UNIVERSITA' DEGLI STUDI DI SASSARI
Corso di Dottorato in Scienze Mediche, Chirurgiche e Sperimentali
Curriculum Biologia e Genetica
38° Ciclo

Identification of microbiota components correlated with host lifestyle, molecular, biochemical, immunophenotypic measurements and genotype in a deeply phenotyped Sardinian cohort

Tutor
Prof. Francesco Cucca

Co-Tutor
Dott. Mauro Pala

Tesi di dottorato di:
Dott.ssa Maria Antonietta Diana

ANNO ACCADEMICO 2024/2025

Abstract

The human gut microbiome plays a crucial role in health and disease, yet the relative contributions of host genetics, environmental exposures, and lifestyle factors to its variation remain incompletely understood. In this thesis, we investigated host–microbiome interactions in a large, well-characterized Sardinian cohort (ProgeNIA, N ≈ 2,650) through integrated analyses of genome-wide genotypes, shotgun metagenomic data, and deep phenotyping.

Using shotgun metagenomic sequencing, we profiled taxonomic and functional features of the gut microbiota and assessed their associations with host traits. Correlation analyses identified age and sex as the strongest determinants of microbiome composition. Smoking emerged as a major lifestyle factor negatively affecting alpha diversity, with current smokers showing reduced microbial diversity compared to non-smokers, while former smokers exhibited levels comparable to never-smokers, suggesting partial recovery following smoking cessation. We also observed strong associations between white blood cell counts and reduced alpha diversity, as well as between wine consumption and increased beta diversity.

We performed genome-wide association studies (GWAS) of microbial taxa, metabolic pathways, and diversity metrics using linear mixed models accounting for family structure. These analyses identified multiple genome-wide significant loci for microbial taxa and pathways, whereas fewer associations were detected for diversity metrics. Notably, one smoking-

associated taxon mapped to a genomic locus previously implicated in nicotine dependence, highlighting potential links between host genetics, lifestyle, and microbiome composition.

To investigate shared genetic architecture with complex diseases, we conducted Bayesian colocalization analyses using publicly available GWAS summary statistics and identified strong evidence of shared causal variants between *Parabacteroides merdae* and coronary artery disease. In parallel, we developed COLSTATS, a scalable web-based platform for harmonizing and interrogating large collections of GWAS summary statistics, enabling researchers to efficiently perform reproducible Bayesian colocalization analyses and systematically explore shared genetic architectures across a wide range of complex traits and diseases.

Heritability analyses indicated that gut microbiome diversity is influenced by both genetic and environmental factors. While additive genetic effects explained a substantial proportion of variance, explicit modeling of cohabitation effects revealed that shared living environment contributes independently to microbiome similarity beyond genetic relatedness.

Additional analyses showed that microbiome genetic associations are highly sensitive to covariate modeling, reinforcing the context-dependent nature of host–microbiome genetic effects.

Overall, our results indicate that host genetic effects on the gut microbiome are modest and strongly modulated by lifestyle, environmental exposures, and shared living conditions, with important implications for clinical and public health research, as they emphasize the need to account for

environmental context when interpreting microbiome–disease associations and translating them into robust biomarkers or risk models.

Sommario

1. INTRODUCTION

1.1. The human microbiota

The human microbiota is the collection of microorganisms that live and interact with the human body. It is composed of bacteria, archaea, viruses, and eukaryotes that are distributed across various body sites. The microbiome refers to the genomic content of these organisms. When we talk about microbiota, the numbers are astounding; approximately one hundred trillion microbes live on and inside the human body. These microorganisms interact with their host in various ways, including commensalistic, mutualistic, and pathogenic relationship [1].

It has been widely demonstrated that the microbiota plays a crucial role in human health and disease through interactions and mechanisms, many of which have not yet been fully clarified [2]. Dysbiosis, an alteration in the composition of the microbiota, can lead to health issues and, in some cases, contribute to disease development. Conversely, a balanced microbiota is important for maintaining health [1]. Research on the microbiome has greatly increased over the past two decades, thanks to advances in sequencing and multi-omic technologies. Studying the microbiota is essential, as it performs many critical (though still not fully understood) functions within the human body and is thus implicated in both health and disease [3]. Microorganisms colonize the entire human body, but the highest density and diversity are found in the gut [4].

1.1.1. Gut microbiota

Most microbiome research has focused on the gut microbiota, as it contains the largest number of microbes and the greatest diversity of species compared to other body sites [3].

The most predominant phyla in the gut are *Bacteroidota* (formerly known as *Bacteroidetes*) and *Bacillota* (formerly known as *Firmicutes*), followed by *Pseudomonadota* (formerly known as *Proteobacteria*), *Fusobacteria*, *Tenericutes*, *Actinobacteria*, and *Verrucomicrobia*. It is well established that the gut microbiota plays a central role in human physiology. Dysbiosis into this community has been associated with various human diseases, including inflammatory bowel diseases, obesity, diabetes, and cardiovascular disease [4]. Several factors influence gut microbiota composition, including lifestyle, diet, antibiotic use, smoking, and living conditions [5]. While the gut microbiota does have a heritable component, environmental factors such as diet, medication use, and anthropometric traits, play a more dominant role in shaping its composition [6].

1.1.2. Host–gut microbiome interactions

The gut microbiome interacts constantly with its human host through multiple layered mechanisms. In health, this interplay supports barrier integrity, immune education, metabolic signaling, and brain–immune crosstalk. However, disruption at any level can contribute to disease.

The first line of defense in host–microbe interaction is the intestinal barrier, which prevents uncontrolled microbial ingress while permitting regulated molecular exchange. This barrier comprises a mucus layer, antimicrobial peptides, secretory IgA, and epithelial cells connected by tight junctions (claudins, occludins, zonula occludens) that regulate paracellular permeability [7], [8]. The mucus layer, secreted by goblet cells, serves as a dynamic buffer: it traps bacteria, limits their access to the epithelium, and provides a habitat for commensals. In germ-free animal models, mucus is thinner and goblet cell density is reduced, demonstrating that microbiota influence mucus production and barrier maintenance [9], [10].

Epithelial cells are not passive: they sense microbial and environmental signals via pattern-recognition receptors (e.g. TLRs, NODs) and mount adaptive responses, such as modulation of tight junction expression or antimicrobial peptide secretion [11]. Immune cells in the lamina propria also contribute: for example, intraepithelial lymphocytes and dendritic cells interact with epithelial cells to maintain barrier function [12]. At the tissue level, organized lymphoid aggregates such as Peyer’s patches and isolated lymphoid follicles serve as key sites for antigen sampling and initiation of mucosal immune responses. When barrier integrity is compromised through inflammation, infection, or dysbiosis, tight junction proteins and mucin production are altered, allowing microbial products (e.g., lipopolysaccharide, peptidoglycan) to translocate, trigger immune activation, and promote systemic inflammation. This phenomenon is implicated in disorders like IBD, obesity, and metabolic disease [13], [14].

Below the barrier, the gut-associated lymphoid tissue (GALT) and mucosal immune network mediate antigen sampling, tolerance induction, and immune surveillance. Microbial colonization early in life is essential to the maturation of GALT: germ-free animals show reduced Peyer's patches, fewer regulatory T cells, and impaired immune architecture [15]. Microbial signals, especially metabolites, condition antigen-presenting cells toward tolerogenic phenotypes, supporting differentiation of regulatory T cells and restraining inflammatory responses against commensals [16]. At the same time, immune outputs such as secretory IgA, antimicrobial peptides, and cytokine milieu shape microbial community composition and spatial organization in the mucus and lumen [17], [18].

A central axis of host–microbe communication runs through microbial metabolites. Among them, short-chain fatty acids (SCFAs: acetate, propionate, butyrate) are generated via fermentation of dietary fiber by gut bacteria. SCFAs support colonocyte energy metabolism, strengthen tight junctions, regulate immune cell populations (particularly regulatory T cells), and influence host metabolism via G-protein–coupled receptors and epigenetic modulation [19], [20]. Other microbial products such as secondary bile acids, tryptophan derivatives (e.g. indoles), and bacterial-derived vitamins interact with host receptors (e.g. AhR, FXR, PXR) to modulate epithelial, immune, and metabolic pathways [21], [22].

The gut–brain axis further extends the reach of microbial influence. Signals from the gut via microbial metabolites, cytokines, and neural (vagal) pathways can affect central nervous system function and behavior. Conversely, neural, hormonal, and autonomic outputs modulate gut

physiology, immune tone, and motility [23]. Microbial effects on microglial development, blood–brain barrier integrity, and neuroinflammation suggest that perturbations in gut ecology may influence neurodegenerative or psychiatric conditions [24], [25].

Maintaining homeostasis in this system requires a careful balance between immune activation and tolerance. Under physiological conditions, regulatory circuits, regulatory T cells, tolerogenic dendritic cells, IL-10–producing macrophages, restrain excessive inflammatory responses toward commensals [15]. If regulatory control fails, due to barrier disruption or microbiome shifts, low-grade chronic inflammation may ensue, contributing to metabolic disorders, cardiovascular disease, and autoimmune syndromes [26], [27].

When external factors such as diet change, antibiotics, infection, stress, aging, perturb the microbial ecosystem, dysbiosis may occur. Dysbiosis encompasses reduced microbial diversity, altered composition, and shifts in functional output, and often precedes barrier compromise and immune activation [26], [28]. In conditions like inflammatory bowel disease (IBD), barrier loss, immune dysregulation, and microbial shifts form a self-reinforcing loop, and transplantation of dysbiotic communities can recapitulate disease phenotypes in experimental models [29]. Emerging evidence also implicates dysbiosis and gut–brain axis disruption in neurodegenerative diseases, via amplified inflammation, metabolite dysregulation, and blood–brain barrier impairment [23], [25].

In summary, the barrier forms the structural gatekeeping layer; immune and epithelial cells mediate controlled sampling and tolerance; microbial

metabolites relay functional cues, and neural circuits integrate these signals into systemic regulation. Disturbance in any of these layers can propagate dysfunction across the system. Understanding these mechanistic threads is essential for translating microbiome research into preventive and therapeutic strategies.

1.1.3. Host genetics and the gut microbiome

While environmental exposures such as diet, antibiotics, lifestyle, and geography exert a dominant influence on the gut microbiome, a non-negligible role of host genetics has emerged from human and animal studies. Twin studies in humans have revealed greater microbiome similarity in monozygotic compared to dizygotic twins, suggesting heritable microbial traits [30], [31]. One of the most consistent findings is the high heritability of the bacterial family *Christensenellaceae*, which in some cohorts accounts for 30–60% of inter-individual variation in relative abundance [32]. This family has also been associated with lean body mass and metabolic health in multiple populations, supporting its functional relevance beyond mere taxonomic variation [30], [32]. Despite these compelling examples, the majority of microbial taxa in humans show little or no measurable heritability. Recent meta-analyses and reviews estimate that solo taxa with non-zero heritability constitute only ~3–13% of the microbiome, and the overall heritable component of microbiome variation may be very small ($h^2 \sim 0.02$) in many populations. Part of this low heritability may reflect methodological constraints: single time-point sampling, unaccounted environmental confounding, and limited taxonomic

resolution (especially in 16S-based studies) reduce power to detect genetic effects [33]. Moreover, human relatives often share environments (diet, household, lifestyle) which can inflate estimates of heritability unless carefully controlled [33], [34]. Genome-wide association and metagenome-wide studies have begun to map how host genetic variation influences microbial taxa and community-level features. For example, a meta-analysis found that genetic diversity affects microbial composition across multiple cohorts, though effect sizes for individual loci are typically modest and replication across studies remains challenging [35]. The general pattern emerging is one of polygenic architecture with small individual effects, meaning that host genetics contributes to microbiome variability but rarely dominates over environmental influences.

1.1.4. Linking the microbiome to human health via genetics

Establishing causal relationships between gut microbiome features and human disease remains a central challenge in microbiome research. While metagenome-wide association studies (MWAS) have identified numerous associations between microbial composition and host phenotypes, these analyses alone cannot determine whether microbiome changes are causal, consequential, or driven by shared environmental or lifestyle factors.

To address these limitations, genome-wide association studies (GWAS) have been applied to microbiome traits, treating microbial taxa and community features as quantitative phenotypes influenced by host genetic variation.

Early evidence for a genetic contribution to microbiome composition came from twin studies. In 2009, Turnbaugh et al. demonstrated that monozygotic twins share more similar gut microbiota than dizygotic twins, suggesting that certain microbial traits are heritable [36]. Subsequent studies expanded these observations by integrating host genotyping and microbiome profiling across larger cohorts.

One of the most consistently replicated host genetic associations with the gut microbiome involves the LCT locus, which encodes lactase, the enzyme responsible for lactose digestion in the small intestine. Variants in LCT determine lactase persistence and influence the amount of undigested lactose reaching the colon, thereby indirectly modulating the abundance of lactose-fermenting bacteria, particularly *Bifidobacterium*. This association was first identified in the Human Microbiome Project, which performed host genotype–microbiome GWAS across multiple body sites [37], and was subsequently replicated in the TwinsUK cohort [31], and in the LifeLines-DEEP cohort using shotgun metagenomics [38]. Large-scale meta-analyses, including the MiBioGen consortium, further confirmed the robustness of this association, although effect sizes varied substantially across populations. Notably, heterogeneity analyses revealed that the effect of LCT variants is modulated by age, ancestry, and dietary context, highlighting the importance of gene–diet and gene–environment interactions in shaping microbiome composition [35].

Additional host genetic loci repeatedly implicated in microbiome variation include FUT2 and ABO. The FUT2 gene encodes an α(1,2)-fucosyltransferase

that determines secretor status by controlling the expression of fucosylated glycans on mucosal surfaces. Functional variation in FUT2 alters the availability of host-derived carbohydrates that serve as substrates for gut microbes, thereby influencing microbial colonization patterns [35], [40]. Similarly, the ABO locus, which governs histo-blood group antigen expression and host glycosylation profiles, has been associated with differences in gut microbiome composition in several population-based studies [39], [35]. Together, these loci illustrate how host genetic variation can influence the microbiome indirectly, by modifying the intestinal environment rather than through direct immune-mediated mechanisms.

Large-scale efforts such as the MiBioGen consortium [35] and the Dutch Microbiome Project [41] have provided a comprehensive assessment of host genetic contributions to microbiome variation across tens of thousands of individuals. These studies consistently report that the overall heritable component of the gut microbiome is modest, with only a small fraction of microbial taxa exhibiting significant heritability. While specific loci such as LCT, FUT2, and ABO show reproducible effects, environmental factors (including diet, medication use, cohabitation, and lifestyle) account for a substantially larger proportion of inter-individual variability [30], [31], [40].

Altogether, these findings indicate that although host genetics can influence specific aspects of gut microbiome composition through well-defined biological mechanisms, its overall contribution is limited and highly context-dependent. The magnitude and detectability of genetic effects vary across populations, reflecting differences in allele frequencies, dietary habits, and

environmental exposures. This underscores the importance of conducting microbiome GWAS in diverse and genetically distinct populations, including isolated cohorts such as Sardinians, and of integrating genetic analyses with detailed environmental and lifestyle data when linking the microbiome to human health.

1.1.5. Social and household transmission (cohabitation effects)

In addition to host genetics and environment, social interactions and cohabitation constitute important drivers of microbiome similarity between individuals. Recent work in isolated villages has demonstrated that microbial strain sharing is structured along the social network: not only household members or kin, but also non-family social contacts share bacterial strains at detectable levels. They profiled strain-level microbiomes and face-to-face social networks in 1,787 adults across 18 isolated Honduran villages, showing that pairs of people with various social ties (family, friends, shared meals, time spent together) share more microbial strains than random pairs in the same village (even after controlling for diet, medication, and co-residence). They also observed that strain sharing extends to second-degree social connections, and that individuals who are more “socially central” have microbiomes more representative of the village-wide microbial milieu. In a longitudinal subset measured 2 years apart, connected individuals converged in strain composition more than

unconnected pairs, providing dynamic evidence that social contact influences microbiome flux over time [42].

These findings suggest that social networks create “microbial niches” within which transmission and homogenization of strains can occur beyond just familial settings. In relatively isolated or stable populations, the effect of cohabitation may confound associations unless explicitly modeled.

1.1.6. Microbiota development over early life and aging

The gut microbial community develops dynamically from birth and continues to evolve throughout life, being shaped by early-life exposures, diet, immune maturation, and environmental factors [43], [44]. Immediately after delivery, the infant gut begins to be colonized by microorganisms derived from maternal and environmental sources, including the mode of birth and the maternal microbiota, which play a key role in establishing the early microbial community. Mode of delivery, breastfeeding, and early environmental contacts profoundly influence the first microbial consortia that will later guide immune system maturation and metabolic programming. During the first months of life, facultative anaerobes dominate the gut ecosystem, but are progressively replaced by obligate anaerobes such as *Bifidobacterium* and *Bacteroides*, particularly in breastfed infants where human milk oligosaccharides act as selective substrates for *Bifidobacterium* growth [44].

This stage is characterized by rapid microbial succession and increasing functional specialization of the intestinal community.

The introduction of solid foods during weaning represents a major ecological transition: new substrates promote the expansion of taxa specialized in the degradation of complex carbohydrates and lipids, driving a marked increase in microbial diversity and a compositional shift toward a more adult-like microbiota, typically achieved by around two to three years of age [43], [44].

Across adulthood, the gut microbiota reaches a relatively stable and functionally resilient state. However, substantial inter-individual variability persists, influenced by long-term dietary patterns, lifestyle, host genetics, and health status [43], [45]. Despite its apparent stability, the adult microbiome remains highly adaptive: changes in diet composition, antibiotic exposure, or chronic diseases can still modulate community structure and function. Longitudinal and population-based studies have consistently shown that microbial diversity is generally higher and more stable in adults than in infants, although compositional plasticity persists throughout life [43].

In older adults, physiological changes such as immune senescence, dietary alterations, polypharmacy, and reduced gut motility are associated with distinct remodeling of the intestinal ecosystem. These alterations often include a decline in beneficial taxa such as *Bifidobacterium* and butyrate-producing *Lachnospiraceae*, coupled with an increased prevalence of potentially pro-inflammatory or opportunistic microorganisms, including *Enterobacteriacea*. Moreover, the inter-individual variability of the gut microbiota tends to increase with age, reflecting divergent aging trajectories influenced by diet, frailty, lifestyle, and underlying health

conditions. Distinct microbial profiles have been associated with healthy aging versus age-related inflammation and metabolic decline [44], [45]. Altogether, evidence from longitudinal and cross-sectional studies depicts the human gut microbiome as a dynamic ecosystem that co-evolves with its host. Its composition and functionality reflect the continuous interplay between biological development, environmental exposures, and host physiology from the formative years of life, through adult stability, to the complex restructuring that accompanies aging [43], [44], [45].

1.2. Brief history of microbiota research methods

1.2.1 Early Foundations of Microbiology and Microbiota Research

The development of the first microscopes, credited to Antonie van Leeuwenhoek around 1675, marked the beginning of microbiology by enabling the observation of microscopic life for the first time. Over the next two centuries, advances in microscopy allowed scientists such as Pasteur and Koch to identify microorganisms as causative agents of infectious diseases. Despite this pathogenic focus, early ideas of beneficial microbes began to emerge, and Pasteur himself recognized that bacteria were essential to life. By the early 20th century, the concept of microbial symbiosis gained traction, and microbes started to be acknowledged for their contributions to soil fertility and food production [46].

1.2.2 The Advent of Molecular Taxonomy: 16S rRNA Gene Sequencing

However, early microbiota research relied on culture-based techniques, which could capture less than 1% of microbial diversity, severely limiting ecological understanding. Things started to change in 1977, when Carl Woese and George Fox introduced the use of the 16s ribosomal RNA (16S rRNA) sequences for the identification and classification of microorganisms by analyzing their genetic material directly from the environment. This breakthrough opened the door to understanding microbial diversity far beyond what could be cultured in a lab [47]. The 16S rRNA gene, which encodes the 30S small subunit of prokaryotic ribosomes, contains highly conserved regions that serve as primer binding sites, along with nine hypervariable regions (V1-V9) that can be used to infer microbial phylogeny. These characteristics allow researchers to group sequences with >=97% similarity into Operational Taxonomic Units (OTUs), which serve as proxies for microbial species in ecological and diversity studies [46]. In the 1990s, the development of universal primers enabled the PCR amplification of hypervariable regions of the 16S rRNA gene, paving the way for large-scale microbial profiling through amplicon sequencing. With the advent of next-generation sequencing (NGS) technologies in the early 2000s, 16S rRNA amplicon sequencing became increasingly high-throughput, cost-effective, and accessible [48]. As a result, this approach has become the most widely used method in microbiome research, thanks to its relative simplicity, affordability, and effectiveness in profiling microbial communities.

1.2.3. Emergence of Metagenomics and Shotgun Sequencing Approaches

The concept of metagenomics, defined as the culture-independent analysis of genetic material recovered directly from environmental samples, emerged in the mid-1990s. While 16S rRNA sequencing provides valuable taxonomic information, it is limited to bacterial and archaeal communities and lacks functional resolution. While amplicon-based sequencing remains widely used, whole metagenome shotgun (WMGS) sequencing has been increasingly adopted since the mid-2000s. Unlike targeted 16S sequencing, shotgun metagenomics sequences all genetic material present in a sample, enabling both taxonomic profiling (at higher resolution, down to species and strain level) and functional characterization of microbial communities. Despite its advantages, shotgun metagenomics remains more expensive and computationally intensive than amplicon-based methods. It also requires higher DNA quality and quantity, and more advanced bioinformatic pipelines, such as MetaPhlAn for taxonomic profiling and HUMAnN for functional pathway reconstruction [7].

1.2.4. Expansion into Multi-Omics: Functional Characterization of Microbiomes

Beyond metagenomics, other “meta-omic” approaches have been developed to capture different functional layers of microbial communities. Metatranscriptomics analyzes gene expression within a microbial community by sequencing microbial RNA, providing insight into actively

expressed genes and identifying functional pathways under specific environmental or host-related conditions. Metaproteomics and metabolomics represent two complementary approaches that extend microbiome analysis beyond genetic potential, allowing insight into its actual functional activity. Metaproteomics focuses on the identification and quantification of proteins expressed by microbial communities, providing direct evidence of biological functions currently active in the system. Unlike metagenomics, which outlines the functional potential encoded in the genome, metaproteomics captures the real-time protein expression profile of the microbiota under specific conditions. Metabolomics, on the other hand, investigates the full set of small-molecule metabolites (typically <2000 Da) produced by the host and its resident microbiota. These metabolic products play a key role in modulating host physiology, including immune responses, nutrient absorption, mental health, and organ function. Despite its potential, interpreting metabolomic data remains complex due to the overlap between host- and microbe-derived metabolites and the intricate interactions among them. Together, metaproteomics and metabolomics contribute to a more dynamic and integrated view of microbial function and its impact on health and disease [49].

1.2.5. Long-Read Sequencing Technologies

More recently, the introduction of third-generation sequencing technologies, such as Pacific Biotechnologies (PacBio) and Oxford Nanopore Technologies (ONT), has allowed the generation of long reads, sometimes exceeding tens of kilobases. These technologies have improved genome assembly and

facilitated strain-level resolution of complex microbial communities. Long-read sequencing is particularly valuable for reconstructing full microbial genomes, resolving repetitive regions, and detecting structural variants, although it currently remains more error-prone and cost-intensive than short-read technologies [50], [51].

1.2.6. Landmark Projects in Microbiome Research

A major catalyst for the expansion of microbiome research was the launch of the Human Microbiome Project (HMP) in 2007 by the U.S. National Institutes of Health. The first phase of the project aimed to characterize the microbial communities inhabiting five major body sites in healthy individuals, using both 16S rRNA and shotgun metagenomics [52]. One of the main findings from HMP1 was that taxonomic composition alone was often a poor predictor of host phenotype, whereas microbial functional potential and strain-specific features showed stronger associations. The second phase, known as the Integrative Human Microbiome Project (iHMP), was launched to explore host–microbiome interactions at a deeper mechanistic level. iHMP employed a multi-omic longitudinal approach, integrating metagenomics, metatranscriptomics, metaproteomics, and metabolomics, along with host data on immune responses, metabolism, and clinical phenotypes. The program focused on conditions with known microbiome involvement, including inflammatory bowel disease (IBD), preterm birth (PTB), and prediabetes, providing not only multi-layered data but also protocols and biospecimens for future research. These efforts helped establish the value of

multi-omics in understanding the dynamic interplay between the human host and its microbiota [53].

2. AIMS OF THE STUDY

The gut microbiome plays a key role in maintaining host health and modulating disease risk through complex interactions with the immune, metabolic, and nervous systems. In recent years, multiple studies have suggested associations between microbiota composition and a broad range of host traits, including metabolic disorders, autoimmune diseases, and behavioral phenotypes. However, most of these studies have relied on small cohorts, 16S rRNA sequencing, and observational designs that are often affected by confounding factors and reverse causation. While several recent efforts have advanced the integration of host genomic data with shotgun metagenomics to map genetic determinants of microbiome composition, this field remains relatively young. Many questions remain open, particularly regarding the reproducibility of associations across populations, their functional interpretation, and their role in disease mechanisms. Given these limitations, the main aim of this work is to provide a comprehensive and statistically robust analysis of host–microbiome interactions by leveraging deep metagenomic and genomic data from the ProgeNIA cohort, a well-characterized population study from Sardinia. This cohort includes 2,650 individuals with fecal shotgun metagenomic sequencing and dense host genotype data imputed on the whole genome. The study design benefits from the unique features of the Sardinian population such as genetic homogeneity, family structure, and extensive phenotyping allowing us to control for familial relationships and reduce confounding noise.

The specific objectives of the project are:

1. To characterize the taxonomic and functional composition of the gut microbiome using high-resolution metagenomic profiling tools such as MetaPhlAn 4 and HUMAnN 3.6, obtaining relative abundances for thousands of microbial taxa and metabolic pathways.
2. To quantify host genetic effects on microbiome variation by performing Genome-Wide Association Studies (GWAS) on taxonomic levels, metabolic pathways, and diversity indices (alpha and beta diversity), using linear mixed models that incorporate kinship matrices to correct for family structure.
3. To test for coincident signals between host genetic loci associated with the microbiome and those reported in public disease GWAS repositories (e.g., GWAS Catalog), using LD-based clumping and colocalization analyses to assess shared genetic architecture with complex traits and diseases.
4. To explore the contribution of environmental and lifestyle factors, such as age, sex, smoking, alcohol intake, and medication use, by correlating microbial traits with host phenotypes from the ProgeNIA database and testing their association with microbiome diversity and specific taxa/pathways.
5. To investigate the effects of cohabitation and host genetic heritability on microbiome similarity between individuals, using kinship-based and cohabitation-based matrices and comparing their impact on microbiome composition.

Through this integrative and multi-level approach, this study aims to provide a comprehensive map of host–microbiome interactions, identifying genetic, environmental, and lifestyle factors that shape gut microbiome composition. The ultimate goal is to contribute to a better understanding of the biological mechanisms underlying complex diseases and to lay the foundation for future microbiome-based precision medicine strategies.

3. MATERIALS AND METHODS

3.1. ProgeNIA cohort

The SardiNIA project is a longitudinal study of 7,730 individuals from the general population (57% females and 43% males), ranging from 14 to 103 years, native of the central east coast of Sardinia, Italy and is deeply phenotyped and genetically profiled. In particular, the volunteers come from a particular area of the Sardinia island, specifically from four towns in the Lanusei Valley in the Ogliastra region of the Sardinian province of Nuoro where there is a strong isolation and so a very high genetic homogeneity [54]. The cohort has been widely studied in many research areas [55], [56], [57].

3.2. Sample collection and processing

We profiled the microbiota of the 2,800 individuals metagenome sequenced from the ProgeNIA cohort. All participants signed informed consent to study protocols approved by the Sardinian Regional Ethics Committee (protocol no. 2171/CE). Volunteers were asked to collect stools in the morning of the visit (or the day before), and to store them at home (in the refrigerator). At the clinic site, stools were frozen at -80 . Then, all at once, were aliquoted (0,25 grams)–avoiding thawing–in safe-lock tubes and frozen at -80 C within 10 minutes, until DNA extraction. Microbial DNA extraction was performed with Bead Beating Plus Column (RBB+C), through high-speed shaking (15 Hz for 10 min with Tissuelyser II, Qiagen) coupled with QIAamp PowerFecal

DNA Kit. About 2,800 samples have been extracted. Each sample was quantified with nanodrop and qubit and DNA integrity has been checked by agarose gel electrophoresis. The library will be made with the Illumina® DNA Prep Tagmentation, according to protocol Illumina. However, a longer time, not foreseen at the beginning of the project, was used to fine-tune the production of the DNA libraries. The protocols were developed under the guidance of Nicola Segata's group.

3.3. Sequencing and pre-processing of raw reads

Shotgun metagenomic sequencing was performed in collaboration with Nicola Segata's team and the NGS facility of the University of Trento, using an Illumina NovaSeq 6000 platform with S4 flow cells, generating an average of 8.68 Gb of 150 bp paired-end reads per sample. Sequenced reads were pre-processed using the pipeline available at https://github.com/SegataLab/preprocessing. First, low quality (quality score <20), fragmented (length <75 bp), and reads with more than 2 ambiguous nucleotides were removed with Trim Galore (v0.6.6) [58]. Contaminant (the bacteriophage phciX174 DNA Illumina spike-in) and host DNA (hg19 human genome) were mapped and removed with Bowtie2 (v2.3.4.3) [59] using the -sensitive-local parameter. The retained high-quality reads were organized into standard forward, reverse, and unpaired reads output files for each metagenome. A total of 7 runs were performed, each containing 384 samples. On average we obtained 58M reads/sample (29M paired-end reads/sample). Each read-pair was 151 + 151 nucleotides

long. We considered only individuals for which at least 10 millions of reads have been produced: only 13 samples were discarded for having less reads. In total 2,688 samples were used for downstream analysis. We considered as ‘‘not healthy’’ those individuals who reported cardiovascular diseases (including infarction, heart failure, and stroke), diabetes (both T1D and T2D), any type of cancer, high cholesterol, Crohn’s disease, peptic ulcer, ulcerative colitis, pancreatitis, and asthma.

3.4. Taxonomic and functional profiling

Read mapping-based profiles were generated using the bioBakery suite [60], and more specifically, SGB-level taxonomic profiles were estimated using MetaPhlAn 4 [61] using the vJan21_CHOCOPhlAnSGB_202103 database, while functional potential profiles were estimated with HUMAnN 3.6 [60]. 35,871 pathways were quantified and 6,503 taxonomic levels, encompassing 26 phyla, 160 classes, 200 orders, 260 forms, 1,035 genera, 2,057 species, and 2,761 taxa.

3.5. Alpha and beta diversity calculation

To derive quantitative measures of the overall composition and variability of the gut microbiota, we calculated alpha and beta diversity indices using species-level relative abundances obtained from MetaPhlAn 4 profiles.

Alpha diversity reflects the within-sample diversity, capturing how many species are present (*richness*) and how evenly their abundances are

distributed (*evenness*). In contrast, beta diversity measures the between-sample dissimilarity, describing how microbial communities differ in composition across individuals.

For alpha diversity, we computed four commonly used indices:

- Observed richness, representing the total number of detected species;
- Shannon index, which combines richness and evenness to quantify community complexity;
- Simpson's index, which emphasizes species dominance and the probability that two randomly selected individuals belong to the same species;
- Gini index, which quantifies inequality in species abundance distribution.

For beta diversity, we estimated several complementary distance metrics capturing distinct aspects of inter-individual microbial variation:

- Bray–Curtis dissimilarity, an abundance-based measure sensitive to shared and unique species;
- Jaccard index, which considers only presence–absence information;
- Aitchison distance, based on Euclidean distances after Centered Log-Ratio (CLR) transformation, suitable for compositional data.

All diversity metrics were computed based on MetaPhlAn 4 profiles generated from shotgun metagenomic data. Species names were

standardized according to MetaPhlAn conventions, and samples with zero total abundance or unknown taxa were excluded from the analysis.

Bray–Curtis, Aitchison, and Jaccard dissimilarities between samples were used to perform Principal Coordinates Analysis (PCoA) using capscale() function from the vegan R package [62], generating individual-level coordinates.

3.6. Correlation analysis

We performed several correlation analyses to investigate the relationship between the gut microbiome and host phenotypic traits in the ProgeNIA cohort (N=2,654). Three main analyses were conducted:
(i) correlation between microbial features (taxa, pathways, and alpha diversity) and a selected panel of 15 host traits/diseases;
(ii) correlation between microbial diversity measures (Shannon index for alpha diversity and Bray–Curtis-based PCoA1 for beta diversity) and all available traits in the ProgeNIA database;
(iii) correlation between a curated set of mucosal immunity-related variants and microbial taxa, pathways, and alpha diversity.

For the first analysis, we selected 15 phenotypes based on prior evidence of their relevance to the gut microbiome: age, sex, antidiabetic drug use, cancer, multiple sclerosis, diabetes, alcohol intake, smoking, proton pump inhibitor use, glycemia, BMI, cholesterol, HDL, triglycerides, and LDL (calculated as: LDL = total cholesterol – HDL – triglycerides/5). Microbial taxa (selected after prevalence and variability filtering) and

microbial functional pathways were analyzed as dependent variables in linear mixed-effects models using the lmekin function from the coxme R package [63]. A kinship matrix derived from pedigree information was included as a random effect to account for family structure. Age and sex were included as fixed covariates. P-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) correction. For the second analysis, we assessed the association between Shannon index (alpha diversity) or Bray–Curtis-based PCoA1 (beta diversity) and all quantitative and categorical traits in the ProgeNIA database. For continuous traits, the same model was applied to three transformations of the data (raw, log1p, and inverse-normal transformed ranks). For each transformation, Pearson and Spearman correlations and p-values were computed. Log-transformation is omitted for PCoA1 as principal coordinate values can be negative. For categorical traits, ANOVA and Kruskal–Wallis tests were used to compare microbial diversity across categories. Age and sex were included as covariates in all models.

For the third analysis, we tested associations between a set of mucosal immunity variants and microbial taxa, pathways, and alpha diversity. These associations were evaluated using linear models on inverse-normal transformed microbial data with age and sex as covariates. P-values were corrected for multiple testing.

3.7. GWAS analysis

We conducted genome-wide association studies (GWAS) to identify host genetic variants associated with microbial features, including taxonomic levels, microbial pathways, alpha diversity, and beta diversity.

Before the analysis, we prepared the host genomic dataset, restricting the analysis to autosomal chromosomes. Variants showing strong deviation from Hardy–Weinberg equilibrium ($p < 1 \times 10^{-6}$) or low imputation quality were excluded. We retained variants with imputation $R^2 > 0.3$ for common variants (MAF > 0.01) and $R^2 > 0.6$ for rare variants (MAF < 0.01).

Each microbial feature was treated as a quantitative trait and normalized by inverse normal transformation before association testing. We handled zero-inflated microbial abundances treating zeros as missing values before applying an inverse normal transformation to the non-zero abundances. To account for sex, age and age2, they were regressed out from the microbial features by using a linear model with a custom R script. The obtained residuals were associated to the genome with GEMMA software (version 0.98.1) [64]. which fits linear mixed models with a kinship matrix – computed with the same GEMMA software - to correct for population structure and relatedness (-lmm 1 GEMMA command). Associations with genetic variants with minor allele frequency (MAF) < 0.05 were excluded with custom scripts. After performing the association analysis, we generated a Manhattan plot using the R package qqman.

Because genome-wide association analyses involve testing millions of correlated variants in linkage disequilibrium (LD), it is necessary to group

those variants that represent the same underlying association into a single representative, or lead, variant. This process, known as LD-based clumping, was carried out using PLINK v1.90b6.20 [61]. Variants were grouped within a 10 Mb window (--clump-kb 10000), applying an r^2 threshold of 0.001 (--clump-r2 0.001) to ensure that only weakly correlated variants were retained as independent. Variants with $p < 5 \times 10^{-6}$ were defined as potential lead SNPs (--clump-p1), while nearby variants with $p < 1 \times 10^{-4}$ (--clump-p2) were assigned to the same clump if in LD with the lead SNP. This procedure generated a set of independent lead variants, each representing a distinct LD-independent genomic locus.

Since the analysis included thousands of metagenomic traits, we further grouped independent loci identified across different traits. Even within the same LD block, the top variant can vary between traits due to differences in association strength. To account for this, a second-stage LD-based clumping was performed using the representative variants from all traits. This additional step produced a comprehensive list of independent genomic loci, each represented by the variant showing the strongest overall association across all tested traits.

3.8. Colocalization analysis with COLOC software

We implemented a colocalization analysis pipeline using the coloc R package [65] to evaluate whether loci identified in microbiome GWAS overlap with loci associated with complex traits or diseases. In the initial phase of the project, we conducted a GWAS on 2,650 individuals using a

broader set of 959 taxa after quality filtering. Independent loci were defined using a two-step LD-based clumping procedure.

We then intersected microbiome-associated loci with sentinel variants reported in the GWAS Catalog, considering variants in linkage disequilibrium (LD) ($r^2 > 0.6$) using the tool LinDA (http://linda.irgb.cnr.it). For each coincident locus, we retrieved publicly available summary statistics for the corresponding complex traits and diseases and performed colocalization using SuSiE-based coloc. Posterior probabilities were calculated for each hypothesis, with H4 indicating a shared causal variant. Only colocalizations with H4 > 0.8 were interpreted as strong evidence for shared causality.

3.9. Implementation of COLSTATS platform

For all studies integrated into COLSTATS, we sought to obtain, whenever available, summary statistics with the following fields: chromosome (chr), position (pos), variant identifier, allele information (effect and other allele), effect size (beta), standard error (SE), p-value, allele frequency (AF), and effective sample size (N). The availability of AF and N was considered crucial, since for quantitative traits it is not always appropriate to assume sdY = 1 in coloc; in such cases, sdY can be estimated from N and the minor allele frequency (MAF). All datasets were harmonized to the GRCh38/hg38 genome build: when summary statistics were provided in other builds, genomic coordinates were converted using LiftOver. The following sections describe the methods applied to each dataset.

We downloaded summary statistics for the following datasets:

- GWAS Catalog – NHGRI-EBI Catalog [66] - from http://ftp.ebi.ac.uk. Metadata were obtained from the studies metadata file (v1.0.2.1) and from per-study YAML files. In total, we downloaded 80,295 harmonized summary statistics files (~24 Tb, human genome build hg38.A total of 19,210 GWAS summary statistics presented the above listed criteria.
- UK Biobank - UKBB [67]. We retrieved UK Biobank metadata by downloading the phenotype flat file from https://pan.ukbb.broadinstitute.org/downloads, for a total of 216,113 GWAS summary statistics.
- Orrù et al. 2020 [68] – This dataset includes summary statistics of immunophenotypes (traits obtained with citofluorimetric analysis including cell counts, parental percentages and surface protein expression), corresponding to 731 GWAS summary statistics. Summary statistics of the study were altready present in the GWAS Catalog but the alleles of the INDEL were not explicited: they were coded as R/D or R/I. From the SardiNIA imputed panel we retrieved the regular name of the indels alleles which we substituted to the summary statistics.
- GTEx eQTLs [69] – We downloaded the summary statistics from (https://gtexportal.org/home/). This dataset includes eQTLs for 44 major human tissues in human genome build 37 including: Adipose Subcutaneous, Adipose Visceral Omentum, Adrenal Gland, Artery Aorta, Artery Coronary, Artery Tibial, Brain Anterior cingulate cortex

BA24, Brain Caudate basal ganglia, Brain Cerebellar Hemisphere, Brain Cerebellum, Brain Cortex, Brain Frontal Cortex BA9, Brain Hippocampus, Brain Hypothalamus, Brain Nucleus accumbens basal ganglia, Brain Putamen basal ganglia, Breast Mammary Tissue, Cells EBV-transformed lymphocytes, Cells Transformed fibroblasts, Colon Sigmoid, Colon Transverse, Esophagus Gastroesophageal Junction, Esophagus Mucosa, Esophagus Muscularis, Heart Atrial Appendage, Heart Left Ventricle, Liver, Lung, Muscle Skeletal, Nerve Tibial, Ovary, Pancreas, Pituitary, Prostate, Skin Not Sun Exposed Suprapubic, Skin Sun Exposed Lower leg, Small Intestine Terminal Ileum, Spleen, Stomach, Testis, Thyroid, Uterus, Vagina, Whole Blood. These data corresponded to 1,086,146 cis-eQTLs summary statistics.

- BLUEPRINT eQTLs [70] – We downloaded the summary statistics from http://blueprint-dev.bioinfo.cnio.es/WP10/qtls. This dataset includes eQTLs for three major human immune cell types (CD14+monocytes, CD16+ neutrophils, and naive CD4+ T cells) from up to 194 individuals (194, 192 and 171 individuals respectively), corresponding to a total of 48,675 cis-eQTLs summary statistics.
- eQTLGen phase 1 [71] – This dataset includes cis- and trans eQTLs from blood-derived expression from 31,684 individuals through the eQTLGen Consortium, corresponding to a total of 39,192 summary statistics (19,250 for cis and 19,942 for trans eQTLs). We downloaded the summary statistics from https://eqtlgen.org/phase1.html.

- Pala et al. 2017 [57] – This dataset includes cis-eQTLs from leukocytes of 606 individuals of the SardiNIA Cohort. We downloaded the summary statistics from https://eqtlsdownload.irgb.cnr.it (full summary statistics with conditional analysis – i.e. primary and after stepwise forward regression) corresponding to a total of 21,183 cis-eQTLs summary statistics.
- Ota et al., 2021 [72]. This dataset consists of 28 distinct immune cell subsets (CD16p Mono, CL Mono, CM CD8, DN B, EM CD8, Fr III T, Fr II eTreg, Fr I nTreg, Int Mono, LDG, Mem CD4, Mem CD8, NC Mono, NK, Naive B, Naive CD4, Naive CD8, Neu, Plasmablast, SM B, TEMRA CD8, Tfh, Th17, Th1, Th2, USM B, mDC, pDC) from 337 patients diagnosed with 10 categories of immune-mediated diseases (systemic lupus erythematosus (SLE), idiopathic inflammatory myopathy (IIM), systemic sclerosis (SSc), mixed connective tissue disease (MCTD), Sjo ¨ gren's syndrome (SjS), rheumatoid arthritis (RA), Behcet's disease (BD), adult-onset Still's disease (AOSD), ANCA-associated vasculitis (AAV), or Ta-kayasu arteritis (TAK)) and 79 healthy volunteers. This datases overall consists of 440,956 cis-eQTLs summary statistics. Summary statistics have been downloaded from NBDC Human Database (https://ddbj.nig.ac.jp).

Summary statistics were harmonized using gwas2vcf (docker image mrcieu/gwas2vcf) with Homo_sapiens.GRCh38.dna.toplevel.fa (from Ensembl) as the reference genome, generating harmonized VCF-formatted results. When original summary statistics were present in other genome

builts, were lifted over to hg38 using LiftoverVcf from GATK v4.1.5.0. We retained only standard chromosomes (1–22, X, Y) VCF files were indexed with tabix. Variant identifiers were coded as chr_pos_ref_alt (chromosome, position, reference allele, alternative allele).

The COLSTATS platform was developed as an R/Shiny web application using a modular client–server architecture. The frontend was built with shinythemes, custom CSS, and JavaScript to guide users through a clear three-step workflow. In Steps 1 and 2, users select traits or diseases of interest via a searchable input field (selectizeInput). After selection, a metadata table is displayed summarizing study details, including identifiers (COLSTATS ID, study ID, PubMed ID), population, consortium, study design (case–control or quantitative), and sample size (with separate counts for cases and controls). This metadata-driven interface allows users to filter and refine their study selection according to relevant criteria. In Step 3, users set the colocalization parameters, such as the genomic region (by coordinates or by gene, based on GENCODE v41), zoom level, and prior probabilities. Once parameters are defined, the colocalization analysis is launched on the server side, and results are returned as interactive tables showing posterior probabilities (PP.H0–H4), top associated variants with effect sizes and p-values, and sensitivity analyses. On the server side, datasets selected in Steps 1 and 2 are passed to the coloc.abf function to compute posterior probabilities for the five hypotheses (H0–H4). For quantitative traits, sdY is automatically set to 1 when applicable or estimated from sample size and allele frequency if needed. The application detects from metadata whether

a trait is quantitative or case–control. COLSTATS integrates several R packages, including: coloc for colocalization analysis, VariantAnnotation and GenomicRanges for handling VCF data, gwasglue for dataset harmonization, shinyjs for interactive features, and ragg for device-independent graphics rendering. Comprehensive error handling was implemented to detect invalid genomic region formats, missing files, and inconsistencies during data harmonization, ensuring robustness and reproducibility of the analyses.

3.10. Additional covariate-adjusted analyses

To further investigate the potential contribution of lifestyle and clinical factors to host–microbiome genetic associations, we performed a set of exploratory analyses including an extended set of covariates.
All primary analyses were initially adjusted for Age, Age², and Sex, which were consistently included in all baseline models. To explore the effect of additional factors potentially influencing gut microbiome composition, we repeated selected analyses by adding alcohol intake, smoking status, diabetes, proton pump inhibitor (PPI) use, and body mass index (BMI) as covariates.

Specifically, we reran GEMMA for the *taxa-level GWAS* using the same kinship matrix and phenotype normalization procedures adopted in the primary analyses, but including the extended covariate set in the model. In parallel, we performed correlation analyses between mucosal variants and metagenomic features (both taxa and pathways) using the same

additional covariates. Finally, we reran the alpha diversity GWAS including these variables to evaluate whether the genetic associations identified under the baseline model remained stable after adjusting for these potential confounders.

3.11. Heritability and cohabitation analysis

To estimate the heritability of microbiome features, we used linear mixed models implemented in SOLAR (Sequential Oligogenic Linkage Analysis Routines) [73], which partition phenotypic variance based on pedigree-derived kinship. The analysis focused on the Shannon alpha diversity index, modeled under the assumption of multivariate normality. Variance components were estimated using the polygenic function, with age, age squared (Age^2), and sex included as fixed-effect covariates. To further explore the sources of phenotypic variance beyond additive genetic effects, we extended the model to include a household random effect, representing shared environmental exposures due to cohabitation. This effect was modeled using a binary matrix derived from household information, distinguishing cohabiting from non-cohabiting pairs. While 2,613 individuals had available metagenomic and pedigree data for the baseline heritability analysis, a subset of 2,019 individuals with complete cohabitation information was used for the extended model including the household effect. These layered models enabled us to dissect the contribution of both genetic and environmental sources of variation in shaping gut microbiome diversity.

4. RESULTS

Here we report the summary of the main findings of the metagenomic analyses conducted within the ProgeNIA cohort, integrating information from host genetics, microbiome composition, and environmental and lifestyle variables.

The results are organized into two main sections: (1) correlation analyses between metagenomic features and phenotypic traits from the ProgeNIA database, and (2) analyses integrating host genomic data, including genome-wide association studies (GWAS), post-GWAS investigations, and heritability estimates.

Part of the results presented in this thesis were selected as platform presentation at the European Society of Human Genetics (ESHG) meeting in 2024.

4.1. Correlation of metagenome composition with ProgeNIA traits, diseases and lifestyle

In this section, we investigated the relationships between gut metagenome composition and a wide range of phenotypic traits available in the ProgeNIA database, including quantitative clinical variables, lifestyle factors, and disease diagnoses.

The analyses were performed at multiple levels of microbial characterization, encompassing taxonomic and functional features (taxa and pathways), as well as overall diversity metrics (alpha and beta

diversity).

Among the examined traits, smoking behavior was further explored in dedicated analyses, given its known influence on microbial composition and its availability in detailed phenotypic form within the cohort.

4.1.1. ProgeNIA database traits vs taxa, pathways

We conducted correlation analyses between microbial taxa and a variety of phenotypic traits and disease-related variables. The results revealed that hundreds of taxa are significantly associated with several traits at a 5% false discovery rate (FDR) threshold. Among all traits analyzed, age and sex displayed the largest number of associations, each correlating with over 400 taxa. These findings highlight the widespread and trait-specific nature of host–microbiota interactions. The most frequently associated microbial groups belong to the *Bacillota* (formerly known as *Firmicutes*) and *Bacteroidota* (formerly known as *Bacteroidetes*) phyla, both of which are well-known dominant components of the human gut microbiota.

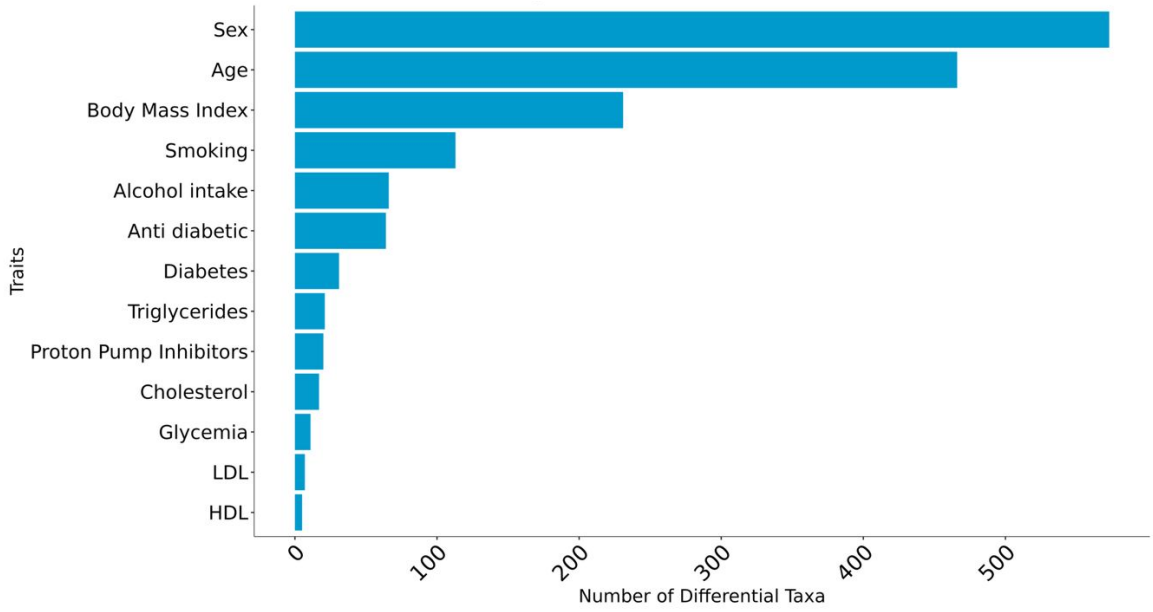

Figure 1. Number of microbial taxa significantly associated (FDR < 0.05) with each phenotypic trait or disease-related variable. Age and sex exhibit the highest number of associations, with more than 400 correlated taxa each.

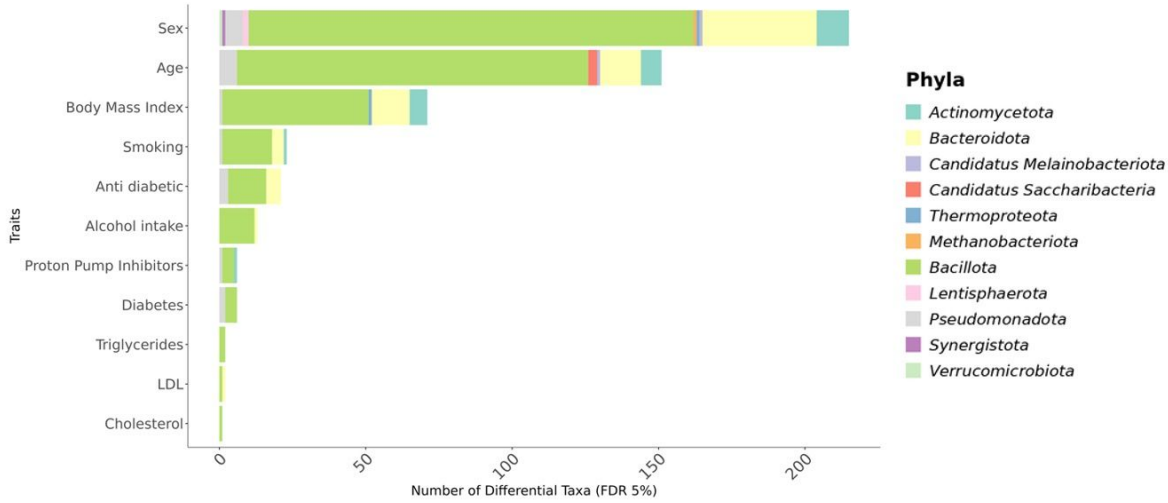

Figure 2. Number of non-redundant taxonomic levels associated with each trait and phylum-level composition of the correlated taxa. *Bacillota* (formerly known as *Firmicutes*) (green) and *Bacteroidota* (formerly known as *Bacteroidetes*) (yellow) are the most represented phyla across traits.

To better visualize the relationship between microbial taxa and specific traits, we employed Graphlan plots [74], which illustrate changes in taxonomic abundance and their distribution across the taxonomic tree.

Graphlan visualizations allow us to observe whether the abundance of taxonomic levels increases or decreases with increasing values of a given trait or between case and control groups. Taxa that show differential abundance with respect to a given trait can be identified based on their direction of change, helping to interpret overall patterns of up- or down-regulation within the microbiome. These representations also demonstrate the depth of taxonomic resolution achievable through shotgun metagenomic sequencing, which enables the identification of microbial signatures across multiple levels of classification.

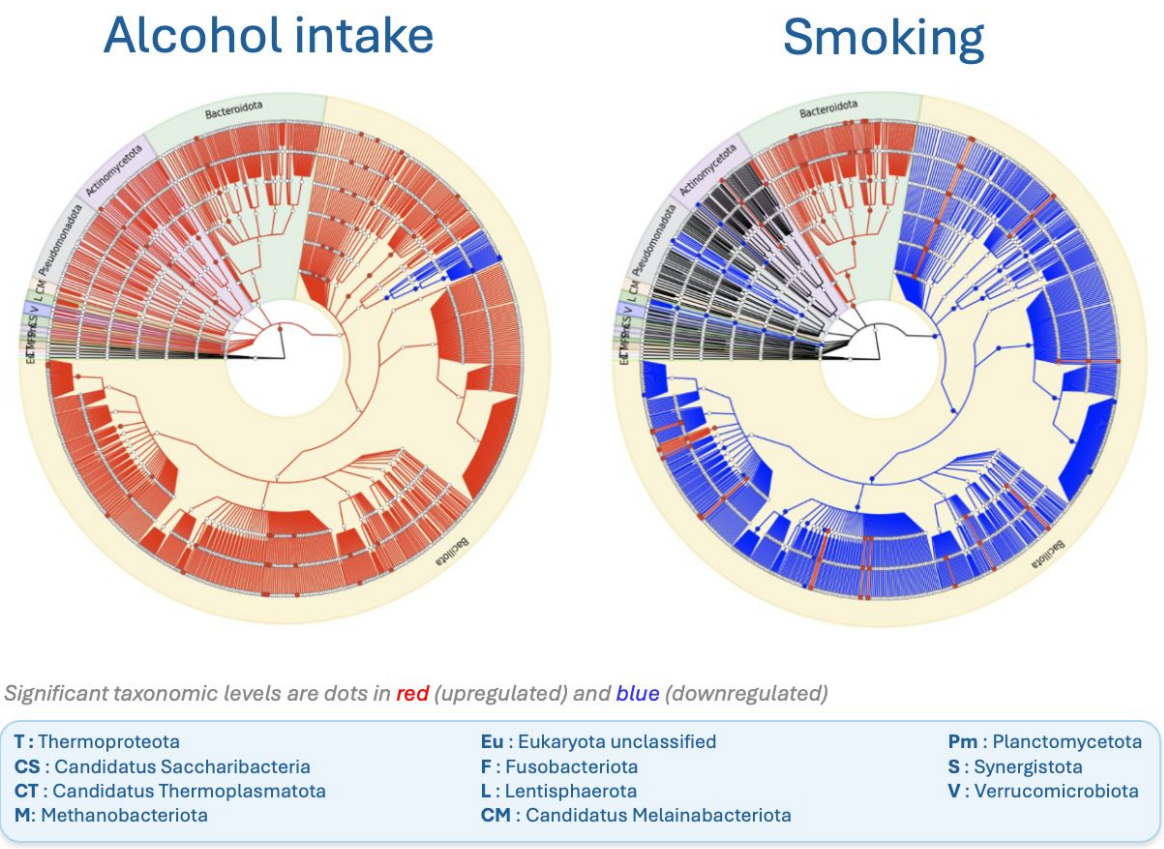

Figure 3. Graphlan plots showing trait-associated changes in taxonomic abundance across the microbial tree. Each plot represents associations with a specific trait: alcohol consumption (left) and smoking status (right). Significant taxa are marked with colored nodes, red indicating upregulation and blue indicating downregulation. Alcohol

consumption is associated with a general increase in microbial abundance, except for the *Lactobacillales* order (*Bacillota*, formerly known as *Firmicutes*), which is downregulated. In contrast, smoking is associated with widespread downregulation, except for several taxa from the *Bacteroidota* (formerly known as *Bacteroidetes*) phylum.

Correlations were also performed for microbial pathways quantified as described in the “Taxonomic and functional profiling” methods session. The pathways represent aggregated functional profiles describing the metabolic potential of the gut microbiome. Pathway abundances were obtained from HUMAnN 3.6, which quantifies the relative abundance of metabolic pathways annotated according to the MetaCyc database. Examples of pathways commonly detected in human gut microbiomes include those involved in short-chain fatty acid (SCFA) biosynthesis (e.g., butyrate or propionate production), vitamin metabolism (e.g., folate or biotin biosynthesis), and bile acid transformation.
Analyzing microbial pathway abundances complements taxonomic correlations by capturing functional aspects of the microbiome that may vary even when taxonomic composition remains stable. This functional perspective enables a more mechanistic interpretation of host–microbiome associations. As observed for taxonomic features, Age and Sex display the highest number of significant associations, confirming their major influence on the gut microbiome functional landscape. Additionally, smoking shows a considerable number of associations with microbial pathways, suggesting that it may also impact the functional potential of the gut microbiota.

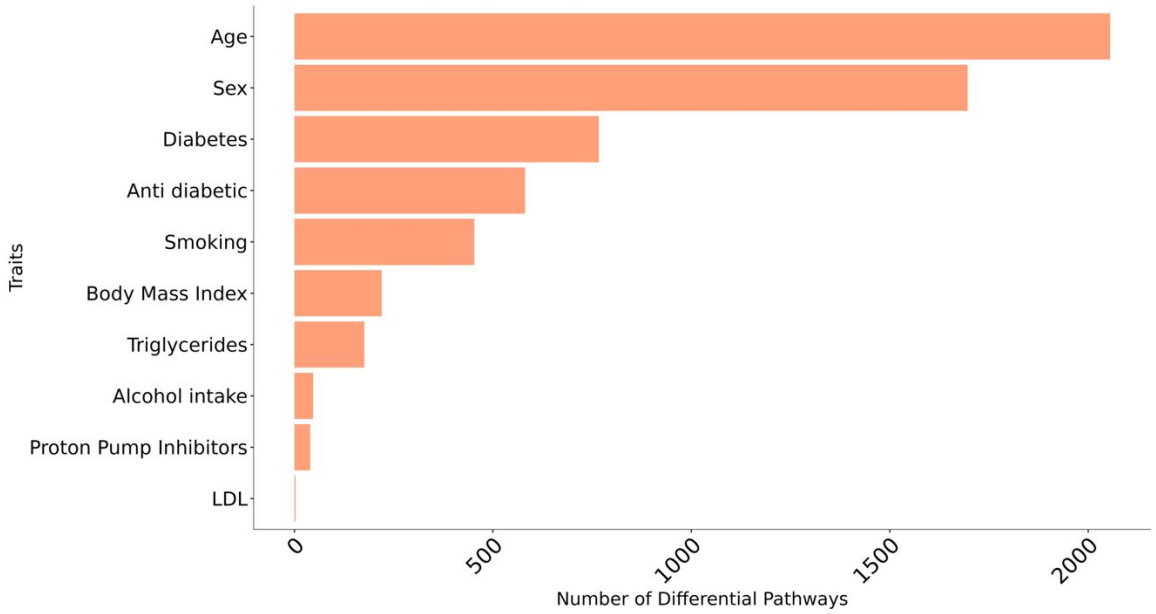

Figure 4. Number of microbial pathways significantly associated (FDR < 0.05) with each trait or condition. Age and sex show the highest number of associations, consistent with their strong influence on gut microbiome composition and function. Smoking also displays a substantial number of associations, suggesting a potential impact on microbial functional pathways.

4.1.2. All ProgeNIA database vs alpha diversity and beta diversity

To explore the relationship between host phenotypic traits and microbial diversity, we tested associations between all the variables available in the SardiNIA database and two diversity measures: Shannon index for alpha diversity and the first unconstrained PCoA axis based on Bray–Curtis distances for beta diversity (see Methods).

For this analysis we used the alpha diversity Shannon index, which reflects both the richness and evenness of species within each sample, defined as

where is the relative abundance of species i and is the total number of species observed.

Beta diversity was assessed through the Bray–Curtis dissimilarity, which quantifies compositional differences between pairs of samples based on species abundances, expressed as

where is the sum of shared abundances between samples i and j, and and are the total abundances in each sample.

Principal Coordinates Analysis (PCoA) was then performed on the Bray–Curtis distance matrix, and the first unconstrained PCoA axis (PCoA1) was used as a summary measure of inter-individual variation in microbiome composition for these correlations.

For continuous variables we applied the statistical model to three different versions of the data: the raw values, log-transformed values (omitted for PCoA1 as principal coordinate values can be negative), and values transformed using the inverse normal method. Results were then ranked based on the strongest association. For categorical variables, both ANOVA and Kruskal–Wallis tests were performed, and the top associations were visualized with violin plots. Among continuous variables, alpha diversity showed the strongest negative correlation with white blood cell count, particularly with neutrophil count, which remained significant across all models (Pearson $r = -0.13$, $p = 1.3 \times 10^{-10}$ for total WBC; $r = -0.12$, $p = 6.9 \times 10^{-9}$ for neutrophils) suggesting an immune cell-mediated regulation of

bacterial community. Another significant inverse association was observed with lifetime exposure to cigarette smoking (Pearson $r = –0.11$, $p = 1.3 \times 10^{-8}$), suggesting that smoking may lead to a long-term reduction in microbial diversity.

Regarding beta diversity, the strongest associations emerged with alcohol consumption, particularly wine intake, which showed a positive correlation with the first Bray–Curtis PCoA axis (Pearson $r = 0.14$, $p = 4.4 \times 10^{-7}$). This suggests that regular wine consumption may be linked to compositional shifts in the gut microbiota. Among categorical variables, alpha diversity was significantly lower in individuals who reported current smoking (Kruskal–Wallis $p = 6.2 \times 10^{-12}$). In addition, reduced diversity was observed in participants with a history of myocardial infarction, cancer, and those taking antidiabetic medications such as biguanides, all showing significant differences in Shannon index values ($p < 0.01$). For beta diversity, the most significant associations were found with sex, which had extremely strong p-values (Kruskal–Wallis $p < 10^{-17}$) with a lower beta diversity is observed in females. The lower beta diversity observed among women suggests a more homogeneous and potentially more stable gut microbiota composition compared to men. This pattern may reflect the combined influence of sex hormones, which modulate mucosal immunity and gut physiology, and more consistent lifestyle factors such as diet or health-related behaviors. Conversely, the higher inter-individual variability observed in men could result from greater heterogeneity in environmental exposures, including dietary habits, smoking, and alcohol consumption. Other significant

variables included polycystic ovary syndrome, myocardial infarction, and coffee consumption, all of which were associated with distinct microbial community structures ($p < 0.001$).

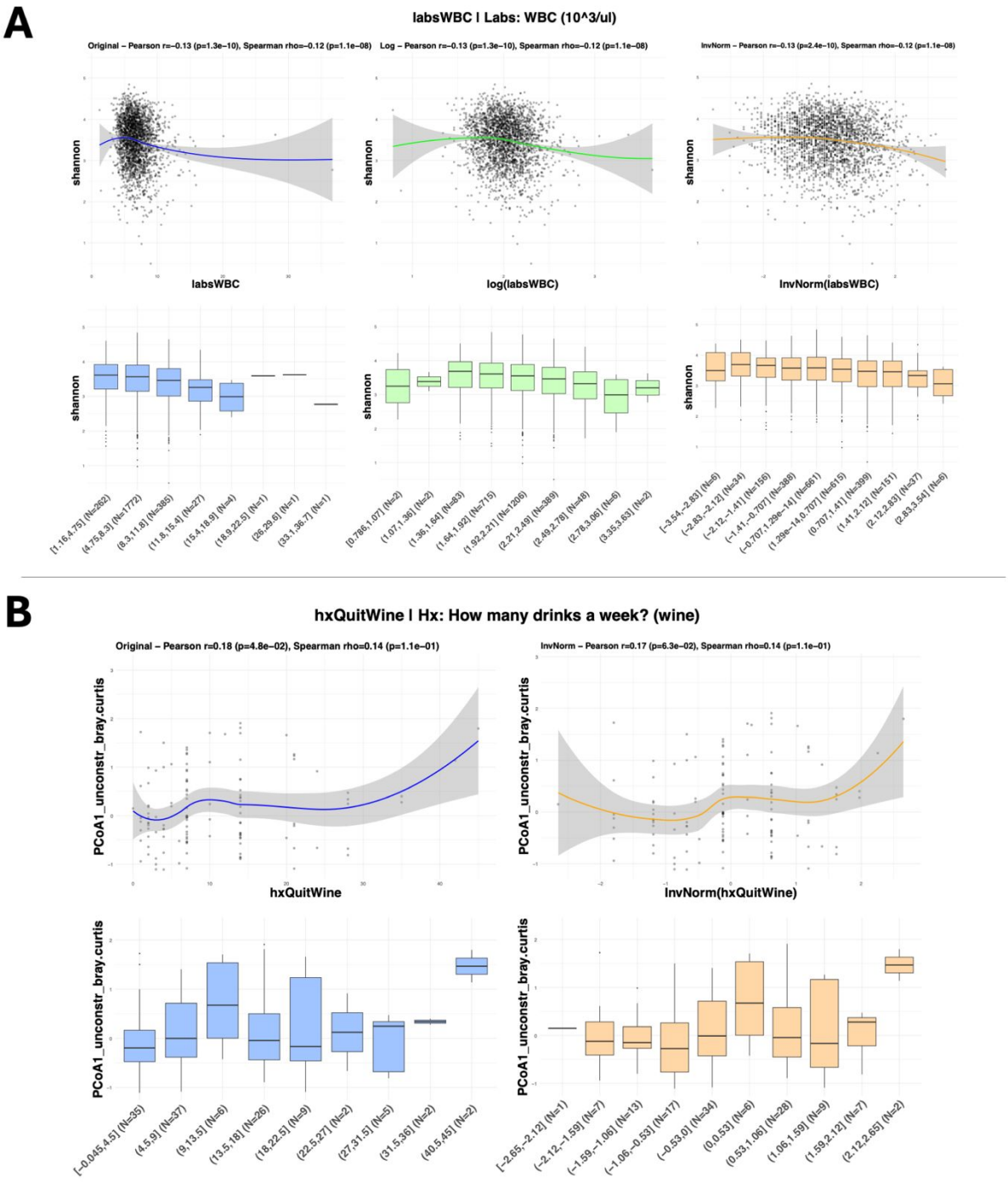

Figure 5. **Strongest associations between microbial diversity and continuous variables**. **A)** Alpha diversity (Shannon index) vs. white blood cell (WBC) count, and **B)**

beta diversity (unconstrained PCoA1 Bray-Curtis) vs. wine consumption (drinks per week). The top row shows scatterplots with LOESS regression curves and 95% confidence intervals for different data transformations: raw values (left), log-transformed (center), and inverse normal transformed (right). Pearson and Spearman correlation coefficients with corresponding p-values are reported above each panel. The bottom row displays the same associations using boxplots across binned values of the independent variable. Log-transformation is omitted for PCoA1 (panel B) as principal coordinate values can be negative.

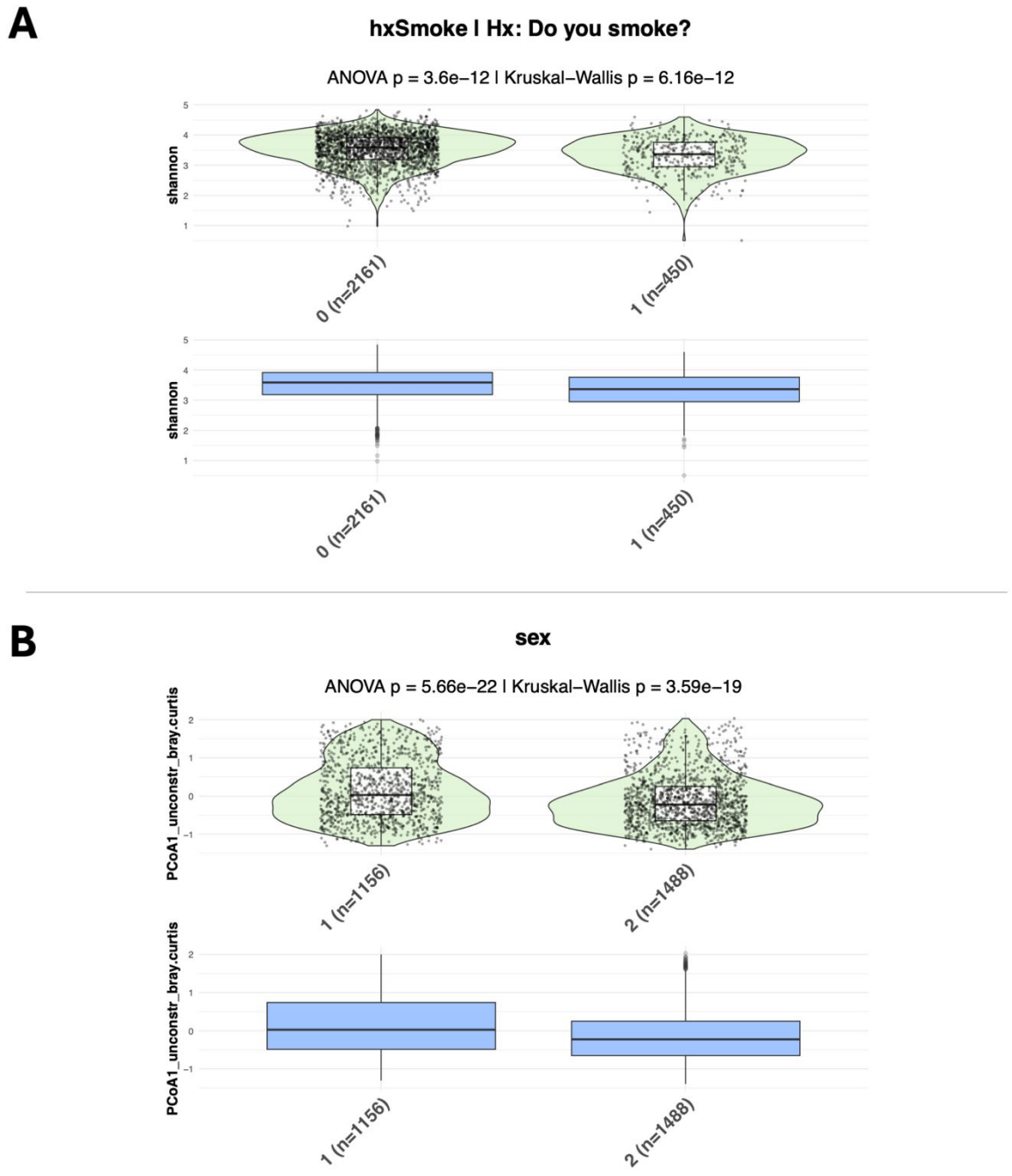

Figure 6. Strongest associations between microbial diversity and categorical variables. For both panels, violin plots (top) show the distribution and density of diversity values, while boxplots (bottom) summarize central tendency and variability. Groups are encoded as 0, 1 for Smoking (1=yes, 0=no) and 1, 2 for Sex: 1=men, 2=women), and statistical significance was assessed using both ANOVA and Kruskal–Wallis tests. **A)** Alpha diversity (Shannon index) in relation to current smoking status. A significant reduction in microbial diversity is observed in current smokers ($p < 10^{-11}$). **B)** Beta diversity (unconstrained PCoA1 Bray–Curtis) in relation to sex. A lower beta diversity is observed in females ($p < 10^{-19}$).

4.1.3. Further analysis of the results: metagenome correlation with smoking

Among the host traits analyzed, smoking showed the strongest negative association with alpha diversity, here represented by the Shannon index, indicating that current smokers tend to have reduced microbial diversity in their gut. This reduction reflects a loss of community evenness, suggesting that smoking may promote the dominance of specific taxa. In line with this, the species *Clostridium phoceensis* displayed one of the strongest positive correlations with smoking among individual taxa, pointing to a potential smoking-associated enrichment of this bacterium.

Interestingly, individuals who quit smoking exhibited alpha diversity levels comparable to those who never smoked, suggesting a potential recovery of microbial diversity following smoking cessation. Overall, these findings indicate that smoking is associated with a measurable alteration in gut microbial structure, characterized by decreased diversity and selective expansion of specific taxa.

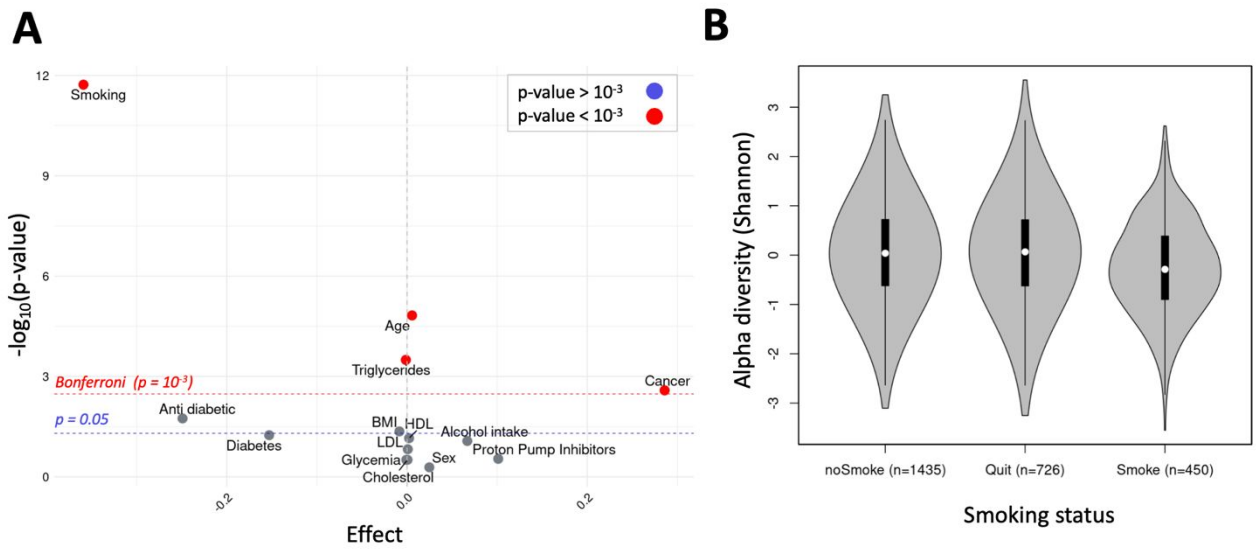

Figure 7. Plots showing the correlation of Shannon alpha diversity metric with ProgeNIA features. A) Among the 15 traits/diseases, smoking shows the most significant associations with alpha-diversity (Shannon index shown). **B)** We identify a decreased alpha diversity in smokers compared to no smokers (p-value : 1.594e-09). Volunteers that quit smoking (former smokers) showed an alpha diversity similar to no smokers.

4.2. Correlation of metagenome composition with host genome

This section describes the integration of host genomic data with metagenomic features to investigate host–microbiome genetic relationships.

We first evaluated associations between variants in mucosal-related genes and microbial features, followed by genome-wide association studies (GWAS) for taxa, pathways, alpha and beta diversity, using mixed linear models implemented in GEMMA [64] and accounting for gender, age, age^2 and relatedness through a kinship matrix (See Methods). Post-GWAS analyses included downstream functional exploration with

colocalization analyses with complex traits and diseases, to assess potential causal relationships.

Finally, we analyzed the contribution of shared living environments and familial relationships to microbial similarity through cohabitation and heritability analyses.

4.2.1. Correlation with mucosal immunity variants associated with MS

To investigate the potential influence of genetic variations involved in mucosal immunity on gut microbiome composition, we selected a set of nine variants. Eight of these were chosen from a published consortium dataset [75], based on their known effect on mucosal immunity and on their location in autosomal, non-HLA regions, including IL7R, encoding the interleukin-7 receptor essential for lymphocyte development and survival; CXCR5, a chemokine receptor guiding B-cell trafficking within lymphoid tissues; VCAM1, an adhesion molecule mediating leukocyte migration across the endothelium; LTBR and CD40, key regulators of lymphoid tissue organization and immune activation; and the transcription factors MAF, BATF, and BCL6, which orchestrate T- and B-cell differentiation and immune cell fate decisions. To this list, we added *BAFF-var*, a variant previously identified in the Sardinian multiple sclerosis GWAS [76], which affects BAFF (B-cell activating factor), a cytokine critical for B-cell survival, maturation, and immune regulation.

The association analysis was conducted on a cohort of 2,654 individuals profiled for gut metagenomic features. To ensure robustness and comparability across individuals, we applied a series of preprocessing steps. First, we filtered the microbial taxa to retain only those present in at least 10% of the samples, in order to focus on consistently detected features. The abundance values of each taxon were then transformed using an inverse normal transformation to approximate a normal distribution. Next, we used a linear model to regress out the effects of age, sex, and age squared from each taxonomic feature, and subsequently applied a second inverse normal transformation to the resulting residuals. This double normalization procedure was applied uniformly across all taxa.

The association between each of the nine genetic variants and the normalized taxonomic traits was then tested using the lmekin function, a linear mixed model that accounts for relatedness among individuals by incorporating a kinship matrix derived from pedigree information. This analytical framework was extended to metagenomic pathway abundances and alpha diversity metrics, such as the Shannon index, using the same covariate adjustment and normalization procedure.

We evaluated microbiome composition in individuals stratified by extreme genotypes related to multiple sclerosis susceptibility. We considered MS-associated variants in GALT genes and compared the gut metagenome composition of individuals carrying extreme GALT–MS associated genotypes, i.e., individuals homozygous for the risk alleles versus

individuals homozygous for the protective alleles. This approach was aimed at assessing whether genetic variants related to mucosal immunity and multiple sclerosis susceptibility are associated with detectable differences in gut microbial composition.

We also performed a complementary analysis in which additional covariates (Age, Sex, Age^2, Alcohol intake, Smoking, Diabetes, Proton Pump Inhibitors, BMI) were removed prior to association testing. Despite the biological relevance of the selected variants, none of the associations reached statistical significance after correction for multiple testing, under different thresholds including 5% and 10% false discovery rate (FDR), and 10% Bonferroni correction.

4.2.2. Correlation with the whole genome (GWAS)

4.2.2.1. GWAS of metagenomic taxa

We conducted a genome-wide association study (GWAS) on microbial taxonomic profiles derived from fecal metagenomes of 2,650 individuals from the ProgeNIA cohort. The analysis was performed using the GEMMA software, which implements a linear mixed model to account for relatedness via a kinship matrix. Age, age squared, and sex were included as covariates in this model. To focus on core components of the gut microbiota, the GWAS was restricted to 133 taxa present in at least 85% of individuals. A total of 10 loci reached genome-wide significance ($p < 5 \times 10^{-8}$), supporting the presence of host genetic effects on the abundance of specific gut microbes.

Among the top association signals, we highlight the most promising loci selected for follow-up analyses:

- *Faecalibacterium prausnitzii* (phylum *Bacillota*), a well-known anti-inflammatory commensal, showed a strong association with a variant on chromosome 9 (9:118047527_C_T, β = 2.260525e-01, p = 4.77×10^{-10}).
- The species *Phocaeicola vulgatus* (formerly *Bacteroides vulgatus*, phylum *Bacteroidota*) was associated with a locus on chromosome 9 (9:31214430_T_C, β = 1.809042e-01, p = 3.50×10^{-8}).
- The species *Clostridium phoceensis* (phylum *Bacillota*) showed an association with the variant 13:74627907_A_T (β = -2.587443e-01, p = 1.80×10^{-8}).
- A less characterized *Clostridiales bacterium* KLE1615 (within the phylum *Bacillota*) was linked to variant 1:179986829_T_G (β = -3.697327e-01, p = 2.80×10^{-8}).
- Finally, a broader *Clostridia*-level taxon (order *Clostridia_unclassified*) was associated with 5:125494872_C_G (β = 3.460079e-01, p = 1.94×10^{-8}).

For the LCT locus, which has been consistently replicated across multiple cohorts, we obtained a weaker association signal involving the genus *Bifidobacterium* (phylum *Actinobacteria*), corresponding to the variant 2:136616754_C_T (β = -3.01×10^{-1}, p = 1.92×10^{-6}).

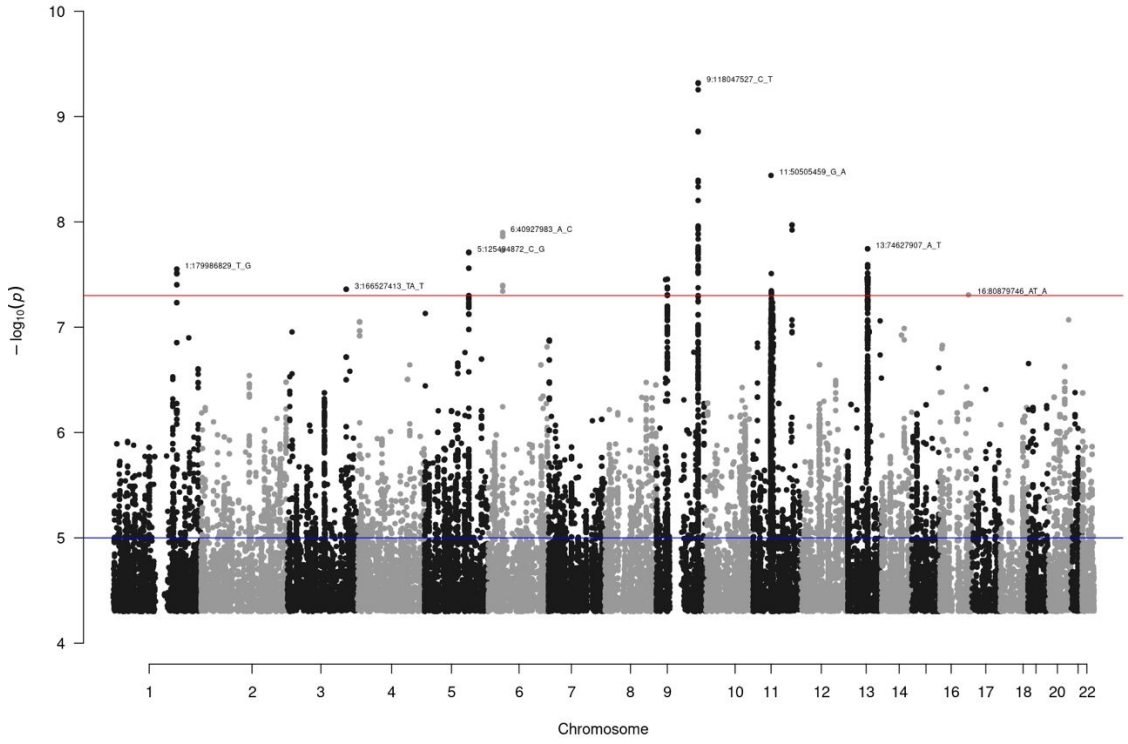

Figure 8. Manhattan plot illustrating the main associations reaching genome-wide significance for the taxa. Red and blue horizontal lines represent the genome-wide ($p < 5 \times 10^{-8}$) and suggestive ($p < 1 \times 10^{-5}$) significance thresholds, respectively. SNPs were filtered for minor allele frequency (MAF) > 0.05. We identified 10 loci associated with microbial taxa, showing p-values below 5×10^{-8}.

4.2.2.1.1. GWAS of metagenomic taxa with additional covariates

In the initial phase of the project, we conducted a genome-wide association study (GWAS) on 2,650 individuals from the ProgeNIA cohort, analyzing a comprehensive set of 959 microbial taxa selected after quality control filtering. This analysis identified 163 loci reaching genome-wide significance ($p < 5 \times 10^{-8}$) and 285 loci with suggestive significance ($p < 1 \times 10^{-7}$). To define independent association signals, we applied a two-step linkage disequilibrium (LD)-based clumping strategy, which resulted in 285 independent loci, each associated with at least one microbial taxon.

To further assess the robustness of these associations, we reran GEMMA including an extended set of covariates in addition to age, age², and sex. The additional variables included alcohol intake, smoking status, diabetes, proton pump inhibitor (PPI) use, and body mass index (BMI).

When comparing the results obtained with this extended model to those of the baseline analysis, we observed notable differences in the strength and significance of some associations. In particular, the association between the bacterial family *Eggerthellaceae* (phylum *Actinobacteria*) and liver fibrosis in non-alcoholic fatty liver disease, as well as the one between the bacterial order *OFGB3047* (class *CFGB3047*, phylum *Bacillota*) and depression, remained largely consistent or slightly increased in significance, indicating that these signals are robust to the inclusion of additional covariates.

Conversely, other associations showed a marked attenuation of the signal. For example, the associations between *Candidatus Saccharibacteria* (unclassified species within the phylum *Candidatus Saccharibacteria*) and primary open-angle glaucoma, and between the bacterial species *GGB6613_SGB9347* (phylum *Pseudomonadota*, formerly known as *Proteobacteria*) and coronary artery disease, substantially decreased in significance and no longer reached genome-wide significance thresholds.

Overall, these exploratory analyses indicate that the inclusion of lifestyle- and medication-related variables can meaningfully influence the observed genetic associations, highlighting the importance of covariate selection in the interpretation of metagenome-wide association results.

4.2.2.2 GWAS of microbial pathways

We performed genome-wide association studies (GWAS) on microbial functional profiles (pathways) inferred from shotgun metagenomic data in 2,650 individuals of the ProgeNIA cohort with HUMAnN software (See Methods). The GWAS analysis was conducted using the GEMMA software, incorporating a linear mixed model with a kinship matrix to account for relatedness among individuals. Age, age squared, and sex were included as covariates in all models (See Methods). To focus on core microbial functions, the analysis was restricted to 833 pathways detected in at least 85% of the samples. A total of 27 independent loci reached genome-wide significance ($p < 5 \times 10^{-8}$), supporting the existence of host genetic effects on the abundance of specific microbial metabolic functions. Among the top signals observed, we highlight two loci of interest that were selected for follow-up analyses:

- 7:7401490_C_A, associated with the pathway THISYNARA-PWY (*superpathway of thiamine diphosphate biosynthesis III*), primarily contributed by *Bacteroides uniformis* (β = 1.862583e-01, $p = 1.11 \times 10^{-8}$).
- 16:80457546_C_G, associated with the pathway PWY0-162 (*superpathway of pyrimidine ribonucleotides de novo biosynthesis*), with β = 1.611843e-01 and $p = 5.17 \times 10^{-9}$.

These loci were prioritized for downstream post-GWAS analyses.

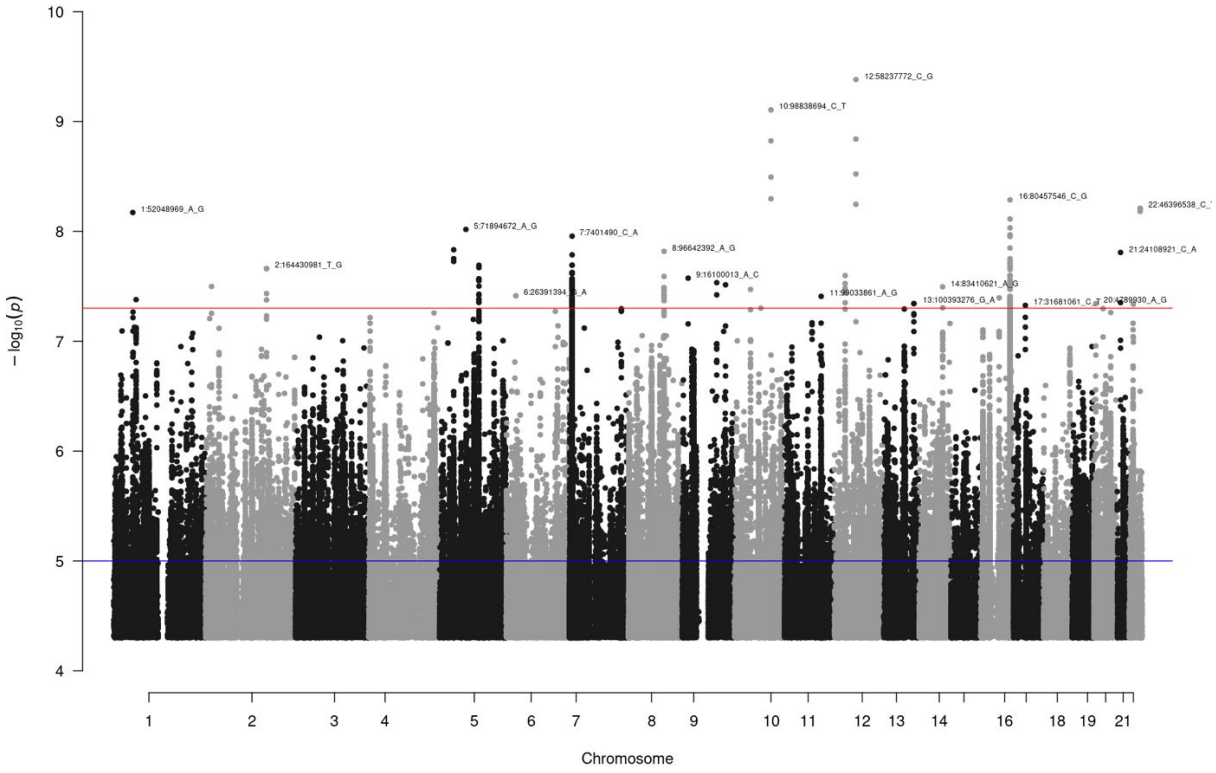

Figure 9. Manhattan plot illustrating the main associations reaching genome-wide significance for the pathways. Red and blue horizontal lines represent the genome-wide ($p < 5 \times 10^{-8}$) and suggestive ($p < 1 \times 10^{-5}$) significance thresholds, respectively. SNPs were filtered for minor allele frequency (MAF) > 0.05. We identified 27 independent loci associated with microbial pathways.

4.2.2.3 GWAS of metagenomic alpha diversity

We conducted genome-wide association studies (GWAS) on multiple indices of metagenomic alpha diversity, including Shannon index, Simpson index, richness, and Gini index, using data from 2,654 individuals in the ProgeNIA cohort. The analyses were performed using GEMMA, which applies linear mixed models to account for population structure and genetic relatedness, incorporating a kinship matrix derived from pedigree information as a random effect. We tested two models. The first included only age and sex as covariates (Figure 10). The second model was adjusted for a broader set of

covariates including alcohol intake, smoking, diabetes, proton pump inhibitor (PPI) use, body mass index (BMI), as well as age, sex, and age squared (Figure 11). No variants reached the genome-wide significance threshold ($p < 5 \times 10^{-8}$) in any of the tested models.

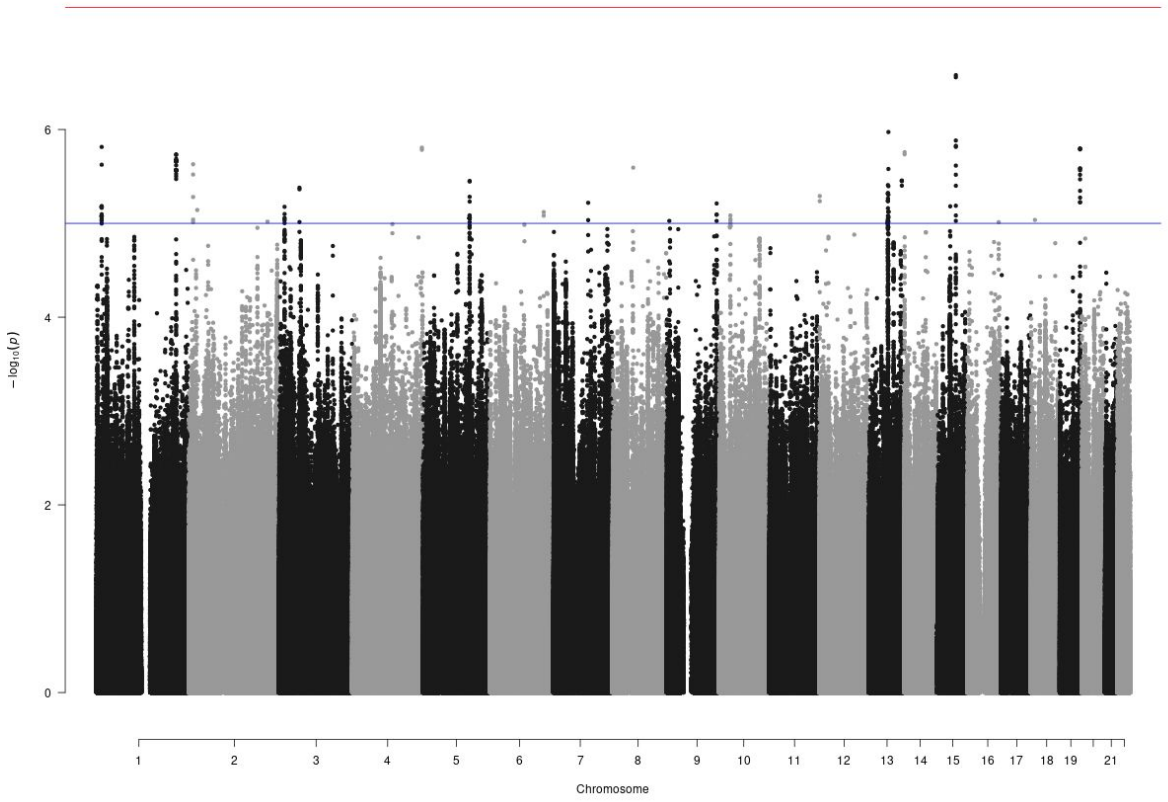

Figure 10. Manhattan plot illustrating the associations for alpha diversity using the covariates Age, Sex, Age². Red and blue horizontal lines represent the genome-wide ($p < 5 \times 10^{-8}$) and suggestive ($p < 1 \times 10^{-5}$) significance thresholds, respectively. SNPs were filtered for minor allele frequency (MAF) > 0.05. No variants reached the genome-wide significance threshold ($p < 5 \times 10^{-8}$).

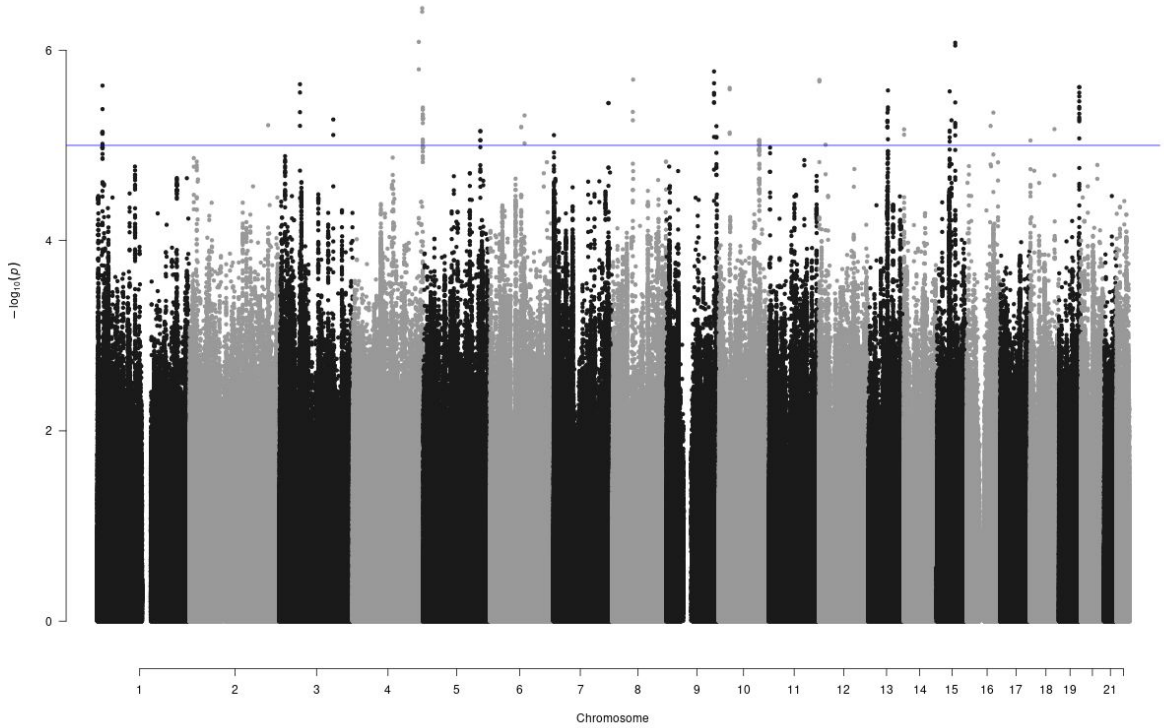

Figure 11. Manhattan plot illustrating the associations for alpha diversity using the covariates Age, Sex, Age², Alcohol intake, Smoking, Diabetes, Proton Pump Inhibitors, BMI. Red and blue horizontal lines represent the genome-wide ($p < 5 \times 10^{-8}$) and suggestive ($p < 1 \times 10^{-5}$) significance thresholds, respectively. SNPs were filtered for minor allele frequency (MAF) > 0.05. No variants reached the genome-wide significance threshold ($p < 5 \times 10^{-8}$).

4.2.2.4 GWAS of metagenomic beta diversity

We performed a genome-wide association study (GWAS) using beta diversity metrics derived from taxonomic profiles. Bray–Curtis, Aitchison, and Jaccard dissimilarities between samples were computed and used to perform Principal Coordinates Analysis (PCoA) using capscale() function from the vegan R package [62], generating individual-level coordinates. For the unconstrained PCoA axes, we applied rank-based inverse normal transformation, removed the effects of age, age squared, and sex via linear

regression, and then applied a second inverse normal transformation to the residuals. For constrained PCoA (dbRDA) axes, already corrected for covariates, we applied only a single inverse normal transformation. The resulting normalized coordinates were used as quantitative phenotypes in GEMMA to assess SNP-level associations across the genome. Among all the tested beta diversity components, only two variants reached genome-wide significance ($p < 5 \times 10^{-8}$) with the top SNP 22:19807953_T_C (β = 1.665392e-01, $p = 9.93 \times 10^{-9}$), which was associated with the first constrained PCoA axis derived from Aitchison distance (*PCoA1_constr_aitchison.InvNorm*).

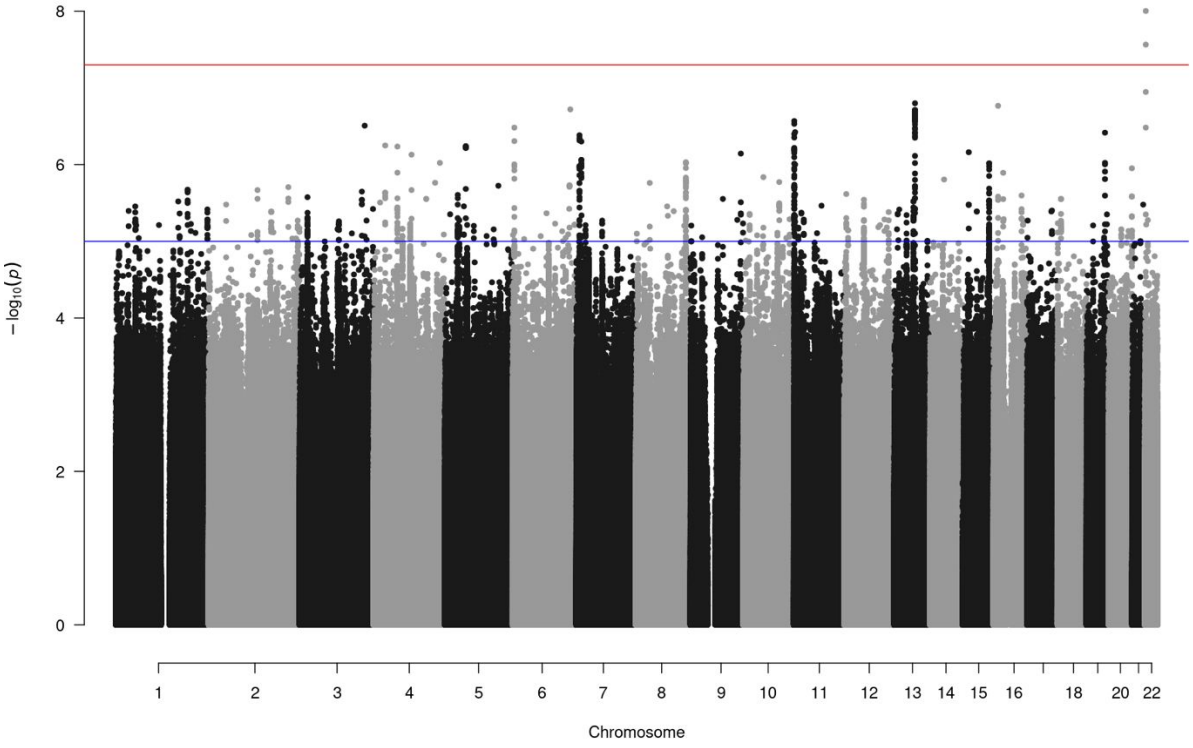

Figure 12. Manhattan plot illustrating the main associations for beta diversity. Red and blue horizontal lines represent the genome-wide ($p < 5 \times 10^{-8}$) and suggestive ($p < 1 \times 10^{-5}$) significance thresholds, respectively. SNPs were filtered for minor allele frequency (MAF) > 0.05. Only two variants reached genome-wide significance ($p < 5 \times 10^{-8}$) with the top SNP 22:19807953_T_C.

4.2.3. Post-GWAS analysis

4.2.3.1. Genomic context of top signals for microbial taxa, pathways, alpha and beta diversity

To further investigate the most prominent signals identified in the GWAS of microbial taxa, we explored the genomic context of the lead SNPs using LocusZoom plots and queried LinDA database for prior evidence of association with molecular traits (eQTLs) or complex phenotypes (GWAS hits). Among the top associations, several loci stood out for the strength and localization of their signals. The variant 1:179986829_T_G (β = -3.697327e-01, $p = 3.26 \times 10^{-8}$), associated with *Clostridiales bacterium KLE1615* (phylum *Bacillota*), shows a well-defined association peak on chromosome 1 (Figure 13A). The LocusZoom plot reveals a cluster of variants in high LD ($r^2 > 0.8$), suggesting a robust haplotype structure. According to LinDA, this SNP has been previously reported as an eQTL for the genes *ABL2*, *QSOX1*, and *RP11-533E19.3*, indicating potential regulatory effects relevant to host–microbe interactions. For the taxon *Clostridia_unclassified*, the variant 5:125494872_C_G (β = 3.460079e-01, $p = 1.94 \times 10^{-8}$) emerges as the lead SNP on chromosome 5 (Figure 13B). This variant is in LD with others near the genes *GRAMD3* and *PHAX*, both reported as eQTL targets in LinDA. The regional signal is compact and suggests a potential regulatory locus

influencing the abundance of this microbial group. Two additional loci were identified on chromosome 9, associated respectively with *Phocaeicola vulgatus* (β = 1.809042e-01, p = 3.50×10^{-8}) and *Faecalibacterium prausnitzii* (β = 2.260525e-01, p = 4.77×10^{-10}). The corresponding LocusZoom plots (Figures 13C and 13D) show localized association peaks with minimal extended LD, and no eQTLs or GWAS associations were reported for these variants in LinDA. These findings may indicate novel loci potentially involved in the modulation of these taxa. Lastly, a noteworthy signal was detected for *Clostridium phoceensis* (phylum *Bacillota*), associated with 13:74627907_A_T (β = -2.587443e-01, p = 1.80×10^{-8}) on chromosome 13 (Figure 13E). This variant overlaps the KLF12 gene, for which it has been identified as an eQTL, and has also been associated with nicotine dependence and refractive error in previous GWAS. Interestingly, in our own cohort, *C. phoceensis* exhibited the strongest correlation with smoking status (β = 0,44, p = 2.2×10^{-16}), supporting a potential host genetic link between this taxon and smoking-related phenotypes.

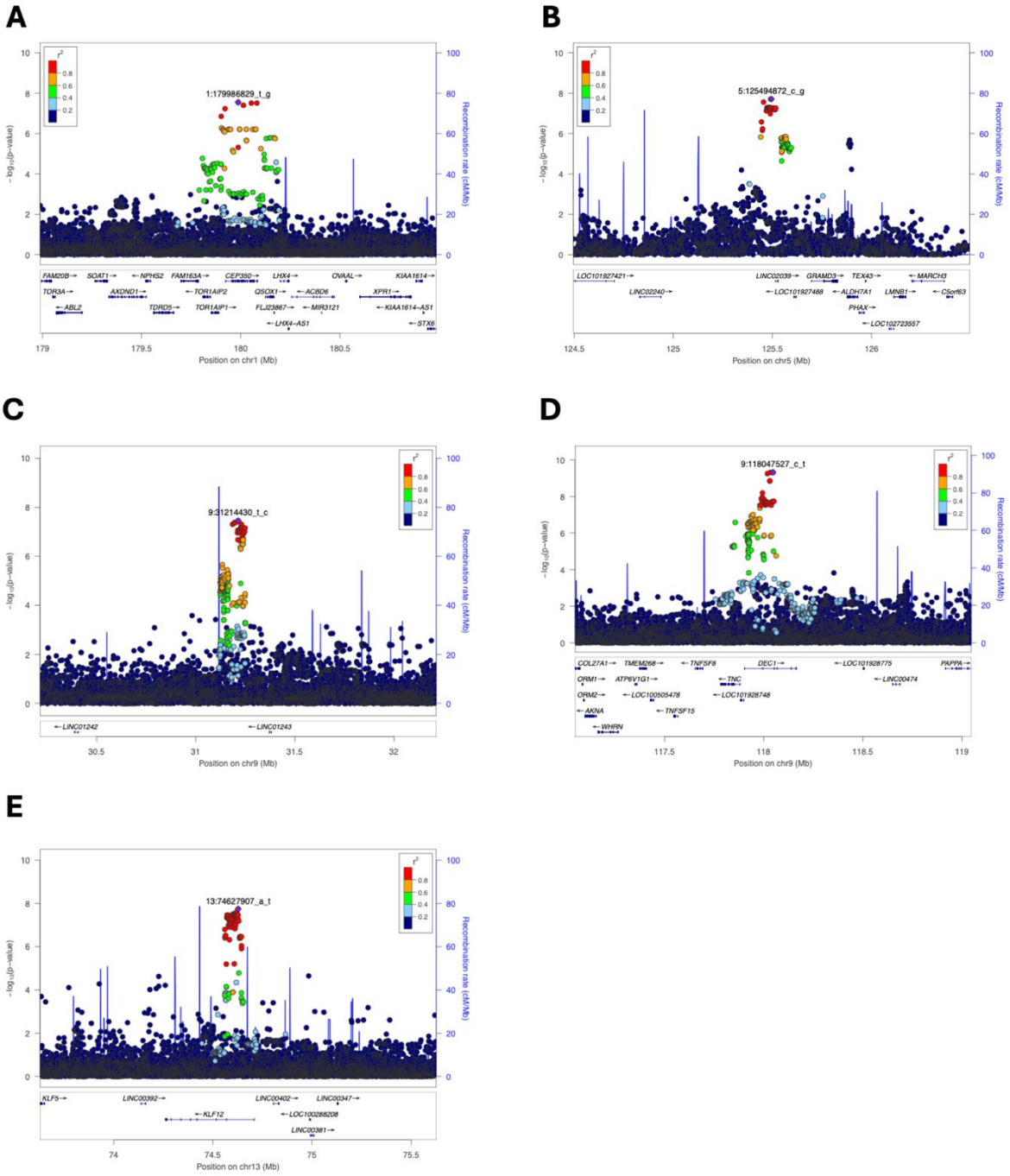

Figure 13. Regional association plots (LocusZoom) for selected genome-wide significant loci associated with microbial taxa. A) Variant 1:179986829_T_G, associated with *Clostridiales bacterium KLE1615* (phylum *Bacillota*). **B)** Variant 5:125505086_C_T, associated with an unclassified taxon in the order *Clostridia* (phylum *Bacillota*). **C)** Variant 9:31214430_T_C, associated with *Phocaeicola vulgatus* (phylum *Bacteroidota*). **C)** Variant 9:118047527_C_T, associated with *Faecalibacterium prausnitzii* (phylum *Bacillota*). **D)** Variant 13:74627907_A_T, associated with *Clostridium phoceensis* (phylum *Bacillota*). Each panel displays the $-\log_{10}$(p-value) for SNPs in a ±500 kb region centered on the lead SNP. SNPs are colored by linkage disequilibrium (r^2) with the lead variant. Blue vertical lines represent local recombination rates, and gene annotations are

shown below each plot. These plots illustrate the genomic architecture of the most promising host–microbe associations observed in our study.

Among the genome-wide significant loci identified for the pathways, two in particular showed strong association signals and clear LD structure: The variant 7:7401490_C_A (β = 1.862583e-01, p = 1.11×10^{-8}) was associated with the pathway THISYNARA-PWY, corresponding to the *superpathway of thiamine diphosphate biosynthesis III (eukaryotes)* and assigned taxonomically to *Bacteroides uniformis*. The LocusZoom plot (Figure 14A) reveals a dense association peak with several variants in high LD ($r^2 > 0.8$), suggesting a robust genetic signal. According to LinDA, this SNP acts as an eQTL for the genes *AC005532.5*, *tcag7.216* (likely a predicted transcript), and *MIOS*. Moreover, it has been reported in a prior GWAS as associated with bipolar disorder and schizophrenia, highlighting a potential host–microbiome link involving vitamin metabolism and neuropsychiatric traits. The second notable signal comes from the variant 16:80457546_C_G (β = 1.611843e-01, p = 5.17×10^{-9}), which showed a genome-wide significant association with the microbial pathway PWY0-162, related to the *superpathway of pyrimidine ribonucleotide biosynthesis de novo*. The corresponding LocusZoom plot (Figure 14B) displays a sharply defined association peak with a cluster of SNPs in tight LD. According to LinDA, this variant has been identified in three GWAS as associated with spherical equivalent (a refractive error phenotype), annual healthcare cost, and photoreceptor cell layer thickness, suggesting potential relevance to complex host traits.

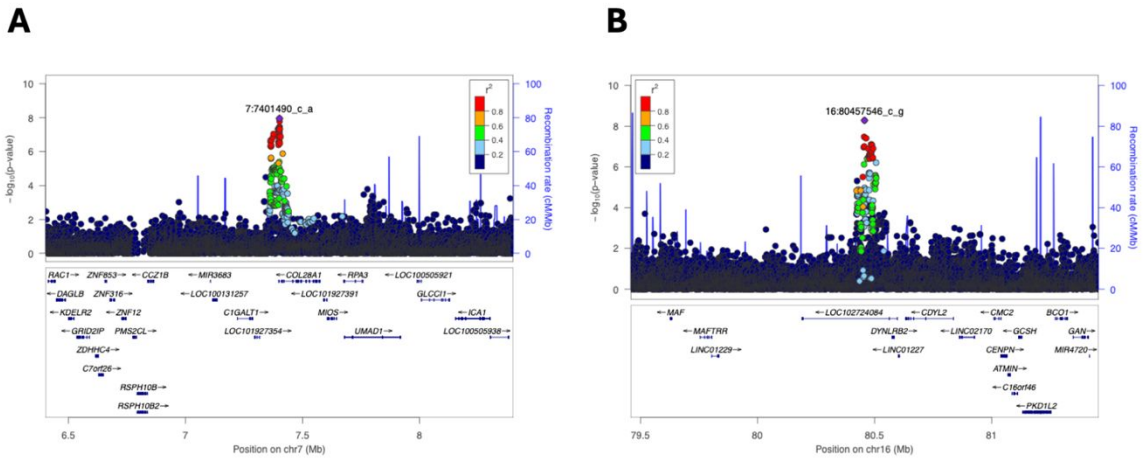

Figure 14. LocusZoom plots for the top genome-wide significant associations identified in the pathway GWAS. A) The variant 7:7401490_C_A (β = 1.862583e-01, p = 1.11 × 10^{-8}) is associated with the microbial pathway THISYNARA-PWY (superpathway of thiamine diphosphate biosynthesis III), which is primarily contributed by *Bacteroides uniformis*.

B) The variant 16:80457546_C_G (β = 1.611843e-01, $p = 5.17 \times 10^{-9}$) is associated with the microbial pathway PWY0-162 (superpathway of pyrimidine ribonucleotides de novo biosynthesis).

In both cases, the plots show regional association peaks with clusters of variants in high linkage disequilibrium (LD) and local recombination rates (blue line). Genes within the region are annotated below the x-axis.

The only variant reaching genome-wide significance in the beta diversity GWAS was 22:19807953_T_C (β = 1.665392e-01, p = 9.93 × 10^{-9}), associated with the first constrained principal coordinate of the aitchison metric (PCoA1_constr_aitchison.InvNorm). The LocusZoom plot (Figure 15) highlights a modest yet distinct association peak, with a few variants in moderate linkage disequilibrium (LD, $r^2 > 0.4$) supporting the signal. The region spans several genes, including *SEPT5*, *COMT*, *DGCR6*, and *GNB1L*, many of which lie within the 22q11.2 locus, known for its involvement in neurodevelopmental and immunological processes. However, no prior

GWAS or eQTL associations involving this specific variant were found in the LinDA database. While this limits functional interpretation, the uniqueness of this signal observed exclusively for a beta diversity axis suggests a potentially novel link between host genotype and overall gut microbial community structure. Further validation and fine-mapping will be necessary to better characterize this locus and determine whether its association reflects a direct host–microbiome interaction or arises from broader host phenotypes impacting microbiome composition.

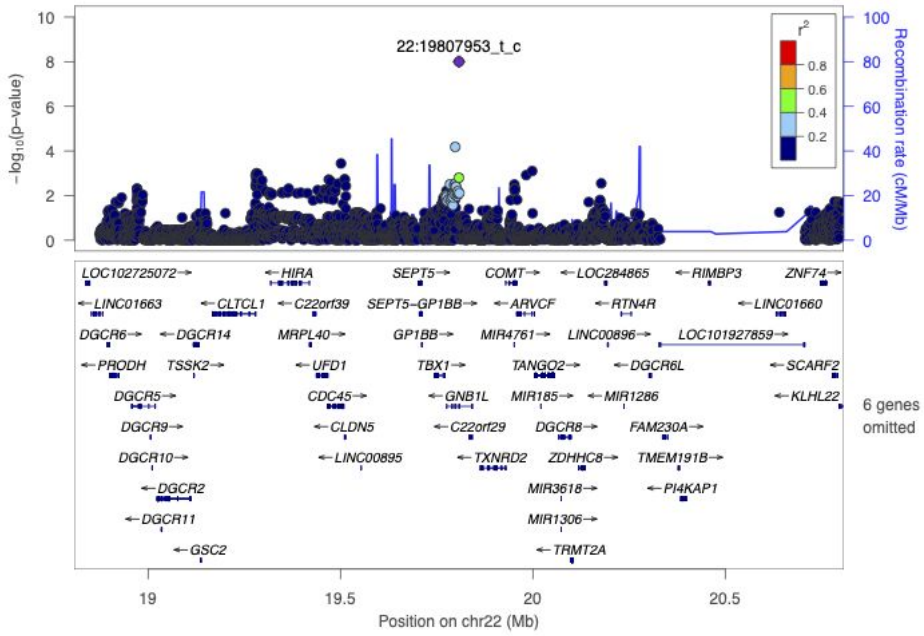

Figure 16. Association signal at the locus of variant 22:19807953_T_C for the first constrained principal coordinate of the Aitchison beta diversity. The LocusZoom plot shows a moderately localized association peak on chromosome 22. The lead variant (22:19807953_T_C, β = 1.665392e-01, $p = 9.93 \times 10^{-9}$) lies within a region spanning several genes including *SEPT5*, *COMT*, *DGCR6*, and *GNB1L*. The color gradient represents linkage disequilibrium (LD, r^2) with the top SNP, and the blue curve denotes local recombination rates. No known eQTLs or previous GWAS associations were reported for this SNP in the LinDA database, suggesting a potentially novel locus influencing beta diversity structure.

4.2.3.2. COLSTATS: a repository and web server for colocalization analysis

Understanding whether two traits share a common genetic basis is a crucial step in uncovering the molecular mechanisms underlying complex phenotypes [77]. Colocalization analysis aims to determine whether two association signals, typically one from a molecular trait and one from a complex trait, are driven by the same causal variant. While statistical frameworks such as the coloc package have enabled rigorous Bayesian inference for colocalization, their application remains limited by several challenges, including the need for harmonized input data, the selection of appropriate priors, and the requirement for advanced computational skills [78].

To overcome these limitations and make colocalization analysis more accessible to a broader scientific community, we developed COLSTATS, an interactive web-based platform and harmonized repository of summary statistics specifically designed for colocalization studies. The tool integrates a comprehensive collection of over two million genome-wide summary statistic datasets; all aligned to the hg38 genome build and formatted in a standardized VCF-like structure. Each variant is uniquely identified by a chr_pos_ref_alt ID, ensuring consistency across datasets and simplifying variant matching. All summary statistics are uniformly annotated with fields such as p-values, effect sizes, standard errors, allele frequencies, and sample sizes, and are accompanied by structured metadata describing

traits, study identifiers, ancestries, tissues or cell types (for molecular traits), and publication information.

The COLSTATS database includes data from several major resources in human genetics and functional genomics. These include:

- NHGRI-EBI GWAS Catalog [66], providing association statistics for thousands of complex traits and diseases across multiple ancestries.
- UK Biobank (UKBB) [67], offering over 200,000 GWAS summary statistics for quantitative and case–control traits in diverse ancestry groups.
- Orrù et al., 2020 [68], including 731 GWAS of immunophenotypes (cell counts, parental percentages, and surface protein expression) from the SardiNIA cohort.
- GTEx Consortium [69], covering cis-eQTLs across 44 human tissues.
- BLUEPRINT [70], reporting eQTLs for three major immune cell types ($CD14^+$ monocytes, $CD16^+$ neutrophils, and naïve $CD4^+$ T cells).
- eQTLGen Consortium, phase 1 [71], providing large-scale cis- and trans-eQTLs from blood-derived expression data in 31,684 individuals.
- Pala et al., 2017 [57], containing cis-eQTLs from peripheral blood leukocytes of 606 Sardinian individuals from the SardiNIA study.
- Ota et al., 2021 [72], comprising cis-eQTLs across 28 immune cell subsets from patients with immune-mediated diseases and healthy controls.

All datasets have been harmonized using a custom pipeline that ensures compatibility across sources and allows for real-time integration into the colocalization workflow (See Methods). Users can interactively search and filter the database by study, ancestry, trait type, or tissue, and select any pair of traits to test for colocalization. After defining a genomic region of interest, either by gene symbol or by custom coordinates, the platform extracts overlapping variants and runs a colocalization analysis using the coloc.abf method. This Bayesian framework estimates the posterior probabilities of five hypotheses, including the presence of shared or distinct causal variants between the two traits.

Summary statistics source	Type	Category	# traits with summary statistics
NHGRI-EBI GWAS Catalog	GWAS	Traits and Disease	80.295
UK Biobank - UKBB	GWAS	Traits and Disease	216.113
Orrù eta al., 2020	GWAS	Traits (immunophenotypes)	731
GTEx	cis-eQTLs	Traits (RNA levels)	1.086.146
BLUEPRINT	cis-eQTLs	Traits (RNA levels)	48.675
eQTLGen	cis- and trans-eQTLs	Traits (RNA levels)	39.192
Pala et al. 2027	cis-eQTLs	Traits (RNA levels)	21.183
Ota et al., 2021	cis-eQTLs	Traits (RNA levels)	440.956
TOTAL			1.933.291

Table 1: Summary of dataset inventory across sources (e.g., number of summary statistics per resource).

The results are returned through an interactive interface that includes summary tables, graphical visualizations such as LocusZoom-style plots, and downloadable output files. A real-time prior sensitivity panel allows users to explore how changes in prior probabilities affect the colocalization

outcome. All analyses are cacheable and reproducible through unique URLs, which facilitates result sharing and collaboration.

COLSTATS thus provides a unified, user-friendly environment for performing robust colocalization analysis on a massive scale. By integrating harmonized summary statistics from multiple sources with an intuitive graphical interface and automated statistical routines, it enables researchers to explore shared genetic architecture across traits without the need for complex preprocessing or programming. This tool is particularly valuable for multi-omics integration and hypothesis generation in human genetics and functional genomics research.

COLSTATS is freely accessible at https://colstats.irgb.cnr.it.

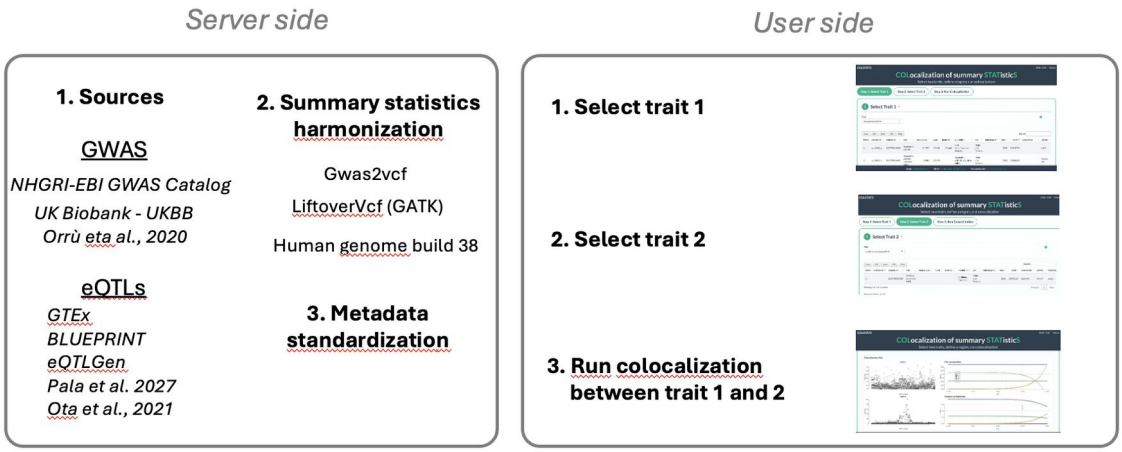

Figure 17. Server-side and user-side workflow of COLSTAT. On the server side, summary statistics from multiple sources (GWAS Catalog, UK Biobank, population cohorts, and eQTL datasets such as GTEx, BLUEPRINT, and eQTLGen) are collected. These data are harmonized through standardized variant representation (e.g., gwas2vcf, liftover to the human reference genome build 38) and metadata unification. On the user side, the web interface guides the analysis in three steps: (1) selection of the first trait of interest, (2) selection of the second trait, and (3) colocalization analysis between the two traits. The results include interactive regional plots and posterior probability estimates that support causal inference.

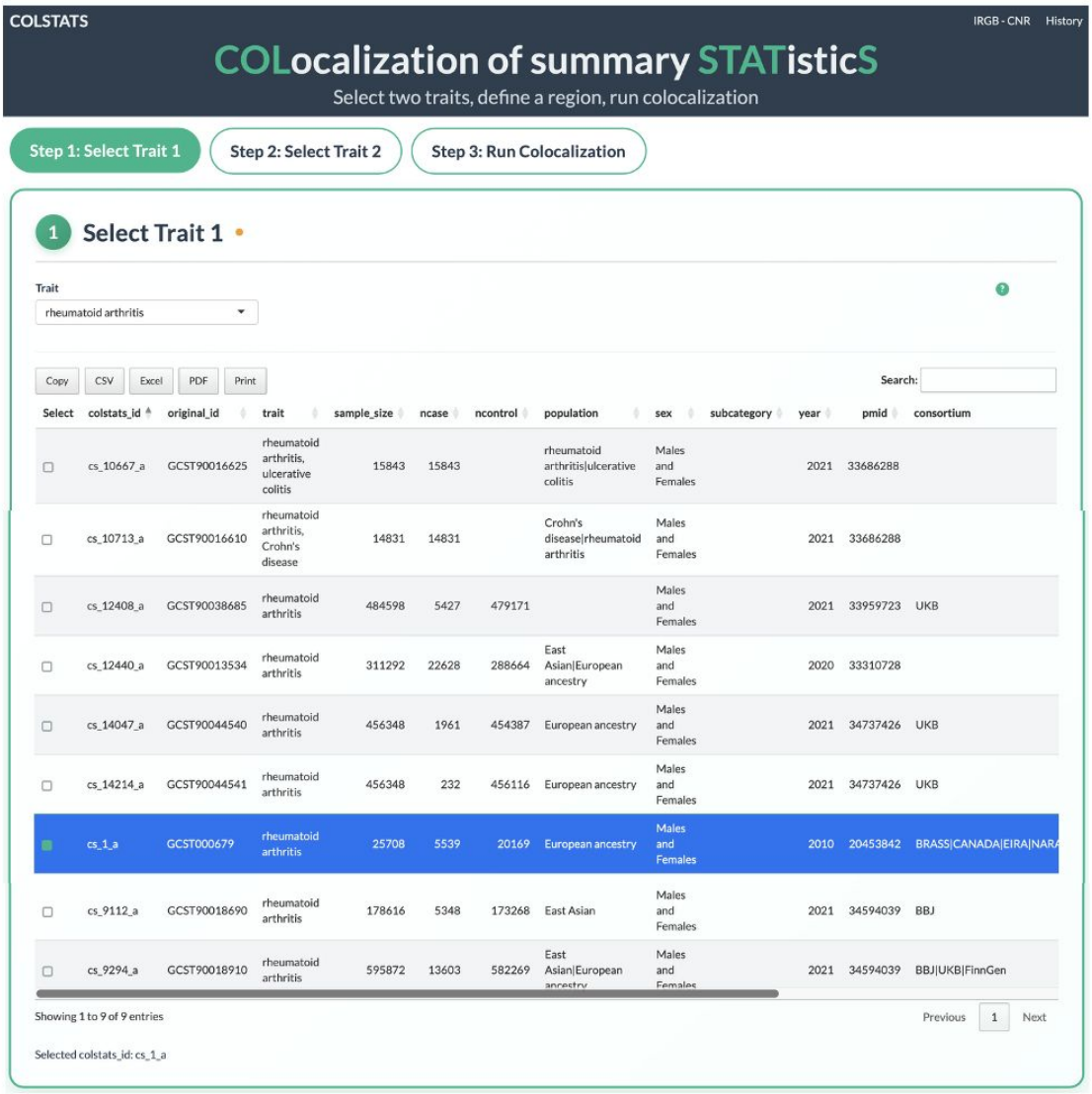

Figure 18. Step 1—Select Trait 1 in COLSTATS. The interface displays a searchable, sortable table of curated GWAS summary-statistics records from which the user chooses the disease/trait to be used in the colocalization workflow. Columns shown include colstats_id, original_id, trait, sample_size, ncase, ncontrol, population, sex, subcategory, year, PMID, and consortium, with one-click export options (CSV/Excel/PDF). In this example, rheumatoid arthritis is selected (blue highlight; GCST000679), which will be carried forward to Steps 2–3 to test colocalization with the CD40 cis-eQTL measured in peripheral blood leukocytes.

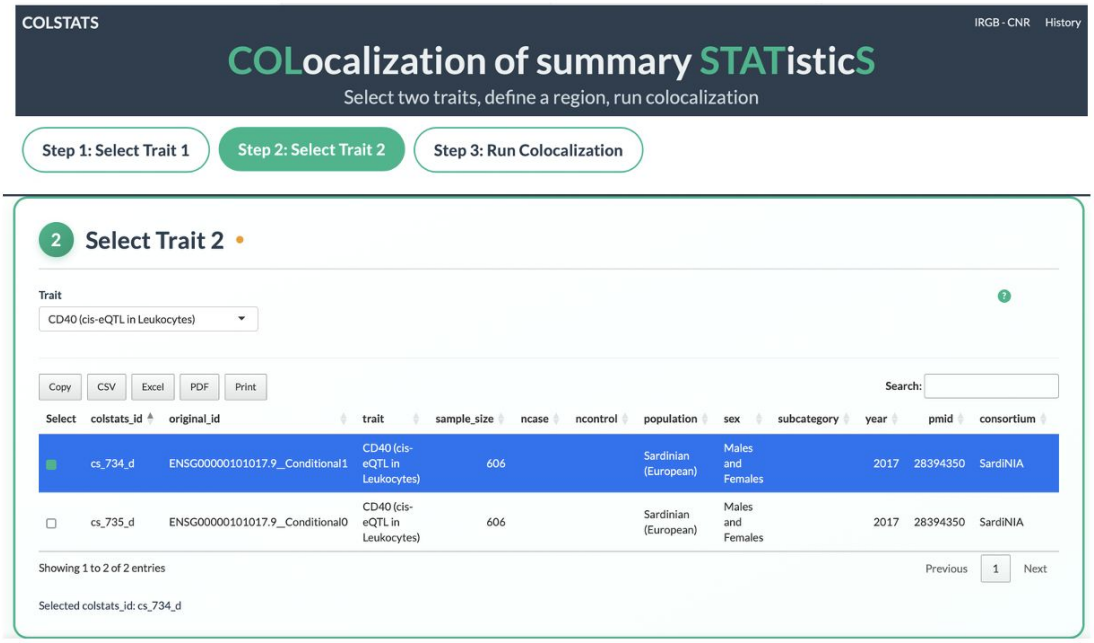

Figure 19. Step 2—Select Trait 2 in COLSTAT. The interface lists molecular traits available for colocalization. Here, CD40 cis-eQTL in leukocytes from Pala et al (14) was chosen. Within the Pala leukocyte eQTL resource, two versions are currently available for CD40: (i) the unconditioned scan (*Conditional0*; no conditional analysis) and (ii) a conditionally independent signal (*Conditional1*) obtained by forward stepwise regression. In this example, the conditional dataset is selected (blue highlight; *colstats_id* cs_734_d, *original_id* ENSG00000101017.9__Conditional1). Metadata shown include sample size (N=606), population (Sardinian/European), sex (males and females), year (2017), PMID (28394350), and consortium (SardiNIA). Other eQTL resources may be added to COLSTAT in future releases; this figure illustrates the options currently available for CD40 in the Pala dataset and defines Trait 2 for downstream colocalization with the rheumatoid arthritis GWAS chosen in Step 1.

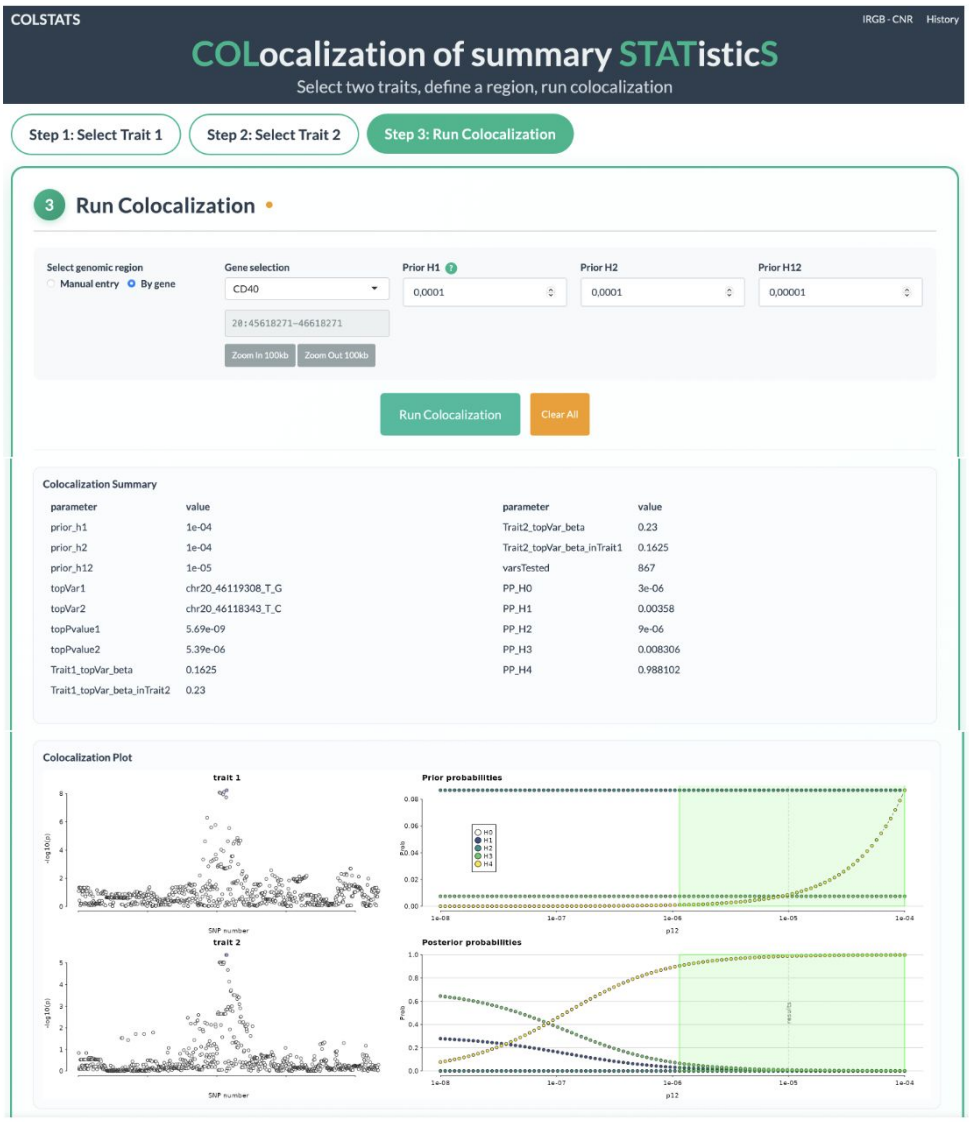

Figure 20. Step 3—Run Colocalization in COLSTATS and visualize results. The CD40 locus is selected because we test a cis-eQTL effect on CD40 for colocalization with rheumatoid arthritis (RA) GWAS (Trait 1). Default priors were used (prior H1 = 1×10^{-4}, prior H2 = 1×10^{-4}, prior H12 = 1×10^{-5}), and 867 variants were evaluated in the region. The strongest RA association is at chr20:46119308_T_G (Trait 1 top P = 5.69×10^{-9}), while the strongest CD40 eQTL is at chr20:46118843_T_C (Trait 2 top P = 5.39×10^{-6}). Colocalization results (coloc): Posterior probabilities are $PP_0=3\times10^{-6}$, $PP_1=0.00358$, $PP_2=9\times10^{-6}$, $PP_3=0.008306$, and $PP_4=0.988102$, providing strong evidence for one shared causal variant driving both the RA signal and the CD40 cis-eQTL. Effect directions for the lead variants are concordant across traits: the RA top variant has a positive effect on CD40expression (β in Trait 2 ≈ +0.23), and the eQTL top variant shows a positive effect on RA (β in Trait 1 ≈ +0.16). This sign concordance indicates that higher CD40 expression associates with increased RA risk. Plots (from the coloc tool): Bottom-left, local association ($-\log_{10}P$) plots for

Trait 1 (top panel) and Trait 2 (bottom panel) show peaks aligned in the same LD block, visually supporting colocalization. Bottom-right, the sensitivity analysis shows prior (top) and posterior (bottom) probabilities across a range of shared-causal priors (p_{12}); PP_4 remains the dominant outcome over a broad, plausible prior range (vertical line marks the chosen prior; shaded area highlights recommended values), indicating that the inference of colocalization is robust to prior choices.

4.2.3.3. Colocalization of Sardinia metagenome GWAS complex traits and diseases

As previously described (in Section 4.2.2.1. GWAS of metagenomic taxa), in the initial phase of the project, we conducted a genome-wide association study (GWAS) on 2,650 individuals from the ProgeNIA cohort, using a broader set of 959 microbial taxa selected after quality control filtering. This analysis led to the identification of 163 loci associated at genome-wide significance ($p < 5 \times 10^{-8}$) and 285 loci with suggestive significance ($p < 1 \times 10^{-7}$). To group association signals into independent loci, we applied a two-step linkage disequilibrium (LD)-based clumping strategy (See GWAS Methods) using plink (v1.90b6.20) [79], which resulted in 285 independent loci, each associated with at least one microbial taxon.

To investigate whether these associations overlapped with those reported for diseases or complex traits, we checked whether the lead variants of our loci were in LD ($r^2 > 0.6$) with sentinel variants reported in the GWAS Catalog. In total, 15 of our loci were found to be in LD with known disease- or trait-associated variants, spanning 30 phenotypes and 7 disease categories. To further assess whether these overlapping signals were driven by a shared causal variant, we performed colocalization analysis using the coloc software and full summary statistics for the corresponding traits and

diseases. Among the most notable findings (posterior probability for hypothesis H4 > 0.8) was a strong colocalization with coronary artery disease (CAD), where we observed a posterior probability H4 = 0.99 for a shared causal variant between CAD and the bacterial species *Parabacteroides merdae*. In our data, the CAD risk allele (T) was associated with increased abundance of *P. merdae*. While this observation points to a potential mechanistic link between host genetics, gut microbiota, and cardiovascular health, the role of *P. merdae* in CAD remains controversial. Previous studies have reported both protective [80], [81], [82]
and detrimental [83],[84] associations, suggesting that its impact may be context-dependent or influenced by population-specific, dietary, or metabolic factors. Our findings contribute to this ongoing debate and highlight the need for further investigation into the causal relationship between *P. merdae* and cardiometabolic diseases.

4.2.4. Cohabitation and heritability

To estimate the contribution of host genetic factors to gut microbiome composition, we performed a heritability analysis using the SOLAR software focusing on the Shannon alpha diversity index. This metric was selected as a representative measure of within-sample microbial diversity. In this first phase, cohabitation was not yet included as a covariate. The analysis was performed on 2,613 individuals with available phenotype and pedigree data. Using a polygenic model that adjusted for age, age squared, and sex, the narrow-sense heritability (h^2) of Shannon diversity was estimated at 0.325

(standard error = 0.048), with a highly significant p-value ($p = 7.56 \times 10^{-14}$). These results indicate that approximately 32.5% of the phenotypic variance in Shannon diversity is attributable to additive genetic effects captured by the pedigree structure. The proportion of variance explained by covariates was modest (0.6%), and the model exhibited acceptable residual kurtosis (0.5911), supporting the reliability of the estimates. Although suggestive of a genetic component influencing microbiome diversity, these results do not account for shared environmental exposures such as household or cohabitation. For this reason, to account for shared environmental effects, we extended the model by including cohabitation as an additional random effect, based on a binary matrix distinguishing cohabiting from non-cohabiting individuals. This analysis was conducted on the subset of 2,019 individuals with both metagenomic and household data. In the extended model, the heritability estimate decreased to 24.9% (standard error = 5.8%, $p = 1.0 \times 10^{-6}$), while the effect of cohabitation (c^2) was estimated at 4.2% (standard error = 2.2%, p = 0.018). These results indicate that a portion of the variance initially attributed to genetic effects may in fact reflect shared environmental exposures within households.

5. DISCUSSION

This study provides a comprehensive investigation of host–microbiome interactions in a well-characterized population, integrating metagenomic profiles with host phenotypes, genotypes, and family structure. Using data from 2,654 individuals from the ProgeNIA cohort, we examined how the gut microbiota is associated with clinical, demographic, and lifestyle traits, how it is influenced by host genetics through genome-wide association studies (GWAS), and whether shared genetic loci underlie associations with human complex diseases through colocalization and Mendelian Randomization analyses. Furthermore, we leveraged the unique family-based design of the cohort to assess the role of cohabitation and genetic relatedness in shaping microbiome variation. The results offer a multilayered view of the host–microbiome axis and provide insights into both environmental and genetic influences on gut microbial features.

5.1. Microbiome–phenotype associations reflect known and population-specific patterns

Our large-scale correlation analyses between gut metagenomic features and phenotypic traits from the ProgeNIA cohort revealed widespread associations across taxonomic, functional, and diversity-related microbial measures. Age and sex emerged as the strongest drivers of microbiome variation, showing associations with more than 400 microbial taxa and the

highest number of pathways, in line with previous studies and reinforcing their fundamental role in shaping gut microbial composition and function. Other traits known to influence the gut microbiota, such as BMI, glycemia, cholesterol levels, and lifestyle exposures including smoking and alcohol consumption also showed numerous significant associations, particularly within the *Bacillota* (formerly known as *Firmicutes*) and *Bacteroidota* (formerly known as *Bacteroidetes*) phyla. The consistency of these findings with established biological mechanisms supports their robustness, yet several associations diverged from those reported in other cohorts, underscoring the complexity and context-dependency of host–microbiota relationships. These discrepancies are likely multifactorial. First, technical and analytical differences such as sequencing depth, reference database used, microbial profiling resolution (species vs genus), normalization methods, and handling of zero-inflated data can greatly affect the detection of associations. Second, population-specific factors, including ancestry, diet, environmental exposures, and cultural habits, may lead to cohort-dependent patterns that limit replication across studies. Finally, variation in statistical modeling, particularly in the selection and inclusion of covariates, can substantially influence both effect estimates and significance, as we also observed in our covariate-adjusted GWAS models [35]. For example, the strong negative correlation we observed between Shannon diversity and current smoking status aligns with prior reports of smoking-related reductions in microbial diversity but also highlights the dynamic nature of this effect and diversity levels appeared to recover in individuals who had

quit smoking. Similarly, alcohol consumption, particularly wine intake, was positively associated with beta diversity metrics, reinforcing previous findings on the role of polyphenol-rich foods in promoting microbial richness and compositional shifts. Altogether, our results reflect both expected and novel patterns of association, emphasizing the need for careful contextual interpretation of microbiome-phenotype findings, especially in population-based studies. Efforts toward standardization, meta-analysis, and replication across ancestries will be crucial to identify robust, generalizable microbiome markers of human traits and diseases.

5.2. Host genetic effects on the gut microbiota: insights from multi-trait GWAS analyses

We conducted a series of genome-wide association studies (GWAS) to investigate the contribution of host genetic variation to gut microbial composition and function in the ProgeNIA cohort. Our analyses encompassed multiple levels of the microbiome: taxonomic profiles, functional pathways, and both alpha and beta diversity using linear mixed models implemented in GEMMA with a kinship matrix to account for relatedness among individuals. The GWAS on taxonomic features identified 10 loci surpassing the genome-wide significance threshold ($p < 5 \times 10^{-8}$) when restricting the analysis to 133 taxa present in at least 85% of individuals. Among the strongest associations were variants linked to well-known commensals such as *Faecalibacterium prausnitzii* and *Phocaeicola vulgatus*. Particularly noteworthy was the variant

13:74627907_A_T, associated with *Clostridium phoceensis*, which overlaps the KLF12 gene and has been previously implicated in GWAS of nicotine dependence. Interestingly, *C. phoceensis* also showed the strongest correlation with smoking status in our cohort, suggesting a potential host genetic component connecting microbial abundance and smoking-related traits. To evaluate the robustness of these results, we re-ran the analyses including additional covariates such as alcohol intake, smoking status, diabetes, proton pump inhibitor (PPI) use, and body mass index (BMI). While several associations remained stable, others were attenuated after covariate adjustment, emphasizing the sensitivity of microbiome GWAS to environmental and lifestyle factors. The GWAS on microbial pathways revealed 27 independent loci reaching genome-wide significance ($p < 5 \times 10^{-8}$) across 833 pathways detected in at least 85% of individuals. Among the most prominent were two loci selected for post-GWAS exploration: 7:7401490_C_A, associated with a thiamine diphosphate biosynthesis superpathway largely contributed by *Bacteroides uniformis*, and 16:80457546_C_G, linked to pyrimidine ribonucleotide biosynthesis. Both loci displayed clear LD structure and overlap with genomic regions previously associated with complex traits in the literature, suggesting possible shared regulatory mechanisms between host metabolism and microbial functional activity. For microbial alpha diversity, we analyzed four indices: Shannon, Simpson, richness, and Gini. In none of the tested models did any variant reach the genome-wide significance threshold ($p < 5 \times 10^{-8}$), suggesting that common host genetic variation does not have a strong or

consistent influence on overall gut microbial diversity. In the analysis of beta diversity, which captured inter-individual dissimilarities in microbial community structure, only one variant reached genome-wide significance: 22:19807953_T_C ($\beta = 1.665392e\text{-}01, p = 9.93 \times 10^{-9}$), associated with the first constrained PCoA axis derived from Aitchison distance. Although this variant has not been previously reported in GWAS or eQTL databases, it lies within the 22q11.2 region, which includes genes involved in immune and neurodevelopmental processes.

While preliminary, this finding suggests a potential genetic influence on overall microbial community variation. Together, these results provide a nuanced picture of host–microbiome interactions. Significant loci were identified for specific microbial taxa and pathways, yet signals for diversity metrics were weaker and highly sensitive to covariate modeling suggest that host genetic effects on the gut microbiota are real but limited in magnitude. This observation aligns with the growing consensus in the field that host genetics explains only a small fraction of the variability in microbial features, with environmental and lifestyle factors playing a predominant role. Importantly, our post-GWAS analyses (including LocusZoom and LinDA queries) allowed us to place several loci in functional and phenotypic context, supporting biological plausibility for some associations and helping to prioritize candidate regions for further investigation. Future studies leveraging larger sample sizes, multi-omics integration, and longitudinal designs will be essential to refine these

associations and to clarify the mechanisms through which host genetics interacts with the gut microbiome.

5.3. Colocalization highlights shared genetic architecture between microbiota and disease

To explore whether host genetic variants may exert pleiotropic effects on both the gut microbiome and human complex traits, we performed colocalization analyses between our metagenome-wide association signals and publicly available GWAS summary statistics. This approach allowed us to test whether a single genetic variant could simultaneously drive variation in microbial abundance and disease risk. Among the most notable results, we identified a strong colocalization (posterior probability for hypothesis H4 = 0.99) between the bacterial species *Parabacteroides merdae* and coronary artery disease (CAD). The lead variant, 21:35605863_C_T, was associated in our data with increased abundance of *P. merdae* and, in external CAD GWAS datasets, with increased disease risk. This observation supports the presence of a shared causal variant influencing both the microbial trait and cardiovascular phenotype. While this finding suggests a potential mechanistic link between host genetics, microbiota, and cardiometabolic outcomes, the biological interpretation remains complex. In fact, *P. merdae* has been variably described in the literature as both protective and detrimental in relation to cardiovascular and metabolic health. Some studies have reported decreased abundance of *P. merdae* in individuals with atherosclerosis or cardiometabolic disease, suggesting a

protective role, while others have observed increased abundance in patients with similar conditions, indicating a possible contribution to disease progression. The conflicting evidence may reflect context-dependent effects mediated by diet, inflammation, medication use, or other environmental exposures. It is also possible that different strains or functional capacities within the *P. merdae* species drive opposing associations in different cohorts. Our results contribute to this ongoing debate by providing genetic evidence of colocalization, but do not alone establish causality or directionality of the effect. Overall, this finding exemplifies the power and complexity of integrating host genetic and microbiome data, and highlights the importance of replication, strain-level resolution, and functional validation in future studies aiming to clarify the causal role of specific microbes in human disease.

5.4. Host genetic contribution to microbial alpha diversity: preliminary evidence and the role of shared environment

In the initial heritability analysis conducted with SOLAR, we estimated that approximately 32.5% of the variance in the Shannon alpha diversity index could be attributed to additive genetic factors ($h^2 = 0.325$, SE = 0.048, $p = 7.56 \times 10^{-14}$). This value appears notably high compared to most previous studies, though the literature reveals considerable heterogeneity in alpha diversity heritability estimates. While some studies report near-zero heritability [30], others have found moderate to substantial genetic

contributions: the Hutterite population showed 52% heritability for Shannon diversity during winter months [85], and the TwinsUK cohort demonstrated 37% heritability for phylogenetic diversity [31], though both studies acknowledged large confidence intervals and seasonal variation. This initial high estimate may be partially explained by unaccounted confounding from environmental factors shared within families, particularly cohabitation. In pedigree-based models, individuals who live together often appear genetically related, making it difficult to separate the contribution of genetics from that of a shared household. Given that diet, hygiene habits, and microbial exchange can all strongly influence the gut microbiota, failing to account for cohabitation may inflate heritability estimates. After incorporating a cohabitation matrix into our model, the estimated heritability decreased to 24.9%, yet remained relatively substantial. This persistent genetic signal, even after controlling for shared household effects, warrants careful interpretation. Several factors may explain this finding. First, the Sardinian population represents a genetic isolate with unique characteristics including reduced genetic heterogeneity, founder effects, and limited gene flow from external populations. Such population structure could amplify genetic signals compared to more heterogeneous cohorts studied previously. Additionally, the relatively homogeneous traditional diet and lifestyle in Sardinia may reduce environmental "noise," allowing genetic effects to emerge more clearly. Second, our methodology using SOLAR with explicit cohabitation control may provide more accurate estimates than traditional twin studies, which often cannot fully disentangle

shared environment from genetic effects. The residual heritability might therefore represent a genuine genetic contribution specific to this population. Nevertheless, we acknowledge that unmeasured environmental factors correlated with kinship, such as long-term dietary traditions, early-life exposures, or transgenerational lifestyle patterns, could still contribute to the observed heritability. The wide variation in alpha diversity heritability across studies (ranging from 0% to over 50%) suggests that this trait's genetic architecture may be highly population-specific and sensitive to both environmental context and measurement methodology. These findings highlight the complexity of parsing genetic from environmental contributions to microbiome diversity and suggest that heritability estimates may be more population-dependent than previously appreciated. Future work will aim to refine the heritability model by incorporating dominance effects and extending the analysis to beta diversity measures and specific microbial taxa, to better characterize the genetic and environmental contributions shaping the human gut microbiome in this unique cohort.

5.5. Strengths and limitations

This study benefits from several strengths: a large sample size, deep metagenomic profiling, high-quality genotypes, and rich phenotypic data. The use of linear mixed models accounting for kinship, the inclusion of multiple covariate models, and the application of both GWAS and post-GWAS analyses (colocalization, MR) enhance the robustness of the

conclusions. However, certain limitations must be acknowledged. First, the power to detect genetic associations remains limited for low-prevalence taxa and rare variants. Second, while WMGS provides detailed resolution, it may miss functionally relevant strains not captured in current reference databases. Third, the cross-sectional study design, with microbiome profiling performed at a single time point per individual, does not capture intra-individual temporal variability and limits the ability to distinguish stable host-associated microbial features from transient fluctuations. Fourth, although WMGS allows inference of the functional potential encoded in microbial genomes, it does not directly inform on which metabolic pathways are transcriptionally or translationally active in vivo, nor on the metabolites ultimately produced. Lastly, while our Mendelian Randomization results are intriguing, they rely on the assumptions of the method and should be interpreted with caution.

5.6. The importance of covariate adjustment in metagenomic GWAS

The exploratory analyses performed with the inclusion of additional covariates emphasize the critical importance of an appropriate covariate adjustment strategy in microbiome GWAS.
Unlike classical genetic studies on homogeneous traits, metagenomic features are strongly influenced by age, sex, and lifestyle factors such as diet, alcohol consumption, smoking, and medication use.

Failure to account for these sources of variability can lead to biased effect estimates or even to spurious genetic associations.

While all primary analyses in this study were adjusted for Age, Age², and Sex, covariates consistently used in microbiome GWAS, we demonstrated that the inclusion of additional factors (e.g., alcohol intake, diabetes, BMI, PPI use) can significantly alter association signals.

This observation underlines the need for comprehensive covariate modelling, especially in studies where host genetics and environmental exposures jointly shape microbial composition.

5.7. Reproducibility of metagenomic GWAS findings

The issue of reproducibility represents one of the main challenges in microbiome–genome research. Despite increasing sample sizes and methodological standardization, results from metagenomic GWAS remain only partially consistent across studies. In most cases, genome-wide significant associations identified in one cohort are not replicated in others, even when similar statistical models, normalization procedures, and covariate adjustments are applied. This limited reproducibility likely reflects a combination of biological, environmental, and methodological factors. Differences in population structure, ancestry, diet, medication use, and other lifestyle-related exposures can strongly influence microbial composition and consequently affect the detection of host–microbiome genetic associations. Moreover, technical aspects such as sequencing

depth, microbial reference databases, taxonomic resolution, and the treatment of zero-inflated abundances introduce further heterogeneity in the results.

Comparing our findings with other published studies confirmed this general trend [35], [39], [41], [86], [87]. While some of the loci reported in previous metagenomic GWAS showed associations in similar genomic regions, most signals were not reproduced with comparable significance. Among the few loci that consistently emerged across independent cohorts, the region near the LCT gene represents the most reproducible example, being repeatedly associated with microbial taxa involved in lactose metabolism. However, in our analysis, this locus did not reach the same significance.

A plausible explanation for this attenuated signal is related to variation in lactase persistence among individuals. Because lactase persistence (LP) and lactose intolerance (LI) influence dietary habits, with lactose-intolerant individuals typically consuming fewer dairy products, differences in lactose intake can strongly affect the abundance of lactose-metabolizing taxa such as *Bifidobacterium*. As a result, the strength of the association at the LCT locus is expected to depend on population-specific dietary patterns.

In addition, the minor allele frequency (MAF) of the LCT variant (2:136616754_C_T) in our Sardinian cohort was approximately 0.056, markedly lower than that reported in most European populations (typically around 0.30–0.45). This lower frequency likely reduced the statistical

power to detect significant associations, further contributing to the weaker signal observed in our analysis.

Altogether, these findings reinforce the idea that host–microbiome genetic associations are highly context-dependent. Achieving robust and replicable results will require large-scale meta-analyses, harmonized analytical pipelines, and standardized definitions of microbial traits to disentangle genuine genetic effects from cohort- or environment-specific influences.

5.8. Conclusions and future directions

Our findings provide evidence that host genetics, environmental exposures, and shared living environments all contribute to shaping gut microbiome composition. The integration of large-scale metagenomic data with advanced statistical models enables a more nuanced dissection of this complex interplay, revealing that genetic effects on the microbiome are generally modest and strongly context-dependent.

Future research should focus on expanding sample sizes, improving taxonomic and functional resolution (e.g., through strain-level analysis), and integrating longitudinal and interventional data to more robustly address causality. In this context, Mendelian Randomization (MR) analyses represent a powerful next step to assess potential causal relationships between host genetic variants and microbiome traits, helping to move beyond correlation and to identify directional effects. The ultimate goal is to pinpoint

microbiome features that may serve as reliable biomarkers or therapeutic targets for precision medicine.

Beyond genetic inference, an important direction for future studies lies in the integration of complementary multi-omics layers. As sequencing technologies continue to evolve and costs decrease, combining metagenomics with metabolomics, metatranscriptomics, and metatranslatomics will allow a more comprehensive characterization of microbiome function. In particular, emerging approaches such as metaRibo-seq offer the opportunity to investigate microbial translational activity and to identify which metabolic processes are actively prioritized by microbes within the host environment, thereby bridging the gap between microbial composition, functional potential, and biological activity.

Finally, future studies should systematically incorporate a comprehensive set of covariates to refine signal interpretation and to disentangle true host–microbiome genetic effects from associations driven by environmental exposures, lifestyle, and shared living conditions. Such integrative modeling will be essential to improve reproducibility across populations and to ensure robust biological interpretation of microbiome–host genetic associations.

Bibliography

[1] G. A. Ogunrinola, J. O. Oyewale, O. O. Oshamika, and G. I. Olasehinde, 'The Human Microbiome and Its Impacts on Health', *Int. J. Microbiol.*, vol. 2020, pp. 1–7, June 2020, doi: 10.1155/2020/8045646.

[2] B. Wang, M. Yao, L. Lv, Z. Ling, and L. Li, 'The Human Microbiota in Health and Disease', *Engineering*, vol. 3, no. 1, pp. 71–82, Feb. 2017, doi: 10.1016/J.ENG.2017.01.008.

[3] J. R. Marchesi *et al.*, 'The gut microbiota and host health: a new clinical frontier', *Gut*, vol. 65, no. 2, pp. 330–339, Feb. 2016, doi: 10.1136/gutjnl-2015-309990.

[4] E. Z. Gomaa, 'Human gut microbiota/microbiome in health and diseases: a review', *Antonie Van Leeuwenhoek*, vol. 113, no. 12, pp. 2019–2040, Dec. 2020, doi: 10.1007/s10482-020-01474-7.

[5] E. Thursby and N. Juge, 'Introduction to the human gut microbiota', *Biochem. J.*, vol. 474, no. 11, pp. 1823–1836, May 2017, doi: 10.1042/BCJ20160510.

[6] A. M. Valdes, J. Walter, E. Segal, and T. D. Spector, 'Role of the gut microbiota in nutrition and health', *BMJ*, p. k2179, June 2018, doi: 10.1136/bmj.k2179.

[7] T. Takiishi, C. I. M. Fenero, and N. O. S. Câmara, 'Intestinal barrier and gut microbiota: Shaping our immune responses throughout life', *Tissue Barriers*, vol. 5, no. 4, p. e1373208, Sept. 2017, doi: 10.1080/21688370.2017.1373208.

[8] M. Vancamelbeke and S. Vermeire, ‘The intestinal barrier: a fundamental role in health and disease’, *Expert Rev. Gastroenterol. Hepatol.*, vol. 11, no. 9, pp. 821–834, Sept. 2017, doi: 10.1080/17474124.2017.1343143.

[9] N. Di Tommaso, A. Gasbarrini, and F. R. Ponziani, ‘Intestinal Barrier in Human Health and Disease’, *Int. J. Environ. Res. Public. Health*, vol. 18, no. 23, p. 12836, Jan. 2021, doi: 10.3390/ijerph182312836.

[10] B. P. G. L. and B. S. of Gastroenterology, ‘Correction: Mucus barrier, mucins and gut microbiota: the expected slimy partners?’, *Gut*, vol. 72, no. 12, pp. e7–e7, Dec. 2023, doi: 10.1136/gutjnl-2020-322260corr1.

[11] B. J. Didriksen, E. M. Eshleman, and T. Alenghat, ‘Epithelial regulation of microbiota-immune cell dynamics’, *Mucosal Immunol.*, vol. 17, no. 2, pp. 303–313, Apr. 2024, doi: 10.1016/j.mucimm.2024.02.008.

[12] Y. Yao, W. Shang, L. Bao, Z. Peng, and C. Wu, ‘Epithelial-immune cell crosstalk for intestinal barrier homeostasis’, *Eur. J. Immunol.*, vol. 54, no. 6, p. e2350631, June 2024, doi: 10.1002/eji.202350631.

[13] C. Chelakkot, J. Ghim, and S. H. Ryu, ‘Mechanisms regulating intestinal barrier integrity and its pathological implications’, *Exp. Mol. Med.*, vol. 50, no. 8, pp. 1–9, Aug. 2018, doi: 10.1038/s12276-018-0126-x.

[14] K. A. Dunleavy, L. E. Raffals, and M. Camilleri, ‘Intestinal Barrier Dysfunction in Inflammatory Bowel Disease: Underpinning Pathogenesis and Therapeutics’, *Dig. Dis. Sci.*, vol. 68, no. 12, pp. 4306–4320, Dec. 2023, doi: 10.1007/s10620-023-08122-w.

[15] Y. Belkaid and T. Hand, ‘Role of the Microbiota in Immunity and inflammation’, *Cell*, vol. 157, no. 1, pp. 121–141, Mar. 2014, doi: 10.1016/j.cell.2014.03.011.

[16] W. Yang and Y. Cong, ‘Gut microbiota-derived metabolites in the regulation of host immune responses and immune-related inflammatory diseases’, *Cell. Mol. Immunol.*, vol. 18, no. 4, pp. 866–877, Apr. 2021, doi: 10.1038/s41423-021-00661-4.

[17] D. Zheng, T. Liwinski, and E. Elinav, ‘Interaction between microbiota and immunity in health and disease’, *Cell Res.*, vol. 30, no. 6, pp. 492–506, June 2020, doi: 10.1038/s41422-020-0332-7.

[18] S. P. Wiertsema, J. van Bergenhenegouwen, J. Garssen, and L. M. J. Knippels, ‘The Interplay between the Gut Microbiome and the Immune System in the Context of Infectious Diseases throughout Life and the Role of Nutrition in Optimizing Treatment Strategies’, *Nutrients*, vol. 13, no. 3, p. 886, Mar. 2021, doi: 10.3390/nu13030886.

[19] T. Takeuchi, Y. Nakanishi, and H. Ohno, ‘Microbial Metabolites and Gut Immunology’, *Annu. Rev. Immunol.*, vol. 42, no. 1, pp. 153–178, June 2024, doi: 10.1146/annurev-immunol-090222-102035.

[20] M. G. Rooks and W. S. Garrett, ‘Gut microbiota, metabolites and host immunity’, *Nat. Rev. Immunol.*, vol. 16, no. 6, pp. 341–352, June 2016, doi: 10.1038/nri.2016.42.

[21] C. H. Kim, ‘Immune regulation by microbiome metabolites’, *Immunology*, vol. 154, no. 2, pp. 220–229, June 2018, doi: 10.1111/imm.12930.

[22] S. Ghosh, C. S. Whitley, B. Haribabu, and V. R. Jala, ‘Regulation of Intestinal Barrier Function by Microbial Metabolites’, *Cell. Mol. Gastroenterol. Hepatol.*, vol. 11, no. 5, pp. 1463–1482, Jan. 2021, doi: 10.1016/j.jcmgh.2021.02.007.

[23] R. Zhang, N. Ding, X. Feng, and W. Liao, ‘The gut microbiome, immune modulation, and cognitive decline: insights on the gut-brain axis’, *Front. Immunol.*, vol. 16, Jan. 2025, doi: 10.3389/fimmu.2025.1529958.

[24] Q. Cao *et al.*, ‘Elucidating the specific mechanisms of the gut-brain axis: the short-chain fatty acids-microglia pathway’, *J. Neuroinflammation*, vol. 22, no. 1, p. 133, May 2025, doi: 10.1186/s12974-025-03454-y.

[25] J. S. Loh *et al.*, ‘Microbiota–gut–brain axis and its therapeutic applications in neurodegenerative diseases’, *Signal Transduct. Target. Ther.*, vol. 9, no. 1, p. 37, Feb. 2024, doi: 10.1038/s41392-024-01743-1.

[26] F. Di Vincenzo, A. Del Gaudio, V. Petito, L. R. Lopetuso, and F. Scaldaferri, ‘Gut microbiota, intestinal permeability, and systemic inflammation: a narrative review’, *Intern. Emerg. Med.*, vol. 19, no. 2, pp. 275–293, Mar. 2024, doi: 10.1007/s11739-023-03374-w.

[27] M. Zhao *et al.*, ‘Immunological mechanisms of inflammatory diseases caused by gut microbiota dysbiosis: A review’, *Biomed. Pharmacother.*, vol. 164, p. 114985, Aug. 2023, doi: 10.1016/j.biopha.2023.114985.

[28] R. Burcelin, ‘Gut microbiota and immune crosstalk in metabolic disease’, *Mol. Metab.*, vol. 5, no. 9, pp. 771–781, Sept. 2016, doi: 10.1016/j.molmet.2016.05.016.

[29] K. A. Dunleavy, L. E. Raffals, and M. Camilleri, ‘Intestinal Barrier Dysfunction in Inflammatory Bowel Disease: Underpinning Pathogenesis and Therapeutics’, *Dig. Dis. Sci.*, vol. 68, no. 12, pp. 4306–4320, Dec. 2023, doi: 10.1007/s10620-023-08122-w.

[30] J. K. Goodrich *et al.*, ‘Human genetics shape the gut microbiome’, *Cell*, vol. 159, no. 4, pp. 789–799, Nov. 2014, doi: 10.1016/j.cell.2014.09.053.

[31] J. K. Goodrich *et al.*, ‘Genetic Determinants of the Gut Microbiome in UK Twins’, *Cell Host Microbe*, vol. 19, no. 5, pp. 731–743, May 2016, doi: 10.1016/j.chom.2016.04.017.

[32] J. L. Waters and R. E. Ley, ‘The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health’, *BMC Biol.*, vol. 17, no. 1, p. 83, Oct. 2019, doi: 10.1186/s12915-019-0699-4.

[33] L. Grieneisen *et al.*, ‘Gut microbiome heritability is near-universal but environmentally contingent’, *Science*, vol. 373, no. 6551, pp. 181–186, July 2021, doi: 10.1126/science.aba5483.

[34] R. Vilchez-Vargas *et al.*, ‘Gut microbial similarity in twins is driven by shared environment and aging’, *EBioMedicine*, vol. 79, p. 104011, Apr. 2022, doi: 10.1016/j.ebiom.2022.104011.

[35] A. Kurilshikov *et al.*, ‘Large-scale association analyses identify host factors influencing human gut microbiome composition’, *Nat. Genet.*, vol. 53, no. 2, pp. 156–165, Feb. 2021, doi: 10.1038/s41588-020-00763-1.

[36] P. J. Turnbaugh *et al.*, ‘A core gut microbiome in obese and lean twins’, *Nature*, vol. 457, no. 7228, pp. 480–484, Jan. 2009, doi: 10.1038/nature07540.

[37] R. Blekhman *et al.*, ‘Host genetic variation impacts microbiome composition across human body sites’, *Genome Biol.*, vol. 16, no. 1, p. 191, Sept. 2015, doi: 10.1186/s13059-015-0759-1.

[38] M. J. Bonder *et al.*, ‘The effect of host genetics on the gut microbiome’, *Nat. Genet.*, vol. 48, no. 11, pp. 1407–1412, Nov. 2016, doi: 10.1038/ng.3663.

[39] M. C. Rühlemann *et al.*, ‘Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome’, *Nat. Genet.*, vol. 53, no. 2, pp. 147–155, Feb. 2021, doi: 10.1038/s41588-020-00747-1.

[40] D. Rothschild *et al.*, ‘Environment dominates over host genetics in shaping human gut microbiota’, *Nature*, vol. 555, no. 7695, pp. 210–215, Mar. 2018, doi: 10.1038/nature25973.

[41] E. A. Lopera-Maya *et al.*, ‘Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project’, *Nat. Genet.*, vol. 54, no. 2, pp. 143–151, Feb. 2022, doi: 10.1038/s41588-021-00992-y.

[42] F. Beghini *et al.*, ‘Gut microbiome strain-sharing within isolated village social networks’, *Nature*, vol. 637, no. 8044, pp. 167–175, Jan. 2025, doi: 10.1038/s41586-024-08222-1.

[43] T. Yatsunenko *et al.*, ‘Human gut microbiome viewed across age and geography’, *Nature*, vol. 486, no. 7402, pp. 222–227, June 2012, doi: 10.1038/nature11053.

[44] T. Odamaki *et al.*, ‘Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study’, *BMC Microbiol.*, vol. 16, no. 1, p. 90, Dec. 2016, doi: 10.1186/s12866-016-0708-5.

[45] M. Valles-Colomer *et al.*, ‘The person-to-person transmission landscape of the gut and oral microbiomes’, *Nature*, vol. 614, no. 7946, pp. 125–135, Feb. 2023, doi: 10.1038/s41586-022-05620-1.

[46] E. C. Gotschlich, R. A. Colbert, and T. Gill, ‘Methods in microbiome research: Past, present, and future’, *Best Pract. Res. Clin. Rheumatol.*, vol. 33, no. 6, p. 101498, Dec. 2019, doi: 10.1016/j.berh.2020.101498.

[47] C. R. Woese and G. E. Fox, ‘Phylogenetic structure of the prokaryotic domain: The primary kingdoms’, *Proc. Natl. Acad. Sci.*, vol. 74, no. 11, pp. 5088–5090, Nov. 1977, doi: 10.1073/pnas.74.11.5088.

[48] A. M. Fricker, D. Podlesny, and W. F. Fricke, ‘What is new and relevant for sequencing-based microbiome research? A mini-review’, *J. Adv. Res.*, vol. 19, pp. 105–112, Sept. 2019, doi: 10.1016/j.jare.2019.03.006.

[49] A. M. Sidebottom, ‘A Brief History of Microbial Study and Techniques for Exploring the Gastrointestinal Microbiome’, *Clin. Colon Rectal Surg.*, vol. 36, no. 2, pp. 98–104, Jan. 2023, doi: 10.1055/s-0042-1760678.

[50] M. Sereika *et al.*, ‘Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing’,

Nat. Methods, vol. 19, no. 7, pp. 823–826, July 2022, doi: 10.1038/s41592-022-01539-7.

[51] C. Kim, M. Pongpanich, and T. Porntaveetus, 'Unraveling metagenomics through long-read sequencing: a comprehensive review', *J. Transl. Med.*, vol. 22, no. 1, p. 111, Jan. 2024, doi: 10.1186/s12967-024-04917-1.

[52] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, 'The Human Microbiome Project', *Nature*, vol. 449, no. 7164, pp. 804–810, Oct. 2007, doi: 10.1038/nature06244.

[53] L. M. Proctor *et al.*, 'The Integrative Human Microbiome Project', *Nature*, vol. 569, no. 7758, pp. 641–648, May 2019, doi: 10.1038/s41586-019-1238-8.

[54] G. Pilia *et al.*, 'Heritability of cardiovascular and personality traits in 6,148 Sardinians', *PLoS Genet.*, vol. 2, no. 8, p. e132, Aug. 2006, doi: 10.1371/journal.pgen.0020132.

[55] V. Orrù *et al.*, 'Genetic Variants Regulating Immune Cell Levels in Health and Disease', *Cell*, vol. 155, no. 1, pp. 242–256, Sept. 2013, doi: 10.1016/j.cell.2013.08.041.

[56] C. Sidore *et al.*, 'Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers', *Nat. Genet.*, vol. 47, no. 11, pp. 1272–1281, Nov. 2015, doi: 10.1038/ng.3368.

[57] M. Pala *et al.*, ‘Population- and individual-specific regulatory variation in Sardinia’, *Nat. Genet.*, vol. 49, no. 5, pp. 700–707, May 2017, doi: 10.1038/ng.3840.

[58] F. Krueger, F. James, P. Ewels, E. Afyounian, and B. Schuster-Boeckler, *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*. (July 23, 2021). Zenodo. doi: 10.5281/zenodo.5127899.

[59] B. Langmead and S. L. Salzberg, ‘Fast gapped-read alignment with Bowtie 2’, *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.

[60] F. Beghini *et al.*, ‘Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3’, *eLife*, vol. 10, p. e65088, May 2021, doi: 10.7554/eLife.65088.

[61] ‘Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4 | Nature Biotechnology’. Accessed: Oct. 02, 2025. [Online]. Available: https://www.nature.com/articles/s41587-023-01688-w

[62] J. Oksanen *et al.*, *vegan: Community Ecology Package*. (Oct. 08, 2025). Accessed: Oct. 12, 2025. [Online]. Available: https://cloud.r-project.org/web/packages/vegan/index.html

[63] T. M. Therneau, *coxme: Mixed Effects Cox Models*. (Aug. 23, 2024). Accessed: Oct. 10, 2025. [Online]. Available: https://cran.r-project.org/web/packages/coxme/index.html

[64] X. Zhou and M. Stephens, ‘Genome-wide Efficient Mixed Model Analysis for Association Studies’, *Nat. Genet.*, vol. 44, no. 7, pp. 821–824, June 2012, doi: 10.1038/ng.2310.

[65] C. Giambartolomei *et al.*, ‘Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics’, *PLOS Genet.*, vol. 10, no. 5, p. e1004383, May 2014, doi: 10.1371/journal.pgen.1004383.

[66] J. MacArthur *et al.*, ‘The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)’, *Nucleic Acids Res.*, vol. 45, no. D1, pp. D896–D901, Jan. 2017, doi: 10.1093/nar/gkw1133.

[67] C. Bycroft *et al.*, ‘The UK Biobank resource with deep phenotyping and genomic data’, *Nature*, vol. 562, no. 7726, pp. 203–209, Oct. 2018, doi: 10.1038/s41586-018-0579-z.

[68] V. Orrù *et al.*, ‘Complex genetic signatures in immune cells underlie autoimmunity and inform therapy’, *Nat. Genet.*, vol. 52, no. 10, pp. 1036–1045, Oct. 2020, doi: 10.1038/s41588-020-0684-4.

[69] GTEx Consortium *et al.*, ‘Genetic effects on gene expression across human tissues’, *Nature*, vol. 550, no. 7675, pp. 204–213, Oct. 2017, doi: 10.1038/nature24277.

[70] L. Chen *et al.*, ‘Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells’, *Cell*, vol. 167, no. 5, pp. 1398-1414.e24, Nov. 2016, doi: 10.1016/j.cell.2016.10.026.

[71] U. Võsa *et al.*, ‘Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene

expression', *Nat. Genet.*, vol. 53, no. 9, pp. 1300–1310, Sept. 2021, doi: 10.1038/s41588-021-00913-z.

[72] M. Ota *et al.*, 'Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases', *Cell*, vol. 184, no. 11, pp. 3006-3021.e17, May 2021, doi: 10.1016/j.cell.2021.03.056.

[73] L. Almasy and J. Blangero, 'Multipoint quantitative-trait linkage analysis in general pedigrees', *Am. J. Hum. Genet.*, vol. 62, no. 5, pp. 1198–1211, May 1998, doi: 10.1086/301844.

[74] F. Asnicar, G. Weingart, T. L. Tickle, C. Huttenhower, and N. Segata, 'Compact graphical representation of phylogenetic data and metadata with GraPhlAn', *PeerJ*, vol. 3, p. e1029, June 2015, doi: 10.7717/peerj.1029.

[75] International Multiple Sclerosis Genetics Consortium, 'Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility', *Science*, vol. 365, no. 6460, p. eaav7188, Sept. 2019, doi: 10.1126/science.aav7188.

[76] M. Steri *et al.*, 'Overexpression of the Cytokine BAFF and Autoimmunity Risk', *N. Engl. J. Med.*, vol. 376, no. 17, pp. 1615–1626, Apr. 2017, doi: 10.1056/NEJMoa1610528.

[77] D. J. Burgess, 'Fine-mapping causal variants — why finding "the one" can be futile', *Nat. Rev. Genet.*, vol. 23, no. 5, pp. 261–261, May 2022, doi: 10.1038/s41576-022-00484-7.

[78] C. Giambartolomei *et al.*, 'Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics', *PLOS*

Genet., vol. 10, no. 5, p. e1004383, May 2014, doi: 10.1371/journal.pgen.1004383.

[79] S. Purcell *et al.*, 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sept. 2007, doi: 10.1086/519795.

[80] S. Qiao *et al.*, 'Gut Parabacteroides merdae protects against cardiovascular damage by enhancing branched-chain amino acid catabolism', *Nat. Metab.*, vol. 4, no. 10, pp. 1271–1286, Oct. 2022, doi: 10.1038/s42255-022-00649-y.

[81] X. Hu *et al.*, 'Multi-omics study reveals that statin therapy is associated with restoration of gut microbiota homeostasis and improvement in outcomes in patients with acute coronary syndrome', *Theranostics*, vol. 11, no. 12, pp. 5778–5793, Mar. 2021, doi: 10.7150/thno.55946.

[82] Y. Zhang, J. Xu, X. Wang, X. Ren, and Y. Liu, 'Changes of intestinal bacterial microbiota in coronary heart disease complicated with nonalcoholic fatty liver disease', *BMC Genomics*, vol. 20, no. 1, p. 862, Nov. 2019, doi: 10.1186/s12864-019-6251-7.

[83] Q. Yan *et al.*, 'Alterations of the Gut Microbiome in Hypertension', *Front. Cell. Infect. Microbiol.*, vol. 7, Aug. 2017, doi: 10.3389/fcimb.2017.00381.

[84] L. Jin *et al.*, 'Gut microbes in cardiovascular diseases and their potential therapeutic applications', *Protein Cell*, vol. 12, no. 5, pp. 346–359, May 2021, doi: 10.1007/s13238-020-00785-9.

[85] E. R. Davenport, D. A. Cusanovich, K. Michelini, L. B. Barreiro, C. Ober, and Y. Gilad, ‘Genome-Wide Association Studies of the Human Gut Microbiota’, *PLoS ONE*, vol. 10, no. 11, p. e0140301, Nov. 2015, doi: 10.1371/journal.pone.0140301.

[86] Y. Qin *et al.*, ‘Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort’, *Nat. Genet.*, vol. 54, no. 2, pp. 134–142, Feb. 2022, doi: 10.1038/s41588-021-00991-z.

[87] X. Liu *et al.*, ‘A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases’, *Cell Discov.*, vol. 7, no. 1, pp. 1–15, Feb. 2021, doi: 10.1038/s41421-020-00239-w.

Acknowledgments

I would like to express my sincere gratitude to my tutor, Prof. Francesco Cucca, for the opportunity to undertake this PhD project and to my co-tutor, Dr. Mauro Pala, for his continuous help, advice, and support during this journey.

I am also grateful to Dr. Federico Santoni, who welcomed me as my supervisor during my year in Switzerland and supported me throughout the entire PhD period.

My appreciation extends to my colleagues at the CNR Institute of Genetic and Biomedical Research (IRGB) for their collaboration, and to all the volunteers and researchers involved in the SardiNIA project, whose contribution made this study possible.

Finally, I would like to thank my family and friends for their constant support and understanding throughout these years.

La borsa di dottorato è stata cofinanziata con risorse del Piano Nazionale di Ripresa e Resilienza presentato alla Commissione europea ai sensi dell'art. 18 e seguenti del Regolamento (UE) 2021/241" Missione 4 ("Istruzione e ricerca") – Componente 1 ("Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università"), Investimento 4.1 ("Estensione del numero di dottorati di ricerca e dottorati innovativi per la pubblica amministrazione e il patrimonio culturale") finalizzate al sostegno di borse per dottorati di ricerca PNRR, per dottorati per la Pubblica Amministrazione e per dottorati per il patrimonio culturale