0 (31.69)	122.60 (32.38)	142.30 (36.27)	139.10 (36.23)
Shannon index	5.27 (0.69)	5.05 (0.92)	5.36 (0.83)	5.29 (0.80)
<b>Chinese</b>				
No. of subjects	946	670	1,826	1,521
Age (yr)	35.16 (6.73)	34.88 (7.07)	56.6 (6.59)	57.36 (6.75)
SV richness	101.80 (27.90)	99.66 (26.48)	99.41 (28.67)	101.4 (28.50)
Shannon index	4.47 (0.85)	4.40 (0.84)	4.36 (0.93)	4.35 (0.95)
<b>Colombian</b>				
No. of subjects	143	133	83	78
Age (yr)	33.83 (7.21)	34.21 (6.98)	52.48 (4.14)	52.90 (4.42)
SV richness	120.41 (30.21)	110.71 (31.06)	123.33 (32.75)	116.13 (33.95)
Shannon index	4.60 (1.05)	4.48 (1.12)	4.73 (0.99)	4.45 (1.13)
Cardiometabolic risk scale	−1.14 (3.07)	0.64 (3.67)	−0.36 (3.06)	1.39 (2.71)

<sup>a</sup>Values are given as the mean (SD).<sup>b</sup>The ages of the Colombian individuals ranged from 20 to 62 years.

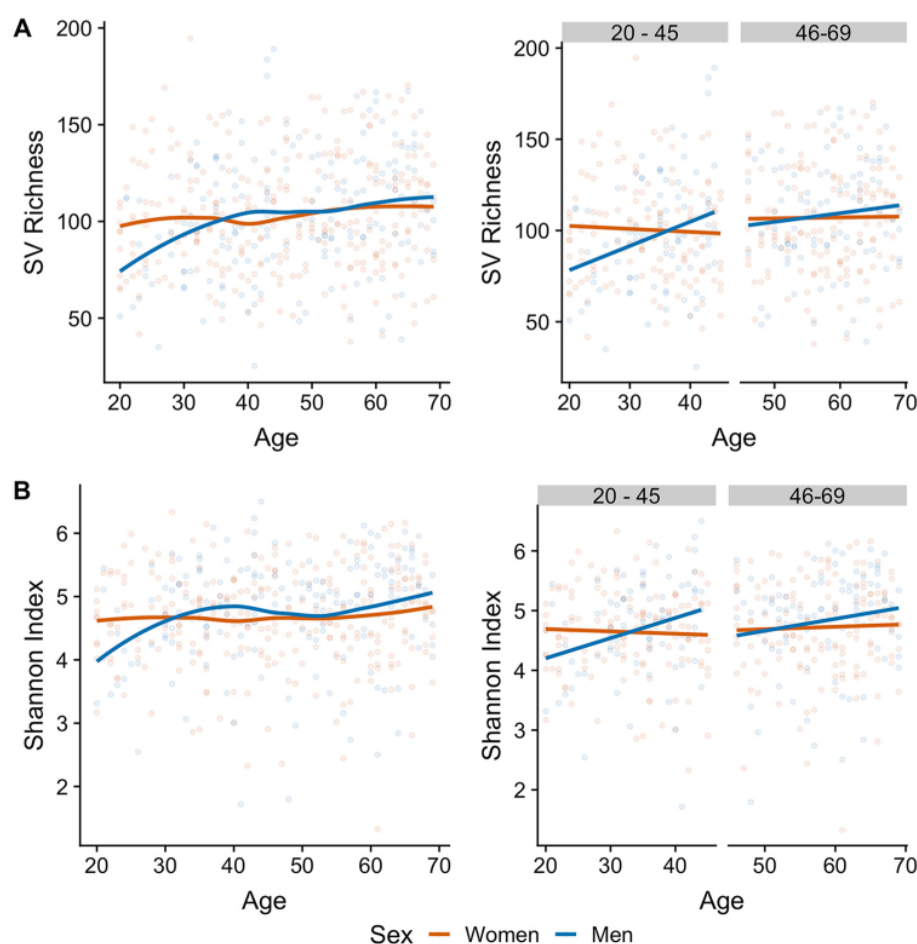
<0.001 for the U.K. cohort), but not in middle-aged adults ( $P$ -adjust, 0.474 for the U.S. cohort and 0.216 for the U.K. cohort) (Fig. 1A and B). In addition, after accounting for age, differences in sequence variant (SV) richness tended to be higher in young adults (for the U.S. cohort, difference between men and women [ $\Delta_{\text{men} - \text{women}}$ ] =  $-3.3$  and  $P$ -adjust = 0.134; for the U.K. cohort,  $\Delta_{\text{men} - \text{women}}$  =  $-9.84$  and  $P$ -adjust = 0.024) than in middle-aged adults (for the U.S. cohort,  $\Delta_{\text{men} - \text{women}}$  =  $-1.3$  and  $P$ -adjust = 0.484; for the U.K. cohort,  $\Delta_{\text{men} - \text{women}}$  =  $-3.7$  and  $P$ -adjust = 0.270). Similar results were observed when we assessed taxon evenness using the Shannon index (Fig. S2). Similar to the U.S. and U.K. cohorts from the AGP, we identified a positive relationship between richness and age in the Colombian cohort in young adults of both sexes ( $P$ -adjust = 0.008) but not in middle-aged adults ( $P$ -adjust = 0.722) (Fig. 1C). Likewise, there was a difference in overall SV richness between the sexes in young adults ( $\Delta_{\text{men} - \text{women}}$  =  $-10.0$ ;  $P$ -adjust = 0.024) but not in middle-aged adults ( $\Delta_{\text{men} - \text{women}}$  =  $-7.3$ ;  $P$ -adjust = 0.225). In contrast to the U.S., U.K., and Colombian cohorts, we observed no association between microbiota alpha diversity and age in young-adult or middle-aged-adult Chinese ( $P$ -adjust > 0.1 for both comparisons) (Fig. 1D). Men in the Chinese cohort tended to have lower SV richness than women as young adults, yet the difference was not significant (for young adults,  $\Delta_{\text{men} - \text{women}}$  =  $-2.14$  and  $P$ -adjust = 0.194; for middle-aged adults,  $\Delta_{\text{men} - \text{women}}$  = 2.04 and  $P$ -adjust = 0.107). We did not find evidence of an interaction between age and sex on microbial diversity in the studied cohorts for young or middle-aged adults with either of the diversity measures, after correcting for multiple comparisons ( $P$ -adjust > 0.15 in all cases). In all cohorts apart from the Chinese, the proportion of variance in alpha diversity explained by age and sex was moderate, yet it was consistently higher in younger adults than in middle-aged adults (Table S1).





**FIG 1** Gut microbiota richness is nonlinearly associated with age and differs between women and men in multiple populations: United States ( $n = 2,588$ ) (A), United Kingdom ( $n = 936$ ) (B), Colombia ( $n = 437$ ) (C), and China ( $n = 4,963$ ) (D). (Left) Sequence variant (SV) richness in adults ages 20 to 69 years (the age of the Colombians ranged from 20 to 62 years); lines indicate the relationship of richness with age after Loess smoothing for women and men separately. (Right) SV richness in young (age, 20 to 45 years) and middle-aged (age, 46 to 69 years) adults; lines indicate the linear regression fit for women and men separately.

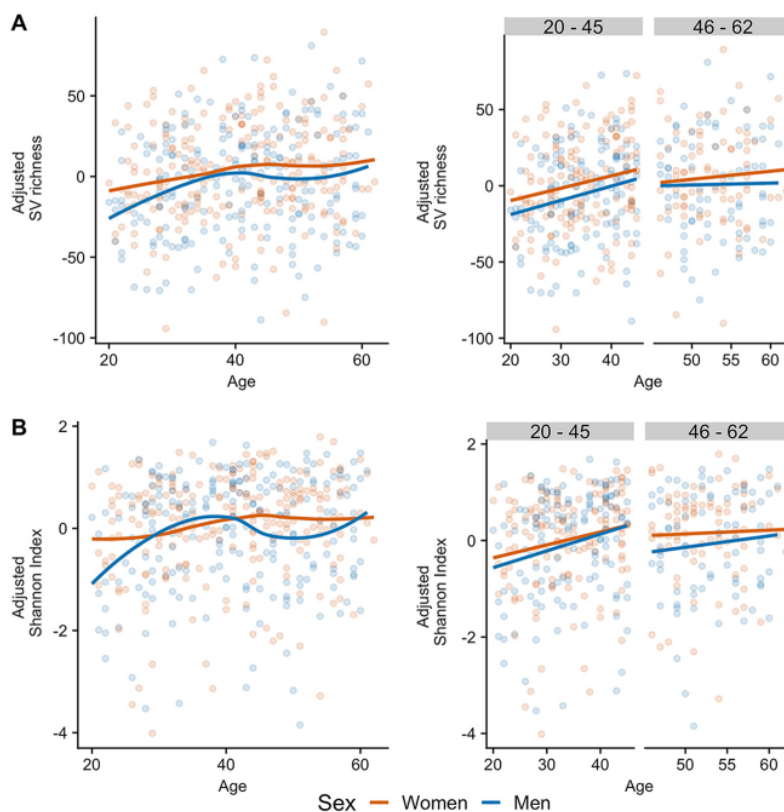
Given that gut microbial diversity may be affected by factors such as antibiotic use or the cardiometabolic health of the host, we replicated the above-described analyses in cohorts in which we observed the patterns, making use of publicly available metadata. To test whether the consumption of antibiotics modified the observed



**FIG 2** Antibiotic consumption has a limited association with the patterns of alpha diversity in U.S. adults that had consumed antibiotics 6 months prior to enrollment ( $n = 457$ ). (A) SV richness; (B) Shannon index. (Left) Alpha-diversity metrics in women and men ages 20 to 62 years; lines indicate the relationships of richness with age after Loess smoothing. (Right) Alpha-diversity metrics in young (age, 20 to 45 years) and middle-aged (age, 46 to 69 years) adults; lines indicate the linear regression fit for women and men separately.

pattern, we performed the above-described analyses on a set of 457 individuals (283 women and 174 men) from the U.S. cohort of the AGP that reported having consumed antibiotics in the 6 months prior to enrollment. We observed a lower SV richness in these individuals than in those that did not consume antibiotics (Table 1). Among the participants that consumed antibiotics in the past 6 months, we observed a similar tendency for alpha diversity to increase in the younger group and plateau in middle-aged individuals, with women having higher diversity than men, although there was a lack of statistical significance (Fig. 2). Likewise, we replicated the analyses in the Colombian cohort after introducing a composite measure of the cardiometabolic health of the subjects as a covariate into the linear models; after we adjusted the analyses for the cardiometabolic health score, the observed patterns were similar (Fig. 3).

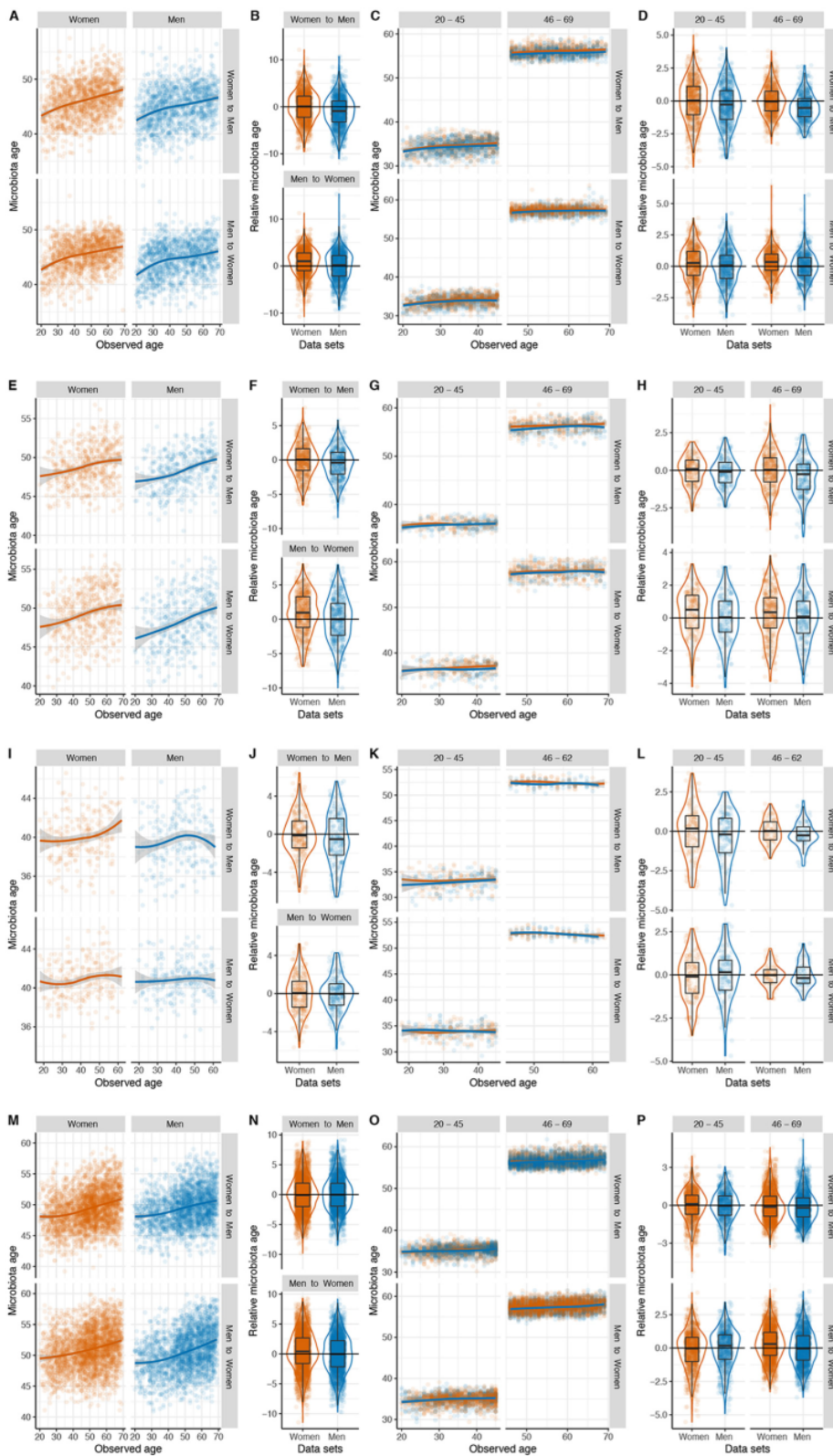
To examine whether similar age- and sex-associated patterns would be observed when analyzing the relative taxon abundance in the gut microbiota, rather than using only alpha-diversity measures, we used a supervised machine-learning approach to compare the composition of the gut microbiota of the subjects of the different populations. We subdivided each cohort by sex, determined the SVs shared by both groups, and used their relative abundances and the chronological age at the time of sample collection of the host to fit a random forest (RF) regression model. Two models



**FIG 3** Adjusting alpha diversity by cardiometabolic health does not affect the observed patterns in Colombian adults ( $n = 437$ ). (A) Residuals of SV richness; (B) residuals of the Shannon index. (Left) Adjusted alpha-diversity metrics in women and men ages 20 to 62 years; lines indicate the relationships of richness with age after Loess smoothing. (Right) Adjusted alpha-diversity metrics in young (age, 20 to 45 years) and middle-aged (age, 46 to 62 years) adults; lines indicate the linear regression fit for women and men separately.

were built for women and men aged 20 to 69 years; each was trained using the data for one sex and tested on the other. For each subject, we calculated the relative microbiota age as the difference between its microbiota age and the microbiota age of the interpolated spline fit of an individual of the opposite sex at the same chronological age. Our results from random forest regressions indicated that the composition of the gut microbiota explained a low to moderate proportion of variance in chronological age, which varied by population and sex (Table S2).

We used 1,494 shared SVs between women and men to build the RF model of the U.S. cohort (Fig. 4A). We found that men exhibited a lower relative microbiota age than women (in the women-to-men model, the difference between women and men [ $\Delta_{\text{women} - \text{men}}$ ] = 0.81 years;  $P$ -adjust < 0.001, Wilcoxon rank-sum test; Fig. 4B, top), suggesting that sex is associated with the adult gut microbial aging process. To validate this finding, we also trained an RF model in the men and then applied it to the women (Fig. 4A, bottom); we found that women had a higher microbiota age (in the men-to-women model,  $\Delta_{\text{women} - \text{men}} = 1.0$  years and  $P$ -adjust < 0.001; Fig. 4B, bottom). To establish whether these trends were present in different age groups, we then examined the sex-dependent association of microbiota age in young and middle-aged adults separately. In the young-adult group, we selected the 1,311 shared SVs between both sexes to build the RF model for women and then applied it to predict the microbiota age of men. We found that young women exhibited a slightly higher relative microbiota age than men ( $\Delta_{\text{women} - \text{men}} = 0.32$  years,  $P$ -adjust < 0.001; Fig. 4C and D, top). Similar results were observed when we assessed the microbiota age in the middle-



**FIG 4** Gut microbiota age differs between women and men in multiple populations: United States ( $n = 2,588$ ) (A to D), United Kingdom ( $n = 936$ ) (E to H), Colombia ( $n = 437$ ) (I to L), and China ( $n = 4,963$ ) (M to P). For each population, the (Continued on next page)

aged-adult group, in which we used the 1,601 shared SVs between sexes to build the RF model as described above. The microbiota age was higher in women than in men ( $\Delta_{\text{women} - \text{men}} = 0.48$  years,  $P\text{-adjust} < 0.001$ ; Fig. 4C and D, top). Furthermore, such sex differences in microbiota age were not affected when we applied the men's model to women (Fig. 4C and D, bottom). Likewise, in the U.K. cohort, we found that the microbiota age was higher in women than in men (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.51$  years and  $P\text{-adjust} = 0.002$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.96$  years and  $P\text{-adjust} < 0.001$ ; Fig. 4E and F), using 1,613 SVs found in either the women's or men's microbiota for building and applying RF models. In addition, we observed significant or borderline significant differences in relative microbiota age between sexes in young adults (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.15$  years and  $P\text{-adjust} = 0.186$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.36$  years and  $P\text{-adjust} = 0.028$ ; Fig. 4G and H) and middle-aged adults (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.41$  years and  $P\text{-adjust} < 0.001$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.25$  years and  $P\text{-adjust} < 0.091$ ; Fig. 4G and H). In the Colombian cohort, we used 1,074 SVs shared between sexes to build the RF model; similar yet nonsignificant trends were observed between microbiota age and sex in the nonstratified analyses (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.38$  years and  $P\text{-adjust} = 0.173$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 3.0e-05$  years and  $P\text{-adjust} > 0.9$ ; Fig. 4I and J) and in the young-adult group (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.18$  years and  $P\text{-adjust} = 0.189$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.02$  years and  $P\text{-adjust} = 0.292$ ; Fig. 4K and L) and the middle-aged-adult group (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.27$  years and  $P\text{-adjust} = 0.186$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.27$  years and  $P\text{-adjust} = 0.873$ ; Fig. 4K and L). We used 1,279 SVs shared between sexes to build the RF models in the Chinese cohort. The association between microbiota age and sex was not consistent when we cross-tested the models (women-to-men model,  $\Delta_{\text{women} - \text{men}} = -0.07$  years and  $P\text{-adjust} = 0.468$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.45$  years and  $P\text{-adjust} < 0.001$ ; Fig. 4M and N). We did not observe significant associations in the young-adult group (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.09$  years and  $P\text{-adjust} = 0.183$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = -0.17$  years and  $P\text{-adjust} = 0.028$ ; Fig. 4O and P), whereas in the middle-aged-adult group, we observed sex-dependent differences in microbiota age, and such differences tended to be consistent in the cross-application of the models (women-to-men model,  $\Delta_{\text{women} - \text{men}} = 0.08$  years and  $P\text{-adjust} = 0.059$ ; men-to-women model,  $\Delta_{\text{women} - \text{men}} = 0.31$  years and  $P\text{-adjust} < 0.01$ ; Fig. 4O and P).

Next, from the RF model trained on each sex to predict age from gut microbial composition, we determined the number of SVs that minimized the 10-fold cross-validation error of the models. We found that the error of the simplified models increased sharply when less than 500 SVs were used (Fig. S3). Finally, we obtained the taxonomic classification of the 500 SVs with the highest RF importance score in at least one of the models (Table S3). Overall, we found that SVs belonging to the families *Ruminococcaceae*, *Bifidobacteriaceae*, *Lachnospiraceae*, *Clostridiaceae*, and *Christensenellaceae* consistently had high RF importance scores, although the values differed between populations and within populations between men and women.

## DISCUSSION

In this study, we analyzed the association of gut microbial alpha diversity with age and sex in three large cross-sectional cohorts encompassing four geographically dis-

### FIG 4 Legend (Continued)

first set of panels (A, E, I, M) shows the microbiota age of women (orange) or men (blue), as calculated by a random forest (RF) model trained on the female (top scatter plots) or male (bottom scatter plots) subsets; lines indicate the spline fit. The second set of panels (B, F, J, N) shows the relative microbiota age (the difference of microbiota age of the interpolated spline fit based on the training data and microbiota age predicted in either training or test data) in women and men, which was derived from either an RF model trained on women and tested on men (top box plot) or an RF model trained on men and tested on women (bottom box plot). The third (C, G, K, O) and fourth (D, H, L, P) sets show results of analyses similar to those for the first two but are stratified by age group.

tinct community-dwelling adult populations. Our analyses indicate that age is positively associated with gut bacterial diversity in men and women, with greater diversity being seen in women than in men. Notably, this association occurs in young but not middle-aged adults. Consistent with these findings, the predicted microbiota age varied based on sex, with stronger associations being seen in young adults. It is worth underscoring that while we did not observe these patterns in all studied cohorts, it was widespread and robust to technical differences, and the alpha-diversity shifts were not modified by the cardiometabolic health of the host or the antibiotic consumption in the cohorts for which this information was available. These findings provide new insights into the development of the human gut microbiome in adulthood according to both age and sex and emphasize the importance of including chronological age and sex as covariates in analyses of the human gut microbiota.

While the most dramatic change in gut microbiota diversity occurs in early childhood (7, 8), its increase in adulthood has also been reported (20, 21). In the cohorts in which the pattern was present, we observed an increase in alpha-diversity measures in young adults; however, this trend halted at about age 40 years (Fig. 1 and Fig. S1 in the supplemental material). This finding agrees with a previous report, in which no significant differences in alpha diversity were found between middle-aged adults and septuagenarians (21). The diversity of the gut microbiota continues changing after the seventh decade of life; it has been shown that centenarians have a higher alpha diversity than middle-aged adults, though it remains unknown whether this is the cause or the effect of healthy aging (20–22). However, gut microbiota diversity in the elderly can differ according to their community residence setting, as community dwellers have been shown to have a higher diversity than individuals in long-term residential care (23, 24).

Interestingly, the relationship between age and diversity was also linked with sex. Multiple studies have reported differences in the diversity and composition of the gut microbiota between female and male mice, which appear to be associated with a sex bias in the incidence of specific diseases, such as type 1 diabetes (12, 15), rheumatoid arthritis (14), and anxiety (25); sex-by-diet interactions have also been reported (26). While differences in alpha diversity between males and females were reported in humans and mice, we showed that the association between sex and alpha diversity was stronger in young adults than in middle-aged adults. In agreement with our results, no differences in alpha diversity were observed between women and men in a recent study in which the mean age of the participants was 60 years (27).

One of the most intriguing findings was the difference in gut microbiota richness between the sexes in young adults. This sex-dependent discrepancy suggests that women may enter adulthood with a more diverse gut microbiota, which plateaus at the same levels in both sexes by approximately age 40 years. The microbiota age models of young adults (ages 20 to 45 years) can explain about 2.5% more of the variance of chronologic age than those of middle-age adults. The establishment of different microbial communities in males and females may be mediated by sex hormones: female mice show a significant increase in alpha diversity during puberty (28), and differences in the composition of the microbiota increase with age but are eliminated by male castration (15). While little is known about the maturation of the human gut microbiota during puberty, we speculate that the differential hormonal milieu between women and men and the earlier timing of puberty in women may result in a more rapid diversification of the gut microbiota in women and that men only achieve the same level of diversification by middle age. Since our findings are based on cross-sectional data, future longitudinal studies are needed to disentangle age and birth cohort effects and the impact of factors such as steroid hormonal levels, pubertal transition, contraceptives, and lifestyle that may vary throughout life. Future research should also investigate specific microbial changes that may influence time-dependent sex differences on the biodiversity of the human gut microbiome.

While 3 of the 4 cohorts had an association between age, sex, and microbial alpha diversity, the Chinese cohort did not (Fig. 1 and Fig. S1), indicating that these associ-

ations are a widespread feature of the human gut microbiota whose universality remains an open question. The overall alpha diversity of this cohort, as measured by SV richness and the Shannon index, was lower than that of the other three cohorts. We also note that the exclusion criteria for this population were not the same as those for the populations in the other studies, with only a 1-month antibiotic exclusion and no stated exclusion of participants with diabetes or inflammatory bowel disease (18).

The striking similarity among the U.S., U.K., and Colombian cohorts with regard to age- and sex-dependent associations with microbial biodiversity arose despite the different geographical origins, sample sizes, and collection protocols of the studies. Moreover, we also found no apparent association of antibiotic use (U.S. or U.K. cohort; Fig. 2) or cardiometabolic health (Colombian cohort; Fig. 3) on the patterns observed in these cohorts, suggesting that the influence of age and sex on the microbiota may be similar in other ethnic and cultural groups beyond the influence of cardiometabolic disease and antibiotic consumption. Nevertheless, similar large-scale population studies should be performed or reanalyzed to determine the extent to which our results are generalizable to other populations, particularly in light of the findings for the Chinese cohort. Indeed, the contrast between the U.K., U.S., and Colombian cohorts and the Chinese cohort highlights the power of using large data sets and comparative analyses across cohorts to uncover subtle patterns and reveal novel insights not discernible in smaller studies. This is of critical importance, given the plausibility of population-specific disease signatures of the microbiome (18).

## MATERIALS AND METHODS

**Cohort description.** Fecal samples were obtained from individuals in three independent cohorts from four geographical locations. (i) The AGP data set is composed of two cohorts with individuals from the United Kingdom (539 women and 397 men) and the United States (1,361 women and 1,227 men) (Table 1) consisting of healthy participants with a self-reported age of between 20 and 69 years, a body mass index (BMI) of between 18.5 and 30 kg/m<sup>2</sup>, and no history of inflammatory bowel disease, diabetes, or antibiotic use in the past year. (ii) A cohort of Chinese individuals comprised 2,772 women and 2,191 men aged 20 to 69 years with a BMI ranging from 18.5 to 30 kg/m<sup>2</sup> and no antibiotic consumption reported 1 month prior to fecal sample collection; pregnant women and hospitalized, disabled, or critically ill individuals were not included in the study. (iii) A cohort of community-dwelling Colombians (226 women and 211 men) consisted of individuals 20 to 62 years of age enrolled in similar proportions according to BMI, city of residence, and age range (20 to 40 and 41 to 62 years); underweight participants, pregnant women, individuals who had consumed antibiotics or antiparasitics in the 3 months prior to enrollment, and individuals diagnosed with neurodegenerative diseases, current or recent cancer (<1 year), and gastrointestinal diseases were excluded. Details on the data acquisition, quality assessment, and processing of fecal samples from these three cohorts were previously described (4, 18, 19).

**16S rRNA gene sequence processing.** The amplicon sequences of all three cohorts were uniformly processed following the same procedures previously described (4). Briefly, the V4 hypervariable region of the 16S rRNA gene was sequenced with the Illumina MiSeq platform. Raw sequences were clustered into sequence variants (SV) with deblur denoising (29) using the QIIME 2 package (30). Sequence counts were rarefied to 1,250 reads per sample across all samples to mitigate uneven sequencing depth. Downstream analyses in the Chinese cohort were replicated using a rarefaction depth of 5,000 reads per sample, and 3,600 reads per sample were used in the Colombian cohort, to exclude the effect of rarefaction depth on alpha-diversity estimation (data not shown). Note, however, that the sample collection and DNA extraction methods differed between the studies.

**Statistical analyses.** SV richness and the Shannon index were calculated using QIIME 2, and statistical analyses were performed using R (v.3.4.3) software. The association of age and alpha diversity was measured with and without separate age groups by fitting linear models with linear splines (lspline [v.1.0] package of R) with a knot at the midpoint of the age range (45 years of age) and simple linear models, respectively. We assessed the goodness of fit of these models by means of the Akaike information criterion (AIC). Next, scatter plots of each alpha-diversity metric according to age were constructed, and then separate Loess curves for women and men were fit using the ggplot2 (v.3.0) package of R. Given the nonlinear association observed between alpha diversity and age, we subdivided the data sets into two separate age groups, 20 to 45 years (young adult) and 46 to 69 years (middle-aged adult), which were then used to fit linear models to test the associations of age (as a continuous variable) and alpha-diversity measures, stratified by sex; *P* values were adjusted for multiple comparisons using the Benjamini-Hochberg method (31).

Additionally, to account for the possible influence of participant antibiotic usage or cardiometabolic health on the observed associations, we conducted the following sensitivity analyses. For the former, we carried out the analyses using a separate group of individuals of the AGP cohort from the United States who had consumed antibiotics during the 6 months prior to their enrollment (283 women and 174 men). For the latter, we performed the analyses by adjusting the linear models for cardiometabolic risk in the Colombian cohort using a risk measure, which we termed the cardiometabolic risk scale (32). This was

calculated using the sum of the z-scores of log-transformed waist circumference, triglyceride levels, insulin levels, diastolic blood pressure, and high-sensitivity C reactive protein levels; positive values of the score are associated with increased cardiometabolic health risk.

Random forest (RF) regression was used to regress the relative abundances of SVs in the gut microbiota of healthy women and men against their chronological age in each data set (randomForest R package of R) using the following parameters:  $n_{tree} = 18,000$  and  $m_{try} = p/3$ , where  $p$  is the number of input features (SVs). The microbiota age model was first trained on the training data set of female adults and was then applied to test the set of male adults, and vice versa. A smoothing spline function was fit between the microbiota age and the chronological age of the hosts for calculation of the relative microbiota age of the adults in the test sets to which the sparse model was applied. For a particular sample, the relative microbiota age was calculated as the difference between the microbiota age of a focal adult and the microbiota age of the interpolated spline fit of healthy female/male adults at the same chronological age. We further employed the Wilcoxon rank-sum test to compare the relative microbiota age between female and male groups in each data set. To determine the sex difference in microbiota age, we subdivided the data sets into the aforementioned age groups and repeated the analyses as described above in all age segments.

**Data availability.** Processed SV tables are publicly available via the Qiita QIIME database (Colombian study, accession number 11993; AGP study, accession number 10317; China study, accession number 11757). The code and data required to reproduce the statistical analyses are available at [https://github.com/jacodela/microbio\\_aDiv](https://github.com/jacodela/microbio_aDiv).

### SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mSystems.00261-19>.

**FIG S1**, PDF file, 2.1 MB.

**FIG S2**, PDF file, 2.7 MB.

**FIG S3**, PDF file, 0.5 MB.

**TABLE S1**, DOCX file, 0.01 MB.

**TABLE S2**, DOCX file, 0.01 MB.

**TABLE S3**, CSV file, 0.7 MB.

### ACKNOWLEDGMENTS

Data acquisition for the Colombian cohort was funded by the Grupo Empresarial Nutresa, Dinámica IPS, and EPS SURA. N.T.M. was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number K01HL141589 and by grants from the Mid-Atlantic Nutrition Obesity Research Center (P30DK072488) and the Foundation for Gender Specific Medicine. V.G.T. was supported by the National Institute of Child Health and Human Development through a cooperative agreement as part of the National Centers for Translational Research in Reproduction and Infertility (P50 HD012303). S.T.K. and V.G.T. received support from the Max Planck Institute for Developmental Biology in Tübingen, Germany. This work is supported by IBM Research AI through the AI Horizons Network and by the UC San Diego Center for Microbiome Innovation.

Some authors of this work collaborate through the Microbiome & Health Network.

While engaged in the research project, J.S.E. was employed by a food company. J.D.L.C.-Z., S.T.K., Y.C., N.T.M., R.E.L., S.H., A.D.S., R.K., D.M., and V.G.T. had no competing interests.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders of this work had no role in the study design, the collection, analysis, or interpretation of the data, the writing of the report, or the decision to submit the paper for publication.

### REFERENCES

- 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* 550:61–66. <https://doi.org/10.1038/nature23889>.
- Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein PC, Knight R. 2016. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 535:94–103. <https://doi.org/10.1038/nature18850>.
- Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. 2017. The evolution of the host microbiome as an ecosystem on a leash. *Nature* 548:43–51. <https://doi.org/10.1038/nature23292>.
- McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dor-



- restien PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciulek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorestein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, et al. 2018. American gut: an open platform for citizen science microbiome research. *mSystems* 3:e00031-18. <https://doi.org/10.1128/mSystems.00031-18>.
5. Hopkins MJ, Sharp R, Macfarlane GT. 2002. Variation in human intestinal microbiota with age. *Dig Liver Dis* 34(Suppl 2):S12–S18. [https://doi.org/10.1016/S1590-8658\(02\)80157-8](https://doi.org/10.1016/S1590-8658(02)80157-8).
  6. Mariat D, Firmesse O, Levenez F, Guimaraes V, Sokol H, Doré J, Corthier G, Furet J-P. 2009. The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol* 9:123. <https://doi.org/10.1186/1471-2180-9-123>.
  7. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* 108(Suppl 1):4578–4585. <https://doi.org/10.1073/pnas.100081107>.
  8. Yatsunenkov T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. *Nature* 486:222–227. <https://doi.org/10.1038/nature11053>.
  9. Maffei VJ, Kim S, Blanchard E, IV, Luo M, Jazwinski SM, Taylor CM, Welsh DA. 2017. Biological aging and the human gut microbiota. *J Gerontol A Biol Sci Med Sci* 72:1474–1482. <https://doi.org/10.1093/gerona/glx042>.
  10. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D'hoë K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaut L, Fu J, Henckaerts L, Zernakova A, Wijmenga C, Raes J. 2016. Population-level analysis of gut microbiome variation. *Science* 352:560–564. <https://doi.org/10.1126/science.aad3503>.
  11. Kozik AJ, Nakatsu CH, Chun H, Jones-Hall YL. 2017. Age, sex, and TNF associated differences in the gut microbiota of mice and their impact on acute TNBS colitis. *Exp Mol Pathol* 103:311–319. <https://doi.org/10.1016/j.yexmp.2017.11.014>.
  12. Markle JGM, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolfe-Kampczyk U, von Bergen M, McCoy KD, Macpherson AJ, Danska JS. 2013. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* 339:1084–1088. <https://doi.org/10.1126/science.1233521>.
  13. Sinha T, Vich Vila A, Garmaeva S, Jankipersadsing SA, Imhann F, Collij V, Bonder MJ, Jiang X, Gurry T, Alm EJ, D'Amato M, Weersma RK, Scherjon S, Wijmenga C, Fu J, Kurilshikov A, Zernakova A. 2018. Analysis of 1135 gut metagenomes identifies sex-specific resistome profiles. *Gut Microbes* 2018:1–9. <https://doi.org/10.1080/19490976.2018.1528822>.
  14. Gomez A, Luckey D, Yeoman CJ, Marietta EV, Berg Miller ME, Murray JA, White BA, Taneja V. 2012. Loss of sex and age driven differences in the gut microbiome characterize arthritis-susceptible 0401 mice but not arthritis-resistant 0402 mice. *PLoS One* 7:e36095. <https://doi.org/10.1371/journal.pone.0036095>.
  15. Yurkovetskiy L, Burrows M, Khan AA, Graham L, Volchkov P, Becker L, Antonopoulos D, Umesaki Y, Chervonsky AV. 2013. Gender bias in autoimmunity is influenced by microbiota. *Immunity* 39:400–412. <https://doi.org/10.1016/j.immuni.2013.08.013>.
  16. Wallis A, Butt H, Ball M, Lewis DP, Bruck D. 2016. Support for the microgenderome: associations in a human clinical population. *Sci Rep* 6:19171. <https://doi.org/10.1038/srep19171>.
  17. Wallis A, Butt H, Ball M, Lewis DP, Bruck D. 2017. Support for the microgenderome invites enquiry into sex differences. *Gut Microbes* 8:46–52. <https://doi.org/10.1080/19490976.2016.1256524>.
  18. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Zheng Z-D-X, Mujagond P, Chen X-J, Rong Z-H, Chen P, Lyu L-Y, Wang X, Wu C-B, Yu N, Xu Y-J, Yin J, Raes J, Knight R, Ma W-J, Zhou H-W. 2018. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med* 24:1532–1535. <https://doi.org/10.1038/s41591-018-0164-x>.
  19. de la Cuesta-Zuluaga J, Corrales-Agudelo V, Velásquez-Mejía EP, Carmona JA, Abad JM, Escobar JS. 2018. Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. *Sci Rep* 8:11356. <https://doi.org/10.1038/s41598-018-29687-x>.
  20. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao J-Z, Abe F, Osawa R. 2016. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol* 16:90. <https://doi.org/10.1186/s12866-016-0708-5>.
  21. Biagi E, Nylund L, Candela M, Ostan R, Bucci L, Pini E, Nikkila J, Monti D, Satokari R, Franceschi C, Brigidi P, De Vos W. 2010. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One* 5:e10667. <https://doi.org/10.1371/journal.pone.0010667>.
  22. Kong F, Hua Y, Zeng B, Ning R, Li Y, Zhao J. 2016. Gut microbiota signatures of longevity. *Curr Biol* 26:R832–R833. <https://doi.org/10.1016/j.cub.2016.08.015>.
  23. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, Harris HMB, Coakley M, Lakshminarayanan B, O'Sullivan O, Fitzgerald GF, Deane J, O'Connor M, Harnedy N, O'Connor K, O'Mahony D, van Sinderen D, Wallace M, Brennan L, Stanton C, Marchesi JR, Fitzgerald AP, Shanahan F, Hill C, Ross RP, O'Toole PW. 2012. Gut microbiota composition correlates with diet and health in the elderly. *Nature* 488:178–184. <https://doi.org/10.1038/nature11319>.
  24. Jeffery IB, Lynch DB, O'Toole PW. 2016. Composition and temporal stability of the gut microbiota in older persons. *ISME J* 10:170–182. <https://doi.org/10.1038/ismej.2015.88>.
  25. Bridgewater LC, Zhang C, Wu Y, Hu W, Zhang Q, Wang J, Li S, Zhao L. 2017. Gender-based differences in host behavior and gut microbiota composition in response to high fat diet and stress in a mouse model. *Sci Rep* 7:10776. <https://doi.org/10.1038/s41598-017-11069-4>.
  26. Org E, Mehrabian M, Parks BW, Shipkova P, Liu X, Drake TA, Lusic AJ. 2016. Sex differences and hormonal effects on gut microbiota composition in mice. *Gut Microbes* 7:313–322. <https://doi.org/10.1080/19490976.2016.1203502>.
  27. Haro C, Rangel-Zúñiga OA, Alcalá-Díaz JF, Gómez-Delgado F, Pérez-Martínez P, Delgado-Lista J, Quintana-Navarro GM, Landa BB, Navas-Cortés JA, Tena-Sempere M, Clemente JC, López-Miranda J, Pérez-Jiménez F, Camargo A. 2016. Intestinal microbiota is influenced by gender and body mass index. *PLoS One* 11:e0154090. <https://doi.org/10.1371/journal.pone.0154090>.
  28. Kelley ST, Skarra DV, Rivera AJ, Thackray VG. 2016. The gut microbiome is altered in a letrozole-induced mouse model of polycystic ovary syndrome. *PLoS One* 11:e0146509. <https://doi.org/10.1371/journal.pone.0146509>.
  29. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. <https://doi.org/10.1128/mSystems.00191-16>.
  30. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus Brown C, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciulek T, et al. 2018. QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ PrePrints* 6:e27295v1. <https://doi.org/10.7287/peerj.preprints.27295v2>.
  31. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57:289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
  32. Guzman-Castaneda SJ, Ortega-Vega EL, de la Cuesta-Zuluaga J, Velásquez-Mejía EP, Rojas W, Bedoya G, Escobar JS. 2018. Gut microbiota composition explains more variance in the host cardiometabolic risk than genetic ancestry. *bioRxiv* <https://doi.org/10.1101/394726>.



Appendix II: Genomic Insights into Adaptations of  
Trimethylamine-Utilizing Methanogens to Diverse Habitats,  
Including the Human Gut



# Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut

 Jacobo de la Cuesta-Zuluaga,<sup>a</sup>  Tim D. Spector,<sup>b</sup>  Nicholas D. Youngblut,<sup>a</sup>  Ruth E. Ley<sup>a</sup>

<sup>a</sup>Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany

<sup>b</sup>Department of Twin Research and Genetic Epidemiology, King's College London, London, UK


**ABSTRACT** *Archaea* of the order *Methanomassiliicoccales* use methylated amines such as trimethylamine as the substrates for methanogenesis. They form two large phylogenetic clades and reside in diverse environments, from soil to the human gut. Two genera, one from each clade, inhabit the human gut: *Methanomassiliicoccus*, which has one cultured representative, and “*Candidatus* Methanomethylophilus,” which has none. Questions remain regarding their distribution across biomes and human populations, their association with other taxa in the gut, and whether host genetics correlate with their abundance. To gain insight into the *Methanomassiliicoccales* clade, particularly its human-associated members, we performed a genomic comparison of 72 *Methanomassiliicoccales* genomes and assessed their presence in metagenomes derived from the human gut ( $n=4,472$ , representing 22 populations), non-human animal gut ( $n=145$ ), and nonhost environments ( $n=160$ ). Our analyses showed that all taxa are generalists; they were detected in animal gut and environmental samples. We confirmed two large clades, one enriched in the gut and the other enriched in the environment, with notable exceptions. Genomic adaptations to the gut include genome reduction and genes involved in the shikimate pathway and bile resistance. Genomic adaptations differed by clade, not habitat preference, indicating convergent evolution between the clades. In the human gut, the relative abundance of *Methanomassiliicoccales* spp. correlated with trimethylamine-producing bacteria and was unrelated to host genotype. Our results shed light on the microbial ecology of this group and may help guide *Methanomassiliicoccales*-based strategies for trimethylamine mitigation in cardiovascular disease.

**IMPORTANCE** *Methanomassiliicoccales* are less-known members of the human gut archaeome. Members of this order use methylated amines, including trimethylamine, in methane production. This group has only one cultured representative; how its members adapted to inhabit the mammalian gut and how they interact with other microbes is largely unknown. Using bioinformatics methods applied to DNA from a wide range of samples, we profiled the abundances of these *Archaea* spp. in environmental and host-associated microbial communities. We observed two groups of *Methanomassiliicoccales*, one largely host associated and one largely found in environmental samples, with some exceptions. When host associated, these *Archaea* have smaller genomes and possess genes related to bile resistance and aromatic amino acid precursors. We did not detect *Methanomassiliicoccales* in all human populations tested, but when present, they were correlated with bacteria known to produce trimethylamine. Due to their metabolism of trimethylamine, these intriguing *Archaea* may form the basis of novel therapies for cardiovascular disease.

**KEYWORDS** *Methanomassiliicoccales*, archaea, comparative genomics, human gut, metagenomes, microbiome

**Citation** de la Cuesta-Zuluaga J, Spector TD, Youngblut ND, Ley RE. 2021. Genomic insights into adaptations of trimethylamine-utilizing methanogens to diverse habitats, including the human gut. *mSystems* 6:e00939-20. <https://doi.org/10.1128/mSystems.00939-20>.

**Editor** Seth Bordenstein, Vanderbilt University  
**Copyright** © 2021 de la Cuesta-Zuluaga et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).  
Address correspondence to Ruth E. Ley, [rley@tuebingen.mpg.de](mailto:rley@tuebingen.mpg.de).

 A meta-analysis of vertebrate and environmental microbiomes shows trimethylamine (TMA)-utilizing methanogens are adapted to the gut and correlate with TMA-producing bacteria

**Received** 17 September 2020

**Accepted** 14 January 2021

**Published** 9 February 2021

**A**rchaea spp. generally make up a tenth or less of the biomass of the human gut microbiota; however, they are widely prevalent and occupy a unique metabolic niche, utilizing byproducts of bacterial metabolism as the substrates for methanogenesis (1). Members of *Methanobacteriales* are the dominant species of the human gut archaeome (1, 2). These include *Methanobrevibacter smithii*, which uses CO<sub>2</sub>, formate and H<sub>2</sub> as the substrates for methane production (3), and *Methanosphaera stadtmanae*, which consumes methanol and H<sub>2</sub> (4). Through methanogenesis, *Archaea* decrease partial pressures of H<sub>2</sub>, potentially increasing the energetic efficiency of primary fermenters and the production of short-chain fatty acids (5).

A second archaeal lineage, the order *Methanomassiliicoccales*, is also found within the human gut, yet its members are less well characterized than those of *Methanobacteriales*. Members of this order, including human-derived *Methanomassiliicoccus luminyensis*, "*Candidatus Methanomassiliicoccus intestinalis*," and "*Candidatus Methanomethylophilus alvus*," perform H<sub>2</sub>-dependent methylotrophic methanogenesis for their sole energy source (6–8). Their genomes encode several methyltransferases and associated proteins that reduce methylamines, methanol, and methylated sulfides to methane (9). Studies based on 16S rRNA and *mcrA* gene diversity analysis indicate that the order *Methanomassiliicoccales* is made up of two large clades, which mostly group species that have either a free-living (FL) or host-associated (HA) lifestyle (10, 11). Based on analyses of the genomes from three human-derived species from both clades, Borrel et al. (9) suggested that each clade colonized the mammalian gut independently. Members of the HA clade, including the human-associated "*Ca. M. alvus*," might be expected to show adaptations similar to those of other methanogens from the gut microbiota (12, 13). How members of the FL clade, including the human-associated *M. luminyensis* and "*Ca. M. intestinalis*," have converged on the gut niche remains to be explored.

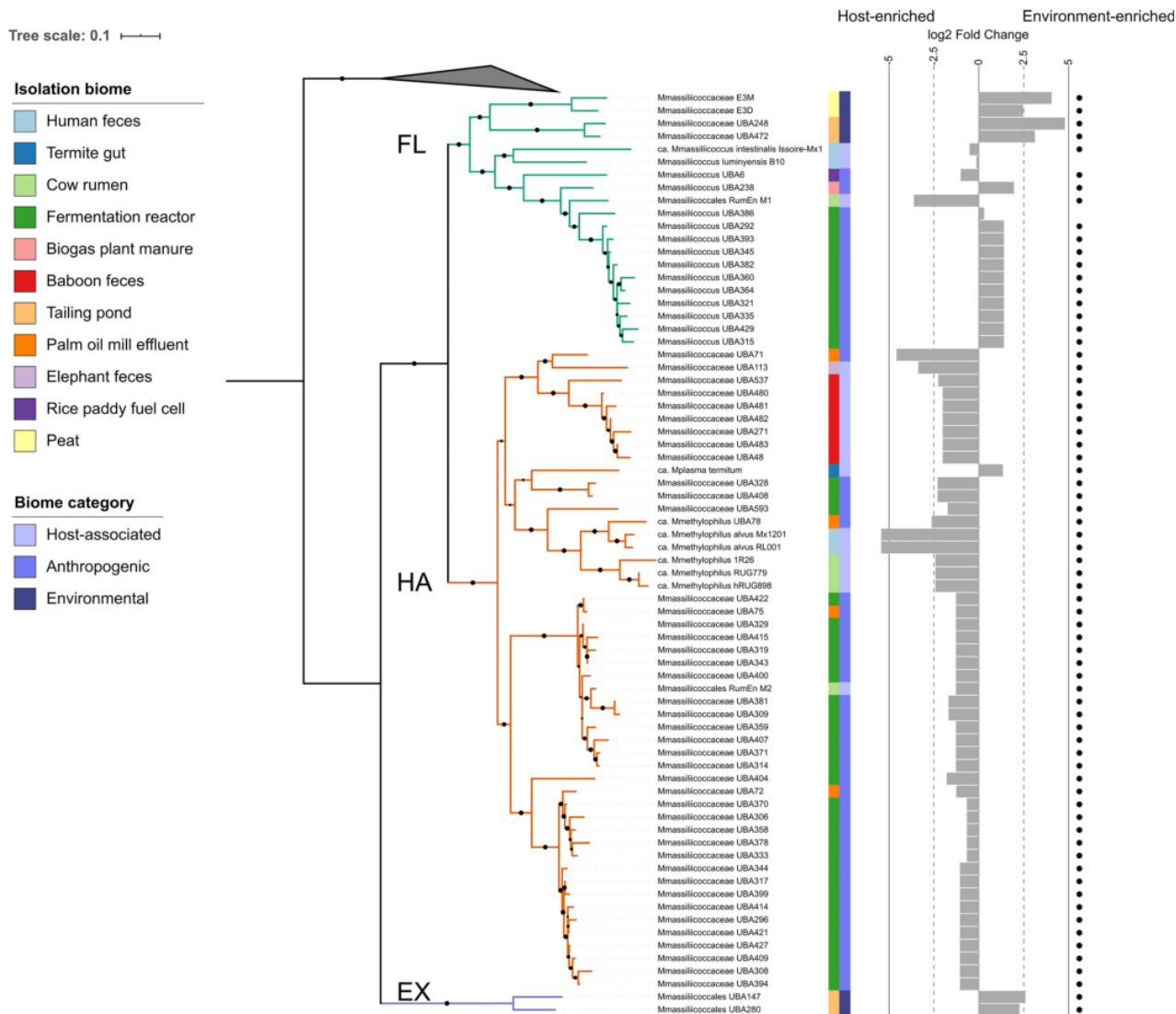
A better understanding of the ecology of *Methanomassiliicoccales* may be of interest to human health, as they can utilize mono-, di-, and trimethylamine (TMA) as the substrates for methanogenesis in the gut (14). TMA, a byproduct of bacterial metabolism of carnitine, choline, and other compounds, is transformed in the liver into trimethylamine N-oxide (TMAO) (15). Circulating TMAO inhibits cholesterol transport and promotes its accumulation in macrophages, inducing the formation of atherosclerotic plaques (16). Decreasing TMA levels in the gut and reducing circulating TMAO levels have been proposed as a therapeutic strategy for cardiovascular disease (17). One way to use the gut microbiome to this end would be to boost levels of *Methanomassiliicoccales* (18). To accomplish this goal requires a deeper understanding of its ecology.

Here, we conducted a comparative analysis of 71 *Methanomassiliicoccales* genomes, together with an additional metagenome-assembled genome (MAG) corresponding to a strain of "*Ca. M. alvus*" that we retrieved by metagenome assembly of gut samples from subjects of the United Kingdom Adult Twin Registry (TwinsUK) cohort (19). We used 305 publicly available metagenomes to assess the prevalence of taxa across various habitat types. While the two large clades grouping host-associated (HA) and free-living (FL) taxa are generally enriched in host-associated and environmental metagenomes, a few exceptions stand out. Our results showed that the repertoire of adhesion proteins encoded by the genomes of taxa from each clade tended to differ. Genes involved in bile resistance and the shikimate pathway are likely involved in the adaptation to the gut environment of members of the HA clade, but not for the FL clade. Thus, gut-adapted members converged on life in the gut using different genomic adaptations. *Methanomassiliicoccales* genera present in the human gut positively correlate with TMA-producing bacteria.

## RESULTS

### Genome-based phylogeny confirms two large *Methanomassiliicoccales* clades.

Based on whole-genome phylogenetic analysis, the order *Methanomassiliicoccales* forms two clades with robust support (Fig. 1). This phylogeny is in agreement with



**FIG 1** The order *Methanomassiliicoccales* forms two large clades that loosely follow the source of isolation. Maximum-likelihood phylogeny of concatenated single-copy marker genes. The gray triangle corresponds to *Thermoplasma acidophilum*, *Picrophilus oshimae*, *Ferroplasma acidarmanus*, *Acidiplasma aeolicum*, and *Cuniculiplasma divulgatum*, which are outgroup taxa from class *Thermoplasmata*. Black circles indicate bootstrap values of >80 (of 1,000 bootstrap permutations), branch color represents the clade, and the scale bar represents the number of amino acid substitutions per site. Colored strips show the source of isolation of each of the included genomes and the general category to which the source belongs. Bar plots show the genome abundance enrichment in gut metagenome samples compared to environmental samples, calculated using DESeq2; dots indicate taxa with significant enrichment in either host or environmental biomes (adj.  $P < 0.05$ ). Mmassiliococcaceae, *Methanomassiliicocceae*; Mmassiliicoccus, *Methanomassiliicoccus*; Mplasma, *Methanoplasma*; Mmethylphilus, *Methanomethylphilus*; Mmassiliococcales, *Methanomassiliicoccales*.

previously reported phylogenies based on 16S rRNA and *mcrA* genes (10, 20, 21). A third distal clade was formed by two closely related MAGs generated in a recent massive meta-genome assembly effort (22), which we labeled external (EX) (Fig. 1). We use the terminology of Borrel et al. (23), as follows: the clade including *Methanomassiliicoccus* is labeled free living (FL), and the clade containing “*Candidatus Methanomethylphilus*” is labeled host associated (HA).

As observed previously (10), the reported source of the genomes was not always consistent with the clade in which it was grouped. For instance, while publicly available genomes originally retrieved from human, baboon, elephant, and cow gastrointestinal tracts were related to “*Candidatus Methanomethylphilus*” (HA), this clade also contained MAGs derived from digestors and reactors (Fig. 1) reportedly not treating

Downloaded from <http://msystems.asm.org/> on February 9, 2021 by guest

animal waste (see Table S1A in the supplemental material). Moreover, MAGs retrieved from pit mud of solid-state fermentation reactors used for the production of Chinese liquor were present in both the HA and FL clades (Table S1A). Similarly, “*Ca. M. intestinalis*” Issoire-Mx1, *M. luminyensis* B10, and *Methanomassiliococcales* archaeon RumEn M1, all retrieved from mammal hosts, grouped in the FL clade.

**Abundance of *Methanomassiliococcales* clades differs in gastrointestinal and environmental samples.** We assessed the abundance of species-level representative *Methanomassiliococcales* taxa in publicly available metagenomes that included 145 samples from gastrointestinal tracts of nonhuman animals, such as cats, pigs, elks, cows, mice, white-throated woodrats, trout, chickens, and geese, and 160 environmental samples from sediment, ice, and diverse water and soil sources (Table S1B).

Taxa from all three clades were detected in a wide range of metagenomes from environmental and gut origin. We observed differences in environmental preference by clade. Abundance of taxa from clade EX was highest in environmental metagenomes ( $0.001\% \pm 0.0012\%$ ) (Fig. 1). These were also detected in gut samples ( $0.0002\% \pm 0.0005\%$ ), albeit with a very low abundance, and in fecal ( $0.0003\% \pm 0.0005\%$ ), large intestine ( $0.0001\% \pm 0.0002\%$ ), and stomach ( $0.0009\% \pm 0.0006\%$ ) metagenomes (Fig. 2). Given their low abundances, further analysis is focused on the FL and HA clades.

The aggregated abundance of clades FL and HA differed across biomes (Fig. 2). In agreement with their names, HA clade members were more abundant in host-associated samples, and FL in non-host-associated samples (Fig. 2). The prevalence and abundance of *Methanomassiliococcales* taxa varied across animal hosts, yet the overall abundance patterns were consistent across hosts and sample types (see Fig. S2 in the supplemental material).

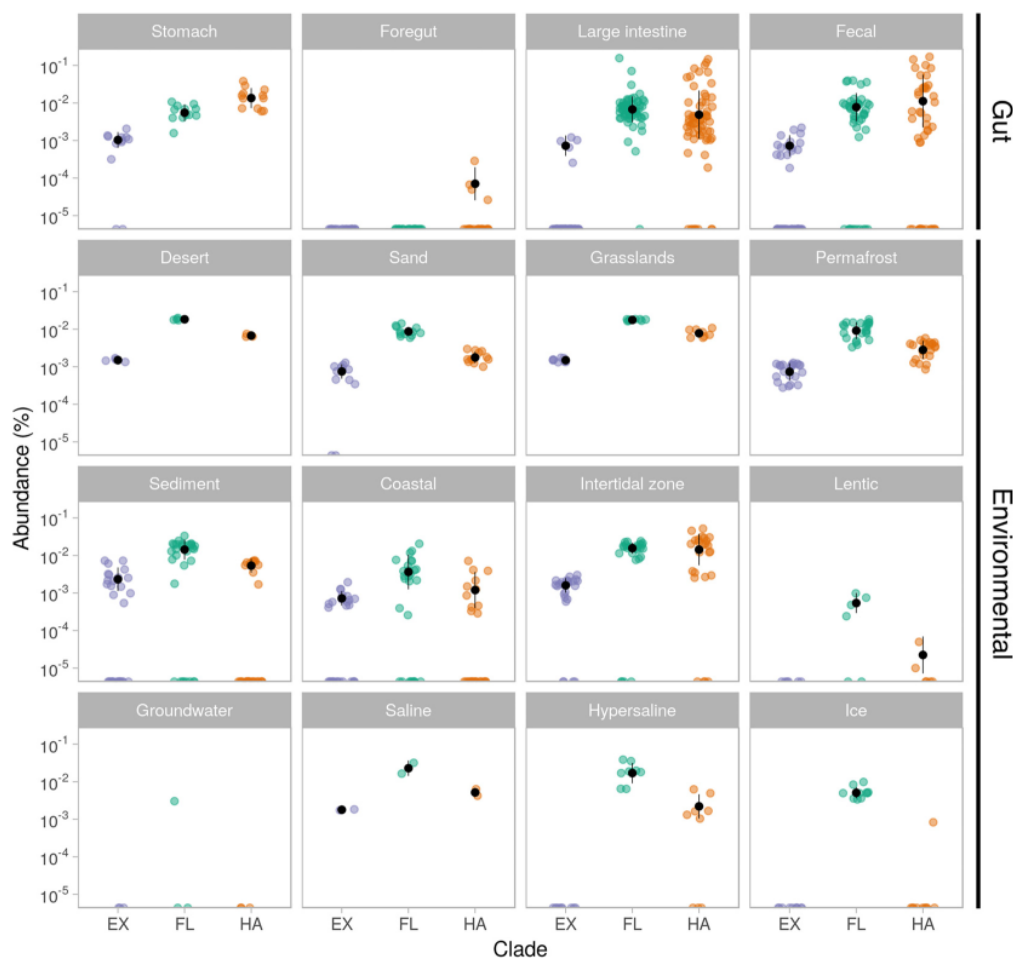
The combined abundance of members of clade FL was higher in samples from environmental biomes ( $0.01\% \pm 0.008\%$ ), although nonzero abundances were observed in digestive system metagenomes ( $0.008\% \pm 0.015\%$ ), with some samples containing levels comparable to that of clade HA (Fig. 2).

The mean abundance of clade HA in aggregate was higher in metagenomes from gut samples ( $0.014\% \pm 0.03\%$ ) compared to those from environmental biomes ( $0.004\% \pm 0.008\%$ ). However, among the environmental biomes, nonzero abundances of clade HA were detected in freshwater ( $0.002\% \pm 0.003\%$ ), marine ( $0.006\% \pm 0.011\%$ ), saline and alkaline ( $0.002\% \pm 0.002\%$ ), and soil ( $0.004\% \pm 0.003\%$ ) samples.

We further validated the differences in clade abundances across biomes by generating a dendrogram of *Methanomassiliococcales* taxa using the DESeq2-based log fold change of individual taxa on gut versus environmental biomes (i.e., the effect size of the test as a measure of enrichment on a given environment). We then compared the structure of this dendrogram with that of the phylogenomic tree and found that they were positively correlated (cophenetic correlation = 0.67;  $P < 0.01$ ).

Overall, we observed a low abundance of individual *Methanomassiliococcales* taxa across all samples, ranging from 0 to 0.15% (Fig. 2 and Fig. S1). Their prevalence across hosts differed; they were prevalent in animals such as elks, pigs, poultry, and cattle, while in others, such as trout and geese, they were largely absent (see Fig. S3 in the supplemental material). The enrichment analysis of individual taxa from clade FL from diverse biomes showed that while most were significantly enriched in environmental metagenomes (adjusted [adj.]  $P < 0.1$ ), some taxa showed the opposite enrichment. *M. luminyensis* and *Methanomassiliococcus* sp. UBA386 were not significantly enriched in gut or environmental biomes. “*Ca. M. intestinalis*” Issoire-Mx1, *Methanomassiliococcales* archaeon RumEn M1, and *Methanomassiliococcus* sp. UBA6 were significantly enriched in gut biomes (Fig. 1), although they were also present in multiple environmental biomes (fig. S1).

When assessed on a per-taxon basis, the vast majority of clade HA taxa were significantly enriched in gut samples (adj.  $P < 0.1$ ), with the exception of “*Candidatus Methanoplasma termitum*.” which was highly abundant in soil samples from grasslands and water samples from intertidal zones (Fig. 1).



**FIG 2** *Methanomassiliicoccales* clades are widespread but not abundant across a range of environments and animal hosts. Combined abundance of representative genomes of the EX (purple), FL (green), and HA (orange) clades in metagenome samples from diverse biomes, as follows: stomach ( $n = 12$ ), foregut ( $n = 23$ ), large intestine ( $n = 66$ ), fecal ( $n = 44$ ), desert ( $n = 4$ ), sand ( $n = 12$ ), grasslands ( $n = 8$ ), permafrost ( $n = 22$ ), sediment ( $n = 31$ ), coastal ( $n = 28$ ), intertidal zone ( $n = 25$ ), lentic ( $n = 6$ ), groundwater ( $n = 3$ ), saline ( $n = 2$ ), hypersaline ( $n = 9$ ), and ice ( $n = 10$ ). Abundances were calculated for individual genomes using KrakenUniq and aggregated by clade. The y axis is in logarithmic scale; black points indicate mean relative abundance in percentage, and black bars indicate standard deviation.

**Genome characteristics and core genes functions differ between *Methanomassiliicoccales* clades.** Given the tendency of clades FL and HA to be enriched in environmental or animal metagenomes, respectively, we searched for genes and genome features linked to putative adaptations of *Methanomassiliicoccales* to an animal gut. For this, we compared 72 genomes from *Methanomassiliicoccales* taxa retrieved from humans, non-human animals, and environmental sources.

We observed that genomes were more similar to others closely located on the phylogeny in terms of genome GC content, genome length and total gene count (local indicator of phylogenetic association [LIPA] adj.  $P < 0.01$  in all cases) (Fig. 3). To determine whether these features differed between clades, while accounting for the autocorrelation due to evolutionary history, we performed a phylogenetic analysis of variance (ANOVA). Clade FL taxa had significantly larger genomes (mean  $\pm$  standard deviation [SD],  $1,985.1 \pm 245.1$  kb) than either clade HA ( $1,318.3 \pm 187.3$  kb) or clade EX ( $1,872.2 \pm 173.8$  kb) (phylogenetic ANOVA adj.  $P = 0.028$ ). In accordance with this, clade FL also had the highest gene count (FL,  $2,153.1 \pm 233.7$  genes; HA,  $1,377.7 \pm 187.7$  genes; EX,  $1,567.0 \pm 90.5$  genes; adj.  $P = 0.025$ ). While this was nonsignificant, clades HA



Tree scale: 0.1

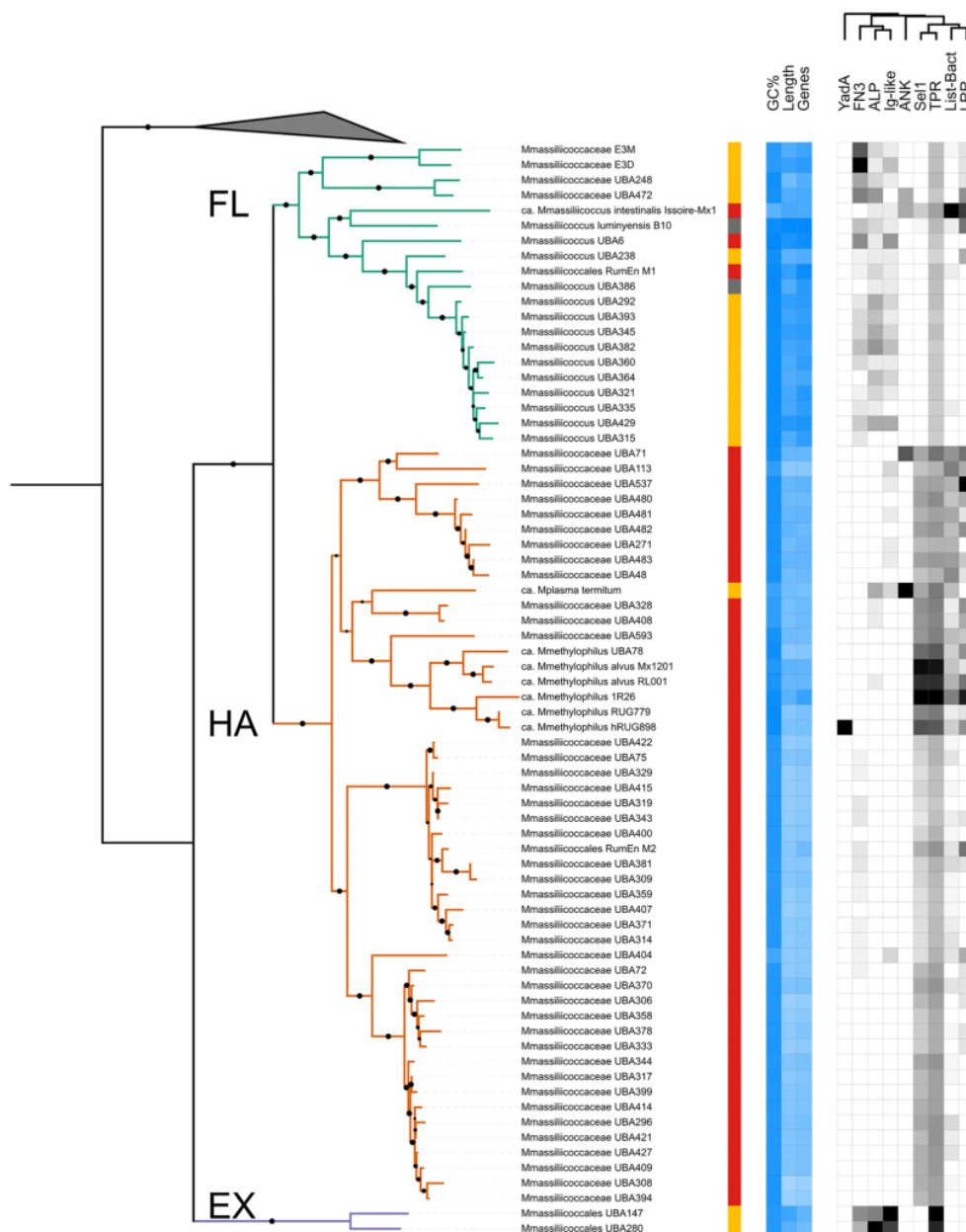
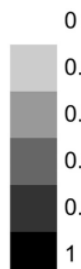
## Biome Enrichment

- Gut
- Non-significant
- Environmental

## Genome Characteristics



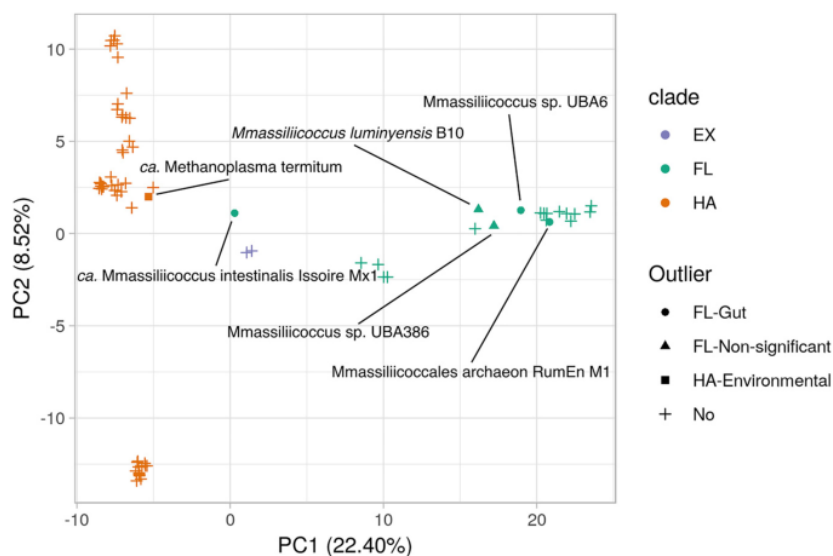
## Adhesion proteins



**FIG 3** Genome characteristics and adhesion proteins of *Methanomassiliicoccales* reflect division of the order into clades. Note that members of clade FL not enriched in environmental biomes resemble those of clade HA. The phylogeny is the same as that shown in Fig. 1. The colored strip summarizes the biome enrichment analysis. Heatmaps show genome features, including genome GC content ("GC%," range, 41.26% to 62.74%), genome length ("Length," 969,311 bp to 2,620,233 bp), and number of predicted genes ("Genes," 1,057 to 2,607) (blue scale), or a repertoire of eukaryote-like proteins: YadA-like domain (YadA; 0, 1), fibronectin type III (FN3; 0, 20) domains, bacterial Ig-like domains (Ig-like; 0, 12), ankyrin repeats (ANK; 0, 3), Sel1-containing proteins (Sel1; 0, 29), tetratricopeptide repeats (TPR; 7, 40), *Listeria-Bacteroides* repeat-containing proteins (List-Bact; 0, 26), and leucine-rich repeats (LRR; 0, 9) (gray scale shows columns ordered by hierarchical clustering); and adhesion-like proteins (ALP; 0, 12). On both heatmaps, the color intensity of each feature is relative to the maximum value of each category.

and EX taxa tended to have a lower GC content than clade FL taxa (FL,  $59.1\% \pm 4.8\%$ ; HA,  $55.8\% \pm 2.8\%$ ; EX,  $54.4\% \pm 0.5\%$ ; adj.  $P=0.6$ ).

To compare gene presence and absence across clades, we performed a pangenome analysis. After identification of orthologous gene clusters based on sequence similarity using panX software, we obtained 13,695 clusters, of which 7,312 were present at least once in clade FL, 6,592 in clade HA, and 1,833 in clade EX. A large proportion of gene



**FIG 4** Ordination of gene content of *Methanomassiliicoccales* group taxa by phylogenetic clade rather than by biome enrichment. Principal-component analysis of the gene cluster presence of taxa from clades FL (green), HA (orange), and EX (purple). Highlighted points correspond to outliers, namely, taxa either not significantly enriched in environmental or gut biomes or with enrichment opposite to expectation given their clade.

clusters were of unknown function according to the clusters of orthologous genes (COG) functional classification ( $38.4\% \pm 4.3\%$ ); gene clusters of unknown function tended to be detected in one or two genomes (see Fig. S4A in the supplemental material).

Principal component analysis (PCA) of gene cluster presence/absence differentiated clades along principal component 1 (PC1) (Fig. 4). We defined outlier taxa as FL taxa enriched in gut biomes (*Methanomassiliicoccales* archaeon RumEn M1, *Methanomassiliicoccus* sp. UBA6, “*Ca. M. intestinalis*” Issoire-Mx1, *M. luminyensis* B10, and *Methanomassiliicoccus* sp. UBA386) and the HA taxon enriched in non-host biomes (“*Ca. M. termitum*”). Outliers mostly clustered with their close relatives, not with the taxa enriched in the same biome (Fig. 4), with the exception of “*Ca. M. intestinalis*” Issoire-Mx1, which did not cluster with either clade.

#### Gene clusters enriched in clade HA evidence adaptation to the gut environment.

Because of the small number of genomes that cluster within clade EX, and because these are largely absent from animal-associated samples, subsequent analyses focus on comparisons between clades FL and HA.

To identify gene clusters potentially involved in the adaptation of members of Clade HA to a host environment, we compared the gene cluster content between clades. The gene cluster frequency spectrum shows many clusters present in few genomes; 7,990 (58.3%) gene clusters were singletons, and 2,002 (14.6%) were doubletons (Fig. S4A and B). After removing rare gene clusters by filtering those with near-zero variance, we included 2,937 clusters, which we then used to perform in phylogenetic ANOVA. Results reveal 14 gene clusters significantly enriched in HA compared to FL (adj.  $P < 0.1$  in all cases) (Table 1). Three gene clusters are involved in detoxification and xenobiotic metabolism, namely, those encoding bile acid: sodium symporter, bleomycin resistance protein and HAD superfamily hydrolase. Two clusters are related to shikimate or chorismate metabolism, namely, those encoding chorismate mutase II and prephenate dehydratase. Other annotated clusters include the small unit of exonuclease VII, Holliday junction resolvase Hjc, nitrogen regulatory protein PII, xylose isomerase-like protein, and metal-binding domain containing protein; four had poor or no

**TABLE 1** InterPro, eggNOG, and Prokka annotations of gene clusters significantly enriched in clade HA compared to clade FL

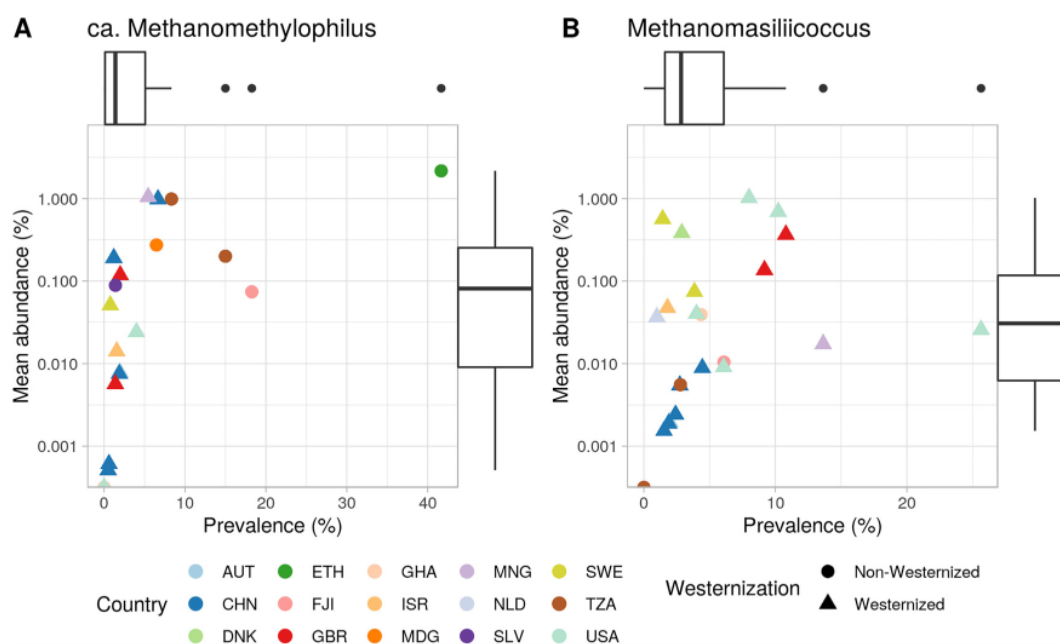
InterPro accession no.	InterPro annotation	NOG accession no.	COG category <sup>a</sup>	Prokka gene name	Prokka annotation
IPR002657	Bile acid:sodium symporter/arsenical resistance protein Acr3	COG0385@NOG	S		Hypothetical protein
IPR029068	Glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase	-	X		Hypothetical protein
IPR006357	HAD superfamily hydrolase, subfamily IIA	COG0647@NOG	G	<i>gph</i>	Glyceraldehyde 3-phosphate phosphatase
IPR002701	Chorismate mutase II, prokaryotic-type	COG1605@NOG	E	<i>aroQ</i>	Chorismate mutase
IPR001086	Prephenate dehydratase	COG0077@NOG	E	<i>pheA</i>	Prephenate dehydratase
IPR003761	Exonuclease VII, small subunit	COG1722@NOG	L	<i>xseB</i>	Exodeoxyribonuclease 7 small subunit
IPR002732	Holliday junction resolvase Hjc	COG1591@NOG	L	<i>rutD</i>	Putative aminoacylate hydrolase RutD
IPR015867	Nitrogen regulatory protein PII/ATP phosphoribosyltransferase, C-terminal	COG3323@NOG	S		Hypothetical protein
IPR013022	Xylose isomerase-like, TIM barrel domain	11IHC@NOG	L		Hypothetical protein
IPR019271	Protein of unknown function DUF2284, metal-binding	11RTN@NOG	S		Hypothetical protein

<sup>a</sup>COG functional classification descriptions: E, amino acid transport and metabolism; L, replication, recombination, and repair; G, carbohydrate transport and metabolism; S, function unknown; X, no annotation retrieved.

annotation (Table 1). Similar results were obtained when we performed this analysis without outlier taxa and when biome enrichment was used as independent variable (not shown), further indicating that genomic adaptations differ by clade, not habitat preference. Likewise, 89 clusters were enriched in clade FL compared to HA; these are presented in Table S1E.

**Genomic adaptations to the gut of members of the FL clade.** To determine whether outlier taxa belonging to clade FL had similar adaptations to the gut to those of members of clade HA, we explored gene clusters that were present in these outliers and in clade HA but were rare in other members of clade FL. We selected gene clusters present in the core genome of clade HA (i.e., present in at least 40 taxa or 80% of this clade, see Text S1 in the supplemental material) and present in less than half of the FL taxa. A total of 15 gene clusters were obtained, most of them encoded by only one of the outlier taxa. Two gene clusters, ferrous iron transport proteins A and B (InterPro accession numbers IPR030389 and IPR007167), were present in three of the outliers (*M. luminyensis*, “*Ca. M. intestinalis*” Isoire-Mx1, and *Methanomassiliicoccales* archaeon RumEn M1). Other clusters detected in more than one outlier included an uncharacterized membrane protein (InterPro number IPR005182, in *Methanomassiliicoccus* sp. UBA6 and *Methanomassiliicoccales* archaeon RumEn M1), a putative nickel-responsive regulator (InterPro number IPR014864, in *M. luminyensis* B10 and *Methanomassiliicoccus* sp. UBA386), and an ABC transporter (InterPro number IPR037294, in *M. luminyensis* B10 and *Methanomassiliicoccus* sp. UBA386). The remaining gene clusters, detected once, corresponded to transcriptional regulators or proteins of unknown function.

**The repertoire of adhesion proteins tended to differ between clades HA and FL.** We compared between FL and HA clades two large groups of membrane proteins involved in adhesion, namely eukaryote-like proteins (ELPs), a series of protein families involved in microbial adherence to the host (24), and adhesin-like proteins (ALPs), a class of proteins hypothesized to be involved in the microbe-microbe interactions of *Methanobacteriales* in the gut (13). We aggregated the counts of gene clusters annotated as the ALP and ELP classes and performed phylogenetic ANOVA. This analysis showed a trend toward differing repertoires of adhesion proteins by clade (Fig. 3), although we did not observe significant differences in the frequency of these factors (adj.  $P > 0.1$  in all cases). Taxa from clade HA tended to have higher mean counts of tetratricopeptide repeats (TPR) (mean  $\pm$  SD counts: HA,  $16.30 \pm 6.56$ , and FL,  $9.55 \pm 1.70$ ), Sel1-containing repeats (Sel1) (HA,  $9.32 \pm 5.69$ , and FL,  $0.35 \pm 1.35$ ), *Listeria-Bacteroides* repeats (List-Bact) (HA,  $3.68 \pm 3.76$ , and FL,  $1.65 \pm 5.78$ ), and leucine-rich repeats (LRR) (HA,  $1.5 \pm 2.15$ , and FL,  $1.1 \pm 2.02$ ) than FL taxa. Conversely, adhesin-like proteins (ALPs) (FL,  $2.25 \pm 1.48$ , and



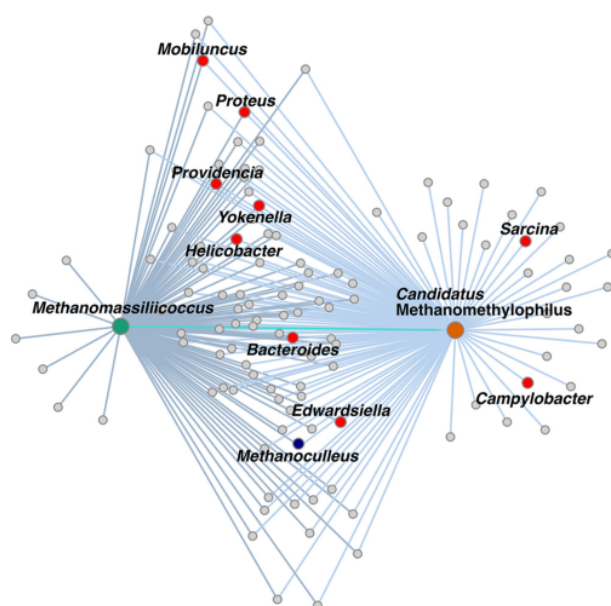
**FIG 5** *Methanomassiliicoccales* are rare members of the human gut microbiota. Scatterplots of the genera (A) “*Ca. Methanomethylophilus*” and (B) *Methanomassiliicoccus* show that their prevalence and mean abundance is low across most studies and populations (35 data sets) with subjects ( $n=4,472$ ) from Austria (AUT), China (CHN), Denmark (DNK), Ethiopia (ETH), Fijo (FJI), Great Britain (GBR), Ghana (GHA), Israel (ISR), Madagascar (MDG), Mongolia (MNG), The Netherlands (NLD), El Salvador (SLV), Sweden (SWE), Tanzania (TZA), and the United States (USA).

HA,  $0.14 \pm 0.61$ ), Ig-like domains (FL,  $1.55 \pm 1.32$ , and HA,  $0.20 \pm 0.53$ ) and fibronectin type III (FN3) domains (FL,  $4.55 \pm 5.09$ , and HA,  $0.32 \pm 0.62$ ) tended to be more abundant in the genomes of members of clade FL. We did not detect invasion protein B (IalB) in any of the analyzed genomes.

Hierarchical clustering based on the presence or absence of adhesion factors largely grouped *Methanomassiliicoccales* taxa by clade (see Fig. S5 in the supplemental material). Additionally, all adhesion factors, with the exceptions of ankyrin repeats (ANK) and the *Yersinia* adhesin A-like domain (YadA), showed a significant phylogenetic signal (adj.  $P < 0.05$  in all cases), further highlighting that closely related taxa had similar counts.

Interestingly, outlier taxa from clade FL had gene counts of several of the adhesion factors higher than the mean of their own clade and more characteristic of clade HA. In some cases, the gene counts were higher than the mean for clade HA. These included *Listeria-Bacteroides* repeats (gene cluster counts: *M. luminyensis*, 2; “*Ca. M. intestinalis*” Isoire-Mx1, 26; *Methanomassiliicoccales* archaeon RumEn M1, 2), Sel1 repeats (*M. luminyensis*, 1; “*Ca. M. intestinalis*” Isoire-Mx1, 6), and leucine-rich repeats (*M. luminyensis*, 5; “*Ca. M. intestinalis*” Isoire-Mx1, 7).

***Methanomassiliicoccales* taxa cooccur with each other, with other Archaea, and with TMA-producing bacteria in the human gut.** We characterized the distribution of *Methanomassiliicoccales* spp. across a collection of human gut metagenomes derived from 34 studies. Together, the combined 4,472 samples represented people from 22 countries, resulting in 35 unique data sets (i.e., study-country combinations). Across the whole set, we detected just two genera, *Methanomassiliicoccus* (clade FL) and “*Ca. Methanomethylophilus*” (clade HA), both rare members of the human gut microbiota (Fig. 5). “*Ca. Methanomethylophilus*” was detectable in 19 out of 35 data sets; in these 19 data sets, it had a prevalence ranging from 0.5% to 41.7%, and mean abundance ranged from  $4.8 \times 10^{-6}\%$  to  $2.2 \times 10^{-2}\%$ . Similarly, *Methanomassiliicoccus* was detectable in 22 of the 35 data sets; in the 22 data sets, it had a prevalence range of 1% to 25.7% and a mean abundance range of  $1.5 \times 10^{-5}\%$  to  $1.0 \times 10^{-2}\%$  (Table S1D).



**FIG 6** Coabundance networks of *Methanomassiliicoccus* (green node, dark edges) and “*Ca.* Methanomethylophilus” (orange node, light edges) in the human gut largely overlap. Both *Methanomassiliicoccales* genera are significantly coabundant (cyan edge). Their abundances are also coordinated with those of another archaeon (blue node) and TMA-producing bacterial taxa (red nodes).

We tested associations of these two genera with age, sex, and Westernization status of the subjects using linear mixed models that included the data set and country as random effects. Subjects from non-Westernized countries had a significantly higher prevalence of “*Ca.* Methanomethylophilus” (mean prevalence  $\pm$  SD: non-Westernized = 8.9%  $\pm$  28.5%, Westernized = 1.1%  $\pm$  10.3%; adj.  $P$  = 0.002). Westernized individuals were more likely to harbor higher prevalences of *Methanomassiliicoccus*, although differences were not significant (non-Westernized = 3.9%  $\pm$  19.4%, Westernized, 5.0%  $\pm$  21.7%; adj.  $P$  > 0.1). The age and sex of the individuals did not explain variance in the prevalence or abundance of either genus (adj.  $P$  > 0.1 in all cases).

To identify other microbial taxa positively associated with members of *Methanomassiliicoccales* in the human gut, we calculated a network of positively associated microorganisms (i.e., coabundant taxa) across samples ( $\rho$  > 0.1 in all cases) (25). In addition, we determined which taxa were present with members of *Methanomassiliicoccales* at a greater prevalence than that expected by chance (i.e., cooccurring taxa) relative to a permuted null model (26). Results showed that both “*Ca.* Methanomethylophilus” and *Methanomassiliicoccus* were part of the same coabundance network, together with a third archaeal genus, *Methanoculleus* (order *Methanomicrobiales*). We did not find evidence of positive or negative abundance associations of either *Methanomassiliicoccales* genus with *Methanobrevibacter*. Cooccurrence analysis showed a random association pattern between these taxa ( $P$  > 0.05 for both “*Ca.* Methanomethylophilus” and *Methanomassiliicoccus*), indicating that their ecological niches do not overlap that of *Methanobrevibacter*.

Analysis of the combined network of “*Ca.* Methanomethylophilus” and *Methanomassiliicoccus* revealed a large overlap between taxa associated with either genus (Fig. 6 and Table S1F): out of 119 taxa in the network, 86 (72.3%) were associated with both. Moreover, 51 taxa (42.9%) also had a significant positive cooccurrence pattern with both genera (adj.  $P$  < 0.05 in all cases). Most bacterial members of this network had low relative abundances; only *Bacteroides* and *Parabacteroides* had a mean relative abundance above 1% (range, 22.7% to 0.0005%). Interestingly, they included taxa that can potentially produce TMA, since their genomes contain genes encoding enzymes involved in its

synthesis; these taxa included *Bacteroides*, *Campylobacter*, *Yokenella*, *Mobiluncus*, *Proteus*, *Providencia*, and *Edwardsiella* (27).

**Abundance of *Methanomassiliicoccales* species is not concordant in monozygotic or dizygotic human twins.** To evaluate whether host genetics influences the abundance of *Methanomassiliicoccales* in the human gut, we compared the intraclass correlation coefficient (ICC) of their abundances at the genus level using a set of 153 monozygotic (MZ) and 200 dizygotic (DZ) twin pairs from the TwinsUK cohort. As a control, we first compared the mean ICC across all taxa between MZ and DZ twins and found that  $ICC_{MZ}$  (0.1) was significantly higher than  $ICC_{DZ}$  (0.03) ( $P < 0.01$ ). In addition, we assessed the ICC values of bacterial (*Christensenella*, *Faecalibacterium*, and *Bifidobacterium*) and archaeal (*Methanobrevibacter*) genera, and consistently found a higher correlation for MZ compared to DZ twins (Table S1G). We were only able to assess ICC values of *Methanomassiliicoccus*, as it was the only *Methanomassiliicoccales* taxon detected in the twins with a prevalence (8.64%) above the 5% cutoff (see Materials and Methods). We did not detect a significant concordance between the abundances of *Methanomassiliicoccus* in MZ ( $ICC_{MZ} = 0.004$ ; adj.  $P = 0.59$ ) or in DZ twins ( $ICC_{DZ} = 0.017$ ; adj.  $P = 0.71$ ). Given the low abundance of *Methanomassiliicoccales* taxa, we performed a sensitivity analysis using samples with a high sequencing depth (>12 million reads/sample); however, we did not observe differences in the abundance and prevalence of the *Methanomassiliicoccales* genera or in the ICC estimates (data not shown).

## DISCUSSION

While the source of the members of the *Methanomassiliicoccales* has been noted in previous surveys of single markers such as 16S rRNA and *mcrA* genes (10, 11), here, we searched metagenomes from host-associated and environmental samples for their relative abundances. Overall, the HA taxa were enriched in host-associated samples and the FL taxa were enriched in environmental samples; intriguingly, all taxa, regardless of clade, were detected in both biomes. This suggests that members of the order *Methanomassiliicoccales* are generalists with an overall habitat preference according to clade, although there were some exceptions to the general pattern. We show that members of *Methanomassiliicoccales* use many of the same adaptations to the gut as other methanogens. These adaptations include genome reduction and genes involved in the shikimate pathway and bile resistance. In addition, gut-enriched taxa tend to have a distinct repertoire of genes encoding adhesion factors. We observed that potential adaptations to the gut differed by clade, not preferred habitat, indicating convergence on a shared niche through different genomic solutions. In the human gut, *Methanomassiliicoccales* taxa correlated with TMA-producing bacteria, rather than host genetics or other host factors.

For members of the HA clade, adaptations to life in the gut included an enrichment of genes involved in bile acid transport, efflux pumps, and hydrolases, which play a role in tolerance to these compounds in the gastrointestinal tract (28). This adaptation is also shared with other members of the gut microbiota, including *Methanobacteriales* taxa; *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* are resistant to bile salts (3, 4). Other gene clusters with known functions enriched in clade HA are involved in metabolism of shikimate and chorismate. The shikimate pathway is involved in the synthesis of aromatic amino acids in plants and microbes, but it is absent in mammals. Shikimate metabolism is carried out by archaeal (29) and bacterial (30, 31) members of the animal gut microbiota and was reported as one of the most conserved metabolic modules in a large-scale gene catalogue from the human gut (32). In turn, aromatic amino acids can be transformed by the gut microbiota into active metabolites, which are involved in diverse physiological processes (33) and conditions such as cardiovascular disease (34). Indeed, plasma concentrations of microbial derivatives of tryptophan have even been shown to negatively correlate with atherosclerosis (35). It remains to be elucidated whether *Methanomassiliicoccales* are involved in human health through the metabolism of aromatic amino acids and associated compounds.

We observed that each clade tended to encode different adhesion factors, although without statistical significance. These factors are involved in the maintenance of syntrophic relationships of the methanogens with bacterial (12, 36) or eukaryotic (37) microorganisms. Two groups of adhesion factors, proteins containing Sel1 domains and *Listeria-Bacteroides* repeats, have been previously studied in *Methanomassiliicoccales* taxa retrieved from the gut (9, 23). Our assessment of these factors in the broader context of the order *Methanomassiliicoccales* showed that these two groups are more likely to be higher in clade HA than clade FL taxa, with the exception of the outlier taxa. Indeed, the repertoire of ELPs and ALPs was similar between species inhabiting the gut, regardless of their clade. This emphasizes the potential involvement of these proteins in the adaptation to the intestinal environment, although the exact mechanisms are yet to be elucidated.

In contrast, members of clade FL appear to be generalists that colonized the animal gut independently from the HA clade. It has been previously noted that *M. luminyensis*, an outlier from clade FL, could have a facultative association to the animal gut. It possesses genes involved in nitrogen fixation, oxidative stress (9), and mercury methylation (23), which are common in soil microorganisms but rare in members of the gut microbiota (38). In accordance with this, we observed that members of clade FL are widespread and abundant in soil, water, and gut metagenomes, with a preference for environmental biomes. Similarities in ELP content between gut-dwelling taxa from both clades indicate that interaction with the host or other members of the gut microbiota might be a key factor in the adaptation of these methanogens.

Analysis of the gene content of outlier taxa from clade FL showed that they tended to be more similar to members of their own clade than to taxa from clade HA, with the exception of “*Ca. M. intestinalis*” Issoire-Mx1, which was distinct from either clade FL and HA. In addition, there was little overlap in gene clusters commonly observed in clade HA and outlier taxa from clade FL, with the exception of the adhesion factors discussed above. These observations support the hypothesis that colonization of animal guts by members of *Methanomassiliicoccales* occurred in two independent events (9, 23), and suggests that there is not one solution to life in the gut for these archaea, as members from two clades seem to have solved the problem with a different set of adaptations.

Characterization of the abundance of *Methanomassiliicoccales* taxa across human populations showed members of this group are rare in the microbiota of healthy adults. We did not detect them in all the studied populations, and, when detected, they had low prevalence and abundance. Our extensive analysis of human gut samples corroborates estimates of *Methanomassiliicoccales* prevalence (up to 11%) (23, 39, 40) and mean abundance (below 1%) (23, 41). Differences in *Methanomassiliicoccales* carriage between Westernized and non-Westernized populations remain to be explained, and may be due to diet. While Westernized diets are richer in TMA precursors than non-Western diets (42), intake varies across populations (43).

Our analysis allowed us to assess whether *Archaea* in the human gut are mutually exclusive. We observed positive correlations of “*Ca. Methanomethylophilus*” and *Methanomassiliicoccus* with each other and with *Methanoculleus*, another rare archaeal member of the gut microbiota (45). We did not find evidence of association between members of *Methanomassiliicoccales* and *Methanobrevibacter*, positive or otherwise, confirming the previous report that these methanogens are not mutually exclusive (39); abundance of H<sub>2</sub> in the gut, together with differences in other substrate utilization, might result in nonoverlapping niches (46).

While genus *Methanobrevibacter* was consistently found to have a moderate heritability in the TwinsUK (19, 39, 47) and other cohorts (48, 49), this was not the case for members of *Methanomassiliicoccales*. Similarly to humans, methane production (50) and abundance of *Methanobrevibacter* (51) are also heritable in bovine cattle, but *Methanomassiliicoccales* taxa are not (51). Thus, host genetics might be linked to particular taxa and methanogenesis pathways, not to all *Archaea* or to methane production as a whole.

Genera “*Ca. Methanomethylophilus*” and *Methanomassiliicoccus* cooccur with TMA-producing bacteria (27), further supporting their potential use as a way of targeting intestinal TMA (52). The exact nature of the ecological relationships each of these taxa establishes with other members of the microbiome remains to be elucidated. In a facilitation scenario between the methanogens and H<sub>2</sub> and TMA producers, freely available TMA and H<sub>2</sub> required for methylotrophic methanogenesis could be utilized by *Methanomassiliicoccales* taxa (53) without cost to the producer. Alternatively, the methanogens could establish syntrophic interactions with other microorganisms, whereby the consumption of these metabolites is also beneficial to the producer (53).

The present study extends our understanding of the order *Methanomassiliicoccales* by revealing genomic adaptations to life in the gut by members of both clades that make up this group. Furthermore, the positive correlation between the relative abundances of these TMA-utilizing *Archaea* with TMA-producing bacteria in the gut is a first step toward understanding how they may be harnessed for therapeutic management of gut TMA levels in the context of cardiovascular disease.

## MATERIALS AND METHODS

For a detailed description of the methods, see Text S1 in the supplemental material.

**Genome annotation and phylogenomic tree reconstruction.** We used 71 substantially complete genomes (completeness,  $\geq 70\%$ ) with low contamination (contamination,  $< 5\%$ ) retrieved from the NCBI Assembly database (<https://www.ncbi.nlm.nih.gov/assembly>), plus an additional high-quality metagenome-assembled genome (MAG) corresponding to “*Candidatus Methanomethylophilus alvus*.” Gene calling was performed using Prokka (54). A maximum-likelihood phylogenomic tree was constructed using PhyloPhlAn (55) with the 72 *Methanomassiliicoccales* genomes plus members of the order *Thermoplasmatales* as an outgroup. We used interactive Tree Of Life (iTOL) (56) to visualize the tree.

**Abundance of *Methanomassiliicoccales* in environmental and animal gastrointestinal metagenomes.** We retrieved 305 publicly available gastrointestinal and environmental metagenome samples (57) (see Table S1B in the supplemental material). To avoid multiple mapping of reads, we dereplicated the 72 genomes at a species level (95% average nucleotide identity [ANI]) using dRep (58), resulting in 29 representative genomes. We quantified the abundance of dereplicated *Methanomassiliicoccales* genomes in the metagenomes using KrakenUniq (59). We estimated the enrichment of each representative *Methanomassiliicoccales* taxon in host or environmental metagenomes using DESeq2 with the Wald test (60) on sequence counts and classifying metagenome samples as either host derived or environmental.

**Comparative genomics.** We grouped the predicted genes into gene clusters using panX (61) and used InterProScan (62) and eggNOG-mapper (63) for annotation. Phylogenetic signal of genome characteristics and gene cluster presence was tested using the *phyloSignal* R package with the local indicator of phylogenetic association (LIPA) (64). The R package *micropan* (65) was used to create a pangenome principal-component analysis (PCA). We performed phylogenetic ANOVA using the R package *phytools* (66) to determine clusters enriched in clades FL or HA. We adjusted *P* values for multiple comparisons with the Benjamini-Hochberg method. Due to the exploratory nature of this work, tests were considered significant if they had an adjusted *P* value (adj. *P*) of  $< 0.1$ ; a false-discovery-rate-adjusted *P* value cutoff of 0.1 implies that 10% of significant tests will result in false-positives. In cases where adjusting *P* values was not necessary, raw *P* values are provided.

We assessed the presence of eukaryote-like proteins (ELPs) (24) by combining the counts of gene clusters classified as Sel1-containing proteins (Sel1), *Listeria-Bacteroides* repeat-containing proteins (List-Bact), tetratricopeptide repeats (TPR), ankyrin repeats (ANK), leucine-rich repeats (LRR), fibronectin type III (FN3) domains, laminin G domains, bacterial Ig-like domains, *Yersinia* adhesin A-like domain (YadA), TadE-like domain, or invasion protein B (IaIB). Likewise, we characterized the presence of parallel beta-helix-repeat-containing proteins, also known as adhesin-like proteins (ALPs).

**Characterization of *Methanomassiliicoccales* distribution across human populations.** We retrieved and quality controlled 4,472 publicly available human gut metagenomes from 34 independent studies (Table S1C). Reads were classified using Kraken (67) and Bracken (68) with custom databases (69). Taxa with  $< 100$  reads in a given sample were considered absent. To determine the cooccurrence patterns of *Methanomassiliicoccales* in the human gut, we used the *cooccur* package (26); to determine their coabundance patterns, we calculated the proportionality of taxa abundance ( $\rho$ ) with the *propr* R package (70). The *lme4* and *lmerTest* R packages (71) were used to fit linear mixed effects models to test differences of *Methanomassiliicoccales* genera abundance by Westernization status, age, and gender. We employed binomial linear mixed models to test differences in genera prevalence. Lists of potential TMA (27, 72, 73) and methanol (74) (see Text S1 in the supplemental materials) producers were compiled from the available literature.

Heritability of *Methanomassiliicoccales* taxa was assessed by comparing relative abundances of taxa within 153 monozygotic (MZ) and 200 dizygotic (DZ) twin pairs from the United Kingdom Adult Twin Registry (TwinsUK) (19, 39, 75). Absolute read counts were transformed using the Yeo-Johnson transformation and adjusted by body mass index (BMI), sex, and sequencing depth (19, 39). We calculated the intraclass correlation coefficient (ICC) in MZ and DZ twins with the *irr* R package, and adjusted *P* values using the Benjamini-Hochberg method. We compared the mean ICC across all taxa between MZ and DZ



twins using the Mann-Whitney test and by assessing the ICC of taxa previously reported as heritable (*Methanobrevibacter*, *Faecalibacterium*, *Christensenella*, and *Bifidobacterium*) (39, 48).

**Data availability.** The raw sequence data are available from the European Nucleotide Archive under study accession number PRJEB40256. Jupyter notebooks are available at <https://github.com/lelyabmpi/Methanomassiliu>. The “*Candidatus* Methanomethylophilus” MAG generated here can be found at <http://ftp.tue.mpg.de/ebio/projects/Mmassiliu/>.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**TEXT S1**, DOCX file, 0.02 MB.

**FIG S1**, TIF file, 2 MB.

**FIG S2**, TIF file, 0.5 MB.

**FIG S3**, TIF file, 2.1 MB.

**FIG S4**, TIF file, 0.5 MB.

**FIG S5**, TIF file, 0.9 MB.

**TABLE S1**, XLSX file, 0.2 MB.

## ACKNOWLEDGMENTS

This work was supported by the Max Planck Society. The study also received support from the National Institute for Health Research (NIHR) BioResource Clinical Research Facility and Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. We thank EMBO and the organizers and participants of the Bioinformatics and Genome Analyses course held at the Fondazione Edmund Mach in San Michele all’Adige, Italy, for sponsoring the attendance of J.D.L.C.-Z. and for their feedback.

We are also grateful to Daphne Welter, Jessica Sutter, and Albane Ruaud for the fruitful discussions and comments.

We declare no competing interests.

## REFERENCES

- Borrel G, Brugère J-F, Gribaldo S, Schmitz RA, Moissl-Eichinger C. 2020. The host-associated archaeome. *Nat Rev Microbiol* 18:622–636. <https://doi.org/10.1038/s41579-020-0407-y>.
- Moissl-Eichinger C, Pausan M, Taffner J, Berg G, Bang C, Schmitz RA. 2018. Archaea are interactive components of complex microbiomes. *Trends Microbiol* 26:70–85. <https://doi.org/10.1016/j.tim.2017.07.004>.
- Miller TL, Wolin MJ. 1982. Enumeration of *Methanobrevibacter smithii* in human feces. *Arch Microbiol* 131:14–18. <https://doi.org/10.1007/BF00451492>.
- Miller TL, Wolin MJ. 1985. *Methanosphaera stadtmaniae* gen. nov., sp. nov.: a species that forms methane by reducing methanol with hydrogen. *Arch Microbiol* 141:116–122. <https://doi.org/10.1007/BF00423270>.
- Horz H-P, Conrads G. 2010. The discussion goes on: what is the role of *Euryarchaeota* in humans? *Archaea* 2010:967271. <https://doi.org/10.1155/2010/967271>.
- Borrel G, Harris HMB, Tottey W, Mihajlovski A, Parisot N, Peyretailade E, Peyret P, Gribaldo S, O’Toole PW, Brugère J-F. 2012. Genome sequence of “*Candidatus* Methanomethylophilus alvus” Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens. *J Bacteriol* 194:6944–6945. <https://doi.org/10.1128/JB.01867-12>.
- Borrel G, Harris HMB, Parisot N, Gaci N, Tottey W, Mihajlovski A, Deane J, Gribaldo S, Bardot O, Peyretailade E, Peyret P, O’Toole PW, Brugère J-F. 2013. Genome sequence of “*Candidatus* Methanomassiliicoccus intestinalis” Issoire-Mx1, a third thermoplasmatales-related methanogenic archaeon from human feces. *Genome Announc* 1:e00453-13.
- Dridi B, Fardeau M-L, Ollivier B, Raoult D, Drancourt M. 2012. *Methanomassiliicoccus luminyensis* gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *Int J Syst Evol Microbiol* 62:1902–1907. <https://doi.org/10.1099/ijs.0.033712-0>.
- Borrel G, Parisot N, Harris HMB, Peyretailade E, Gaci N, Tottey W, Bardot O, Raymann K, Gribaldo S, Peyret P, O’Toole PW, Brugère J-F. 2014. Comparative genomics highlights the unique biology of *Methanomassiliicoccales*, a *Thermoplasmatales*-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* 15:679. <https://doi.org/10.1186/1471-2164-15-679>.
- Söllinger A, Schwab C, Weinmaier T, Loy A, Tveit AT, Schleper C, Urich T. 2016. Phylogenetic and genomic analysis of *Methanomassiliicoccales* in wetlands and animal intestinal tracts reveals clade-specific habitat preferences. *FEMS Microbiol Ecol* 92:fiv149. <https://doi.org/10.1093/femsec/fiv149>.
- Speth DR, Orphan VJ. 2018. Metabolic marker gene mining provides insight in global diversity and, coupled with targeted genome reconstruction, sheds further light on metabolic potential of the. *PeerJ* 6:e5614. <https://doi.org/10.7717/peerj.5614>.
- Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, Fulton R, Latreille P, Kim K, Wilson RK, Gordon JL. 2007. Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. *Proc Natl Acad Sci U S A* 104:10643–10648. <https://doi.org/10.1073/pnas.0704189104>.
- Hansen EE, Lozupone CA, Rey FE, Wu M, Guruge JL, Narra A, Goodfellow J, Zaneveld JR, McDonald DT, Goodrich JA, Heath AC, Knight R, Gordon JL. 2011. Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc Natl Acad Sci U S A* 108 Suppl 1:4599–4606. <https://doi.org/10.1073/pnas.1000071108>.
- Söllinger A, Urich T. 2019. Methylophilic methanogens everywhere—physiology and ecology of novel players in global methane cycling. *Biochem Soc Trans* 47:1895–1907. <https://doi.org/10.1042/BST20180565>.
- Brown JM, Hazen SL. 2018. Microbial modulation of cardiovascular disease. *Nat Rev Microbiol* 16:171–181. <https://doi.org/10.1038/nrmicro.2017.149>.
- Geng J, Yang C, Wang B, Zhang X, Hu T, Gu Y, Li J. 2018. Trimethylamine *N*-oxide promotes atherosclerosis via CD36-dependent MAPK/JNK pathway. *Biomed Pharmacother* 97:941–947. <https://doi.org/10.1016/j.biopha.2017.11.016>.
- Wang Z, Roberts AB, Buffa JA, Levison BS, Zhu W, Org E, Gu X, Huang Y, Zamanian-Daryoush M, Culley MK, DiDonato AJ, Fu X, Hazen JE, Krajcik D, DiDonato JA, Lusis AJ, Hazen SL. 2015. Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. *Cell* 163:1585–1595. <https://doi.org/10.1016/j.cell.2015.11.055>.
- Brugère J-F, Borrel G, Gaci N, Tottey W, O’Toole PW, Malpuech-Brugère C. 2014. Archaeobiotics: proposed therapeutic use of archaea to prevent

- trimethylaminuria and cardiovascular disease. *Gut Microbes* 5:5–10. <https://doi.org/10.4161/gmic.26749>.
19. Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, Ward KJ, Jackson MA, Xia Y, Chen X, Chen B, Xia H, Xu C, Li F, Xu X, Al-Aama JY, Yang H, Wang J, Kristiansen K, Wang J, Steves CJ, Bell JT, Li J, Spector TD, Jia H. 2016. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst* 3:572–584.e3. <https://doi.org/10.1016/j.cels.2016.10.004>.
  20. Paul K, Nonoh JO, Mikulski L, Brune A. 2012. “*Methanoplasmatales*,” *Thermoplasmatales*-related archaea in termite guts and other environments, are the seventh order of methanogens. *Appl Environ Microbiol* 78:8245–8253. <https://doi.org/10.1128/AEM.02193-12>.
  21. Borrel G, O’Toole PW, Harris HMB, Peyret P, Brugère J-F, Gribaldo S. 2013. Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis. *Genome Biol Evol* 5:1769–1780. <https://doi.org/10.1093/gbe/evt128>.
  22. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2:1533–1542. <https://doi.org/10.1038/s41564-017-0012-7>.
  23. Borrel G, McCann A, Deane J, Neto MC, Lynch DB, Brugère J-F, O’Toole PW. 2017. Genomics and metagenomics of trimethylamine-utilizing *Archaea* in the human gut microbiome. *ISME J* 11:2059–2074. <https://doi.org/10.1038/ismej.2017.72>.
  24. Alex A, Antunes A. 2018. Genus-wide comparison of *Pseudovibrio* bacterial genomes reveal diverse adaptations to different marine invertebrate hosts. *PLoS One* 13:e0194368. <https://doi.org/10.1371/journal.pone.0194368>.
  25. Quinn TP, Erb I, Richardson MF, Crowley TM. 2018. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34:2870–2878. <https://doi.org/10.1093/bioinformatics/bty175>.
  26. Griffith DM, Veech JA, Marsh CJ. 2016. cooccur: probabilistic species co-occurrence analysis in R. *J Stat Softw* 69:1–17.
  27. Fennema D, Phillips IR, Shephard EA. 2016. Trimethylamine and trimethylamine *N*-oxide, a flavin-containing monooxygenase 3 (FMO3)-mediated host-microbiome metabolic axis implicated in health and disease. *Drug Metab Dispos* 44:1839–1850. <https://doi.org/10.1124/dmd.116.070615>.
  28. Begley M, Gahan CGM, Hill C. 2005. The interaction between bacteria and bile. *FEMS Microbiol Rev* 29:625–651. <https://doi.org/10.1016/j.femsr.2004.09.003>.
  29. Hovey R, Lentens S, Ehrenreich A, Salmon K, Saba K, Gottschalk G, Gunsalus RP, Deppenmeier U. 2005. DNA microarray analysis of *Methanosarcina mazei* Gö1 reveals adaptation to different methanogenic substrates. *Mol Genet Genomics* 273:225–239. <https://doi.org/10.1007/s00438-005-1126-9>.
  30. Kamke J, Kittelmann S, Soni P, Li Y, Tavendale M, Ganesh S, Janssen PH, Shi W, Froula J, Rubin EM, Attwood GT. 2016. Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a *Sharpea*-enriched microbiome characterised by lactic acid formation and utilisation. *Microbiome* 4:56. <https://doi.org/10.1186/s40168-016-0201-2>.
  31. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. 2013. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr Opin Biotechnol* 24:160–168. <https://doi.org/10.1016/j.copbio.2012.08.005>.
  32. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2020. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 49:55.
  33. Lin R, Liu W, Piao M, Zhu H. 2017. A review of the relationship between the gut microbiota and amino acid metabolism. *Amino Acids* 49:2083–2090. <https://doi.org/10.1007/s00726-017-2493-3>.
  34. Liu Y, Hou Y, Wang G, Zheng X, Hao H. 2020. Gut microbial metabolites of aromatic amino acids as signals in host-microbe interplay. *Trends Endocrinol Metab* 31:818–834. <https://doi.org/10.1016/j.tem.2020.02.012>.
  35. Cason CA, Dolan KT, Sharma G, Tao M, Kulkarni R, Helenowski IB, Doane BM, Avram MJ, McDermott MM, Chang EB, Ozaki CK, Ho KJ. 2018. Plasma microbiome-modulated indole- and phenyl-derived metabolites associate with advanced atherosclerosis and postoperative outcomes. *J Vasc Surg* 68:1552–1562.e7. <https://doi.org/10.1016/j.jvs.2017.09.029>.
  36. Ruaud A, Esquivel-Elizondo S, de la Cuesta-Zuluaga J, Waters JL, Angenent LT, Youngblut ND, Ley RE. 2020. Syntrophy via interspecies H<sub>2</sub> transfer between and underlies their global cooccurrence in the human gut. *mBio* 11:e03235-19. <https://doi.org/10.1128/mBio.03235-19>.
  37. Ng F, Kittelmann S, Patchett ML, Attwood GT, Janssen PH, Rakonjac J, Gagic D. 2016. An adhesin from hydrogen-utilizing rumen methanogen *Methanobrevibacter ruminantium* M1 binds a broad range of hydrogen-producing microorganisms. *Environ Microbiol* 18:3010–3021. <https://doi.org/10.1111/1462-2920.13155>.
  38. Podar M, Gilmour CC, Brandt CC, Soren A, Brown SD, Crable BR, Palumbo AV, Somenahally AC, Elias DA. 2015. Global prevalence and distribution of genes and microorganisms involved in mercury methylation. *Sci Adv* 1:e1500675. <https://doi.org/10.1126/sciadv.1500675>.
  39. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human genetics shape the gut microbiome. *Cell* 159:789–799. <https://doi.org/10.1016/j.cell.2014.09.053>.
  40. Dridi B, Henry M, Richet H, Raoult D, Drancourt M. 2012. Age-related prevalence of *Methanomassiliicoccus luminyensis* in the human gut microbiome. *APMIS* 120:773–777. <https://doi.org/10.1111/j.1600-0463.2012.02899.x>.
  41. Vanderhaeghen S, Lacroix C, Schwab C. 2015. Methanogen communities in stools of humans of different age and health status and co-occurrence with bacteria. *FEMS Microbiol Lett* 362:fnv092. <https://doi.org/10.1093/femsle/fnv092>.
  42. Zeisel SH, Mar M-H, Howe JC, Holden JM. 2003. Concentrations of choline-containing compounds and betaine in common foods. *J Nutr* 133:1302–1307. <https://doi.org/10.1093/jn/133.5.1302>.
  43. Wiedeman AM, Barr SI, Green TJ, Xu Z, Innis SM, Kitts DD. 2018. Dietary choline intake: current state of knowledge across the life cycle. *Nutrients* 10:1513. <https://doi.org/10.3390/nu10101513>.
  44. Reference deleted.
  45. Horz H-P. 2015. Archaeal lineages within the human microbiome: absent, rare or elusive? *Life (Basel)* 5:1333–1345. <https://doi.org/10.3390/life5021333>.
  46. Feldewert C, Lang K, Brune A. 2020. The hydrogen threshold of obligately methyl-reducing methanogens. *FEMS Microbiol Lett* 367:fnaa137. <https://doi.org/10.1093/femsle/fnaa137>.
  47. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. 2016. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* 19:731–743. <https://doi.org/10.1016/j.chom.2016.04.017>.
  48. Goodrich JK, Davenport ER, Clark AG, Ley RE. 2017. The relationship between the human genome and microbiome comes into view. *Annu Rev Genet* 51:413–433. <https://doi.org/10.1146/annurev-genet-110711-155532>.
  49. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, Le Roy CI, Raygoza Garay JA, Finnicum CT, Liu X, Zhernakova DV, Bonder MJ, Hansen TH, Frost F, Rühlemann MC, Turpin W, Moon J-Y, Kim H-N, Lüll K, Barkan E, Shah SA, Fornage M, Szopinska-Tokov J, Wallen ZD, Borisevich D, Agreus L, Andreasson A, Bang C, Bedrani L, Bell JT, Bisgaard H, Boehnke M, Boomsma DI, Burk RD, Claringbould A, Croitoru K, Davies GE, van Duijn CM, Duijts L, Falony G, Fu J, van der Graaf A, Hansen T, Homuth G, Hughes DA, Ijzerman RG, Jackson MA, Jaddoe VWW, Joossens M, Jørgensen T, Keszhelyi D, Knight R, Laakso M, Laudes M, et al. 2020. Genetics of human gut microbiome composition. *bioRxiv* <https://doi.org/10.1101/2020.06.26.173724>.
  50. Roehe R, Dewhurst RJ, Duthie C-A, Rooke JA, McKain N, Ross DW, Hyslop JJ, Waterhouse A, Freeman TC, Watson M, Wallace RJ. 2016. Bovine host genetic variation influences rumen microbial methane production with best selection criterion for low methane emitting and efficiently feed converting hosts based on metagenomic gene abundance. *PLoS Genet* 12:e1005846. <https://doi.org/10.1371/journal.pgen.1005846>.
  51. Difford GF, Plichta DR, Lovendahl P, Lassen J, Noel SJ, Højberg O, Wright A-DG, Zhu Z, Kristensen L, Nielsen HB, Guldbandsen B, Sahana G. 2018. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLoS Genet* 14:e1007580. <https://doi.org/10.1371/journal.pgen.1007580>.
  52. Hania WB, Ballet N, Vandekerckove P, Ollivier B, O’Toole PW, Brugère J-F. 2017. Archaeobiotics: archaea as pharmabiotics for treating chronic disease in humans? p 42–62. *In* Sghaier H, Najjari A, Ghedira K (ed), *Archaea: new biocatalysts, novel pharmaceuticals and various biotechnological applications*. InTech Open, London, UK.
  53. Douglas AE. 2020. The microbial exometabolome: ecological resource and architect of microbial communities. *Philos Trans R Soc Lond B Biol Sci* 375:20190250. <https://doi.org/10.1098/rstb.2019.0250>.
  54. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
  55. Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* 4:2304. <https://doi.org/10.1038/ncomms3304>.
  56. Letunic I, Bork P. 2016. Interactive Tree Of Life (iTOL) v3: an online tool for

- the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44:W242–W245. <https://doi.org/10.1093/nar/gkw290>.
57. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter FMI, Ten Hoopen P, Alako B, Amid C, Wilkinson DJ, Curtis TP, Cochrane G, Finn RD. 2018. EBI metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res* 46:D726–D735. <https://doi.org/10.1093/nar/gkx967>.
  58. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 11:2864–2868. <https://doi.org/10.1038/ismej.2017.126>.
  59. Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol* 19:198. <https://doi.org/10.1186/s13059-018-1568-0>.
  60. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550. <https://doi.org/10.1186/s13059-014-0550-8>.
  61. Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. *Nucleic Acids Res* 46:e5. <https://doi.org/10.1093/nar/gkx977>.
  62. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang H-Y, Dosztányi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Xenarios I, Yeh L-S, Young S-Y, Mitchell AL. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* 45: D190–D199. <https://doi.org/10.1093/nar/gkx1107>.
  63. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* 34:2115–2122. <https://doi.org/10.1093/molbev/msx148>.
  64. Keck F, Rimet F, Bouchez A, Franc A. 2016. phylSignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol* 6:2774–2780. <https://doi.org/10.1002/ece3.2051>.
  65. Snipen L, Liland KH. 2015. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* 16:79. <https://doi.org/10.1186/s12859-015-0517-0>.
  66. Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223. <https://doi.org/10.1111/j.2041-210X.2011.00169.x>.
  67. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20:257. <https://doi.org/10.1186/s13059-019-1891-0>.
  68. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* 3:e104. <https://doi.org/10.7717/peerj-cs.104>.
  69. de la Cuesta-Zuluaga J, Ley RE, Youngblut ND. 2020. Struo: a pipeline for building custom databases for common metagenome profilers. *Bioinformatics* 36:2314–2315. <https://doi.org/10.1093/bioinformatics/btz899>.
  70. Quinn TP, Richardson MF, Lovell D, Crowley TM. 2017. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci Rep* 7:16252. <https://doi.org/10.1038/s41598-017-16520-0>.
  71. Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48.
  72. Rath S, Heidrich B, Pieper DH, Vital M. 2017. Uncovering the trimethylamine-producing bacteria of the human gut microbiota. *Microbiome* 5:54. <https://doi.org/10.1186/s40168-017-0271-9>.
  73. Rath S, Rud T, Pieper DH, Vital M. 2019. Potential TMA-producing bacteria are ubiquitously found in Mammalia. *Front Microbiol* 10:2966. <https://doi.org/10.3389/fmicb.2019.02966>.
  74. Dorokhov YL, Shindyapina AV, Sheshukova EV, Komarova TV. 2015. Metabolic methanol: molecular pathways and physiological roles. *Physiol Rev* 95:603–644. <https://doi.org/10.1152/physrev.00034.2014>.
  75. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohny RP, Li W, de Rinaldis E, Bell JT, Venter JC, Nelson KE, Spector TD, Falchi M. 2019. Interplay between the human gut microbiome and host metabolism. *Nat Commun* 10:4505. <https://doi.org/10.1038/s41467-019-12476-z>.



## Appendix III: Struo - a pipeline for building custom databases for common metagenome profilers

## Databases and ontologies

# Struo: a pipeline for building custom databases for common metagenome profilers

Jacobo de la Cuesta-Zuluaga, Ruth E. Ley and Nicholas D. Youngblut  \*

Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen 72076, Germany

\*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

Received on September 24, 2019; revised on November 18, 2019; editorial decision on November 24, 2019; accepted on November 26, 2019

## Abstract

**Summary:** Taxonomic and functional information from microbial communities can be efficiently obtained by metagenome profiling, which requires databases of genes and genomes to which sequence reads are mapped. However, the databases that accompany metagenome profilers are not updated at a pace that matches the increase in available microbial genomes, and unifying database content across metagenome profiling tools can be cumbersome. To address this, we developed Struo, a modular pipeline that automatizes the acquisition of genomes from public repositories and the construction of custom databases for multiple metagenome profilers. The use of custom databases that broadly represent the known microbial diversity by incorporating novel genomes results in a substantial increase in mappability of reads in synthetic and real metagenome datasets.

**Availability and implementation:** Source code available for download at <https://github.com/leylabmpi/Struo>. Custom genome taxonomy database databases available at <http://ftp.tue.mpg.de/ebio/projects/struo/>.

**Contact:** [nicholas.youngblut@tuebingen.mpg.de](mailto:nicholas.youngblut@tuebingen.mpg.de)

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

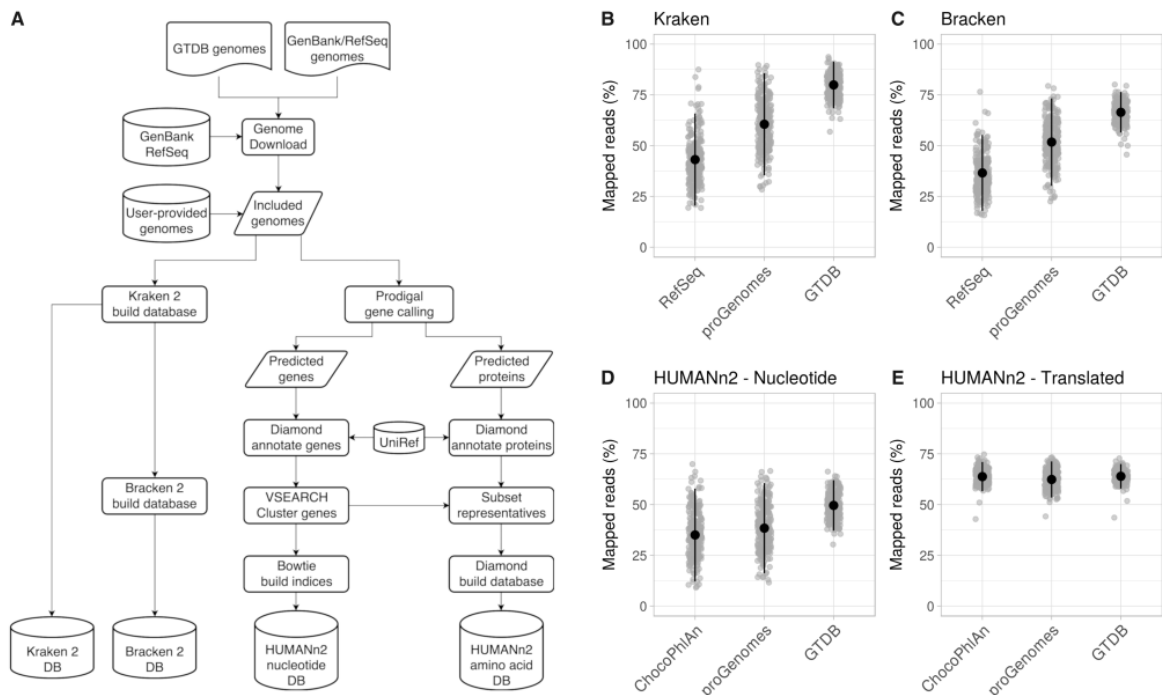
## 1 Introduction

Advances in metagenome sequencing and culturing methods have led to the recovery of thousands of genomes from as-of-yet uncultured microorganisms (Pasolli *et al.*, 2019). In turn, these genomes can be used to obtain comprehensive taxonomic and functional data from microbiomes by metagenome profiling, where reads from shotgun metagenome sequencing are mapped to databases of genes or genomes. A successful exploration of the diversity present in a microbial community depends on the selected database, the content of which will heavily influence the outcome of the profiling (Nasko *et al.*, 2018). Databases used as defaults in metagenome profilers are often not updated at a pace that reflects the increase in microbial genomics data or might not suit the particular needs of researchers, who may wish to expand the existing databases with their genomes of interest. Whereas up-to-date databases customized to match the question at hand would improve metagenome analysis, their creation is cumbersome due to the complexity and high computational requirements of retrieving appropriate genomes, and of configuring and executing the software. Thus, many metagenomic analyses fail to include the most up-to-date microbial data, leading to oversights. We address this problem with the development of Struo (from the Latin: ‘I build’ or ‘I gather’), an automated and modular pipeline that assists in the retrieval of genomes and in the construction of databases for Kraken2 (Wood and Salzberg, 2014), Bracken2 (Lu *et al.*, 2017) and HUMANn2 (Franzosa *et al.*, 2018).

## 2 Results

Struo uses the workflow engine Snakemake (Köster and Rahmann, 2012) and the Conda package manager to install the required software and build databases in a reproducible manner on Unix-based high-performance compute clusters (Fig. 1A). By default, the pipeline uses the genome taxonomy database (GTDB; Parks *et al.*, 2018) to retrieve taxonomic classifications and assembly statistics of 127 318 publicly available genomes (v.03-RS86), which are then filtered by completeness and contamination to retain medium and high quality representative genomes. This results in 21 276 non-redundant genomes that broadly encompass known microbial diversity. Users can also provide their own genome files in fasta format or a list of NCBI assembly IDs as input, making Struo compatible with other collections of curated genomes. Kraken2 and Bracken2 are the taxonomic profilers currently supported; they require genome sequences in fasta format with their corresponding NCBI taxID. HUMANn2 is the functional profiler implemented, which requires gene and protein sequence calling, annotation and clustering to create nucleotide and amino acid sequence databases (see [Supplementary Methods](#)).

We compared the mappability of reads from synthetic and real metagenomes between the default databases included with the supported profilers (RefSeq Bacteria and Archaea for Kraken2/Bracken2 and ChocoPhlAn for HUMANn2) and custom databases created with proGenomes and GTDB. We used five simulated communities (CAMI High Complexity test dataset) rarefied at 5 million



**Fig. 1.** (A) Struo's workflow encompasses the steps from genome download to database construction. The use of custom databases created using the proGenomes or GTDB collections of genomes increased the mappability of reads from 250 human gut metagenomes compared with the default databases of Kraken2 (B), Bracken2 (C) and HUMANn2 after nucleotide search (D) but not after translated search (E). Black points indicate mean proportion of mapped reads; black bars indicate standard deviation

reads per sample (Szczyrba *et al.*, 2017), and 250 human gut metagenomes rarefied at 2 million reads per sample (Xie *et al.*, 2016). Kraken2 was executed using default parameters; for Bracken, we filtered reads assigned to taxa with <100 hits. In both the real and simulated datasets, the use of custom databases resulted in more mapped reads compared with the default databases. The proportion of reads mapped by Kraken2 was highest when using the GTDB-derived database (Fig. 1B and Supplementary Fig. S1A), with similar results observed on the number of reads kept by Bracken2 (Fig. 1C and Supplementary Fig. S1B). We did not observe differences in mappability of reads in the functional profile obtained with HUMANn2 between ChocoPhlAn or the custom databases after nucleotide or translated searches in the synthetic dataset (Supplementary Fig. S1C and D), which can be explained by the presence of the genomes that comprise the simulated metagenomes in public repositories; however, we observed an increase in the percentage of mapped reads after nucleotide search in the human gut metagenomes (Fig. 1D). This is particularly important for HUMANn2, which first maps reads against the nucleotide database, allowing to quantify the contribution of different microbial species to the abundance of the detected gene families.

### 3 Conclusion

A careful yet broad selection of genomes to be included in databases for metagenome profiling can shed light on the so-called microbial dark matter, allowing to study the contribution of otherwise inaccessible microbial species (Pasolli *et al.*, 2019). We expect Struo, and the databases we provide, to enable a greater community of researchers to engage in a more comprehensive analysis of microbial communities, an imperative step in the study of the hidden microbial diversity (Thomas and Segata, 2019).

### Acknowledgements

We thank Albane Ruaud, Jessica Sutter, Jillian Waters and Shrikant Mantri for providing feedback and testing the built databases.

### Funding

This work was supported by the Max Planck Society.

*Conflict of Interest:* none declared.

### References

- Franszosa, E.A. *et al.* (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods*, **15**, 962–968.
- Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
- Lu, J. *et al.* (2017) Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.*, **3**, e104.
- Nasko, D.J. *et al.* (2018) RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.*, **19**, 165.
- Parks, D.H. *et al.* (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.
- Pasolli, E. *et al.* (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, **176**, 649–662.e20.
- Szczyrba, A. *et al.* (2017) Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063.
- Thomas, A.M. and Segata, N. (2019) Multiple levels of the unknown in microbiome research. *BMC Biol.*, **17**, 48.
- Wood, D.E. and Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
- Xie, H. *et al.* (2016) Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.*, **3**, 572–584.e3.





Appendix IV: Obesity is the main driver of functional alterations of the gut microbiome in cardiometabolic disease

## **Title**

Obesity is the main driver of functional alterations of the gut microbiome in cardiometabolic disease

## **Authors**

Jacobo de la Cuesta-Zuluaga<sup>1</sup>, Nicholas D. Youngblut<sup>1</sup>, Juan S. Escobar<sup>2</sup>, Ruth E. Ley<sup>1#</sup>

<sup>1</sup>Department of Microbiome Science, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany.

<sup>2</sup>Vidarium–Nutrition, Health and Wellness Research Center, Grupo Empresarial Nutresa, 050023 Medellin, Colombia.

# correspondence: rley@tuebingen.mpg.de.

## **Abstract**

The discovery of distinct links between obesity (OB) and the cardiometabolic health status (CHS) with the gut microbiome is hindered by the overlap between these conditions. Moreover, differences in study design and covariates used encumber the comparison of study outcomes. Here, we describe features of gut microbiome function associated independently with OB or CHS in a cohort of adults; and test for the replication of associations previously reported for microbiome and OB/CHS. We enrolled 459 deeply-phenotyped Colombians from whom we obtained 408 gut metagenomes. We measured three OB indices and classified individuals according to their CHS using blood biochemistry and anthropometric data. We evaluated the association of 136 KEGG modules and 2 653 orthologs previously linked with OB, cardiovascular disease or diabetes. Medication use, city, sex and age were included as covariates. We found that metagenome sequence diversity negatively correlated with OB; subjects with CHS had lower diversity than healthy subjects with similar OB levels. OB explained a higher

proportion of variance for sequence diversity and functional beta-diversity. Similarly, more modules and orthologs were uniquely associated with OB than with CHS or shared by both conditions. The microbiome potential of diseased individuals in both conditions showed a decreased fermentative ability and an increased response to oxygen. Disease-linked features were mainly contributed by members of *Proteobacteria*. Our results suggest that OB drives the microbiome associations with CHS when both are present.

## Introduction

The human gut microbiome, the microbial community colonizing the gastrointestinal tract, is a major participant in multiple metabolic, nutritional and immune processes of the host (1). As such, it is involved in the etiology of obesity (2) (OB), the abnormal or excessive fat accumulation that presents a risk to health (3). The microbiome also plays a role in the maintenance of the cardiometabolic health status (CHS), the presence of factors that increase the risk of heart disease, stroke and type 2 diabetes (T2D), namely, increased blood pressure, elevated glucose levels or insulin resistance, excess of fat around the waist, and abnormal concentrations of triglycerides or cholesterol(4,5). The incidence of these non-communicable diseases has been steadily increasing worldwide: as of 2016, 1.9 billion adults were overweight and 650 million were obese (6). Obesity leads to a loss of disease-free years owing to its associated conditions (7). The gut microbiome has therefore become the focus of study for the discovery of biomarkers, and the target of therapies or dietary interventions (8). However, OB and CHS are often confounded, obese individuals often suffer of other cardiometabolic affections (e.g., T2D, cardiovascular disease, liver disease), therefore decoupling CHS- and OB-specific associations with the gut microbiome remains a challenge (9).

At the taxonomic level, the composition of the microbiome shows high heterogeneity, with dominant species varying between individuals according to their geographical origin, dietary patterns and disease status, among other factors (10). This heterogeneity often results in contradictory reports of correlation of microbial taxa with host phenotypes, which hinders the translation

of gut microbiome findings across populations (11). Conversely, the potential metabolic capacity of the gut microbiome is highly conserved across individuals (12). The metabolic potential encoded by the members of the microbial community is largely redundant, which results in comparable ecosystem functioning (12). Despite this redundancy, modest changes in microbial metabolism can affect the functioning of the microbial community and the interaction with its host. Indeed, shifts in a small number of key metabolites with major biological relevance may be sufficient to induce alterations in gut homeostasis or microbiome functionality (13).

Focusing on the metabolic and signaling functions encoded by the microbiome might help overcome the contradictory associations commonly reported at the taxonomic level (11). In the case of OB and CHS, the identification of gut microbiome functions shared between them would help uncover disease-specific functional disturbances (9). To achieve this, well-characterized cohorts with relevant host phenotypic data are required. Moreover, to identify microbial features uniquely associated with each condition or shared by both, it is possible to incorporate the results of previous studies into the analysis of novel cohorts. This way, the set of microbial functions evaluated is restricted to those that have been reported to be linked to OB or CHS in different populations (14).

In a previous study of the gut microbiota of a community-dwelling cohort of Colombian men and women by means of 16S rRNA gene sequencing, we found that multiple microbial configurations were associated with OB and related conditions. These configurations were defined as taxonomic profiles characterized by a high abundance of different consortia of co-abundant microorganisms (15). In the same cohort, we explored a classification method that allowed us to differentiate subjects by their cardiometabolic health and obesity status (5). We observed that microbial species richness negatively correlated with OB, and that cardiometabolically unhealthy subjects showed lower microbiome diversity than healthy subjects with similar OB levels (16). Studying the functional profile of the microbiome would allow the identification of

a set of genes or pathways that drive the association with a given condition, even if the microorganisms that encode said features differ.

Here, we studied the functional profile of the gut microbiome of a well-phenotyped cohort of 459 Colombian men and women from five cities by means of shotgun metagenome sequencing. To identify microbiome functions robustly associated with OB and CHS, we selected functional features linked to obesity, cardiovascular disease or T2D from published studies performed in diverse populations, and tested for their replication in this Colombian cohort. In our analyses, we included host variables known to confound host-microbe associations but that are often overlooked in cross-sectional studies. We also sought to disentangle microbial features uniquely associated with OB or CHS by classifying individuals according to both factors and determining which functions are associated with one condition while accounting for the other. This allowed us to discriminate between microbiome associations unique to OB or CHS, or shared by both.

## **Materials and Methods**

### ***Ethics approval***

This cross-sectional study was conducted in accordance with the principles of the Declaration of Helsinki 2013 and had minimal risk according to the Colombian Ministry of Health (Resolutions 8430 of 1993 and 2378 of 2008). All the participants were thoroughly informed about the study and procedures before signing consent forms. Participants were assured of anonymity and confidentiality. Written informed consent was obtained from all the participants before beginning the study. The Bioethics Committee of SIU—University of Antioquia (Medellin, Colombia) reviewed the protocol and the consent forms and approved the procedures described here (approbation act 14-24-588 dated 28 May 2014).

### ***Study population***

We enrolled 459 community-dwelling adults from Colombia, South America, which has been previously described (15). Briefly, between July and

November 2014, adult men and women aged 18 to 62 insured by the health insurance provider EPS Sura were enrolled from five Colombian cities as part of a cross-sectional study. Underweight participants ( $\text{BMI} < 18.5 \text{ kg m}^{-2}$ ), pregnant women and individuals who had consumed antibiotics or antiparasitics in the three months prior to enrollment were excluded. The consumption of other medications did not warrant exclusion from the study. We also excluded subjects diagnosed with any of the following diseases: Alzheimer's disease, Parkinson disease or any other neurodegenerative disease; current or recent cancer ( $< 1$  year); and gastrointestinal diseases (Crohn's disease, ulcerative colitis, short bowel syndrome, diverticulosis or celiac disease).

### ***Blood biochemistry, anthropometric evaluation, diet assessment and medication use***

The assessment of host parameters, including the measurement of clinical variables in blood serum, anthropometric characteristics, blood pressure, short-chain fatty acids (SCFAs), dietary parameters and medication use is described in detail elsewhere (15,17,18). Briefly, peripheral venous blood was used to measure total cholesterol, high density lipoprotein (HDL) cholesterol, low density lipoprotein (LDL) cholesterol, very low density lipoprotein (VLDL) cholesterol, triglycerides, fasting glucose, fasting insulin, glycated hemoglobin (HbA1C), adiponectin, lipopolysaccharide-binding protein (LBP) and high-sensitivity C-reactive protein (hsCRP). The insulin resistance index using the homeostasis model assessment (HOMA-IR) was calculated from fasting insulin and glucose. Trained evaluators measured weight, height, waist circumference (WC), four skin folds (biceps, triceps, subscapular and ileocrestal), and systolic and diastolic blood pressures (15). We calculated the body mass index (BMI) as weight (kg)/height squared ( $\text{m}^2$ ); participants were classified as lean ( $18.5 \leq \text{BMI} < 25.0 \text{ kg m}^{-2}$ ), overweight ( $25.0 \leq \text{BMI} < 30.0 \text{ kg m}^{-2}$ ) or obese ( $\text{BMI} \geq 30.0 \text{ kg m}^{-2}$ ). Body fat percentage (BF%) was calculated from the skin folds (19). To quantify calories and diet quality in the habitual diet of participants, we performed 24-hour dietary recall interviews (17). Pharmacological treatments were registered by the participants in specific

questionnaires (15). The SCFAs acids butyrate, propionate, acetate, and the branched-chain fatty acid isobutyrate were measured from feces using gas chromatography–mass spectrometry (18).

We classified subjects by their cardiometabolic health status (CHS) (16) as follows. Participants were considered cardiometabolically unhealthy when they presented 2 or more of the following conditions: systolic/diastolic blood pressure  $\geq 130/85$  mm Hg or consumption of antihypertensive medication; fasting triglycerides  $\geq 150$  mg dl<sup>-1</sup>; HDL  $> 40$  mg/dl (men),  $> 50$  mg dl<sup>-1</sup> (women) or consumption of lipid-lowering medication; fasting glucose  $\geq 100$  mg dl<sup>-1</sup> or consumption of antidiabetic medication; HOMA-IR  $> 3$ , and hsCRP  $> 3$  mg dl<sup>-1</sup>.

### ***DNA extraction and sequencing***

Fecal sample collection was performed by each participant, who kept the sample refrigerated in household freezers and brought it to a collection center in each city within 12 h. Upon arrival at the collection center, samples were stored on dry ice and shipped by courier to the Colombian Institute of Tropical Medicine (ICMT) in Medellin, Colombia, for DNA extraction (20). We extracted total DNA from human fecal samples of 430 out of the 459 enrolled subjects using the QIAamp DNA Stool Mini Kit (Qiagen, Hilden, Germany). DNA was quantified with a NanoDrop spectrophotometer (Nyxor Biotech, Paris, France) and stored at -80°C.

We prepared shotgun metagenome libraries with a modified Nextera protocol, as described elsewhere (21). Briefly, we used 1 ng of total fecal DNA for Nextera Tn5 tagmentation. After purification with Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA), we normalized and pooled the samples. Next, we performed size selection of the pooled samples using BluePippin (Sage Sciences, Beverly, MA, USA) to restrict fragment sizes to 400 to 700 bp. Barcoded pools were sequenced using the Illumina HiSeq 3000 platform with 2x150 bp paired-end sequencing. Library preparation and sequencing was performed at the Max Planck Institute for Biology — Tübingen, Tübingen, Germany.

### ***Sequence quality control***

We validated raw sequence reads with fqtools v.2.0 (22) and de-duplicated with the “clumpify” command of bbtools v37.78 (<https://jgi.doe.gov/data-and-tools/bbtools/>). Adapters were trimmed and read quality control was performed using the “bbduk” command of bbtools and skewer v0.2.2 (23). We removed human genome reads *in silico* by mapping them to the hg19 assembly with the “bbmap” command of bbtools. Quality reports for each sample were created with fastqc v0.11.7 (<https://github.com/s-andrews/FastQC>) and multiQC v1.5a (24). Metagenome coverage for each sample was estimated using Nonpareil v.3.3.4 (25). Samples with a sequencing depth < 500 000 reads or a metagenome coverage < 60 % were discarded from downstream analyses.

### ***Metagenome profiling***

Filtered reads were used to obtain the functional profile using HUMANN2 v.2.8.1. (26) prior subsampling to a maximum of 10 million reads per sample with seqtk v.1.3. We mapped reads against custom databases of archaeal and bacterial genes and genomes generated using Struo v.0.1.6 (27) based on the release 89 of the Genome Taxonomy Database (28) (available at <http://ftp.tue.mpg.de/ebio/projects/struo/>).

### ***Selection of functional features for analysis***

We focused our analyses on a set of protein orthologs or metabolic modules previously reported to be associated with obesity, T2D and cardiovascular disease. The selection of features to include was systematized by retrieving the results of studies reporting analyses of novel populations or meta-analyses which used similar databases to group functional features, and that were easily accessible in the original publication. For coarse-level analyses, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) modules, as reported by Jie et al. (29) and Wu et al. (30). For fine-level analyses, we used KEGG orthogroups as reported by Jie et al. (29) and Armour et al. (14). Selection was performed in three steps (figure S1A): first, for each study we



selected features that were consistently associated with diseased subjects or healthy controls (e.g., by removing features enriched in both lean and diabetic subjects within the same study). Next, we merged the resulting lists of features and performed a similar selection of features with consistent associations across studies. Finally, the shortlisted functional features that were detected in the Colombian subjects were further filtered to include only those present in at least 50% of individuals. This resulted in 2 653 KEGG orthologs and 136 KEGG modules, which were used for further analyses (tables S1 and S2).

### ***Statistical analyses***

Statistical analyses were performed using R v.4.0.2 (31) unless stated otherwise. We adjusted P values for multiple comparisons using the Benjamini-Hochberg method, with a significance threshold of 0.1.

### ***Transformation of functional feature abundances***

Functional profile data was transformed using the centered log-ratio (clr). For this, we first replaced zero values with pseudocounts in a compositionally aware manner using the zCompositions v.1.3.4 package of R (32), and used the propr v.4.2.6 (33) and compositions v.2.0 (34) packages to compute the clr transformation. Positive clr values imply that the feature is more abundant than the average feature, conversely, negative values imply that the feature is less abundant than the average (35).

### ***Validation of covariates***

We evaluated the individual association of age, sex, city, socioeconomic status (according to the official Colombian strata division, from 1 [lowest income] to 6 [highest income]), and the consumption of medications for diabetes, hypertension, dyslipidemia and proton pump inhibitors (PPIs) with the overall composition of the microbiome. For this, we used k-mer-based richness calculated using Nonpareil for alpha-diversity, and Aitchison's distance (Euclidean distance using clr-transformed abundances) calculated using KEGG module abundances as measure of beta-diversity. For each medication

category, we performed a 1:3 nearest neighbor propensity score matching without replacement using the MatchIt v.4.1.0 package of R (36), matching consumers with non-consumers by age, sex, city and BF%.

We tested differences in alpha-diversity by sex and medication consumption using Welch's t-test, ANOVA for city and socioeconomic status, and Spearman's rank correlation for age. Next, we assessed multivariate homogeneity of groups dispersions, and performed a permutational multivariate analysis of variance (PERMANOVA) on Aitchison distance matrices using the vegan v.2.5-7 package of R (37). Host factors that were significantly associated with at least one of the two diversity metrics were selected for inclusion as covariates in downstream analyses.

#### *Functional diversity analyses*

For each sample, we estimated k-mer based sequence richness using Nonpareil (25). We then performed linear regression analyses to test the association of BMI, BF% or WC (as continuous variables), and cardiometabolic status of the subjects with sequence richness after adjusting for age, city, sex and medication usage. Goodness-of-fit of each model was assessed using Akaike's information criterion (AIC), and the proportion of variance was estimated by means of the adjusted coefficient of determination ( $R^2$ ). Next, we assessed differences in beta-diversity estimates of functional features by BMI, BF% or WC (as categorical variables), and cardiometabolic status. For this, we performed a permutational multivariate analysis of variance (PERMANOVA) on a matrix of Aitchison's distances of KEGG modules, as implemented in the vegan v.2.5-7 package of R (37). The proportion of variance was estimated using the adjusted coefficient of determination ( $R^2$ ). We calculated the unique and shared contributions of obesity measures and cardiometabolic status to the functional beta-diversity by variance partitioning.

*Identification of a set of shared functions between obesity and cardiometabolic health status, and association of functional features with host parameters*

We identified individual functional features (i.e. KEGG modules and orthologs) uniquely associated with obesity or cardiometabolic status using the MaAsLin2 v.1.0.0 package (38) on R v.3.6.2. For this, we fitted Gaussian linear models and assessed each feature's clr-transformed abundance including both a measure of obesity (e.g. BMI, WC or BF%) and the cardiometabolic health status as main effects. We used type II ANOVAs to test for significance. A functional feature was considered to be uniquely associated with a given condition if it was significant after regressing out the other condition. Oppositely, we considered a feature as part of the set of shared functions between OB and CHS if it was significantly associated with a single condition (e.g. in a model only including OB) but was non-significant after the other was included in the model (e.g. in a model including both OB and CHS). All models included age, sex, city and the consumption of the aforementioned medications as covariates; other host factors evaluated above were not included. We also performed a sensitivity analysis by removing all subjects who consumed the medications in question and repeating the analysis; results were largely consistent with the main analysis (data not shown).

Next, the association of blocks of related microbial features and host biochemical and anthropometric parameters was tested using HALLA v.0.8.19 (hierarchical all-against-all association testing; <https://github.com/biobakery/halla>). We used matrices of residuals of the microbial features and host parameters after adjusting each for age, sex, city and medication consumption. Spearman's rank correlation was used as the similarity measure and hierarchical clustering was performed with Ward's method.

*Quantification of contributory diversity and abundance of functional features*

We used HUMAnN2's tiered search (26), combined with the Struo-generated custom databases (27), to quantify the contribution of different members of the microbial community to the total abundance of the detected

modules and orthologs. In addition, we used these data to quantify the contributonal diversity of the functional features, measured as the number of taxa that contributed to its total abundance, that is, the contributonal richness (26). We restricted our analyses to the set that were uniquely associated to any of the obesity measures, the CHS, or that were shared between conditions. For each set, we compared the contributonal richness between modules associated with health and disease.

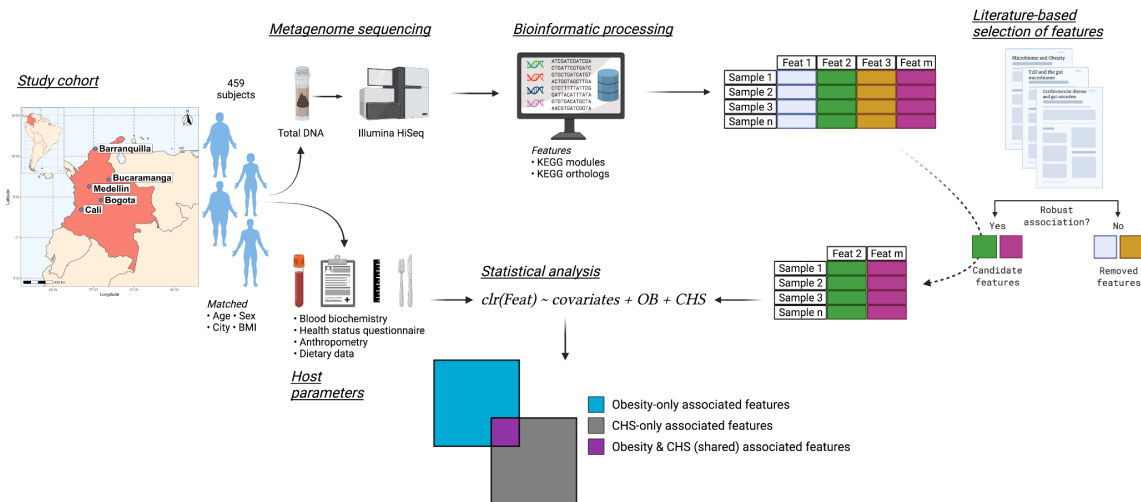
### ***Data and code availability***

The metagenomic sequence data will be deposited in the European Nucleotide Archive upon submission of this manuscript to a preprint server and/or a peer-reviewed journal. The R notebooks and associated data will be made available at [https://github.com/leylabmpi/Colombian\\_Cardiometabolic](https://github.com/leylabmpi/Colombian_Cardiometabolic).

## **Results**

### ***Overview of a deeply-characterized human cohort***

An overview of the cohort, host and microbial data, and the analyses performed is provided in figure 1. We carried out a cross-sectional study, in which we recruited 459 community-dwelling adults living in five large Colombian cities (15), attempted microbiome sequencing in 430 samples and succeeded in the shotgun characterization in 408 of them. We enrolled subjects in similar proportions by body mass index category (BMI: lean, overweight, obese), city of origin (Bogotá, Medellín, Cali, Barranquilla and Bucaramanga), sex at birth (man, woman) and age group (18 to 40 and 41 to 62 years).



**Figure 1. Overview of the Colombian cohort, study design, metagenome profiling, functional feature selection and data analysis.** Figure made with BioRender.

We assessed diverse demographic, health-related and dietary parameters from these subjects, and used DNA from fecal samples for gut metagenome shotgun analyses. Summary statistics of subjects from the studied cohort are presented in table 1. After metagenomic library construction, sequencing, and bioinformatic curation of sequencing reads, we retained 408 samples which had a sequencing depth  $> 5.0 \times 10^5$  reads (mean  $\pm$  SD: 6 719 985 reads sample<sup>-1</sup>  $\pm$  8 960 996) or a metagenome coverage calculated using Nonpareil  $> 60\%$  (82.36%  $\pm$  8.38).

	Cardiometabolic health status				Body mass index category			
	Overall	Healthy	Unhealthy	P val.	Lean	Overweight	Obese	P val.
n	408	153	255		125	161	122	
Age (years)	40.74 (11.11)	38.35 (10.60)	42.18 (11.18)	0.001	39.25 (11.24)	40.39 (11.18)	42.75 (10.68)	0.04
Sex—Woman (%)	205 (50.2)	84 (54.9)	121 (47.5)	0.175	64 (51.2)	76 (47.2)	65 (53.3)	0.58
City (%)								
Barranquilla	80 (19.6)	27 (17.6)	53 (20.8)		20 (16.0)	31 (19.3)	29 (23.8)	
Bogotá	77 (18.9)	31 (20.3)	46 (18.0)	0.831	22 (17.6)	34 (21.1)	21 (17.2)	0.677
Bucaramanga	72 (17.6)	30 (19.6)	42 (16.5)		27 (21.6)	29 (18.0)	16 (13.1)	
Cali	88 (21.6)	33 (21.6)	55 (21.6)		27 (21.6)	32 (19.9)	29 (23.8)	
Medellín	91 (22.3)	32 (20.9)	59 (23.1)		29 (23.2)	35 (21.7)	27 (22.1)	
BMI (kg/m <sup>2</sup> )	27.94 (5.00)	25.33 (3.88)	29.51 (4.95)	<0.001	22.65 (1.60)	27.45 (1.37)	34.00 (3.59)	<0.001
Body fat (%)	37.18 (5.44)	34.76 (5.70)	38.64 (4.73)	<0.001	33.34 (5.01)	36.95 (4.46)	41.43 (3.75)	<0.001
Waist circumference (cm)	92.87 (13.15)	85.06 (10.28)	97.56 (12.46)	<0.001	80.54 (7.51)	91.98 (6.93)	106.69 (10.48)	<0.001
Systolic BP (mm Hg)	124.64 (18.52)	116.92 (16.08)	129.28 (18.37)	<0.001	116.93 (17.64)	126.20 (17.72)	130.52 (17.88)	<0.001
Diastolic BP (mm Hg)	80.28 (12.25)	75.27 (11.36)	83.28 (11.80)	<0.001	74.82 (11.23)	81.23 (12.05)	84.61 (11.53)	<0.001
Medication usage (%)								
Hypertension	71 (17.4)	8 (5.2)	63 (24.7)	<0.001	12 (9.6)	25 (15.5)	34 (27.9)	0.001
Diabetes	19 (4.7)	0 (0.0)	19 (7.5)	0.001	3 (2.4)	7 (4.3)	9 (7.4)	0.174
Dyslipidemia	41 (10.0)	2 (1.3)	39 (15.3)	<0.001	8 (6.4)	17 (10.6)	16 (13.1)	0.206
PPI	19 (4.7)	5 (3.3)	14 (5.5)	0.43	8 (6.4)	4 (2.5)	7 (5.7)	0.236
Total cholesterol (mg/dL)	186.63 (34.80)	182.14 (33.25)	189.33 (35.48)	0.043	183.82 (38.90)	188.09 (33.86)	187.59 (31.54)	0.553

HDL (mg/dL)	45.67 (13.71)	53.18 (14.59)	41.17 (10.92)	<0.001	50.78 (15.13)	43.68 (12.53)	43.07 (12.28)	<0.001
LDL (mg/dL)	115.44 (30.66)	112.56 (30.16)	117.18 (30.88)	0.141	111.98 (32.13)	119.30 (31.18)	113.86 (27.94)	0.107
Triglycerides (mg/dL)	147.16 (102.46)	94.99 (39.69)	178.47 (115.14)	<0.001	122.26 (111.64)	147.37 (83.70)	172.42 (109.51)	0.001
ApoB (mg/dL)	96.14 (39.98)	87.05 (27.06)	101.60 (45.20)	<0.001	92.44 (30.48)	99.55 (54.14)	95.45 (22.74)	0.322
LpA (mg/dL)	2.08 (4.00)	2.57 (4.30)	1.79 (3.79)	0.056	2.33 (4.29)	2.40 (4.30)	1.40 (3.14)	0.08
Adiponectin (µg/ml)	6.65 (3.90)	8.09 (4.36)	5.79 (3.31)	<0.001	7.83 (4.28)	6.18 (3.51)	6.06 (3.74)	<0.001
Glucose (mmol/L)	89.78 (22.19)	83.13 (6.24)	93.77 (26.89)	<0.001	85.86 (20.75)	87.63 (12.50)	96.65 (30.64)	<0.001
HbA1c (%)	5.56 (0.64)	5.38 (0.29)	5.67 (0.76)	<0.001	5.43 (0.48)	5.50 (0.61)	5.78 (0.77)	<0.001
Insulin (µU/ml)	13.52 (9.07)	8.47 (3.12)	16.55 (10.07)	<0.001	8.52 (3.99)	12.65 (5.96)	19.79 (12.13)	<0.001
HOMA-IR	3.15 (3.19)	1.74 (0.65)	4.00 (3.75)	<0.001	1.83 (1.09)	2.75 (1.41)	5.02 (4.97)	<0.001
Leptin (ng/mL)	7.00 (6.28)	5.43 (4.91)	7.94 (6.81)	<0.001	4.20 (4.09)	5.98 (5.31)	11.22 (7.13)	<0.001
LBP (µg/ml)	4.48 (1.58)	4.02 (1.49)	4.77 (1.57)	<0.001	4.11 (1.61)	4.45 (1.58)	4.92 (1.46)	<0.001
hs-CRP (mg/L)	3.17 (4.61)	1.74 (1.80)	4.03 (5.49)	<0.001	1.69 (1.50)	3.23 (5.10)	4.62 (5.52)	<0.001
Acetate (µmol/g)	3.81 (4.85)	3.05 (3.55)	4.27 (5.44)	0.014	3.00 (3.23)	3.20 (5.10)	5.45 (5.49)	<0.001
Propionate (µmol/g)	1.18 (1.97)	0.90 (1.18)	1.36 (2.30)	0.021	0.86 (1.16)	1.17 (2.61)	1.54 (1.56)	0.023
Butyrate (µmol/g)	0.61 (0.94)	0.47 (0.59)	0.69 (1.09)	0.024	0.49 (0.69)	0.61 (1.23)	0.72 (0.68)	0.157
Isobutyrate (µmol/g)	0.04 (0.15)	0.03 (0.04)	0.04 (0.18)	0.25	0.03 (0.03)	0.05 (0.23)	0.04 (0.05)	0.454

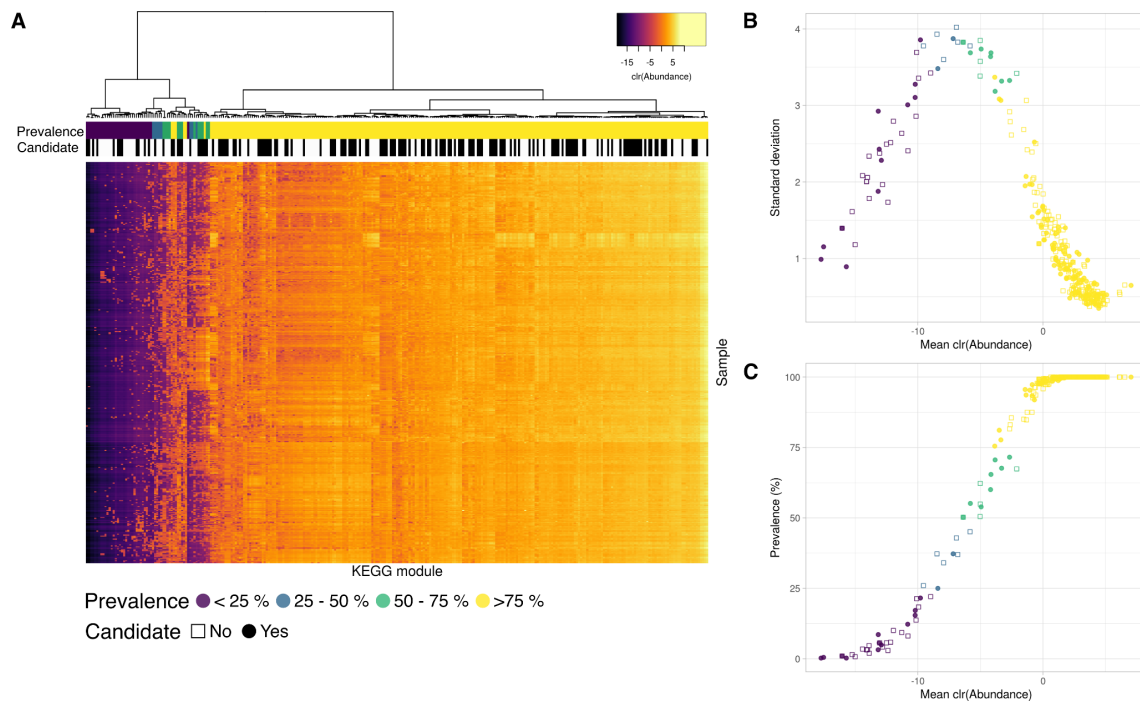
*Table legend in next page*

**Table 1. General, anthropometric, health-related and dietary characteristics of subjects with microbiome data included in the present study (n = 408).** Data presented as the mean (SD) or count (%). BMI: body mass index, HDL: high density lipoprotein cholesterol, LDL: low density lipoprotein cholesterol, hsCRP: high-sensitivity C-reactive protein, HOMA-IR: homeostatic model assessment–insulin resistance. P values from ANOVA or Chi-squared test.

***The functional redundancy of the microbiome is evidenced by the low variability of the features detected***

To assess the metabolic potential of the gut microbiome of the subjects, we used centered log-ratio (clr) transformed abundances of 3 303 detected KEGG orthologs and 301 KEGG modules; this transformation allowed us to account for the compositional nature of the data. Hereafter, we will refer to KEGG modules and orthologs as functional features. For visualization, we grouped the KEGG modules according to their clr-transformed abundance using hierarchical clustering, and observed that they formed three groups, which roughly correspond to high, medium and low mean abundances (figure 2A). Likewise, the clustering of modules closely followed their prevalence and abundance, as we found a strong positive correlation between these two variables (Spearman's  $\rho = 0.97$ ,  $P < 0.001$ ). Overall, we observed little variation on the abundance of the features across subjects. Features with mean abundances located on the extremes of the distribution, that is, low and high mean abundance, showed the smallest standard deviation. Conversely, features with intermediate prevalence had the highest variance (figure 2B and C). To test whether these patterns are specific to the Colombian cohort, we performed a similar analysis on a set of 147 publicly available metagenome samples from the Human Microbiome Project retrieved using the CuratedMetagenomeData Bioconductor package (39), and observed comparable patterns (not shown).





**Figure 2. Functional redundancy of the microbiome is evidenced by high prevalence and low variance of features across subjects of the Colombian cohort.** A) Heatmap of clr-transformed abundance of KEGG modules across subjects (n = 408). Dendrogram on top was obtained by hierarchical Ward-linkage clustering based on transformed abundances. Color bands represent prevalence quartiles of modules and whether they were selected for further analyses based on literature search. B) Scatter plot of mean transformed abundance and standard deviation of all modules in the Colombian cohort. Point color represents prevalence quartiles and shape indicates inclusion for further analyses. C) Scatter plot of mean transformed abundance and prevalence of all modules in the Colombian cohort. Shape and color same as in B).

### ***Potential confounding host factors associate with functional diversity of the microbiome***

Multiple host parameters are associated with the structure of the gut microbiome and cardiometabolic health, including age (40), sex (41), geographical origin (42), and the consumption of medication for conditions such as T2D, hypertension, dyslipidemia and PPIs (43–45). Accounting for such covariates can reduce the risk of identifying spurious correlations (40,44,46), however, the inclusion of all covariates might result in overfitting of statistical models. Therefore, before testing the association of CHS or OB with individual functional features, we selected covariates to include in the linear models. For this, we validated the association of each of the aforementioned host parameters, in addition to the total number of medications and the

socioeconomic status of the subjects with the overall composition of the functional profile as measured by alpha- and beta-diversity metrics.

We tested the association of T2D, hypertension, dyslipidemia and PPI medication by matching subjects consuming and not consuming by age, sex, city and BF% in a 1:3 ratio for each medication separately (figure S2). We tested differences in alpha-diversity using a Welch's t-test and observed a significantly lower metagenome richness in consumers of drugs for T2D ( $n_{\text{consumers}} = 19$ ,  $n_{\text{non-consumers}} = 57$ ,  $t_{25.95} = 1.522$ ,  $P = 0.070$ ), hypertension ( $n_{\text{consumers}} = 71$ ,  $n_{\text{non-consumers}} = 213$ ,  $t_{110.74} = 2.186$ ,  $P = 0.016$ ) and dyslipidemia ( $n_{\text{consumers}} = 41$ ,  $n_{\text{non-consumers}} = 123$ ,  $t_{58.11} = 2.017$ ,  $P = 0.024$ ). We did not observe differences between PPI consumers and non-consumers ( $n_{\text{consumers}} = 19$ ,  $n_{\text{non-consumers}} = 57$ ,  $t_{25.20} = 0.031$ ,  $P = 0.489$ ). We did not observe differences in beta-diversity between consumers and non-consumers of diabetes ( $F_{1,406} = 1.066$ ,  $R^2 = 0.015$ ,  $P = 0.341$ ), hypertension ( $F_{1,406} = 0.858$ ,  $R^2 = 0.002$ ,  $P = 0.967$ ), dyslipidemia ( $F_{1,406} = 0.866$ ,  $R^2 = 0.006$ ,  $P = 0.423$ ) and PPI ( $F_{1,406} = 1.021$ ,  $R^2 = 0.013$ ,  $P = 0.469$ ) medication using PERMANOVA on Aitchison's distances.

We observed significant differences in beta-diversity by sex using PERMANOVA ( $F_{1,406} = 1.908$ ,  $R^2 = 0.005$ ,  $P = 0.012$ ), but not in alpha-diversity using Welch's t-test ( $n_{\text{women}} = 205$ ,  $n_{\text{men}} = 203$ ,  $t_{402.99} = -0.56$ ,  $P = 0.65$ ). We did not observe a significant correlation between age and alpha-diversity using Spearman's correlation coefficient ( $\rho = 0.002$ ,  $P = 0.964$ ), nor in beta-diversity by age group ( $n_{18-40} = 195$ ,  $n_{41-60} = 213$ ,  $F_{1,406} = 0.791$ ,  $R^2 = 0.002$ ,  $P = 0.764$ ). We observed significant differences in alpha-diversity by city using ANOVA ( $F_{4,403} = 8.949$ ,  $P < 0.001$ ) and beta-diversity using PERMANOVA ( $F_{4,403} = 6.154$ ,  $R^2 = 0.058$ ,  $P = 0.001$ ). We did not observe differences in alpha- or beta-diversity by the socioeconomic status of the subjects (ANOVA  $F_{5,402} = 0.455$ ,  $P = 0.809$ , PERMANOVA  $F_{5,402} = 0.976$ ,  $R^2 = 0.012$ ,  $P = 0.556$ ) (figure S2). We did not observe differences in beta-dispersion in any of the factors tested ( $P > 0.05$ ), that is, groups within each of the variables showed homogeneous variance.

Based on their association with functional diversity metrics, we kept the city of origin and consumption of medications for hypertension, T2D and

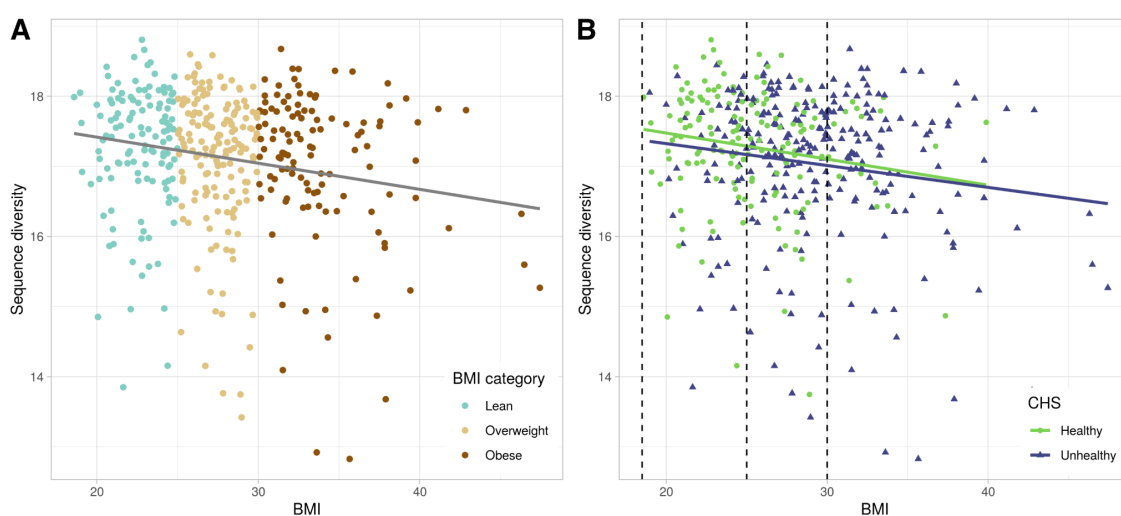
dyslipidemia as covariates to include in downstream statistical models. While non-significant, we also included PPI consumption and sex; we consider that the risk of not including these variables, given what is known of their influence over the microbiome in other populations and conditions outweighs the risk of overfitting the linear models (47).

***Obesity better explains differences in functional richness and beta-diversity than cardiometabolic health status***

We evaluated the association of OB and CHS with metagenome sequence richness using a k-mer-based diversity index, fitting linear models adjusting for age, sex, city of origin, and the consumption of medications. We used body mass index (BMI), waist circumference (WC) and percentage of body fat (BF%) as measures of obesity. As reported in other cohorts, we observed that in this population there was a negative association between k-mer-based sequence diversity and host health. Sequence diversity was negatively correlated with two obesity measures (BMI adj.  $P = 0.008$ , WC adj.  $P = 0.009$ , BF% adj.  $P = 0.540$ ). Similarly, metabolically healthy individuals had significantly higher richness than unhealthy individuals (CHS adj.  $P = 0.092$ ). Moreover, it was possible to differentiate subjects by their cardiometabolic status within BMI categories (figure 3 and figure S3). OB measures better explained differences in sequence richness than CHS. BMI had the lowest Akaike information criterion (AIC) value and explained the highest proportion of variance (AIC = 1162.796, Adj.  $R^2 = 0.021$ ), followed by WC (AIC = 1164.281, Adj.  $R^2 = 0.017$ ). Only BF% (AIC = 1172.110, Adj.  $R^2 = -0.002$ ) was a worse predictor than CHS (CHS AIC = 1169.153, Adj.  $R^2 = 0.006$ ).

Next, we assessed the unique and shared contributions of OB and CHS to the functional diversity of the microbiome by calculating the proportion of variance attributed to each. For this, we calculated Euclidean distances on clr-transformed abundances of KEGG modules and calculated the proportion of variance explained using the adjusted  $R^2$  from the PERMANOVA test. We found that BMI explained a higher total and unique proportion of variance for functional beta-diversity than CHS (PERMANOVA BMI: total adj.  $R^2 = 0.0045$ ,

unique adj.  $R^2 = 0.0025$ ,  $P = 0.0037$ ; CHS: total adj.  $R^2 = 0.0021$ , unique adj.  $R^2 = 0.00014$ ,  $P = 0.02$ ). The proportion of variance contributed by both BMI category and CHS, that is, the variance resulting from the correlation of both conditions, was higher than the unique proportion contributed by CHS but lower than that of BMI (shared adj.  $R^2 = 0.0020$ ). We did not observe significant results for BF% or WC ( $P > 0.05$  in all cases). Similar results were obtained when the analyses were performed using KEGG ortholog tables.



**Figure 3. Metagenome sequence richness is negatively correlated with obesity (A) and can be further differentiated by including cardiometabolic status of subjects (B).** Scatter plots of sequence richness and BMI, regression lines are shown. This pattern is consistent when WC or BF% are used (see figure S3)

### ***Literature-based selection of candidate features allows to test robust associations in a novel population***

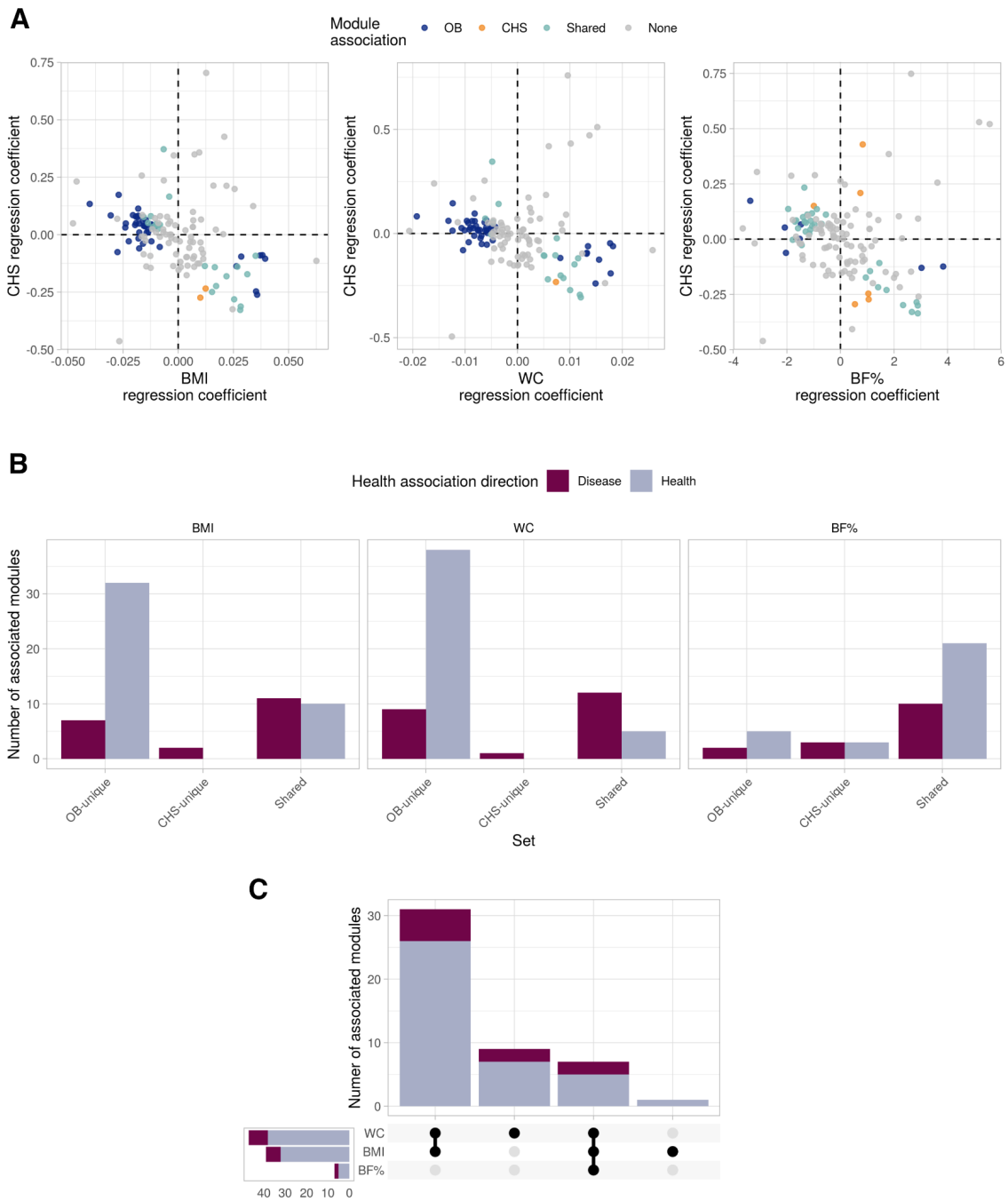
To narrow down the set of functional features that might be associated with OB and/or CHS, that is, protein orthologs and metabolic modules to test, we used data from publicly available studies looking into cardiometabolic diseases that used the same databases to annotate the microbiome features. We obtained KEGG orthologs from Jie et al. (2017) (29) and Wu et al. (2020) (30), and KEGG modules from Jie et al. (2017) (29) and Armour et al. (2019) (14). We filtered the retrieved data to select functional features that were consistently associated with disease or with healthy controls within and between studies (see methods and figure S1). We used these lists of candidate features

to filter the functional profile data generated for the Colombian cohort. After removal of functional features present in < 50% of our cohort, 2 653 KEGG orthologs and 136 KEGG modules were retained for downstream analyses (tables S1 and S2).

### ***Most differentially-abundant functions negatively correlate with obesity***

We sought to identify which of the selected candidate functional features were uniquely linked with OB or CHS, that is, were associated with one phenotype while accounting for the influence of the other. For this, we fitted linear models of the clr-transformed abundance using MaAsLin2 adjusting for the selected host variables. We considered a feature to be uniquely associated with a given main effect or condition (e.g. OB) if it had a FDR-adjusted  $P < 0.1$  in models with and without adjusting for the other condition (e.g. CHS). Conversely, a feature was considered as shared between OB and CHS if it was significant in a model with a single condition but non-significant after both conditions were considered.

Only a fraction of the candidate features included in our analyses were significantly associated with CHS or an OB measure (figure S1B): 67/136 (49.3%) KEGG modules and 1 417/2 653 (53.4%) orthologs were uniquely associated with one of the conditions or belonged to a set of shared features between CHS and OB (tables S3 and S4). Most features were uniquely associated with OB measures and showed a negative correlation with them (figure 4A), with the exception of BF%, where the set of shared features with CHS was the largest (figure 4B). Regardless of the OB measure used in the models, the set of features uniquely associated with CHS was small (figure 4B).



**Figure 4. Obesity, rather than cardiometabolic health status, drives associations with functional features of the microbiome.** A) Scatterplots of linear regression coefficient of KEGG module abundance in models including CHS and a measure of OB (BMI: left panel, WC: center panel, BF%: right panel). Colors represent whether a module is uniquely associated with OB or CHS, shared or not significant. B) Barplots show the number of KEGG modules belonging to the unique or shared sets shown in A) for each OB measure. Colors represent the association with health (i.e., health-associated modules are enriched in lean or cardiometabolically healthy individuals). C) Upset plot indicates the overlap in KEGG modules associated with each OB measure.

Most significant features were positively associated with host health, independent of the set to which they belonged. In the case of OB-associated features, their abundance was higher in lean individuals compared to obese; for CHS-linked features, they were enriched in cardiometabolically healthy compared to unhealthy subjects (figure 4B). This reflects the actual metabolic functions in which they are involved. Disease-associated modules and orthologs evidence a metabolic configuration of the microbial community with increased capacity of metabolizing sulfur and nitrogen compounds through an increased abundance of thioredoxins, thiosulfate sulfurtransferases, thioesterases, nitrate reductases, nitrate/nitrite transporters and sensor proteins, and other nitrogen regulatory proteins. They are also indicative of a metal acquisition response, with features related to the synthesis and export of the siderophore enterobactin, in addition to zinc uptake regulators and manganese/iron transport systems. Other signatures of the gut of diseased subjects include an increased tolerance to reactive oxygen species by glutathione-regulated efflux systems and glutathione transferases; the transport of simple sugars via the PTS system; the metabolism of choline, glycine betaine and trimethylamine; the synthesis of lipopolysaccharide and polyamine production; and the degradation of epithelial cells and the utilization of resulting ethanolamine. Conversely, health-related features included orthologs and modules involved in microbial energy and fermentative pathways such as glycolysis, the pentose phosphate pathway, the bifidobacterium shunt and methanogenesis; acetogenesis and the production of propionate and succinate; degradation of mucin; the biosynthesis of vitamins, and degradation of amino acids (tables S3 and S4).

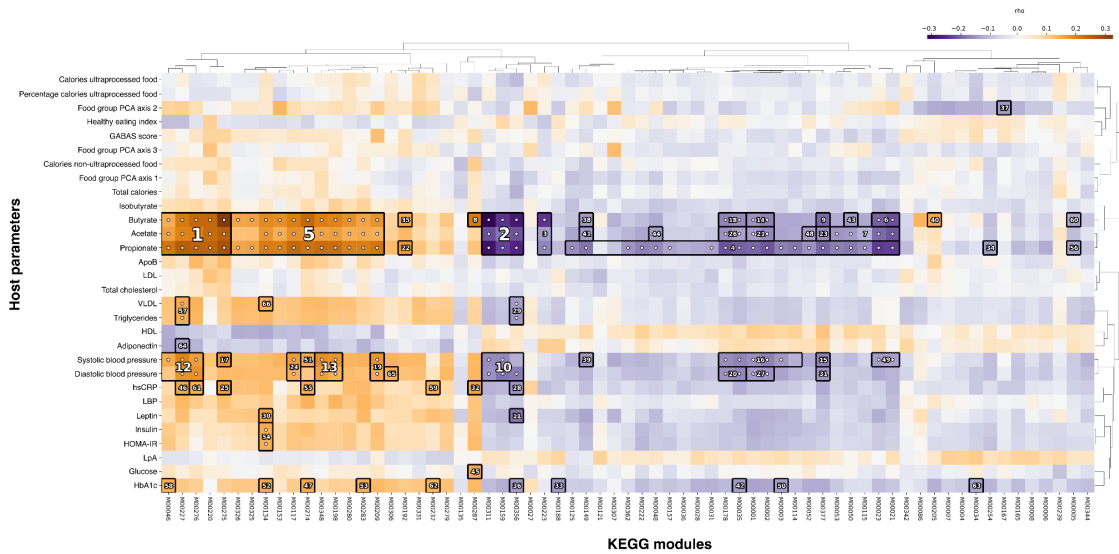
We observed a high overlap between functions uniquely associated with any of the OB measures (BMI, WC or BF%); figure 4C), as expected given the positive correlation between them (figure S4). Most features belonging to the unique sets correlated with at least two of the OB metrics and had a consistent link with host health (figure 4C).

***Functional features significantly associated with OB or CHS also correlate with blood pressure and SCFA excretion***

Next, we sought to identify groups of correlated host parameters and microbial functional features. For this, we used Hierarchical All-against-All Association testing (HAAA) on the set of functional features significantly associated with OB, CHS or both, and biochemical and dietary variables. Both data matrices were adjusted for age, sex, city and medication consumption, and P values were adjusted for multiple comparisons. We found 1 573 correlation blocks when using KEGG orthologs and 66 blocks when using KEGG modules. We observed a large overlap between the functions that formed these blocks, that is, multiple functions were linked to several host parameters. The association of these features was consistently related to improved or deteriorated host health.

The set of host variables forming the largest correlation blocks were the fecal concentration of SCFAs: butyrate, acetate or propionate. Twenty-five module blocks (comprising 49 modules) and 244 ortholog blocks (comprising 1073 orthologs) significantly correlated with at least one SCFAs (figure 5). Positively correlated blocks encompassed modules predominantly involved in simple sugar and osmoprotectant transport systems, including components of the PTS system, and glutamine, glutathione, thiamine, sn-glycerol-phosphate, rhamnose, RTX toxin, N-acetylglucosamine and alpha-hemolysin/cyclolysin transport systems (tables S5 and S6). Conversely, modules comprising negatively correlated blocks were involved in methanogenesis and the Wood–Ljungdahl pathway; the citrate cycle, glycolysis and gluconeogenesis; and the biosynthesis of ascorbate, NAD and several nucleotides and amino acids (tables S5 and S6). It is worth noting that in our cohort, higher fecal SCFA levels have been linked to obesity and altered cardiometabolic status (18).





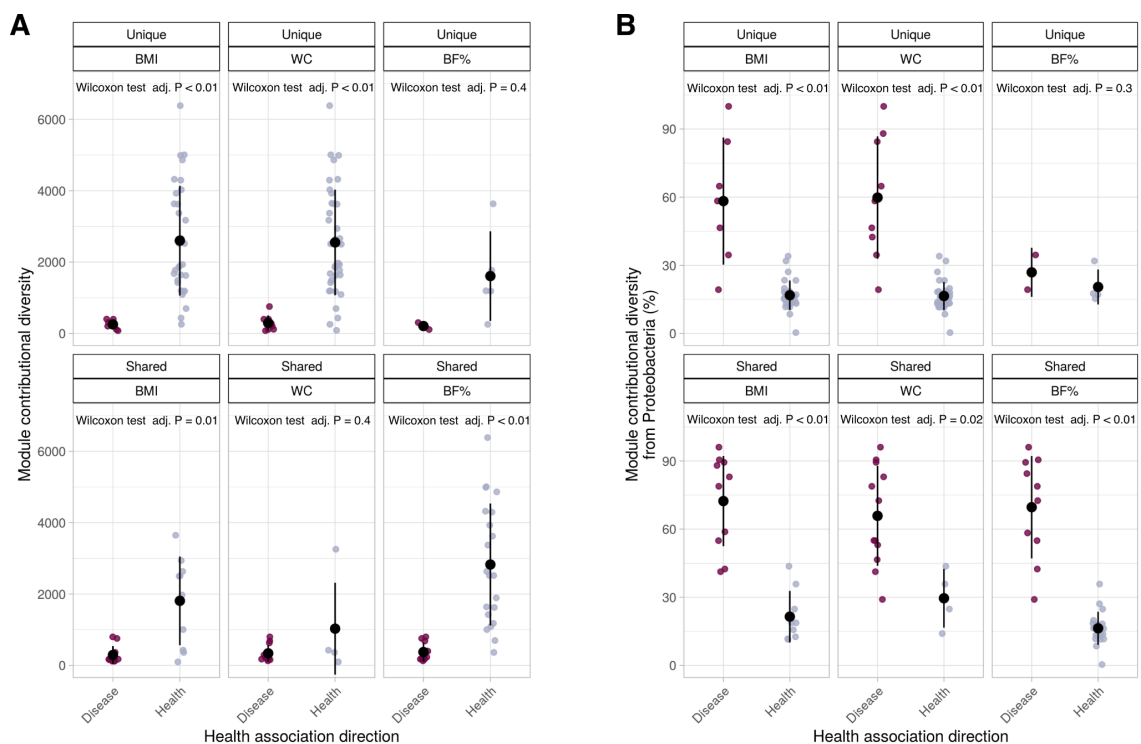
**Figure 5. Functional features of the microbiome form correlation blocks with host factors, including blood pressure and fecal short chain fatty acid levels.** Heatmap showing the correlations between modules significantly associated with OB measures, CHS or shared between both. Association blocks were obtained using hierarchical all-against-all association testing (HAIIA) and are demarcated by black borders; significance of individual associations are denoted by white dots (adj.  $P < 0.1$ ) and color indicates Spearman's correlation coefficient ( $\rho$ ).

After SCFAs, diastolic and systolic blood pressure were the host factors with the largest blocks of correlated features; we detected 15 blocks (comprising 23 modules) and 434 ortholog blocks (comprising 602 orthologs) correlated with either. Most of these blocks overlapped with the large correlation blocks formed by SCFAs. The direction of the association was largely the same as above, with higher abundance of transport of simple sugars and osmoprotectants linked to increased blood pressure, while the opposite was true for the abundance of methanogenesis, glycolysis, gluconeogenesis and associated biosynthesis modules (figure 5). Smaller correlation blocks were also detected for inflammation measured by hsCRP, HbA1c, HOMA-IR, leptin, insulin, triglycerides, adiponectin and VLDL.

***Disease-associated features are less diverse and mainly contributed by members of Proteobacteria***

The above results indicate a scenario where the loss of gut homeostasis is driven by obesity, and which is characterized by an increase of microbial functions related to transport of simple sugars and tolerance to reactive oxygen

species, where fermentative functions are reduced. This pattern closely follows the germ-organ theory of non-communicable diseases (48), wherein inflammation in the gut is linked to a disruption of anaerobiosis in the gut and an expansion of facultative anaerobes from the phylum *Proteobacteria* (49). Therefore, we assessed the contribution of microbial taxa to the functional profile of the microbiome of Colombians using HUMAnN2's tiered search, focusing on members of *Proteobacteria*.



**Figure 6. *Proteobacteria* are the main contributors to disease-linked modules.** A) Contributinal diversity of modules significantly associated with obesity measures, CHS or shared by both. B) Relative contribution of *Proteobacteria* taxa to the total module diversity. Black points and bars represent mean and standard deviations.

We calculated the total contributinal diversity of the modules significantly associated with the OB metrics, CHS or both. We measured the contributinal diversity as the number of taxa identified that contributed to the total abundance of said module, and compared the values between health- and disease-linked modules. Overall, we observed that modules associated with health were more diverse than those associated with disease (figure 6A). For

example, in the case of modules unique to BMI,  $253.43 \pm 129.54$  taxa contributed on average to disease-associated modules, whereas  $2600.94 \pm 1537.17$  contributed to health-associated modules across all subjects (Mann-Whitney U = 4, adj. P < 0.01). Very similar patterns were observed with other obesity measures, both in uniquely associated or shared modules with CHS (figure 6A).

Next, we assessed the proportion of the contributory diversity that could be attributed only to members of the phylum *Proteobacteria*. The pattern was opposite to the one described above: *Proteobacteria* taxa contributed more to the total abundance of disease-linked modules than to health-associated ones (figure 6B). In the case of modules uniquely associated to BMI, the mean contribution of *Proteobacteria* to disease-linked modules was  $58.29 \% \pm 27.96$ , while only of  $16.83 \% \pm 6.49$  in the case of health-related modules (Mann-Whitney U = 204, adj. P < 0.01). This was consistent across obesity measures and unique or shared sets of modules (figure 6B).

## Discussion

The consolidation of microbiome science as a framework to understand health and disease processes depends on the generalizability of identified links between the microbiome and host phenotypes across human populations. In this study, we looked into the functional profile of the gut metagenome of a deeply-phenotyped cohort from a non-Westernized population, Colombian adults, and its association with obesity and cardiometabolic health. We incorporated data from published studies to narrow the set of hypotheses to test. Our rich data set allowed us to decouple obesity and the cardiometabolic health status of the subjects while also accounting for several confounding factors, including age, sex, the city of origin and medication consumption. Our analyses indicate that alterations to the diversity and composition of the gut microbiome functional profile are mainly driven by obesity, not by the cardiometabolic status of the individuals. Nevertheless, the microbiome of diseased individuals in both conditions left a footprint characterized by a lower sequence richness, a decreased fermentative ability and an increased response

to higher levels of oxygen in the gut lumen. Consistent with these findings, functional features enriched in both diseased states, for obesity and cardiometabolic status, were mainly contributed by members of the phylum *Proteobacteria*, many of which are considered as pathogens or pathobionts (50). Our results validate in this population the previously reported loss of beneficial functions and diversity, and emphasize the importance of using well-characterized cohorts to disentangle overlapping associations of host phenotypes with the microbiome.

Easily accessible data from large cohorts with host and metagenome information comparable to those used in the present study are scarce. Thus, performing a multi-population analysis to decouple the association of the microbiome with obesity and cardiometabolic status, while controlling confounding factors, was not feasible. Nevertheless, we were able to incorporate findings from published studies into our analyses by focusing on a set of functional features that were previously linked to obesity or cardiometabolic conditions. This allowed us to assess the robustness and generalizability of the associations on an understudied population from South America (51) using a rigorous statistical framework.

Even though only a fraction of the features evaluated were found to be significantly associated with obesity or cardiometabolic status, our results underscore the link between these conditions and the gut microbiome through inflammation, both systemic and epithelial. Multiple inflammatory mechanisms have been associated with the development of obesity, insulin resistance and cardiovascular disease (52). Recent evidence underscores the importance of intestinal inflammation in the development of obesity (53). In murine models, gut epithelial inflammation precedes and correlates with diet-induced obesity and insulin resistance, and interactions between the gut microbiome and diet are required for the induction of inflammation (54). In turn, epithelial inflammation affects intestinal permeability, allowing the translocation of bacterial antigens such as lipopolysaccharide (LPS) to circulation, resulting in metabolic endotoxemia (55). The low-grade systemic inflammation induced by increased

LPS causes an increase in proinflammatory cytokines, compromising insulin signaling and culminating in insulin resistance and glucose intolerance (56).

In our cohort, functions associated with adverse health outcomes indicate that the configuration of the gut microbiome in diseased subjects favors the growth of facultative anaerobes due to a loss of gut epithelial hypoxia (49). It has been suggested that the expansion of members of the phylum *Proteobacteria*, in particular of the family *Enterobacteriaceae*, are a signature of epithelial dysfunction (48,57). Colonocytes utilize microbiome-produced butyrate as a source of energy by mitochondrial beta-oxidation, depleting oxygen in the epithelium surface; thus, the gut environment is dominated by obligate anaerobes, such as bacteria from the phyla *Bacteroidota*, *Firmicutes*, *Actinobacteriota* or *Verrucomicrobiota*, and methanogenic archaea from the phyla *Methanobacteriota* and *Thermoplasmata* (58). These microorganisms ferment complex carbohydrates that escape host digestion, or utilize its by-products as energy substrate (59). Gut epithelium disruption by pro-inflammatory stimuli, such as the use of antibiotics (60) or the consumption of a low-fiber, high-fat diet (61–63), promotes a shift in the energy metabolism of colonocytes from beta-oxidation of butyrate towards anaerobic glycolysis, which requires a higher consumption of glucose and does not result in oxygen depletion (58). In turn, such conditions present a selective advantage to facultative anaerobes, thus favoring their expansion (64,65). The expansion of facultative anaerobes is detrimental to the health of the host, since members of *Proteobacteria* are not fiber degraders, and their presence hinders the nutrition of the host by metabolizing products of microbial fermentation to carbon dioxide in the presence of oxygen (66,67). Moreover, the disruption of the epithelial barrier results in translocation of microbial antigens, such as LPS, which promote low-grade inflammation, exacerbating epithelial dysfunction (68) and inducing insulin resistance (69).

We observed that functions enriched in both obesity and unhealthy cardiometabolic status display the hallmarks of the germ-organ theory of non-communicable disease (48), where obesity-associated inflammation results in a loss of anaerobiosis in the gut and an expansion of *Proteobacteria*, which is

aggravated by the subsequent appearance of cardiometabolic disease. The enrichment of functions involved in the tolerance to reactive oxygen species, and the utilization of sulfur and nitrogen in alternative energy metabolism pathways suggest the loss of the hypoxic status of the gut epithelium (60). This might be due to a degradation of the gut epithelium, as evidenced by orthologs involved in the utilization of ethanolamine, derived from the cell membrane of dead epithelial cells (65,70). Inflammation of the gut epithelium causes an increase in neutrophil transmigration to the gut lumen (71), where reactive oxygen species and iron and zinc sequesters are released (72), hence the increase in siderophores and zinc transporters we observed in obese subjects. Moreover, the potential synthesis of beneficial metabolites, such as acetate, propionate and succinate was reduced, while bacterial antigen biosynthesis, such as lipopolysaccharide, together with the capacity for metabolizing trimethylamine (TMA) from trimethylamine-N-oxide (TMAO), choline or glycine betaine was increased. TMAO can be used by diverse *Enterobacteriaceae* taxa as electron acceptor, converting it to TMA in the gut (65,73). In turn, TMA is absorbed by the host, converted into TMAO in the liver and enters circulation; TMAO inhibits cholesterol transport and promotes its accumulation in macrophages, inducing the formation of atherosclerotic plaques (74).

Compared to disease-associated modules, health-associated modules had a higher contributory diversity, that is, more taxa encoded such functions. The contribution of members of *Proteobacteria* to the total abundance of the studied features was higher for disease- than for health-associated modules and orthologs. This could help explain the negative correlation we observed between the database-independent sequence diversity and obesity. We shown in this cohort that individuals with higher abundance of pathobionts, including *Proteobacteria*, had worse obesity and cardiometabolic health outcomes than subjects with higher counts of consortia that included *Ruminococcaceae*, *Bacteroidales*, *Christensenellaceae*, *Methanobacteriaceae* or *Akkermansiaceae* (15). Additionally, we observed that the disease-associated modules were positively correlated with other host parameters such as LBP, a marker of translocation of bacterial antigens to the host bloodstream (75); hsCRP, a

marker of systemic inflammation (76); and fecal short chain fatty acids levels (18,77). All the aforementioned host factors were significantly different by cardiometabolic status or body mass index categories (table 1). Conversely, features enriched in healthy subjects highlight the loss of the fermentative capacity of the microbiome, such as methanogenesis, central energy metabolism pathways, the production of SCFAs such as acetate and propionate, and the synthesis of succinate and ascorbate.

We did not observe complete replication of the functional features tested, that is, not all features reached statistical significance in our tests. There are multiple non-exclusive explanations for this. First, we used a statistical framework in which feature abundance was transformed to take into account the compositional nature of sequencing data. In addition, we accounted for host factors known to influence the composition of the microbiome; namely, the age, sex, local geographical origin, and the medication consumption of the subjects. Incorporating known confounders can reduce the risk of obtaining false positives in cross-sectional population studies (40,46). In other words, accounting for common confounders might facilitate the comparison between studies by ameliorating biases introduced by the confounding variables (46). Second, some of the non-significant results could be due to differences in the composition of the microbiome between human populations. Indeed, whether certain characteristics of the microbiome are universal and other are population-specific, and which associations of the microbiome with human health are conserved across populations are still open questions in microbiome research (78). Lastly, the effect size of the correlation between certain human phenotypes and the gut microbiome tends to be small. Therefore, some associations can only be recovered by studying cohorts with even larger sample size than the one used here (46).

Our study is not without limitations. As discussed above, the inclusion of host covariates such as age, sex, city of origin and the consumption of certain medications allowed us to reduce the potential for confounding. Although *in vitro* studies demonstrated that drugs with human targets of all therapeutic classes inhibit the growth of human commensals (79), we only considered medications

with direct relevance to obesity and cardiometabolic health, with no dosage information. We are, thus, unable to rule out residual confounding from variables not measured in our study (47). In addition, our findings are based on cross-sectional data from a single human population. Therefore, we are only able to report associations between microbial features and host phenotypes, not provide causal inference. Finally, the present study used shotgun metagenomics to assess the functional profile of the microbiome. This approach does not directly measure the transcripts or proteins expressed by the microbiota, but rather the genomic potential (80). Thus, the activity of the community at a specific point in time as a response to the studied conditions cannot be measured (81,82). Nevertheless, this approach provides useful information insofar as the abundance of genes in a metagenome is positively correlated with their mRNA expression (81) and the protein levels (83).

The present study strengthens our understanding of the metabolic potential of the gut microbiome in obesity and cardiometabolic disease in an understudied population from a middle-income country (78). The scarcity of easily accessible datasets with as much information as ours makes the direct comparison between studies difficult, but does not prevent us from evaluating the generalizability of previously reported patterns by other means. We expect that the robust associations we report will serve to inform mechanistic studies of the role of the microbiome in disease and guide the development of microbiome-based interventions for personalized nutrition and medicine.

## **Acknowledgements**

This work was supported by the Max Planck Society and Vidarium–Nutrition, Health and Wellness Research Center. We thank the participants who took part in the study, and the Vidarium, EPS SURA and Dinámica IPS staff that helped with recruitment and field work. Some authors of this work collaborate through the Microbiome & Health Network. We are grateful to Alejandra Duque, Albane Ruaud, Taichi Suzuki, William Walters and Laura Salazar-Jaramillo for the fruitful discussions and comments.



While engaged in the research project, J.S.E. was employed by a food company. J.D.L.C.-Z., N.D.Y. and R.E.L. had no competing interests.

## References

1. Suzuki TA, Ley RE. The role of the microbiota in human genetic adaptation. *Science* [Internet]. 2020 Dec 4;370(6521). Available from: <http://dx.doi.org/10.1126/science.aaz6827>
2. Rojo D, Méndez-García C, Raczkowska BA, Bargiela R, Moya A, Ferrer M, et al. Exploring the human microbiome from multiple perspectives: factors altering its composition and function. *FEMS Microbiol Rev* [Internet]. 2017 Jul 1;41(4):453–78. Available from: <http://dx.doi.org/10.1093/femsre/fuw046>
3. World Health Organization. Obesity [Internet]. [cited 2022 Feb 23]. Available from: [https://www.who.int/en/health-topics/obesity#tab=tab\\_1](https://www.who.int/en/health-topics/obesity#tab=tab_1)
4. Grundy SM, Brewer HB Jr, Cleeman JI, Smith SC Jr, Lenfant C, American Heart Association, et al. Definition of metabolic syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation* [Internet]. 2004 Jan 27;109(3):433–8. Available from: <http://dx.doi.org/10.1161/01.CIR.0000111245.75752.C6>
5. Wildman RP, Muntner P, Reynolds K, McGinn AP, Rajpathak S, Wylie-Rosett J, et al. The obese without cardiometabolic risk factor clustering and the normal weight with cardiometabolic risk factor clustering: prevalence and correlates of 2 phenotypes among the US population (NHANES 1999-2004). *Arch Intern Med* [Internet]. 2008 Aug;168(15):1617–24. Available from: <http://dx.doi.org/10.1001/archinte.168.15.1617>
6. World Health Organization. Obesity and overweight [Internet]. 2021 [cited 2021 Sep 22]. Available from: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>
7. Nyberg ST, Batty GD, Pentti J, Virtanen M, Alfredsson L, Fransson EI, et al. Obesity and loss of disease-free years owing to major non-communicable diseases: a multicohort study. *Lancet Public Health* [Internet]. 2018 Oct;3(10):e490–7. Available from: [http://dx.doi.org/10.1016/S2468-2667\(18\)30139-7](http://dx.doi.org/10.1016/S2468-2667(18)30139-7)
8. Sorbara MT, Pamer EG. Microbiome-based therapeutics. *Nat Rev Microbiol* [Internet]. 2022 Jan 6; Available from: <http://dx.doi.org/10.1038/s41579-021-00667-9>

9. Thingholm LB, Rühlemann MC, Koch M, Fuqua B, Laucke G, Boehm R, et al. Obese Individuals with and without Type 2 Diabetes Show Different Gut Microbial Functional Capacity and Composition. *Cell Host Microbe* [Internet]. 2019 Aug 14;26(2):252–64.e10. Available from: <http://dx.doi.org/10.1016/j.chom.2019.07.004>
10. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* [Internet]. 2017 Oct 5;550(7674):61–6. Available from: <http://dx.doi.org/10.1038/nature23889>
11. Wilmanski T, Rappaport N, Diener C, Gibbons SM, Price ND. From taxonomy to metabolic output: what factors define gut microbiome health? *Gut Microbes* [Internet]. 2021 Jan;13(1):1–20. Available from: <http://dx.doi.org/10.1080/19490976.2021.1907270>
12. The Human Microbiome Project Consortium, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. Structure, function and diversity of the healthy human microbiome. *Nature* [Internet]. 2012 Jun;486(7402):207–14. Available from: <http://dx.doi.org/10.1038/nature11234>
13. Rojo D, Hevia A, Bargiela R, López P, Cuervo A, González S, et al. Ranking the impact of human health disorders on gut metabolism: systemic lupus erythematosus and obesity as study cases. *Sci Rep* [Internet]. 2015 Feb 6;5:8310. Available from: <http://dx.doi.org/10.1038/srep08310>
14. Armour CR, Nayfach S, Pollard KS, Sharpton TJ. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* [Internet]. 2019 Jul;4(4). Available from: <http://dx.doi.org/10.1128/mSystems.00332-18>
15. de la Cuesta-Zuluaga J, Corrales-Agudelo V, Velásquez-Mejía EP, Carmona JA, Abad JM, Escobar JS. Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. *Sci Rep* [Internet]. 2018 Jul 27;8(1):11356. Available from: <http://dx.doi.org/10.1038/s41598-018-29687-x>
16. de la Cuesta-Zuluaga J, Corrales-Agudelo V, Carmona JA, Abad JM, Escobar JS. Body size phenotypes comprehensively assess cardiometabolic risk and refine the association between obesity and gut microbiota. *Int J Obes* [Internet]. 2018 Mar;42(3):424–32. Available from: <http://dx.doi.org/10.1038/ijo.2017.281>
17. García-Vega ÁS, Corrales-Agudelo V, Reyes A, Escobar JS. Diet Quality, Food Groups and Nutrients Associated with the Gut Microbiota in a Nonwestern Population. *Nutrients* [Internet]. 2020 Sep 25;12(10). Available from: <http://dx.doi.org/10.3390/nu12102938>

18. de la Cuesta-Zuluaga J, Mueller NT, Álvarez-Quintero R, Velásquez-Mejía EP, Sierra JA, Corrales-Agudelo V, et al. Higher Fecal Short-Chain Fatty Acid Levels Are Associated with Gut Microbiome Dysbiosis, Obesity, Hypertension and Cardiometabolic Disease Risk Factors. *Nutrients* [Internet]. 2019;11(1):51. Available from: <http://dx.doi.org/10.3390/nu11010051>
19. Siri WE. Body composition from fluid spaces and density: analysis of methods. *Techniques for measuring body composition*. 1961;61:223–44.
20. de la Cuesta-Zuluaga J, Mueller NT, Corrales-Agudelo V, Velásquez-Mejía EP, Carmona JA, Abad JM, et al. Metformin Is Associated With Higher Relative Abundance of Mucin-Degrading *Akkermansia muciniphila* and Several Short-Chain Fatty Acid – Producing Microbiota in the Gut. *Diabetes Care* [Internet]. 2017;40(1):54–62. Available from: <http://dx.doi.org/10.2337/dc16-1324>
21. Karasov TL, Almario J, Friedemann C, Ding W, Giolai M, Heavens D, et al. *Arabidopsis thaliana* and *Pseudomonas* Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host Microbe* [Internet]. 2018 Jul 11;24(1):168–79.e4. Available from: <http://dx.doi.org/10.1016/j.chom.2018.06.011>
22. Droop AP. fqtools: an efficient software suite for modern FASTQ file manipulation. *Bioinformatics* [Internet]. 2016 Jun 15;32(12):1883–4. Available from: <http://dx.doi.org/10.1093/bioinformatics/btw088>
23. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* [Internet]. 2014 Jun 12;15:182. Available from: <http://dx.doi.org/10.1186/1471-2105-15-182>
24. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* [Internet]. 2016 Oct 1;32(19):3047–8. Available from: <http://dx.doi.org/10.1093/bioinformatics/btw354>
25. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *mSystems* [Internet]. 2018 May;3(3). Available from: <http://dx.doi.org/10.1128/mSystems.00039-18>
26. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* [Internet]. 2018 Nov;15(11):962–8. Available from: <http://dx.doi.org/10.1038/s41592-018-0176-y>
27. de la Cuesta-Zuluaga J, Ley RE, Youngblut ND. Struo: a pipeline for building custom databases for common metagenome profilers.

- Bioinformatics [Internet]. 2019 Nov 28; Available from: <http://dx.doi.org/10.1093/bioinformatics/btz899>
28. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* [Internet]. 2018 Nov;36(10):996–1004. Available from: <http://dx.doi.org/10.1038/nbt.4229>
  29. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun* [Internet]. 2017 Oct 10;8(1):845. Available from: <http://dx.doi.org/10.1038/s41467-017-00900-1>
  30. Wu H, Tremaroli V, Schmidt C, Lundqvist A, Olsson LM, Krämer M, et al. The Gut Microbiota in Prediabetes and Diabetes: A Population-Based Cross-Sectional Study. *Cell Metab* [Internet]. 2020 Sep 1;32(3):379–90.e3. Available from: <http://dx.doi.org/10.1016/j.cmet.2020.06.011>
  31. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: <https://www.R-project.org/>
  32. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics Intellig Lab Syst* [Internet]. 2015 Apr;143:85–96. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0169743915000490>
  33. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci Rep* [Internet]. 2017 Nov 24;7(1):16252. Available from: <http://dx.doi.org/10.1038/s41598-017-16520-0>
  34. van den Boogaart KG, Tolosana-Delgado R. “compositions”: A unified R package to analyze compositional data. *Comput Geosci* [Internet]. 2008 Apr 1;34(4):320–38. Available from: <https://www.sciencedirect.com/science/article/pii/S009830040700101X>
  35. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. *Gigascience* [Internet]. 2019 Sep 1;8(9). Available from: <http://dx.doi.org/10.1093/gigascience/giz107>
  36. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Polit Anal* [Internet]. 2007 [cited 2021 Jan 18];15(3):199–236. Available from: [https://www.cambridge.org/core/product/identifier/S1047198700006483/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S1047198700006483/type/journal_article)

37. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. vegan: Community Ecology Package [Internet]. R package version 2.3-1. 2015. p. 264. Available from: <http://dx.doi.org/10.4135/9781412971874.n145>
38. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, et al. Multivariable Association Discovery in Population-scale Meta-omics Studies [Internet]. bioRxiv. 2021 [cited 2021 Mar 11]. p. 2021.01.20.427420. Available from: <https://www.biorxiv.org/content/10.1101/2021.01.20.427420v1>
39. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods [Internet]. 2017 Oct 31;14(11):1023–4. Available from: <http://dx.doi.org/10.1038/nmeth.4468>
40. Ghosh TS, Das M, Jeffery IB, O'Toole PW. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. Elife [Internet]. 2020 Mar 11;9. Available from: <http://dx.doi.org/10.7554/eLife.50240>
41. Sinha T, Vich Vila A, Garmaeva S, Jankipersadsing SA, Imhann F, Collij V, et al. Analysis of 1135 gut metagenomes identifies sex-specific resistome profiles. Gut Microbes [Internet]. 2019;10(3):358–66. Available from: <http://dx.doi.org/10.1080/19490976.2018.1528822>
42. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med [Internet]. 2018 Oct;24(10):1532–5. Available from: <http://dx.doi.org/10.1038/s41591-018-0164-x>
43. Jackson MA, Goodrich JK, Maxan M-E, Freedberg DE, Abrams JA, Poole AC, et al. Proton pump inhibitors alter the composition of the gut microbiota. Gut [Internet]. 2016 May;65(5):749–56. Available from: <http://gut.bmj.com/lookup/doi/10.1136/gutjnl-2015-310861>
44. Vich Vila A, Collij V, Sanna S, Sinha T, Imhann F, Bourgonje AR, et al. Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. Nat Commun [Internet]. 2020 Jan 17;11(1):362. Available from: <http://dx.doi.org/10.1038/s41467-019-14177-z>
45. Mueller NT, Differding MK, Zhang M, Maruthur NM, Juraschek SP, Miller ER 3rd, et al. Metformin Affects Gut Microbiome Composition and Function and Circulating Short-Chain Fatty Acids: A Randomized Trial. Diabetes Care [Internet]. 2021 May 18; Available from: <http://dx.doi.org/10.2337/dc20-2257>
46. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. Nature

- [Internet]. 2020 Nov;587(7834):448–54. Available from: <http://dx.doi.org/10.1038/s41586-020-2881-9>
47. Forslund SK, Chakaroun R, Zimmermann-Kogadeeva M, Markó L, Aron-Wisnewsky J, Nielsen T, et al. Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* [Internet]. 2021 Dec;600(7889):500–5. Available from: <http://dx.doi.org/10.1038/s41586-021-04177-9>
  48. Byndloss MX, Bäumlér AJ. The germ-organ theory of non-communicable diseases. *Nat Rev Microbiol* [Internet]. 2018 Feb;16(2):103–10. Available from: <http://dx.doi.org/10.1038/nrmicro.2017.158>
  49. Bäumlér AJ, Sperandio V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* [Internet]. 2016 Jul;535(7610):85–93. Available from: <http://www.nature.com/doi/10.1038/nature18849>
  50. Rizzatti G, Lopetuso LR, Gibiino G, Binda C, Gasbarrini A. Proteobacteria: A Common Factor in Human Diseases. *Biomed Res Int* [Internet]. 2017 Nov 2;2017:9351507. Available from: <http://dx.doi.org/10.1155/2017/9351507>
  51. Abdill RJ, Adamowicz EM, Blekhman R. Public human microbiome data dominated by highly developed countries [Internet]. *bioRxiv*. 2021 [cited 2021 Sep 6]. p. 2021.09.02.458641. Available from: <https://www.biorxiv.org/content/10.1101/2021.09.02.458641v1>
  52. Scheithauer TPM, Rampanelli E, Nieuwdorp M, Vallance BA, Verchere CB, van Raalte DH, et al. Gut Microbiota as a Trigger for Metabolic Inflammation in Obesity and Type 2 Diabetes. *Front Immunol* [Internet]. 2020 Oct 16;11:571731. Available from: <http://dx.doi.org/10.3389/fimmu.2020.571731>
  53. Cox AJ, West NP, Cripps AW. Obesity, inflammation, and the gut microbiota. *Lancet Diabetes Endocrinol* [Internet]. 2015 Mar;3(3):207–15. Available from: [http://dx.doi.org/10.1016/S2213-8587\(14\)70134-2](http://dx.doi.org/10.1016/S2213-8587(14)70134-2)
  54. Ding S, Chi MM, Scull BP, Rigby R, Schwerbrock NMJ, Magness S, et al. High-fat diet: bacteria interactions promote intestinal inflammation which precedes and correlates with obesity and insulin resistance in mouse. *PLoS One* [Internet]. 2010 Aug 16;5(8):e12191. Available from: <http://dx.doi.org/10.1371/journal.pone.0012191>
  55. Gomes JMG, Costa J de A, Alfenas R de CG. Metabolic endotoxemia and diabetes mellitus: A systematic review. *Metabolism* [Internet]. 2017 Mar;68:133–44. Available from: <http://dx.doi.org/10.1016/j.metabol.2016.12.009>
  56. Ding S, Lund PK. Role of intestinal inflammation as an early event in

- obesity and insulin resistance. *Curr Opin Clin Nutr Metab Care* [Internet]. 2011 Jul;14(4):328–33. Available from: <http://dx.doi.org/10.1097/MCO.0b013e3283478727>
57. Litvak Y, Byndloss MX, Tsohis RM, Bäumlér AJ. Dysbiotic Proteobacteria expansion: a microbial signature of epithelial dysfunction. *Curr Opin Microbiol* [Internet]. 2017 Oct;39:1–6. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S136952741630220X>
  58. Litvak Y, Byndloss MX, Bäumlér AJ. Colonocyte metabolism shapes the gut microbiota. *Science* [Internet]. 2018 Nov 30 [cited 2021 Dec 2]; Available from: <https://science.sciencemag.org/content/362/6418/eaat9076.abstract>
  59. Makki K, Deehan EC, Walter J, Bäckhed F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host Microbe* [Internet]. 2018 Jun 13;23(6):705–15. Available from: <http://dx.doi.org/10.1016/j.chom.2018.05.012>
  60. Palleja A, Mikkelsen KH, Forslund SK, Kashani A, Allin KH, Nielsen T, et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat Microbiol* [Internet]. 2018 Nov;3(11):1255–65. Available from: <http://dx.doi.org/10.1038/s41564-018-0257-9>
  61. Statovci D, Aguilera M, MacSharry J, Melgar S. The Impact of Western Diet and Nutrients on the Microbiota and Immune Response at Mucosal Interfaces. *Front Immunol* [Internet]. 2017 Jul 28;8:838. Available from: <http://dx.doi.org/10.3389/fimmu.2017.00838>
  62. Martínez-Medina M, Denizot J, Dreux N, Robin F, Billard E, Bonnet R, et al. Western diet induces dysbiosis with increased *E coli* in CEABAC10 mice, alters host barrier function favouring AIEC colonisation. *Gut* [Internet]. 2014 Jan;63(1):116–24. Available from: <http://dx.doi.org/10.1136/gutjnl-2012-304119>
  63. O’Keefe SJD, Li JV, Lahti L, Ou J, Carbonero F, Mohammed K, et al. Fat, fibre and cancer risk in African Americans and rural Africans. *Nat Commun* [Internet]. 2015 Apr 28;6:6342. Available from: <http://dx.doi.org/10.1038/ncomms7342>
  64. Shin N-R, Whon TW, Bae J-W. Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol* [Internet]. 2015 Sep;33(9):496–503. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0167779915001390>
  65. Zeng MY, Inohara N, Nuñez G. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. *Mucosal Immunol* [Internet]. 2017 Jan;10(1):18–26. Available from: <http://dx.doi.org/10.1038/mi.2016.75>
  66. Conway T, Cohen PS. Commensal and Pathogenic *Escherichia coli*

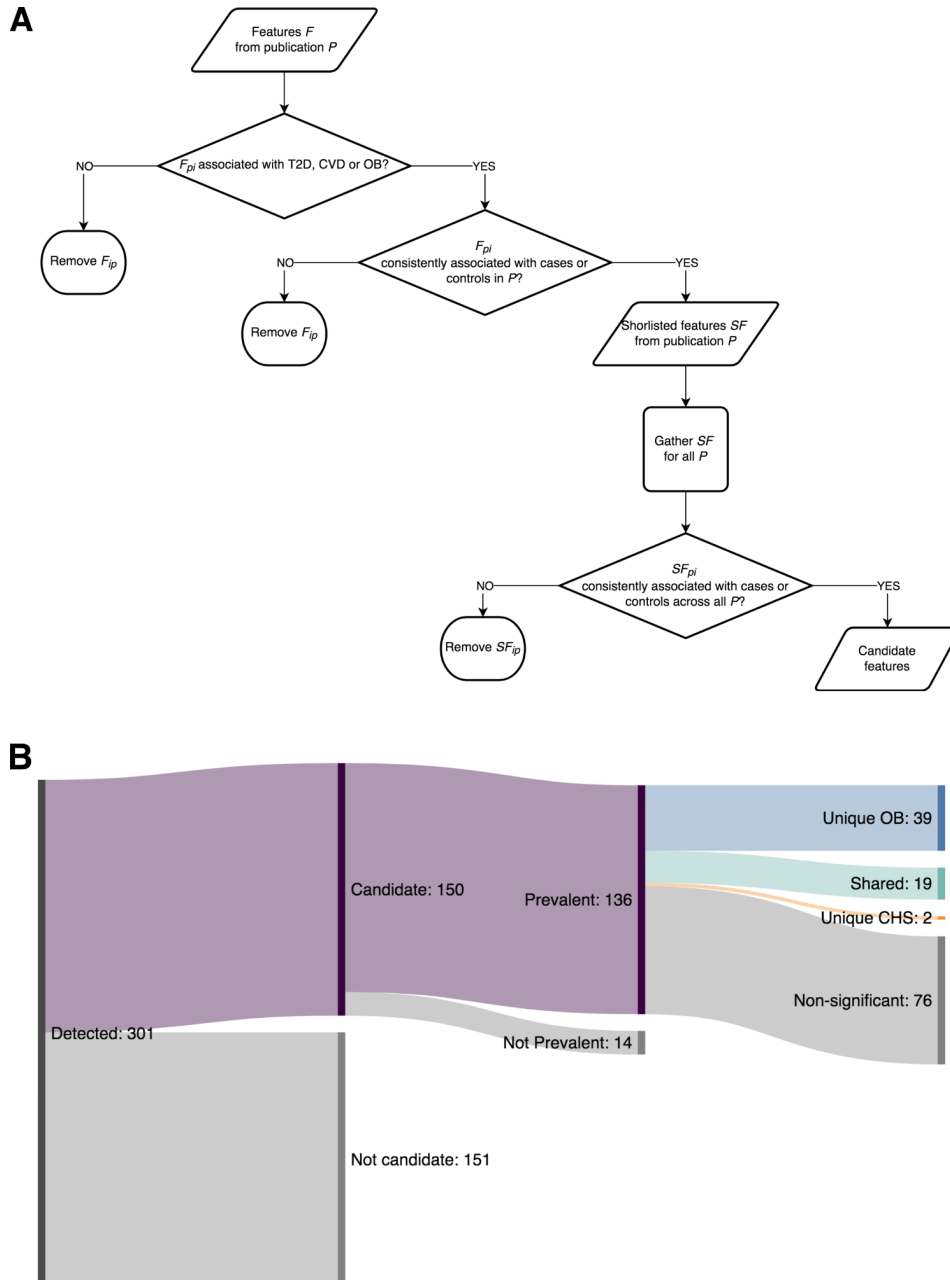
- Metabolism in the Gut. *Microbiol Spectr* [Internet]. 2015 Jun;3(3). Available from: <http://dx.doi.org/10.1128/microbiolspec.MBP-0006-2014>
67. Spiga L, Winter MG, Furtado de Carvalho T, Zhu W, Hughes ER, Gillis CC, et al. An Oxidative Central Metabolism Enables Salmonella to Utilize Microbiota-Derived Succinate. *Cell Host Microbe* [Internet]. 2017 Sep 13;22(3):291–301.e6. Available from: <http://dx.doi.org/10.1016/j.chom.2017.07.018>
  68. Mohammad S, Thiemermann C. Role of Metabolic Endotoxemia in Systemic Inflammation and Potential Interventions. *Front Immunol* [Internet]. 2020;11:594150. Available from: <http://dx.doi.org/10.3389/fimmu.2020.594150>
  69. Rorato R, Borges B de C, Uchoa ET, Antunes-Rodrigues J, Elias CF, Elias LLK. LPS-Induced Low-Grade Inflammation Increases Hypothalamic JNK Expression and Causes Central Insulin Resistance Irrespective of Body Weight Changes. *Int J Mol Sci* [Internet]. 2017 Jul 4;18(7). Available from: <http://dx.doi.org/10.3390/ijms18071431>
  70. Garsin DA. Ethanolamine utilization in bacterial pathogens: roles and regulation. *Nat Rev Microbiol* [Internet]. 2010 Apr;8(4):290–5. Available from: <http://dx.doi.org/10.1038/nrmicro2334>
  71. Fournier BM, Parkos CA. The role of neutrophils during intestinal inflammation. *Mucosal Immunol* [Internet]. 2012 Jul;5(4):354–66. Available from: <http://dx.doi.org/10.1038/mi.2012.24>
  72. Zhang D, Frenette PS. Cross talk between neutrophils and the microbiota. *Blood* [Internet]. 2019 May 16;133(20):2168–77. Available from: <http://dx.doi.org/10.1182/blood-2018-11-844555>
  73. Barrett EL, Kwan HS. Bacterial reduction of trimethylamine oxide. *Annu Rev Microbiol* [Internet]. 1985;39:131–49. Available from: <http://dx.doi.org/10.1146/annurev.mi.39.100185.001023>
  74. Geng J, Yang C, Wang B, Zhang X, Hu T, Gu Y, et al. Trimethylamine N-oxide promotes atherosclerosis via CD36-dependent MAPK/JNK pathway. *Biomed Pharmacother* [Internet]. 2018 Jan;97:941–7. Available from: <http://dx.doi.org/10.1016/j.biopha.2017.11.016>
  75. Guo S, Al-Sadi R, Said HM, Ma TY. Lipopolysaccharide causes an increase in intestinal tight junction permeability in vitro and in vivo by inducing enterocyte membrane expression and localization of TLR-4 and CD14. *Am J Pathol* [Internet]. 2013 Feb;182(2):375–87. Available from: <http://dx.doi.org/10.1016/j.ajpath.2012.10.014>
  76. Gentile M, Panico S, Rubba F, Mattiello A, Chiodini P, Jossa F, et al. Obesity, overweight, and weight gain over adult life are main determinants



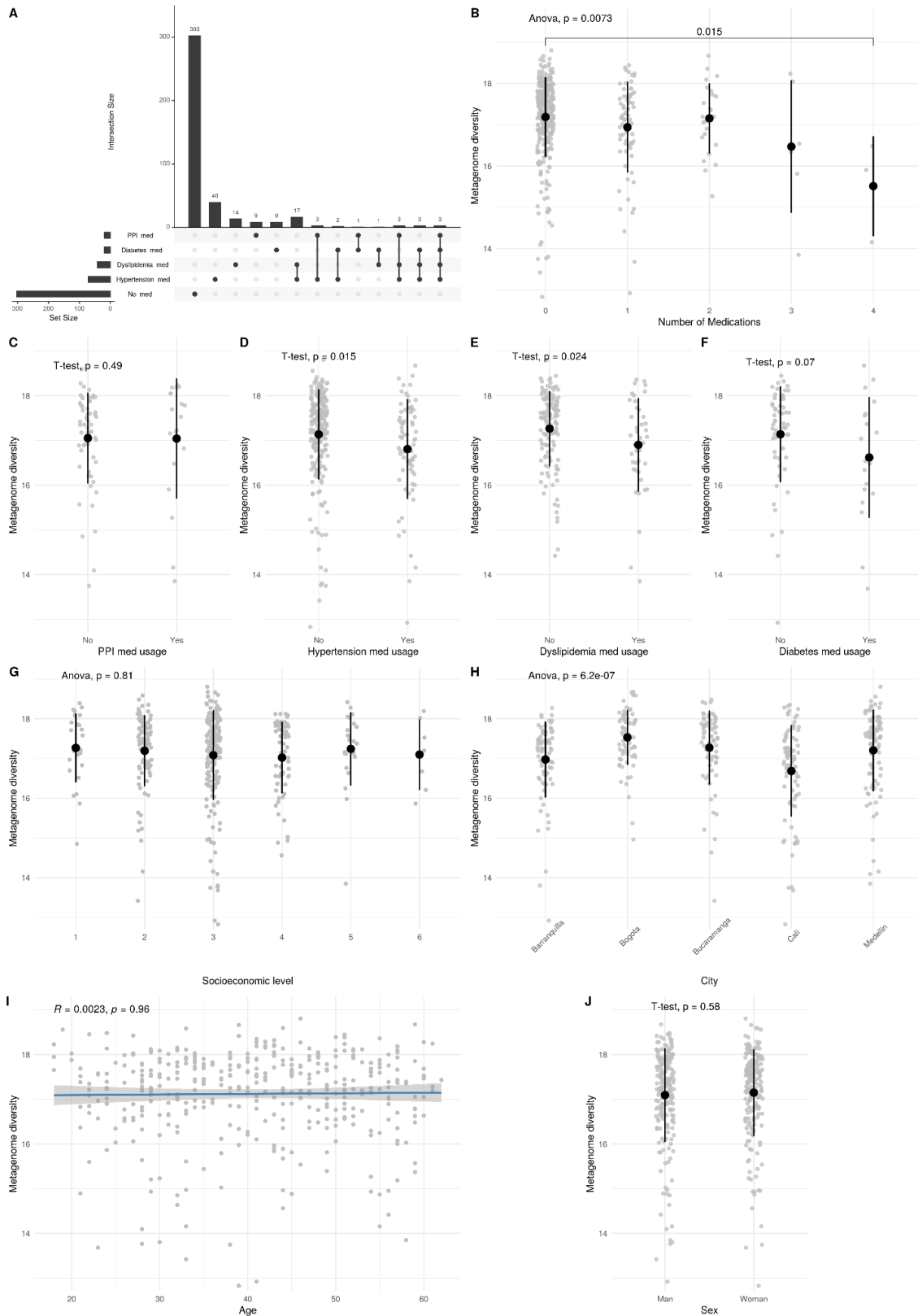
of elevated hs-CRP in a cohort of Mediterranean women. *Eur J Clin Nutr* [Internet]. 2010 Aug;64(8):873–8. Available from: <http://dx.doi.org/10.1038/ejcn.2010.69>

77. Kim KN, Yao Y, Ju SY. Short Chain Fatty Acids and Fecal Microbiota Abundance in Humans with Obesity: A Systematic Review and Meta-Analysis. *Nutrients* [Internet]. 2019 Oct 18;11(10). Available from: <http://dx.doi.org/10.3390/nu11102512>
78. Porras AM, Brito IL. The internationalization of human microbiome research. *Curr Opin Microbiol* [Internet]. 2019 Aug;50:50–5. Available from: <http://dx.doi.org/10.1016/j.mib.2019.09.012>
79. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* [Internet]. 2018 Mar 29;555(7698):623–8. Available from: <http://dx.doi.org/10.1038/nature25979>
80. Shakya M, Lo C-C, Chain PSG. Advances and Challenges in Metatranscriptomic Analysis. *Front Genet* [Internet]. 2019 Sep 25;10:904. Available from: <http://dx.doi.org/10.3389/fgene.2019.00904>
81. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences* [Internet]. 2014 Jun;111(22):E2329–38. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.1319284111>
82. Maurice CF, Haiser HJ, Turnbaugh PJ. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell* [Internet]. 2013 Jan 17;152(1-2):39–50. Available from: <http://dx.doi.org/10.1016/j.cell.2012.10.052>
83. Tanca A, Abbondio M, Palomba A, Fraumene C, Manghina V, Cucca F, et al. Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* [Internet]. 2017 Jul 14;5(1):79. Available from: <http://dx.doi.org/10.1186/s40168-017-0293-3>

## Supplementary figures

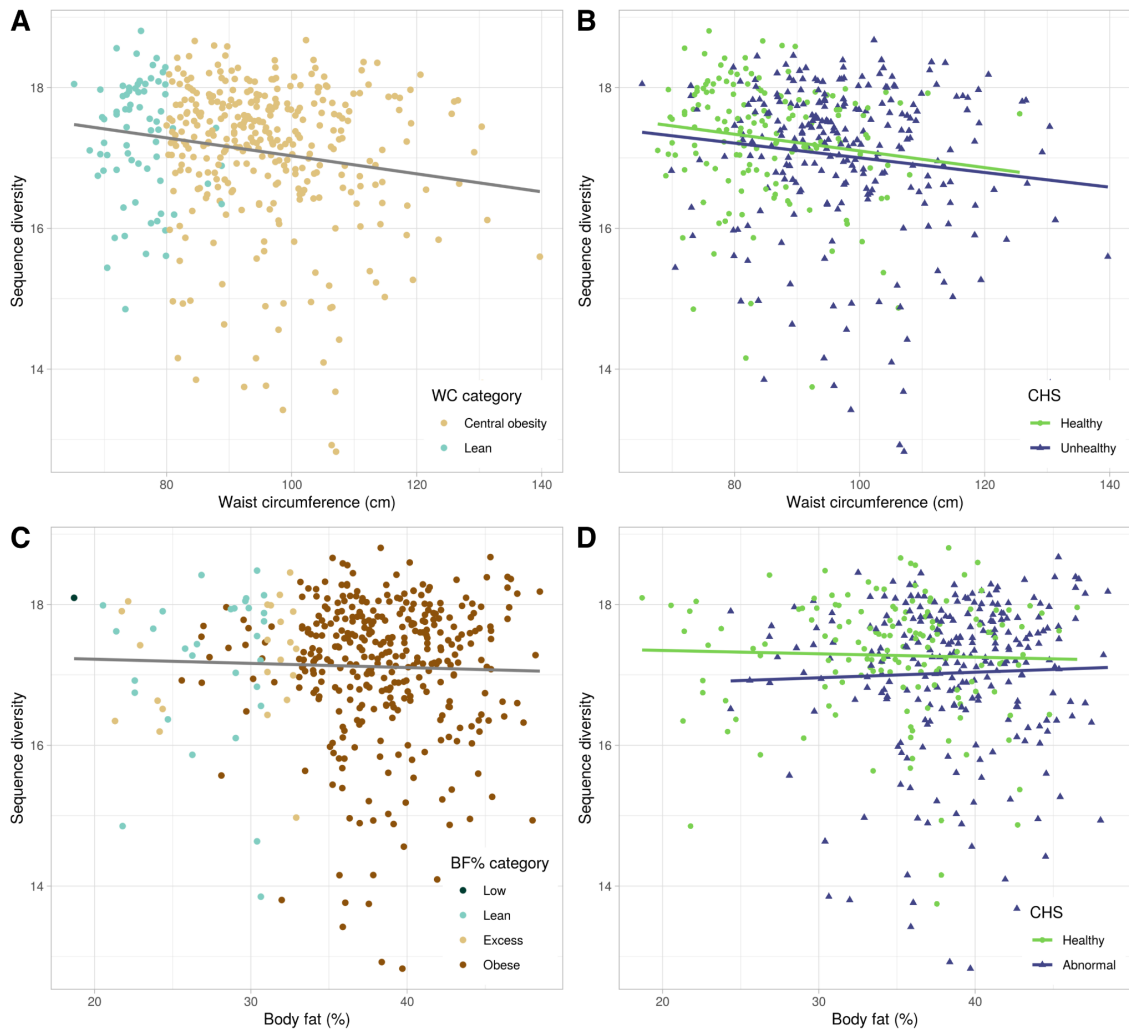


**Figure S1. Functional features included in the analyses were a subset of the detected features, and complete replication was not observed in our analyses.** A) Decision tree of candidate feature selection. KEGG modules and orthologs were included according to their association with cardiovascular disease, T2D or obesity in multiple studies. B) Alluvial plot indicating the selection of KEGG modules evaluated using BMI as OB measure, from detection and selection to significance.

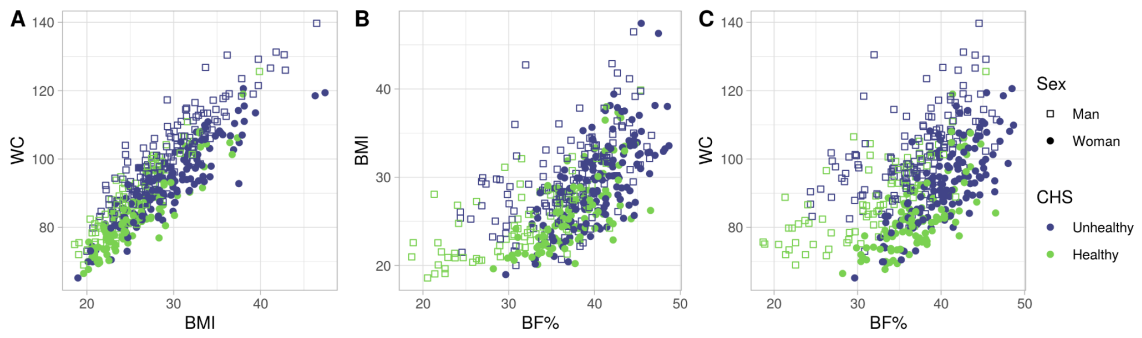


**Figure S2. Host covariables are associated with metagenome sequence richness.** A) UpSet plot of medication usage across all subjects. B) Sequence richness by number of medications consumed across all subjects ( $n = 408$ ). P values from ANOVA and the only significant contrast after Tukey's test. C) to F) Sequence richness by individual medication

usage after matching consumers with non-consumers by age, sex, city and BF% in a 1:3 ratio. C) proton pump inhibitors (n consumers, n non-consumers = 19, 57), D) hypertension (71, 213), E) dyslipidemia (41, 123), F) diabetes (19, 57). G) and H) Sequence richness by number socioeconomic level and city across all subjects. P values from ANOVA. I) Scatter plot of sequence richness by age across all subjects. Regression line, regression coefficient, P value and 95% confidence interval are shown. J) Sequence richness by sex across all subjects. Black points and bars represent mean and standard deviations.



**Figure S3. Patterns of metagenome sequence richness are conserved when (A) WC and (B) BF% are evaluated.** Scatter plots of sequence richness for both OB measures, regression lines are shown.



**Figure S4. Scatter plots showing the association between OB measures and CHS across all subjects (n = 408). Shapes indicate sex and color the CHS of the individuals.**

## Supplementary tables

Supplementary tables are available online as excel spreadsheets at <https://figshare.com/s/38d798ca39eb89e34750>. For table legends, see below.

### **Table S1. List of candidate KEGG modules used in downstream analyses.**

The final 136 modules were selected based on literature review and had a prevalence  $\geq 50\%$  in the Colombian cohort. Feature: KEGG ID. Annotation: feature description. Enrichment: condition in which the feature is enriched according to Jie et al. (2017) and Armour et al. (2019). (Controls: healthy subjects, T2D: type 2 diabetes, Shared: multiple conditions).

### **Table S2. List of candidate KEGG orthologs used in downstream analyses.**

The final 2 653 modules were selected based on literature review and had a prevalence  $\geq 50\%$  in the Colombian cohort. Feature: KEGG ID. Annotation: feature description. Enrichment: condition in which the feature is enriched according to Jie et al. (2017) and Wu et al. (2020). (Controls: healthy subjects, ACVD: atherosclerotic cardiovascular disease, T2D: type 2 diabetes, Shared: multiple conditions).

### **Table S3. KEGG modules uniquely associated with a given OB measure, CHS or shared by both conditions.**

Feature: KEGG ID. Annotation: feature description Health\_association: health association direction of feature. Main\_Effect\_1: Main effect (OB/CHS) associated with module abundance. value: reference level. lm\_coefficient: coefficient of linear model. raw\_P\_value: raw P value. Adj\_P\_value: P value adjusted for multiple comparisons. Geometric\_mean\_abundance: mean abundance in Colombian cohort Standard\_Deviation: standard deviation in Colombian cohort Prevalence: prevalence in Colombian cohort Main\_Effect\_2: Main effect (OB/CHS) adjusted model. Set: feature uniquely associated with OB, CHS or shared.

### **Table S4. KEGG orthologs uniquely associated with a given OB measure, CHS or shared by both conditions.**

Feature: KEGG ID. Annotation: feature

description Health\_association: health association direction of feature. Main\_Effect\_1: Main effect (OB/CHS) associated with module abundance. value: reference level. lm\_coefficient: coefficient of linear model. raw\_P\_value: raw P value. Adj\_P\_value: P value adjusted for multiple comparisons. Geometric\_mean\_abundance: mean abundance in Colombian cohort Standard\_Deviation: standard deviation in Colombian cohort Prevalence: prevalence in Colombian cohort Main\_Effect\_2: Main effect (OB/CHS) adjusted model. Set: feature uniquely associated with OB, CHS or shared.

**Table S5. Correlation blocks of KEGG modules and host parameters obtained using Hierarchical All-against-All Association testing (HAIA).**

Cluster\_rank: cluster order based on minimum adjusted P value. Cluster\_X: KEGG modules included in the association block. Cluster\_Y: host factors included in the association block. Best\_adjusted\_pvalue: minimum P value of all pairwise correlations of block.

**Table S6. Correlation blocks of KEGG orthologs and host parameters obtained using Hierarchical All-against-All Association testing (HAIA).**

Cluster\_rank: cluster order based on minimum adjusted P value. Cluster\_X: KEGG orthologs included in the association block. Cluster\_Y: host factors included in the association block. Best\_adjusted\_pvalue: minimum P value of all pairwise correlations of block.



Appendix V: Gut metagenomes and assembled microbial genomes  
from human adults from urban cohorts from Colombia, South  
America

## **Title**

Gut metagenomes and assembled microbial genomes from human adults from urban cohorts from Colombia, South America.

## **Authors**

Jacobo de la Cuesta-Zuluaga<sup>1</sup>, Nicholas D. Youngblut<sup>1</sup>, Juan S. Escobar<sup>2</sup>, Ruth E. Ley<sup>1#</sup>

<sup>1</sup>Department of Microbiome Science, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany.

<sup>2</sup>Vidarium–Nutrition, Health and Wellness Research Center, Grupo Empresarial Nutresa, 050023 Medellin, Colombia.

# correspondence: rley@tuebingen.mpg.de.

## **Abstract**

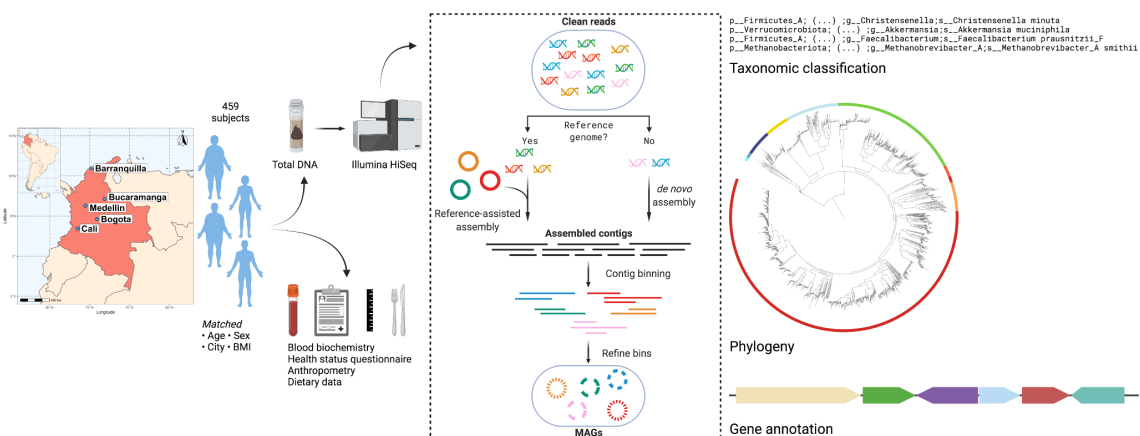
The human gut microbiome is an important mediator of multiple physiological processes. The identification of generalizable associations and mechanistic links between this microbial community and human health requires the study of diverse human populations. Yet the microbiomes of subjects from low- and middle-income countries are understudied. Here, we present a set of shotgun gut metagenomes of 459 deeply-phenotyped men and women (18-62 years old) living in geographically distinct urban areas of Colombia (South America), studied in the context of westernization and the epidemiological transition. We assembled these metagenomes and retrieved 2 266 medium- and high-quality metagenome-assembled genomes (MAGs), which we annotated, classified taxonomically, and compared to large collections of microbial genomes. The metagenomes, MAGs, and accompanying host data presented here will benefit initiatives looking into the human microbiome's

diversity and its role in westernization, nutrition, obesity and cardiometabolic disease.

## Background & Summary

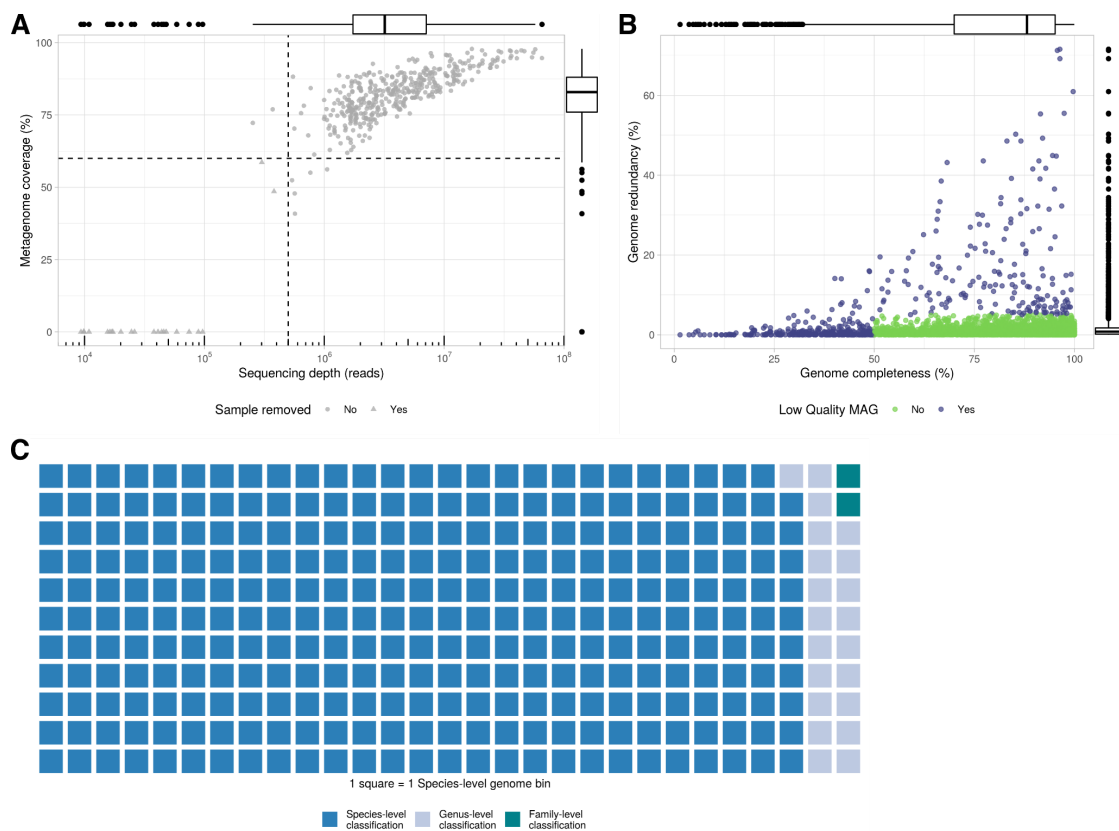
The study of the human gut microbiome has profoundly transformed the interpretation of multiple physiological processes, from health and disease to metabolism and nutrient absorption (1). In recent years, major advances in the establishment of large collections of microbial genomes have allowed us to gain insights into the taxonomic and functional repertoire of the intestinal microbial ecosystem (2).

However, the vast majority of studies from where these collections stem have been performed in subjects from industrialized countries such as the United States, China or members of the European Union (3,4). Large-scale studies from low- and middle-income countries that aimed to describe gut microbial diversity and its association with human health are sparse (4). The lack of studies in such populations makes it difficult to determine the generality of many of the previously reported links between the microbiome and the host. Therefore, calls for initiatives that encompass populations with socioeconomic and environmental factors beyond high-income countries have been made, so that a universal understanding of the human microbiome and its effect on host health can be achieved (3).



**Figure 1. Overview of the population, study design, metagenome assembly workflow and generated data sets.** Figure made with BioRender.

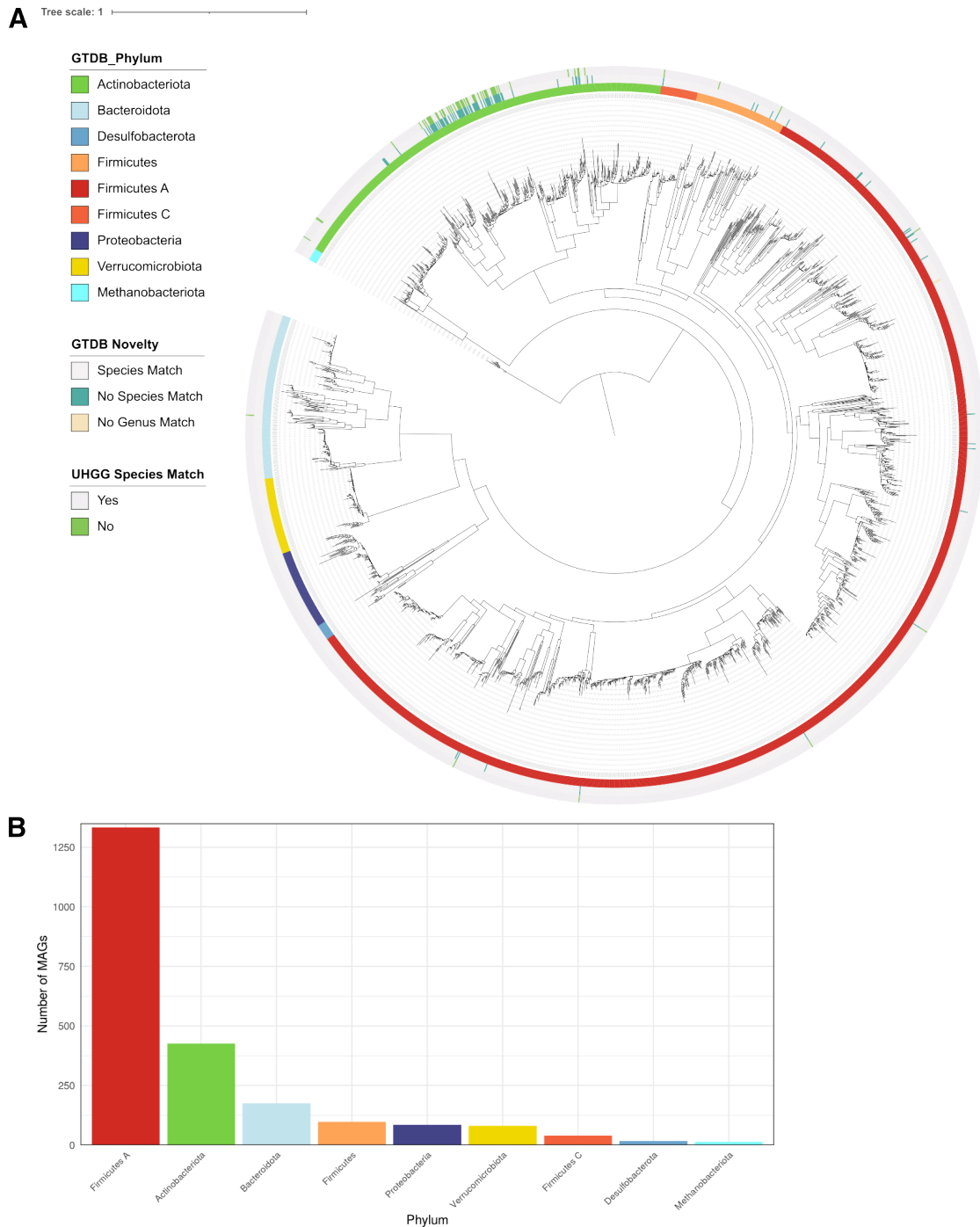
In the present Data Descriptor, we present a set of shotgun metagenome samples and metagenome-assembled genomes (MAGs) from the gut of a cohort of 459 community-dwelling human adults from five cities in Colombia, South America (figure 1). These subjects were sampled as part of a research project aimed at characterizing the gut microbiota of a population undergoing westernization, and to determine variation associated with obesity and cardiometabolic disease. This cohort has been previously studied using 16S rRNA gene sequencing in the context of the epidemiological transition (5), the association with cardiometabolic disease (6), the genetic ancestry (7,8), age and sex (9), and nutritional patterns of the host (10).



**Figure 2. Summary of shotgun metagenome sequencing and metagenome assembly from 408 successfully sequenced human gut samples.** A) Sequencing depth and metagenome coverage of metagenome samples; dashed lines represent inclusion thresholds for each parameter. B) Completeness and redundancy of each of the obtained MAGs. Green points correspond to the 2 266 MAGs above quality cut-off values. C) Waffle plot of taxonomic classification of the 355 SGBs, colored according to the lowest taxonomic level assigned.

Each subject provided a stool sample, from which we extracted total DNA. Shotgun metagenome sequencing was performed using the Illumina HiSeq 3000 platform. We retained 408 samples which had a sequencing depth of  $>5.0 \times 10^5$  reads (mean  $\pm$  SD 6 719 985 reads/sample  $\pm$  8 960 996) or a metagenome coverage calculated using nonpareil  $>60$  % (mean  $\pm$  SD 82.36 %  $\pm$  8.38) (figure 2A). We performed metagenome assembly and binning on a per sample basis and retrieved 2 797 MAGs. After removal of low-quality genomes (completeness  $< 50$  % or contamination  $\geq 5$  %) and dereplication at 99.9 % average nucleotide identity (ANI), we retained 2 266 MAGs, with an average ( $\pm$  standard deviation) completeness and contamination of 85.57 %  $\pm$  13.11 and 1.07 %  $\pm$  1.04, respectively (figure 2B). Further dereplication at 95 % ANI resulted in 358 species-level genome bins (SGBs). Downstream analyses of MAGs usually involve the inference of their phylogenetic relations and the prediction of the functional information encoded in the genome. We performed gene calling and annotation using Prokka. The predicted proteome was used to infer a multilocus phylogeny using PhyloPhlAn.

We used release 202 of the Genome Taxonomy Database (GTDB) to assign a taxonomic classification to each genome. The set of SGBs comprised 296 known species belonging to 172 genera. A total of 57 SGBs (15.92 %) lacked a species-level classification and 2 SGBs (0.56 %) lacked genus classification (figure 2C). We compared our MAGs to the set of representative genomes of the Unified Human Gut Genomes catalog (UHGG v.1.0). The vast majority of MAGs corresponded to novel strains of the identified species: 2 132 MAGs had ANI values between 95 % and 99 % to their closest match in the UHGG. Remarkably, we found that 62 (2.74 %) of our MAGs did not have a species-level match in the UHGG. Mapping the taxonomic novelty onto the MAG phylogeny revealed that it tended to group within a few clades. Most novel MAGs belonged to the orders *Coriobacteriales* (72 MAGs), *Oscillospirales* (10 MAGs) and *Christensenellales* (8 MAGs) (figure 3A).



**Figure 3.** Phylogeny and taxonomy of the retrieved 2 266 MAGs. A) Phylogenetic tree built from multilocus sequence alignment. From innermost to outermost rings, the data mapped onto the phylogeny are: GTDB r202 phylum-level taxonomic classification; taxonomic novelty of the MAGs compared to GTDB at the species, genus or family level; presence of a genome with ANI < 95% at the UHGG catalog v.1.0. The phylogeny was inferred from multiple conserved loci using PhyloPhlAn. The phylogeny is rooted on the last common ancestor of Archaea and Bacteria. Scale bar represents the number of amino acid substitutions per site. B) Barplot showing the number of MAGs classified on each bacterial and archaeal phylum. Colors same as in A).

## **Methods**

### ***Ethics approval***

This cross-sectional study was conducted in accordance with the principles of the Declaration of Helsinki 2013 and had minimal risk according to the Colombian Ministry of Health (Resolutions 8430 of 1993 and 2378 of 2008). All the participants were thoroughly informed about the study and procedures before signing consent forms. Participants were assured of anonymity and confidentiality. Written informed consent was obtained from all the participants before beginning the study. The Bioethics Committee of SIU—University of Antioquia (Medellin, Colombia) reviewed the protocol and the consent forms and approved the procedures described here (approbation act 14-24-588 dated 28 May 2014).

### ***Study population***

As part of a cross-sectional study aiming to characterize the gut microbiota of Colombians, we enrolled a cohort of 459 community-dwelling men and women, aged 18 to 62, from five Colombian cities, with body mass index (BMI)  $\geq 18.5$  kg/m<sup>2</sup> and without report of cancer, neurodegenerative or gastrointestinal disease. For each participant, total DNA extracted from fecal samples, as well as anthropometric, socioeconomic, biochemical and dietary information were collected. This information included personal medical history and medication use; dietary intake and physical activity levels; blood pressure and several measures of adiposity such as BMI, waist circumference and percentage of body fat; blood biochemistry parameters, namely, high density lipoprotein (HDL) cholesterol, low density lipoprotein (LDL) cholesterol, very low density lipoprotein (VLDL) cholesterol, triglycerides, fasting glucose and insulin, glycated hemoglobin (HbA1c), leptin, adiponectin and high-sensitive C-reactive protein. A detailed description of data acquisition can be found elsewhere (6).

### ***DNA extraction and sequencing***

Each participant performed a stool sample collection. We extracted total DNA from 430 fecal samples utilizing the QIAamp DNA Stool Mini Kit (Qiagen).

We prepared metagenome libraries with a modified Nextera protocol, as described elsewhere (11). Briefly, we used 1 ng of total stool DNA for Nextera Tn5 tagmentation. After purification with Agencourt AMPure XP beads (Beckman Coulter), samples were normalized and pooled. Next, we performed size selection of the pooled samples using BluePippin (Sage Sciences) to restrict fragment sizes to 400 to 700 bp. Barcoded pools were sequenced using the Illumina HiSeq 3000 platform with 2x150 bp paired-end sequencing.

### ***Sequence quality control***

Raw sequencing reads were validated using fqtools v.2.0 (12), and the “clumpify” command of bbtools v37.78 (<https://jgi.doe.gov/data-and-tools/bbtools/>) was used to deduplicate them. We removed adapters and performed read quality control with skewer v0.2.2 (13) and the “bbduk” command of bbtools. To remove human genome reads, we mapped them to the hg19 assembly using the “bbmap” command of bbtools. After each step, we generated data quality reports using fastqc v0.11.7 (<https://github.com/s-andrews/FastQC>) and multiQC v1.5a (14). To estimate the metagenome coverage, we used Nonpareil v.3.3.4. A total of 408 samples that had a sequencing depth above 1 million reads or a metagenome coverage over 60 % were retained for downstream analyses.

### ***Metagenome assembly***

We performed metagenome assembly following the workflow developed by Youngblut et al. (15). Briefly, each sample was assembled separately, prior subsampling to a maximum of 20 million reads per sample using seqtk v.1.3. A reference-based metagenome assembly was performed using metacompass v.1.2 (16) based on each sample’s taxonomic profile. We profiled each sample using centrifuge v.1.0.3 (17) to select reference genomes, which we then downloaded using ncbi-genome-download v.0.2.1. Reads that did not map to any reference genome were used for de-novo assembly with metaSPAdes v.3.12.0 (18). Contigs with a minimum length of 2 000 bp from the reference-based and *de novo* assemblies were combined and de-replicated



using bbttools. To bin contigs, we used MaxBin2 v.2.2.4 and MetaBAT2 v.2.12.1, each executed with 2 parameter settings for a total of 4 bin collections per sample. Per-sample binning of contigs was performed, but we utilized reads from all metagenome samples to calculate differential coverage with Bowtie2 v.2.3.5 (19). The best non-redundant set of contig bins (i.e. metagenome-assembled genomes; MAGs) was selected with DAS-Tool v.1.1.1 (20) based on estimates of completeness and contamination from CheckM. For downstream analyses, MAGs from all samples were combined.

We calculated completeness and contamination of each MAG using CheckM v.1.0.13. MAGs with completeness of <50 % or contamination of  $\geq 5$  % were discarded. Taxonomic classification of the MAGs was obtained using GTDB-Tk v.0.3.3 (21) against the release 202 of the Genome Taxonomy Database (GTDB). We used dRep to collapse clonal genomes at an average nucleotide identity (ANI) of 99.9 %. Using PhyloPhlAn v.0.41 (22), we constructed a maximum-likelihood phylogenomic tree of all de-replicated MAGs from a concatenated alignment of multiple universally distributed single copy marker genes. We assessed the taxonomic novelty of the non-redundant MAGs against the Unified Human Gut Genomes catalog v.1.0 (UHGG) (2) by calculating the Average Nucleotide Identity (ANI) between each of our MAGs to each of the 4 644 representative genomes of the UHGG using FastANI v.1.31 (23). We considered a MAG to be novel if it did not have a species level match in the UHGG (< 95 % ANI). Gene calling, proteome prediction and annotation was performed on each genome using Prokka 1.12 (24).

## **Technical validation**

We processed stool samples using sterilized equipment and standard laboratory procedures following manufacturer instructions. The quality of the extracted DNA was measured prior to the construction of metagenomic libraries. Negative extraction and library construction controls were sequenced; a mock community was included as positive sequencing control. We processed raw sequencing data to remove host and poor quality reads. We restricted the metagenome assembly only to samples with adequate sequencing coverage

and depth. Quality of the MAGs was assessed using CheckM; redundant and low-quality MAGs were removed from the data set. Whenever possible, we used up-to-date versions of the databases and software for metagenome assembly and MAG characterization.

## **Usage notes**

Initiatives to expand the understanding of human-associated microbial diversity across populations benefit from well-annotated, accessible sequences that are rich in host data, such as those presented here. This Data Descriptor is useful to researchers studying the diversity of the human microbiome and the role it plays in westernization, nutrition, obesity and cardiometabolic disease. As such, we expect the bulk of the data we generated to contribute to large cataloguing efforts, such as the GTDB or the UHGG. Likewise, selected genomes could be utilized to broaden the comprehension of particular clades by comparative genomics (25).

## **Data records**

The annotated MAGs, the taxonomic classification tables and phylogenetic tree, as well as summaries of the data described here will be released upon submission of this manuscript to a preprint server and/or a peer-reviewed journal. The raw metagenome sequence data, and the 2 266 non-redundant MAGs, will be submitted to the European Nucleotide Archive.

Host anthropometric, biochemical and dietary data that has been made available as part of previously published works can be found at:

<https://github.com/jsescobar/westernization>, <https://github.com/jsescobar/bsp>  
and [https://github.com/Vidarium/diet\\_microbiota\\_MiSalud1.0](https://github.com/Vidarium/diet_microbiota_MiSalud1.0)

## **Code availability**

The code used for processing the data will be made available at [https://github.com/leylabmpi/Colombian\\_MAGs](https://github.com/leylabmpi/Colombian_MAGs)

## Author contributions

Sample processing and metagenome sequencing: JdlCZ. Sequence data processing, metagenome assembly, genome quality assessment and taxonomic characterization: JdlCZ. Bioinformatics pipelines development and support: NDY. Project supervision and discussion of results: REL, JSE, NDY. Manuscript preparation: JdlC. Manuscript review: REL, JSE, NDY. Comments: all authors.

## Competing interests

While engaged in the research project, JSE was employed by a food company. JdlCZ, NDY and REL declare no competing interests.

## Acknowledgements

This work was supported by the Max Planck Society and Vidarium–Nutrition, Health and Wellness Research Center. We thank the participants who took part in the study, and the Vidarium, EPS SURA and Dinámica IPS staff that helped with recruitment and field work. Some authors of this work collaborate through the Microbiome & Health Network. We are grateful to Liam Fitzstevens, Taichi Suzuki and Laura Salazar-Jaramillo for the fruitful discussions and comments.

## References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* [Internet]. 2012 Jun 13;486(7402):207–14. Available from: <http://dx.doi.org/10.1038/nature11234>
2. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* [Internet]. 2020 Jul 20;490:55. Available from: <http://dx.doi.org/10.1038/s41587-020-0603-3>
3. Porras AM, Brito IL. The internationalization of human microbiome research. *Curr Opin Microbiol* [Internet]. 2019 Aug;50:50–5. Available from: <http://dx.doi.org/10.1016/j.mib.2019.09.012>
4. Abdill RJ, Adamowicz EM, Blekhman R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol* [Internet]. 2022

Feb;20(2):e3001536. Available from:  
<http://dx.doi.org/10.1371/journal.pbio.3001536>

5. de la Cuesta-Zuluaga J, Corrales-Agudelo V, Velásquez-Mejía EP, Carmona JA, Abad JM, Escobar JS. Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. *Sci Rep* [Internet]. 2018 Jul 27;8(1):11356. Available from: <http://dx.doi.org/10.1038/s41598-018-29687-x>
6. de la Cuesta-Zuluaga J, Corrales-Agudelo V, Carmona JA, Abad JM, Escobar JS. Body size phenotypes comprehensively assess cardiometabolic risk and refine the association between obesity and gut microbiota. *Int J Obes* [Internet]. 2018 Mar;42(3):424–32. Available from: <http://dx.doi.org/10.1038/ijo.2017.281>
7. Ortega-Vega EL, Guzmán-Castañeda SJ, Campo O, Velásquez-Mejía EP, de la Cuesta-Zuluaga J, Bedoya G, et al. Variants in genes of innate immunity, appetite control and energy metabolism are associated with host cardiometabolic health and gut microbiota composition. *Gut Microbes* [Internet]. 2020 May 3;11(3):556–68. Available from: <http://dx.doi.org/10.1080/19490976.2019.1619440>
8. Guzmán-Castañeda SJ, Ortega-Vega EL, de la Cuesta-Zuluaga J, Velásquez-Mejía EP, Rojas W, Bedoya G, et al. Gut microbiota composition explains more variance in the host cardiometabolic risk than genetic ancestry. *Gut Microbes* [Internet]. 2020;11(2):191–204. Available from: <http://dx.doi.org/10.1080/19490976.2019.1634416>
9. de la Cuesta-Zuluaga J, Kelley ST, Chen Y, Escobar JS, Mueller NT, Ley RE, et al. Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults. *mSystems* [Internet]. 2019 Jul;4(4). Available from: <http://dx.doi.org/10.1128/mSystems.00261-19>
10. García-Vega ÁS, Corrales-Agudelo V, Reyes A, Escobar JS. Diet Quality, Food Groups and Nutrients Associated with the Gut Microbiota in a Nonwestern Population. *Nutrients* [Internet]. 2020 Sep 25;12(10). Available from: <http://dx.doi.org/10.3390/nu12102938>
11. Karasov TL, Almario J, Friedemann C, Ding W, Giolai M, Heavens D, et al. *Arabidopsis thaliana* and *Pseudomonas* Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host Microbe* [Internet]. 2018 Jul 11;24(1):168–79.e4. Available from: <http://dx.doi.org/10.1016/j.chom.2018.06.011>
12. Droop AP. fqtools: an efficient software suite for modern FASTQ file manipulation. *Bioinformatics* [Internet]. 2016 Jun 15;32(12):1883–4. Available from: <http://dx.doi.org/10.1093/bioinformatics/btw088>
13. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC*

- Bioinformatics [Internet]. 2014 Jun 12;15:182. Available from: <http://dx.doi.org/10.1186/1471-2105-15-182>
14. Ewels P, Magnusson M, Lundin S, Källér M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* [Internet]. 2016 Oct 1;32(19):3047–8. Available from: <http://dx.doi.org/10.1093/bioinformatics/btw354>
  15. Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, Walzer C, et al. Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. *mSystems* [Internet]. 2020 Nov 3;5(6). Available from: <http://dx.doi.org/10.1128/mSystems.01045-20>
  16. Cepeda V, Liu B, Almeida M, Hill CM, Koren S, Treangen TJ, et al. MetaCompass: Reference-guided Assembly of Metagenomes [Internet]. *bioRxiv*. 2017 [cited 2020 Jan 10]. p. 212506. Available from: <https://www.biorxiv.org/content/10.1101/212506v1>
  17. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* [Internet]. 2016 Dec;26(12):1721–9. Available from: <http://dx.doi.org/10.1101/gr.210641.116>
  18. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* [Internet]. 2017 May;27(5):824–34. Available from: <http://dx.doi.org/10.1101/gr.213959.116>
  19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* [Internet]. 2012 Mar 4;9(4):357–9. Available from: <http://dx.doi.org/10.1038/nmeth.1923>
  20. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* [Internet]. 2018 Jul;3(7):836–43. Available from: <http://dx.doi.org/10.1038/s41564-018-0171-1>
  21. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* [Internet]. 2019 Nov 15; Available from: <http://dx.doi.org/10.1093/bioinformatics/btz848>
  22. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* [Internet]. 2013;4:2304. Available from: <http://dx.doi.org/10.1038/ncomms3304>
  23. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* [Internet]. 2018 Nov 30;9(1):5114. Available from: <http://dx.doi.org/10.1038/s41467-018-07641-9>

24. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* [Internet]. 2014 Jul 15;30(14):2068–9. Available from: <http://dx.doi.org/10.1093/bioinformatics/btu153>
25. de la Cuesta-Zuluaga J, Spector TD, Youngblut ND, Ley RE. Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut. *mSystems* [Internet]. 2021 Feb 9;6(1). Available from: <http://dx.doi.org/10.1128/mSystems.00939-20>