


ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

INTEGRATION OF FUNGAL AND BACTERIAL MICROBIOME SEQUENCE DATA

A doctoral thesis presented by
Zixuan Xie to aim for the degree of Doctor

Supervisor: Dr. Chaysavanh Manichanh

Tutor: Dr. Victor Manuel Vargas Blasco

**Doctoral Program in Medicine
Department of Medicine
Barcelona, 2023**



**Universitat Autònoma
de Barcelona**

*“Always go too far,
because that's where
you will find the truth.”*

— Albert Camus.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everyone who has played a role in making my PhD journey a memorable and enriching experience.

First and foremost, I am incredibly grateful to my supervisor, Dr. Chaysavanh Manichanh, for her unwavering support, patience, and guidance throughout my research. Her dedication to academic excellence and her constant motivation and encouragement have been invaluable to me. I could not have imagined a better mentor for my PhD study.

I would also like to extend my sincere thanks to my colleagues: Andrea Elias Vidal, Andrea Martinez, Blanca Ruiz de la Torres Reyes, Francisca Yanez, Gerard Serrano Gomez, Guillaume Sarrabayrouse, Iñigo Oyarzun, Leidy Alejandra-Gonzales Molano, Luis Mayorga, Marc Pons, who have been an integral part of my journey. Their collaboration, support, and camaraderie have made my research experience a truly fulfilling one. I would like to express my special gratitude to Aleix Canalda-Baltrons for the contribution in estimating ITS copy numbers, Encarna Varela Castro and Zaida Soler Luque for their help on all the PCR and DNA extraction experiments, Dr. Xavier Martínez and Sara Vega-Abellaneda for their help in construction of the MycoDM web server.

Furthermore, I would like to acknowledge the invaluable support and encouragement provided by the members of the FunHoMic consortium. I am grateful to the principal investigators: Alan Walker, Alister J P Brown, Bernard Hube, Carol Munro, Ilse Jacobsen, Karine Roget, Karla Queiroz, Marie-Elisabeth Bougnoux, Mihai Netea, Peter Warn, Salome Leibundgut-Landmann, Vincent Thomas, and the FunHomies: Ann-kristin Kaune, Benoit Marsaux, Daria Kosmala, Diletta Rosati, Leovigildo-Rey Alaban, Manjyot Kaur, Margot Delavy, Marisa Valentine, Moran Morelli, Nathaniel Cole, Sayoni Chakraborty, Ricardo Martins, for their contributions to my research. In particular, I am grateful to Dr. Christophe d'Enfert for the enlightened suggestions and help in estimating the fungal ITS copy

numbers.

Finally, I would like to thank my parents, Xiaohong Zhu and Xingmin Xie, for bringing me to this beautiful world and providing me their unwavering spiritual and material support, love, and encouragement throughout my life. I am also grateful to my boyfriend, Jujian Huang, for his love, accompany and support during all the challenging times.

To all those who have helped me during my PhD journey, I express my sincere gratitude.

ABBREVIATIONS

ITS: Internal transcribed spacer
SCMG: Single copy marker gene
BE: Barrett's esophagus
GERD: gastroesophageal reflux disease
SCFAs: short-chain fatty acids
MUC2 : glycoprotein Mucin 2
AMPs: antimicrobial peptides
IECs: intestinal epithelial cells
IBD: inflammatory bowel diseases
CSTs: community state types
H₂O₂: hydrogen peroxide
EoE: eosinophilic esophagitis
DFUs: diabetic foot ulcers
VVC: vulvovaginal candidiasis
IUA: Intrauterine adhesion
RVC: recurrent vaginal candidiasis
HMMs: hidden Markov models
EC: enzyme commission
UHGG: Unified Human Gastrointestinal Genome
CD: Crohn's disease
UC: ulcerative colitis
ESRD: end-stage renal disease
T1D: type 1 diabetes
T2D: type 2 diabetes
MAGs: metagenome-assembled genomes
MSA: multiple sequence alignment
FDR: false discovery rate
KEGG: Kyoto Encyclopedia of Genes and Genomes
RMT: random matrix theory

INDEX

ABSTRACT.....	17
RESUMEN.....	21
1. INTRODUCTION	25
1.1 Human microbiome	27
1.1.1 Definition.....	27
1.1.2 Bacterial community in the human microbiome	28
1.1.3. Mycobiome	37
1.1.4 Interaction between the fungal and bacterial microbiome.....	45
1.2. Metagenomics in studying the microbiome.....	46
1.2.1 Amplicon metagenomic sequencing	46
1.2.2 Shotgun metagenomic sequencing.....	47
1.2.3 Databases and pipelines for the human mycobiome profiling when applying shotgun metagenomic sequencing.....	48
2. HYPOTHESIS.....	51
Hypothesis	52
3. OBJECTIVES.....	53
3.1 Main objective	55
3.2 Secondary objectives	55
4. METHODS	57
4.1 Construction of the FunOMIC databases and pipeline for human mycobiome profiling using shotgun metagenomics	59
4.1.1Collection of fungal genomes.....	59
4.1.2 Construction of the taxonomic and functional FunOMIC database	59
4.1.3 Validation of the FunOMIC databases and the pipeline.....	60
4.1.4 Collection of metagenomic data.....	61

4.1.5	Aligning human metagenomic sequencing reads onto the FunOMIC database	62
4.1.6	Prokaryotic taxonomic and functional profilings of human metagenomic data.....	62
4.1.7	Statistical Analysis	62
4.2	Update of the FunOMIC and establishment of the MycoDM web server..	63
4.2.1	DATABASES CONTENT AND CONSTRUCTION.....	63
4.2.2	Design and construction of the Web server	65
4.3	Robust integration of fungal and bacterial gut microbiome with dietary data in a longitudinal setting.....	66
4.3.1	Fungal genomes collection	66
4.3.2	Estimation of the ITS copy numbers	67
4.3.3	<i>In silico</i> comparison of ITS and shotgun methods	69
4.3.4	Fungal enrichment protocol	72
4.3.5	Collection and processing of habitual diet information.....	73
4.3.6	Sample collection and DNA extraction.....	73
4.3.7	Shotgun metagenomic sequencing and profiling	74
4.3.8	Keystone species analysis.....	74
4.3.9	Statistical analysis	75
5.	RESULTS	77
5.1	Validation and application of the FunOMIC databases and pipeline	79
5.1.1	Characteristics of the taxonomic and functional FunOMIC database	79
5.1.2	Characteristics of the 2,679 metagenomes.....	83
5.1.3	Fungal community structure, diversity, and functions of the 1950 metagenomes.....	83
5.1.4	Association between metadata and mycobiome composition and	

functions	86
5.1.5 Core taxonomic fungal microbiomes of different body sites and different countries	88
5.1.6 Bacterial and fungal microbiome interaction	89
5.2 Description and usage of MycoDM web server	92
5.2.1 Home page	92
5.2.2 Download page	93
5.2.3 Analysis platform	94
5.2.4 MycobialMarkers page.....	95
5.3 Robust integration of fungal and bacterial gut microbiome with dietary data in a longitudinal setting.....	97
5.3.1 Shotgun metagenomics sequencing provides higher accuracy than ITS amplicon sequencing in mycobiome profiling at the species level	97
5.3.2 A fungal enrichment protocol effectively concentrates fungal cells in human fecal samples.....	101
5.3.3 Keystone bacterial and fungal species in the human gut.....	106
5.3.4 Short-term dynamics of the human gut microbiome	109
5.3.5 Microbial diversity and composition are associated with habitual diet	111
6. DISCUSSION.....	115
7. CONCLUSION.....	129
8. FUTURE LINES.....	133
9. BIBLIOGRAPHY	137
10. ANNEXES.....	159
ANNEX 1. Link to all the supplementary tables.....	161
ANNEX 2. Fungal species and pathway classes in different groups of	

mycobiomes.	162
ANNEX3. Interaction of fungal and bacterial functions in the gut. microbiome of healthy individuals.	164
ANNEX4. Enrichment efficiency in bacteria.	165
ANNEX5. Bacterial taxonomic compositions are associated with habitual diet.	167
ANNEX6. Fungal functional compositions are associated with habitual diet.	169
ANNEX7. Bacterial functional compositions are associated with habitual diet.	171
ANNEX8. Publication related to this thesis: Xie Z, Manichanh C. FunOMIC: Pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling. Comput Struct Biotechnol J. 2022;20:3685-94.	173

ABSTRACT

The analysis of the bacterial microbiome has become routine, but the study of the fungal microbiome, or mycobiome, is still hindered by a lack of robust databases and bioinformatic pipelines. To address this challenge, we developed FunOMIC and its updated version, a pipeline with built-in taxonomic and functional databases for identifying fungi from the human microbiome using shotgun sequencing. The pipeline includes raw sequence quality control, removal of human and bacterial DNA, and comprehensive taxonomic and functional mycobiome profiling. We validated the pipeline using *in silico*-generated mock communities and over 2,600 real human metagenomic samples. Our findings show that shotgun sequencing combined with FunOMIC outperforms the commonly used internal transcribed spacer (ITS) sequencing in terms of accuracy and cost-effectiveness. We proposed the application of shotgun sequencing with a new enrichment protocol to provide a cost-effective approach to perform gut mycobiome profiling at the species level.

We also investigated the relationship between microbial diversity, composition, and functions with habitual diet composition. Our study showed that microbial diversity and composition were associated with specific diet composition instead of driven by global dietary changes.

Furthermore, we proposed a web server called MycoDM, which provides searching of mycobial markers, online data analysis, and visualization platform to investigate the relationship of human gut mycobiome with various diseases using shotgun metagenomic data. This platform will help researchers study the role of the fungal community associated with disease, which is still unclear. But growing evidence suggests that mycobiome dysbiosis can be related to various conditions and human immune function and metabolism malfunction.

Our work provides a comprehensive description of the inter-kingdom interaction between bacteria and fungi integrating dietary data. We believe that our proposed workflow will be a valuable resource for mycobiome studies.

RESUMEN

El análisis del microbioma bacteriano se ha vuelto rutinario, pero el estudio del microbioma fúngico, o micobioma, aún se ve obstaculizado por la falta de bases de datos y pipelines bioinformáticos robustos. Para abordar este desafío, desarrollamos FunOMIC y su versión actualizada, un pipeline con bases de datos taxonómicas y funcionales integradas para identificar hongos del microbioma humano utilizando secuenciación shotgun. El pipeline incluye control de calidad de secuencias en bruto, eliminación de ADN humano y bacteriano, y perfilado taxonómico y funcional completo del micobioma. Validamos el pipeline utilizando comunidades simuladas in silico y más de 2.600 muestras de metagenómica humana reales. Nuestros hallazgos muestran que la secuenciación shotgun combinada con FunOMIC supera a la secuenciación del espacio transrito interno (ITS) comúnmente utilizada en términos de precisión y rentabilidad. Propusimos la aplicación de la secuenciación shotgun con un nuevo protocolo de enriquecimiento para proporcionar un enfoque rentable para realizar el perfilado del micobioma intestinal a nivel de especie.

También investigamos la relación entre la diversidad, composición y funciones microbianas con la composición habitual de la dieta. Nuestro estudio mostró que la diversidad y composición microbiana se asociaban con una composición de dieta específica en lugar de estar impulsadas por cambios dietéticos globales.

Además, propusimos un servidor web llamado MycoDM, que proporciona la búsqueda de marcadores micobiales, análisis de datos en línea y plataforma de visualización para investigar la relación del micobioma humano con diversas enfermedades utilizando datos de metagenómica shotgun. Esta plataforma ayudará a los investigadores a estudiar el papel de la comunidad fúngica asociada con la enfermedad, que aún no está claro. Pero hay evidencia creciente que sugiere que la disbiosis del micobioma puede estar relacionada con diversas condiciones y el mal funcionamiento del sistema inmunológico y del metabolismo humano.

Nuestro trabajo proporciona una descripción integral de la interacción entre los reinos bacteriano y fúngico, integrando datos dietéticos. Creemos que nuestro flujo de trabajo propuesto será un recurso valioso para los estudios de microbioma.

1. INTRODUCTION

1.1 Human microbiome

1.1.1 Definition

The term “microbiome” was first introduced by Legerberg and McCray in 2001, to describe the ecological community of all the microorganisms that inhabit a specific niche, such as the human body. Since then, during the past decades, this word has been mostly defined as the collection of genes and genomes of members of the assemblage of microorganisms present in a defined environment.

Recent discussions have argued that the term “biome” refers to the combination of both biotic and abiotic members of a habitat; thus, the surrounding conditions of the microorganisms should also be taken into account (1). In 2020, Berg et al. proposed a more comprehensive definition of the human microbiome and microbiota (1).

“The human microbiome is defined as a characteristic microbial community occupying different sites of human bodies. The microbiome not only refers to the microorganisms involved but also encompasses their theatre of activity, which results in the formation of specific niches. The microbiome, which forms a dynamic and interactive micro-ecosystem prone to change in time and scale, is integrated into macro-ecosystems, including the hosts, and hence crucial for their functioning and health.

The human microbiota consists of the assembly of microorganisms belonging to different kingdoms (Prokaryotes [Bacteria, Archaea], Eukaryotes [e.g., Protozoa, Fungi, and Algae]), while “their theatre of activity” includes microbial structures, metabolites, mobile genetic elements (e.g., transposons, phages, and viruses), and relic DNA embedded in the human bodies.” (Fig. 1) (1)

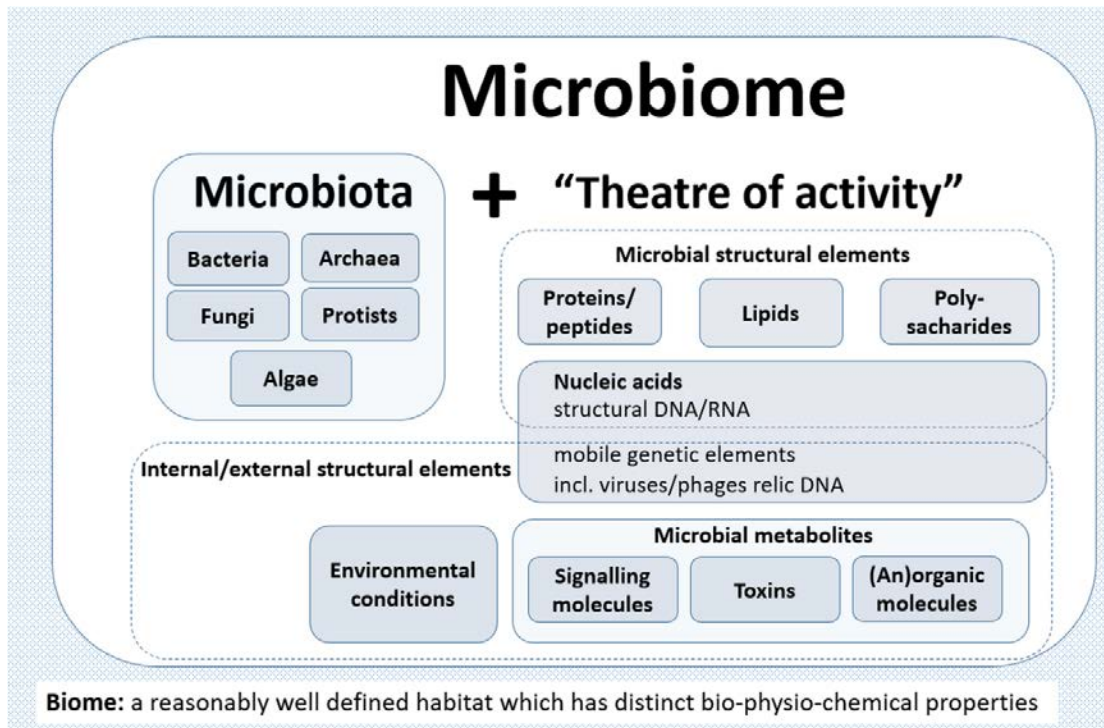


Figure1. Schematic definition of the microbiome.

1.1.2 Bacterial community in the human microbiome

As defined above, the human microbiota contains microbes from different kingdoms – Archaea, Bacteria, Fungi, Protozoa, etc. The size of the bacterial community overwhelmingly outnumbers other microbial members (2, 3). The total wet weight of the bacteria in the human body is about 0.2kg, and the ratio of bacterial cells to human cells is around 1.3 - 2.3 depending on different gender and life circumstances (4). It was estimated that the vast majority of the bacterial community inhabits the large intestine or colon, with a proximate order of magnitude of 10^{14} bacterial cells, followed by dental plaque, which is around 10^{12} bacteria. The bacterial communities present in the ileum, saliva, and skin are approximately bound by the same order of magnitude, which is 10^{11} . The stomach and the upper small intestine (duodenum and jejunum) harbor the least number of bacteria, with an order of magnitude around 10^7 (1, 4, 5, 6, 7, 8).

As the physiological condition varies widely in different body sites, it is intuitively that the compositions of the bacterial microbiome also differ greatly.

1.1.2.1 Gastrointestinal (GI) tract

The compartments of the human GI tract, including the oral cavity, esophagus, stomach, duodenum, jejunum, cecum, colon, and rectum, have variable physiology, therefore, it is not surprising that the GI tract consists of a heterogeneous collection of distinct habitats (9, 10).

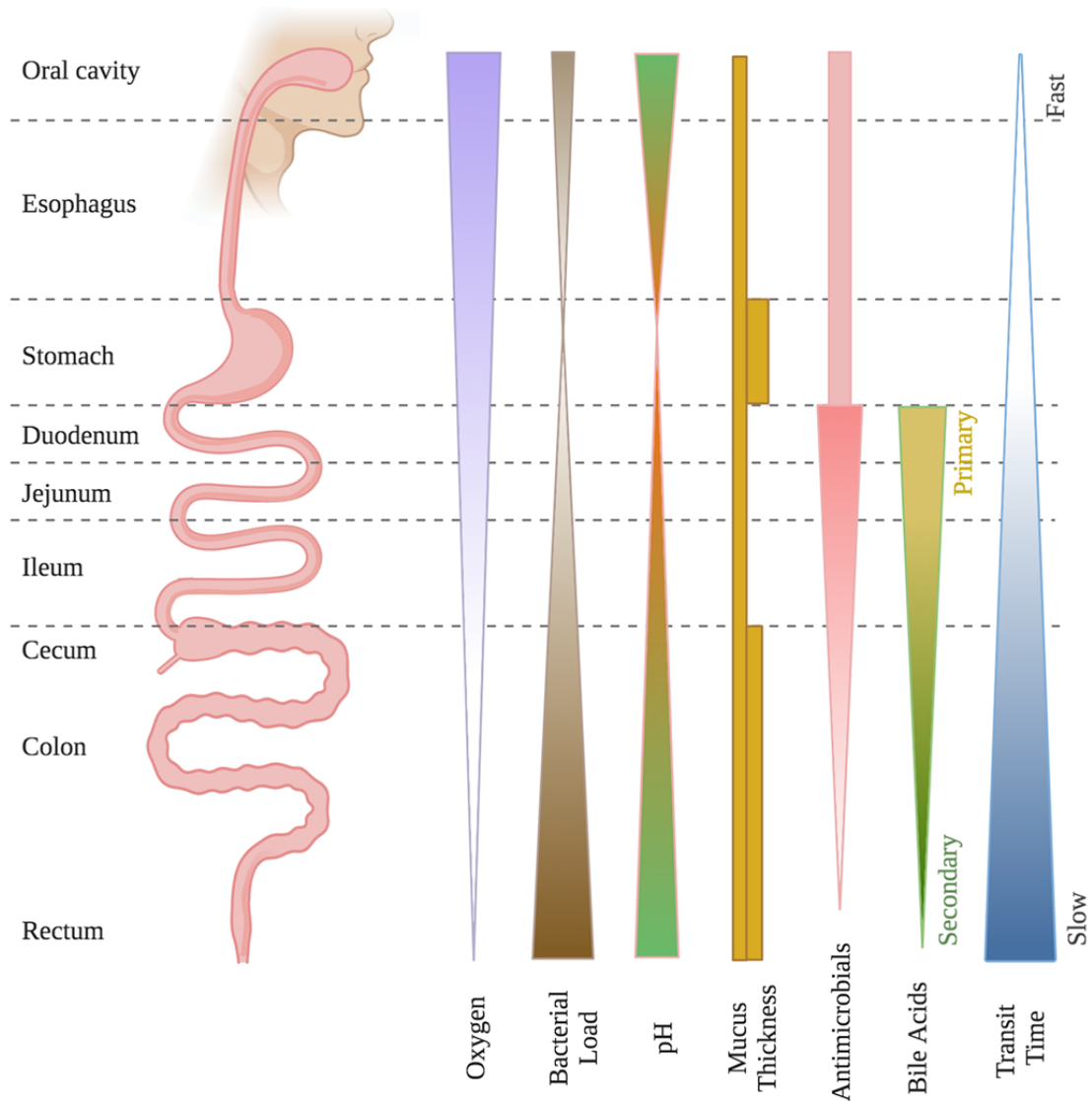


Figure 2. Compartments of the GI tract along the rostral-caudal axis. (10)

The **oral cavity** is a complex ecosystem that harbors a diverse community of around 700 bacterial species (11). The conditions in the oral cavity, including

physical and chemical parameters, are subject to continuous change due to external environmental exposure. The primary source of nutrition for indigenous microbes is saliva, along with food consumed by the host and byproducts produced by interspecies interactions, which support their growth and multiplication. However, saliva has poor nutrient availability and variable flow rates. Thus, the bacterial microbiome of the oral cavity mainly comprises microbes with the ability to adhere to surfaces like gums and teeth, thus resisting removal. Non-adherent microbes are removed by mechanical flushing of the mouth, which occurs during chewing and talking and are subsequently destroyed in the stomach. However, defining the exact composition of the oral microbiome is challenging due to the mouth's exposure to exogenous bacteria in food, water, and air. Social contact and kissing can also result in changes in the microbial community (12). The populations of invading microbes initially consist mostly of aerobes and obligate anaerobes, mainly related to the genera *Streptococcus*, *Actinomyces*, *Veillonella*, and *Neisseria*. Later, after the eruption of teeth, anaerobic forms such as *Prevotella*, *Fusobacterium*, and others dominate due to the presence of an anaerobic environment between gums and teeth. *Streptococcus* spp., such as *S. parasanguis* and *S. mutans*, grow on enamel and colonize gingival epithelial surfaces and saliva by producing various adherence factors that facilitate their attachment and colonization (13).

Gene expression and metabolic pathways of the oral bacterial microbiome play a crucial role in maintaining oral health and preventing oral diseases. Individuals who have not suffered from dental caries possess genes responsible for antimicrobial peptides and quorum sensing (14). *Fusobacterium nucleatum* is a keystone species in periodontal disease, and lysine fermentation is the major metabolic pathway in the diseased condition, as *F. nucleatum* degrades lysine into butyrate. Transcriptome-based analysis in periodontal disease showed increased expression of butyrate production genes in *F. nucleatum*, resulting in disease promotion (15). Microbes in periodontal disease can produce intracellular toxins that accumulate in periodontal pockets due to reduced ability for their decomposition (16).

The oral cavity acts as a gateway to different organs of the body and acts as a reservoir for different diseases associated with various organs. Many studies have reported on the direct influence of the oral microbiome on prominent systemic diseases such as cardiovascular diseases (17), diabetes (17), stroke (18), and pneumonia (19).

Under normal physiologic conditions, the **esophagus** acts as a conduit and does not retain food contents, which is in contrast to the oral cavity or the stomach and colon(20). Culture studies based on washings from the esophagus suggested that bacteria obtained from the esophagus were either swallowed from the oral cavity or reached the distal esophagus during reflux from the stomach (20, 21). A study of the bacterial microbiome of the oral cavity and the upper and lower esophagus, obtained by esophageal brushings and biopsy samples, revealed that *Streptococcus viridans* is the most common bacterium (22), with the most prevalent organisms being *Streptococcus*, *Prevotella*, and *Veillonella* (10, 23, 24, 25).

Several studies have reported changes in the microbiota of the lower esophagus in a variety of diseases, including reflux disease, Barrett's esophagus (BE), and esophageal carcinoma, in addition to eosinophilic esophagitis in a pediatric population. For example, one group suggests that the pathogenesis of gastroesophageal reflux disease (GERD) might be driven by alterations of the esophageal microbiome with increasing gram-negative bacteria in esophagitis and BE (26). With this increase in gram-negative bacteria, their lipopolysaccharide can upregulate gene expression and, through the TLR4 and NFB pathway, proinflammatory cytokine production can also be increased.

The **stomach** environment is extremely acidic and detrimental for the colonization by most of the bacteria, which for a long time led people to believe that it was sterile. In 1982, Marshall and Warren introduced *Campylobacter pyloridis*, which was later renamed *Helicobacter pylori* in 1989. This discovery had a significant impact on how we understand the role of bacteria in the stomach (27).

The bacterium *H. pylori* has adapted to survive in the low pH environment of the stomach by producing urease and ammonia (28, 29), which help it to alkalinize the immediate surroundings. This enables *H. pylori* to survive through the variable acidity of the gastric juice and reach the higher pH of the mucous layer in close apposition to the surface epithelial cells. In response to *H. pylori* infection, the acute inflammatory response initiates the release of interleukin-8 and the recruitment of inflammatory cells, leading to a chronic active gastritis (30, 31).

Although *H. pylori* is the most well-known and studied of the gastric bacteria, other bacteria have been identified in the stomach, such as *Veillonella*, *Lactobacillus*, and *Clostridium*, soon after the discovery of *H. pylori* (32). *Lactobacillus* species convert lactose to lactic acid, acidifying the surface of the gastric mucous layer (33), which explains its adaptation to the acidic environment and colonization of the stomach (34, 35, 36). *Yersinia enterocolitica* and *Vibrio cholera* are two additional species that survive gastric acidity, with *Yersinia* having an acid-activated urease mechanism and *Vibrio* expressing an acid tolerance mechanism that maintains the cytoplasm at a pH of 4 to 5, although growth does not occur (33, 37).

With the development of culture-independent methods, other bacteria have been identified in the stomach, including *Neisseria*, *Haemophilus*, *Prevotella*, *Streptococcus*, and *Porphyromonas* (38, 39, 40). In healthy individuals, the predominant bacteria are Actinobacteria (*Rothia*, *Actinomyces*, and *Micrococcus*), Bacteroidetes (*Prevotella* species), Firmicutes (*Streptococcus* and *Bacillus*), and Proteobacteria (which include *H. pylori* as well as *Haemophilus*, *Actinobacillus*, and *Neisseria*). The predominant genus is *Streptococcus*, which may originate from the oral cavity (38, 40, 41).

The **colon** is a highly anaerobic environment that serves as the final site of digestion and absorption in the human gastrointestinal tract (10). Digesta that pass through the colon consist of complex polysaccharides and fibers that could not be digested by host processes, as well as trace nutrients and any remaining bile acids that were not absorbed in the ileum. Motility in the colon is much slower, with a

typical transit time of up to 30 hours, creating an optimal environment for microbial growth and metabolism (10, 42).

Microbes thrive in the colon due to the anaerobic environment and abundance of fibrous substrates, which are ideal conditions for fermentative metabolism, a widespread process in bacteria. Additionally, the slow transit time through the colon allows microbes plenty of time to adhere, consume, multiply, and expand in physical space, resulting in higher levels of microbial accumulation (10).

One of the most important physiological functions of the gut microbiome is the microbial synthesis of short-chain fatty acids (SCFAs) through fermentative metabolism of complex polysaccharides. This process releases SCFAs like acetate, propionate, and butyrate (43), which are salvaged by host tissues and contribute to an estimated 5-15% of the total caloric requirement for humans (44). Butyrate, in particular, serves as the preferred energy source for colonocytes (45, 46), which oxidize butyrate into CO₂, promoting an anaerobic state that is important for pathogen resistance, immune homeostasis, and the growth of butyrate-producing anaerobic microbial populations, in a classic positive feedback loop (44, 45, 47, 48).

The microbial composition, and therefore fermentation and metabolism, in the colon can be dramatically affected by the components of the host's diet. Studies have shown that microbial composition changes rapidly and reversibly in response to dietary components (49), with high-fat diets shifting communities towards a greater ratio of Firmicutes:Bacteroidetes taxa (50). Recent research has shown that administration of specific complex polysaccharides can promote the growth of specific *Bacteroides* species, indicating that the host diet can have a significant impact on the microbial composition and function in the colon (51, 52).

The availability of resources in the colonic lumen, which is the open space within the colon, varies depending on dietary intake. However, many colonic microbes reside in and consume host-derived components of the mucosa, which may remain more stable across dietary behavioral patterns (52).

The glycoprotein Mucin 2 (MUC2) is the major component of the colonic mucus, which is a gel-like substance that protects the underlying epithelial cells from

mechanical and chemical damage. MUC2 is coated with a diverse assortment of O-linked glycans, which can be cleaved from MUC2 and metabolized to support the growth of various specialized microbial taxa, such as *Akkermansia muciniphila* (53, 54). Different microbes vary in their ability to penetrate and adhere to the mucus layer, as well as in their tolerance to antimicrobial peptides (AMPs) and oxygen that diffuse outward through the mucus from the underlying epithelial cells. As a result, the mucosal niche experiences unique selective pressures compared to the luminal niche in the colon and supports an enrichment of aerotolerant, asaccharolytic protein-metabolizing species from the phyla Actinobacteria and Proteobacteria compared to the lumen (55, 56). Researchers have also documented further spatial niche partitioning within the mucosa. For example, *Acinetobacter* species are particularly effective at navigating through the mucus layer and its associated biochemical gradients to bind directly to the intestinal epithelial cells (IECs) of the colonic crypts (57).

In addition to the mucosa, the lumen itself harbors significant spatial heterogeneity, with distinct "inter-fold" regions in the intestinal lumen that are enriched for certain taxa in the families Lachnospiraceae and Ruminococcus. These taxa are thought to benefit both from the local accumulation of mucus from the epithelia and from an environment that is relatively protected from the flow of other luminal contents (58, 59). By contrast, the central lumen is dominated by strictly anaerobic, saccharolytic taxa from the families Bacteroidaceae, Enterococcaceae, Prevotellaceae, and Rikenellaceae (59). It is worth noting that much of this heterogeneity has been overlooked by the use of fecal sampling in the majority of colonic microbiome studies, which may not accurately reflect the microbial composition of different niches within the colon (10).

The interplay between the host's immune system and the colonic microbiome is critical in maintaining a healthy gut. Commensal microbes have evolved various strategies to evade the immune system and avoid triggering inflammation (10). For instance, some commensal strains have modified their outer membrane to evade host-derived antimicrobial peptides (60), while others employ immunomodulatory strategies to colonize the mucosal niche. A component of the

polysaccharide capsule of *Bacteroides fragilis*, for example, stimulates regulatory T cells to produce immunosuppressive interleukin-10, allowing *B. fragilis* to colonize the mucosal niche (61, 62). Dysbiosis and inflammation are interrelated, with inflammatory conditions resulting in a microbiome characterized by reduced Firmicutes and Bacteroidetes phyla and increased Actinobacteria and Proteobacteria (63, 64, 65, 66). The mucus degrader *A. muciniphila* has been linked to wound healing by promoting enterocyte proliferation and migration, while inflammatory bowel diseases (IBD) are associated with a compromised mucosal barrier and inappropriate immune activation by commensals. *B. fragilis* biofilms have been observed in IBD patients, although it remains unclear whether this is a cause or effect of the disease (61, 62). Moreover, mucus-consuming bacteria with increased prevalence in IBD patients may play an essential role in mucus utilization, mucosal proximity, and disease.

1.1.2.2 Skin

As the largest and most exposed organ in humans, the skin has the feature of very low dispersal limitation, niche differentiation, and high perturbation. The human skin is also a complex ecosystem that provides diverse microenvironments, such as variation in pH, moisture, temperature, and sebum content (67). The composition of microbial communities was found to be primarily dependent on the physiology of the skin site in sequencing surveys of healthy adults (67, 68, 69, 70). Changes in the relative abundance of bacterial taxa were associated with moist, dry, and sebaceous microenvironments, with lipophilic *Propionibacterium* species dominating sebaceous sites, while bacteria such as *Staphylococcus* and *Corynebacterium* species, which thrive in humid environments, were preferentially abundant in moist areas such as the bends of the elbows and the feet.

Individuals were found to be colonized by different multi-phyletic communities of *Propionibacterium acnes* and *Staphylococcus epidermidis* strains across body sites (68). Compared with the richer environment of our intestines, skin lacks many nutrients beyond basic proteins and lipids. However, the resident microbiota of our skin has adapted to utilize the resources present in sweat, sebum, and the stratum corneum to survive in such a cool, acidic, and desiccated environment (71).

For instance, the facultative anaerobe *Propionibacterium acnes* thrives in the anoxic sebaceous gland by using proteases to liberate the amino acid arginine from skin proteins (72) and lipases to degrade triglyceride lipids in sebum (73), releasing free fatty acids that promote bacterial adherence (74, 75, 76). Sebum levels of the cheek were shown to positively correlate with *Propionibacterium* spp. abundance in facial samples (77).

Auxotrophic *Corynebacterium* species that are unable to produce their own lipids utilize the lipids of sebum and the stratum corneum to generate the corynemycolic acids that coat their cell surface (71). *Staphylococcus* spp., on the other hand, have evolved many strategies for surviving on the skin, including the ability to be halotolerant (withstanding the high salt content of sweat) and utilize the urea present in sweat as a nitrogen source. Moreover, various *Staphylococcus* spp. can produce adhesins that promote attachment to the skin and proteases that liberate nutrients from the stratum corneum to further promote colonization (71).

1.1.2.3 Vagina

The human vaginal mucosa is a stratified squamous nonkeratinized epithelium covered by cervicovaginal secretion, which acquires oxygen, glucose, and other nutrients from underlying submucosal tissues through diffusion due to the limited blood supply (78, 79). In women of reproductive age, physiological changes, such as fluctuations in hormone levels, cause marked differences in the vaginal microbiome (80, 81). Notably, pregnant women experience a sharp decline in the diversity and abundance of the vaginal microbiome, with the predominance of *Lactobacillus* spp., *Actinomycetales*, *Clostridiales*, and *Bacteroidales*. In contrast, non-pregnant women display the predominance of *Lactobacillus* spp., *Actinobacteria*, *Prevotella*, *Veillonellaceae*, *Streptococcus*, *Proteobacteria*, *Bifidobacteriaceae*, *Bacteroides*, and *Burkholderiales* (82). However, the vaginal microbiome differs largely among individuals, with variations in sexual activity (83), douching (84), chronic stress (85), regional disparity (86), race (87), and other factors (88). High-throughput sequencing studies have identified five community state types (CSTs) of the vaginal microbiome. These configurations can be represented by five CSTs, four of which are dominated by single species of

Lactobacillus (CST I-*L. crispatus*, CST II-*L. gasseri*, CST III-*L. iners*, CST V-*L. jensenii*). A fifth configuration, CST IV, represents the more proportionally even collection of facultative and obligate anaerobes, including *Gardnerella*, *Atopobium*, *Prevotella*, *Candidatus Lachnocurva vaginae*, *Sneathia*, *Peptoniphilus*, *Fingoldia*, and *Megasphaera* (89, 90). CSTs I, III, and IV are the most prevalent and account for around 90% of reproductive-age women (87).

Lactobacillus species flourish in the vaginal anaerobic environment and produce various antimicrobial compounds, such as lactic acid, hydrogen peroxide (H₂O₂), and bacteriocins, thereby contributing to a healthy vaginal microbiome and establishing a defense against invading pathogens. *Lactobacillus* species are the main source of l-lactic acid and d-lactic acid that keep the pH value of the habitat lower than 4.5 (91, 92). The dominant *Lactobacillus* species determines the extent of vaginal ecosystem protection. For instance, dysbiosis and low stability are usually related to the vaginal microbiota dominated by *L. iners*. On the contrary, health and high stability of the vaginal community are enhanced by *L. crispatus* that provides d- and l-lactic acids (93). Different from other *Lactobacillus* species, *L. iners* cannot generate d-lactic acid, which plays a more important role than l-lactic acid (91, 94, 95).

The composition of the vaginal microbiota has been associated with increased risk for non-sexually transmitted infections, including urinary tract infections (96, 97), vulvovaginal candidiasis (98, 99, 100), and pelvic inflammatory disease (101, 102, 103). There is evidence supporting an association between the composition of the vaginal microbiota and reproductive health, including the risk for spontaneous preterm birth.

1.1.3. Mycobiome

The term “mycobiome” was first introduced in 2010 by Ghannoum (104), which referred to the fungal community of the microbiome. Compared with the bacterial community in the human microbiome, the mycobiome has only been partially investigated. The fungal species only make up a small proportion of all the microbes residing in the human body. The cultivable fungi in feces range from 10² to 10⁷ cfu/g (105, 106, 107), indicating that the ratio of the fungal cells against the

bacterial cells is between 10^{-9} and 10^{-4} (4). Moreover, the proportion of the fungal genes in the human gut is reported to be less than 0.08% of the whole microbiome based on metagenome analysis (108, 109, 110). In general, this numerical inferiority, making the human gut mycobiome a subdominant community, places more obstacles for scientists to study due to the high sequencing costs in the shotgun metagenomic approach.

1.1.3.1 Gastrointestinal (GI) tract

The **oral cavity** is home to a diverse mycobiome, with *Candida* species being the most prevalent and responsible for various oral infections. *Candida albicans*, a normal inhabitant of the oral cavity, can form biofilms on solid surfaces and invade adjoining cells, leading to infections. Culture-independent studies carried out in 20 healthy hosts have reported the presence of 85 fungal genera in the oral cavity, with the main species observed being those belonging to *Candida*, *Cladosporium*, *Aureobasidium*, *Saccharomycetales*, *Aspergillus*, *Fusarium*, and *Cryptococcus* (3, 104, 111, 112, 113). However, further studies are needed to confirm the presence of these genera in the mouth and determine whether they are transient or permanent members of the normal microbiota.

While *C. albicans* is isolated in association with oral candidiasis 70-80% of the time, other *Candida* species such as *C. glabrata* and *C. tropicalis* are associated with a minority of such infections (114). Other invasive fungal organisms that have been identified as potential members of the oral mycobiota may cause oral disease, albeit rarely. For instance, *C. neoformans* may produce oral lesions in the form of superficial ulcerations, nodules, or granulomas (114). *Aspergillus* species may cause palatal or other oral disease, which often appear as black or yellow necrotic lesions following progression from infection in the maxillary sinuses. Additionally, saprophytic Mucoraceae has been cultured from healthy oral cavities and may cause necrosis or ulceration of the palate in immunocompromised individuals via extension from paranasal infection (114). *Geotrichum* is an uncommon oral mycobiota member and has been reported to cause oral disease in immunocompromised patients or diabetics (115).

The oral microbiome mainly exists in the form of a biofilm, which plays a crucial

role in maintaining oral homeostasis, protecting the oral cavity, and preventing disease development. *Streptococci* associated with *Candida* in oral biofilms may promote the invasive properties of the latter (116, 117, 118). *Candida* species typically cause most infections of the oral mucosa, mainly in immunocompromised individuals due to local overgrowth of the organisms (116). The common underlying link between all known host systemic conditions associated with oral *Candida* overgrowth is a functional immunodeficiency in the Th17 CD4+ cell subset (116).

The mycobiome of the human **esophagus** has received relatively little attention compared to other body sites. However, recent studies have shown that fungi such as *Candida albicans*, *Candida glabrata*, and *Saccharomyces cerevisiae* are commonly found in the esophagus of healthy individuals (119). The presence of these fungi in the esophagus can lead to various diseases such as oral and esophageal candidiasis when there is a deficiency of CD4+ Th1 lymphocytes and reduced formation of proinflammatory cytokines (IL-12, INF-gamma) that prevent effective defense against fungi (120, 121). In a study of 69 eosinophilic esophagitis (EoE) patients and 10 non-EoE healthy controls (122), fungal taxa commonly present in esophageal samples included *Candida*, *Cladosporiaceae*, and *Malassezia*. Interestingly, *Agaricomycetes*, *Candida*, *Cladosporiaceae*, and *Peniophora* were seen most often in healthy samples. Another study of 106 subjects who underwent upper gastrointestinal endoscopy using shotgun sequencing found *Candida albicans*, *Candida glabrata*, *Saccharomyces cerevisiae*, and other fungi (0.0097–1.08%) in approximately 20% of subjects (119). These findings suggest that the fungal microbiome of the esophagus may play a role in health and disease, and further research is needed to fully understand its composition and function.

The human **stomach**, as discussed before, which was previously thought to be hostile to microorganisms due to its acidic environment, has been found to harbor *Candida* and *Phialemonium*. These two groups of fungi are able to colonize and

survive in the low pH environment in gastric fluids (111). Within the stomach, the growth of *Candida* can be antagonized by the presence of *Lactobacillus* (123). In some cases, erosions and ulcerations of the mucosal surfaces in the stomach and intestinal tract can create a favorable environment for *C. albicans* to colonize and grow (124). *C. albicans* has been associated with gastric ulcers in humans, although its role as an etiologic agent of gastric ulceration is not widely recognized (125).

The **colon** harbors a more diverse and highly variable mycobiota compared with other body sites, which makes it difficult to define a standard healthy composition (126, 127). However, two dominant phyla, Ascomycota and Basidiomycota (3, 126, 128), are prevalent in the gut. *Candida*, *Saccharomyces*, *Galactomyces*, *Penicillium*, *Aspergillus*, *Malassezia*, and *Debaryomyces* are among the most frequently identified genera, although some may not be permanent colonizers of the gut. An individual's lifestyle plays a significant role in the variability of their gut mycobiota (127, 129). Ascomycota encompasses several classes, including Saccharomycetes, Dothideomycetes, Sordariomycetes, and Eurotiomycetes, with Saccharomycetaceae, Aspergillaceae, Cladosporiaceae, Debaryomycetaceae, Dipodascaceae, and Pichiaceae dominating at the family level (127, 130). Other families, such as Ceratocystidaceae, Hypocreaceae, Metschnikowiaceae, Nectriaceae, Thermoascaceae, or Microascaceae, are less abundant. The most commonly described fungi genus are *Candida* and *Saccharomyces* (3, 126, 131, 132, 133, 134, 135, 136, 137), but various other genera have been reported, including *Debaryomyces*, *Meyerozyma*, *Toluraspota*, *Pichia*, *Clavispora*, *Cyberlidnera*, *Hanseniaspora*, *Geotrichum*, *Galactomyces*, and *Zygosaccharomyces* (3, 127, 130). Filamentous genera such as *Paecilomyces*, *Cladosporium*, *Aspergillus*, and *Penicillium* are also prevalent (126, 131, 134, 136, 137, 138), with some studies reporting *Claviceps*, *Fonsecaea*, *Exophiala*, *Eurotium*, *Phialophora*, and *Scopulariopsis* (127, 129, 131, 139). In Basidiomycota, the most common yeasts in the gut are from the families Malasseziaceae, Cryptococcaceae, Corticiaceae, Sporidiobolaceae, and

Erythrobasidiaceae. *Malassezia*, *Cryptococcus*, and *Rhodotorula* are the most commonly described genera, but *Filobasidium* and *Trichosporon* have also been reported. The least abundant phylum, Mucoromycota, is represented by the class Mucoromycetes, family Mucoraceae, and filamentous genera *Mucor* and *Rhizopus* (127, 130, 131). A previous study has suggested that the gut mycobiota can be classified into two mycotypes, with Mycotype 1 characterized by a high abundance of *Saccharomyces* and other unclassified genera and Mycotype 2 predominantly consisting of *Penicillium*, *Malassezia*, and *Mucor* (137).

Throughout life, the diet is a crucial factor that can influence the gut mycobiome, as numerous food products harbor food-borne fungi, such as vegetables, fruits, fermented dairy products, meat, and fermented beverages (49, 140, 141). Abundances of several fungal taxa have been reported to be related with food categories. For example, the genera *Saccharomyces* and *Hannaella* have positive correlations with butter and animal fats, while the genus *Aspergillus* shows a positive correlation with eggs and refined grains. In contrast, *Saccharomyces* and *Aspergillus* are negatively correlated with whole grains, while *Hannaella* is negatively correlated with fish and shellfish (141). Additionally, the genus *Fusarium* is identified in 88% of vegetarians but only in 3% of omnivores (142, 143). A switch to a strictly animal-based diet for a short period increases the relative abundance of the genus *Penicillium* but decreases the genera *Debaryomyces* and *Candida* (49). Of the microscopic fungi that colonize the gut, only approximately 20% permanently inhabit this environment, such as the genus *Candida* and species *Geotrichum candidum* and *Rhodotorula mucilaginosa*. The other 80% are considered allochthonous environmental and food-borne fungi, such as *Aspergillus* and *Penicillium* (127, 143).

Various fungal taxa show positive and negative correlations with lipid and carbohydrate metabolism, which can alter the gut mycobiome. Lipid metabolic factors, including body mass index, body fat mass, fasting triglycerides, serum total cholesterol, low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol, can shape the gut mycobiome (130). Carbohydrate metabolic parameters, such as fasting glycated hemoglobin, insulin, and fasting glucose,

also play a role in determining the gut mycobiome composition (130, 144). The gut mycobiomes of overweight and obese individuals differ from those of healthy eutrophic controls and have their specific composition (130, 131). Overweight patients tend to have yeast *Candida* and *Pichia* and filamentous *Bipolaris*, *Beauveria*, *Exophiala*, *Syncephalastrum*, and *Helminthosporium* as the predominant genera (131). In contrast, obese patients tend to have *Candida*, *Nakaseomyces*, *Penicillium*, *Chaetomium*, and *Emmonsia* as the dominant genera (130, 131).

Identifying and influencing the onset of pathogenic processes is crucial for maintaining a healthy body state. Intestinal dysbiosis can affect microbial penetration across the gut barrier and is associated with various conditions such as IBD, IBS, colorectal cancer, obesity, diabetes, multiple sclerosis, atopic dermatitis, Parkinson's disease, and schizophrenia. The gut mycobiome composition differs in these conditions, and certain species like *C. albicans*, *C. glabrata*, and *C. tropicalis* are more abundant in patients with IBD (128, 145, 146). Overweight and obese individuals show decreased biodiversity, while anorexic patients have unique mycobiome species (147). Modulating gut microbial colonizers through diet could have anti-diabetic effects (148). In multiple sclerosis patients, there is higher alpha diversity and over-representation of *Saccharomyces* and *Aspergillus* (149). Fungal dysbiosis may contribute to various disorders, and understanding fungal-bacterial interactions could lead to novel therapeutic strategies in the future.

1.1.3.2 Skin

In contrast to the diverse bacterial microbiome inhabiting different niches of skin, the fungal community present a more homogenous pattern regardless the versatile physiology (68, 150). *Malassezia* species predominate the majority of the skin, while foot sites host a more diverse combination of *Malassezia* spp., *Aspergillus* spp., *Cryptococcus* spp., *Rhodotorula* spp., *Epicoccum* spp., and others (150). This fungal community composition was found to be similar across core body sites regardless of physiology, unlike bacterial communities which can vary greatly depending on the location. *Malassezia* species, which are auxotrophic

and rely on the lipids of sebum and the stratum corneum, are enriched for lipase genes and depleted for carbohydrate-utilizing enzyme genes compared to other sequenced fungi, which may explain their predominance in the adult skin mycobiome (71, 151).

In a cohort of diabetic foot ulcers (DFUs), fungal diversity was explored using amplicon sequencing of the ITS1 region (152), revealing fungi in 80% of the 100 DFUs analyzed. *Cladosporium herbarum* and *Candida albicans* were identified as the most abundant species. Interestingly, chronic wounds with poor clinical outcomes had increased fungal diversity and commonly exhibited polymicrobial biofilms of fungi and bacteria. These findings highlight the importance of understanding the fungal component of the skin microbiome in the context of disease and wound healing (152).

1.1.3.3 Vagina

Until recently, research on the vaginal mycobiome has predominantly focused on *Candida albicans*, a leading cause of vaginal infection (153). However, recent studies have revealed that other non-albicans species (153), including *C. krusei*, *C. parapsilosis*, *C. tropicalis*, *C. glabrata*, *C. guilliermondii*, *C. pseudotropicalis*, and *C. stellatoidea* (153), along with *Saccharomycetales*, *Davidiellaceae*, *Cadosporium*, and *Pichia*, are also present in the vaginal mycobiome, albeit in smaller numbers (154, 155). Strain tropism for *Candida*-induced infections has not been supported by existing studies, as identical strains have been isolated from patients with and without vulvovaginal candidiasis (VVC). It has been estimated that 10-20% of healthy women have commensal *Candida* fungal colonies in the vaginal area that do not cause physical symptoms (156, 157).

A 2012 study using 18S rRNA gene clone sequence libraries identified three phyla of fungi in the vaginal area: Ascomycota (78.6%), Basidiomycetes (17.8%), and Oomycetes (3.6%) (158). *Candida* was the primary genus of Ascomycota. The study concluded that women with recurrent vaginal candidiasis (RVC) and allergic rhinitis (AR) had higher populations of *C. albicans* in the vaginal area than healthy women, and women with RVC had lower populations of *S. cerevisiae* than healthy women. The study also found a general increase in the diversity of vaginal fungal

flora in women with RVC and AR, indicating that allergic reactions in the vagina could alter the fungal flora of affected patients (158).

A study conducted in Estonia, found the prevalence of *Candida* to be significantly higher (36.9%) than in earlier studies, with 64.5% of women showing vaginal colonization with *Candida*. The study identified two phyla of fungi: Ascomycota (58%) and Basidiomycota (3%). Of the Ascomycota OTUs identified as *Candida*, 82% belonged to *C. albicans*. A large portion of the data (38%) was comprised of unspecified OTUs, highlighting the low number of fungal species represented in reference databases, which is a significant issue in the study of mycobiota (154, 159).

The human vaginal microbiota, including the mycobiome, is a crucial aspect of women's health. Intrauterine adhesion (IUA) disease has been linked to certain fungal genera, such as *Filobasidium* and *Exophiala*, which are enriched in IUA samples versus healthy subjects (160). Furthermore, studies have demonstrated correlations between certain fungal and bacterial genera in the cervical canal. Ascomycota and Basidiomycota, for instance, have been found to correlate with Proteobacteria, while a negative association was observed between *Prevotella bivia* and *Candida maltose* (160). In healthy subjects, but not IUA subjects, a negative correlation was found between *C. parapsilosis* and *Cutaneotrichosporon jirovecii* (160). Interestingly, it has been observed that the presence of certain fungal strains, such as *C. parapsilosis*, has a protective effect against IUA progression (160). A reduction in inflammation and fibrosis was observed in a rat model of IUA in the presence of *C. parapsilosis* (160). In addition, *C. parapsilosis* has been shown to protect against *Candida albicans*-induced damage in intestinal epithelial cells (161). The role of fungi in the pathogenesis or protection of IUA could have significant implications for women's health, as IUA is linked to infertility, pregnancy terminations, hypomenorrhea, and amenorrhea (160, 162).

Changes in the vaginal mycobiome composition can lead to the destruction of important bacterial normal flora, such as Lactobacilli, which have antifungal properties and antagonistic competition (155). Such changes can lead to the development of candidiasis and other complications. Dysbiosis and synergistic

bacterial interactions with native vaginal *Candida*, specifically with *Streptococcus* group B and *E. coli*, have been linked to preterm birth, low birth weight, and sepsis (163, 164). Lactobacilli-containing probiotics have been proposed as a potential treatment and preventative supplement for fungal vaginal dysbiosis, as they have proven useful for bacterial vaginosis. However, more research is needed before they can be widely adopted in the pharmaceutical industry (158).

1.1.4 Interaction between the fungal and bacterial microbiome

In a healthy microbiome, bacteria and fungi coexist in balance and contribute to various essential functions, such as digestion, nutrient absorption, and immune system regulation. However, disruptions in this balance, such as an overgrowth of harmful fungi or bacteria, can lead to dysbiosis, which is known to associate with a variety of diseases, including inflammatory bowel disease, diabetes, irritable bowel syndrome, obesity, etc (165). Members of the human microbiome, especially in the vaginal or GI tract, interact with each other as well as with the host to maintain homeostasis. It has been reported that commensal or pathogenic bacteria can modulate the pathogenicity of fungi by affecting their ability to thrive in the host. For example, the abilities of *C. albicans* to adapt to various environmental perturbations can be influenced by their interactions with the bacterial community. Lactic acid, which is secreted by *Lactobacillus* spp., helps to maintain the vaginal pH at a level unfavorable to the growth of potentially pathogenic microorganisms such as *C. albicans* (166, 167). *Lactobacillus* spp. were also found to secrete cyclic dipeptides and hydrogen peroxide, which are thought to have a direct inhibitory effect on *C. albicans* (168, 169). The interaction between different kingdoms, as observed in these cases, appears to play a crucial role in preserving a healthy physiological condition. Those who have lower levels of colonization by certain *Lactobacillus* spp. are more susceptible to developing vulvovaginal candidiasis (170, 171).

One notable interaction between fungi and bacteria in the gut is cross-feeding, where fungi and bacteria exchange nutrients and byproducts, promoting the growth and survival of each other. For instance, certain bacteria can break down complex carbohydrates into smaller molecules that are then consumed by fungi.

In turn, fungi can also induce the production of byproducts of bacteria such as short-chain fatty acids (172, 173). On the other hand, some bacteria and fungi can also compete for resources in the gut, leading to an imbalance in their populations. For example, the overgrowth of certain fungi, such as *C. albicans*, can lead to the suppression of beneficial bacteria, which can result in the development of gut dysbiosis (125, 174, 175). Overall, the interaction between fungi and bacteria in the human microbiome is a dynamic and intricate process that plays a crucial role in maintaining health.

1.2. Metagenomics in studying the microbiome

Taxonomic and functional profiling of the microbial communities is the most important step to study the human microbiome. When this discipline was in its infancy, the culture-dependent method was the most utilized tool. However, the ability of this technique was limited due to the high risk of contamination, high requirements for researchers' skill level, and the difficulty to implement with high throughput. In recent decades, since the Sanger sequencing technique was invented in 1977, DNA sequencing technologies have developed rapidly, which enables researchers to study human microbiome in a high-resolution and culture-independent manner. Other new techniques, such as culturomics and in vitro modeling approaches, also contribute to elaborate the complexity of the study of the human microbiome. Metagenome is the recovery and sequencing of targeted or whole genetic material extracted from all biological samples in an environment, and this process of creating a metagenome is referred to as metagenomics. Metagenome analysis is usually carried out by either the amplicon or shotgun metagenomic sequencing.

1.2.1 Amplicon metagenomic sequencing

Amplicon metagenomic sequencing is the most popular approach in the microbiome field because of its low-cost and low-complexity approach (Fig. 3). In the typical workflow of amplicon sequencing, DNA is first extracted, then a specific region, mostly inside the ribosomal DNA (the 16S rRNA gene (16S) for bacteria, the internal transcribed spacer (ITS) region for fungi), is amplified, sequenced, and

then identified by mapping to a reference database, such like Greengenes database (176) for 16S and UNITE database (177) for ITS. However, amplicon sequencing has its major limitations: first, the primers used for amplification can introduce bias as they bind to regions that are not 100% conserved across all taxa, especially in the fungi kingdom; second, in most cases, the profiling resolution can be accurate until genus level due to high intraspecific similarity between 16S rRNA gene or ITS region; third, the copy numbers of the ribosomal region in different bacterial or fungal clade are not unique, for example, in one of our unpublished study, we found that the ITS copy numbers of 32 tested *Saccharomyces cerevisiae* strains range from 15 to 137. This high variation challenges the quantitative taxonomic profiling of the human microbiome. Furthermore, this method is not able to provide functional information about the microbes, given the fact that it only sequences the ribosomal region.

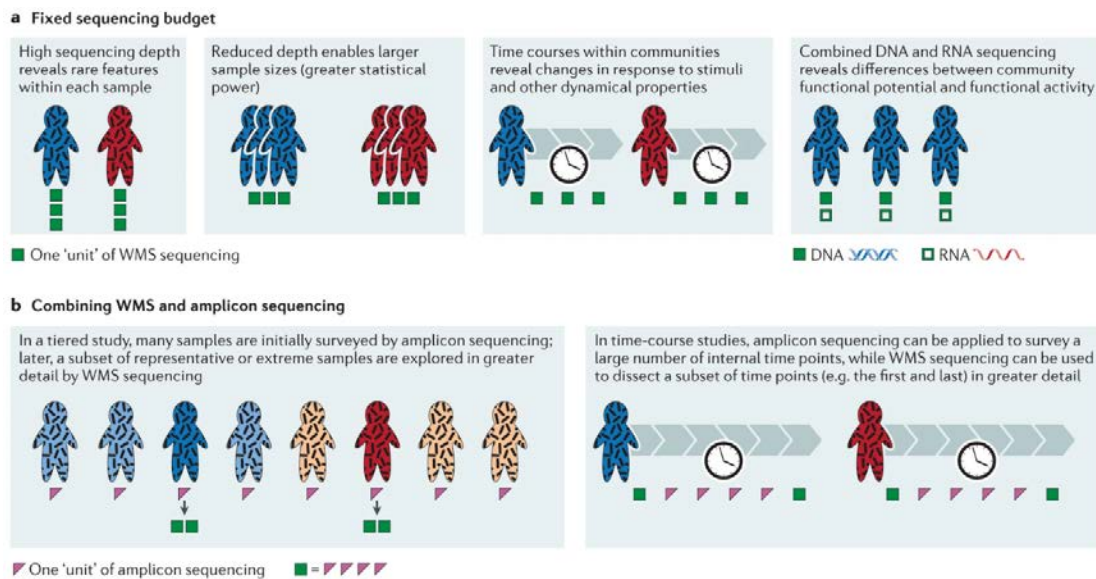


Figure 3. Cost comparison of metagenomic techniques (178).

1.2.2 Shotgun metagenomic sequencing

Shotgun metagenomic sequencing provides the solution to most of the deficiencies of amplicon sequencing. Instead of sequencing only the ribosomal DNA region, the shotgun metagenomic method sequences the DNA fragments generated by randomly breaking the long DNA molecules, such as complete

genomes. Consequently, it provides the full genome data for functional profiling and also avoids the bias introduced by non-universal primers. By mapping the sequencing reads to a single-copy marker gene reference database, it also circumvents the bias of variable copy numbers (3). Moreover, the decreasing cost of sequencing has resulted in the trend of shifting from amplicon analyses towards shotgun metagenomic sequencing, shotgun sequencing has been used substantially more recent years in the analysis of bacterial microbiome. However, due to the numerical inferiority of the fungal community, potential source of bias when using shotgun sequencing to recover fungal sequences has to be considered. To address this problem, either deep shotgun sequencing should be applied or the fungal community has to be enriched.

1.2.3 Databases and pipelines for the human mycobiome profiling when applying shotgun metagenomic sequencing

While many bioinformatics tools have been developed to identify the bacterial community compositions from metagenomic data, such as Metaphlan (179), HuMANn (180), mOTUs (181), few of them focus on the fungi present in the human microbiome. There are currently only a few databases and bioinformatics pipelines that are specifically designed for the analysis of the mycobiome using metagenomic reads from shotgun sequencing data.

FindFungi (182), published in 2018, is the earliest pipeline designed for the identification of fungal species in shotgun metagenomics datasets without relying on rDNA amplicons. Its built-in database includes whole genomes of 949 fungal species for taxonomic profiling. It integrates read identification through the use of Kraken (183) with an analysis of how the reads are distributed across the target genome. Then in 2019, Soverini et al. announced the tool HumanMycobiomeScan (184), which leverages a fungal database that includes 265 fungal genomes to assign reads to specific fungal species with greater accuracy and speed, more than 10,000 times faster than BlastN (185) and MG-RAST (186). The tool was the first human mycobiome profiling pipeline that embedded a decontamination step to remove the bacterial reads.

The above two pipelines use the whole fungal genomes as the content of their

reference databases. Whole genome databases provide a comprehensive catalog of the genomes of different microorganisms. However, these databases have several drawbacks that can limit their utility. High computation time is one of the major problems; the analysis of microbiome samples, which typically involves the simultaneous analysis of multiple genomes, can be particularly computationally intensive. This is due to the large amounts of data generated by high-throughput sequencing, and the need to compare the genomes of thousands of organisms to identify their presence and relative abundance. Another drawback of the whole genome databases is the bias of the relative abundance quantification. Since not all of the DNA regions are single-copy, the multi-copy genes, such as the ribosomal DNA, mentioned above, are likely to wrongly drive the estimated abundance of an organism from the real world.

Single copy marker genes (SCMGs) databases can be the alternative that solves the drawbacks of the whole genome marker genes. SCMGs are genes that are present in only one copy per genome and are, therefore, useful for quantifying the relative abundance of different species in a sample. The SCMGs also substantially reduce the database size, which utterly saves computational time and resources. In 2021, Pollard et al. published the pipeline EukDetect, which includes 214 SCMGs from 2010 fungal sequences. However, this pipeline was designed for detecting the microbial eukaryotes in the microbiome, instead of targeting the fungi kingdom specifically. Besides, all the aforementioned pipelines only aimed at taxonomic profiling, ignoring the seeking of the functional potential of the human mycobiome. Therefore, as the first section of this thesis, we have developed the first version of FunOMIC (3), which is the bioinformatics pipeline with built-in databases for profiling the human mycobiome. FunOMIC contains both the taxonomic database FunOMIC-T which consists of more than 1.6 million fungal SCMGs that covers 1916 fungal species and the functional database FunOMIC-P which encompasses more than three million fungal proteins sequences. In 2022, we have updated the FunOMIC to expand the FunOMIC-T to more than 2 million of SCMGs that covers 3062 fungal species, and the FunOMIC-P to more than 21 million of fungal protein sequences. The pipeline of FunOMIC2 was also upgraded

to involve a bacterial contamination removal step by discarding the sequencing reads that mapped to the UHGG database (108).

Table 1. Summary of the existing reference databases and pipelines for annotating mycobiome.

Tool	Number of fungal species	Type	Algorithm included	Functional database	Reference
FindFungi	949	Genomes	k-mers	No	(182)
HumanMycobiomeScan	Not specified	Genomes	Alignment	No	(184)
EukDetect	1904	SCMGs	Alignment	No	(187)
FunOMICv1	1916	SCMGs	Alignment	Yes	(3)
FunOMICv2	3062	SCMGs	Alignment	Yes	Unpublished

2. HYPOTHESIS

Hypothesis

The human mycobiome, and its interaction with the human bacterial microbiome is still understudied due to various reasons including the challenge associated with unculturable microorganisms, the extremely low abundance among the human microbiome community, inter-individual variability, and the lack of a comprehensive database. This PhD thesis was based on the hypothesis that promoting the fungal taxonomic and functional profiling through experimental and bioinformatics methods will help to unravel the role of human mycobiome in the human microbiome context.

3. OBJECTIVES

3.1 Main objective

The major objective of this thesis is to develop databases and pipelines for the simultaneous analysis of bacterial and fungal components of the microbiome. These databases will allow the realization of applying shotgun sequencing to comprehensively and unbiasedly analyze the human mycobiome.

3.2 Secondary objectives

- To develop experimental method to enrich the proportion of the fungal community in human fecal samples as an assistant step to be combined with the application of shotgun sequencing.
- To explore the fungal-host-microbiome interplay by integrating both bacterial and fungal components in the human microbiome.
- To evaluate whether specific fungal or bacterial signatures in the GI tract correlate with host health status.

4. METHODS

4.1 Construction of the FunOMIC databases and pipeline for human mycobiome profiling using shotgun metagenomics

4.1.1 Collection of fungal genomes

In total, 9,401 publicly available strain-level fungal genomes or draft genomes were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>) and JGI MycoCosm (<https://mycocosm.jgi.doe.gov/mycocosm/home>) (188) before January 25th, 2021. All fungal genomes with more than 500 contigs and N50 < 10 kbp were filtered out (189), which led to a final set of 4,331 high-quality genomes and draft genomes. Genomic shotgun data from 508 *Candida* isolates were downloaded from 30 unique bioprojects from the NCBI SRA before February 4th, 2021 (<https://www.ncbi.nlm.nih.gov/sra/>). The accession numbers of the 4,839 combined reference fungal genomes are listed in Supplementary Table 1.

4.1.2 Construction of the taxonomic and functional FunOMIC database

4.1.2.1 Identification of marker genes for establishing a taxonomic fungal database

Assembling genomic sequencing reads of the 508 *Candida* isolates was performed as described in the study of Montoliu-Nerin et al (190). Basically, each of the *Candida* genomic sequencing reads was normalised by BBNorm v38.9021 of BBtools (<https://jgi.doe.gov/data-and-tools/bbtools/>) with a target average depth of 100x. Then, normalized data were assembled by SPAdes v3.15.2 (191)(<https://cab.spbu.ru/software/spades/>). BUSCO (Benchmarking Universal Single-Copy Orthologs) version 5.0.0 (105) was used to identify marker genes using Fungi OrthoDB version 10.1 (106) in the pool of 4,839 fungal genomes. BUSCO made use of 758 HMMs (hidden Markov models) of fungal single-copy marker genes and was run using default parameters with the AUGUSTUS gene predictor (105). Genomes with less than 30 single-copy

marker genes identified were discarded, resulting in a final set of 4,816 genomes. Clustering with a 99% identity threshold (107, 187) was applied using CD-HIT (192) to remove redundancies, which led to a final set of 1.69 million fungal marker genes, referred here as FunOMIC-T.

4.1.2.2 Establishment of a functional fungal database.

A protein database for fungal functional analysis was also constructed by collecting the corresponding amino-acid sequences that were available for 2,967 of the 4,331 genomes cited above and the 35,360 reviewed fungal proteins from UniProt (<https://www.uniprot.org/>), both before January 2022. Then, the proteins without an explicit annotation were discarded (1.5 million), leading to a total of 4.9 million genes. Redundancy was removed with a 95% identity clustering using CD-HIT (2). Finally, 3,413,239 non-redundant fungal proteins, referred to as FunOMIC-P, were obtained for fungal functional profiling. These protein accessions (from JGI, NCBI, UniProt) were then linked to EC numbers and KEGG pathways.

4.1.3 Validation of the FunOMIC databases and the pipeline

To verify the absence of bacterial contamination [14] in the fungal database and to ensure specificity for fungal detection, we applied three different validation methods. Firstly, we mapped the 1.69 million fungal single-copy marker genes to the Unified Human Gastrointestinal Genome (UHGG), which is a gene catalog that comprises 204,938 non-redundant genomes from 4,644 gut prokaryotes (108) using bowtie2. Because of the memory limitation of our computers (44 CPUs), we simulated sequencing reads of all the marker gene sequences (22 million paired reads, 1-fold coverage, 11.2 GB out of 4.6 GB) to perform the alignment to the UHGG. Secondly, we simulated Illumina formatted sequencing output reads from a set of 903 bacterial genomes from 458 species that inhabit the human body collected from the NCBI to create a mock community for a bacterial community (Supplementary Table 2). The simulation was carried out by ART, a set of simulation tools that generate synthetic next-

generation sequencing reads (109). The simulated reads were then aligned to FunOMIC-T. Thirdly, another mock community was created with the top 20 fungal species and top 20 bacterial species identified in the 2,679 human metagenomes collected (cited below). The genomes of these 40 species were used to simulate Illumina formatted sequencing output reads, which were then mapped to the constructed database. The lists of genomes used for creating the mock communities and the number of simulated reads can be found in Supplementary Table 2.

To validate the FunOMIC-P database, a mixed mock community was created with the available coding gene sequences of the aforementioned top fungal and bacterial species. Again, the coding gene sequences collected from NCBI were used to simulate Illumina formatted sequencing output reads, which were then mapped to the FunOMIC-P database using Diamond blastx function v2.0.8 with an e-value $< 10e-10$ to recover the fungal functional profiling. To optimize the alignment parameters, we tested nine different combinations using three different percentages of coverage (>90%, >95%, >99%) and three different percentages of identity (>90%, >95%, >99%).

4.1.4 Collection of metagenomic data

We downloaded 2,679 public human shotgun metagenomic sequencing data from NCBI SRA before February 4th, 2021 (193) (<https://www.ncbi.nlm.nih.gov/sra/>). The 2,679 public human metagenomic data derive from 27 unique bioprojects, two of which were published in our previous studies (PRJNA514452, PRJEB1220). The metadata of all the human metagenomic data can be found in Supplementary Table 3. This metadata contains available information such as continent, country, city, latitude, longitude, sample source, gender, age, extraction procedure, and use of mechanical lysis during extraction.

4.1.5 Aligning human metagenomic sequencing reads onto the FunOMIC database

After quality control and decontamination using KneadData v0.7.7-alpha (<https://huttenhower.sph.harvard.edu/kneaddata/>), Bowtie2 v2.3.4.3 was used to map the 2,679 metagenomic data to the FunOMIC-T database for fungal taxonomic annotation. Mapped reads were kept if more than 80% of the length aligned to the reference sequence with a q-score of over 30 (2, 187, 194) by using Samtools v1.9. Diamond blastx function v2.0.8 was used to map the metagenomic data to the FunOMIC-P database (read coverage > 95% and identity percentage > 99%, and e-value < 10e-10) for fungal functional annotation. An in-house script, which is freely available on our GitHub (<https://github.com/ManichanhLab/FunOMIC>), was used to recover the final fungal taxonomic and functional profiling.

4.1.6 Prokaryotic taxonomic and functional profilings of human metagenomic data

After quality control and decontamination using KneadData v0.7.7-alpha (<https://huttenhower.sph.harvard.edu/kneaddata/>), we used MetaPhlAn v3.0.9 for profiling the composition of prokaryotic communities in the 2,679 human metagenomic data. Then, the HUMAnN v3.0 (180) (<https://huttenhower.sph.harvard.edu/humann/>) and the UniRef90 database (195) were used to profile the abundance of prokaryotic metabolic pathways and other molecular functions.

4.1.7 Statistical Analysis

All statistical analyses, except for SparCC correlation, were performed using R software 4.1.2 (2021-11-01). Alpha and beta diversity were calculated using the Phyloseq package. Beta diversity was compared between different disease groups using the UniFrac distance metric with permutational multivariate analysis of variance (PERMANOVA) to identify significance ($p \leq 0.05$). The

associations between fungal profiling with variables from the metadata were measured using the MaAsLin2 package with age as the random effect (results were considered significant if FDR (false discovery rate) < 0.05). The correlations of taxonomic profiling or functional profiling between bacteria and fungi were performed using the Python script SparCC (196).

4.2 Update of the FunOMIC and establishment of the MycoDM web server

4.2.1 DATABASES CONTENT AND CONSTRUCTION

4.2.1.1 Expansion and update of the FunOMIC database

The update of the FunOMIC database is split into two works: update of the taxonomic database FunOMIC-T and update of the functional database FunOMIC-P. For the second version of FunOMIC-T, we collected 3847 newly published high-quality fungal genomes or draft genomes until June 2022 from NCBI (<https://www.ncbi.nlm.nih.gov/>) and JGI MycoCosm (<https://mycocosm.jgi.doe.gov/mycocosm/home>). Single-copy marker genes of each genome were extracted following the method of the previous version, which is based on HMM (3). For the expansion of the FunOMIC-P, the available amino acid sequences of the coding genes of the newly collected fungal genomes were appended to the first version of FunOMIC-P. Then, both the newly collected single-copy marker genes and the amino acids were concatenated to the previous version of the FunOMIC databases. After this, clustering was performed by CD-HIT (192) to remove redundancies with a 99 % identity threshold (3).

Consequently, the updated version of FunOMIC-T, namely FunOMIC2-T, included more than 2 million fungal single-copy marker genes, while the FunOMIC2-P included more than 21 million fungal protein sequences. The newly constructed database covers twelve fungal phyla, among which three

(Ascomycota, Basidiomycota, and Mucoromycota) represented more than 98% of the genomes. At lower taxonomic levels, they encompassed 1,080 genera, 3,032 species, and 8,686 strains. The taxonomy of the FunOMIC database was based on the NCBI taxa system, with manual curation for the unknown taxonomic levels and revised fungal names.

4.2.1.2 Construction of the phylogeny

Together with the updated FunOMIC2, we constructed the phylogeny of all the fungal species included in the database, to let users conveniently perform the diversity analysis, for example, calculating an UniFrac distance (197). For this purpose, we translated the 758 single-copy BUSCO genes into amino acid sequences as described in the method section of Chapter 1 for each of the 3,032 fungal species. We then aligned each of the 758 groups of amino acid sequences using MAFFT v7.471 (198) with options “—auto —maxiterate 1000.” The regions that were suitable for inferring phylogenetic trees were selected from each multiple sequence alignment (MSA) using BMGE with the command: `bmge -i input -t AA -h 0.4 -m BLOSUM62 -of output` (199). The trimmed alignments of these 758 BUSCO genes, of which 96.6% (732 out of 758) had more than 50% of taxon occupancy, were then concatenated into one MSA. The inference of the phylogeny was then performed by Fasttree v2.1.10 using CAT+IG model (200).

4.2.1.3 Update of the FunOMIC pipeline

A step of prokaryotic reads removal was newly implemented in the FunOMIC pipeline by using the Unified Human Gastrointestinal Genome (UHGG) prokaryotic database (108). For more accurate taxonomic and functional profilings, the quality-controlled metagenomic reads were mapped to the UHGG prokaryotic database using Bowtie2 v2.3.4.3 with default settings to remove the prokaryotic contamination. The unmapped reads were kept as the non-prokaryotic reads in each of the metagenomic samples and used as input in the FunOMIC profiling pipeline. By testing this step using the mixed mock

community, which we used in the previous paper (3), we noticed that more than 80% of the prokaryotic reads were removed without affecting the fungal reads.

4.2.1.4 Collection of the human gut shotgun metagenomic data

We used a subset of 1,147 public human shotgun metagenomic sequencing data from the aforementioned 2,679 collected human metagenomes (29) (<https://www.ncbi.nlm.nih.gov/sra/>). The 1,147 public human metagenomic data were all collected from gut samples of healthy controls or patients in CD, UC, T1D, T2D, ESRD, and PSO. These samples were from six unique bioprojects, one of which was published in our previous studies (PRJNA514452). The metadata of all the human metagenomic data can be found in Supplementary Table 9. This metadata contains available information such as continent, country, city, latitude, longitude, sample source, gender, age, extraction procedure, and use of mechanical lysis during extraction.

4.2.1.5 Statistical Analysis

All statistical analyses were performed using R software 4.1.2 (2021-11-01). Alpha and beta diversity were calculated using the Phyloseq package. Beta diversity was compared between different disease groups using the UniFrac distance metric with permutational multivariate analysis of variance (PERMANOVA) to identify significance ($p \leq 0.05$). The associations between microbiome profilings with different diseases from the metadata were measured using the MaAsLin2 package with country, study, and BMI as the random effects (results were considered significant if FDR (false discovery rate) < 0.05).

4.2.2 Design and construction of the Web server

The MycoDM web server uses JavaScript as client language, PHP (v7.2.24-0; <http://www.php.net>) as server language and MySQL (v15.1 Distrib 10.1.48-MariaDB; <http://www.mysql.com>) for storing data. The web pages were written in Hypertext Markup Language (HTML), JavaScript, and PHP. They provide a user-friendly interface for locating and retrieving information from the database. The interactive APPs embedded were created using R ([https://www.r-](https://www.r-65)

project.org) code and the Shiny package (<https://shiny.rstudio.com>), and were hosted using Shinyapps.io servers (<https://www.shinyapps.io>).

The web server includes a total of four subpages: Home, Downloads, DiseaseMarker, and online analysis platform. The website features a prominent header that makes it easy to explore and discover relevant information.

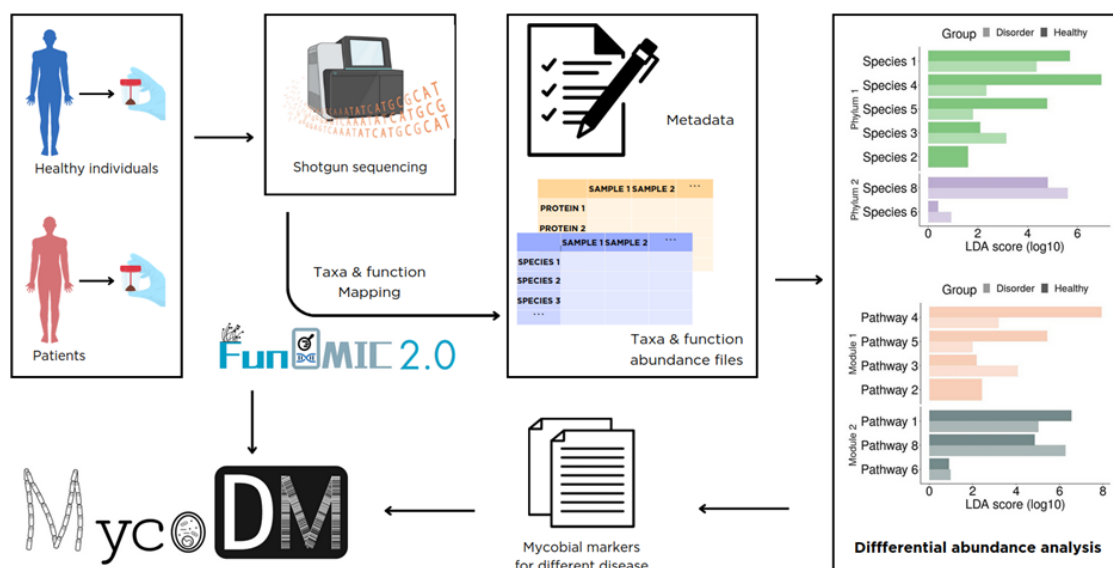


Figure 4. The workflow of constructing the MycoDM web server. The raw shotgun sequencing reads were annotated to get the taxonomic and functional profilings using FunOMIC2 database and pipeline. Then, the generated profiling together with the metadata were used for performing the differential abundance analysis. Finally, the FunOMIC2 database and the detected significant mycobial taxonomic and functional markers associated with diseases were implemented into the MycoDM web server.

4.3 Robust integration of fungal and bacterial gut microbiome with dietary data in a longitudinal setting

4.3.1 Fungal genomes collection

All species and strains used in this project were collected from the FunOMIC-T database (3) in order to use the marker genes. Similar to the analysis of the human gut bacterial microbiome, studies of the human gut mycobiome have

typically used amplicon sequencing of ribosomal DNA, particularly the internal transcribed spacer (ITS) region. However, it has been shown that the copy numbers of the ribosomal DNA in fungi can vary widely at the species and strain levels. This high variation poses a challenge to the quantitative taxonomic profiling of fungal communities in an environment. Additionally, ITS or 18S sequencing methods have been criticized for having low phylogenetic resolution at the species level and for not providing functional information. To address these issues, we evaluated the ITS copy numbers of a set of fungal genomes. To analyze the ITS copy number, 260 fungal strains covered by seven species (*Aspergillus flavus*, *Candida albicans*, *Candida glabrata*, *Cryptococcus neoformans*, *Rhizopus oryzae*, *Rhodotorula mucilaginosa* and *Saccharomyces cerevisiae*) were chosen, as explained in the Results section. For each of the selected strains, both its genome assembly and its corresponding raw shotgun sequencing reads were downloaded from the NCBI (201) or JGI (188) databases. Moreover, genome assemblies of the 14 most abundant species in the human healthy gut mycobiome were also downloaded to create *in silico* mock communities (3).

4.3.2 Estimation of the ITS copy numbers

Two methods were used to estimate copy numbers of fungal ITS regions: hidden Markov models (HMM) and mapping depth (202, 203). To determine the ITS copy numbers using HMM, we created two HMMs respectively for predicting the flanking sequences of the two ends of the whole ITS region. These two HMMs are located separately in the large subunit (LSU) and small subunit (SSU) of the rRNA gene. For creating the HMM of the LSU, a total of 97 sequences in FASTA format were obtained from NCBI (Supplementary Table 10), which were then aligned using the "MUSCLE" tool (204) in Stockholm format. This alignment was then used to create the HMM profiles with the hmmbuild function in UNIX (Wheeler and Eddy, 2013; Baum and Petrie, 1966, doi: 10.1214/aoms/1177699147). For creating the HMM of the

SSU, 100 sequences (Supplementary Table 11) were recovered and used through the same process as with the LSU. The resulting hmm profile had a length of 588 bp for LSU and 142 bp for SSU. With the obtained HMM profiles, we searched the DNA homologies of the beginning and the ending of the ITS regions in each of the 260 genome assemblies mentioned above, then estimated the number of copies by using the nhmmer function. Then, we applied a filtering step to eliminate those matches that presented an E-value greater than 0.001 (205). The copy number was then determined using an in-house Python script that evaluated the distance between the start and end of the ITS region. As most ITS lengths were reported in a range of 400 and 800 bp with an average of 550 bp (206, 207), the script accepted a distance between 400 and 800 bp. Moreover, if there was an HMM match for one end of the ITS but not the other, the script determined whether the prospective ITS sequence was at the end of a chromosome/scaffold, in which case the other end of the region could not be found, thus the script counted another copy. Once the copy number was determined for a genome, the complete ITS sequence was extracted by means of the BEDTools tool, which will be used in the mapping depth method mentioned below (208). The resulting file contained three columns for each genome: genome ID, ITS copy number, and ITS sequence. Summary statistics were calculated for the seven species analyzed.

To estimate CNV using the mapping depth approach, we calculated the ratio of the ITS depth against the single-copy marker genes' depth (203, 209). The ITS sequence for each genome was obtained from the HMM method mentioned above, while the single-copy marker genes were obtained from the FunOMIC-T database (3). The raw reads of the 260 genomes previously downloaded from the NCBI and JGI were filtered and trimmed using Trimmomatic (version 0.36) with the default settings (210) to obtain reads with higher qualities. The filtered reads were then mapped respectively to the corresponding ITS sequence and the set of single-copy marker genes by using Bowtie2 (211). Once the mapping

was finished, samtools (38, 212) was used to convert the resulting SAM file into a BAM file. Next, those reads mapped with a q-score inferior to 30 were filtered out, and the depth at each base was calculated. The resulting file was then analyzed using R (version 4.0.2), where the mapping depth, which was determined as the mean depth of all the bases of each gene, was calculated and normalized by gene lengths. Prior to calculating the mean, the positions at the two ends (which present fewer reads and lead to a bias on the real mapping depth) were trimmed by deleting the first and last 50 base pairs, as done in the report of Lofgren et al. The copy number was finally estimated as the ratio of the mapping depth of the ITS region by the median of the mapping depths of all the single-copy marker genes. We used the median in order to avoid a bias of outlier single-copy genes that had a higher-than-usual mapping depth. The whole pipeline for this analysis is summarized in Fig. 12a.

To validate the above mapping pipeline, we selected 10 *S. cerevisiae* genome assemblies out of the 260 assemblies downloaded from NCBI. We recovered the ITS sequences from each of these assembled genomes using the CN-HMM method. Then we queried the ITS sequences in the chromosome XII of each genome assembly to get the total number of hits, which is used as the reference ITS copy numbers. Then for each of the 10 assemblies, we generated 15 million sequence reads using the InSilicoSeq tool (213). Those reads were then mapped to their respective ITS and single-copy fungal marker genes' sequences using the read mapping pipeline to identify the copy number estimated by mapping depth (CN-MD). The student t-test was done to compare the CN-MD with the references, with a significant threshold p-value < 0.05.

4.3.3 *In silico* comparison of ITS and shotgun methods

To compare the accuracy of the ITS and shotgun methods in detecting the relative abundance of fungi at strain, species, or genus level in environmental samples, genomes assemblies from 27 strains were used, for which the ITS sequences (obtained by CN-HMM method) and copy numbers of ITS (CN-MD)

were extracted. These were used as artificial genomes for creating mock communities. Five different *in silico* mock fungal communities were generated with randomly selected strains for each species. The strains and their randomly attributed abundances are shown in Supplementary Table 12. With the InSilicoSeq tool, we created 15 million reads (214) from the whole genomes (for the shotgun simulation) and another 15 million from the ITS sequences (for the ITS sequencing simulation) for each community (Supplementary Figure 5). After obtaining the reads, the shotgun reads were mapped to the FunOMIC-T database, while the ITS reads were mapped to an in-house ITS database (Fig. 5). The ITS database was created by integrating the UNITE version 8.2 (215) and the RefSeq database (data downloaded before 09/12/2020) (216), as well as the sequences extracted from the HMM, in total 96,388 sequences. The post-mapping processing was the same as the CN-MD pipeline. Then, an extra filtering step was taken: the filtering of those genes that presented less than 15 mapping depth, as they were possible off-target mappings and could be a cause of bias. This filtering was based on previous publications (217, 218), where various tests were undertaken to determine the optimal filtering value.



Figure 5. Workflow of generating simulated sequencing reads for mock communities. The colored dots represent the species in the *in silico* mock community. The genomes of the species were used directly as the input of InSilicoSeq for mimicking the shotgun sequencing. The ITS sequences of the species were replicated with their corresponding CN-MD before input to the InSilicoSeq for mimicking the ITS sequencing.

A lower mapping depth filter would result in the introduction of more off-target species, while a higher filtering depth would reduce low-abundant but relevant species (217, 218). Relative abundance for each mapping hit inside each mock population was calculated by dividing the mean depth of each hit by the sum of all mean depths. Then, the relative abundance of each species was retrieved

by summing all the hits that corresponded to the relevant species; all other species were marked as off-target. The expected and observed relative abundances were then compared using R (v4.0.2), as described in the statistical analysis section.

To further evaluate a more diverse fungal community and to mimic a gut microbiome sample, an additional mock community, consisting of the 14 most prevalent fungal species with their relative abundance found in healthy gut controls, was created (Supplementary Table 12) (3).

4.3.4 Fungal enrichment protocol

Centrifugation was used to separate fungal and bacterial cells based on their heterogeneous cell sizes. Stoke's law was used to estimate the centrifugation speed and time, in which D is the particle diameter (cm), η is the fluid viscosity (poise), R_f and R_o are the final and initial radius of rotation respectively (cm), ρ_p and ρ_f are the density of the particle and fluid respectively (g/ml), ω is the rotational velocity (radians/sec) and t is the required time for sedimentation from R_o to R_f (sec) (equation 1).

$$D = \left(\frac{18\eta \ln(R_f / R_o)}{(\rho_p - \rho_f)\omega^2 t} \right)^{0.5} \quad (1)$$

Briefly, 15 ml of 1X PBS solution was added to 500 mg fecal samples together with 10 1mm glass beads to homogenize the feces into fecal suspension. The suspension was then passed through a 40-micron cell strainer to remove large-size undigested particles. We then centrifuged the filtered suspension for 3 minutes at 201 g using an Eppendorf A-4-62 centrifuge. The supernatant was collected in a 50 ml falcon tube, and the pellet was resuspended in 15 ml of 1X PBS. The resuspended solution was centrifuged again for 3 min at 201 g with the same centrifuge to reduce the remaining number of bacterial cells from the fungi-enriched fraction, then the supernatant was collected and combined with the previous supernatant. We then resuspended the pellet with 1 ml of 1X PBS and centrifuged it for 20 min at 10000 g using an Eppendorf Centrifuge 5427R

to collect the pellet containing the enriched fungal cell fraction. The combined supernatant was centrifuged parallelly in Fiberlite™ F14-6 x 250LE for 30 min at 10000 g, the pellet was collected as the enriched bacterial cell fraction.

4.3.5 Collection and processing of habitual diet information

Six volunteers, free of diagnosed diseases, were recruited between August 2021 and October 2021 by disseminating an announcement. The study was approved by the local Ethics Committee of the Vall d'Hebron University Hospital, Barcelona (Project identification code: PR(AG)84/2020). All participants signed a consent form. During two months, each volunteer filled a sFFQ that recorded the previous month's dietary information (219); in total, 12 sFFQs from the six volunteers were collected.

Nutrients were adjusted by energy using the residual method (220) to control the confounding effect of calories. We then used the Wilcoxon test and the intraclass correlation coefficient (ICC) (220, 221) to evaluate the reproducibility of the sFFQ by comparing both the energy-adjusted nutrient data and the food groups extracted from the sFFQ administered on two-time points.

4.3.6 Sample collection and DNA extraction

Each of the above-mentioned volunteers donated one fecal sample per week for two months, in total, 48 fecal samples were collected. The fecal samples were frozen immediately at -20 °C then transferred to -80 °C within the month. For each of the 48 samples, two aliquots of 500 mg were taken, one was used directly for the DNA extraction, and the other was separated into the fungal enriched partition and enriched bacterial partition by applying the fungal enrichment protocol before the DNA extraction. Thus, three partitions per sample (enriched in fungi, enriched in bacteria, and control without enrichment), in total 143 samples (volunteer No.4 did not provide enough feces for time point 1, so only one aliquot was obtained for getting the enriched fungal and bacterial partitions) were processed for genomic DNA extraction as previously described (222).

4.3.7 Shotgun metagenomic sequencing and profiling

Shotgun metagenomic sequencing was applied to the 143 extracted genomic DNA using the Illumina Novaseq 6000 platform. The average reading depth was 6.45 Giga base pairs. For each of the sequencing samples, we used the KneadData v0.7.7-alpha tool (<https://huttenhower.sph.harvard.edu/kneaddata/>) for trimming out low-quality reads and decontaminating human sequences. Then, the FunOMIC2 database, which contains 2 million single-copy marker genes and 21 million protein sequences extracted from more than 3,000 fungal species (3), was used for getting the raw reads of taxonomic and functional mycobiome profiling. Then the raw reads were normalized with the TMM method (223, 224) using the R package “edgeR”. The MetaPhlAn v3.0.9 and the HUMAnN v3.0 (180) (<https://huttenhower.sph.harvard.edu/humann/>) were used respectively for the taxonomic and functional prokaryotic microbiome profiling. The functional profiling output by HUMAnN was annotated using the MetaCyc pathway database (225), while that of FunOMIC2 was using the KEGG pathway database (226). To make the annotations consistent, we regrouped the prokaryotic functional profiling into the KEGG annotation style by using the function “humann_regroup_table” embedded in HUMAnN and the function “keggLink” under R package “KEGGREST” (Dan Tenenbaum and Bioconductor Package Maintainer (2021). KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)).

4.3.8 Keystone species analysis

The network was constructed based on the species-level SparCC correlation matrix measured using the SparCC tool which uses logarithmically scaled variances to calculate correlations between species (196). We inferred and removed the indirect effects from the observed correlation matrix by using the network deconvolution algorithm as previously proposed (227, 228). Then based on the random matrix theory (RMT), we determined a threshold of $\rho =$

0.78. All correlations that had an absolute value lower than 0.78 were discarded (229). The p-values for all the correlations were adjusted using the Benjamini and Hochberg false discovery rate (FDR), and a cutoff of FDR=0.001 was applied to remove the non-relevant correlations. The resulting correlation matrix was then used to construct the network using the “igraph” R package (230). After network construction, the topological indices, including the degree, betweenness centrality, and closeness centrality of each node, were calculated by using functions developed in igraph.

4.3.9 Statistical analysis

To compare the ITS and the shotgun approaches, weighted UniFrac distances (231) were calculated using the phyloseq package (232). Distances were compared between methods using a Student t-test (233), as the values belonged to a normal distribution, proved beforehand by doing a Shapiro test (234). Spearman correlations of dietary data or metadata with microbiome alpha diversities or microbiome taxonomic and functional compositions were computed using the cor.test from the stats R package (v4.0.2). The p-values for all the correlations were adjusted using the Benjamini and Hochberg FDR. We considered significant correlations with an FDR < 0.05. In the heatmaps for partial correlations, the asterisk indicates that the correlation index for the corresponding species metadata pair is significant.

5. RESULTS

5.1 Validation and application of the FunOMIC databases and pipeline

5.1.1 Characteristics of the taxonomic and functional FunOMIC database

To build a database for taxonomic profiling of environmental fungal species, more than 1.6 million fungal single-copy marker genes were extracted from 4,816 fungal high-quality genomes and draft genomes by aligning them to a set of 758 fungal universal orthologs from OrthoDB (Fig. 6). The newly constructed database, FunOMIC-T, covers eight fungal phyla, among which three (Ascomycota, Basidiomycota, and Mucoromycota) represented more than 98% of the genomes (Fig. 6A). At lower taxonomic levels, they encompassed 475 genera, 1,916 species, and 4,537 strains.

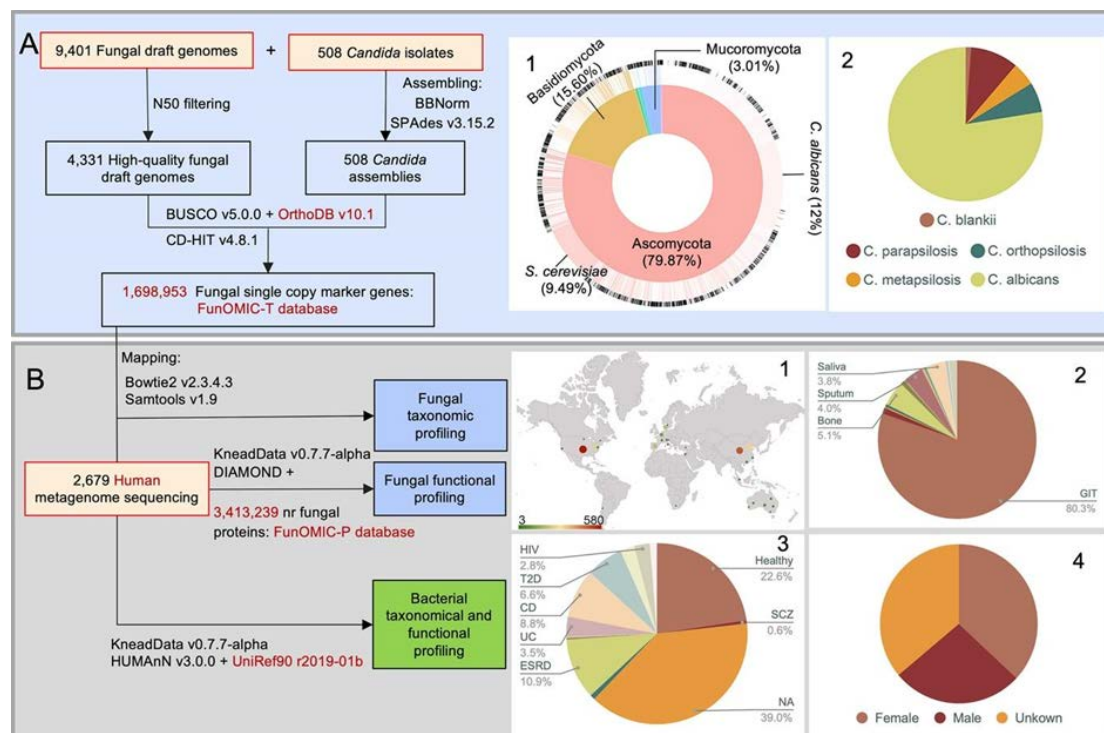


Figure 6. Workflow of the construction of the FunOMIC database and its application in metagenomic analysis. (A) Recovery of fungal single-copy marker genes from fungal draft genomes and *Candida* isolate sequencing

reads downloaded from NCBI and JGI. A.1) Distribution of the fungal draft genomes at the phylum and species levels in FunOMIC-T (Taxonomy). A.2) Distribution of *Candida* assemblies at the species level. (B) Fungal and bacterial taxonomic and functional profiling of the 2,679 metagenomic datasets downloaded from NCBI. B.1) Geographical location of the collected human metagenomes. B.2) Proportions of the collected human metagenomes by body sites. B.3) Proportions of human metagenomes by disease type (HIV=human immunodeficiency virus; T2D=type 2 diabetes; CD=Crohn's disease; UC=ulcerative colitis; ESRD=end-stage renal disease; SCZ=schizophrenia). B.4) Distribution of the collected human metagenomes by gender.

Table 2. Summary of the characteristics of the 1,950 human metagenomes.

Body site	Country	Health status	Number of samples	Mechanical Lysis
Blood	USA	Filariasis	1	no
		Lyme disease	1	no
Bone and joint	France	Infections	24	no
Conjunctiva	China	HC	100	no
Gallstones	Australia	NA	8	no
Gut	Australia	HC	56	yes
		T1D	60	yes
	Belgium	CD	92	yes
	Canada	PLWH	10	na
	China	HC	204	yes
		CD	38	yes
		ESRD	208	yes
		T2D	89	yes
		NA	15	NA
	Denmark	HC	165	no
	Israel	NA	20	na
	Italy	HC	18	yes
	Spain	HC	63	yes
		CD	50	yes
		UC	69	yes
	Sweden	T2D	10	yes
	USA	HC	11	no
		CD	13	na
		HIV	3	na
		PSO	24	no
UC		10	na	
Infant-preterm		140	na	
NA		272	na	
Nasal mucosa	Chile	HC	6	no
		Asthma	5	no
Oropharyngeal	South Africa	TB	4	na
Saliva	USA	NA	61	na
Skin	Italy	HC	3	yes
Sputum	Singapore	NA	30	na
	South Africa	TB	10	no
Throat swab	USA	HC	16	no
		SCZ	14	no
Tongue	Italy	HC	12	yes
NA	USA	mock communities	15	na

It has been reported that 99.9% of human metagenome sequences are from bacteria (2) and that, bacterial sequences are ubiquitous in eukaryotic genomes (187). Validation of the absence of bacterial sequence contamination in the fungal database is, therefore, critical. To address this requirement, the FunOMIC-T database was mapped to the UHGG dataset, which contains 204,938 non-redundant genomes from 4,644 gut prokaryotes (109). Only less than 0.01% of the fungal marker genes mapped to the UHGG, demonstrating

that this fungal taxonomic database was specific enough to detect mostly fungal sequences.

A bacterial environmental mock community was also created. For this, we collected 903 genomes from 458 bacterial species found to inhabit human bodies (Supplementary Table 2). These genomes were then simulated into 19,301,201 Illumina formatted sequencing output reads and mapped to the FunOMIC-T database. The mapping rate of this artificial community to the database was also less than 0.01%. Lastly, a mixed mock community was also created comprising the top 20 bacterial species and top 20 fungal species identified during the taxonomic profiling of the metagenomes (Supplementary Table 4). To better mimic real human metagenomes, the ratio of the number of simulated bacterial reads over fungal reads was set to nearly 1,000 (999,021 bacterial reads and 1,046 fungal reads) (2). As expected, none of the 999,021 bacterial reads aligned against FunOMIC-T, leading to a specificity (false positive / (false positive + true negative)) of 0.9999.

Given the numerically small proportion of fungal sequences in human metagenomes, the fungal functional analysis was not relevant in almost all the published human mycobiome studies. To address this knowledge gap, in the present work, we also proposed a protein database specifically for environmental fungal functional profiling. The FunOMIC-P database consists of 3,413,239 non-redundant fungal protein sequences integrated from NCBI, JGI, and UniProt (see Methods section above, Fig. 6B). Evaluation and validation were also performed by a mixed mock community constituted of the top species mentioned above. The available coding gene sequences of these species were simulated into 439,798 Illumina formatted sequencing output reads and mapped to the FunOMIC-P database. We tuned the Diamond blastx function with nine different combinations of parameters to optimize mapping performance. With the threshold of read coverage > 95%, identity percentage > 99%, and an e-value < 10e-10, we obtained the highest mapping rate of the

fungal reads, where around 70% of the hits passed this threshold. More than 50% of the mis-mapped bacterial genes were related to ATP synthase (Supplementary Table 2).

5.1.2 Characteristics of the 2,679 metagenomes

A set of 2,679 metagenomes, which encompassed a total of 9077.12 Gb, collected from 27 bioprojects are listed in Supplementary Table 3. Taxonomic profiling of the metagenomes against FunOMIC-T detected fungal DNA sequences in 1,950 metagenomes (72.9%) which was much higher than the ratio reported in previous shotgun sequencing studies analyzing the human mycobiome. Lind et al., reported a detection rate of less than 20%, and Olm et al. found 6% in their cohorts (infant) (235). The 1,950 metagenomes were collected from 14 countries, 12 body sites, and 19 health and disease conditions (Table 1). The average mapping rate was 4.72E-05 (8.16E-09 min, 1.1E-02 max).

Gut samples comprised the majority of the dataset (84%), followed by conjunctiva (5%), saliva (3%), and throat swab (1.5%). Among the diseases evaluated, Crohn's disease (CD), ulcerative colitis (UC), end-stage renal disease (ESRD), type 1 diabetes (T1D), and type 2 diabetes (T2D) accounted for 779 fecal samples, whereas 500 fecal samples were obtained from healthy individuals.

All biological specimens were extracted by at least 10 different protocols, for which mechanical lysis, previously reported as a crucial step during the DNA extraction process to recover an optimum microbial diversity (9), was applied in 1,049 samples (53.8%).

5.1.3 Fungal community structure, diversity, and functions of the 1950 metagenomes

Five phyla, 232 genera, and 475 species were identified in the 1,950 metagenomes. More than 80% of the sequences were represented by two phyla (Ascomycota and Basidiomycota), two genera (*Saccharomyces*,

Candida), and three species (*Saccharomyces cerevisiae*, *Candida albicans*, *Malassezia restricta*) (Fig. 7). Under healthy conditions, the gut mycobiome was dominated, in terms of relative abundance, by *Saccharomyces cerevisiae*, which was detected in 52.4% of the samples, while *Dacryopinax primogenitus* was found in 23.6%, *Yarrowia lipolytica* in 13.6%, and *Candida parapsilosis* in 11% of the samples. *C. albicans*, known as an opportunistic pathogenic yeast (6), was found in only 4% of the GI tract samples of healthy individuals. The fungal species profiling data can be found in Supplementary Table 4. *Malassezia* predominated conjunctiva samples, whereas *Aspergillus* predominated the saliva mycobiome.

The number of observed fungal species in the 1,950 metagenomic samples ranged from 1 to 40 (median of 2), Chao1 index (236) varied between 1 and 76.1 (median of 3), and the Shannon index (237) ranged from 0 to 3.36 (median of 0.62) (Supplementary Table 5). These three measurements indicated that the fungal community in humans is, in general, of very low diversity compared with the bacterial community, which could reach an average of 70 in terms of the Chao1 index (236).

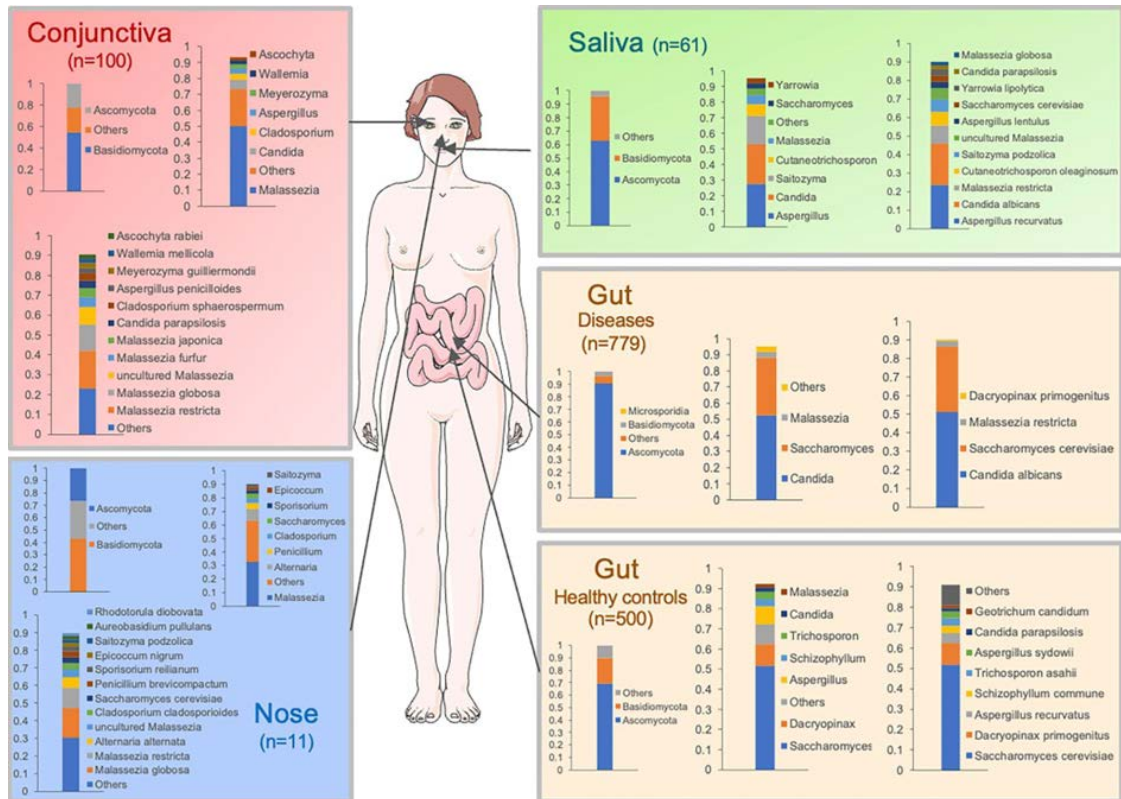


Figure 7. Fungal taxonomic profiling of several human body sites based on the 1950 shotgun metagenomic data using the FunOMIC-T database. Taxonomic profiling is displayed at the phylum, genus, and species levels. Only the mean relative abundance of the genera and species summing 90% of the sequence data is exhibited. Gut taxonomic profiling was performed for diseases including Crohn’s disease (CD, n=193; from the USA, Europe, and Asia), ulcerative colitis (UC, n=79 from Europe and the USA), end-stage renal disease (ESRD, n=208, from Asia), type 1 diabetes (T1D, n=60 from Australia), and type 2 diabetes (T2D, n=99 from Asia). 468 fecal samples did not have health status information in the metadata files. The health status and geo-localization of the conjunctiva, nasal, and saliva samples are described in Table 1.

While fungal taxonomic profiling of human microbial communities has increased considerably over the last 10 years through the sequencing of phylogenetic marker genes such as ITS2/18S, the fungal community function was scarcely investigated mainly due to, again, the lack of a comprehensive

database. Using FunOMIC-P, we annotated the sequencing reads of the 1,950 human metagenomes using the DIAMOND aligner. In total, 1,948 metagenomes were successfully mapped to the database, and the average mapping rate was 0.088% (5.42E-04% min, 1.2% max), consistent with that previously reported in Qin et al., for eukaryotic DNA (2).

Sixteen pathway classes and 120 pathways were detected from the metagenomes. Five pathway classes (Amino Acid Metabolism, Carbohydrate Metabolism, Nucleotide Metabolism, Energy Metabolism, Metabolism of Cofactors and Vitamins) and 29 pathways (list in Supplementary Table 6), along with unidentified pathways and pathway classes represented more than 80% of the sequences. The pattern of fungal functional structure indicated higher evenness compared with fungal taxonomic structure, i.e., the relative abundances of the pathways are closer instead of being dominated by one or two pathways.

5.1.4 Association between metadata and mycobiome composition and functions

Next, we evaluated the contribution of available variables, collected from the metadata files, to the mycobiome variations using the `adonis2` function from the `vegan` R package (Fig. 8). These variables included countries, health status, body sites, ages, gender, and bead-beating. Individually, countries and health status were the factors that contributed most to fungal composition and function variations; body sites and the bead-beating step also contributed to these variations, but to a lesser extent (FDR < 0.01, Fig. 8).

Associations between these variables and individual taxa were then examined using generalized linear models implemented in the `MaAsLin2` (Microbiome Multivariable with Linear Models) package. Five fungal species (*Aspergillus recurvatus*, *Malassezia restricta*, *Saccharomyces cerevisiae*, uncultured *Malassezia* spp., *Yarrowia lipolytica*), which were among the 10 most prevalent and abundant fungal species (Supplementary Table 4), were found associated

with health status, country, and body sites (Supplementary Figure 1A). This finding suggests that the high variability of the human mycobiome could be linked to these five species. Interestingly, *Yarrowia lipolytica* was found positively associated with the employment of the bead-beating step during DNA extraction from the biological specimens (Supplementary Figure 1B), which could be explained by its relatively higher fraction of chitin (10.3-18.9%) in the cell wall compared with *S. cerevisiae*, *C. albicans*, and *M. restricta* (238, 239, 240).

We found that geography, health status, and body sites had marked effects on the variability of most of the fungal pathway classes among the 16 that we recovered from all samples, yet bead-beating did not impact the compositions of fungal pathways, as reported for fungal taxa (Supplementary Figure 1).

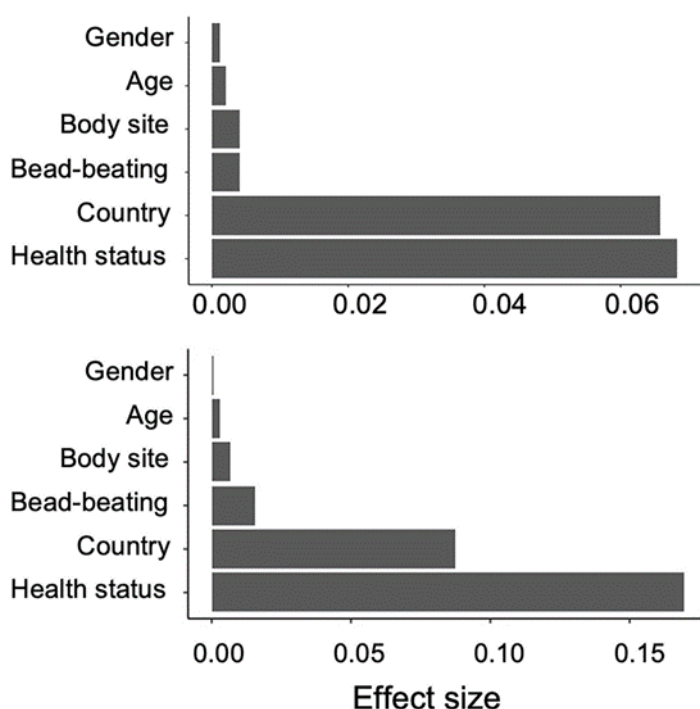


Figure 8. Effect size of variables on the mycobiome community. The impact of the covariates on mycobiome composition (A) and function (B) was tested by performing a univariate analysis (adonis2) on the 1,950 metagenomes. The effect was considered significant when $FDR < 0.05$.

5.1.5 Core taxonomic fungal microbiomes of different body sites and different countries

To identify groups of key taxa that may influence the microbiome community, we applied the concept of core microbiome across body sites and geography, taking into account health status. For this purpose, fungal species with an occurrence of over 50% in the respective set of metagenomes of interest, in which fungi were detected, were defined as the core mycobiome. The 50% occurrence threshold was chosen based on the review of the core bacterial microbiome published by Neu et al., in 2021 (241), but an abundance cut-off was not applied to avoid missing any lowly abundant fungal species. We summarized the core mycobiome for body sites (Table 2) and countries (Table 3). In the human gut mycobiome of non-infants, *S. cerevisiae* was found to be the only member of the core gut mycobiome, except for CD and T1D patients who were dominated by *Aspergillus recurvatus*. The core gut mycobiome of infants consisted of only one species from the *Malassezia* genera, in accordance with several previous studies (235, 242). In other body sites, except saliva, several *Malassezia* species were the most detected members of the core mycobiome. The saliva mycobiome was driven by *Aspergillus recurvatus*.

Table 3. Core fungal species of different body sites.

Bodysite	Health status	Core fungal species (>50% prevalence)
Gut	HC (n=262)	<i>Saccharomyces cerevisiae</i>
	CD (n=109)	<i>Aspergillus recurvatus</i>
	ESRD (n=106)	<i>Saccharomyces cerevisiae</i>
	UC (n=55)	<i>Saccharomyces cerevisiae</i>
	T1D (n=40)	<i>Aspergillus recurvatus</i>
	T2D (n=50)	<i>Saccharomyces cerevisiae</i>
	PSO (n=16)	<i>Saccharomyces cerevisiae</i>
	PLWH (n=7)	<i>Saccharomyces cerevisiae</i>
	Infant (n=14)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>
	Nasal mucosa	HC (n=5)
Conjunctiva	HC (n=76)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>
Saliva	NA (n=38)	<i>Aspergillus recurvatus</i>
Throat swab	HC (n=14)	<i>Schizophyllum commune</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>
	SCZ (n=12)	<i>Candida albicans</i> , <i>Malassezia restricta</i>
Tongue dorsum	Infant (n=8)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>
Bones and joints	BJIs (n=24)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>
Gallstone	GS (n=8)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>

Table 4. Core fungal species of different countries.

Country	Health status	Core fungal species (>50% prevalence)
Australia	HC (n=46)	<i>Aspergillus recurvatus</i>
	T1D (n=40)	<i>Aspergillus recurvatus</i>
Belgium	CD (n=76)	<i>Aspergillus recurvatus</i> , <i>Saccharomyces cerevisiae</i>
China	ESRD (n=106)	<i>Yarrowia lipolytica</i>
	T2D (n=49)	<i>Saccharomyces cerevisiae</i>
Canada	PLWH (n=7)	<i>Saccharomyces cerevisiae</i>
Denmark	HC (n=118)	<i>Saccharomyces cerevisiae</i>
Italy	Infant (n=14)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i>
Spain	HC (n=57)	<i>Dacryopinax primogenitus</i> , <i>Saccharomyces cerevisiae</i>
	CD (n=38)	<i>Saccharomyces cerevisiae</i>
	UC (n=52)	<i>Saccharomyces cerevisiae</i>
USA	PSO (n=16)	<i>Saccharomyces cerevisiae</i>

Given that geographical difference contributes the most to fungal taxonomic structure variations, we also defined the core mycobiome for gut samples collected in different countries. We focused only on gut samples, as they represented the most available samples. *S. cerevisiae* appeared as a member of the core gut mycobiome in most countries (Table 3), which is in agreement with the aforementioned core mycobiome (Table 2). *A. recurvatus* was the only core fungal species among all the gut samples with different health status collected from Australia, whereas *Y. lipolytica* was that of the gut samples collected from end-stage renal disease (ESRD) patients in China (Table 3).

Core biochemical pathways, defined as pathways that have occurrences over 99% among all the samples with a relative abundance of over 1% (13), were also summarized for each body site and country with different health status (Supplementary Table 7). For countries, only gut samples, as the most available sample type, were considered. The majority of core fungal pathways were related to nucleotides, amino acids, energy, and carbohydrate metabolisms, which are essential functions, indicating that the functionality of the human mycobiome is maintained across body niches and populations.

5.1.6 Bacterial and fungal microbiome interaction

Next, we sought to evaluate the correlations between fungal and bacterial

taxonomic composition in gut samples under healthy conditions, especially concentrating on core fungal species. Because of the failure in detecting the core mycobiome under healthy conditions from China, we focused on the healthy conditions of Denmark and Spain. To address this aim, we first performed bacterial taxonomic and functional profiling of the metagenomic data. Due to a very extensive computational time requirement (6 hours/40 CPUs/sample on average), only a subset of 1,485 of the 2,679 metagenomic samples was processed (Fig. 6). We then carried out a correlation analysis with the SparCC correlation method, which handles compositional data (196). In total, 4,184 significant ($p < 0.05$) inter-kingdom correlations were found in the Danish cohort, while 3,471 significant inter-kingdom correlations were found in the Spanish cohort, (Supplementary Table 8). In the Spanish cohort, the two core fungal species, *S. cerevisiae* and *D. primogenitus*, were found to correlate with the bacterial species *Haemophilus pittmaniae* positively and negatively, respectively (Fig. 9A). Beyond that, in the Spanish cohort, *C. albicans* was found to negatively correlate with *Megasphaera sp MJR8396C*, which was positively correlated with *D. primogenitus*. *C. albicans* was also found negatively correlated with *Lactobacillus sanfranciscensis*, *Bifidobacterium scardovii*, *Desulfovibrio fairfieldensis*, *Ruminococcus sp CAG563*, *Coprococcus catus*, and *Roseburia sp CAG309* (Supplementary Table 8, Fig. 9A), many of which are potential short-chain fatty acid (SCFA) producers (243). In the Danish cohort, significant correlations were found between the only core fungal species, *S. cerevisiae*, and seven bacterial species, of which five were negative (*Tropheryma whipplei*, *Prevotella sp CAG1124*, *Firmicutes bacterium CAG24*, *Gemella sanguinis*, and *Sutterella parvirubra*) and two were positive (*Bacteroides nordii* and *Prevotella stercorea*) (Fig. 9B).

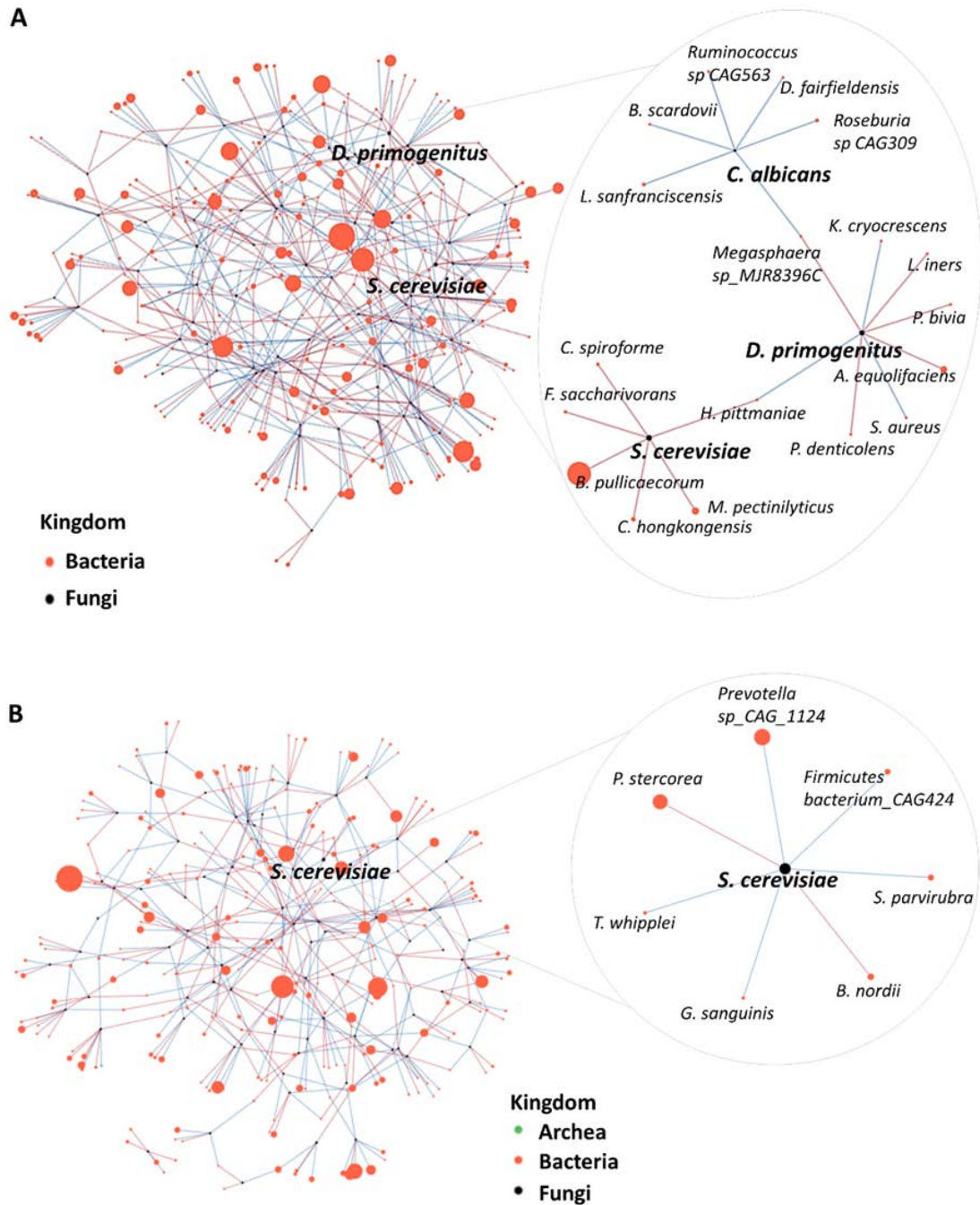


Figure 9. Interaction of fungal and bacterial communities in gut microbiome under healthy conditions. Correlation network between the relative abundance of fungal and bacterial species in the gut mycobiome under healthy conditions from Spain (A) and Denmark (B) using the SparCC algorithm. Each node represents a fungal/bacterial/archaeal species, and their sizes are determined by relative abundances. The colors of the edges connecting two nodes represent positive (red) and negative (blue) correlations.

For a better visual effect, only correlations with p-values less than 0.001 and an absolute correlation coefficient over 0.05 are represented.

We also applied SparCC to analyze correlations between fungal and bacterial functions in gut samples under healthy conditions. In the Danish cohort, 93 significant correlations were detected (Supplementary Table 8, Supplementary Figure 2A), of which the strongest was the positive correlation ($\rho=0.06$, $p<0.001$) between the biosynthesis of secondary metabolites in fungi and the endocrine system in bacteria. In the Spanish cohort, 76 significant correlations were detected (Supplementary Figure 2B), the strongest was a negative correlation ($\rho=-0.13$, $p<0.001$) between carbohydrate metabolism in fungi and signal transduction in bacteria. These functional inter-kingdom correlations could explain how bacteria and fungi interact in the microbiome community.

5.2 Description and usage of MycoDM web server

5.2.1 Home page

The home page is split into two sections (Fig. 10A). The upper section consists of three compartments. The left includes a list of checkboxes based on the metadata of the built-in metagenomes where users can select a subset of the samples by checking the boxes. Then the top right panel with an interactive world map will show the highlight of the corresponding geographical spots. In the meantime, the bottom right panel will display the descriptive summary of the selected subset metagenomes, including the distribution of the sequencing depth, the number of samples per age zone, the pie chart of different genders, and the bar plots of different health status. In the lower section, users can find a brief description of the MycoDM web server, the link to our GitHub page, and the citation and license information.

5.2.2 Download page

The Download page provides the downloading function of both the taxonomic and functional databases of FunOMIC2. The selection bar on the top section lets users choose which database they want. The search box beside the selection bar allows querying the database by taxon name. The lower section is split into two panels. The left one includes a hierarchical file system based on the taxonomic tree that allows a simple navigation by clicking the desired clade name. Once the user has selected a clade, the right panel will display the corresponding pie chart of the lower phylogenetic composition of the selected clade. An example is shown in Fig. 10B; by searching “Candida”, all the clades that match this string are highlighted in the tree-view file system on the left, and all the species under the genera *Candida* are shown proportionally in the right panel. A download button is also provided to let users download the ready-to-use database as well as the full phylogenetic tree of all the 3,031 fungal species that FunOMIC2 covers.

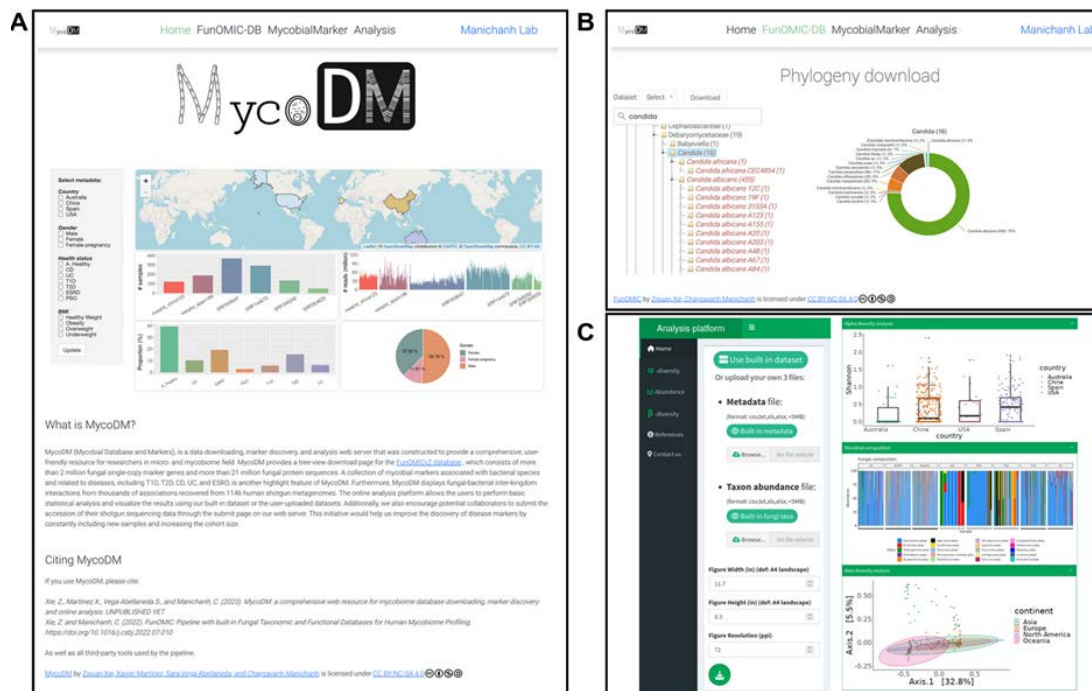


Figure 10. Screenshots of the home page, download page, and the analysis platform.

5.2.3 Analysis platform

We created a user-friendly web platform (Fig. 10C) for performing statistical analysis and visualizing both the fungal and bacterial microbiome features. Currently, MycoDM provides three web apps and the underlying R code is available on GitHub (<https://github.com/ManichanhLab/mycodm>), allowing users to download and run locally on their own datasets. Users can choose to use our built-in dataset from the 1,146 gut metagenomes or upload their own dataset and metadata. Once users select a dataset, detailed information will be provided, including:

Taxa bar plots. Stacked taxa bar plots for both fungal and bacterial microbial abundance. Users can choose their desired taxonomic level, from phylum to species, and choose any one of the metadata variables to group the samples. Each stacked bar represents a metagenomic sample, and different colors represent different taxa. Users can download the generated taxa bar plot by in jpg, peg, or tiff format by clicking the provided download button.

Alpha diversity. Box plots of alpha diversities of both fungal and bacterial taxonomy in different groups based on the selection of metadata variables. Our platform provides choices of three alpha diversity indices: observed number of species, Chao1, and Shannon. The diversity indices and metadata grouping variables are optional so that users can adjust the visualization according to their needs to compare the alpha diversity among different groups. Wilcoxon test results are also provided under the picture with the corresponding p-value. The diversity table and the boxplot of the alpha diversities can be downloaded by clicking the provided download button.

Beta diversity. Principal coordinate analysis (PCoA) plots are displayed for both the fungal and bacterial microbiomes. Our platform provides Bray-Curtis dissimilarity as the default distance metric. The current computational capability does not allow the use of UniFrac distance, however, users can choose to

download the phylogenetic tree from the Download page for calculating the UniFrac distance matrix locally. The distance metrics and metadata grouping variables are optional so that users can adjust the visualization according to their needs to compare the beta diversity among different groups. PERMANOVA analysis results are also provided under the picture with the corresponding p-value. Users can download the generated PCoA plot by in jpg, peg, or tiff format by clicking the provided download button.

5.2.4 MycobialMarkers page

We annotated the 1,146 collected human gut metagenomes using FunOMIC2 and HuMANn3 for the fungal and bacterial microbiome profiling, respectively. Then we searched for the fungal markers associated with variables, especially health status, from the metadata by using the random models fit by the MaAsLin2 and the ANCOMBC R packages. The variables country, study, and BMI were used as the random effect for all models. We also sought the interaction between fungal and bacterial microbiomes in terms of composition and function by performing the SparCC correlations. For both analyses, results were considered significant if the false discovery rate (FDR) was lower than 0.05. All the significant fungal markers associated with a certain disease are reported on this MycobialMarkers page. This page provides one home page and two sub-pages. The main page provides a short description of the mycobial markers and a submission section (Fig. 11A).

The two subpages were installed with a search engine and interactive tables to query detailed information about associations between human mycobioime and diseases or human bacterial microbiome. The interactive table has five attributes: "Mycobial-Marker", "Disease", "Alteration", "Inter-kingdom", and "Sample source". "Mycobial-Marker" lists all the significant mycobial markers, either taxonomic or functional, "Disease" lists the associated disease with the corresponding mycobial marker, "Alteration" lists the change of the certain mycobial marker in the microbiome of the patient with the corresponding

disease compared to the healthy controls. The “Inter-kingdom” provides links to a sub-table which contains the bacterial species that is correlated with the mycobial marker if it exists, as well as the correlation coefficient. The “Sample source” lists the type of the samples where the specific microbial marker was detected. Under each column title, a filtration bar is provided for users to select a specific category and check the subset. The search engine also offers the possibility to display associations by querying a keyword of a microbe or a disease. After entering the input query, a subset of the entries in the database will be returned if the search engine finds match with the item. For example, in Fig. 11B we show all the associations between ESRD and gut mycobiome taxonomy and functions. A download button is also provided for users to get the table of the markers.

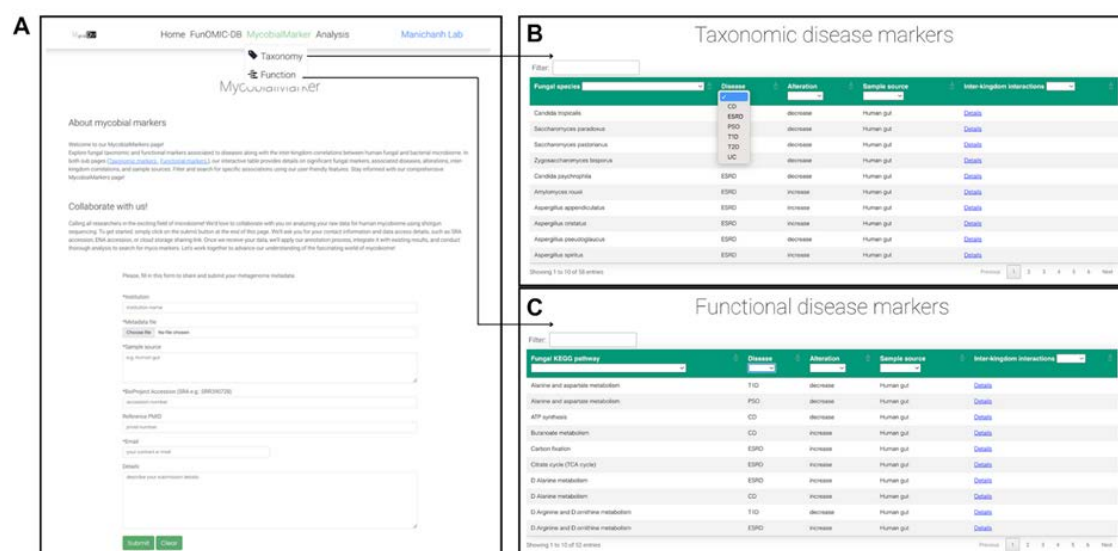


Figure 11. Screenshots of the MycobialMarker page. (A) The main page that includes a short description and a submission section. (B) The subpage that includes the taxonomic mycobial disease markers. (C) The subpage that includes the functional mycobial disease markers.

5.3 Robust integration of fungal and bacterial gut microbiome with dietary data in a longitudinal setting

5.3.1 Shotgun metagenomics sequencing provides higher accuracy than ITS amplicon sequencing in mycobiome profiling at the species level

Inaccuracy in the genome assembly of the ribosomal region of the fungal genomes. To analyze the copy number variability of the ITS region, we recovered 260 assembled fungal genomes covering seven fungal species, known to be relevant in human microbiome studies: *Saccharomyces cerevisiae*, *Aspergillus flavus*, *Candida albicans*, *Candida glabrata*, *Cryptococcus neoformans*, *Rhodotorula mucilaginosa*, and *Rhizopus oryzae* (3). From each of these genomes, we calculated the copy number of the ITS regions using hidden Markov models (CN-HMM) and an in-house bioinformatic pipeline. All species, except *S. cerevisiae*, presented a very low copy number of ITS (average of 2) and this did not vary much across species. This observation is not in agreement with a previous study that reported a high number of ITS copies, ranging from 14 to 144,216 (Supplementary Table 13). Furthermore, we recovered only one copy for most of the *C. albicans* strains. However, *C. albicans* is well-studied and has been shown to carry 21 to 200 copy numbers (CN) per genome (244, 245, 246). The other five species presented also a very low mean CN-HMM value (Supplementary Table 13). Together, these results suggest an inaccuracy in the assembly of fungal genomes, at least in the region of the ribosomal genes.

High inter- and intra-species variability in the ITS copy number. Genomes that contain many repetitive sequences have usually been difficult to assemble when short sequence reads have been generated. Indeed, during assembly, repetitive regions such as the ITS regions are algorithmically collapsed into only a few sequences due to their similarity, leading to a potential bias in the CN-HMM estimation. To circumvent the bias introduced by the incomplete fungal

genomes, we used a mapping depth method for estimating ITS copy numbers (CN-MD) (Fig. 12a), for which more details are described in the Methods section. Before estimating the ITS copy number, we first validated the mapping pipeline. Ten genome assemblies of *S. cerevisiae* were selected for this validation, and their ITS copy numbers were retrieved from the NCBI nucleotide database to work as the expected ITS copy number. At the same time, we generated 15 million simulated shotgun sequencing reads of the 10 assemblies using the InSilicoSeq tool (30016412). The simulated reads were then used as the input of the mapping pipeline for calculating the CN-MDs for each of the genomes. At last, the calculated CN-MDs were compared with the reference copy numbers, by applying the Student t-test. The comparison between the two values did not show significant differences (Supplementary Table 14, p-value = 0.28), which indicates that the pipeline could reliably recover the expected ITS copy numbers from whole genome shotgun sequencing reads.

Next, we applied the mapping pipeline to estimate the CN-MDs of the 260 assembled fungal genomes using their shotgun sequencing reads downloaded from NCBI or JGI. The resulting CN-MD ranged from 7 to 170, with an average of 60 (Supplementary Table 15). We observed that both the intra- and inter-species variability was high for the ITS copy numbers of the analyzed genomes which cover seven species and three phyla (Figure 1B). The copy numbers of the ITS region of the 32 collected *S. cerevisiae* strains were widely distributed, ranging from 15 to 137 (Figure 1C) and those of the 182 *C. albicans* strains varied from 11 to 74. The variance between *C. albicans* and *S. cerevisiae* was significantly different (p-value = $8e-10$; Levene's test). These findings indicate that possible bias could be introduced when profiling the fungal community by using ITS amplicons without normalizing by the actual strain level ITS copy numbers.

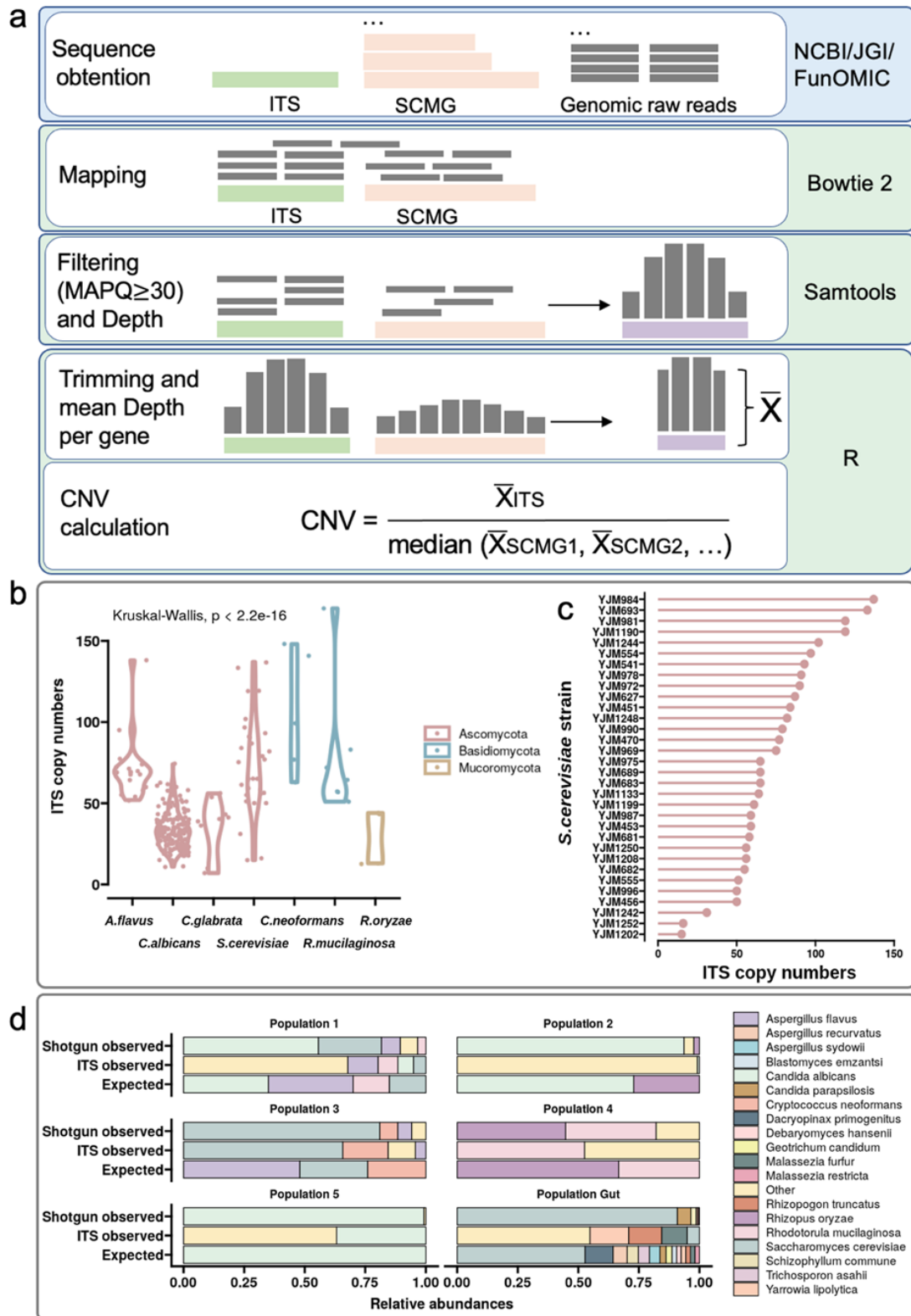


Figure 12. Shotgun sequencing provides higher accuracy than ITS sequencing in mycobiome profiling at the species level. a, Workflow of the mapping depth approach to recover ITS copy numbers. b, The distribution of

CN-MD (y axis) of the ITS across the strains of the 7 analyzed species (x axis) (n = 260). c, The intraspecific CN-MD of *S. cerevisiae* (x axis) for the 32 strains analyzed (y axis) (n = 32). d, *In silico* comparison at species level between expected abundance (“Expected”) and observed abundance by using the shotgun method (“Shotgun observed”) and the ITS method (“ITS observed”).

Shotgun data are more accurate than ITS data for taxonomic profiling at the species level. To compare the accuracy of the species-level mycobiome profiling generated by ITS sequencing and shotgun sequencing, we created *in silico* mock communities with different groups of fungal species. A total of 27 artificial fungal genomes with known ITS copy numbers were used to create five *in silico* mock communities. We randomly generated relative abundances for the species in each of the five communities (Supplementary Table 12). The artificial genomes and their ITS sequences were used to simulate the shotgun sequencing reads and the ITS sequencing reads, respectively. An additional mock community mimicking the gut mycobiome was also created using the 14 most abundant gut fungal species and their observed relative abundance based on a previous study (3) (Supplementary Table 12). The annotations for both sequencing methods were done using QIIME2 and FunOMIC pipelines for ITS and shotgun reads, respectively.

To compare the efficiency of the two methods in performing taxonomic profiling, we calculated weighted UniFrac metrics, which then allowed us to test whether phylogenetic lineages between samples were significantly different. The metrics were calculated between the observed taxonomic profiling generated from both sequencing methods and the fixed relative abundances of the six mock communities, at the species and genus levels. At the genus level, the results showed that the two methods were not significantly different (p -value = 0.623, Student t-test). The ITS method exhibited a mean distance of 0.263 and the shotgun method exhibited a mean distance of 0.213 (Supplementary Table

16), which indicates that both methods showed similar accuracy in taxonomic profiling at the genus level. However, at the species level, the mapping results (Fig. 12d) showed that the two methods differed significantly (p -value = 0.005, Student t -test). The ITS method exhibited a mean distance of 0.616 and the shotgun method a mean distance of 0.237 (Supplementary Table 16), indicating that the shotgun method was able to recover the expected fungal community compositions more reliably at the species level. The same analysis at the strain level was also employed, however, the results revealed that neither shotgun nor ITS sequencing was accurate enough to detect the specific strains.

5.3.2 A fungal enrichment protocol effectively concentrates fungal cells in human fecal samples

As demonstrated by previous studies (2, 3), the proportion of fungal sequences obtained upon shotgun sequencing of DNA prepared from human fecal samples consists of less than 0.08% of the total sequences, which would likely introduce bias to the shotgun sequencing results if the sequencing depth is not high enough. However, the cost of deep shotgun sequencing is still not easily affordable by all researchers. We thus proposed an enrichment protocol based on a series of centrifugations to separate fungal and bacterial cells prior to the regular DNA extraction method.

To evaluate the practical efficiency of this enrichment protocol, we collected fecal samples from six healthy volunteers that included three females and three males. Each of the volunteers donated their fecal samples weekly during an eight-week span, making up a batch of 48 fecal samples. Then for each of the 48 samples, two aliquots of 500 mg were kept, from which one aliquot underwent the enrichment protocol to be separated into a fungal enriched partition and a bacterial enriched partition, while the other aliquot did not pass any further operation and was used as the unenriched control. Finally, a total of 143 partitions (one of the volunteers did not provide enough feces for two aliquots) of fecal samples were sent for shotgun sequencing using the Illumina

Novaseq 6000 platform (Fig. 13a). The sequencing provided an average of 6.4 Gb, and 21.5 million pair reads which are comparable with other studies using shotgun sequencing (126, 247, 248). Next, we annotated the bacterial and fungal communities for all 143 samples using HUMANN (180) and FunOMIC pipelines(3). A total of 411 bacterial species and 208 KEGG pathways were found in the bacterial community, and 91 fungal species and 154 KEGG pathways were found in the fungal community. To assess whether the sequencing depth was sufficient to recover the majority of both fungal and bacterial richness, we selected eight samples that had the highest number of Gb to perform rarefaction curves. Each sample was subsampled and annotated with a gradient of sequencing depths. With the cutoff of the 6.4 Gb, around 80% of fungal taxonomic richness, more than 70% of fungal functional richness, 100% of bacterial richness, and almost 100% of bacterial functional richness were recovered, showing that our shotgun sequencing run was able to capture most of the microbiome information (Fig. 13B). The plateau was reached at 7.5 Gb for fungal taxonomy, 15 Gb for fungal functions, and 6.7 Gb for bacterial functions. Together, these results showed that a sequencing depth of 15 Gb would allow the capture of the entire bacterial and fungal communities.

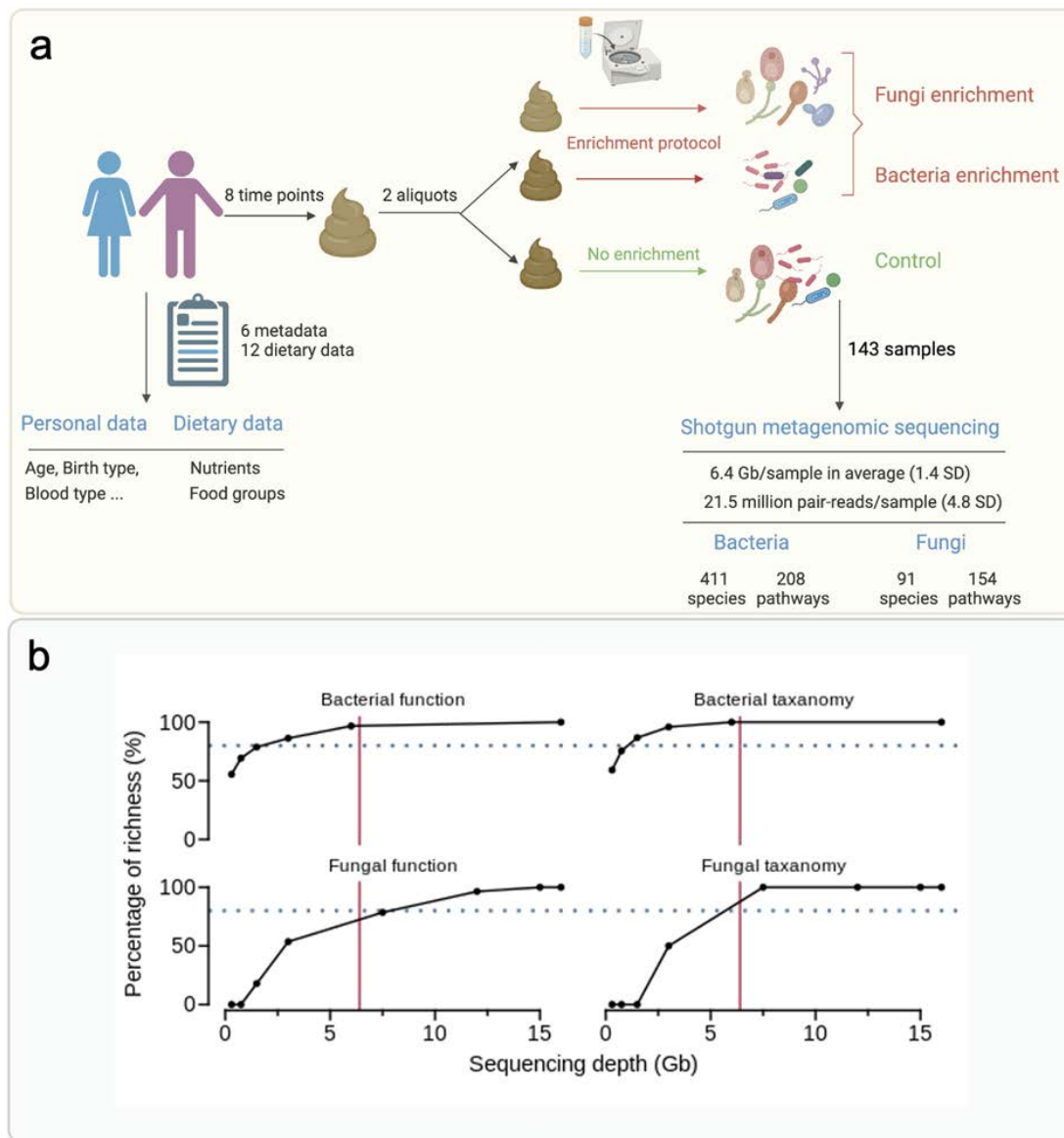


Figure 13. Study design and quality of shotgun sequencing. a, Design and workflow of this study. b, Rarefaction curves of the shotgun sequencing, x-axis represents the depth of sequencing, y-axis represents the percentage of richness. The rows of the panel are different microbial communities, the columns of the panel represent the taxonomic or functional level richness. Each dot is the average percentage of richness of the 8 samples at the specific depth of subsampling. The red solid lines represent the average sequencing depth 6.4Gb, the blue dotted lines represent the threshold of 80% of richness.

Then, we mapped each of the 143 samples to fungal and bacterial databases

(3, 180) to calculate the enrichment efficiencies and meanwhile get their fungal and bacterial, taxonomic, and functional profiling. Notably, the fungal profiling was annotated with the unpublished updated version of the FunOMIC database that contains 2 million single-copy marker genes and 21 million protein sequences extracted from more than 3000 fungal species. In total, we have detected fungi in 96 samples out of 143 (67%). We observed that by applying the enrichment protocol, the proportion of samples that have fungi detected have increased from 58.3% (28 out of 48) to 95.8% (46 out of 48). We then calculated the ratio of the reads that were mapped to the fungal database against the reads that were mapped to the bacterial database for all 143 samples. Then, we used this ratio in the fungal partition divided by this ratio into their corresponding control partitions to estimate the extent to which the fungal sequences have been enriched. The ratio increased on average 18.47 times (ranging from 0.07 to 235) after applying the enrichment protocol, and the fungal alpha diversity in fungal enriched partitions was found significantly higher than in both bacterial ($q = 4.3e-5$ Shannon index, $q = 2e-7$ Chao1 index) and control partitions ($q = 5.2e-5$ Shannon index, $q = 3.7e-6$ Chao1 index) (Fig. 14a, b).

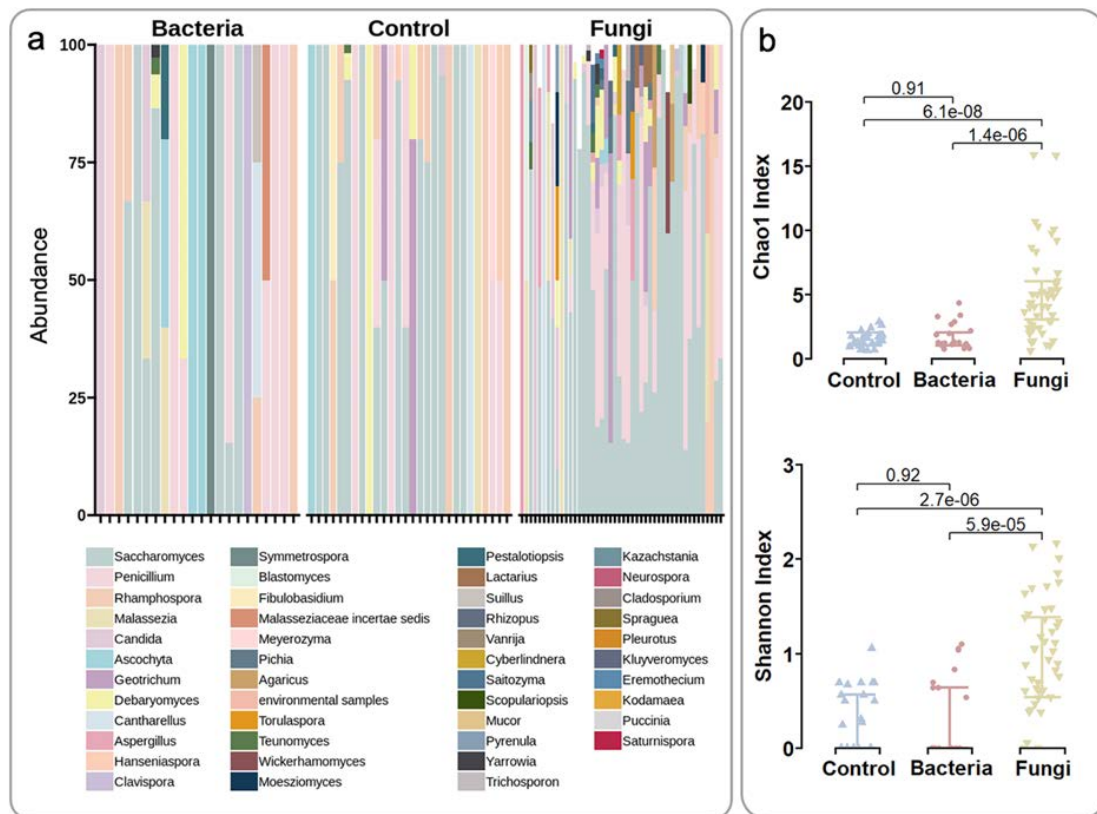


Figure 14. Enrichment efficiency in fungi. (A), The genus-level taxa bar plot of the fungal community compositions in bacterial, control, and fungal partitions. (B), Boxplots of the species-level fungal community alpha diversities (Shannon and Chao1 indices) in control, bacterial, and fungal partitions (n = 96), ordered by their mean from smallest to largest (left to right).

Similar but less significant results were found in the bacterial partitions (Supplementary Figure 3a, b). Since bacterial reads were still present in the fungal partitions, they were merged with those in their corresponding bacterial partitions. After applying a paired Wilcoxon test between the bacterial alpha diversities before and after merging, we found the Chao1 index after merging had a trend of being higher than that of before merging ($p = 0.06$, Supplementary Figure 3c). This result was not observed for the fungal alpha diversity. For the purpose of capturing more information, in the subsequent analysis, the merged bacterial microbiome profiling was used to represent the bacterial community, and the fungal microbiome profiling in fungal partitions

was used to represent the fungal community in each sample.

5.3.3 Keystone bacterial and fungal species in the human gut

To determine the keystone species in the human gut microbiome, we constructed networks based on the SparCC correlation matrix and the corresponding BH-adjusted P-values matrix. In each network, the nodes represent the microbial species that were included in this network, and the edges connecting the nodes represent the significant correlations (FDR < 0.001). This network captured 625 associations among 199 microbial species which includes 111 bacterial species, 87 fungal species, and one Archaea species (Fig. 15a). Among the 625 associations 349 were positive and 276 negative associations. This network consisted of only one large connected group (199 out of 199 microbial species (100%)). The global network had an average node degree (number of edges adjacent to the node) of 6.28 (7.66 for bacteria and 4.5 for fungi), and it perfectly followed a scale-free degree distribution (power law) (Fig. 15b), indicating that most nodes had low-degree values, and only a few nodes had the highest degree values, which are often called "hubs", and are thought to serve specific purposes in the networks.

Candida albicans, a known fungal pathogen (9), was found to form in the gut microbiome seven cross-domain associations with *Ruminococcus gnavus*, *Firmicutes bacterium CAG 110*, *Streptococcus salivarius*, *Holdemanella biformis*, *Eubacterium sp CAG 274*, *Proteobacteria bacterium CAG 139*, and *Alistipes inops*. It also had the highest betweenness centrality (the number of shortest paths going through a node) (n=945), and high node degree (n=11) among the fungal species (Fig. 15c), suggesting a critical role in the gut microbial community. From the species analysis (using betweenness centrality and node degree), we identified one fungal species and 13 bacterial species as potential gut keystone species (Table 1) as they were the species that had both the highest node degree and betweenness centrality (top 20).

Table 4. Keystone microbial species in the human gut

Keystone species	betweenness centrality	node degree
<i>Faecalibacterium prausnitzii</i>	1199.25	25
<i>Bacteroides fragilis</i>	864.43	15
<i>Enorma massiliensis</i>	1100.07	17
<i>Alistipes inops</i>	995.29	17
<i>Prevotella sp AM42 24</i>	654.86	18
<i>Collinsella aerofaciens</i>	789.85	20
<i>Akkermansia muciniphila</i>	1334.21	22
<i>Alistipes putredinis</i>	821.04	16
<i>Dorea formicigenerans</i>	829.1	18
<i>Coprococcus comes</i>	1286.18	20
<i>Holdemanella biformis</i>	1279.61	18
<i>Prevotella copri</i>	817.2	21
<i>Bifidobacterium pseudocatenulatum</i>	961.73	15
<i>Debaryomyces hansenii</i>	832.15	16

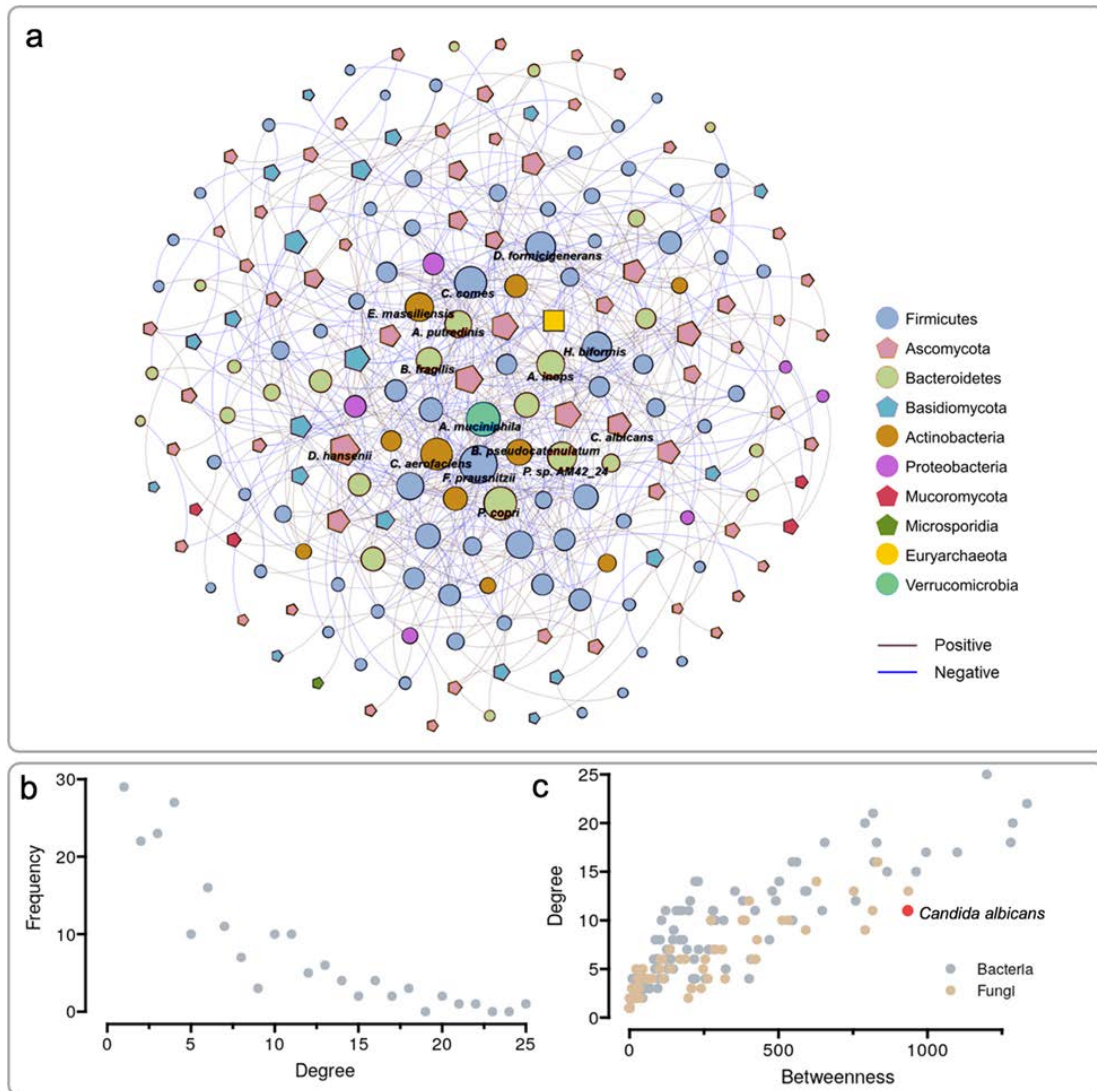


Figure 15. Inter-kingdom network and keystone species. a, Network of the SparCC correlation between the fungal and bacterial taxonomic composition, at the species level (FDR < 0.01). Each node shows a unique microbial species, each edge showing the SparCC correlation between the two nodes linked. The edges connecting the nodes represent significant correlations (FDR < 0.001). The shape of the nodes represents the kingdom of the specific species, and the color represents the phyla. The color of the edges represents the symbol of the correlation. b, Degree distribution of the network following a scale-free distribution. c, *Candida albicans*, one of the fungal species that has high node degree and the highest betweenness centrality.

5.3.4 Short-term dynamics of the human gut microbiome

To determine the intra- and inter-individual variability of the volunteers' gut microbiome, we measured the pairwise dissimilarities using the Bray-Curtis dissimilarity values between longitudinal samples donated by the same volunteer and between samples donated by different volunteers for both fungal and bacterial microbiomes. The results revealed that both bacterial and fungal communities exhibited higher inter-individual than intra-individual dissimilarities (Fig. 16a), while this difference was significantly more pronounced in the bacterial community. We then compared the dissimilarity between fungal and bacterial communities; the variabilities in the fungal community were significantly higher than in the bacterial community (Fig. 16b).

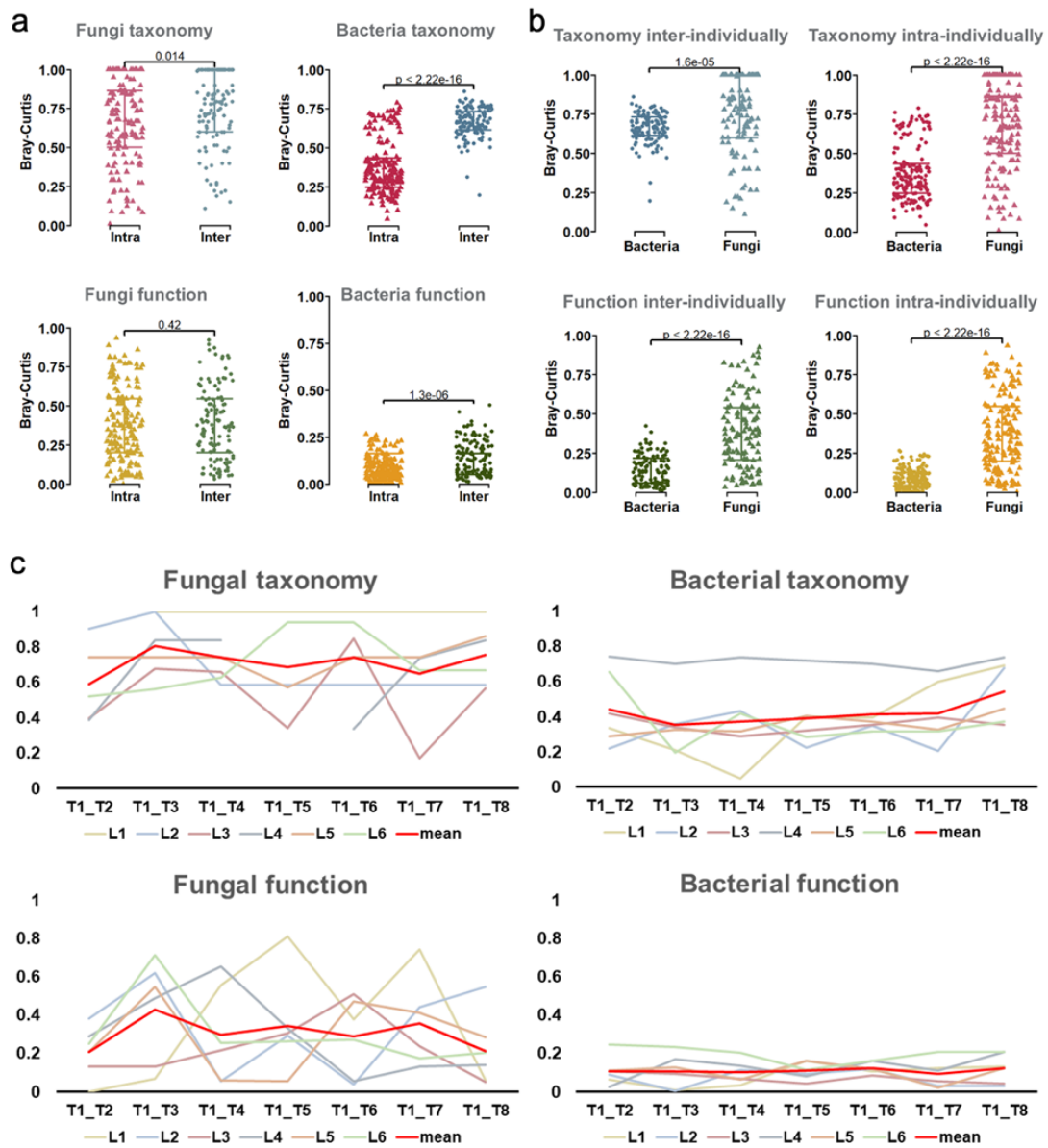


Figure 16. Dynamics of the human gut microbiome. a, Intra- and inter-individual beta diversity (Bray-Curtis) in fungal and bacterial communities at taxonomic and functional levels. b, Comparison of the beta diversities between fungal and bacterial communities intra- and inter-individually at taxonomic and functional levels. c, Dynamics of fungal and bacterial communities at taxonomic and functional levels. The x-axis represents different time points, the y-axis represents Bray-Curtis values.

Then, to investigate the stability of the gut microbiome over time, we considered

the first time point as a baseline, and for each of the individuals, we measured the Bray-Curtis dissimilarities of other time points against the baseline. In both taxonomy and function, despite the high degree of short-term longitudinal change in both communities, we found that the fungal community displayed increased dynamics as compared to the bacterial community (Fig. 16c). Notably, the mean Bray-Curtis values calculated from data of the six individuals were significantly higher for the fungal microbiome than bacterial microbiome (Wilcoxon, $p < 0.01$ for both taxonomy and functions). To determine whether dietary changes drove the dynamics of the gut microbiome, we correlated the pairwise Bray-Curtis dissimilarity of the microbiome (fungal and bacterial, taxonomy and function) with the pairwise Bray-Curtis dissimilarity of dietary data (nutrient macromolecules and food groups). Unfortunately, no significant correlations were found, which may indicate the absence of an effect of the diet on the microbiome composition and function at the global level, but does not exclude the effect of specific food groups or nutrients.

5.3.5 Microbial diversity and composition are associated with habitual diet

We then assessed the correlation between habitual diet (nutrients and food groups) and the alpha diversity of the human gut microbiome to get a broad view of how habitual diet could modulate microbial communities. Interestingly, using the Spearman correlation coefficient, we detected 21 significant associations ($FDR < 0.05$) with fungal taxa while no significant associations were found with bacterial taxa (Fig. 17a). Furthermore, 31 significant associations were found with fungal functions, and five significant associations with bacterial functions (Fig. 17b, c). The overlapped associations found with fungal taxa and fungal function were all consistent, whereas the overlapped associations detected with fungal function and bacterial function were all opposite, indicating that fungal and bacterial communities are likely to act competitively for some dietary products (Fig. 17d).

Next, we calculated the Spearman correlation coefficients between a habitual diet and specific gut microbiome components and functional pathways. We found 23 fungal species were significantly correlated with one or more dietary categories. Among them, the strongest correlations were *Lactarius pseudohatsudake* with biscuits ($\rho = -0.32$, FDR = 0.027), *Penicillium lancoscoeruleum* with fish ($\rho = 0.31$, FDR = 0.027), *Candida albicans* with iron ($\rho = -0.29$, FDR = 0.038) (Fig. 18a). More significant correlations were detected in the bacterial community. At a broad level, we found three apparent groups of species clustered to a group of foods mainly classified as related to more animal-based foods (fish, sauces, sausages, processed food, dairy products) and two others related to less animal-based foods (fruit, vegetables) (Fig. 18b). Similar but less obvious groupings were also found when correlating habitual diet with microbial functions (Supplementary Figure 5, 6).

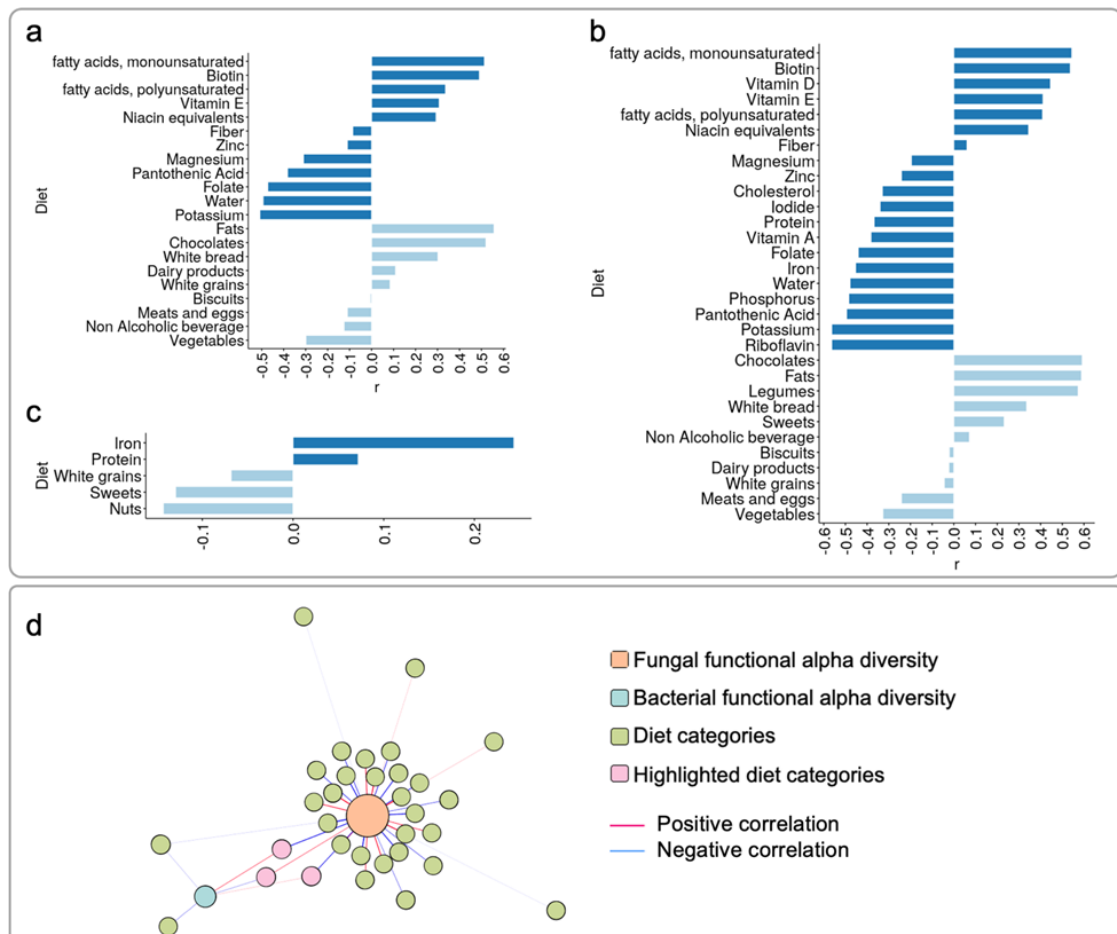


Figure 17. Microbial taxonomic and functional alpha diversities are associated with habitual diet. a, Significant (FDR < 0.05) Spearman correlations found between the fungal taxonomic alpha diversity and diet categories. The x-axis is the value of the correlation coefficient, the y-axis is the name of the diet categories. b, Significant (FDR < 0.05) Spearman correlations found between the fungal functional alpha diversity and diet categories. The x-axis is the value of the correlation coefficient, the y-axis is the name of the diet categories. c, Significant (FDR < 0.05) Spearman correlations found between the bacterial functional alpha diversity and diet categories. The x-axis is the value of the correlation coefficient, the y-axis is the name of the diet categories. d, Network of the fungal functional alpha diversity, bacterial functional alpha diversity, and diet categories detected to be significantly (FDR < 0.01) correlated with them. The edges are the Spearman correlation coefficients between the two nodes linked. The color of the edges represents the symbol of the correlation.

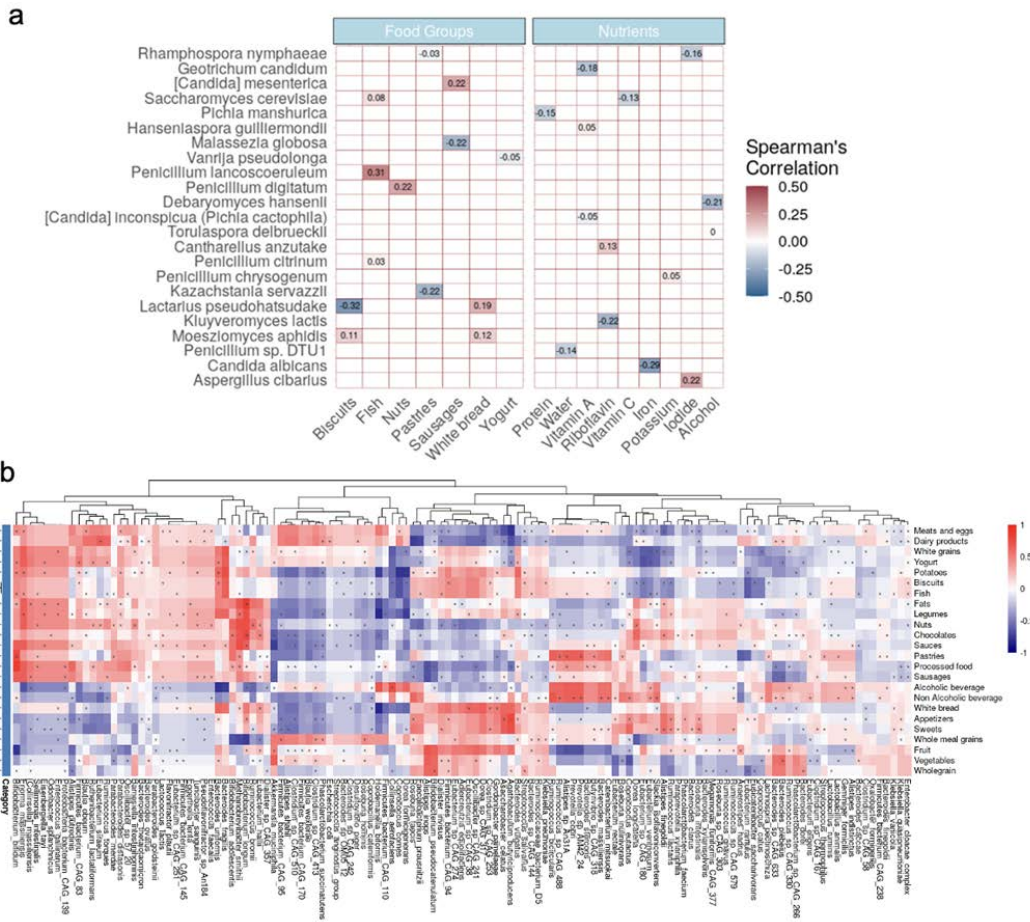


Figure 18. Microbial taxonomic compositions are associated with habitual diet. a, Heatmap of all the detected significant correlations between fungal taxonomic compositions and diet categories. b, Heatmap of all the detected significant correlations between bacterial taxonomic compositions and diet categories. The asterisk indicates that the correlation index for the corresponding species metadata pair is significant. For better visualization, this plot with higher resolutions can be found in Supplementary Figure 4.

6. DISCUSSION

The major objective of this thesis is to develop bioinformatic database and pipeline for the analysis of mycobiome and to apply the developed tool to characterize the human mycobiome, especially in the context of the interplay with the bacterial microbiome.

At the start of this work, very few studies were published in terms of analyzing the human mycobiome using shotgun sequencing data, let alone much less about the corresponding workflow integrating wet and dry lab works and the bioinformatics developed for this field (182, 184). Further, the only two tools we could find for profiling human mycobiome, FindFungi and HumanMycobiomScan, had a common and serious issue in that they utilize the whole fungal genomes as the reference databases, which not only consume substantial computational resources but also have the tendency to introduce bias when quantifying the mycobiome relative abundance by mapping reads to multi-copy reference genes. During the establishment of the database and pipeline of our FunOMIC tool, there was a new tool named EukDetect published. This tool was also constructed based on SCMG, however, it utilized only around 200 SCMGs that were found conserved in all Eukaryotes, which was not as comprehensive as our database. Further, this tool did not incorporate a protein database for annotating human mycobiome functions. Based on several tools designed for profiling the bacterial microbiome, such as MetaPhlan2 (178) and mOTUs2 (194), we designed and implemented our own tool FunOMIC, using SCMGs as the content of the reference database. This new tool represents a significant improvement over previously published mycobiome annotation pipelines, owing to the incorporation of a taxonomic database that encompasses the most comprehensive fungal SCMGs available at the time of publication, in addition to the first protein reference database for functional annotation of mycobiome using shotgun metagenomics.

Besides the databases, we also designed and validated the metagenomic pipeline that integrates quality control, taxonomic profiling (FunOMIC-T), and functional profiling (FunOMIC-P) for a comprehensive analysis of fungi in environmental samples, and, particularly, in humans. The FunOMIC pipeline exploited the read-mapping approach, which, in brief, is a bioinformatics approach to identify and quantify the abundance of microbial species in a metagenomic sample. This approach maps the fastq format short reads generated from a metagenomic sample to reference genomes or databases. The mapped reads and their counts can then be used to estimate the abundance of each microbial species in the sample. The approach used in FunOMIC has a significant advantage in that it preserves all the information from the input files, which would otherwise be lost when using a metagenome-assembled genomes (MAGs) strategy. This advantage is particularly important in metagenomic communities where the amount of fungal DNA is typically low relative to bacterial and human DNA. Assembling such communities is a difficult task, and this approach helps to overcome that challenge.

MAGs approach has become a new trend in microbiome studies, especially for the profiling of bacterial communities for the ability they can provide insights into the unknown of uncultured microbes in the absence of genome isolates (249, 250, 251, 252). Despite the fact that successful MAGs provide a more comprehensive understanding of the functional capabilities of a microbial community, some studies have reported the failure to recover MAGs due to the high intrapopulation strain heterogeneity (253). Meziti et al., have demonstrated that MAGs recovered from their samples missed 25% to 50% of the population core and variable genes on average, which corroborated with another study assessing the bias in genome reconstruction from metagenomic data (200, 254). Beyond that, they also highlighted that the common tools used for checking contaminations in MAGs, such as CheckM (255) and MiGA (256), failed to estimate the consistent contamination of sequences from other

bacterial families recovered by MAGs. We thus doubted the feasibility of applying the MAG approach for profiling human mycobiome, considering that MAGs may be incomplete or contain errors even for bacterial communities which encompass the majority of the sequences. Before the decision of using the read-mapping method, we also tested the possibility of applying the MAG approach in recovering the fungal genomes from human shotgun metagenomics data by using a tool published in 2018 for reconstructing eukaryotic genomes from complex natural microbial communities (257). However, as we expected, all the tested samples did not have successfully recovered fungal MAGs. Our own test proved again our concerns about the proper application of the MAG approach to identify and annotate fungal communities. Therefore, we have determined the combination of the read-mapping method with our comprehensive SCMG database as the pipeline to study human mycobiome. We then validated the efficiency and accuracy of our pipeline in discovering fungal sequences by using different *in silico* shotgun metagenomics mock communities.

To the best of our knowledge, at the time of this writing, FunOMIC offers the most comprehensive coverage of the reference fungal species and functions compared with other existing databases that we have listed in the introduction section for profiling the human mycobiome. Indeed, FunOMIC-T, which contains more than 1.6million fungal single-copy marker genes and covers 1,916 fungal species, exceeds the fungal spectrum of other similar tools. In addition to performing taxonomic profiling with FunOMIC-T, another advance is that we also proposed FunOMIC-P which includes more than 3 million non-redundant fungal proteins, which is, to our knowledge, the first protein database proposed for analyzing human mycobiome functions. Additionally, via the utilization of SCMG, FunOMIC-T provided a smaller-sized taxonomic database that requires less computational resources with more accurate taxonomic

quantification than the full genome-based fungal reference database. The FunOMIC pipeline can also be combined with tools for characterizing bacterial reads, such as MetaPhlAn, to jointly study the taxonomic and functional association of fungal and bacterial microbiome.

We have shown an example of such analysis by applying it to around 2700 real human shotgun metagenomic samples, which represent human microbiomes of different body sites from individuals with different health statuses and from different geographical regions. We corroborated previous human mycobiome results showing that the species *S. cerevisiae*, *C. albicans*, and *M. restricta* dominate the fungal communities in different human body sites (104, 258, 259, 260). We found that geography and health status were the two most important factors contributing to the variabilities of human mycobiome taxonomic and functional compositions. Five fungal species (*A. recurvatus*, *M. restricta*, *S. cerevisiae*, uncultured *Malassezia* spp., *Y. lipolytica*) varied along with different countries, health status, and body sites. *C. albicans*, one of the most common human fungal pathogens (261), negatively correlated with bacterial species that are mainly SCFA producers (243). This finding suggests that therapeutic strategies based on SCFA administration or inducing SCFA producers could be implemented to control *C. Albicans* infection. Although we focused on the human mycobiome and only applied FunOMIC databases and pipeline to human metagenomes, however, the FunOMIC databases cover a wide range of fungal genomes including 1566 Sordariomycetes, 1359 Eurotiomycetes, 843 Dothideomycetes, 681 Agaricomycetes, and 164 Leotiomycetes, which are also found to be the major components of soil mycobiome (262, 263) and marine mycobiome (264, 265). Moreover, the increasing number of downloads (more than five hundreds since 2023) of the FunOMIC databases and pipeline from the built-in downloading counter of the FunOMIC website, along with the high number of visits to the pipeline github page (more than six hundreds since

2023), indicates the growing interest in our database and pipeline.

Along the course of this dissertation, we have noticed a great increase in the public deposition of new fungal genomes. To keep the database up-to-date and to adapt to the newly emerged analysis requirements, we have generated an updated version of the databases and pipeline, FunOMIC2. In this version, the taxonomic database is increased to more than 2 million fungal SCMGs, while the functional database is increased to more than 21 million fungal protein sequences. After the expansion, the FunOMIC2 database now covers more than 3000 fungal species. This great expansion makes the FunOMIC2 not only the more powerful tool for profiling human mycobiome but also broadens its possibility of application to other ecological shotgun metagenomic samples. Along with the updates of the databases, we also included a new step in the pipeline to remove as many bacterial reads as possible prior to the carry out of the mycobiome profiling. Through this step, we can further ensure that the contamination that may be caused by the bacterial sequence in subsequent steps is removed. We also launched the corresponding web server MycoDM at the same time. This web server contains a visualization and tree download page for the FunOMIC2 database, our collection of disease-associated fungal markers discovered through analysis of human metagenomics, and an online analysis platform. We also encourage researchers in the same field who want to cooperate with us to upload their metagenomic data accession numbers to the submission platform of our server, so that we can utilize more shotgun metagenomic data to validate the existing mycobial markers and to discover novel human disease-associated mycobial markers. As far as we know, MycoDM is the first server focused on discovering and sharing fungal markers, and with the upcoming collaboration, we look forward to building a fungal marker retrieval platform with the widest coverage of related diseases in the world.

Among the detected mycobial markers (https://manichanh.vhir.org/mycodm/taxa_marker.php), *Saccharomyces pastorianus* and *Saccharomyces paradoxus* have been found to be negatively correlated with CD, which supports previous findings that the depletion of *Saccharomyces* has been observed in feces from patients with inflammatory bowel disease (IBD). In these patients, *Saccharomyces* was positively correlated with the abundance of bacteria depleted in IBD, such as the butyrate-producing *Roseburia*, *Blautia*, and *Ruminococcus* genera (266, 267). This inter-kingdom correlation was also confirmed in Chapter 1 of this thesis. Additionally, *S. cerevisiae*, which is closely related to *S. paradoxus*, has been shown to have anti-inflammatory effects (267), which may explain the negative correlation between *S. paradoxus* and CD. Our findings show positive associations between several *Aspergillus* species and ESRD, which is a risk factor for developing fungal infections such as invasive aspergillosis. These results were reported previously (268, 269). In addition, *Candida albicans*, which can trigger in vivo inflammatory responses, was found to be enriched in patients with ESRD and T2D, as previously reported by multiple studies (270, 271, 272). In terms of mycobial functional markers (https://manichanh.vhir.org/mycodm/functional_marker.php), our study found an increase in the glutamate metabolism pathway in CD. Glutamate, which is the immediate product of glutamine metabolism (273), plays multiple roles in cells such as being an excitatory neurotransmitter, participating in oxidative metabolism, and regulating metabolic pathways (274, 275, 276, 277). In an inflammation model, the glutamate was demonstrated to improve intestinal barrier function, alleviate inflammation, and inhibit protein degradation through various signaling pathways (278). Given its unique role in the gastrointestinal tract, glutamate may be an adjuvant treatment for IBD with broad application, as confirmed by Li et al (189). These findings suggest that gut fungal

communities may play an important role in reducing the inflammatory response in CD.

Due to the relatively high cost of shotgun sequencing service, most scientists that study mycobiome are still using ITS sequencing metagenomics for research. Also, because of the high variability of ITS copy numbers, we had planned to build an ITS reference database containing different species-specific copy numbers and the corresponding profiling pipeline that includes the normalization of the read counts based on the copy number. However, we have realized that this plan was not feasible since we have evidenced the high intra- and interspecific variabilities of the fungal ITS region at the strain level. Similar results have been reported previously, where 14 to 1,442 ITS copies were found in 91 fungal taxa (203), 22 to 227 copies across the 788 *S. cerevisiae* isolates (209), and 38 to 91 18S rRNA gene copies in 8 *Aspergillus fumigatus* strains (279). Given that the highest resolution of the ITS region barely reaches the species level (280), normalization of the ITS counts cannot reach the strain level, thus, accurate quantification of the fungal community in a complex ecology is impossible. Then based on a series of *in silico* simulations, we have concluded that shotgun sequencing provides higher accuracy than ITS sequencing in mycobiome profiling at the species level. Indeed, the shotgun metagenomic sequencing plus the annotation using FunOMIC2 databases and pipeline offers a relevant alternative. Our comparison of the performance of the ITS sequencing and the shotgun sequencing with *in silico* simulated mock community reads has supported this hypothesis. Though ITS sequencing is always considered a more cost-effective approach in performing taxonomic profiling, with the rapid development of next-generation sequencing technologies, the cost of shotgun sequencing has dropped to a more affordable level, taking into account that shotgun sequencing skips the amplification and amplicon purification steps. In sum, the total cost of both sequencing methods

can differ slightly, while shotgun sequencing is able to capture more information, including the functions of the fungal communities, and the taxonomy and functions of the bacterial communities. Thus, we strongly recommend that researchers in this field switch to the usage of shotgun metagenomic sequencing when studying the fungal microbiome in the future, not only in basic research but also in clinical studies.

The scarcity of fungal populations in the environment has always been a thorny issue in the study of human mycobiome. In addition to the aforementioned method of utilizing the removal of bacterial reads in the post-experimental step, we also performed experimental methods to remove bacteria. To reduce the bias introduced by the low proportion of fungal cells in human fecal samples, we proposed a fungal enrichment protocol, based on a centrifugation approach, that effectively concentrated fungal cells. This protocol successfully increased the detected fungal counts and richness. A membrane filter approach was also tested parallelly when testing the centrifugation method. Both methods utilized the nature that most bacterial cells are smaller than fungal cells. Several cellulose nitrate filters with different pore sizes (0.65 microns, 3 microns, and 5 microns) were used individually to intercept the fungal cells and let go of the bacterial cells. Nonetheless, practically the membrane method was time-consuming and burdensome to implement, for the intercept fungal cells and other impurities that failed to be removed immediately blocked the pores. We therefore terminated and discarded the trial of membrane enrichment.

By implementing a final optimized mycobiome research workflow including centrifugation-based enrichment methods, shotgun metagenomics, and the FunOMIC2 database and pipeline on human fecal samples, we found that our method outperformed other methods in the sensitivity of detecting fungi. For example, in another study that compares amplicon sequencing with shotgun

sequencing (281), the authors only detected fungi in 3.83% of the samples when using MetaPhlan4 pipeline. When applying the FindFungi pipeline, they recovered more fungal reads, however, most of the mapped reads were later evaluated as bacterial contamination. Furthermore, the tool FindFungi identified most of the reads as belonging to *Melampsora pini*7, which is a probable contaminant found commonly in public genome assemblies. This was further investigated and revealed that most of these reads were actually bacterial, specifically those of the genus *Bacteroides*. Therefore, the output of FindFungi is likely inaccurate (281). In contrast, our results detected fungi in 58.3% of the samples without enrichment, and 95.8% of the samples with enrichment.

To the same cohort, we have also applied the HuMANn3 pipeline to get the bacterial taxonomic and functional profiling for studying the interkingdom correlation. The network analysis has suggested candidate keystone microbial species, including 13 bacterial species and 1 fungal species in the human gut environment. The only fungus identified as the keystone species, *Debaryomyces hansenii*, has been implicated as a fungus that is found in Crohn's disease tissue and can lead to dysregulated healing. Crohn's disease is usually characterized by the dysbiosis of the gut microbiome. Bacterial species correlating with *D. hansenii* might play crucial roles in keeping the gut microbiome in a healthy balance. *Faecalibacterium prausnitzii*, *Enorma massiliensis*, *Collinsella aerofaciens*, and *Prevotella copri* were also identified as the keystone bacterial species correlating with *D.hansenii*. *F. prausnitzii* is well known as one of the most abundant and important bacterial species and is also an important butyrate and other short-chain fatty acid producer in the gut microbiome.

We found that the mycobiome was much more dynamic than the bacterial

community at the taxonomic and functional levels, which is consistent with the results found in other studies (126, 282), indicating that the fungal compositions in human gut shift rapidly instead of level off to a stable status like the bacterial community. Since the habitual diet was found to have an influence on the composition of the fungal microbiome in both human and mice models (49, 129, 141, 144), we sought the relationship between the dynamics of habitual diet and the dynamic of the gut microbiome. Although the microbiome changes were not driven by global dietary changes, we showed that microbial diversity, composition, and functions were associated with specific habitual diet composition. We found that bacterial and fungal alpha diversity were oppositely correlated with three diet categories: sweets, protein, and iron. The level of iron in the habitual diet was found to correlate negatively with the fungal functional alpha diversity. To the best of our knowledge, this is the first observation of the effect of iron on the fungal functional alpha diversity, though some studies have discussed that high iron levels promote the growth of specific fungal species (283). Our study highlights a competitive inter-kingdom interaction between bacteria and fungi for nutrients utilization. This finding aligns with a recent study that suggests potential inhibitory actions between gut fungal and bacterial communities during low-carbohydrate diet-induced weight loss (284).

However, it is essential to consider the limitations of this study when interpreting the results, engaging in discussions, and drawing conclusions. Therefore, it is advisable to approach the findings with caution. Important limitations of the FunOMIC pipeline remain in the efficiency of the DNA extraction method and the quality of single-copy marker genes. The latter limitation relies on the completeness of the available fungal genomes, which may result in a lower coverage of fungal taxonomies compared with the fungal amplicon databases (107, 215). Also, one has to be aware that only fungal species with sequenced genomes can be detected. According to the study of Lloyd et al., in 2018, which

claimed that phylogenetically novel uncultured microbial cells dominate earth microbiomes, we have reason to think that the FunOMIC databases does not cover the full fungi kingdom. Improvements in MAGs approach of eukaryotes from human metagenomic samples would enhance the portrayal of un-cultured fungi species in genome databases. As FunOMIC incorporates universal genes, expanding its database as more genomes are sequenced would be a simple process. Another limitation comes from the high inter-kingdom conservation of a portion of protein-coding genes. As a consequence, bacterial contamination was not totally preventable, even after applying an exceedingly strict mapping threshold to the fungal functional annotation with the filtration to remove the majority of bacterial reads before functional annotation. Beyond that, in this study, FunOMIC was only applied to human microbiome data; in the future, applications with soil microbiome, marine microbiome, or other different environmental samples will be launched with FunOMIC to test its ability to handle other environmental data.

Overall, in this dissertation we have assessed the existing analysis approaches for the research of the human mycobiome and its interaction with the human bacterial microbiome. The databases, pipeline, enrichment protocol, and web server that we proposed have together filled the blank of a robust workflow for integrating bacterial and fungal shotgun metagenomics data to characterize the human microbiome and its modulation by dietary components.

7. CONCLUSION

The results obtained in this dissertation allowed us to address the following conclusions:

1, We developed FunOMIC databases and pipeline for the human mycobiome profiling. Our taxonomic database, FunOMIC-T, stands out as the most comprehensive fungal database for mycobiome profiling. Additionally, we have provided the FunOMIC-P, the first database designed specifically for functional mycobiome analysis.

2, By integrating the enrichment protocol, shotgun sequencing, and FunOMIC pipeline, we have successfully expanded the scope of fungi detection within the population and enhanced the alpha diversity of the recovered mycobiome. This workflow has proven to be highly effective in increasing the overall understanding of fungal communities and their impact.

3, Combining the aforementioned workflow with an analysis of bacterial microbiome and dietary data collected using our food frequency questionnaire, allowed us to establish an efficient routine for studying the human gut microbiome integrating robustly fungal, bacterial, and dietary data. This integrated workflow provides a valuable tool for the new era of human microbiome investigation, enabling researchers to gain insights into the complete context rather than solely focusing on the bacterial community.

8. FUTURE LINES

The future directions for this thesis can be focused on three aspects that align with the three main achievements of this thesis.

First, when considering fungal species, it will be important to note that virulence is often limited to specific strains within a given species (285, 286, 287). This highlights the need to focus on strain-level analysis in the study of fungal communities, which can be challenging when using culture-independent metagenomics. However, as similar attempts have been made for bacterial microbiomes, there is potential for developing new pipelines and algorithms to enable strain-level annotation of the fungal microbiome, utilizing the comprehensive collection of the strain-level fungal genomes in FunOMIC databases. Therefore, in the future, an upgraded version of the FunOMIC tool could be developed to provide feasible solutions for the analysis of mycobiome at the strain-level.

Second, the current enrichment protocol has only been tested to a small group of samples collected from human feces. As the field of human mycobiome research continues to grow, there will likely be a need for enrichment protocols suitable for other types of samples, particularly those with low fungal biomass such as vaginal swabs and blood. Therefore, the development of such experimental protocols will be necessary to further expand our understanding of the human mycobiome.

Last, the MycoDM web server will be improved to enhance the ease and user-friendliness of extracting mycobial markers. Strategies include bolstering computational capabilities, updating MycoDM and FunOMIC annually, and incorporating newly discovered markers or fungal genomes. The advancement of high-throughput sequencing technologies will also be incorporated, and additional links between human mycobiome, diseases, and relevant resources

will be provided. Data visualization and analysis platforms will be refined, and new apps such as differential abundance analysis will be added. Mining and integrating metatranscriptomics, metabolomics, or other meta-omics data of human mycobiome to the FunOMIC tool and the MycoDM web server will also be a focus. These results will constitute a first step toward a full-featured, open-source web platform for a systemic view of fungal communities and their interactions with bacterial communities.

We believe these advancements will enhance our understanding of the human mycobiome and its implications in health and disease.

9. BIBLIOGRAPHY

1. Berg G, Rybakova D, Fischer D, Cernava T, Verges MC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*. 2020;8(1):103.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464(7285):59-65.
3. Xie Z, Manichanh C. FunOMIC: Pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling. *Comput Struct Biotechnol J*. 2022;20:3685-94.
4. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol*. 2016;14(8):e1002533.
5. Schiller C, Frohlich CP, Giessmann T, Siegmund W, Monnikes H, Hosten N, et al. Intestinal fluid volumes and transit of dosage forms as assessed by magnetic resonance imaging. *Aliment Pharmacol Ther*. 2005;22(10):971-9.
6. Eve IS. A review of the physiology of the gastrointestinal tract in relation to radiation doses from radioactive materials. *Health Phys*. 1966;12(2):131-61.
7. Wang Y, Moss J, Thisted R. Predictors of body surface area. *J Clin Anesth*. 1992;4(1):4-10.
8. Leyden JJ, McGinley KJ, Nordstrom KM, Webster GF. Skin microflora. *J Invest Dermatol*. 1987;88(3 Suppl):65s-72s.
9. d'Enfert C, Kaune AK, Alaban LR, Chakraborty S, Cole N, Delavy M, et al. The impact of the Fungus-Host-Microbiota interplay upon *Candida albicans* infections: current knowledge and new perspectives. *FEMS Microbiol Rev*. 2021;45(3).
10. Kennedy MS, Chang EB. The microbiome: Composition and locations. *Prog Mol Biol Transl Sci*. 2020;176:1-42.
11. Deo PN, Deshmukh R. Oral microbiome: Unveiling the fundamentals. *J Oral Maxillofac Pathol*. 2019;23(1):122-8.
12. Kort R, Caspers M, van de Graaf A, van Egmond W, Keijser B, Roeselers G. Shaping the oral microbiota through intimate kissing. *Microbiome*. 2014;2:41.
13. Sharma N, Bhatia S, Sodhi AS, Batra N. Oral microbiome and health. *AIMS Microbiol*. 2018;4(1):42-66.
14. Xie G, Chain PS, Lo CC, Liu KL, Gans J, Merritt J, et al. Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol Oral Microbiol*. 2010;25(6):391-405.
15. Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the human oral microbiome during health and disease. *mBio*. 2014;5(2):e01012-14.
16. Wang J, Qi J, Zhao H, He S, Zhang Y, Wei S, et al. Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci Rep*. 2013;3:1843.
17. Beck JD, Offenbacher S. Systemic effects of periodontitis: epidemiology of periodontal disease and cardiovascular disease. *J Periodontol*. 2005;76(11

Suppl):2089-100.

18. Joshipura KJ, Hung HC, Rimm EB, Willett WC, Ascherio A. Periodontal disease, tooth loss, and incidence of ischemic stroke. *Stroke*. 2003;34(1):47-52.
19. Awano S, Ansai T, Takata Y, Soh I, Akifusa S, Hamasaki T, et al. Oral health and mortality risk from pneumonia in the elderly. *J Dent Res*. 2008;87(4):334-9.
20. Hunt RH, Yaghoobi M. The Esophageal and Gastric Microbiome in Health and Disease. *Gastroenterol Clin North Am*. 2017;46(1):121-41.
21. Gagliardi D, Makihara S, Corsi PR, Viana Ade T, Wiczner MV, Nakakubo S, et al. Microbial flora of the normal esophagus. *Dis Esophagus*. 1998;11(4):248-50.
22. Norder Grusell E, Dahlen G, Ruth M, Ny L, Quiding-Jarbrink M, Bergquist H, et al. Bacterial flora of the human oral cavity, and the upper and lower esophagus. *Dis Esophagus*. 2013;26(1):84-90.
23. Pei Z, Bini EJ, Yang L, Zhou M, Francois F, Blaser MJ. Bacterial biota in the human distal esophagus. *Proc Natl Acad Sci U S A*. 2004;101(12):4250-5.
24. Fillon SA, Harris JK, Wagner BD, Kelly CJ, Stevens MJ, Moore W, et al. Novel device to sample the esophageal microbiome--the esophageal string test. *PLoS One*. 2012;7(9):e42938.
25. Pei Z, Yang L, Peek RM, Jr Levine SM, Pride DT, Blaser MJ. Bacterial biota in reflux esophagitis and Barrett's esophagus. *World J Gastroenterol*. 2005;11(46):7277-83.
26. Yang L, Francois F, Pei Z. Molecular pathways: pathogenesis and clinical implications of microbiome alteration in esophagitis and Barrett esophagus. *Clin Cancer Res*. 2012;18(8):2138-44.
27. Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet*. 1984;1(8390):1311-5.
28. Wen Y, Feng J, Scott DR, Marcus EA, Sachs G. The HP0165-HP0166 two-component system (ArsRS) regulates acid-induced expression of HP1186 alpha-carbonic anhydrase in *Helicobacter pylori* by activating the pH-dependent promoter. *J Bacteriol*. 2007;189(6):2426-34.
29. Bauerfeind P, Garner R, Dunn BE, Mobley HL. Synthesis and activity of *Helicobacter pylori* urease and catalase at low pH. *Gut*. 1997;40(1):25-30.
30. Kusters JG, van Vliet AH, Kuipers EJ. Pathogenesis of *Helicobacter pylori* infection. *Clin Microbiol Rev*. 2006;19(3):449-90.
31. Crowe SE, Alvarez L, Dytoc M, Hunt RH, Muller M, Sherman P, et al. Expression of interleukin 8 and CD54 by human gastric epithelium after *Helicobacter pylori* infection in vitro. *Gastroenterology*. 1995;108(1):65-74.
32. Zilberstein B, Quintanilha AG, Santos MA, Pajecki D, Moura EG, Alves PR, et al. Digestive tract microbiota in healthy volunteers. *Clinics (Sao Paulo)*. 2007;62(1):47-54.
33. Carr FJ, Chill D, Maida N. The lactic acid bacteria: a literature survey. *Crit Rev Microbiol*. 2002;28(4):281-370.

34. Fujimura S, Kato S, Oda M, Miyahara M, Ito Y, Kimura K, et al. Detection of *Lactobacillus gasseri* OLL2716 strain administered with yogurt drink in gastric mucus layer in humans. *Lett Appl Microbiol*. 2006;43(5):578-81.
35. Azcarate-Peril MA, Altermann E, Hoover-Fitzula RL, Cano RJ, Klaenhammer TR. Identification and inactivation of genetic loci involved with *Lactobacillus acidophilus* acid tolerance. *Appl Environ Microbiol*. 2004;70(9):5315-22.
36. Cotter PD, Hill C. Surviving the acid test: responses of gram-positive bacteria to low pH. *Microbiol Mol Biol Rev*. 2003;67(3):429-53, table of contents.
37. Sachs G, Weeks DL, Wen Y, Marcus EA, Scott DR, Melchers K. Acid acclimation by *Helicobacter pylori*. *Physiology (Bethesda)*. 2005;20:429-38.
38. Li XX, Wong GL, To KF, Wong VW, Lai LH, Chow DK, et al. Bacterial microbiota profiling in gastritis without *Helicobacter pylori* infection or non-steroidal anti-inflammatory drug use. *PLoS One*. 2009;4(11):e7985.
39. Fraher MH, O'Toole PW, Quigley EM. Techniques used to characterize the gut microbiota: a guide for the clinician. *Nat Rev Gastroenterol Hepatol*. 2012;9(6):312-22.
40. Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, Francois F, et al. Molecular analysis of the bacterial microbiota in the human stomach. *Proc Natl Acad Sci U S A*. 2006;103(3):732-7.
41. Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P, Engstrand L. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One*. 2008;3(7):e2836.
42. Metcalf AM, Phillips SF, Zinsmeister AR, MacCarty RL, Beart RW, Wolff BG. Simplified assessment of segmental colonic transit. *Gastroenterology*. 1987;92(1):40-7.
43. Sommer F, Backhed F. Know your neighbor: Microbiota and host epithelial cells interact locally to control intestinal function and physiology. *Bioessays*. 2016;38(5):455-64.
44. Bergman EN. Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiol Rev*. 1990;70(2):567-90.
45. Hamer HM, Jonkers D, Venema K, Vanhoutvin S, Troost FJ, Brummer RJ. Review article: the role of butyrate on colonic function. *Aliment Pharmacol Ther*. 2008;27(2):104-19.
46. Donohoe DR, Wali A, Brylawski BP, Bultman SJ. Microbial regulation of glucose metabolism and cell-cycle progression in mammalian colonocytes. *PLoS One*. 2012;7(9):e46589.
47. Kelly CJ, Zheng L, Campbell EL, Saeedi B, Scholz CC, Bayless AJ, et al. Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and Intestinal Epithelial HIF Augments Tissue Barrier Function. *Cell Host Microbe*. 2015;17(5):662-71.
48. Rivera-Chavez F, Zhang LF, Faber F, Lopez CA, Byndloss MX, Olsan EE, et al. Depletion of Butyrate-Producing Clostridia from the Gut Microbiota Drives

an Aerobic Luminal Expansion of Salmonella. *Cell Host Microbe*. 2016;19(4):443-54.

49. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*. 2014;505(7484):559-63.

50. Turnbaugh PJ, Backhed F, Fulton L, Gordon JI. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe*. 2008;3(4):213-23.

51. Sonnenburg ED, Zheng H, Joglekar P, Higginbottom SK, Firkbank SJ, Bolam DN, et al. Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell*. 2010;141(7):1241-52.

52. Kashyap PC, Marcobal A, Ursell LK, Smits SA, Sonnenburg ED, Costello EK, et al. Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc Natl Acad Sci U S A*. 2013;110(42):17059-64.

53. Larsson JM, Karlsson H, Sjoval H, Hansson GC. A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MSn. *Glycobiology*. 2009;19(7):756-66.

54. Png CW, Linden SK, Gilshenan KS, Zoetendal EG, McSweeney CS, Sly LI, et al. Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. *Am J Gastroenterol*. 2010;105(11):2420-8.

55. Albenberg L, Esipova TV, Judge CP, Bittinger K, Chen J, Laughlin A, et al. Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology*. 2014;147(5):1055-63 e8.

56. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. *Science*. 2005;308(5728):1635-8.

57. Pedron T, Mulet C, Dauga C, Frangeul L, Chervaux C, Grompone G, et al. A crypt-specific core microbiota resides in the mouse colon. *mBio*. 2012;3(3).

58. Tropini C, Earle KA, Huang KC, Sonnenburg JL. The Gut Microbiome: Connecting Spatial Organization to Function. *Cell Host Microbe*. 2017;21(4):433-42.

59. Nava GM, Friedrichsen HJ, Stappenbeck TS. Spatial organization of intestinal microbiota in the mouse ascending colon. *ISME J*. 2011;5(4):627-38.

60. Cullen TW, Schofield WB, Barry NA, Putnam EE, Rundell EA, Trent MS, et al. Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science*. 2015;347(6218):170-5.

61. Round JL, Mazmanian SK. Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. *Proc Natl Acad Sci U S A*. 2010;107(27):12204-9.

62. Round JL, Lee SM, Li J, Tran G, Jabri B, Chatila TA, et al. The Toll-like

receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science*. 2011;332(6032):974-7.

63. Halfvarson J, Brislawn CJ, Lamendella R, Vazquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*. 2017;2:17004.

64. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ, et al. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A*. 2008;105(43):16731-6.

65. Willing B, Halfvarson J, Dicksved J, Rosenquist M, Järnerot G, Engstrand L, et al. Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn's disease. *Inflamm Bowel Dis*. 2009;15(5):653-60.

66. Rajca S, Grondin V, Louis E, Vernier-Massouille G, Grimaud JC, Bouhnik Y, et al. Alterations in the intestinal microbiome (dysbiosis) as a predictor of relapse after infliximab withdrawal in Crohn's disease. *Inflamm Bowel Dis*. 2014;20(6):978-86.

67. Grice EA, Segre JA. The skin microbiome. *Nat Rev Microbiol*. 2011;9(4):244-53.

68. Oh J, Byrd AL, Deming C, Conlan S, Program NCS, Kong HH, et al. Biogeography and individuality shape function in the human skin metagenome. *Nature*. 2014;514(7520):59-64.

69. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, et al. Topographical and temporal diversity of the human skin microbiome. *Science*. 2009;324(5931):1190-2.

70. Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. *Nat Rev Microbiol*. 2018;16(3):143-55.

71. Scharschmidt TC, Fischbach MA. What Lives On Our Skin: Ecology, Genomics and Therapeutic Opportunities Of the Skin Microbiome. *Drug Discov Today Dis Mech*. 2013;10(3-4).

72. Holland KT, Greenman J, Cunliffe WJ. Growth of cutaneous propionibacteria on synthetic medium; growth yields and exoenzyme production. *J Appl Bacteriol*. 1979;47(3):383-94.

73. Bruggemann H, Henne A, Hoster F, Liesegang H, Wiezer A, Strittmatter A, et al. The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin. *Science*. 2004;305(5684):671-3.

74. Marples RR, Downing DT, Kligman AM. Control of free fatty acids in human surface lipids by *Corynebacterium acnes*. *J Invest Dermatol*. 1971;56(2):127-31.

75. Ingham E, Holland KT, Gowland G, Cunliffe WJ. Partial purification and characterization of lipase (EC 3.1.1.3) from *Propionibacterium acnes*. *J Gen Microbiol*. 1981;124(2):393-401.

76. Gribbon EM, Cunliffe WJ, Holland KT. Interaction of *Propionibacterium acnes* with skin lipids in vitro. *J Gen Microbiol*. 1993;139(8):1745-51.

77. Mukherjee S, Mitra R, Maitra A, Gupta S, Kumaran S, Chakraborty A, et al. Sebum and Hydration Levels in Specific Regions of Human Face Significantly Predict the Nature and Diversity of Facial Skin Microbiome. *Sci Rep.* 2016;6:36062.
78. Pekmezovic M, Mogavero S, Naglik JR, Hube B. Host-Pathogen Interactions during Female Genital Tract Infections. *Trends Microbiol.* 2019;27(12):982-96.
79. Linhares IM, Summers PR, Larsen B, Giraldo PC, Witkin SS. Contemporary perspectives on vaginal pH and lactobacilli. *Am J Obstet Gynecol.* 2011;204(2):120 e1-5.
80. Hickey RJ, Zhou X, Pierson JD, Ravel J, Forney LJ. Understanding vaginal microbiome complexity from an ecological perspective. *Transl Res.* 2012;160(4):267-82.
81. Chen X, Lu Y, Chen T, Li R. The Female Vaginal Microbiome in Health and Bacterial Vaginosis. *Front Cell Infect Microbiol.* 2021;11:631972.
82. Aagaard K, Riehle K, Ma J, Segata N, Mistretta TA, Coarfa C, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One.* 2012;7(6):e36466.
83. Noyes N, Cho KC, Ravel J, Forney LJ, Abdo Z. Associations between sexual habits, menstrual hygiene practices, demographics and the vaginal microbiome as revealed by Bayesian network analysis. *PLoS One.* 2018;13(1):e0191625.
84. Schwebke JR, Richey CM, Weiss HL. Correlation of behaviors with microbiological changes in vaginal flora. *J Infect Dis.* 1999;180(5):1632-6.
85. Culhane JF, Rauh V, McCollum KF, Elo IT, Hogan V. Exposure to chronic stress and ethnic differences in rates of bacterial vaginosis among pregnant women. *Am J Obstet Gynecol.* 2002;187(5):1272-6.
86. Gupta VK, Paul S, Dutta C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Front Microbiol.* 2017;8:1162.
87. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1(Suppl 1):4680-7.
88. Nelson TM, Borgogna JC, Michalek RD, Roberts DW, Rath JM, Glover ED, et al. Cigarette smoking is associated with an altered vaginal tract metabolomic profile. *Sci Rep.* 2018;8(1):852.
89. France MT, Ma B, Gajer P, Brown S, Humphrys MS, Holm JB, et al. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. *Microbiome.* 2020;8(1):166.
90. Srinivasan S, Fredricks DN. The human vaginal bacterial biota and bacterial vaginosis. *Interdiscip Perspect Infect Dis.* 2008;2008:750479.
91. Witkin SS, Mendes-Soares H, Linhares IM, Jayaram A, Ledger WJ, Forney LJ. Influence of vaginal bacteria and D- and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection

- against upper genital tract infections. *mBio*. 2013;4(4).
92. Witkin SS, Linhares IM. Why do lactobacilli dominate the human vaginal microbiota? *BJOG*. 2017;124(4):606-11.
93. Petrova MI, Lievens E, Malik S, Imholz N, Lebeer S. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. *Front Physiol*. 2015;6:81.
94. Amabebe E, Anumba DOC. The Vaginal Microenvironment: The Physiologic Role of Lactobacilli. *Front Med (Lausanne)*. 2018;5:181.
95. Edwards VL, Smith SB, McComb EJ, Tamarelle J, Ma B, Humphrys MS, et al. The Cervicovaginal Microbiota-Host Interaction Modulates Chlamydia trachomatis Infection. *mBio*. 2019;10(4).
96. Gupta K, Stapleton AE, Hooton TM, Roberts PL, Fennell CL, Stamm WE. Inverse association of H₂O₂-producing lactobacilli and vaginal Escherichia coli colonization in women with recurrent urinary tract infections. *J Infect Dis*. 1998;178(2):446-50.
97. Kirjavainen PV, Pautler S, Baroja ML, Anukam K, Crowley K, Carter K, et al. Abnormal immunological profile and vaginal microbiota in women prone to urinary tract infections. *Clin Vaccine Immunol*. 2009;16(1):29-36.
98. McClelland RS, Richardson BA, Hassan WM, Graham SM, Kiarie J, Baeten JM, et al. Prospective study of vaginal bacterial flora and other risk factors for vulvovaginal candidiasis. *J Infect Dis*. 2009;199(12):1883-90.
99. Zhou X, Westman R, Hickey R, Hansmann MA, Kennedy C, Osborn TW, et al. Vaginal microbiota of women with frequent vulvovaginal candidiasis. *Infect Immun*. 2009;77(9):4130-5.
100. Brown SE, Schwartz JA, Robinson CK, O'Hanlon DE, Bradford LL, He X, et al. The Vaginal Microbiota and Behavioral Factors Associated With Genital Candida albicans Detection in Reproductive-Age Women. *Sex Transm Dis*. 2019;46(11):753-8.
101. Ness RB, Kip KE, Hillier SL, Soper DE, Stamm CA, Sweet RL, et al. A cluster analysis of bacterial vaginosis-associated microflora and pelvic inflammatory disease. *Am J Epidemiol*. 2005;162(6):585-90.
102. Haggerty CL, Ness RB, Totten PA, Farooq F, Tang G, Ko D, et al. Presence and Concentrations of Select Bacterial Vaginosis-Associated Bacteria Are Associated With Increased Risk of Pelvic Inflammatory Disease. *Sex Transm Dis*. 2020;47(5):344-6.
103. Haggerty CL, Totten PA, Tang G, Astete SG, Ferris MJ, Norori J, et al. Identification of novel microbes associated with pelvic inflammatory disease and infertility. *Sex Transm Infect*. 2016;92(6):441-6.
104. Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, et al. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog*. 2010;6(1):e1000713.
105. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral

Genomes. *Mol Biol Evol.* 2021;38(10):4647-54.

106. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;47(D1):D807-D11.

107. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41(Database issue):D590-6.

108. 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.* 2021;39(1):105-14.

109. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics.* 2012;28(4):593-4.

110. Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H, et al. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* 2010;4(5):642-7.

111. Schulze J, Sonnenborn U. Yeasts in the gut: from commensals to infectious agents. *Dtsch Arztebl Int.* 2009;106(51-52):837-42.

112. Witchley JN, Penumetcha P, Abon NV, Woolford CA, Mitchell AP, Noble SM. *Candida albicans* Morphogenesis Programs Control the Balance between Gut Commensalism and Invasive Infection. *Cell Host Microbe.* 2019;25(3):432-43 e6.

113. Ward TL, Dominguez-Bello MG, Heisel T, Al-Ghalith G, Knights D, Gale CA. Development of the Human Mycobiome over the First Month of Life and across Body Sites. *mSystems.* 2018;3(3).

114. Muzyka BC, Epifanio RN. Update on oral fungal infections. *Dent Clin North Am.* 2013;57(4):561-81.

115. Bonifaz A, Vazquez-Gonzalez D, Macias B, Paredes-Farrera F, Hernandez MA, Araiza J, et al. Oral geotrichosis: report of 12 cases. *J Oral Sci.* 2010;52(3):477-83.

116. Xu H, Dongari-Bagtzoglou A. Shaping the oral mycobiota: interactions of opportunistic fungi with oral bacteria and the host. *Curr Opin Microbiol.* 2015;26:65-70.

117. Marsh PD, Zaura E. Dental biofilm: ecological interactions in health and disease. *J Clin Periodontol.* 2017;44 Suppl 18:S12-S22.

118. Shirliff ME, Peters BM, Jabra-Rizk MA. Cross-kingdom interactions: *Candida albicans* and bacteria. *FEMS Microbiol Lett.* 2009;299(1):1-8.

119. Deshpande NP, Riordan SM, Castano-Rodriguez N, Wilkins MR, Kaakoush NO. Signatures within the esophageal microbiome are associated with host genetics, age, and disease. *Microbiome.* 2018;6(1):227.

120. Raska M, Belakova J, Krupka M, Weigl E. Candidiasis--do we need to fight or to tolerate the *Candida* fungus? *Folia Microbiol (Praha).* 2007;52(3):297-312.

121. Farah CS, Elahi S, Drysdale K, Pang G, Gotjamanos T, Seymour GJ,

et al. Primary role for CD4(+) T lymphocytes in recovery from oropharyngeal candidiasis. *Infect Immun.* 2002;70(2):724-31.

122. Benitez AJ, Tanes C, Mattei L, Hofstaedter CE, Kim DK, Gross J, et al. Effect of topical swallowed steroids on the bacterial and fungal esophageal microbiota in eosinophilic esophagitis. *Allergy.* 2021;76(5):1549-52.

123. Huffnagle GB, Noverr MC. The emerging world of the fungal microbiome. *Trends Microbiol.* 2013;21(7):334-41.

124. Kumamoto CA. Inflammation and gastrointestinal *Candida* colonization. *Curr Opin Microbiol.* 2011;14(4):386-91.

125. Mason KL, Erb Downward JR, Falkowski NR, Young VB, Kao JY, Huffnagle GB. Interplay between the gastric bacterial microbiota and *Candida albicans* during postantibiotic recolonization and gastritis. *Infect Immun.* 2012;80(1):150-8.

126. Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, et al. The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome.* 2017;5(1):153.

127. Raimondi S, Amaretti A, Gozzoli C, Simone M, Righini L, Candelieri F, et al. Longitudinal Survey of Fungi in the Human Gut: ITS Profiling, Phenotyping, and Colonization. *Front Microbiol.* 2019;10:1575.

128. Chehoud C, Albenberg LG, Judge C, Hoffmann C, Grunberg S, Bittinger K, et al. Fungal Signature in the Gut Microbiota of Pediatric Patients With Inflammatory Bowel Disease. *Inflamm Bowel Dis.* 2015;21(8):1948-56.

129. Auchtung TA, Fofanova TY, Stewart CJ, Nash AK, Wong MC, Gesell JR, et al. Investigating Colonization of the Healthy Adult Gastrointestinal Tract by Fungi. *mSphere.* 2018;3(2).

130. Mar Rodriguez M, Perez D, Javier Chaves F, Esteve E, Marin-Garcia P, Xifra G, et al. Obesity changes the human gut mycobiome. *Sci Rep.* 2015;5:14600.

131. Borges FM, de Paula TO, Sarmiento MRA, de Oliveira MG, Pereira MLM, Toledo IV, et al. Fungal Diversity of Human Gut Microbiota Among Eutrophic, Overweight, and Obese Individuals Based on Aerobic Culture-Dependent Approach. *Curr Microbiol.* 2018;75(6):726-35.

132. Chen Y, Chen Z, Guo R, Chen N, Lu H, Huang S, et al. Correlation between gastrointestinal fungi and varying degrees of chronic hepatitis B virus infection. *Diagn Microbiol Infect Dis.* 2011;70(4):492-8.

133. Hamad I, Sokhna C, Raoult D, Bittar F. Molecular detection of eukaryotes in a single human stool sample from Senegal. *PLoS One.* 2012;7(7):e40888.

134. Motooka D, Fujimoto K, Tanaka R, Yaguchi T, Gotoh K, Maeda Y, et al. Fungal ITS1 Deep-Sequencing Strategies to Reconstruct the Composition of a 26-Species Community and Evaluation of the Gut Mycobiota of Healthy Japanese Individuals. *Front Microbiol.* 2017;8:238.

135. Pandey PK, Siddharth J, Verma P, Bavdekar A, Patole MS, Shouche YS. Molecular typing of fecal eukaryotic microbiota of human infants and their

- respective mothers. *J Biosci.* 2012;37(2):221-6.
136. Kabwe MH, Vikram S, Mulaudzi K, Jansson JK, Makhalanyane TP. The gut mycobiota of rural and urban individuals is shaped by geography. *BMC Microbiol.* 2020;20(1):257.
137. Botschuijver S, Roeselers G, Levin E, Jonkers DM, Welting O, Heinsbroek SEM, et al. Intestinal Fungal Dysbiosis Is Associated With Visceral Hypersensitivity in Patients With Irritable Bowel Syndrome and Rats. *Gastroenterology.* 2017;153(4):1026-39.
138. Gouba N, Raoult D, Drancourt M. Plant and fungal diversity in gut microbiota as revealed by molecular and culture investigations. *PLoS One.* 2013;8(3):e59474.
139. Richard ML, Lamas B, Liguori G, Hoffmann TW, Sokol H. Gut fungal microbiota: the Yin and Yang of inflammatory bowel disease. *Inflamm Bowel Dis.* 2015;21(3):656-65.
140. Hoffmann C, Dollive S, Grunberg S, Chen J, Li H, Wu GD, et al. Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. *PLoS One.* 2013;8(6):e66019.
141. Sun Y, Zuo T, Cheung CP, Gu W, Wan Y, Zhang F, et al. Population-Level Configurations of Gut Mycobiome Across 6 Ethnicities in Urban and Rural China. *Gastroenterology.* 2021;160(1):272-86 e11.
142. Suhr MJ, Banjara N, Hallen-Adams HE. Sequence-based methods for detecting and evaluating the human gut mycobiome. *Lett Appl Microbiol.* 2016;62(3):209-15.
143. Hallen-Adams HE, Suhr MJ. Fungi in the healthy human gastrointestinal tract. *Virulence.* 2017;8(3):352-8.
144. Mims TS, Abdallah QA, Stewart JD, Watts SP, White CT, Rousselle TV, et al. The gut mycobiome of healthy mice is shaped by the environment and correlates with metabolic outcomes in response to diet. *Commun Biol.* 2021;4(1):281.
145. Hoarau G, Mukherjee PK, Gower-Rousseau C, Hager C, Chandra J, Retuerto MA, et al. Bacteriome and Mycobiome Interactions Underscore Microbial Dysbiosis in Familial Crohn's Disease. *mBio.* 2016;7(5).
146. Liguori G, Lamas B, Richard ML, Brandi G, da Costa G, Hoffmann TW, et al. Fungal Dysbiosis in Mucosa-associated Microbiota of Crohn's Disease Patients. *J Crohns Colitis.* 2016;10(3):296-305.
147. Gouba N, Raoult D, Drancourt M. Gut microeukaryotes during anorexia nervosa: a case report. *BMC Res Notes.* 2014;7:33.
148. Li BY, Xu XY, Gan RY, Sun QC, Meng JM, Shang A, et al. Targeting Gut Microbiota for the Prevention and Management of Diabetes Mellitus by Dietary Natural Products. *Foods.* 2019;8(10).
149. Shah S, Locca A, Dorsett Y, Cantoni C, Ghezzi L, Lin Q, et al. Alterations of the gut mycobiome in patients with MS. *EBioMedicine.* 2021;71:103557.
150. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, et al.

Topographic diversity of fungal and bacterial communities in human skin. *Nature*. 2013;498(7454):367-70.

151. Wu G, Zhao H, Li C, Rajapakse MP, Wong WC, Xu J, et al. Genus-Wide Comparative Genomics of *Malassezia* Delineates Its Phylogeny, Physiology, and Niche Adaptation on Human Skin. *PLoS Genet*. 2015;11(11):e1005614.

152. Kalan L, Loesche M, Hodkinson BP, Heilmann K, Ruthel G, Gardner SE, et al. Redefining the Chronic-Wound Microbiome: Fungal Communities Are Prevalent, Dynamic, and Associated with Delayed Healing. *mBio*. 2016;7(5).

153. Bradford LL, Ravel J. The vaginal mycobiome: A contemporary perspective on fungi in women's health and diseases. *Virulence*. 2017;8(3):342-51.

154. Drell T, Lillsaar T, Tummeleht L, Simm J, Aaspollu A, Vain E, et al. Characterization of the vaginal micro- and mycobiome in asymptomatic reproductive-age Estonian women. *PLoS One*. 2013;8(1):e54379.

155. Mishra K, Bukavina L, Ghannoum M. Symbiosis and Dysbiosis of the Human Mycobiome. *Front Microbiol*. 2021;12:636131.

156. Odds FC. Genital candidosis. *Clin Exp Dermatol*. 1982;7(4):345-54.

157. Chatzivasileiou P, Vyzantiadis TA. Vaginal yeast colonisation: From a potential harmless condition to clinical implications and management approaches-A literature review. *Mycoses*. 2019;62(8):638-50.

158. Guo R, Zheng N, Lu H, Yin H, Yao J, Chen Y. Increased diversity of fungal flora in the vagina of patients with recurrent vaginal candidiasis and allergic rhinitis. *Microb Ecol*. 2012;64(4):918-27.

159. Ackerman AL, Underhill DM. The mycobiome of the human urinary tract: potential roles for fungi in urology. *Ann Transl Med*. 2017;5(2):31.

160. Liu NN, Zhao X, Tan JC, Liu S, Li BW, Xu WX, et al. Mycobiome Dysbiosis in Women with Intrauterine Adhesions. *Microbiol Spectr*. 2022;10(4):e0132422.

161. Gonia S, Archambault L, Shevik M, Altendahl M, Fellows E, Bliss JM, et al. *Candida parapsilosis* Protects Premature Intestinal Epithelial Cells from Invasion and Damage by *Candida albicans*. *Front Pediatr*. 2017;5:54.

162. Min J, Lu N, Huang S, Chai X, Wang S, Peng L, et al. Phenotype and biological characteristics of endometrial mesenchymal stem/stromal cells: A comparison between intrauterine adhesion patients and healthy women. *Am J Reprod Immunol*. 2021;85(6):e13379.

163. Cools P, Jaspers V, Hardy L, Crucitti T, Delany-Moretlwe S, Mwaura M, et al. A Multi-Country Cross-Sectional Study of Vaginal Carriage of Group B Streptococci (GBS) and *Escherichia coli* in Resource-Poor Settings: Prevalences and Risk Factors. *PLoS One*. 2016;11(1):e0148052.

164. van de Wijgert J, Verwijs MC. Lactobacilli-containing vaginal probiotics to cure or prevent bacterial or fungal vaginal dysbiosis: a systematic review and recommendations for future trial designs. *BJOG*. 2020;127(2):287-99.

165. Kabi F. Advances in

Microbiota, Nutrition
and Treatment in IBD2021.

166. Sherrington SL, Sorsby E, Mahtey N, Kumwenda P, Lenardon MD, Brown I, et al. Adaptation of *Candida albicans* to environmental pH induces cell wall remodelling and enhances innate immune recognition. *PLoS Pathog.* 2017;13(5):e1006403.
167. Lourenco A, Pedro NA, Salazar SB, Mira NP. Effect of Acetic Acid and Lactic Acid at Low pH in Growth and Azole Resistance of *Candida albicans* and *Candida glabrata*. *Front Microbiol.* 2018;9:3265.
168. Kwak MK, Liu R, Kim MK, Moon D, Kim AH, Song SH, et al. Cyclic dipeptides from lactic acid bacteria inhibit the proliferation of pathogenic fungi. *J Microbiol.* 2014;52(1):64-70.
169. Sobel JD. Recurrent vulvovaginal candidiasis. *Am J Obstet Gynecol.* 2016;214(1):15-21.
170. Osset J, Garcia E, Bartolome RM, Andreu A. [Role of *Lactobacillus* as protector against vaginal candidiasis]. *Med Clin (Barc).* 2001;117(8):285-8.
171. Seelig MS. The role of antibiotics in the pathogenesis of *Candida* infections. *Am J Med.* 1966;40(6):887-917.
172. Moens F, Duysburgh C, van den Abbeele P, Morera M, Marzorati M. *Lactobacillus rhamnosus* GG and *Saccharomyces cerevisiae* boulardii exert synergistic antipathogenic activity in vitro against enterotoxigenic *Escherichia coli*. *Benef Microbes.* 2019;10(8):923-35.
173. Schneider SM, Girard-Pipau F, Filippi J, Hebuterne X, Moyses D, Hinojosa GC, et al. Effects of *Saccharomyces boulardii* on fecal short-chain fatty acids and microflora in patients on long-term total enteral nutrition. *World J Gastroenterol.* 2005;11(39):6165-9.
174. Erb Downward JR, Falkowski NR, Mason KL, Muraglia R, Huffnagle GB. Modulation of post-antibiotic bacterial community reassembly and host response by *Candida albicans*. *Sci Rep.* 2013;3:2191.
175. Zuo T, Wong SH, Cheung CP, Lam K, Lui R, Cheung K, et al. Gut fungal dysbiosis correlates with reduced efficacy of fecal microbiota transplantation in *Clostridium difficile* infection. *Nat Commun.* 2018;9(1):3663.
176. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol.* 2006;72(7):5069-72.
177. Koljalg U, Nilsson HR, Schigel D, Tedersoo L, Larsson KH, May TW, et al. The Taxon Hypothesis Paradigm-On the Unambiguous Detection and Communication of Taxa. *Microorganisms.* 2020;8(12).
178. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat Rev Microbiol.* 2015;13(6):360-72.
179. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 2017;27(4):626-38.

180. Franzosa EA, Mclver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 2018;15(11):962-8.
181. Ruscheweyh HJ, Milanese A, Paoli L, Sintsova A, Mende DR, Zeller G, et al. mOTUs: Profiling Taxonomic Composition, Transcriptional Activity and Strain Populations of Microbial Communities. *Curr Protoc*. 2021;1(8):e218.
182. Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K. Identification of fungi in shotgun metagenomics datasets. *PLoS One*. 2018;13(2):e0192898.
183. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15(3):R46.
184. Soverini M, Turrone S, Biagi E, Brigidi P, Candela M, Rampelli S. HumanMycobiomeScan: a new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. *BMC Genomics*. 2019;20(1):496.
185. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 2004;32(Web Server issue):W20-5.
186. Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol*. 2016;1399:207-33.
187. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome*. 2021;9(1):58.
188. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otilar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res*. 2014;42(Database issue):D699-704.
189. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol*. 2014;32(8):834-41.
190. Montoliu-Nerin M, Sanchez-Garcia M, Bergin C, Grabherr M, Ellis B, Kutschera VE, et al. Building de novo reference genome assemblies of complex eukaryotic microorganisms from single nuclei. *Sci Rep*. 2020;10(1):1303.
191. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77.
192. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-2.
193. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. *Nucleic Acids Res*. 2011;39(Database issue):D19-21.
194. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun*. 2019;10(1):1014.
195. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving

- sequence similarity searches. *Bioinformatics*. 2015;31(6):926-32.
196. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*. 2012;8(9):e1002687.
197. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*. 2005;71(12):8228-35.
198. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059-66.
199. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol*. 2010;10:210.
200. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5(3):e9490.
201. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2000;28(1):10-4.
202. Shen Y, Gu Y, Pe'er I. A hidden Markov model for copy number variant prediction from whole genome resequencing data. *BMC Bioinformatics*. 2011;12 Suppl 6(Suppl 6):S4.
203. Lofgren LA, Uehling JK, Branco S, Bruns TD, Martin F, Kennedy PG. Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Mol Ecol*. 2019;28(4):721-30.
204. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. 2019;47(W1):W636-W41.
205. Alam I, Hubbard SJ, Oliver SG, Rattray M. A kingdom-specific protein domain HMM library for improved annotation of fungal genomes. *BMC Genomics*. 2007;8:97.
206. Nilsson RH, Tedersoo L, Ryberg M, Kristiansson E, Hartmann M, Unterseher M, et al. A Comprehensive, Automatically Updated Fungal ITS Sequence Dataset for Reference-Based Chimera Control in Environmental Sequencing Efforts. *Microbes Environ*. 2015;30(2):145-50.
207. Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, Irinyi L, et al. Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database (Oxford)*. 2014;2014.
208. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.
209. Sharma D, Denmat SH, Matzke NJ, Hannan K, Hannan RD, O'Sullivan JM, et al. A new method for determining ribosomal DNA copy number shows differences between *Saccharomyces cerevisiae* populations. *Genomics*. 2022;114(4):110430.
210. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20.

211. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. 2019;35(3):421-32.
212. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
213. Gourle H, Karlsson-Lindsjo O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*. 2019;35(3):521-2.
214. Liu J, Wang X, Xie H, Zhong Q, Xia Y. Analysis and evaluation of different sequencing depths from 5 to 20 million reads in shotgun metagenomic sequencing, with optimal minimum depth being recommended. *Genome*. 2022;65(9):491-504.
215. Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res*. 2019;47(D1):D259-D64.
216. O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-45.
217. Patel ZH, Kottyan LC, Lazaro S, Williams MS, Ledbetter DH, Tromp H, et al. The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. *Front Genet*. 2014;5:16.
218. Song K, Li L, Zhang G. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Sci Rep*. 2016;6:35736.
219. Yanez F, Soler Z, Olierio M, Xie Z, Oyarzun I, Serrano-Gomez G, et al. Integrating Dietary Data into Microbiome Studies: A Step Forward for Nutri-Metaomics. *Nutrients*. 2021;13(9).
220. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-8.
221. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15(2):155-63.
222. Serrano-Gomez G, Mayorga L, Oyarzun I, Roca J, Borrueal N, Casellas F, et al. Dysbiosis and relapse-related microbiome in inflammatory bowel disease: A shotgun metagenomic approach. *Comput Struct Biotechnol J*. 2021;19:6481-9.
223. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
224. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics*. 2018;19(1):274.

225. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2016;44(D1):D471-80.
226. Kanehisa M. KEGG Bioinformatics Resource for Plant Genomics and Metabolomics. *Methods Mol Biol.* 2016;1374:55-70.
227. Feizi S, Marbach D, Medard M, Kellis M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat Biotechnol.* 2013;31(8):726-33.
228. Ma B, Wang H, Dsouza M, Lou J, He Y, Dai Z, et al. Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J.* 2016;10(8):1891-901.
229. Luo F, Zhong J, Yang Y, Zhou J. Application of random matrix theory to microarray data for discovering functional gene modules. *Phys Rev E Stat Nonlin Soft Matter Phys.* 2006;73(3 Pt 1):031924.
230. Csárdi G. The igraph software package for complex network research. In: Center for Complex Systems Studies KC, editor.: R; 2006.
231. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 2011;5(2):169-72.
232. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One.* 2013;8(4):e61217.
233. Student. The probable error of a mean. *Biometrika* 1908;6(1):1-25.
234. S. S. SHAPIRO MBW. An analysis of variance test for normality (complete samples). *Biometrika.* 1965;52(3-4):591-611.
235. Boutin RCT, Sbihi H, McLaughlin RJ, Hahn AS, Konwar KM, Loo RS, et al. Composition and Associations of the Infant Gut Fungal Microbiota with Environmental Factors and Childhood Allergic Outcomes. *mBio.* 2021;12(3):e0339620.
236. Chao A. Nonparametric estimation of the number of classes in a population. *Scand J Stat.* 1984;11(4):6.
237. Shannon CE. The mathematical theory of communication. 1963. *MD Comput.* 1997;14(4):306-17.
238. Chaffin WL, Lopez-Ribot JL, Casanova M, Gozalbo D, Martinez JP. Cell wall and secreted proteins of *Candida albicans*: identification, function, and expression. *Microbiol Mol Biol Rev.* 1998;62(1):130-80.
239. Chattaway FW, Holmes MR, Barlow AJ. Cell wall composition of the mycelial and blastospore forms of *Candida albicans*. *J Gen Microbiol.* 1968;51(3):367-76.
240. Stalhaberger T, Simenel C, Clavaud C, Eijsink VG, Jourdain R, Delepierre M, et al. Chemical organization of the cell wall polysaccharide core of *Malassezia restricta*. *J Biol Chem.* 2014;289(18):12647-56.
241. Neu AT, Allen EE, Roy K. Defining and quantifying the core

microbiome: Challenges and prospects. *Proc Natl Acad Sci U S A*. 2021;118(51).

242. Ventin-Holmberg R, Eberl A, Saqib S, Korpela K, Virtanen S, Sipponen T, et al. Bacterial and Fungal Profiles as Markers of Infliximab Drug Response in Inflammatory Bowel Disease. *J Crohns Colitis*. 2021;15(6):1019-31.

243. Parada Venegas D, De la Fuente MK, Landskron G, Gonzalez MJ, Quera R, Dijkstra G, et al. Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Front Immunol*. 2019;10:277.

244. Rustchenko EP, Curran TM, Sherman F. Variations in the number of ribosomal DNA units in morphological mutants and normal strains of *Candida albicans* and in normal strains of *Saccharomyces cerevisiae*. *J Bacteriol*. 1993;175(22):7189-99.

245. Pendrak ML, Roberts DD. Ribosomal RNA processing in *Candida albicans*. *RNA*. 2011;17(12):2235-48.

246. Freire-Beneitez V, Price RJ, Tarrant D, Berman J, Buscaino A. *Candida albicans* repetitive elements display epigenetic diversity and plasticity. *Sci Rep*. 2016;6:22989.

247. Liu NN, Jiao N, Tan JC, Wang Z, Wu D, Wang AJ, et al. Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. *Nat Microbiol*. 2022;7(2):238-50.

248. Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med*. 2021;27(2):321-32.

249. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. *Nature*. 2019;568(7753):505-10.

250. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. *Nature*. 2019;568(7753):499-504.

251. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*. 2019;176(3):649-62 e20.

252. Ramos-Barbero MD, Martin-Cuadrado AB, Viver T, Santos F, Martinez-Garcia M, Anton J. Recovering microbial genomes from metagenomes in hypersaline environments: The Good, the Bad and the Ugly. *Syst Appl Microbiol*. 2019;42(1):30-40.

253. Meziti A, Rodriguez RL, Hatt JK, Pena-Gonzalez A, Levy K, Konstantinidis KT. The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol*. 2021;87(6).

254. Nelson WC, Tully BJ, Mobberley JM. Biases in genome reconstruction

- from metagenomic data. *PeerJ*. 2020;8:e10119.
255. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043-55.
256. Rodriguez RL, Gunturu S, Harvey WT, Rossello-Mora R, Tiedje JM, Cole JR, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res*. 2018;46(W1):W282-W8.
257. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res*. 2018;28(4):569-80.
258. Gupta S, Hjelmsø MH, Lehtimäki J, Li X, Mortensen MS, Russel J, et al. Environmental shaping of the bacterial and fungal community in infant bed dust and correlations with the airway microbiota. *Microbiome*. 2020;8(1):115.
259. Hamad I, Ranque S, Azhar EI, Yasir M, Jiman-Fatani AA, Tissot-Dupont H, et al. Culturomics and Amplicon-based Metagenomic Approaches for the Study of Fungal Population in Human Gut Microbiota. *Sci Rep*. 2017;7(1):16788.
260. Zhang E, Tanaka T, Tajima M, Tsuboi R, Nishikawa A, Sugita T. Characterization of the skin fungal microbiota in patients with atopic dermatitis and in healthy subjects. *Microbiol Immunol*. 2011;55(9):625-32.
261. Kim J, Sudbery P. *Candida albicans*, a major human fungal pathogen. *J Microbiol*. 2011;49(2):171-7.
262. Naumova N, Barsukov P, Baturina O, Rusalimova O, Kabilov M. Soil Mycobiome Diversity under Different Tillage Practices in the South of West Siberia. *Life (Basel)*. 2022;12(8).
263. Tedersoo L, Bahram M, Polme S, Koljalg U, Yorou NS, Wijesundera R, et al. Fungal biogeography. Global diversity and geography of soil fungi. *Science*. 2014;346(6213):1256688.
264. Zäncker B, Cunliffe M, Engel A. Eukaryotic community composition in the sea surface microlayer across an east–west transect in the Mediterranean Sea. *Biogeosciences*. 2021;18(6):2107-18.
265. Amend A, Burgaud G, Cunliffe M, Edgcomb VP, Ettinger CL, Gutierrez MH, et al. Fungi in the Marine Environment: Open Questions and Unsolved Problems. *mBio*. 2019;10(2).
266. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, et al. Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease. *Digestion*. 2016;93(1):59-65.
267. Sokol H, Leducq V, Aschard H, Pham HP, Jegou S, Landman C, et al. Fungal microbiota dysbiosis in IBD. *Gut*. 2017;66(6):1039-48.
268. Cicek N, Yildiz N, Kadayifci EK, Gokce I, Alpay H. Invasive aspergillosis in a patient with end stage renal disease. *Med Mycol Case Rep*. 2017;18:12-4.
269. Sato N, Yokoi H, Ichioka M, Ishii A, Matsubara T, Yanagita M. Invasive aspergillosis in the patient with focal segmental glomerulosclerosis initiating

hemodialysis: a case report and mini-review. *Renal Replacement Therapy*. 2022;8(1).

270. Gosiewski T, Salamon D, Szopa M, Sroka A, Malecki MT, Bulanda M. Quantitative evaluation of fungi of the genus *Candida* in the feces of adult patients with type 1 and 2 diabetes - a pilot study. *Gut Pathog*. 2014;6(1):43.

271. Jayasudha R, Das T, Kalyana Chakravarthy S, Sai Prashanthi G, Bhargava A, Tyagi M, et al. Gut mycobiomes are altered in people with type 2 Diabetes Mellitus and Diabetic Retinopathy. *PLoS One*. 2020;15(12):e0243077.

272. Bhute SS, Suryavanshi MV, Joshi SM, Yajnik CS, Shouche YS, Ghaskadbi SS. Gut Microbial Diversity Assessment of Indian Type-2-Diabetics Reveals Alterations in Eubacteria, Archaea, and Eukaryotes. *Front Microbiol*. 2017;8:214.

273. Newsholme P, Lima MM, Procopio J, Pithon-Curi TC, Doi SQ, Bazotte RB, et al. Glutamine and glutamate as vital metabolites. *Braz J Med Biol Res*. 2003;36(2):153-63.

274. Tapiero H, Mathe G, Couvreur P, Tew KD. II. Glutamine and glutamate. *Biomed Pharmacother*. 2002;56(9):446-57.

275. Blachier F, Boutry C, Bos C, Tome D. Metabolism and functions of L-glutamate in the epithelial cells of the small and large intestines. *Am J Clin Nutr*. 2009;90(3):814S-21S.

276. Jiao N, Wu Z, Ji Y, Wang B, Dai Z, Wu G. L-Glutamate Enhances Barrier and Antioxidative Functions in Intestinal Porcine Epithelial Cells. *J Nutr*. 2015;145(10):2258-64.

277. Wu M, Xiao H, Ren W, Yin J, Tan B, Liu G, et al. Therapeutic effects of glutamic acid in piglets challenged with deoxynivalenol. *PLoS One*. 2014;9(7):e100591.

278. Liu Y, Wang X, Hu CA. Therapeutic Potential of Amino Acids in Inflammatory Bowel Disease. *Nutrients*. 2017;9(9).

279. Herrera ML, Vallor AC, Gelfond JA, Patterson TF, Wickes BL. Strain-dependent variation in 18S ribosomal DNA Copy numbers in *Aspergillus fumigatus*. *J Clin Microbiol*. 2009;47(5):1325-32.

280. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A*. 2012;109(16):6241-6.

281. Usyk M, Peters BA, Karthikeyan S, McDonald D, Sollecito CC, Vazquez-Baeza Y, et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. *Cell Rep Methods*. 2023;3(1):100391.

282. Amenyogbe N, Adu-Gyasi D, Enuameh Y, Asante KP, Konadu DG, Kaali S, et al. Bacterial and Fungal Gut Community Dynamics Over the First 5 Years of Life in Predominantly Rural Communities in Ghana. *Front Microbiol*. 2021;12:664407.

283. Santus W, Rana AP, Devlin JR, Kiernan KA, Jacob CC, Tjokrosurjo J, et al. Mycobiota and diet-derived fungal xenosiderophores promote *Salmonella* gastrointestinal colonization. *Nat Microbiol.* 2022;7(12):2025-38.
284. Yu D, Xie L, Chen W, Qin J, Zhang J, Lei M, et al. Dynamics of the Gut Bacteria and Fungi Accompanying Low-Carbohydrate Diet-Induced Weight Loss in Overweight and Obese Adults. *Front Nutr.* 2022;9:846378.

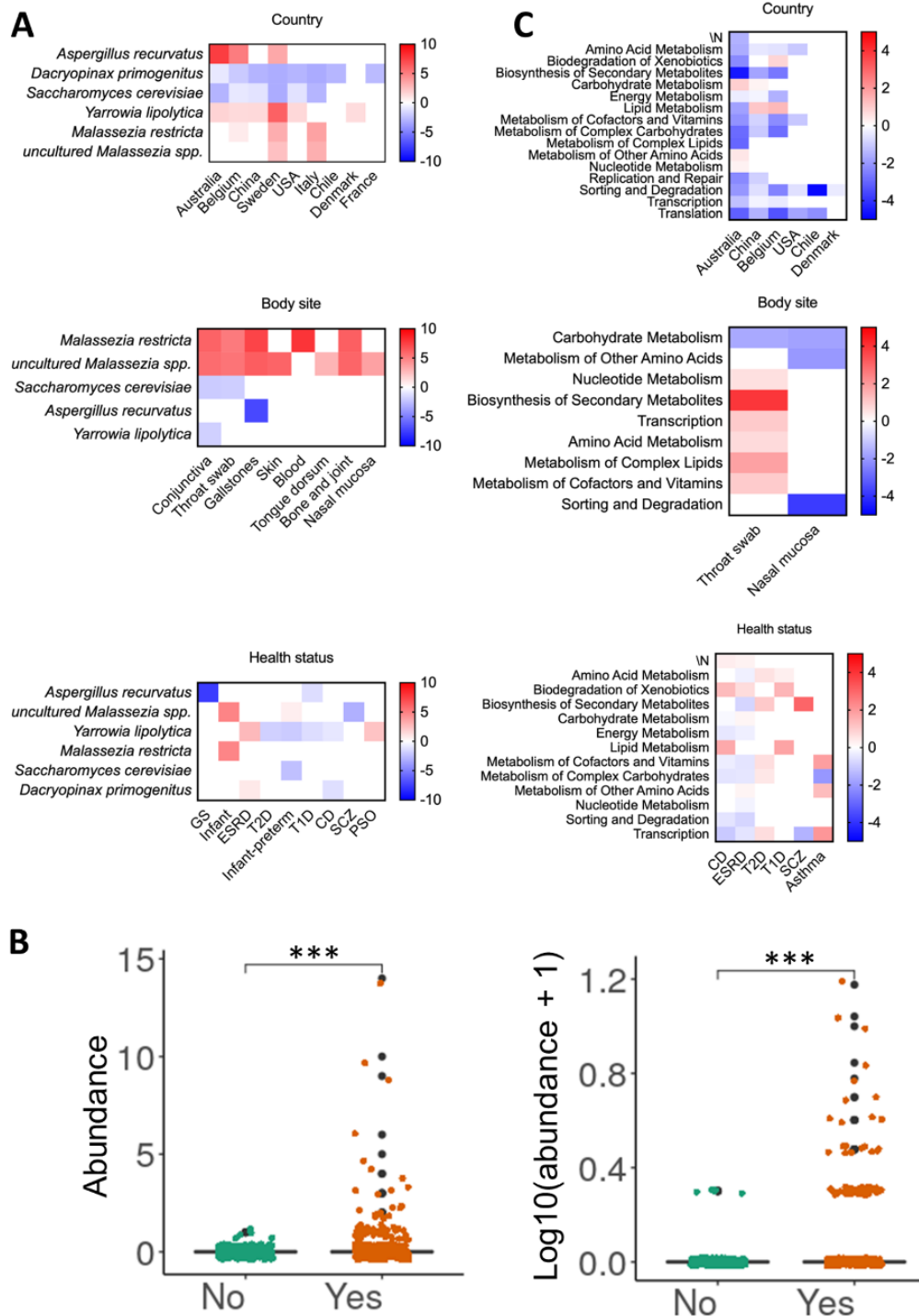
10. ANNEXES

ANNEX 1. Link to all the supplementary tables

For better visualization, all supplementary tables are stored in drive, and can be freely accessed through this link:

https://drive.google.com/drive/folders/1j-7m73_koRJWuW2EVIHwtyayhJsRH3-s?usp=sharing

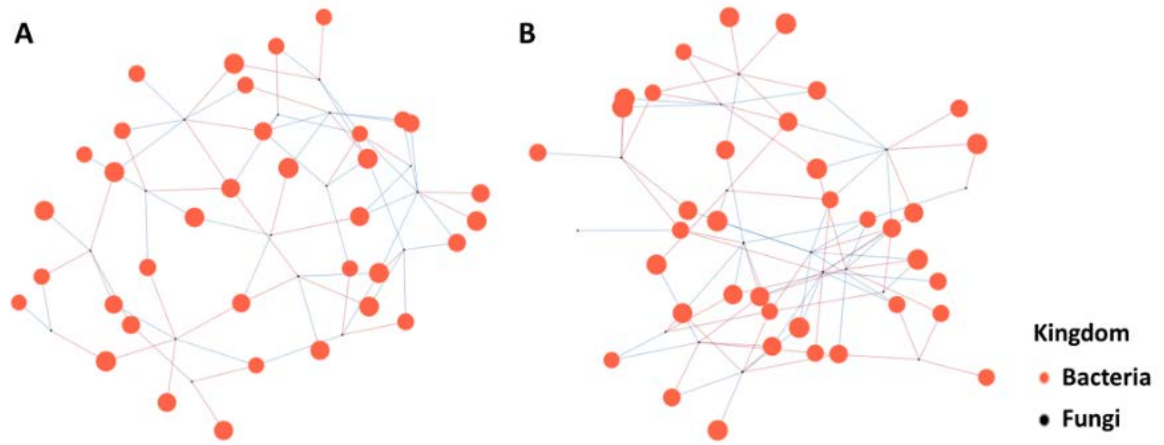
ANNEX 2. Fungal species and pathway classes in different groups of mycobiomes.



ANNEX2, Supplementary Figure 1. Fungal species (A) and pathway classes (B) associated with different countries, health status and body sites. The

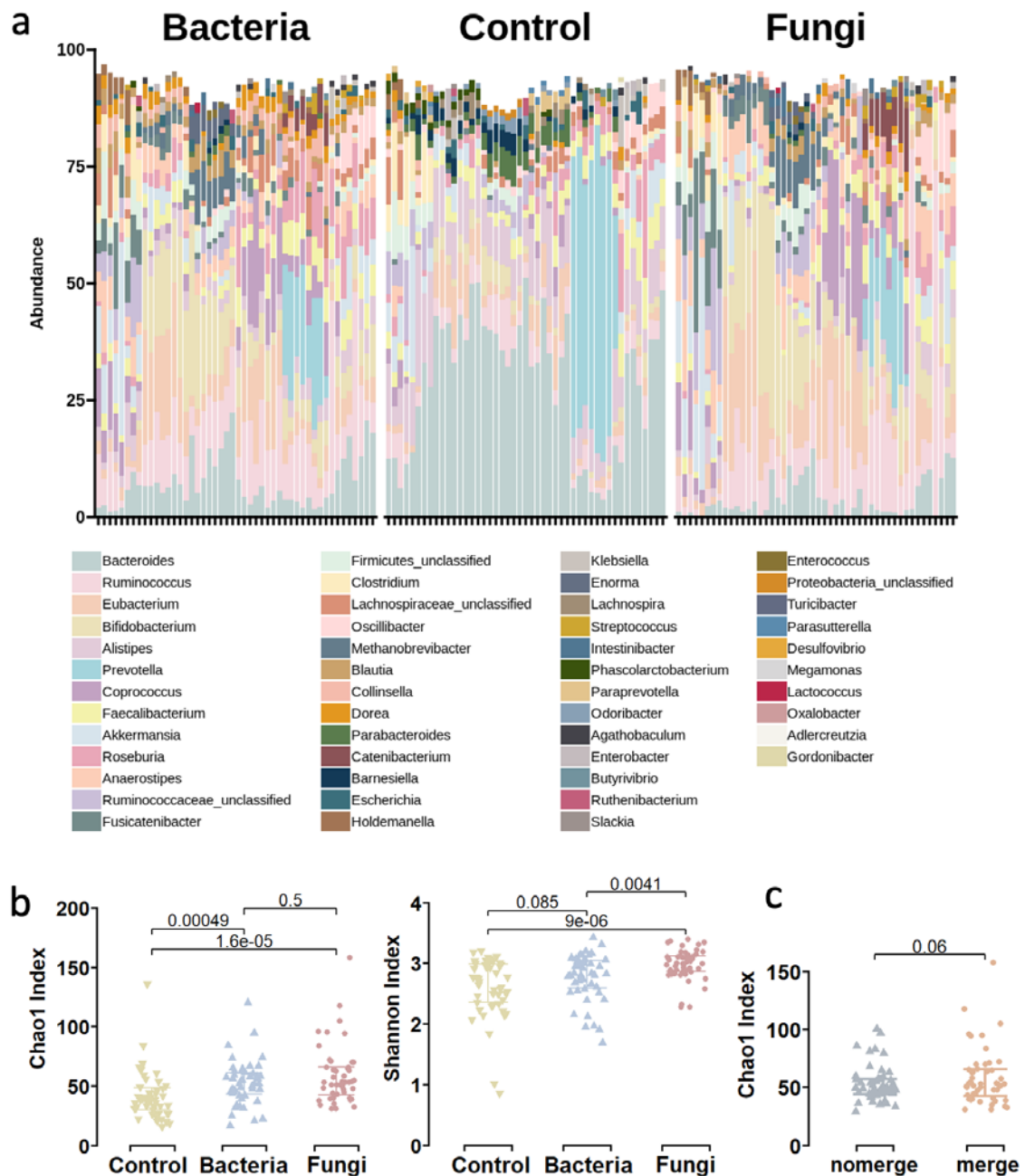
colours represent the sign of the association (red means positive, blue means negative). The intensity of the colours represents the degree of association (darker means stronger association). Spain, healthy and gut were set as the reference level for each of the models, respectively. (C) Significant differences in *Yarrowia lipolytica* abundances or log₁₀ transformed abundances between bead-beaten samples and non-bead beaten samples.

ANNEX3. Interaction of fungal and bacterial functions in the gut microbiome of healthy individuals.



ANNEX3, Supplementary Figure 2. Significant correlation ($p < 0.05$) network between the relative abundance of fungal and bacterial functions in the gut microbiome of healthy individuals from Spain (A) and Denmark (B) using SparCC algorithm. Each node represents a fungal/bacterial pathway class and their sizes are determined by the relative abundances. The colors of the edges connecting two nodes represent the positive (red) and negative (blue) correlations. For a better visual effect, only the correlations with p-values smaller than 0.001 and absolute correlation coefficient over 0.05 are represented

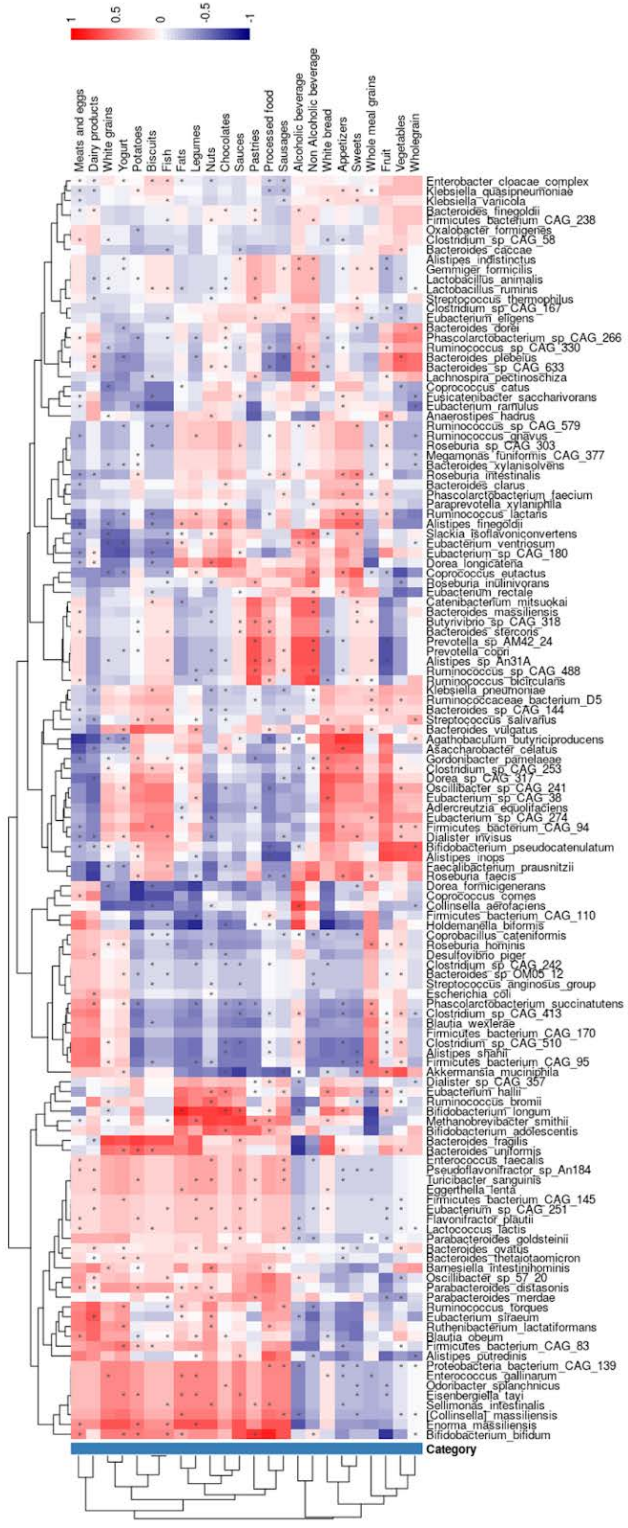
ANNEX4. Enrichment efficiency in bacteria.



ANNEX4, Supplementary Figure 3. a, The genus-level taxa bar plot of the bacterial community compositions in bacterial, control, and fungal partitions. b, Boxplots of the species-level bacterial community alpha diversities (Shannon and Chao1 indices) in fungal, control, and bacterial partitions (n=143) ordered by their mean from smallest to largest (left to right). c, Boxplots of the species-level bacterial community alpha diversity (Chao1 index) before and after

merging bacterial reads in the bacterial and fungal partitions (n=48) ordered by their mean from smallest to largest (left to right).

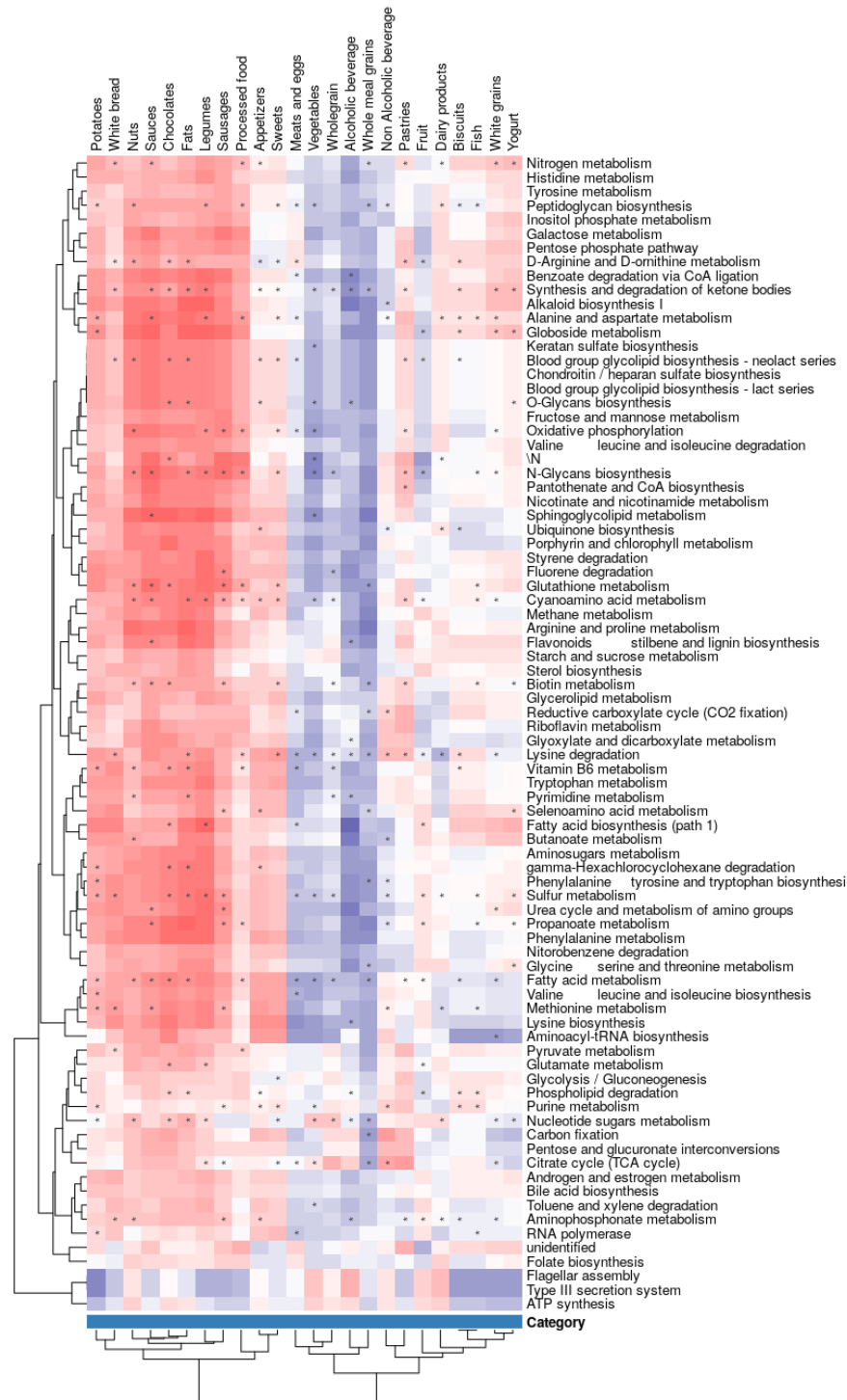
ANNEX5. Bacterial taxonomic compositions are associated with habitual diet.



ANNEX5, Supplementary Figure 4. Heatmap of all the detected significant

correlations between bacterial taxonomic compositions and diet categories. The asterisk indicates that the correlation index for the corresponding species metadata pair is significant.

ANNEX6. Fungal functional compositions are associated with habitual diet.



ANNEX6, Supplementary Figure 5. Heatmap of all the detected significant correlations between fungal functional compositions and diet categories. The

asterisk indicates that the correlation index for the corresponding species metadata pair is significant.

asterisk indicates that the correlation index for the corresponding species metadata pair is significant.

ANNEX8. Publication related to this thesis: Xie Z, Manichanh C. FunOMIC: Pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling. *Comput Struct Biotechnol J.* 2022;20:3685-94.

Computational and Structural Biotechnology Journal 20 (2022) 3685–3694



COMPUTATIONAL
AND STRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj



FunOMIC: Pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling

Zixuan Xie^{a,b}, Chaysavanh Manichanh^{a,b,*}

^a Microbiome Lab, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain
^b Departament de Medicina, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain



ARTICLE INFO

Article history:

Received 30 May 2022
Received in revised form 4 July 2022
Accepted 4 July 2022
Available online 11 July 2022

Keywords:

Mycobiome
Fungal databases
Taxonomy and functions
Shotgun metagenomics
Inter-kingdom interactions

ABSTRACT

While analysis of the bacterial microbiome has become routine, that of the fungal microbiome is still hampered by the lack of robust databases and bioinformatic pipelines. Here, we present FunOMIC, a pipeline with built-in taxonomic (1.6 million marker genes) and functional (3.4 million non-redundant fungal proteins) databases for the identification of fungi. Applied to more than 2,600 human metagenomic samples, the tool revealed fungal species associated with geography, body sites, and diseases. Correlation network analysis provided new insights into inter-kingdom interactions. With this pipeline and two of the most comprehensive fungal databases, we foresee a fast-growing resource for mycobiome studies.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Fungi ubiquitously exist as commensals in various body sites of humans, including the gastrointestinal tract (GIT), oral cavity, vagina, and skin [1]. Under certain circumstances, some of these fungal commensals, identified as pathobionts, could cause harm [1,2]. Also, bacterial-fungal interactions have been reported to exacerbate, reduce, or resist disease caused by fungal infection [3,4]. The colonised fungi are highly variable across populations [5], which may prevent the establishment and, thereby, the identification of key players among the fungal community in humans. It is therefore critical to investigate commensal fungi and their interactions with the host and commensal bacteria in a large-scale study.

Unlike the prokaryotic community in the human microbiome, the fungal population, known as mycobiome, is still understudied due to various reasons, including the challenge associated with unculturable microorganisms, the extremely low abundance

among the human microbiome community [6], inter-individual variability, and the lack of a comprehensive database. Over the last decades, along with the rapid development of high-throughput sequencing (HTS) technology, the study of the human bacterial and fungal microbiome has gradually moved from culture-dependent towards culture-independent methods [1].

The characterization of the mycobiome has been catalysed by targeted HTS of the internal transcribed spacer (ITS) or the 18S rRNA (18S) region located inside the ribosomal region. Similar to the 16S rRNA (16S) gene in prokaryotes, the ITS and 18S regions have conserved and highly variable segments among different fungal organisms. Moreover, the ITS has been recognised as a universal DNA barcode marker for fungi [7]. The current knowledge of human mycobiome derives mostly from the analysis of ITS and 18S amplicon sequencing [8,9]. However, as for the 16S amplicon sequencing approach [10], ITS and 18S approaches can introduce biases due to variability in amplification efficiency [11], problems related to species delineation, and the large variations in gene copy numbers, which limits the relative abundance analysis between closely related species [12]. As an alternative to ribosomal DNAs, a set of single-copy marker genes can be candidates for taxonomically annotating the microbiome. They have been shown to provide higher resolution than 16S in prokaryotic species delineation [13] and have been used to estimate relative abundances and richness of bacterial members in human faecal microbiomes.

Abbreviations: CD, Crohn's disease; ESRD, End-stage renal disease; FDR, False discovery rate; GS, Gallstones; HC, Healthy control; HTS, High throughput sequencing; ITS, internal transcribed spacer; NA, Not applicable; PLWH, People live with HIV; PSO, Psoriasis; SCFA, Short chain fatty acid; SCZ, Schizophrenia; TB, Tuberculosis; T1D, Type 1 diabetes; T2D, Type 2 diabetes; UC, Ulcerative colitis.

* Corresponding author at: Microbiome Lab, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain.

E-mail address: cmancha@gmail.com (C. Manichanh).

<https://doi.org/10.1016/j.csbj.2022.07.010>

2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

With the decreasing cost of sequencing, the shotgun approach, which can capture more unbiased information from the gene pool of microbial genomes within an environment than the amplicon approach, has emerged as a more attractive tool in microbiome research. Various strategies and databases have been developed to determine eukaryotic community compositions from metagenomic data [14–16], yet few of them tackled fungi in the context of the human microbiome.

To enable a more precise analysis of the human mycobiome, we propose herein two built-in fungal databases, FunOMIC-T and FunOMIC-P, integrated into an automated pipeline for taxonomic and functional profiling, respectively. The functionality of the pipeline is achieved by mapping next-generation sequencing reads to the two FunOMIC databases. FunOMIC-T contains more than 1.6 million single-copy marker genes from 4,839 high-quality fungal genome data. FunOMIC-P includes more than 3 million fungal proteins, being an integration of the corresponding coding genes of the collected fungal genomes with the fungal subset of the Uniprot database. FunOMIC was used to analyse a publicly accessible set of 2,679 human metagenome samples, which revealed fungal taxonomic and functional signatures associated with clinical and demographic metadata.

2. Methods

2.1. Aim, design and setting of the study

FunOMIC is a pipeline implemented with two fungal databases FunOMIC-T and FunOMIC-P aiming at providing automatic mycobiome analysis. Shotgun sequencing reads are directly mapped to the databases to obtain mycobiome taxonomic and functional profiles via the main program FunOMIC.sh. The main program and two databases can be downloaded from Manichanh Lab (vhir.org). Detailed establishment steps can be found below.

2.2. Collection of fungal genomes

In total, 9,401 publicly available strain-level fungal genomes or draft genomes were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>) and JGI MycoCosm (<https://mycocosm.jgi.doe.gov/mycocosm/home>) [17] before January 25th, 2021. All fungal genomes with more than 500 contigs and N50 < 10 kbp were filtered out [18], which led to a final set of 4,331 high-quality genomes and draft genomes. Genomic shotgun data from 508 *Candida* isolates were downloaded from 30 unique bioprojects from the NCBI SRA before February 4th, 2021 (<https://www.ncbi.nlm.nih.gov/sra/>). The accession numbers of the 4,839 combined reference fungal genomes are listed in Supplementary Table S8.

2.3. Construction of the taxonomic and functional FunOMIC database

2.3.1. Identification of marker genes for establishing a taxonomic fungal database

Assembling *Candida* genomic sequencing reads was performed as described in the study of Montoliu-Nerin *et al.* [19]. Basically, each of the *Candida* genomic sequencing reads was normalised by BBNorm v38.9021 of BBtools (<https://jgi.doe.gov/data-and-tools/bbtools/>) with a target average depth of 100x. Then, normalised data were assembled by SPAdes v3.15.2 [20] (<https://cab.spbu.ru/software/spades/>). BUSCO (Benchmarking Universal Single-Copy Orthologs) version 5.0.0 (22) was used to identify marker genes using Fungi OrthoDB version 10.1 [21] in the pool of 4,839 fungal genomes. BUSCO makes use of 758 HMMs (hidden Markov models) of fungal single-copy marker genes and was run using default parameters with the AUGUSTUS gene predictor

[22]. Genomes with <30 single-copy marker genes identified were discarded, resulting in a final set of 4,816 genomes. Clustering with a 99 % identity threshold [14,23] was applied using CD-HIT [24] to remove redundancies, which led to a final set of 1.69 million fungal marker genes, referred here as FunOMIC-T.

2.3.2. Establishment of a functional fungal database.

A protein database for fungal functional analysis was also constructed by collecting the corresponding amino-acid sequences that were available for 2,967 of the 4,331 genomes cited above and the 35,360 reviewed fungal proteins from UniProt (<https://www.uniprot.org/>), both before January 2022. Then, the proteins without an explicit annotation were discarded (1.5 million) leading to a total of 4.9 million genes. Redundancy was removed with a 95 % identity clustering using CD-HIT [6]. Finally, 3,413,239 non-redundant fungal proteins, referred to as FunOMIC-P, were obtained for fungal functional profiling. These protein accessions (from JGI, NCBI, UniProt) were then linked to EC numbers and KEGG pathways.

2.4. Validation of the FunOMIC databases and the pipeline

To verify the absence of bacterial contamination [14] in the fungal database and to ensure specificity for fungal detection, we applied three different validation methods. Firstly, we mapped the 1.69 million fungal single-copy marker genes to the Unified Human Gastrointestinal Genome (UHGG), which is a gene catalogue that comprises 204,938 non-redundant genomes from 4,644 gut prokaryotes [25] using bowtie2. Because of the memory limitation of our computers (44 CPUs), we simulated sequencing reads of all the marker gene sequences (22 million paired-reads, 1-fold coverage, 11.2 GB out of 4.6 GB) to perform the alignment to the UHGG. Secondly, we simulated Illumina formatted sequencing output reads from a set of 903 bacterial genomes from 458 species that inhabit the human body collected from the NCBI to create a mock community for a bacterial community (Supplementary Table S1). The simulation was carried out by ART, a set of simulation tools that generate synthetic next-generation sequencing reads [26]. The simulated reads were then aligned to FunOMIC-T. Thirdly, another mock community was created with the top 20 fungal species and top 20 bacterial species identified in the 2,679 human metagenomes collected (cited below). The genomes of these 40 species were used to simulate Illumina formatted sequencing output reads, which were then mapped to the constructed database. The lists of genomes used for creating the mock communities and the number of simulated reads can be found in Supplementary Table S1.

To validate the FunOMIC-P database, a mixed mock community was created with the available coding gene sequences of the aforementioned top fungal and bacterial species. Again, the coding gene sequences collected from NCBI were used to simulate Illumina formatted sequencing output reads, which were then mapped to the FunOMIC-P database using Diamond blastx function v2.0.8 with an e-value < 10e-10 to recover the fungal functional profiling. To optimise the alignment parameters, we tested nine different combinations using three different percentages of coverage (>90 %, >95 %, >99 %) and three different percentages of identity (>90 %, >95 %, >99 %).

2.5. Collection of metagenomic data

We downloaded 2679 public human shotgun metagenomic sequencing data from NCBI SRA before February 4th, 2021 [27] (<https://www.ncbi.nlm.nih.gov/sra/>). The 2679-public human metagenomic data derive from 27 unique bioprojects, two of which were published in our previous studies (PRJNA514452,

PRJEB1220). The metadata of all the human metagenomic data can be found in [Supplementary Table S2](#). This metadata contains available information such as continent, country, city, latitude, longitude, sample source, gender, age, extraction procedure, and use of mechanical lysis during extraction.

2.6. Aligning human metagenomic sequencing reads onto the FunOMIC database

After quality control and decontamination using KneadData v0.7.7-alpha (<https://huttenhower.sph.harvard.edu/kneaddata/>), Bowtie2 v2.3.4.3 was used to map the 2,679 metagenomic data to the FunOMIC-T database for fungal taxonomic annotation. Mapped reads were kept if more than 80 % of the length aligned to the reference sequence with a q-score of over 30 [6,14,28] by using Samtools v1.9. Diamond blastx function v2.0.8 was used to map the metagenomic data to the FunOMIC-P database (read coverage >95 % and identity percentage >99 % and e-value < 10e-10) for fungal functional annotation. An in-house script, which is freely available at our GitHub (<https://github.com/ManichanhLab/FunOMIC>), was used to recover the final fungal taxonomic and functional profiling.

2.7. Prokaryotic taxonomic and functional profiling of human metagenomic data

After quality control and decontamination using KneadData v0.7.7-alpha (<https://huttenhower.sph.harvard.edu/kneaddata/>), we used MetaPhlan v3.0.9 for profiling the composition of prokaryotic communities in the 2,679 human metagenomic data. Then, the HUMAnN v3.0 [29] (<https://huttenhower.sph.harvard.edu/humann/>) and the UniRef90 database [30] were used to profile the abundance of prokaryotic metabolic pathways and other molecular functions.

2.8. Statistical analysis

All statistical analyses, except for SparCC correlation, were performed using R software 4.1.2 (2021-11-01). Alpha- and beta-diversity were calculated using the Phyloseq package. Beta-diversity was compared between different disease groups using the UniFrac distance metric with permutational multivariate analysis of variance (PERMANOVA) to identify significance ($p \leq 0.05$). The associations between fungal profilings with variables from the metadata were measured using the MaAsLin2 package with age as the random effect (results were considered significant if FDR (false discovery rate) < 0.05). The correlations of taxonomic profilings or functional profilings between bacteria and fungi were performed using the Python script SparCC [31].

3. Results

3.1. Characteristics of the taxonomic and functional FunOMIC database

To build a database for taxonomic profiling of environmental fungal species, more than 1.6 million fungal single-copy marker genes were extracted from 4,816 fungal high-quality genomes and draft genomes by aligning them to a set of 758 fungal universal orthologs from OrthoDB (Fig. 1). The newly constructed database, FunOMIC-T, covers eight fungal phyla, among which three (Ascomycota, Basidiomycota, and Mucoromycota) represented more than 98 % of the genomes (Fig. 1A). At lower taxonomic levels, they encompassed 475 genera, 1,916 species, and 4,537 strains.

It has been reported that 99.9 % of human metagenome sequences are from bacteria [6] and that, bacterial sequences are ubiquitous in eukaryotic genomes [14]. Validation of the absence of bacterial sequence contamination in the fungal database is, therefore, critical. To address this requirement, the FunOMIC-T database was mapped to the UHGG dataset, which contains 204,938 non-redundant genomes from 4,644 gut prokaryotes [25]. Only <0.01 % of the fungal marker genes mapped to the UHGG, demonstrating that this fungal taxonomic database was specific enough to detect mostly fungal sequences.

A bacterial environmental mock community was also created. For this, we collected 903 genomes from 458 bacterial species found to inhabit human bodies ([Supplementary Table S1](#)). These genomes were then simulated into 19,301,201 Illumina formatted sequencing output reads and mapped to the FunOMIC-T database. The mapping rate of this artificial community to the database was also <0.01 %. Lastly, a mixed mock community was also created comprising the top 20 bacterial species and top 20 fungal species identified during the taxonomic profiling of the metagenomes ([Supplementary Table S1](#)). To better mimic real human metagenomes, the ratio of the number of simulated bacterial reads over fungal reads was set to nearly 1000 (999,021 bacterial reads and 1046 fungal reads) [6]. As expected, none of the 999,021 bacterial reads aligned against FunOMIC-T, leading to a specificity (false positive / (false positive + true negative)) of 0.9999.

Given the numerically small proportion of fungal sequences in human metagenomes, the fungal functional analysis was not relevant in almost all the published human mycobiome studies. To address this knowledge gap, in the present work, we also proposed a protein database specifically for environmental fungal functional profiling. The FunOMIC-P database consists of 3,413,239 non-redundant fungal protein sequences integrated from NCBI, JGI, and UniProt (see Methods section above, Fig. 1B). Evaluation and validation were also performed by a mixed mock community constituted by the top species mentioned above. The available coding gene sequences of these species were simulated into 439,798 Illumina formatted sequencing output reads and mapped to the FunOMIC-P database. We tuned the Diamond blastx function with nine different combinations of parameters to optimize mapping performance. With the threshold of read coverage >95 %, identity percentage >99 %, and an e-value < 10e-10, we obtained the highest mapping rate of the fungal reads, where around 70 % of the hits passed this threshold. More than 50 % of the mis-mapped bacterial genes were related to ATP synthase ([Supplementary Table S1](#)).

3.2. Characteristics of the 2679 metagenomes

A set of 2679 metagenomes, which encompassed a total of 9077.12 Gb, collected from 27 bioprojects are listed in [Supplementary Table S2](#). Taxonomic profiling of the metagenomes against FunOMIC-T detected fungal DNA sequences in 1950 metagenomes (72.9 %) which was much higher than the ratio reported in previous shotgun sequencing studies analysing the human mycobiome. Lind *et al.*, reported a detection rate of <20 % and Olm *et al.*, found 6 % in their cohorts (infant). The 1,950 metagenomes were collected from 14 countries, 12 body sites, and 19 health and disease conditions (Table 1). The average mapping rate was 4.72E-05 (8.16E-09 min, 1.1E-02 max).

Gut samples comprised the majority of the dataset (84 %), followed by conjunctiva (5 %), saliva (3 %), and throat swab (1.5 %). Among the diseases evaluated, Crohn's disease (CD), ulcerative colitis (UC), end-stage renal disease (ESRD), type 1 diabetes (T1D), and type 2 diabetes (T2D) accounted for 779 faecal samples, whereas 500 faecal samples were obtained from healthy individuals.

All biological specimens were extracted by at least 10 different protocols, for which mechanical lysis, previously reported as a cru-

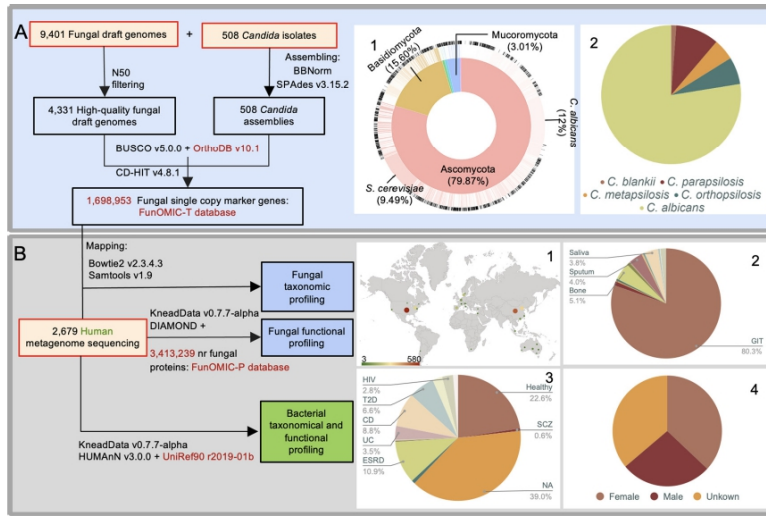


Fig. 1. Workflow of the construction of the FunOMIC database and its application in metagenomic analysis. (A) Recovery of fungal single-copy marker genes from fungal draft genomes and *Candida* isolate sequencing reads downloaded from NCBI and JGI. A.1) Distribution of the fungal draft genomes at the phylum and species levels in FunOMIC-T (Taxonomy). A.2) Distribution of *Candida* assemblies at the species level. (B) Fungal and bacterial taxonomic and functional profiling of the 2,679 metagenomic datasets downloaded from NCBI. B.1) Geographical location of the collected human metagenomes. B.2) Proportions of the collected human metagenomes by body sites. B.3) Proportions of human metagenomes by disease type (HIV = human immunodeficiency virus; T2D = type 2 diabetes; CD = Crohn's disease; UC = ulcerative colitis; ESRD = end-stage renal disease; SCZ = schizophrenia). B.4) Distribution of the collected human metagenomes by gender.

cial step during the DNA extraction process to recover an optimum microbial diversity [32], was applied in 1,049 samples (53.8 %).

3.3. Fungal community structure, diversity, and functions of the 1950 metagenomes

Five phyla, 232 genera and 475 species were identified in the 1,950 metagenomes. More than 80 % of the sequences were represented by two phyla (Ascomycota and Basidiomycota), two genera (*Saccharomyces*, *Candida*), and three species (*Saccharomyces cerevisiae*, *Candida albicans*, *Malassezia restricta*) (Fig. 2). Under healthy conditions, the gut mycobiome was dominated, in terms of relative abundance, by *Saccharomyces cerevisiae*, which was detected in 52.4 % of the samples, while *Dacryopinax primogenitus* was found in 23.6 %, *Yarrowia lipolytica* in 13.6 %, and *Candida parapsilosis* in 11 % of the samples. *C. albicans*, known as an opportunistic pathogenic yeast [33], was found in only 4 % of the GI tract samples of healthy individuals. The fungal species profiling data can be found in Supplementary Table S3. *Malassezia* predominated conjunctiva samples, whereas *Aspergillus* predominated the saliva mycobiome.

The number of observed fungal species in the 1950 metagenomic samples ranged from 1 to 40 (median of 2), Chao1 index [34] varied between 1 and 76.1 (median of 3), and the Shannon index [35] ranged from 0 to 3.36 (median of 0.62) (Supplementary Table S4). These three measurements indicated that the fungal community in humans is, in general, of very low diversity compared with the bacterial community, which could reach an average of 70 in terms of the Chao1 index [36].

While fungal taxonomic profiling of human microbial communities has increased considerably over the last 10 years through the sequencing of phylogenetic marker genes such as ITS2/18S, the fungal community function was scarcely investigated mainly

due to, again, the lack of a comprehensive database. Using FunOMIC-P, we annotated the sequencing reads of the 1,950 human metagenomes using the DIAMOND aligner. In total 1,948 metagenomes successfully mapped to the database, the average mapping rate was 0.088 % (5.42E-04 % min, 1.2 % max), consistent with that previously reported in Qin *et al.*, for eukaryotic DNA [6].

Sixteen pathway classes and 120 pathways were detected from the metagenomes. Five pathway classes (Amino Acid Metabolism, Carbohydrate Metabolism, Nucleotide Metabolism, Energy Metabolism, Metabolism of Cofactors and Vitamins) and 29 pathways (Supplementary Table S5), along with unidentified pathways and pathway classes represented more than 80 % of the sequences. The pattern of fungal functional structure indicated higher evenness compared with fungal taxonomic structure, i.e., the relative abundances of the pathways are closer instead of being dominated by one or two pathways.

3.4. Association between metadata and mycobiome composition and functions

Next, we evaluated the contribution of available variables, collected from the metadata files, to the mycobiome composition variations using the *adonis2* function from the *vegan* R package (Fig. 3). These variables included countries, health status, body sites, ages, gender, and bead-beating. Individually, countries and health status were the factors that contributed most to fungal composition and function variations; body sites and the bead-beating step also contributed to these variations, but to a lesser extent (FDR < 0.01, Fig. 3).

Associations between these variables and individual taxa were then examined using generalised linear models implemented in the *MaAsLin2* (Microbiome Multivariable with Linear Models)

Table 1
Summary of the characteristics of the 1,950 human metagenomes.

Body site	Country	Health status	Number of samples	Mechanical Lysis
Blood	USA	Filaria	1	no
		Lyme disease	1	no
Bone and joint	France	Infections	24	no
		HC	100	no
Conjunctiva	China	HC	100	no
Gallstones	Australia	NA	8	no
Gut	Australia	HC	56	yes
		T1D	60	yes
	Belgium	CD	92	yes
		PLWH	10	na
	Canada	HC	204	yes
		CD	38	yes
	China	ESRD	208	yes
		T2D	89	yes
	Denmark	NA	15	NA
		HC	165	no
	Israel	NA	20	na
		HC	18	yes
	Spain	HC	63	yes
		CD	50	yes
	Sweden	UC	69	yes
		T2D	10	yes
	USA	HC	11	no
		CD	13	na
		HIV	3	na
		PSO	24	no
		UC	10	na
		Infant-preterm	140	na
		NA	272	na
		HC	6	no
Nasal mucosa	Chile	Asthma	5	no
Oropharyngeal	South Africa	TB	4	na
Saliva	USA	NA	61	na
Skin	Italy	HC	3	yes
Sputum	Singapore	NA	30	na
		TB	10	no
Throat swab	USA	HC	16	no
		SCZ	14	no
Tongue	Italy	HC	12	yes
NA	USA	mock communities	15	na

PLWH = People live with HIV patients, PSO = Psoriasis, TB = Tuberculosis.

package. Five fungal species (*Aspergillus recurvatus*, *Malassezia restricta*, *Saccharomyces cerevisiae*, uncultured *Malassezia* spp., *Yarrowia lipolytica*), which were among the 10 most prevalent and abundant fungal species (Supplementary Table S3), were found associated with health status, country, and body sites (Supplementary Fig. S1A). This finding suggests that the high variability of the human mycobiome could be linked to these five species. Interestingly, *Yarrowia lipolytica* was found positively associated with bead-beating (Supplementary Fig. S1B), which could be explained by its relatively higher fraction of chitin (10.3–18.9 %) in the cell wall compared with *S. cerevisiae*, *C. albicans*, and *M. restricta* [37–39].

We found that geography, health status, and body sites had marked effects on the variability of most of the fungal pathway classes among the 16 that we recovered from all samples, yet bead-beating did not impact the compositions of fungal pathways, as reported for fungal taxa (Supplementary Fig. S1).

3.5. Core taxonomic fungal microbiomes of different body sites and different countries

To identify groups of key taxa that may influence the microbiome community, we applied the concept of core microbiome across body sites and geography, taking into account health status. For this purpose, fungal species with an occurrence of over 50 % in the respective set of metagenomes of interest, in which fungi were detected, were defined as the core mycobiome. The 50 % occur-

rence threshold was chosen based on the review of the core bacterial microbiome published by Neu *et al.* [40], but an abundance cutoff was not applied to avoid missing any lowly abundant fungal species. We summarised the core mycobiome for body sites (Table 2) and countries (Table 3). In the human gut mycobiome of non-infants, *S. cerevisiae* was found to be the only member of the core gut mycobiome, except for CD and T1D patients who were dominated by *Aspergillus recurvatus*. The core gut mycobiome of infants consisted of only species from the *Malassezia* genera, in accordance with several previous studies [41,42]. In other body sites, except saliva, several *Malassezia* species were the most detected members of the core mycobiome. The saliva mycobiome was driven by *Aspergillus recurvatus*.

Given that geographical difference contributes the most to fungal taxonomic structure variations, we also defined the core mycobiome for gut samples collected in different countries. We focused only on gut samples, as they represented the most available samples. *S. cerevisiae* appeared as a member of the core gut mycobiome in most countries (Table 3), which is in agreement with the aforementioned core mycobiome (Table 2). *A. recurvatus* was the only core fungal species among all the gut samples with different health status collected from Australia, whereas *Y. lipolytica* was that of the gut samples collected from end-stage renal disease (ESRD) patients in China (Table 3).

Core biochemical pathways, defined as pathways that have occurrences over 99 % among all the samples with a relative abundance of over 1 % [40], were also summarised for each body site

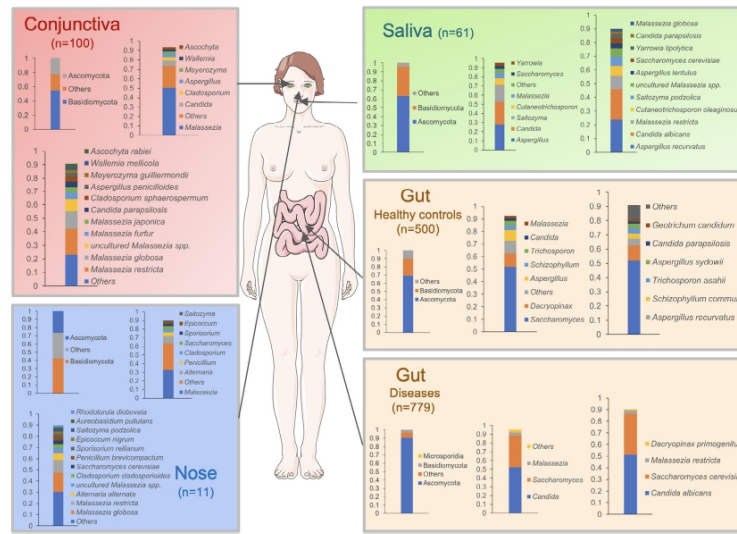


Fig. 2. Fungal taxonomic profiling of several human body sites based on the 1950 shotgun metagenomic data using the FunOmic-T database. Taxonomic profiling is displayed at the phylum, genus, and species levels. Only the mean relative abundance of the genera and species summing 90% of the sequence data is exhibited. Gut taxonomic profiling was performed for diseases including Crohn's disease (CD, $n = 193$; from the USA, Europe, and Asia), ulcerative colitis (UC, $n = 79$ from Europe and the USA), end-stage renal disease (ESRD, $n = 208$, from Asia), type 1 diabetes (T1D, $n = 60$ from Australia), and type 2 diabetes (T2D, $n = 99$ from Asia). 468 faecal samples did not have health status information in the metadata files. Health status and geo-localization of conjunctiva, nasal, and saliva samples are described in Table 1.

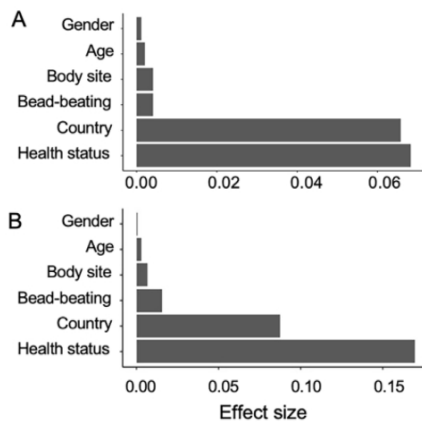


Fig. 3. Effect size of variables on the mycobiome community. The impact of the covariates on mycobiome composition (A) and function (B) was tested by performing a univariate analysis (adonis2) on the 1,950 metagenomes. The effect was considered significant when $FDR < 0.05$.

and country with different health status (Supplementary Table S6). For countries, only gut samples, as the most available sample type,

were considered. The majority of core fungal pathways were related to nucleotides, amino acids, energy and carbohydrate metabolisms, which are essential functions, indicating that the functionality of the human mycobiome is maintained across body niches and populations.

3.6. Bacterial and fungal microbiome interaction

Next, we sought to evaluate the correlations between fungal and bacterial taxonomic composition in gut samples under healthy conditions, especially concentrating on core fungal species. Because of the failure in detecting the core mycobiome under healthy conditions from China, we focused on the healthy conditions of Denmark and Spain. To address this aim, we first performed a bacterial taxonomic and functional profiling of the metagenomic data. Due to a very extensive computational time requirement (6 h/40 CPUs/sample on average), only a subset of 1,485 of the 2,679 metagenomic samples was processed (Fig. 1). We then carried out a correlation analysis with the SparCC correlation method, which handles compositional data [31] (34). In total, 4,184 significant ($p < 0.05$) inter-kingdom correlations were found in the Danish cohort, while 3,471 significant inter-kingdom correlations were found in the Spanish cohort, (Supplementary Table S7). In the Spanish cohort, the two core fungal species, *S. cerevisiae* and *D. primigenius*, were found to correlate with the bacterial species *Haemophilus pittmaniae* positively and negatively, respectively (Fig. 4A). Beyond that, in the Spanish cohort, *C. albicans* was found to negatively correlate with *Megasphaera sp MJR8396C*, which was positively correlated with *D. primigenius*. *C. albicans* was also found negatively correlated with *Lactobacillus sanfranciscensis*, *Bifidobacterium scardovii*, *Desulfovibrio fairfielden-*

Table 2
Core fungal species of different body sites.

Body site	Health status	Core fungal species (>50 % prevalence)
Gut	HC (n = 262)	<i>Saccharomyces cerevisiae</i>
	CD (n = 109)	<i>Aspergillus recurvatus</i>
	ESRD (n = 106)	<i>Saccharomyces cerevisiae</i>
	UC (n = 55)	<i>Saccharomyces cerevisiae</i>
	T1D (n = 40)	<i>Aspergillus recurvatus</i>
	T2D (n = 50)	<i>Saccharomyces cerevisiae</i>
	PSO (n = 16)	<i>Saccharomyces cerevisiae</i>
	PLWH (n = 7)	<i>Saccharomyces cerevisiae</i>
	Infant (n = 14)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.
	HC (n = 5)	<i>Alternaria alternata</i> , <i>Malassezia globosa</i>
Nasal mucosa	HC (n = 76)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.
Conjunctiva	NA (n = 38)	<i>Aspergillus recurvatus</i>
Saliva	HC (n = 14)	<i>Schizophyllum commune</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.
Throat swab	SCZ (n = 12)	<i>Candida albicans</i> , <i>Malassezia restricta</i>
Tongue dorsum	Infant (n = 8)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.
Bones and joints	BjJs (n = 24)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.
Gallstone	GS (n = 8)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.

Table 3
Core fungal species of different countries.

Country	Health status	Core fungal species (greater than 50 % prevalence)
Australia	HC (n = 46)	<i>Aspergillus recurvatus</i>
	T1D (n = 40)	<i>Aspergillus recurvatus</i>
Belgium	CD (n = 76)	<i>Aspergillus recurvatus</i> , <i>Saccharomyces cerevisiae</i>
China	ESRD (n = 106)	<i>Yarrowia lipolytica</i>
	T2D (n = 49)	<i>Saccharomyces cerevisiae</i>
Canada	PLWH (n = 7)	<i>Saccharomyces cerevisiae</i>
Denmark	HC (n = 118)	<i>Saccharomyces cerevisiae</i>
Italy	Infant (n = 14)	<i>Malassezia globosa</i> , <i>Malassezia restricta</i> , uncultured <i>Malassezia</i> spp.
Spain	HC (n = 57)	<i>Dacryopinax primigenitus</i> , <i>Saccharomyces cerevisiae</i>
	CD (n = 38)	<i>Saccharomyces cerevisiae</i>
USA	UC (n = 52)	<i>Saccharomyces cerevisiae</i>
	PSO (n = 16)	<i>Saccharomyces cerevisiae</i>

sis, *Ruminococcus* sp CAG563, *Coprococcus catus*, and *Roseburia* sp CAG309 (Supplementary Table S7, Fig. 4A), many of which are potential short-chain fatty acid (SCFA) producers [43]. In the Danish cohort, significant correlations were found between the only core fungal species, *S. cerevisiae*, and seven bacterial species, of which five were negative (*Tropheryma whipplei*, *Prevotella* sp CAG1124, *Firmicutes bacterium* CAG24, *Gemella sanguinis*, and *Sutterella parvirubra*) and two were positive (*Bacteroides nordii* and *Prevotella stercorea*) (Fig. 4B).

We also applied SparCC to analysing correlations between fungal and bacterial functions in gut samples under healthy conditions. In the Danish cohort, 93 significant correlations were detected (Supplementary Table S7, Supplementary Fig. S2A), of which the strongest was the positive correlation ($\rho = 0.06$, $p < 0.001$) between the biosynthesis of secondary metabolites in fungi and the endocrine system in bacteria. In the Spanish cohort, 76 significant correlations were detected (Supplementary Fig. S2B), the strongest was a negative correlation ($\rho = -0.13$, $p < 0.001$) between carbohydrate metabolism in fungi and signal transduction in bacteria. These functional inter-kingdom correlations could explain how bacteria and fungi interact in the microbiome community.

4. Discussion

Here, we have designed and validated FunOMIC, a metagenomic pipeline that integrates quality control, taxonomic profiling (FunOMIC-T), and functional profiling (FunOMIC-P) for a compre-

hensive analysis of fungi in environmental samples, and, particularly, in humans. First, to the best of our knowledge, FunOMIC offers the most comprehensive coverage of the reference fungal species and functions compared with other existing databases for profiling the human mycobiome. Indeed, FunOMIC-T, which contains more than 1.6 million fungal single-copy marker genes and covers 1,916 fungal species, exceeds the fungal spectrum of other similar tools [14,44,45]. We also proposed FunOMIC-P which includes more than 3 million non-redundant fungal proteins, which is, to our knowledge, the first protein database proposed for analysing human mycobiome functions. Second, FunOMIC-T provided a smaller-sized taxonomic database with more accurate mapping possibilities for mycobiome profiling using universal conserved fungal genes instead of the full genome-based fungal reference database. Third, validations with different mock communities mimicking the human gut microbiome ensured extremely low bacterial read mis-mapping.

In this study, we applied the FunOMIC pipeline to a set of nearly 2,700 metagenomic human samples representing human microbiomes of different body sites from individuals with different health status and from different geographical regions. We corroborated previous human mycobiome results showing that the species *S. cerevisiae*, *C. albicans*, and *M. restricta* dominate the fungal communities in different human body sites [46–49]. We found that geography and health status were the two most important factors contributing to the variabilities of human mycobiome taxonomic and functional compositions. Five fungal species (*A. recurvatus*, *M. restricta*, *S. cerevisiae*, uncultured *Malassezia* spp., *Y. lipolytica*) varied

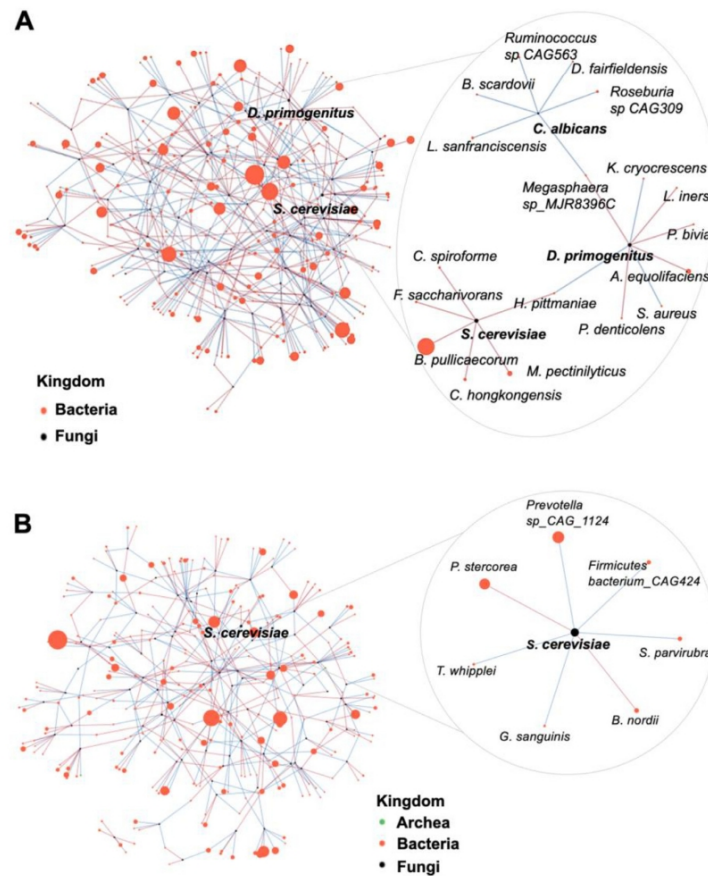


Fig. 4. Interaction of fungal and bacterial communities in gut microbiome under healthy conditions. Correlation network between the relative abundance of fungal and bacterial species in the gut mycobiome under healthy conditions from Spain (A) and Denmark (B) using the SparCC algorithm. Each node represents a fungal/bacterial/archaeal species and their sizes are determined by relative abundances. The colours of the edges connecting two nodes represent positive (red) and negative (blue) correlations. For a better visual effect, only correlations with p -values < 0.001 and an absolute correlation coefficient over 0.05 are represented. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

along with different countries, health status, and body sites. *C. albicans*, one of the most common human fungal pathogens [50], negatively correlated with bacterial species that are mainly SCFA producers [43]. This finding suggests that therapeutic strategies based on SCFA administration or on inducing SCFA producers could be implemented to control *C. Albicans* infection.

One important limitation of this pipeline is that the extraction and quality of single-copy marker genes rely on the completeness of the available fungal genomes, which may result in a lower coverage of fungal taxonomies compared with the fungal amplicon databases [23,51]. Another limitation comes from the high inter-

kingdom conservation of a portion of protein-coding genes. As a consequence, bacterial contamination was not totally preventable, even after applying an exceedingly strict mapping threshold to the fungal functional annotation. To overcome this drawback, filtration to remove the majority of bacterial reads before functional annotation could be included in the future update of this tool. Beyond that, in this study, FunOMIC was only applied to human microbiome data; in the future, applications with soil microbiome, marine microbiome, or other different environmental samples will be launched with FunOMIC to test its ability to handle other microbiome data.

5. Conclusions

Taken together, our work presented here demonstrates that the proposed taxonomic database FunOMIC-T can effectively detect fungal species from shotgun metagenomic sequencing data. Together with FunOMIC-P, which to our knowledge, the first proposed functional database for mycobiome analysis, we believe that more mycobiome findings will be revealed in the future.

6. Data access

The two built-in databases, FunOMIC-T and FunOMIC-P, are freely available at <https://manichanh.vhir.org/funomic/>. The source code of pipeline FunOMIC is freely available at our GitHub (<https://github.com/ManichanhLab/FunOMIC>).

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Action, Innovative Training Network [grant number 812969] and by the Instituto de Salud Carlos III / FEDER, a government agency (Grant No: P117/00614; PI20/00130).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2022.07.010>.

References

- [1] Seed PC. The human mycobiome. *Cold Spring Harb Perspect Med* 2014;5(5):a019810.
- [2] Uguori G, Lamas B, Richard ML, Brandi G, da Costa G, Hoffmann TW, et al. Fungal Dysbiosis in Mucosa-associated Microbiota of Crohn's Disease Patients. *J Crohns Colitis* 2016;10(3):296–305.
- [3] Santus W, Devlin JR, Behnsen J. Crossing Kingdoms: How the Mycobiota and Fungal-Bacterial Interactions Impact Host Health and Disease. *Infect Immun* 2021;89(4).
- [4] van Tilburg BE, Pettersen VK, Gutierrez MW, Laforest-Lapointe I, Jendzjowsky NG, Cavin JB, et al. Intestinal fungi are causally implicated in microbiome assembly and immune development in mice. *Nat Commun* 2020;11(1):2577.
- [5] Sun Y, Zuo T, Cheung CP, Gu W, Wan Y, Zhang F, et al. Population-Level Configurations of Gut Mycobiome Across 6 Ethnicities in Urban and Rural China. *Gastroenterology* 2021;160(1):272–86 e11.
- [6] Qin J, Li R, Raes J, Arunugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464(7285):59–65.
- [7] Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A* 2012;109(16):6241–6.
- [8] Andersen LO, Vedel Nielsen H, Stensvold CR. Waiting for the human intestinal Eukaryotome. *ISME J* 2013;7(7):1253–5.
- [9] Del Campo J, Pons MJ, Herranz M, Wakeman KC, Del Valle J, Vermeij MJA, et al. Validation of a universal set of primers to study animal-associated microeukaryotic communities. *Environ Microbiol* 2019;21(10):3855–61.
- [10] Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* 2018;6(1):41.
- [11] Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H, et al. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* 2010;4(5):642–7.
- [12] Lofgren LA, Uehling JK, Branco S, Bruns TD, Martin F, Kennedy PG. Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Mol Ecol* 2019;28(4):721–30.
- [13] Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. *Nat Methods* 2013;10(9):881–4.
- [14] Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* 2021;9(1):58.
- [15] Marcelino VR, Clausen P, Buchmann JP, Wille M, Iredell JR, Meyer W, et al. CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol* 2020;21(1):103.
- [16] West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* 2018;28(4):569–80.
- [17] Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otiillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* 2014;42(Database issue):D699–704.
- [18] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;32(8):834–41.
- [19] Montoliu-Nerin M, Sanchez-Garcia M, Bergin C, Grabherr M, Ellis B, Kutschera VE, et al. Building de novo reference genome assemblies of complex eukaryotic microorganisms from single nuclei. *Sci Rep* 2020;10(1):1303.
- [20] Bankveich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19(5):455–77.
- [21] Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2019;47(D1):D807–11.
- [22] Manni M, Berkeley MR, Seppely M, Simao FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* 2021;38(10):4647–54.
- [23] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;41(Database issue):D590–6.
- [24] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150–2.
- [25] 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* 2021;39(1):105–14.
- [26] Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;28(4):593–4.
- [27] Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database C. The sequence read archive. *Nucleic Acids Res* 2011;39(Database issue):D19–21.
- [28] Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 2019;10(1):1014.
- [29] Franzosa EA, Mciver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 2018;15(11):962–8.
- [30] Suzek BE, Wang Y, Huang H, McCarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31(6):926–32.
- [31] Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012;8(9):e1002687.
- [32] Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat Biotechnol* 2017;35(11):1069–76.
- [33] d'Enfert C, Kaune AK, Alaban LR, Chakraborty S, Cole N, Delavy M, et al. The impact of the Fungus-Host-Microbiota interplay upon *Candida albicans* infections: current knowledge and new perspectives. *FEMS Microbiol Rev* 2021;45(3).
- [34] Chao A. Nonparametric estimation of the number of classes in a population. *Scand J Stat* 1984;11(4):6.
- [35] Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;27(3):379–423.
- [36] Serrano-Gomez G, Mayorga L, Oyazun I, Roca J, Borruel N, Casellas F, et al. Dysbiosis and relapse-related microbiome in inflammatory bowel disease: A shotgun metagenomic approach. *Comput Struct Biotechnol J* 2021;19:6481–9.
- [37] Chaffin WL, Lopez-Ribot JL, Casanova M, Gozalbo D, Martinez JP. Cell wall and secreted proteins of *Candida albicans*: identification, function, and expression. *Microbiol Mol Biol Rev* 1998;62(1):130–80.
- [38] Chattaway FW, Holmes MR, Barlow AJ. Cell wall composition of the mycelial and blastospore forms of *Candida albicans*. *J Gen Microbiol* 1968;51(3):367–76.
- [39] Stahlberger T, Simenel C, Clavaud C, Eijnsink VG, Jourdain R, Delepierre M, et al. Chemical organization of the cell wall polysaccharide core of *Malassezia restricta*. *J Biol Chem* 2014;289(18):12647–56.
- [40] Neu AT, Allen EE, Roy K. Defining and quantifying the core microbiome: challenges and prospects. *Proc Natl Acad Sci U S A* 2021;118(51).
- [41] Boutin RCT, Sbihi H, McLaughlin RJ, Hahn AS, Konwar KM, Luo RS, et al. Composition and Associations of the Infant Gut Fungal Microbiota with Environmental Factors and Childhood Allergic Outcomes. *mBio*. 2021;12(3):e0339620.

- [42] Ventin-Holmberg R, Eberl A, Saqib S, Korpela K, Virtanen S, Sipponen T, et al. Bacterial and fungal profiles as markers of infliximab drug response in inflammatory bowel disease. *J Crohns Colitis* 2021;15(6):1019–31.
- [43] Parada Venegas D, De la Fuente MK, Landskron G, Gonzalez MJ, Quera R, Dijkstra G, et al. Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Front Immunol* 2019;10:277.
- [44] Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K. Identification of fungi in shotgun metagenomics datasets. *PLoS ONE* 2018;13(2):e0192898.
- [45] Soverini M, Turroni S, Biagi E, Brigidi P, Candela M, Rampelli S, HumanMycobiomeScan: a new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. *BMC Genomics* 2019;20(1):496.
- [46] Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, et al. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog* 2010;6(1):e1000713.
- [47] Gupta S, Hjelmsø MH, Lehtimäki J, Li X, Mortensen MS, Russel J, et al. Environmental shaping of the bacterial and fungal community in infant bed dust and correlations with the airway microbiota. *Microbiome* 2020;8(1):115.
- [48] Hamad I, Ranque S, Azhar El, Yasir M, Jiman-Fatani AA, Tissot-Dupont H, et al. Culturomics and Amplicon-based Metagenomic Approaches for the Study of Fungal Population in Human Gut Microbiota. *Sci Rep* 2017;7(1):16788.
- [49] Zhang E, Tanaka T, Tajima M, Tsuboi R, Nishikawa A, Sugita T. Characterization of the skin fungal microbiota in patients with atopic dermatitis and in healthy subjects. *Microbiol Immunol* 2011;55(9):625–32.
- [50] Kim J, Sudbery P. *Candida albicans*, a major human fungal pathogen. *J Microbiol* 2011;49(2):171–7.
- [51] Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 2019;47(D1):D259–64.