HUMANS AS HOLOBIONTS: SYSTEMS-LEVEL APPROACHES FOR DISEASE PREVENTION AND THERAPY

Dissertation

in Partial Fulfilment of the Requirements for the Degree of

“Doctor of Philosophy” (PhD)

Submitted to the Council of

the Faculty of Biological Sciences

of Friedrich Schiller University Jena

by M.Sc. Sara Leal Siliceo
born on July 6th, 1994 in Madrid

This dissertation was prepared at the Leibniz Institute for Natural Product Research and Infection Biology e.V – Hans Knöll Institute (HKI), Jena in the Microbiome Dynamics Group under the supervision of Prof. Dr. Gianni Panagiotou. This project has received funding from the European Union Horizon 2020 Research and Innovation programme under the Marie Sklodowska-Curie grant agreement No 813781.

Reviewers:

Prof. Dr. Gianni Panagiotou
Leibniz Institute for Natural Product Research and Infection Biology. Hans-Knöll-Institut
Jena

Prof. Dr. Michael Bauer
Universitätsklinikum Jena

Prof. Dr. Susanne Brix Pedersen
Technical University of Denmark

Date of public defense: 11th April 2024, online

SUMMARY

The microbes that live in the gut, also known as the gut microbiota, play an important role in the well-being of the host. In the last years, the development of metagenomics and metabolomics have helped to better understand the vital role of the gut microbiome in human health and disease; however, the mechanisms and its implication are still not fully clarified. Therefore, more research and improved pipelines, protocols, and tools are needed to investigate gut microbiome in more depth and understand its connection with host health.

This dissertation aimed to develop and implement bioinformatic and statistical analyses to improve our understanding of the role of the gut microbiome in non-alcoholic fatty liver disease (NAFLD) pathogenesis. In addition, during my Ph.D. I focused on investigating novel microbiome-based therapeutic strategies.

NAFLD is the hepatic manifestation of a metabolic syndrome, and its prevalence has reached epidemic proportions with a global prevalence of about 32.4%. Previous works have demonstrated a link between gut microbiome dysbiosis and NAFLD progression. In this dissertation, three different projects have focused on investigating the role of the gut microbiome in NAFLD (**manuscripts I, III, and IV**). Due to the lack of pharmaceutical treatment for NAFLD, new strategies are being studied. In **manuscript I,** the effect of 4-month resistant starch (RS) supplementation as one type of microbiome-directed food was investigated in a cohort of NAFLD individuals. We showed that 4-month RS intervention ameliorates NAFLD, and multi-omics profiling provided an integrated understanding of how RS and associated alterations in the gut microbiota or metabolites contributed to NAFLD improvement. In addition, whole microbiota changes, the potential RS-targeted single species, and microbial metabolites were validated *in vivo* and *in vitro* for causal insights. The study has been accepted for publication in the journal **Cell Metabolism**, and will be published as the cover article of the journal.

In **manuscript IV**, the potential value of the gut microbiome in NAFLD prognosis was investigated, where differences in the microbiome signature and metabolic shifts in subjects that will develop NAFLD compared to controls were shown. In this project, potential microbial markers for NAFLD prognosis were identified and a machine learning model able to predict the development of NAFLD was developed. The study was published in the journal **Science Translational Medicine.**

Concerning the mycobiome, very little is known about how the fungal composition contributes to NAFLD progression. In **manuscript III,** a possible antifungal immunity and potential mycobiome dysbiosis were explored to investigate how intestinal fungi contribute to NAFLD development. Our results showed that NAFLD patients harboring a genetic variation in their IL-17A gene also presented increased *Candida* CTG species levels, and these two factors predispose to disease progression up to steatohepatitis (NASH) and advanced fibrosis. All these three projects together provide a more comprehensive understanding of the connection of the gut bacteriome, mycobiome, and metabolome changes with NAFLD development. Furthermore, novel microbiome-based therapeutic

techniques for NAFLD were explored, including a microbiota-directed food intervention and a NAFLD prognostic assessment tool.

Lastly, the utilization of lifestyle interventions targeting the gut microbiome to improve host health is a commonly used practice nowadays. A fourth study in this dissertation aimed to explore the gut microbiome dynamics in response to lifestyle microbiome-targeted therapies. In **manuscript II,** robust and generalizable biomarkers within the gut microbial communities that are associated with resistance to gut microbial community change were identified. Moreover, a machine learning model able to predict the gut microbiome resistance to change in response to lifestyle interventions using the baseline microbiome composition was developed. The study was published in the journal **Microbiome.**

In conclusion, in this dissertation, I have shown that the human body and its microbiome form a unity of life or holobiont that is indispensable for the well-functioning of the organism. The different bioinformatic analyses performed during my Ph.D. have highlighted the important role of the gut microbiome in human health and disease, especially giving new insights in relation to NAFLD pathogenesis and management. In addition, I have explored different microbiome-based strategies showing the high potential of the gut microbiome in the development of new therapies. Therefore, the use of microbiome-related information for patient therapeutics needs to be further explored and applied to improve and develop new and more personalized treatments.

ZUSAMMENFASSUNG

Die im Darm lebenden Mikroben, auch bekannt als Darmmikrobiota, spielen eine wichtige Rolle für das Wohlbefinden des Wirts. In den letzten Jahren hat die Entwicklung der Metagenomik und der Metabolomik dazu beigetragen, die wichtige Rolle des Darmmikrobioms für die menschliche Gesundheit und Krankheit besser zu verstehen. Dennoch sind die Mechanismen und ihre Auswirkungen noch nicht vollständig geklärt. Daher sind weitere Forschungsarbeiten und verbesserte Pipelines, Protokolle und Instrumente erforderlich, um das Darmmikrobiom eingehender zu untersuchen und seinen Zusammenhang mit der Gesundheit des Wirts zu verstehen.

Ziel dieser Dissertation war es, bioinformatische und statistische Analysen zu entwickeln und zu implementieren, um unser Verständnis der Rolle des Darmmikrobioms bei der Pathogenese der nichtalkoholischen Fettlebererkrankung (NAFLD aus dem Englischen) zu verbessern. Darüber hinaus konzentrierte ich mich während meiner Doktorarbeit auf die Erforschung neuer mikrobiombasierter therapeutischer Strategien.

Die NAFLD ist die hepatische Manifestation eines metabolischen Syndroms und hat mit einer weltweiten Prävalenz von etwa 32,4 % epidemische Ausmaße erreicht. Frühere Arbeiten haben einen Zusammenhang zwischen einer Dysbiose des Darmmikrobioms und dem Fortschreiten der NAFLD nachgewiesen. In dieser Dissertation wurde in drei verschiedenen Projekten die Rolle des Darmmikrobioms bei NAFLD untersucht (**Manuskripte I, III und IV**). Da es keine pharmazeutische Behandlung für NAFLD gibt, werden neue Strategien untersucht. In **Manuskript I** wurde die Wirkung einer 4-monatigen Supplementierung mit resistenter Stärke (RS) (eine Art mikrobiomgesteuerter Nahrungsmittel) in einer Kohorte von Personen mit NAFLD untersucht. Wir konnten zeigen, dass eine 4-monatige RS-Intervention die NAFLD verbessert, und die Multi-omics-Profilierung lieferte ein integriertes Verständnis dafür, wie RS und die damit verbundenen Veränderungen in der Darmmikrobiota oder den Metaboliten zur Verbesserung der NAFLD beitragen. Darüber hinaus wurden die Veränderungen der gesamten Mikrobiota, die potenziell auf RS abzielenden einzelnen Spezies und die mikrobiellen Stoffwechselprodukte in vivo und in vitro validiert, um kausale Erkenntnisse zu gewinnen. Die Studie ist zur Veröffentlichung in der Zeitschrift **Cell Metabolism** angenommen worden, und wird als Titelartikel in der Zeitschrift veröffentlicht.

In **Manuskript IV** wurde der potenzielle Wert des Darmmikrobioms für die NAFLD-Prognose untersucht, indem Unterschiede in der Mikrobiomsignatur und Stoffwechselverschiebungen bei Betroffenen, die NAFLD entwickeln werden, im Vergleich zu Kontrollen aufgezeigt wurden. Potenzielle mikrobielle Marker für die NAFLD-Prognose wurden im Rahmen dieses Projekt identifiziert und ein maschinelles Lernmodell vorgestellt, das die Entwicklung von NAFLD vorhersagen kann. Die Studie wurde in der Zeitschrift **Science Translational Medicine** veröffentlicht.

In Bezug auf das Mykobiom ist nur sehr wenig darüber bekannt, wie die Pilzzusammensetzung zum Fortschreiten der NAFLD beiträgt. In **Manuskript III** wurden

eine mögliche antimykotische Immunität und eine mögliche Mykobiom-Dysbiose untersucht, um herauszufinden, wie Darmpilze zur Entwicklung der NAFLD beitragen. Unsere Ergebnisse zeigten, dass NAFLD-Patienten mit einer genetischen Variation in ihrem IL-17A-Gen auch erhöhte Werte von Candida-CTG-Speziesaufweisen, und diese beiden Faktoren prädisponieren für ein Fortschreiten der Krankheit bis hin zu Steatohepatitis (NASH) und fortgeschrittener Fibrose. Unsere Ergebnisse zeigten, dass NAFLD-Patienten mit einer genetischen Variation in ihrem IL-17A-Gen auch erhöhte Candida-CTG-Spezies-Spiegel aufwiesen, und diese beiden Faktoren prädisponieren für ein Fortschreiten der Krankheit bis hin zu Steatohepatitis (NASH) und fortgeschrittener Fibrose. Alle drei Projekte zusammen ermöglichen ein umfassenderes Verständnis des Zusammenhangs zwischen dem Darmbakteriom, dem Mykobiom und den Veränderungen des Metaboloms mit der Entwicklung der NAFLD. Außerdem wurden neuartige mikrobiombasierte therapeutische Verfahren für die NAFLD erforscht, darunter eine auf das Mikrobiom ausgerichtete Ernährungsintervention und ein Instrument zur prognostischen Bewertung der NAFLD.

Schließlich ist der Einsatz von auf das Darmmikrobiom abzielenden Lebensstilinterventionen heutzutage eine verbreitete angewandte Praxis zur Verbesserung der Gesundheit des Wirts. Eine vierte Studie in dieser Dissertation zielte darauf ab, die Dynamik des Darmmikrobioms als Reaktion auf mikrobiomorientierte Lifestyle-Therapien zu untersuchen. In **Manuskript II** wurden robuste und verallgemeinerbare Biomarker innerhalb der Darmmikrobiota identifiziert, die mit der Resistenz gegen Veränderungen der mikrobiellen Darmgemeinschaft in Verbindung gebracht wurden. Darüber hinaus wurde ein maschinelles Lernmodell entwickelt, das die Resistenz des Darmmikrobioms gegen Veränderungen als Reaktion auf Lebensstilmaßnahmen anhand der Zusammensetzung des Ausgangsmikrobioms vorhersagen kann. Die Studie wurde in der Zeitschrift **Microbiome** veröffentlicht.

Abschließend habe ich in dieser Dissertation gezeigt, dass der menschliche Körper und sein Mikrobiom eine Lebenseinheit oder einen Holobionten bilden, der für das gute Funktionieren des Organismus unerlässlich ist. Die verschiedenen bioinformatischen Analysen, die ich während meiner Doktorarbeit durchgeführt habe, haben die wichtige Rolle des Darmmikrobioms für die menschliche Gesundheit und Krankheit hervorgehoben und insbesondere neue Erkenntnisse in Bezug auf die Pathogenese und das Management der NAFLD geliefert. Darüber hinaus habe ich verschiedene mikrobiombasierte Strategien erforscht, die das große Potenzial des Darmmikrobioms für die Entwicklung neuer Therapien aufzeigen. Daher muss die Nutzung mikrobiombezogener Informationen für Patiententherapien weiter erforscht und angewendet werden, um neue und stärker personalisierte Behandlungen zu verbessern und zu entwickeln.

TABLE OF CONTENTS

SUMMARY .... **I**

ZUSAMMENFASSUNG .... **III**

TABLE OF CONTENTS .... **V**

ABBREVIATIONS .... **VII**

INTRODUCTION **1**

1. The human microbiome 1
    - 1.1 The gut bacteriome and mycobiome 2
2. The role of gut bacteria and fungi in metabolic diseases 2
    - 2.1 Non-alcoholic fatty liver disease and gut dysbiosis 3
3. Gut microbiome modulation 6
    - 3.1 Resistant starch as microbiota-directed food intervention 7
4. Resistance and resilience of the human gut microbiome 8
5. Bioinformatics approaches for studying the microbiome 9
    - 5.1 Amplicon sequencing 10
    - 5.2 Whole-genome shotgun sequencing 11
    - 5.3 Metabolomics 12
6. Statistical and machine learning techniques in metagenomics 12
    - 6.1 Methods for abundance comparisons 13
    - 6.2 Methods for correlation analysis 14
    - 6.3 Methods for network analysis 14
    - 6.4 Machine learning 16

OBJECTIVES **19**

RESEARCH PUBLICATIONS **20**

- Manuscript I 21
    - Overview 21
    - Manuscript 23
- Manuscript II 61
    - Overview 61
    - Manuscript 63

Manuscript III 79
    Overview 79
    Manuscript 81
Manuscript IV 105
    Overview 105
    Manuscript 107
DISCUSSION 121
1. Importance of confounders in microbiome-based data analyses 122
2. Using metagenomics and metabolomics analyses to investigate potential therapies and diagnostic strategies for NAFLD 124
3. Bioinformatic and machine learning approaches allowed the development of a microbiome-based stratification model 127
4. Limitations in mycobiome analyses 129
5. Synthesis of the four studies comprising my dissertation 131
REFERENCES 133
DECLARATION 149
ACKNOWLEDGMENTS 150
CURRICULUM VITAE 151
APPENDIX 153
Appendix 1. Form 2 of the own contribution to the manuscripts 153

ABBREVIATIONS

AAAs	Aromatic Amino Acids
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
AUC	Area Under Curve
BCAAs	Branched-Chain Amino Acids
BCFAs	Branched-Chain Fatty Acids
BMI	Body Mass Index
CAP	Controlled Attenuation Parameter
CT	Computed Tomography
FGF21	Fibroblast Growth Factor 21
FIB-4	Fibrosis-4
FMT	Fecal Microbiome Transplant
GO	Gene Ontology
HMP	Human Microbiome Project
IBD	Inflammatory Bowel Disease
ITS	Internal Transcribed Spacer
KOs	KEGG Orthologs
LDL	Low-Density Lipoprotein
LPS	Lipopolysaccharide
MDFs	Microbiota-Directed Foods
ML	Machine Learning
MetaHIT	METAgenomics of the Human Intestinal Tract
MRS	Magnetic Resonance Spectrometry
NAFLD	Non-Alcoholic Fatty Liver Disease
NASH	Non-Alcoholic Steatohepatitis
NDCs	Non-Digestible Carbohydrates
NGS	Next Generation Sequencing
NIH	National Institutes of Health
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction

ROC	Retriever Operating Characteristic
rRNA	Ribosomal Ribonucleic Acid
RS	Resistant Starch
SCFAs	Short-Chain Fatty Acids
T2D	Type 2 Diabetes
TE	Transient Elastography
TG	Triglycerides
WGS	Whole Genome Sequencing

INTRODUCTION

1. The human microbiome

Microbes are microscopic organisms that are found in almost every environment and are essential to life. In the popular imaginary, they are generally linked with disease, however, most microbes are beneficial. They modulate key ecosystem processes and participate in functions such as plant growth, marine biogeochemical cycles, or food digestion (Malla et al. 2019). The human body is colonized by trillions of commensal, symbiotic, and pathogenic microorganisms that constitute the human microbiota, and the collection of genomes of an organism is known as the microbiome. The assembled of the host organism together with its microbiota is known as "holobiont" (Greek, from holos, whole; bios, life; -ont, to be; whole unit of life) a concept that was proposed by Lynn Margulis in 1991 (Thomas et al. 2017). Investigating the holobiont by exploring the interactions between the hosts and their associated microbial communities is crucial as it has been shown that the microbiota and the host mutually affect each other, being involved in health and disease development (Postler and Ghosh 2017). Therefore, the microbiome has an important role in regulating human health and functioning. This complex community is composed of bacteria (bacteriome), archaea (archaeome), fungi (mycobiome), and viruses (virome) (Lloyd-Price et al. 2016). However, to date, most of the studies have focused on the composition of the bacterial microbiota and their implication for human health, leaving the mycobiome, archaeome, and virome poorly understood (Matijašić et al. 2020).

Some of the large-scale international projects that strongly impacted microbiome research are the Human Microbiome Project (HMP) launched in 2007 and the EU FP7 METAgenomics of the Human Intestinal Tract (MetaHIT) project launched in 2008, financed by the National Institutes of Health (NIH) and the European Commission, respectively (Turnbaugh et al. 2007; Ehrlich 2011). Both projects helped to characterize the human-associated microbial communities and their alterations in different human pathologies. The HMP initially aimed to characterize the 'healthy' human microbiome, as well as the characterization of the microbiome of the different sites of the human body such as the skin, oral cavity, respiratory tract, gastrointestinal tract, urinary tract, etc (Nash et al. 2017). Since then, many studies have focused on investigating these two objectives. Numerous studies have shown that the different sites in the human body differ greatly in terms of their microbiome composition and functions (Lloyd-Price et al. 2016; Ward et al. 2018; Dekaboruah et al. 2020). In addition, characterizing the healthy microbiome and investigating the link between disease and the imbalance in the composition and function of microbial taxa, also known as dysbiosis, is crucial to determine the role of the microbiome in contributing to health and disease (Lloyd-Price et al. 2016; Koh and Kim 2017).

1.1 The gut bacteriome and mycobiome

The human gastrointestinal tract is a diverse and abundant microbial community made up of more than 100 trillion microorganisms, being the colon one of the most densely populated microbial habitats known on Earth (Rinninella et al. 2019). These digestive tract-associated microbes are known as the gut microbiome, and it is predominantly composed of bacteria. The mycobiome has been considered a minor component of the gut microbiota representing approximately less than 0.1% of the microbial community in the gut (Fotis et al. 2022). However, in the last years, the mycobiome has gained recognition as a fundamental part of the gut microbiome community (Zhang et al. 2021).

Due to its essential role in human health, the microbial community in the gastrointestinal tract has been widely studied as it is involved in host metabolism, immune system education and regulation, maintenance of structural integrity of the gut mucosal barrier, and protection against pathogen invasion (Jandhyala et al. 2015).

Several factors are involved in shaping the gut microbiota composition. For example, age has been proven to be a major contributor to differences in gut microbiota (Bosco and Noti 2021). Diet and the use of medication (e.g., antibiotics) are also considered to be key modulators of the gut microbial community (Valdes et al. 2018). The use of active microorganisms that colonize the human intestines and change the composition of the flora in particular parts of the host, also called probiotics, have been proven to play important roles in the gut microbiota community. Probiotics inhibit the colonization of pathogenic bacteria in the intestine, help the host to build a healthy intestinal mucosa protective layer, and enhance the host's immune system (Wang et al. 2021). In early stages of life, environmental factors such as the delivery mode or breastfeeding seem to have important effects on the microbial colonization of infants (Arrieta et al. 2014; Zhuang et al. 2019; Korpela 2021; Henderickx et al. 2022), and previous studies suggest that aberrant early microbial exposures have long-term immunological and metabolic consequences in the future development of the newborns (Yao et al. 2021).

2. The role of gut bacteria and fungi in metabolic diseases

In the last decades, the prevalence of metabolic disorders such as disturbed glucose metabolism, general and abdominal obesity, elevated blood pressure, dyslipidemia, insulin resistance, hyperglycemia, and hyperuricemia; has dramatically increased. These conditions are all risk factors for numerous serious diseases such as type 2 diabetes mellitus (T2D), non-alcoholic fatty liver disease (NAFLD), and inflammatory bowel disease (IBD), among others. The abnormally elevated levels of lipids in the blood or hyperlipidemia is also considered a high-risk factor and a key indicator of many metabolic diseases, and it has been reported to play a vital role in regulating host lipid metabolism (Jia et al. 2021). Therefore, in the last years, metabolic disorders have become a growing worldwide health challenge (Stephens et al. 2020).

The gut microbiota plays a role in regulating the host metabolism, and alteration of the gut microbiota's taxonomic composition and functions are associated with metabolic disorders (Magro et al. 2019; Glassner et al. 2020; Li et al. 2020; He et al. 2021). Microbiota changes including a decrease in diversity have been found in IBD subjects over time. A decrease in *Firmicutes* species and an increase in *Proteobacteria* species were seen in association with IBD (Glassner et al. 2020). In relation to fungal composition, *Candia* species were found to increase in IBD individuals compared to controls, whereas *Saccharomyces cerevisiae* levels decreased (Glassner et al. 2020). In subjects with Crohn’s disease, a type of IBD, a lower microbial diversity compared to controls was also found (Magro et al. 2019). Furthermore, greater abundance in the Proteobacteria phylum and a reduction in *Akkermansia* and *Oscillospira* bacterial genera and *Saccharomyces cerevisiae* fungal species was identified in subjects with Crohn’s disease (Magro et al. 2019). Obesity has also been associated with gut bacteriome and mycobiome dysbiosis (Rodríguez et al. 2015; Pinart et al. 2022). Lower relative proportions of *Bifidobacterium* and *Eggerthella*, and higher *Acidaminococcus*, *Dialister*, *Dorea*, *Prevotella*, and *Roseburia* were found in obese versus non-obese adults (Pinart et al. 2022). Concerning the fungal composition, an increased presence of the phylum Ascomycota and families Dipodascaceae and Saccharomycetaceae as well as an increase in the relative abundance of fungi belonging to the class Tremellomycetes were found in obese compared with non-obese subjects (Rodríguez et al. 2015). Along with bacteria and fungi, microbial communities like the virome are also altered in metabolic disorders including hypertension and T2D (Ma et al. 2018; Han et al. 2018). Therefore, investigating the complex interactions between different gut microbial communities may enhance our comprehension of disease development and progression.

Studies combining metabolomics and metagenomics are being used to elucidate the link between gut microbiota dysbiosis and metabolic disturbances, becoming a key focus of study in the characterization and progression of metabolic diseases (Agus et al. 2021). Specific kinds of microbiota-derived metabolites, such as short-chain fatty acids (SCFAs), branched-chain amino acids (BCAAs), aromatic amino acids (AAAs), or tryptophan have been implicated in the pathogenesis of metabolic disorders (Schnabl and Brenner 2014; Lavelle and Sokol 2020; Xiao et al. 2021). Ejtahed et al. found that the gut microbiota of obese individuals compared with lean controls may have a higher capacity for production of some SCFAs, branched-chain fatty acids (BCFAs), and AAAs, known as risk factors for some metabolic-related diseases (Ejtahed et al. 2020). Plasma concentrations of BCAAs are also frequently elevated in obesity and T2D (Cuomo et al. 2022). In addition, BCAAs (valine, isoleucine, and leucine) concentrations in plasma and urine were found to be associated with insulin resistance (Cuomo et al. 2022).

2.1 Non-alcoholic fatty liver disease and gut dysbiosis

Non-alcoholic fatty liver disease (NAFLD) is the hepatic manifestation of a combination of metabolic dysfunctions mainly characterized by insulin resistance, dyslipidemia, impaired glucose tolerance, abdominal adiposity, and hypertension, collectively known as cardiometabolic syndrome (Younossi 2019). The main liver condition that characterizes NAFLD, as the name indicates, consists of too much fat stored in the liver cells in individuals who drink very little or no alcohol. NAFLD can evolve into non-alcoholic steatohepatitis (NASH), characterized by inflammation and fibrosis in the liver, and progressively lead to liver cirrhosis and hepatocellular carcinoma (Figure 1) (Younossi 2019). NAFLD is the leading cause of liver-related morbidity and mortality, being the most prevalent liver disease with an estimated global prevalence of up to 32.4% (Riazi et al. 2022). In addition, awareness about the disease among the general population is very low, and NAFLD is creating an extraordinary burden of clinical- and economic-related factors (Lazarus et al. 2022).

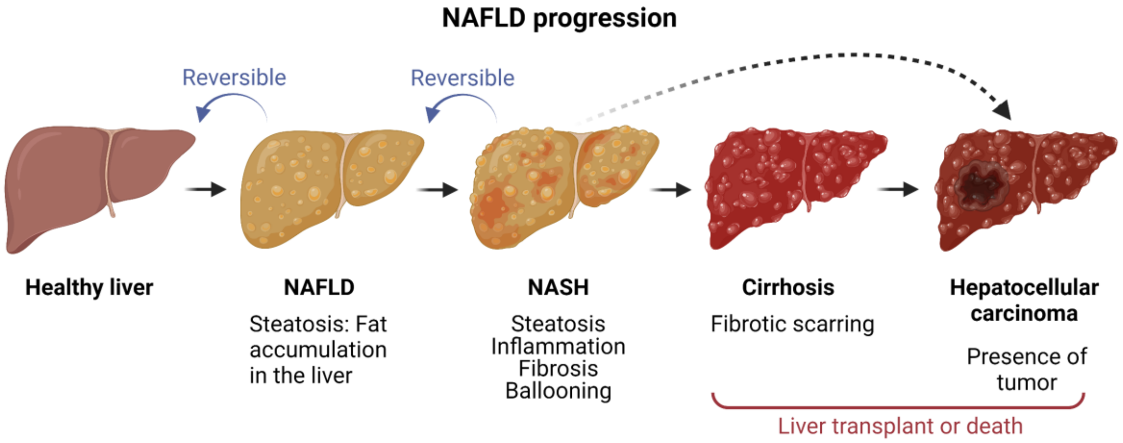


Figure 1 | NAFLD liver progression. Figure created with Biorender.

Even though the pathogenesis of NAFLD is not fully clarified, it is thought that dysbiosis, diet, genetics, and changes in intestinal permeability are risk factors that drive the progression from simple steatosis to NAFLD (Dongiovanni and Valenti 2017; Kolodziejczyk et al. 2019; Hu et al. 2020). Evidence has shown that the gut microbiome is closely related to the pathogenesis of NAFLD and contributes to the development of the disease via the gut-liver axis (Tripathi et al. 2018; Bauer et al. 2022). Boursier et al. investigated the changes in the microbiota in subjects with and without NASH and found that the bacterial genera *Bacteroides* and *Ruminococcus* were substantially increased, and *Prevotella* was reduced in patients with NASH (stage 2 fibrosis or higher) (16S) (Boursier et al. 2016). Loomba et al. studied the gut microbiota in patients with NAFLD with and without advanced fibrosis (stages 3 and 4) and showed an increased abundance of *Escherichia coli* and *Bacteroides vulgatus* in patients with advanced fibrosis (Loomba et al. 2017). Studies characterizing the intestinal microbiome in NAFLD have mostly focused on bacteria. Only one recently published study investigated the fungal community in the progression of NAFLD suggesting the mycobiome as a novel and relevant modulator of the development of the disease (Demir et al. 2022). In this study, Demir et al. identified that advanced NAFLD severity in non-obese subjects was associated with distinct fecal mycobiome signatures. In addition, there was an increased systemic immune response to *Candida albicans* in patients with NAFLD and advanced fibrosis. In **manuscript III** we investigated the fungal changes and genetic variations in antifungal immunity in a NAFLD cohort, whereas in **manuscripts I and IV** bacterial changes together with metabolome shifts were evaluated in two different NAFLD cohorts.

Evidence of hepatic steatosis is needed in order to diagnose NAFLD. Histology (liver biopsy) and no-histology techniques (e.g., magnetic resonance spectroscopy, computed tomography, and ultrasonography) can be used to identify hepatic steatosis in the liver. Liver biopsy is considered the gold standard for NAFLD diagnosis; however, it can cause severe complications due to its invasive nature, and it is also prone to sampling error due to the unevenly distributed histological lesions (Herrema and Niess 2020). Some of the image-based diagnostic tools currently in use are magnetic resonance spectroscopy (MRS), computed tomography (CT), and ultrasonography. These techniques also have limitations such as high cost, radiation exposure, or limited accuracy. Among them, MRS has been considered the non-invasive standard method due to its highly accurate and reproducible diagnostic performance for evaluating NAFLD, and the no exposure to radiation (Lee 2017). However, the high cost and low availability make MRS remain primarily as a research tool not commonly used for clinical practice (Kechagias et al. 2022). Transient elastography (TE, FibroScan®) is an ultrasound-based technique to assess the liver stiffness (Piazzolla and Mangia 2020). TE with controlled attenuation parameter (CAP) simultaneously measures liver stiffness and fibrosis (Piazzolla and Mangia 2020). TE has been shown to have the best performance for the diagnosis and exclusion of advanced fibrosis when compared to liver fibrosis (Tovo et al. 2019).

New tools for NAFLD diagnosis combining non-invasive biomarkers are under investigation, as these seem to be the most promising cost-effective strategy. Numerous studies have explored non-invasive diagnostic approaches using clinically relevant biomarkers including non-invasive fibrosis models (e.g., fibrosis-4 index, NAFLD fibrosis score or stiffness, and AST/ALT ratio), clinical parameters (e.g., age, diabetes, and BMI), blood-based biomarkers (e.g., PRO-C3 and platelet count), and omics approaches (biomarkers that stem from genomics, transcriptomics, epigenomics, proteomics, lipidomics, and metabolomics). These emerging biomarkers can potentially be used in clinical practice and serve to develop novel diagnostic tools (Piazzolla and Mangia 2020; Hernandez Roman and Siddiqui 2020; Masoodi et al. 2021). **Manuscript IV** aimed to identify potential microbial biomarkers for early NAFLD detection and to develop a machine learning model that predicts the development of NAFLD 4 years before integrating metagenomics, metabolomics, and clinical data.

Regarding NAFLD medical strategies, there have been clinical trials investigating the effects of some drugs to treat NAFLD; however, to date there is no approved pharmacological treatment. Therefore, NAFLD pharmacological treatment used nowadays focuses on associated diseases such as diabetes, obesity, or lipid disorders to control patient glycemic status, liver injury, and lipid profiles (Jeznach-Steinhagen et al. 2019). Lifestyle interventions are currently the most effective strategy for managing NAFLD (Jeznach-Steinhagen et al. 2019). The importance of lifestyle has been recognized, and while potential pharmacological treatments are being tested in clinical trials, finding new strategies to stop or slow down the progression of NAFLD is crucial. In **manuscript I**, a potential dietary intervention treatment was investigated in a cohort of NAFLD subjects.

3. Gut microbiome modulation

Given the significant contribution of the gut microbiome to a wide range of diseases, the human gut microbiota has become an attractive target for novel therapeutics, and the mechanisms shaping the gut microbiome are being studied. Determining cause-effect relationships and designing microbiome-based therapies that can produce specific outcomes on the microbial community and host health, are some of the biggest challenges in microbiome research (Wong and Levy 2019).

Different mechanisms may modulate the gut microbiota including clinical treatments (antibiotics, fecal microbiome transplant, probiotics, and pharmabiotics) and lifestyle interventions (exercise and diet) (Quigley and Gajula 2020).

Infections caused by pathogenic bacterial species are treated with antibiotics. Unfortunately, the current generation of antibiotics is broad-spectrum, which has a devastating impact on the commensal microbiota (Avis et al. 2021). Some of the major effects caused by antibiotics on the gut microbiota are the reduction of the species diversity, the alteration of the metabolic activity, and the development of bacterial antibiotic resistance (Ramirez et al. 2020). A meta-analysis of randomized controlled trials in children performed by McDonnell et al. showed that antibiotic exposure was associated with reduced microbiome diversity and richness, and with changes in bacterial abundance (McDonnell et al. 2021). Palleja et al. showed that 4 days of antibiotic treatment induced large shifts in bacterial abundances in adults (Palleja et al. 2018). Rashidi et al. demonstrated that specific microbiota signatures at baseline determine personalized microbiota responses to antibiotic perturbations in humans (Rashidi et al. 2021). In relation to the mycobiome, little is known about the effect of antibiotics on the fungal community. Antibiotic treatment has been shown to eliminate bacterial species that promote resistance against fungal colonization during homeostasis, leading to yeast overgrowth and fungal dysbiosis (Li et al. 2018). The overgrowth of *Candida albicans* species has also been linked to antibiotic intake (Shankar et al. 2015; Fan et al. 2015; Gutierrez et al. 2020). Interestingly, changes produced by antibiotics were found to be recovered in the bacterial community mostly over three months, while alterations in the fungal community were long-lasting (Seelbinder et al. 2020).

Fecal microbiome transplant (FMT) is a therapeutic strategy that involves administering specially prepared stool material from a donor into the intestinal tract of a recipient, aiming to alter the gut microbiota composition and improve the individual’s health (Gupta et al. 2016). FMT has been used successfully as a treatment option in recurrent *Clostridium difficile* infection (Rohlke and Stollman 2012). Recently, the U.S. Food and Drug Administration (FDA) approved the first orally administered fecal microbiota product for the prevention of recurrence of *C. difficile* infection in individuals 18 years old and older (Carvalho 2023). The use of FMT in other microbiota-associated conditions that also experience gut microbiota dysbiosis such as NAFLD, diabetes, obesity, or IBD seems to be a promising therapy but needs to be further investigated (Smits et al. 2013; Gupta et al. 2016; Napolitano and Covasa 2020).

Besides invasive solutions, exercise and microbiota-directed food interventions (MDFs) are the major gut microbial-modulation strategies that are been investigated for the treatment of a variety of human diseases associated with gut microbiome dysbiosis (Conlon and Bird 2015). In relation to exercise activity, Ni et al. showed that exercise in subjects with prediabetes was associated with differential gut microbiota changes, and these alterations were found to be linked with improvements in glucose homeostasis and insulin sensitivity. In addition, subjects that responded to the exercise intervention were found to have an enhanced capacity for generating SCFAs and an increased breakdown of BCAAs, whereas an increased production of metabolically detrimental compounds was associated with the microbiome of non-responders subjects (Liu et al. 2020). A different study in obese children showed that exercise training modulates positively the gut microbiota profile and produces changes reducing inflammatory signaling pathways induced by obesity (Quiroga et al. 2020). A recent investigation in subjects with NAFLD and prediabetes who underwent aerobic exercise combined with dietary intervention found that hepatic liver fat decreased in the intervention groups while increased in the control group; even more, the authors identified changes in the microbial alpha diversity and changes in the gut microbiota co-occurrence network (Cheng et al. 2022).

Regarding diet, microbiota-directed foods are known as aliments that aim to alter the structure or function of the gut microbiome and promote the growth of beneficial microbes associated with good health (Barratt et al. 2017). Numerous strategies making use of MDFs are being explored, for example, high fiber diet or low carbohydrate diet among others. Mardinoglu et al. found that the use of an isocaloric low-carbohydrate diet with increased protein promotes multiple metabolic benefits in obese humans with NAFLD (Mardinoglu et al. 2018). High-fiber diet intervention promoted the growth of SCFA-producing organisms in diabetic humans and induced changes in the microbiome community correlated with elevated levels of glucagon-like peptide-1, reduction in acetylated hemoglobin levels, and improved blood-glucose regulation (Zhao et al. 2018). A combination of low-fat/high-fiber diet was shown to improve the overall quality of life and decrease the inflammatory markers and dysbiosis in patients with ulcerative colitis (Fritsch et al. 2021).

3.1 Resistant starch as microbiota-directed food intervention

Non-digestible carbohydrates (NDCs) are fibers that are fermented by the gut microbes into SCFAs (Dobranowski and Stintzi 2021). Resistant starch (RS) is a type of NDCs that escapes digestion and survives passage through the stomach and small intestine to reach the colon where it is fermented by microorganisms (DeMartino and Cockburn 2020). Depending on their source and processing procedure, resistant starches are classified into five types (RS1–RS5). RS1 is physically unreachable starch, such as whole grains; RS2 is native granular starches, such as raw potatoes, green bananas, or high-amylose maize; RS3 is retrograded amylose starch or crystallized starch, such as cooked and cooled starchy foods; RS4 is chemically modified starch; and RS5 is amylose-lipid complex (Zhu et al. 2022).

The use of resistant starch as microbiota-directed foods (MDFs) has become one of the focuses of non-digestible carbohydrate therapies for the prevention and treatment of obesity and related diseases (Zhang et al. 2015). The regulatory effects of RS supplementation on NAFLD mainly occur in the gut, where RS contributes to the restoration of the gut microbiota structure, the increase in SCFAs release, and the enhancement of the gut barrier integrity (Zhu et al. 2022). A recent meta-study showed that especially for diabetic overweight or obese subjects, RS supplementation can improve fasting glucose, fasting insulin, insulin resistance, and insulin sensitivity (Wang et al. 2019). RS supplementation showed improvements in clinical remission in patients with IBD (Montroy et al. 2020). Animal studies in mice also demonstrated RS to decrease adiposity, reduce insulin levels, and exert metabolic benefits (Higgins et al. 2011; Polakof et al. 2013; Rosado et al. 2020; Zhang et al. 2020; Montroy et al. 2020). Nevertheless, up to now, no clinical study has explored RS as a potential therapeutic treatment for NAFLD. In **manuscript I** the effects of RS2 from high-amylose maize in NAFLD subjects were investigated.

4. Resistance and resilience of the human gut microbiome

Gut microbiome responses to microbiome-targeted interventions are highly individual-specific (Olsson et al. 2022). This highlights the importance of investigating gut microbiome dynamics to better understand the mechanisms causing the microbiome to remain unaltered after the same perturbation and to determine whether is a microbial signature and specific taxa that are important contributors that contribute to the overall stability of the community (Risely 2020). Microbial stability is known as the property of maintaining a state of equilibrium and resist to perturbations to the community (Fassarella et al. 2021).

The use of metabolic engineering and synthetic biology will allow us to develop personalized therapies that target the gut microbiota. Identifying when a microbiome is not going to be affected by a microbiome-targeted therapy is crucial to develop personalized therapies to first alter the gut microbiome stability, so that it is sufficiently plastic to conduct a well-defined microbiome modulation treatment afterward. In addition, characterizing gut microbiome signatures associated with microbiome dynamics will also help to develop more precise patient stratification strategies combining host phenotype and microbiome stratification.

The gut microbiome is constantly fluctuating and trying to maintain a dynamic equilibrium over time while being exposed to external perturbations such as diet, medications, and the environment; that can disrupt the stability of the gut microbial ecosystem (Fassarella et al. 2021). However, to date, it is not clear how an individual microbial community responds to perturbations. External perturbations can lead to a transient dysbiotic state that will recover over time, but it can also lead to a stable dysbiotic state with negative implications for the host (Fassarella et al. 2021). To understand the response to perturbations in the gut as a complex ecosystem, it is important to distinguish two concepts: resilience, which is the property of how fast or to what extent an ecosystem will recover its initial state after a perturbation; and resistance, that is the ability to remain unchanged during a perturbation (Sommer et al. 2017). An appropriate equilibrium state of resilience and resistance of the healthy gut microbiota protects us from dysbiotic-associated diseases (Sommer et al. 2017).

Even though the abundance of specific bacteria fluctuates over time, it has been shown that the gut microbial community in healthy subjects can be stable for years (Faith et al. 2013). In addition, gut microbial communities with higher diversity at the baseline showed more microbial stability over time (Chen et al. 2021). Chen et al. also observed that the genetic stability of gut microbes varies substantially across different species, and some species including *Ruminococcus torques*, *Streptococcus parasanguinis*, and *Faecalibacterium prausnitzii* were identified to be genetically unstable over time. A recent study in a Swedish healthy cohort showed that the gut microbiota functional potential is more stable than the species profile, and they identified that intra-individual compositional variability was negatively associated with the abundance of *Faecalibacterium prausnitzii* and two *Bifidobacterium* species (Olsson et al. 2022). Concerning diet, the stability of the microbiome composition was found to be correlated to dietary diversity, and food-microbiome interactions were identified to be highly personalized (Johnson et al. 2019). Hutchison et al. investigated the effect of a fermentable fiber diet in mice with different microbial communities, and their findings suggest that the effectiveness of a fermentable fiber diet in protecting against atherosclerosis is different for each animal and influenced by the composition of the gut microbiome (Hutchison et al. 2023). However, even though much research has been done on the effects of lifestyle interventions on the gut microbiome and human health, very little is known about how these lifestyle perturbations impact the stability, resistance, and resilience of the gut microbial community.

In **manuscript II** we explored the microbiome response to antibiotics and lifestyle treatments in order to characterize signatures associated with the gut microbial dynamics. We also developed a machine learning model that predicts the microbiome responsiveness in response to lifestyle interventions.

5. Bioinformatics approaches for studying the microbiome

The fast development of next-generation sequencing (NGS) technologies in the last decades has facilitated the rapid expansion of the microbiome field. Omics analyses including metatranscriptomics, metagenomics, proteomics, and metabolomics are some of the bioinformatic fields that have helped to understand and elucidate the role of the human microbiome in health and disease.

Metagenomic approaches allow the identification of microorganisms present in a sample. Thanks to the advances in NGS technologies, numerous metagenomic approaches are nowadays available to identify the composition of microbial populations. Amplicon sequencing and whole-genome shotgun (WGS) metagenomics are the two major methodologies for researching the microbiome utilizing high-throughput sequencing

(Figure 2) (Agus et al. 2021). In addition, the development of analyzing tools and resources, and the creation and curation of metagenome databases have widely increased in the last decade. Nowadays, a combination of different omics disciplines such as metagenomics and metabolomics has been suggested as a promising approach to study host–microbiome interactions (Turnbaugh and Gordon 2008), with a high potential to investigate metabolic-related disorders.

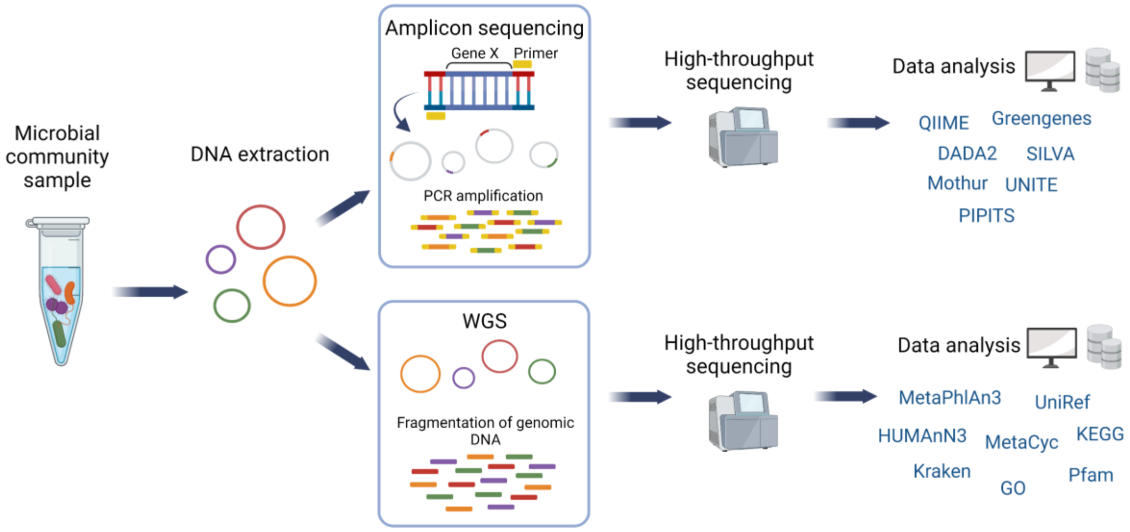


Figure 2 | Amplicon sequencing and whole-genome-shotgun sequencing (WGS) overview showing some of the most used tools and databases for each sequencing strategy. Figure created with Biorender.

5.1 Amplicon sequencing

To identify the microbial composition of a sample using sequencing strategies, it is not needed to sequence the full metagenome. A common practice, called amplicon sequencing, is to target a fragment of the genome. Amplicon sequencing uses generic primers designed to amplify a particular gene or gene fragment from all genomes present in a sample, and then the resulting product is sequenced. The rRNA regions commonly amplified for amplicon sequencing profiling are the 16S rRNA gene for bacteria/archaea, the 18S rRNA gene for eukaryotes, and the internal transcribed spacer regions 1 or 2 (ITS1/ITS2) of the fungal ribosome for fungi. This strategy generally involves four steps: DNA extraction, amplification of the marker gene through polymerase chain reaction (PCR) using target primers, barcoding of the amplicons of each sample with a short "barcode" sequence unique to each sample, and high throughput "multiplexed" sequencing of the combined amplicons from all samples in a single sequencing run (Dong et al. 2017).

Some of the generally used open-source amplicon analysis tools are QIIME2, DADA2, Mothur, and PIPITS. QIIME2 is a frequently used bioinformatics pipeline for performing microbiome analysis from raw sequencing data including demultiplexing and quality filtering, OTU picking, taxonomic assignment, phylogenetic reconstruction, diversity analysis, and visualizations (Bolyen et al. 2019). DADA2 is an R package that implements the full amplicon workflow such as filtering, dereplication, chimera identification, and merging of paired-end reads (Callahan et al. 2016). Mothur is a software package for bioinformatics data processing that includes pre-processing, OTU picking, join reads, contig screening, taxonomic assignment, sequence filtering, chimera removal, etc (Schloss et al. 2009). Lastly, PIPITS is a stand-alone suite of software specifically for ITS data for automated processing of Illumina MiSeq sequences for fungal community analysis (Gweon et al. 2015). In order to perform the taxonomic assignment, it is essential the availability of reference databases. Some of the most used public reference databases for amplicon taxonomy assignment are SILVA for 16S/18S rRNA data (Quast et al. 2013), Greengenes for 16S rRNA data (DeSantis et al. 2006), and UNITE for ITS data (Nilsson et al. 2019).

In **manuscript III,** I used the PIPITS pipeline and UNITE database to process Illumina ITS1 sequencing data, and 16S sequencing data from **manuscript III** were processed using DADA2, QIIME2, and SILVA databases.

5.2 Whole-genome shotgun sequencing

In contrast to amplicon sequencing, shotgun metagenomics allows the sequencing of all accessible genomic DNA present in a given sample. WGS commonly consists of five stages: DNA extraction, random fragmentation of genomic DNA, genomic library preparation, paired-end sequencing, and genome assembly (Fuentes-Pardo and Ruzzante 2017). This strategy is a good taxonomic classification option due to the alignment of the whole genome and allows functional annotation providing an integrated understanding of the community structure, genetic population heterogeneity, and potential metabolism pathways (Niu et al. 2018).

Two popular tools for metagenomic taxonomic profiling are MetaPhlAn (Beghini et al. 2021) and Kraken (Wood and Salzberg 2014). Both tools are open source and include their curated database. MetaPhlAn uses clade-specific marker genes to study the microbiome taxonomic composition, while Kraken uses a K-mer based searching algorithm to assign taxonomic labels to the reads.

Functional profiling to identify the presence/expression of bacterial genes in the microbial community is also possible when analyzing shotgun sequencing data. HUMAnN (Beghini et al. 2021) is one of the most used tools for analyzing bacterial gene expression profiles. It allows to access different levels of information, being the reads first assigned to bacterial taxa and then searched against the protein databases for gene assignments. UnifRef database (Suzek et al. 2007) is used to identify gene families that then are mapped to different systems using the specific databases including MetaCyc reactions (Caspi et al. 2014), KEGG Orthologs (KOs) (Kanehisa et al. 2016), Pfam domains (Mistry et al. 2021), and Gene Ontology (GO) (The Gene Ontology Consortium 2021).

Whole-genome sequencing data were analyzed in **manuscripts I, II, and IV**. This allowed the taxonomical characterization of the microbial community as well as the functional profile. MetaPhlAn (versions 2 and 3) and HUMAnN (versions 2 and 3) tools were used to perform the taxonomic and functional profiling respectively.

5.3 Metabolomics

Metabolites are the intermediates or end products of multiple enzymatic reactions and therefore are the most informative proxies of the biochemical activity of an organism (Alonso et al. 2015). Metabolomics, known as the study of the metabolite composition within cells, biofluids, tissues, or organisms; has become an emerging technology in the last decades (Newgard 2017). There are two different metabolomic approaches: targeted and untargeted metabolomics, also known as validation-based or discovery-based, respectively (Schrimpe-Rutledge et al. 2016). Untargeted metabolomics focuses on global detection and relative quantitation of small molecules in a sample. In contrast, targeted metabolomics aims to measure quantitatively specific groups of metabolites; for this reason, prior knowledge of metabolites of interest and known compounds is needed (Schrimpe-Rutledge et al. 2016). Therefore, to perform targeted metabolomics analysis it is required to have a previously developed analytical method to measure the concentration of the specific metabolite in the sample (Shulaev 2006). Numerous metabolites cannot be identified with the currently available analytical techniques and purification standards, so targeted metabolomics cannot be used for novel metabolic markers identification (Shulaev 2006). Thus, targeted metabolomics is more quantitative, whereas untargeted often provides more information (Zhang et al. 2016).

Metabolomics has a wide range of applications being involved in numerous research areas including plant biotechnology, food technology, human diseases, and toxicology, among others (Gomez-Casati et al. 2013). One of the most growing areas is the biomedical field and the research of the metabolome in the development of human diseases, especially in metabolic-related disorders (Alonso et al. 2015). Metabolomics is facilitating the discovery of metabolite-disease biomarkers and in practice, it will enhance diagnosis, prognosis, surveillance, and personalized drug treatments (Gonzalez-Covarrubias et al. 2022). Albeit knowledge about the metabolites is crucial, integration of metabolomics with different omics disciplines may be a better approach to understanding many as yet undetermined disease mechanisms (Cambiaghi et al. 2017).

In this dissertation, targeted metabolomics data were analyzed in **manuscripts I and IV** from the two different NAFLD cohorts. Metabolomics changes and signatures were investigated, and metabolome-microbiome integration analyses were performed for better insights.

6. Statistical and machine learning techniques in metagenomics

Statistical analyses and machine learning workflows in this dissertation were performed using R programming language (R Core Team 2022) and R studio software (RStudio Team 2022). R is a very popular open-source programming language and environment, commonly used in statistical computing, data analytics, and scientific research, supported by the R Core Team and the R Foundation for Statistical Computing. RStudio is an open-source integrated development environment for the R programming language. It includes a console, syntaxhighlighting editor that supports direct code execution, as well as tools for plotting, history, debugging, and workspace management.

In order to perform comprehensive analyses and to have a full understanding of the different microbial communities analyzed in this dissertation, numerous bioinformatics data analysis techniques were applied to investigate and elucidate the different project objectives. Some of the main data analysis approaches that I applied during my Ph.D. are described below.

6.1 Methods for abundance comparisons

Identifying differentially abundant features such as bacteria, functions, or fungi is a common goal of metagenomics studies. New methods have been developed to identify changes in microbial abundances between different comparisons or conditions. Some of the commonly used tests or tools for differential abundance analysis are Wilcoxon test, generalized linear models (GLM), and metagenomeSeq.

Wilcoxon test is a non-parametric alternative to a t-test, appropriate when working with microbial data as commonly are non-normalized. Wilcoxon rank-sum test is used to compare two independent samples, while Wilcoxon signed-rank test is used to compare two related samples, matched samples, or to conduct a paired difference test of repeated measurements on a single sample (Schwaid 2017). Wilcoxon test can be performed in R using the function *wilcoxon.test* from the R package stats. Generalized linear model (GLM) is an advanced statistical modeling technique and is a basic method for advanced testing of differential abundance in sequencing data. GLM can model a mean response under non-linear, non-symmetric, and non-gaussian association conditions, where discrete and continuous data distributions can be fitted (Lu et al. 2019). The function *glm* from R package stats is used to fit generalized linear models. metagenomeSeq is a tool that was developed specifically for microbial datasets (Paulson et al. 2013) and was designed to determine microbial features that are differentially abundant between two or more groups of multiple samples. metagenomeSeq addresses the effects of both normalization and under-sampling of microbial communities on disease association detection and the testing of feature correlations. It is available in metagenomeSeq R package.

In addition, some of these methods allow accounting for covariates. It is known that the gut microbiome undergoes significant changes through age (Bosco and Noti 2021), and adjusting for age has been shown to improve the identification of gut microbial alterations (Ghosh et al. 2020). Other confounders affecting the microbiome composition that are commonly used when studying gut microbial changes are gender, ethnicity, or smoking among others (Chua et al., 2017; Jian et al., 2022; Loftfield et al., 2020; Si et al., 2021). The importance of considering covariates in microbiome studies will be addressed in the discussion section. From the previously mentioned methods, GLM and metagenomeSeq can perform differentially abundant analysis adjusting for covariates.

6.2 Methods for correlation analysis

Correlation is a statistical method used to study a possible linear association, connection, or relationship between two continuous variables. The statistic is called correlation coefficient, and it measures the strength and the direction of the relationship (Mukaka 2012). It ranges from –1 to 1, with +1 indicating a perfect direct association, –1, a perfect inverse association, and 0, no relationship between the two variables, respectively (Mukaka 2012).

Three frequently used correlation methods in biostatistics are Pearson correlation, Kendall rank correlation, and Spearman correlation. These methods can detect linear or non-linear monotonic (strictly increasing or strictly decreasing function) relationships (Santos et al. 2014). We use Pearson correlation when both variables are normally distributed. For example, when we want to know if two clinical variables that are normally distributed have a linear association, we can perform a Pearson correlation between both variables. Pearson's correlation coefficient *r* is calculated as the covariance of the two variables divided by the product of their standard deviations. Unlike Pearson's correlation coefficient, Spearman's correlation rho (ρ) and Kendall's tau (τ) do not require the assumption of normality of the variables. Thus, these two correlation methods are more used when working with metagenomics data, as microbial data are rarely normally distributed. Spearman's correlation is a non-parametric test that does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. It measures the degree of association between two variables. This method is simply the application of Pearson's correlation in the data converted to ranks before calculating the coefficient. Kendall's tau (τ) is another non-parametric test that measures the strength of dependence between two variables. In R, the function cor.test from the R package stats (R Core Team 2022) computes the Pearson's, Spearman's, or Kendall's correlation of two variables provided to the function.

In addition, it is also possible to compute correlation analysis between variables eliminating the effect of other covariates using the partial correlation approach. Partial correlation is defined as the association between two random variables after eliminating the effect of one or more other random variables. In R, the function pcor from the R package ppcor (Kim 2015) calculates the partial correlations of all pairs of two random variables of a matrix or a data frame for the different correlation methods previously described (Pearson's, Spearman's, and Kendall's).

6.3 Methods for network analysis

Network-based approaches are commonly used to explore -omics data. In metagenomics, microbiome network analyses allow us to understand community dynamics and explore the interactions and dependencies between the different members of the gut microbial community (Matchado et al. 2021). The complex interactions between thousands of individual taxa or functions and between different communities (e.g., bacteria, fungi, and metabolites), suggest network analysis as a powerful method in the microbiome field

(Matchado et al. 2021). Taxa and functions are common components modeled when performing gut microbial network analysis. These components in the network are known as nodes. When incorporating other types of data, nodes can also be host features, metabolites, genes, or proteins. The presence of an edge means that two nodes are connected, indicating an association between the two nodes. Correlation-based approaches, including Pearson or Spearman correlation previously described, are popular methods for studying these interactions (Jiang et al. 2019). However, in order to account for the compositionality of the microbial data, more specific tools such as SparCC and SPIEC-EASI have been developed to explore co-occurrence microbial networks. SparCC (Sparse Correlations for Compositional data) is one widely used method to build microbial community networks (Friedman and Alm 2012). This method was developed for estimating correlation values from compositional data. SparCC estimates the linear Pearson correlations between abundances in microbiome data accounting for their inherent sparsity and compositionality. It uses the centered log-ratio transformation to address data compositionality (Friedman and Alm 2012). SparCC was developed as a Python module, but there is also a reimplementation of the SparCC algorithm available using the R function sparcc. Another method developed to explore microbial networks is SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) (Kurtz et al. 2015). This method infers an ecological network (inverse covariance matrix) from compositional data using the log-ratio transformation and performs neighborhood selection and sparse inverse covariance selection (Kurtz et al. 2015). SPIEC-EASI pipeline was developed in R and can be run using the R function spiec.easi from the SpiecEasi package.

Once the microbial community network is built, we can explore the different characteristics of the network topology to investigate the connectivity and structure of the microbial ecosystem. Network metrics, such as degree centrality, betweenness centrality, closeness centrality, and hub score can be used to quantitatively describe these communities and identify the most important nodes (i.e., taxa) of a given community (Zamkovaya et al. 2021). Degree centrality measures the number of connections of a node (i.e., taxa or metabolite), and determines the level of co-occurance of a node. Betweenness centrality is a measure based on the shortest paths and computes the extent to which a node lies on paths between others (Zamkovaya et al. 2021). Closeness centrality measures how far a node is from all other nodes and can be used to find the most central taxa of a given community network (Zamkovaya et al. 2021). Nodes with high degree and betweenness centrality are typically the most connected taxa within the community and are also known as "hubs" (Zamkovaya et al. 2021). These network characteristics provide essential insights into how specific features may contribute to ecosystem functioning.

Lastly, network clustering analysis allows the identification of densely connected nodes that form network subcommunities (or clusters) and reveals relationships among nodes in the community network. Clusters are powerful topological features to reflect network differences (Pan et al. 2021). Some clustering algorithms available in the R package igraph are Louvain, Walktrap, and Greedy clustering.

6.4 Machine learning

Machine learning (ML) is a branch of artificial intelligence that develops, analyzes, and implements predictive methods through the use of dynamic algorithms capable of data-driven decisions (El Bouchefry and de Souza 2020).

There are two main types of ML algorithms commonly applied to microbial datasets: supervised and unsupervised learning. In supervised learning algorithms, the output has been given a priori labels or the learner has some prior knowledge of the data, while in unsupervised learning algorithms, hidden patterns are identified in unlabeled data (Auslander et al. 2021). Some commonly used unsupervised algorithms include k-means clustering, hierarchical clustering, and principal component analysis. In the case of supervised learning, some examples of algorithms are random forest, support vector machines, and gradient boosting machines.

The development of a machine learning pipeline can be summarized in four main steps: data handling, model training, evaluation, and development (De Souza Nascimento et al. 2019). The data handling step includes different approaches such as data preprocessing and feature selection. Due to the importance of data quality in the model performance, it is crucial to implement an appropriate data handling approach when building a model. Feature selection is a key step to obtain an optimal and non-redundant subset of the initial features due to the extremely large number of features when working with microbiome data (e.g., species, genes, metabolites, etc.). Moreover, to evaluate and test the performance of a machine learning model, a resampling method called cross-validation is usually applied. Cross-validation uses different portions of the data to test and train a model on different iterations. Cross-validation together with feature selection are useful techniques that help to prevent one of the main problems when building a model, namely overfitting (when the model cannot generalize and fits too closely to the training dataset instead). To evaluate the performance of a machine learning model, different measurements are frequently computed including confusion matrix, accuracy, precision, specificity, sensitivity, receiver operating characteristic (ROC), and area under ROC curve (AUC).

The application of ML in the microbiome field is relatively new but it has shown a lot of potential for sophisticated analyses and generating new knowledge from the vast amount of omics data produced. Clinical data and gut microbial profiles from multi-omics analyses are used to develop ML models with different applications including phenotypic prediction, patient stratification, biomarker discovery, treatment outcome evaluation, and personalized treatment and nutrition (Li et al. 2022a). For instance, Franzosa et al. developed a machine-learning model using gut microbiome and metabolic profiles to classify subjects according to IBD phenotype (Franzosa et al. 2019). Zeevi et al. showed the power of the microbiome in personalized medicine and found that dietary, clinical, and anthropometric information together with microbial profiles can successfully predict postprandial glucose responses using a gradient boosting model (Zeevi et al. 2015). A recent study showed the potential of the combination of conventional risk factors and gut microbiome data for early risk stratification for liver disease (Liu et al. 2022). In this study, Liu et al. developed a gradient-boosting model able to predict liver disease 15 years before.

In this dissertation, caret R package has been used to implement machine learning pipelines. Caret is a powerful machine learning package that provides methods for common ML steps, such as preprocessing, training, tuning, and evaluating predictive models.

OBJECTIVES

This dissertation aimed to implement bioinformatic analyses making use of multi-omics techniques to expand the knowledge about the human gut microbiome and mycobiome and its implication for human health and disease. Complete study designs were set up for the different projects included in this dissertation to establish and perform meticulous analyses making use of state-of-the-art bioinformatics and statistical approaches to achieve the research objectives. During my research, I focused on investigating the role and connection of the gut microbiome and mycobiome in non-alcoholic fatty liver disease (NAFLD), and the effect of lifestyle interventions on the gut microbiome composition and dynamics.

Three research projects of this dissertation aimed to study the gut microbiome, mycobiome, and metabolome in NAFLD, and to evaluate different novel microbiome-based strategies for NAFLD. Using shotgun metagenomics (microbiome), ITS sequencing (mycobiome), and metabolomics; the characterization of the microbiome, mycobiome, and metabolome signatures in NAFLD progression was investigated. These projects served to ask the questions mentioned below:

1. Is resistant starch (RS) a beneficial microbiome-directed food intervention to treat NAFLD?
2. How the metabolome and gut microbiome of patients with NAFLD are altered after 4-month RS intervention?
3. Are there potential RS-targeted species or RS-targeted microbial metabolites?
4. In healthy people, is there a microbiome-metabolome signature able to predict the development of NAFLD and microbial biomarkers that serve as early detectors of NAFLD development?
5. How do intestinal fungi contribute to NAFLD progression?

A fourth study focused on investigating the resistance potential of an individual microbial ecosystem to lifestyle interventions. We performed a large-scale meta-analysis of metagenomic samples to better understand gut microbial dynamics and to identify potential microbial biomarkers associated with the microbiome resistance to lifestyle interventions. This project aimed to shed light on the following research questions:

1. How does the stability of the microbial community or the stability of the specific species respond to different antibiotic or lifestyle interventions?
2. Are there microbial biomarkers of community response that characterize the stability of the microbial composition of an individual?
3. Is it possible to predict the resistance of a microbiome community to be changed in response to a lifestyle intervention using the baseline gut microbial composition?

RESEARCH PUBLICATIONS

This cumulative dissertation consists of four research publications. Three first-author publications of which one is published in the journal Microbiome (**manuscript II**), another one is accepted for publication and will be the cover article in the issue of September of Cell Metabolism journal, and a third publication in preparation (**manuscript III**). Lastly, one co-author publication is published in the journal Science Translational Medicine (**manuscript IV**). Supplemental information of all published manuscripts can be downloaded from the websites of the respective publishers. Additionally, copies of these files are also included in the digital version of this dissertation. This dissertation comprises the following manuscripts:

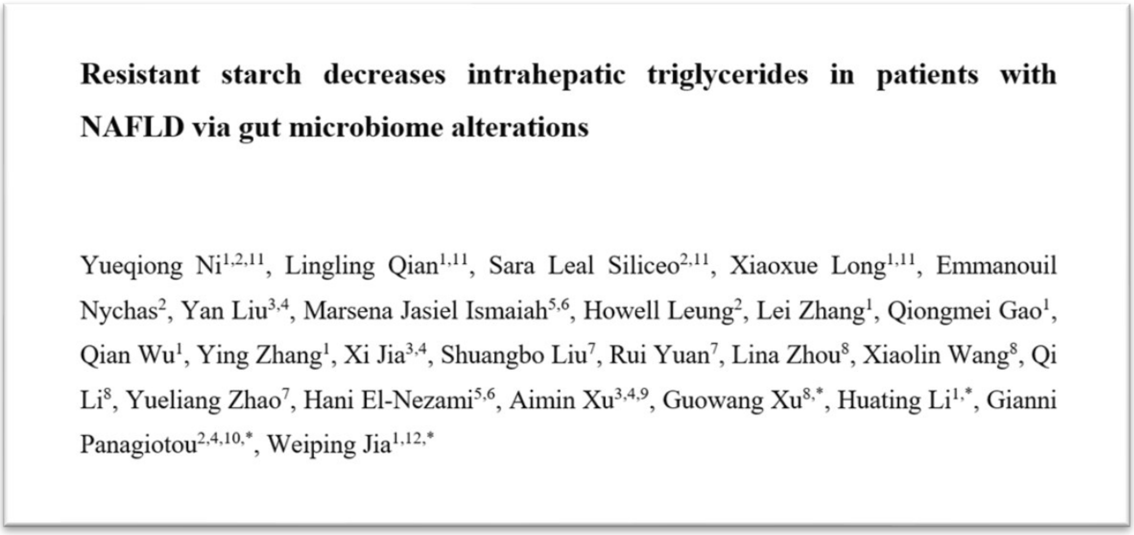
- **Manuscript I:**
  [co-first author] Ni Y, Qian L, Siliceo SL, Long X, Nychas E, Liu Y, Ismaiah MJ, Leung H, Zhang L, Gao Q, Wu Q, Zhang Y, Jia X, Liu S, Yuan R, Zhou L, Wang X, Li Q, Zhao Y, El-Nezami H, Xu A, Xu G, Li H, Panagiotou G, and Jia W., (in press). **Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations.** *Cell Metabolism*. [Accepted for publication in Cell Metabolism on June 13th].

- **Manuscript II:**
  [co-first author] Chen J, Siliceo SL, Ni Y, Nielsen HB, Xu A, and Panagiotou G., (2023). **Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions**. *Microbiome*, **11**, 178. https://doi.org/10.1186/s40168-023-01604-z

- **Manuscript III:**
  [co-first author] Thielemann N, Siliceo SL, Rau M, Herz M, Nieuwenhuizen N, Aldejohann AM, Hermanns HM, Mirhakkak MH, Löffler J, Dandekar T, Martin R, Panagiotou G, Geier A, and Kurzai O. **Genetic variation in IL-17A regulation and mycobiome dysbiosis contribute to non-alcoholic fatty liver disease**. [Manuscript in preparation].

- **Manuscript IV:**
  [co-author] Leung H, Long X, Ni Y, Qian L, Nychas E, Siliceo SL, Pohl D, Hanhineva K, Liu Y, Xu A, Nielsen HB, Belda E, Clément K, Loomba R, Li H, Jia W, and Panagiotou G. (2022) **Risk assessment with gut microbiome and metabolite markers in NAFLD development**. *Science translational medicine*, 14(648):eabk0855. https://doi.org/10.1126/scitranslmed.abk0855

Manuscript I

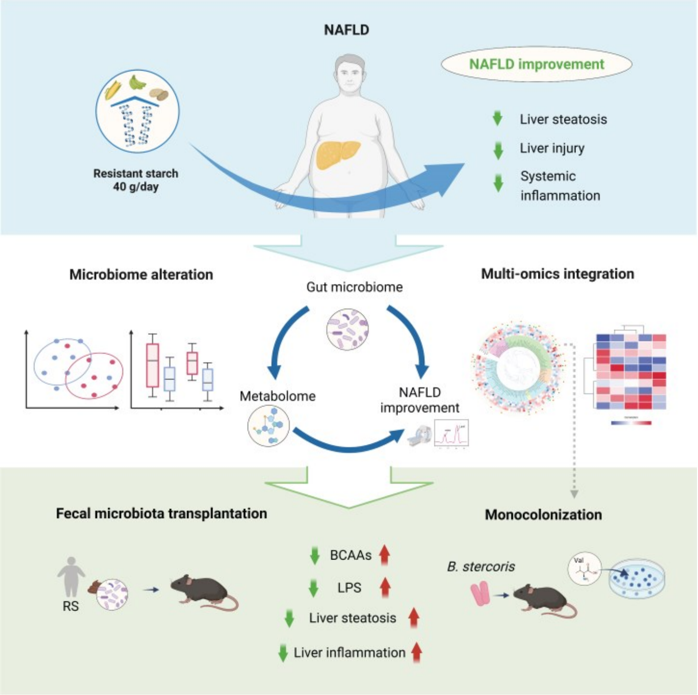


Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations

Yueqiong Ni[1,2,11], Lingling Qian[1,11], Sara Leal Siliceo[2,11], Xiaoxue Long[1,11], Emmanouil Nychas[2], Yan Liu[3,4], Marsena Jasiel Ismaiah[5,6], Howell Leung[2], Lei Zhang[1], Qiongmei Gao[1], Qian Wu[1], Ying Zhang[1], Xi Jia[3,4], Shuangbo Liu[7], Rui Yuan[7], Lina Zhou[8], Xiaolin Wang[8], Qi Li[8], Yueliang Zhao[7], Hani El-Nezami[5,6], Aimin Xu[3,4,9], Guowang Xu[8,*], Huating Li[1,*], Gianni Panagiotou[2,4,10,*], Weiping Jia[1,12,*]

Overview

In manuscript I, we aimed to elucidate the effect of a resistant starch (RS) supplementation as one type of microbiome-directed foods (MDFs) to treat NAFLD and characterize the changes in the gut microbiome and metabolome during the intervention. Therefore, we conducted a randomized, double-blinded, placebo-controlled clinical trial of 4-month RS supplementation in individuals with NAFLD. Multi-omics profiling was used to provide an integrated understanding of how RS and associated alterations in the gut microbiota or metabolites contributed to NAFLD improvement. Our results demonstrated the efficacy of RS as a novel microbiota-targeted intervention for NAFLD. Moreover, by performing multi-omics analyses tackling the complexity and heterogeneity of NAFLD pathogenesis, we identified possible mediators of the beneficial effect of RS. In addition, whole microbiota changes, the potential RS-targeted single species, and microbial metabolites were validated in mice and cell lines for causal insights.



Graphical abstract

FORM I

Manuscript No: 1

Manuscript title: Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations

Authors: Yueqiong Ni*, Lingling Qian*, **Sara Leal Siliceo***, Xiaoxue Long*, Emmanouil Nychas, Yan Liu, Marsena Jasiel Ismaiah, Howell Leung, Lei Zhang, Qiongmei Gao, Qian Wu, Ying Zhang, Xi Jia, Shuangbo Liu, Rui Yuan, Lina Zhou, Xiaolin Wang, Qi Li, Yueliang Zhao, Hani El-Nezami, Aimin Xu, Guowang Xu, Huating Li, Gianni Panagiotou, Weiping Jia

Bibliographic information (if published or accepted for publication: Citation): Ni et al., (in press). Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations. *Cell Metabolism*.

The candidate is (Please tick the appropriate box)**:**

☐ First author, ☒ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status (if not published; "submitted for publication", "in preparation".): published (accepted for publication in Cell Metabolism on 13th June 2023)

Authors' contributions (in %) to the given categories of the publication

Author	**Conceptual**	**Data analysis**	**Experimental**	**Writing the manuscript**	**Provision of material**
Ni, Y.*	20%	10%		35%	
Qian, L.*	5%	10%	30%	10%	
Leal Siliceo, S.*	15%	45%		5%	
Long, X.*	5%	10%	30%	15%	
Nychas, E.		10%			
Leung, H.		10%			
Panagiotou, G.	20%			5%	30%
Others	35%	5%	40%	30%	70%
Total:	100%	100%	100%	100%	100%

*Authors contributed equally

Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations

Yueqiong Ni[1,2,11], Lingling Qian[1,11], Sara Leal Siliceo[2,11], Xiaoxue Long[1,11], Emmanouil Nychas[2], Yan Liu[3,4], Marsena Jasiel Ismaiah[5,6], Howell Leung[2], Lei Zhang[1], Qiongmei Gao[1], Qian Wu[1], Ying Zhang[1], Xi Jia[3,4], Shuangbo Liu[7], Rui Yuan[7], Lina Zhou[8], Xiaolin Wang[8], Qi Li[8], Yueliang Zhao[7], Hani El-Nezami[5,6], Aimin Xu[3,4,9], Guowang Xu[8,*], Huating Li[1,*], Gianni Panagiotou[2,4,10,*], Weiping Jia[1,12,*]

Affiliations

[1] Shanghai Key Laboratory of Diabetes Mellitus, Department of Endocrinology and Metabolism, Shanghai Diabetes Institute, Shanghai Clinical Center for Diabetes, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200233, China

[2] Microbiome Dynamics, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstraße 11A, 07745 Jena, Germany

[3] State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong, China

[4] Department of Medicine, The University of Hong Kong, Hong Kong, China

[5] Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio 70211, Finland

[6] School of Biological Sciences, Faculty of Science, The University of Hong Kong, Hong Kong S.A.R., China

[7] College of Food Science and Technology, Shanghai Ocean University, Shanghai, 201306, China

[8] CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, 116023, China

[9] Department of Pharmacology and Pharmacy, the University of Hong Kong, Hong Kong, China

[10] Friedrich Schiller University, Faculty of Biological Sciences, Jena, Germany

[11] These authors contributed equally to this work.

[12] Lead contact

*Correspondence: xugw@dicp.ac.cn (G.X.); huarting99@sjtu.edu.cn (H.L.); gianni.panagiotou@leibniz-hki.de (G.P.); wpjia@sjtu.edu.cn (W.J.)

SUMMARY

Non-alcoholic fatty liver disease (NAFLD) is a hepatic manifestation of metabolic dysfunctions for which effective interventions are lacking. To investigate the effects of resistant starch (RS) as a microbiota-directed dietary supplement for NAFLD treatment, we coupled a 4-month randomized placebo-controlled clinical trial in individuals with NAFLD (ChiCTR-IOR-15007519), with metagenomics and metabolomics analysis. Relative to the control (n=97), the RS intervention (n=99) resulted in a 9.08% absolute reduction of intrahepatic triglyceride content (IHTC), which was 5.89% after adjusting for weight loss. Serum branched chain amino acids (BCAAs) and gut microbial species, in particular *Bacteroides stercoris*, significantly correlated with IHTC and liver enzymes, and were reduced by RS. Multi-omics integrative analyses revealed the interplay among gut microbiota changes, BCAA availability, and hepatic steatosis, with causality supported by fecal microbiota transplantation and monocolonization in mice. Thus, RS dietary supplementation might be a strategy for managing NAFLD by altering gut microbiota composition and functionality.

Keywords: non-alcoholic fatty liver disease; gut microbiota; resistant starch; microbiota-directed foods; microbiota transplantation; BCAAs

INTRODUCTION

An estimated 30% of the world's population currently has nonalcoholic fatty liver disease (NAFLD), which has reached epidemic proportions globally [1,2]. It is a multisystem disease that may not only develop into severe chronic hepatic diseases but also contribute to extrahepatic diseases such as type 2 diabetes, cardiovascular disease, and chronic kidney disease, causing a tremendous clinical and economic burden [3,4]. A recent large nationwide cohort study with long-term follow-up showed significantly increased overall mortality with all NAFLD histological stages including steatosis [5], thus it is suggested that steatosis can no longer be ignored as 'benign and an incidental finding' [6]. Although there have been clinical trials exploring drug candidates, no pharmacological treatments have been approved for NAFLD so far [7]. Hence, effective intervention strategies are urgently needed to delay or halt its progression to related hepatic and extrahepatic diseases.

Accumulating evidence suggests that NAFLD is a disease closely related to gut microbiota via the gut-liver axis [8,9], which stimulates the efforts to explore therapeutic interventions to improve NAFLD by modulating the gut microbiota [10]. The safety and persistence needed for microbiota-targeted intervention in humans highlights the importance of exploring microbiota-directed foods (MDFs), which by definition can elicit a targeted metabolic response in specific indigenous microbiota that confer a health benefit on the host [11]. Prebiotics and synbiotics like oligofructose and yogurt, which can manipulate gut microbiota, were found to reduce insulin resistance, intrahepatic lipids, liver enzymes, and histologically confirmed steatosis in patients with NAFLD/NASH [12-14]. Despite the promise of MDFs in patients with NAFLD, such studies are in an early stage [15]. According to the NAFLD practice guidance from the American Association for the Study of Liver Diseases, rigorous, prospective, longer-term trials are required before making recommendations about specific diets [16]. The complexity and heterogeneity of the NAFLD pathogenesis calls for deep and extensive phenotyping to evaluate the multiple effects of an intervention on NAFLD and their molecular mediators [17]. Furthermore, subsequent causal investigations are needed to verify the specific microbial signatures identified in clinical studies and their metabolites, which represent the highest levels in the chain of evidence in microbiome-linked disease [18].

In line with the above we performed here a randomized, double-blinded, placebo-controlled clinical trial in individuals with NAFLD that lasted for 4 months to enable a relatively long-term observation. We used as MDF resistant starch (RS), a prebiotic of nondigestible fibers that are fermented in the large intestine [19], which has been shown to reduce adiposity and exert metabolic benefits in previous animal studies [20-23], however so far, no clinical study has investigated the therapeutic effect of RS on NAFLD. Comprehensive clinical measurements were conducted to evaluate the changes in metabolic phenotypes of NAFLD during the intervention. Multi-omics profiling was used to provide an integrated understanding on how RS and associated alterations in the gut microbiota or metabolites contributed to NAFLD improvement. In addition, the potential RS-targeted gut species and microbial metabolites revealed by multi-omics analysis were validated in mice and cell lines for causal insights.

RESULTS

Four-month RS intervention alleviates NAFLD in Chinese adults

To investigate the effects of RS on NAFLD, we conducted a randomized, double-blinded, placebo-controlled clinical trial in Shanghai, China from 2016-March to 2017-October (ChiCTR-IOR-15007519). A total of 200 participants with NAFLD (145 male and 55 female) were recruited and randomized with a 1:1 allocation to the 4-month administration of RS type 2 from high-amylose maize (HAM-RS2, 40 g/day) or control starch (CS) with equal energy supply (**Figures 1A** and **S1A**). The average age of all the participants receiving randomization was 39.1 ± 9.1 years (mean ± SD), while the average intrahepatic triglyceride content (IHTC) was 24.12% ± 14.64% (mean ± SD). Both groups were counseled to manage their diet following the standard menu designed by nutritionists. We measured the IHTC by magnetic resonance spectroscopy (MRS) during interventions along with anthropometric parameters and biochemical indexes. Four participants (3 in the CS group and 1 in the RS group) did not receive the corresponding intervention after randomization and were therefore excluded from the primary analysis (**Figure S1A**). Baseline anthropometric and clinical characteristics of participants were balanced between the two groups (**Table 1**). During the 4-month (120-day) intervention, the mean (± SD) percentage of the meals for which participants adhered to starch intake was 84.0 ± 16.1% in CS and 83.8 ± 12.6% in the RS group, with no significant difference (**Figure S1B**). Similarly, no significant difference was found in the adherence to diet (**Figure S1C, Table S1**). Dietary intake of energy and macronutrients except fiber were not significantly different between the two groups (**Table S1**).

After the 4-month intervention, the primary outcome IHTC was significantly decreased in the RS group compared to the CS group (-13.29% vs. -6.32%, $P < 0.0001$) (**Figure 1B)**. The net absolute and relative change of IHTC in the RS group relative to the CS group was -9.08% (95% CI: -11.91% to -6.26%) and -39.42% (95% CI: -56.13% to -22.72%), respectively (**Table 1**). Together with the alleviation of steatosis, we observed significant reduction of body weight and BMI in the RS group compared to the CS group (**Figure 1C**). The waist circumference, hip circumference, and waist-hip ratio (WHR) in the RS group were all lower compared to the CS group. Regarding the body composition, the reduction of fat percentage (FAT%) and fat mass (FM) were all significantly higher in the RS group compared with the CS group (**Table 1**). The reduction of visceral fat areas (VFA) and subcutaneous fat areas (SFA) evaluated by abdominal magnetic resonance imaging (MRI) were significantly higher after RS consumption compared to CS consumption (**Figures 1D and 1E**).

Furthermore, we observed significant reductions in alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma-glutamyl transpeptidase (GGT) after RS intervention (**Figures 1F-1H; Table 1**), which indicate the improvements of liver injury.

The dyslipidemia was also alleviated by the RS intervention as shown in the improvement of total cholesterol (TC), triglyceride (TG), low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C), which was absent after CS intervention (**Table 1**). Notably, fibroblast growth factor 21 (FGF21), a generally acknowledged NAFLD biomarker [24], was reduced after RS consumption (**Figure 1I**). The level of CK18 M65ED, which correlates with hepatocyte apoptosis and independently predicts the presence of NASH [24], was also significantly lower after RS compared to CS consumption (**Table 1**). In addition, the circulating levels of lipopolysaccharides (LPS) and other inflammatory markers including MCP-1, IL-1β and TNFα, were all significantly reduced after RS intervention in comparison to CS intervention (**Figure 1J-1M**).

While the fasting blood glucose level was reduced in both groups after the 4-month intervention, neither fasting nor postprandial glucose levels during the meal tolerance test demonstrated significant difference between the RS and CS intervention. Both the fasting and postprandial insulin levels were significantly decreased in the RS compared to CS group, as well as the insulin resistance evaluated by homeostasis model assessment (HOMA-IR) and insulin resistance index of adipose tissue (Adipo-IR) (**Table 1**). Besides, the 4-month RS consumption also resulted in cardiovascular improvements as the blood pressure was significantly decreased compared to the CS consumption (**Table 1**). Due to dietary management, significant reductions of adiposity and several metabolic parameters were also observed in the CS group, albeit significantly smaller than the RS group.

We also performed a secondary analysis to adjust for the effect of weight loss. The net absolute change of IHTC in the RS group relative to CS group after adjusting for weight loss was -5.89% (95% CI: -8.87% to -2.91%), corresponding to a relative change of -24.30% (95% CI: -42.42% to -6.18%), and remained statistically significant ($P = 0.0001$) (**Table 1**). A regression analysis associating absolute change of IHTC with weight loss showed an R^2 of 23%, suggesting only a small part of the RS effect was mediated by weight loss. While some clinical parameters showed weight loss-dependent changes, the changes of other parameters related to adiposity (VFA), glucose metabolism (insulin levels, HOMA-IR, Adipo-IR), lipid metabolism (TG, TC, HDL-C, LDL-C), hypertension (SBP, DBP), and ALT all remained significant (**Table 1**). Moreover, we included all randomized participants including four participants who did not receive the corresponding intervention in our sensitivity analysis, and the conclusion remains the same (**Table S2**). Collectively, a 4-month RS intervention reduced IHTC and improved liver injury and related metabolic disorders in patients with NAFLD, even after adjusting for weight loss.

RS intervention alters both fecal and serum metabolites in patients with NAFLD

To investigate how the 4-month RS intervention affected the metabolism of the human host and the commensal intestinal microbiota, we performed targeted metabolomics on serum and fecal samples of participants in both the RS and CS groups before and after the intervention. In total, we measured 30 amino acids (AAs) and 26 bile acids (BAs) in serum, and 10 short chain fatty acids (SCFAs) and 18 BAs in feces. The RS and CS interventions had different effects on the overall changes in measured metabolites ($P < 0.05$, PERMANOVA) (**Figure 2A**). Examination of the different categories of metabolites showed a small but significant change in both serum and fecal BA profiles ($P < 0.05$, PERMANOVA).

At the level of individual metabolites, 13 metabolites among fecal BAs, serum BAs and serum AAs were significantly changed ($P < 0.05$, Wilcoxon signed-rank test) by the RS intervention but not the CS intervention (**Figure 2B**). No differences were observed in fecal SCFA metabolites. All serum AAs with significant changes showed different directions of change between the RS and CS groups. Interestingly, the serum levels of all three branched-chain amino acids (BCAAs) (valine, leucine and isoleucine) decreased after the RS intervention. The glutamate–serine–glycine (GSG) index, a possible marker of liver disease severity that is independent of BMI [25], was also significantly reduced after the RS intervention. In addition, we found 10 metabolites showing no differences at baseline to be significantly different between the RS and CS groups at the end of the intervention ($P < 0.05$, Wilcoxon rank-sum test), including valine, phenylalanine, and alpha-aminobutyric acid.

Spearman's correlation analyses showed multiple strong, significant correlations between the identified significant metabolites and patients' clinical parameters (**Figure S2A**). To identify key metabolites and their possible relationships with NAFLD that were independent of body weight, we repeated the correlation analyses, adjusting for clinical parameters related to obesity, including BMI, waist circumference, VFA, SFA and body fat percentage. This revealed multiple metabolites that significantly positively (alanine, valine, leucine, and tyrosine) or negatively (aminobutyric acid) correlated with levels of human IHTC (false discovery rate [FDR]-corrected $q < 0.1$) (**Figure 2C**). BCAAs and some BAs including serum taurocholic acid (TCA) and serum glycocholic acid (GCA) were significantly correlated with three NAFLD-relevant liver enzymes ALT, AST and GGT (FDR-corrected $q < 0.1$). The serum levels of alanine, α-aminobutyric acid and valine ($P=0.062$) also correlated with serum triglycerides (**Figure 2C**). Furthermore, the correlations between BCAAs and IHTC, the primary outcome in our trial, remained significant after controlling for obesity-related measures and insulin resistance (HOMA-IR).

In summary, the RS intervention may exert its beneficial effects on patients with NAFLD by altering the levels of microbial metabolic products, specifically the AAs pool and BCAA levels available for the human host.

The changes of gut microbiota upon RS intervention are associated with NAFLD alleviation

To investigate changes in the gut microbiota, we performed shotgun metagenomic sequencing on fecal samples before and after the 4-month intervention for 50 participants randomly selected from each group (matched with the full analysis set), generating 6.1 Gbp of sequencing data on average (*s.d.* 1.3 Gbp per sample). While similar at baseline in alpha (richness, Simpson index, and Faith's phylogenetic diversity) and beta diversity (weighted or generalized UniFrac) based on MetaPhlAn2 taxonomic profiling, significant differences between the RS and CS groups were observed after the 4-month intervention ($P < 0.05$, Wilcoxon rank-sum test for alpha and PERMANOVA for beta diversity) (**Figures 3A and 3B**). This result suggested different effects of RS on the overall gut microbiota community compared to CS. Specifically, the RS group had lower alpha diversity than the CS group after the intervention. This is consistent with many human and animal studies into the effects of RS2 consumption, as reviewed before [26]. Bendiks *et al.* also suggested the enrichment of particular taxa, which can efficiently metabolize RS and its degradation products, as the possible reason of decreased alpha diversity. In addition to the MetaPhlAn2 profiling, we used an approach relying on co-abundance gene groups (CAGs) to quantify the gut microbiota composition. This led to the same findings for comparisons of microbiota alpha and beta diversity (**Figure S2B and S2C**).

To uncover the bacterial species that were potentially associated with the beneficial effects of the RS intervention, we adopted two approaches: a non-parametric Wilcoxon test and generalized linear models. We focused on species that either significantly changed their abundance after the RS treatment (but did not change after CS intervention) (**Figure 3C**) or became significantly different in abundance between the two groups after the intervention (with no differences at baseline). The non-parametric test revealed that the relative abundances of 31 species significantly changed compared to the baseline or control group

($P < 0.05$, Wilcoxon signed-rank test or Wilcoxon rank-sum test). The generalized linear model found microbiota species that were significantly associated with the intervention, while controlling for the effect of obesity-related measures. This analysis led to the identification of species including *Bacteroides stercoris*, whose abundance was significantly lower after the RS compared to the CS intervention (FDR-corrected $q < 0.2$) (**Figures 3C and Table S3**). We correlated the abundances of all significant bacterial species with a panel of clinical parameters, adjusting for obesity-related measurements, to pinpoint the key species that were relevant to NAFLD (**Figure 3C**). We focused on the bacteria that significantly correlated with important clinical features in NAFLD (IHTC, ALT, AST, GGT and FGF21), and found *Bacteroides stercoris* correlated positively with IHTC, ALT and AST. Significance remained (except $P = 0.054$ for AST) after further adjusting for insulin resistance (HOMA-IR).

We next sought a deeper understanding of the bacterial-phenotype associations by integrating them with the metabolomic profiles. We observed significant correlations between the gut microbial community and the overall fecal BA and fecal SCFA profiles, as well as the serum AA profile ($P < 0.05$, Mantel test). Moreover, significant associations were found between microbiota composition and serum levels of valine, isoleucine and leucine ($P < 0.01$, PERMANOVA). To further disentangle the interplay between gut microbiota taxonomy and serum or fecal metabolite pools, we used Spearman's correlations to link microbial species and metabolites with significantly differential abundances. *Parabacteroides merdae*, whose abundance was significantly lower in RS than CS group after intervention, had the highest number of significant correlations (mostly positive) with multiple metabolites (**Figure 3D**). The RS-depleted intestinal microbe *B. stercoris* correlated positively with serum valine level ($P < 0.05$, Spearman's correlation), which also showed significant positive correlations with IHTC, ALT, AST, GGT and TG (**Figure 2C**).

Transplantation of RS-altered gut microbiota alleviates NAFLD in mice

To investigate the potential causality between RS-induced broad gut microbiota alteration and reduction of hepatic steatosis, we performed fecal microbiota transplantation (FMT) (**Figure 4A**) into conventional antibiotics-treated mice fed with high-fat high-cholesterol (HFHC) diet, using samples from human donors after RS or CS intervention (whose changes in IHTC after the intervention were close to the corresponding group average; $n = 2$ per group). Compared to CS donors, FMT from RS donors led to significant reduction of the body weight and liver weight (**Figures S3A-S3C**). Serum level of FGF21 was significantly lower in the RS group, which was accompanied by the increased expression of FGF21 receptor, co-receptor, and adiponectin in the adipose tissue (**Figures S3D-S3F**). Improvement of glucose metabolism, especially a significant increased insulin sensitivity, was also observed in mice receiving fecal microbiota from RS donors (**Figures S3G and S3H**). Histological assessments demonstrated significant decrease in hepatic steatosis, ballooning, inflammation and NAFLD activity score after FMT from RS donors (**Figure 4B and 4C**). Moreover, the RS group had lower levels of liver enzymes ALT and AST, hepatic TG, and total cholesterol in the liver (**Figure 4D-4G**). At the molecular level, FMT from RS donors reduced the expression of marker genes in liver related to inflammation, macrophage, and neutrophil recruitment (**Figure 4H**). It also reduced gene expression in liver for lipogenesis and promoted the expression of genes related to lipolysis (**Figures 4I and 4J**). Moreover, we also observed the improvement of gut barrier integrity as reflected by the increased expression of genes encoding tight junction proteins (**Figure 4K**), together with a significant reduction of serum LPS suggesting a possible alleviation of systemic inflammation (**Figure 4L**). The levels of BCAAs in the colon content were also significantly reduced in mice receiving FMT from RS donors than from CS donors (**Figure 4M**).

In addition to the wild-type mice, we also performed the experiment using a genetic model of NAFLD, where *ApoE*$^{-/-}$ mice were fed with HFHC diet followed by the same FMT procedure. The causal effect of RS-mediated microbiome changes was successfully replicated in the *ApoE*$^{-/-}$ mice, including changes in body weight, liver weight, histological scores, liver enzymes and serum FGF21 (**Figures S4A-S4H**). Consistent with the wild-type mice, the *ApoE*$^{-/-}$ mice receiving FMT from RS donors had decreased expression of lipogenesis-related genes and increased expression of lipolysis-related genes in the liver, as well as lower levels of colonic BCAAs (**Figures S4I-S4K**). Moreover, serum LPS was also significantly reduced in the RS compared to the CS group, coupled by increased expression of genes related to gut barrier integrity in the ileum (**Figures S4L and S4M**). In line with the histological changes in inflammation, expression of inflammation-related genes in the liver were effectively reduced (**Figure S4N**).

Multi-omics integration analysis identifies key species associated with NAFLD alleviation

We profiled the functional potential of the gut microbiota and examined functional differences in the RS and CS intervention groups. We found using the MetaCyc database the relative abundances of 8 pathways to be significantly altered after the RS intervention ($P < 0.05$, Wilcoxon signed-rank test) (**Figure S5A**). The microbiota functional potential for starch degradation (MetaCyc PWY-6731) significantly increased after the RS intervention ($P = 0.038$, Wilcoxon signed-rank test), but not in the CS group (**Figure S5B**). In a particular category of gene families responsible for carbohydrate metabolism, we found that 14 CAZy families were significantly altered after both the RS (**Figure 5A)** and CS (**Figure S5C**) interventions ($P < 0.05$, Wilcoxon signed-rank test), with no common families between RS and CS. Interestingly, a significant decrease was observed only after the RS intervention in the abundances of the KEGG functional modules M00060 (lipopolysaccharide [LPS] biosynthesis, KDO2-lipid A, $P = 0.024$) and M00320 (LPS export system, $P = 0.012$, Wilcoxon signed-rank tests) (**Figure 5B**). Another two LPS-biosynthesis-related KEGG modules (M00063 and M00064) had significantly increased abundances only after the CS intervention ($P = 0.017$ and $P = 0.023$, respectively, Wilcoxon signed-rank test). As a proinflammatory bacterial compound, LPS can reduce intestinal barrier function and increase translocation, and is demonstrated to accelerate hepatic steatosis in NAFLD development [27]. Moreover, *B. stercoris*-specific LPS biosynthesis potential (M00060) was also significantly lower in RS than CS group after intervention ($P = 0.012$, Wilcoxon rank-sum test).

The analyses above revealed several potential intestinal species/functions markers, and signature metabolites related to NAFLD improvement after the RS intervention. To uncover potential mechanistic links between changes in gut microbiota and NAFLD alleviation, we applied a recently developed computational framework to integrate various data types [28]. Initially, we used a three-tiered analysis to screen out microbiota functions (KEGG modules) that were significantly correlated with metabolites and important phenotypes (IHTC, ALT, AST, GGT, and FGF21), while adjusting for obesity-related parameters. These functions may serve as a bridge between the gut microbiota and host metabolism and thus could be potentially related to NAFLD progression, such as the biosynthesis and transport of various amino acids (tryptophan, histidine, lysine), cobalamin (vitamin B12) and lipopolysaccharide (**Figure S6**). In the functional modules significantly correlated with IHTC ($P < 0.05$, Spearman's correlation coefficient ≥ 0.2) in the RS group, we identified four KEGG modules related to BCAA biosynthesis (**Figures 5C and 5D**), emphasizing the strong relevance of BCAAs in NAFLD pathogenesis. Subsequently, gut microbiota driver species analysis was performed to determine which species were the main contributors of the function-phenotype associations. Overall, many more potential KEGG modules (correlated to FGF21 or IHTC) and driver species were observed in the RS than in the CS group, adding evidence that RS shaped the gut microbiome composition and activity in a directed way (**Table S4**). In particular, we found three highly contributing driver species involved in the correlations between the four BCAAs modules and IHTC (**Figure 5D**). *B. stercoris*, the abundance of which was correlated with IHTC, ALT, AST and serum valine, had the strongest driving effect on average of the four modules, especially M00019 for valine/isoleucine biosynthesis.

To validate the positive association between *B. stercoris* and NAFLD, we re-analyzed the metagenomic data from two published cohorts involving NAFLD (see Methods). In a Chinese cohort [29], the abundance of *B. stercoris* was significantly higher in patients with NAFLD than in NAFLD-free participants (**Figure 5E**). In a European cohort, patients diagnosed by liver biopsy [30] with moderate or severe steatosis also had higher levels of *B. stercoris* compared to mild steatosis or control (**Figure 5F**).

B. stercoris promotes NAFLD progression partially through LPS and BCAA production

To examine a potential causal effect of *B. stercoris* on NAFLD progression, mice were given a HFHC diet for 8 weeks to induce NAFLD, together with daily oral gavage of live or heat-killed *B. stercoris* at 5×10^9 cfu/day along with the HFHC feeding (**Figure 6A**). Real-time PCR showed a significantly increased amount of *B. stercoris* in the feces of the live bacteria group compared with mice on the HFHC diet only (**Figure 6B**). *B. stercoris* treatment showed no obvious effects on body weight or fat mass percentage (**Figures S7A-S7B**), but significantly increased the liver weight percentage compared to control (**Figure S7C**). Despite no obvious effects on either glucose or insulin levels in both fasting and fed status, *B. stercoris* intervention for 8 weeks led to an impaired ability of the mice to dispose glucose and decreased insulin sensitivity (**Figures S7D-S7G**). The serum level of ALT increased 1.8-fold in mice gavaged with live *B. stercoris* while AST did not change significantly (**Figure 6C and 6D**). By histological assessment, hepatic lipid accumulation, inflammatory cell infiltration and fibrogenesis were all markedly enhanced in mice gavaged with live *B. stercoris* (**Figures 6E-6J**) compared to mice only fed with HFHC. Consistent with the histological observations, the level of hepatic TG was more than 2-fold higher in mice gavaged with live *B. stercoris* (**Figure 6K**). Serum level of LPS was also significantly increased after 8-week oral gavage (**Figure 6L**). To determine the specific impact of *B. stercoris* at the molecular level, we explored the transcription of inflammatory and fibrogenesis markers in liver tissues. In line with the higher histological scores, genes involved in pro-inflammatory response, inflammatory cell infiltration, and collagen formation were significantly higher in mice gavaged with live *B. stercoris* (**Figures 6M and 6N**). Collectively, these findings suggest that increased abundance of *B. stercoris* contributed to NAFLD progression.

Besides live *B. stercoris*, heat-killed *B. stercoris* also showed ability to aggravate NAFLD and elicited similar effects on inflammation, as reflected in the histological score, ALT, serum LPS, and expression of genes involved in inflammation activation (**Figure 6C, 6H, 6L-6M**). Yet, the content of hepatic TG in mice with heat-killed *B. stercoris* showed the trend to be lower than that in the live group (P=0.054) (**Figure 6K**).

Given the strong driving effect of *B. stercoris* in the correlations between gut microbial BCAA biosynthesis and IHTC (**Figures 5C and 5D**), we then measured the fecal levels of BCAAs in mice. We found the 8-week oral administration of live *B. stercoris* significantly increased levels of fecal valine and isoleucine (**Figure 6O**). Unlike live *B. stercoris*, mice daily gavaged with heat-killed *B. stercoris* showed only a minimal effect on accumulation of fecal BCAAs, with similar pattern as steatosis score and hepatic TG (**Figure 6F and 6K**). To further demonstrate a direct metabolic production of BCAA by *B.*

stercoris, we performed various cell cultures, with and without *B. stercoris* for different time periods, followed by targeted metabolomics analysis of the cultured supernatant. Compared with other groups, the cultured supernatant of live *B. stercoris* showed a remarkable accumulation of BCAAs in a time-dependent manner, especially for valine (**Figures 6P and S7H**).

The significant correlation identified in our clinical study between valine and IHTC after controlling for obesity-related measures and HOMA-IR suggested a possible direct influence of valine on liver fat accumulation and thus NAFLD pathogenesis. We therefore investigated the direct *in vitro* effect of valine, which can be derived from NAFLD-promoting *B. stercoris*, on lipid metabolism in HepG2 cells. Compared with incubation with only fatty acid (FA), we observed a significant increase in intracellular TG content (**Figure S7I**), and a dose-dependent increase in expression of the transcription factor SREBP1 and lipogenic genes following incubation with valine (**Figure S7J**). Expression of FA transporters and their corresponding transcription factors demonstrated similar dose-dependent increases (**Figure S7K**). CPT1A, a gene involved in beta oxidation and lipid catabolism, had lower expression following incubation with valine (**Figure S7K**).

DISCUSSION

The vital role of gut microbiota in liver diseases has been demonstrated by studies involving FMT [31] or single species such as *Roseburia intestinalis* [32] and *Klebsiella pneumoniae* [33]. Such bidirectional relationship between the gut (and its resident microbiota) and the liver, i.e., the gut-liver axis, has gained attention in the last several years with the hope of developing microbiome-based strategies for diagnosis, prognosis and therapeutics of liver diseases [9,34,35]. However, the efficacy of most of the potential therapeutics for NAFLD needs confirmation in well-designed human studies [10]. Previous clinical trials have demonstrated the ability of MDFs to modulate human immune status [36] and to contribute to healthier metabolic and growth profiles of undernourished children [37]. In our randomized clinical trial, we evaluated the effects of RS as a MDF for NAFLD treatment. To quantify changes in liver fat content, we used MRI, a highly reproducible and the most accurate non-invasive approach to detect hepatic steatosis [24,38]. Several studies have also confirmed the superiority of MRI over liver histology in assessing liver fat [39,40]. It is more sensitive than histological grading in detecting changes in liver fat over time [41]. Four-month intervention with this MDF was effective in reducing IHTC in patients with NAFLD by an absolute reduction of -5.89% and a relative reduction of -24.30% after adjusting for weight loss. Such effect was partly mediated by altered composition and metabolic profile of gut microbiota. Indeed, transfer of fecal microbiota from human donors receiving 4-month RS into mice fed with HFHC diet reduced hepatic steatosis, lobular inflammation, and expression of lipogenesis- and inflammation-related genes, suggesting a causal role of gut microbiota in alleviating NAFLD. Moreover, expression of genes related to gut barrier integrity were enhanced while the level of serum LPS was reduced in mice receiving FMT from RS donors. This is consistent with decreased circulating level of LPS and the lower microbiota functional potential for LPS biosynthesis in human participants after RS intake.

Amino acids were also identified as possible molecular mediators of the RS beneficial effects. Perturbation in AA metabolism, especially aromatic amino acids (AAAs), GSG index and BCAAs, has been shown to be involved in NAFLD and NASH pathogenesis [25,30]. Serum levels of two AAAs, phenylalanine and tyrosine, were significantly lower after RS than CS intervention; serum glutamic acid for GSG index calculation was significantly reduced after RS intake (**Figure 2B**). Serum BCAAs has been associated with gut microbiome alteration and insulin resistance [42], which represents a NAFLD pathophysiology. Here we observed consistent correlations between BCAAs and insulin resistance, and the 4-month RS intervention in humans could significantly reduce the serum levels of BCAAs. Furthermore, serum BCAAs were positively correlated with IHTC, ALT, AST, and GGT. Importantly, the correlations between BCAAs and the primary outcome IHTC remain significant after adjusting for obesity-related parameters and insulin resistance, suggesting a direct influence of BCAAs on hepatic steatosis and thus NAFLD pathogenesis. In the FMT experiment where transfer of RS-altered microbiota into mice alleviated NAFLD, the colonic levels of BCAAs were also decreased, suggesting that the change of gut microbiota caused by RS led to the change in BCAAs. The role of BCAAs in hepatic steatosis was also supported by *in vitro* experiments investigating the direct effect of valine on intracellular TG levels, through modulation of lipogenic transcription factors, increased lipogenesis and decreased FA oxidation. Amino acids may modulate lipogenic transcription factors through participating in the processing of enzymes and transcriptional regulators as well as acting as substrates for lipid synthesis [43-45]. Elevated hepatic lipogenesis is intimately involved in pathological consequences [45].

Apart from the causality between RS-induced broad gut microbiota alteration and reduction of hepatic steatosis, we also attempted to pinpoint specific microbial species involved in NAFLD development though multi-omics integration analysis. Among them, we found RS reduced the abundance in the gut of *B. stercoris*, which is one of the species highly correlated with IHTC, ALT and AST. These positive correlations remained significant after controlling for obesity-related parameters and HOMA-IR, suggesting a body weight- and insulin resistance-independent effect of *B. stercoris* on NAFLD aggravation. The positive association of *B. stercoris* in the gut with NAFLD was further validated in two independent external case-control cohorts from Asia and Europe (**Figures 5E and 5F**). In addition, *B. stercoris* was selected as a feature in a metagenome-based model for predicting advanced fibrosis in US patients with NAFLD [46]. Furthermore, we conducted a monocolonization study to confirm the NAFLD-promoting effect of *B. stercoris* and to explore the possible mechanisms involved. Oral gavage of both live and heat-killed *B. stercoris* into mice could lead to increased lobular inflammation and enhanced expression of genes involved in inflammation activation, which might be explained by the increased serum level of LPS in both groups. On the other hand, considerably higher hepatic lipid accumulation was only observed in the mice gavaged with live *B. stercoris*, which was accompanied by the significantly higher levels of fecal BCAAs. Notably, the abundance of *B. stercoris* in human participants was also found to positively correlate with BCAAs (statistically significant for valine), and targeted measures of BCAAs in the monoculture supernatant of live *B. stercoris* substantiated its BCAA-releasing activity. Altogether, it suggests that *B. stercoris* can promote NAFLD progression, at least partially through LPS and BCAA production.

The serum level of FGF21 was found to be significantly reduced after the 4-month RS intervention. A number of preclinical and clinical studies demonstrate the robust effects of FGF21 on alleviation of dyslipidemia and NAFLD [47,48]. Contrary to the multiple metabolic benefits of FGF21, circulating FGF21 is paradoxically elevated in individuals with NAFLD [49,50]. The concept of ‘FGF21 resistance’ was proposed to explain the paradoxical changes of plasma FGF21 levels, in analogy to obesity-associated insulin and leptin resistance [51]. Based on animal studies, aberrant FGF21 signaling has been suggested as a key pathological step in the development and progression of NAFLD [52]. Notably, both circulating and hepatic levels of FGF21 in obese mice were markedly reduced by exercise training, where the FGF21 sensitivity in adipose tissue was enhanced [53]. Besides engineered human FGF21 analogues, the sensitization of the actions of FGF21 may represent an alternative strategy for treatment of metabolic disorders [48]. In line with this, here we observed decreased serum level of FGF21 in participants after RS intervention and in mice receiving feces from RS-fed donors, as well as increased expression of its receptor complex and downstream effector in adipose tissue. Our findings suggested that RS-induced microbiome changes might also lead to the sensitization of FGF21 actions.

Altogether, our study provides evidence that RS could be a novel, relatively simple and inexpensive microbiota-targeted therapeutic option for NAFLD, which can reduce IHTC by 5.89% in a weight loss-independent manner and decrease the liver enzymes indicative of liver injury and markers for systemic inflammation. The change of gut microbiota composition and functionality is an important mediator of the beneficial effect of RS on NAFLD amelioration, including one gut microbe *B. stercoris* that aggravates NAFLD at least partially through LPS and BCAA production. Our findings might contribute to further understanding of NAFLD pathogenesis and the development of innovative microbiome-based therapeutics or MDFs.

Limitations of the study

First, due to the lack of liver biopsy, we could not evaluate whether there were beneficial histological changes in the liver, such as biopsy-proven steatosis, NASH or fibrosis. However, our primary outcome was the change of liver fat content (hepatic steatosis) and IHTC is considered to be more sensitive than the histological steatosis grades in quantifying such changes, which has been recommended for clinical trial usage [54] and adopted by other NAFLD intervention studies [55,56]. Notably, the limitations of liver biopsy, including invasiveness, sampling error, poor acceptability, and only moderate reproducibility, also constrain its use as a repeat measurement to investigate histological changes in intervention studies [57]. Therefore, liver biopsy is not suitable for widespread use to assess disease stage or determine progression or response to therapy [58]. Second, in our randomized clinical trial, dietary guidelines were offered to the enrolled patients and information on their dietary intake was collected through questionnaires and further compared between the two intervention groups. Similar studies in the future may use a standard identical diet to directly control for the effect of diet as a potential confounding factor. Further research may reveal other possible molecular mechanisms by which the RS-altered metabolites or gut microbes lead to the accumulation or reduction of liver fat, the change of inflammation and fibrosis in the liver.

ACKNOWLEDGEMENTS

We would like to thank all the medical staffs and study participants who took part in the trial. The authors thank Prof. Jiarui Wu and Prof. Rong Zeng for the fruitful discussions, and Dr. Ruben V. Uribe and Prof. Morten Sommer for help with the design of qPCR primers. This work was supported by the National Natural Science Foundation of China (NSFC) major international (regional) joint research project (81220108006), Shanghai Municipal Key Clinical Specialty, Shanghai Research Center for Endocrine and Metabolic Diseases (2022ZZ01002), and the National Key Research and Development Program of China (2018YFA0800402) to W.J; the Marie Skłodowska-Curie Actions (MSCA) and Innovative Training Networks (H2020-MSCA-ITN-2018 813781) “BestTreat” (G.P., S.L.S., E.N., H.Leung.), and DFG under Germany’s Excellence Strategy – EXC 2051 – Project ID 390713860 (G.P., Y.N.); Excellent Young Scholars of NSFC (82022012), National Key Research and Development Program of China (2022YFA1004804), General Program of NSFC (82270907), Two Hundred Program from Shanghai Jiao Tong University School of Medicine (20191830) and Innovative research team of high-level local universities in Shanghai (SHSMU-ZDCX20212700) to H.Li.; the Innovation Program (DICP I202019) of Science and Research from the Dalian Institute of Chemical Physics, Chinese Academy of Sciences, and Key Foundation (21934006) from the National Natural Science Foundation of China to G.X.

AUTHOR CONTRIBUTIONS

W.J., G.P., H.Li., G.X. and Y.N. conceived and designed the study. L.Q., X.L., L.Zhang., Q.G., Q.W., and H.Li. recruited participants, collected and analyzed clinical data. L.Q. collected serum and fecal samples and extracted DNA from feces. L.Zhou, X.W. and Q.L. generated the targeted metabolomics data. Y.N., S.L.S., E.N. and H.Leung. performed bioinformatics analyses. L.Q., X.L., Y.L., X.J., S.L., R.Y. and Y.Z. conducted animal experiments. X.L., Q.W. and Y.Z. performed *in vitro* monoculture experiments. M.J.I. performed HepG2 cell line experiments. Y.N., L.Q., X.L., Y.L. and M.J.I. wrote the manuscript. Y.N., G.P. and H.Li. coordinated and supervised the study. W.J., G.P., H.Li., G.X., A.X. and H.E reviewed and edited the manuscript. All authors made substantial contributions and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Figure titles and legends

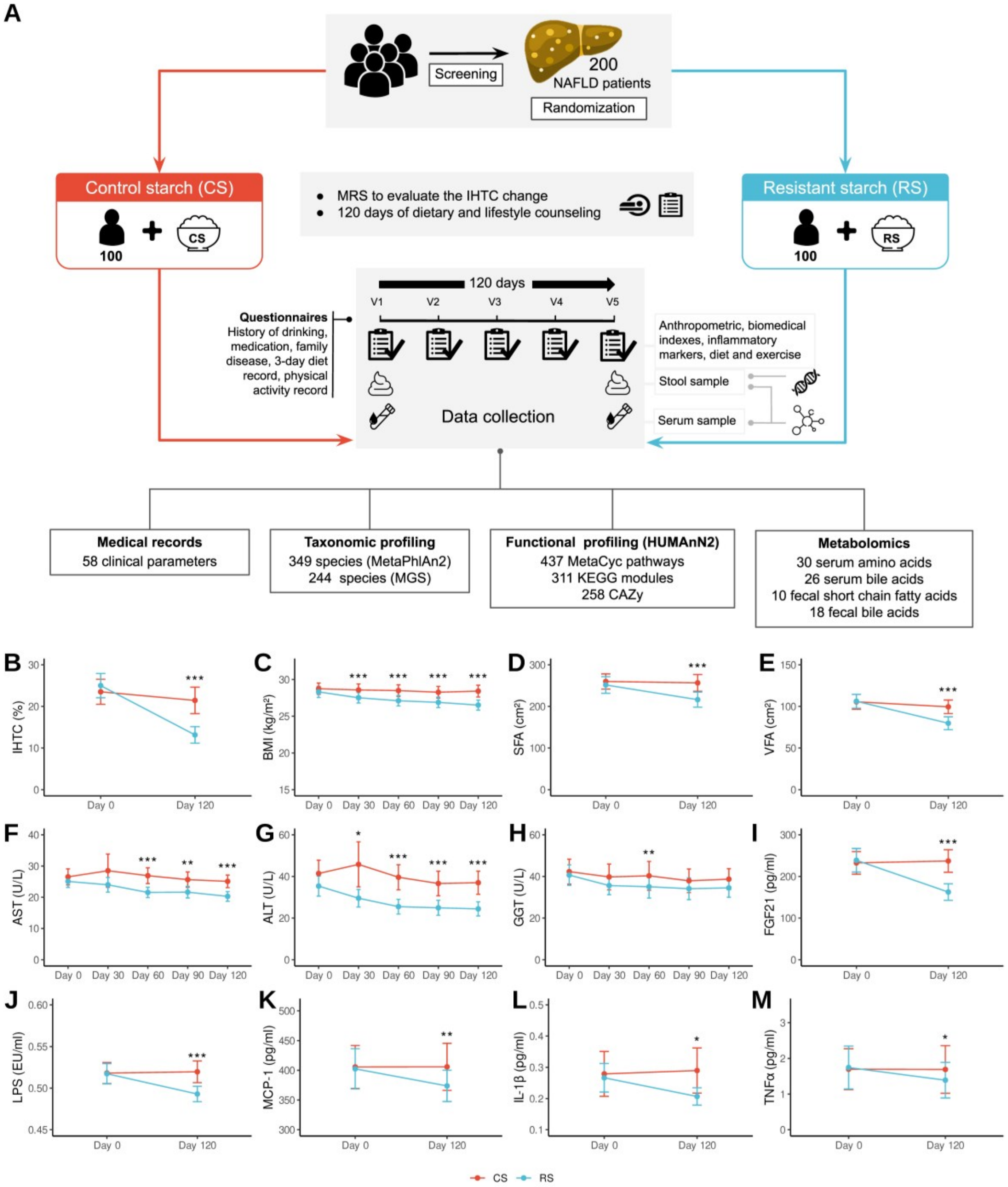


Figure 1: Resistant starch (RS) intervention for 4 months alleviates nonalcoholic fatty liver disease (NAFLD). (**A**) Overall study flow. RS intervention significantly reduced (**B**) intrahepatic triglyceride content (IHTC), (**C**) body mass index (BMI), and changed abdominal fat distribution during the study, including (**D**) subcutaneous fat area (SFA) and (**E**) visceral fat area (VFA). (**F-I**) liver enzymes (ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, gamma-glutamyl transpeptidase) and the serum NAFLD biomarker FGF21 changed during the intervention. (**J-M**) The reduction of serum levels of lipopolysaccharides (LPS) and other inflammatory markers including MCP-1, IL-1β and TNFα, after RS intervention in comparison to CS intervention. Analysis of covariance adjusted by baseline value was used for comparison between RS and CS at each visit. Red: control starch (CS) group; blue: RS group. Data are mean ± 95%CI. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ RS vs. CS.

See also Figure S1 and Table S1-S2.

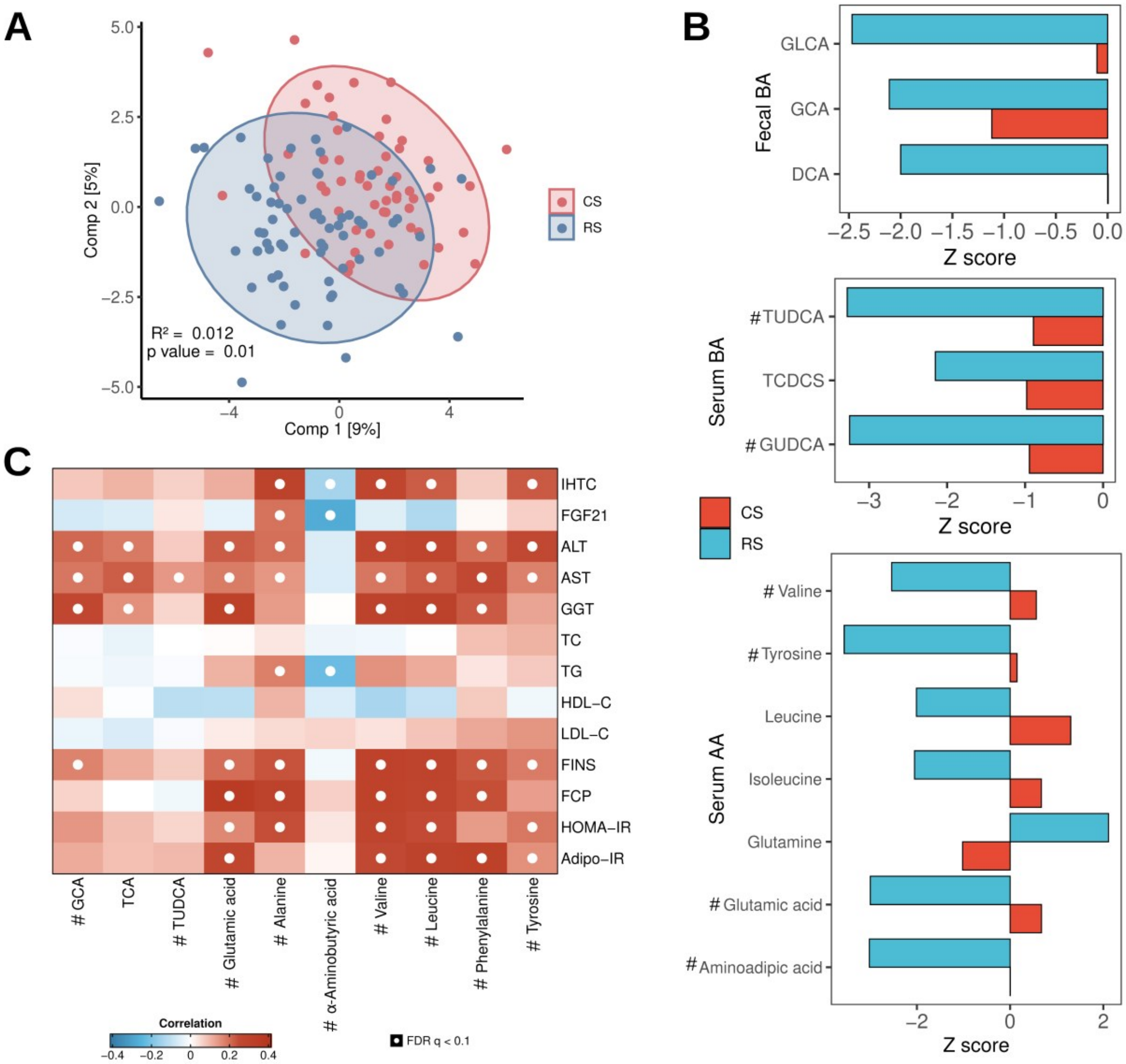


Figure 2: Fecal and serum metabolomic changes after 4 months of resistant starch (RS) and control starch (CS) interventions. (**A**) Difference in overall changes of measured metabolome between RS and CS. Log2-transformed fold-change profiles for all individual metabolites (serum and fecal) were used in partial least-squares discriminant analysis (PLSDA) and used to derive a Euclidean distance for statistical comparison between RS and CS with PERMANOVA. (**B**) Significantly changed fecal bile acids, serum bile acids and serum amino acids after the RS intervention (without significant changes after CS intervention) are shown. $P < 0.05$, Wilcoxon signed-rank test. X-axis represents z scores derived from Wilcoxon signed-rank test, for which positive and negative values indicate higher and lower abundance after the intervention, respectively. (**C**) Partial Spearman's correlation analyses between significant metabolites and patient clinical measures, adjusting for obesity-related clinical data (body mass index, waist circumference, visceral fat areas, subcutaneous fat areas, and fat percentage). Only serum metabolites significantly different between RS and CS groups at intervention end with no differences at baseline and with at least one significant correlation are shown. #, FDR-corrected $q < 0.1$ for individual metabolites. Circles, $P < 0.05$; solids dots, FDR-corrected $q < 0.1$. BA; bile acids; AA, amino acids; GLCA, glycolithocholic acid; GCA, glycocholic acid; DCA, deoxycholic acid; TUDCA, tauroursodeoxycholic acid; TCDCS, sulfated taurochenodeoxycholic acid; GUDCA, glycoursodeoxycholic acid; FINS: fasting insulin; FCP: fasting C-peptide.
See also Figure S2.

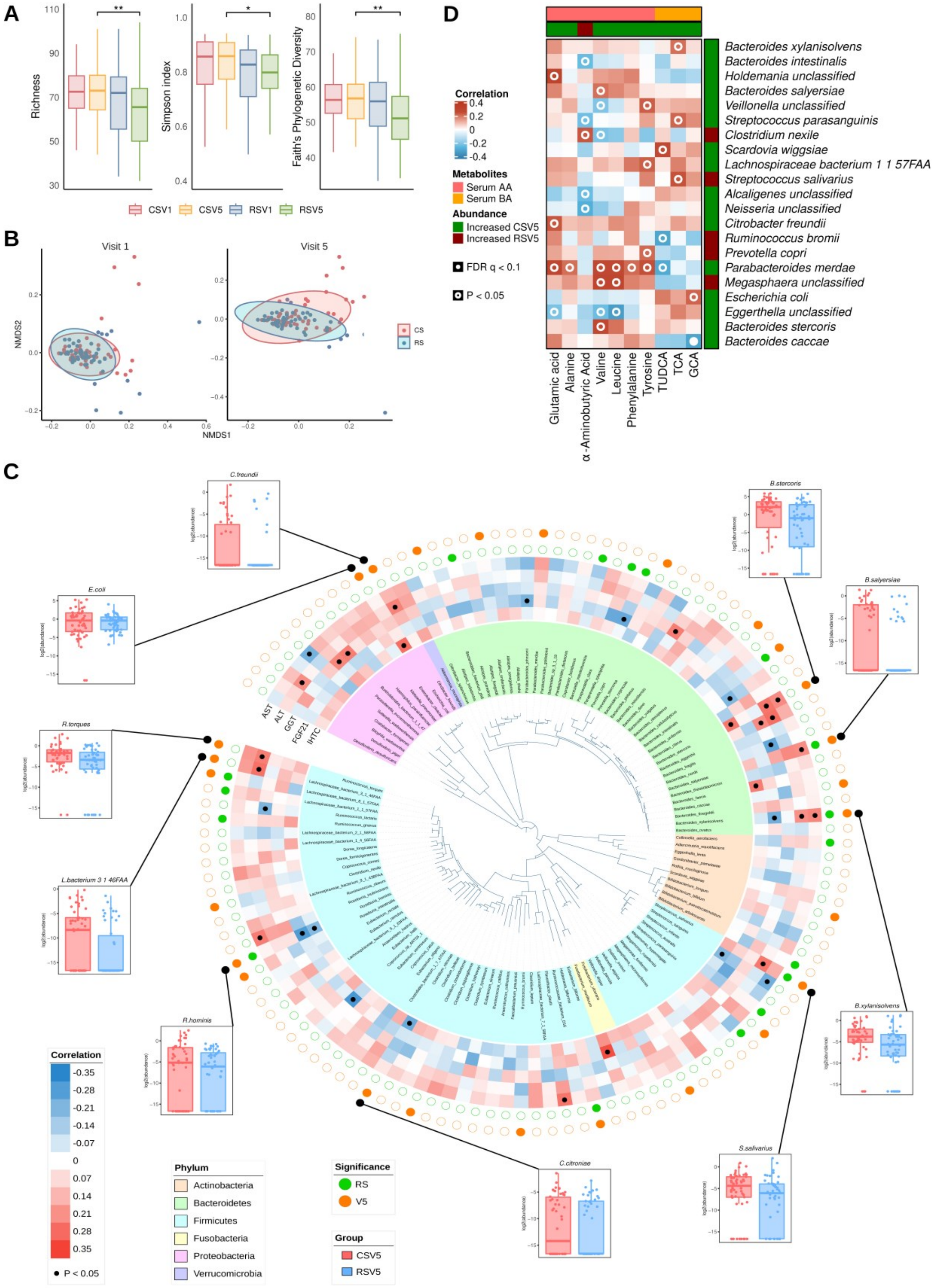


Figure 3: Compositional changes of gut microbiota after resistant starch (RS) intervention are associated with improvements in clinical phenotypes. Comparison of microbiota (**A**) alpha (richness, Simpson index, and Faith's phylogenetic diversity) and (**B**) beta diversity (weighted UniFrac distance) based on MetaPhlAn2-derived taxonomic profiles. *, P < 0.05; **, P < 0.01. Wilcoxon signed-rank or rank-sum test was used for alpha

diversity comparisons; PERMANOVA was used to assess the statistical significance of beta diversity comparisons. V1: visit 1 or baseline; V5: visit 5 or after 4-month intervention. (**C**) Circos plot showing significant species in a phylogenetic tree and correlations with liver-related parameters or biomarker. Significant species (solids dots, $P < 0.05$) refer to (i) microbial species with significantly changed abundances after 4-month RS intervention but not control starch (CS) ($P < 0.05$, Wilcoxon signed-rank test); or (ii) significantly different species after 4-month RS intervention, controlling for effects of obesity-related measures (BMI, waist circumference, visceral fat areas, subcutaneous fat areas, and fat percentage), using generalized linear models. Partial Spearman's correlations were calculated between significant species and patient clinical measures, adjusting for obesity-related clinical measures. Surrounding boxplots show the abundances after RS and CS intervention, for species with at least one significant correlation. (**D**) Spearman's rank-based correlations between significant metabolites (from **Figure 2C)** and significant species (from **Figure 3C**). After-intervention (V5) samples from both groups were used for correlation calculations. Metabolites and species enriched in RS or CS groups are respectively, dark red or green. Circles, $P < 0.05$; solids dots, FDR-corrected $q < 0.1$. Boxplots in (**A**) and (**C**) show median (centerlines), lower/upper quartiles (box limits) and whiskers (the last data points 1.5 times interquartile range (IQR) from the lower or upper quartiles). BA; bile acids; AA, amino acids.

See also Figure S2 and Table S3.

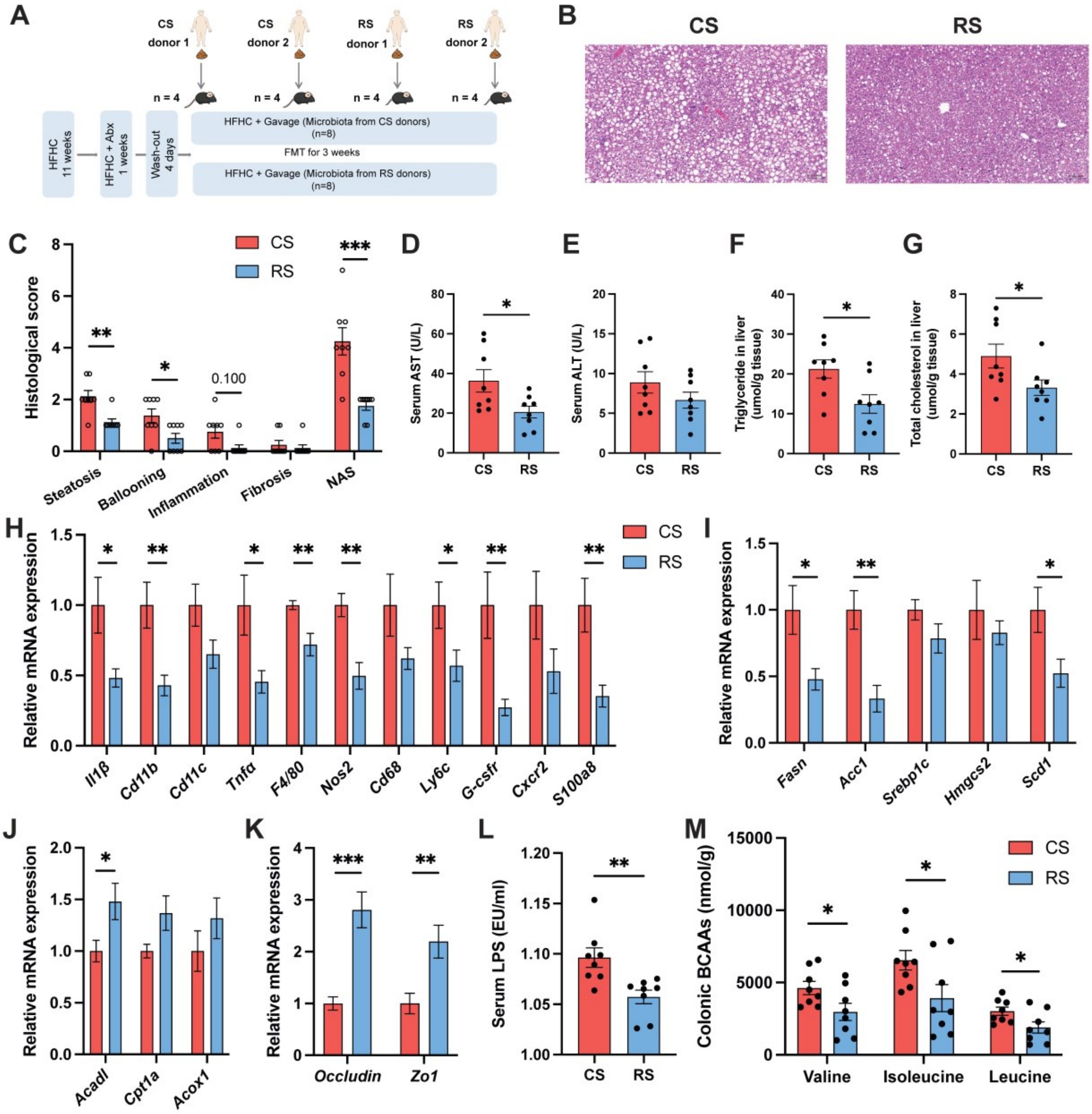


Figure 4: Transplant of RS-altered gut microbiota into mice alleviates diet-induced NAFLD. (**A**) Schematic diagram showing the study design for fecal microbiome transplantation from human donors to mice. (**B**) Representative images of liver sections stained with H&E (scale bar, 100 μm). (**C**) Quantification of histological scores. (**D-G**) Quantification of serum levels of AST, ALT, hepatic cholesterol, and hepatic triglyceride. (**H-J**) Expression levels of genes involved in inflammation, lipogenesis, and lipolysis in the liver. (**K**) Expression levels of genes related to tight junction protein in ileum. (**L**) Serum level of LPS. (**M)** Levels of valine, isoleucine, and leucine in colon content. Data are mean ± SEM. For C-G and L-M, n = 8 biological replicates for each group; For H-K, n = 2 technical replicates from 8 biological replicates for each group. *P < 0.05, **P < 0.01 and ***P < 0.001, two-tailed Student's unpaired *t* test (normally distributed) or non-parametric Wilcoxon rank-sum test (non-normally distributed) was used for statistical comparison. RS, resistant starch; CS, control starch; AST, aspartate aminotransferase; ALT, alanine aminotransferase; LPS, lipopolysaccharide.

See also Figure S3-S4 and Table S5.

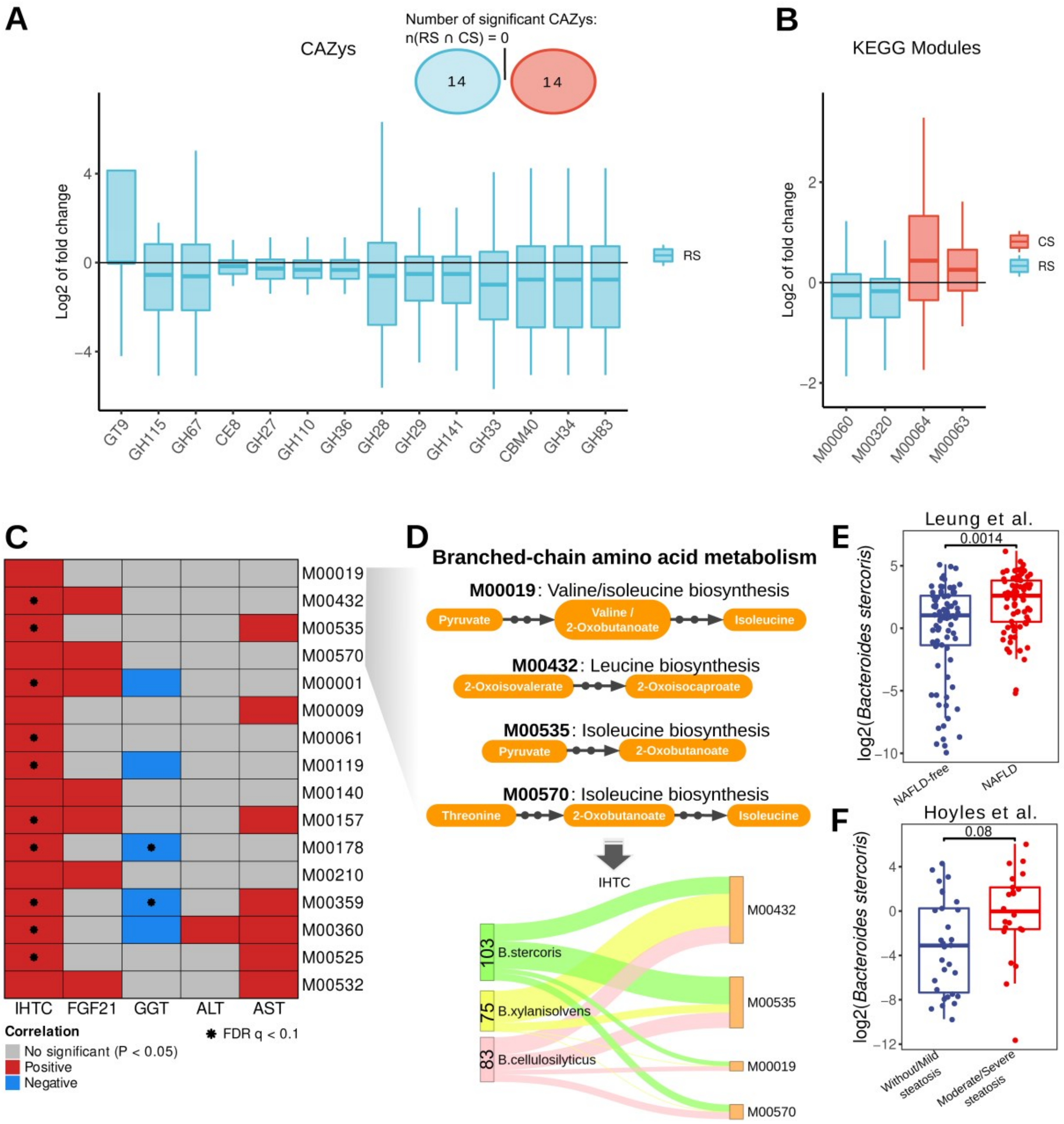


Figure 5: Gut microbiota functional changes after resistant starch (RS) intervention and driver species analysis linking microbiota function and phenotype alteration. (**A**) RS intervention for 4 months significantly altered the abundances of 14 CAZy families that did not change significantly with CS intervention. $P < 0.05$, Wilcoxon signed-rank test. (**B**) Abundance changes of KEGG functional modules related to lipopolysaccharide biosynthesis (M00060, M00063 and M00064) or export (M00320) in RS and CS groups. Significant changes in RS: M00060 and M00320; in CS: M00063 and M00064. $P < 0.05$, Wilcoxon signed-rank test. Boxplots in (**A**) and (**B**) show median (centerlines), lower/upper quartiles (box limits) and whiskers (the last data points 1.5 times interquartile range (IQR) from the lower or upper quartiles). (**C**) Correlations between gut microbiota functional modules and key clinical parameters or biomarkers of NAFLD. KEGG modules shown have strong correlations (absolute Spearman's correlation coefficients ≥0.2) with intrahepatic triglyceride content (IHTC) in the RS group. Red and blue indicate, respective, significantly ($P < 0.05$) positive and negative correlations. *, FDR-corrected $q < 0.1$. (**D**) Driver species for correlations between BCAA-related microbiota functions and IHTC. Species with contributions higher than 10% effect towards function-IHTC correlations when removed in leave-one-species-out analysis (pctSCCeffect.bgadj > 10%) were considered driver species.

Only top contributing species for BCAA module-IHTC correlations are shown with their driving effects illustrated in the Sankey plot. (**E**) Comparison of *B. stercoris* abundance in a published Chinese cohort. (**F**) Comparison of *B. stercoris* abundance in a published European cohort with liver biopsy. Data in (**E**) and (**F**) are log2-transformed bacterial relative abundances and compared with generalized linear model. Boxplots show median (centerlines), lower/upper quartiles (box limits) and whiskers (the last data points 1.5 times interquartile range (IQR) from the lower or upper quartiles).

See also Figure S5-S6 and Table S4.

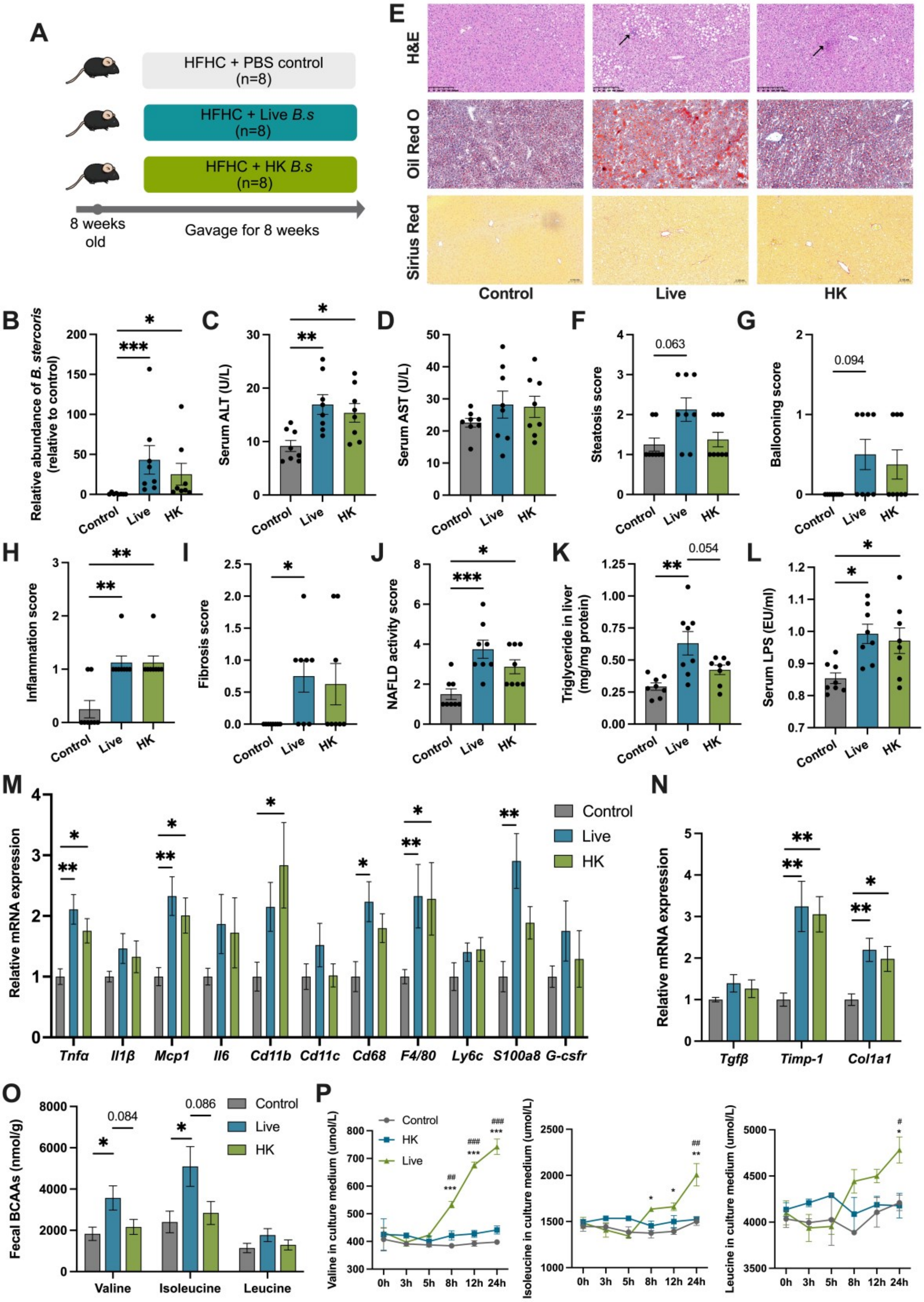


Figure 6: Supplementation with *Bacteroides stercoris* promotes NAFLD progression in mice partially through BCAA production. (**A**) Schematic diagram showing the study design for *B. stercoris* intervention. (**B**) Abundance of *B. stercoris* quantified by qPCR with specific primers to determine fecal content for groups. Bacterial abundances are relative to control group. (**C, D**) Serum levels of ALT and AST. (**E**) Representative images of liver sections stained with H&E (scale bar, 200 µm), Oil Red O (scale bar, 100 µm) and Sirius Red (scale bar, 100 µm). Arrows indicate inflammatory foci. (**F-J**) Quantification of

histological scores. (**K**) Quantification of hepatic triglycerides. (**L**) Serum level of LPS. (**M, N**) Expression levels of genes involved in inflammation (M) and fibrogenesis (N) in the liver. (**O**) Fecal levels of valine, isoleucine, and leucine by group. (**P**) Targeted quantification of BCAAs by UPLC-MS/MS in culture supernatants of live or heat-killed *B. stercoris* or culture media at different time points. For mice, data are mean ± SEM. For B and M-N, n = 2 technical replicates from 8 biological replicates for each group; For C-L and O, n = 8 biological replicates for each group; for P, n = 3 biological replicates for each group. *P < 0.05, **P < 0.01 and ***P < 0.001, based on one-way ANOVA (normally distributed) followed by Tukey's post hoc test, or Kruskal-Wallis test (non-normally distributed) followed by Dunn's test. In (P), * indicates the comparison between live and control while # (#P < 0.05, ##P < 0.01 and ###P < 0.001) indicates the comparison between live and heat-killed group. HFHC, high-fat, high cholesterol; BS, *Bacteroides stercoris*; HK, heat-killed; AST, aspartate aminotransferase; ALT, alanine aminotransferase; LPS, lipopolysaccharide.

See also Figure S7 and Table S5.

Table 1. Summary of clinical characteristics of the CS group and RS group.

Variables	CS group (n=97)			RS group (n=99)			P value[b]	P value[c]	P value[d]
	day 0	day 120	P value[a]	day 0	day 120	P value[a]			
NAFLD severity and biomarkers									
IHTC (%)	23.51 (20.49, 26.52)	21.44 (18.26, 24.62)	0.0780	24.99 (22.07, 27.91)	13.14 (11.16, 15.13)	<0.0001	0.4823	<0.0001	0.0001
CK18 M65ED (U/L)	362.20 (268.05, 456.35)	187.03 (141.26, 232.80)	<0.0001	304.77 (234.10, 375.45)	118.23 (61.01, 175.46)	<0.0001	0.3341	0.0071	0.2009
FGF21 (pg/ml)	232.64 (205.46, 259.83)	237.27 (210.17, 264.36)	0.8761	238.96 (210.69, 267.23)	162.56 (142.68, 182.43)	<0.0001	0.7497	<0.0001	<0.0001
Pro-C3 (pg/ml)	11.42 (10.49, 12.22)	10.2 (9.02, 13.66)	0.3292	10.96 (9.55, 12.29)	10.15 (9.27, 12.2)	0.401	0.1662	0.8646	0.6837
Anthropometric parameters									
Age (years)	38.91 (37.00, 40.82)	38.91 (37.00, 40.82)	-	39.20 (37.50, 40.90)	39.20 (37.50, 40.90)	-	0.8189	-	-
Female, No. (%)	28 (28.87)	28 (28.87)	-	26 (26.26)	26 (26.26)	-	0.6834	-	-
Weight (kg)	84.24 (81.22, 87.27)	83.29 (80.13, 86.46)	<0.0001	83.52 (80.67, 86.38)	78.05 (75.40, 80.70)	<0.0001	0.7307	<0.0001	-
BMI (kg/m^2)	28.74 (27.96, 29.52)	28.41 (27.61, 29.22)	<0.0001	28.31 (27.55, 29.07)	26.51 (25.82, 27.20)	<0.0001	0.4306	<0.0001	-
Waist circumference (cm)	97.80 (95.89, 99.71)	95.54 (93.45, 97.62)	<0.0001	97.43 (95.55, 99.31)	90.48 (88.81, 92.15)	<0.0001	0.7813	<0.0001	0.0095
Hip circumference (cm)	105.19 (103.74, 106.64)	104.56 (103.18, 105.95)	<0.0001	104.42 (103.01, 105.83)	101.95 (100.66, 103.25)	<0.0001	0.4528	0.0046	0.5009
Waist-to-hip ratio	0.93 (0.92, 0.94)	0.91 (0.90, 0.92)	<0.0001	0.93 (0.92, 0.94)	0.89 (0.88, 0.90)	<0.0001	0.6906	<0.0001	0.0063
Fat percentage	30.36 (28.92, 31.80)	29.08 (27.45, 30.72)	<0.0001	29.73 (28.20, 31.25)	26.56 (25.04, 28.07)	<0.0001	0.5478	<0.0001	0.0803
FM (kg)	25.65 (24.05, 27.25)	24.18 (22.51, 25.85)	<0.0001	24.96 (23.29, 26.64)	20.83 (19.33, 22.32)	<0.0001	0.5558	<0.0001	0.1012
FFM (kg)	58.30 (55.97, 60.62)	58.68 (56.26, 61.10)	0.8691	58.56 (56.40, 60.73)	57.23 (55.12, 59.35)	<0.0001	0.8681	0.0320	0.6661
TBW (kg)	40.94 (39.49, 42.39)	41.25 (39.72, 42.78)	0.8528	40.95 (39.53, 42.37)	40.16 (38.72, 41.60)	<0.0001	0.9907	0.0067	0.5086
VFA (cm^2)	105.51 (96.49, 114.54)	99.49 (91.35, 107.63)	0.0022	106.14 (97.80, 114.49)	79.75 (71.97, 87.52)	<0.0001	0.9192	<0.0001	<0.0001
SFA (cm^2)	259.90 (241.68, 278.13)	256.59 (236.46, 276.73)	0.0114	251.53 (231.61, 271.46)	216.19 (198.13, 234.26)	<0.0001	0.5397	<0.0001	0.1276
SBP (mmHg)	120.82 (118.79, 122.86)	117.25 (115.09, 119.41)	<0.0001	120.23 (118.29, 122.17)	113.32 (111.51, 115.12)	<0.0001	0.6760	0.0007	0.0127
DBP (mmHg)	80.30 (78.83, 81.76)	79.04 (77.53, 80.54)	0.0977	81.01 (79.63, 82.39)	76.24 (74.96, 77.51)	<0.0001	0.4837	<0.0001	0.0002
Liver enzymes and renal function									
ALT (U/L)	41.43 (35.02, 47.85)	37.00 (31.45, 42.55)	0.1183	35.37 (30.56, 40.19)	24.41 (21.03, 27.80)	<0.0001	0.1354	0.0002	0.0021
AST (U/L)	26.54 (23.97, 29.10)	25.10 (23.07, 27.12)	0.4042	25.08 (23.17, 26.99)	20.28 (18.74, 21.83)	<0.0001	0.3675	0.0098	0.1296
GGT (U/L)	42.32 (36.35, 48.30)	38.70 (33.64, 43.75)	0.0320	40.65 (35.75, 45.56)	34.51 (29.98, 39.03)	<0.0001	0.6680	0.0156	0.1851
TBIL (μmol/L)	12.39 (11.38, 13.40)	12.08 (11.01, 13.16)	0.5535	12.13 (11.21, 13.05)	13.11 (12.11, 14.11)	0.0429	0.7105	0.1435	0.5291
DBIL (μmol/L)	4.11 (3.80, 4.41)	4.05 (3.71, 4.39)	0.6838	3.99 (3.69, 4.29)	4.50 (4.17, 4.83)	<0.0001	0.5982	0.0300	0.2540
TBA (μmol/L)	3.52 (3.01, 4.02)	3.65 (3.12, 4.17)	0.6297	3.06 (2.59, 3.53)	3.02 (2.61, 3.43)	0.9908	0.1901	0.2527	0.5754
BUN (mmol/L)	4.74 (4.51, 4.97)	4.82 (4.55, 5.09)	0.0087	4.77 (4.55, 5.00)	4.62 (4.44, 4.79)	0.0002	0.8526	0.0634	0.2451
Cr (μmol/L)	70.61 (67.57, 73.65)	71.98 (68.61, 75.34)	0.1074	71.04 (68.06, 74.02)	71.42 (68.34, 74.51)	0.4352	0.8404	0.9393	0.8814
UA (μmol/L)	407.29 (388.70, 425.89)	405.07 (383.95, 426.20)	0.2524	399.40 (383.45, 415.35)	383.93 (366.81, 401.06)	0.0581	0.5222	0.5753	0.6491
RBP (mg/L)	51.18 (49.51, 52.84)	51.95 (50.05, 53.85)	0.1339	50.83 (49.35, 52.31)	51.48 (49.65, 53.31)	0.8875	0.7572	0.3742	0.7447
Lipid profiles									
TC (mmol/L)	5.09 (4.93, 5.25)	5.09 (4.90, 5.28)	0.4252	5.08 (4.91, 5.25)	4.87 (4.70, 5.05)	<0.0001	0.9513	<0.0001	0.0008
TG (mmol/L)	1.88 (1.69, 2.07)	1.97 (1.78, 2.17)	0.0232	2.01 (1.74, 2.28)	1.56 (1.40, 1.71)	<0.0001	0.4571	<0.0001	0.0011
HDL-C (mmol/L)	1.13 (1.08, 1.17)	1.13 (1.08, 1.18)	0.1136	1.13 (1.09, 1.17)	1.20 (1.15, 1.24)	0.0152	0.8337	0.0071	0.0029
LDL-C (mmol/L)	3.23 (3.08, 3.37)	3.17 (2.99, 3.34)	0.1224	3.13 (2.97, 3.29)	3.00 (2.85, 3.15)	0.0017	0.3682	0.0174	0.0399
NEFA (μEq/L)	616.30 (576.05, 656.56)	580.57 (544.00, 617.15)	<0.0001	649.22 (600.71, 697.73)	582.37 (549.31, 615.43)	<0.0001	0.3013	0.6823	0.4019
Glucose parameters during MTT									
PG 0min (mmol/L)	5.24 (5.14, 5.34)	5.00 (4.86, 5.15)	0.0029	5.12 (5.00, 5.25)	4.96 (4.85, 5.06)	<0.0001	0.1432	0.8178	0.4890
PG 30min (mmol/L)	8.02 (7.76, 8.28)	7.89 (7.60, 8.18)	0.4375	8.08 (7.77, 8.40)	7.87 (7.61, 8.13)	0.1334	0.7691	0.2909	0.3043
PG 60min (mmol/L)	8.26 (7.95, 8.58)	8.24 (7.90, 8.58)	0.4134	8.32 (8.02, 8.62)	8.22 (7.96, 8.47)	0.7323	0.8076	0.3217	0.5207
PG 120min (mmol/L)	7.32 (7.08, 7.56)	7.16 (6.87, 7.45)	0.1923	7.51 (7.26, 7.77)	7.13 (6.86, 7.40)	0.0114	0.2806	0.6327	0.8167
AUC (min*mmol/L)	911.44 (885.80, 937.07)	886.82 (857.88, 915.76)	0.1888	903.40 (878.75, 928.05)	881.03 (859.41, 902.64)	0.1910	0.6565	0.4098	0.7397

Insulin 0min (uU/ml)	19.95 (17.82, 22.07)	18.37 (16.43, 20.31)	0.0031	17.99 (16.33, 19.65)	13.38 (12.26, 14.51)	<0.0001	0.1503	<0.0001	0.0184
Insulin 120min (uU/ml)	100.79 (87.71, 113.86)	92.89 (80.78, 105.00)	0.2019	105.21 (90.07, 120.35)	79.30 (70.21, 88.40)	<0.0001	0.6618	0.0039	0.0257
C-peptide 0min (ng/ml)	3.27 (3.06, 3.48)	3.25 (3.01, 3.48)	0.3882	3.17 (2.99, 3.36)	2.69 (2.56, 2.82)	<0.0001	0.4952	<0.0001	0.0007
C-peptide 120min (ng/ml)	10.97 (9.88, 12.07)	10.60 (9.89, 11.32)	0.4786	11.17 (9.75, 12.60)	11.66 (8.43, 14.90)	0.4691	0.8242	0.6878	0.5418
HOMA-IR	4.70 (4.16, 5.23)	4.17 (3.67, 4.68)	0.0137	4.17 (3.73, 4.61)	2.96 (2.70, 3.23)	<0.0001	0.1302	<0.0001	0.0330
Adipo-IR	12.15 (10.63, 13.67)	10.76 (9.38, 12.13)	<.0001	11.27 (10.04, 12.51)	7.78 (7.00, 8.57)	<0.0001	0.3759	0.0001	0.0089
Inflammation-related factors									
LPS (EU/ml)	0.52 (0.51, 0.53)	0.52 (0.51, 0.53)	0.2953	0.52 (0.51, 0.53)	0.49 (0.48, 0.50)	0.0002	0.9284	0.0007	0.0028
MCP-1 (pg/ml)	405.61 (369.52, 441.71)	405.84 (366.19, 445.49)	0.4606	402.50 (368.60, 436.40)	373.75 (347.40, 400.10)	0.0086	0.9008	0.0068	0.0410
IL-1β (pg/ml)	0.28 (0.21, 0.35)	0.29 (0.22, 0.36)	0.9013	0.27 (0.22, 0.31)	0.21 (0.18, 0.23)	0.025	0.7718	0.0143	0.0405
TNFα (pg/ml)	1.70 (1.12, 2.27)	1.69 (1.02, 2.36)	0.2162	1.74 (1.14, 2.35)	1.39 (0.89, 1.89)	0.0398	0.9119	0.0168	0.0248
IL-6 (pg/ml)	1.48 (1.08, 1.88)	1.72 (1.23, 2.20)	0.2521	1.14 (0.83, 1.45)	1.17 (0.92, 1.41)	0.9643	0.1823	0.1070	0.4047

Data are presented as mean (95%CIs). Four participants (3 in the CS group and 1 in the RS group) did not receive the corresponding intervention after randomization and were therefore excluded from analysis.

CS, control starch; RS, resistant starch; IHTC, intra-hepatic triglyceride content; CK 18, Cytokeratin 18; FGF21, fibroblast growth factor 21; BMI, body mass index; FM, fat mass; FFM, free fat mass; TBW, total body water; VFA, visceral fat area; SFA, subcutaneous fat area; SBP, systolic blood pressure; DBP, diastolic blood pressure; ALT, alanine transaminase; AST, aspartate transaminase; GGT, gamma-glutamyl transferase; TBIL, total bilirubin; DBIL, direct bilirubin; TBA, total bile acid.; BUN, blood urea nitrogen,; Cr, creatinine; UA, uric acid; RBP, retinol binding protein; TC, total cholesterol; TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NEFA, non-esterified fatty acid; MTT, meal tolerance test; PG, plasma glucose; AUC, area under curve; HOMA, homeostasis model assessment; Adipo-IR, adipose tissue insulin resistance index; LPS, lipopolysaccharide; MCP-1, monocyte chemoattractant protein-1; IL, interleukin; TNFα, tumor necrosis factor alpha.

P value^a: Within-group differences were assessed using a linear mixed model.
P value^b: Differences in baseline variables between the RS and CS groups were assessed using Student's unpaired t-test.
P value^c: Differences in outcomes between the RS and CS groups were assessed using a linear mixed model.
P value^d: Differences in outcomes between the RS and CS group were assessed using a linear mixed model adjusted by weight loss and the baseline values of the variable assessed.

STAR ★ METHODS

RESOURCE AVAILABILITY

Lead Contact

- Further information and requests for resources and reagents should be directed to the Lead Contact, Weiping Jia (wpjia@sjtu.edu.cn).

Materials Availability

- This study did not generate new unique reagents.

Data and Code Availability

- The raw metagenomic sequencing data for all samples have been deposited into NCBI Sequencing Read Archive under accession number PRJNA703757.
- Computational analyses were performed using the bioBakery suite of tools including MetaPhlAn2 (https://github.com/biobakery/MetaPhlAn; Methods) for microbiota taxonomic profiling and HUMAnN2 (https://github.com/biobakery/humann; Methods) for profiling of functional potential (ECs, pathways and modules).
- Source values used to create figures in the manuscript are available as Data S1.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND PARTICIPANT DETAILS

Study participants

A total of 200 individuals who met all the following eligibility criteria were recruited in the Shanghai Jiao Tong University Affiliated Sixth People's Hospital. Inclusion criteria were: (i) ethnic Chinese, (ii) liver steatosis diagnosed by ultrasonography, (iii) aged 18-70 years old, and (iv) written informed consent obtained.

Exclusion criteria were: (i) participants with diabetes mellitus; (ii) alcohol consumption history of more than 20 g per day for men and more than 10 g per day for women; (iii) acute or chronic gastrointestinal diseases (including diarrhea, gastrointestinal infection, inflammatory bowel disease), malignant tumor or severe renal dysfunction; (iv) pregnancy, breastfeeding or planning to get pregnant; (v) consuming antibiotics within the last 3 weeks or during the study; (vi) viral hepatitis, drug-induced liver disease, total parenteral nutrition, Wilson's disease, autoimmune liver disease or other specific diseases that can lead to fatty liver; (vii) routine use of prescription medicines or adjuvant Chinese and Western medicines (except regular contraceptives); (viii) expected poor compliance; (ix) use of weight loss medication or participation in weight-loss program in the past 3 months; (x) mental disorder preventing cooperation; or (xi) wearing pacemaker or metallic implants, claustrophobia or other conditions that would be unable to undergo magnetic resonance examinations.

The study was approved by the Ethics Committee of Shanghai Jiao Tong University Affiliated Sixth People's Hospital, following the principles of the declaration of Helsinki. All relevant ethical regulations were followed during the study. Written informed consent was obtained from all participants. Complete clinical trial registration is deposited in the WHO International Clinical Trials Registry Platform and Chinese Clinical Trial Registry (http://www.chictr.org.cn/showproj.aspx?proj=12353; ChiCTR-IOR-15007519). The primary indication was change in intrahepatic triglyceride content (IHTC), with changes in anthropometric indicators, body composition, glycemic control and insulin sensitivity, liver and renal function, lipid profiles, cytokines, multi-omic parameters, Single nucleotide polymorphisms (SNPs), NAFLD remission rate, and percent change in intrahepatic triglyceride content as secondary outcomes.

Animal model

All mice were housed in a specific-pathogen-free facility with a 12-h/12-h light/dark cycle and given free access to food and water. All protocols for mouse experiments were approved by the Committee on the Use of Live Animals for Teaching and Research of the University of Hong Kong (CULATR No. 4361-17). All relevant ethical regulations were complied. Mice involved in all experiments were given a HFHC diet (D12079B, Research Diets, New Brunswick, NJ, USA) to induce NAFLD. For FMT experiments, 5-week-old male C57BL/6J wild-type mice and 6-week-old male C57BL/6J *ApoE*$^{-/-}$ mice were purchased from GemPharmatech (Nanjing, China). Mice were randomly divided into two groups: RS group and CS group (n =8 per group in the dietary model; n=6 per group in the genetic model). Both were fed the HFHC diet for 11 weeks before fecal microbiota transplantation (diet, water, and bedding were all sterilized). For *B. stercoris* gavage, eight-week-old male C57BL/6J mice were randomly divided into three groups: HFHC with PBS group, HFHC with live *B. stercoris* group, and HFHC with heat-killed *B. stercoris* group, and were treated for 8 weeks. Body composition was assessed with a Minispec LF90 body composition analyzer (Bruker, Billerica, MA, USA) after 8 weeks of feeding. Glucose and insulin levels in both fasting and fed status, intraperitoneal glucose tolerance test, and insulin tolerance tests were performed after 8 weeks of daily gavage [59].

Culture and administration of *B. stercoris*

B. stercoris (catalog No. 19555, DSMZ-German Collection of Microorganisms and Cell Cultures GmbH, Germany) was cultured in chopped meat medium (Hardy Diagnostics, USA) at 37 °C in an anaerobic workstation (Gene Science AG300, China) with a gas mix containing 10% hydrogen, 10% carbon dioxide and 80% nitrogen. The concentration of bacteria was calculated by measuring the absorbance at the wavelength of 600nm. A fresh culture containing 5×10^9 cfu of *B. stercoris* in 200μL PBS was orally gavaged daily to C57BL/6J mice, with sterile PBS as control. For one experimental group, *B. stercoris* was heat killed at 121℃ under 225-kPa pressure for 15 min.

Cell culture and treatment with valine

Human hepatocellular carcinoma cells from the HepG2 cell line (ATCC, Manassas, VA, USA) were cultured in Dulbecco's Modified Eagle Medium (Gibco, NY, USA) with 10% fetal bovine serum (Gibco, NY, USA). This cell model demonstrated comparable results to primary human hepatocytes in terms of lipid accumulation [60], which was the scope of our *in vitro* study. Cells were kept in a 37°C incubator with 5% CO_2. Culture medium was replaced every 2–3 days, and cells were sub-cultured upon reaching 80% confluence.

L-valine (TCI Chemicals, Portland, OR, USA) was dissolved in Milli-Q water to form a stock solution and diluted with serum-free medium to working concentration (30-750 μM). The concentration range for cell experiments was determined by the serum concentration range from human clinical samples and preliminary cytotoxicity assays. Sodium oleate (Sigma-Aldrich, St. Louis, MO, USA) was dissolved in water and sodium palmitate (Sigma-Aldrich, St. Louis, MO, USA) in methanol for stock solutions. The stock solution was further diluted with serum-free medium supplemented with 1% FA-free bovine serum albumin (Gibco, NY, USA) to working concentration 1000 μM. FA-BSA complex was prepared fresh before treatment.

For experiments, cells were plated on 6-well plates at 1 x 10^6 cells/well and allowed to adhere overnight. Cells were incubated with valine for 24 hours followed by fatty acid incubation for another 24 hours. Cells were then collected for further assays.

METHOD DETAILS

Study design

The study procedure has been detailed in the study protocol (**Methods S1**). Briefly, the study was a randomized, double-blinded, placebo-controlled trial conducted at Shanghai Jiao Tong University Affiliated Sixth People's Hospital from 2016-March to 2017-October. Participants were randomized into two groups with an allocation ratio of 1:1 and consumed either HAM-RS2 (Ingredion Inc., Bridgewater, NJ, USA) at 255.4 kcal/day (2.8 kcal/g, 91.2 g, containing 40 g RS) or matched CS (Ingredion Inc., USA) at 255.6 kcal/day (3.55 kcal/g, 72 g, containing 0 g RS) for 4 months (120 days). RS and CS were packaged in sealed bags that were identical in appearance. During the entire trial, participants received dietary and lifestyle counselling. All were engaged in light physical labor or had a sedentary lifestyle, and were advised to keep their usual physical activity habits. Dietary counseling was conducted by a trained dietitian. Standard menus with targeted dietary caloric restrictions and macronutrient intake designed by the dietitian from the Department of Clinical Nutrition, Shanghai Jiao Tong University Affiliated Sixth People's Hospital were provided to participants, as well as the oilcan and scale. Participants were asked to fill in three consecutive 24-hour dietary records (2 weekdays and 1 weekend day) at each visit period and were encouraged to weigh foods to ensure they accurately reported their caloric intake. In each visit, participants were met with a nutritionist individually for assessment of their adherence to both the diet and the starches (adherence to diet was evaluated as whether the total energy intake according to the 24-h dietary recalls met the requirement of diet management; adherence to starch was evaluated by counting the empty packaging bags of starch participants returned at each visit).

At each visit, participants came to the Department of Endocrinology and Metabolism in the morning for collection of blood, urine and stool samples, and for the measurement of anthropometric and biochemical indexes. Abdominal magnetic resonance imaging (MRI) scan and MRS were conducted at V1 and V5, whereas meal tolerance tests were conducted at V1, V3 and V5. The primary outcome was the change in IHTC evaluated by MRS. Secondary outcomes were changes in anthropometric indexes, body composition, body fat analysis by MRI, glycemic control, insulin sensitivity, liver and renal function, lipid profiles, measurement of serum biomarkers, and other tests.

We recorded the combined medication during the follow-up visits and no gastrointestinal drugs such as antacid were used. No serious adverse events were reported throughout the study. Other potential intervention-related adverse events, including constipation (8 participants in RS and 15 participants in CS, $P = 0.108$) and flatulence (20 participants in RS and 19 participants in CS, $P = 0.914$), were equally distributed between the two groups, except the intestinal exhaust (35 participants in RS compared with 8 in CS, $P < 0.001$).

Anthropometric and biochemical measurements

Blood pressure, body weight, height, waist circumference, and biomedical indices were measured according to the study protocol (**Methods S1**). BMI (weight [kg]/ height^2 [m^2]) was also calculated. Blood samples were collected from participants after an overnight fast of at least 10 hours and were used to measure serum ALT, AST, GGT, TG, TC, HDL-C, LDL-C, and non-esterified FA (NEFA). To assess the glucose metabolism, serial blood samples were taken in a fasting state and at postprandial time points for laboratory tests of plasma glucose, insulin and c-peptide after a standardized meal tolerance test (85 g of non-fried instant noodles without soup: 376.98 kcal including 68.6 g carbohydrate, 9.4 g protein and 6.8 g fat) (China Oil & Foodstuffs Corporation, China). Insulin resistance indexes were calculated as follows: HOMA-IR = FPG (mmol/L) × FINS (mU/L)/22.5; Adipo-IR = fasted insulin (mmol/L) × fasted NEFA (pmol/L).

MRS examination

Participants underwent liver MRS using the 3.0-T Philips Ingenia medical system (Philips Healthcare, The Netherlands). Sagittal, coronal, and axial slices through the right lobe of the liver were acquired, and regions of interest were selected by an experienced radiologist, who avoided visible blood vessels and bile ducts. IHTC was measured in a single voxel ($2 \times 2 \times 2$ cm^3) and calculated by dividing the integral of the methylene groups in fatty acid chains of the hepatic triglyceride by the sum of methylene groups and water. The experienced radiologists who performed the test were blinded to the clinical data.

MRI examination

Levels of SFA and VFA were determined by MRI using a 3.0-T Philips Ingenia medical system (Philips Healthcare, The Netherlands) with spin echo sequences: 500/20 (TR/TE) and matrix size = 256 × 25,659. Scan time was approximately 180 seconds. MRI scans were obtained at the abdominal level between L4 and L5 vertebrae in the prone position. Analysis of images was performed on a workstation provided by the manufacturer. MRI was performed by experienced radiologists who were blinded to clinical presentation and laboratory findings. Acquired images underwent measurement of SFA and VFA using a semiautomated segmentation method. According to the signal intensity of adipose tissue, SFA and VFA outlines were manually traced with a graphic user interface. The area inside the outline was automatically labelled and calculated by the software SliceOmatic (Version 5.0, TomoVision, Canada).

Diagnostic criteria for NAFLD

We followed guidelines for the assessment and management of NAFLD in the Asia-Pacific region. For all participants, NAFLD was diagnosed by B ultrasonography (detailed in study protocol), ruling out secondary causes of hepatic fat accumulation including acute infectious disease, biliary obstructive diseases, alcohol abuse (more than 20 g per day for men and more than 10 g per day for women), acute or chronic cholecystitis, acute or chronic viral hepatitis.

Measurement of FGF21 and cytokeratin 18 M65ED

Concentration of FGF21 in human serum was quantified using an enzyme-linked immunosorbent assay (ELISA) kit from Antibody and Immunoassay Services, the University of Hong Kong (AIS, HKU, China). Human serum cytokeratin 18 (CK18) M65ED concentration was quantified with the M65 EpiDeath ELISA kit (Peviva AB, Bromma, Sweden). Intra-assay variations for the measurement of FGF21 and CK18 M65ED were 1.89% and 0.77%, respectively, and for inter-assay variations, these values were 4.08% and 8.23%.

Measurement of LPS and pro-inflammatory factors

Human serum LPS was measured by the Limulus Amebocyte Lysate assay (Hycult Biotech, The Netherlands). Concentration of pro-inflammatory factors including IL6, IL1β, TNFα, and MCP1 were quantified with ELISA kit (Invitrogen, USA). Intra-assay variations and inter-assay variations for the measurements were all below 10%.

Targeted metabolomics analysis of human fecal bile acids, serum bile acids and amino acids

Sample pre-treatment

For fecal samples, about 20-30 mg freeze-dried sample was added to 2 mL Eppendorf tubes. One mL ethanol solution containing internal standards (CA-d5 0.3 µg/mL, CDCA-d4 0.9 µg/mL, GCA-d5 0.6 µg/mL, GCDCA-d4 0.6 µg/mL, TCA-d5 0.3 µg/mL, TDCA-d5 0.3 µg/mL) was added and vortexed. Subsequently, samples were ground with zirconia beads (30 Hz, 1 min). After centrifugation (14,000 x g, 4°C for 10 min), 800 µL supernatant was transferred for freeze-drying. Samples were then dissolved in 800 µL aqueous solution containing 25% acetonitrile and filtered through a 0.22 µm filter membrane. For serum samples*,* 50 µL sample was fully mixed with 200 µL acetonitrile solution containing internal standards (CA-d5 0.1 µg/mL, CDCA-d4 0.3 µg/mL, GCA-d5 0.2 µg/mL, GCDCA-d4 0.2 µg/mL, GCDCS-d5 0.2 µg/mL, TCA-d5 0.1 µg/mL, TCDCA-d5 0.1 µg/mL, TDCA-d5 0.1 µg/mL, alanine-d3 3 µg/mL, phe-d5 3 µg/mL, histine-$^{13}C_6$ 1 µg/mL) for protein precipitation and metabolite extraction. Supernatants were pipetted for freeze-drying. Finally, the powder was dissolved in 70 µL aqueous solution containing 25% acetonitrile at 70 µL and 50 µL redissolved solution was transferred into sample bottles. Another 10 µL was used for freeze-drying for AA analysis.

Using the above 10 µL freeze-dried sample, derivative reactions were performed with AccQTag derivatization kits (Waters, USA) before AA liquid chromatography (LC)-MS analysis. Derivative reactions were performed according to the protocol and briefly described as follows: 70 µL AccQ·Tag ^{TM} ultra-borate buffer (pH 8.8) was added to freeze-dried samples and mixed for 30 seconds and 20 µL AccQ·Tag ^{TM} derivative reagent was added after 10 seconds of vortex. The mixture was kept at room temperature for 1 min and heated for 10 min at 55°C for derivatization reaction.

LC-MS analysis

For both BA and AA profiling, a high-performance liquid chromatograph Nexera X2 (Shimadzu, Japan) and triple quadrupole mass spectrometer (MS) 8050 (Shimadzu, Japan) system equipped with electron spray ionization (ESI) ion source was employed. The main MS parameters were: nebulizing gas flow at 3 L/min, heating gas flow at 10 L/min, interface temperature at 300°C, DL temperature at 250°C, heat block temperature at 400°C, drying gas flow at 10 L/min. Multiple reaction monitoring (MRM) was used to detect BAs and AAs. ACQUITY UPLC C18 columns (100 mm × 2.1 mm, 1.7 µm) were used for chromatograph separation.

Elution conditions for BA analysis were: Mobile phase A was 10 mM ammonium bicarbonate aqueous solution and mobile phase B was pure acetonitrile. The gradient started from 25% B and was maintained for 0.5 minutes, then linearly increased to 40% B in 12.5 minutes and 90% B in another 1 minute. The gradient was maintained at 90% B for 3 minutes, returning to 25% B in 0.5 minutes. The initial pre-equilibrium time was 2.5 minutes. Column temperature was 35°C. Flow rate was 0.35 mL/min. Injection volume was 5 µL.

Elution conditions for AA analysis were: The gradient started from 1% B, was maintained for 1.08 min, and increased to 9.1% B in 10.4 min. At 16.3 min, the gradient was linearly increased to 21.2% B, then quickly to 59.6% B in 0.6 min, and maintained for 1.2 min. The gradient was returned to 1% B in 0.18 min and maintained for 3.72 min for initial pre-equilibrium. Column temperature was 55°C. Flow rate was 0.35 mL/min. Injection volume was 0.1 µL.

Targeted metabolomics analysis of fecal SCFAs

Sample processing

About 20 mg feces sample and 200 µL 50% acetonitrile/Milli Q water were mixed in an Eppendorf tube. Samples were ground twice with zirconia bead (30 Hz for 1 min). After centrifugation (14,000 x g, 4°C for 10 min), supernatants were collected and filtered, and an aliquot of 40 µL was transferred into 1.5 mL Eppendorf tubes following addition of 10 µL hexanoic acid -d11 (50 µg/mL in 50% acetonitrile/MilliQ water). After vortexing, 20 µL 3-nitrophenyl hydrazine (200 mM in 50% acetonitrile/MilliQ water) and 20 µL EDC (120 mM in 50% acetonitrile/MilliQ water containing 6% pyridine) were added. Tubes were incubated in a water bath (40°C) for 30 min and placed on ice for 1 min to stop derivative reactions. Before LC-MS analysis, 910 µL 10% acetonitrile/MilliQ water was used for dilution.

LC-MS analysis

Quantitative analysis used an AB SCIEX ExionLC AD UPLC coupled with AB SCIEX triplequadrupole 6500 plus MS (AB SCIEX, Framingham, US). An ESI ion source was used. MRM scan was operated in negative ionization mode. Ion source parameters were capillary temperature 325°C, capillary voltage 49V and sheath gas 40 arb. An ACQUITY UPLC C18 column (100 mm × 2.1 mm, 1.7 µm) was used for separation. Mobile phase A was 0.1% formic acid in MilliQ water. Mobile phase B was 0.1% formic acid in acetonitrile. The total run time was 11 min per sample. The gradient started from 15% B and was increased to 27% B in 4 min, then to 42% B in 4 min, then 100% B in 0.5 min, maintained for 1 min before returning to 15% B in 0.5 min and maintained for 1 min. Column temperature was 40°C. Flow rate was 0.35 mL/min. Injection volume was 5 µL.

Fecal sample collection and DNA extraction

Fecal samples were collected using a commercial tube with DNA stabilizer (STRATEC Molecular, Berlin, Germany) and stored at -80°C. Stool DNA was extracted using PSP Spin Stool DNA Kits (STRATEC Molecular, Berlin, Germany) according to the manufacturer's instructions. Fecal DNA extracts were used to construct shotgun metagenomic libraries using the KAPA soil kit following the standard protocol. The Novaseq 6000 platform was used for 150 bp paired-end sequencing at Novogene, China.

Quality control and taxonomic profiling

For quality control of raw reads, human DNA contamination was removed using BWA *mem* version 0.7.4 [61] against human reference genome ucsc.hg19 and adaptors, low quality reads, bases or PCR duplicates were filtered as previously described [35]. High-quality reads were taxonomically profiled at different taxonomic levels using MetaPhlAn2 [62] version 2.7.7 with default settings, generating taxonomic relative abundances (total sum scaling normalization). For the CAGs-based approach, genes obtained from HUMAnN2 ("Functional profiling" below) were clustered into CAGs and then metagenomic species (MGS, referring to CAGs with >700 genes) as described before [63] using default algorithm options. MGS were assigned a species-level annotation if more than 50% of genes were assigned the same species level taxonomy and if the second-most assigned taxonomy was <10% or unclassified.

Microbial community diversity analysis

The alpha diversity was calculated using the R package *vegan* [64] and *picante*. Statistical comparisons of alpha diversity between groups were by Wilcoxon rank-sum test or signed-rank test using R package *stats*. Beta diversity (Bray-Curtis dissimilarity, weighted UniFrac and generalized UniFrac) was calculated with the R packages *phyloseq* and *GUniFrac*. Statistical comparison between groups was by the function adonis to perform a permutational multivariate analysis using R package *vegan* with 999 permutations. $P < 0.05$ was considered significant.

Functional profiling

Microbial gene families and pathway abundances were determined using HUMAnN2 software [65] version 0.11.2 and the UniRef90 and MetaCyc databases. Gene families were mapped to Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology database included in HUMAnN2 to obtain KEGG modules and KEGG pathway abundances. Gene families were also mapped to level-4 enzyme commission (EC) categories using the EC database included in HUMAnN2. Carbohydrate-Active enZYmes (CAZy) [66] were obtained by annotating ECs to the CAZy database. Tables of pathway and gene family abundance obtained using HUMAnN2 were normalized to copies per million, including unmapped and unintegrated reads.

Integrating microbiome, metabolome and phenotypes

The omics computational framework described before [28] was used to perform a three-way analysis to screen potential KEGG modules that significantly correlated with metabolites and phenotypes, and leave-one-species-out analysis to determine the driver species of KEGG modules and important phenotypes. In the three-way analysis, correlations between the functional potential and phenotypes were by partial Spearman's correlation adjusting for obesity-related parameters. In cases of correlations between microbiota functional potential and metabolites, Spearman's correlation was used. In the leave-one-species-out analysis, we checked for KEGG modules with strong correlations (absolute Spearman's correlation ≥ 0.2) with IHTC or FGF21. Species with > 10% effect on the correlation after removal were deemed driver species.

Validation of *B. stercoris* in external cohorts

Two independent external cohorts of different ethnicity were used. The Chinese cohort included 100 patients with NAFLD (diagnosed with ultrasonography) and 90 NAFLD-free control [29]. The European cohort had different degrees of steatosis confirmed by liver biopsy controls [30], including 32 participants with no or mild steatosis and 24 participants with moderate or severe steatosis. The metagenomic data were processed with the above pipeline for quality control and taxonomic profiling.

FMT Experiment

Mice were treated with antibiotics cocktail (ampicillin 1g/l, neomycin 1g/l, metronidazole 1g/l, vancomycin 0.5g/l) for 7 days for microbiota depletion, followed by a 4-day wash-out period to eliminate antibiotics before fecal microbial transplantation as described previously [30,67]. Two human donors from RS or CS group who were close to the average change of IHTC within RS and CS intervention were selected for fecal microbial transplantations. Approximately 500 mg fresh human stools from each donor were collected in the anaerobic workstation and suspended in 5 ml PBS buffer containing 0.2 g/l Na_2S and 0.5 g/l cysteine. The mixture was homogenized and centrifuged, and the supernatant was collected under a stream of nitrogen. Stool samples from participants were not pooled, and fecal slurry from each donor was transferred into 4 conventional antibiotic-treated mice housed in one cage. The mice were colonized by oral gavage with 200 µl of RS or CS fecal slurry. Mice were treated once daily for three consecutive days by gavage in the first week of colonization, then fecal slurries were introduced every other day to reinforce colonization during the remaining days of the 3 weeks.

Histological examination

Mouse liver samples were resected and fixed with 10% formaldehyde phosphate-buffered saline (pH 7.4), embedded in paraffin, sectioned, stained with hematoxylin/eosin (H&E, Sigma, USA) for morphology and Sirius Red (Abcam, UK) for fibrosis, followed by analysis with a Nikon DS-Ri2 microscope (Nikon Instruments Inc., Melville, USA). For detection of neutral lipids, liver cryosections embedded in OCT were stained with Oil Red O (Sigma, USA). Histology was evaluated by two independent researchers who were blinded to the experimental design and treatment groups according to the NAFLD scoring system as previously reported [68]. In brief, 4 histological features were evaluated semi-quantitatively including steatosis (0-3), lobular inflammation (0-3), hepatocellular ballooning (0-2), and fibrosis (0-4). The unweighted sum of first three features was defined as NAFLD activity score (NAS).

Biochemical assays in mice

Serum levels of AST, ALT, TC, TG, and glucose in mice were measured with commercial kits from Stanbio Laboratory (Boerne, TX, USA) or Nanjing Jiancheng Bioengineering Institute (Nanjing, China). Serum lipopolysaccharide (LPS) was measured by the Limulus Amebocyte Lysate assay (Hycult Biotech, The Netherlands). Insulin level was determined by immunoassay from Immunodiagnostics (AIS, HKU, China). Serum FGF21 in mice were measured by ELISA (AIS, HKU, China).

Quantification of mice hepatic lipids

TG content of livers was determined by a modified Folch method [69]. Briefly, 50 mg of liver tissue was homogenized in chloroform/methanol (2/1; v/v). After extraction at room temperature overnight, the organic phase was used to measure hepatic TG with commercial kits from Stanbio Laboratory (Boerne, TX, USA) or Nanjing Jiancheng Bioengineering Institute (Nanjing, China).

RNA preparation and real-time quantitative polymerase chain reaction

Total RNA from livers or ileum was extracted with TRIzol reagent (Invitrogen, CA, USA), and total RNA from cells was extracted by RNAIso Plus reagent (Takara, Japan) according to the manufacturer's manual. RNA concentration was determined using a NanoDrop ND-1000 Spectrophotometer (Nano-Drop Technologies, Wilmington, DE, USA) and RNA quality was determined by the A260/A280 ratio of 1.8-2.1. RNA integrity was also checked by 1% agarose gel electrophoresis to ensure 2 intact bands of 28S and 18S RNA. Genomic DNA digestion and reverse transcription yielded cDNA from 1 μg RNA template using the HiScript RT SuperMix for qPCR kit (Vazyme Biotech, Nanjing, China) or PrimeScript™ RT reagent Kit (Takara, Japan) according to the manufacturer's manual. Quantitative real-time PCR was performed using SYBR Green master mix on StepOnePlus Real-Time PCR system (Applied Biosystems, Foster City, CA, USA) or Light Cycler 480 system (Roche, USA). The mouse glyceraldehyde-3-phosphate dehydrogenase gene and human beta-actin gene were the reference for tests in mice or cell line, respectively. Relative changes in gene expression were calculated using the $2^{-\triangle\triangle CT}$ method. Primers used for PCR are listed in **Table S5.**

In vitro testing of *B. stercoris* for valine releasing activity

B. stercoris was grown in the same culture medium as mentioned above to stationary phase and was inoculated to fresh medium that had been sterilized by an autoclave. Then the mixture was aliquoted and incubated under anaerobic conditions for 3, 5, 8, 12, 24 hours, respectively. At each time point, an aliquot was removed for centrifuging at 4500rpm (4°C for 15 min) to obtain culture supernatants. Following filtration (pore size 0.22 µm; Millipore, USA), the samples were stored at -80 °C until use. In parallel, fresh medium was aliquoted and incubated under the same conditions for the same period of time as blank control. At each time point, the supernatant was obtained from aliquot centrifuged at 4500rpm (4°C for 15 min) as control.

Experiments seeking to test whether heat-killed *B. stercoris* could produce valine were carried out by first incubating *B. stercoris* in medium at 37 °C under anaerobic conditions to stationary phase. Stationary phase culture was then treated at 121°C under pressure of 225 kPa for 15 minutes. And the culture was added to fresh medium. The mixture was incubated under anaerobic condition for 8h (logarithmic phase of growth of live *B. stercoris*) and then centrifuged to obtain culture supernatants.

All cultures in each condition were performed in triplicate. An ultraperformance liquid chromatography coupled to tandem mass-spectrometry (UPLC-MS/MS) system (ACQUITY UPLC-Xevo TQ-S, Waters, USA) was used to quantitate all targeted metabolites by Metabo-Profile Biotechnology Co., Ltd [70].

Intracellular TG assays

Intracellular TG assays were performed using Triglyceride Colorimetric Assay Kits (Cayman Chemicals, Ann Arbor, MI, USA) according to the manufacturer's manual. Intracellular TG content was normalized to cellular protein quantified using Bradford Protein Assay (Bio-Rad, Hercules, CA, USA).

Data visualization

Circos plots were made using interactive Tree of Life (https://itol.embl.de/). All other figures were generated by R software 3.6.3, using ggplot2 and ComplexHeatmap packages, or by GraphPad Prism 9.0.

QUANTIFICATION AND STATISTICAL ANALYSIS

Clinical and experimental data

For clinical data, analyses were performed with SPSS 25.0 (Chicago, IL, USA) and SAS version 9.4 (SAS Institute, Cary, NC, USA) as 2-sided with a significance level of $\alpha = 0.05$. Analyses were performed mainly in the full analysis set, which included all randomized patients who received at least one dose of study medication and had at least one post-intervention assessment of effectiveness. Numerical variables were expressed as mean (95% CIs). Categorical variables were expressed as percentages. Student's unpaired t-tests and chi-square tests were used for comparison between two groups at baseline. Linear mixed model was used to assess within-group differences. Comparison between the RS and CS groups at each visit was through analysis of covariance with treatment group as a factor and baseline value as a covariate. Differences in outcomes between the RS and CS groups were assessed using a linear mixed model. Fixed effects included baseline values of the assessed variable; treatment group (RS vs. CS as a categorical variable); categorical time points represented by 5 visits at days 0, 30, 60, 90, 120; and the interaction term of visit × treatment group. Repeated measures were added as a random effect. Weight loss was used as an additional fixed effect when the weight loss-independent effect was assessed.

For animal and cell line experiments, all analyses were performed with GraphPad Prism 9.0 (GraphPad Prism, USA). Data were shown as mean ± SEM. Two-tailed Student's unpaired *t* test (normally distributed) or non-parametric Wilcoxon rank-sum test (non-normally distributed) was used for comparisons between two groups. Comparisons among more than two groups were performed using one-way ANOVA (normally distributed) followed by Tukey's post hoc test, or Kruskal-Wallis test (non-normally distributed) followed by Dunn's test.

Metabolomics and metagenomics data

We performed partial least squares discriminant analysis using metabolite fold-change (log2-transformed) profiles with the R package *mixOmics* [71]. Statistical comparison between groups was based on Bray Curtis dissimilarity using the function "adonis" in R package *vegan* with 999 permutations. Taxonomic and metabolite variations were further calculated as the ratio between microbial relative abundance or metabolite abundance at week 16 against abundance at baseline. Log2 transformation was applied to fold-changes. For taxonomic variation, before deriving fold-changes, zero values were additively smoothed by minimal nonzero abundance among all observed measurements. Differentially abundant species, functions and metabolites were identified by two-sided Wilcoxon rank-sum test or Wilcoxon signed-rank test, when appropriate, using R package *stats*. Z scores for metabolites variation were calculated using R package *rcompanion*. Generalized linear models were used to obtain differentially abundant species after adjusting for obesity-related parameters (species ~ group + VFA + SFA + BMI + Waist circumference + FAT%) with *glm* function from R package *stats*. To determine if metabolome and microbiome were associated, a mantel test was performed using the *mantel* function from the R package *vegan*. Bray-Curtis dissimilarity matrices based on the species and metabolite abundance tables were computed to perform this test.

Spearman's correlation analysis was performed using R package *stats*. Partial Spearman correlation adjusting for obesity-related parameters (VFA + SFA + Waist circumference + BMI + FAT%) was performed between metabolites/species and clinical data using the R package *ppcor*. All statistical analyses were performed with R software 3.6.3 and $P < 0.05$ was deemed significant unless otherwise stated. P values were adjusted by an FDR method [72] using R package *stats*.

Data S1. Source data underlying the display items in the manuscript, related to Figures 2–6, S2-S5, and S7.

Table S4. Species contributions to correlations between microbiota functional modules and intracellular hepatic triglyceride content (IHTC) or fibroblast growth factor 21 (FGF21) in resistant starch (RS) and control starch (CS) groups, related to Figure 5.

Methods S1: Study Protocol and statistical analysis plan, related to Figures 1-3, S1-2, and STAR Methods.

REFERENCES

1. Le, M.H., Le, D.M., Baez, T.C., Wu, Y., Ito, T., Lee, E.Y., Lee, K., Stave, C.D., Henry, L., Barnett, S.D., et al. (2023). Global incidence of non-alcoholic fatty liver disease: a systematic review and meta-analysis of 63 studies and 1,201,807 persons. J Hepatol. 10.1016/j.jhep.2023.03.040.
2. Younossi, Z., Tacke, F., Arrese, M., Chander Sharma, B., Mostafa, I., Bugianesi, E., Wai-Sun Wong, V., Yilmaz, Y., George, J., Fan, J., and Vos, M.B. (2019). Global Perspectives on Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis. Hepatology (Baltimore, Md.) *69*, 2672-2682. 10.1002/hep.30251.

3. Adams, L.A., Anstee, Q.M., Tilg, H., and Targher, G. (2017). Non-alcoholic fatty liver disease and its relationship with cardiovascular disease and other extrahepatic diseases. Gut *66*, 1138-1153. 10.1136/gutjnl-2017-313884.
4. Lazarus, J.V., Mark, H.E., Anstee, Q.M., Arab, J.P., Batterham, R.L., Castera, L., Cortez-Pinto, H., Crespo, J., Cusi, K., Dirac, M.A., et al. (2022). Advancing the global public health agenda for NAFLD: a consensus statement. Nature reviews. Gastroenterology & hepatology *19*, 60-78. 10.1038/s41575-021-00523-4.
5. Simon, T.G., Roelstraete, B., Khalili, H., Hagstrom, H., and Ludvigsson, J.F. (2021). Mortality in biopsy-confirmed nonalcoholic fatty liver disease: results from a nationwide cohort. Gut *70*, 1375-1382. 10.1136/gutjnl-2020-322786.
6. Tilg, H., and Targher, G. (2021). NAFLD-related mortality: simple hepatic steatosis is not as 'benign' as thought. Gut *70*, 1212-1213. 10.1136/gutjnl-2020-323188.
7. Neuschwander-Tetri, B.A. (2020). Therapeutic Landscape for NAFLD in 2020. Gastroenterology *158*, 1984-1998.e1983. 10.1053/j.gastro.2020.01.051.
8. Leung, C., Rivera, L., Furness, J.B., and Angus, P.W. (2016). The role of the gut microbiota in NAFLD. Nature reviews. Gastroenterology & hepatology *13*, 412-425. 10.1038/nrgastro.2016.85.
9. Tripathi, A., Debelius, J., Brenner, D.A., Karin, M., Loomba, R., Schnabl, B., and Knight, R. (2018). The gut-liver axis and the intersection with the microbiome. Nat Rev Gastroenterol Hepatol *15*, 397-411. 10.1038/s41575-018-0011-z.
10. Aron-Wisnewsky, J., Warmbrunn, M.V., Nieuwdorp, M., and Clement, K. (2020). Nonalcoholic Fatty Liver Disease: Modulating Gut Microbiota to Improve Severity? Gastroenterology *158*, 1881-1898. 10.1053/j.gastro.2020.01.049.
11. Michalak, L., Gaby, J.C., Lagos, L., La Rosa, S.L., Hvidsten, T.R., Tétard-Jones, C., Willats, W.G.T., Terrapon, N., Lombard, V., Henrissat, B., et al. (2020). Microbiota-directed fibre activates both targeted and secondary metabolic shifts in the distal gut. Nature communications *11*, 5773. 10.1038/s41467-020-19585-0.
12. Y, C., R, F., X, Y., J, D., M, H., X, J., Y, L., AP, O., G, G., JU, O., et al. (2019). Yogurt improves insulin resistance and liver fat in obese women with nonalcoholic fatty liver disease and metabolic syndrome: a randomized controlled trial. The American journal of clinical nutrition *109*, 1611-1619. 10.1093/ajcn/nqy358.
13. Bakhshimoghaddam, F., Shateri, K., Sina, M., Hashemian, M., and Alizadeh, M. (2018). Daily Consumption of Synbiotic Yogurt Decreases Liver Steatosis in Patients with Nonalcoholic Fatty Liver Disease: A Randomized Controlled Clinical Trial. The Journal of nutrition *148*, 1276-1284. 10.1093/jn/nxy088.
14. Bomhof, M.R., Parnell, J.A., Ramay, H.R., Crotty, P., Rioux, K.P., Probert, C.S., Jayakumar, S., Raman, M., and Reimer, R.A. (2019). Histological improvement of non-alcoholic steatohepatitis with a prebiotic: a pilot clinical trial. European journal of nutrition *58*, 1735-1745. 10.1007/s00394-018-1721-2.
15. Schupack, D.A., Mars, R.A.T., Voelker, D.H., Abeykoon, J.P., and Kashyap, P.C. (2022). The promise of the gut microbiome as part of individualized treatment strategies. Nature reviews. Gastroenterology & hepatology *19*, 7-25. 10.1038/s41575-021-00499-1.
16. Chalasani, N., Younossi, Z., Lavine, J.E., Charlton, M., Cusi, K., Rinella, M., Harrison, S.A., Brunt, E.M., and Sanyal, A.J. (2018). The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology (Baltimore, Md.) *67*, 328-357. 10.1002/hep.29367.
17. Mardinoglu, A., Wu, H., Bjornson, E., Zhang, C., Hakkarainen, A., Räsänen, S.M., Lee, S., Mancina, R.M., Bergentall, M., Pietiläinen, K.H., et al. (2018). An Integrated Understanding of the Rapid Metabolic Benefits of a Carbohydrate-Restricted Diet on

Hepatic Steatosis in Humans. Cell metabolism *27*, 559-571.e555. 10.1016/j.cmet.2018.01.005.

18. Chaudhari, S.N., McCurry, M.D., and Devlin, A.S. (2021). Chains of evidence from correlations to causal molecules in microbiome-linked diseases. Nature chemical biology *17*, 1046-1056. 10.1038/s41589-021-00861-z.
19. Zhang, L., Li, H.T., Shen, L., Fang, Q.C., Qian, L.L., and Jia, W.P. (2015). Effect of Dietary Resistant Starch on Prevention and Treatment of Obesity-related Diseases and Its Possible Mechanisms. Biomedical and environmental sciences : BES *28*, 291-297. 10.3967/bes2015.040.
20. Keenan, M.J., Zhou, J., Hegsted, M., Pelkman, C., Durham, H.A., Coulon, D.B., and Martin, R.J. (2015). Role of resistant starch in improving gut health, adiposity, and insulin resistance. Advances in nutrition (Bethesda, Md.) *6*, 198-205. 10.3945/an.114.007419.
21. Barouei, J., Bendiks, Z., Martinic, A., Mishchuk, D., Heeney, D., Hsieh, Y.H., Kieffer, D., Zaragoza, J., Martin, R., Slupsky, C., and Marco, M.L. (2017). Microbiota, metabolome, and immune alterations in obese mice fed a high-fat diet containing type 2 resistant starch. Molecular nutrition & food research *61*. 10.1002/mnfr.201700184.
22. Tachon, S., Zhou, J., Keenan, M., Martin, R., and Marco, M.L. (2013). The intestinal microbiota in aged mice is modulated by dietary resistant starch and correlated with improvements in host responses. FEMS microbiology ecology *83*, 299-309. 10.1111/j.1574-6941.2012.01475.x.
23. Rosado, C.P., Rosa, V.H.C., Martins, B.C., Soares, A.C., Santos, I.B., Monteiro, E.B., Moura-Nunes, N., da Costa, C.A., Mulder, A., and Daleprane, J.B. (2020). Resistant starch from green banana (Musa sp.) attenuates non-alcoholic fat liver accumulation and increases short-chain fatty acids production in high-fat diet-induced obesity in mice. International journal of biological macromolecules *145*, 1066-1072. 10.1016/j.ijbiomac.2019.09.199.
24. Wong, V.W., Adams, L.A., de Lédinghen, V., Wong, G.L., and Sookoian, S. (2018). Noninvasive biomarkers in NAFLD and NASH - current progress and future promise. Nature reviews. Gastroenterology & hepatology *15*, 461-478. 10.1038/s41575-018-0014-9.
25. Gaggini, M., Carli, F., Rosso, C., Buzzigoli, E., Marietti, M., Della Latta, V., Ciociaro, D., Abate, M.L., Gambino, R., Cassader, M., et al. (2018). Altered amino acid concentrations in NAFLD: Impact of obesity and insulin resistance. Hepatology (Baltimore, Md.) *67*, 145-158. 10.1002/hep.29465.
26. Bendiks, Z.A., Knudsen, K.E.B., Keenan, M.J., and Marco, M.L. (2020). Conserved and variable responses of the gut microbiome to resistant starch type 2. Nutr Res *77*, 12-28. 10.1016/j.nutres.2020.02.009.
27. Fukunishi, S., Sujishi, T., Takeshita, A., Ohama, H., Tsuchimoto, Y., Asai, A., Tsuda, Y., and Higuchi, K. (2014). Lipopolysaccharides accelerate hepatic steatosis in the development of nonalcoholic fatty liver disease in Zucker rats. J Clin Biochem Nutr *54*, 39-44. 10.3164/jcbn.13-49.
28. Pedersen, H.K., Forslund, S.K., Gudmundsdottir, V., Petersen, A.O., Hildebrand, F., Hyotylainen, T., Nielsen, T., Hansen, T., Bork, P., Ehrlich, S.D., et al. (2018). A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. Nat Protoc *13*, 2781-2800. 10.1038/s41596-018-0064-z.
29. Leung, H., Long, X., Ni, Y., Qian, L., Nychas, E., Siliceo, S.L., Pohl, D., Hanhineva, K., Liu, Y., Xu, A., et al. (2022). Risk assessment with gut microbiome and metabolite markers in NAFLD development. Sci Transl Med *14*, eabk0855. 10.1126/scitranslmed.abk0855.

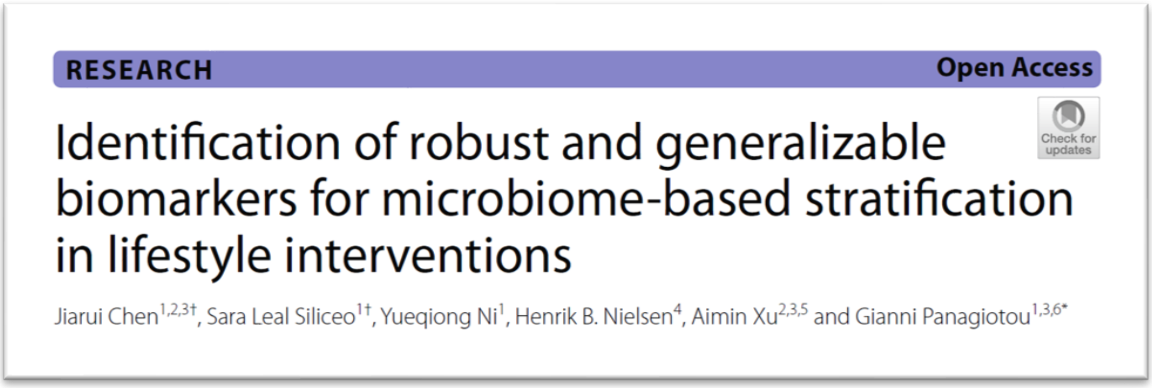
30. Hoyles, L., Fernández-Real, J.-M., Federici, M., Serino, M., Abbott, J., Charpentier, J., Heymes, C., Luque, J.L., Anthony, E., Barton, R.H., et al. (2018). Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. Nat. Med. *24*, 1070-1080. 10.1038/s41591-018-0061-3.
31. Le Roy, T., Llopis, M., Lepage, P., Bruneau, A., Rabot, S., Bevilacqua, C., Martin, P., Philippe, C., Walker, F., Bado, A., et al. (2013). Intestinal microbiota determines development of non-alcoholic fatty liver disease in mice. Gut *62*, 1787-1794. 10.1136/gutjnl-2012-303816.
32. Seo, B., Jeon, K., Moon, S., Lee, K., Kim, W.K., Jeong, H., Cha, K.H., Lim, M.Y., Kang, W., Kweon, M.N., et al. (2020). Roseburia spp. Abundance Associates with Alcohol Consumption in Humans and Its Administration Ameliorates Alcoholic Fatty Liver in Mice. Cell Host Microbe *27*, 25-40 e26. 10.1016/j.chom.2019.11.001.
33. Yuan, J., Chen, C., Cui, J., Lu, J., Yan, C., Wei, X., Zhao, X., Li, N., Li, S., Xue, G., et al. (2019). Fatty Liver Disease Caused by High-Alcohol-Producing Klebsiella pneumoniae. Cell Metab *30*, 675-688.e677. 10.1016/j.cmet.2019.08.018.
34. Albillos, A., de Gottardi, A., and Rescigno, M. (2020). The gut-liver axis in liver disease: Pathophysiological basis for therapy. J Hepatol *72*, 558-577. 10.1016/j.jhep.2019.10.003.
35. Li, J., Sung, C.Y., Lee, N., Ni, Y., Pihlajamaki, J., Panagiotou, G., and El-Nezami, H. (2016). Probiotics modulated gut microbiota suppresses hepatocellular carcinoma growth in mice. Proc Natl Acad Sci U S A *113*, E1306-1315. 10.1073/pnas.1518189113.
36. Wastyk, H.C., Fragiadakis, G.K., Perelman, D., Dahan, D., Merrill, B.D., Yu, F.B., Topf, M., Gonzalez, C.G., Van Treuren, W., Han, S., et al. (2021). Gut-microbiota-targeted diets modulate human immune status. Cell *184*, 4137-4153.e4114. 10.1016/j.cell.2021.06.019.
37. Gehrig, J.L., Venkatesh, S., Chang, H.W., Hibberd, M.C., Kung, V.L., Cheng, J., Chen, R.Y., Subramanian, S., Cowardin, C.A., Meier, M.F., et al. (2019). Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. Science (New York, N.Y.) *365*. 10.1126/science.aau4732.
38. Li, G., Zhang, X., Lin, H., Liang, L.Y., Wong, G.L., and Wong, V.W. (2022). Non-invasive tests of non-alcoholic fatty liver disease. Chin Med J (Engl) *135*, 532-546. 10.1097/CM9.0000000000002027.
39. Castera, L., Friedrich-Rust, M., and Loomba, R. (2019). Noninvasive Assessment of Liver Disease in Patients With Nonalcoholic Fatty Liver Disease. Gastroenterology *156*, 1264-1281 e1264. 10.1053/j.gastro.2018.12.036.
40. Le, T.A., Chen, J., Changchien, C., Peterson, M.R., Kono, Y., Patton, H., Cohen, B.L., Brenner, D., Sirlin, C., and Loomba, R. (2012). Effect of colesevelam on liver fat quantified by magnetic resonance in nonalcoholic steatohepatitis: a randomized controlled trial. Hepatology *56*, 922-932. 10.1002/hep.25731.
41. Wong, V.W., Wong, G.L., Yeung, D.K., Lau, T.K., Chan, C.K., Chim, A.M., Abrigo, J.M., Chan, R.S., Woo, J., Tse, Y.K., et al. (2015). Incidence of non-alcoholic fatty liver disease in Hong Kong: a population study with paired proton-magnetic resonance spectroscopy. J Hepatol *62*, 182-189. 10.1016/j.jhep.2014.08.041.
42. Pedersen, H.K., Gudmundsdottir, V., Nielsen, H.B., Hyotylainen, T., Nielsen, T., Jensen, B.A., Forslund, K., Hildebrand, F., Prifti, E., Falony, G., et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. Nature *535*, 376-381. 10.1038/nature18646.
43. Metallo, C.M., Gameiro, P.A., Bell, E.L., Mattaini, K.R., Yang, J., Hiller, K., Jewell, C.M., Johnson, Z.R., Irvine, D.J., Guarente, L., et al. (2011). Reductive glutamine

metabolism by IDH1 mediates lipogenesis under hypoxia. Nature *481*, 380-384. 10.1038/nature10602.

44. Bar-Peled, L., and Sabatini, D.M. (2014). Regulation of mTORC1 by amino acids. Trends Cell Biol *24*, 400-406. 10.1016/j.tcb.2014.03.003.
45. Jeon, Y.G., Kim, Y.Y., Lee, G., and Kim, J.B. (2023). Physiological and pathological roles of lipogenesis. Nat Metab. 10.1038/s42255-023-00786-y.
46. Loomba, R., Seguritan, V., Li, W., Long, T., Klitgord, N., Bhatt, A., Dulai, P.S., Caussy, C., Bettencourt, R., Highlander, S.K., et al. (2017). Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab *25*, 1054-1062.e1055. 10.1016/j.cmet.2017.04.001.
47. Sanyal, A., Charles, E.D., Neuschwander-Tetri, B.A., Loomba, R., Harrison, S.A., Abdelmalek, M.F., Lawitz, E.J., Halegoua-DeMarzio, D., Kundu, S., Noviello, S., et al. (2019). Pegbelfermin (BMS-986036), a PEGylated fibroblast growth factor 21 analogue, in patients with non-alcoholic steatohepatitis: a randomised, double-blind, placebo-controlled, phase 2a trial. Lancet (London, England) *392*, 2705-2717. 10.1016/s0140-6736(18)31785-9.
48. Geng, L., Lam, K.S.L., and Xu, A. (2020). The therapeutic potential of FGF21 in metabolic diseases: from bench to clinic. Nat Rev Endocrinol *16*, 654-667. 10.1038/s41574-020-0386-0.
49. Li, H., Fang, Q., Gao, F., Fan, J., Zhou, J., Wang, X., Zhang, H., Pan, X., Bao, Y., Xiang, K., et al. (2010). Fibroblast growth factor 21 levels are increased in nonalcoholic fatty liver disease patients and are correlated with hepatic triglyceride. J Hepatol *53*, 934-940. 10.1016/j.jhep.2010.05.018.
50. Li, H., Dong, K., Fang, Q., Hou, X., Zhou, M., Bao, Y., Xiang, K., Xu, A., and Jia, W. (2013). High serum level of fibroblast growth factor 21 is an independent predictor of non-alcoholic fatty liver disease: a 3-year prospective study in China. J Hepatol *58*, 557-563. 10.1016/j.jhep.2012.10.029.
51. Fisher, F.M., Chui, P.C., Antonellis, P.J., Bina, H.A., Kharitonenkov, A., Flier, J.S., and Maratos-Flier, E. (2010). Obesity is a fibroblast growth factor 21 (FGF21)-resistant state. Diabetes *59*, 2781-2789. 10.2337/db10-0193.
52. Tucker, B., Li, H., Long, X., Rye, K.A., and Ong, K.L. (2019). Fibroblast growth factor 21 in non-alcoholic fatty liver disease. Metabolism: clinical and experimental *101*, 153994. 10.1016/j.metabol.2019.153994.
53. Geng, L., Liao, B., Jin, L., Huang, Z., Triggle, C.R., Ding, H., Zhang, J., Huang, Y., Lin, Z., and Xu, A. (2019). Exercise Alleviates Obesity-Induced Metabolic Dysfunction via Enhancing FGF21 Sensitivity in Adipose Tissues. Cell Rep *26*, 2738-2752 e2734. 10.1016/j.celrep.2019.02.014.
54. Noureddin, M., Lam, J., Peterson, M.R., Middleton, M., Hamilton, G., Le, T.A., Bettencourt, R., Changchien, C., Brenner, D.A., Sirlin, C., and Loomba, R. (2013). Utility of magnetic resonance imaging versus histology for quantifying changes in liver fat in nonalcoholic fatty liver disease trials. Hepatology *58*, 1930-1940. 10.1002/hep.26455.
55. Schwimmer, J.B., Ugalde-Nicalo, P., Welsh, J.A., Angeles, J.E., Cordero, M., Harlow, K.E., Alazraki, A., Durelle, J., Knight-Scott, J., Newton, K.P., et al. (2019). Effect of a Low Free Sugar Diet vs Usual Diet on Nonalcoholic Fatty Liver Disease in Adolescent Boys: A Randomized Clinical Trial. JAMA *321*, 256-265. 10.1001/jama.2018.20579.
56. Zhang, H.J., He, J., Pan, L.L., Ma, Z.M., Han, C.K., Chen, C.S., Chen, Z., Han, H.W., Chen, S., Sun, Q., et al. (2016). Effects of Moderate and Vigorous Exercise on Nonalcoholic Fatty Liver Disease: A Randomized Clinical Trial. JAMA Intern Med *176*, 1074-1082. 10.1001/jamainternmed.2016.3202.

57. Tamaki, N., Ajmera, V., and Loomba, R. (2022). Non-invasive methods for imaging hepatic steatosis and their clinical importance in NAFLD. Nat Rev Endocrinol *18*, 55-66. 10.1038/s41574-021-00584-0.
58. Powell, E.E., Wong, V.W., and Rinella, M. (2021). Non-alcoholic fatty liver disease. Lancet *397*, 2212-2224. 10.1016/S0140-6736(20)32511-3.
59. Hui, X., Gu, P., Zhang, J., Nie, T., Pan, Y., Wu, D., Feng, T., Zhong, C., Wang, Y., Lam, K.S., and Xu, A. (2015). Adiponectin Enhances Cold-Induced Browning of Subcutaneous Adipose Tissue via Promoting M2 Macrophage Proliferation. Cell Metab *22*, 279-290. 10.1016/j.cmet.2015.06.004.
60. Gomez-Lechon, M.J., Donato, M.T., Martinez-Romero, A., Jimenez, N., Castell, J.V., and O'Connor, J.E. (2007). A human hepatocellular in vitro model to investigate steatosis. Chem Biol Interact *165*, 106-116. 10.1016/j.cbi.2006.11.004.
61. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN].
62. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods *12*, 902-903. 10.1038/nmeth.3589.
63. Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol *32*, 822-828. 10.1038/nbt.2939.
64. Dixon, P. (2003). VEGAN, A Package of R Functions for Community Ecology. J. Veg. Sci. *14*, 927-930.
65. Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N., and Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods *15*, 962-968. 10.1038/s41592-018-0176-y.
66. Drula, E., Garron, M.L., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res *50*, D571-D577. 10.1093/nar/gkab1045.
67. Liu, Y., Wang, Y., Ni, Y., Cheung, C.K.Y., Lam, K.S.L., Wang, Y., Xia, Z., Ye, D., Guo, J., Tse, M.A., et al. (2020). Gut Microbiome Fermentation Determines the Efficacy of Exercise for Diabetes Prevention. Cell Metab. *31*, 77-91.e75. 10.1016/j.cmet.2019.11.001.
68. Kleiner, D.E., Brunt, E.M., Van Natta, M., Behling, C., Contos, M.J., Cummings, O.W., Ferrell, L.D., Liu, Y.C., Torbenson, M.S., Unalp-Arida, A., et al. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology (Baltimore, Md.) *41*, 1313-1321. 10.1002/hep.20701.
69. Folch, J., Lees, M., and Sloane Stanley, G.H. (1957). A simple method for the isolation and purification of total lipides from animal tissues. The Journal of biological chemistry *226*, 497-509.
70. Xie, G., Wang, L., Chen, T., Zhou, K., Zhang, Z., Li, J., Sun, B., Guo, Y., Wang, X., Wang, Y., et al. (2021). A Metabolite Array Technology for Precision Medicine. Anal Chem *93*, 5709-5717. 10.1021/acs.analchem.0c04686.
71. Rohart, F., Gautier, B., Singh, A., and Le Cao, K.A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol *13*, e1005752. 10.1371/journal.pcbi.1005752.
72. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289-300. 10.2307/2346101.

Manuscript II



RESEARCH Open Access

Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions

Jiarui Chen[1,2,3†], Sara Leal Siliceo[1†], Yueqiong Ni[1], Henrik B. Nielsen[4], Aimin Xu[2,3,5] and Gianni Panagiotou[1,3,6*]

Overview

In manuscript II, we aimed to investigate gut microbiome dynamics in response to lifestyle microbiome-targeted therapies and to identify robust and generalized biomarkers among the gut microbial communities associated with the resistance to change of the gut microbial community. Therefore, we performed longitudinal shotgun metagenomic analysis from a wide range of lifestyle interventions and defined a criterion to classify individuals based on the microbiome response using as a point of departure the natural fluctuation of a healthy microbiome without any intervention. We identified microbial biomarkers of microbiota’s resistance to structural changes, and we found amino acid biosynthesis as an important regulator of microbiome dynamics. Lastly, we developed a machine learning model able to predict the gut microbiome resistance to change in response to lifestyle interventions using the baseline microbiome composition.

FORM I

Manuscript No: 2

Manuscript title: Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions

Authors: Jiarui Chen*, **Sara Leal Siliceo***, Yueqiong Ni, Henrik B. Nielsen, Aimin Xu, Gianni Panagiotou

Bibliographic information (if published or accepted for publication: Citation): Chen et al., (2023). Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions. *Microbiome*, 11, 178. https://doi.org/10.1186/s40168-023-01604-z

The candidate is (Please tick the appropriate box)**:**

☐ First author, ☒ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status (if not published; "submitted for publication", "in preparation".): published

Authors' contributions (in %) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Chen, J.*	30%	50%		35%	
Leal Siliceo, S.*	30%	50%		35%	
Ni, Y.	10%				
Nielsen, H.B.	5%				
Xu, A.	5%				50%
Panagiotou, G.	20%			30%	50%
Total:	100%	100%	NA	100%	100%

*Authors contributed equally

Chen *et al. Microbiome* (2023) 11:178
https://doi.org/10.1186/s40168-023-01604-z

Microbiome

RESEARCH **Open Access**

Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions

Jiarui Chen[1,2,3†], Sara Leal Siliceo[1†], Yueqiong Ni[1], Henrik B. Nielsen[4], Aimin Xu[2,3,5] and Gianni Panagiotou[1,3,6*]

Abstract

Background A growing body of evidence suggests that the gut microbiota is strongly linked to general human health. Microbiome-directed interventions, such as diet and exercise, are acknowledged as a viable and achievable strategy for preventing disorders and improving human health. However, due to the significant inter-individual diversity of the gut microbiota between subjects, lifestyle recommendations are expected to have distinct and highly variable impacts to the microbiome structure.

Results Here, through a large-scale meta-analysis including 1448 shotgun metagenomics samples obtained longitudinally from 396 individuals during lifestyle studies, we revealed *Bacteroides stercoris*, *Prevotella copri*, and *Bacteroides vulgatus* as biomarkers of microbiota's resistance to structural changes, and aromatic and non-aromatic amino acid biosynthesis as important regulator of microbiome dynamics. We established criteria for distinguishing between significant compositional changes from normal microbiota fluctuation and classified individuals based on their level of response. We further developed a machine learning model for predicting "responders" and "non-responders" independently of the type of intervention with an area under the curve of up to 0.86 in external validation cohorts of different ethnicities.

Conclusions We propose here that microbiome-based stratification is possible for identifying individuals with highly plastic or highly resistant microbial structures. Identifying subjects that will not respond to generalized lifestyle therapeutic interventions targeting the restructuring of gut microbiota is important to ensure that primary end-points of clinical studies are reached.

Keywords Gut microbiome, Microbiome dynamics, Resistance, Lifestyle intervention, Machine learning

---

†Jiarui Chen and Sara Leal Siliceo contributed equally.

*Correspondence:
Gianni Panagiotou
Gianni.Panagiotou@leibniz-hki.de
[1] Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute -Microbiome Dynamics, Jena, Germany
[2] State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong S.A.R., China
[3] Department of Medicine, The University of Hong Kong, Hong Kong S.A.R., China
[4] Clinical Microbiomics, Fruebjergvej 3, 2100 Copenhagen, Denmark
[5] Department of Pharmacology and Pharmacy, The University of Hong Kong, Hong Kong S.A.R., China
[6] Faculty of Biological Sciences, Friedrich Schiller University, Jena, Germany

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The human gut microbiome is a complex ecosystem made up of trillions of bacteria, viruses, archaea, and eukaryotic microbes contributing to essential functions in the host. Emerging studies have shown the close connection between the gut microbiome and human health and disease [1], such as influencing host nutrition and metabolism, training and modulating immune function, and contributing to patterns of brain development and behavior. Gut microbiota dysbiosis has been associated with several highly prevalent chronic diseases including gastrointestinal and neurological [2–4] disorders, metabolic diseases, as well as cardiovascular and respiratory illnesses [5–7]. Therefore, targeting the gut microbiome seems to be a promising strategy for restoring balance in the gut in order to improve the host's health. However, unhealthy gut microbiota states can result in a recurring susceptibility to chronic illnesses and resistance to treatment efficacy [8].

Lifestyle interventions targeting the gut microbiota have been explored as a therapeutic treatment for numerous diseases. For example, prebiotic consumption, diet, and exercise have been associated with alterations in the gut microbiota structure and a positive impact on the host's phenotype [9–11]. In most trials, large inter-individual differences in the treatment response have been observed [8], and some of these differences may depend on subject-specific microbiome response to the perturbation. In most cases, the microbiome response is currently unpredicted. Consequently, gut microbiota stability, resilience, and resistance are crucial ecological features [12]. Therefore, it is urgent to understand the potential mechanisms involved in gut microbiome resistance that may govern the response to perturbations and to determine whether lifestyle interventions can shape gut microbiota composition towards resilient healthy states.

In order to shed light on the resistance potential presented by an individual gut microbial ecosystem, we performed a large-scale meta-analysis of metagenomics samples obtained from longitudinal lifestyle interventions and compared the responses with no-intervention and antibiotic treatment studies. Groups of "responders" and "non-responders" were defined by their magnitude of taxonomic changes to a diverse set of lifestyle interventions and characterized by distinct gut microbiota compositions and functional profiles. From a clinical and translational perspective, the ability to predict microbiome resistance to perturbation offers significant advantages to further optimize disease therapies through microbiome-informed patients' stratification and possibly restore plasticity in patients with resilient dysbiosis microbiomes.

Results

The extent of microbiome compositional changes depends on the environmental stimuli and varies between individuals

In order to elucidate the compositional and functional characteristics of the gut microbiome that may predict the personalized responses of the microbial communities to lifestyle, we collected metagenomic shotgun sequencing data from 10 studies covering 467 subjects sampled longitudinally (1590 total in total) (Table 1).

These included 1118 samples from subjects that did not undergo intervention. This allowed us to set a "response threshold" to differentiate between microbiome changes that could simply be considered as natural fluctuation, and significant alterations following various interventions. We also retrieved five cohorts with lifestyle-based treatment (165 subjects, 330 samples), including a low-carbohydrate diet with increased protein content (I_LCD); a high-fiber diet (I_HFD); a highly resistant starch type II (HRS); a multidisciplinary weight-loss program (I_MWP); and an exercise training program (I_ETP) (Table 1). Moreover, the dataset contains four cohorts with different antibiotic treatments (71 subjects, 142 samples): a cocktail of meropenem, gentamicin, and vancomycin (referred to from now on as A_MER–GEN–VAN); cefprozil (A_CEF); ciprofloxacin (A_CIP); and cotrimoxazole (A_COT). Taxonomic and functional profiling was performed with all samples from different cohorts simultaneously after passing through the quality control. Intraclass correlation coefficient (ICC) is a measure of reliability or reproducibility that can be used to quantify the biological variability of the microbiome structure, previously used by Sinha et al. [22] to compute the microbiome temporal stability. The genus-level ICC was calculated for different estimates of alpha (Shannon, Simpson, and Chao1 Index) and beta diversity (using the top principal coordinates analysis (PCoA) scores based on Bray–Curtis dissimilarity, and unweighted or weighted UniFrac distances) for every cohort in our study (Table S1). ICCs range from 0 (no stability) to 1 (perfect stability), where values below 0.5 indicate poor microbiome stability and above 0.5, high microbiome stability [23].

We observed significantly higher mean ICCs values of Shannon and Simpson diversities for the two no-intervention cohorts (that did not include any intervention) compared to the four cohorts treated with antibiotics as well as the five cohorts with lifestyle interventions (Student t test, $p < 0.05$, Fig. 1a). The average ICC values of the two no-intervention cohorts remained high (> 0.50) for all diversity indexes, suggesting a stable gut microbiome alpha diversity in the

Table 1 Description of the study cohorts used in the meta-analysis

Study	Disease	Intervention	Intervention information	Duration (days)	Number subjects/ samples
Mehta et al., 2018 [13]	Healthy	No intervention	-	-	140/560
Poyet et al., 2019 [14]	Healthy	No intervention	-	-	91/558
Palleja et al., 2018 [15]	Healthy	Antibiotics	Meropenem, Gentamicin, and Vancomycin	4	12/24
Raymond et al., 2015 [16]	Healthy	Antibiotics	Cefprozil	7	18/36
Willmann et al., 2019 [17]	Hematological-Oncological disease	Antibiotics	Ciprofloxacin	6	20/40
			Cotrimoxazole		21/42
Louis et al., 2016 [18]	Obesity	Exercise/Dietary	Multidisciplinary weight-loss program (OPTI-FAST® 52, Nestlé Inc.): psychology, medicine, dietetics (very low-calorie diet), and exercise	84	14/28
Mardinoglu et al., 2018 [19]	Obesity with NAFLD	Dietary	Low-carbohydrate diet with increased protein content	14	10/20
Zhao et al., 2018 [20]	T2D	Dietary	High fiber diet composed of whole grains, traditional Chinese medicinal foods, and prebiotics	84	71/142
Ni et al., (in press) [21]	NAFLD	Dietary	Diet with high resistant starch type II content	120	50/100
Liu et al., 2020 [9]	Prediabetes	Exercise	Exercise activity 3 days/week as a combined aerobic and strength training program	84	20/40

absence of external disturbances. Interestingly, there is no significant difference in mean ICCs values of any alpha diversity index when comparing the four cohorts treated with antibiotics and the five cohorts undergoing lifestyle interventions (Student *t* test, $p = 0.14$, 0.240, 0.052 for Shannon index, Simpson index, and Chao1 index, respectively, Fig. 1a). Moreover, despite a clear trend of decreased ICCs for all beta diversity indexes in the 4 antibiotics cohorts compared to the two no-intervention cohorts, the result was not statistically significant (Student *t* test, $p \geq 0.05$) except PCoA1 of Weighted Unifrac index and the average ICC value of PCoA1-5 of Unweighted Unifrac index. These results are probably due to the high variability observed between antibiotic types and personalized responses to each antibiotic. Nevertheless, the overall diversity ICC values of the cohort treated with a combination of meropenem, gentamicin, and vancomycin (A_MER–GEN–VAN) were extremely low with an average of 0.183, indicating a severe disturbance of the microbiome structure (Fig. 1a), while the ICC values of cohorts treated with cefprozil or cotrimoxazole were significantly higher than those of A_MER–GEN–VAN (paired *t*-test, adjusted $p < 0.1$) with an average of 0.453 and 0.368, respectively. On the contrary, the differences in the ICC values for beta diversity between no and lifestyle interventions were less obvious and again they were characterized by high variability among different types of intervention and of participants' responses in each study group (Fig. 1a). By comparing the differences in the ICC values among the lifestyle intervention cohorts, we found that the I_MWP study, which used a multidisciplinary weight-loss program combining psychology, medicine, dietetics, and exercise, had average ICC values of 0.173 and 0.237, for alpha and beta diversity, respectively, significantly lower compared to all other single interventions (either dietetics or exercise) (paired *t*-test, adjusted $p < 0.1$). The comparisons among other cohorts in the lifestyle intervention category showed no significant differences (paired *t*-test, adjusted $p \geq 0.1$). We further compared the beta diversity ICC values of the I_MWP with the four antibiotic-treated cohorts and interestingly, we found that its impact on microbial stability was higher than the A_CEF and A_COT studies (paired *t*-test, adjusted $p < 0.1$).

In summary, we generally observed that the microbial stability estimated as ICCs of alpha and beta diversity is disturbed by antibiotics and lifestyle interventions, but the extent depends on the specifics of each environmental stressor. Furthermore, the insignificance of beta dissimilarity between interventions and no intervention, which may be due to the variability of responsiveness among individuals, supports the notion that generic approaches to altering the microbiome structure in an unbalanced state may not bring the desired structural changes. The baseline microbiome could potentially define the magnitude of the response of the community structure to external stimuli, something that we further explored below.

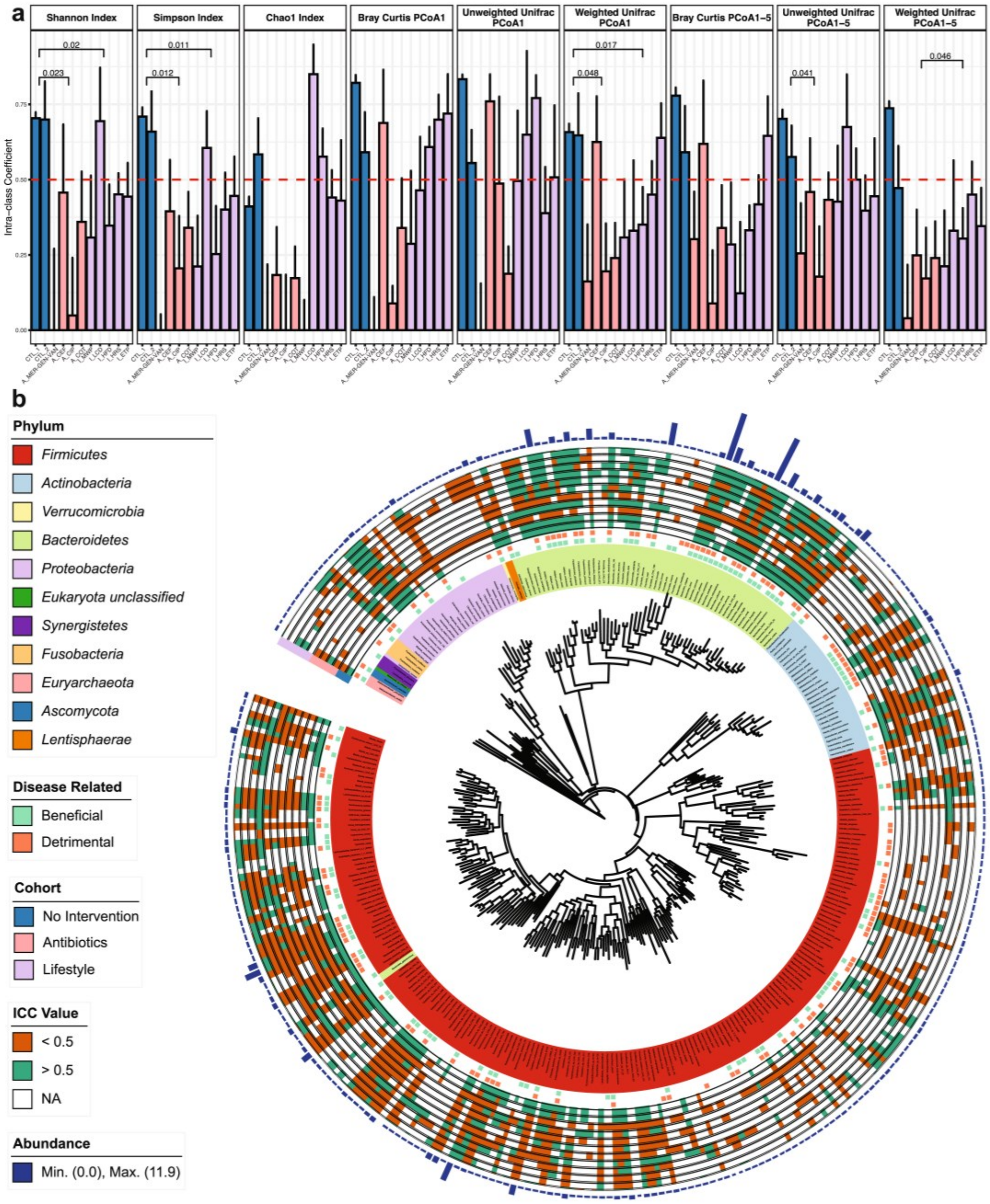


Fig. 1 ICC evaluation of taxonomic profiles among study cohorts. **a** ICC values of alpha and beta diversity indexes at the species level in each study cohort. The error bars represent 95% confidence intervals. Cohort type is indicated by blue, pink, and lilac colors for no intervention, antibiotic intervention, and lifestyle intervention, respectively. Only significant *p* values are shown (Student *t* test, $p < 0.05$). The red dash line indicates an ICC value of 0.5. **b** Circos plot showing the annotated species in our metagenomics datasets in a phylogenetic tree. In the inner circles, disease-related species are shown in light green (beneficial) and light orange (detrimental). Species ICC values are indicated in orange (ICC < 0.5), green (ICC ≥ 0.5), and white (non-valid ICC). Barplots represent the median value of species abundance

Lifestyle interventions could have a comparable impact with antibiotics on individual species' stability

Looking at the ICCs of the 309 individual species annotated in our metagenomics dataset, we observed a clear stratification among the no-intervention cohorts and the other two study groups (Fig. 1b). In the no-intervention cohorts, 79.6% of the detected species were regarded as stable (mean ICCs > 0.5), while this percentage dropped to an average of 27.3 and 43.6% for antibiotics and lifestyle intervention cohorts, respectively. When looking into the individual studies, we observed a similar tendency for the species stability as for the community diversity. The ICCs of 99% of the species in the A_MER–GEN–VAN study were < 0.5 indicating that almost all bacteria present in the microbial community were affected. The percentage of species having ICCs < 0.5 was high for all antibiotics (78.4, 62.3, and 50.3% for A_CIP, A_COT, and A_CEF, respectively). Interestingly, three of the lifestyle interventions had a similar or even higher impact than the administration of single antibiotics on the stability of individual species. The I_MWP intervention resulted in the highest percentage of species with ICCs < 0.5, affecting 71.6% of the community members. The studies using a high-fiber diet (I_HFD) and a high-resistant starch diet (I_HRS) were also characterized by a high percentage of species with ICCs < 0.5 (68.6 and 61.1%, respectively).

Looking for global taxonomic patterns in the lifestyle intervention cohorts, the statistical comparisons among the major phylum showed that the ICCs of *Bacteroidetes* species were significantly higher compared to *Firmicutes* and *Proteobacteria* species (Student *t* test, $p < 0.05$). Using the lifestyle intervention cohorts, we also examined whether the stability of species is correlated with their relative abundance at baseline. However, only 16 out of 309 species showed a significant correlation (Spearman correlation, adjusted $p < 0.05$) between the ICC and relative abundance. By extracting information from the Disbiome Database, we were able to retrieve disease associations for 162 species annotated in our metagenomics datasets. The stability of 115 out of the 162 disease-associated species could be influenced by at least one of the lifestyle interventions. The I_HFD study resulted in ICC values < 0.5 for 83 species associated with a wide range of metabolic diseases (obesity, type 2 diabetes, and hypertension) confirming the potential of a high-fiber diet as a way to target dysbiotic microbiome states. Disease-associated species, whose stability was uniquely influenced by particular lifestyle interventions, were also found. The I_HFD showed specificity towards 12 disease-associated species, whereas I_LCD, I_MWP, and I_HRS showed specificity towards 8, 6, and 3 species, respectively (Table S2).

In order to perform a comparative analysis among all interventions, we extracted 65 species with valid ICCs (not NULL value of ICCs) in every individual study. Out of the 65 species, 47 were stable (ICCs > 0.5) in the no-intervention cohorts; however, all of them lost their highly stable status in at least one study of either antibiotic treatment or lifestyle intervention. Interestingly, among these 47 species, we found 5 species, *Bacteroides massiliensis, Bacteroides stercoris, Barnesiella intestinihominis, Parabacteroides merdae,* and *Parasutterella excrementihominis,* that remained stable in all the lifestyle interventions (ICCs > 0.5). These 5 species further showed resistance (ICCs > 0.5) to two of the antibiotic treatments, cefprozil (A_CEF) and cotrimoxazole (A_COT), suggesting that these species are highly stable. The aforementioned 5 species have a conditional effect on human metabolic diseases, playing either beneficial (non-alcoholic fatty liver disease, cirrhosis, multiple sclerosis, etc.) or detrimental (autism, Parkinson's disease, colon polyps, etc.) roles (Fig. 1b) [24]. We further explored the relative abundance of these 5 species across all samples and observed that they were all low-abundant species (< 0.92%), further confirming that stability and abundance are not correlated. *Alistipes indistinctus*, a species that has been associated with hypertension and autism, also showed an interesting stability pattern. *A. indistinctus* had a high prevalence of 32% but a low abundance of 0.25% on average. *A. indistinctus* showed high stability (ICC > 0.5) not only in the no-intervention cohorts but also in the cohorts with antibiotic treatments (except where the subjects were administered a cocktail dose of antibiotics, A_MER–GEN–VAN). Interestingly, a low-carbohydrate diet (I_LCD) and exercise (I_ETP) could result in an ICC < 0.5 for *A. indistinctus*, suggesting the potential of using specific lifestyle interventions to target highly stable and disease-associated species.

In summary, by evaluating the ICC value of each species across studies, we have identified both species that are highly resistant to any lifestyle and antibiotics intervention and species whose stability pattern can only be affected by specific lifestyle interventions. Interestingly, we also observed that lifestyle interventions can reach similar or even higher capability to impact the stability of microbial species as single antibiotics administration, questioning the broad characterization of antibiotics treatment as a more intense intervention compared to lifestyle interventions.

Identification of species associated with microbiome responsiveness

By calculating the day-to-day Bray–Curtis dissimilarities of each subject from the two longitudinal no-intervention cohorts, we established the criteria to differentiate effective response to a microbiome-targeted intervention from normal fluctuation of the microbial community composition. We used the mean + SD (68% population) and mean + 2*SD (95% population) of the Bray–Curtis dissimilarity (see "Methods" for details) as the two cutoffs to distinguish individual responses and formed the following groups for downstream analysis: (i) non-responders (< mean + SD), (ii) partial-responders ([mean + SD, mean + 2SD]), and (iii) responders (> mean + 2SD) as shown in Fig. 2a. By evaluating the dissimilarity before and after intervention of each subject among the five lifestyle intervention cohorts, 47.3% of individuals were classified as responders, while 24.2% were partial-responders, and the remaining 28.5% were grouped as non-responders. We calculated the species ICCs before and after intervention in each study and compared them based on the responder classification. The species ICCs were significantly lower in the responders compared to the non-responders (paired t test, $p < 0.05$), confirming the grouping.

We subsequently used the baseline microbiome samples of each subject to perform a principal coordinates analysis (PCoA) based on the Bray–Curtis dissimilarities. The microbiome composition of the subjects grouped by the newly constructed classification from non-responder to responder was significantly different (PERMANOVA, $p < 0.05$, $R^2 = 0.05$, Fig. 2b). By comparing the species abundances between the non-responder, partial-responder, and responder groups, we found 41 species with significant differences among the groups (ordinal logistic regression, adjusted $p < 0.2$, Fig. 2c, Table S6). Interestingly, 37 out of the 41 species from the ordinal regression are highly stable species in the absence of interventions (ICCs > 0.5 in the no-intervention cohort), an important property for serving as biomarkers of community response. Among these 37 species, only 3 species were significantly enriched in the non-responder group, namely *Bacteroides stercoris*, *Prevotella copri*, and *Bacteroides vulgatus*. These 3 species remained significantly enriched in the non-responders group even when the lifestyle grouped subjects were combined with non-responders, partial-responders, and responders from the antibiotics cohorts. Similarly, 17 out of the 37 species were found significantly enriched in the responders group even when combining the lifestyle with the antibiotics cohorts, including *Collinsella aerofaciens*, *Gordonibacter pamelaeae*, *Ruthenibacterium lactatiformans*, *Turicibacter sanguini, Fusicatenibacter saccharivorans, Dorea longicatena,* and *Eubacterium hallii,* which were highly stable species (ICCs > 0.5).

Biosynthesis of amino acids and their taxonomic contributors as mediators of microbiome dynamic responses

Subsequently, we compared the MetaCyc pathway abundances among the three response groups using their baseline samples and found 116 pathways with significantly different abundances (Ordinal regression, adjusted $p < 0.1$), indicating a clear baseline stratification also at the functional level. Among the 116 pathways, enrichment of 34 was associated with non-responders and 82 with responders (Fig. S1). As observed with the species biomarkers of responsiveness, 97 out of the 116 pathways from the ordinal regression were highly stable in the absence of interventions (ICCs > 0.5 in the no-intervention cohort). We then investigated the contributions of the 41 significant species (Fig. 2c) to the 116 significant pathways using the stratified output of HUMAnN3. At least one of *B. stercoris*, *P. copri,* and *B. vulgatus*, the 3 species significantly enriched in non-responders, was taxonomically linked to 33 out of the 34 pathways enriched in non-responders, and all 3 species were contributing to the abundances of 24 pathways enriched in non-responders. Interestingly, when exploring the 82 pathways enriched in responders, we found only 4 pathways to have contributions from these non-responders associated species. Similarly, the 38 species enriched in responders were found to contribute to 51 out of the 82 pathways enriched in this response group.

In order to further investigate the relationship between species, pathways, and microbiome response, we performed Spearman correlation analysis between the 41 significant species and the 84 significant pathways that they contributed to (Fig. 3a). A consistent

(See figure on next page.)

Fig. 2 Microbiome compositional differences of responders, partial-responders, and non-responders to lifestyle interventions. **a** Bray–Curtis dissimilarity of longitudinal samples in subjects from no and lifestyle interventions. The two longitudinal no-intervention cohorts were combined in the first box. Bray–Curtis indexes dot colors indicate the microbiome response classification group by coral, blue, and green for non-responders, partial-responders, and responders, respectively. The two red dash lines represent the mean + SD and mean + 2*SD of the Bray–Curtis dissimilarities in the no-intervention cohorts as cutoffs to differentiate significant microbiome compositional changes from normal microbiome fluctuation. (CTL: study with no intervention; I_MWP: intervention study with multidisciplinary weight-loss program; I_LCD: intervention study with low-carbohydrate diet; I_HFD: intervention study with high-fiber diet; I_HRS: intervention study with high-resistant starch; I_ETP: intervention study with exercise training program). **b** Principal coordinate analysis of Bray–Curtis dissimilarity in non-responders, partial-responders, and responders to lifestyle interventions. **c** Relative abundances of the significant species using ordinal regression among non-responders, partial-responders, and responders groups ($p < 0.05$)

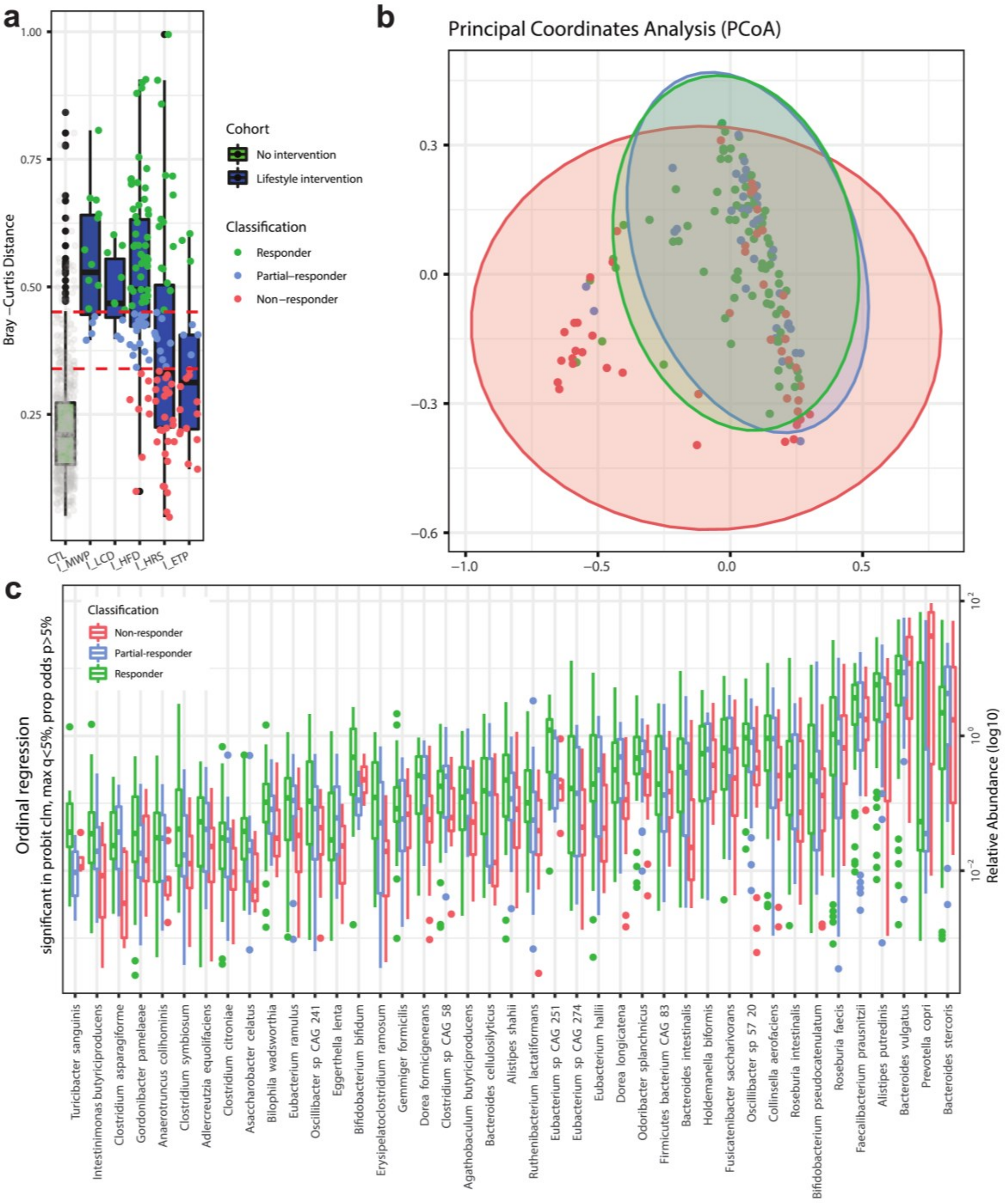


Fig. 2 (See legend on previous page.)

pattern between the 3 species enriched in non-responders and specific functional groups was not observed, besides the significant positive correlations (Spearman, adjusted $p < 0.1$) with fucofuranose biosynthesis, flavin biosynthesis, and its precursors (Fig. 3a). On the contrary, a significantly larger and consistent pattern of positive associations between species and pathways was observed in the responders' enriched taxonomic and functional signatures. Some of the strongest positive correlations were observed between the responders' enriched species, including *C. aerofaciens*, *F. saccharivorans*, *E. hallii*, *Gemmiger formicilis*, and *G. pamelaeae*, and several pathways related to the biosynthesis of amino acids, e.g., arginine, isoleucine, and ornithine biosynthesis, among others (Fig. 3a). Metabolic cross-feeding of the aforementioned biosynthetically costly amino acids has been shown to promote stronger cooperative microbial interactions and drastically impact the community dynamics [25].

We subsequently performed differential abundance comparisons between responders and non-responders for the 2768 detected KEGG Orthology (KOs) and found 395 as significantly different (Wilcoxon rank sum test, adjusted $p<0.1$). Among them, only 11 were more abundant in non-responders, whereas the remaining 384 KOs were highly abundant in responders (Fig. 3b). By mapping the significant KOs to the KEGG pathway database, the biosynthesis of secondary metabolites and biosynthesis of amino acids were two of the pathways with the highest KO contribution in responders, while very limited results were obtained for the non-responders (Fig. 3b). A significant enrichment of KOs related to the biosynthesis of amino acids in the responders compared to the non-responders was also observed (chi-square test, $p<0.01$) with 43 KOs found to be significantly higher in responders, whereas none were higher in non-responders. Moreover, we investigated the species contributions to these amino acid-related KOs that were significantly enriched in the responders. We pinpointed 10 species that were top contributors to multiple significant KOs including *Faecalibacterium prausnitzii*, *Bacteroides cellulosilyticus*, *Fusicatenibacter saccharivorans*, *Eubacterium ramulus*, and *Eubacterium hallii* (Fig. 3c).

We subsequently built species co-abundance networks for responders and non-responders using the baseline samples, in order to further investigate the mechanisms by which responders' enriched species regulate microbial community structural changes. We explored two commonly used centrality measures that reflect the flow of information in the network, the degree and closeness centrality. Responders have a more interconnected community network (Student *t* test, $P<0.001$; Fig. 3d) and higher closeness centrality compared to non-responders (Student *t* test, $P<0.001$; Fig. 3d). When the responder network was investigated in detail, we observed that 11 out of the 15 amino acid auxotroph (AAA) species identified recently in the study of Yu et al. [26] were present in the community network. These 11 AAA species had positive interactions with 30 species found from the ordinal regression to be highly abundant in responders (Fig. S2). Lastly, by integrating the species co-abundance network with the amino acid KO profile, we identified 6 significantly enriched species in responders (*C. symbiosum*, *B. cellulosilyticus*, *G. formicilis*, *F. saccharivorans*, *E. ramulus*, *D. longicatena*, and *E. hallii*) contributing to 22 amino acid-related KOs, which were further positively correlated with 6 AAA species (*R. gnavus*, *B. wadsworthia*, *B. adolescentis*, *D. formicigenerans*, *C. aerofaciens*, and *E. eligens*) (Fig. 3e).

In summary, our analysis revealed signature species in responders and non-responders that could serve as biomarkers of microbiome's resistance to lifestyle interventions. Furthermore, the functional capacity of enriched species in responders suggest that amino acid biosynthesis is playing an important role in regulating microbiome dynamics.

Development of a machine learning model to predict microbiome responsiveness

We then explored whether a machine learning (ML) model can be developed for predicting the degree of responsiveness of a microbiome community to lifestyle interventions. We used the abundance of bacterial species, genera, and pathways from the baseline samples among the cohorts with lifestyle intervention as features for training the model. We used the baseline samples of subjects classified above as responders ($N=78$) and non-responders ($N=47$, Table S8). We built a total of four different gradient boosting machine (gbm)

(See figure on next page.)

Fig. 3 Microbiome functional differences of responders, partial-responders, and non-responders to lifestyle interventions. **a** Heatmap showing Spearman's rank-based correlations between species and pathways with significantly different abundance (using ordinal regression among non-responders, partial-responders, and responders groups; adjusted $p<0.1$). Only pathways with contributions from at least one of the species enriched in the same condition are shown. FDR-corrected $p<0.1$ was deemed significant. The condition where the species or pathways are enriched is shown in coral and green for non-responders and responders, respectively. **b** Barplots showing the number of significant KOs mapped to each enriched pathway in responder and non-responder in green and coral, respectively. **c** Volcano plot of differentially abundant KOs based on the comparison between responders and non-responders. The log2 fold change and the log10 *p* values adjusted for multiple testing are plotted for each of the KOs. The dots marked with green represent significant KOs and the dots marked with red represent significant KOs involved in the biosynthesis of amino acids. The significant species which contributed to these KOs were annotated in the plot. **d** Comparison of the degree and closeness centrality between responders (R) and non-responders (NR) SparCC networks (Student *t* test, ***: $p<0.001$). **e** Co-abundance network among species enriched in responders, the amino acid-related KOs, and amino acids (AA) auxotroph species (only significant correlations are considered, $p<0.05$). Color intensity of the edges refers to the correlation value. The green, blue, and red color of the nodes represents species enriched in responders, the amino acid-related KOs, and AA auxotroph species, respectively

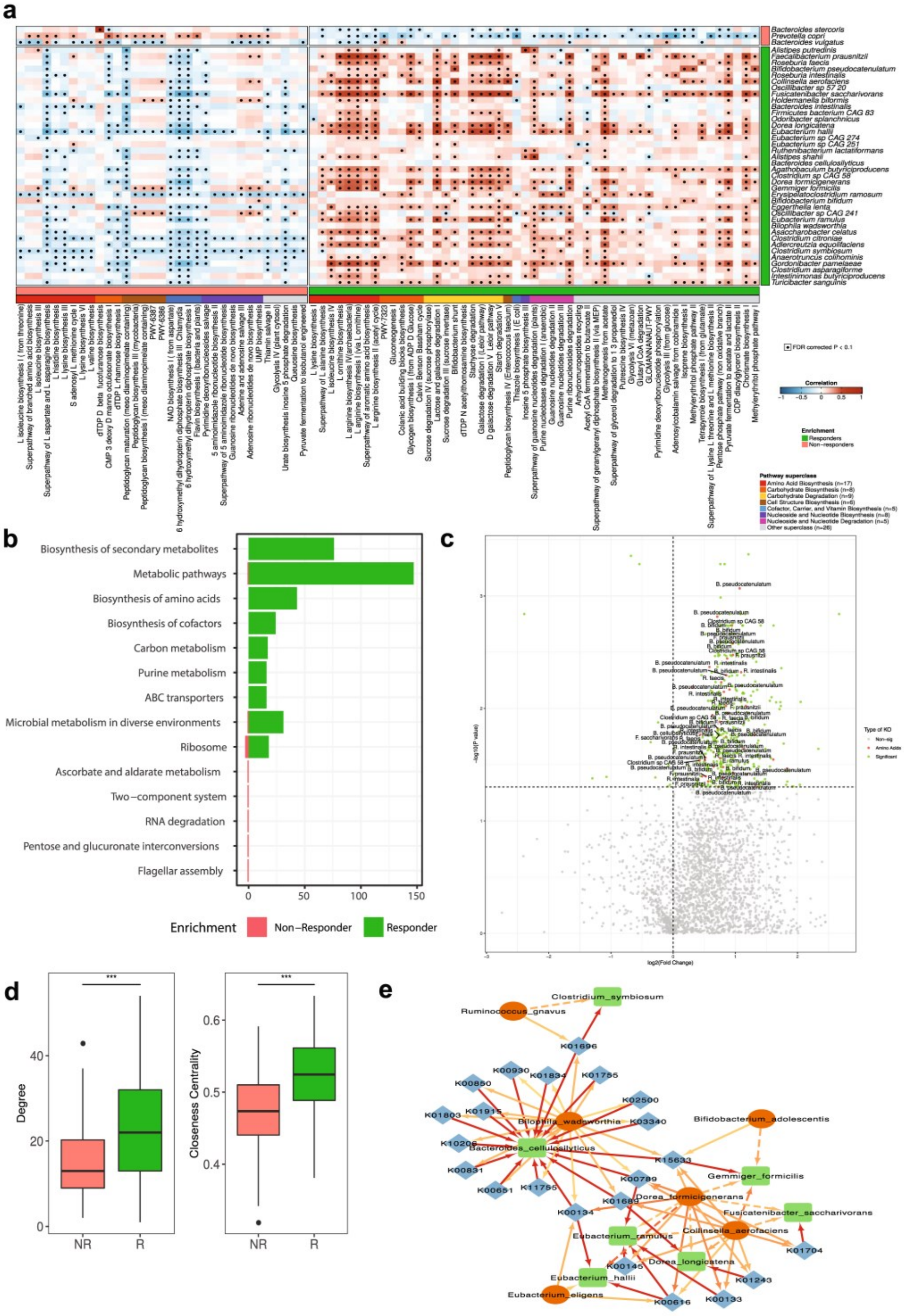


Fig. 3 (See legend on previous page.)

models depending on the input data to classify patients as responders or non-responders: a species, a genus, a taxonomic (with genus and species), and a hybrid model using pathways and taxa (Table S3).

We found that the species-based model classified the responders vs non-responders correctly, with an AUC of 0.75 ± 0.10. Eight species were selected in more than 70% of the 100 gbm species-based models, and *P. copri* (a significantly enriched species in non-responders) was selected in all the models. The classification performance was slightly increased in the genus-based model with an AUC of 0.79 ± 0.09. In the case of the genus-based model, 16 genera were consistently selected (> 70% of the 100 gbm models), and 2 were selected in all the models (*Bacteroides* and *Prevotella*). Similar classification performance was obtained when combining species and genus together (0.78 ± 0.08 AUC) or combining species with pathways (AUC of 0.74 ± 0.10). We built a final taxonomic-based model (see "Methods" for details) and obtained an AUC of 0.81 for the training set (sensitivity = 0.81 and specificity = 0.78, Fig. 4a). Recursive feature elimination was performed to reduce the dimensionality of the dataset to select the most important taxa for the classification of the responsiveness of the microbiome community. Figure 4b shows the feature importance of each of the selected features, or in other words, a score that measures how powerful is each feature in classifying the microbiome responsiveness. The final model consisted of 18 species and 12 genera as the top features, including 13 species and 6 genera that were significantly associated with responsiveness in the ordinal regression analysis (Fig. 4B, Table S4). The final model was then validated in two different external cohorts. The first cohort of subjects with inflammatory bowel disease (IBD) underwent an IBD-anti-inflammatory diet (IBD-AID, consumption of prebiotics, probiotics, and beneficial foods) for a period of 8 weeks, and the second cohort underwent a whole grain-rich diet (WGD) intervention for 8 weeks and consisted of overweight subjects. A total of 6 responders and 6 non-responders were identified in the IBD cohort, and 14 responders and 25 non-responders in the overweight cohort based on the same criteria established. The predictive power of the model in the external cohorts remained high with an AUC of 0.86

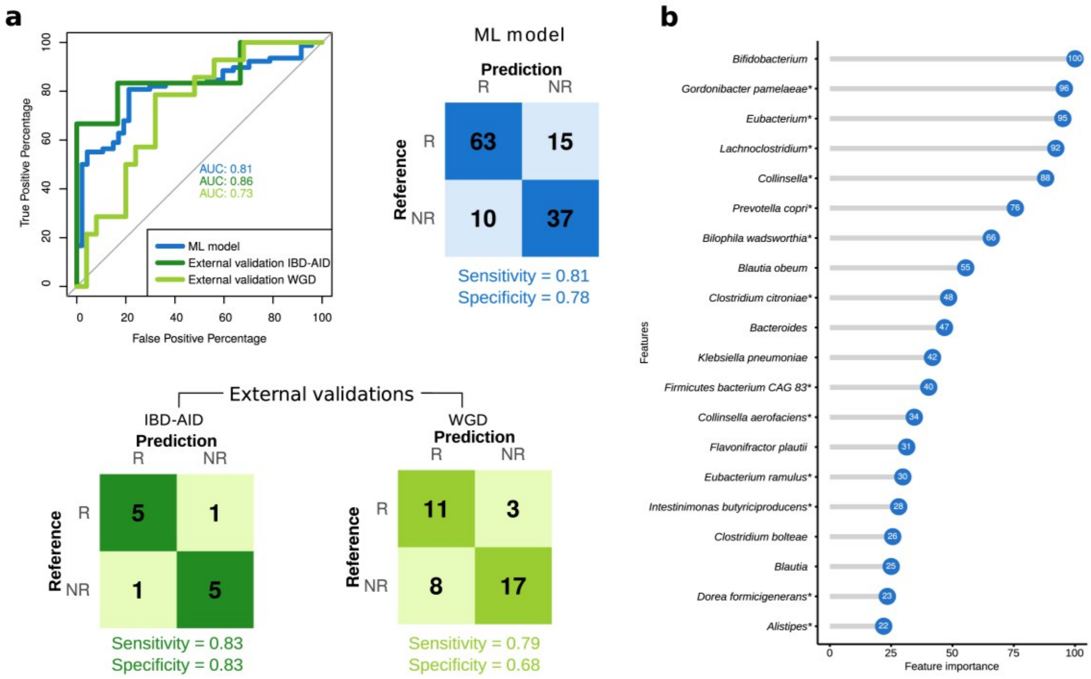


Fig. 4 Performance of the machine learning model to classify individuals as responders and non-responders based on the degree of microbiome response. **a** Receiver operating characteristic curves (ROC) for the final model and external validation. Confusion matrix of the training model and external validation cohorts. IBD-AID: Inflammatory bowel disease-anti-inflammatory diet; WGD: whole grain diet. **b** Variance importance of the top 20 features selected by the final model. Significantly different in abundance species using ordinal regression are marked with *. The importance score of each feature is indicated inside the blue circles

(sensitivity = 0.83 and specificity = 0.83) for the IBD-AID intervention and an AUC of 0.73 (sensitivity = 0.79 and specificity = 0.68) for the WGD intervention (Fig. 4A).

In summary, a gradient boosting model based on taxonomic data was developed achieving a good prediction of the microbiome response in two external cohorts including individuals from different ethnic backgrounds with metabolic and non-metabolic diseases that underwent lifestyle interventions.

Discussion

From metabolic to immune to neurological disorders, the microbiome influences the development, progression, and therapeutic outcomes of diseases [27–30]. A novel treatment approach for both disease control and disease prevention involves altering host–microbiota interactions through tailored lifestyle interventions. Unlike antibiotics usage, which is broadly reported to have a negative impact on healthy host by significantly decreasing the overall gut microbiome diversity, lifestyle interventions are regarded as a beneficial strategy to improve the metabolic performance by modulating the host-gut microbiome. Changes in the composition of the bacterial consortia in the gut from a disease-associated to a more homeostatic state are one of the desired effects of lifestyle interventions. Furthermore, comparing with the dramatical dysregulation of gut microbiome after receiving high doses of antibiotics [15], although the response to lifestyle interventions may have a common signature within the population, heterogeneous and highly personalized shifts in the human microbiota have been confirmed in several studies [31–35].

Here we attempted to identify robust and generalizable biomarkers among the gut microbial communities associated with the degree of change in the microbiome structure. We performed longitudinal shotgun metagenomics analysis from a wide range of lifestyle interventions, and established criteria to classify individuals as responders and non-responders based on their gut microbiome restructuring, using as a point of departure the natural fluctuation of a healthy gut microbiome without any intervention. We identified *P. copri*, *B. stercoris*, and *B. vulgatus* to be highly abundant in the baseline microbiomes of individuals in whom lifestyle interventions had only a minor impact on the microbial community's structure. Similarly, we found these 3 species enriched in the microbiome of individuals that were resistant to antibiotics treatment in line with recent evidence [36] from a 16S rRNA-based analysis in which the response to antibiotics in humans is determined by specific genera in the pre-treatment microbiota. Interestingly, *P. copri*, *B. stercoris*, and *B. vulgatus* are highly stable in the absence of interventions (ICCs > 0.5) suggesting their potential as biomarkers for microbiome stratification.

In contrast to the low number of species found enriched in the resistant microbiomes, we found 38 species to be highly abundant in the microbiomes that were significantly re-structured in response to lifestyle interventions. Interestingly, almost all species enriched in responders were positively correlated with at least one amino acid biosynthesis pathway. The interchange of vital metabolites, also known as metabolic cross-feeding, is a crucial process that controls the development and composition of microbial communities. Case-by-case explanations of the significance of amino acids in natural interkingdom and interspecies exchange networks have been provided by entomological investigations [37, 38]. Furthermore, a considerable proportion of all bacteria, according to comparative analysis of microbial genomes, lack crucial pathways for amino acid production [39]. Therefore, amino acid auxotrophy may promote cooperative interactions between different bacteria in the microbiome [25]. Our findings here suggest that microbiomes with a high abundance of amino acid biosynthesis pathways are also more likely to respond to different lifestyle interventions including both dietary and exercise interventions, targeting the restructuring of gut microbial communities. This finding is consistent with previous studies that amino acid biosynthesis is enriched in elite athletes [40] and decreased with high-fat diet treatment [41], highlighting the inner correlation between exercise and diet interventions. Therefore, supplementation with *F. prausnitzii*, *F. saccharivorans*, *E. ramulus*, and *E. hallii*, or other species both enriched and identified here as major taxonomic drivers of amino acid biosynthesis in the responder group, should be explored as a way to restore the metabolic flexibility required prior to microbiome-targeted lifestyle interventions. Importantly, among these potentially beneficial species, *F. prausnitzii*, *F. saccharivorans*, and *E. hallii* were reported to be enriched after exercise and diet intervention across multiple studies [42–47].

Similar to personalized medicine, personalized lifestyle approaches look for critical microbiome characteristics that can predict how an individual will react to specific lifestyle components. This information can then be used to help design a lifestyle that will have positive effects. Identifying the interactions between the host, the microbiome, and lifestyle exposures that influence lifestyle responses is the fundamental difficulty in realizing the potential of a microbiome-informed customized lifestyle. Whereas previous studies have demonstrated that the microbiome composition can be used to classify individuals into responders and non-responders on the basis of the health improvements from lifestyle interventions [48, 49], predictive models of personalized microbiota response have not yet been developed. We demonstrated here that it is possible to develop a generic ML model covering diverse lifestyle exposures that predicts the scale of microbiome change using only the baseline microbiome composition. Our model, which achieved AUCs up to 0.86 in external validation cohorts, can potentially be used for individual microbiome-based stratification, as an intermediate step towards personalized recommendations for improving the success rates of certain lifestyle interventions.

Our study has limitations. Even though several comparative analyses among studies have been performed using the ICC values [22], it is possible that ICCs may be affected by the general setup of each study, including the storage and sampling procedures, which may influence the outcome of the comparative analysis. Nevertheless, previous studies suggested a relatively stable bacterial community evaluated by ICCs with limited impact by the processing speed and storage duration [50, 51]. Furthermore, DNA extraction methods have been shown to influence the microbiome community results [52], and remained inconsistent across different cohorts in our study. Nevertheless, the impact of DNA extraction methods on metagenomic shotgun sequencing analysis of stool samples was reported to be the lowest compared to other tissue [53]. Lastly, following a strict filtering criterion, only two large-scale studies, both with Caucasian subjects, were selected to represent the healthy gut microbiome with high confidence of disease absence. Analysis of a larger cohort, well-balanced in gender and ethnicity, would allow to establish a more generalized baseline of microbiome variation in healthy individuals. The number of studies with dense longitudinal characterization of the microbiome upon lifestyle interventions is also limited and in most cases the clinical and biochemical data of the subjects are not available. Larger, more complete, and balanced datasets would allow to increase the statistical power of the data analysis and use of advanced algorithms, like deep learning, to investigate the correlation between microbiome and host response to lifestyle interventions. Nevertheless, our study offered novel insight into the microbial species and functions that may determine microbiome dynamics in response to lifestyle interventions.

Conclusions

Human gut microbiome serves as a therapeutic target for multiple diseases through lifestyle interventions. However, subjects may have different treatment efficacy which may be due to the response of gut microbiota towards the interventions. In this study, we observe individuals with either highly plastic or resistant microbial composition with the stress of lifestyle interventions. We further identify key species and functions such as *Bacteroides stercoris*, *Prevotella copri*, and amino acid biosynthesis regulating the responsiveness of the gut microbiota. Last but not least, we demonstrate with our machine learning model that it is possible to predict microbiome resistance to change in response to lifestyle interventions using the baseline microbiome composition. In summary, this study shows that the composition and function of the gut microbiome are important to determine their response to lifestyle interventions and this knowledge may help to improve the design of personalized lifestyle approaches.

Methods

Data collection and availability

In this study, we collected shotgun metagenomic sequencing data from 10 publicly open available microbiome projects. These projects included (i) 2 longitudinal cohorts of healthy subjects ($N=231$); (ii) 4 antibiotic intervention cohorts ($N=71$); and (iii) 5 lifestyle intervention cohorts ($N=165$) with metabolically diseased subjects that underwent dietary and/or exercise interventions (Table 1 and Table S5). The 2 longitudinal studies of healthy subjects with no intervention applied, abbreviated as CTL_1 and CTL_2, respectively, served as controls of normal gut microbiota fluctuation. In both studies, the selected subjects were not asked to follow diet or lifestyle recommendations and they followed their own lifestyle habits. From CTL_1, two pairs of samples taken 6 months apart from 140 subjects were used. In CTL_2, we used data from 78 subjects with one pair of samples and 4 subjects with a dense long-term time series. We used pair samples with a time interval between pairs of 2–3 months. We also selected samples that were taken 4 days apart (12 such pair samples were included). For the antibiotic intervention cohorts, the study of Palleja et al. [15] provides a cohort of healthy subjects that underwent a 4-day intervention with a cocktail of 3 last-resort antibiotics: meropenem, gentamicin, and vancomycin (A_MER–GEN–VAN). The Raymond et al. [54] cohort is composed of healthy participants that were treated twice a day with an oral dose of cefprozil for 7 days (A_CEF). The Willmann et al. [17] study provides two different cohorts of hematological patients receiving prophylactic antibiotics during a mean period of 6 days. One cohort was treated with ciprofloxacin (A_CIP) and the other with cotrimoxazole (A_COT). Regarding the lifestyle intervention cohorts, the first cohort was obtained from the study of Louis et al. [18] in which obese patients were involved in a multidisciplinary weight-loss program for 3 months (I_MWP). In Mardinoglu et al. [19], Non-alcoholic fatty liver disease (NAFLD) obese subjects underwent a low-carbohydrate diet with increased protein content during a 2-week period (I_LCD). The cohort of Zhao et al. [20] is composed of participants diagnosed with type 2 diabetes (T2D) that were administered a high-fiber diet for 3 months (I_HFD). The Ni et al. study provides data from NAFLD patients that were involved in a diet with high-resistant starch type II content for 4 months (I_HRS). The last cohort, from Liu et al. [9], is composed of prediabetes patients that enrolled in an exercise training program 3 days/week for a period of 3 months (I_ETP). More information and the number of samples used in each cohort are shown in Table 1 and Table S5. The Olendzki et al. [11] cohort was used as external validation of the machine learning predictive final model of response to lifestyle interventions. It is an IBD-anti-inflammatory dietary intervention (IBD-AID) for 8 weeks in a total of 15 subjects with inflammatory bowel disease. A second external validation cohort from Nielsen et al. [55] composed of 50 overweight subjects that underwent a whole grain dietary intervention for 8 weeks was used.

Quality control and taxonomic profiling

For the quality control of the raw reads, human DNA contaminations were removed using bwa mem against the human reference genome ucsc.hg19, and adaptors, low-quality reads, bases, or PCR duplicates were filtered as previously described [56]. The high-quality reads were taxonomically profiled at different taxonomic levels using MetaPhlAn 3.0 [57]. Default settings were used to generate taxonomic relative abundances (total sum scaling normalization).

Functional profiling

Microbial gene family abundances in metagenomic DNA reads were estimated using HUMAnN 3.0 [58]. Gene families were further mapped to the MetaCyc metabolic pathway database included in HUMAnN3 to obtain the MetaCyc pathway abundances. KOs with the species contribution were obtained in HUMAnN3 by KEGG database. Tables of pathway and gene family abundance obtained using HUMAnN3 were normalized to copies per million (CPM), including unmapped and unintegrated read mass.

Microbiome diversity measurements

Microbiome diversity was calculated based on the species, phylum, and KO gene abundance profiles, respectively. For taxonomic diversity, 3 alpha diversity indexes (including Shannon diversity, Simpson diversity, Chao1 diversity) and 3 beta diversity indexes (including Bray–Curtis dissimilarity, Weighted and Unweighted UniFrac distance) were analyzed by the vegan package [59] and phyloseq package [60] in R, respectively. For functional diversity, the 3 mentioned alpha diversity indexes and Bray–Curtis dissimilarity were calculated. Principal coordinate analysis (PCoA) based on the beta diversity was performed, and the top 5 axes were included for follow-up analyses.

ICCs of microbiome measurements

The intraclass correlation coefficient (ICC) that ranges from 0 to1 was used to represent the microbial stability (and resistance to perturbations) from totally unstable (ICC = 0) to perfectly stable (ICC = 1). We evaluated the ICC value for each diversity measurement described above and for the species, genus and MetaCyc pathway profiles using relative abundances to investigate the microbial stability within individuals of the no-intervention cohort and the resistance to perturbations within individuals from intervention cohorts. Diversity indexes were transformed into Gaussian distribution with bestNormalize package in R and the arcsine square-root transformation was implemented to the relative abundances of taxonomic and functional profiles as proposed previously [61]. After the metric transformation, ICC estimates and their 95% confident intervals were calculated using the rptR [62] package in R based on a mean-rating, absolute-agreement, 2-way random-effects model with 1000 bootstraps. The statistical comparisons of ICC values among cohort types were performed with Student *t* test. False discovery rate (FDR) correction was implemented to adjust *p* value for multiple comparisons.

Defining degree of response to perturbation

We first calculated the Bray–Curtis dissimilarity of the microbiome composition between samples within 1–2 days for each individual from the longitudinal cohorts with no intervention, which we used to estimate the daily fluctuation of the microbiome without disturbance. We then evaluated the degree of response towards lifestyle (and antibiotic) interventions of each subject by calculating the Bray–Curtis dissimilarity between baseline and each time point after the intervention and selected the time point with the first peak value of Bray–Curtis distance to baseline. The information of the selected time point for each subject among studies are shown in Table S7. The mean + SD and the mean + 2SD of the Bray–Curtis dissimilarity calculated from the control cohorts (no-intervention) were further used as the two cut-offs in the lifestyle interventions for distinguishing between responders (> mean + 2SD), partial-responders ([mean + SD, mean + 2SD]) and non-responders (< mean + SD). PERMANOVA tests were performed among responders, partial-responders, and non-responders of the lifestyle interventions using the

Bray–Curtis dissimilarity of baseline microbiome, and an ordinal regression model was used to find statistically significant taxonomic and functional differences among the three groups by applying the ordinal package in R. FDR correction was implemented to adjust *p* value for multiple comparisons.

Network analysis

In order to investigate the differences and the role of specific taxa in the microbiome community between the subjects with different responses, network analyses were performed using taxonomic data of the baseline samples. To build SparCC correlation networks for the responder and non-responder subjects, FastSpar R package was used. Only significant correlations between species were considered (adjusted $p < 0.1$). Cytoscape version 3.9.0 [63] was used to analyze the networks. Statistical comparisons between the degree and closeness centrality of the taxonomic networks of responders and non-responders were performed using the t.test function from R package stats. Furthermore, the Spearman correlation among species significantly enriched in responders, their contributed KOs which were related to the biosynthesis of amino acid and AA auxotroph species were performed in R. Only significant correlations were considered (adjusted $p < 0.1$).

Development of machine learning models

The Caret [64] R package was used to build a gradient boosting machine (gbm) model to train and classify responders and non-responders based on the baseline microbiome. We built 4 different models depending on the input data provided: a species model, a genus model, a taxonomic model using species and genus data, and a hybrid model using species, genus, and pathways data. To obtain a learning model with good interpretability and generalizability, we built a final model that included not only internal validation but also external validations, as it is critical to developing quality machine learning models [65]. The following approach was applied to build the model which included the following steps: (1) loaded the specific data (depending on the model species, genera, or pathways); (2) used the createdatapartition function from caret package to select 80% of the samples as training set; (3) performed feature selection in the training set selecting the top 30 features by applying recursive feature elimination using the rfe R function; (4) trained the model after centering, scaling the data, and removing variables with near-zero variance, using leave-one-out cross-validation (LOOCV) as a resampling method. Leave-one-out cross-validation (LOOCV) is a special case of *K*-fold cross-validation, where *K* equals the number of observations in the dataset [66]. Cross-validation techniques are used for evaluating ML models protecting the model against overfitting or selection bias and giving insights on how the model will generalize when an independent dataset is provided to the model. GBM was used as a machine learning model method and grid search to tune the hyperparameters. “Interaction.depth”, “n.trees”, “shrinkage”, and “n.minobsinnode” were applied by the expand.grid R function; (5) tested the training model in the 20% of the data. Doing only one partition may provide biased results depending on the data split (“lucky” or “unlucky” split) [67]. Therefore, in order to perform a robust interpretation of the model’s performance, the machine learning algorithm was applied 100 times using different random training-test splits; (6) steps 2–5 were repeated 99 times to obtain the overall testing performances. Model performance was assessed using the evalm function from Mleval R package, and receiver operating characteristic curve (ROC) was obtained using the R package pROC; (7) then applied steps 3–4 to the entire dataset to obtain the final machine learning model; (8) evaluated the model’s performance in external cohorts (information about the external cohorts is found in the “Supplementary Information” section).

Data visualization

The circos plot was made using iTOL (interactive Tree of Life) v6 [68]. Network visualizations were made by using the software Cytoscape version 3.9.0 [63]. All the other figures were generated by R software 3.6.3, using ggplot2, ggcorrplot, and pROC packages.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-023-01604-z.

Additional file 1: Figure S1. Related to Figure 3. Relative abundances of the significant pathways using ordinal regression among non-responders, partly-responders and responders groups ($p < 0.05$). **Figure S2.** Related to Figure 3. Correlation network of responders showing the positive correlations between enriched in responders species and auxotroph species (only significant correlations are considered, $p < 0.05$). Width and color intensity or the edges refers to the correlation value. Blue nodes are species significantly enriched in responders, yellow nodes are AA auxotroph species and orange nodes are AA auxotroph and significantly enriched in responders species. **Table S1.** Related to Figure 1. Detailed ICCs value of different diversity indexes for each cohort. **Table S2.** Related to Figure 4. Statistics of the ICCs value of each species. Uniquely influenced disease related species of each cohort. **Table S3.** Related to Figure 4. Model performance results of the 100 different splits. Mean and standard deviation of sensitivity, specificity, and AUC for the 100 models. **Table S4.** Related to Figure 4. Species and genus selected by the final model. Significance from the ordinal regression comparing response groups. No: non-significant, Enriched R: significant and enriched in responders, Enriched NR: significant and enriched in non-responders. **Table S5.** Related to Table 1. Summary of sequencing and microbiome information of the studies used in the meta-analysis. **Table S6.** Related to Figure 2. Significant species between responder and non-responder from ordinal regression. **Table S7.** Related to Table 1. Information of the time point selected for each subject for responsiveness classification. **Table S8.** Related to Figure 4. Count of each category among discovery and validation cohorts.

Acknowledgements
G.P. would like to thank the members of the Cluster of Excellence Balance of the Microverse (DFG, German Research Foundation) for intellectual input.

Authors' contributions
GP, JC, and SLS conceived and designed the study. SLS and JC performed the bioinformatics analyses. GP and YN coordinated the study. GP, YN, HBN, and AX supervised the research work. GP, SLS, and JC wrote the manuscript. YN, HBN, and AX reviewed the manuscript. All authors made substantial contributions and approved the final version of the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This project was funded by the Marie Sklodowska-Curie Actions (MSCA), Innovative Training Networks, H2020-MSCA-ITN-2018, 813781 "BestTreat" (G.P., H.B.N., S.L.S.), the European Union (EU) – Hong Kong (HK) Research and Innovation Cooperation Co-funding Mechanism by the RGC (E-HKU703/20) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Germany's Excellence Strategy—EXC 2051 – Project-ID 390713860 (G.P, Y.N). G.P would like to thank the German Bundesministerium für Bildung und Forschung (BMBF, Project 01KD2101) for their support.

Availability of data and materials
The raw sequence data used in the current study are available in the NCBI Sequencing Read Archive under the accession numbers: PRJNA354235, PRJNA544527, PRJEB20800, PRJEB8094, PRJEB28058, PRJNA290729, PRJNA420817, PRJEB15179, PRJNA703757, PRJNA454826, PRJNA642308, and PRJNA395744.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 September 2022 Accepted: 19 June 2023
Published online: 08 August 2023

References

1. Wang B, et al. The human microbiota in health and disease. Engineering. 2017;3(1):71–82.
2. Asnicar F, et al. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med. 2021;27(2):321–32.
3. Chakrabarti A, et al. The microbiota–gut–brain axis: pathways to better brain health. Perspectives on what we know, what we need to investigate and how to put knowledge into practice. Cell Mol Life Sci. 2022;79(2):1–15.
4. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. Cell Res. 2020;30(6):492–506.
5. Chunxi L, et al. The gut microbiota and respiratory diseases: new evidence. J Immunol Res. 2020;2020:2340670.
6. Cryan JF, et al. The gut microbiome in neurological disorders. Lancet Neurol. 2020;19(2):179–94.
7. Durack J, Lynch SV. The gut microbiome: relationships with disease and opportunities for therapy. J Exp Med. 2019;216(1):20–40.
8. Fassarella M, et al. Gut microbiome stability and resilience: elucidating the response to perturbations in order to modulate gut health. Gut. 2021;70(3):595–605.
9. Liu Y, et al. Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. Cell Metabol. 2020;31(1):77-91.e5.
10. Roager HM, et al. Whole grain-rich diet reduces body weight and systemic low-grade inflammation without inducing major changes of the gut microbiome: a randomised cross-over trial. Gut. 2019;68(1):83–93.
11. Olendzki B, et al. Dietary manipulation of the gut microbiome in inflammatory bowel disease patients: pilot study. Gut Microbes. 2022;14(1):2046244.
12. Lozupone CA, et al. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220–30.
13. Raaj S, et al. Stability of the human faecal microbiome in a cohort of adult men. Nature Microbiology. 2018;3(3):347–55.
14. Poyet M, et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nature Medicine. 2019;25(9):1442–52.
15. Palleja A, et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. Nat Microbiol. 2018;3(11):1255–65.
16. Raymond F, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. The ISME Journal. 2016;10(3):707–20.
17. Willmann M, et al. Distinct impact of antibiotics on the gut microbiome and resistome: a longitudinal multicenter cohort study. BMC Biol. 2019;17(1):1–18.
18. Louis S, et al. Characterization of the gut microbial community of obese patients following a weight-loss intervention using whole metagenome shotgun sequencing. PLoS ONE. 2016;11(2): e0149564.
19. Mardinoglu A, et al. An integrated understanding of the rapid metabolic benefits of a carbohydrate-restricted diet on hepatic steatosis in humans. Cell Metabolism. 2018;27(3):559-571.e5.
20. Zhao L, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science. 2018;359(6380):1151–6.
21. Ni Y, et al., Resistant starch decreases intrahepatic triglycerides in NAFLD patients via gut microbiome alterations. Cell Metabolism. (in press).
22. Sinha R, et al. Quantification of human microbiome stability over 6 months: implications for epidemiologic studies. Am J Epidemiol. 2018;187(6):1282–90.
23. Bobak CA, Barr PJ, O'Malley AJ. Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. BMC Med Res Methodol. 2018;18(1):1–11.
24. Janssens Y, et al. Disbiome database: linking the microbiome to disease. BMC Microbiol. 2018;18(1):1–6.
25. Mee MT, et al. Syntrophic exchange in synthetic microbial communities. Proc Natl Acad Sci. 2014;111(20):E2149–56.
26. Yu JS, et al. Microbial communities form rich extracellular metabolomes that foster metabolic interactions and promote drug tolerance. Nat Microbiol. 2022;7(4):542–55.
27. Cantoni C, et al. Alterations of host-gut microbiome interactions in multiple sclerosis. EBioMedicine. 2022;76: 103798.
28. Jiang X, et al. Advances in the involvement of gut microbiota in pathophysiology of NAFLD. Front Med. 2020;7:361.
29. Vatanen T, et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. Nature. 2018;562(7728):589–94.
30. Vogt NM, et al. Gut microbiome alterations in Alzheimer's disease. Sci Rep. 2017;7(1):1–11.
31. Cotillard A, et al. Dietary intervention impact on gut microbial gene richness. Nature. 2013;500(7464):585–8.
32. Korpela K, et al. Gut microbiota signatures predict host and microbiota responses to dietary interventions in obese individuals. PLoS ONE. 2014;9(3): e90702.
33. Salonen A, et al. Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. ISME J. 2014;8(11):2218–30.
34. Tap J, et al. Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults. Environ Microbiol. 2015;17(12):4954–64.
35. Walker AW, et al. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J. 2011;5(2):220–30.
36. Rashidi A, et al. Gut microbiota response to antibiotics is personalized and depends on baseline microbiota. Microbiome. 2021;9(1):1–11.
37. McCutcheon JP, Von Dohlen CD. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol. 2011;21(16):1366–72.

38. Russell CW, et al. Shared metabolic pathways in a coevolved insect-bacterial symbiosis. Appl Environ Microbiol. 2013;79(19):6117–23.
39. Mee MT, Wang HH. Engineering ecosystems and synthetic ecologies. Mol BioSyst. 2012;8(10):2470–83.
40. Clauss M, et al. Interplay between exercise and gut microbiome in the context of human health and performance. Front Nutr. 2021;8: 637010.
41. Guo K, et al. Gut microbiota in a mouse model of obesity and peripheral neuropathy associated with plasma and nerve lipidomics and nerve transcriptomics. Microbiome. 2023;11(1):1–17.
42. Mokhtarzade M, et al. Home-based exercise training influences gut bacterial levels in multiple sclerosis. Complement Ther Clin Pract. 2021;45: 101463.
43. Janabi A, et al. The effects of acute strenuous exercise on the faecal microbiota in Standardbred racehorses. Comparative Exercise Physiology. 2017;13(1):13–24.
44. Qiu L, et al. Exercise interventions improved sleep quality through regulating intestinal microbiota composition. Int J Environ Res Public Health. 2022;19(19):12385.
45. Maioli TU, et al. Possible benefits of Faecalibacterium prausnitzii for obesity-associated gut disorders. Front Pharmacol. 2021;12: 740636.
46. Duan M, et al. Characteristics of gut microbiota in people with obesity. PLoS ONE. 2021;16(8): e0255446.
47. Engels C, et al. The common gut microbe Eubacterium hallii also contributes to intestinal propionate formation. Front Microbiol. 2016;7:713.
48. Chumpitazi BP, et al. Gut microbiota influences low fermentable substrate diet efficacy in children with irritable bowel syndrome. Gut microbes. 2014;5(2):165–75.
49. Chumpitazi BP, et al. Randomised clinical trial: gut microbiome biomarkers are associated with clinical response to a low FODMAP diet in children with the irritable bowel syndrome. Aliment Pharmacol Ther. 2015;42(4):418–27.
50. Allegretti JR, et al. Stool processing speed and storage duration do not impact the clinical effectiveness of fecal microbiota transplantation. Gut microbes. 2020;11(6):1806–8.
51. Flores R, et al. Collection media and delayed freezing effects on microbial composition of human stool. Microbiome. 2015;3(1):1–11.
52. Greathouse KL, Sinha R, Vogtmann E. DNA extraction for human microbiome studies: the issue of standardization. Genome Biol. 2019;20:1–4.
53. Sui H-Y, et al. Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. Front Microbiol. 2020;11:953.
54. Raymond F, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. 2016;10(3):707–20.
55. Nielsen RL, et al. Data integration for prediction of weight loss in randomized controlled dietary trials. Sci Rep. 2020;10(1):20103.
56. Li J, et al. Probiotics modulated gut microbiota suppresses hepatocellular carcinoma growth in mice. Proc Natl Acad Sci. 2016;113(9):E1306–15.
57. Beghini F, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021;10:e65088.
58. Franzosa EA, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15(11):962–8.
59. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14(6):927–30.
60. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE. 2013;8(4): e61217.
61. Nakagawa S, Schielzeth H. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev. 2010;85(4):935–56.
62. Stoffel MA, Nakagawa S, Schielzeth H. rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods Ecol Evol. 2017;8(11):1639–44.
63. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
64. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1–26.
65. Ho SY, et al. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns. 2020;1(8): 100129.
66. Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn. 2015;48(9):2839–46.
67. Topçuoğlu BD, et al. A framework for effective application of machine learning to microbiome-based classification problems. MBio. 2020;11(3):e00434-e520.
68. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293–6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

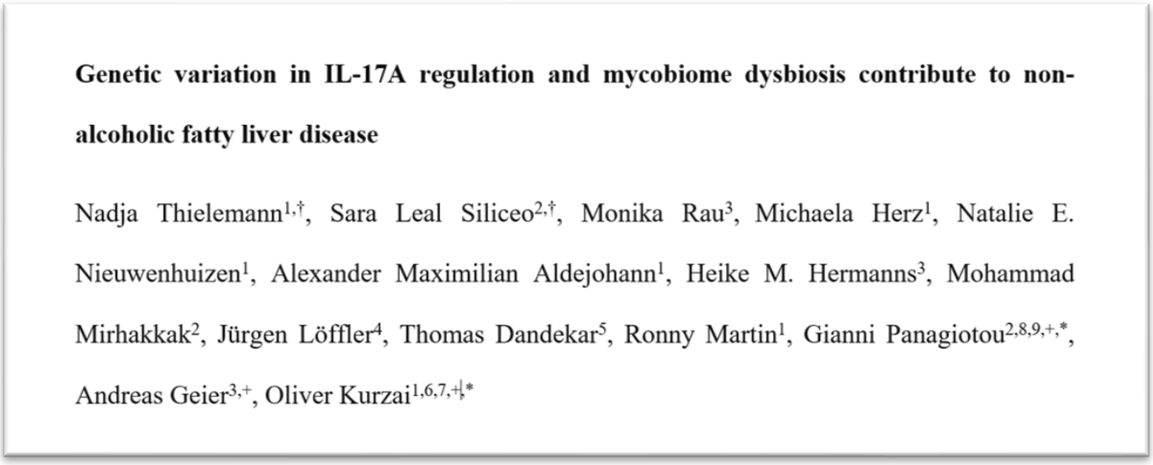
- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Manuscript III



Genetic variation in IL-17A regulation and mycobiome dysbiosis contribute to non-alcoholic fatty liver disease

Nadja Thielemann[1,†], Sara Leal Siliceo[2,†], Monika Rau[3], Michaela Herz[1], Natalie E. Nieuwenhuizen[1], Alexander Maximilian Aldejohann[1], Heike M. Hermanns[3], Mohammad Mirhakkak[2], Jürgen Löffler[4], Thomas Dandekar[5], Ronny Martin[1], Gianni Panagiotou[2,8,9,+,*], Andreas Geier[3,+], Oliver Kurzai[1,6,7,+,*]

Overview

In manuscript III, we aimed to investigate how intestinal fungi contribute to NAFLD development by looking into a possible antifungal immunity and a potential mycobiome dysbiosis. Therefore, we characterized the fecal mycobiome of a NAFLD cohort to elucidate if potential alterations in Th-17 (T-helper cells that produce interleukin-17) signaling are accompanied by mycobiome dysbiosis. In addition, we further investigated how the combination of genetic variation in Th-17 signaling and mycobiome dysbiosis contributes to inflammation in steatohepatitis (NASH). These results showed that NAFLD patients harboring a genetic variation in their IL-17A gene concomitantly present increased levels of *Candida* CTG species, and these factors predispose to develop disease progression up to NASH and advanced fibrosis.

FORM I

Manuscript No: 3

Manuscript title: Genetic variation in IL-17A regulation and mycobiome dysbiosis contribute to non-alcoholic fatty liver disease

Authors: Nadja Thielemann*, **Sara Leal Siliceo***, Monika Rau, Michaela Herz, Natalie Nieuwenhuizen, Alexander Maximilian Aldejohann, Heike M. Hermanns, Mohammad H. Mirhakkak, Jürgen Löffler, Thomas Dandekar, Ronny Martin, Gianni Panagiotou, Andreas Geier, Oliver Kurzai

Bibliographic information (if published or accepted for publication: Citation)**:**

The candidate is (Please tick the appropriate box)**:**

☐ First author, ☒ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status (if not published; "submitted for publication", "in preparation".): in preparation

Authors' contributions (in %) to the given categories of the publication

Author	**Conceptual**	**Data analysis**	**Experimental**	**Writing the manuscript**	**Provision of material**
Thielemann, N.*	20%	15%	65%	40%	
Leal Siliceo, S.*	20%	75%		40%	
Panagiotou, G.	15%			5%	25%
Geier, A.	15%			5%	25%
Kurzai, O.	15%			5%	25%
Others	15%	10%	35%	5%	25%
Total:	100%	100%	100%	100%	100%

*Authors contributed equally

Genetic variation in IL-17A regulation and mycobiome dysbiosis contribute to non-alcoholic fatty liver disease

Nadja Thielemann[1,†], Sara Leal Siliceo[2,†], Monika Rau[3], Michaela Herz[1], Natalie E. Nieuwenhuizen[1], Alexander Maximilian Aldejohann[1], Heike M. Hermanns[3], Mohammad Mirhakkak[2], Jürgen Löffler[4], Thomas Dandekar[5], Ronny Martin[1], Gianni Panagiotou[2,8,9,+,*], Andreas Geier[3,+], Oliver Kurzai[1,6,7,+,*]

Affiliations

[1]Institute for Hygiene and Microbiology, University of Würzburg, Würzburg, Germany
[2]Microbiome Dynamics, Leibniz Institute of Natural Product Research and Infection Biology-Hans Knöll Institute, Jena, Germany.
[3]Department of Medicine II, Division of Hepatology, University Hospital Würzburg, Germany
[4]Department of Internal Medicine II, University Hospital Würzburg, Germany
[5]Functional Genomics & Systems Biology, Department of Bioinformatics, University Würzburg, Germany
[6]Research Group Fungal Septomics, Leibniz Institute for Natural Product Research and Infection Biology- Hans Knöll Institute, Jena, Germany
[7]National Reference Center for Invasive Fungal Infections, Leibniz Institute for Natural Product Research and Infection Biology- Hans Knöll Institute, Jena, Germany
[8] Department of Medicine, The University of Hong Kong, Hong Kong S.A.R., China
[9] Friedrich Schiller University, Faculty of Biological Sciences, Jena, Germany

[†] These authors contributed equally
[+] These authors contributed equally

***Correspondence:** Gianni.Panagiotou@leibniz-hki.de (G.P.), okurzai@hygiene.uniwuerzburg.de (O.K.)

ABSTRACT

Non-alcoholic fatty liver disease (NAFLD) is the leading cause of chronic liver disease in Western countries. In non-alcoholic steatohepatitis (NASH), fat accumulation triggers inflammatory processes with a central role of Th17 responses. We show that the *IL-17A* rs2275913 minor allele variant is associated with fibrosis progression in NAFLD patients, indicating a genetic pre-disposition to NASH-associated inflammatory processes. Fungal gut commensals including *Candida albicans* are potent activators of Th17 responses. To investigate if alterations in Th17 signaling are accompanied by mycobiome dysbiosis, we characterized the fecal mycobiome in our NAFLD cohort. In NAFLD patients with advanced fibrosis, we observed an increased abundance of *Candida* CTG-clade species. In addition, T cells from donors carrying the minor allele variant secreted significantly higher IL-17A levels in response to stimulation with *Candida* CTG-clade species. This combination of increased IL-17A release and mycobiome dysbiosis may thus result in enhanced inflammation, revealing a significant role of intestinal fungi in NAFLD.

Keywords: Mycobiome, Intestinal Fungi, Th17 Signaling, IL-17A, NAFLD, NASH, *Candida*, liver fibrosis, liver inflammation

MAIN

Non-alcoholic fatty liver disease (NAFLD) has emerged as the major cause of chronic liver disease in recent years, reaching a global prevalence of around 25%[1]. The accumulation of excess fat in the liver in the absence of relevant alcohol consumption is functionally linked to obesity and risk factors like type 2 diabetes and metabolic syndrome[2]. Fat accumulation in hepatocytes is the main driver of NAFLD pathogenesis and constitutes a non-alcoholic fatty liver (NAFL)[3]. Ongoing fat accumulation and emerging lipotoxicity trigger the onset of inflammation, characterizing non-alcoholic steatohepatitis (NASH). Inflammatory processes eventually lead to the development of fibrosis which could ultimately result in a cirrhotic liver[4]. It is unclear why some patients progress to NASH and others do not, but there is some indication that Th17 responses are associated with progression to NASH[5-7].

The liver receives approximately 75% of its blood supply via the portal vein and is thus closely connected to the human intestinal tract, which is massively colonized by microorganisms – bacteria, viruses, fungi – collectively known as the microbiome[8]. Importantly, gut microbiota dysbiosis has been repeatedly observed in obesity and type 2 diabetes mellitus[9,10] and recent data provide clear evidence that the composition of gut microbiota also has a direct impact on the pathogenesis of NAFLD[11-14]. A previous study characterized a significantly higher abundance of short chain fatty acid (SCFA)-producing bacteria such as *Fusobacteriaceae*, *Prevotellaceae*, and *Ruminococcaceae* in the gut of patients with advanced NAFLD[15]. In contrast to the gut bacteriome, the composition of intestinal fungi is less well characterized and mycobiome studies are hampered by non-standardized protocols, technical difficulties and incomplete reference databases[16,17]. Although no “core gut mycobiome” has been defined so far, some species like *Candida albicans* have been identified as key colonizers and are known for their multiple interactions with the human host in health and disease[18,19]. Recent work identified *C. albicans* as the major direct inducer of human antifungal Th17 cell responses[20]. Recognition of fungal β-1,3-glucan by dectin-1 receptors can promote Th17 signaling[21]. Downstream signaling of this pattern recognition receptor involves the formation of a complex of CARD9, Bcl10 & MALT1, which ultimately leads to Th17 cell differentiation and IL-17 secretion[22]. Thus, immune activation induced by this single component of the mycobiome could be a central mechanism for systemic induction of human Th17 responses that have a broad impact upon the human body[20]. Recently, Demir et al. characterized a distinct fecal mycobiome signature in non-obese NAFLD patients to be associated with liver disease severity. The abundance of e.g. *Malassezia* sp. was increased in NAFL patients, whereas *C. albicans* and *Penicillium* spp. abundance were increased in NASH patients. Increased intestinal *C. albicans* numbers were mirrored by increased levels of systemic antibodies against *C. albicans* in NAFLD patients with advanced fibrosis[23]. Interestingly, intestinal *C. albicans* has also been linked to the pathogenesis of alcoholic liver disease via its exotoxin candidalysin[24,25]. In both cases however, a link between Th17 activation and intestinal fungi has not been addressed.

The aim of our study was to clarify how associations between genetic variations in antifungal immunity and the resident intestinal mycobiome contribute to NAFLD pathogenesis. We have identified a novel risk variant in *IL-17A* and show that intestinal colonization with *C. albicans* and related species (*Candida* CTG-clade) might contribute to enhanced inflammation in the presence of this risk genotype.

RESULTS

Study population

A total of 482 European subjects were recruited for this study including 230 histology-proven NAFLD patients (89 NAFL and 141 NASH). Figure 1 illustrates the clinical and histological phenotypes of the study participants in a flow diagram. Stool samples were collected from a sub-cohort of subjects (42 NAFL, 79 NASH, 100 NAFLD). Due to the high frequency and often diverse nature of NAFLD we mainly aimed for comparisons within the disease group. However, as an additional control a group of healthy individuals (HC) was included. Patients with 6 months antibiotic-free intervals were analyzed separately.

Genetic variation in IL-17A predisposes patients to develop fibrotic NAFLD

Th17 responses and IL-17A signaling have been shown to be an important part of inflammatory activation in NASH[7]. In an extensive dbSNP database search for genetic variants in Th17 signaling associated genes linked to gastrointestinal disease with inflammatory properties, we identified the rs2275913 *IL-17A* SNP as one suitable candidate[26]. TaqMan SNP genotyping of our 451-patient NAFLD cohort identified 175 G/G (homozygous for major allele variant, 38.8%), 55 A/A (homozygous for minor allele variant, 12.2%), and 221 A/G (heterozygous, 49%) genotypes (Fig. 2a). Thus, genotype frequencies are in Hardy-Weinberg equilibrium and selection for specific genotypes was excluded (Extended Data Fig. 1). The calculated minor allele frequency (MAF) of 36.7% was just slightly elevated in comparison to the published ALFA European cohort MAF of 34.85%. Statistical analysis of genotyping data revealed an association between the *IL-17A* rs2275913 genotype and fibroscan liver stiffness values ($P_{\text{Kruskal-Wallis}}$=0.368; P_{glm}=0.029; GLM adjusted; Fig. 2c). Patients carrying the SNPs minor allele variant (A/A & A/G) had more severe fibrosis than homozygous major allele variant carriers (G/G; P_{glm}=0.0292, GLM adjusted). Due to its potential as a major genetic risk factor for NAFLD[27], we validated the association of available *PNPLA3* rs738409 genotyping data from a previous study[28] to fibroscan stiffness values of patients in this cohort (P_{Kruskal}Wallis=0.105; P_{glm}=0.028; GLM adjusted; Fig. 2b) and adjusted all SNP-based generalized linear model (glm) calculations for the *PNPLA3* rs738409 risk genotype. We also analyzed the rs16910526 SNP in the *CLEC7A* gene (coding for Dectin-1) and the rs4077515 SNP in the *CARD9* gene, both also related to Th17 response, but did not find an association with NAFLD disease parameters (Extended Data Fig. 2).

Alpha diversity of patients with NASH, but not NAFL, is affected by prior use of antibiotics

Intestinal colonization by *Candida* is a major inducer of Th17 responses[20]. Thus, we built ITS1 libraries for 145 subjects from our study cohort to estimate the fungal genus and species abundance and explore the possible role of fungi in NAFLD progression and liver damage. On average, we generated 15,500 high-quality, non-chimeric reads per sample while fungal annotation identified 29 genera and 223 species in total. Investigating genus-level fungal profiles showed that *Saccharomyces*, *Penicillium*, and *Candida* CTG-clade were the top most abundant fungi among our study participants, at 16.7%, 16.1%, and 12.5%, respectively. We used the *Candida* CTG-clade for genus clustering as *Candida* is a polyphyletic genus comprising a large variety of phylogenetically distant species. To account for this, we clustered only *Candida spp.* characterized by an alternative decoding of the CTG codon leading to a serine amino acid instead of leucine[29]. Members of the CTG-clade are pathogenic *C. albicans*, *Candida tropicalis*, *Candida dubliniensis*, *Candida parapsilosis* and non-pathogenic *D. hansenii.* Although still commonly referred to as *Candida,* other species are only distantly related to *Candida* CTG. Important examples include *Candida glabrata, Kluyveromyces marxianus* (formerly *Candida kefyr*) and *Pichia kudriavzevii* (formerly *Candida krusei*)[29,30].

In total, 76 NAFL, NASH, and NAFLD subjects in our cohort reported antibiotic use 6 months before the stool collection. A recent study showed that antibiotics might have a longterm influence on the human gut mycobiome[31]. Therefore, we investigated whether antibiotics had a noticeable impact on the mycobiome profiles of the different disease groups. We found that the mycobiome alpha-diversity measured by the Shannon and Simpson index at genus level was significantly increased in NASH subjects who used antibiotics compared to the antibiotic-free subjects (Wilcoxon rank-sum test, P=0.028 and P=0.025, Shannon and Simpson respectively; Extended Data Fig. 3). However, no differences were found in the NAFL and NAFLD groups between antibiotic and antibiotic-free subjects. Using Aitchison distance to compute beta-diversity, no differences were found between the antibiotic and antibiotic-free subjects in any of the disease groups (PERMANOVA adjusted for age, gender and obesity-related parameters, P>0.05). Nevertheless, to avoid a possible impact of antibiotics on the downstream analysis, two approaches were used for the mycobiome comparisons. For all the main results, unless specified, a dataset with only the long-term antibiotic-free samples was used. Alternatively, mycobiome analysis was performed using all samples, adjusting for antibiotic intake when appropriate (see Methods for details).

A distinct mycobiome structure characterizes NASH patients

To study the mycobiome changes related to NAFLD progression, we first performed pairwise comparisons between NAFL, NASH, NAFLD and HC in alpha diversity measured by the Shannon and Simpson index and we found no significant differences between the four diagnosed groups (Wilcoxon rank-sum test, P>0.05 for all pair group comparisons for Shannon and Simpson index). Beta diversity analysis using Aitchison distance to assess the overall mycobiome community differences showed that the fungal composition was significantly different between NASH and HC subjects (PERMANOVA adjusted for age, gender and obesity-related parameters, P=0.01, Fig. 3a), but no significant differences were found in any other pairwise comparison (PERMANOVA adjusted, P>0.05).

We then explored the fungal abundance differences between the diseased groups (NAFL, N=31; NASH, N=64; NAFLD, N=50) and healthy controls (HC, N=25). Again, fungi were grouped according to genus, except for the *Candida* CTG clade. The most abundant genus in the HC group was *Penicillium* (22.2%), followed by *Saccharomyces*

(20.9%), and the *Candida* CTG-clade (12.2%) (Fig. 3b) and a similar abundance pattern was observed for the NAFL and NAFLD groups (Fig. 3b). However, in NASH, the most abundant genus was the *Candida* CTG-clade (17.8%), followed by *Saccharomyces* (14.1%) and *Penicillium* (12.5%) (Fig. 3b). From the most abundant genera, we found *Saccharomyces* significantly decreased in abundance in NAFL and NASH in comparison to HC (Wilcoxon rank-sum test, $P_{HCvsNAFL}$=0.026, $P_{HCvsNASH}$=0.027, Fig. 3c). *Penicillium* was also found significantly decreased in abundance in all disease stages in comparison to HC even though it did not reach statistical significance in comparison to NAFL (Wilcoxon rank-sum test, $P_{HCvsNASH}$=0.038, $P_{HCvsNAFLD}$=0.023, $P_{HCvsNAFL}$=0.051, Fig. 3c). However, the statistical significance was lost for both genera when accounting for age, gender, and obesity-related parameters (Generalized Linear Model (GLM) adjusted, $P>0.05$, Fig. 3c), suggesting a potential confounding effect in the abundances of the two genera by these factors.

We subsequently repeated all the analytical steps using the full cohort and not only the antibiotics-free subjects and confirmed the significant differences in beta diversity between the NASH and HC groups (PERMANOVA adjusted, P=0.034, Extended Data Fig. 4a) and the significant decrease in abundance of *Penicillium* in the NAFLD and NASH groups compared to HC ($P_{HCvsNASH}$=0.03, $P_{HCvsNAFLD}$=0.02, Wilcoxon rank-sum test; $P_{HCvsNASH}$=0.007, $P_{HCvsNAFLD}$=0.42, GLM adjusted). Even though it did not reach statistical significance as it did when using the antibiotic-free set of samples, the same trend was observed for *Saccharomyces,* having lower abundance in the NASH and NAFLD groups compared to HC (Wilcoxon rank-sum test, $P_{HCvsNASH}$ and $P_{HCvsNAFLD} < 0.1$).

We then used 16S data from our cohort in order to build a microbial community network to identify possible associations between fungal and bacterial genera and NAFLD progression. Using all cohort samples, we built a community network using FastSpar[32], and identified a total of 5,848 significant correlations (SparCC, $P<0.05$) from which 4,017 remained significant after multiple testing correction (FDR correction, $q<0.1$). Using greedy modularity optimization, a total of 4 subcommunities were identified in the full network (Fig. 3d). We then studied the associations between these subcommunity modules and NAFLD and identified one module that consists of 2 fungal (*Candida* CTG-clade and *Saccharomyces*) and 9 bacterial genera (including *Ruminococcus*, *Dialister*, and *Parasutterella* amongst others) that were significantly associated with NAFLD-related parameters (fibroscan, AST, ALT, and GGT) (Fisher's Exact test, P=0.049, odds ratio=3.580), suggesting the interplay of the two microbial kingdoms and NAFLD.

High levels of the *Candida* CTG-clade in advanced fibrosis patients

In order to investigate whether changes in the mycobiome composition may be associated with liver fibrosis progression in more advanced stages of the disease, we classified the subjects into early- or advanced- fibrosis groups using a cutoff fibroscan value of 9.7kPa[33]. We performed diversity analyses and found no significant differences (Wilcoxon rank-sum test, $P>0.05$) in alpha diversity (Shannon and Simpson index) at the genus level. Beta diversity analysis using Aitchison distance showed that the mycobiome composition at the genus level between early and advanced fibrosis groups was significantly different (PERMANOVA adjusted for age, gender, and obesity-related parameters, P=0.007, Fig. 4a). We explored further the mycobiome composition profiles (Fig. 4b) and discovered that the *Candida* CTG-clade was significantly increased in the advanced compared to the early fibrosis group even when accounting for age, gender, and obesity-related parameters ($P_{wilcoxon}$=0.0009, Wilcoxon rank-sum test; P_{glm}=0.002, GLM adjusted, Fig. 4c). The trend of increasing *Candida* CTG-clade abundance was also visible grouped by fibrosis stage as obtained by histology (Extended Data Fig. 5). However, for fibrosis stage F3 and F4 the sample size was relatively small for biopsied patients and thus this trend was not significant (Kruskal-Wallis, P=0.07 for the antibiotic-free sample set and P=0.086 for the full cohort).

We calculated the beta diversity (Aitchison distance) of early and advanced fibrosis groups using all subjects and not only the antibiotic-free subjects, and significant differences were again identified (PERMANOVA adjusted, P=0.007, Extended Data Fig. 4b). A significant increase in *Candida* CTG-clade abundance was also found in advanced- compared to early fibrosis (Wilcoxon rank-sum test, $P_{wilcoxon}$=0.0007; GLM adjusted, P_{glm}=0.002) when analyzing the full cohort (Extended Data Fig. 4c). Our findings suggest that in both, antibiotic-free and total study cohorts, *Candida* CTG-clade abundance is significantly higher in the advanced fibrosis group, suggesting that this clade may contribute to the progression of the disease.

We further explored the *Candida* CTG-clade imbalance in an advanced fibrosis stage in the antibiotic-free set of samples and found an association between the presence/absence of the *Candida* CTG-clade and the fibrosis stage (Fisher's Exact test, P=0.006, odds ratio=3.097). We then performed regression analysis between the fibroscan stiffness values and *Candida* CTG-clade abundance and found a significant association (GLM adjusted for age, gender, and obesity-related parameters, P=0.001, estimate=0.22). Correlation analysis also showed a positive significant correlation between fibroscan values and *Candida* CTG-clade abundances (Spearman's correlation adjusted, P=0.026, ρ=0.23). We then evaluated this association for the complete cohort of samples and the same results were obtained (presence/absence of *Candida* CTG-clade associated with the fibrosis stage, P=0.002, odds ratio=2.73, Fisher's Exact test; *Candida* CTG-clade abundances and fibroscan, significant positive correlation, Spearman's correlation adjusted, P=0.01, ρ=0.20 and GLM adjusted, P=0.002, estimate=0.23, accounting for age, gender, obesity-related parameters and antibiotic use).

The IL-17A rs2275913 SNP influences IL-17 production in response to CTG-*Candida*

To determine whether the *IL-17A* rs2275913 SNP has functional relevance in responses to the highly abundant *Candida* CTG-clade representatives and may be involved in NAFLD pathogenesis, we stimulated freshly isolated T cells from rs2275913-genotyped healthy donors with fungal lysates and measured the resulting IL-17A secretion. To ensure that differences in T cell proportions among PBMCs of individual donors did not influence IL-17A levels, we first isolated T cells and then used equal numbers of T cells in the stimulation assays. An age-dependent influence on $CD4^+$ T cell frequency was excluded due to similar mean age of donors in genotype groups. T cells were stimulated with fungal lysates of pathogenic (*C. albicans*) and non-pathogenic (*D. hansenii*) representatives of the *Candida* CTG-clade[34,35] as well as with PepTivator® *C. albicans*, a peptide pool of the major T cell antigen MP65 of *C. albicans* (Extended Data Fig. 6b). IL-17A secretion was measured by ELISA and calculated with a 4parameter standard fit curve (Extended Data Fig. 7). Both fungal lysates induced IL-17A production, with T cells from individuals with the rs2275913 A/A genotype having significantly increased IL-17A levels in comparison to those with the rs2275913 G/G and heterozygous genotypes (*C. albicans*: $P_{Kruskal\text{-}Wallis}$=0.022, P_{glm}=0.088; *D. hansenii*: $P_{Kruskal\text{-}Wallis}$=0.025; P_{glm}=0.040, Fig. 5a-b). Thus, the IL-17A rs2275913 genotype modifies the amount of IL-17A produced in response to the *Candida* CTG-clade species. Together with the elevated *Candida* CTG-clade abundance in NAFLD patients, this suggests a combinatory effect of dysregulated antifungal immunity and imbalance in *Candida* CTG-clade species on fibrosis development in patients with NASH.

DISCUSSION

In this study we investigated genetic variation in IL-17A and its interaction with the intestinal mycobiome as a potential risk factor in NAFLD. Previous data from our group indicate that the NAFL to NASH progression is marked by an increased frequency of IL-17 producing cells among intrahepatic CD4$^+$ T cells and higher Th17/resting regulatory T cells (rTreg) ratio in peripheral blood[7]. Genotyping of our NAFLD cohort revealed a significant association between the *IL-17A* rs2275913 variant and fibrosis severity. IL-17 is known as a profibrotic cytokine especially for liver fibrosis[36], and our data clearly suggest that this may play a role in the pathogenesis of NASH. These data provide additional insight into genetic risk factors that promote inflammation and fibrosis in NAFLD.

C. albicans is a potent trigger of Th17 responses. Increased intestinal *C. albicans* abundance was positively correlated to systemic levels of fungal-specific Th17 inflammation measured by IL-17 producing CD4$^+$ T cells. In this context, *C. albicans* commensal gut colonization triggers a host defense that is cross-reactive against other pathogens[37]. Our results suggest that in combination with the *IL-17A* risk genotype, increased *C. albicans* abundance could contribute to inflammation-driven liver fibrosis.

Intestinal mycobiome analysis additionally identified other *Candida* CTG-clade species to be highly abundant in advanced fibrosis with similarly decreased *Saccharomyces* abundance in these NAFLD patients. Importantly, a genus base taxon analysis should not be used for polyphyletic genera such as *Candida*[38]. Our results confirm recent mycobiome analysis data generated by ITS2 sequencing but extend these with the taxonomically relevant *Candida* CTG-clade grouping[23]. Primer bias does not seem to have influenced overall highly abundant fungal species but in our case ITS1 sequencing might have led to missing e.g., *Malassezia* species detection[39]. In the context of the overall intestinal microbiome, microbial communities rather than single species can contribute to NAFLD development as shown by our interaction analysis where we identified a subcommunity module that includes *Candida* CTG-clade together with one more fungal genus and nine bacterial genera that are jointly associated with NAFLD progression. In addition to mycobiome dysbiosis in NASH patients, the composition of intestinal bacteria is altered in NAFLD, suggesting the interaction of multiple components of the intestinal microbiome. Future work should focus on characterization of possible interaction mechanisms to further elucidate the role of intestinal fungi in NAFLD pathogenesis in relation to the multifactorial nature of this disease. However due to the high intestinal mycobiome variability, longitudinal and well-monitored studies are essential to exclude possible diet, antibiotic or environmental-mediated effects and identify only causal intestinal mycobiome changes associated with NAFLD pathogenesis.

Gastrointestinal or liver disease-associated mycobiome studies commonly involve members of the *Candida* CTG-clade. Studies investigating alcohol-associated liver disease (ALD) identified elevated abundance of *C. albicans* and *Debaryomyces* sp. in alcohol use disorder patients in comparison to healthy controls. Interestingly, this fecal mycobiome dysbiosis was improved by two weeks of abstinence[40]. The *C. albicans*-derived exotoxin candidalysin has been associated with the severity of liver disease in ALD patients[24], and a recent study revealed that *C. albicans* strain diversity regulates the immune response in inflammatory bowel disease[41] indicating a crucial role for highly-abundant *C. albicans* in other liver and gastrointestinal diseases. In mouse models of disease associated with mycobiome dysbiosis, amphotericin B was identified as a promising treatment as it counteracted mycobiome dysbiosis involving elevated abundance of *C. albicans* and *D. hansenii*. In fecalmicrobiome humanized gnotobiotic mice, amphotericin treatment led to decreased intestinal *C. albicans* abundance and improved diet-induced liver fibrosis and steatohepatitis[23]. In mice with Crohn's disease, *D. hansenii* was isolated in high abundance directly from mucosal wound tissue and amphotericin B treatment not only reduced *D. hansenii* abundance but also reversed the impaired crypt regeneration after injury[42].

Mycobiome dysbiosis involving increased intestinal abundance of *D. hansenii* is of particular interest, as the food-borne *D. hansenii* is often found on cheese as well as processed meat in Western-style diet and is therefore often seen as a transient mycobiome component. Although the possible probiotic properties of *D. hansenii* have been studied intensively[43], its functional role in the context of human disease is still unknown and needs to be characterized. However, it seems to possess Th17-stimulating potential, as our *ex vivo* T cell stimulation assay demonstrated elevated IL-17A levels in response to *D. hansenii* in rs2275913 minor allele variant carriers, similar to the increased IL-17A levels after *C. albicans* stimulation. Therefore, our results suggest a combinatory effect of risk variant-driven increased antifungal IL-17A response and elevated intestinal *Candida* CTG-clade species abundance that may promote fibrosis in NASH patients and thereby further elucidate the role of intestinal fungi in inflammatory-driven liver disease.

METHODS

Patients (NAFLD cohort)

In this prospective study, 451 NAFLD patients were enrolled between 2016-2019 in the division of hepatology of the department of Medicine II, University Hospital Würzburg, Germany. All study participants were >18 years old and diagnosed with NAFLD either by histology (n=230) or clinically by transient elastography (TE; fibroscan & controlled attenuation parameter (CAP) (n= 350). We included all clinically characterized NAFLD subjects in our cohort irrespective of histological characterization to investigate associations between genetic variations in antifungal immunity and gut mycobiome imbalance with the largest possible sample size. Although liver histology is considered the gold standard of NAFLD diagnosis, the more easily accessible TE is a widely used and validated technique that has shown a high performance for the diagnosis and exclusion of advanced fibrosis when compared to liver biopsy[44]. Additionally, it reduces the imminent risk of sampling error due to heterogeneous distribution of fibrosis when assessing liver biopsy specimens[45].

Clinical and anthropometric characteristics of the study cohort are shown in Table 1. A cutoff for daily alcohol consumption was set (<20g/d for female and <30g/d for male subjects) and further underlying liver disease (e.g. autoimmune liver disease or chronic viral hepatitis) was excluded. Information on patients' last antibiotic treatment was documented. Fecal, serum and whole blood samples were immediately snap-frozen and stored in the local biobank.

Ethics approval & consent to participate

This study, involving the NAFLD patient cohort (University of Würzburg: EK 96/12, 05.09.2012; EK 188/17, 13.01.2020) and healthy volunteers (University of Würzburg: EK 191/21, 16.08.2021) was approved by the local ethics committee and conforms to the ethical guidelines of the 1975 Declaration of Helsinki. We obtained written informed consent from all patients and healthy volunteers included in this study.

DNA extraction from blood and PBMCs and TaqMan SNP Genotyping

DNA was extracted from frozen blood or PBMC samples using the Roche High Pure PCR Template Preparation Kit (Sigma Aldrich, #11796828001) according to the manufacturer's instructions. Isolated DNA was then used in TaqMan SNP Genotyping Assays (ThermoFisher, CN #4351376; CARD9 (ID: C__25956930_20), CLEC7A (ID: C__33748481_10), IL-17A rs2275913 (ID: C__15879983_10) according to manufacturer's instructions. Assays were conducted with the qTower^3 (Analytik Jena) and analyzed with the qPCRsoft 3.4 software (Analytik Jena). The functionality of TaqMan

SNP Genotyping was confirmed by additional sequencing of 5% samples and validating the obtained genotypes. For sequencing, a 414 bp part of interest in the IL-17A gene was amplified (5´: ATATGATGGGAACTTGAGTAGTTTCCG, 3´: CTCCTTCTGTGGTCACTTACGTGG) with 2x Q5 polymerase master mix according to the manufacturer's instructions (NEB, #M0492L). PCR samples were purified with the PCR & Gel Clean-Up Kit (Macherey-Nagel, #740609.50) according to the manufacturer's instructions and sent to LGC genomics for sequencing with the 5´ primer. DNA sequences were evaluated with ApE (v.3.0.8).

PBMC and T cell isolation

Freshly drawn blood from healthy volunteers was diluted 1:1 in PBS / 1 mM EDTA (Invitrogen, ThermoFisher Scientific: #AM9260G) containing 1% inactivated human AB serum (Sigma Aldrich, #H4522-100ML) and separated via Biocoll density gradient medium (Bio&SELL, #BS.L 6115) in SepMate tubes (Stemcell Technologies, #85460) according to the manufacturer's instructions. Afterwards, PBMCs were washed 3 times with PBS-EDTA-human serum mix. As T cell proportions vary strongly between individual PMBC donors, we additionally isolated T cells before stimulation and IL-17A secretion measurement.

T cells were isolated from freshly isolated PBMCs by negative selection with the human Pan T Cell Isolation Kit (Miltenyi, #130-096-535) according to manufacturer's instructions and the purity of >90% was assessed by flow cytometry (Miltenyi MACSQuant ®).

PBMC and T cell number and cell viability were measured directly after isolation with the LUNA automated cell counter (Logos Biosystems) and the viability was in each case >99%.

Preparation of fungal lysates

50ml inoculated YPD medium (20g/L glucose, 20g/L peptone, 10g/L yeast extract) was cultured overnight at 25°C (*D. hansenii* CBS767) and 37°C (*C. albicans* SC5314). The overnight culture was diluted 1:50 in 50ml YPD medium and thereafter cultured for 5h. Cells were harvested by centrifugation at 4.000g for 10min and the cell pellet was resuspended in lysis buffer (50mM Tris-HCl, 150mM NaCl, 0.1% Triton X-100, 1mM DTT, 10% glycerol) with freshly adjusted proteinase inhibitor (Sigma, #S8820-20TAB). For lysis, 500µl glass beads were added per tube and five 1min vortexing steps were followed by five 1min cooling steps on ice. After centrifugation at 20000g for 5min the supernatant was transferred to a fresh reaction tube and stored in aliquots at -80°C. The protein concentration was measured with the Qubit protein assay kit (Invitrogen, ThermoFisher Scientific: #Q33211).

T cell stimulation and measurement of IL-17A

Freshly isolated T cells were plated at $2*10^6$ cells/well in 48-well plates and stimulated with 40µg/ml *C. albicans* SC5314 lysate, 40µg/ml *D. hansenii* CBS767 lysate, 1µg/ml PepTivator® *C. albicans* mp65 peptide pool mix (Miltenyi, #130-096-776), or medium as a negative control, in a final volume of 500µl. For the positive control, wells were precoated with 1µg/ml antihuman CD3 antibody (Miltenyi, #130-093-387) at 37°C for 2h. All wells were supplemented with 1µg/ml anti-human CD28 antibody (Miltenyi, #130-093-375). The plates were incubated for 48h at 37°C with 5% CO2. All samples were prepared in duplicates. After incubation, supernatants were frozen at -80°C until IL-17A measurement. Antigen-specific IL-17A levels were measured in thawed supernatants in duplicate using the IL-17A ELISA kit (Invitrogen, ThermoFisher Scientific: #BMS2017) according to the manufacturer's instructions. The standard curve was calculated from blank-curated mean standard values with a 4-parameter curve fit (R package dr4pl, v2.0.0). All sample values were blank-curated before concentration calculation via the standard curve formula.

Fecal DNA extraction, internal transcribed spacer 1 and 16S rRNA sequencing

Microbial DNA was extracted from stool samples using the DNeasy PowerSoil Kit (Qiagen, #12888-100) according to the manufacturer´s instructions. We divided the sample into 4 subsamples to increase efficiency of the beat-beating step.

The Illumina platform Miseq V3 with paired-end reads of 300 bp was used for all samples. For the ITS sequencing samples were processed by LGC Genomics GmbH. The ITS1 region was amplified using ITS1F/ITS2R primers. The total read count was on average 54,000 reads/sample. From the 246 total 16S rRNA sequencing samples, 149 16S rRNA sequencing samples that had not been previously analyzed were processed by LGC Genomics GmbH using sequencing primers 341F-785R, targeting the V3-V4 region. The total read count was on average 56,000 reads/sample. 97 16S rRNA sequencing samples from a previous study were processed as described in Rau et al[15].

Taxonomic profiling

Taxonomic annotation of fungal Internal Transcribed Spacer (ITS) was performed using the PIPITS pipeline[46] version 2.4, with default parameters including quality filtering, read-pair merging, ITS1 extraction and chimera removal. Remaining reads were binned based on 97% similarity as operational taxonomic units (OTUs) and aligned with QIIME[47] to the UNITE fungi database[48] using mothur classifier. Samples were then normalized by cumulative sum scaling using the R package metagenomeSeq. Due to the complex fungal taxonomy, we grouped fungi according to genus but used the CTG-clade to characterize the *Candida* genus.

For the 16S rRNA sequencing data, quality control to remove low-quality reads and taxonomic annotation was performed using QIIME[47]. Raw reads were joined and trimmed with cutadapt to remove the primer sequences. Deblur workflow was used for filtering and denoising the joined reads. Assigning taxonomic information to each amplicon sequence variant (ASV) was performed using a Naive Bayes classifier with 99% similarity in QIIME. The classifier was fitted to the appropriate rRNA gene region (V3-V4) with the SILVA 132 database[49].

Diversity analysis

Alpha diversity indices detailing mycobiome community composition within samples were calculated using the R package vegan. Testing for significant differences in alpha diversity was performed using Wilcoxon rank-sum test. For estimating beta diversity reflecting community dissimilarities, *cmultRepl* function from R package zCompositions was first used to perform Bayesian-Multiplicative replacement of count zeros to the raw OTU table. Aitchison distances were calculated using *aDist* function from the R package robCompositions. We performed Partial Least Squares Discriminant Analysis (PLS-DA) using the mycobiome Aitchison distance matrix with the R package mixOmics. To test for significant differences in the mycobiome composition, permutational multivariate analysis of variance (PERMANOVA) as implemented in the function *adonis* from R package vegan adjusting for age, gender, obesityrelated parameters (age, gender, BMI, DM, aHT and hyperlipidemia) was used. Mycobiome community and clinical data (age, gender, height, weight, BMI, AST and ALT) were fit onto the ordination using the function *envfit* from vegan R package.

Statistics

Associations between the SNP genotypes and fibroscan values or grouped fibrosis (cutoff 9,7kPa) were investigated with generalized linear models adjusting for age, BMI, gender and PNPLA3 rs738409 with the *glm* function of the R package stats. Due to its potential as a genetic risk factor for NAFLD[27], we additionally adjusted for the PNPLA3 rs738409 genotype in all SNP-based generalized linear model (glm) calculations. The PNPLA3 rs738409 genotyping data was available for samples of our NAFLD patient cohort data but were generated in a previous study[28]. Statistical analysis of this data with the *glm* function confirmed primary findings from this study (Fig. 2c).

Correlations between mycobiome and clinical data were assessed by Spearman's correlation adjusting for age, gender, and obesity-related parameters (age, sex, BMI, DM, aHT and hyperlipidemia) using the function *pcor.test* from R package ppcor. Differentially abundant genera were identified by the Wilcoxon rank-sum test using R package stats, and by a generalized linear model adjusting for previously mentioned parameters (genus ~ fibroscan.group + age + gender + BMI + DM + aHT + hyperlipidemia), with *glm* function from R package stats. Association between the presence or absence of *Candida* CTG-clade and the fibrosis state was calculated by the Fisher test, using the *fisher.test* function from R package stats. A generalized linear model adjusting for previously mentioned parameters was used to study the association between *Candida* CTG-clade and fibroscan value (genus ~ fibroscan + age + gender + BMI + DM + aHT + hyperlipidemia), with *glm* function from R package stats. When exploring all data, the antibiotic intake was included for adjustment when appropriate.

Microbiome community network analysis

In order to explore the associations between microbial genera and NAFLD progression, we built a correlation network using mycobiome and bacteriome data. To build the correlation network, the *correlate_fastspar* function from R was used. The SparCC method was used to calculate the interactions between pairs of co-abundant taxa. The significance threshold for correlations was set to 0.05. Clustering of the network taxa was performed using *cluster_fast_greedy* function from igraph R package. Correlations between taxa and clinical variables (fibroscan, weight, AST, ALT and GGT) were calculated using *pcor.test* adjusting for age, gender and obesity-related parameters. To explore the association between clusters and NAFLD, the *fisher.test* function from R package stats was used. A fungal/bacterial genus was considered to be correlated with NAFLD progression if there was at least one significant correlation with fibroscan, AST, ALT or GGT.

Data visualization

Figures were generated by R software 3.6.3, using ggplot2 package.

DATA AVAILABILITY

Raw sequences from ITS1 gene sequencing were registered at NCBI under BioProject PRJNA834619.

ACKNOWLEDGMENTS

This project was funded by the IZKF Würzburg (project A-401), the Marie Sklodowska-Curie Actions (MSCA), and Innovative Training Networks, H2020-MSCA-ITN-2018, 813781 "BestTreat". GP would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy for intellectual input. We want to thank Carolin Spielau-Romer and Lea Strobel for their contribution to this project.

AUTHOR CONTRIBUTIONS

O.K., A.G., and T.D. conceived and designed the study. A.G., H.H., and M.R. recruited the participants and were responsible for clinical data collection. M.R. and H.H. collected fecal samples and extracted DNA from feces together with N.T.. A.M.A., and M.H. collected blood samples for the T cell stimulation assay, J.L. provided additional samples for this assay. H.H. and N.T. extracted DNA from blood and PBMC samples. R.M. and N.E.N. were involved in planning of experimental analyses. N.T. performed and analyzed the experimental analyses. S.L.S. and M.M. performed the metagenomics analyses. O.K., G.P., and A.G. led and supervised the research work. N.T. and S.L.S. wrote the manuscript. G.P., O.K., A.G., and M.R. edited the manuscript. All authors reviewed and made substantial contributions and approved the final version of the manuscript.

COMPETING INTERESTS

The authors declared no conflict of interest.

ABBREVIATIONS

aHT: arterial hypertension, ALD: alcohol-associated liver disease, ALT: alanine aminotransferase, AST: aspartate aminotransferase, BCAA: branched-chain amino acid, *C. albicans*: *Candida albicans*, *D. hansenii*: *Debaryomyces hansenii*, GLM: generalized linear model, HC: healthy control, MAF: minor allele frequency, NAFL: non-alcoholic fatty liver, NAFLD: non-alcoholic fatty liver disease, NASH: non-alcoholic steatohepatitis, OTU: operational taxonomic unit, TE; transient elastography, rTreg: resting regulatory T cells, *S.cerevisiae*: *Saccharomyces cerevisiae*, SCFA: short-chain fatty acid

FIGURES AND LEGENDS

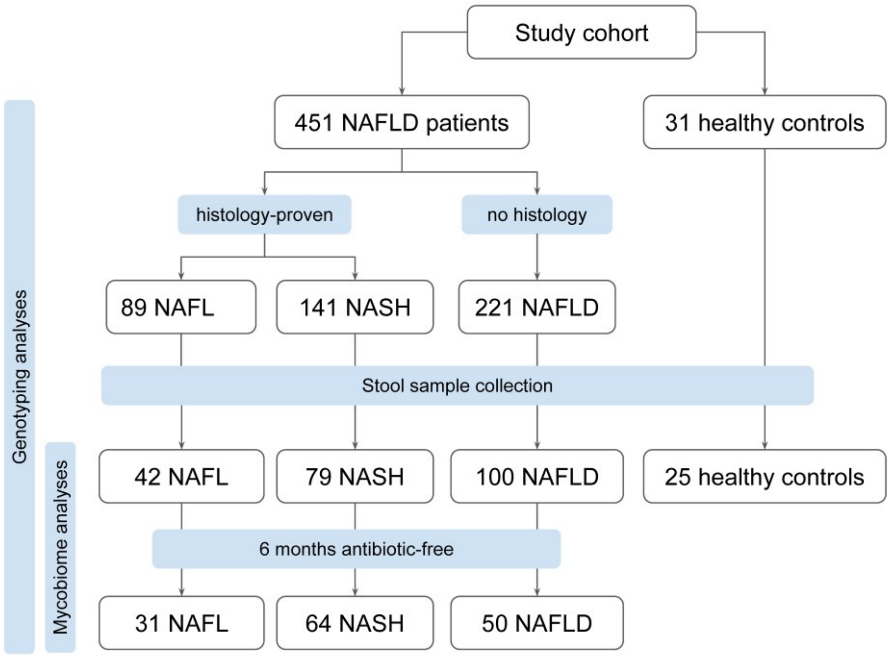


Fig. 1| Flow diagram with the overview of the study participants.

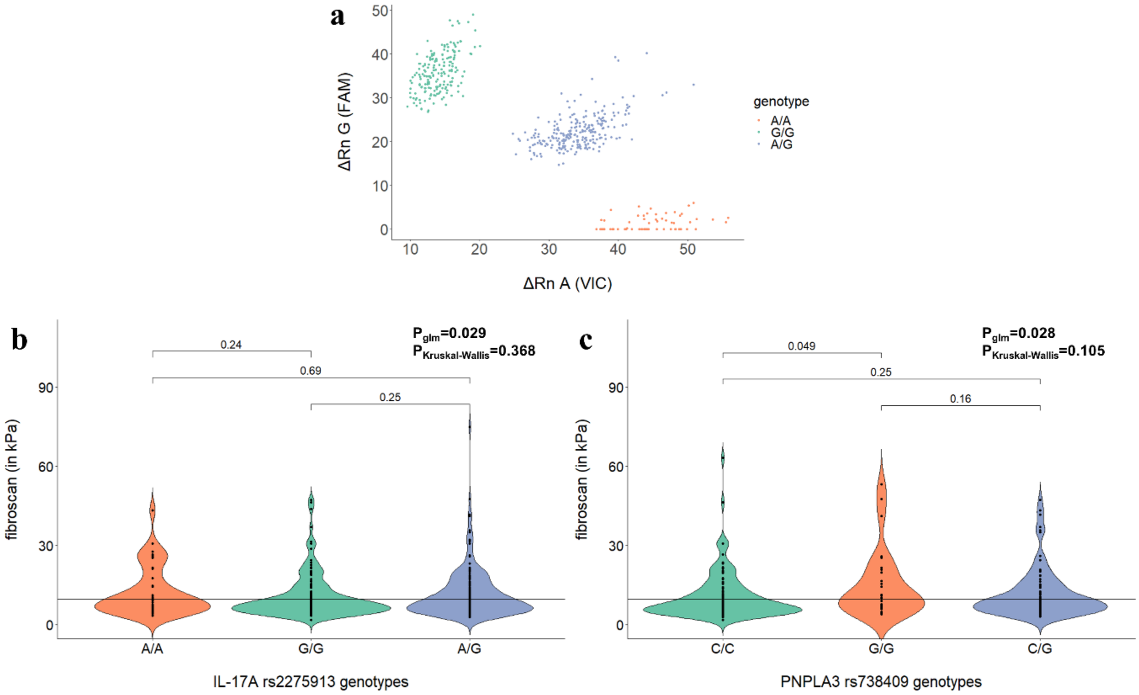


Fig. 2| The IL-17A rs2275913 genotype is associated with liver stiffness in NAFLD. a, Allelic discrimination Plot after TaqMan SNP Genotyping. **b**, Violin Plot for visualization of genotype association with fibrosis as assessed by fibroscan. Statistical comparison was performed using Kruskal-Wallis Test ($P_{\text{Kruskal-Wallis}}$) and generalized linear models adjusted for age, gender, BMI, PNPLA3 rs738409 genotype (P_{glm}). **c**, Violin Plot for visualization of known PNPLA3 risk variant rs738409 association with fibrosis as assessed by fibroscan. Statistical comparison was performed using Kruskal-Wallis Test ($P_{\text{Kruskal-Wallis}}$) and generalized linear models adjusted for age, gender and BMI (P_{glm}).

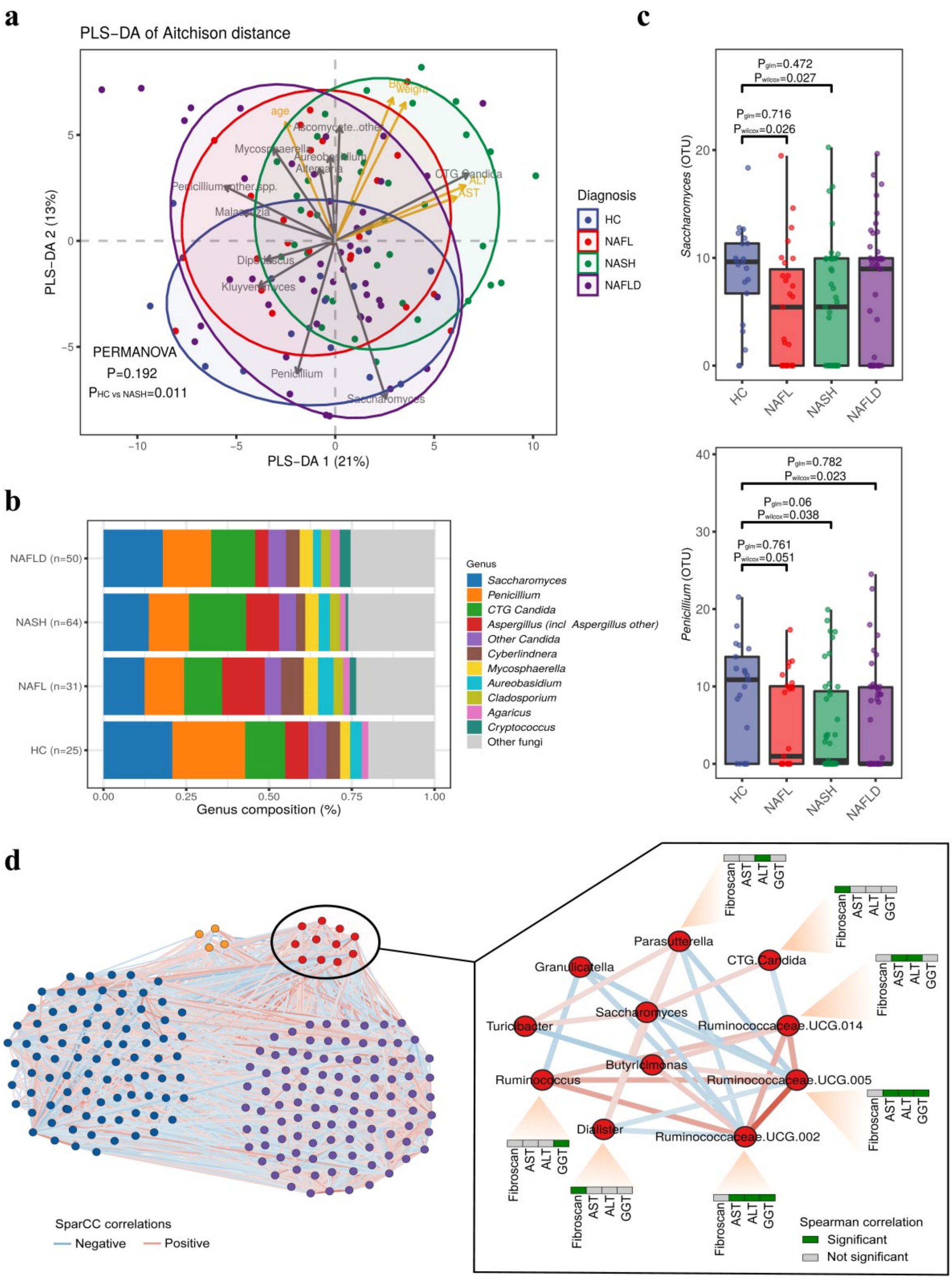


Fig. 3| Mycobiome changes in the different diagnosed groups and healthy controls and microbial community network. a, Beta diversity. PLS-DA of Aitchison distance of the mycobiome composition by diagnosis. **b**, Overview of mycobiome composition at genus level in NAFLD, NAFL, NASH, and HC groups. **c**, Boxplot of Saccharomyces and Penicillium abundances. Statistical comparison between groups (HC, NAFL, NASH, and NAFLD) was performed using Wilcoxon rank-sum test (Pwilcoxon) and generalized linear models adjusting for age, gender and obesity-related parameters (Pglm). **d**, Microbial community network showing the 4 subcommunity modules. Significant negative correlations are shown in blue and positive in red. The module significantly associated with NAFLD-related parameters is shown with red nodes and significant correlations between the genera and fibroscan, AST, ALT, and GGT are shown in green.

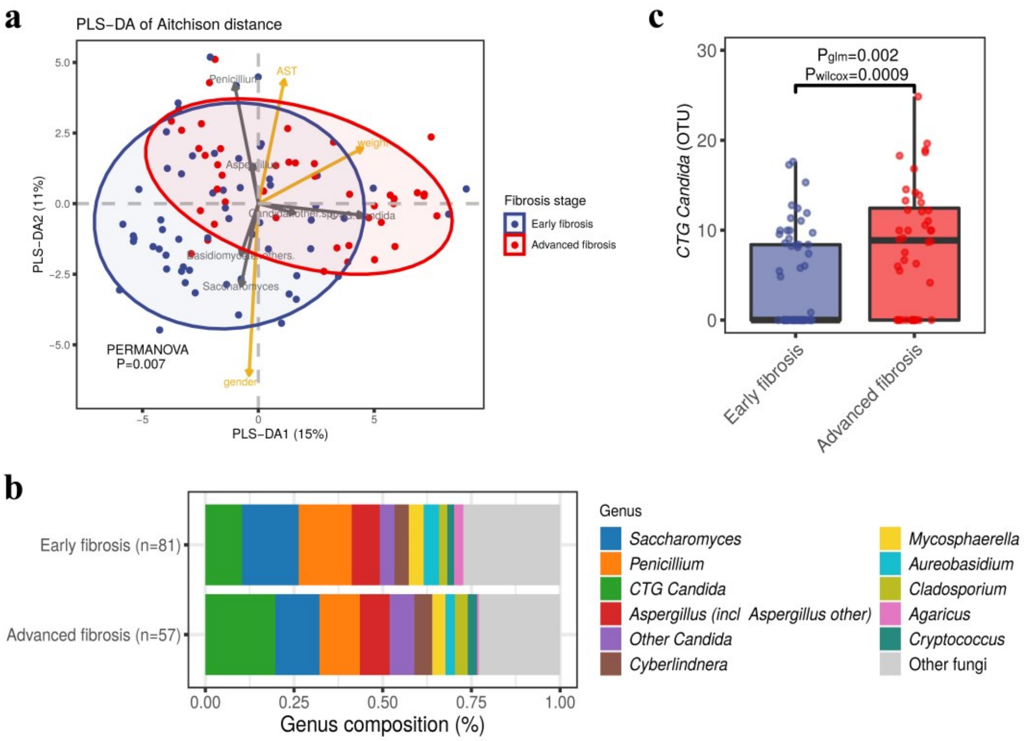


Fig. 4| Mycobiome changes by fibroscan-based fibrosis groups. a, Beta diversity. PLS-DA of Aitchison distance of the mycobiome composition by fibrosis stage group. **b**, Overview of mycobiome composition at genus level in early and advanced fibrosis groups (cut-off </> 9.7 kPa). **c**, Boxplot of *Candida* CTG-clade abundances. Statistical comparison between early and advanced fibrosis was performed using Wilcoxon rank-sum test ($P_{wilcoxon}$) and generalized linear models adjusting for age, gender and obesity-related parameters (P_{glm}).

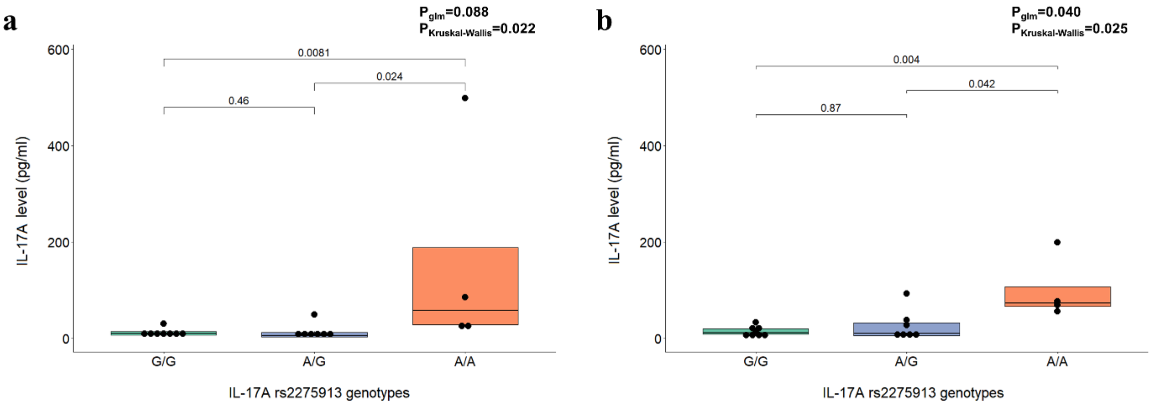
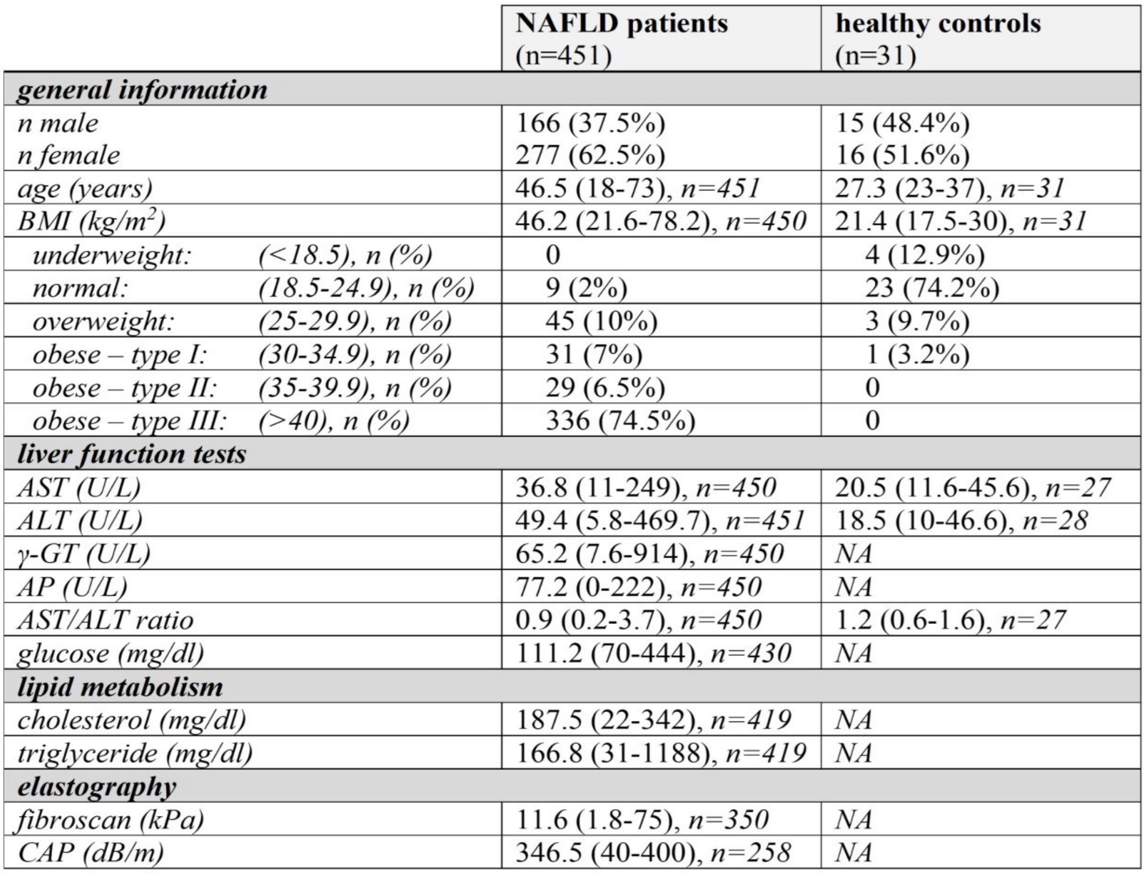


Fig. 5| Impaired IL-17A production in T cells from subjects homozygous for the rs2275913 minor allele variant. T cells were stimulated with fungal lysates and IL-17A concentrations in samples were measured by ELISA and calculated with a 4-parameter standard fit curve. 19 subjects were included in this assay (G/G: n=8, A/G: n=7, A/A: n=4). **a**, IL-17A secretion after stimulation with *C. albicans* lysate. **b**, IL-17A secretion after stimulation with *D. hansenii* lysate. Statistical comparisons for **a** and **b** were performed using Kruskal-Wallis Test ($P_{Kruskal}$Wallis) and generalized linear models (P_{glm}).

TABLES WITH TITLES AND LEGENDS

Table 1| NAFLD patient cohort characteristics. Values are shown as means and range.



	NAFLD patients (n=451)	**healthy controls** (n=31)
general information		
n male *n female*	166 (37.5%) 277 (62.5%)	15 (48.4%) 16 (51.6%)
age (years)	46.5 (18-73), *n=451*	27.3 (23-37), *n=31*
BMI (kg/m^2)	46.2 (21.6-78.2), *n=450*	21.4 (17.5-30), *n=31*
underweight: (<18.5), n (%)	0	4 (12.9%)
normal: (18.5-24.9), n (%)	9 (2%)	23 (74.2%)
overweight: (25-29.9), n (%)	45 (10%)	3 (9.7%)
obese – type I: (30-34.9), n (%)	31 (7%)	1 (3.2%)
obese – type II: (35-39.9), n (%)	29 (6.5%)	0
obese – type III: (>40), n (%)	336 (74.5%)	0
liver function tests		
AST (U/L)	36.8 (11-249), *n=450*	20.5 (11.6-45.6), *n=27*
ALT (U/L)	49.4 (5.8-469.7), *n=451*	18.5 (10-46.6), *n=28*
γ-GT (U/L)	65.2 (7.6-914), *n=450*	*NA*
AP (U/L)	77.2 (0-222), *n=450*	*NA*
AST/ALT ratio	0.9 (0.2-3.7), *n=450*	1.2 (0.6-1.6), *n=27*
glucose (mg/dl)	111.2 (70-444), *n=430*	*NA*
lipid metabolism		
cholesterol (mg/dl)	187.5 (22-342), *n=419*	*NA*
triglyceride (mg/dl)	166.8 (31-1188), *n=419*	*NA*
elastography		
fibroscan (kPa)	11.6 (1.8-75), *n=350*	*NA*
CAP (dB/m)	346.5 (40-400), *n=258*	*NA*

REFERENCES

1 Younossi, Z. *et al.* Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol* **15**, 11-20, doi:10.1038/nrgastro.2017.109 (2018).

2 Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of NAFLD development and therapeutic strategies. *Nat Med* **24**, 908-922, doi:10.1038/s41591-018-0104-9 (2018).

3 Donnelly, K. L. *et al.* Sources of fatty acids stored in liver and secreted via lipoproteins in patients with nonalcoholic fatty liver disease. *J Clin Invest* **115**, 1343-1351, doi:10.1172/JCI23621 (2005).

4 Buzzetti, E., Pinzani, M. & Tsochatzis, E. A. The multiple-hit pathogenesis of nonalcoholic fatty liver disease (NAFLD). *Metabolism* **65**, 1038-1048, doi:10.1016/j.metabol.2015.12.012 (2016).

5 Tang, Y. *et al.* Interleukin-17 exacerbates hepatic steatosis and inflammation in nonalcoholic fatty liver disease. *Clin Exp Immunol* **166**, 281-290, doi:10.1111/j.13652249.2011.04471.x (2011).

6 Harley, I. T. *et al.* IL-17 signaling accelerates the progression of nonalcoholic fatty liver disease in mice. *Hepatology* **59**, 1830-1839, doi:10.1002/hep.26746 (2014).

7 Rau, M. *et al.* Progression from Nonalcoholic Fatty Liver to Nonalcoholic Steatohepatitis Is Marked by a Higher Frequency of Th17 Cells in the Liver and an Increased Th17/Resting Regulatory T Cell Ratio in Peripheral Blood and in the Liver. *J Immunol* **196**, 97-105, doi:10.4049/jimmunol.1501175 (2016).

8 Tripathi, A. *et al.* The gut-liver axis and the intersection with the microbiome. *Nat Rev Gastroenterol Hepatol* **15**, 397-411, doi:10.1038/s41575-018-0011-z (2018).

9 Sharma, S. & Tripathi, P. Gut microbiome and type 2 diabetes: where we are and where to go? *J Nutr Biochem* **63**, 101-108, doi:10.1016/j.jnutbio.2018.10.003 (2019).

10 Palmas, V. *et al.* Gut microbiota markers associated with obesity and overweight in Italian adults. *Sci Rep* **11**, 5532, doi:10.1038/s41598-021-84928-w (2021).

11 Lang, S. & Schnabl, B. Microbiota and Fatty Liver Disease-the Known, the Unknown, and the Future. *Cell Host Microbe* **28**, 233-244, doi:10.1016/j.chom.2020.07.007 (2020).

12 Sharpton, S. R., Schnabl, B., Knight, R. & Loomba, R. Current Concepts, Opportunities, and Challenges of Gut Microbiome-Based Personalized Medicine in Nonalcoholic Fatty Liver Disease. *Cell Metab* **33**, 21-32, doi:10.1016/j.cmet.2020.11.010 (2021).

13 Leung, H. *et al.* Risk assessment with gut microbiome and metabolite markers in NAFLD development. *Sci Transl Med* **14**, eabk0855, doi:10.1126/scitranslmed.abk0855 (2022).

14 Liu, Y. *et al.* Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. *Cell Metab* **34**, 719-730 e714, doi:10.1016/j.cmet.2022.03.002 (2022).

15 Rau, M. *et al.* Fecal SCFAs and SCFA-producing bacteria in gut microbiome of human NAFLD as a putative link to systemic T-cell activation and advanced disease. *United European Gastroenterol J* **6**, 1496-1507, doi:10.1177/2050640618804444 (2018).

16 Nash, A. K. *et al.* The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome* **5**, 153, doi:10.1186/s40168-017-0373-4 (2017).

17 Thielemann, N., Herz, M., Kurzai, O. & Martin, R. Analyzing the human gut mycobiome - A short guide for beginners. *Comput Struct Biotechnol J* **20**, 608-614, doi:10.1016/j.csbj.2022.01.008 (2022).

18 Naglik, J. R., Konig, A., Hube, B. & Gaffen, S. L. Candida albicans-epithelial interactions and induction of mucosal innate immunity. *Curr Opin Microbiol* **40**, 104112, doi:10.1016/j.mib.2017.10.030 (2017).

19 Patin, E. C., Thompson, A. & Orr, S. J. Pattern recognition receptors in fungal immunity. *Semin Cell Dev Biol* **89**, 24-33, doi:10.1016/j.semcdb.2018.03.003 (2019).

20 Bacher, P. *et al.* Human Anti-fungal Th17 Immunity and Pathology Rely on CrossReactivity against Candida albicans. *Cell* **176**, 1340-1355 e1315, doi:10.1016/j.cell.2019.01.041 (2019).

21 Plato, A., Hardison, S. E. & Brown, G. D. Pattern recognition receptors in antifungal immunity. *Semin Immunopathol* **37**, 97-106, doi:10.1007/s00281-014-0462-4 (2015).

22 Drummond, R. A., Saijo, S., Iwakura, Y. & Brown, G. D. The role of Syk/CARD9 coupled C-type lectins in antifungal immunity. *Eur J Immunol* **41**, 276-281, doi:10.1002/eji.201041252 (2011).

23 Demir, M. *et al.* The fecal mycobiome in non-alcoholic fatty liver disease. *J Hepatol* **76**, 788-799, doi:10.1016/j.jhep.2021.11.029 (2022).

24 Chu, H. *et al.* The Candida albicans exotoxin candidalysin promotes alcohol-associated liver disease. *J Hepatol* **72**, 391-400, doi:10.1016/j.jhep.2019.09.029 (2020).

25 Moyes, D. L. *et al.* Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature* **532**, 64-68, doi:10.1038/nature17625 (2016).

26 Sarlos, P. *et al.* Genetic update on inflammatory factors in ulcerative colitis: Review of the current literature. *World J Gastrointest Pathophysiol* **5**, 304-321, doi:10.4291/wjgp.v5.i3.304 (2014).

27 Salari, N. *et al.* Association between PNPLA3 rs738409 polymorphism and nonalcoholic fatty liver disease: a systematic review and meta-analysis. *BMC Endocr Disord* **21**, 125, doi:10.1186/s12902-021-00789-4 (2021).

28 Arslanow, A. *et al.* The common PNPLA3 variant p.I148M is associated with liver fat contents as quantified by controlled attenuation parameter (CAP). *Liver Int* **36**, 418-426, doi:10.1111/liv.12937 (2016).

29 Turner, S. A. & Butler, G. The Candida pathogenic species complex. *Cold Spring Harb Perspect Med* **4**, a019778, doi:10.1101/cshperspect.a019778 (2014).

30 Borman, A. M. & Johnson, E. M. Name Changes for Fungi of Medical Importance, 2018 to 2019. *J Clin Microbiol* **59**, doi:10.1128/JCM.01811-20 (2021).

31 Seelbinder, B. *et al.* Antibiotics create a shift from mutualism to competition in human gut communities with a longer-lasting impact on fungi than bacteria. *Microbiome* **8**, 133, doi:10.1186/s40168-020-00899-6 (2020).

32 Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064-1066, doi:10.1093/bioinformatics/bty734 (2019).

33 Eddowes, P. J. *et al.* Accuracy of FibroScan Controlled Attenuation Parameter and Liver Stiffness Measurement in Assessing Steatosis and Fibrosis in Patients With Nonalcoholic Fatty Liver Disease. *Gastroenterology* **156**, 1717-1730, doi:10.1053/j.gastro.2019.01.042 (2019).

34 Butler, G. *et al.* Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* **459**, 657-662, doi:10.1038/nature08064 (2009).

35 Ramos-Moreno, L., Ruiz-Perez, F., Rodriguez-Castro, E. & Ramos, J. Debaryomyces hansenii Is a Real Tool to Improve a Diversity of Characteristics in Sausages and DryMeat Products. *Microorganisms* **9**, doi:10.3390/microorganisms9071512 (2021).

36 Ramani, K. & Biswas, P. S. Interleukin-17: Friend or foe in organ fibrosis. *Cytokine* **120**, 282-288, doi:10.1016/j.cyto.2018.11.003 (2019).

37 Shao, T. Y. *et al.* Commensal Candida albicans Positively Calibrates Systemic Th17 Immunological Responses. *Cell Host Microbe* **25**, 404-417 e406, doi:10.1016/j.chom.2019.02.004 (2019).

38 Diezmann, S., Cox, C. J., Schonian, G., Vilgalys, R. J. & Mitchell, T. G. Phylogeny and evolution of medical species of Candida and related taxa: a multigenic analysis. *J Clin Microbiol* **42**, 5624-5635, doi:10.1128/JCM.42.12.5624-5635.2004 (2004).

39 Frau, A. *et al.* DNA extraction and amplicon production strategies deeply influence the outcome of gut mycobiome studies. *Sci Rep* **9**, 9328, doi:10.1038/s41598-019-44974-x (2019).

40 Hartmann, P. *et al.* Dynamic Changes of the Fungal Microbiome in Alcohol Use Disorder. *Front Physiol* **12**, 699253, doi:10.3389/fphys.2021.699253 (2021).

41 Li, X. V. *et al.* Immune regulation by fungal strain diversity in inflammatory bowel disease. *Nature* **603**, 672-678, doi:10.1038/s41586-022-04502-w (2022).

42 Jain, U. *et al.* Debaryomyces is enriched in Crohn's disease intestinal tissue and impairs healing in mice. *Science* **371**, 1154-1159, doi:10.1126/science.abd0919 (2021).

43 Ochangco, H. S. *et al.* In vitro investigation of Debaryomyces hansenii strains for potential probiotic properties. *World J Microbiol Biotechnol* **32**, 141, doi:10.1007/s11274-016-2109-1 (2016).

44 Tovo, C. V. *et al.* Transient hepatic elastography has the best performance to evaluate liver fibrosis in non-alcoholic fatty liver disease (NAFLD). *Ann Hepatol* **18**, 445-449, doi:10.1016/j.aohep.2018.09.003 (2019).

45 European Association for Study of, L. & Asociacion Latinoamericana para el Estudio del, H. EASL-ALEH Clinical Practice Guidelines: Non-invasive tests for evaluation of liver disease severity and prognosis. *J Hepatol* **63**, 237-264, doi:10.1016/j.jhep.2015.04.006 (2015).

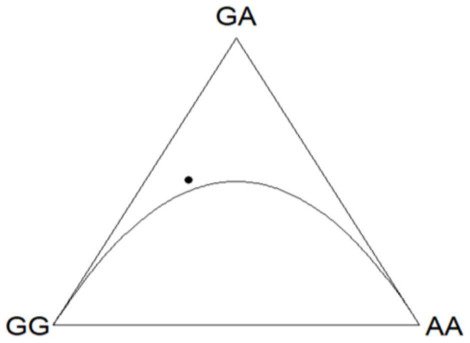
46 Gweon, H. S. *et al.* PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods Ecol Evol* **6**, 973-980, doi:10.1111/2041-210X.12399 (2015).

47 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).

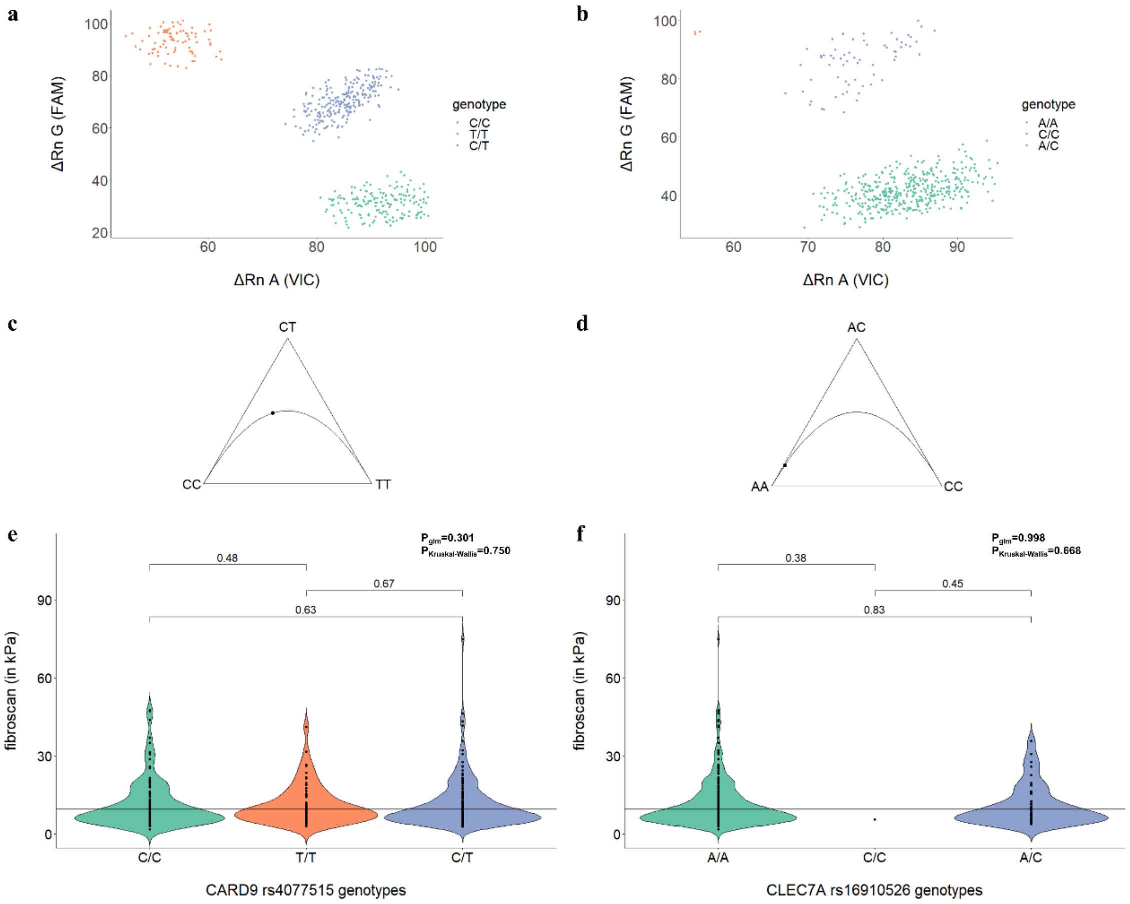
48 Nilsson, R. H. *et al.* The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* **47**, D259-D264, doi:10.1093/nar/gky1022 (2019).

49 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).

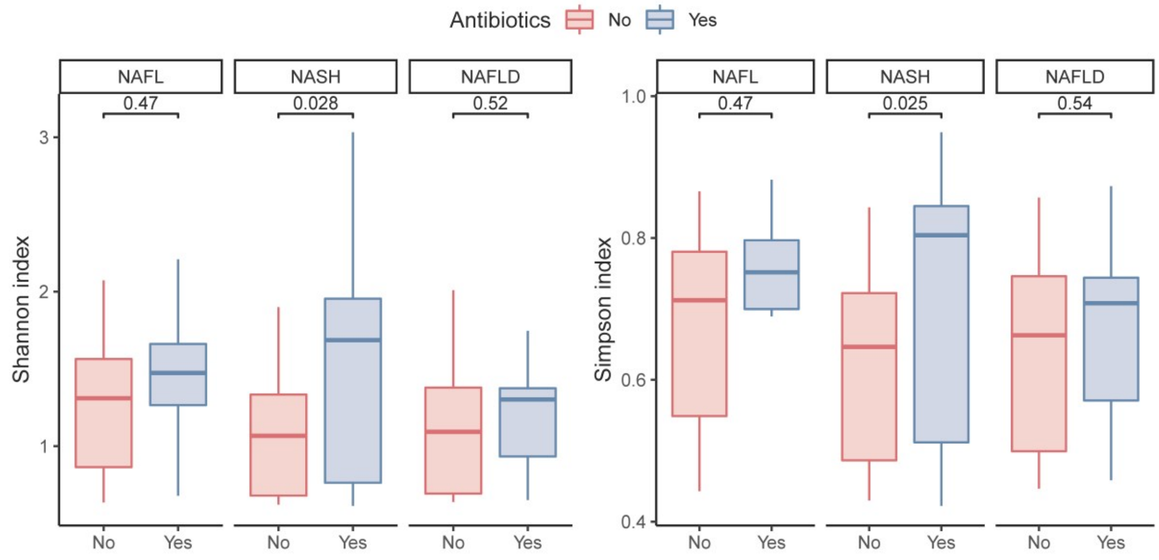
EXTENDED DATA



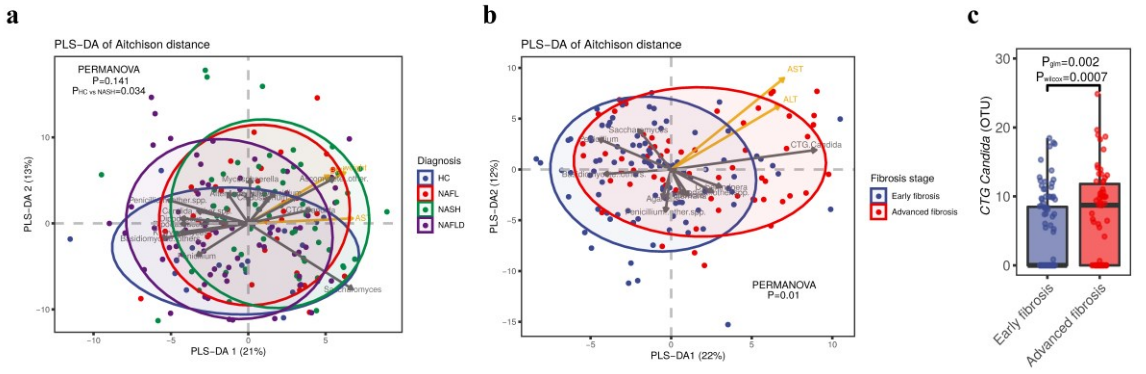
Extended Data Fig. 1| Ternary Plot of IL-17A data. IL-17A data are in Hardy-Weinberg equilibrium and thereby selection for specific genotypes was excluded.



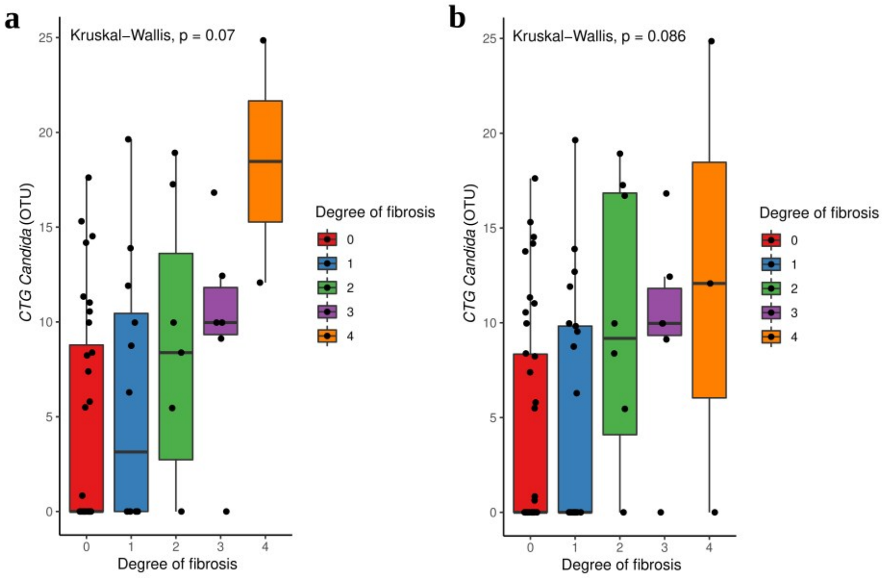
Extended Data Fig. 2| TaqMan SNP genotyping data for CARD9 rs4077515 & CLEC7A rs16910526. Allelic discrimination plots after genotyping for CARD9 rs4077515 **a**, and CLEC7A rs16910526 **b**, Ternary Plot for evaluation of Hardy-Weinberg equilibrium for CARD9 rs4077515 **c**, and CLEC7A rs16910526 **d**, Violin Plot for visualization of genotype association for CARD9 rs4077515 **e**, and CLEC7A rs16910526 **f**, to fibroscan values. Statistical comparisons were performed using generalized linear models adjusted for age, gender, BMI, PNPLA3 rs738409 genotype but were not significant (P_{glm} (rs4077515)=0.3, P_{glm}(rs16910526=0.5).



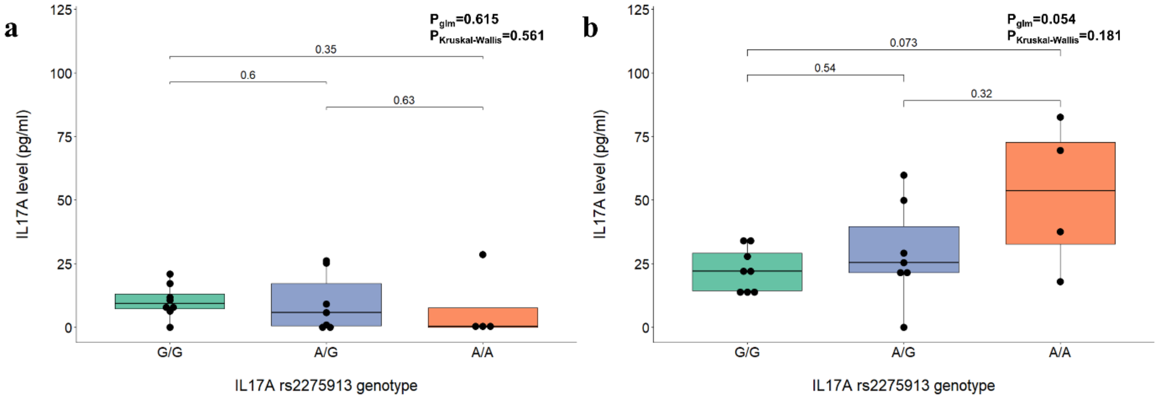
Extended Data Fig. 3| Comparison of Shannon and Simpson indexes between antibiotic-free subjects (No) and subjects that used antibiotics more than six months before the sample collection (Yes) in NAFL, NASH and NAFLD groups.



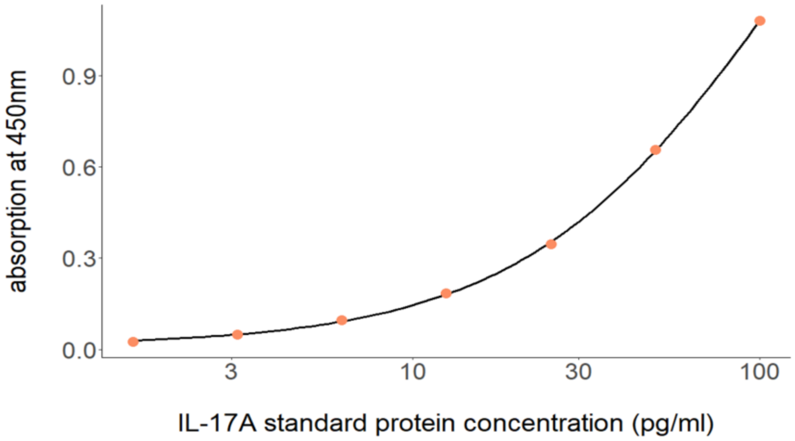
Extended Data Fig. 4| Mycobiome changes using the full cohort of samples. **a**, Beta diversity. PLS-DA of Aitchison distance of the mycobiome composition by diagnosis. **b**, Beta diversity. PLS-DA of Aitchison distance of the mycobiome composition by fibrosis stage group. **c**, Boxplot of *Candida* CTG-clade abundances. Statistical comparison between early and advanced fibrosis was performed using: Wilcoxon rank-sum test ($P_{wilcoxon}$) and generalized linear models adjusting for age, gender and obesity-related parameters and antibiotic intake (P_{glm}).



Extended Data Fig. 5| Boxplot of *Candida* CTG-clade abundances. **a**, Antibiotic-free set of samples. **b**, Full cohort. Statistical comparison between fibrosis stages (obtained by biopsy) were performed using Kruskal-Wallis test.

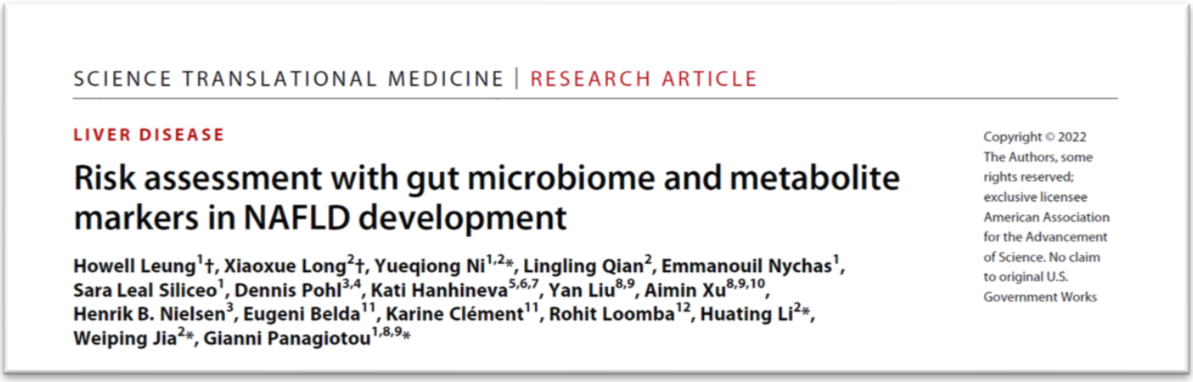


Extended Data Fig. 6| IL-17A production in T cells after stimulation with control samples. IL17A concentrations in samples were measured by ELISA and calculated with a 4-parameter standard fit curve (Extended Data Fig. 7). 19 subjects were included in this assay (G/G: n=8, A/G: n=7, A/A: n=4). **a**, IL-17A secretion without any additional stimulus (medium control). **b**, IL-17A secretion after stimulation with PepTivator® *C. Dalbicans* mp65 peptide pool mix. Statistical comparisons for **a** and **b** were performed using Kruskal-Wallis Test ($P_{Kruskal-Wallis}$) and generalized linear models (P_{glm}).



Extended Data Fig. 7| Standard curve for calculation of IL-17A secretion using 4-parameter curve fit. The standard curve was calculated with dr4pl R package and parameters were used to calculate the secreted IL-17A levels (y= ((122,9292*(((0,0124-2,4283)/(x-2,4283))1))^(1/1,1346))*2).

Manuscript IV



SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

LIVER DISEASE

Risk assessment with gut microbiome and metabolite markers in NAFLD development

Howell Leung[1]†, Xiaoxue Long[2]†, Yueqiong Ni[1,2]*, Lingling Qian[2], Emmanouil Nychas[1], Sara Leal Siliceo[1], Dennis Pohl[3,4], Kati Hanhineva[5,6,7], Yan Liu[8,9], Aimin Xu[8,9,10], Henrik B. Nielsen[3], Eugeni Belda[11], Karine Clément[11], Rohit Loomba[12], Huating Li[2]*, Weiping Jia[2]*, Gianni Panagiotou[1,8,9]*

Copyright © 2022
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim
to original U.S.
Government Works

Overview

In manuscript IV, we aimed to explore the potential value of the gut microbiome in the development of NAFLD and to build a machine learning model able to predict individuals at risk to develop NAFLD four years later. We demonstrated differences in the microbiome signature and metabolic shifts in subjects that will develop NAFLD compared to controls. In addition, we presented a machine learning model able to predict the progression to NAFLD with an auROC of 0.80. These results showed the biological relevance of the gut microbiome and potential microbial markers for early NAFLD diagnosis.

FORM I

Manuscript No: 4

Manuscript title: Risk assessment with gut microbiome and metabolite markers in NAFLD development

Authors: Howell Leung*, Xiaoxue Long*, Yueqiong Ni, Lingling Qian, Emmanouil Nychas, **Sara Leal Siliceo**, Dennis Pohl, Kati Hanhineva, Yan Liu, Aimin Xu, Henrik B. Nielsen, Eugeni Belda, Karine Clément, Rohit Loomba, Huating Li, Weiping Jia, Gianni Panagiotou

Bibliographic information (if published or accepted for publication: Citation): Leung et al. (2022). Risk assessment with gut microbiome and metabolite markers in NAFLD development. *Science translational medicine,* 14(648):eabk0855. https://doi.org/10.1126/scitranslmed.abk0855

The candidate is (Please tick the appropriate box)**:**

□ First author, □ Co-first author, □ Corresponding author, ☒ Co-author.

Status (if not published; "submitted for publication", "in preparation".): published

Authors' contributions (in %) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Leung, H.*	20%	70%		30%	
Long, X.*	20%		70%	5%	
Ni, Y.	20%			30%	
Nychas, E.		10%			
Leal Siliceo, S.		10%			
Panagiotou, G.	20%			30%	40%
Others	20%	10%	30%	5%	60%
Total:	100%	100%	100%	100%	100%

*Authors contributed equally

SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

LIVER DISEASE

Copyright © 2022
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim
to original U.S.
Government Works

Risk assessment with gut microbiome and metabolite markers in NAFLD development

Howell Leung[1]†, Xiaoxue Long[2]†, Yueqiong Ni[1,2]*, Lingling Qian[2], Emmanouil Nychas[1], Sara Leal Siliceo[1], Dennis Pohl[3,4], Kati Hanhineva[5,6,7], Yan Liu[8,9], Aimin Xu[8,9,10], Henrik B. Nielsen[3], Eugeni Belda[11], Karine Clément[11], Rohit Loomba[12], Huating Li[2]*, Weiping Jia[2]*, Gianni Panagiotou[1,8,9]*

A growing body of evidence suggests interplay between the gut microbiota and the pathogenesis of nonalcoholic fatty liver disease (NAFLD). However, the role of the gut microbiome in early detection of NAFLD is unclear. Prospective studies are necessary for identifying reliable, microbiome markers for early NAFLD. We evaluated 2487 individuals in a community-based cohort who were followed up 4.6 years after initial clinical examination and biospecimen sampling. Metagenomic and metabolomic characterizations using stool and serum samples taken at baseline were performed for 90 participants who progressed to NAFLD and 90 controls who remained NAFLD free at the follow-up visit. Cases and controls were matched for gender, age, body mass index (BMI) at baseline and follow-up, and 4-year BMI change. Machine learning models integrating baseline microbial signatures (14 features) correctly classified participants (auROCs of 0.72 to 0.80) based on their NAFLD status and liver fat accumulation at the 4-year follow up, outperforming other prognostic clinical models (auROCs of 0.58 to 0.60). We confirmed the biological relevance of the microbiome features by testing their diagnostic ability in four external NAFLD case-control cohorts examined by biopsy or magnetic resonance spectroscopy, from Asia, Europe, and the United States. Our findings raise the possibility of using gut microbiota for early clinical warning of NAFLD development.

INTRODUCTION

Since the 1980s, the prevalence of obesity, insulin resistance, type 2 diabetes mellitus, and obesity-associated nonalcoholic fatty liver disease (NAFLD) has grown worldwide (*1*–*3*). The occurrence of these interconnected diseases is partly driven by consumption of high-energy food and a sedentary lifestyle, and these diseases are considered critical global health and socioeconomic problems (*4*). Apart from associations with liver-related diseases, epidemiological studies have associated NAFLD with increased risk of developing extrahepatic chronic diseases, such as type 2 diabetes, cardiovascular disease, and chronic kidney disease (*5*, *6*). A recent cohort study showed that overall mortality risk increases progressively with worsening NAFLD histology, and even simple steatosis increases mortality risk by 71% (*7*), thus simple steatosis can no longer be considered as benign as previously thought (*8*). Although NAFLD affects about 25% of the world's population (*9*) and has a high disease burden, awareness of NAFLD is low. In a cross-sectional analysis (n = 2788) in four U.S. cities, NAFLD prevalence was 23.9%, whereas awareness of NAFLD was 2.4% in study participants with computed tomography (CT)–defined NAFLD (*10*). One important reason for low awareness is that most patients with NAFLD are largely asymptomatic in the disease course, where disease is mainly detected through an incidental finding of fatty liver on ultrasound or an imagining modality or routine laboratory testing (*11*, *12*). Diagnosis by liver biopsy or imaging is reliable but difficult for large-scale screening and monitoring. Thus, the need to identify individuals who are at high risk of developing NAFLD or are at an early stage of the disease is urgent, as lifestyle interventions can reverse the disease when it is in the first stages (*13*). According to one study (*14*), weight loss and healthy diet might be sufficient to reverse simple steatosis, whereas intensified lifestyle intervention coupled with pharmacological treatment might be necessary for more advanced stages of liver diseases. Exercise programs (*15*), low-carbohydrate diet (*16*), and various types of gut microbiota–targeted treatments (*17*) have demonstrated their ability to prevent steatosis development and improve NAFLD outcomes in human or preclinical models. Early diagnosis and interventions to prevent NAFLD progression can also greatly reduce future health care cost, as most economic costs associated with NAFLD are incurred in advanced stages (*18*). Currently available methods (*19*–*21*) for early prediction of NAFLD are limited and use only a few clinical parameters or biomarkers that may not reflect the heterogeneity and complexity of NAFLD (*22*, *23*). Thus, more convenient noninvasive alternatives are needed.

In the last 10 years, the gut microbiome has emerged as a major regulator of host energy homeostasis and substrate metabolism (*24*–*26*). The human gastrointestinal tract is colonized with 4644

---

[1]Systems Biology and Bioinformatics Unit, Leibniz Institute for Natural Product Research and Infection Biology–Hans Knöll Institute, Beutenbergstraße 11A, 07745 Jena, Germany. [2]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai Diabetes Institute, Shanghai Clinical Center for Diabetes, Shanghai Key Laboratory of Diabetes Mellitus, 200233 Shanghai, China. [3]Clinical Microbiomics, Fruebjergvej 3, 2100 Copenhagen, Denmark. [4]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Building 220, 2800 Kgs. Lyngby, Denmark. [5]Department of Life Technologies, Food Chemistry and Food Development Unit, University of Turku, 20014 Turku, Finland. [6]Department of Biology and Biological Engineering, Division of Food and Nutrition Science, Chalmers University of Technology, 412 96 Gothenburg, Sweden. [7]School of Medicine, Institute of Public Health and Clinical Nutrition, University of Eastern Finland, 70211 Kuopio, Finland. [8]The State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong SAR, China. [9]Department of Medicine, The University of Hong Kong, Hong Kong SAR, China. [10]Department of Pharmacology and Pharmacy, The University of Hong Kong, Hong Kong SAR, China. [11]Sorbonne Université, INSERM, NutriOmics Research Unit, Nutrition Department, Pitié-Salpêtrière Hospital, Assistance Publique-Hôpitaux de Paris, 75013 Paris, France. [12]NAFLD Research Center, Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA.
*Corresponding author. Email: yueqiong.ni@leibniz-hki.de (Y.N.); huarting99@sjtu.edu.cn (H.L.); wpjia@sjtu.edu.cn (W.J.); gianni.panagiotou@hki-jena.de (G.P.)
†These authors contributed equally to this work.

bacterial species encoding 171 million genes (*27*). Therefore, it is not unexpected that abnormalities in gut microbiome structure and especially function might affect the brain, adipose tissue, muscle, and liver metabolism. Microbial components or metabolites such as lipopolysaccharides, secondary bile acids, dimethyl- and trimethyl-amines, and compounds derived from carbohydrate and protein fermentation appear to be strongly involved in the gut host-microbiome metabolic axis and the occurrence of metabolic diseases (*28–31*).

Human cross-sectional studies have delineated the role of gut bacteria in the development of NAFLD. An increased ratio of Bacteroidetes to Firmicutes phyla and a decrease in butyrate-producing *Ruminococcaceae* are suggested to be involved in NAFLD progression; however, the data are not always consistent (*32–35*). Furthermore, whether NAFLD causes taxonomic and functional changes in the microbiome or the observed dysbiosis in patients with NAFLD leads to progression of the disease is not clear. For a possible causal role in NAFLD development, gut microbiota alteration should take place long before disease is diagnosed, which would suggest prognostic value in evaluating the gut microbiome in individuals with a high risk of developing NAFLD. To assess this potential value, we conducted a 4-year prospective study in a community-based cohort of 2487 Chinese individuals. We profiled 180 matched case-control individuals who were NAFLD free at baseline using well-documented clinical information and comprehensive metagenomic and metabolomic analysis. We developed machine learning models integrating baseline microbial signatures to classify individuals based on their NAFLD status 4 years after baseline (either remaining disease free or diagnosed with the disease). We also examined whether the selected features in the model were biologically relevant to NAFLD development by exploring the diagnostic power of the model in several case-control cohorts from Asia, the United States, and Europe.

RESULTS

Characterization of the study cohort

To develop a microbiome-based prognostic model for long-term development of NAFLD, we designed a nested case-control study within a community-based prospective cohort study of Chinese adults. About 2500 participants were screened in 2014 with ultrasonography, which is recommended as the first-line diagnostic test for NAFLD (*36*); 1216 participants were determined as NAFLD free using criteria proposed by the Asian Pacific Association for the Study of the Liver (*37*). Participant enrolment is outlined in fig. S1. Stool and serum samples were obtained from participants at baseline. At the follow-up visit in 2018, after a strict exclusion process, 90 participants (38 males and 52 females) were identified as having NAFLD (NAFLD$^{-/+}$) (Fig. 1). The participants in the NAFLD$^{-/+}$ group were matched with 90 controls who did not have NAFLD at baseline or at the follow-up visit (NAFLD$^{-/-}$). The two groups were matched in gender, age, and body mass index (BMI) at both the baseline and follow-up visits and 4-year change in BMI. There were no differences between the two groups in the prevalence of type 2 diabetes, hypertensive disease, metabolic syndrome, and medication usage at both baseline and follow-up in the cohort, apart from a significantly higher metabolic syndrome ratio in NAFLD$^{-/+}$ at follow-up as expected (chi-square test, $P < 0.05$; table S1).

Detailed baseline anthropometric parameters, glucose homeostasis parameters, serum liver enzymes and renal function, lipid profiles, and cytokines are shown in Table 1. No significant differences (*t* test, $P > 0.05$) were seen for most clinical parameters between the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups at baseline. Fasting insulin (FINS), homeostasis model assessment for insulin resistance (HOMA-IR), triglycerides (TGs), and high-sensitivity C-reactive protein (hs-CRP) in the NAFLD$^{-/+}$ group were slightly higher than in the NAFLD$^{-/-}$ group

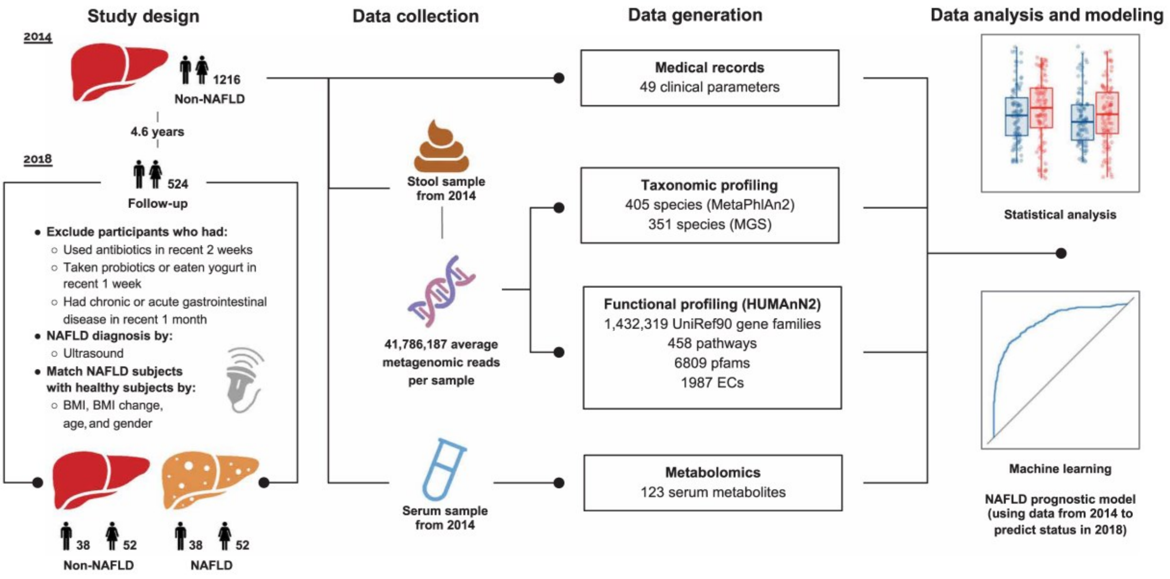

Study design
2014
1216
Non-NAFLD
4.6 years
2018
524
Follow-up
Exclude participants who had:
Used antibiotics in recent 2 weeks
Taken probiotics or eaten yogurt in recent 1 week
Had chronic or acute gastrointestinal disease in recent 1 month
NAFLD diagnosis by:
Ultrasound
Match NAFLD subjects with healthy subjects by:
BMI, BMI change, age, and gender
38
52
Non-NAFLD
38
52
NAFLD
Data collection
Stool sample from 2014
41,786,187 average metagenomic reads per sample
Serum sample from 2014
Data generation
Medical records
49 clinical parameters
Taxonomic profiling
405 species (MetaPhlAn2)
351 species (MGS)
Functional profiling (HUMAnN2)
1,432,319 UniRef90 gene families
458 pathways
6809 pfams
1987 ECs
Metabolomics
123 serum metabolites
Data analysis and modeling
Statistical analysis
Machine learning
NAFLD prognostic model (using data from 2014 to predict status in 2018)

Fig. 1. Overview of the prospective study design. A graphical representation summarizing the study design, data collection, and the methodologies of data generation and analysis. Further details of the study design can be found in fig. S1.

Table 1. Baseline characteristics of participants in NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups. Data are expressed as means ± SD or median (lower quartile and upper quartile) for continuous variables, and *n* represents percentage for categorical variables. BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting blood glucose; hs-CRP, high-sensitivity C-reactive protein; TC, total cholesterol; TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; FGF-21, fibroblast growth factor 21; HbA1c, hemoglobin A1C; HOMA-IR, homeostasis model assessment insulin resistance; apo, apolipoprotein; PG30, 30-min postprandial plasma glucose; PG120, 120-min postprandial plasma glucose; INS30, 30-min postprandial insulin; INS120, 120-min postprandial insulin; FINS, fasting insulin; Cr, creatinine; UAlb/Ucr, urinary albumin to creatinine ratio; UA, uric acid; FFA, free fatty acid; TBIL, total bilirubin; GA, glycated albumin; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transferase.

Characteristic	Total ($n = 180$)	NAFLD$^{-/-}$ ($n = 90$)	NAFLD$^{-/+}$ ($n = 90$)	*P* value*	*P* value†
Anthropometric parameters					
Sex (male)	76 (42.22%)	38 (42.22%)	38 (42.22%)	–	–
Age (years)	62.51 ± 3.81	62.03 ± 3.78	62.99 ± 3.81	0.0921	0.1481
Weight (kg)	62.82 ± 8.07	62.47 ± 7.53	63.16 ± 8.6	0.5688	0.3617
BMI (kg/m^2)	24.55 ± 2.13	24.35 ± 2	24.75 ± 2.25	0.2059	0.0681
SBP (mmHg)‡	130 (120, 140)	130 (120, 140)	130 (120, 140)	0.2854	0.3363
DBP (mmHg)‡	80 (80, 86)	80 (80, 84)	80 (80, 86)	0.7579	0.7304
Glucose homeostasis parameters					
FBG (mM)‡	5.93 (5.57, 6.36)	5.86 (5.54, 6.33)	6.02 (5.67, 6.39)	0.2866	0.8033
PG30 (mM)‡	10.29 (9.3, 11.45)	10.12 (9.19, 11.16)	10.34 (9.37, 11.78)	0.2219	0.4460
PG120 (mM)‡	7.72 (6.63, 9.3)	7.49 (6.17, 8.96)	7.85 (7.12, 9.58)	0.0612	0.2829
FINS (uU/ml)‡	5.2 (4.03, 7.11)	5.06 (3.99, 6.06)	5.42 (4.11, 8.08)	0.0266	0.8033
INS30 (uU/ml)‡	38.82 (26.11, 57.25)	36.25 (25.73, 50.64)	40.39 (26.77, 60.31)	0.2956	0.5751
INS120 (uU/ml)‡	37.62 (25.13, 57.42)	34.59 (21.78, 53.46)	41.86 (29.18, 60.03)	0.0562	0.3623
GA (%)‡	0.61 (0.57, 0.67)	0.62 (0.57, 0.68)	0.61 (0.57, 0.67)	0.6013	0.1853
HbA1c (%)‡	5.5 (5.2, 5.9)	5.4 (5.2, 5.7)	5.6 (5.3, 6)	0.0604	0.2707
HOMA-IR‡	1.35 (1.05, 1.99)	1.33 (1.02, 1.69)	1.5 (1.09, 2.21)	0.0266	–
HOMA-β‡	40.23 (31.6, 54.5)	37.68 (31, 52.44)	42.13 (32.76, 56.04)	0.1038	0.8752
Serum liver enzymes and renal function indexes					
ALT (IU/liter)‡	15 (12, 18)	15 (12, 17)	15 (13, 18)	0.2112	0.2477
AST (IU/liter)‡	21 (19, 23)	21 (19, 23)	21 (19, 23)	0.9994	0.8621
GGT (IU/liter)‡	18 (15, 25)	17 (14, 22)	20 (16, 27)	0.2764	0.2087
TBIL (μM)‡	10.7 (9, 14.4)	10.9 (9, 14)	10.7 (9, 14.5)	0.9518	0.7712
Cr (μM)‡	64 (56, 73)	64 (57, 73)	64 (55, 76)	0.8085	0.4501
UAlb/Ucr‡	6.79 (5.12, 12.29)	6.62 (5.16, 13.03)	6.84 (5.02, 11.91)	0.5568	0.4701
UA (μM)‡	295 (249, 341)	289.50 (243, 334)	302 (258, 342)	0.1618	0.0913
Lipid profiles					
TG (mM)‡	1.20 (0.86, 1.66)	1.07 (0.80, 1.53)	1.34 (0.90, 1.80)	0.0051	0.0193
TC (mM)‡	4.97 (4.44, 5.58)	4.80 (4.41, 5.57)	5.01 (4.46, 5.58)	0.4808	0.7310
FFA (μM)‡	497 (377, 666)	497.5 (371, 688)	497 (395, 650)	0.3526	0.5783
HDL-C (mM)‡	1.35 (1.14, 1.53)	1.40 (1.19, 1.64)	1.30 (1.09, 1.48)	0.0895	0.1451
LDL-C (mM)	3.08 ± 0.72	2.99 ± 0.74	3.17 ± 0.69	0.0965	0.1969
apoA-1 (g/liter)	1.49 ± 0.26	1.51 ± 0.28	1.46 ± 0.24	0.2124	0.3083
apoB (g/liter)	0.91 ± 0.16	0.9 ± 0.17	0.93 ± 0.16	0.1899	0.3731
apoE (mg/dl)‡	3.92 (3.28, 4.67)	3.84 (3.22, 4.56)	4.09 (3.4, 4.87)	0.0609	0.1709
Lipoprotein (a) (mg/dl)‡	14.05 (5.81, 25.07)	15.33 (6.12, 25.89)	12.51 (5.78, 23.24)	0.7654	0.6452
Cytokines					
FGF21 (pg/ml)‡	302.06 (180.19, 429.52)	268.84 (171.83, 389.11)	332.27 (236.74, 452.81)	0.9538	0.9304
hs-CRP (μg/ml)‡	0.63 (0.35, 1.17)	0.53 (0.28, 1.17)	0.72 (0.43, 1.16)	0.0351	0.0758

**P* value denotes differences between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ analyzed by *t* test without adjustment. †*P* value denotes differences between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ analyzed by analysis of covariance with HOMA-IR adjusted. ‡Log-transformed before analysis.

(t test, $P < 0.05$); however, their mean or median values were within reference ranges in both groups (FINS: 5.1 to 11.2 uU/ml, HOMA-IR < 2.5, TG < 1.70 mM, and hs-CRP < 1 μg/ml) (*38*–*41*). Only TGs remained significantly different after adjusting for HOMA-IR (1.55 ± 0.90 mM versus 1.23 ± 0.61 mM; Table 1).

Modest but distinguishable differences in baseline gut microbiome between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ individuals

We assessed the gut microbiome structure of the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups at baseline via shotgun metagenomic sequencing, generating 1128 gigabase pairs of high-quality reads with an average of 41,786,187 reads per sample (Fig. 1). Taxonomic profiling with MetaPhlAn2 (*42*) led to the identification of 405 species. Community alpha diversity measured as richness, and Shannon and Simpson indexes showed no significant differences (Wilcoxon rank-sum test, $P > 0.05$) at the species, genus, or family levels between the two groups (fig. S2A). Bray-Curtis, unweighted UniFrac, and weighted UniFrac distance comparisons indicated that the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups did not have significant community dissimilarities [permutational multivariate analysis of variance (PERMANOVA), $P > 0.05$; fig. S2, B and C]. The same patterns were observed when using a metagenomic species approach (*43*) for the taxonomic annotation (table S2).

In addition, we sequenced the baseline gut microbiota from 66 participants who were diagnosed as NAFLD in both 2014 and 2018 (NAFLD$^{+/+}$) and 34 participants who were diagnosed as NAFLD in 2014 but not in 2018 (NAFLD$^{+/-}$). These two groups were also matched with the other two groups described above by age, gender, BMI, and 4-year change in BMI. A thorough comparison of microbiota alpha and beta diversity among the four groups at baseline indicated that the two non-NAFLD groups were distinguishable from the two NAFLD groups ($P < 0.05$, Wilcoxon rank-sum test for alpha diversity comparisons and PERMANOVA for beta diversity comparisons using Bray-Curtis distances) (fig. S3). Moreover, the gut microbiota of NAFLD$^{-/+}$ subjects was different from that of NAFLD$^{+/+}$ and NAFLD$^{+/-}$ individuals. This argues that the NAFLD$^{-/+}$ group was not already diseased at the baseline because they clustered with NAFLD$^{-/-}$ subjects at baseline. Because our focus was to identify gut microbiota signatures in disease-free individuals suggestive of NAFLD predisposition, only the NAFLD$^{-/-}$ and NAFLD$^{-/+}$ groups were further analyzed.

A compositional analysis found that several of the 10 most abundant genera and species (Fig. 2A) were significantly associated (envfit from R package vegan, $P < 0.05$) with observed variation in the taxonomic profile of the study participants (fig. S2, B and C). However, their relative abundances were not significantly different (zero-inflated Gaussian mixture model, $P > 0.05$) between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups. Nevertheless, the relative abundances of 8 and 21 less-abundant genera and species, respectively, were significantly different (zero-inflated Gaussian mixture model, $P < 0.05$) between the two groups (fig. S2D). *Methanobrevibacter* [false discovery rate (FDR) = 0.01] was decreased in NAFLD$^{-/+}$ compared to NAFLD$^{-/-}$ (a reduction in *Phascolarctobacterium* was insignificant at FDR = 0.2). Lower abundances of these two genera have been observed in cohort studies in obese individuals compared to lean individuals (*44*, *45*). *Slackia* has been reported to be more abundant in individuals with moderate-to-severe fibrosis than in individuals with absent-to-mild fibrosis (*46*), and this genus was increased in the NAFLD$^{-/+}$ compared to the NAFLD$^{-/-}$ group (FDR = 0.06). The relative abundance of *Dorea formicigenerans*, a species that is highly abundant in people with obesity (*47*), was higher in the NAFLD$^{-/+}$ than the NAFLD$^{-/-}$ group (FDR = 0.17). Differences in the relative abundances of *Methanobrevibacter*, *Phascolarctobacterium*, *Slackia*, and *D. formicigenerans* between the two study groups remained significant even after adjusting for age, gender, BMI, and HOMA-IR (zero-inflated Gaussian mixture model, $P < 0.05$). Because the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups had no difference in BMI and in the aforementioned cohort studies, the liver status of the obese individuals was not evaluated, and our prospective design suggested that *Methanobrevibacter*, *Phascolarctobacterium*, *Slackia*, and *D. formicigenerans* could be signatures of NAFLD risk in addition to being obesity-related signatures.

We used HUMAnN2 (*48*) for functional profiling of the microbial communities and identified 458 pathways. Likewise, the taxonomic profile and the microbiota functional potential could not differentiate between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups by alpha and beta diversity (fig. S4, A and B). Four of the most abundant pathways detected, uridine monophosphate biosynthesis I, uridine diphosphate–*N*–acetylmuramoyl-pentapeptide biosynthesis I and II, and peptidoglycan biosynthesis I (Fig. 2A), were significantly associated with observed variation in the functional profiles of study participants (envfit from R package vegan, $P < 0.05$; fig. S4B). These pathways were proposed to be discriminatory for NAFLD cirrhosis against control groups in a recent U.S. cohort study (*49*); however, their relative abundances were not significantly different (zero-inflated Gaussian mixture model, $P > 0.05$) between the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups in our prospective study. Nevertheless, we found 19 biosynthetic pathways significantly different in relative abundance between the two groups (zero-inflated Gaussian mixture model, $P < 0.05$) (fig. S4C). We observed a significantly higher relative abundance of geranylgeranyl diphosphate biosynthesis and the mevalonate pathway in the NAFLD$^{-/-}$ group. These pathways are dysregulated in mice and humans with nonalcoholic steatohepatitis (NASH) (*50*). Two genes encoding enzymes involved in these pathways, hydroxymethylglutaryl–coenzyme A (CoA) reductase (EC 1.1.1.34) and mevalonate kinase (EC 2.7.1.36), were significantly enriched in the NAFLD$^{-/-}$ group (zero-inflated Gaussian mixture model, $P < 0.05$; table S3). *Methanobrevibacter smithii* was the major contributor of gene expression abundance of hydroxymethylglutaryl-CoA reductase (95%) and mevalonate kinase (40%). In contrast, the NAFLD$^{-/+}$ group had a higher relative abundance of phosphatidate metabolism and cholic acid degradation. Cholic acid is a primary bile acid that decreases substantially in rats on a Western diet and is proposed as an early marker of NAFLD development (*51*). Genes encoding phospholipase D (EC 3.1.4.4) and bile-acid-7-alpha-dehydratase (EC 4.2.1.106) were also significantly enriched in the NAFLD$^{-/+}$ group (zero-inflated Gaussian mixture model, $P < 0.05$, table S3). The four significant pathways above (geranylgeranyl diphosphate biosynthesis, mevalonate pathway, phosphatidate metabolism, and cholic acid degradation) remained significantly different (zero-inflated Gaussian mixture model, $P < 0.05$) between the two groups after adjusting for age, gender, BMI, and HOMA-IR, except for cholic acid degradation that was marginally significant ($P = 0.050$).

Metabolite enrichment and metabolic shifts in NAFLD$^{-/+}$ versus NAFLD$^{-/-}$ groups

We next performed targeted metabolomic analysis of serum samples collected at baseline to interrogate whether differences in species and pathway abundance of gut microbiota led to distinct profiles of microbial metabolites in the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups. We

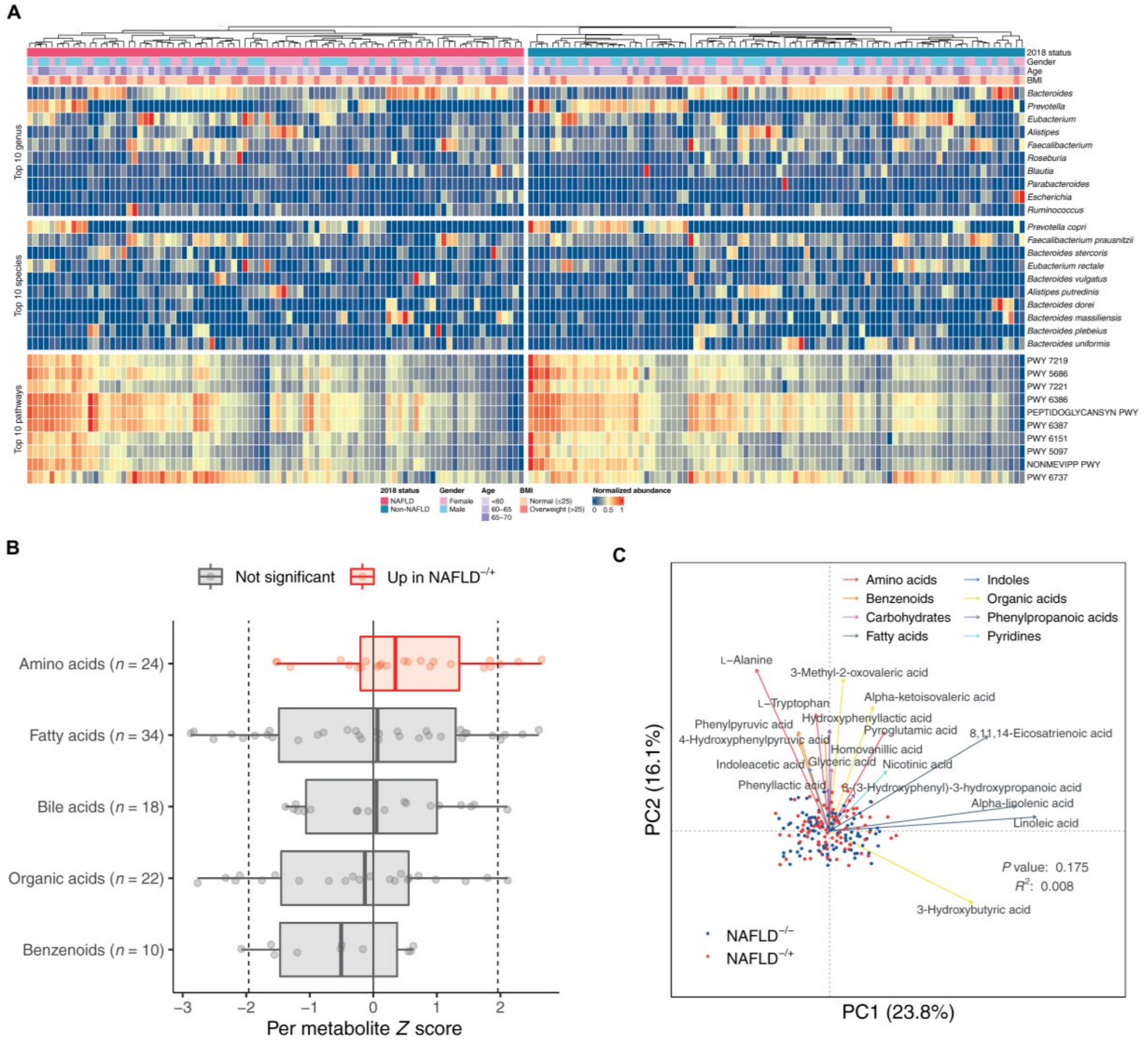


Fig. 2. Global characteristics of gut microbiome and serum metabolome. (**A**) Relative abundance of the 10 most abundant genera, species, and pathways for the 180 participants at baseline, grouped by NAFLD status at the follow-up visit. Anthropometric characteristics of the participants at baseline are also shown. Abundance values are normalized to the range of 0 and 1. (**B**) Changes of metabolites (μM) in metabolite classes containing at least 10 metabolites. Each point represents a metabolite and its *z* score from Wilcoxon rank-sum test comparing the two groups (negative indicates higher abundance in NAFLD$^{-/-}$; positive indicates higher abundance in NAFLD$^{-/+}$). Dotted lines at −1.96 and 1.96 denote the significance threshold. Colors indicate comparisons between *z* scores of metabolites in a metabolite class against the *z* scores of metabolites in all other classes. Box plots show median, lower/upper quartiles, and whiskers (the last data points 1.5 times interquartile range from the lower or upper quartiles). (**C**) Principal coordinates analysis for 180 participants based on Bray-Curtis distances using baseline serum concentrations of 123 metabolites. For each metabolite class, the top 3 (or fewer) metabolites that were significantly associated with the metabolome variation in the study cohort are shown. PC, principal coordinates.

detected 123 metabolites grouped into nine metabolite classes (Fig. 2B). We performed enrichment analysis to identify metabolite classes that were significantly overabundant or underabundant in the NAFLD$^{-/+}$ or the NAFLD$^{-/-}$ group, and found amino acids were significantly elevated in the NAFLD$^{-/+}$ group (Wilcoxon rank-sum test, $P < 0.05$; table S4). We further analyzed the untargeted metabolomic data of a European case-control cohort (MICROBARIA) involving 52 obese women including 26 with biopsy-confirmed NAFLD and 26 non-NAFLD (*52*). Two amino acids positively correlated with NAFLD-related liver enzymes, including the branched-chain amino acid valine with alanine transaminase (ALT) ($P < 0.05$, Spearman correlation) and the aromatic amino acid tyrosine with aspartate transaminase

(AST) ($P < 0.05$, Spearman correlation). Our findings in the Asian prospective and European cohort–based datasets are further supported by recent metabolomic-based studies, suggesting that perturbations in amino acid metabolism are involved in NAFLD and NASH pathogenesis (*53–55*).

Of the 15 significantly different metabolites between the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups at baseline (generalized linear model, $P < 0.05$; fig. S5A) and the metabolites that were significantly associated with the observed metabolomic variation (envfit from R package vegan, $P < 0.05$; Fig. 2C), several are reported to be involved in NAFLD in case-control human or animal studies. For example, 3-chlorotyrosine, arachidonic acid, and oxoglutaric acid are markers, respectively, of liver damage and NAFLD development in mouse (*56*) and rat (*57*) models and a human NAFLD study (*58*). Tryptophan was also significantly associated with the metabolome variation (envfit from R package vegan, $P < 0.05$; Fig. 2C) in our cohort, and aromatic amino acids have been associated with NAFLD (*54*). These metabolites were higher in the NAFLD$^{-/+}$ group than the NAFLD$^{-/-}$ group. Concentration of a gut microbiota–regulated fatty acid, 8,11,14-eicosatrienoic acid, linked to obesity and insulin resistance (*59*, *60*), was also significantly higher (generalized linear model, $P < 0.05$; fig. S5A) in the NAFLD$^{-/+}$ group in our prospective study. Phenyllactic acid, produced by lactic acid bacteria and suggested to reduce reactive oxygen species production in rodents (*61*), was significantly higher in the NAFLD$^{-/-}$ group (generalized linear model, $P < 0.05$; fig. S5A). On the contrary, the direction of concentration differences in the two study groups for isovaleric and docosahexaenoic acids (both higher in NAFLD$^{-/+}$) (generalized linear model, $P < 0.05$; fig. S5A) was inconsistent with proposals in the literature from case-control NAFLD studies about the possible roles of these compounds (*62–64*). These agreements and discrepancies in metabolite abundances in our prospective study with case-control cohort and mouse studies in the literature should help to narrow the metabolic marker possibilities for NAFLD progression. The concentrations of additional fatty acids were significantly different between the NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups (fig. S5A), but the functional significance of these metabolites in NAFLD is relatively unknown. Last, the concentrations of measured serum metabolites such as 3-chlorotyrosine and phenyllactic acid were significantly associated with gut microbiota species composition (Mantel test, $P < 0.05$; fig. S5B and table S5).

A machine learning prospective model to detect early signatures of NAFLD

We built a noninvasive risk assessment model (random forest algorithm) to classify healthy subjects based on their NAFLD status after 4.6 years, using a combination of baseline metagenomic and metabolomic features. A leave-one-out iterative approach was applied to build and evaluate our model due to the relatively small cohort size ($n = 180$). We built a prospective model using 14 taxonomic, functional, and metabolomic features of the study participants at baseline that enabled classification based on their NAFLD status 4.6 years later with an area under the receiver operating characteristic curve (auROC) of 0.72 (Fig. 3A). The performance of the model was significantly improved to an auROC of 0.79 (DeLong test, P value for difference < 0.05) with the addition of only two more noninvasive clinical features (Fig. 3B). We then slightly improved our model by also including the most accessible anthropometric parameters, BMI and age, to obtain our final model (auROC, 0.80; Fig. 3C and fig. S6).

We evaluated the biological relevance of the selected features by testing the diagnostic ability of components of our model to distinguish between healthy individuals and patients with NAFLD in two publicly available independent external Asian case-control cohorts; one cohort was diagnosed by biopsy and the other was diagnosed by magnetic resonance spectroscopy (MRS). This allowed us to further explore whether the patient diagnosis method had an impact on the model performance. We built a new prospective model using only nine features from the final model that were available in these external cohorts (fig. S6). The new model derived based on our study cohort discriminated healthy and NAFLD groups in the two external cohorts with auROCs of 0.78 and 0.72 (Fig. 3, D and E), showcasing that the features we identified were closely related to NAFLD development or pathophysiology. Besides the Asian cohorts, we further validated our prospective model in other case-control cohorts of different ethnicity. In the European cohort FLORINASH (*54*), the model (with the same nine features as in the Asian cohorts) reached an auROC of 0.76 (Fig. 3F), whereas in a U.S. cohort (*49*), the validation auROC (with seven available features) was 0.78 (Fig. 3G). Taking into consideration that only no more than half of features in our original prospective model were available in the external cohorts, we expect that the true accuracy may be higher.

Previous clinical prospective NAFLD studies demonstrated that fibroblast growth factor 21 and BMI (FGF21 + BMI), fatty liver index (FLI), and TG and glucose index (TyG) predict NAFLD development from 3 up to 9 years before diagnosis (auROCs of 0.71 to 0.82) (*19–21*). We compared the performance of our prospective model with FGF21 + BMI, FLI, and TyG to predict NAFLD occurrence in our cohort with matched baseline characteristics. The performance of our final model (auROC, 0.80) was significantly better than all three clinical models (auROCs of 0.58 to 0.60, P values for difference < 0.01; Fig. 3, H to J). To confirm the importance of metagenomic and metabolomic information in prospective NAFLD prediction, we added metagenomic and metabolomic features from our final prospective model to the clinical models (fig. S6) and observed significant improvements in all (auROCs of 0.73 to 0.75, P values for difference < 0.05; Fig. 3, H to J); however, none of the models reached the auROCs of our final model.

In total, 18 features were used in the final model: two genera, three pathways, nine metabolites, and four anthropometric and clinical parameters (Fig. 4, A and B). Our analysis revealed that the most important feature of our risk assessment model was phenyllactic acid (Fig. 4A). By analyzing the untargeted metabolomic data from the European MICROBARIA cohort (*52*), we found that phenyllactic acid negatively correlated with ALT, AST, and gamma-glutamyl transferase (correlation coefficients = −0.35, −0.45, and −0.45; $P = 0.004$, 0.14, and 0.053; Pearson's correlation adjusted for age, BMI, fasting glucose, and insulin).

SHapley Additive exPlanations (SHAP) (*65*) analysis also revealed that *Methanobrevibacter* was associated with NAFLD$^{-/-}$, and *Slackia* was associated with NAFLD$^{-/+}$ (Fig. 4, A and B). These genera were differentially abundant in our two study groups (fig. S2D). Furthermore, 8,11,14-eicosatrienoic acid, hydrocinnamic acid, and oxoglutaric acid are associated with type 2 diabetes, obesity, insulin resistance, and NAFLD (*58*, *60*, *61*, *66*), and our model revealed similar trends (Fig. 4, A and B).

The feature set contribution was also computed by summing the SHAP values per category. Metabolites were the most important in the model, contributing 44.6% to model performance, followed by

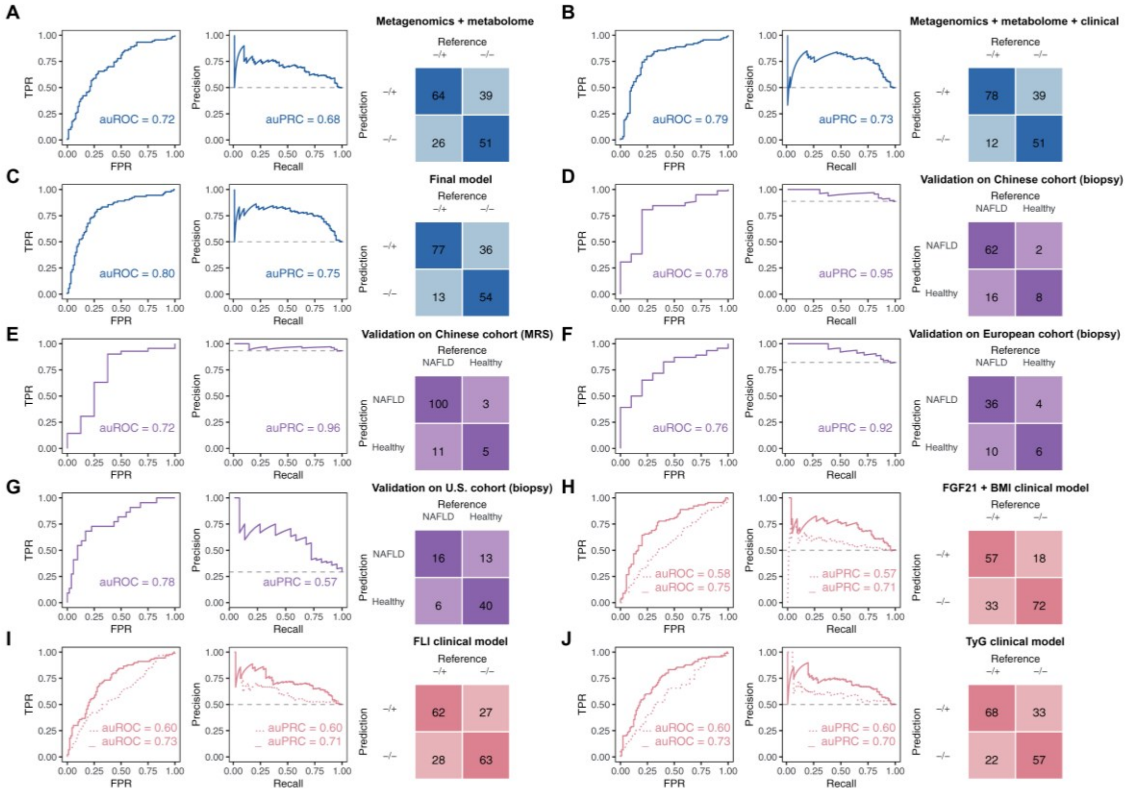


Fig. 3. Predictive performance of machine learning models in the study cohort and diagnostic performance of the final model in external cohorts. (**A** to **C**, in blue) Performance of leave-one-out iterative machine learning models discriminating between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups using features of the following: (A) metagenomics + metabolome, (B) metagenomics + metabolome + 2 clinical parameters (HDL and fasting insulin), and (C) metagenomics + metabolome + 2 clinical parameters (HDL and fasting insulin) + anthropometrics (BMI and age). (**D** to **G**, in purple) Diagnostic performances of a model built based on subsets of the selected features to discriminate between participants who were healthy or had NAFLD in four external cohorts: (D) a Chinese cohort in which NAFLD diagnosis was determined with biopsy, (E) a Chinese cohort in which NAFLD diagnosis was based on MRS, (F) a biopsy-diagnosed European NAFLD cohort, and (G) a biopsy-diagnosed U.S. cirrhosis cohort. (**H** to **J**, in peach) Leave-one-out iterative machine learning performance to discriminate between NAFLD$^{-/+}$ and NAFLD$^{-/-}$ groups in models of: (H) FGF21 + BMI clinical model, with and without metagenomics + metabolome features; (I) FLI clinical model, with and without metagenomics + metabolome features; and (J) TyG clinical model, with and without metagenomics + metabolome features. (H to J) Models without metagenomics and metabolome features were trained by logistic regression (dotted lines); models including metagenomics and metabolome features were trained by random forest (solid lines). Confusion matrices in (F) to (H) are from models with metagenomics and metabolome features. The figure colors represent the purpose of the model: blue, model construction; purple, external validation in cohorts of different ethnicity; peach, testing performance of previous clinical models in our cohort. Further details of the overall machine learning analysis framework can be found in fig. S6. auROC, area under the receiver operating characteristics curve; auPRC, area under the precision-recall curve; TPR, true-positive rate; FPR, false-positive rate.

the microbiome and nonmicrobiome features, with contributions of 31.2 and 24.1%, respectively (Fig. 4C).

Dependence plots were built to reveal the nonlinear correlations of features and risk of NAFLD. The optimal thresholds of each feature were identified (fig. S7). We found that high-density lipoprotein (HDL) was associated with NAFLD occurrence after 4.6 years when <1.39 mM, which is close to the diagnostic criteria for metabolic syndrome when HDL was <1.0 mM (male) or <1.3 mM (female) (*67*). We also examined the dependence plots of microbial metabolite phenyllactic acid, hydrocinnamic acid, and 8,11,14-eicostrienoic acid (Fig. 4, D to F). Phenyllactic acid was associated with protection against NAFLD at a concentration of >0.25 μM. The concentration of 8,11,14-eicosatrienoic acid increased the risk of NAFLD at >51.5 μM, and hydrocinnamic acid was associated with NAFLD at a concentration of <0.39 μM. Visual inspection of the dependence plots did not indicate any differences by sex. We converted the features into binary variables (≥ or < thresholds) according to their optimal threshold and found that 12 of 18 features showed significant association with NAFLD progression (chi-square test, $P < 0.05$; table S6). These results demonstrated the importance of including an interpretable machine learning framework, such as SHAP, to provide insights when analyzing microbiome data.

We further examined whether the features of our risk assessment model could be used to classify subjects of the NAFLD$^{-/+}$ group based

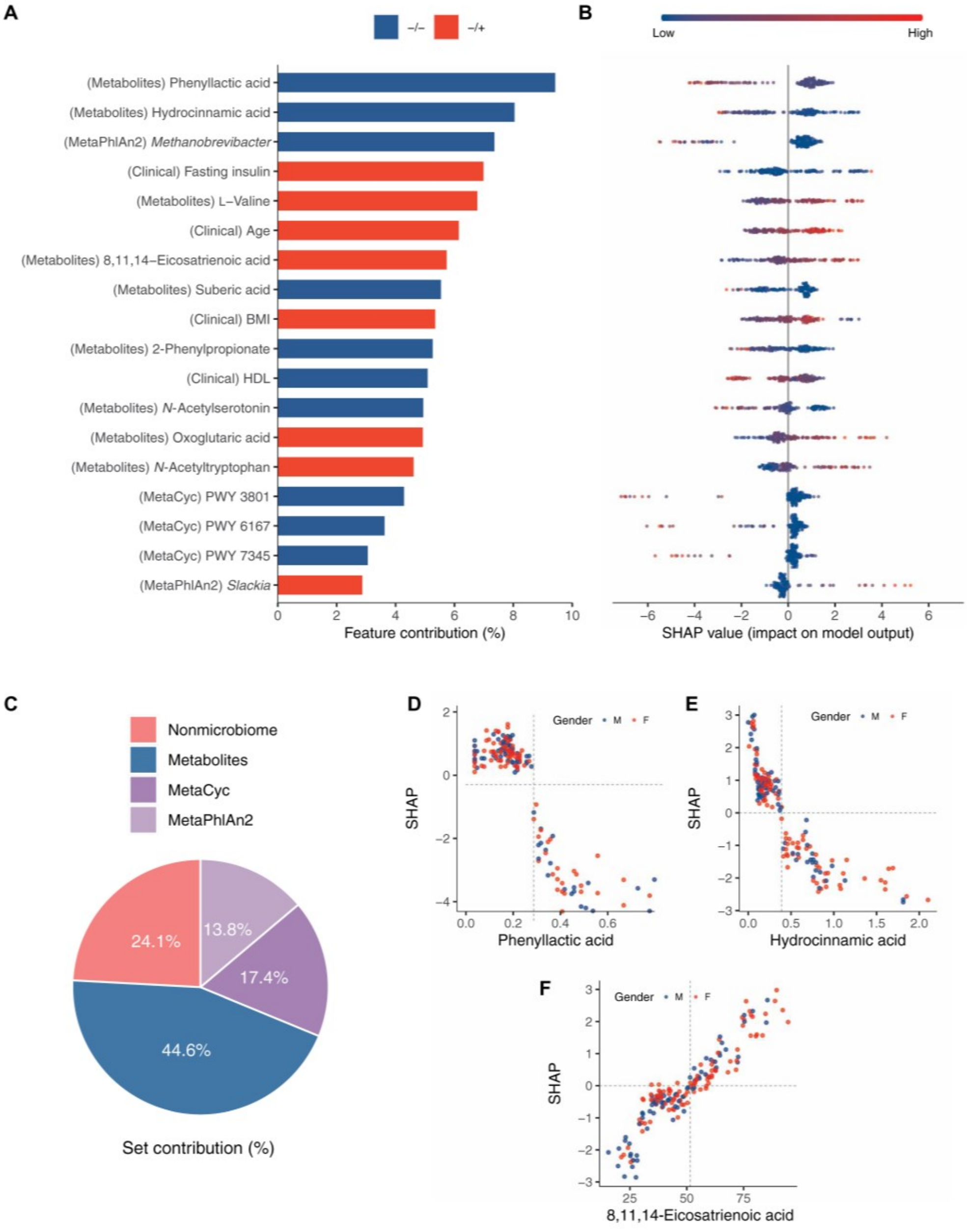


Fig. 4. SHAP-based model interpretation. (**A**) Bar plot of selected features and their contribution in the NAFLD prediction model. Features are in descending order by contribution (also known as importance) in the model. Blue bar, higher value of the feature for association with NAFLD$^{-/-}$; red bar, higher value of the feature for association with NAFLD$^{-/+}$. Details of associations are shown in (**B**) a bee swarm plot in which each point represents a participant ($n = 180$). Color indicates the value of the feature, with red higher and blue lower. Negative SHAP value indicates the feature attribution for prediction of NAFLD$^{-/-}$; Positive SHAP value indicates the feature attribution for prediction of NAFLD$^{-/+}$. (**C**) Feature category contribution calculated by summing the SHAP values per set. (**D** to **F**) Examples of SHAP dependence plots, showing the effect the feature has on model prediction. Each point represents a participant ($n = 180$). Color indicates sex with blue for male and red for female. *X* axis is the feature value, and *y* axis is the SHAP value for the feature. The optimal thresholds for features are indicated by the vertical dotted lines.

on different degrees of steatosis. We initially divided the NAFLD$^{-/+}$ group based on their liver fat percentage at the time of diagnosis (4.6 years after enrolment). Subsequently, using the values of the 18 features at baseline, we built a model classifying mild and severe steatosis cases. This new random forest model had an auROC of 0.78 (fig. S8A). Similarly as above, we attempted to confirm the biological relevance of the selected features by testing the diagnostic power of our prospective model in an independent external case-control cohort from the United States (*49*). Despite the lack of absolute quantification of metabolomic data, our model showed an accuracy of 71.4% to correctly identify severe steatosis cases with only gut microbial and clinical features.

Previous work has demonstrated the value of gut microbiome-based diagnostic tests for advanced fibrosis (*68*). The participants in our cohort were unlikely to develop advanced fibrosis after 4 years, starting as NAFLD free at baseline. Nevertheless, the prospective design of our study enabled us to explore whether the baseline microbiota is associated with the change or deterioration of fibrosis. Grouping our NAFLD$^{-/+}$ participants by the change of fibrosis 4 (FIB-4) index from 2014 to 2018, we built a new risk assessment model using five gut microbiota functional pathways, classifying subjects by the fibrosis deterioration with an auROC of 0.72 (fig. S8B). In a U.S. case-control cohort (*49*), the pathway with the highest importance in our model, phosphopantothenate biosynthesis, was significantly higher (zero-inflated Gaussian mixture model, $P < 0.05$) in the cirrhosis group than in non-NAFLD controls. *Methanobrevibacter*, which was the top taxonomic feature in the prospective model, was also significantly lower (zero-inflated Gaussian mixture model, $P < 0.05$) in patients with cirrhosis.

DISCUSSION

NAFLD prevalence has rapidly increased over a short time, especially in China (*69*). China is projected to have the largest number of liver-related deaths among the most economically developed countries by 2030 (*70*). Accumulating evidence suggests that the gut microbiome may emerge as an active player in NAFLD development (*71*). Human studies demonstrated different gut microbiota profiles among individuals with NAFLD and those without, as well as in individuals at different stages of NAFLD (*68*, *72*). In the recently proposed concept of metabolic-associated fatty liver disease (MAFLD) that extends beyond NAFLD (*23*), gut microbiota is suggested to be a major factor related to the heterogeneous phenotype of MAFLD. In both NAFLD and MAFLD, the disease complexity and heterogeneity may be better resolved by the inclusion of omics technologies that integrate patient clinical phenotypes and molecular phenomics and gut microbial features. This approach has shown its potential in the classification of hepatic (*73*) and, more recently, extrahepatic diseases including ischemic heart disease (*74*) and coronary artery disease (*75*). Both studies of cardiovascular diseases suggested that major alterations of the gut microbiome and metabolome might occur earlier than clinical onset of disease, suggesting the utility of gut microbiota–based risk assessment. A recent prospective study extended cross-sectional evidence and demonstrated that gut microbiota composition is predictive of incident type 2 diabetes after 15.8 years (*76*). Our study comprehensively characterized the gut microbiome of Chinese participants using stool samples taken 4.6 years before the NAFLD diagnosis and matched controls. We assessed the ability of metagenomic and metabolomic features as a risk assessment tool of NAFLD occurrence within 4.6 years and developed a random forest machine learning model that distinguished individuals at risk for NAFLD from controls with a performance of 0.80 auROC. The final model consisted of 18 features of mainly bacterial genera, pathways, and metabolites, with two clinical and two anthropometric parameters. Using subsets of those features available in external case-control cohorts also showed good ability (auROC of 0.73 to 0.78) to classify individuals with and without NAFLD, including in cohorts with the biopsy-confirmed present/absence of NAFLD and of different ethnicities, supporting the biological relevance and generalizability of our prospective model.

Diagnosis of NAFLD requires evidence of hepatic steatosis, either by histology or imaging. Liver biopsies have a risk of severe complications, and the sampling procedure may leave some people with NAFLD undiagnosed if they have unevenly distributed histological lesions (*13*). Steatosis evaluation based on imaging such as MRS, CT, or ultrasonography has limitations in clinical practice such as high price, radiation exposure, and limited sensitivity. Numerous research efforts have searched for other reliable, cost-effective, noninvasive diagnostic approaches, including using features that are clinical (age, gender, diabetes, and BMI), biochemical (aminotransferases, bilirubin, and ferritin), metabolic (glycated hemoglobin, insulin, and HOMA-IR), or lipid (TG and cholesterol) parameters or other markers such as FGF21 and adiponectin (*77*–*79*). A few prospective studies have also attempted to predict the development of NAFLD over the long term (*19*–*21*). However, the predictive power of these models was evaluated in study groups with unmatched baseline characteristics, which may have led to overestimation of model performance. In our community-based prospective study, these models showed limited performance (auROC in the range of 0.58 to 0.60) when our nested case-control design included matching for gender, age, BMI, and 4-year BMI change. This matching is particularly important for removing confounding effects and to uncover microbiome-related risk factors for NAFLD development, given that obesity is a major risk factor for NAFLD. Our microbiome-based model demonstrated a good performance (auROC of 0.72) for predicting the NAFLD status of NAFLD-free individuals after 4.6 years.

Our study has limitations. The classification of patients into two groups was not based on liver biopsy, which remains the gold standard for NAFLD diagnosis. However, this method is impractical in a community study with thousands of participants, as in our study, and is unethical for participants who do not show any sign of the disease (matched controls). Moreover, according to guidelines from the European Association for the Study of the Liver, European Association for the Study of Diabetes, and European Association for the Study of Obesity, ultrasound is the first-line diagnostic test for NAFLD (*36*), especially for large-scale screening studies. This diagnostic criterion has been extensively used in previous studies, such as the Rotterdam cohort (*80*), the Golestan cohort (*81*), and the Kangbuk Samsung Health Study (*82*). We note that our cohort is of high quality, with relatively comprehensive indexes acquired in a large population. For example, in the measurement of glucose metabolism, oral glucose tolerance tests were conducted for all participants. This test is usually replaced by fasting glucose or FINS tests in many population-based studies. Second, we could not predict the development of more severe outcomes such as fibrosis because of their low incidence. This was mainly due to the nature of our community-based epidemiological investigation. However, using baseline microbiota, we were able to classify subjects by fibrosis deterioration with an auROC of 0.72.

Furthermore, serum ferritin was not measured in our study, although several studies have indicated its relevance in NAFLD (*83–85*). Therefore adding ferritin in our prospective model could potentially enhance performance. The predictive power of our prospective model (auROC of 0.80) was an advance compared to existing clinical models (auROCs of 0.58 to 0.60). However, further improvements, for example, integrating additional biochemical parameters, will be necessary for clinical applications. Our metabolic signatures and their taxonomic drivers revealed by our prospective model imply but do not prove causality; thus, additional studies are required to clarify the molecular mechanisms involved in NAFLD development.

Integrating bacterial species and functions in machine learning models for predicting host response to treatment or lifestyle interventions and disease progression has shown great potential (*86–89*). For NAFLD and its complications, gut microbiota changes can independently predict the risk of short-term hospitalizations (90 days) in patients with cirrhosis with an auROC of 0.83 (*90*). Elucidating the importance of the gut microbiome as a long-term risk assessment tool in NAFLD is important because of the current limited therapeutic landscape for NAFLD and findings that early detection can substantially improve outcomes for patients with NAFLD (*91*, *92*). Our proof-of-concept study identified a microbiome signature in participants at risk of developing NAFLD in the next 4 years and points to the potential of noninvasive diagnostic tests to complement existing clinical screening tools for NAFLD. Moreover, identifying microbiome signatures also opens a window of opportunities for microbiome-based prophylactic and therapeutic interventions such as the utility of propionic acid as a potent immunomodulatory supplement to multiple sclerosis drugs (*93*), which is not offered by a clinical predictive model built upon only a few clinical parameters or other features. Evaluation and further improvement of our NAFLD risk assessment model using larger prospective studies that are heterogeneous for ethnicity and lifestyle patterns will increase the model's generalizability and obtain more refined estimations of its accuracy.

MATERIALS AND METHODS

Study design

The aim of this study was to identify potential predictive signatures for early clinical warning of NAFLD and to develop a prognostic risk assessment model for long-term NAFLD development. For this purpose, we conducted a nested case-control study within a 4.6-year prospective study in 2487 Chinese individuals, and we profiled 180 individuals from 1216 NAFLD-free participants at baseline, including 90 that were diagnosed with NAFLD in the follow-up visit (NAFLD$^{-/+}$), which were matched with 90 controls without NAFLD (NAFLD$^{-/-}$) by gender, age, BMI, and 4.6-year BMI change. We performed comprehensive metagenomic and metabolomic analyses using stool and serum samples taken at baseline, including taxonomic diversity and profiles at family, genus and species levels, microbial enzymes, metabolic pathways, and metabolites. An interpretable machine learning model integrating baseline microbial signatures was built to predict NAFLD development after 4 years. The biological relevance of selected features in the model to NAFLD development was further validated in external cohorts, including three cohorts with the biopsy-confirmed presence/absence of NAFLD. New models were built for validation, given that some features were not available in the external cohorts. All validation models were trained on our cohort and tested in the external cohorts. Further materials and methods details are available in the Supplementary Materials.

Study participants

All participants were from the Nicheng Diabetes Screening Project (also called the Shanghai Nicheng Cohort Study) previously described (*94*, *95*). This population-based, prospective study was designed to assess the prevalence, incidence, and factors related to cardiometabolic diseases among adults in Nicheng County, a suburb of Shanghai, China. On the basis of the project, we designed a nested case-control study to explore the potential causal role of the gut microbiome in NAFLD in three randomly selected Nicheng communities (involving 2487 participants). Figure S1 outlines study enrolment. Of 2487 participants, 1216 were identified as not having NAFLD at baseline; among them, 524 completed a follow-up visit 4.6 years after baseline and were screened by ultrasonography. Incident cases of NAFLD ($n = 146$) were identified at the 4.6-year follow-up visit, of which 90 participants were eligible for this study involving gut microbiota, according to the following criteria to exclude participants: existed fatty liver, acute infectious disease, biliary obstructive diseases, alcohol abuse (more than 140 g of ethanol/week for men or 70 g of ethanol/week for women), acute or chronic cholecystitis, acute or chronic viral hepatitis, cirrhosis, diarrhea, known hyperthyroidism or hypothyroidism, chronic renal insufficiency, heart failure, presence of cancer, pregnancy, stroke in acute phase, receipt of any antibiotic treatment within 2 weeks or receipt of any probiotic or prebiotic within 1 week before sample collection, and suffering from chronic or acute gastrointestinal diseases (including diarrhea, gastrointestinal infection, and inflammatory bowel disease) in recent 1 month before sample collection. Controls ($n = 90$ for a case-control ratio of 1:1) were chosen from the remaining participants who did not develop NAFLD by the follow-up visit. To control for the risk profiles in patients who developed NAFLD and those who did not, controls were matched for age (±3 years), sex (male and female), BMI (±3 kg/m^2) at both baseline and follow-up, and BMI change (±0.5 kg/m^2). The study was approved by the ethics committee of the Shanghai Sixth People's Hospital (approval no: 2014-27), following the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants.

Evaluating the diagnostic ability of the model in external cohorts

To our knowledge, no similar studies have conducted long-term follow-up of NAFLD development in healthy individuals using a combination of gut metagenome, metabolome, and clinical features as a risk assessment tool. Thus, we were unable to test our prospective model directly in an external cohort. Instead, we used external case-control cohorts to examine the ability of our final prognostic model to classify correctly NAFLD and healthy participants. Four cohorts were used, including two cohorts of Chinese: (i) 78 patients with NAFLD and 10 controls without NAFLD, as diagnosed with biopsy (BioProject ID: PRJNA732131), and (ii) 111 MRS-diagnosed NAFLD patients and 8 controls (BioProject IDs: PRJNA703757 and PRJNA414688); and two biopsy-diagnosed cohorts of other ethnicity: (iii) a European cohort of 46 patients with NAFLD and 10 controls (*54*) and (iv) a U.S. cohort of 26 cirrhosis patients and 54 controls (*49*). For further additional data (e.g., anthropometric and/or available clinical data) for the two Chinese validation cohorts besides grouping information, please contact the corresponding author.

Because some selected features included in the final model were not available in the external cohorts, we were unable to test our model directly. Instead, we built a new prognostic model based on the $\text{NAFLD}^{-/+}$ and $\text{NAFLD}^{-/-}$ groups using a subset of the 18 selected features that were available in the external cohorts. In the model for the two Chinese cohorts and the European cohort, 9 of the 18 selected features were used: two genera, three pathways, two anthropometric parameters, and two noninvasive clinical metadata; whereas 7 of the 18 selected features were used in the model for the U.S. cohort: two genera, one pathway, two anthropometric parameters, and two noninvasive clinical metadata. Performances of models, including ROC curves, precision-recall curves, and confusion matrices (generated with the optimal probability cutoff of the ROC curve), were produced by applying the model to the unseen external cohort data.

Statistical analysis

Statistical analyses of clinical data were performed with SAS version 9.4 (SAS Institute Inc.). Normally distributed data were expressed as means ± SD. Data that were not normally distributed, as determined using the Kolmogorov-Smirnov test, were logarithmically transformed before analysis and expressed as median with lower and upper quartiles. Student's *t* test and chi-square tests were used to assess differences between two groups for continuous and categorical variables, respectively. In addition, analysis of covariance was used for continuous variables to assess the difference between the two groups after adjusting for HOMA-IR.

Metagenomic data, including taxonomy and functional data, and metabolomic data were analyzed in R software version 3.6.3. Metagenomic data were analyzed with the zero-inflated Gaussian mixture model, using the function fitZig from R package metagenomeSeq (*96*) with the default settings; metabolomic data were analyzed using the generalized linear model with inverse gamma distribution. Wilcoxon rank-sum tests were used to test for significant differences in alpha diversity. PERMANOVA was used to analyze beta diversity with adonis function from R package vegan. A Mantel test, implemented in mantel from R package vegan, using Spearman's correlation coefficient was used to analyze the associations between microbiome and metabolites. Bray-Curtis dissimilarity matrices based on taxonomic relative abundance and Euclidean dissimilarity matrix for each metabolite were computed to perform this test. The auROCs of different models were compared with the DeLong test, using the roc.test function from R package pROC (*97*). Data were considered statistically significant at P value < 0.05. The Benjamini-Hochberg procedure was applied to calculate the FDR to adjust P values for multiple hypothesis testing.

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scitranslmed.abk0855
Figs. S1 to S8
Tables S1 to S6
Data file S1
MDAR Reproducibility Checklist
References (*98–110*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. Y. Zheng, S. H. Ley, F. B. Hu, Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat. Rev. Endocrinol.* **14**, 88–98 (2018).
2. Z. M. Younossi, A. B. Koenig, D. Abdelatif, Y. Fazel, L. Henry, M. Wymer, Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* **64**, 73–84 (2016).
3. T. Seuring, O. Archangelidi, M. Suhrcke, The economic costs of type 2 diabetes: A global systematic review. *Pharmacoeconomics* **33**, 811–831 (2015).
4. World Health Organization, "Diet, nutrition and the prevention of chronic diseases" (Technical Report series 916, World Health Organization, 2003).
5. L. A. Adams, Q. M. Anstee, H. Tilg, G. Targher, Non-alcoholic fatty liver disease and its relationship with cardiovascular disease and other extrahepatic diseases. *Gut* **66**, 1138–1153 (2017).
6. R. Loomba, S. L. Friedman, G. I. Shulman, Mechanisms and disease consequences of nonalcoholic fatty liver disease. *Cell* **184**, 2537–2564 (2021).
7. T. G. Simon, B. Roelstraete, H. Khalili, H. Hagström, J. F. Ludvigsson, Mortality in biopsy-confirmed nonalcoholic fatty liver disease: Results from a nationwide cohort. *Gut* **70**, 1375–1382 (2021).
8. H. Tilg, G. Targher, NAFLD-related mortality: Simple hepatic steatosis is not as 'benign' as thought. *Gut* **70**, 1212–1213 (2021).
9. Z. Younossi, F. Tacke, M. Arrese, B. Chander Sharma, I. Mostafa, E. Bugianesi, V. Wai-Sun Wong, Y. Yilmaz, J. George, J. Fan, M. B. Vos, Global perspectives on nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Hepatology* **69**, 2672–2682 (2019).
10. E. R. Cleveland, H. Ning, M. B. Vos, C. E. Lewis, M. E. Rinella, J. J. Carr, D. M. Lloyd-Jones, L. B. VanWagner, Low awareness of nonalcoholic fatty liver disease in a population-based cohort sample: The CARDIA study. *J. Gen. Intern. Med.* **34**, 2772–2778 (2019).
11. E. K. Spengler, R. Loomba, Recommendations for diagnosis, referral for liver biopsy, and treatment of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Mayo Clin. Proc.* **90**, 1233–1246 (2015).
12. R. Loomba, Role of imaging-based biomarkers in NAFLD: Recent advances in clinical application and future research directions. *J. Hepatol.* **68**, 296–304 (2018).
13. N. Chalasani, Z. Younossi, J. E. Lavine, M. Charlton, K. Cusi, M. Rinella, S. A. Harrison, E. M. Brunt, A. J. Sanyal, The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the american association for the study of liver diseases. *Hepatology* **67**, 328–357 (2018).
14. N. Stefan, H. U. Haring, K. Cusi, Non-alcoholic fatty liver disease: Causes, diagnosis, cardiometabolic consequences, and treatment strategies. *Lancet Diabetes Endocrinol.* **7**, 313–324 (2019).
15. H. J. Zhang, J. He, L. L. Pan, Z. M. Ma, C. K. Han, C. S. Chen, Z. Chen, H. W. Han, S. Chen, Q. Sun, J. F. Zhang, Z. B. Li, S. Y. Yang, X. J. Li, X. Y. Li, Effects of moderate and vigorous exercise on nonalcoholic fatty liver disease: A randomized clinical trial. *JAMA Intern. Med.* **176**, 1074–1082 (2016).
16. A. Mardinoglu, H. Wu, E. Bjornson, C. Zhang, A. Hakkarainen, S. M. Rasanen, S. Lee, R. M. Mancina, M. Bergentall, K. H. Pietilainen, S. Soderlund, N. Matikainen, M. Stahlman, P. O. Bergh, M. Adiels, B. D. Piening, M. Graner, N. Lundbom, K. J. Williams, S. Romeo, J. Nielsen, M. Snyder, M. Uhlen, G. Bergstrom, R. Perkins, H. U. Marschall, F. Backhed, M. R. Taskinen, J. Boren, An integrated understanding of the rapid metabolic benefits of acarbohydrate-restricted diet on hepatic steatosis in humans. *Cell Metab.* **27**, 559–571.e5 (2018).
17. J. Aron-Wisnewsky, M. V. Warmbrunn, M. Nieuwdorp, K. Clement, Nonalcoholic fatty liver disease: Modulating gut microbiota to improve severity? *Gastroenterology* **158**, 1881–1898 (2020).
18. J. M. Schattenberg, J. V. Lazarus, P. N. Newsome, L. Serfaty, A. Aghemo, S. Augustin, E. Tsochatzis, V. de Ledinghen, E. Bugianesi, M. Romero-Gomez, H. Bantel, S. D. Ryder, J. Boursier, V. Leroy, J. Crespo, L. Castera, L. Floros, V. Atella, J. Mestre-Ferrandiz, R. Elliott, A. Kautz, A. Morgan, S. Hartmanis, S. Vasudevan, L. Pezzullo, A. Trylesinski, S. Cure, V. Higgins, V. Ratziu, Disease burden and economic impact of diagnosed non-alcoholic steatohepatitis in five European countries in 2018: A cost-of-illness analysis. *Liver Int.* **41**, 1227–1242 (2021).
19. H. Li, K. Dong, Q. Fang, X. Hou, M. Zhou, Y. Bao, K. Xiang, A. Xu, W. Jia, High serum level of fibroblast growth factor 21 is an independent predictor of non-alcoholic fatty liver disease: A 3-year prospective study in China. *J. Hepatol.* **58**, 557–563 (2013).
20. N. Motamed, A. H. Faraji, M. R. Khonsari, M. Maadi, F. S. Tameshkel, H. Keyvani, H. Ajdarkosh, M. H. Karbalaie Niya, N. Rezaie, F. Zamani, Fatty liver index (FLI) and prediction of new cases of non-alcoholic fatty liver disease: A population-based study of northern Iran. *Clin. Nutr.* **39**, 468–474 (2020).
21. R. Zheng, Z. Du, M. Wang, Y. Mao, W. Mao, A longitudinal epidemiological study on the triglyceride and glucose index and the incident nonalcoholic fatty liver disease. *Lipids Health Dis.* **17**, 262 (2018).
22. A. Lonardo, F. Nascimbeni, M. Maurantonio, A. Marrazzo, L. Rinaldi, L. E. Adinolfi, Nonalcoholic fatty liver disease: Evolving paradigms. *World J. Gastroenterol.* **23**, 6571–6592 (2017).
23. M. Eslam, A. J. Sanyal, J. George; International Consensus Panel, MAFLD: A consensus-driven proposed nomenclature for metabolic associated fatty liver disease. *Gastroenterology* **158**, 1999–2014.e1 (2020).
24. F. Bäckhed, H. Ding, T. Wang, L. V. Hooper, G. Y. Koh, A. Nagy, C. F. Semenkovich, J. I. Gordon, The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15718–15723 (2004).

25. R. E. Ley, P. J. Turnbaugh, S. Klein, J. I. Gordon, Human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
26. E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jørgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clément, J. Doré, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J.-D. Zucker, J. Raes, T. Hansen; MetaHIT consortium, P. Bork, J. Wang, S. D. Ehrlich, O. Pedersen, Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
27. A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, R. D. Finn, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
28. A. Wahlström, S. I. Sayin, H.-U. Marschall, F. Bäckhed, Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab.* **24**, 41–50 (2016).
29. E. D. Sonnenburg, S. A. Smits, M. Tikhonov, S. K. Higginbottom, N. S. Wingreen, J. L. Sonnenburg, Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).
30. P. D. Cani, M. Osto, L. Geurts, A. Everard, Involvement of gut microbiota in the development of low-grade inflammation and type 2 diabetes associated with obesity. *Gut Microbes* **3**, 279–288 (2012).
31. E. E. Canfora, R. C. R. Meex, K. Venema, E. E. Blaak, Gut microbial metabolites in obesity, NAFLD and T2DM. *Nat. Rev. Endocrinol.* **15**, 261–273 (2019).
32. L. Zhu, S. S. Baker, C. Gill, W. Liu, R. Alkhouri, R. D. Baker, S. R. Gill, Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology* **57**, 601–609 (2013).
33. B. Wang, X. Jiang, M. Cao, J. Ge, Q. Bao, L. Tang, Y. Chen, L. Li, Altered fecal microbiota correlates with liver biochemistry in nonobese patients with non-alcoholic fatty liver disease. *Sci. Rep.* **6**, 32002 (2016).
34. M. Mouzaki, E. M. Comelli, B. M. Arendt, J. Bonengel, S. K. Fung, S. E. Fischer, I. D. McGilvray, J. P. Allard, Intestinal microbiota in patients with nonalcoholic fatty liver disease. *Hepatology* **58**, 120–127 (2013).
35. H. E. Da Silva, A. Teterina, E. M. Comelli, A. Taibi, B. M. Arendt, S. E. Fischer, W. Lou, J. P. Allard, Nonalcoholic fatty liver disease is associated with dysbiosis independent of body mass index and insulin resistance. *Sci. Rep.* **8**, 1466 (2018).
36. European Association for the Study of the Liver (EASL); European Association for the Study of Diabetes (EASD); European Association for the Study of Obesity (EASO), EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. *J. Hepatol.* **64**, 1388–1402 (2016).
37. S. Chitturi, G. C. Farrell, E. Hashimoto, T. Saibara, G. K. Lau, J. D. Sollano; Asia-Pacific Working Party on NAFLD, Non-alcoholic fatty liver disease in the Asia-Pacific region: Definitions and overview of proposed guidelines. *J. Gastroenterol. Hepatol.* **22**, 778–787 (2007).
38. Y. Gallois, S. Vol, E. Caces, B. Balkau; DESIR Study Group, Distribution of fasting serum insulin measured by enzyme immunoassay in an unselected population of 4,032 individuals. Reference values according to age and sex. *Diabetes Metab.* **22**, 427–431 (1996).
39. R. Muniyappa, S. Lee, H. Chen, M. J. Quon, Current approaches for assessing insulin sensitivity and resistance in vivo: Advantages, limitations, and appropriate usage. *Am. J. Physiol. Endocrinol. Metab.* **294**, E15–E26 (2008).
40. A. L. Catapano, I. Graham, G. De Backer, O. Wiklund, M. J. Chapman, H. Drexel, A. W. Hoes, C. S. Jennings, U. Landmesser, T. R. Pedersen, Z. Reiner, G. Riccardi, M. R. Taskinen, L. Tokgozoglu, W. M. M. Verschuren, C. Vlachopoulos, D. A. Wood, J. L. Zamorano, M. T. Cooney; ESC Scientific Document Group, 2016 ESC/EAS guidelines for the management of dyslipidaemias. *Eur. Heart J.* **37**, 2999–3058 (2016).
41. T. A. Pearson, G. A. Mensah, R. W. Alexander, J. L. Anderson, R. O. Cannon III, M. Criqui, Y. Y. Fadl, S. P. Fortmann, Y. Hong, G. L. Myers, N. Rifai, S. C. Smith Jr., K. Taubert, R. P. Tracy, F. Vinicor; Centers for Disease Control and Prevention; American Heart Association, Markers of inflammation and cardiovascular disease: Application to clinical and public health practice: A statement for healthcare professionals from the centers for disease control and prevention and the american heart association. *Circulation* **107**, 499–511 (2003).
42. D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, N. Segata, MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
43. H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Q. D. Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, H. I. T. C. Meta, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, S. D. Ehrlich; MetaHIT Consortium, Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
44. M. Million, M. Maraninchi, M. Henry, F. Armougom, H. Richet, P. Carrieri, R. Valero, D. Raccah, B. Vialettes, D. Raoult, Obesity-associated gut microbiota is enriched in Lactobacillus reuteri and depleted in Bifidobacterium animalis and Methanobrevibacter smithii. *Int. J. Obes. (Lond)* **36**, 817–825 (2012).
45. D. A. Muñiz Pedrogo, M. D. Jensen, C. T. Van Dyke, J. A. Murray, J. A. Woods, J. Chen, P. C. Kashyap, V. Nehra, Gut microbial carbohydrate metabolism hinders weight loss in overweight adults undergoing lifestyle intervention with a volumetric diet. *Mayo Clin. Proc.* **93**, 1104–1110 (2018).
46. J. B. Schwimmer, J. S. Johnson, J. E. Angeles, C. Behling, P. H. Belt, I. Borecki, C. Bross, J. Durelle, N. P. Goyal, G. Hamilton, M. L. Holtz, J. E. Lavine, M. Mitreva, K. P. Newton, A. Pan, P. M. Simpson, C. B. Sirlin, E. Sodergren, R. Tyagi, K. P. Yates, G. M. Weinstock, N. H. Salzman, Microbiome signatures associated with steatohepatitis and moderate to severe fibrosis in children with nonalcoholic fatty liver disease. *Gastroenterology* **157**, 1109–1122 (2019).
47. L. K. Brahe, E. Le Chatelier, E. Prifti, N. Pons, S. Kennedy, T. Hansen, O. Pedersen, A. Astrup, S. D. Ehrlich, L. H. Larsen, Specific gut microbiota features and metabolic markers in postmenopausal women with obesity. *Nutr. Diabetes* **5**, e159 (2015).
48. E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, C. Huttenhower, Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
49. T. G. Oh, S. M. Kim, C. Caussy, T. Fu, J. Guo, S. Bassirian, S. Singh, E. V. Madamba, R. Bettencourt, L. Richards, R. T. Yu, A. R. Atkins, T. Huan, D. A. Brenner, C. B. Sirlin, M. Downes, R. M. Evans, R. Loomba, A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* **32**, 901 (2020).
50. J. Liu, S. Jiang, Y. Zhao, Q. Sun, J. Zhang, D. Shen, J. Wu, N. Shen, X. Fu, X. Sun, D. Yu, J. Chen, J. He, T. Shi, Y. Ding, L. Fang, B. Xue, C. Li, Geranylgeranyl diphosphate synthase (GGPPS) regulates non-alcoholic fatty liver disease (NAFLD)-fibrosis progression by determining hepatic glucose/fatty acid preference under high-fat diet conditions. *J. Pathol.* **246**, 277–288 (2018).
51. D. Gabbia, M. Roverso, M. Guido, D. Sacchi, M. Scaffidi, M. Carrara, G. Orso, F. P. Russo, A. Floreani, S. Bogialli, S. De Martin, Western diet-induced metabolic alterations affect circulating markers of liver function before the development of steatosis. *Nutrients* **11**, 1602 (2019).
52. J. Aron-Wisnewsky, E. Prifti, E. Belda, F. Ichou, B. D. Kayser, M. C. Dao, E. O. Verger, L. Hedjazi, J. L. Bouillot, J. M. Chevallier, N. Pons, E. Le Chatelier, F. Levenez, S. D. Ehrlich, J. Dore, J. D. Zucker, K. Clement, Major microbiota dysbiosis in severe obesity: Fate after bariatric surgery. *Gut* **68**, 70–82 (2019).
53. A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen, J. Nielsen, Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014).
54. L. Hoyles, J.-M. Fernández-Real, M. Federici, M. Serino, J. Abbott, J. Charpentier, C. Heymes, J. L. Luque, E. Anthony, R. H. Barton, J. Chilloux, A. Myridakis, L. Martinez-Gili, J. M. Moreno-Navarrete, F. Benhamed, V. Azalbert, V. Blasco-Baque, J. Puig, G. Xifra, W. Ricart, C. Tomlinson, M. Woodbridge, M. Cardellini, F. Davato, I. Cardolini, O. Porzio, P. Gentileschi, F. Lopez, F. Foufelle, S. A. Butcher, E. Holmes, J. K. Nicholson, C. Postic, R. Burcelin, M.-E. Dumas, Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat. Med.* **24**, 1070–1080 (2018).
55. M. Gaggini, F. Carli, C. Rosso, E. Buzzigoli, M. Marietti, V. Della Latta, D. Ciociaro, M. L. Abate, R. Gambino, M. Cassader, E. Bugianesi, A. Gastaldelli, Altered amino acid concentrations in NAFLD: Impact of obesity and insulin resistance. *Hepatology* **67**, 145–158 (2018).
56. A. C. Koop, N. D. Thiele, D. Steins, E. Michaëlsson, M. Wehmeyer, L. Scheja, B. Steglich, S. Huber, J. Schulze Zur Wiesch, A. W. Lohse, J. Heeren, J. Kluwe, Therapeutic targeting of myeloperoxidase attenuates NASH in mice. *Hepatol. Commun.* **4**, 1441–1458 (2020).
57. K. Sztolsztener, A. Chabowski, E. Harasim-Symbor, P. Bielawiec, K. Konstantynowicz-Nowicka, Arachidonic acid as an early indicator of inflammation during non-alcoholic fatty liver disease development. *Biomolecules* **10**, 1133 (2020).
58. E. Rodríguez-Gallego, M. Guirro, M. Riera-Borrull, A. Hernández-Aguilera, R. Mariné-Casadó, S. Fernández-Arroyo, R. Beltrán-Debón, F. Sabench, M. Hernández, D. del Castillo, J. A. Menendez, J. Camps, R. Ras, L. Arola, J. Joven, Mapping of the circulating metabolome reveals α-ketoglutarate as a predictor of morbid obesity-associated non-alcoholic fatty liver disease. *Int. J. Obes.* **39**, 279–287 (2014).
59. A. Kindt, G. Liebisch, T. Clavel, D. Haller, G. Hörmannsperger, H. Yoon, D. Kolmeder, A. Sigruener, S. Krautbauer, C. Seeliger, A. Ganzha, S. Schweizer, R. Morisset, T. Strowig, H. Daniel, D. Helm, B. Küster, J. Krumsiek, J. Ecker, The gut microbiota promotes hepatic fatty acid desaturation and elongation in mice. *Nat. Commun.* **9**, 1–15 (2018).

60. Y. Tsurutani, K. Inoue, C. Sugisawa, J. Saito, M. Omura, T. Nishikawa, Increased serum dihomo-γ-linolenic acid levels are associated with obesity, body fat accumulation, and insulin resistance in japanese patients with type 2 diabetes. *Intern. Med.* **57**, 2929–2935 (2018).
61. N. Beloborodova, I. Bairamov, A. Olenin, V. Shubina, V. Teplova, N. Fedotcheva, Effect of phenolic acids of microbial origin on production of reactive oxygen species in mitochondria and neutrophils. *J. Biomed. Sci.* **19**, 89 (2012).
62. S. Takada, T. Matsubara, H. Fujii, M. Sato-Matsubara, A. Daikoku, N. Odagiri, Y. Amano-Teranishi, N. Kawada, K. Ikeda, Stress can attenuate hepatic lipid accumulation via elevation of hepatic β-muricholic acid levels in mice with nonalcoholic steatohepatitis. *Lab. Invest.* **101**, 193–203 (2021).
63. S. Chashmniam, M. Ghafourpour, A. Rezaei Farimani, A. Gholami, B. F. N. M. Ghoochani, Metabolomic biomarkers in the diagnosis of non-alcoholic fatty liver disease. *Hepat. Mon.* **19**, e92244 (2019).
64. L. Hodson, L. Bhatia, E. Scorletti, D. E. Smith, N. C. Jackson, F. Shojaee-Moradie, M. Umpleby, P. C. Calder, C. D. Byrne, Docosahexaenoic acid enrichment in NAFLD is associated with improvements in hepatic metabolism and hepatic insulin sensitivity: A pilot study. *Eur. J. Clin. Nutr.* **71**, 1251 (2017).
65. S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. arXiv:1705.07874 (2017).
66. C. Menni, J. Zhu, C. I. Le Roy, O. Mompeo, K. Young, C. M. Rebholz, E. Selvin, K. E. North, R. P. Mohney, J. T. Bell, E. Boerwinkle, T. D. Spector, M. Mangino, B. Yu, A. M. Valdes, Serum metabolites reflecting gut microbiome alpha diversity predict type 2 diabetes. *Gut Microbes* **11**, 1632–1642 (2020).
67. S. M. Grundy, J. I. Cleeman, S. R. Daniels, K. A. Donato, R. H. Eckel, B. A. Franklin, D. J. Gordon, R. M. Krauss, P. J. Savage, S. C. Smith Jr., J. A. Spertus, F. Costa; American Heart Association; National Heart, Lung, and Blood Institute, Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation* **112**, 2735–2752 (2005).
68. R. Loomba, V. Seguritan, W. Li, T. Long, N. Klitgord, A. Bhatt, P. S. Dulai, C. Caussy, R. Bettencourt, S. K. Highlander, M. B. Jones, C. B. Sirlin, B. Schnabl, L. Brinkac, N. Schork, C.-H. Chen, D. A. Brenner, W. Biggs, S. Yooseph, J. C. Venter, K. E. Nelson, Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. *Cell Metab.* **25**, 1054–1062.e5 (2017).
69. H. W. Lee, V. W.-S. Wong, Changing NAFLD epidemiology in China. *Hepatology* **70**, 1095–1098 (2019).
70. C. Estes, Q. M. Anstee, M. T. Arias-Loste, H. Bantel, S. Bellentani, J. Caballeria, M. Colombo, A. Craxi, J. Crespo, C. P. Day, Y. Eguchi, A. Geier, L. A. Kondili, D. C. Kroy, J. V. Lazarus, R. Loomba, M. P. Manns, G. Marchesini, A. Nakajima, F. Negro, S. Petta, V. Ratziu, M. Romero-Gomez, A. Sanyal, J. M. Schattenberg, F. Tacke, J. Tanaka, C. Trautwein, L. Wei, S. Zeuzem, H. Razavi, Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030. *J. Hepatol.* **69**, 896–904 (2018).
71. A. Albillos, A. de Gottardi, M. Rescigno, The gut-liver axis in liver disease: Pathophysiological basis for therapy. *J. Hepatol.* **72**, 558–577 (2020).
72. J. Aron-Wisnewsky, C. Vigliotti, J. Witjes, P. Le, A. G. Holleboom, J. Verheij, M. Nieuwdorp, K. Clément, Gut microbiota and human NAFLD: Disentangling microbial signatures from metabolic disorders. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 279–297 (2020).
73. L. Hoyles, J. M. Fernandez-Real, M. Federici, M. Serino, J. Abbott, J. Charpentier, C. Heymes, J. L. Luque, E. Anthony, R. H. Barton, J. Chilloux, A. Myridakis, L. Martinez-Gili, J. M. Moreno-Navarrete, F. Benhamed, V. Azalbert, V. Blasco-Baque, J. Puig, G. Xifra, W. Ricart, C. Tomlinson, M. Woodbridge, M. Cardellini, F. Davato, I. Cardolini, O. Porzio, P. Gentileschi, F. Lopez, F. Foufelle, S. A. Butcher, E. Holmes, J. K. Nicholson, C. Postic, R. Burcelin, M. E. Dumas, Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat. Med.* **24**, 1070–1080 (2018).
74. S. Fromentin, S. K. Forslund, K. Chechi, J. Aron-Wisnewsky, R. Chakaroun, T. Nielsen, V. Tremaroli, B. Ji, E. Prifti, A. Myridakis, J. Chilloux, P. Andrikopoulos, Y. Fan, M. T. Olanipekun, R. Alves, S. Adiouch, N. Bar, Y. Talmor-Barkan, E. Belda, R. Caesar, L. P. Coelho, G. Falony, S. Fellahi, P. Galan, N. Galleron, G. Helft, L. Hoyles, R. Isnard, E. Le Chatelier, H. Julienne, L. Olsson, H. K. Pedersen, N. Pons, B. Quinquis, C. Rouault, H. Roume, J. E. Salem, T. S. B. Schmidt, S. Vieira-Silva, P. Li, M. Zimmermann-Kogadeeva, C. Lewinter, N. B. Sondertoft, T. H. Hansen, D. Gauguier, J. P. Gotze, L. Kober, R. Kornowski, H. Vestergaard, T. Hansen, J. D. Zucker, S. Hercberg, I. Letunic, F. Backhed, J. M. Oppert, J. Nielsen, J. Raes, P. Bork, M. Stumvoll, E. Segal, K. Clement, M. E. Dumas, S. D. Ehrlich, O. Pedersen, Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat. Med.* **28**, 303–314 (2022).
75. Y. Talmor-Barkan, N. Bar, A. A. Shaul, N. Shahaf, A. Godneva, Y. Bussi, M. Lotan-Pompan, A. Weinberger, A. Shechter, C. Chezar-Azerrad, Z. Arow, Y. Hammer, K. Chechi, S. K. Forslund, S. Fromentin, M. E. Dumas, S. D. Ehrlich, O. Pedersen, R. Kornowski, E. Segal, Metabolomic and microbiome profiling reveals personalized risk factors for coronary artery disease. *Nat. Med.* **28**, 295–302 (2022).
76. M. O. Ruuskanen, P. P. Erawijantari, A. S. Havulinna, Y. Liu, G. Meric, J. Tuomilehto, M. Inouye, P. Jousilahti, V. Salomaa, M. Jain, R. Knight, L. Lahti, T. J. Niiranen, Gut microbiome composition is predictive of incident type 2 diabetes in a population cohort of 5,572 finnish adults. *Diabetes Care* **45**, 811–818 (2022).
77. E. Tas, S. Bai, X. Ou, K. Mercer, H. Lin, K. Mansfield, R. Buchmann, E. C. Diaz, J. Oden, E. Børsheim, S. H. Adams, J. Dranoff, Fibroblast growth factor-21 to adiponectin ratio: A potential biomarker to monitor liver fat in children with obesity. *Front. Endocrinol.* **11**, 654 (2020).
78. H. Li, Q. Fang, F. Gao, J. Fan, J. Zhou, X. Wang, H. Zhang, X. Pan, Y. Bao, K. Xiang, A. Xu, W. Jia, Fibroblast growth factor 21 levels are increased in nonalcoholic fatty liver disease patients and are correlated with hepatic triglyceride. *J. Hepatol.* **53**, 934–940 (2010).
79. E. Vilar-Gomez, N. Chalasani, Non-invasive assessment of non-alcoholic fatty liver disease: Clinical prediction rules and blood-based biomarkers. *J. Hepatol.* **68**, 305–315 (2018).
80. L. J. Alferink, J. C. Kiefte-de Jong, N. S. Erler, B. J. Veldt, J. D. Schoufour, R. J. de Knegt, M. A. Ikram, H. J. Metselaar, H. Janssen, O. H. Franco, S. Darwish Murad, Association of dietary macronutrient composition and non-alcoholic fatty liver disease in an ageing population: The Rotterdam Study. *Gut* **68**, 1088–1098 (2019).
81. M. Hashemian, S. Merat, H. Poustchi, E. Jafari, A. R. Radmard, F. Kamangar, N. Freedman, A. Hekmatdoost, M. Sheikh, P. Boffetta, R. Sinha, S. M. Dawsey, C. C. Abnet, R. Malekzadeh, A. Etemadi, Red meat consumption and risk of nonalcoholic fatty liver disease in a population with low meat consumption: The golestan cohort study. *Am. J. Gastroenterol.* **116**, 1667–1675 (2021).
82. H. S. Jung, Y. Chang, M. J. Kwon, E. Sung, K. E. Yun, Y. K. Cho, H. Shin, S. Ryu, Smoking and the risk of non-alcoholic fatty liver disease: A cohort study. *Am. J. Gastroenterol.* **114**, 453–463 (2019).
83. V. W. Wong, L. A. Adams, V. de Ledinghen, G. L. Wong, S. Sookoian, Noninvasive biomarkers in NAFLD and NASH - Current progress and future promise. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 461–478 (2018).
84. J. Y. Jung, J.-J. Shim, S. K. Park, J.-H. Ryoo, J.-M. Choi, I.-H. Oh, K.-W. Jung, H. Cho, M. Ki, Y.-J. Won, C.-M. Oh, Serum ferritin level is associated with liver steatosis and fibrosis in Korean general population. *Hepatol. Int.* **13**, 222–233 (2019).
85. J. Mayneris-Perxachs, M. Cardellini, L. Hoyles, J. Latorre, F. Davato, J. M. Moreno-Navarrete, M. Arnoriaga-Rodriguez, M. Serino, J. Abbott, R. H. Barton, J. Puig, X. Fernandez-Real, W. Ricart, C. Tomlinson, M. Woodbridge, P. Gentileschi, S. A. Butcher, E. Holmes, J. K. Nicholson, V. Perez-Brocal, A. Moya, D. M. Clain, R. Burcelin, M. E. Dumas, M. Federici, J. M. Fernandez-Real, Iron status influences non-alcoholic fatty liver disease in obesity through the gut microbiome. *Microbiome* **9**, 104 (2021).
86. Y. Heshiki, R. Vazquez-Uribe, J. Li, Y. Ni, S. Quainoo, L. Imamovic, J. Li, M. Sørensen, B. K. C. Chow, G. J. Weiss, A. Xu, M. O. A. Sommer, G. Panagiotou, Predictable modulation of cancer treatment outcomes by the gut microbiota. *Microbiome* **8**, 28 (2020).
87. D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalová, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav, E. Segal, Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
88. R. Z. Gharaibeh, C. Jobin, Microbiota and cancer immunotherapy: In search of microbial signals. *Gut* **68**, 385–388 (2019).
89. Y. Liu, Y. Wang, Y. Ni, C. K. Y. Cheung, K. S. L. Lam, Y. Wang, Z. Xia, D. Ye, J. Guo, M. A. Tse, G. Panagiotou, A. Xu, Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. *Cell Metab.* **31**, 77–91.e75 (2020).
90. J. S. Bajaj, N. S. Betrapally, P. B. Hylemon, L. R. Thacker, K. Daita, D. J. Kang, M. B. White, A. B. Unser, A. Fagan, E. A. Gavis, M. Sikaroodi, S. Dalmet, D. M. Heuman, P. M. Gillevet, Gut microbiota alterations can predict hospitalizations in cirrhosis independent of diabetes mellitus. *Sci. Rep.* **5**, 18559 (2015).
91. T. Wong, R. J. Wong, R. G. Gish, Diagnostic and treatment implications of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Gastroenterol. Hepatol.* **15**, 83–89 (2019).
92. J.-Z. Zhang, J.-J. Cai, Y. Yu, Z.-G. She, H. Li, Nonalcoholic fatty liver disease: An update on the diagnosis. *Gene Expr.* **19**, 187–198 (2019).
93. A. Duscha, B. Gisevius, S. Hirschberg, N. Yissachar, G. I. Stangl, E. Eilers, V. Bader, S. Haase, J. Kaisler, C. David, R. Schneider, R. Troisi, D. Zent, T. Hegelmaier, N. Dokalis, S. Gerstein, S. Del Mare-Roumani, S. Amidror, O. Staszewski, G. Poschmann, K. Stuhler, F. Hirche, A. Balogh, S. Kempa, P. Trager, M. M. Zaiss, J. B. Holm, M. G. Massa, H. B. Nielsen, A. Faissner, C. Lukas, S. G. Gatermann, M. Scholz, H. Przuntek, M. Prinz, S. K. Forslund, K. F. Winklhofer, D. N. Muller, R. A. Linker, R. Gold, A. Haghikia, Propionic acid shapes the multiple sclerosis disease course by an immunomodulatory mechanism. *Cell* **180**, 1067–1080.e1016 (2020).
94. P. Chen, X. Hou, G. Hu, L. Wei, L. Jiao, H. Wang, S. Chen, J. Wu, Y. Bao, W. Jia, Abdominal subcutaneous adipose tissue: A favorable adipose depot for diabetes? *Cardiovasc. Diabetol.* **17**, 93 (2018).

95. X. Hou, P. Chen, G. Hu, L. Wei, L. Jiao, H. Wang, Y. Liang, Y. Bao, W. Jia, Abdominal subcutaneous fat: A favorable or nonfunctional fat depot for glucose metabolism in chinese adults? *Obesity* **26**, 1078–1087 (2018).
96. J. N. Paulson, O. C. Stine, H. C. Bravo, M. Pop, Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
97. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
98. L. Zhao, Y. Ni, M. Su, H. Li, F. Dong, W. Chen, R. Wei, L. Zhang, S. P. Guiraud, F. P. Martin, C. Rajani, G. Xie, W. Jia, High throughput and quantitative measurement of microbial metabolome by gas chromatography/mass spectrometry using automated alkyl chloroformate derivatization. *Anal. Chem.* **89**, 5565–5577 (2017).
99. J. Li, C. Y. J. Sung, N. Lee, Y. Ni, J. Pihlajamäki, G. Panagiotou, H. El-Nezami, Probiotics modulated gut microbiota suppresses hepatocellular carcinoma growth in mice. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1306–E1315 (2016).
100. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (13 March 2013).
101. P. Dixon, VEGAN, A package of r functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
102. M. J. Vavrek, Fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontol. Electronica* **14**, 16 (2011).
103. M. Kuhn, Building predictive models inRUsing thecaretPackage. *J. Stat. Softw.* **28**, 1–26 (2008).
104. M. N. Wright, A. Ziegler, Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
105. C. R. John, MLeval: Machine Learning Model Evaluation. R package version 0.3, (2020); https://CRAN.R-project.org/package=MLeval.
106. S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
107. N. S. Artzi, S. Shilo, E. Hadar, H. Rossman, S. Barbash-Hazan, A. Ben-Haroush, R. D. Balicer, B. Feldman, A. Wiznitzer, E. Segal, Prediction of gestational diabetes based on nationwide electronic health records. *Nat. Med.* **26**, 71–76 (2020).
108. B. Greenwell, fastshap: Fast Approximate Shapley Values. R package version 0.0.7, (2021); https://CRAN.R-project.org/package=fastshap.
109. A. Kotronen, M. Peltonen, A. Hakkarainen, K. Sevastianova, R. Bergholm, L. M. Johansson, N. Lundbom, A. Rissanen, M. Ridderstrale, L. Groop, M. Orho-Melander, H. Yki-Jarvinen, Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. *Gastroenterology* **137**, 865–872 (2009).
110. R. K. Sterling, E. Lissen, N. Clumeck, R. Sola, M. C. Correa, J. Montaner, M. S. Sulkowski, F. J. Torriani, D. T. Dieterich, D. L. Thomas, D. Messinger, M. Nelson; APRICOT Clinical Investigators, Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology* **43**, 1317–1325 (2006).

Acknowledgments: We thank K. Zych (Clinical Microbiomics) for reviewing the machine learning analysis. **Funding:** This work was supported by Marie Sklodowska-Curie Actions (MSCA) and Innovative Training Networks, H2020-MSCA-ITN-2018 813781 "BestTreat" (G.P., H. Leung, S.L.S., D.P., E.N., and H.B.N.); the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC 2051, Project ID 390713860 (G.P.); the Hong Kong Research Grants Council/Area of Excellence (AoE/M/707-18) (A.X.); the EU-HK Co-funding Mechanism (E/HKU703/20) (A.X.); the National Natural Science Foundation of China (NSFC)–NHMRC joint research grant (81561128016) (W.J.); the Shanghai Municipal Key Clinical Specialty (W.J.); the National Key Research and Development Program of China (2018YFA0800402) (W.J.); the General Program of NSFC (81870598) (H. Li); the Excellent Young Scholars of NSFC (82022012) (H. Li); the Two Hundred Program from Shanghai Jiao Tong University School of Medicine (20191830) (H. Li); the Innovative Research Team of High-level Local Universities in Shanghai (SHSMU-ZDCX20212700) (H. Li); the Assistance Publique Hôpitaux de Paris (APHP) DGOS for the Programme hospitalier de Recherche clinique (microbaria) predictive value in identifying individuals with a high risk of developing NAFLD and EU transatlantic project Leducq and JPI-Microdiet (K.C.); the NCATS (5UL1TR001442) (R.L.); the NIDDK (U01DK061734, U01DK130190, R01DK106419, R01DK121378, R01DK124318, and P30DK120515) (R.L.); the NHLBI (P01HL147835) (R.L.); and the NIAAA (U01AA029019) (R.L.). **Author contributions:** G.P., W.J., H. Li, and Y.N. conceptualized and designed experiments. X.L. and L.Q. acquired the clinical samples and performed the experiment. H. Leung, Y.N., X.L., E.N., S.L.S., and D.P. processed and analyzed the data. H. Leung, Y.N., and G.P. wrote the original draft. Y.L., A.X., K.C., E.B., R.L., H. Li, and W.J. provided data for analysis and external validation. G.P., W.J., H. Li, Y.N., and K.H. lead and supervised the research work. H. Leung, Y.N., G.P., H.B.N., K.H., Y.L., A.X., K.C., R.L., H. Li, and W.J. reviewed and edited the manuscript. **Competing interests:** R.L. serves as a consultant to Aardvark Therapeutics, Altimmune, Anylam/Regeneron, Amgen, Arrowhead Pharmaceuticals, AstraZeneca, Bristol-Myer Squibb, CohBar, Eli Lilly, Galmed, Gilead, Glympse Bio, Hightide, Inipharma, Intercept, Inventiva, Ionis, Janssen Inc., Madrigal, Metacrine Inc., NGM Biopharmaceuticals, Novartis, Novo Nordisk, Merck, Pfizer, Sagimet, Theratechnologies, 89 Bio, Terns Pharmaceuticals, and Viking Therapeutics. In addition, his institutions received research grants from Arrowhead Pharmaceuticals, AstraZeneca, Boehringer-Ingelheim, Bristol-Myers Squibb, Eli Lilly, Galectin Therapeutics, Galmed Pharmaceuticals, Gilead, Intercept, Hanmi, Intercept, Inventiva, Ionis, Janssen, Madrigal Pharmaceuticals, Merck, NGM Biopharmaceuticals, Novo Nordisk, Merck, Pfizer, Sonic Incytes, and Terns Pharmaceuticals. R.L. is a co-founder of LipoNexus Inc. K.H. is the founder of the company Afekta Technologies providing metabolomic services. All other authors declare that they have no competing interests. **Data and materials availability:** All data associated with this study are in the paper or the Supplementary Materials. Raw metagenomic sequencing data for all samples have been deposited in NCBI Sequencing Read Archive under BioProject IDs PRJNA728908 and PRJNA686835 (only SRR13279648, SRR13279753, SRR13279664, and SRR13279666). Metabolite profiles have been deposited in MetaboLights (www.ebi.ac.uk/metabolights/) with accession MTBLS2615. The data used to generate main and supplementary figures are in data file S1. Further information and requests for resources and reagents should be directed to G.P. (gianni.panagiotou@leibniz-hki.de).

Submitted 21 June 2021
Resubmitted 1 February 2022
Accepted 4 May 2022
Published 8 June 2022
10.1126/scitranslmed.abk0855

DISCUSSION

The gut microbiome is considered a key element contributing to the regulation of human health and disease, being associated with the development and severity of numerous highly prevalent and chronic diseases (Thomas et al. 2017). Through the different manuscripts of this thesis, I have provided evidence that the human body together with its microbiome forms a unity of life or holobiont indispensable for the well-functioning of the organism.

During my Ph.D., I have applied state-of-the-art bioinformatics and statistical analyses to better understand and to give new insights into the human gut bacteriome and mycobiome and their implication in non-alcoholic fatty liver disease (NAFLD) progression. In addition, the potential of the gut microbiome to develop new clinical tools was also explored by applying novel machine learning approaches to metagenomic data. In this section I am going to further discuss the following five topics:

- Importance of confounders in microbiome-based data analyses (**manuscripts I, II, III, and IV**)
- Using metagenomics and metabolomics analyses to investigate potential therapies and diagnostic strategies for NAFLD (**manuscripts I, II, and IV**)
- Bioinformatic and machine learning approaches allowed the development of a microbiome-based stratification model (**manuscript II**)
- Limitations in mycobiome analyses (**manuscript III**)
- Synthesis of the four studies comprising this dissertation (**manuscripts I, II, III, and IV**)

Table 1| Summary table of the manuscripts that comprise this dissertation.

Manuscript	Title
I	Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations
II	Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions
III	Genetic variation in IL-17A regulation and mycobiome dysbiosis contribute to non-alcoholic fatty liver disease
IV	Risk assessment with gut microbiome and metabolite markers in NAFLD development

1. Importance of confounders in microbiome-based data analyses

The gut microbiome is a complex ecosystem exposed to numerous covariates and confounding variables (e.g., age, gender, BMI, antibiotics, diet, geography, etc.) (Hasan and Yang 2019). Therefore, the use of analytic approaches to address confounding is needed to obtain more reliable statistical associations to better evaluate the relationship between gut microbiome and disease.

Previous investigations have shown age to affect the gut microbial composition (Odamaki et al. 2016; Ghosh et al. 2020; Wilmanski et al. 2021). Elderly subjects have been found to have distinct gut microbial community characteristics compared to young/middle-aged individuals (Ghosh et al. 2020). In addition, gender has also been mentioned as an important variable affecting the gut microbiome, and it has been suggested that dysbiosis may drive sex differences in the progression of some diseases (Haro et al. 2016; Ahmed and Spence 2021). Haro et al. found that gender differences in the gut microbiome may be influenced by the grade of obesity. Ethnicity is another important variable to consider in microbiome analyses. Different investigations have explored the influence of ethnicity on the gut microbiota, and differences in the diversity of the microbial composition and in the abundances of some gut microbes have been shown between different ethnicities (Deschasaux et al. 2018; Dwiyanto et al. 2021; Boulund et al. 2022). Diet is also an important factor to take into account in microbiome studies as some foods may produce significant alterations in the gut microbial community (Johnson et al. 2019). In addition, medication is also a major confounder when analyzing the gut microbiome, as some medications including antibiotics and non-antibiotics drugs may alter the microbial community in the gut (Weersma et al. 2020). For example, commonly used drugs such as proton pump inhibitors (drugs used to treat acid-related disorders), laxatives (drugs to treat constipation), and metformin (oral blood glucose-lowering compound used in the treatment of T2D) have been shown to influence gut microbiome composition and function (Weersma et al. 2020). Antibiotics also have a drastic effect on the gut microbiota community (Taur et al. 2018; Palleja et al. 2018; Gutierrez et al. 2020; Seelbinder et al. 2020; Ramirez et al. 2020). Availability of the medication patient history during and some months before the clinical trial allows us to determine, when designing the analysis strategy, the most appropriate statistical analysis accounting for these drugs when convenient. Therefore, considering and accounting for these confounders in microbiome-based analyses has become a recommended practice in the last years (Gao et al. 2018; Zhong et al. 2019; Jie et al. 2021; Zuo et al. 2022).

In different projects that comprise this dissertation (**manuscripts I, III, and IV**), I made use of statistical approaches that allowed accounting for gut microbiome influential variables such as age, gender, and BMI, as they influence the statistical results when analyzing the gut microbiome community (Ghosh et al. 2020; Tierney et al. 2022). For each manuscript, apart from the main variables of age, gender, and BMI; the most appropriate confounders were chosen depending on the cohort, research question, and objectives of each project. In **manuscript III**, where we investigated changes in the fungal community in NAFLD subjects, our NAFLD cohort was free of antibiotic intake for at least 6 months before the sample collection. A recent study showed that antibiotics might have a long-term influence on the human mycobiome compared to the bacterial community (Seelbinder et al. 2020). Therefore, I investigated if antibiotics had a noticeable long-term effect in the fungal community of our cohort and I found a significant impact in alpha diversity between antibiotic-free and the individuals that took antibiotics more than six months before the sample collection in the NASH group. Knowing this, in the analyses performed with the full cohort, the antibiotic intake was considered as a confounder. This allowed me to keep the full cohort of samples to perform the analyses, without removing samples considering the antibiotic intake. In addition, analyses were also performed in the sub-cohort of antibiotic-free samples and the same results were found providing confident and reliable conclusions.

On the other side, accounting for specific confounders may be useful when investigating a certain disease in order to adjust for the major confounders linked to some conditions that may affect the analyses. To determine the major confounders that are of interest to be included, it is important to understand the disease pathogenesis and its implication with other complications. NAFLD is associated with metabolic and cardiovascular disorders such as T2D, obesity, and hypertension (Kolodziejczyk et al. 2019). Therefore, in the NAFLD-related projects of this dissertation (**manuscripts I, III, and IV**) differential abundance and correlation analyses were performed accounting for obesity- and diabetes- related parameters (e.g., BMI, VFA, SFA, adipo-IR or HOMA-IR) when appropriate. Adjusting for obesity- and diabetes-related variables allowed us to obtain conclusions linked directly with NAFLD progression and not to these conditions. In addition, in **manuscript I** where we explored RS supplementation as a potential microbiota-directed food intervention, we accounted for weight loss when we investigated the clinical changes that underwent the RS intervention group compared to the control group. We found that the amelioration of the disease progression in the RS intervention group, showed by a reduction of intrahepatic triglyceride content (IHTC, our primary outcome) and an improvement in liver injury and related metabolic disorders, was independent of the weight loss. One limitation when performing these types of analyses adjusting for confounders, is the possibility of having numerous missing values in the confounder variables. Missing or incomplete data on the confounding factors may lead to a result bias (Lin and Chen 2014). In addition, the available functions and tools to adjust for confounding do not usually allow to provide the factors for the adjustment with missing values. Therefore, missing samples need to be removed or methods for data imputation should be used to perform the analyses.

Another fact that I considered during my research to have more generalizable and confident conclusions is the validation of the results in different population cohorts when convenient, and for example in cohorts from different ethnicities. Nowadays, this task has become easier to put into practice thanks to the large amount of openly available data. In **manuscripts I and IV**, we investigated the role of the gut microbiome in two Chinese NAFLD cohorts. In **manuscript I**, I also explored the abundance of the identified detrimental species *B. stercoris* in two external cohorts, one Chinese and one European, confirming our results in different population cohorts including different ethnicities. In **manuscript IV**, we developed an early NAFLD detection machine learning model and the performance of the final model was also validated in several external case-control cohorts from Asia, the United States, and Europe. Apart from the non-invasive risk assessment model for NAFLD development, a model to classify different degrees of steatosis and a model to predict the fibrosis-4 (FIB-4) index change four years later were built and validated in the external cohorts. Lastly, in **manuscript II**, different cohorts with Asian and Caucasian ethnicities were used to build the microbiome-based stratification model for lifestyle interventions, and the final model was validated in two external cohorts with Caucasian and American ethnicities. Therefore, the microbiome-based machine learning models developed in **manuscripts II and IV** were validated in different cohorts including different ethnicities obtaining a good model performance supporting the biological relevance and generalizability of the models and the biomarkers identified.

In conclusion, the different manuscripts that comprise this dissertation have shown the importance of the covariates in microbiome-related data analysis to obtain generalizable and more comprehensive conclusions. Taking into consideration microbiome-related cofounder factors and including statistical approaches that allow accounting for confounders in microbiome-based studies, helps and improves the identification of disease-specific microbiome alterations, and it is encouraged to be applied for further microbiome-based studies. In addition, validation of the results in different population cohorts, when possible, is suggested to obtain more confident and generalizable results.

2. Using metagenomics and metabolomics analyses to investigate potential therapies and diagnostic strategies for NAFLD

Previous research has investigated the beneficial effects of resistant starch (RS) supplementation on host health (Zhu et al. 2022). One of the advantages of RS is that it is a relatively simple and inexpensive dietary strategy. RS has been found to decrease fat accumulation by improving insulin sensitivity and maintaining lipid metabolic homeostasis (Zhang et al. 2015; Maier et al. 2017). In a previous study, RS significantly improved insulin and low-density lipoprotein (LDL) in obese individuals (Eshghi et al. 2019). RS also showed a significant reduction of liver steatosis and in the serum levels of ALT, TG, and HOMA-IR in NAFLD mice (Shou et al. 2021). However, RS supplementation has never been explored in a clinical study on NAFLD patients. Therefore, in **manuscript I**, I made use of bioinformatic and statistical techniques in order to investigate RS supplementation as a potential microbiota-directed foods (MDFs) therapy to treat NAFLD. We performed a randomized clinical trial where we identified RS as having a beneficial effect with a reduction of the IHTC independently of weight loss in the intervention group. RS also reduced body weight, body fat, abdominal fat, liver enzymes, liver profiles, fasting insulin, insulin resistance, and serum fibroblast growth factor (FGF21). Combining shotgun metagenomics sequencing and targeted metabolomic profiling we discovered that improvements produced by RS intervention may lead to beneficial changes in the gut microbiota composition and the metabolic profile. We found decreased levels of microbial metabolic products, especially the AAs pool and BCAAs levels. Previous studies have demonstrated an increase in the circulating levels of BCAAs (leucine, isoleucine, and valine) in NAFLD and NASH (Kalhan et al. 2011; Goffredo et al. 2017; Gaggini et al. 2018). Correlation analysis showed BCAAs levels to be significantly positively correlated with IHTC, ALT, AST, and GGT. In addition, the correlation with IHTC was found to be independent of body weight and insulin resistance, suggesting a direct influence on BCAAs with NAFLD pathogenesis. We deeper investigated the effect of RS on the microbial composition, and we identified a significantly reduced abundance of *Bacteroides stercoris*, a species that was also correlated with IHTC, ALT, and AST levels when controlling for age, gender, HOMA-IR, and obesity-related parameters. This finding suggested that the effect of *B. stercoris* on NAFLD aggravation was independent of body weight and insulin resistance.

The availability of WGS data allowed us to profile the functional potential of the gut microbiota community. We found that the phenotypic alleviation of NAFLD may be potentially linked to a reduction in the intervention group of lipopolysaccharide (LPS) biosynthesis and export. Previous studies have associated an increase in LPS production with NAFLD and the progression of hepatic steatosis (Soares et al. 2010; Fukunishi et al. 2014; Carpino et al. 2020).

In **manuscript I**, a sophisticated computational framework was applied to integrate microbial species, functions, and host phenotypes. From this comprehensive analysis, *B. stercoris* was identified to contribute to the correlations of several BCAAs biosynthesis modules and the clinical phenotypes IHTC and FGF21. To date, no research has been done investigating the effect of *B. stercoris* on the aggravation of diseases. In this project, we validated the positive association of *B. stercoris* with NAFLD in mice by conducting a monocolonization study to confirm the NAFLD-promoting effect of *B. stercoris* and to explore the possible mechanisms involved.

Bioinformatic analyses performed suggested *B. stercoris* as a good candidate to investigate its role in NAFLD aggravation and it was validated in different experimental studies by our collaborators. Of note, experiments were carried out using a specific *B. stercoris* strain, despite the fact that the taxonomic profiling only reached the species level. In some cases, subtle genetic differences between strains within a single bacterial species can have profound impacts to induce different behaviors (Ghazi et al. 2022). For example, *Escherichia coli* strains have been shown to be commensal, pathogenic, host-associated, or environmental (Leimbach et al. 2013). Yet methods for strain-level assignment are insufficient, therefore numerous efforts are being done to improve and develop tools for more precise strain-level analyses. Especially, the main challenges consist of dealing with reference databases including highly similar reference strain genomes and the possibility of having multiple strains under one species in a sample (Liao et al. 2022). Advances in metagenomic strain-level population genomics will allow researchers to perform more accurate and high-resolution microbiome analyses.

The results presented up to here demonstrate the role of a single species, namely *B. stercoris*, in facilitating the progression of NAFLD through BCAA production. These results have helped to understand some mechanisms involved in the disease pathology. However, NAFLD is a complex disease caused by multiple mechanisms and not by a single species, but rather by, most probably, broad microbiome changes. To provide evidence about the causal role of whole microbiota changes in NAFLD, in **manuscript I** we performed FMT in mice fed with a Western diet. Previous mice studies have also explored gut microbiota's role in NAFLD through FMT (García-Lezana et al. 2018; Xue et al. 2019; Shou et al. 2021). Our study showed that FMT of RS-altered microbiota profile alleviates NAFLD by, among other findings, reducing hepatic steatosis, improving gut barrier integrity, promoting the expression of lipolysis, and reducing inflammation-related genes, suggesting a causal role of gut microbiota in attenuating NAFLD.

To help to understand the implication of not only bacteria but also the fungal community in NAFLD pathogenesis, in **manuscript III** we performed genotyping and mycobiome analyses in a NAFLD cohort and investigated associations between genetic variations in antifungal immunity and fungal changes in NAFLD. A previous study showed an increase in the frequency of IL-17-producing cells among intrahepatic $CD4^+$ T cells and a higher Th17/resting regulatory T cells (rTreg) ratio in peripheral blood in NAFLD progression to NASH (Rau et al. 2016). From the genotyping and bioinformatics analyses, we identified an association between the IL-17A rs2275913 variant and fibrosis severity that is accompanied by mycobiome dysbiosis, where we found increased abundances of the *Candida* CTG-clade in patients with advanced fibrosis. Previous studies have found an increase of *Candida* CTG-clade species such as *Candida albicans* and *Debaryomyces hansenii*, that were found higher in liver- and gastrointestinal-related diseases (Jain et al. 2021; Hartmann et al. 2021; Li et al. 2022b). In **manuscript III**, *ex vivo* T-cell stimulation assay performed by our collaborators demonstrated elevated IL-17A levels in response to *D. hansenii* and *C. albicans* lysates in rs2275913 minor allele variant, suggesting a Th17-stimulating potential. Together with the elevated *Candida* CTG-clade abundance in NAFLD patients identified from the bioinformatic analyses, our results suggest a combinatory effect of dysregulated antifungal immunity and an imbalance in *Candida* CTG-clade species on fibrosis development in patients with NASH. After investigating the fungal changes in our NAFLD cohort, I performed community network analysis that allowed the integration of bacterial and fungal genera and the identification of one subcommunity module containing *Candida* CTG-clade together with one more fungal and nine bacterial genera, that was found to be associated with NAFLD progression, suggesting multiple microbial factors contributing to NAFLD.

Trying to address the current challenge of finding new clinically reliable and cost-effective strategies for diagnosis and early detection of NAFLD mentioned in the introduction section 2.1, in **manuscript IV,** the potential value of the gut microbiome in NAFLD diagnosis was explored. The use of shotgun metagenomics and targeted metabolomics combined with clinical information allowed the development of a machine learning model able to early predict NAFLD development. A recent study has shown the predictive capacity of the gut microbiome for incident liver diseases (Liu et al. 2022). Liu et al. showed that adding microbiome information to conventional risk factors improved the model performance that successfully predicted the development of liver disease 15 years later. In relation to NAFLD, up to now, some studies have built predictive models for early detection of NAFLD using clinical biomarkers such as fatty liver index (FLI), FGF21, BMI, or TG, and glucose index (TyG) (Li et al. 2013; Zheng et al. 2018; Motamed et al. 2020). In **manuscript IV**, we characterized the gut microbiome of the cohort and we found at the baseline differences in the microbiome signature and metabolic shifts in subjects that will develop NAFLD compared to controls. Novel machine learning approaches were used to build a non-invasive risk assessment model to predict NAFLD progression and to identify microbial signatures in participants at risk of developing NAFLD in the next four and a half years. Feature selection approaches identified 18 features including bacterial genera, pathways, metabolites, and clinical parameters; as potential early NAFLD biomarkers having a relevant role in early NAFLD detection. In addition, incorporating the final model microbiome features into previous clinical models (Li et al. 2013; Zheng et al. 2018; Motamed et al. 2020), produced significant improvements demonstrating the potential of the microbial community for early NAFLD detection. Finally, to further explore the relevance of the selected features, a new model to classify mild or severe steatosis was built using the 18 selected features obtaining a good model performance.

In conclusion, the metagenomic and metabolomic approaches used in this thesis have allowed a better understanding of the implication of the gut microbiome in NAFLD development and progression. Comprehensive analyses and application of bioinformatic pipelines have helped to identify microbial biomarkers and to validate the vital role of the bacteriome, mycobiome, and metabolome in NAFLD pathogenesis. These new findings may direct the development of new microbiome-based therapies and diagnostic strategies for NAFLD, that will help to improve the population's quality of life and to reduce the healthcare cost for NAFLD management.

3. Bioinformatic and machine learning approaches allowed the development of a microbiome-based stratification model

In the last years, the application of machine learning approaches in microbiome-related studies has been widely applied. Machine learning has become a powerful tool to identify microbiome signatures and microbial biomarkers. It is also employed in model development for diagnostics, biotherapeutic selection, patient stratification, and early disease detection (Richens et al. 2020; Spiga et al. 2021; Huang et al. 2021). In this thesis, machine learning approaches were applied in **manuscripts II and IV** to develop models for microbiome-based stratification and early NAFLD detection respectively. In this section, I am going to focus on **manuscript II** where I applied machine learning approaches to build a microbiome-based patient stratification model.

Nowadays novel microbiome-targeted interventions that aim to target and alter the gut microbial community to improve the host health are widely investigated and applied (e.g., **manuscript I**). However, clinical trials have demonstrated that there are large interindividual differences in the treatment response and some of these differences may depend on subject-specific gut microbial composition (Cotillard et al. 2013; Korpela et al. 2014; Tap et al. 2015). The gut microbiota is resistant to changes and some microbiome-targeted interventions are not going to affect the microbial composition of some individuals. Knowing in advance if a microbiome community will react to specific lifestyle components may help to optimize personalized lifestyle approaches. Therefore, in **manuscript II,** we aimed to identify microbial biomarkers among the gut microbial community associated with the degree of change in the microbiome structure and develop a machine learning model able to predict microbial responsiveness.

In order to build the stratification model, we first applied bioinformatic analyses to investigate the resistance potential of the gut microbial ecosystem of an individual using a stability metric called intraclass correlation coefficient (ICC), and established a criterion for distinguishing significant changes from natural fluctuations of the gut microbiota community. Differentially abundant analysis was performed to identify highly abundant species in the individuals that showed minor impact on the microbial's community structure also named as non-responders, and in the individuals that showed significant changes in the microbial community in response to lifestyle interventions or responders. From the differentially abundant and microbiome stability analyses we identified 3 species, *Prevotella copri*, *Bacteroides stercoris,* and *Bacteroides vulgatus,* to be potential biomarkers of microbiome resistance. We also explored the species highly abundant in responders and a total of 38 species were identified. Interestingly, most of these species were correlated with at least one amino acid biosynthesis pathway. We further investigated amino acid biosynthesis-related pathways and explored the top species that contributed to these functions. Network analysis performed showed differences between responders' and non-responders' community networks. We identified in the responders network that amino acid auxotroph species suggested by Yu et al. study (Yu et al. 2022) were also correlated with amino acid-related KOs when integrating the functional profile into the network. Previous studies have shown that metabolic cross-feeding is a crucial process involved in the development and composition of the microbial community (Henriques et al. 2020; Lopez and Wingreen 2021). In addition, Mee et al. observed that microbial genomes for amino acid biosynthesis are absent in a substantial proportion of all bacteria (Mee and Wang 2012), therefore, amino acid auxotrophy may have an important role in promoting cooperative interactions between different bacteria in the microbiome (Mee et al. 2014).

The different analyses performed in the first part of the study suggested that amino acid biosynthesis has an important role in microbiome dynamics; plus, comprehensive analyses identified signature species in responders' and non-responders' microbial profiles that may serve as potential microbial biomarkers of the microbiome's resistance to lifestyle interventions. These differences found at the baseline microbial community between responders and non-responders suggested that machine learning approaches may be able to successfully predict the gut microbiome responsiveness to a lifestyle intervention using the gut microbial profile before starting the intervention.

Previous studies have developed machine learning models that predict host phenotype changes such as insulin resistance (Liu et al. 2020), fat change (Jian et al. 2022), or significant weight loss (Dong et al. 2021) in response to lifestyle interventions using the microbial community at the baseline. However, few studies have focused on investigating microbiome resistance in response to lifestyle interventions, with only one recent study that proposed a method for preliminary assessment of the microbiome response using amplicon data (Klimenko et al. 2022).

In **manuscript II**, I made use of gradient boosting, a widely used method for regression and classification, to build a model that predicts the microbiome resistance to lifestyle interventions using the microbial composition of the individuals before they start the intervention. Gradient boosting is a popular machine learning algorithm that has been used across different domains including metagenomics (Ryan et al. 2020; Liu et al. 2022). In this project, I investigated how different microbiome-related data (species, genera, and pathways) were able to predict the microbiome resistance to change after lifestyle interventions. A final taxonomic-based model was built achieving a good performance with an AUC up to 0.86 in two validation population cohorts from different ethnicities. Recursive feature elimination was used as a feature selection method to subset the top 30 features used by the model. A total of 12 genera and 18 species contributing to the prediction of microbiome resistance to interventions were selected from which 19 were identified in previous analyses of the study as associated with microbiome responsiveness, confirming the relevance of these taxa in the microbiome dynamics in response to lifestyle interventions. Therefore, the predictive model built may potentially serve as an initial step to personalized lifestyle intervention therapies by applying a microbiome-based stratification strategy.

Despite our promising results, we are aware that a larger number of samples would allow us to perform more advanced machine learning algorithms, such as deep learning. In this sense, the availability of more longitudinal cohorts with clinical and biochemical characterization would be beneficial for future studies. In addition, the integration of clinical and biochemical profiles may help to build more precise and personalized models.

In conclusion, this study demonstrated that there are differences in the baseline microbial community signatures that characterize the microbiome's resistance to lifestyle interventions, and we present a machine learning model that predicts the microbiome resistance to interventions using the baseline microbiome composition. The application of machine learning approaches is helping to improve the design of personalized lifestyle approaches and to develop new models for therapeutic strategies. Future approaches combining stratification by host phenotype and by gut microbial characterization will optimize the intervention response of individual patients, achieving a greater success in treating patients.

4. Limitations in mycobiome analyses

Even though the mycobiome constitutes only a small proportion of the microbes living in the gastrointestinal tract, the potential role of the fungal community in the gut is increasingly recognized (Huseyin et al. 2017). Yet, the gut mycobiome remains underinvestigated, being bacteria the main focus of study in the last decades. Therefore, mycobiome data analyses are still limited remaining one step behind compared to bacteriome analyses (Thielemann et al. 2022). Apart from challenges in the sample collection, sample processing, and storage; several limitations still arise concerning the bioinformatic analyses including the lack of reference databases, standardization procedures, and analysis pipelines or tools.

The availability of well-curated and high-quality databases is needed for comprehensive and more robust taxonomic classification, and fungal databases are often being updated (Lücking et al. 2020). The choice of the database has been shown to clearly impact the study results with considerable differences in the community compositions (Huseyin et al. 2017; Thielemann et al. 2022). The curation of databases and fungal profiling tools need to deal with the complexity of the fungal taxonomy and the sparse databases. Due to the complex fungal taxonomy, in **manuscript III**, fungi were grouped according to genus except for the characterization of *Candida*. *Candida* is a polyphyletic genus comprising a large variety of phylogenetically distant species, and to account for this, we grouped it in the *Candida* CTG-clade. This clade belongs to the order *Saccharomycetales* and clusters the *Candida spp.* characterized by an alternative decoding of the CTG codon leading to a serine amino acid instead of the canonical leucine (Mühlhausen and Kollmar 2014). Some of the species belonging to this clade are *C. albicans*, C. *tropicalis*, *C. parapsilosis*, and *Debaryomyces hansenii* (formerly known as *C. famata*). However, other species are distantly related to *Candida* CTG-clade and were partially renamed and regrouped in other genera, for example, *Nakaseomyces glabrata* (commonly known as *Candida glabrata*) and *Kluyveromyces marxianus* (formerly *Candida kefyr*) (Borman and Johnson 2020).

The choice of the profiling strategy and the primer may influence the analysis results. Nuclear ribosomal DNA internal transcribed spacer (ITS) has been consensually selected as the formal barcode to assess the fungal composition as this region can amplify the fungal rDNA from the majority of the species (Huseyin et al. 2017). High throughput sequencing strategies focused on two different ITS subloci that are separated by the conserved 5.8S region, the ITS1 and ITS2 regions, however, it is still not clear which of them has the best taxonomic resolution (Yang et al. 2018). In addition, the amplicon strategy used may influence the study outcome due to different fungal taxonomic identification (Monard et al. 2013; Yang et al. 2018; Frau et al. 2019). In **manuscript III**, ITS1 data was used to access the fungal profile of our NAFLD cohort. Even though ITS2 may have slightly different taxonomic identification, our results go in line with the only previous NAFLD mycobiome study where they used a different primer strategy (ITS2). Therefore, primer bias seems not to have a major influence on the results for overall highly abundant fungal species, despite the fact that previous research has reported ITS1 probably lead to missing species detection from Basidiomycetes phylum (i.e. *Malassezia* and *Cryptococcus*) (Frau et al. 2019). In addition, amplicon sequencing strategies remain to be improved. A study investigating the identification of Basidiomycota species using different primer strategies showed that, neither the complete ITS region nor the sub-regions successfully identified 11 of the 113 Basidiomycota genera (Badotti et al. 2017). Amplicon sequencing has been shown to be helpful for identifying low-abundance fungi, even though it also has some disadvantages as it is subject to selective primer bias or lower taxonomic resolution (Tiew et al. 2020). An alternative is the use of whole-genome shotgun (WGS) metagenomics to study the mycobiome. Different from amplicon sequencing, WGS metagenomics sequences the total DNA from a given sample. However, the availability of tools to characterize the fungal community based on metagenomic reads is almost inexistent. The low abundance of fungi relative to bacterial DNA across a variety of human samples is a challenge for WGS mycobiome profiling. High sequencing depth that goes together with high cost is likely required to determine the fungal composition using WGS metagenomics (Tiew et al. 2020). Much work remains to improve and develop new strategies and protocols for more accurate and sensitive fungal WGS metagenomic profiling.

Finally, more efforts to develop standardized protocols need to be done for fungi. For example, studies have found that the DNA extraction method influences the microbial community obtained (Sui et al. 2020). The impact has been shown to be lower when extracting bacterial DNA, as most of the protocols have been optimized for bacteria, leading to more biased results for other microbes such as fungi (Leigh Greathouse et al. 2019).

Therefore, more investigation is required to evaluate optimal DNA extraction methods and protocols to study fungal communities.

In conclusion, knowledge about the gut mycobiome and how it affects human health and disease is still limited, and new advances and research need to be done in this field. The development of novel tools, standardized protocols, and more complete and curated databases is crucial for achieving more comprehensive mycobiome analyses. Only one previous study together with **manuscript III** has investigated the fungal community in NAFLD development and has demonstrated an important role of the mycobiome in NAFLD progression. Therefore, more studies are required to deeper elucidate the fungal role and its link with bacterial changes in NAFLD.

5. Synthesis of the four studies comprising my dissertation

Numerous studies have found that different types of environmental and lifestyle factors have a clear impact on the gut microbiome of individuals. An important gut microbiome modulator is diet, which has a lot of potential for developing microbiome-based therapies. Consequently, much research has focused on understanding how different types of foods modulate the gut microbiome. For instance, Barone et al. conducted a study demonstrating that a Mediterranean diet promotes a gut microbiome composition associated with health benefits, while a high-fat Western diet leads to gut microbiome alterations with negative implications for metabolic health, including a greater relative abundance of asaccharolytic bacteria as well as of fat- and bile-loving microorganisms (Barone et al. 2019). Ketogenic diet was also found to produce changes in the gut microbial composition associated with an improvement of metabolic health in different diseases including epilepsy, obesity, and dyslipidemia (Attaye et al. 2022). In this dissertation, my manuscript published in Cell Metabolism (**manuscript I, Ni et al. in press**) investigated the potential use of RS as microbiome-directed food to treat NAFLD. In this study, we found that RS supplementation modulates the gut microbiome composition, and possible mediators related to the beneficial effects of RS were evaluated. However, as demonstrated in my publication in Microbiome (**manuscript II, Chen et al. 2023**), each individual's microbiome community response is different and predictable. The ML model that I developed in this study can be used for patient stratification in dietary studies. Therefore, if I had to repeat **manuscript I** study, besides the randomization based on age, gender, and other factors, I could use this ML model to stratify patients including the same proportion of individuals with resistant and flexible gut microbiomes in the two groups.

Apart from the use of machine learning for patient stratification, this novel discipline has a variety of applications in the microbiome field including phenotypic prediction, biomarker discovery, and treatment outcome evaluation among others (Li et al. 2022a). For instance, Heshiki et al. developed a ML model that successfully predicts cancer treatment output using the microbiome information prior to therapy initiation (Heshiki et al. 2020). Another study by Chen et al. constructed a highly accurate age prediction model using gut microbiome data (Chen et al. 2022). My publication in Science Translational Medicine (**manuscript IV, Leung et al. 2022**) presents, for the first time, a microbiome-based prognostic model for NAFLD. This model incorporates clinical data, metabolites, and gut bacteria community information. However, as demonstrated in my collaborative manuscript with the University of Würzburg (**manuscript III**), not only bacteria but also fungi are playing an important role in the development of NAFLD. Taking into consideration the importance of the fungal community in NAFLD and the recent development of new methods for fungal identification using shotgun metagenomics data (Xie and Manichanh 2022; Narunsky-Haziza et al. 2022; Salem-Bango et al. 2023), it would be very interesting to repeat **manuscript IV** study incorporating the fungal profile in the model. In this way, it would be possible to evaluate the prognostic power and potential of the fungal composition for NAFLD prognosis

REFERENCES

Agus A, Clément K, Sokol H (2021) Gut microbiota-derived metabolites as central regulators in metabolic disorders. Gut 70:1174–1182. https://doi.org/10.1136/gutjnl-2020-323071

Ahmed S, Spence JD (2021) Sex differences in the intestinal microbiome: interactions with risk factors for atherosclerosis and cardiovascular disease. Biol Sex Differ 12:1–12. https://doi.org/10.1186/s13293-021-00378-z

Alonso A, Marsal S, Julià A (2015) Analytical methods in untargeted metabolomics: State of the art in 2015. Front Bioeng Biotechnol 3:23. https://doi.org/10.3389/fbioe.2015.00023

Arrieta MC, Stiemsma LT, Amenyogbe N, et al (2014) The intestinal microbiome in early life: Health and disease. Front Immunol 5:427. https://doi.org/10.3389/fimmu.2014.00427

Attaye I, van Oppenraaij S, Warmbrunn M V., Nieuwdorp M (2022) The role of the gut microbiota on the beneficial effects of ketogenic diets. Nutrients 14:191. https://doi.org/10.3390/NU14010191/S1

Auslander N, Gussow AB, Koonin E V. (2021) Incorporating Machine Learning into Established Bioinformatics Frameworks. International Journal of Molecular Sciences 2021, Vol 22, Page 2903 22:2903. https://doi.org/10.3390/IJMS22062903

Avis T, Wilson FX, Khan N, et al (2021) Targeted microbiome-sparing antibiotics. Drug Discov Today 26:2198–2203. https://doi.org/10.1016/j.drudis.2021.07.016

Badotti F, De Oliveira FS, Garcia CF, et al (2017) Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi). BMC Microbiol 17:1–12. https://doi.org/10.1186/S12866-017-0958-X

Barone M, Turroni S, Rampelli S, et al (2019) Gut microbiome response to a modern Paleolithic diet in a Western lifestyle context. PLoS One 14:. https://doi.org/10.1371/JOURNAL.PONE.0220619

Barratt MJ, Lebrilla C, Shapiro HY, Gordon JI (2017) The gut microbiota, food science, and human nutrition: A timely marriage. Cell Host Microbe 22:134–141. https://doi.org/10.1016/j.chom.2017.07.006

Bauer KC, Littlejohn PT, Ayala V, et al (2022) Nonalcoholic fatty liver disease and the gut-liver axis: Exploring an undernutrition perspective. Gastroenterology 162(7):1858-1875.e2. https://doi.org/10.1053/j.gastro.2022.01.058

Beghini F, McIver LJ, Blanco-Míguez A, et al (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. Elife 10:. https://doi.org/10.7554/ELIFE.65088

Bolyen E, Rideout JR, Dillon MR, et al (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9

Borman AM, Johnson EM (2020) Name changes for fungi of medical importance, 2018 to 2019. J Clin Microbiol 59:1811–1831. https://doi.org/10.1128/JCM.01811-20

Bosco N, Noti M (2021) The aging gut microbiome and its impact on host immunity. Genes Immun 22:289–303. https://doi.org/10.1038/s41435-021-00126-8

Boulund U, Bastos DM, Ferwerda B, et al (2022) Gut microbiome associations with host genotype vary across ethnicities and potentially influence cardiometabolic traits. Cell Host Microbe. https://doi.org/10.1016/j.chom.2022.08.013

Boursier J, Mueller O, Barret M, et al (2016) The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. Hepatology 63:764–775. https://doi.org/10.1002/hep.28356

Callahan BJ, McMurdie PJ, Rosen MJ, et al (2016) DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods 13:581. https://doi.org/10.1038/NMETH.3869

Cambiaghi A, Ferrario M, Masseroli M (2017) Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief Bioinform 18:498–510. https://doi.org/10.1093/BIB/BBW031

Carpino G, del Ben M, Pastori D, et al (2020) Increased liver localization of lipopolysaccharides in human and experimental NAFLD. Hepatology 72:470–485. https://doi.org/10.1002/HEP.31056

Carvalho T (2023) First oral fecal microbiota transplant therapy approved. Nat Med. https://doi.org/10.1038/D41591-023-00046-2

Caspi R, Altman T, Billington R, et al (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42:D459–D471. https://doi.org/10.1093/NAR/GKT1103

Chen L, Wang D, Garmaeva S, et al (2021) The long-term genetic stability and individual specificity of the human gut microbiome. Cell 184:2302-2315.e12. https://doi.org/10.1016/J.CELL.2021.03.024

Chen Y, Wang H, Lu W, et al (2022) Human gut microbiome aging clocks based on taxonomic and functional signatures through multi-view learning. Gut Microbes 14:. https://doi.org/10.1080/19490976.2021.2025016

Cheng R, Wang L, Le S, et al (2022) A randomized controlled trial for response of microbiome network to exercise and diet intervention in patients with nonalcoholic fatty liver disease. Nat Commun 13:1–13. https://doi.org/10.1038/s41467-022-29968-0

Chua LL, Rajasuriar R, Azanan MS, et al (2017) Reduced microbial diversity in adult survivors of childhood acute lymphoblastic leukemia and microbial associations with increased immune activation. Microbiome 5:1–14. https://doi.org/10.1186/S40168-017-0250-1

Conlon MA, Bird AR (2015) The Impact of diet and lifestyle on gut microbiota and human health. Nutrients 7:17. https://doi.org/10.3390/NU7010017

Cotillard A, Kennedy SP, Kong LC, et al (2013) Dietary intervention impact on gut microbial gene richness. Nature 500:585–588. https://doi.org/10.1038/nature12480

Cuomo P, Capparelli R, Iannelli A, Iannelli D (2022) Role of Branched-Chain Amino Acid Metabolism in Type 2 Diabetes, Obesity, Cardiovascular Disease and Non-Alcoholic Fatty Liver Disease. International Journal of Molecular Sciences 2022, Vol 23, Page 4325 23:4325. https://doi.org/10.3390/IJMS23084325

De Souza Nascimento E, Ahmed I, Oliveira E, et al (2019) Understanding Development Process of Machine Learning Systems: Challenges and Solutions. International Symposium on Empirical Software Engineering and Measurement 2019-Septemer: https://doi.org/10.1109/ESEM.2019.8870157

Dekaboruah E, Suryavanshi MV, Chettri D, Verma AK (2020) Human microbiome: an academic update on human body site specific surveillance and its possible role. Arch Microbiol 202:2147–2167. https://doi.org/10.1007/S00203-020-01931-X

DeMartino P, Cockburn DW (2020) Resistant starch: impact on the gut microbiome and health. Curr Opin Biotechnol 61:66–71. https://doi.org/10.1016/J.COPBIO.2019.10.008

Demir M, Lang S, Hartmann P, et al (2022) The fecal mycobiome in non-alcoholic fatty liver disease☆. J Hepatol 76:788–799. https://doi.org/10.1016/J.JHEP.2021.11.029

DeSantis TZ, Hugenholtz P, Larsen N, et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072. https://doi.org/10.1128/AEM.03006-05

Deschasaux M, Bouter KE, Prodan A, et al (2018) Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med 24:1526–1531. https://doi.org/10.1038/S41591-018-0160-1

Dobranowski PA, Stintzi A (2021) Resistant starch, microbiome, and precision modulation. Gut Microbes 13:1926842. https://doi.org/10.1080/19490976.2021.1926842

Dong TS, Luu K, Lagishetty V, et al (2021) The intestinal microbiome predicts weight loss on a calorie-restricted diet and is associated with improved hepatic steatosis. Front Nutr 8:718661. https://doi.org/10.3389/FNUT.2021.718661/FULL

Dong X, Kleiner M, Sharp CE, et al (2017) Fast and simple analysis of MiSeq amplicon sequencing data with MetaAmp. Front Microbiol 8:1461. https://doi.org/10.3389/FMICB.2017.01461

Dongiovanni P, Valenti L (2017) A nutrigenomic approach to non-alcoholic fatty liver disease. Int J Mol Sci 18:. https://doi.org/10.3390/IJMS18071534

Dwiyanto J, Hussain MH, Reidpath D, et al (2021) Ethnicity influences the gut microbiota of individuals sharing a geographical location: a cross-sectional study from a middle-income country. Sci Rep 11:1–10. https://doi.org/10.1038/s41598-021-82311-3

Ehrlich SD (2011) MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract. Metagenomics of the Human Body 307–316. https://doi.org/10.1007/978-1-4419-7089-3_15

Ejtahed HS, Angoorani P, Soroush AR, et al (2020) Gut microbiota-derived metabolites in obesity: a systematic review. Biosci Microbiota Food Health 39:65–76. https://doi.org/10.12938/BMFH.2019-026

El Bouchefry K, de Souza RS (2020) Learning in Big Data: Introduction to Machine Learning. Knowledge Discovery in Big Data from Astronomy and Earth Observation: Astrogeoinformatics 225–249. https://doi.org/10.1016/B978-0-12-819154-5.00023-0

Eshghi F, Bakhshimoghaddam F, Rasmi Y, Alizadeh M (2019) Effects of resistant starch supplementation on glucose metabolism, lipid profile, lipid peroxidation marker, and oxidative stress in overweight and obese adults: Randomized, double-blind, crossover trial. Clin Nutr Res 8:318. https://doi.org/10.7762/CNR.2019.8.4.318

Faith JJ, Guruge JL, Charbonneau M, et al (2013) The long-term stability of the human gut microbiota. Science (1979) 341:. https://doi.org/https://doi.org/10.1126/science.1237439

Fan D, Coughlin LA, Neubauer MM, et al (2015) Activation of HIF-1α and LL-37 by commensal bacteria inhibits Candida albicans colonization. Nat Med 21:808. https://doi.org/10.1038/NM.3871

Fassarella M, Blaak EE, Penders J, et al (2021) Gut microbiome stability and resilience: elucidating the response to perturbations in order to modulate gut health. Gut 70:595–605. https://doi.org/10.1136/GUTJNL-2020-321747

Fotis D, Liu J, Dalamaga M (2022) Could gut mycobiome play a role in NAFLD pathogenesis? Insights and therapeutic perspectives. Metabol Open 14:100178. https://doi.org/10.1016/J.METOP.2022.100178

Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al (2019) Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 4:293. https://doi.org/10.1038/S41564-018-0306-4

Frau A, Kenny JG, Lenzi L, et al (2019) DNA extraction and amplicon production strategies deeply influence the outcome of gut mycobiome studies. Scientific Reports 2019 9:1 9:1–17. https://doi.org/10.1038/s41598-019-44974-x

Friedman J, Alm EJ (2012) Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol 8:e1002687. https://doi.org/10.1371/JOURNAL.PCBI.1002687

Fritsch J, Garces L, Quintero MA, et al (2021) Low-fat, high-fiber diet reduces markers of inflammation and dysbiosis and improves quality of life in patients with ulcerative colitis. Clinical Gastroenterology and Hepatology 19:1189-1199.e30. https://doi.org/10.1016/J.CGH.2020.05.026

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol 26:5369–5406. https://doi.org/10.1111/MEC.14264

Fukunishi S, Sujishi T, Takeshita A, et al (2014) Lipopolysaccharides accelerate hepatic steatosis in the development of nonalcoholic fatty liver disease in Zucker rats. J Clin Biochem Nutr 54:39. https://doi.org/10.3164/JCBN.13-49

Gaggini M, Carli F, Rosso C, et al (2018) Altered amino acid concentrations in NAFLD: Impact of obesity and insulin resistance. Hepatology 67:145–158. https://doi.org/10.1002/HEP.29465

Gao X, Zhang M, Xue J, et al (2018) Body mass index differences in the gut microbiota are gender specific. Front Microbiol 9:. https://doi.org/10.3389/FMICB.2018.01250/FULL

García-Lezana T, Raurell I, Bravo M, et al (2018) Restoration of a healthy intestinal microbiota normalizes portal hypertension in a rat model of nonalcoholic steatohepatitis. Hepatology 67:1485–1498. https://doi.org/https://doi.org/10.1002/hep.29646

Ghazi AR, Münch PC, Chen D, et al (2022) Strain identification and quantitative analysis in microbial communities. J Mol Biol 434:167582. https://doi.org/10.1016/J.JMB.2022.167582

Ghosh TS, Das M, Jeffery IB, O'Toole PW (2020) Adjusting for age improves identification of gut microbiome alterations in multiple diseases. Elife 9:. https://doi.org/10.7554/ELIFE.50240

Glassner KL, Abraham BP, Quigley EMM (2020) The microbiome and inflammatory bowel disease. J Allergy Clin Immunol 145:16–27. https://doi.org/10.1016/J.JACI.2019.11.003

Goffredo M, Santoro N, Tricò D, et al (2017) A branched-chain amino acid-Related metabolic signature characterizes obese adolescents with non-alcoholic fatty liver disease. Nutrients 9:. https://doi.org/10.3390/NU9070642

Gomez-Casati DF, Zanor MI, Busi M v. (2013) Metabolomics in plants and humans: Applications in the prevention and diagnosis of diseases. Biomed Res Int 2013:. https://doi.org/10.1155/2013/792527

Gonzalez-Covarrubias V, Martínez-Martínez E, Bosque-Plata L del (2022) The potential of metabolomics in biomedical applications. Metabolites 12:. https://doi.org/10.3390/METABO12020194

Gupta S, Allen-Vercoe E, Petrof EO (2016) Fecal microbiota transplantation: in perspective. Therap Adv Gastroenterol 9:229. https://doi.org/10.1177/1756283X15607414

Gutierrez D, Weinstock A, Antharam VC, et al (2020) Antibiotic-induced gut metabolome and microbiome alterations increase the susceptibility to Candida albicans colonization in the gastrointestinal tract. FEMS Microbiol Ecol 96:. https://doi.org/10.1093/FEMSEC/FIZ187

Gweon HS, Oliver A, Taylor J, et al (2015) PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods Ecol Evol 6:973–980. https://doi.org/10.1111/2041-210X.12399

Han M, Yang P, Zhong C, Ning K (2018) The Human Gut Virome in Hypertension. Front Microbiol 9:. https://doi.org/10.3389/FMICB.2018.03150/FULL

Haro C, Rangel-Zúñiga OA, Alcalá-Díaz JF, et al (2016) Intestinal microbiota is influenced by gender and body mass index. PLoS One 11:e0154090. https://doi.org/10.1371/JOURNAL.PONE.0154090

Hartmann P, Lang S, Zeng S, et al (2021) Dynamic Changes of the Fungal Microbiome in Alcohol Use Disorder. Front Physiol 12:. https://doi.org/10.3389/FPHYS.2021.699253

Hasan N, Yang H (2019) Factors affecting the composition of the gut microbiota, and its modulation. PeerJ 7:. https://doi.org/10.7717/PEERJ.7502

He LH, Yao DH, Wang LY, et al (2021) Gut microbiome-mediated alteration of immunity, inflammation, and metabolism involved in the regulation of non-alcoholic fatty liver disease. Front Microbiol 12:3262. https://doi.org/10.3389/FMICB.2021.761836

Henderickx JGE, de Weerd H, Groot Jebbink LJ, et al (2022) The first fungi: mode of delivery determines early life fungal colonization in the intestine of preterm infants. Microbiome Research Reports 1:7. https://doi.org/10.20517/MRR.2021.03

Henriques SF, Dhakan DB, Serra L, et al (2020) Metabolic cross-feeding in imbalanced diets allows gut microbes to improve reproduction and alter host behaviour. Nat Commun 11:1–15. https://doi.org/10.1038/s41467-020-18049-9

Hernandez Roman J, Siddiqui MS (2020) The role of noninvasive biomarkers in diagnosis and risk stratification in nonalcoholic fatty liver disease. Endocrinol Diabetes Metab 3:e00127. https://doi.org/10.1002/EDM2.127

Herrema H, Niess JH (2020) Intestinal microbial metabolites in human metabolism and type 2 diabetes. Diabetologia 63:2533–2547. https://doi.org/10.1007/S00125-020-05268-4

Heshiki Y, Vazquez-Uribe R, Li J, et al (2020) Predictable modulation of cancer treatment outcomes by the gut microbiota. Microbiome 8:. https://doi.org/10.1186/s40168-020-00811-2

Higgins JA, Jackman MR, Brown IL, et al (2011) Resistant starch and exercise independently attenuate weight regain on a high fat diet in a rat model of obesity. Nutr Metab (Lond) 8:1–15. https://doi.org/10.1186/1743-7075-8-49

Hu H, Lin A, Kong M, et al (2020) Intestinal microbiome and NAFLD: molecular insights and therapeutic perspectives. J Gastroenterol 55:142. https://doi.org/10.1007/S00535-019-01649-8

Huang K, Xiao C, Glass LM, et al (2021) Machine learning applications for therapeutic tasks with genomics data. Patterns 2:100328. https://doi.org/10.1016/J.PATTER.2021.100328

Huseyin CE, O’Toole PW, Cotter PD, Scanlan PD (2017) Forgotten fungi-the gut mycobiome in human health and disease. FEMS Microbiol Rev 41:. https://doi.org/10.1093/FEMSRE/FUW047

Hutchison ER, Kasahara K, Zhang Q, et al (2023) Dissecting the impact of dietary fiber type on atherosclerosis in mice colonized with different gut microbial communities. npj Biofilms and Microbiomes 2023 9:1 9:1–12. https://doi.org/10.1038/s41522-023-00402-7

Jain U, ver Heul AM, Xiong S, et al (2021) Debaryomyces is enriched in Crohn’s disease intestinal tissue and impairs healing in mice. Science 371:1154–1159. https://doi.org/10.1126/SCIENCE.ABD0919

Jandhyala SM, Talukdar R, Subramanyam C, et al (2015) Role of the normal gut microbiota. World Journal of Gastroenterology : WJG 21:8787. https://doi.org/10.3748/WJG.V21.I29.8787

Jeznach-Steinhagen A, Ostrowska J, Czerwonogrodzka-Senczyna A, et al (2019) Dietary and Pharmacological Treatment of Nonalcoholic Fatty Liver Disease. Medicina (B Aires) 55:. https://doi.org/10.3390/MEDICINA55050166

Jia X, Xu W, Zhang L, et al (2021) Impact of gut microbiota and microbiota-related metabolites on hyperlipidemia. Front Cell Infect Microbiol 11:634780. https://doi.org/10.3389/FCIMB.2021.634780

Jian C, Silvestre MP, Middleton D, et al (2022) Gut microbiota predicts body fat change following a low-energy diet: a PREVIEW intervention study. Genome Med 14:1–18. https://doi.org/10.1186/S13073-022-01053-7

Jiang D, Armour CR, Hu C, et al (2019) Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. Front Genet 10:995. https://doi.org/10.3389/FGENE.2019.00995

Jie Z, Liang S, Ding Q, et al (2021) A transomic cohort as a reference point for promoting a healthy human gut microbiome. Medicine in Microecology 8:100039. https://doi.org/10.1016/J.MEDMIC.2021.100039

Johnson AJ, Vangay P, Al-Ghalith GA, et al (2019) Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. Cell Host Microbe 25:789-802.e5. https://doi.org/10.1016/J.CHOM.2019.05.005

Kalhan SC, Guo L, Edmison J, et al (2011) Plasma metabolomic profile in nonalcoholic fatty liver disease. Metabolism 60:404–413. https://doi.org/10.1016/J.METABOL.2010.03.006

Kanehisa M, Sato Y, Kawashima M, et al (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462. https://doi.org/10.1093/NAR/GKV1070

Kechagias S, Ekstedt M, Simonsson · Christian, Nasr P (2022) Non-invasive diagnosis and staging of non-alcoholic fatty liver disease. Hormones 2022 1:1–20. https://doi.org/10.1007/S42000-022-00377-8

Kim S (2015) ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun Stat Appl Methods 22:665. https://doi.org/10.5351/CSAM.2015.22.6.665

Klimenko NS, Odintsova VE, Revel-Muroz A, Tyakht A v. (2022) The hallmarks of dietary intervention-resilient gut microbiome. npj Biofilms and Microbiomes 2022 8:1 8:1–11. https://doi.org/10.1038/s41522-022-00342-8

Koh JH, Kim WU (2017) Dysregulation of gut microbiota and chronic inflammatory disease: from epithelial defense to host immunity. Exp Mol Med 49:e337. https://doi.org/10.1038/EMM.2017.55

Kolodziejczyk AA, Zheng D, Shibolet O, Elinav E (2019) The role of the microbiome in NAFLD and NASH. EMBO Mol Med 11:. https://doi.org/10.15252/EMMM.201809302

Korpela K (2021) Impact of delivery mode on infant gut microbiota. Ann Nutr Metab 77:11–19. https://doi.org/10.1159/000518498

Korpela K, Flint HJ, Johnstone AM, et al (2014) Gut Microbiota Signatures Predict Host and Microbiota Responses to Dietary Interventions in Obese Individuals. PLoS One 9:e90702. https://doi.org/10.1371/JOURNAL.PONE.0090702

Kurtz ZD, Müller CL, Miraldi ER, et al (2015) Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Comput Biol 11:e1004226. https://doi.org/10.1371/JOURNAL.PCBI.1004226

Lavelle A, Sokol H (2020) Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 17:223–237. https://doi.org/10.1038/s41575-019-0258-z

Lazarus J v., Mark HE, Villota-Rivas M, et al (2022) The global NAFLD policy review and preparedness index: Are countries ready to address this silent public health challenge? J Hepatol 76:771–780. https://doi.org/10.1016/J.JHEP.2021.10.025

Lee DH (2017) Imaging evaluation of non-alcoholic fatty liver disease: focused on quantification. Clin Mol Hepatol 23:290. https://doi.org/10.3350/CMH.2017.0042

Leigh Greathouse K, Sinha R, Vogtmann E (2019) DNA extraction for human microbiome studies: The issue of standardization. Genome Biol 20:1–4. https://doi.org/10.1186/S13059-019-1843-8/TABLES/1

Leimbach A, Hacker J, Dobrindt U (2013) E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Curr Top Microbiol Immunol 358:3–32. https://doi.org/10.1007/82_2012_303

Li H, Dong K, Fang Q, et al (2013) High serum level of fibroblast growth factor 21 is an independent predictor of non-alcoholic fatty liver disease: A 3-year prospective study in China. J Hepatol 58:557–563. https://doi.org/10.1016/J.JHEP.2012.10.029

Li P, Luo H, Ji B, Nielsen J (2022a) Machine learning for data integration in human gut microbiome. Microbial Cell Factories 2022 21:1 21:1–16. https://doi.org/10.1186/S12934-022-01973-4

Li X v., Leonardi I, Putzel GG, et al (2022b) Immune regulation by fungal strain diversity in inflammatory bowel disease. Nature 2022 603:7902 603:672–678. https://doi.org/10.1038/s41586-022-04502-w

Li W-Z, Stirling K, Yang J-J, Zhang L (2020) Gut microbiota and diabetes: From correlation to causality and mechanism. World J Diabetes 11:293. https://doi.org/10.4239/wjd.v11.i7.293

Li X, Leonardi I, Semon A, et al (2018) Response to Fungal Dysbiosis by Gut Resident CX3CR1+ Mononuclear Phagocytes Aggravates Allergic Airway Disease. Cell Host Microbe 24:847. https://doi.org/10.1016/j.chom.2018.11.003

Liao H, Ji Y, Sun Y (2022) Accurate strain-level microbiome composition analysis from short reads. bioRxiv 01.26.477962. https://doi.org/10.1101/2022.01.26.477962

Lin HW, Chen YH (2014) Adjustment for Missing Confounders in Studies Based on Observational Databases: 2-Stage Calibration Combining Propensity Scores From Primary and Validation Data. Am J Epidemiol 180:308–317. https://doi.org/10.1093/AJE/KWU130

Liu Y, Méric G, Havulinna AS, et al (2022) Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. Cell Metab 34:719. https://doi.org/10.1016/j.cmet.2022.03.002

Liu Y, Wang Y, Ni Y, et al (2020) Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. Cell Metab 31:77-91.e5. https://doi.org/10.1016/j.cmet.2019.11.001

Lloyd-Price J, Abu-Ali G, Huttenhower C (2016) The healthy human microbiome. Genome Med 8:1–11. https://doi.org/10.1186/s13073-016-0307-y

Loftfield E, Herzig KH, Caporaso JG, et al (2020) Association of Body Mass Index with Fecal Microbial Diversity and Metabolites in the Northern Finland Birth Cohort. Cancer Epidemiol Biomarkers Prev 29:2289. https://doi.org/10.1158/1055-9965.EPI-20-0824

Loomba R, Seguritan V, Li W, et al (2017) Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. Cell Metab 25:1054-1062.e5. https://doi.org/10.1016/j.cmet.2017.04.001

Lopez JG, Wingreen NS (2021) Noisy metabolism can drive the evolution of microbial cross-feeding. bioRxiv 2021.06.02.446805. https://doi.org/10.1101/2021.06.02.446805

Lu J, Shi P, Li H (2019) Generalized linear models with linear constraints for microbiome compositional data. Biometrics 75:235–244. https://doi.org/10.1111/biom.12956

Lücking R, Aime MC, Robbertse B, et al (2020) Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? IMA Fungus 11:1–32. https://doi.org/10.1186/s43008-020-00033-z

Ma Y, You X, Mai G, et al (2018) A human gut phage catalog correlates the gut phageome with type 2 diabetes. Microbiome 6:. https://doi.org/10.1186/S40168-018-0410-Y

Magro DO, Santos A, Guadagnini D, et al (2019) Remission in Crohn's disease is accompanied by alterations in the gut microbiota and mucins production. Sci Rep 9:1–10. https://doi.org/10.1038/s41598-019-49893-5

Maier T v., Lucio M, Lee LH, et al (2017) Impact of dietary resistant starch on the human gut microbiome, metaproteome, and metabolome. mBio 8:. https://doi.org/10.1128/mBio.01343-17

Malla MA, Dubey A, Kumar A, et al (2019) Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment. Front Immunol 10:2868. https://doi.org/10.3389/fimmu.2018.02868

Mardinoglu A, Wu H, Bjornson E, et al (2018) An Integrated understanding of the rapid metabolic benefits of a carbohydrate-restricted diet on hepatic steatosis in humans. Cell Metab 27:559. https://doi.org/10.1016/j.cmet.2018.01.005

Masoodi M, Gastaldelli A, Hyötyläinen T, et al (2021) Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. Nature Reviews Gastroenterology & Hepatology 2021 18:12 18:835–856. https://doi.org/10.1038/s41575-021-00502-9

Matchado MS, Lauber M, Reitmeier S, et al (2021) Network analysis methods for studying microbial communities: A mini review. Comput Struct Biotechnol J 19:2687–2698. https://doi.org/10.1016/j.csbj.2021.05.001

Matijašić M, Meštrović T, Paljetak HČ, et al (2020) Gut microbiota beyond bacteria—mycobiome, virome, archaeome, and eukaryotic parasites in IBD. Int J Mol Sci 21:. https://doi.org/10.3390/ijms21082668

McDonnell L, Gilkes A, Ashworth M, et al (2021) Association between antibiotics and gut microbiome dysbiosis in children: systematic review and meta-analysis. Gut Microbes 13:1–18. https://doi.org/10.1080/19490976.2020.1870402

Mee MT, Collins JJ, Church GM, Wang HH (2014) Syntrophic exchange in synthetic microbial communities. Proc Natl Acad Sci U S A 111:E2149–E2156. https://doi.org/10.1073/PNAS.1405641111

Mee MT, Wang HH (2012) Engineering ecosystems and synthetic ecologies. Mol Biosyst 8:2470–2483. https://doi.org/10.1039/C2MB25133G

Mistry J, Chuguransky S, Williams L, et al (2021) Pfam: The protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913

Monard C, Gantner S, Stenlid J (2013) Utilizing ITS1 and ITS2 to study environmental fungal diversity using pyrosequencing. FEMS Microbiol Ecol 84:165–175. https://doi.org/10.1111/1574-6941.12046

Montroy J, Berjawi R, Lalu MM, et al (2020) The effects of resistant starches on inflammatory bowel disease in preclinical and clinical settings: a systematic review and meta-analysis. BMC Gastroenterol 20:1–14. https://doi.org/10.1186/S12876-020-01516-4

Motamed N, Faraji AH, Khonsari MR, et al (2020) Fatty liver index (FLI) and prediction of new cases of non-alcoholic fatty liver disease: A population-based study of northern Iran. Clinical Nutrition 39:468–474. https://doi.org/10.1016/j.clnu.2019.02.024

Mühlhausen S, Kollmar M (2014) Predicting the fungal CUG codon translation with Bagheera. BMC Genomics 15:1–10. https://doi.org/10.1186/1471-2164-15-411

Mukaka MM (2012) A guide to appropriate use of Correlation coefficient in medical research. Malawi Med J 24:69

Napolitano M, Covasa M (2020) Microbiota transplant in the treatment of obesity and diabetes: current and future perspectives. Front Microbiol 11:2877. https://doi.org/10.3389/fmicb.2020.590370

Narunsky-Haziza L, Sepich-Poore GD, Livyatan I, et al (2022) Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. Cell 185:3789-3806.e17. https://doi.org/10.1016/J.CELL.2022.09.005

Nash AK, Auchtung TA, Wong MC, et al (2017) The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome 5:153. https://doi.org/10.1186/s40168-017-0373-4

Newgard CB (2017) Cell metabolism review metabolomics and metabolic diseases: Where do we stand? Cell Metab 25:43–56. https://doi.org/10.1016/j.cmet.2016.09.018

Nilsson RH, Larsson KH, Taylor AFS, et al (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res 47:D259. https://doi.org/10.1093/NAR/GKY1022

Niu SY, Yang J, McDermaid A, et al (2018) Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. Brief Bioinform 19:1415–1429. https://doi.org/10.1093/BIB/BBX051

Odamaki T, Kato K, Sugahara H, et al (2016) Age-related changes in gut microbiota composition from newborn to centenarian: A cross-sectional study. BMC Microbiol 16:1–12. https://doi.org/10.1186/S12866-016-0708-5

Olsson LM, Boulund F, Nilsson S, et al (2022) Dynamics of the normal gut microbiota: A longitudinal one-year population study in Sweden. Cell Host Microbe 30:1–14. https://doi.org/10.1016/j.chom.2022.03.002

Palleja A, Mikkelsen KH, Forslund SK, et al (2018) Recovery of gut microbiota of healthy adults following antibiotic exposure. Nat Microbiol 3:1255–1265. https://doi.org/10.1038/s41564-018-0257-9

Pan Z, Chen Y, Zhou M, et al (2021) Microbial interaction-driven community differences as revealed by network analysis. Comput Struct Biotechnol J 19:6000. https://doi.org/10.1016/J.CSBJ.2021.10.035

Paulson JN, Colin Stine O, Bravo HC, Pop M (2013) Differential abundance analysis for microbial marker-gene surveys. Nature Methods 2013 10:12 10:1200–1202. https://doi.org/10.1038/nmeth.2658

Piazzolla VA, Mangia A (2020) Noninvasive Diagnosis of NAFLD and NASH. Cells 9:. https://doi.org/10.3390/CELLS9041005

Pinart M, Dötsch A, Schlicht K, et al (2022) Gut microbiome composition in obese and non-obese persons: A systematic review and meta-analysis. Nutrients 14:12. https://doi.org/10.3390/NU14010012/S1

Polakof S, Díaz-Rubio ME, Dardevet D, et al (2013) Resistant starch intake partly restores metabolic and inflammatory alterations in the liver of high-fat-diet-fed rats. J Nutr Biochem 24:1920–1930. https://doi.org/10.1016/J.JNUTBIO.2013.05.008

Postler TS, Ghosh S (2017) Understanding the Holobiont: How Microbial Metabolites Affect Human Health and Shape the Immune System. Cell Metab 26:110–130. https://doi.org/10.1016/j.cmet.2017.05.008

Quast C, Pruesse E, Yilmaz P, et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/NAR/GKS1219

Quigley EMM, Gajula P (2020) Recent advances in modulating the microbiome. F1000Res 9:46. https://doi.org/10.12688/F1000RESEARCH.20204.1

Quiroga R, Nistal E, Estébanez B, et al (2020) Exercise training modulates the gut microbiota profile and impairs inflammatory signaling pathways in obese children. Exp Mol Med 52:1048–1061. https://doi.org/10.1038/s12276-020-0459-0

R Core Team (2022) R: A Language and Environment for Statistical Computing

Ramirez J, Guarner F, Bustos Fernandez L, et al (2020) Antibiotics as major disruptors of gut microbiota. Front Cell Infect Microbiol 10:731. https://doi.org/10.3389/fcimb.2020.572912

Rashidi A, Ebadi M, Rehman TU, et al (2021) Gut microbiota response to antibiotics is personalized and depends on baseline microbiota. Microbiome 9:1–11. https://doi.org/10.1186/S40168-021-01170-2

Rau M, Schilling A-K, Meertens J, et al (2016) Progression from nonalcoholic fatty liver to nonalcoholic steatohepatitis Is marked by a higher frequency of Th17 cells in the liver and an increased Th17/resting regulatory T cell ratio in peripheral blood and in the liver. J Immunol 196:97–105. https://doi.org/10.4049/JIMMUNOL.1501175

Riazi K, Azhari H, Charette JH, et al (2022) The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 7:851–861. https://doi.org/10.1016/S2468-1253(22)00165-0

Richens JG, Lee CM, Johri S (2020) Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun 11:1–9. https://doi.org/10.1038/s41467-020-17419-7

Rinninella E, Raoul P, Cintoni M, et al (2019) What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. Microorganisms 7:14. https://doi.org/10.3390/MICROORGANISMS7010014

Risely A (2020) Applying the core microbiome to understand host–microbe systems. Journal of Animal Ecology 89:1549–1558. https://doi.org/10.1111/1365-2656.13229

Rodríguez MM, Pérez D, Javier Chaves F, et al (2015) Obesity changes the human gut mycobiome. Scientific Reports 2015 5:1 5:1–15. https://doi.org/10.1038/srep14600

Rohlke F, Stollman N (2012) Fecal microbiota transplantation in relapsing Clostridium difficile infection. Therap Adv Gastroenterol 5:403. https://doi.org/10.1177/1756283X12453637

Rosado CP, Rosa VHC, Martins BC, et al (2020) Resistant starch from green banana (Musa sp.) attenuates non-alcoholic fat liver accumulation and increases short-chain fatty acids production in high-fat diet-induced obesity in mice. Int J Biol Macromol 145:1066–1072. https://doi.org/10.1016/J.IJBIOMAC.2019.09.199

Ryan FJ, Ahern AM, Fitzgerald RS, et al (2020) Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. Nat Commun 11:1–12. https://doi.org/10.1038/s41467-020-15342-5

Salem-Bango Z, Price TK, Chan JL, et al (2023) Fungal Whole-Genome Sequencing for Species Identification: From Test Development to Clinical Utilization. Journal of fungi 9:. https://doi.org/10.3390/JOF9020183

Santos S de S, Takahashi DY, Nakata A, Fujita A (2014) A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief Bioinform 15:906–918. https://doi.org/10.1093/BIB/BBT051

Schloss PD, Westcott SL, Ryabin T, et al (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/https://doi.org/10.1128/AEM.01541-09

Schnabl B, Brenner DA (2014) Interactions between the intestinal microbiome and liver diseases. Gastroenterology 146:1513–1524. https://doi.org/10.1053/J.GASTRO.2014.01.020

Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA (2016) Untargeted metabolomics strategies – Challenges and emerging directions. J Am Soc Mass Spectrom 27:1897. https://doi.org/10.1007/S13361-016-1469-Y

Schwaid G (2017) Epidemiology and Biostatistics. Board Review in Preventive Medicine and Public Health 79–185. https://doi.org/10.1016/B978-0-12-813778-9.00003-7

Seelbinder B, Chen J, Brunke S, et al (2020) Antibiotics create a shift from mutualism to competition in human gut communities with a longer-lasting impact on fungi than bacteria. Microbiome 8:1–20. https://doi.org/10.1186/S40168-020-00899-6

Shankar J, Solis N v., Mounaud S, et al (2015) Using Bayesian modelling to investigate factors governing antibiotic-induced Candida albicans colonization of the GI tract. Sci Rep 5:. https://doi.org/10.1038/SREP08131

Shou D, Cao C, Xu H, et al (2021) Type 2 resistant starch improves liver steatosis induced by high-fat diet relating to gut microbiota regulation and concentration of propionic acid in portal vein blood in C57BL/6J mice. Gut 70:A13–A14. https://doi.org/10.1136/GUTJNL-2021-IDDF.20

Shulaev V (2006) Metabolomics technology and bioinformatics. Brief Bioinform 7:128–139. https://doi.org/10.1093/BIB/BBL012

Si J, Lee G, You HJ, et al (2021) Gut microbiome signatures distinguish type 2 diabetes mellitus from non-alcoholic fatty liver disease. Comput Struct Biotechnol J 19:5920–5930. https://doi.org/10.1016/J.CSBJ.2021.10.032

Smits LP, Bouter KEC, de Vos WM, et al (2013) Therapeutic potential of fecal microbiota transplantation. Gastroenterology 145:946–953. https://doi.org/10.1053/J.GASTRO.2013.08.058

Soares JB, Pimentel-Nunes P, Roncon-Albuquerque R, Leite-Moreira A (2010) The role of lipopolysaccharide/toll-like receptor 4 signaling in chronic liver diseases. Hepatol Int 4:659. https://doi.org/10.1007/S12072-010-9219-X

Sommer F, Anderson JM, Bharti R, et al (2017) The resilience of the intestinal microbiota influences health and disease. Nat Rev Microbiol 15:630–638. https://doi.org/10.1038/nrmicro.2017.58

Spiga O, Cicaloni V, Dimitri GM, et al (2021) Machine learning application for patient stratification and phenotype/genotype investigation in a rare disease. Brief Bioinform 22:. https://doi.org/10.1093/BIB/BBAA434

Stephens CR, Easton JF, Robles-Cabrera A, et al (2020) The impact of education and age on metabolic disorders. Front Public Health 8:180. https://doi.org/10.3389/FPUBH.2020.00180

Sui HY, Weil AA, Nuwagira E, et al (2020) Impact of DNA Extraction Method on Variation in Human and Built Environment Microbial Community and Functional Profiles

Assessed by Shotgun Metagenomics Sequencing. Front Microbiol 11:953. https://doi.org/10.3389/fmicb.2020.00953

Suzek BE, Huang H, McGarvey P, et al (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288. https://doi.org/10.1093/BIOINFORMATICS/BTM098

Tap J, Furet JP, Bensaada M, et al (2015) Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults. Environ Microbiol 17:4954–4964. https://doi.org/10.1111/1462-2920.13006

Taur Y, Coyte K, Schluter J, et al (2018) Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant. Sci Transl Med 10:eaap9489. https://doi.org/10.1126/scitranslmed.aap9489

The Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49:D325–D334. https://doi.org/10.1093/NAR/GKAA1113

Thielemann N, Herz M, Kurzai O, Martin R (2022) Analyzing the human gut mycobiome – A short guide for beginners. Comput Struct Biotechnol J 20:608–614. https://doi.org/10.1016/J.CSBJ.2022.01.008

Thomas S, Izard J, Walsh E, et al (2017) The host microbiome regulates and maintains human health: A primer and perspective for non-microbiologists. Cancer Res 77:1783. https://doi.org/10.1158/0008-5472.CAN-16-2929

Tierney BT, Tan Y, 1¤ ID, et al (2022) Systematically assessing microbiome-disease associations identifies drivers of inconsistency in metagenomic research. PLoSBiol 20:e3001556. https://doi.org/10.1371/journal.pbio.3001556

Tiew PY, mac Aogain M, Ali NABM, et al (2020) The Mycobiome in Health and Disease: Emerging Concepts, Methodologies and Challenges. Mycopathologia 2020 185:2 185:207–231. https://doi.org/10.1007/S11046-019-00413-Z

Tovo C v., Villela-Nogueira CA, Leite NC, et al (2019) Transient hepatic elastography has the best performance to evaluate liver fibrosis in non-alcoholic fatty liver disease (NAFLD). Ann Hepatol 18:445–449. https://doi.org/10.1016/J.AOHEP.2018.09.003

Tripathi A, Debelius J, Brenner DA, et al (2018) The gut-liver axis and the intersection with the microbiome. Nat Rev Gastroenterol Hepatol 15:397–411. https://doi.org/10.1038/S41575-018-0011-Z

Turnbaugh PJ, Gordon JI (2008) An Invitation to the Marriage of Metagenomics and Metabolomics. Cell 134:708–713. https://doi.org/10.1016/J.CELL.2008.08.025

Turnbaugh PJ, Ley RE, Hamady M, et al (2007) The Human Microbiome Project. Nature 449:804–810. https://doi.org/10.1038/nature06244

Valdes AM, Walter J, Segal E, Spector TD (2018) Role of the gut microbiota in nutrition and health. BMJ 361:36–44. https://doi.org/10.1136/BMJ.K2179

Wang X, Zhang P, Zhang X (2021) Probiotics regulate gut microbiota: An effective method to improve immunity. Molecules 26:6076. https://doi.org/10.3390/MOLECULES26196076

Wang Y, Chen J, Song YH, et al (2019) Effects of the resistant starch on glucose, insulin, insulin resistance, and lipid parameters in overweight or obese adults: a systematic review and meta-analysis. Nutr Diabetes 9:19. https://doi.org/10.1038/S41387-019-0086-9

Ward TL, Dominguez-Bello MG, Heisel T, et al (2018) Development of the human mycobiome over the first month of life and across body sites. mSystems 3:140–157. https://doi.org/10.1128/MSYSTEMS.00140-17

Weersma RK, Zhernakova A, Fu J (2020) Interaction between drugs and the gut microbiome. Gut 69:1510–1519. https://doi.org/10.1136/GUTJNL-2019-320204

Wilmanski T, Diener C, Rappaport N, et al (2021) Gut microbiome pattern reflects healthy ageing and predicts survival in humans. Nat Metab 3:274–286. https://doi.org/10.1038/S42255-021-00348-0

Wong AC, Levy M (2019) New approaches to microbiome-based therapies. mSystems 4:e00122-19. https://doi.org/10.1128/mSystems.00122-19

Wood DE, Salzberg SL (2014) Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:1–12. https://doi.org/10.1186/GB-2014-15-3-R46

Xiao L, Liu Q, Luo M, Xiong L (2021) Gut microbiota-derived metabolites in irritable bowel syndrome. Front Cell Infect Microbiol 11:880. https://doi.org/10.3389/FCIMB.2021.729346

Xie Z, Manichanh C (2022) FunOMIC: Pipeline with built-in fungal taxonomic and functional databases for human mycobiome profiling. Comput Struct Biotechnol J 20:3685–3694. https://doi.org/10.1016/J.CSBJ.2022.07.010

Xue L-F, Luo W-H, Wu L-H, et al (2019) Fecal microbiota transplantation for the treatment of nonalcoholic fatty liver disease. http://www.xiahepublishing.com/ 4:12–18. https://doi.org/10.14218/ERHM.2018.00025

Yang RH, Su JH, Shang JJ, et al (2018) Evaluation of the ribosomal DNA internal transcribed spacer (ITS), specifically ITS1 and ITS2, for the analysis of fungal diversity by deep sequencing. PLoS One 13:. https://doi.org/10.1371/JOURNAL.PONE.0206428

Yao Y, Cai X, Ye Y, et al (2021) The role of microbiota in infant health: From early life to adulthood. Front Immunol 12:4114. https://doi.org/10.3389/FIMMU.2021.708472

Younossi ZM (2019) Non-alcoholic fatty liver disease - A global public health perspective. J Hepatol 70:531–544. https://doi.org/10.1016/J.JHEP.2018.10.033

Yu JSL, Correia-Melo C, Zorrilla F, et al (2022) Microbial communities form rich extracellular metabolomes that foster metabolic interactions and promote drug tolerance. Nat Microbiol 7:542–555. https://doi.org/10.1038/s41564-022-01072-5

Zamkovaya T, Foster JS, de Crécy-Lagard V, Conesa A (2021) A network approach to elucidate and prioritize microbial dark matter in microbial communities. ISME J 15:228. https://doi.org/10.1038/S41396-020-00777-X

Zeevi D, Korem T, Zmora N, et al (2015) Personalized Nutrition by Prediction of Glycemic Responses. Cell 163:1079–1094. https://doi.org/10.1016/j.cell.2015.11.001

Zhang L, Li HT, Shen L, et al (2015) Effect of dietary resistant starch on prevention and treatment of obesity-related diseases and its possible mechanisms. Biomed Environ Sci 28:291–297. https://doi.org/10.3967/BES2015.040

Zhang L, Zhan H, Xu W, et al (2021) The role of gut mycobiome in health and diseases. Therap Adv Gastroenterol 14:. https://doi.org/10.1177/17562848211047130

Zhang X, Zhu X, Wang C, et al (2016) Non-targeted and targeted metabolomics approaches to diagnosing lung cancer and predicting patient prognosis. Oncotarget 7:63437. https://doi.org/10.18632/ONCOTARGET.11521

Zhang Y, Chen L, Hu M, et al (2020) Dietary type 2 resistant starch improves systemic inflammation and intestinal permeability by modulating microbiota and metabolites in aged mice on high-fat diet. Aging 12:9173–9187. https://doi.org/10.18632/AGING.103187

Zhao L, Zhang F, Ding X, et al (2018) Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science (1979) 359:1151–1156. https://doi.org/10.1126/SCIENCE.AAO5774/

Zheng R, Du Z, Wang M, et al (2018) A longitudinal epidemiological study on the triglyceride and glucose index and the incident nonalcoholic fatty liver disease. Lipids Health Dis 17:1–9. https://doi.org/10.1186/S12944-018-0913-3

Zhong H, Penders J, Shi Z, et al (2019) Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. Microbiome 7:1–14. https://doi.org/10.1186/S40168-018-0608-Z

Zhu W, Zhou Y, Tsao R, et al (2022) Amelioratory effect of resistant starch on non-alcoholic fatty liver disease via the gut-liver axis. Front Nutr 0:969. https://doi.org/10.3389/FNUT.2022.861854

Zhuang L, Chen H, Zhang S, et al (2019) Intestinal microbiota in early life and its implications on childhood health. Genomics Proteomics Bioinformatics 17:13–25. https://doi.org/10.1016/J.GPB.2018.10.002

Zuo W, Michail S, Sun F (2022) Metagenomic analyses of multiple gut datasets revealed the association of phage signatures in colorectal cancer. Front Cell Infect Microbiol 12:749. https://doi.org/10.3389/FCIMB.2022.918010/BIBTEX

DECLARATION

I, Sara Leal Siliceo, as a doctoral student, hereby confirm that:

- I am familiar with the valid doctoral examination regulations.
- I produced this doctoral thesis myself, I neither used any text passages from third parties nor their own previous final theses without citing them.
- I cited the tools, personal information, and sources having been used in this thesis.
- I provide the names of the persons who assisted the applicant in selecting and analyzing materials and supported them in writing the manuscript.
- I did not receive any assistance from specialized consultants and that any third party did not receive either direct or indirect financial benefits from me for the work connected to the doctoral thesis submission.
- I have not already submitted the doctoral thesis project as my final thesis for a state examination or other scientific examination.
- I did not submit the same, a substantially similar, or another scientific paper to any other institution of higher education or to any other faculty.

Place, date	Sara Leal Siliceo

ACKNOWLEDGMENTS

Firstly, special thanks to my supervisor Prof. Dr. Gianni Panagiotou for his guidance, encouragement, and valuable insights. His support and supervision helped me to have the right direction for my research and to develop my career. I also truly appreciate the guidance and technical support during these four years from Dr. Yueqiong Ni.

I would also like to express my sincere gratitude to my colleagues and friends from the SBI group who have supported me throughout the journey of completing this dissertation and make me also enjoy my time in Jena during these four years.

Thanks to my salseros friends from Jena for giving me enjoyable and funny memories.

Thanks also to all my friends from Spain who always supported me from the distance.

Last but not least, I would like to thank all my family who gave me love and helped me during this journey. Special thanks to my parents, sister, and granny. Truly thanks to my love Alejandro that always encourages and supports my dreams.

CURRICULUM VITAE

Personal Details

Name:	Sara Leal Siliceo
Date and place of birth:	06.07.1994, Madrid, Spain
ORCID:	0000-0002-9076-3962

Education

2019 – present

Doctor of Philosophy (Ph.D.)
Faculty of Biological Sciences, Friedrich Schiller University, Jena, Germany

2017 – 2018

M.Sc. in Bioinformatics and Computational Biology
Autonoma University of Madrid, Higher Polytechnic University College, Madrid, Spain

2012 – 2016

B.Sc. in Biomedical Engineering
Technical University of Madrid, Higher Technical School of Telecommunication Engineering and Center for Biomedical Technology, Madrid, Spain

Scholarships

2019 – 2021

Marie Curie Ph.D. Grant, Marie Skłodowska-Curie Actions (MSCA), European Union

Jun – Nov 2018

Swiss European Mobility Program (SEMP) Scholarship, Yverdon-les-Bains, Switzerland

2012 – 2013

Scholarship for Academic Excellence, Community of Madrid, Spain

Professional experience

Aug 2023 – present

Global Health Biostatistics/Bioinformatics
GlaxoSmithKline (GSK), Tres Cantos, Spain

2019 – 2022

Bioinformatician – Ph.D. Student / Marie Curie Early Stage Researcher of ITN "Building a Gut Microbiome Engineering Toolbox for In-Situ Therapeutic Treatments for Non-alcoholic Fatty Liver Disease (BestTreat)"
Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute (Leibniz-HKI), Group of System Biology and Bioinformatics (SBI), Jena, Germany

Sep – Nov 2020

Ph.D. secondment in industry
Clinical Microbiomics, Copenhagen, Denmark

Jun – Nov 2018

Bioinformatician master thesis student
Computational Intelligence for Computational Biology Group (CI4CB) at the School of Engineering and Management (HEIG-vd), Yverdon-les-Bains, Switzerland

2017

Data analyst intern
Philips Healthcare, department of Healthcare Informatics, Solutions & Services and Patient Care & Monitoring Systems, Madrid, Spain

Sep – Nov 2016

Biomedical engineer student intern
Unit of pediatrics endocrinology and diabetes, Biomedical Research Foundation of Ramón y Cajal Hospital, Madrid, Spain

Publications

Ni Y., Qian L., Siliceo S.L., Long X. *et al.*, (in press). **Resistant starch decreases intrahepatic triglycerides in patients with NAFLD via gut microbiome alterations**. *Cell Metabolism*.

Chen J., and Siliceo S.L. *et al.*, (2023). **Identification of robust and generalizable biomarkers for microbiome-based stratification in lifestyle interventions**. *Microbiome*, 11, 178.

Leung H., Long X. *et al.*, (2022). **Risk assessment with gut microbiome and metabolite markers in NAFLD development**. *Science Translational Medicine*, *14*(648).

Fournier B. *et al.*, (2022). **Toward the use of protists as bioindicators of multiple stresses in agricultural soils: A case study in vineyard ecosystems**. *Ecological Indicators*, *139*, 108955.

APPENDIX

Appendix 1. Form 2 of the own contribution to the manuscripts

FORM 2

Manuscript No. 1

Short reference Ni, Qian, Siliceo, and Long *et al.*, (in press), Cell Metabolism.

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 3, S2, S5	☒	Approximate contribution:	**90%**
Figure(s) # 2, 5	☒	Approximate contribution:	**70%**
Figure(s) # 1, S6	☒	Approximate contribution:	**50%**
Figure(s) # S7	☒	Approximate contribution:	**5%**
Figure(s) # 4, 6, S1, S3, S4	☒	**0%**	

FORM 2

Manuscript No. 2

Short reference Chen and Siliceo *et al.*, (2023), Microbiome.

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 4, S2	☒	Approximate contribution:	**90%**
Figure(s) # 3	☒	Approximate contribution:	**50%**
Figure(s) # 1, 2, S1	☒	Approximate contribution:	**10%**

FORM 2

Manuscript No. 3

Short reference Thielemann and Siliceo *et al*., in preparation.

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 1, 3, 4, S3, S4, S5	☒	**100%**
Figure(s) # 2, 5, S1, S2, S6, S7	☒	**0%**

FORM 2

Manuscript No. 4

Short reference Leung and Long *et al*., (2022), Science Translational Medicine.

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 2, 3, 4, S4	☒	Approximate contribution: **5%**
Figure(s) # 1, S1, S2, S3, S5, S6, S7	☒	**0%**