RESEARCH

Open Access

Functional archetypes in the human gut microbiome reveal metabolic diversity, stability, and influence disease-associated signatures

Mohamed Meawad[1], Dalwinder Singh[1], Alice Deng[1,2], Rohan Sonthalia[1], Evelyn Cai[1,3] and Vanessa Dumeaux[1,2,4,5*]

Abstract

Background Understanding the functional diversity of the gut microbiome is critical for elucidating its roles in human health and disease. While traditional approaches focus on taxonomic composition, functional configurations of the microbiome remain understudied. This study introduces a deep-learning framework combined with archetypal analysis to identify and characterize functional archetypes in the adult human gut microbiome. This approach aims to provide insights into interindividual variability, function-driven microbiome stability, and the potential confounding role of functional diversity in disease-associated microbial signatures.

Results Analyzing 9838 whole-genome metagenomic samples from healthy adults across 29 countries, we identified three distinct functional archetypes that define the boundaries of the gut microbiome's functional space. Each archetype is characterized by unique metabolic potentials: Archetype 1 is enriched in sugar metabolism, branched-chain amino acid biosynthesis, and cell wall synthesis; Archetype 2 is dominated by fatty acid metabolism and TCA cycle pathways; and Archetype 3 is defined by amino acid and nitrogen metabolism. While most gut microbiome communities are a blend of these archetypes, some align closely with a single archetype, potentially reflecting adaptation to host factors such as distinct dietary patterns. Proximity to these archetypes correlates with microbiome stability, with Archetype 2 representing the most resilient state, likely due to its metabolic flexibility and diversity.

Functional archetypes emerged as a potential confounder in disease-associated microbial signatures, including in type-2 diabetes, colorectal cancer, and inflammatory bowel disease (IBD). In IBD, archetype-specific shifts were observed: Archetype 1-dominant samples exhibited increased carbohydrate metabolism, while Archetype 3-dominant samples showed enrichment in inflammatory pathways. These findings highlight the potential for archetype-specific functional changes to inform microbiome-targeted interventions.

Conclusions The identified functional archetypes provide a robust framework for addressing interindividual variability and potential confounding in gut microbiome-based disease studies. By incorporating archetypes as potential confounders or stratification factors, researchers can reduce variability, uncover novel pathways, and improve the precision of microbiome-targeted interventions. The deep-learning framework can be applied to other host-associated microbial ecosystems, providing new insights into microbial functional dynamics and their implications for the host's health.

*Correspondence:
Vanessa Dumeaux
vdumeaux@uwo.ca
Full list of author information is available at the end of the article

Keywords Microbiome, Whole-genome metagenomics, Functional profiling, Functional archetypes, Deep learning, Gut microbiome, Health, Disease confounding, Inflammatory bowel disease

Background

Host-adapted microbial communities and their collective genomes, termed microbiomes, are shaped by host anatomy and physiology, inhabiting diverse ecological niches across the body [1–3]. Within these niches, microbiomes engage in dense and dynamic interactions—ranging from mutualism to competition—which critically influence their composition, functionality, stability, and resilience [4–9].

Microbiomes exhibit variability at multiple scales, including differences between host species, body sites, and individuals [1, 10, 11]. This variability poses significant challenges for experimental design and clinical interpretation, hindering reproducibility and limiting biological discovery [12]. Furthermore, identifying the magnitude and patterns of these variations is essential for understanding microbiome dynamics and their functional consequences.

In the human gut, the microbiome undergoes dramatic changes from infancy to adulthood, ultimately assembling into a relatively stable "climax community" [10, 13]. Despite broad interindividual variation, healthy adult gut microbiota have been broadly categorized into three "enterotypes" dominated by *Bacteroides*, *Prevotella*, or *Ruminococcus* [14–17]. These compositional configurations are associated with host factors such as diet, health, and metabolic characteristics [18, 19]. Similar patterns, which we term Microbial Configurations (MCs), have been identified among non-bacterial members of the gut microbiome, in other human body sites and across animal species, underscoring the generality of this phenomenon [1, 20–27].

Early efforts to identify such compositional MCs relied on simple clustering techniques, but these methods often oversimplify the complexity of microbial ecosystems. Continuous or'quasi-discrete'gradients have since been proposed as more accurate representations of microbial variation [18, 28–31]. While the recognition of quasi-discrete compositional MCs and their associated factors represents an important step forward, it raises critical questions about their origins and implications. Are these configurations primarily driven by community composition, or do they reflect functional characteristics that transcend taxonomic variation? Moreover, it remains unclear whether compositional MCs sufficiently capture overall community function or if functional MCs diverge from compositional patterns altogether.

These questions are fundamental to not only understand microbiome inter-individual variability and function but also underlying parameters defining their stability and resilience. Functional redundancy—where multiple species within a community perform overlapping roles—has emerged as a key mechanism underlying microbiome resilience, ensuring the preservation of essential functions despite compositional fluctuations [32–34]. Early work showed that metabolic module redundancy varies significantly across enterotypes and correlates with ecosystem stability [33]. Furthermore, metabolic independence and resilience in stressed gut environments have been identified as crucial factors for maintaining host health [35].

Recent work using nonnegative matrix factorization (NMF) successfully identified four functional profiles for fiber and mucin metabolism in the human gut microbiome by focusing on curated metabolic processes and highlighted their potential as markers of diet, dysbiosis, inflammation, and disease [36]. These insights demonstrate the value and need for analytical frameworks that can capture the functional dynamics of microbial ecosystems. While broader pathway databases like MetaCyc provide comprehensive coverage of metabolic functions with extensive literature curation, the challenge remains in extracting interpretable patterns from these complex, high-dimensional datasets.

We explore Archetypal Analysis (AA), which decomposes functional profiles into mixtures of extreme patterns found within the dataset. Unlike NMF, AA requires that these patterns correspond to real samples (archetypes) and that their contributions sum to 1, ensuring biological interpretability. By extending AA with a deep-learning framework [37–40], we capture non-linear relationships in our data. It is reasonable to expect functional interactions within complex microbial communities to be non-linear, and non-linear methods have previously been successfully applied to study compositional enterotypes [39, 41]. Using this framework, we define the functional variability of global healthy adult gut microbiomes and investigate the interplay between composition, function, and stability. This approach addresses key challenges in microbiome research, including inter-individual variability and potential confounding, providing robust insights into disease-associated functional shifts. Beyond the gut microbiome, this framework establishes a foundation for studying microbial ecosystems in diverse contexts, with broad implications for understanding microbial community dynamics and developing microbiome-based interventions targeting specific functions.

Methods

Compendium data curation, preprocessing, transformation, and filtering

We curated 11,309 publicly available whole genome sequencing (WGS) samples' data from healthy adult (age ≥ 18) stool samples using several large dataset repositories such as GMrepo [42] and curatedMetagenomicData R package [43] as well as datasets under controlled access, namely LifeLines-DEEP [44] and Milieu Interieur [45] (Fig. 1A, Table 1, Table S1).

Each sample underwent processing with a computational pipeline designed to use the latest genome annotations and minimize false positives, converting raw reads into relative pathway abundances (Fig. S1). First, the raw reads were preprocessed for quality control and adapter removal using fastp [47] with the following parameters: `--trim_poly_x --trim_poly_g -p --length_required 40 --cut_front --cut_tail --cut_mean_quality 25`. Based on our experience and corroborated by previous studies [48–50], when using properly defined thresholds and a comprehensive reference database, the Kraken2 + Bracken [51, 52] toolset delivers superior performance to estimate microbial composition of gut microbiomes. We therefore used Kraken2 (v2.1.3) with a confidence threshold of 0.15, which has been shown to minimize both false positives and false negatives [51], and Bracken (v2.8) to align reads to the HumGut database [53] (downloaded on 11/2021) and the human genome downloaded from NCBI to identify and remove contamination.

Biochemical pathways/functions were then identified using HUMAnN3 v3.7 [54] based on species identified using Kraken2 + Bracken and HUMAnN3 default reference databases including the Chocophlan (v201901_v31), UniRef90 (EC-filtered 201901b), and MetaCyc (release 24.0) (downloaded on 11/2021). HUMAnN3 quantified pathways in units of RPKs (reads per kilobase) and counts were further normalized where read counts for each sample are constrained to sum to 1 million (copies per millions, CoPM).

Samples were filtered if their total pathway abundance (sum of reads per kilobase across all pathways) was outside the range of 100,000–2,000,000. We removed rare pathways that were detected in fewer than 10% of samples and excluded studies with fewer than 30 samples. A total of 9,838 samples from 48 studies across 29 countries were left after filtering (library size mean: 40.5 M; range 5.1 M—183.5 M) (Fig. 1B, C, Fig. S2A, B).

As expected, we observed a significant batch effect associated with sample study source (Fig. S2C),and applied batch correction using ComBat-seq [55] (Fig. S2D). To assess the effectiveness of the batch correction, we examined the second-largest study in the dataset (LifeLines DEEP, $n = 1131$), which exhibited the most pronounced batch effect. After correction, the batch effect associated with the study source was no longer observed (Fig. S2C, D, Fig. S3A, B), confirming the appropriateness of the ComBat -seq method. Details on the archetype assignments used in Fig. S3A, B are provided in the subsequent section. Finally, to ensure that this correction did not alter the distribution or sparsity of the functional profiles, a comparison between values in the original and batch-corrected matrices was conducted. Overall, the values in both datasets were overwhelmingly similar, and the sparsity (zero counts) of the data remained identical (Fig. S3C–E).

To determine the compositional enterotypes of our samples, we used two recent approaches: Enterosignatures [28] and Enterotyper [29]. For both methods, we submitted filtered and batch-corrected count data using their respective web applications. Enterosignatures required normalized genus-level counts, which we obtained by summing species counts within each genus and adjusting for samples' total counts.

Deep archetypal analysis

Following batch correction, non-linear archetypal analysis was performed by integrating archetypal analysis with a deep autoencoder [40] (Fig. S1). Prior to model training, the distribution of the count data was assessed to inform the parameterization of the decoder. Pathway abundance data (RPK) were modeled using the generalized additive model for location, scale, and shape (GAMLSS) as implemented in the `scDesign3` R package [56]. The Akaike Information Criterion (AIC) values were calculated for Poisson, zero-inflated Poisson (ZIP), Gaussian, negative binomial (NB), and zero-inflated negative binomial (ZINB) distributions to evaluate their fit to the data. NB and ZINB had the lowest AIC values (Fig. S4A, B), and simulated data using these distributions (Fig. S4C) produced UMAP visualizations with a structure and distribution qualitatively similar to the original dataset. Ultimately, we selected the NB distribution to parameterize our decoder, as it is less complex, supports ComBat-seq batch correction assumptions, and provides a comparable fit to the data than the ZINB distribution.

The scAAnet model was configured with a batch size of 64 and a hidden layer width of 128 (see Hyperparameter search section below). The activation function used was a rectified linear unit (ReLU). The model was

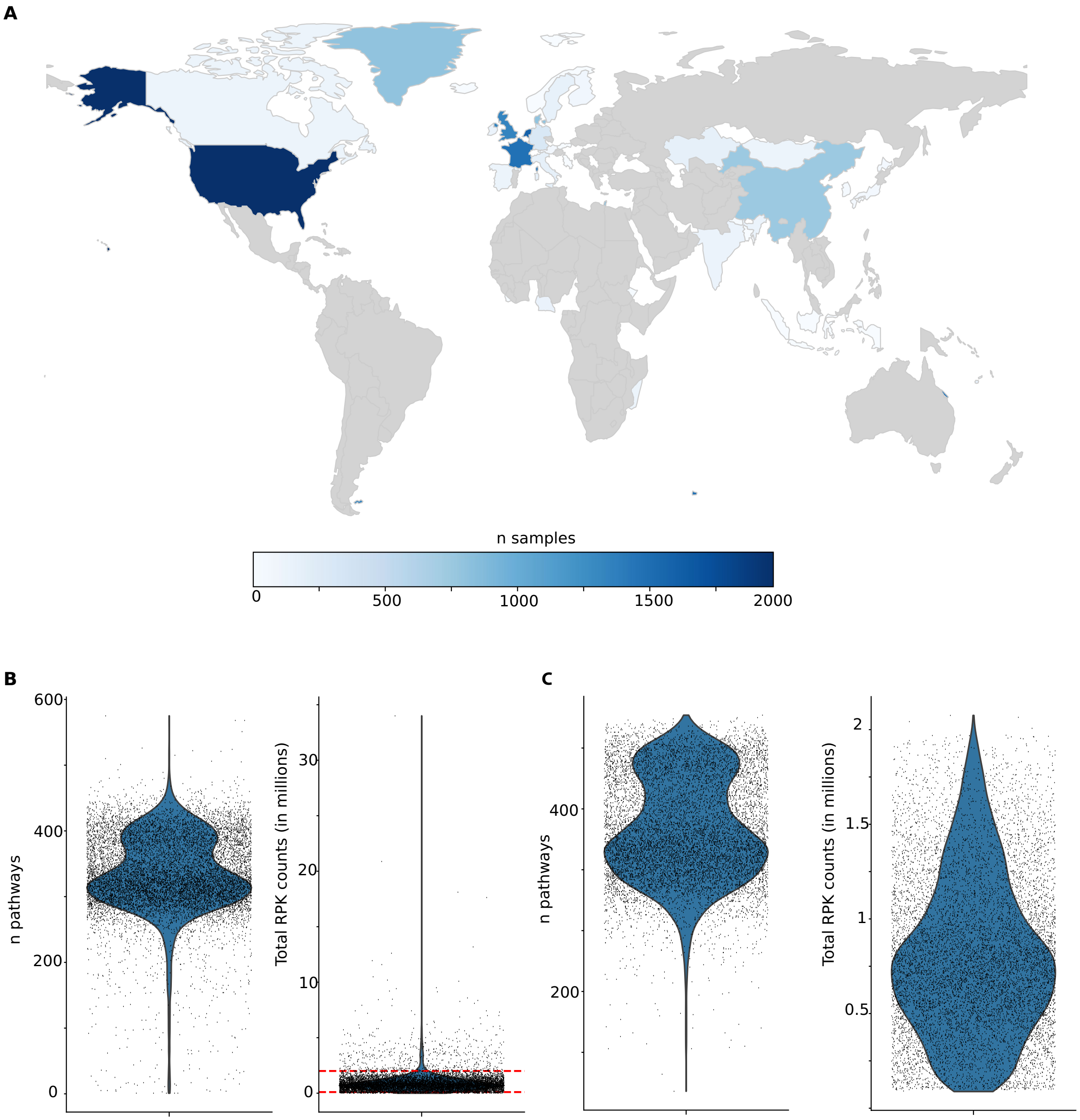


Fig. 1 A large compendium of healthy adult microbiomes for robust functional archetype identification. **A** A world map illustrating the distribution of microbiome samples across countries before filtering, with the number of samples per country indicated by color. Countries without available samples are shown in gray. **B**, **C** Violin plots displaying the distribution of the number of functional pathways (left) and total RPK counts (right) for **B** the raw functional pathway data and **C** the filtered dataset. Filtering criteria included retaining samples with a total RPK counts between 100,000 and 2,000,000, removing pathways present in fewer than 10% of samples, and excluding studies with fewer than 30 samples

trained for up to 1000 epochs (each epoch representing one full training cycle), with 20 warm-up epochs. Early stopping was enabled if the loss did not improve for 100 epochs, and the learning rate was reduced if the loss failed to improve for 10 epochs. The initial learning rate was set to 0.01.

Number of archetypes (*K* value) selection

Two metrics were used to determine the optimal number of archetypes (*K* value): archetypes' stability and

Table 1 Geographic distribution of curated human gut microbiome datasets. Summary of curated healthy adult gut microbiome datasets after quality filtering based on total RPK counts and cohort size. The table shows the number and percentage of profiles from the Global North and Global South regions for each data source

Data source	Global North *N* (%)	Global South *N* (%)
CuratedMetagenomicData[a]	5319 (82)	1137 (18)
GMrepo V2[a]	603 (79)	162 (21)
EGA (controlled-access)		
LifeLines DEEP [44]	1108 (100)	0 (0%)
Milieu Interieur [45]	1351 (100)	0 (0%)
PubMed		
PRJEB49206 [46]	0 (0)	158 (100)
Total	**8381 (85%)**	**1457 (15%)**

[a] Individual BioProject IDs are provided in Table S1

samples' archetype usage consistency. Archetype stability was assessed by performing K-means clustering on the archetype spectra derived from multiple random initializations ($n = 8$) and computing the Euclidean distance silhouette score to evaluate the quality and robustness of the inferred archetypes at each K value. This stability metric ranges from 0 to 1, with higher values indicating more robust archetypes.

To assess consistency in sample archetype usage, we identified the dominant archetype for each sample across multiple random states ($n = 8$). For each pair of samples, we calculated the proportion of times they were assigned to the same archetype, resulting in a consensus matrix C that indicates the probability of two samples being assigned to the same archetype. Using these pairwise probabilities as measures of similarity, we performed average linkage hierarchical clustering. We then calculated the cophenetic distance for each pair of samples based on the hierarchical clustering dendrogram, reflecting how similar two samples are within the dendrogram. The cophenetic correlation coefficient (ranging from 0 to 1) compares these cophenetic distances with the original assignment probabilities, providing an overall measure of clustering stability.

In our analyses, both stability metrics peaked at $K = 3$ (Fig. S5A, B), which was therefore selected for downstream analyses.

Determining the most representative random state

We identified the most representative state for each archetype ($K = 3$) from 100 random states using two key metrics: cosine similarity and archetype distinctiveness. Cosine similarity was used to assess how closely each state's archetypes aligned with the mean archetype configuration, ensuring consistency. Simultaneously, we measured the distinctiveness of archetypes by measuring the euclidean distances between the 3 archetypes in each random state, to uphold the core principle of archetypal analysis—identifying extreme points. Our final scoring formula combined similarity to the mean with half-weighted distinctiveness, striking a balance between avoiding outlier states and maintaining representativeness. The selected state is visualized in the K-means clustering plot (Fig. S6), highlighting its optimal fit for subsequent analyses.

Hyperparameter search

We performed a systematic search over the following hyperparameters:

- Number of archetypes (k): [3, 4, 5, 6, 7]
- Width of hidden layers: [32, 64, 128, 256, 512]
- Batch size: [32, 64, 128, 256, 512]
- Learning rate: [0.1, 0.01, 0.001]

The search was conducted using 8 random states initially, with extended validation using 150 random states for selected hyperparameter combinations. Our primary objective was to identify hyperparameter combinations that yield reproducible and distinct archetypes with stable sample assignments, assessed using the usage and archetype stability metrics described above.

Because many combinations achieved satisfactory sample usage stability (> 0.75), we focused on optimizing archetype stability and distinctness. One combination (batch size = 256, hidden layers = 512, learning rate = 0.01) achieved the highest stability (0.77 across 150 random states) but frequently produced redundant archetypes (31 out of 130 runs identified at least 2 archetypes belonging to the same cluster).

We therefore selected an alternative configuration (batch size = 64, hidden layers = 128, learning rate = 0.01) that achieved comparable stability (0.75 across 150 random states) with minimal redundancy (2/150 runs) and better computational efficiency. Notably, both hyperparameter sets produced very similar pathway archetype scores, with Spearman correlations of 0.96, 0.98, and 0.93 for archetypes 1, 2, and 3, respectively, demonstrating the robustness of the identified functional patterns regardless of specific hyperparameter values within a reasonable range.

Pathways' compound visualization using graphs

Network graphs were created to represent pathway interactions. The input data was sourced from the MetaCyc database [57] and consisted of edges annotated with pathway-specific metabolites. Nodes corresponding to compounds shared across multiple pathways were connected using dashed gray edges to highlight duplication and identify clusters of shared compounds. Common compounds, such as H+, phosphate, ATP, ADP, H2O, NADP+, NADPH, NADH, NAD+, CO2, coenzyme A, AMP, dioxygen, hydrogen carbonate, and diphosphate, were excluded to reduce visual noise. A small subset of pathways did not have graph data available and therefore were not included in the visualization; this was the case for 4 pathways in the top 20 defining archetype 2: PRPP-PWY: superpathway of histidine, purine, and pyrimidine biosynthesis; TCA-GLYOX-BYPASS: superpathway of glyoxylate bypass and TCA; TCA: TCA cycle I (prokaryotic); METSYN-PWY: superpathway of L-homoserine and L-methionine biosynthesis; however, they often had other highly similar pathways presented in the graph.

The finalized graphs were exported as HTML files, allowing interactive exploration.

Archetype stability analysis

To assess the stability of the archetypes, we filtered our dataset to identify healthy subjects with at least two unique samples, which yielded seven studies ($n = 656$ subjects, 1557 samples). These subjects had 2–6 visits over a 2–730 day span (Table S2). We subtracted the archetype values between consecutive visits to calculate absolute differences. For analyses based on archetype usage, samples were categorized as high usage (≥ 0.66) or medium usage (0.33–0.66).

Assigning archetypal scores in external datasets

A pre-trained model learned using our large compendium of healthy samples can be used to infer archetypal scores in external datasets.

Using this approach, we assessed whether the archetypal space could capture functional variation across gut microbiome profiles from diseased individuals. We curated metagenomic profiles from studies of patients with type 2 diabetes (T2D; 3 studies), colorectal cancer (CRC; 5 studies), and inflammatory bowel disease (IBD; 3 studies) (Table S3A). Raw fastq files were processed similarly using Kraken2/Bracken followed by HUMAnN3 while using the same reference databases and parameters as mentioned above. Pathways were further filtered to only include the 436 pathways used to train the model. Counts and metadata were then concatenated with the original dataset to the pre-batch correction count matrix and metadata files, respectively. Batch correction was then performed using ComBat-seq with batch = study. Finally, the trained model was applied to this dataset extended with these new profiles to map the disease-associated samples and their respective controls onto the archetypal space.

For external datasets, we recommend following the same preprocessing and batch correction steps described above. The robustness of archetypal scores across alternative preprocessing approaches using different taxonomic read classification tools and databases was tested (Table S4). Overall, we found only small differences in archetypal scores between the different preprocessing pipelines (Fig. S7) but the batch correction using ComBat-seq is required. Detailed scripts and files are available in the GitHub repository https://github.com/dumeaux-lab/deep-fMC_paper.

Differential analysis of disease samples using MaAsLin2

Using MaAsLin2 [58], we identified significant differences in pathway relative abundances within gut microbiomes between healthy and diseased individuals. All analyses were performed using the negative binomial model to address the compositional nature of the microbiome and account for overdispersion in the data, cumulative sum scaling normalization to adjust for library size, and Benjamini–Hochberg FDR to adjust for multiple testing, as implemented in MaAsLin2 R/Bioconductor (V1.18.0). In all analyses, subject ID and study were modeled as random effects to account for inter-individual variability and differences between datasets, respectively.

We conducted the following analyses:

1. Model 1 without archetype adjustment: Disease status (disease vs. healthy) was included as the fixed effect without any adjustment for archetypes.
2. Analyses with archetype adjustment: These models extended the Model 1 analysis by adding three fixed effects to adjust for samples' archetype scores (archetype1, archetype2, and archetype3)
3. Stratified analyses with archetype adjustment: Samples were stratified based on their dominant archetype (highest archetype value), and each stratified group was analyzed separately. For each group, the model included disease status as a fixed effect along with archetype scores for archetype1, archetype2, and archetype3.

Results

A comprehensive compendium of healthy adult gut microbiomes for robust functional archetype identification

We developed a comprehensive compendium to integrate metagenomic profiles from large dataset repositories and curated datasets [42–46] (Table 1; Table S1). This compendium represents the largest repository of whole-genome metagenomic (WGM) profiles from healthy adults, comprising 11,309 samples from 76 studies conducted across 31 countries, primarily from the Global North (Fig. 1A**,** Table 1). Each sample was processed using a computational pipeline designed to incorporate the latest genome annotations and minimize false positives, converting raw reads into relative pathway abundances (Fig. S1).

The final dataset includes 436 pathways, with samples expressing an average of 333 pathways (range, 118–427), across 9838 samples available for downstream analyses (Fig. S2A, B). We note that much of the mapping to functional databases likely derives from common housekeeping and well-characterized genes present in reference databases.

A significant batch effect associated with study origin was detected and corrected using ComBat-seq [55] (Fig. S2C, D). Importantly, no substantial batch effect was observed at the level of Global Region beyond the study of origin (Fig. S2E, F). Despite the compendium including only 15% of samples from the Global South, this finding suggests that, after adjustment for study effects, the results are likely generalizable to these populations.

Following batch correction, non-linear archetypal analysis revealed that each microbiome functional profile could be decomposed into a mixture of three archetypes, representing extreme profiles (Fig. S5-6). High stability values and consistent archetype patterns across random states confirmed the robustness of the three-archetype solution (Fig. S5-S6).

To visualize the relationship between samples and their archetypal compositions, we used multidimensional scaling (MDS), where the three archetypes defined the vertices of the solution space. The MDS projection showed that while some samples were closely aligned with pure archetypes (vertices), the majority exhibited varying contributions from multiple archetypes, indicating that most human gut microbiomes represent functional gradients rather than discrete states (Fig. 2A). A density gradient in archetype usage was observed, with an enrichment of samples showing mid to high usage of Archetypes 1 and 3 (Fig. 2B).

Interestingly, samples with a stronger affinity to Archetype 2 had a higher number of detected pathways but showed no corresponding increase in the number of identified species in compositional data (Fig. 2C, D). Differences in the distribution of Archetype 2 scores were observed across age groups, while Archetypes 1 and 3 showed differences across sexes (ANOVA or *t*-test, $p<0.001$; Fig. S8A). However, these differences may be partially influenced by variations in country or regional distributions (Fig. S8B, C), as sex and age are not equally represented across countries (Fig. S2A).

Gut microbiome functional states represent blends of three distinct metabolic archetypes

To characterize the functional profiles of each archetype, we identified the top 20 pathways that most strongly distinguished them (Fig. 2E**,** Table S5). Each archetype exhibited distinct metabolic signatures, highlighting their unique functional roles. To visualize how these signature pathways vary across samples, we mapped their abundance patterns relative to archetypal classifications (Fig. 2F). The heatmap revealed smooth transitions in pathway abundances across samples, consistent with their archetypal scores.

Archetype 1 is characterized by high potential for carbohydrate sugar metabolism and biosynthesis of branched-chain amino acid (BCAA) and microbial cell wall component biosynthesis (Fig. 2E**,** Table S5). The dTDP-β-L-rhamnose biosynthesis pathway showed maximal enrichment, accompanied by other carbohydrate-processing pathways, including glycolysis IV, glucose and

(See figure on next page.)

Fig. 2 Pathway usage and abundance patterns across archetypal states. **A** Multidimensional scaling (MDS) representation of samples based on their archetypal compositions. Each panel shows the same MDS coordinates, with color intensity indicating the relative contribution (usage) of each archetype (scale 0–1). The three archetypes (labeled 1–3) define the vertices of the solution space, and each sample's position reflects its mixture of archetypal contributions, which sum to 1. Left, middle, and right panels highlight the usage of Archetype 1, 2, and 3, respectively. **B**–**D** Same MDS coordinates as in **A**, with color intensity indicating **B** the sample density or the number of unique **(C)** functional pathways and **(D)** species present in each sample. **E** Heatmap showing the relative contribution (usage) of the top 20 pathways in defining each archetype identified by the deep archetypal analysis model. Color intensity indicates the degree of pathway usage (scale 0–1) for each archetype (type 1–3). **F** Heatmap displays pathway abundances across samples for the top 20 pathways defining each archetype. Values were scaled and winsorized at the 2nd and 98th percentiles (z-score). Samples are ordered by relative pathway abundance average ranksum plotted on top of the heatmap. Bottom annotations show archetypal scores and dominant archetype classification for each sample, where dominant type is determined by the highest score among the three archetypes

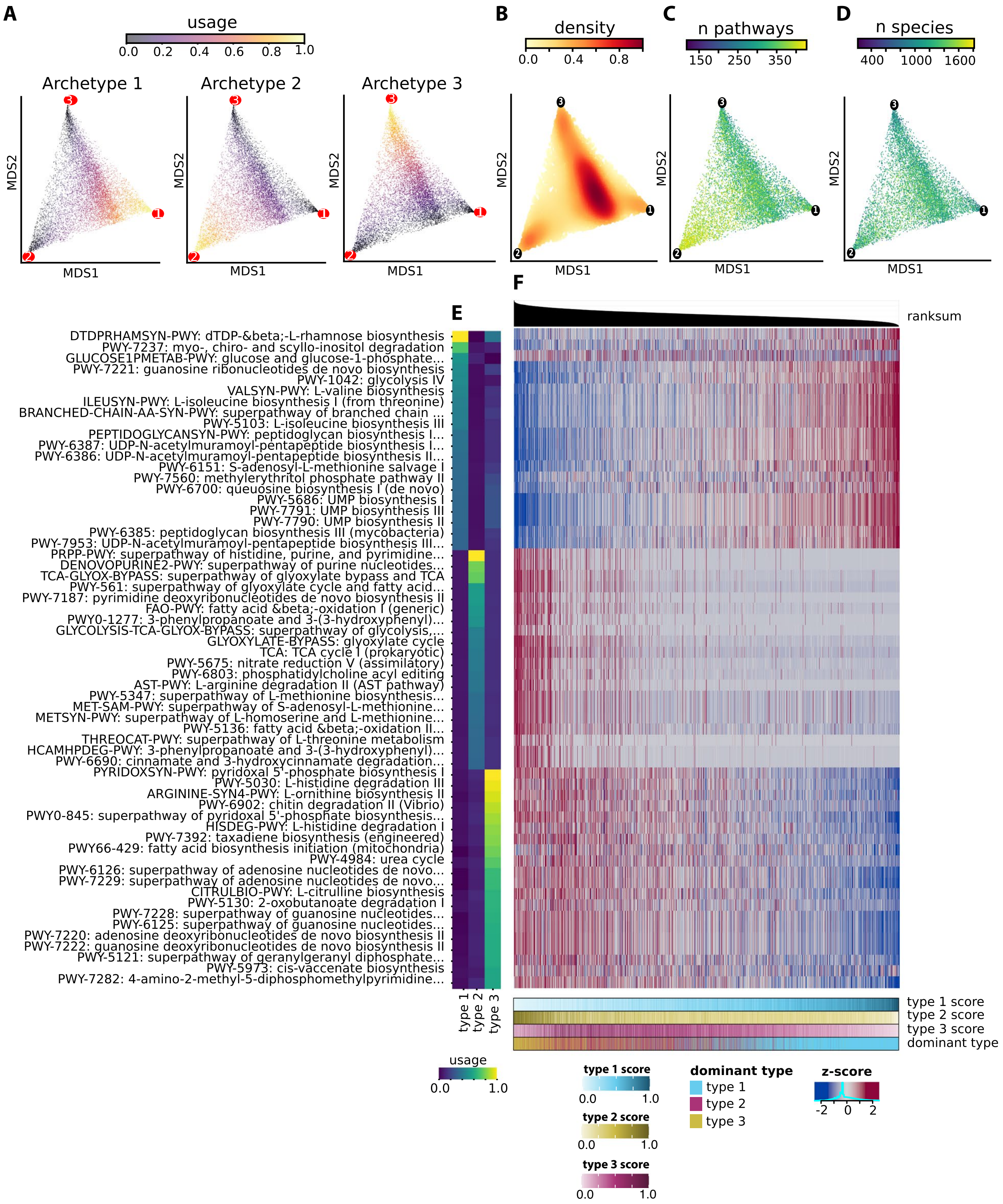


Fig. 2 (See legend on previous page.)

glucose-1-phosphate degradation, and myo-, chiro-, and scyllo-inositol degradation, which collectively constituted four of the top five pathways defining this archetype (Fig. 2E, Table S5). These sugar metabolism pathways supply key intermediates—such as D-glyceraldehyde 3-phosphate, pyruvate, and phosphoenolpyruvate—that

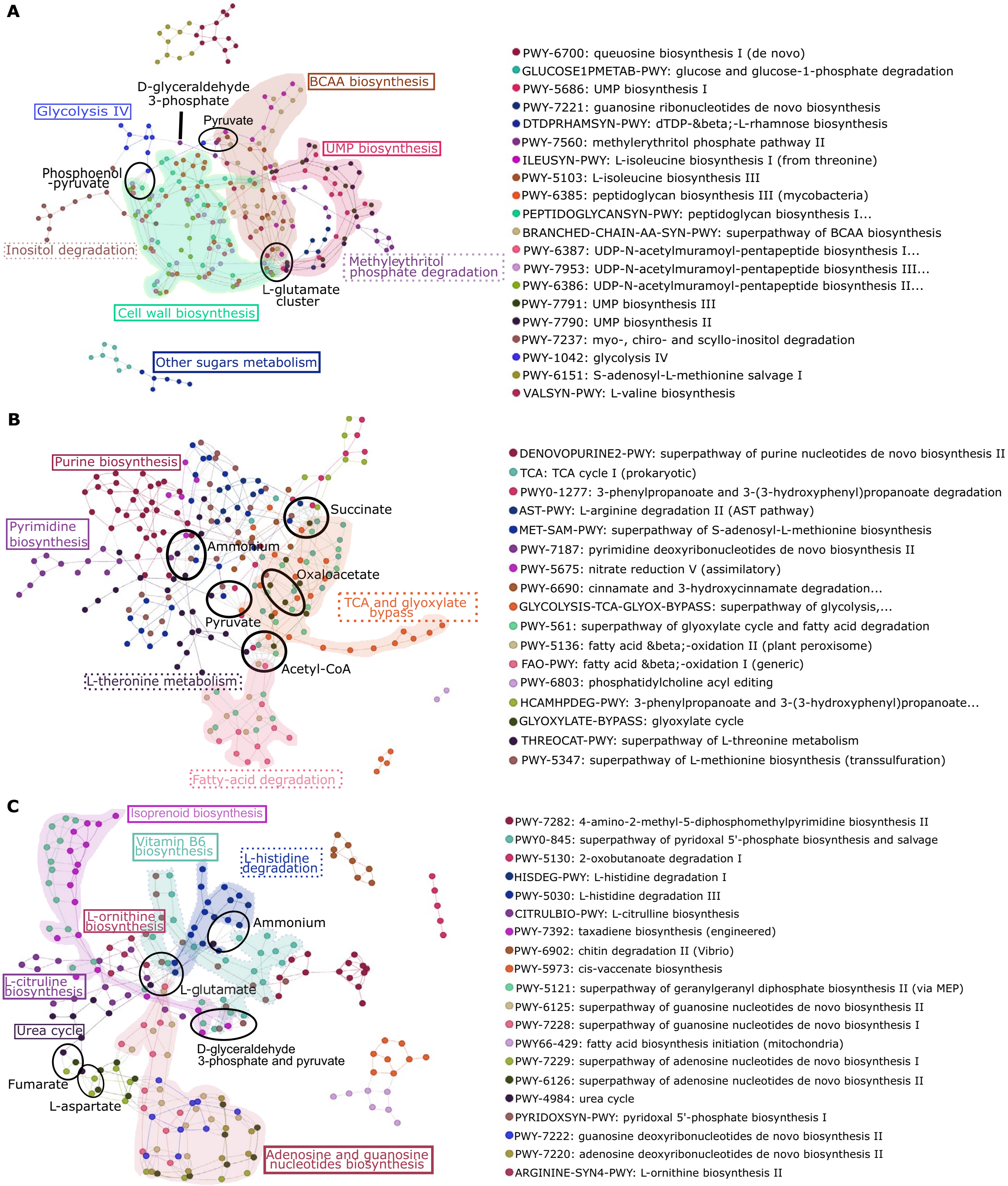

A
D-glyceraldehyde 3-phosphate
BCAA biosynthesis
Glycolysis IV
Pyruvate
UMP biosynthesis
Phosphoenol -pyruvate
Inositol degradation
Methylethritol phosphate degradation
L-glutamate cluster
Cell wall biosynthesis
Other sugars metabolism
PWY-6700: queuosine biosynthesis I (de novo)
GLUCOSE1PMETAB-PWY: glucose and glucose-1-phosphate degradation
PWY-5686: UMP biosynthesis I
PWY-7221: guanosine ribonucleotides de novo biosynthesis
DTDPRHAMSYN-PWY: dTDP-&beta;-L-rhamnose biosynthesis
PWY-7560: methylerythritol phosphate pathway II
ILEUSYN-PWY: L-isoleucine biosynthesis I (from threonine)
PWY-5103: L-isoleucine biosynthesis III
PWY-6385: peptidoglycan biosynthesis III (mycobacteria)
PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I...
BRANCHED-CHAIN-AA-SYN-PWY: superpathway of BCAA biosynthesis
PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I...
PWY-7953: UDP-N-acetylmuramoyl-pentapeptide biosynthesis III...
PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II...
PWY-7791: UMP biosynthesis III
PWY-7790: UMP biosynthesis II
PWY-7237: myo-, chiro- and scyllo-inositol degradation
PWY-1042: glycolysis IV
PWY-6151: S-adenosyl-L-methionine salvage I
VALSYN-PWY: L-valine biosynthesis
B
Purine biosynthesis
Succinate
Pyrimidine biosynthesis
Ammonium
Oxaloacetate
TCA and glyoxylate bypass
Pyruvate
Acetyl-CoA
L-theronine metabolism
Fatty-acid degradation
DENOVOPURINE2-PWY: superpathway of purine nucleotides de novo biosynthesis II
TCA: TCA cycle I (prokaryotic)
PWY0-1277: 3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate degradation
AST-PWY: L-arginine degradation II (AST pathway)
MET-SAM-PWY: superpathway of S-adenosyl-L-methionine biosynthesis
PWY-7187: pyrimidine deoxyribonucleotides de novo biosynthesis II
PWY-5675: nitrate reduction V (assimilatory)
PWY-6690: cinnamate and 3-hydroxycinnamate degradation...
GLYCOLYSIS-TCA-GLYOX-BYPASS: superpathway of glycolysis,...
PWY-561: superpathway of glyoxylate cycle and fatty acid degradation
PWY-5136: fatty acid &beta;-oxidation II (plant peroxisome)
FAO-PWY: fatty acid &beta;-oxidation I (generic)
PWY-6803: phosphatidylcholine acyl editing
HCAMHPDEG-PWY: 3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate...
GLYOXYLATE-BYPASS: glyoxylate cycle
THREOCAT-PWY: superpathway of L-threonine metabolism
PWY-5347: superpathway of L-methionine biosynthesis (transsulfuration)
C
Isoprenoid biosynthesis
Vitamin B6 biosynthesis
L-histidine degradation
Ammonium
L-ornithine biosynthesis
L-citruline biosynthesis
L-glutamate
Urea cycle
D-glyceraldehyde 3-phosphate and pyruvate
Fumarate
L-aspartate
Adenosine and guanosine nucleotides biosynthesis
PWY-7282: 4-amino-2-methyl-5-diphosphomethylpyrimidine biosynthesis II
PWY0-845: superpathway of pyridoxal 5'-phosphate biosynthesis and salvage
PWY-5130: 2-oxobutanoate degradation I
HISDEG-PWY: L-histidine degradation I
PWY-5030: L-histidine degradation III
CITRULBIO-PWY: L-citrulline biosynthesis
PWY-7392: taxadiene biosynthesis (engineered)
PWY-6902: chitin degradation II (Vibrio)
PWY-5973: cis-vaccenate biosynthesis
PWY-5121: superpathway of geranylgeranyl diphosphate biosynthesis II (via MEP)
PWY-6125: superpathway of guanosine nucleotides de novo biosynthesis II
PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I
PWY66-429: fatty acid biosynthesis initiation (mitochondria)
PWY-7229: superpathway of adenosine nucleotides de novo biosynthesis I
PWY-6126: superpathway of adenosine nucleotides de novo biosynthesis II
PWY-4984: urea cycle
PYRIDOXSYN-PWY: pyridoxal 5'-phosphate biosynthesis I
PWY-7222: guanosine deoxyribonucleotides de novo biosynthesis II
PWY-7220: adenosine deoxyribonucleotides de novo biosynthesis II
ARGININE-SYN4-PWY: L-ornithine biosynthesis II

Fig. 3 Pathways defining each archetype focus on specific metabolic strategies favoring different cellular functions. Graphs of the compounds involved in the reactions of the top 20 pathways characterizing **A** Archetype 1, **B** Archetype 2, and **C** Archetype 3. Each pathway and its nodes are given a single color that matches the legend on the right. Each node represents a single compound in a pathway's reactions. Nodes/compounds that are duplicates, i.e. found in multiple pathways, are connected with a dashed black line to help visualize core compounds in the archetype and similarities between the pathways' reactions. Label boxes were added to highlight the core components defining each archetype. Common compounds, such as H+, phosphate, ATP, ADP, H_2O, NADP+, NADPH, NADH, NAD+, CO_2, coenzyme A, AMP, dioxygen, hydrogen carbonate, and diphosphate, were excluded to reduce visual noise. Interactive graph html files can be found in https://osf.io/tvu52

can fuel multiple downstream processes also highly represented in this archetype (Fig. 3A). For instance, pyruvate serves as a critical substrate for highly expressed BCAA biosynthesis pathways and the methylerythritol phosphate pathway. Similarly, phosphoenolpyruvate contributes to peptidoglycan biosynthesis pathways essential for the microbial cell wall. In addition to sugars, this archetype relies on L-glutamate as a key substrate for several highly expressed pathways involved in BCAA biosynthesis and peptidoglycan production (Fig. 3A). L-glutamate is supplied by pathways such as S-adenosyl-L-methionine salvage I, guanosine ribonucleotides de novo biosynthesis, and UMP biosynthesis pathways. Together, these metabolic features highlight Archetype 1's specialized role in carbohydrate utilization and biosynthesis of key structural and functional components, including BCAAs and components of the microbial cell wall such as peptidoglycans.

Archetype 2 is defined by pathways integrating glycolysis, the TCA cycle, glyoxylate bypass, and fatty acid metabolism, driving the production of acetyl-CoA, succinate, oxaloacetate, and pyruvate (Fig. 2E**,** Fig. 3B**,** Table S5). Central carbon metabolism integrates glycolysis, the TCA cycle, and related pathways to process carbon substrates into energy (ATP) and biosynthetic precursors such as acetyl-CoA and oxaloacetate. Archetype 2's elevated potential activity within central carbon metabolism, including pyruvate dehydrogenase, underscores the integration of multiple pathways to meet both energy and biosynthetic demands. Acetyl-CoA is both produced and consumed across diverse pathways, including fatty acid degradation, L-threonine metabolism, and aromatic compound degradation, while its consumption in the glyoxylate bypass generates succinate as a key downstream product (Fig. 3B). Succinate production is a recurring feature of Archetype 2, supported by amino acid degradation, methionine biosynthesis, and central carbon metabolism, with some pathways demonstrating a metabolite-dependent ability to both produce and consume succinate, while others, such as nucleotide biosynthesis, exclusively consume it (Fig. 3B). Finally, ammonium is produced through pathways linked to amino acid and nucleotide metabolism, while oxaloacetate and pyruvate can be both produced and consumed depending on availability (Fig. 3B). Unlike Archetype 1, which supports branched-chain amino acid biosynthesis, Archetype 2 prioritizes energy metabolism and nucleotide biosynthesis, emphasizing its role in driving core biosynthetic and energy-yielding processes.

Finally, the top 20 pathways representing Archetype 3 reveal links between the urea cycle (including the biosynthesis and recycling of intermediates like L-ornithine and L-citrulline), nucleoside biosynthesis, isoprenoid biosynthesis, and vitamin B6 biosynthesis (Fig. 2E**,** Fig. 3C**,** Table S5). Specifically, the urea cycle and histidine degradation pathways are indirectly connected to nucleoside biosynthesis through key intermediates such as fumarate/L-aspartate and L-glutamate, respectively, which integrate nitrogen and carbon metabolism (Fig. 3C). Histidine degradation produces ammonium, which is further processed in the urea cycle, creating a direct link between these pathways in nitrogen metabolism. Isoprenoid biosynthesis and vitamin B6 pathways share the same initial compounds such as D-glyceraldehyde 3-phosphate and pyruvate (Fig. 3C). Vitamin B6 is a cofactor essential for numerous enzymatic processes, including those in the urea cycle, amino acid metabolism, and nitrogen metabolism (Fig. 3C). Overall, this network demonstrates the tight metabolic integration between nitrogen metabolism through the urea cycle and histidine degradation and the production of essential components and cofactors (nucleosides and vitamin B6).

Together, these archetypes highlight specific metabolic strategies which might favor different ecosystems and functions with different impacts on the host's physiology and health: Archetype 1 may favor microbes that thrive on abundant carbohydrates supporting structural and essential cellular component biosynthesis for microbial growth (microbial cell wall, BCAA). Archetype 2 may favor microbes that use multiple energy harvesting routes supporting metabolic flexibility with specialization in fatty-acid metabolism. Finally, Archetype 3 may favor microbes with enhanced nitrogen metabolism capabilities, particularly in processing nitrogen compounds (including amino-acids) through urea cycle and related pathways.

Functional archetypes display distinct associations with gut microbiome compositional enterotypes

To explore potential relationships between functional archetypes and compositional community structures, we analyzed enterotype distributions across our cohort and functional archetypes using two recent approaches: Enterosignatures (ES) [28] and Enterotyper [29]. Both methods acknowledge the non-discrete nature of microbial communities, using probabilistic approaches to characterize gradual shifts in community structures captured by classification confidences or probabilities. Both methods consistently identified three major adult gut MCs associated with *Bacteroides*, *Prevotella*, or *Firmicutes* genera (Fig. 4A, B). Additional enterosignatures, ES-Esch (*Escherichia* and *Citrobacter*) and ES-Bifi (*Bifidobacterium* and *Streptococcus*) were minimally represented in our dataset, as expected, since they are primarily found in young children [28]. While the Firmicutes enterotype/ES was the most prevalent in our dataset, samples classified as the *Prevotella* enterotype exhibited notably stronger classification scores, suggesting more distinct compositional profiles (Fig. 4B).

Our analysis revealed significant, but not exclusive, associations between functional archetypes and the three compositional enterotypes. Like enterotypes, functional archetypes represent a continuous spectrum rather than discrete categories, with samples showing varying degrees of similarity to each archetypal state. Samples with high scores for Archetype 1 (characterized by enhanced carbohydrate metabolism, BCAA and cell wall component biosynthesis potential) were significantly associated with the *Prevotella* enterotype (Fig. 4B, C). While most samples classified as *Prevotella* enterotype show high Archetype 1 scores, not all samples with high Archetype 1 scores belong to the *Prevotella* enterotype (Fig. 4C-F). Similarly, Archetype 3 (elevated nitrogen and amino acid metabolism potential) showed an association with the *Bacteroides*/*Phocaeicola* enterotype (Fig. 4C-F) though this relationship was less exclusive when considering samples with dominant but not necessarily high Archetype 3 scores (Fig. 4C-F). Most Archetype 2 (High or Dominant) are *Firmicutes*-enterotype samples but not all *Firmicutes* are Archetype 2 with many samples clustered in the densely populated region of the archetypal space, characterized by medium Archetype 1 scores (Fig. 4C-F).

Analysis of the most abundant species and genera across dominant archetypes further supported these patterns (Fig. S9-10, Table S6). Samples with more extreme functional profiles (particularly for archetypes 1 and 3) showed more distinct compositional signatures: samples with high Archetype 3 scores showed higher relative abundance of *Bacteroides* and *Phocaeicola* compared to samples that were Archetype 3 dominant (Fig. S9-10). Similarly, *Prevotella* was significantly enriched in samples with high Archetype 1 scores compared to Archetype 1 dominant samples (Fig. S9-10). Finally, high Archetype 2 samples tended to have higher relative abundance of *Escherichia* genus compared to dominant Archetype 2 (Fig. S9-11).

Overall, these results indicate that while specific associations exist between functional archetypes and compositional MCs, the relationship is not strictly deterministic. The observed patterns suggest that similar functional capabilities can be maintained across different compositional configurations in the adult human gut microbiome, though certain extreme functional profiles may correspond to more distinct compositional patterns.

Functionally diverse archetype 2 shows enhanced temporal stability

To assess the stability of archetypal states, we analyzed the variation in archetype contributions of samples collected across consecutive visits from the same individuals (7 studies; *n* subjects = 656; *n* samples = 1557; range visits 2–6 in the span of 2—730 days; **Table S2**). Notably, Archetype 2, which is characterized by high pathway diversity and enrichment in energy-yielding processes, showed more consistent scores between consecutive samples, with changes distributing more tightly around zero compared to the broader variations seen in Archetypes 1 and 3 (Fig. 5A).

We further classified samples into high (≥ 0.66) and medium (0.33–0.66) archetype usage categories to examine state transitions. Individuals exhibiting high usage of a particular archetype showed substantial persistence of that state, with maintenance percentages of 41.6%, 49.5%, and 45.3% for Archetypes 1, 2, and 3, respectively (Fig. 5B, C). In contrast, samples with medium usage levels showed lower state stability, with persistence percentages ranging from 8% to 27.5% (Fig. 5B). These findings suggest that while strongly committed archetypal states tend to persist over time, microbiomes with intermediate archetype contributions display greater temporal flexibility in their metabolic configurations.

When classifying samples by their highest archetype score (dominant archetype), Archetype 3 showed the greatest stability (Fig. 5D). While transitions between archetypes can occur bidirectionally, some showed directional preferences: for example, Dominant Archetype 3 was more likely to transition to Dominant Archetype 2 than vice versa. These patterns suggest potential preferential metabolic state transitions, though these exist along a continuous spectrum rather than as discrete categories.

(See figure on next page.)

Fig. 4 Functional archetypes display distinct associations with gut microbiome compositional enterotypes. **A** Number of samples and classification probability for each enterosignature and **B** enterotype. **C** Relationships between archetype usage and enterotype classification probability, with lowess curves shown in red. **D** MDS plot of the archetypal space, colored by the classification probability of each enterotype. **E** Comparison of functional archetype and enterotype classifications for samples assigned by the highest archetypal score (dominant archetype). **F** Comparison for samples with a minimum score of 0.66 for one archetype (high archetype)

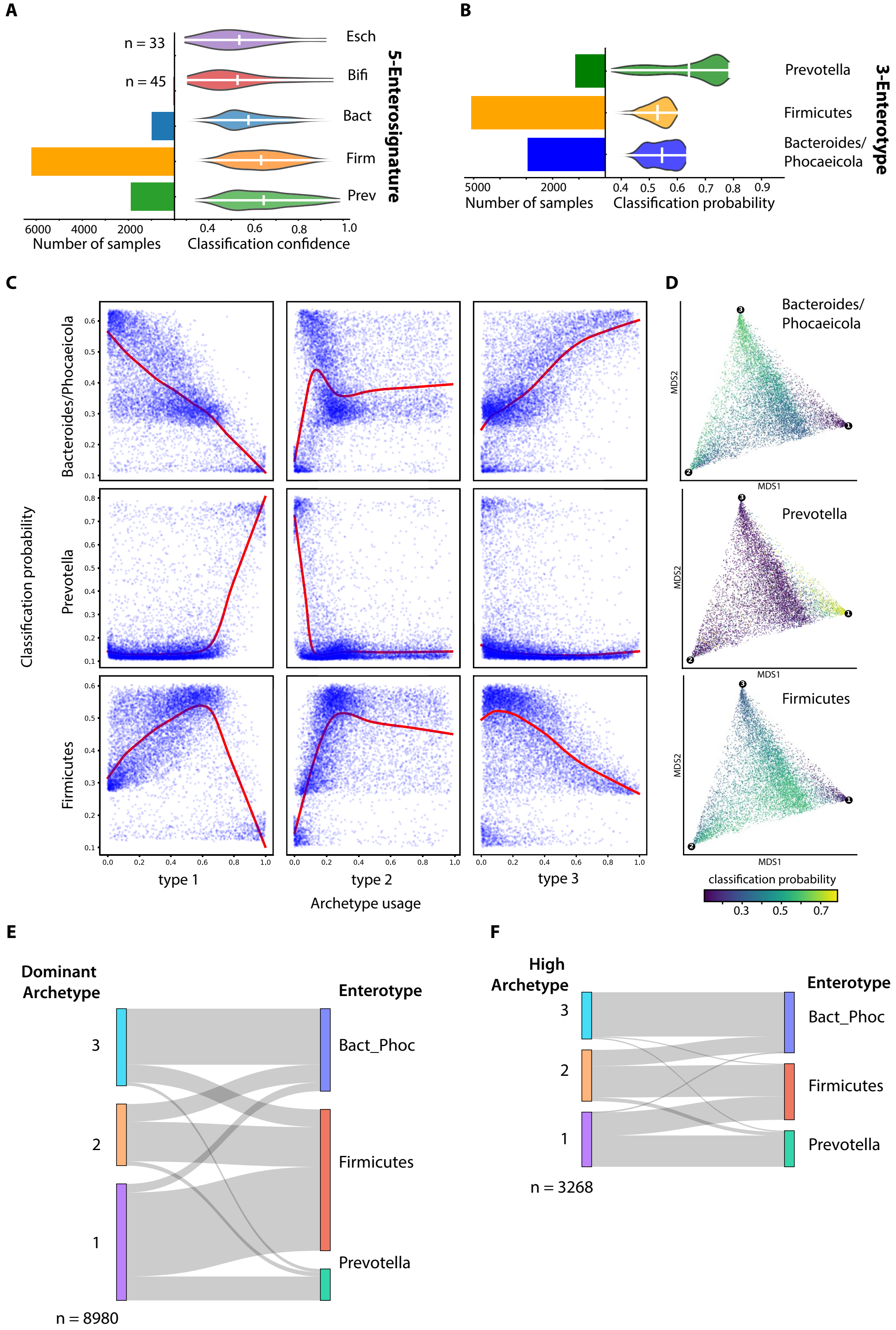
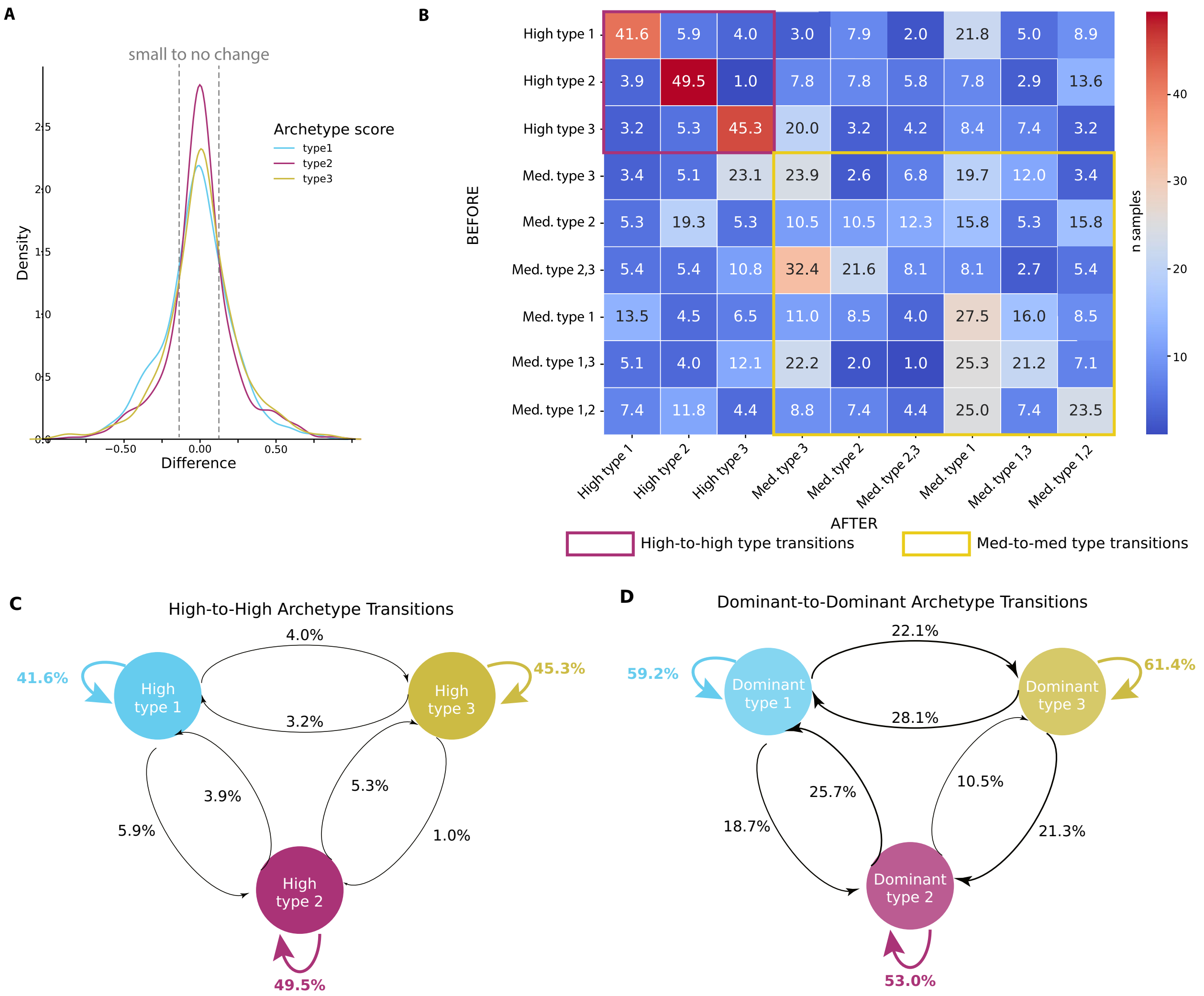


Fig. 4 (See legend on previous page.)



BEFORE \ AFTER	High type 1	High type 2	High type 3	Med. type 3	Med. type 2	Med. type 2,3	Med. type 1	Med. type 1,3	Med. type 1,2
High type 1	41.6	5.9	4.0	3.0	7.9	2.0	21.8	5.0	8.9
High type 2	3.9	49.5	1.0	7.8	7.8	5.8	7.8	2.9	13.6
High type 3	3.2	5.3	45.3	20.0	3.2	4.2	8.4	7.4	3.2
Med. type 3	3.4	5.1	23.1	23.9	2.6	6.8	19.7	12.0	3.4
Med. type 2	5.3	19.3	5.3	10.5	10.5	12.3	15.8	5.3	15.8
Med. type 2,3	5.4	5.4	10.8	32.4	21.6	8.1	8.1	2.7	5.4
Med. type 1	13.5	4.5	6.5	11.0	8.5	4.0	27.5	16.0	8.5
Med. type 1,3	5.1	4.0	12.1	22.2	2.0	1.0	25.3	21.2	7.1
Med. type 1,2	7.4	11.8	4.4	8.8	7.4	4.4	25.0	7.4	23.5

Fig. 5 Temporal dynamics of functional archetype scores in the human gut microbiome. **A** Density plot showing the difference between samples' archetype values that are from the same subject. Colors reflect scores for the different archetypes: type 1 (blue), type 2 (red), and type 3 (yellow). **B** Heatmap showing the probability of transitioning between archetype states for subjects with multiple samples. Sample states were categorized based on their archetype usage levels: high type usage (≥ 0.66 archetype value); if they did not have a high type usage, then they were set to medium type usage (0.33–0.66 archetype value). **C** High-to-high archetype transitions map. **D** Dominant-to-dominant archetype transitions map

Functional archetypal space captures disease-associated microbiomes and reveals context-dependent pathway differences

To determine whether the archetypal space captures disease-associated gut microbiomes, we curated metagenomic profiles from studies of patients with type 2 diabetes (T2D; 3 studies), colorectal cancer (CRC; 5 studies), and inflammatory bowel disease (IBD; 3 studies) (Table S3A). Using the model trained on healthy samples, we mapped the disease-associated samples and their respective controls onto the archetypal space (Fig. 6A). Notably, these samples aligned within the same archetypal framework, but their distributions differed significantly from those of healthy controls included in the same studies (Fig. 6B). IBD samples exhibited higher usage of Archetype 2 and lower usage of Archetype 3 compared to their controls, while CRC samples also showed significantly higher usage of Archetype 2 and lower usage of Archetype 1 (Kolmogorov–Smirnov test, $p \leq 0.05$).

Given these differences in archetype usage patterns between disease and healthy states, we investigated how accounting for archetype usage affects the identification of differentially represented pathways. Across all diseases (T2D, CRC, and IBD), many pathways identified as differentially represented between disease and healthy samples were also strongly associated with archetype usage. Adjusting for archetype usage significantly reduced the number of differentially enriched pathways (Fig. 6C;

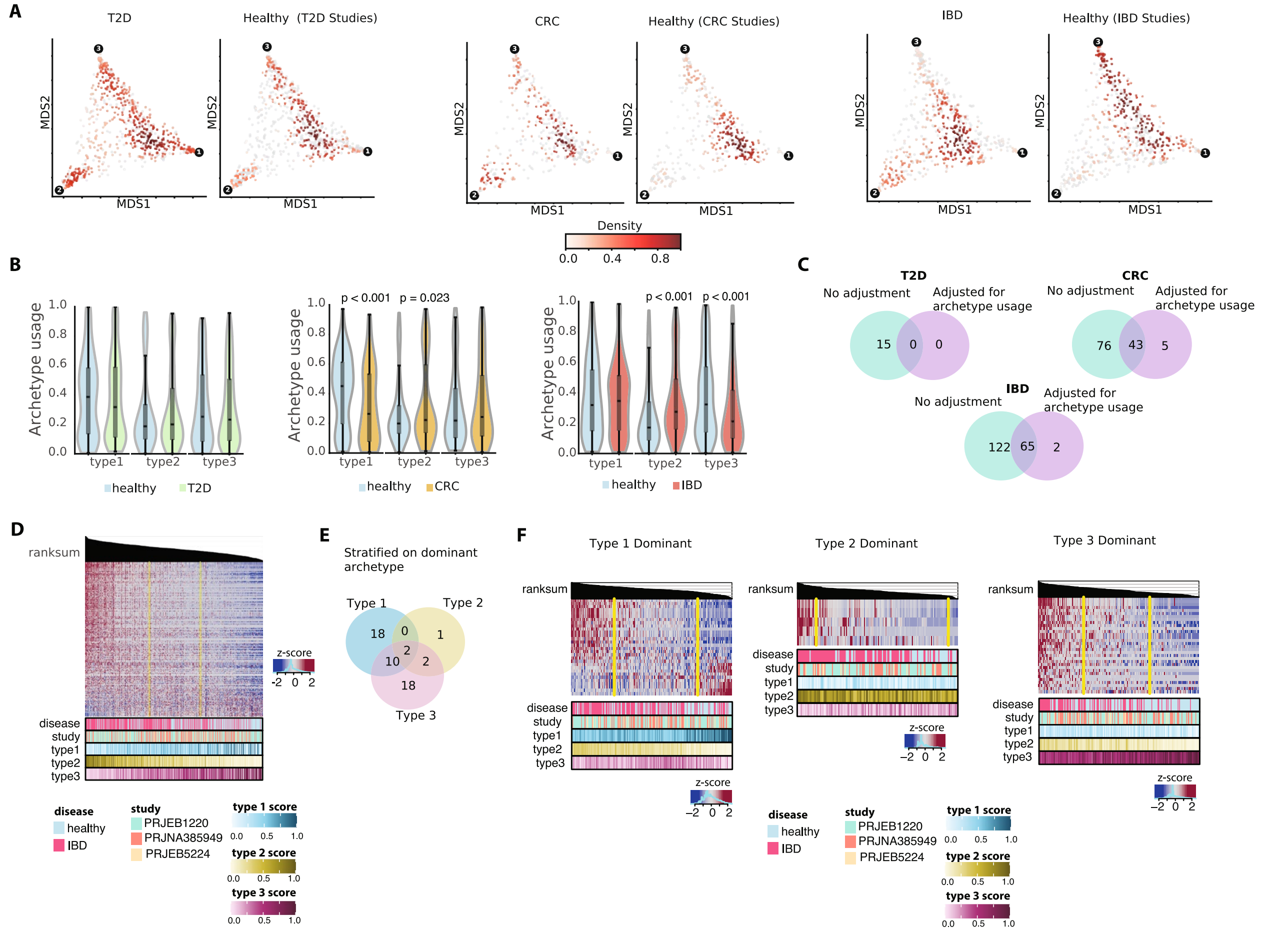


Fig. 6 **A** Density plots of healthy samples (left) and diseased samples (right) from type 2 diabetes (T2D), colorectal cancer (CRC) and inflammatory bowel disease (IBD) studies, shown in archetypal space after MDS. Samples of the opposite health status are displayed in grey. **B** Violin plot of archetype usage in diseased study samples, stratified by dominant archetype and colored by disease status. **C** Venn diagrams comparing the number of differentially expressed pathways between healthy and diseased samples from T2D, CRC, and IBD, with and without adjustment for archetypes usage. Pathway names are listed in Table S3B–F. **D** Heatmaps of pathways differentially expressed (FDR < 0.01, Table S3E) between healthy and IBD samples. Samples are ordered based on the ranksum of pathways relative abundance (z-score) depicted on top of the heatmap. Metadata at the bottom of the heatmap provide health status and sample's usage scores for each archetype. **E** Venn diagram showing the count of differentially expressed pathways (FDR < 0.01) between healthy and IBD samples, stratified by dominant archetype. **F** Heatmaps of pathways differentially expressed (FDR < 0.01) between healthy and IBD samples, stratified by dominant archetypes 1 (left), 2 (middle), and 3 (right). Pathway names are listed in Table S3G–I

Table S3B–F). For instance, when comparing IBD to healthy samples, pathways initially identified as differentially enriched ($n = 187$, FDR < 0.01) also largely reflected Archetype 2 usage (Fig. 6D). A similar effect was not observed after adjustment for enterotypes (Figure S12).

Stratified analysis based on archetype dominance defined by maximum usage score revealed distinct disease-associated pathways varying by dominant archetype (Fig. 6E, F; Table S3G–I). Sample sizes remained relatively balanced across strata, with Archetype 1 having 138 healthy and 146 IBD samples, Archetype 3 having 85 healthy and 84 IBD samples, while Archetype 2 showed fewer healthy samples (65) compared to IBD samples (98), consistent with the preferential association between IBD and this archetype.

In IBD samples dominated by Archetype 1—characterized by high carbohydrate metabolism potential—we identified 32 significant pathways when compared to healthy samples dominated by Archetype 1, including 18 pathways unique to this archetype comparison (Fig. 6E; Table S3). Among these unique pathways, we found processes related to thiamine biosynthesis, a process critical for carbohydrate metabolism and the production of short-chain fatty acids (SCFAs). Additionally, molybdopterin biosynthesis, essential for anaerobic respiration in bacteria, was significantly overrepresented in the gut microbiome of these IBD patients compared to controls. This pathway is particularly relevant given its potential role in the overgrowth of *Enterobacteriaceae* in the inflamed gut [59].

Conversely, IBD and healthy samples dominated by Archetype 3—associated with high processing capabilities of nitrogen compounds (including amino-acids)—differed in their pathway enrichment patterns. These differences included unique pathways involved in NAD salvage, the biosynthesis of amino-acid (L-cysteine, L-glutamine, arginine and polyamines), and fatty-acid-related pathways, including one specific to *Escherichia coli*. Fatty acids play important roles for bacterial membrane synthesis, while peptidoglycan biosynthesis, essential for bacterial cell wall integrity, was also uniquely overrepresented in gut microbiomes of IBD patients in this archetype (Fig. 6F).

Notably, ten differential pathways were shared among IBD samples closer to Archetype 1 or 3, including those involved in the TCA cycle, nucleotide degradation, starch metabolism, and heme b biosynthesis. In the unstratified analysis, differential pathways between IBD and controls captured Archetype 2, with IBD samples more likely to have high Archetype 2 usage characterized by high energy-harvesting potential through the TCA cycle. The consistent identification of TCA cycle pathway suggests that dysregulation of central energy metabolism is a common feature across IBD samples, regardless of archetype dominance.

Overall, these findings reveal substantial inter-individual variability in the functional metabolic landscape of gut microbiomes, presenting a potential confounding factor in differential analyses when comparing disease and control groups. The archetypal framework we developed provides a robust approach to mitigating these potential confounding effects, reducing false positives, and uncovering archetype-specific functional changes. These insights, for example, offer a novel perspective on IBD gut microbiome subtypes and their specific or common metabolic alterations depending on their overall archetypal functional landscape.

Discussion

Our deep archetypal model revealed that global gut microbiome functional potential can be represented by three archetypes, each defined by a high potential to express pathways within specific metabolic frameworks. Specifically, the three archetypes are skewed toward distinct metabolic features: sugar-related metabolism, whose products feed into branched-chain amino acid (BCAA) and cell wall component biosynthesis (Archetype 1); fatty acid metabolism, whose products fuel the TCA and glyoxylate cycles (Archetype 2); and amino acid metabolism and nitrogen metabolism through the urea cycle and related pathways (Archetype 3).

While most gut microbiome communities are a blend of these archetypes, some communities align closely with a single archetype, potentially reflecting adaptation to specific host factors such as physiology or diet. For example, a diet consistently rich in complex carbohydrates might promote a community resembling Archetype 1, while a diet rich in fatty acids or protein might promote a community resembling Archetypes 2 and 3, respectively. Similarly, host genetic, physiological factors or disease could create conditions that favor one archetype over others. In some cases, a community dominated by a single archetype could represent a functionally specialized microbiome optimized to meet the host's needs.

Our findings reveal a nuanced relationship between functional archetypes and compositional enterotypes in the human gut microbiome. While specific associations exist between functional archetypes and the three dominant adult enterotypes, the relationship is not strictly deterministic. The strong association between Prevotella enterotype and Archetype 1 reflects biological patterns consistent with the literature, as *Prevotella*-dominant communities are known for their enhanced capacity to metabolize plant-derived carbohydrates and produce SCFAs from diets rich in complex carbohydrates and dietary fibers [19, 60]. Similarly, the connection between Bacteroides/Phocaeicola enterotype and Archetype 3's elevated amino acid metabolism potential corresponds with this enterotype's known association with protein-rich Western diets [19, 60]. However, compositional configurations still show some flexibility in their functional archetypal profiles, supporting the concept of functional redundancy in microbial ecosystems, where different community structures can achieve similar metabolic capabilities [28, 33, 34]. These insights have important implications for microbiome-based interventions, suggesting that targeting functional capabilities rather than specific taxonomic compositions might be a more robust approach for therapeutic strategies.

Our findings provide new insights into the relationship between functional diversity and ecosystem stability. The identification of distinct functional archetypes, each characterized by different metabolic capabilities, demonstrates how gut microbiomes can achieve functional stability through various configurations. Furthermore, Archetype 2's enhanced stability, likely due to its diverse energy harvesting capabilities, exemplifies how metabolic flexibility can contribute to ecosystem resilience. These insights advance our understanding of the relationship between community composition, function, and stability in the gut microbiome.

The relationship between archetypal extremes and ecosystem stability presents intriguing questions about the functional stability of disease states. While disease samples are represented within the same archetypal space, they do not necessarily occupy extreme positions nor reflect simple imbalances in archetype usage. Instead, we observed disease-specific enrichments, such as IBD's association with Archetype 2. This finding aligns with previous analysis finding enrichment of TCA cycle pathways in IBD [61] and findings linking disruptions in the TCA cycle and its intermediates, such as succinate, to heightened inflammation and IBD pathogenesis [62, 63]. However, these patterns warrant careful interpretation, as they might reflect either true functional shifts favoring disease states or sampling biases due to the relatively limited scale and diversity of existing case–control studies.

Importantly, across all curated disease studies, the distributions of disease samples within the archetypal space consistently differed significantly from those of healthy groups. These differences, rooted in the distinct metabolic profiles of the archetypes, can introduce potential confounding when directly comparing disease and healthy samples. Our findings further revealed archetype-specific differences across IBD samples, consistent with previous findings identifying subtype-specific gut microbiome signatures in IBD patients [64–67]. While stratification by dominant archetype reduces sample sizes for individual comparisons, particularly for less prevalent archetypes, these archetype-specific signatures may reflect meaningful differences in clinical presentation (e.g., constipated vs non-constipated) or disease stages (e.g., quiescent vs inflamed). For instance, Gargari et al. identified a subgroup of non-constipated IBD patients with higher levels of SCFAs [65] which aligns with our findings in Archetype 1-dominant IBD samples. Archetype 1 is characterized by high carbohydrate metabolism potential and these IBD samples exhibited enrichment for thiamin biosynthesis, a pathway critical for SCFA production through the bacterial fermentation of carbohydrates; this enrichment was also seen in previous analysis in the literature [61]. These archetype-specific functional changes could inform microbiome-targeted interventions, such as dietary strategies aimed at lowering the fiber-fermenting microbial components of the gut microbiome. Notably, low-fiber diets have previously been shown to be potentially more effective in IBD patients with high fecal SCFA levels [65, 67]. Similarly, our findings highlight the potential for targeting specific metabolic pathways in Archetype 1-dominant IBD samples. For example, tungstate has been shown to prevent *Enterobacteriaceae* overgrowth in the inflamed gut by replacing molybdenum in the molybdopterin cofactor [59]. This substitution disrupts molybdopterin-dependent enzymatic pathways, which are essential for anaerobic respiration in *Enterobacteriaceae*—a pathway we identified as uniquely enriched in Archetype 1-dominant IBD samples.

In contrast, Archetype 3-dominant IBD samples exhibited enrichment of pathways involved in known inflammatory processes, including immune responses linked to *Escherichia coli* adherence. These pathways include NAD salvage [68, 69], amino acid biosynthesis [70–72], and the production of immunogenic cell wall-derived molecules, which may also promote the proliferation of adherent-invasive *Escherichia coli* [73–75]. These findings underscore the distinct functional and inflammatory mechanisms associated with different archetypes in IBD microbiomes.

To reduce dimensionality of functional profiles including > 1 million gene families and gain a more interpretable and biologically meaningful result, we aggregated gene families into higher-level functional categories using MetaCyc metabolic pathways [57]. MetaCyc contains experimentally determined pathways curated across a wide variety of organisms, though with enrichment for model organisms (*E. coli* is associated with 12% of pathways) [57]. Rather than modeling complete metabolism of specific organisms, MetaCyc serves as a general reference for metabolic pathways and enzymes [57]. HUMAnN3 leverages MetaCyc pathway definitions and MinPath (Minimal set of Pathways) to identify a parsimonious set of pathways that explain observed community reactions [76].

While pathway-level profiles provide a broad view of metagenomic functional capacity and can reduce noise from sequencing and annotation errors [76, 77], this higher-level aggregation may mask biologically relevant variation at finer functional resolutions [77]. Previous studies have demonstrated the value of manually curated, biome-specific pathway modules [33, 36, 78] as well as broader functional characterizations [54, 79, 80]—each approach offering distinct and complementary insights. To enable exploration across multiple functional resolutions, we have made both pathway-level profiles and gene family counts available, which allow users to investigate specific curated pathways of interest within this large collection of gut microbiome profiles.

Furthermore, the identification of functional genes through metagenomics does not necessarily equate to their expression or activity within the microbial community. Therefore, functional annotation of metagenomic data may not always reflect the active metabolic pathways or functions occurring within the community and needs to be complemented with additional multi-omics approaches, such as metaproteomics, metatranscriptomics, and metabolomics.

These methodological considerations, combined with our findings on functional archetypes and their relationship to disease states, underscore the need for large-scale studies with comprehensive functional profiling of diverse microbial communities. Such studies could improve our understanding of the functional landscape and help address potential sampling biases in case–control studies. Functional archetypes can serve dual purposes in microbiome analyses—as predictive markers of disease states, demonstrated by the distinct distributions of disease samples in the archetypal space, and as important functional contexts for differential analyses. Importantly, our work demonstrates that incorporating archetype values as a covariate or stratification factor in differential analyses reduces inter-individual variability and reveals novel pathways specifically altered in IBD gut microbiomes when accounting for their broader metabolic context, as captured by archetypal scores.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-025-02240-5.

Additional file 1: Fig. S1–S3 andTables S1, S2, S4, S5

Additional file 2: Table S3. Disease-associated pathway enrichment analyses. Excel file containing: (A) Summary statistics of disease cohorts. Differentially enriched pathways between healthy and disease samples: (B) T2D without archetype adjustment, (C) CRC without archetype adjustment, (D) CRC with archetype adjustment, (E) IBD without archetype adjustment, (F) IBD without archetype adjustment. (G-I) IBD-associated pathways stratified by dominant archetype (G: archetype 1, H: archetype 2, I: archetype 3), related to Fig. 6.

Acknowledgements

Not applicable

Authors' contributions

Conceptualization, V.D.; methodology, M.M, R.S., V.D.; software, M.M, D.S., V.D.; investigation, M.M., R.S., A.D., V.D.; visualization, M.M., V.D.; funding acquisition, V.D.; data curation, M.M., D.S., E.C., V.D.; supervision, V.D.; writing – original draft, M.M., V.D.; writing – review & editing, M.M., A.D., R.S., E.C., A.D., and V.D.

Funding

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant #391682. We thank the Canadian Foundation for Innovation J. Evans Leaders Fund grant #43481 for the support of the computing infrastructure used throughout our analyses.

MM received the Vector Scholarship in Artificial Intelligence, provided through the Vector Institute, and RS received the MITACS Globalink Summer Internship Award.

Data Availability

The code used for processing of the raw data is available at https://github.com/dumeaux-lab/compendium-fMC.The functional profile dataset and the code used for data analyses and generation of the figures in this manuscript is available through the Open Science Foundation (OSF) repository ([https://osf.io/tvu52/](https://osf.io/et4w9)) and its associated github repository https://github.com/dumeaux-lab/deep-fMC_paper.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Anatomy and Cell Biology, Western University, London, Canada. [2]Department of Biochemistry, Western University, London, Canada. [3]Department of Physiology & Pharmacology, Western University, London, Canada. [4]Department of Oncology, Western University, London, Canada. [5]Lawson Research Institute, London, Canada.

Received: 14 March 2025 Accepted: 21 August 2025
Published online: 26 November 2025

References

1. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol. 2008;6:776–88.
2. Tropini C, Earle KA, Huang KC, Sonnenburg JL. The gut microbiome: connecting spatial organization to function. Cell Host Microbe. 2017;21:433–42.
3. Conwill A, Kuan AC, Damerla R, Poret AJ, Baker JS, Tripp AD, et al. Anatomy promotes neutral coexistence of strains in the human skin microbiome. Cell Host Microbe. 2022;30:171-182.e7.
4. Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: networks, competition, and stability. Science. 2015;350:663–6.
5. Heintz-Buschart A, Wilmes P. Human gut microbiome: function matters. Trends Microbiol. 2018;26:563–74.
6. Hu J, Amor DR, Barbier M, Bunin G, Gore J. Emergent phases of ecological diversity and dynamics mapped in microcosms. Science. 2022;378:85–9.
7. McKinlay JB. Are bacteria leaky? Mechanisms of metabolite externalization in bacterial cross-feeding. Annu Rev Microbiol. 2023. https://doi.org/10.1146/annurev-micro-032521-023815. (**null**).
8. Seth EC, Taga ME. Nutrient cross-feeding in the microbial world. Front Microbiol. 2014 [cited 2022 Sep 20];5. Available from: https://www.frontiersin.org/articles/doi.org/10.3389/fmicb.2014.00350
9. Hibbing ME, Fuqua C, Parsek MR, Peterson SB. Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol. 2010;8:15–25.
10. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326:1694–7.
11. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. Science. 2016;352:560–4.
12. Rothschild D, Leviatan S, Hanemann A, Cohen Y, Weissbrod O, Segal E. An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. PLoS One. 2022;17:e0265756.
13. Shenhav L, Furman O, Briscoe L, Thompson M, Silverman JD, Mizrahi I, et al. Modeling the temporal dynamics of the gut microbial community in adults and infants. PLoS Comput Biol. 2019;15:e1006960.
14. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473:174–80.
15. Gibson TE, Bashan A, Cao H-T, Weiss ST, Liu Y-Y. On the origins and control of community types in the human microbiome. PLoS Comput Biol. 2016;12:e1004688.
16. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

17. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486:222–7.
18. Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol. 2018;3:8–16.
19. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334:105–8.
20. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014;509:357–60.
21. Hildebrand F, Nguyen TLA, Brinkman B, Yunta RG, Cauwe B, Vandenabeele P, et al. Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. Genome Biol. 2013;14:R4.
22. Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. Chimpanzees and humans harbour compositionally similar gut enterotypes. Nat Commun. 2012;3:1179.
23. Zhou Y, Mihindukulasuriya KA, Gao H, La Rosa PS, Wylie KM, Martin JC, et al. Exploration of bacterial community classes in major human habitats. Genome Biol. 2014;15:R66.
24. Kim JY, Whon TW, Lim MY, Kim YB, Kim N, Kwon M-S, et al. The human gut archaeome: identification of diverse haloarchaea in Korean subjects. Microbiome. 2020;8:114.
25. Lai S, Yan Y, Pu Y, Lin S, Qiu J-G, Jiang B-H, et al. Enterotypes of the human gut mycobiome. Microbiome. 2023;11:179.
26. Larzul C, Estellé J, Borey M, Blanc F, Lemonnier G, Billon Y, et al. Driving gut microbiota enterotypes through host genetics. Microbiome. 2024;12:116.
27. Huang G, Shi W, Wang L, Qu Q, Zuo Z, Wang J, et al. PandaGUT provides new insights into bacterial diversity, function, and resistome landscapes with implications for conservation. Microbiome. 2023;11:221.
28. Frioux C, Ansorge R, Özkurt E, Nedjad CG, Fritscher J, Quince C, et al. Enterosignatures define common bacterial guilds in the human gut microbiome. Cell Host Microbe. 2023;31:1111-1125.e6.
29. Keller MI, Nishijima S, Podlesny D, Kim CY, Robbani SM, Schudoma C, et al. Refined Enterotyping Reveals Dysbiosis in Global Fecal Metagenomes. bioRxiv; 2024 [cited 2025 Jan 11]. p. 2024.08.13.607711. Available from: https://www.biorxiv.org/content//10.1101/2024.08.13.607711v2
30. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "Enterotypes." Cell Host Microbe. 2014;16:433–7.
31. Levy R, Magis AT, Earls JC, Manor O, Wilmanski T, Lovejoy J, et al. Longitudinal analysis reveals transition barriers between dominant ecological states in the gut microbiome. Proc Natl Acad Sci U S A. 2020;117:13839–45.
32. Moya A, Ferrer M. Functional redundancy-induced stability of gut microbiota subjected to disturbance. Trends Microbiol. 2016;24:402–13.
33. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, et al. Species–function relationships shape ecological properties of the human gut microbiome. Nat Microbiol. 2016;1:1–8.
34. Reichardt N, Vollmer M, Holtrop G, Farquharson FM, Wefers D, Bunzel M, et al. Specific substrate-driven changes in human faecal microbiota composition contrast with functional redundancy in short-chain fatty acid production. ISME J. 2018;12:610–22.
35. Watson AR, Füssel J, Veseli I, DeLongchamp JZ, Silva M, Trigodet F, et al. Metabolic independence drives gut microbial colonization and resilience in health and disease. Genome Biol. 2023;24:78.
36. Labarthe S, Plancade S, Raguideau S, Plaza Oñate F, Le Chatelier E, Leclerc M, et al. Four functional profiles for fibre and mucin metabolism in the human gut microbiome. Microbiome. 2023;11:231.
37. Cutler A, Breiman L. Archetypal analysis. Technometrics. 1994;36:338–47.
38. Keller SM, Samarin M, Arend Torres F, Wieser M, Roth V. Learning extremal representations with deep archetypal analysis. Int J Comput Vis. 2021;129:805–20.
39. van Dijk D, Burkhardt D, Amodio M, Tong A, Wolf G, Krishnaswamy S. Finding Archetypal Spaces Using Neural Networks. ArXiv190109078 Cs Stat. 2019 [cited 2022 Oct 11]; Available from: http://arxiv.org/abs/1901.09078
40. Wang Y, Zhao H. Non-linear archetypal analysis of single-cell RNA-seq data by deep autoencoders. PLoS Comput Biol. 2022;18:e1010025.
41. Bulygin I, Shatov V, Rykachevskiy A, Raiko A, Bernstein A, Burnaev E, et al. Absence of enterotypes in the human gut microbiomes reanalyzed with non-linear dimensionality reduction methods. PeerJ. 2023;11:e15838.
42. Dai D, Zhu J, Sun C, Li M, Liu J, Wu S, et al. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. Nucleic Acids Res. 2021;50:D777–84.
43. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through experimenthub. Nat Methods. 2017;14:1023–4.
44. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: lifelines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015;5:e006772.
45. Scepanovic P, Hodel F, Mondot S, Partula V, Byrd A, Hammer C, et al. A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. Microbiome. 2019;7:130.
46. Carter MM, Olm MR, Merrill BD, Dahan D, Tripathi S, Spencer SP, et al. Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. Cell. 2023;186:3111-3124.e13.
47. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
48. Wright RJ, Comeau AM, Langille MGI. From defaults to databases: parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. Microb Genom. 2023;9:000949.
49. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell. 2019;178:779–94.
50. Pereira-Marques J, Ferreira RM, Figueiredo C. A metatranscriptomics strategy for efficient characterization of the microbiome in human tissues with low microbial biomass. Gut Microbes. 2024;16:2323235.
51. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;3:e104.
52. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20:257.
53. Hiseni P, Rudi K, Wilson RC, Hegge FT, Snipen L. Humgut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. Microbiome. 2021;9:165.
54. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife. 2021;10:e65088.
55. Zhang Y, Parmigiani G, Johnson WE. Combat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2:lqaa078.
56. Song D, Wang Q, Yan G, Liu T, Sun T, Li JJ. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nat Biotechnol. 2023 [cited 2023 May 18]; Available from: https://www.nature.com/articles/s41587-023-01772-1
57. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014;42:D459–71.
58. Mallick H, Rahnavard A, McIver L. Maaslin2: Multivariable Association Discovery in Population-scale Meta-omics Studies. Bioconductor version: Release (3.15); 2022 [cited 2022 Sep 8]. Available from: https://bioconductor.org/packages/Maaslin2/
59. Zhu W, Winter MG, Byndloss MX, Spiga L, Duerkop BA, Hughes ER, et al. Precision editing of the gut microbiota ameliorates colitis. Nature. 2018;553:208–11.
60. De Filippis F, Pasolli E, Tett A, Tarallo S, Naccarati A, De Angelis M, et al. Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. Cell Host Microbe. 2019;25:444-453.e3.
61. Zheng J, Sun Q, Zhang M, Liu C, Su Q, Zhang L, et al. Noninvasive, microbiome-based diagnosis of inflammatory bowel disease. Nat Med. 2024;30:3555–67.
62. Connors J, Dawe N, Van Limbergen J. The role of succinate in the regulation of intestinal inflammation. Nutrients. 2018;11:25.
63. Fernández-Veledo S, Vendrell J. Gut microbiota-derived succinate: friend or foe in human metabolic diseases? Rev Endocr Metab Disord. 2019;20:439–47.
64. Garcia-Mazcorro JF, Amieva-Balmori M, Triana-Romero A, Wilson B, Smith L, Reyes-Huerta J, et al. Fecal microbial composition and predicted functional profile in irritable bowel syndrome differ between subtypes and geographical locations. Microorganisms. 2023;11:2493.

65. Gargari G, Mantegazza G, Taverniti V, Gardana C, Valenza A, Rossignoli F, et al. Fecal short-chain fatty acids in non-constipated irritable bowel syndrome: a potential clinically relevant stratification factor based on catabotyping analysis. Gut Microbes. 2023;15:2274128.
66. Su Q, Tun HM, Liu Q, Yeoh YK, Mak JWY, Chan FK, et al. Gut microbiome signatures reflect different subtypes of irritable bowel syndrome. Gut Microbes. 2023;15:2157697.
67. Vervier K, Moss S, Kumar N, Adoum A, Barne M, Browne H, et al. Two microbiota subtypes identified in irritable bowel syndrome with distinct responses to the low FODMAP diet. Gut. 2022;71:1821–30.
68. Chen C, Yan W, Tao M, Fu Y. NAD+ metabolism and immune regulation: new approaches to inflammatory bowel disease therapies. Antioxidants. 2023;12:1230.
69. Gerner RR, Klepsch V, Macheiner S, Arnhard K, Adolph TE, Grander C, et al. NAD metabolism fuels human and mouse intestinal inflammation. Gut. 2018;67:1813–23.
70. Li P, Yin Y-L, Li D, Kim SW, Wu G. Amino acids and immune function. Br J Nutr. 2007;98:237–52.
71. Liu Y, Wang X, Hou Y, Yin Y, Qiu Y, Wu G, et al. Roles of amino acids in preventing and treating intestinal diseases: recent studies with pig models. Amino Acids. 2017;49:1277–91.
72. Ruth MR, Field CJ. The immune modifying effects of amino acids on gut-associated lymphoid tissue. J Anim Sci Biotechnol. 2013;4:27.
73. Kittana H, Gomes-Neto JC, Heck K, Juritsch AF, Sughroue J, Xian Y, et al. Evidence for a causal role for *Escherichia coli* strains identified as adherent-invasive (AIEC) in intestinal inflammation. mSphere. 2023;8:e00478–22.
74. Palmela C, Chevarin C, Xu Z, Torres J, Sevrin G, Hirten R, et al. Adherent-invasive *Escherichia coli* in inflammatory bowel disease. Gut. 2018;67:574–87.
75. Yin R, Wang T, Dai H, Han J, Sun J, Liu N, et al. Immunogenic molecules associated with gut bacterial cell walls: chemical structures, immune-modulating functions, and mechanisms. Protein Cell. 2023;14:776–85.
76. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput Biol. 2009;5:e1000465.
77. Manor O, Borenstein E. Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome. Microbiome. 2017;5:19.
78. Darzi Y, Falony G, Vieira-Silva S, Raes J. Towards biome-specific analysis of meta-omics data. ISME J. 2016;10(5):1025–8.
79. Björk JR, Bolte LA, Maltez Thomas A, Lee KA, Rossi N, Wind TT, et al. Longitudinal gut microbiome changes in immune checkpoint blockade-treated advanced melanoma. Nat Med. 2024;30:785–96.
80. Keshet A, Segal E. Identification of gut microbiome features associated with host metabolic health in a large population-based cohort. Nat Commun. 2024;15:9358.

Publisher's Note

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