Four functional profiles for fibre and mucin metabolism in the human gut microbiome

Simon Labarthe[1,2,3], Sandra Plancade[1,4], Sebastien Raguideau[1,5], Florian Plaza Oñate[6], Emmanuelle Le Chatelier[6], Marion Leclerc[7,8], and Beatrice Laroche[1,9]

[1]Université Paris-Saclay, INRAE, MaIAGE,78350 Jouy-en-Josas, France
[2]Univ. Bordeaux, INRAE, BIOGECO, 33610 Cestas, France
[3]Inria, INRAE, Pléiade, 33400 Talence, France
[4] UR875 MIAT, Université fédérale de Toulouse, INRAE, Castanet-Tolosan, France
[5]Earlham Institute, Organisms and Ecosystems, NR4 7UZ Norwich, UK
[6]Université Paris-Saclay, INRAE, MGP, 78350 Jouy-en-Josas, France
[7]Université Paris-Saclay, INRAE, Micalis, 78350 Jouy-en-Josas, France
[8] Pendulum Therapeutics, San Francisco, USA
[9]Inria, INRAE, Musca, 91120 Palaiseau, France

January 2023

Abstract

Background With the emergence of metagenomic data, multiple links between the gut microbiome and the host health have been shown. Deciphering these complex interactions require evolved analysis methods focusing on the microbial ecosystem functions. Despite the fact that host or diet-derived fibres are the most abundant nutrients available in the gut, the presence of distinct functional traits regarding fibre and mucin hydrolysis, fermentation and hydrogenotrophic processes has never been investigated.

Results After manually selecting 91 KEGG orthologies and 33 glycoside hydrolases further aggregated in 101 functional descriptors representative of fibre and mucin degradation pathways in the gut microbiome, we used non-negative matrix factorization to mine metagenomic datasets. Four distinct metabolic profiles were further identified on a training set of 1153 samples and thoroughly validated on a large database of 2571 unseen samples from 5 external metagenomic cohorts. Profiles 1 and 2 are the main contributors to the fibre-degradation-related metagenome: they present contrasted involvement in fibre degradation and sugar metabolism and are differentially linked to dysbiosis, metabolic disease and inflammation. Profile 1 takes over Profile 2 in healthy samples, and unbalance of these profiles characterize dysbiotic samples. Furthermore, high fibre diet favours a healthy balance between Profiles 1 and Profile 2. Profile 3 takes over Profile 2 during Crohn's disease, inducing functional reorientations towards unusual metabolism such as fucose and H2S degradation or propionate, acetone and butanediol production. Profile 4 gathers under-represented functions, like methanogenesis. Two taxonomic makes up of the profiles were investigated, using either the covariation of 203 prevalent genomes or metagenomic species, both providing consistent results in line with their functional characteristics. This taxonomic characterization showed that Profiles 1 and 2 were respectively mainly composed of bacteria from the phyla *Bacteroidetes* and *Firmicutes* while Profile 3 is representative of *Proteobacteria* and Profile 4 of methanogens.

Conclusions Integrating anaerobic microbiology knowledge with statistical learning can narrow down the metagenomic analysis to investigate functional profiles. Applying this approach to fibre degradation in the gut ended with

4 distinct functional profiles that can be easily monitored as markers of diet, dysbiosis, inflammation and disease.

Background

The generalization of metagenome sequencing 15 years ago has provided ample evidence of the complex interactions between the gut microbiota and its host health [1]. Since then, a large number of new links between the function and composition of the microbiota and the host health have been consistently discovered. Significant efforts put into the recruitment of large cohorts to constitute reference datasets made it possible to explore the high inter-individual variability of the microbial communities in the gut [2, 3, 4, 5].

Metabarcoding methods have been first popularized. Amplification of universal taxonomic marker gene before sequencing allows the construction of taxonomic entities (Operational Taxonomic Units, OTUs [6], or Amplicon Sequence Variants, ASVs [7]) informing on the phylogenetic composition of the microbial community [8] and on ecological biomarkers such as diversity indices [9]. Additional analysis can show co-occurrence networks [10, 11] or dynamical interactions in time-series [12, 13] both informing on ecological interactions. However, as the functional potential of the microbial populations remains unknown with metabarcoding techniques, the functional mechanisms that drive these interactions cannot be identified, even if tools leveraging reference databases of known genomes partially mitigate this issue [14].

With the development of metagenomic Next Generation Sequencing (mNGS) techniques [15], the entire functional information contained in the metagenomes became accessible. Shotgun sequencing together with bioinformatics methods identifying contigs between the sequenced fragments [16] and the constitution of massive catalogs of annotated genes [4] provide decisive tools for the study of the functional ecology in the gut microbiome. Multi-omics studies including metatranscriptomics or metabolomics give complementary information on the microbial functions actually activated in the gut [3]. Taxonomic and functional ecology can be addressed simultaneously with mNGS with the identification of entire microbial genomes in the metagenomes, such as Metagenomic Species (MGS [17]) or Metagenome-Assembled Genomes (MAG [18]). Statistical analysis makes it possible to decipher universal MGS patterns in both metabarcoding and metagenomic cohorts, termed enterotypes, that are linked to different physiopathological status [19].

However, despite the massive amount of metagenomic data that were gathered by the microbial ecology community and the sophisticated agnostic data-driven analysis methods that were developed, the understanding of the mechanisms involved in the gut microbiota regulation and dynamics remains scarce. This observation calls for the development of new approaches operating a shift from descriptive ecology towards functional ecology [20] by leveraging existing knowledge in microbiology to explore the links between community structure and functions [21].

Dietary and host-derived fibres are the main primary substrate for the gut microbiota [22] so that anaerobic hydrolysis and consecutive downstream sugar degradation towards short-chain fatty acids (SCFAs) are the most common microbial functions in the colon, the distribution of which reflects the fibre intake [22]. The corresponding metabolic pathways are very well characterized [23], hence providing suitable candidate functions for pattern identification and differential analysis. Considering the well-defined framework of fibre anaerobic hydrolysis, we hypothesize that (H1) functional invariants can be deciphered, defining 'universal' functional profiles shared by all individuals, describing fibre degradation in the microbiota, (H2) functional and taxonomic interpretation of these profiles can be obtained and (H3) these profiles characterize the metagenomic samples and are related to dysbiosis or disease.

In this study, we build on a method proposed in [24], which informs a data-driven dimension reduction technique termed nonnegative matrix factorization (NMF) with the well-established knowledge of fibre degradation pathways in the gut to analyse fibre-degradation-related metagenomic count matrix. The method is trained on a database of 1152 samples and validated on 5 external databases gathering 2571 unseen samples, allowing to identify four functional profiles the mixture of which reconstruct

the metagenomes. Extensive functional and taxonomic characterization of the profiles is performed and systematic differential analysis is conducted to identify possible links between the profiles and the sample physiopathological status. The microbiota simplification provided by the method allows in-depth biological interpretations of the differential analysis.

Methods

We first introduce the different datasets that are considered in this study. We then describe the rationale of the function selection and the pooling of the corresponding genes related to dietary and host-derived fibre degradation pathways, and the subsequent bioinformatics, from the samples to the frequencies matrix. We finally detail the NMF decomposition of the frequencies matrices to identify functional profiles in the metagenomes. Finally, we present the differential analysis method, based on profile weights in samples.

Training and external validation datasets

A training set was assembled with $n_s = 1126$ samples covering a balanced mix of health status, including healthy samples, inflammatory diseases (Crohn Disease -CD-, Ulcerative Colitis -UC-) and metabolic diseases (obese, type 2 diabetes) taken from 7 cohorts (accession ID PRJEB1220 [18], PRJEB4336 [25], PRJEB5224[4, 26], PRJNA48479 [27], PRJNA422434 [28], PRJEB6337 [29], PRJNA375935 [30]) and 5 countries (USA, China, Spain, Denmark, France) to avoid potential study or country effects. External validation datasets were taken from studies selected for their focus on a specific effect. We selected two cohorts dedicated to IBD – hmp2 (PRJNA398089 [3], $n_s = 1266$ samples) and CD (PRJEB15371 [31], $n_s = 119$ samples) –, one cohort to obesity –metacardis (accession ID PRJEB37249 [32], $n_s = 883$ samples)–, one cohort to mediterranean diet (accession ID PRJEB33500 [33], $n_s = 244$) and one to Parkinson disease (accession ID PRJEB17784 [34], $n_s = 59$ samples) since this disease is associated to a longer transit time and microbial modifications. Note that 3 and 5 samples, respectively, have been removed from cohorts PRJEB15371 and PRJEB37249 after quality checks. All together, these datasets make it possible to consider a large variety of co-variables, including Dysbiosis index (DI, see subsection Statistical treatment), Body Mass Index (BMI) used to define obesity, statin treatment against hypercholesterolemia, the four enterotypes *Bacteroides* 1 (Bact1), *Ruminococcaceae* (Rum), *Prevotella* (Prev) and *Bacteroides* 2 (Bact2) [19, 32] and Bristol score [35] used to determine stool appearance. Dataset overview can be found in Table 1. Dataset homogeneity has been assessed by computing intra and inter-variability of pairwise Bray-Curtis distance (pBCd, see subsection Statistical treatment and Figure 2). The complete list of samples and their corresponding metadata can be found in **Additional file 9 — Dataset count matrices, profile decomposition and metadata**.

A functional view of fibre degradation in metagenomes

Following the method that was previously used in [24], we assembled a simplified view of the metabolic network of fibre degradation (see Fig. 1.a and Methods sec. GH, PL and KO Graphical representation). Briefly, the first metabolic step was the hydrolysis of fibre, performed by specialized multimodular enzymes belonging to the CAZymes [36, 37]. We selected the main Glycosyl hydrolases (GH) and Pectin Lyases (PL) involved in the catabolism of the main dietary fibre consumed as part of a balanced diet: cellulose, hemicellulose, xylan, resistant starch and pectin [38, 39, 36, 40, 41, 37, 42]. Furthermore, since mucin can be used as a substrate by both pathogens and commensals, we included the beta-N-acetyl-glucosaminidase (GH84), fucosidase (GH29 and GH95), Neuraminidase/Sialidase (GH33) that cleave endogenous mucins and release galactose (GH2), glucose, fucose, or sialic acid moieties [43, 44] (Table 2 and Fig. 1.A). Pectate lyases PL1, PL9 and PL12 were also added. The hydrolysis of fibre and mucin releases oligosides and sugars that are subsequently subjected to anaerobic fermentation. The known fermentation pathways of glucose, fructose, mannose,

galactose, L-Arabinose, Xylose, L-Fucose and L-Rhamnose were recapitulated using bibliographic ressources [45, 23, 46] and Metacyc database (`https://metacyc.org/`) guided by expertise [47, 48, 49, 50, 51]. We included the Embden-Meyerhoff-Parnas (EMP), Entner-Doudoroff (ED) and semi-phosphorylative Entner-Doudoroff (SP-ED) pathways and the Bifidobacterium shunt. The downstream SCFA producing reactions were added i) the three known propionate pathways, including lactate pathways and the propanediol one which is not commonly found in commensals ii) butyrate produced from acetate and lactate-utilizing species, iii) acetate produced through the main pathways but also by some human GI tract pathogens. Finally, H2S, butanediol and acetone production pathways were added, together with the three hydrogen hydrogenotrophic utilization pathways : methanogenesis, sulfate reduction and the Wood-Ljundhal pathway of acetate production from H2/CO2 and glucose (see Fig. 1.a). For each pathway, KEGG Orthology (KO) were selected as being representative (KO not involved in other pathway) and essential (the corresponding function is needed for the completion of the pathway) to the given metabolism with the method detailed in [24]. We note that H2S production pathway has been added compared to [24]. See Additional file 10 — Supplementary materials for additional precisions on KO selection.

From the IGC 9.9M genes catalog [4], we extracted the resulting 129 352 selected genes (SG) included in the KO, GH and PL, that were further pooled in aggregated functional traits (AFT, see Fig. 1.b for a sketch of the selection and aggregation steps). A final list of 101 AFTs characterizing the fibre degradation process in the human gut microbiome was obtained, comprising 33 GH and PL and 68 KOs or KO aggregations (See Table 2 for the complete list of KOs, GHs and PLs that were conserved, and the file *List_of_Reactions.xlsx* in the Additional file 9 — Dataset count matrices, profile decomposition and metadata for the complete list of reactions).

Metagenomic Data and gene frequencies.

Gene abundance tables were generated with the METEOR software suite [52]. First, reads were mapped with bowtie2 [53] (parameters: –trim 80 -k 1000) to the integrated gene catalog (IGC) of the human gut microbiome [4], comprising 9.9 million of genes. Alignments with nucleotide identity less than 95% were discarded and gene counts were computed with a two-step procedure previously described that handles multi-mapped reads[29]. Finally, raw gene counts were normalized according to gene length and total number of mapped reads per sample, reported in relative frequency (FPKM normalization).

The IGC KO annotation was used to map the genes to their corresponding AFTs. The GHs and PLs were re-annotated in the IGC using Hmmer [54] and dbCan version 3 [55] with default parameters, after assessment of dbCan annotation quality on 145 manually annotated protein sequences as previously described [24], and the corresponding genes were mapped to their AFTs. The AFT frequencies were obtained by summing the FPKMs of all genes with the corresponding annotation, handling for multiple annotations as previously described [24].

At end, a AFT frequency matrix $X_i^{(AFT)}$ of dimension $n_{s,i} \times 101$ is built for each dataset $i \in \{train, hmp2, CD, metacardis, med.diet, Parkinson\}$, where $n_{s,i}$ is the number of samples of dataset i. The 9.9M genes frequencies are also used to compute pBCd between samples at the three aggregation levels, on the 9.9M genes, on the SGs and on the AFTs as displayed in Fig. 2.c (see Fig. 1.b for a sketch of the different aggregation levels and Methods sec. Statistical treatment for methods).

GH, PL and KO Graphical representation

GH and PL were distributed according to the dietary fibre type they degrade. Some GH or PL appear in several arrows because GH or PL CAZymes classification does not represent a unique substrate uptake and fibre degradation modular enzymes are usually not substrate specific. KO were represented by directed arrows linking metabolites together on a graph (Fig. 1.a). Note that each array of this graph represent a full metabolic pathway between metabolites, represented by the specific KOs collected for this pathway. Reaction cofactors such as CO2, ATP and others, were left out of

this representation. Extracellular compounds, which micro-organisms can uptake or excrete, were identified with black contours. Functional modules were identified from KEGG and expert knowledge. The metabolic network has been displayed with Pathvisio [56] (Fig. 1.a) and further annotated (functional blocks) with Inkscape [57].

Prevalent genome selection and function frequencies computation in prevalent genomes.

A list of 203 genomes (see Additional file 9 — Dataset count matrices, profile decomposition and metadata, *Genome_list.xlsx*) was built by selecting prevalent genomes from [26] and [58], taking care that the main phyla are represented. The genes involved in the 101 AFTs were recovered in 191 genomes (see Additional file 9 — Dataset count matrices, profile decomposition and metadata, *Genome_list.xlsx* for subset list): KEGG Orthology annotation was carried out using diamond (0.7.11) [59] and default parameters on the KEGG database from 2016 [60]. If a query was found to have multiple hits, only the best hit was kept, any hit with bitscore under 60 was discarded [28]. GH and PL annotations were obtained using Hmmer [54] and dbCan version 3 [55] with default parameters. The resulting presence/absence annotation is given in (see Additional file 9 — Dataset count matrices, profile decomposition and metadata, *Genome_list.xlsx*) and used for clustering in Fig. S7 (see sec. Statistical treatment).

Taxonomic count matrices

Two different taxonomic informations were derived by counting in the samples either the 203 PGs through annotation of taxonomic marker genes or metagenomic species. 40 taxonomic marker genes (TMG) [61, 62, 63] were extracted from each 203 gut microbiota PGs using fetchMG (`http://vm-lux.embl.de/~mende/fetchMG/about.html`) [64] with default parameters. These genes were annotated in the IGC catalog using diamond (0.7.11) [59] and default parameters. Any hit with bitscore, percent identity or alignment length under respectively 60, 97 and 45 was discarded as indicated in [64] for correct taxonomic annotation. TMGs frequencies in each sample were pooled by PG to assemble a genome frequency matrix $X^{(PG)}$ (see Additional file 9 — Dataset count matrices, profile decomposition and metadata, *X_PG.xlsx*). Metagenomic species (MGS) [17] were recovered in the *train* dataset. Genus abundance was computed according to MGS abundance in order to assemble a MGS-derived genus frequency matrix $X^{(mgs)}$ (see Additional file 9 — Dataset count matrices, profile decomposition and metadata, *X_mgs.xlsx*)

Inference of functional profiles

The inference method was thoroughly detailed in [24]. Briefly, starting from the frequence matrix $X_{train}^{(AFT)}$ of the 101 AFTs of the *train* dataset, we used a constrained Nonnegative Matrix Factorization (NMF) to decompose $X_{train}^{(AFT)}$ as the product of two nonnegative matrices, the profile matrix $H^{(AFT)}$ of dimension $k \times 101$ and the weight matrix $W_{train}^{(AFT)}$ of dimension $n_{s,train} \times k$ where k is the number of profiles, an hyperparameter to be tuned (see below). Each line of $H^{(AFT)}$ represents a functional profile, characterized by a vector of co-varying AFT frequencies: $H_{i,j}^{(AFT)}$ is the frequency of AFT j in profile i. The columns of $W_{train}^{(AFT)}$ represent the weights of the corresponding profiles in the different samples: $W_{train}^{(AFT)}{}_{i,j}$ represents the weight of profile j in the i-th sample of the *train* dataset $X_{train}^{(AFT)}$.

Matrices $W_{train}^{(AFT)}$ and $H^{(AFT)}$ are inferred by solving the optimization problem

$$(W_{train}^{(AFT)}, H^{(AFT)}) = \underset{\substack{W \geq 0 \\ H \geq 0 \\ FH^T \leq 0}}{\arg\min} \quad \|(X_{train}^{(AFT)} - WH)D^{-1}\|_F^2 + \alpha \left(\|W\|_F^2 + \|HD^{-1}\|_{1,2}^2\right) \qquad (1)$$

In this equation, D is a diagonal scaling matrix, so that $D_{ii} = \|X_{train}^{(AFT)}{}_i\|_2$ is the l_2 norm of the i-th column. The matrix F is a constraint matrix designed to favour

the presence in the profile of complete metabolic pathways linking two extracellular compounds in Fig. 1.a so that a given profile carries the whole set of reactions needed for intracellular metabolism (see Additional file 10 — Supplementary materials for additional precisions on the construction of F, Additional file 9 — Dataset count matrices, profile decomposition and metadata, *F.xlsx* for the constraint matrix and [24] for more details). Finally, 155 constraints were implemented so that F has dimension 155×101. The parameter α is a tuning parameter that sets up the impact of the regularization penalties $\|W\|_F^2 + \|HD^{-1}\|_{1,2}^2$ on the NMF. The Froebenius norm in penalty $\|W\|_F^2$ tends to standardize the profile weights in a given sample while the $l_{1,2}$ norm on H tends to assign each AFT to a limited number of profiles by inducing sparsity on the rows of H. The resulting profiles are not exclusive, meaning that a given AFT can be represented in several profiles.

The selection of the regularization parameter α and the number of profiles k was performed using the same triple criterion approach as in [24] providing the best trade-off between internal data reconstruction (reconstruction error criterion), reconstruction of external samples (bi-cross validation) and profile stability, while avoiding over-fitting. See Additional file 10 — Supplementary materials for precise definitions of the hyperparameter selection criteria.

Implementation of the NMF inference in python based on OSQP solver [65] is available at `https://forgemia.inra.fr/nmf4metagenomics/pynmf` and is based on a block coordinate descent algorithm consisting in alternatively solving the nonnegative least-square problems inferring $W_{train}^{(AFT)}$ knowing $H^{(AFT)}$ with

$$W_{train}^{(AFT)} = \underset{W \geq 0}{\arg\min} \ \|(X_{train}^{(AFT)} - WH^{(AFT)})D^{-1}\|_F^2 + \alpha \left(\|W\|_F^2\right) \tag{2}$$

and inferring $H^{(AFT)}$ knowing $W_{train}^{(AFT)}$

$$H^{(AFT)} = \underset{\substack{H \geq 0 \\ FH^T \leq 0}}{\arg\min} \ \|(X_{train}^{(AFT)} - W_{train}^{(AFT)}H)D^{-1}\|_F^2 + \alpha \left(\|HD^{-1}\|_{1,2}^2\right). \tag{3}$$

Average profile weights $\bar{W}_{train}^{(AFT)}$ and AFT counts $\bar{X}_{train}^{(AFT)}$ of the training set are defined. Namely, average profile weights $\bar{W}_{train}^{(AFT)} = \frac{1}{n_s}\sum_{i=1}^{n_s} W_{train,i}^{(AFT)}$ are computed by averaging W on the train set. Average AFT counts $\bar{X}_{train}^{(AFT)} = \frac{1}{n_s}\sum_{i=1}^{n_s} X_{train,i}^{(AFT)}$ are obtained in the same manner.

Profiles validation

The matrix $H^{(AFT)}$ whose lines are the 4 functional profiles obtained after NMF on $X_{train}^{(AFT)}$ was held fixed, and the positive least square regression (2) was performed on the validation datasets $X_d^{(AFT)}$, for $d \in \{hmp2, CD, metacardis, med.diet, Parkinson\}$ to determine the corresponding weight matrices $W_d^{(AFT)}$. Relative reconstruction error distributions $\|X_d^{(AFT)}{}_i - W_d^{(AFT)}{}_i H^{(AFT)}\|_F / \|X_d^{(AFT)}{}_i\|_F$, for $i = 1, \ldots, n_{s,d}$ are computed for validation assessment.

Genomes and MGS affectation to profiles

To affect genomes to the functional profiles, we assumed that the weights predicting profile assemblage to reconstruct $X_{train}^{(AFT)}$ were also a suitable predictor to reconstruct genome frequencies. In other words, we search for genomes that co-vary with the functional profiles, with the implicit assumption that the genes included in a functional profile will vary proportionally with the genomes that carry them. Hence, knowing the (1153×4) matrix $W_{train}^{(AFT)}$, the unconstrained positive least square regression (3) was solved on respectively the prevalent genomes and the MGS frequency matrices $X_{train}^{(PG)}$ and $X_{train}^{(mgs)}$ to infer $H^{(PG)}$ the 4×203 prevalent genome and $H^{(mgs)}$ the 4×217 MGS-derived genus count matrices. Note that the same $L_{1,2}$ regularization penalty as in equation (1) was applied to favour unique allocation to the profiles, together with the same penalty coefficient α.

Statistical treatment

All the computations and statistics have been performed with custom scripts using the standard python libraries numpy [66], scipy [67], pandas [68] and matplotlib [69]. Ternary plots, that are plots in barycentric coordinates of normalized W_1, W_2, and W_3 values, i.e. $W_i/(W_1 + W_2 + W_3)$ for $i = 1, 2, 3$, are produced with the Ternary python package [70] (Fig. 6.a, c and e, Fig. 7, Fig. S5.a, c and e, Fig. S8.c and d).

pBCd have been computed with scikit-learn [71] (see fig. 2). Intra-cohort pBCd refers to dissimilarities obtained with two samples of the same cohort while inter-cohort pBCd distribution of the dataset $i \in \{train, hmp2, CD, metacardis, med.diet, Parkinson\}$ refers to dissimilarities obtained with a sample from the dataset i and another sample from dataset $j \neq i$.

Dysbiosis index (DI) has been computed following [3]: a reference set has been set up with non-IBD samples of the *'hmp2'* cohort obtained after the 20-th weeks from the patient enrolment and DI is defined as the median pBCd with the reference dataset, excluding samples from the same individual. A dysbiotic threshold is defined as the quantile 0.9 of the DI in healthy samples: samples with DI above this threshold are tagged as dysbiotic [3].

To avoid statistical bias (individual effect) due to over-representations of the same individuals, only the first time point of each individual is included in differential analysis involving the 'hmp2' cohort, i.e. for BMI (Fig. 5.a and b), CD and dysbiosis analysis (Fig. 6.a to d).

PERMANOVA (fig. 2.d) has been performed on the intra-cohort pBCd matrices obtained from the different levels of aggregation (9.9M genes, SGs and AFTs, see Method sec. A functional view of fibre degradation in metagenomes) with scikit-bio [72] using 10000 permutations and default parameters, respectively to the following structuring co-variables: individual, sex, age, Body Mass Index (BMI), diagnosis, study and nationality.

All the statistical tests have been performed with the scipy.stats module [67] (Fig. 5, 6 and S5). Multiple test corrections were made with statsmodels.stats.multitest [73] (Fig. 7 and S6). In all graphs, significant p-values are indicated with one star if $1e-2 < p \leq 5e-2$, 2 stars if $1e-3 < p \leq 1e-2$, 3 stars if $1e-4 < p \leq 5e-3$ and 4 stars if $p \leq 1e-4$, non significant p-values are indicated with $n.s.$. The test name is indicated with the significance level. MW stands for the 'two-sided' Mann-Withney U test, levene for the levene test for the variance.

Support Vector Machine (SVM) classification has been made with scikit-learn [71] using 'rbf' kernel after cross-validation of the hyperparameters C and γ and min-max scaling normalization. The SVM classifier was trained on the 'hmp2' cohort, by classifying CD against healthy samples (Fig. S6).

Hierarchical clustering has been performed with the package scipy.cluster.hierarchy using a pairwise Jaccard distance matrix computed on the AFT presence-absence in the 191 genomes and the 4 profiles(see Prevalent genome selection and function frequencies computation in prevalent genomes), Ward algorithm and 4 clusters (Fig. S7).

Results

Assessment of dataset and gene selection

Upstream to any data analysis, we first assess that the training set is representative of the whole set of metagenomes included in the study by computing pBCd on the 9.9M genes, focusing on intra and inter-cohort distributions (see Methods Statistical treatment). The training set shows nearly identical intra- and inter-cohort pBCd distributions, that are also very close to the pBCd distribution obtained when the whole set of sample pairs are pooled (Fig. 2.a, dashed and plain blue curves superimposed with dotted red curve), indicating that the training set is representative of the gene diversity observed in the metagenomes of the different datasets. The intra- and inter-cohort pBCd of the CD cohort show a pick of high dissimilarities (Fig. 2.a, red curve), showing a higher prevalence of dissimilar samples in agreement with the over-representation of dysbiotic samples in this cohort (Fig. 2.b, red). A similar ob-

servation can be done for the hmp2 cohort, with slighter effects, that can be related to the over-representation of inflammatory bowel diseases (IBD) in these cohorts. On the contrary, the Mediterranean diet cohort presents a higher fraction of samples with low dysbiosis index (Fig. 2.b, purple).

We next check that the functional simplification operated in this study by selecting genes related to fibre degradation does not strongly bias the functional variability observed in the metagenome. Indeed, as fibres are the main substrate in the gut, fibre-related pathways are expected to be observed in all the metagenomes, inducing less variable counts that could impair sample differentiation. We then assess the impact of the different levels of aggregation and simplification of the metagenome performed in the study (see Fig. 1.b and Sec. A functional view of fibre degradation in metagenomes). The pBCd obtained on the SG frequencies (Fig.2.c, plain lines) show very similar distributions to the pBCd computed on the 9.9M genes (Fig.2.c, dotted lines), indicating that the functional simplification resulting from the gene selection allows to reproduce the same sample stratification as the one obtained from the whole metagenome. As expected, dissimilarities are strongly reduced when pooling the SGs in AFTs shifting pBCd towards lower values (Fig. 2.c, colored distributions), but AFT-based pBCd captures the over-representation of dissimilar samples in the CD and hmp2 cohorts. Furthermore, PERMANOVA shows that the main part of dataset structures with respect to co-variables are correctly reproduced by AFT-based pBCd (Fig. 2.d), indicating that the functions related to fibre degradation selected for the AFTs are suitable to capture stratifications observed in the whole metagenome.

Fibre degradation process is accurately described by 4 universal functional profiles

Statistical inference of the functional profiles.

Co-varying AFTs are identified in the training dataset using the NMF method (see Methods sec. Inference of functional profiles), resulting in 4 distinct functional profiles (matrix $H^{(AFT)}$) whose wheighted mixture with weights $W^{(AFT)}_{train}$ allows to reconstruct the training AFT counts $X^{(AFT)}_{train}$: $X^{(AFT)}_{train} \simeq W^{(AFT)}_{train} H$ (mean relative error : 17 %, see Fig. S1.a). We recall that the NMF method was specifically constrained by a metabolic-based constraint F favouring in practice the clustering in the same profile of AFTs belonging to the same metabolic pathways [24]. This constraint results in the distribution of the different metabolic pathways of the fibre degradation network among the 4 profiles.

Validation on external datasets

To assess the ability of the profiles to reconstruct external datasets, i.e. to validate the universality of the functional profiles, the nonnegative least square regressions (2) is performed on the AFT count matrix $X^{(AFT)}_d$ in order to identify the best weight matrix $W^{(AFT)}_d$ so that $X^{(AFT)}_d \simeq W^{(AFT)}_d H^{(AFT)}$ with $d \in \{hmp2, CD, metacardis, med.diet, Parkinson\}$. The relative reconstruction error distributions are very homogeneous across datasets, except for the CD dataset where increased reconstruction errors are observed (Fig. S1.a). This is probably induced by an over-representation of dysbiotic and CD samples in this dataset, that are less acurately reconstructed (Fig. S1 d and g). Stucturing variables such as study, health or weight status, drug administration, diet or dysbiosys do not strongly affect reconstructions (Fig. S1). In the worst case (dysbiotic samples), the mean relative error is kept under 27 % and the 0.95 quantile is kept under a relative error of 44 %.

We note a strong discrepancy in the four profile weights in the samples (Fig. S1 j). The weights W_1 and W_2 of profiles 1 and 2 are significantly higher than W_3 and W_4 in all datasets (paired t-test, p$< 1e-6$). This observation suggests that Profiles 1 and 2 carry characteristic gut microbiota fibre degradation functions dominant in the majority of metagenomes whereas Profiles 3 and 4 indicate specific functional variations.

To investigate the contribution of the different profiles to metagenome reconstruction, we compare the pBCd obtained on reconstructed counts with AFT-based pBCd

when the number of profiles is increased. Namely, we compute the reconstructed count matrices $\sum_{m=1}^{K} W_{d,m}^{(AFT)} H_m^{(AFT)}$ for $K = 1$ to 4, and compared the resulting AFT-based pBCd with the AFT-based pBCd computed on the original count matrix $X_d^{(AFT)}$, for $d \in \{hmp2, CD, metacardis, med.diet, Parkinson\}$ (Fig. 2. e). We can see that the first profile alone does not provide an accurate reconstruction of the pBCd distribution. Interestingly, adding the second profile allows to reconstruct the main part of the pBCd distributions (until approximatively the 80th centile in the worst case, Fig. 2.e, CD, orange line), except when the dataset involves over-representation of highly dissimilar samples (HMP2 and CD datasets, Fig. 2, c, orange and red distributions). However, in these cases, adding the third profile (and even the fourth for the *hmp2* dataset, Fig. 2.e, HMP2, green and pink lines) makes it possible to reconstruct higher pBCds. These observations suggest that Profiles 1 and 2 carry sufficient information to describe the AFT-related metagenomic variability in the main part of the population, except in dysbiotic situations, that are correctly rendered by adding Profiles 3 and 4 in the reconstruction. We also note that the reconstructed pBCds are slightly uniformly underestimated, the qq-plot lying slightly under the bisector line.

The four profiles present contrasted functional characteristics

To dig into the intrinsic functional characteristics of the different profiles, we plot their AFTs distributions (Fig. 3.a and S2). We first observe that the different profiles do not exhibit the same balance between GHs, i.e. AFTs involved in complex molecule cleavage like fibres, and KOs, i.e. AFTs taking in charge the downstream part of fibre degradation, from simple sugars to end products (Fig 1.a). Profile 1 carries the largest set of GH (70%), reflecting a very broad capacity to breakdown fibre, resistant starch and diverse plant cell wall polymers, unlike Profile 2 (38%), Profile 3 (23%) and Profile 4 (22%). Profile 1 main GHs are related to mucin (GH2, GH43, GH29, GH95), protein and xylan (GH3), pectin and plant cell wall (GH 43, GH28), and to a less extent to starch degradation (GH13) as shown in Fig.3.a (GH pie chart) and Fig. S2. Profiles 2, 3 and 4 are shifted towards sugar fermentation rather than hydrolysis. They are preponderantly characterized by starch degradation and amylase (GH13), with secondary GH activity related to protein and xylan degradation (GH3) and mucin (GH2) for Profile 2, fructan and inulin degradation (GH32) for Profile 3 and cellulose degradation (GH5) for Profile 4. Profiles 2 and 4 present high proportions of GH involved in protein degradation. In contrast, Profile 3 has noticeably low proportions of GH involved in plant cell wall breakdown compared to other profiles, but presents high proportions of GH2 releasing galactose from N acetyl-galactosamine moieties and GH29 and GH95 releasing fucose, suggesting a shift from polymers hydrolysis towards unusual sugar fermentation.

In the downstream part of fibre degradation, Profile 1 and Profile 2 are very similar (Fig 3.a, KO pie charts and S2). The main differences are related to galactose pathway (AFT 21 is more present in Profile 2) and in the propanoate pathway where Profile 1 takes in charge AFT 48 linking lactate to propanoate while Profile 2 is involved upstream in AFT 47 linking pyruvate to lactate. Profiles 3 and 4 present more dissimilarities: EMP proportion is reduced in Profile 3 while fucose (AFTs 22, 23 and 24) and propanoate (AFT 48 and 50) pathways are enhanced (Fig. 3.a and S2). Profile 3 is also the unique profile providing AFT 19 in galactose pathway. Profile 4 is characterized by a higher proportion of AFTs of the pyruvate pathway and the presence of the methanogenesis.

Profile contribution to the microbiota functional potential

These intrinsic characteristics functionally characterize each profile, but do not give insight into its importance in the metagenomes. We assess the relative contribution of each Profile i to the total count of AFT j by computing $\bar{W}_{train,i}^{(AFT)} H_{ij}^{(AFT)} / \bar{X}_{train,j}^{(AFT)}$, where $\bar{W}_{train}^{(AFT)}$ and $\bar{X}_{train}^{(AFT)}$ are the average weights and counts in the training set as defined in Methods sec. Inference of functional profiles. The four profiles have different ecological contributions in the metagenomes (Fig. 3.b and S3.a). As expected, Profile 1 is the main provider of GH counts, except for GH with the lowest

counts (GH44 and 48 for plant cell wall degradation, GH 101 and 129 for protein cleavage). It is also particularly involved in some pathways such as bifidobacterium shunt, butyrate production, WL, SPED, EMP, ED, fructose and fucose pathways. Profile 2 has a major contribution in the pyruvate, butanoate, acetone pathway and some specific KOs (K00882 and K01786 in the fructose pathway, K00965 for galactose metabolism, K13788 for acetate pathway). Profile 3 is the unique provider of some KOs such as K03080 in the fructose pathway, K01690 in ED or K04020 in acetate production. It is also particularly present in galactose, fucose, SPED and propionate production. Profile 4 is the main contributor for methanogenesis, and has otherwise small to marginal contributions in EMP, pyruvate or sulfur pathways.

Taxonomic make up of the 4 profiles

A natural question at this point is to wonder which taxonomic units could provide the AFT of each functional profiles. We selected 203 genomes among the top-prevalent strains in metagenomes, covering the main phyla found in the gut microbiota (see Methods sec. Prevalent genome selection and function frequencies computation in prevalent genomes) and assembled TMG count matrix $X^{(PG)}$ for the different metagenomic datasets (see Methods sec. Taxonomic count matrices). Under the assumption that a genome providing a specific AFT in a functional profile $H_i^{(AFT)}$, $i = 1, \cdots, 4$, should co-vary with the profile, we search by nonnegative inference the best $H^{(PG)}$ so that $X_{train}^{(PG)} \simeq W_{train}^{(AFT)} H^{(PG)}$. In this equation, $W_{train}^{(AFT)}$ is the weight matrix of the functional profiles (see Fig. 1.b and Methods sec. Genomes and MGS affectation to profiles). Hence, if $H^{(PG)}$ is consistent, we should also have for each external dataset $d \in \{hmp2, CD, metacardis, med.diet, Parkinson\}$ $X_d^{(PG)} \simeq W_d^{(AFT)} H^{(PG)}$. This is actually the case since the reconstruction errors at the phyla levels (Fig. S4) follow similar characteristics to the reconstruction of the AFT counts (Fig. S1). The same inference procedure is performed to reconstruct the training MGS count matrix $X_{train}^{(mgs)}$ resulting in the MGS profile matrix $H^{(mgs)}$ with similar reconstruction error distributions (Fig. S4.j).

Marked taxonomic structure of the profiles

The taxonomic profiling obtained with the MGS or the 203 PGs are particularly consistent (Fig 4.a and 4.b). Profile 1 is dominated by *Bacteroidetes* species belonging to the genera *Bacteroides* and *Prevotella*. In contrast, Profile 2 has a high diversity of *Firmicutes* species, with butyrate-producing species from the Cluster IV *Faecalibacterium prausnitzii* species, *Roseburia intestinalis*, *Ruminococcus bromii* which is a mucin degrader, and cluster XIVa *Eubacterium rectale* such as *Eubacterium eligens*. *Anaerostipes putredinis* is the main representer of the *Bacteroidetes* phylum. *Actinobacteria*, including the bifidobacteria and the *Verromicrobia* species *Akkermansia muciniphila* are also present in Profile 2. Profile 3 is strikingly distinct from the two first profiles. It has a major proportion of commensals of the *Proteobacteria* phylum (*Escherichia coli* K12 and *Klebsiella pneumoniae*) but also marginaly the multi-drug resistant *Escherichia coli* SMS-3-5 strain and *Citrobacter sp.* The mucin degrader *Ruminococcus gnavus* is the main representer of the *Firmicutes*. Within the *Bacteroidetes*, the main fibre hydrolysing species are not contributing but the *Bacteroides fragilis* are dominant. *Bifidobacteria* and *Akkermansia muciniphila* are also part of Profile 3 taxonomic contribution but more marginally. Profile 4, is significantly distinct regarding its taxonomic representation. The *Euryarchaeota* domain, and specifically with hydrogenotrophic methanogenic strains from *Methanobrevibacter smithii* species, are over-represented. Then follow *Firmicutes*, *Verrucomicrobia* (*Akkermansia Muciniphila*), *Bacteroidetes* and *Actinoacteria*. The MGS profiling of Profile 4 is rather different: it also includes the methanogens but otherwise gathers unclassified genus. These discrepancies can be related to the low amount of signal carried by Profile 4 (Fig. S1.j).

We now wonder how consistent are the profiles with the enterotypes obtained from the analysis of the taxonomic compositions of large metagenomic datasets [19, 32]. Profile 1 and 2 present contrasted distribution among enterotypes (fig. S5 c and d): if Profile 1 is over-represented in Bact2 and Prevotela enterotypes, higher

weights $W_2^{(AFT)}$ are observed for Bact1 and Ruminoccocus enterotypes. Interestingly, Profile 3 is almost only observed in Bact2 enterotypes and Profile 4 in Ruminoccocus enterotype (Fig. S5 d).

The profiles link the taxonomic and functional composition of the microbiota

Compared to the functional contribution of the profiles (Fig. 3.b), their taxonomic contribution is very structured (Fig. 4.c-d and S3.b-c). Profile 1 is the main contributor for *Bacteroidetes*, Profile 2 for *Firmicutes* and *Actinobacteriota*, Profile 3 for the *Proteobacteria* and some *Firmicutes* and Profile 4 for the *Euryarchaeota*. Repeating this analysis on MGS clustered by genus (Fig. 4.d and S3.c) leads to consistent results, despite the very different nature of the taxonomic data, i.e. targeted PGs versus untargeted MGS. This clear structure is particularly strinking since the taxonomic profiling is indirect and based on the profile weights obtained on the AFT counts, indicating that these specific phyla may carry specific AFTs of the different profiles, linking taxonomic composition and functional contribution to the metagenome.

To check this hypothesis, we blasted the genes involved in the AFTs in 191 PGs (fig. S7), and clustered the genomes by their similarity in carying AFT genes, adding the four profiles to the clustering process (see Methods sec. Statistical treatment). The *bacteroidota*, main carrier of GH genes, clustered with Profile 1 as expected. *Actinobacteria* clustered together, characterized by the *Bifidobacterium* shunt and one function involved in acetate production (AFT 60). *Firmicutes* are splitted in two groups: the first group characterized by the absence of fucose-related genes and little presence of fructose and mannose pathways clustered with Profile 4, while the others clustered with Profile 2 and 3. Profile 3 clustered with the *Proteobacteria* characterized by a strong representation of fucose, fructose, manose and propionate pathways. This clustering is very consistent with the taxonomic profiling, even though derived from very different biological signals. This repeated consistency (profiling with targeted PGs, untargeted MGS, clustering based on AFT presence/absence in genomes) suggests that the functional stratification described by the different profiles actually reflects co-variations of microbial entities. These covarying taxons, characterized by within-group functional similarities and between-group functional discrepancies, are the taxonomic support of the covarying AFTs defining the functional profiles.

Balance of profiles 1 and 2 reflects metabolic status and dysbiosis.

Profiles 1 and 2 particularly contributing to GH production and sugar metabolism AFTs, we therefore wondered if Body Mass Index (BMI) structured the samples in the W_1-W_2 space (Fig. 5.a). When $W_1^{(AFT)}$ is high and $W_2^{(AFT)}$ is low, higher BMIs are preponderant (light green dashed confidence ellipse), whereas lower BMIs are over-represented in the region defined by low $W_1^{(AFT)}$ and high $W_2^{(AFT)}$ (green confidence ellipse). Plotting $W_1^{(AFT)}$ and $W_2^{(AFT)}$ distributions stratified by obesity levels (Fig. 5.b) shows that $W_1^{(AFT)}$ values are significantly higher and $W_2^{(AFT)}$ significantly lower for class 3 obesity compared to healthy samples. Interestingly, the shifts are significantly reversed under statin treatment (Fig. S5.f), a drug used against hypercholesterolemia, suggesting metabolism-driven modifications of the microbiota. Statin is known to impact the microbial composition, reducing the prevalence of Bact2 enterotype in patients under treatment [32], consistently with the statin-induced reduction of $W_1^{(AFT)}$ since Profile 1 is over-represented in Bact2 enterotype (Fig. S5 d).

As profiles 1 and 2 are preponderant in the samples, we investigated if their respective weights are impacted during dysbiosis. To quantify the balance between profiles 1 and 2 in the microbiota, we introduce the barycentric coordinate $W^* = W_2^{(AFT)}/(W_1^{(AFT)} + W_2^{(AFT)})$ that we plot with stratification by Dysbiosis Index (DI, see Sec. Statistical treatment for DI definition). For balanced microbiota (Fig. 5.c, blue, DI $<$ dysbiotic threshold), the barycentric coordinates are tightened around an average ratio of 0.2, meaning that Profile 1 and 2 are mixed with a respective ratio

4:1 in non dysbiotic samples. On the contrary, W^* is significantly higher in dysbiotic samples (DI>dysbiotic threshold, orange, $p < 1e{-}5$, two-sided Mann-Withney (MW) test), with significantly increased dispersion around the mean ($p < 1e{-}5$, levene test). Shrinkage around $W^* = 0.2$ is enhanced for the first DI decile (gray) and W^* is more dispersed in the last decile (pink) compared to the set of dysbiotic samples. All together, these observations suggest that dysbiosis is characterised by unbalanced profiles 1 and 2. Furthermore, unbalance is induced by both a significant depletion of Profile 1 (Fig. S5 b, MW test) and a significant increase of Profile 2 (MW test).

Profile 1 main characteristic is its preponderant contribution in GH-related AFTs, involved in fibre cleavage. We then hypothesized that high fibre diet may impact Profile 1 and 2 balance. In an interventional study comparing mediterranean diet (considered as a high fibre diet) to a control diet, the distribution of the barycentric coordinate W^* are similar in the mediterranean diet and control groups at baseline (Fig. 5 d). Four weeks after intervention, W^* is tightened around the value 0.2 in the mediterranean diet group and this shrinkage is maintained eight weeks after intervention, whereas the dispersion is similar to the baseline in the control group (Fig. 5 d). Furthermore, the variance of W^* is significantly reduced after intervention in the mediterranean diet (Fig. 5 e, levene test) unlike the control group. The shift of W^* between four and eight weeks are higher in the control group compared to the mediterranean diet (Fig. 5 d.) with slight significance ($p = 0.06$, one-sided MW test). These observations suggest that the higher fibre intake in the Mediterranean diet contributes to the stabilization of Profile 1 and 2, particularly equipped with fibre degradation functions, around a non-dysbiotic ratio.

Profiles 3 is associated to Crohn's Disease and Profile 4 to slow transit.

When plotting the weight of the 3 first profiles in a ternary plot in the $W_1 - W_2 - W_3$ space (Fig. 6 a), Crohn's disease (CD) samples (red dots, red line : 95% confidence) are mainly shifted towards the W_1 and the W_3 corners whereas healthy samples (green dots, green line: 95% confidence) are kept near the basis of the triangle, around the ratio 0.2 between profiles 1 and 2 previously identified as a marker of healthy samples. This means that Crohn's disease is characterized by unbalanced profiles 1 and 2 and over represented Profile 3. Bar plots (Fig. 6 b) shows that the unbalance is driven by a very significant (MW test) depletion of $W_2^{(AFT)}$ in CD samples while $W_1^{(AFT)}$ is not significantly modified and a very significant increase of $W_3^{(AFT)}$ is observed (MW test). Hence, in CD samples, a shift in the profile weights occurs from Profile 2 towards Profile 3. This shift carries enough signal to correctly classify CD and healthy samples using SVM classifier with high accuracy (Fig. S6 e., recall:0.94, precision:0.81, AUC:0.92 for the unseen test cohort).

This observation is unexpected since dysbiotic samples ought to be over-represented in CD samples and we just saw that dysbiosis is characterized by an increase of $W_2^{(AFT)}$ (Fig. S5 a and b). We then color-coded dysbiotic and not dysbiotic samples in the ternary plot (Fig. 6 c) and stratified accordingly the bar plots (Fig. 6 d). In not dysbiotic samples, the weight of profiles 1 is increased in CD compared to healthy population whereas Profile 2 drops, with high significance. During dysbiosis, usual shifts occur: $W_1^{(AFT)}$ is reduced while $W_2^{(AFT)}$ is increased in both CD and normal populations, but $W_2^{(AFT)}$ remains lower for dysbiotic CD compared to healthy dysbiotic samples, with high significance (MW test). We also observe that Profile 3 is not a strong marker of dysbiosis since in healthy populations, a dysbiosis only triggers a limited increase of Profile 3 weight, while CD induces a strong increase of $W_3^{(AFT)}$ whatever the dysbiotic status with a strong enhancement during dysbiosis (Fig. 6 d). Interestingly, Profile 3 is mitigated by mediterranean diet (Fig. 6 f). After 8 weeks of high fibre diet, Profile 3 is significantly reduced (MW test) together with its variance (levene test) compared to baseline and to control (MW test $p = 3e - 3$, levene test $p = 4e - 2$) so that samples are kept near the basis of a ternary plot (Fig. 6 e). Mediterranean diet has been shown to improve the inflammatory status of patients experiencing an increase of microbial richness after diet change[33], suggesting that $W_3^{(AFT)}$ reduction after intervention could be linked with the inflammation reduc-

tion. This would be consistent with the taxonomic composition of Profile 3, carrying *Proteobacteria* known to bloom during inflammation. The association of $W_3^{(AFT)}$ with the CD inflammative disease and the over-representation of higher $W_3^{(AFT)}$ in Bact2 enterotype (Fig. S5 d) are also consistent with the previous identification of Bact2 as a dysbiotic microbiome [32].

The weight transfer from Profile 2 towards Profile 3 reflects functional shifts in CD compared to healthy samples. Functional modules are significantly over-represented in CD samples compared to healthy ones, in particular ED, fucose, galactose GH, sulfur and propionate pathways (Fig. S6. a and c, fdr 0.05, Benjamini-Hochberg correction). A closer look to the metabolic pathaways during CD and dysbiosis indicates a shift towards non typical metabolic pathways in the metagenome (Fig. 7). If some GHs are shifted, mainly involved in cellulose (GHs 44 and 48), xylan (GH 8) and protein (GH 101) degradations, the most interesting modifications occur in the KOs. First, the downstream part of fucose pathway including the propane1-2 diol production from L-lactadehyde (AFT 58; K13922) and propionate production through AFT 59 is particulary marked in CD samples: it is a propionate production pathway distinct to the usual one based on lactate transformation, which is reduced in CD (AFT 50). Consistently, the genes involved in acetate production through AFT 36 (K04020) and 60 (K13788) are non typical for anaerobic pathways and are over-represented in CD samples. Further shifts are observed during CD presenting alternatives in sulfur (AFT 62), SP-ED (AFT 9) and pyruvate (AFT 33) pathways.

During dysbiosis, these shifts are further enhanced. Fucose fermentation pathway is exacerbated with the increase of AFT 23 encoding for fucK which is present in *Proteobacteria* and *Akkermansia muciniphila* genomes (Fig. S7), which complete a pathway from fucose to propionate and enforces the availability of the corresponding genes. AFT 67 encoding for sulfite and NAPDH from hydrogen sulfide (Fig. 7) driving hydrogen removal from dissimilatory sulfate reduction is also increased: these functions are characteristic of *Proteobacteria* and *Bacteroidota* (Fig. S7) and are an alternative to AFTs 66 and 68 more present in *Firmicutes* and Profile 2 in healthy samples (Fig. 7, S7 and S2). Further modifications occur during dysbiosis and dysbiotic CD such as AFT 43 (acetone production) or GH 74 (hemicellulose degradation). Alternatively, some shifts are preponderant in healthy dysbiotic samples but do not belong to the main modifications in dysbiotic CD. Among them, alpha-galactose fermentation as characterized by AFT 19 (including gene dgoK; 2-dehydro-3-deoxygalactonokinase [EC:2.7.1.58]) involved in galactose to pyruvate metabolism is an alternative to galactose transformation towards glucose during dysbiosis. The alternative ED pathways for glucose fermentation is also enhanced (AFTs 12, 13 and 15) compared to EMP pathway in healthy dysbiotic samples, since these AFTs are over-represented in Profile 3 (fig. 3.a and Fig. S2).

If the functional count changes are relatively limited (fig S6 c.), the taxonomic changes are massive (fig S6 d.) supporting the fact that the observed functional shifts are carried by modifications in the taxonomic composition of the microbiota during CD. Functional redundancies across micro-organisms (fig. S7) lead to limited changes in the functional composition of the fibre-related metagenome, with more marked modifications in a limited number of functions involved in species functional specialization in alternative pathways. For example, *Proteobacteria* are characterized by the presence of propionate-related AFTs (fig S7), which relates the preponderance of *Proteobacteria* in CD samples (fig S6 d) to the shift towards $W_3^{(AFT)}$ in the distribution of propionate-related AFTs during CD (AFT 58, 59, 50, fig. 7).

Regarding Profile 4, $W_4^{(AFT)}$ weight is significantly reduced for higher Bristol scores (3 to 7), associated to more fluid stools, compared to low Bristol scores (1-2) associated to hard stools (Fig. S8, a). As fluid stools are often related to lower retention times in the gut, we wondered if larger retention times would favour Profile 4 and investigated a cohort including patients suffering Parkinson's disease, a disease associated to constipation, reported in 80–90% of PD patients [74]. As expected, $W_4^{(AFT)}$ is higher in PD samples compared to control with slight significance (MW test, p: 5.3e-2, Fig. S8,b). This relation of $W_4^{(AFT)}$ with low transit time can be linked to the taxonomic composition of Profile 4, mainly marked by the presence of methanogen archae, characterized by low growth rates.

Discussion

We used the NMF method previously introduced [24] to analyse metagenomic gene count matrix taking into account prior knowledge on fibre degradation. Our approach is based on a two-step microbiota simplification. In the first step, functional marker genes of interest are selected to build the AFT count matrix while providing a simplified view of the metagenome focused on fibre degradation. In the second step co-varying AFT are identified using NMF, leading to 4 universal functional profiles that can be used to reconstruct external samples. This double simplification is crucial to decipher changes among the very high dimensional metagenomic data and to provide extensive biological interpretations of the different profiles and their shifts during diseases. This functional viewpoint is supplemented by a taxonomic make up of the 4 profiles. Several external datasets were further studied and profile weight variations were linked to obesity, dysbiosis, mediterranean diet, statin intake and Crohn's disease.

Screening the profiles weights allows the identification of global shifts in the microbiota induced by conjoint changes in the co-varying genes of the profiles. We emphasize that this differential analysis relies on four quantities only (the weights of the four profiles), representing a dramatic reduction of the dimensionality. Furthermore, as the profiles take in charge specific parts of the metabolic network of fibre degradation, our framework is very suitable for functional interpretation: the profile weight variations are directly linked to functional variations that can be mapped to specific metabolic pathways of the fibre degradation network. Finally, the profile functional potentials are particularly consistent with their taxonomic composition and the functional peculiarities of the genomes they include.

In particular, new biomarkers were identified for dysbiosis and CD. A healthy microbiota is characterized by a balance of Profiles 1 and 2 around a proportion 4:1 while microbiota diverging from this 4:1 proportion are over-represented in dysbiotic samples. As Profiles 1 and 2 mainly differ by their GHs, these shifts reflect preponderantly changes in fibre cleavage. In the same way, Profile 2 and Profile 3 are sufficient to classify CD samples with high accuracy and reflect functional shift from usual to unusual pathways for fucose, propionate, H2S, SPED, acetone or butanediol, together with a bloom of *Proteobacteria*. These biomarkers give in themselves new insights on the underlying ecology during these pathological events. However, due to our focus on fibre degradation, we only capture changes inside fibre cleavage and fermentation pathways of fibre-derived sugars: our methodology is missing all the functional shifts outside this scope, which can be important in particular in pathological situations. This limitation could explain why many samples are tagged as dysbiotic with the 9.9M genes pBCd-derived classification, but display a healthy ratio of 4:1 between Profiles 1 and 2 for fibre-related genes. Hence, our methodology could be extended to other metabolic functions, such as respiration functions in micro-aerophilic environments during inflammation or protein degradation, or to non-metabolic functions such as antibiotic production or bile salt hydrolysis.

AFT selection is a crucial step of this methodology. Narrowing down the number of genes in the metagenome is needed for microbiota simplification. Furthermore, the careful selection of specific genes allows to link an AFT count to specific metabolic pathways despite ubiquitous genes: enlarging too much the set of selected genes would have blurred the biological interpretation by adding genes involved in very different pathways. However, some of these genes had to be added in the selection to allow certain degradation pathways. Selection step is then a trade-off between specificity and completeness of the global network, in the context of ubiquitous enzymes. Again, this modelling option can be seen as a bias of the present study that could be corrected by enlarging the functional scope of the method by enrolling other functions. We consider it as a necessary bias intrinsic to microbiota simplification, a price to pay for facilitated biological interpretations.

Another ambition of microbiota simplification is to decipher universal pattern, or functional invariant that can be searched for in a metagenomic sample. In the present study, four functional profiles are identified, that structure the main part of the metagenomic samples. In the inference procedure, strong caution has been put in hyperparameter selection and inference validation, with a particular criterion on the

stability of the inferred matrix H: the selected hyperparameters reduced the sensitivity of the inferred H to subsampling of the training set, enforcing the universality of the inferred profiles. Furthermore, the training set has been carefully constituted by enrolling a large panel of healthy, inflammatory disease, metabolic disorder, with a strong caution not to introduce age, sex, study or origin bias. The representativeness of the training set has been validated *a posteriori* by checking that its intrinsic pBCd distribution was identical to the overall pBCd distribution. We also stress that the ability of the profile to accurately reconstruct external samples has been widely validated by applying them to 2571 unseen samples from 5 external studies. However, other inference settings such as other regularization penalty, or a different learning set, could bring slightly different profiles. This drawback is inherent to dimension reduction strategies, also present in other strategies such as enterotyping: the microbiota simplification allows to decipher general features but the statistical method itself comes with intrinsic bias that introduces peculiarities.

The NMF method was previously used for metagenomic data analysis [75, 76, 77]. It can also be related to other dimension reduction or soft clustering techniques. NMF is comparable from a modelling point of view to mixture models such as DDM, that were used to identify enterotypes [19]: the metagenomic counts are seen as a mixture of different populations the composition and weight of which is unknown. The inference setting is however very different, and NMF suggests a continuous interpretation of the weights, by comparison to discrete allocation to an enterotype in DMM. NMF method can also be interpreted as a PCA-like method, constrained by the positiveness of the weights and the direction. The very specific added-value of our approach compared to previous microbiota reduction methods is the inclusion of prior knowledge on microbial physiology and bio-chemistry in the inference process through the functional constraint F (see eq.(1)). This introduction, deeply discussed in [24], facilitates the biological interpretation of the profiles, compared to completely agnostic approaches. We believe that adding such modelling overlay on statistical learning methods could be decisive in facilitating the integration of the wealth of knowledge acquired during decades by microbiologists before the omics revolution in the analysis of the high-throughput data of NGS methods.

Conclusion

In this paper, we analysed a large amount of data coming from various mNGS studies. From a training dataset with 1153 samples from 7 cohorts, we performed a two step microbiota simplification method based on AFT selection and NMF dimension reduction technique. We identified four universal functional profiles that were thoroughly validated on 2571 external samples from 5 independent studies and further characterized in term of functional capabilities related to fibre degradation and taxonomic composition. Profile 1 is strongly equipped in GH, making hydrolysis of a large variety of carbohydrates its main characteristic, and is mainly composed of *Bacteroidetes*. By contrast, Profile 2 is more directed towards starch or protein degradation and is mainly composed of *Firmicutes*. Profile 1 and 2 balance of roughly 4:1 is associated with a healthy microbiota while unbalance are associated with dysbiotic events. A Mediterranean diet can help stabilizing the microbiota around this healthy equilibrium. Profiles 1 and 2 unbalances mainly reflect shifts in fibre cleavage towards simple sugars, GHs distribution being the principal difference between these profiles.

Profile 3 takes over Profile 2 during CD, making shifts between both profiles a biomarker able to correctly classify CD patients. This ecological unbalance reflects functional reorientations towards unusual metabolism, in particular for fucose and H2S degradation and propionate, acetone and butanediol production. These alternative pathways are carried by *Proteobacteria*, the main phylum involved in Profile 3. Profile 4 is mainly marked by rare metabolism, such as methanogenesis, and is favoured by slow transit.

Integrating anaerobic microbiology knowledge into statistical learning methods narrows down the metagenomic analysis to investigating ecosystem traits and identifying functional invariants that can be easily monitored to identify markers of diet, dysbiosis, inflammation and disease.

Funding

This publication has been written with the support of the AgreenSkills+ fellowship programme which has received funding from the EU's Seventh Framework Programme under grant agreement Number FP7-609398 (AgreenSkills+ contract), and of the French National Research Agency under grant agreement number ANR-11-DPBS-0001.

Availability of data and materials

All data and codes are provided in the supplementary materials and the Methods section.

Authors' contributions

S.L. designed the study, performed the analysis and interpretation, performed additional bioinformatics, made the graphs and drafted the paper. S.P. and S.R. developed the NMF methodology and reviewed the paper. F.P.O. and E.L-C. provided the data, made the bioinformatics and participated to the dataset selection. B.L. and M.L. developed the NMF methodology, designed the study, made the result interpretation, participated to the graphs, drafted and reviewed the paper.

References

[1] C. R. Armour, S. Nayfach, K. S. Pollard, and T. J. Sharpton, "A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome," *MSystems*, vol. 4, no. 4, pp. e00332–18, 2019.

[2] H. Integrative, L. M. Proctor, H. H. Creasy, J. M. Fettweis, J. Lloyd-Price, A. Mahurkar, W. Zhou, G. A. Buck, M. P. Snyder, J. F. Strauss III, *et al.*, "The integrative human microbiome project," *Nature*, vol. 569, no. 7758, pp. 641–648, 2019.

[3] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, *et al.*, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, 2019.

[4] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, *et al.*, "An integrated catalog of reference genes in the human gut microbiome," *Nature biotechnology*, vol. 32, no. 8, pp. 834–841, 2014.

[5] S. D. Ehrlich, M. Consortium, *et al.*, "Metahit: The european union project on metagenomics of the human intestinal tract," in *Metagenomics of the human body*, pp. 307–316, Springer, 2011.

[6] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd, and E. Abebe, "Defining operational taxonomic units using dna barcode data," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1935–1943, 2005.

[7] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *The ISME journal*, vol. 11, no. 12, pp. 2639–2643, 2017.

[8] J. Rajendhran and P. Gunasekaran, "Microbial phylogeny and diversity: small subunit ribosomal rna sequence analysis and beyond," *Microbiological research*, vol. 166, no. 2, pp. 99–110, 2011.

[9] B.-R. Kim, J. Shin, R. B. Guevarra, J. H. Lee, D. W. Kim, K.-H. Seol, J.-H. Lee, H. B. Kim, and R. E. Isaacson, "Deciphering diversity indices for a better understanding of microbial communities," *Journal of Microbiology and Biotechnology*, vol. 27, no. 12, pp. 2089–2093, 2017.

[10] K. Faust and J. Raes, "Microbial interactions: from networks to models," *Nature Reviews Microbiology*, vol. 10, no. 8, pp. 538–550, 2012.

[11] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," 2012.

[12] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Rätsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota," *PLoS computational biology*, vol. 9, no. 12, p. e1003388, 2013.

[13] V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, *et al.*, "Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses," *Genome biology*, vol. 17, no. 1, pp. 1–17, 2016.

[14] M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, *et al.*, "Predictive functional profiling of microbial communities using 16s rrna marker gene sequences," *Nature biotechnology*, vol. 31, no. 9, pp. 814–821, 2013.

[15] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz, "A bioinformatician's guide to metagenomics," *Microbiology and molecular biology reviews*, vol. 72, no. 4, pp. 557–578, 2008.

[16] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nature biotechnology*, vol. 35, no. 9, pp. 833–844, 2017.

[17] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, *et al.*, "Metagenomic species profiling using universal phylogenetic marker genes," *Nature methods*, vol. 10, no. 12, pp. 1196–1199, 2013.

[18] H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, *et al.*, "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes," *Nature biotechnology*, vol. 32, no. 8, pp. 822–828, 2014.

[19] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, *et al.*, "Enterotypes of the human gut microbiome," *nature*, vol. 473, no. 7346, pp. 174–180, 2011.

[20] J. I. Prosser, B. J. Bohannan, T. P. Curtis, R. J. Ellis, M. K. Firestone, R. P. Freckleton, J. L. Green, L. E. Green, K. Killham, J. J. Lennon, *et al.*, "The role of ecological theory in microbial ecology," *Nature Reviews Microbiology*, vol. 5, no. 5, pp. 384–392, 2007.

[21] C. R. Tiffany and A. J. Bäumler, "Dysbiosis: from fiction to function," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 317, no. 5, pp. G602–G608, 2019.

[22] K. P. Scott, S. H. Duncan, and H. J. Flint, "Dietary fibre and the gut microbiota," *Nutrition bulletin*, vol. 33, no. 3, pp. 201–211, 2008.

[23] J. Cummings and G. Macfarlane, "The control and consequences of bacterial fermentation in the human colon," *Journal of Applied Bacteriology*, vol. 70, no. 6, pp. 443–459, 1991.

[24] S. Raguideau, S. Plancade, N. Pons, M. Leclerc, and B. Laroche, "Inferring aggregated functional traits from metagenomic data using constrained non-negative matrix factorization: Application to fiber degradation in the human gut microbiota," *PLoS computational biology*, vol. 12, no. 12, p. e1005252, 2016.

[25] E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto, S. Kennedy, *et al.*, "Richness of human gut microbiome correlates with metabolic markers," *Nature*, vol. 500, no. 7464, pp. 541–546, 2013.

[26] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[27] J. Lloyd-Price, A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A. B. Hall, A. Brady, H. H. Creasy, C. McCracken, M. G. Giglio, *et al.*, "Strains, functions and dynamics in the expanded human microbiome project," *Nature*, vol. 550, no. 7674, pp. 61–66, 2017.

[28] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[29] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J. M. Batto, S. P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S. D. Ehrlich, S. Zheng, and L. Li, "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, pp. 59–64, Sept. 2014.

[30] C. Wen, Z. Zheng, T. Shao, L. Liu, Z. Xie, E. Le Chatelier, Z. He, W. Zhong, Y. Fan, L. Zhang, *et al.*, "Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis," *Genome biology*, vol. 18, no. 1, pp. 1–13, 2017.

[31] Q. He, Y. Gao, Z. Jie, X. Yu, J. M. Laursen, L. Xiao, Y. Li, L. Li, F. Zhang, Q. Feng, *et al.*, "Two distinct metacommunities characterize the gut microbiota in crohn's disease patients," *Gigascience*, vol. 6, no. 7, p. gix050, 2017.

[32] S. Vieira-Silva, G. Falony, E. Belda, T. Nielsen, J. Aron-Wisnewsky, R. Chakaroun, S. K. Forslund, K. Assmann, M. Valles-Colomer, T. T. D. Nguyen, *et al.*, "Statin therapy is associated with lower prevalence of gut microbiota dysbiosis," *Nature*, vol. 581, no. 7808, pp. 310–315, 2020.

[33] V. Meslier, M. Laiola, H. M. Roager, F. De Filippis, H. Roume, B. Quinquis, R. Giacco, I. Mennella, R. Ferracane, N. Pons, *et al.*, "Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake," *Gut*, vol. 69, no. 7, pp. 1258–1268, 2020.

[34] J. R. Bedarf, F. Hildebrand, L. P. Coelho, S. Sunagawa, M. Bahram, F. Goeser, P. Bork, and U. Wüllner, "Functional implications of microbial and viral gut metagenome changes in early stage l-dopa-naïve parkinson's disease patients," *Genome medicine*, vol. 9, no. 1, pp. 1–13, 2017.

[35] M. Blake, J. Raker, and K. Whelan, "Validity and reliability of the bristol stool form scale in healthy adults and patients with diarrhoea-predominant irritable bowel syndrome," *Alimentary pharmacology & therapeutics*, vol. 44, no. 7, pp. 693–703, 2016.

[36] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat, "The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D233–D238, 2009.

[37] M.-L. Garron and B. Henrissat, "The continuing expansion of cazymes and their families," *Current opinion in chemical biology*, vol. 53, pp. 82–87, 2019.

[38] Z. Xu, C. Hu, M. Xia, X. Zhan, and M. Wang, "Effects of dietary fructooligosaccharide on digestive enzyme activities, intestinal microflora and morphology of male broilers," *Poultry science*, vol. 82, no. 6, pp. 1030–1036, 2003.

[39] P. J. Turnbaugh, B. Henrissat, and J. I. Gordon, "Viewing the human microbiome through three-dimensional glasses: integrating structural and functional studies to better define the properties of myriad carbohydrate-active enzymes," *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, vol. 66, no. 10, pp. 1261–1264, 2010.

[40] B. L. Cantarel, V. Lombard, and B. Henrissat, "Complex carbohydrate utilization by the healthy human microbiome," *PloS one*, vol. 7, no. 6, p. e28742, 2012.

[41] A. E. Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat, "The abundance and variety of carbohydrate-active enzymes in the human gut microbiota," *Nature Reviews Microbiology*, vol. 11, no. 7, pp. 497–504, 2013.

[42] W. Helbert, L. Poulet, S. Drouillard, S. Mathieu, M. Loiodice, M. Couturier, V. Lombard, N. Terrapon, J. Turchetto, R. Vincentelli, *et al.*, "Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space," *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, pp. 6063–6068, 2019.

[43] E. C. Martens, H. C. Chiang, and J. I. Gordon, "Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont," *Cell host & microbe*, vol. 4, no. 5, pp. 447–457, 2008.

[44] L. E. Tailford, E. H. Crost, D. Kavanaugh, and N. Juge, "Mucin glycan foraging in the human gut microbiome," *Frontiers in genetics*, vol. 6, p. 81, 2015.

[45] G. Gottschalk, "Bacterial fermentations," in *Bacterial metabolism*, pp. 208–282, Springer, 1986.

[46] S. Macfarlane and G. T. Macfarlane, "Regulation of short-chain fatty acid production," *Proceedings of the Nutrition Society*, vol. 62, no. 1, pp. 67–72, 2003.

[47] M. Leclerc, A. Bernalier, G. Donadille, and M. Lelait, "H2/co2metabolism in acetogenic bacteria isolated from the human colon," *Anaerobe*, vol. 3, no. 5, pp. 307–315, 1997.

[48] L. Tasse, J. Bercovici, S. Pizzut-Serin, P. Robe, J. Tap, C. Klopp, B. L. Cantarel, P. M. Coutinho, B. Henrissat, M. Leclerc, *et al.*, "Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes," *Genome research*, vol. 20, no. 11, pp. 1605–1612, 2010.

[49] P. Lepage, M. C. Leclerc, M. Joossens, S. Mondot, H. M. Blottière, J. Raes, D. Ehrlich, and J. Doré, "A metagenomic insight into our gut's microbiome," *Gut*, vol. 62, no. 1, pp. 146–158, 2013.

[50] D. A. Cecchini, E. Laville, S. Laguerre, P. Robe, M. Leclerc, J. Dore, B. Henrissat, M. Remaud-Siméon, P. Monsan, and G. Potocki-Véronèse, "Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria," *PloS one*, vol. 8, no. 9, p. e72766, 2013.

[51] R. Muñoz-Tamayo, B. Laroche, É. Walter, J. Doré, S. H. Duncan, H. J. Flint, and M. Leclerc, "Kinetic modelling of lactate utilization and butyrate production by key human colonic bacterial species," *FEMS microbiology ecology*, vol. 76, no. 3, pp. 615–624, 2011.

[52] F. Gauthier and N. Pons, "Meteor."

[53] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, Apr. 2012.

[54] "Hmmer."

[55] Y. Yin, X. Mao, J. Yang, X. Chen, F. Mao, and Y. Xu, "dbcan: a web resource for automated carbohydrate-active enzyme annotation," *Nucleic acids research*, vol. 40, no. W1, pp. W445–W451, 2012.

[56] M. Kutmon, M. P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A. R. Pico, and C. T. Evelo, "Pathvisio 3: an extendable pathway analysis toolbox," *PLoS computational biology*, vol. 11, no. 2, p. e1004085, 2015.

[57] B. Harrington and t. Inkscape development team, "Inkscape."

[58] J. Tap, S. Mondot, F. Levenez, E. Pelletier, C. Caron, J.-P. Furet, E. Ugarte, R. Muñoz-Tamayo, D. L. Paslier, R. Nalin, *et al.*, "Towards the human intestinal microbiota phylogenetic core," *Environmental microbiology*, vol. 11, no. 10, pp. 2574–2584, 2009.

[59] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using diamond," *Nature methods*, vol. 12, no. 1, pp. 59–60, 2015.

[60] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D457–D462, 2016.

[61] D. R. Mende, S. Sunagawa, G. Zeller, and P. Bork, "Accurate and universal delineation of prokaryotic species," *Nature methods*, vol. 10, no. 9, pp. 881–884, 2013.

[62] R. Sorek, Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork, and E. M. Rubin, "Genome-wide experimental determination of barriers to horizontal gene transfer," *Science*, vol. 318, no. 5855, pp. 1449–1452, 2007.

[63] F. D. Ciccarelli, T. Doerks, C. Von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *science*, vol. 311, no. 5765, pp. 1283–1287, 2006.

[64] J. R. Kultima, S. Sunagawa, J. Li, W. Chen, H. Chen, D. R. Mende, M. Arumugam, Q. Pan, B. Liu, J. Qin, *et al.*, "Mocat: a metagenomics assembly and gene prediction toolkit," 2012.

[65] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "Osqp: An operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.

[66] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, p. 357–362, 2020.

[67] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[68] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, Austin, TX, 2010.

[69] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[70] Marc, B. Weinstein, tgwoodcock, C. Simon, chebee7i, W. Morgan, V. Knight, N. Swanson-Hysell, M. Evans, jl bernal, ZGainsforth, T. G. Badger, SaxonAnglo, M. Greco, and G. Zuidhof, "marcharper/python-ternary: Version 1.0.6," Apr. 2019.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[72] T. scikit-bio development team, "scikit-bio: A bioinformatics library for data scientists, students, and developers," 2020.

[73] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[74] A. Fasano, N. P. Visanji, L. W. C. Liu, A. E. Lang, and R. F. Pfeiffer, "Gastrointestinal dysfunction in parkinson's disease," *The Lancet Neurology*, vol. 14, no. 6, pp. 625–639, 2015.

[75] Y. Baran and E. Halperin, "Joint analysis of multiple metagenomic samples," *PLoS computational biology*, vol. 8, no. 2, p. e1002373, 2012.

[76] X. Jiang, J. S. Weitz, and J. Dushoff, "A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data," *Journal of Mathematical Biology*, vol. 64, no. 4, pp. 697–711, 2012.

[77] X. Jiang, M. G. I. Langille, R. Y. Neches, M. Elliot, S. A. Levin, J. A. Eisen, J. S. Weitz, and J. Dushoff, "Functional biogeography of ocean microbes revealed through non-negative matrix factorization," *PLOS ONE*, vol. 7, pp. 1–9, 09 2012.

Figures

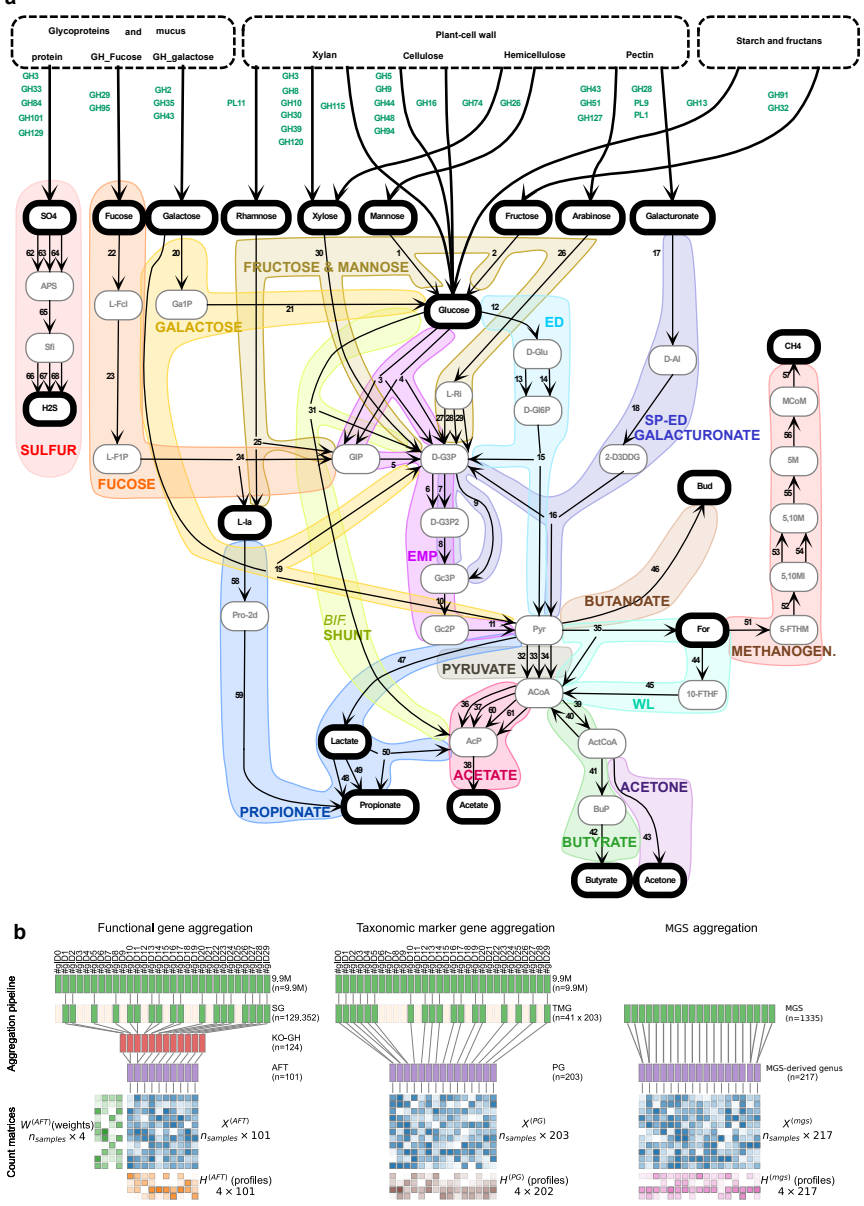


Figure 1: **Modeling overview**. **a**) *Schematic metabolic network of fibre degradation in the gut.* The metabolic network used to model fibre degradation in the gut is represented from complex dietary and host-derived fibres to terminal metabolites. Dashed boxes in the upper part represent fibre pools that are linked to fibre-derived sugars by GH and PL. Intra- and extra-cellular metabolites are respectively represented by gray and black boxes. Metabolic pathways linking metabolites are numbered from 1 to 68 (see Table 2): representative KOs are selected for each pathway, checking for specificity (KO are not involved in other metabolic reactions) and essentialness (essential reactions for the completion of the pathway). Functional blocks are represented by coloured shapes. GH_Fucose and GH_galactose: complex carbohydrate involving respectively fucose and galactose. ED: Entner-Doudoroff, SP-ED: semi-phosphorylative Entner-Doudoroff, EMP: Embden-Meyerhoff-Parnas, Bif. shunt: Bifidobacterium shunt, WL: Wood-Ljundhal. Complete list of reactions and abbreviations can be found in the Additional file 9 — Dataset count matrices, profile decomposition and metadata. **b**) *Gene count aggregation pipelines.* The pipelines used to build the count matrices are sketched. To build $X^{(AFT)}$, KO, GH and PL are first selected according to the metabolic network in a), leading to a list of Selected Genes (SG) that are annotated in the 9.9M gene catalog and pooled into their respective KO, GH or PL. Some KOs are gathered according to functional proximity, leading to Aggregated Functional Trait (AFT). This aggregation scheme allows to transform sample gene frequencies into AFT frequencies in $X^{(AFT)}$ by pooling SG counts. For Prevalent Genome (PG) counts, Taxonomic Marker Genes (TMG) are extracted from the genomes with FetchMg and annotated in the 9.9M catalog: the aggregated TMG are next counted in the samples to build $X^{(PG)}$. MGS are reconstructed from the metagenomes, directly counted in the samples and pooled by genus to build $X^{(mgs)}$. A NMF is performed on $X^{(AFT)}$ to obtain $W^{(AFT)}$ (weights) and $H^{(AFT)}$ (functional profiles). Then, nonnegative least square inference is conducted on $X^{(PG)}$ and $X^{(mgs)}$ using $W^{(AFT)}$ as regressor to obtain $H^{(PG)}$ and $H^{(mgs)}$ (PG and MGS taxonomic profiles).

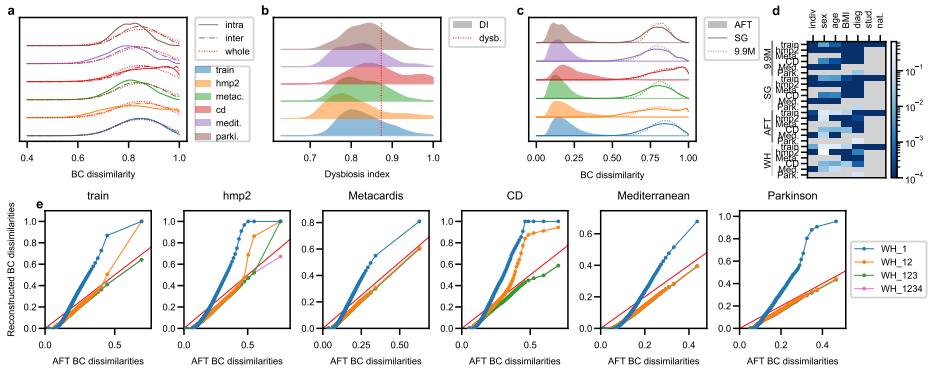


Figure 2: **Samples representation with AFT. a**) Intra and inter datasets pBCd distributions are computed on the 9.9M genes for each cohort dataset and compared with pBCd distributions among all samples. Little discrepancies are observed except for the Metacardis and Mediterranean diet cohorts, where intra pBCd is shifted towards lower values, and the CD cohort, where the shift is towards higher values. **b**) The dysbiosis index distribution of each dataset is displayed, together with the dysbiosis threshold (red dotted line). Dysbiotic samples are over-represented in the CD cohort. **c**) Comparison of different aggregation levels. pBCd distributions are displayed for each dataset, computed both on the 9.9M gene counts, on the subset of SGs or on the AFT counts (see Figure 1.b). pBCd with AFT are strongly decreased. HMP2 and CD distributions are wider than other datasets for all aggregation levels. **d**) Permanova p-values after variance decomposition analysis of pBCd matrices respectively to structuring co-variables. The permanova was performed for the different levels of aggregation and for the WH reconstruction. We can see that significance tends to decrease for higher aggregation levels, but the same level of significance is kept between AFT and WH, indicating that the same level of structure is kept after NMF decomposition. **e**) Qq-plots of AFT and reconstructed pBCd distributions. The dots indicate the distribution centiles. The reconstructed pBCd are computed on WH reconstructions including 1 $\left(w_{H_1} = w_1^{(AFT)} H_1^{(AFT)}\right)$, 2 $\left(w_{H_{12}} = \sum_{i=1}^{2} w_i^{(AFT)} H_i^{(AFT)}\right)$, 3 $\left(w_{H_{123}} = \sum_{i=1}^{3} w_i^{(AFT)} H_i^{(AFT)}\right)$ or the 4 profiles $\left(w_{H_{1234}} = \sum_{i=1}^{4} w_i^{(AFT)} H_i^{(AFT)}\right)$. The red line indicates the bisector line. We observe that profile 1 alone is not sufficient to reconstruct accurate pBCd, but that profiles 1 and 2 together allow the reconstruction of the main part of the pBCd distribution, for the lowest pBCd values. We can see that higher pBCd are not correctly rendered by the 2 profiles, especially for the CD cohort where dysbiotic samples are over represented. Adding the third and the fourth profile enables a correct reconstruction of the whole distribution, but with a homogeneous bias among the whole distribution.

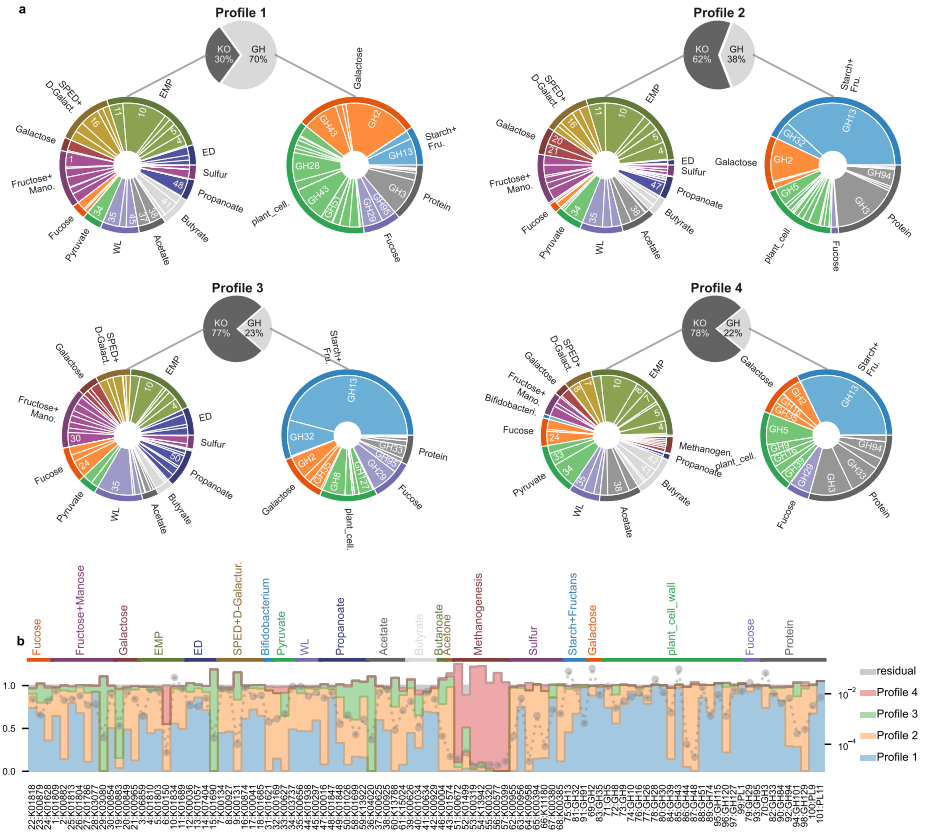


Figure 3: **Functional profile characterization.** **a**) KO and GH-related AFT frequencies are first gathered to show the distribution of KO and GH in each profile (top central pie chart). Then, the frequency of each AFT is renormalized by KO or GH/PL total frequency, and displayed in pie-charts for KO (left) and GH/PL (right) after clustering by functional modules (color coded, name displayed radially in the outer zone, see Fig. 1.a for the functional modules). The number of the KO or GH-related AFT is displayed in its corresponding pie-chart sector (radially, inner zone) when its frequency is higher than 3% in the profile. **b**) Average profile contribution in AFT counts. Average profile contribution for AFT j and profile i is computed as the proportion of average AFT counts provided by the profile i with $\bar{W}^{(AFT)}_{train,i} H^{(AFT)}_{ij} / \bar{X}^{(AFT)}_{train,j}$, where $\bar{W}^{(AFT)}_{train}$ and $\bar{X}^{(AFT)}_{train}$ are introduced in Methods sec. Inference of functional profiles. Finally, contributions are stacked by AFT in bar plots and ordered by functional modules. The residual $1 - \sum_{i=1}^{4} \bar{W}^{(AFT)}_{train,i} H^{(AFT)}_{ij} / \bar{X}^{(AFT)}_{train,j}$ is plotted in gray. Dotted gray lines indicate the value of $\bar{X}^{(AFT)}_{train,j}$ measuring the average AFT frequency (y log-scale on the right).

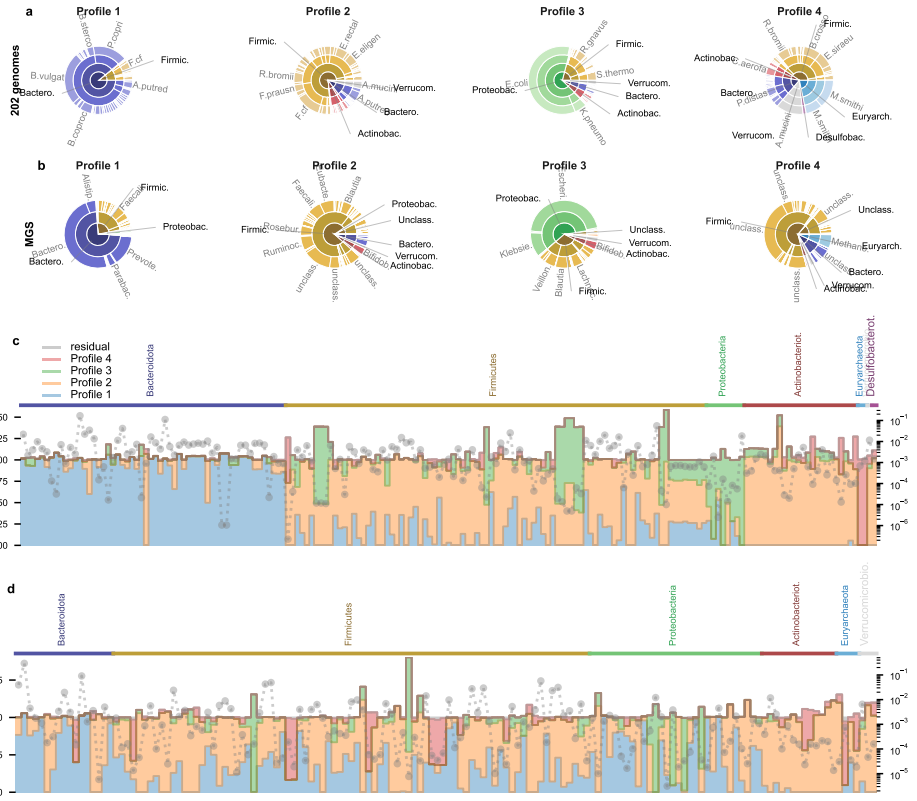


Figure 4: **Taxonomic profile characterization.** **a**) The 203 genomes frequencies in $H^{(PG)}$ are displayed in pie-charts and clustered by successive taxonomic levels, i.e. taxa (outer ring), genus, class and phyla (inner ring), color-coded by phyla (phyla name displayed radially in the outermost zone). Taxa names are displayed radially when their frequency is higher than 1% in the profile. **b**) The same procedure is applied on MGS clustered at the genus level. Taxonomic levels are genus, class and phyla. **c**) Average profile contribution in the 203 genomes counts. Namely, the same average profile weight $\bar{W}^{(PG)}_{train}$ as in Fig. 3 is computed together with $\bar{X}^{(PG)}_{train}$. Then, average profile contribution for genome j and profile i is computed with $\bar{W}^{(PG)}_{train,i} H^{(PG)}_{ij} / \bar{X}^{(PG)}_{train,j}$. Finally, contributions are stacked by genome in bar plots and ordered by phyla. The residual $1 - \sum_{i=1}^{4} \bar{W}^{(PG)}_{train,i} H^{(PG)}_{ij} / \bar{X}^{(PG)}_{train,j}$ is plotted in gray. Dotted gray lines indicate the value of $\bar{X}^{(PG)}_{train,j}$ measuring the average AFT frequency (y log-scale on the right). **d**) The same procedure is repeated on the MGS clustered at the genus level.

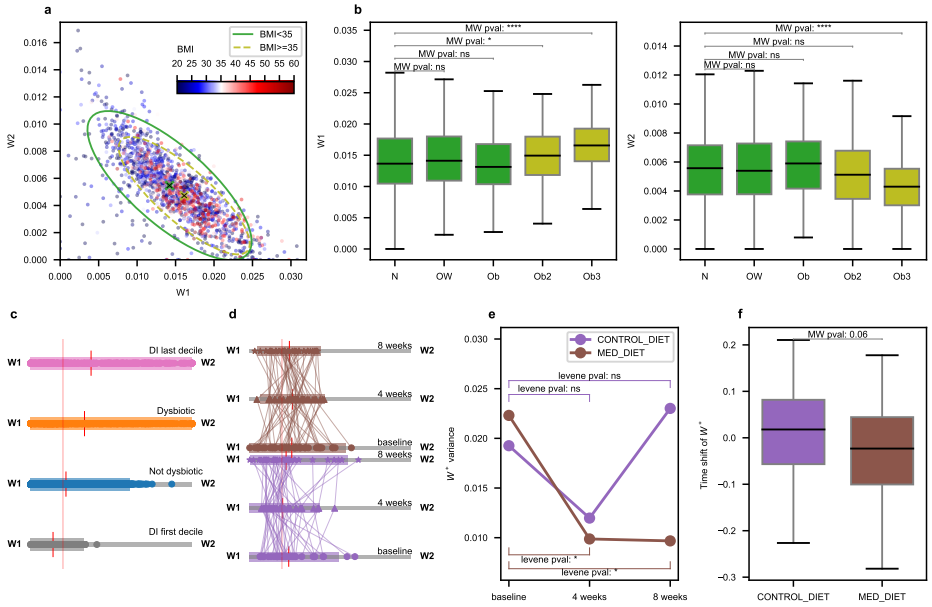


Figure 5: **W1 and W2 profile variations.** **a**) BMI. When BMI is available, samples are displayed in the W1-W2 space, coloured by BMI. 95% confidence ellipse are indicated for BMI lower and higher than 35 (class 2, severe obesity threshold). **b**) Obesity status. Boxplot of W1 and W2 levels structured by obesity status. We can observe that for highest obesity classes, W1 is significantly higher whereas W2 is significantly lower (MW = Mann-Withney test). This shift can be also observed in the confidence ellipse centroid in subfigure a. **c**) Dysbiosis index. All samples are displayed in barycentric coordinates in the W1-W2 space. Barycentric coordinates are equivalent to compute $W^* = W_2^{(AFT)}/(W_1^{(AFT)} + W_2^{(AFT)})$. The extremity $W_1^{(AFT)}$ corresponds to $W^* = 0$, i.e. when only Profile 1 is present in the sample, and the extremity $W_2^{(AFT)}$ corresponds to $W^* = 1$, i.e. when only Profile 2 is present. Samples are stratified by DI: the first DI decile (gray), non dysbiotic samples (blue, DI<dysbiosis threshold), dysbiotic samples (orange, DI>dysbiosis threshold) and last DI decile (pink) are displayed. The red ticks indicate the group mean, and confidence interval (mean +/- 2*standard deviation) is displayed with a coloured bar. The dotted red line indicate the value $W^* = 0.2$. We note a higher $W_1 - W_2$ unbalance for increasingly dysbiotic groups. **d**) Mediterranean diet. Samples are displayed in barycentric coordinates in the $W_1 - W_2$ space for the Mediterranean Diet cohort at baseline (circles) , 4 weeks (trianges) and 8 weeks (stars) after intervention for control (mauve) and Mediterranean diet (brown). The mean of each category is displayed with a red vertical line and confidence intervals are indicated as in plot c). The dotted red line indicate the value $W^* = 0.2$. We can observe that sample variability around the mean is strongly shrunk for the Mediterranean diet group after 4 weeks. **e**) Mediterranean diet stabilises the microbiota. The variance of $W_2/(W_1+W_2)$ in the control and Mediterranean diet groups is displayed at baseline, 4 weeks and 8 weeks. The variance decreases for Mediterranean diet after 4 and 8 weeks is significant (levene test). **e**) Time shifts of W^*. Time shifts, defined as the difference of $W^* = W_2/(W_1 + W_2)$ between 4 weeks and 8 weeks, are displayed with boxplots, for the Mediterranean and control diet groups. Time shifts are reduced after intervention for the Mediterranean group, with low significance ($p = 0.06$, Mann-Withney one-sided test).

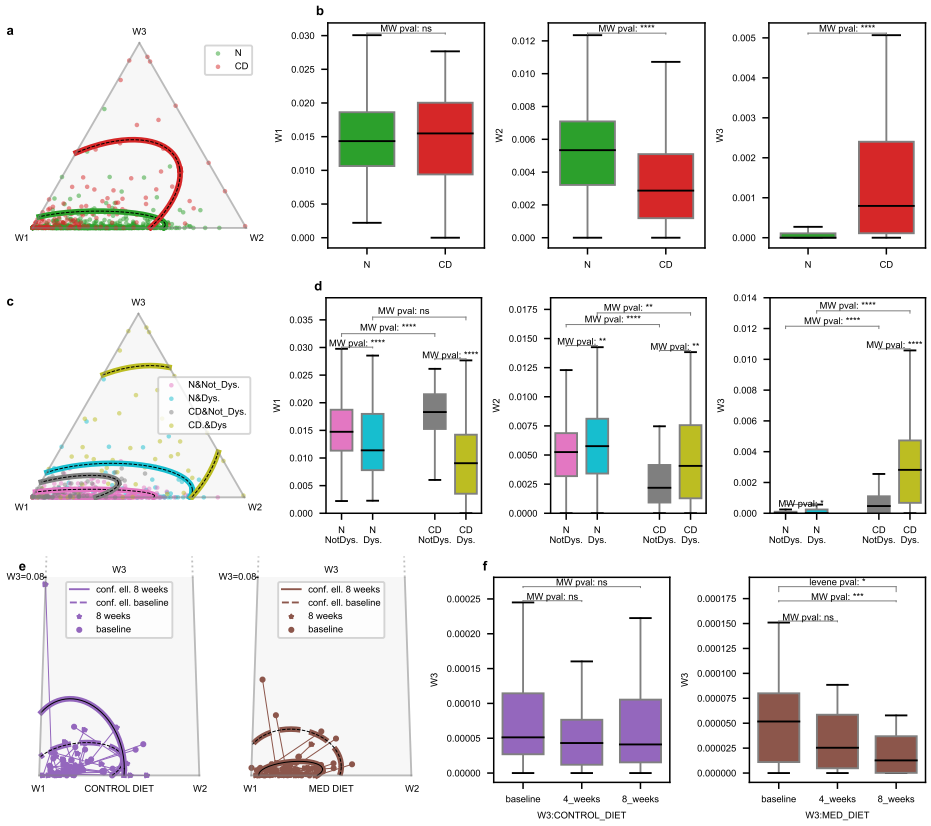


Figure 6: W_3 **profile variations associated to inflammatory status. a**) Crohn's disease. Ternary plot in the $W_1 - W_2 - W_3$ space of samples coloured by disease status (red: Crohn's disease (CD), green: Non-CD). 95% confidence area of each category are displayed with plain lines. **b**) Boxplot of W_1, W_2 and W_3 levels, structured by disease status. We can observe that CD samples have no marked difference in W_1 levels, but are characterized by significantly lower W_2 and strongly higher W_3 levels. This pattern differs from dysbiotic samples where W_2 were over-represented. This observation corroborates the shift of the confidence area in the ternary plot c). **c**) Unraveling dysbiotic and CD profiles. CD and non-CD (N) dysbiotic samples (left panel) and CD dysbiotic and not dysbiotic samples (right panel) are displayed in a ternary plot in the $W_1 - W_2 - W_3$ space. **d**) Boxplot of the W_1, W_2 and W_3 levels, structured by dysbiotic and CD status. **e**) Mediterranean diet. Ternary plots in the $W_1 - W_2 - W_3$ space of samples of the Mediterranean diet cohort. Control and Mediterranean diet groups are displayed in separated ternary plots. For a same individual, samples at baseline (circles) and 8 weeks after intervention (stars) are represented and linked by a line, showing the individual time trajectory. 95% confidence areas are displayed for baseline and 8 weeks groups. Ternary plots are clipped in the W_3 direction at W3=0.08. **f**) Boxplots of W_3 levels in the control and Mediterranean diet groups at baseline, 4 weeks and 8 weeks after intervention. W_3 mean and variance are significantly reduced after 8 weeks of Mediterranean diet.

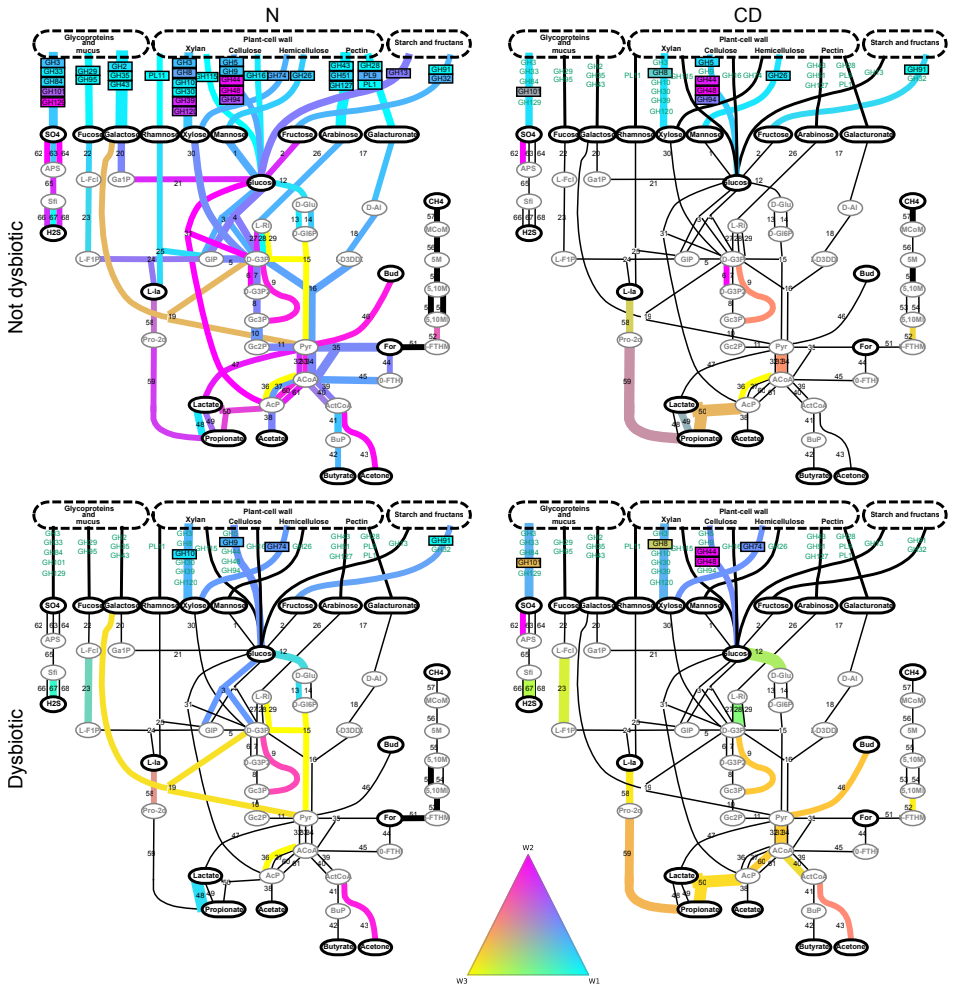


Figure 7: **Variation of profile contributions in healthy vs CD, and dysbiotic vs not dysbiotic samples.** The metabolic network of fibre degradation is displayed, and profile contribution in GH/PL and KO counts is colour coded on the corresponding arrows of the network. Profile contributions are displayed for healthy (N, left panel) and CD (right panel) samples, and dysbiotic (lower panel) and not dysbiotic (upper panel) samples. Namely, we compute CD and healthy average profiles weight $\bar{W}^{(AFT)}_{train,g}$ by averaging $W^{(AFT)}_{train}$ on the sample group g (N and dysbiotic, N and not dysbiotic, CD and dysbiotic, CD and not dysbiotic). Average AFT counts $\bar{X}^{(AFT)}_{train,g}$ are obtained in the same manner for each group. Then, average profile contribution for AFT j and profile i is computed with $\bar{W}^{(AFT)}_{train,g,i} H^{(AFT)}_{ij} / \bar{X}^{(AFT)}_{train,g,j}$. The respective relative contribution of Profile 1, 2 and 3 is then mapped into a ternary colour map (central triangle) and displayed on the corresponding arrow or GH/PL box. Black arrows indicate AFT the main contribution of which is given by Profile 4. Arrow widths are proportional to AFT counts in $\bar{X}^{(AFT)}_{train,g}$. For N & Not dysbiotic graph, all the AFTs are represented (control situation). For the other graph, the AFTs that significantly changed compared to N & Not dysbiotic group (t-test and Benjamini Hochberg correction with $FDR < 0.05$) were filtered; we then ordered AFTs by compositional changes compared to N&Not Dysbiotic group (L2 difference on $\bar{W}^{(AFT)}_{train,g,i} H^{(AFT)}_{ij} / \bar{X}^{(AFT)}_{train,g,j}$ computed on both groups) and kept the top 20 AFTs in order to highlight the main changes in microbiota composition.

Tables

Table 1: **Dataset overview.** We indicate for each dataset the number of samples n_s, individuals n_i, and if the dataset is used for DI, BMI, CD, statin, enterotypes, bristol score, diet or Parkinson studies.

	train	hmp2	CD	metacardis	med.diet	Parkin.
n_s	1126	1266	119	883	244	59
n_i	1126	106	119	883	82	59
DI	x	x	x	x	x	x
BMI	x	x	x	x		
CD	x	x	x			
statin				x		
enter.				x		
Bristol				x		
diet					x	
Parkin.						x

Additional Files

Additional file 1 — reconstruction error distribution of the AFTs

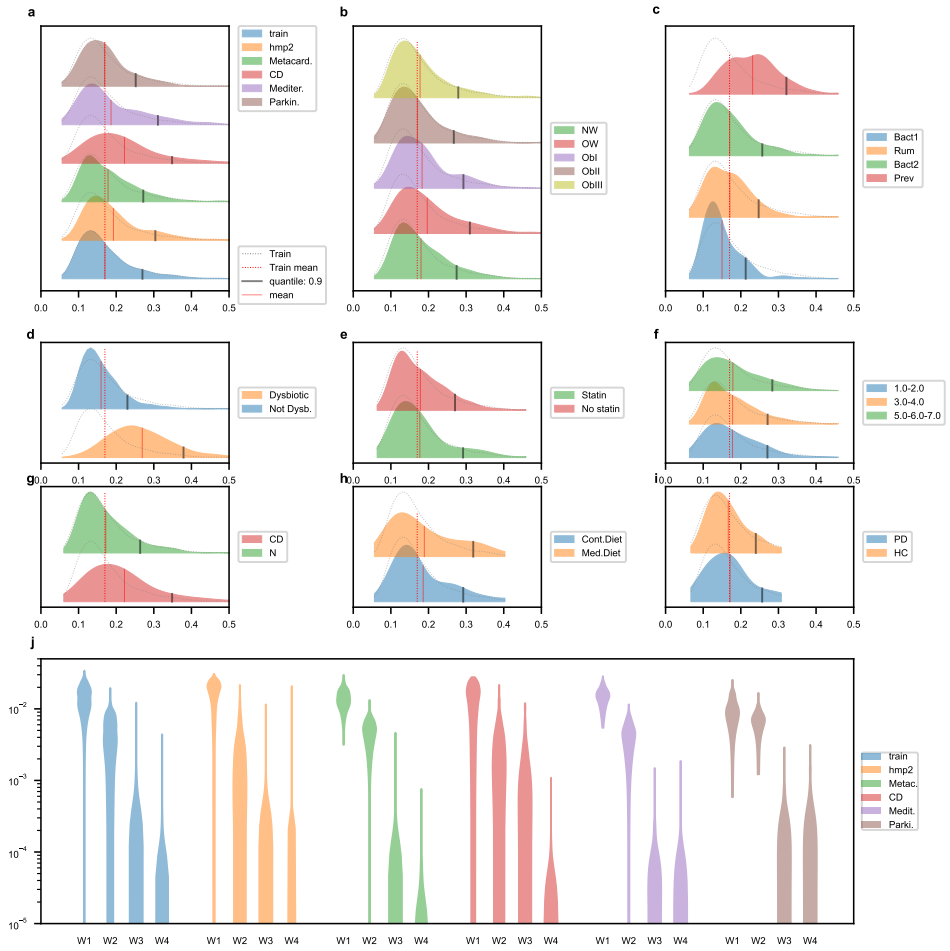


Figure S1: **AFT reconstruction error distribution and weight distribution.** The relative reconstruction error distribution among samples defined as $\|X_{g,i}^{(AFT)} - W_{g,i}^{(AFT)} H^{(AFT)}\| / \|X_{g,i}^{(AFT)}\|$ is displayed, and structured according to the different groups g encountered along the study i.e. **a**) datasets, **b**) obesity status, **c**) enterotypes, **d**) dysbiotic status, **e**) statin intake, **f**) Bristol score, **g**) Crohn disease status, **h**) mediterranean or control diet and **i**) parkinson disease. For comparison, the distribution observed in the train dataset is displayed in all graphs (gray dash lines), together with its mean relative reconstruction error (red dashed line). The mean and quantile 90% of each distribution are displayed with the vertical red and black lines. We can see that the relative reconstruction error distributions are very homogenenous along every structuring variables, except for dysbiotic and CD samples and Prevotela enterotypes, where relative reconstruction error is increased, but keeping the 95% quantile under 44% of reconstruction error. All together, the functional profiles allow to reconstruct the large majority of external samples with a level of accuracy comparable to the training dataset reconstruction, with a higher bias for dysbiotic, CD and Prevotela samples. **j**) The distribution of the weights $W_i^{(AFT)}$ are displayed for each dataset, with violin plots in log scale. Profiles 1 and 2 are preponderant, and Profile 3 and 4 are associated with lower weights.

Table 2: KO, GH, PL lists and dataset characteristics. The list of reactions corresponding to Figure 1 is displayed (top), with their corresponding KO (KEGG nomenclature). Then, GH and PL are listed (bottom).

Id	KO
1	K01809
2	K00882
3	K06859
4	K01810
5	K01803
6	K00150
7	K00134
8	K00927
9	K00131
10	K01834
11	K01689
12	K00036
13	K01057
14	K07404
15	K01690
16	K00874
17	K00041
18	K01685
19	K00883
20	K00849
21	K00965
22	K01818
23	K00879
24	K01628
25	K01813
26	K01804
27	K01786
28	K03077
29	K03080
30	K00854
31	K01621
32	K00169, K00170, K00171, K00172
33	K00627
34	K03737
35	K00656
36	K04020
37	K00625
38	K00925
39	K00626
40	K01034, K01035
41	K00634
42	K00929
43	K01574
44	K01938, K00288, K01491
45	K00297
46	K00004, K03366
47	K00016
48	K01847
49	K01848, K01849
50	K01026
51	K00672
52	K01499
53	K00319
54	K13942
55	K00320
56	K00577, K00578, K00579, K00580, K00581, K00582, K00583, K00584
57	K00399, K00401, K00402
58	K01699, K13919, K13920
59	K13922
60	K13788
61	K15024
62	K00955
63	K00956, K00957
64	K00958
65	K00394, K00395
66	K11180, K11181
67	K00380, K00381
68	K00385

GH/PL		
GH2	GH3	GH5
GH8	GH9	GH10
GH13	GH16	GH26
GH28	GH29	GH30
GH32	GH33	GH35
GH39	GH43	GH44
GH48	GH51	GH74
GH84	GH91	GH94
GH95	GH101	GH115
GH120	GH127	GH129
PL1	PL9	PL11

Additional file 2 — within-profile AFT frequencies.

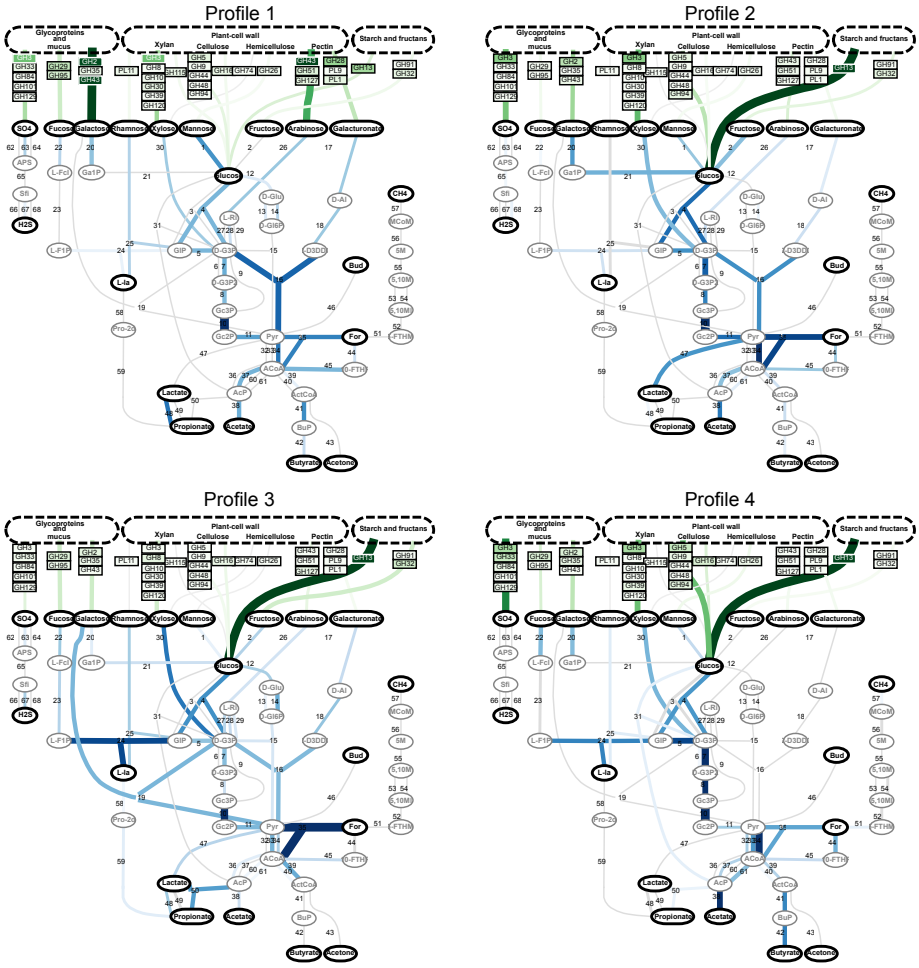


Figure S2: **Within-profile KO and GH frequencies mapped on the metabolic network.** For each profile, i.e. each row of the profile matrix $H^{(AFT)}$, we display the within-profile AFT frequencies, expressed for profile i and AFT j as $H_{ij}^{(AFT)} / \sum_{k \in KO} H_{ik}^{(AFT)}$ or $H_{ij}^{(AFT)} / \sum_{k \in GH} H_{ik}^{(AFT)}$ depending if j is a KO or a GH. The resulting value is plotted in green and blue color scales for GH and KO respectively, with edge width proportional to the value. These plots represent the intrinsic functional potential of the different profiles, like Fig. 3.a and unlike Fig. S3 or 3.b which represent the profile contributions to the metagenome.

Additional file 3 — Top functional and taxonomic profile contribution to metagenome.

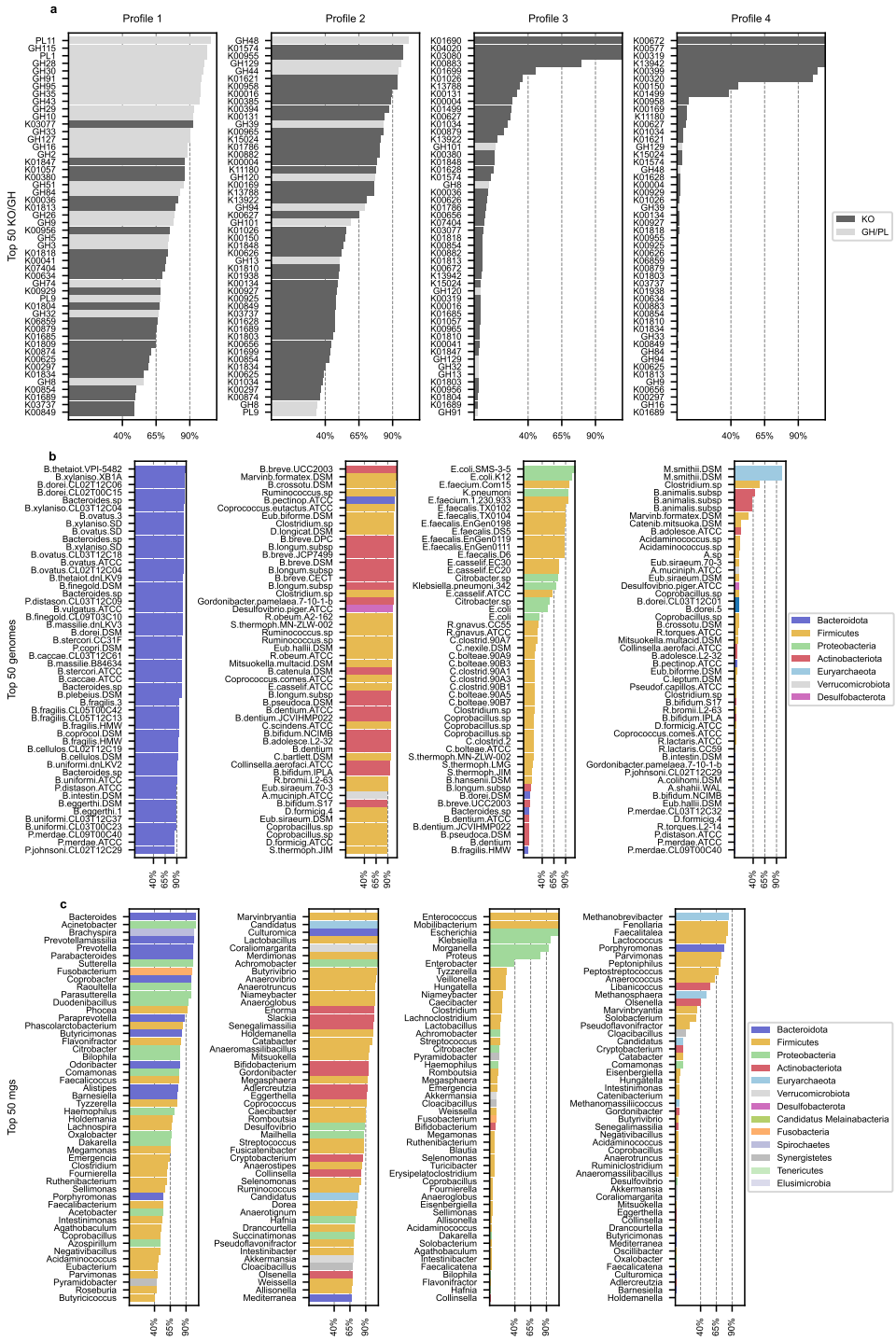


Figure S3: **Top functional and taxonomic profile contribution to metagenome.** The top 50 relative profile contribution to **a**) AFTs **b**) PGs and **c**) MGS-derived genus reconstruction are displayed. Namely, we compute for profile i and AFT or genome j the profile contribution $\bar{W}^{(AFT)}_{train,i} H^{(AFT)}_j / \bar{X}^{(AFT)}_{train,j}$ where $\bar{W}^{(AFT)}_{train}$ and $\bar{X}^{(AFT)}_{train}$ are averaged among the training samples. Then, contributions are sorted and top 50 contributions are kept and colorcoded by KO or GH for AFT, and phylum for PGs and MGS clustered by genus. Profile 1 is characterized by an over-representation of GH and *Bacteroidetes*, while Profile 2 is characterized by more KOs, and *Firmicutes* and *Actinobacteriota*.

Additional file 4 — Phyla-level reconstruction error distribution.

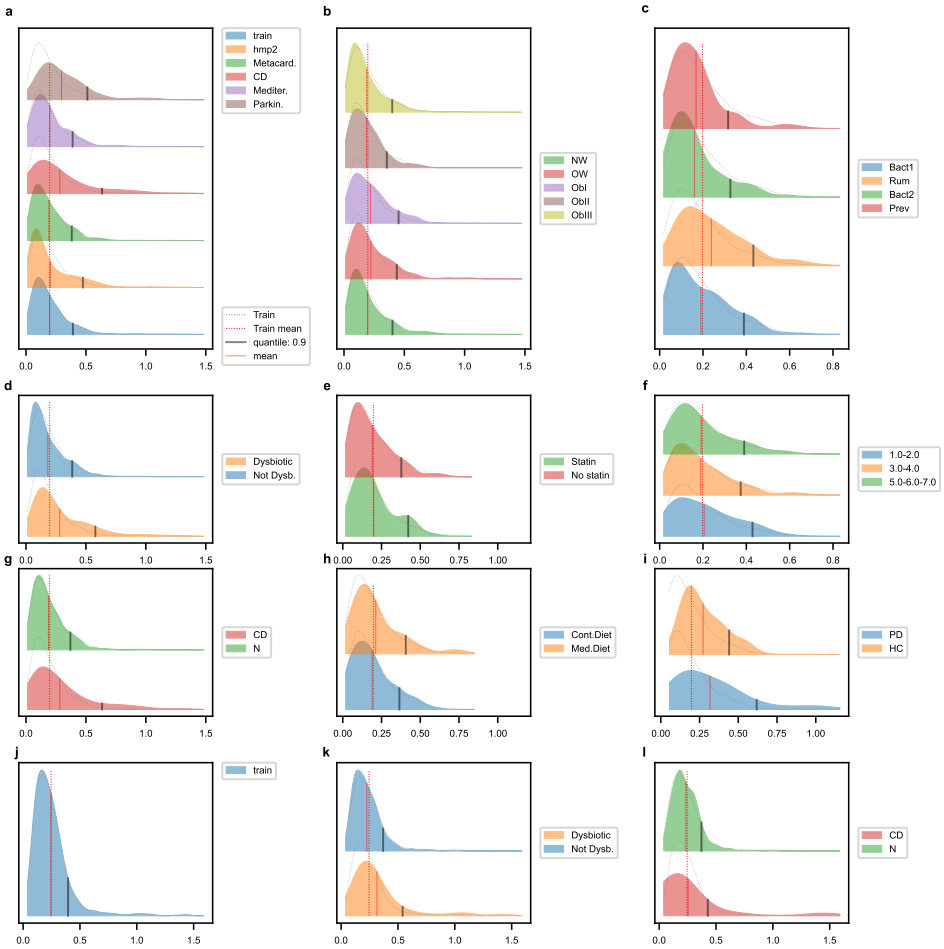


Figure S4: **Phyla reconstruction error distribution when reconstructing the PG counts.** The phyla relative reconstruction error distribution among samples defined as $\|(X_i^{(PG)} - W_i^{(AFT)} H^{(PG)}).A_{phyla}\| / \|X_i^{(PG)}.A_{phyla}\|$ is displayed, where $X^{(PG)}$ is the count matrix of the 203 representative genomes and A_{phyla} is an allocation matrix of each genome to its phyla, and structured according to the different classes encountered along the study, i.e. **a**) datasets, **b**) obesity status, **c**) enterotypes, **d**) dysbiotic status, **e**) statin intake, **f**) Bristol score, **g**) Chron disease status, **h**) mediterranean or control diet and **i**) parkinson disease. For comparison, the distribution observed in the train dataset is displayed in all graphs (gray dash lines), together with its mean relative reconstruction error (red dashed line). The mean and quantile 90% of each distribution are displayed with the vertical red and black lines. We can see that the relative reconstruction error distributions of the phyla are very homogenenous along every structuring variables, except for dysbiotic, CD and Parkinson disease samples, where relative reconstruction error is increased. Like for AFTs, the functional profiles allow to reconstruct the taxonomic composition of the large majority of external samples at the phyla level with a level of accuracy comparable to the training dataset reconstruction. **j**) MGS. The same procedure is repeated with MGS. Namely, $\|(X_i^{(mgs)} - W_i^{(AFT)} H^{(mgs)}).A_{phyla}\| / \|X_i^{(mgs)}.A_{phyla}\|$ is displayed, where $X^{(mgs)}$ is the MGS count matrix and A_{phyla} is an allocation matrix of each MGS to its phyla and structured according to the different classes encountered in the 'train' test, i.e. **k**) dysbiotic status and **l**) Chron disease status. The MGS count matrix are correctly reconstructed, whatever the structuring variable.

Additional file 5 — Characterization of dysbiosis, enterotypes and statin related samples using the profiles.

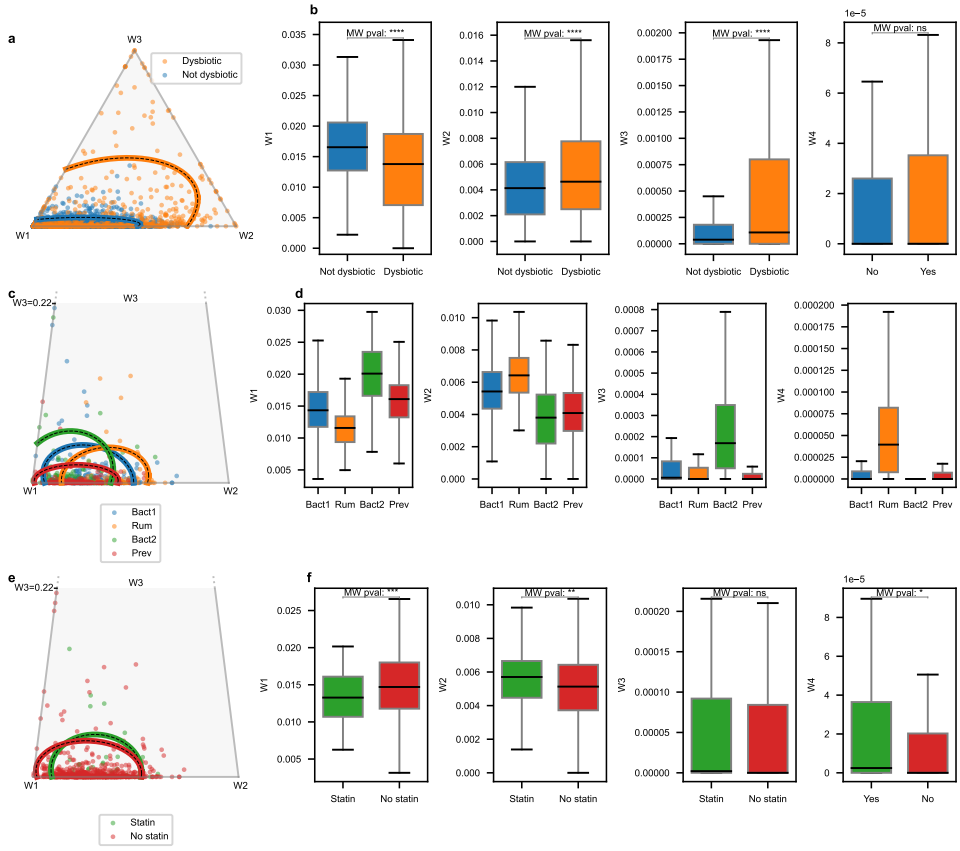


Figure S5: **Characterization of dysbiosis, enterotypes and statin related samples. a**) Dysbiosis. Ternary plot in the $W_1 - W_2 - W_3$ space of samples coloured by dysbiotic status. We also display the 95% confidence area for each category (coloured line). **b**) Boxplot of $W_1^{(AFT)}$, $W_2^{(AFT)}$ and $W_3^{(AFT)}$ levels, structured by dysbiotic status. We can observe that dysbiotic samples are characterized by significantly lower $W_1^{(AFT)}$, higher $W_2^{(AFT)}$ and strongly higher $W_3^{(AFT)}$ levels. This information corroborates the much wider confidence area for dysbiotic samples in the ternary plot a). **c**) Enterotypes. Samples are displayed in a ternary plot in the $W_1 - W_2 - W_3$ space, coloured by enterotype, when available. 95% confidence ellipses of each class are displayed. **d**) Boxplot of $W_1^{(AFT)}$, $W_2^{(AFT)}$ and $W_3^{(AFT)}$ levels, structured by enterotypes. We can observe that Ruminococcus enterotype is overrepresented for higher $W_2^{(AFT)}$ and lower $W_1^{(AFT)}$. The reverse observation can be made for Bact2 enterotype. To a lower extent, Bact1 enterotype is more prevalent for lower $W_1^{(AFT)}$ and higher $W_2^{(AFT)}$, which is the inverse of Prevotella enterotype. High $W_3^{(AFT)}$ counts are related to Bact2 enterotype. **e**) Statin. Ternary plot in the $W_1 - W_2 - W_3$ space, coloured by statin intake, together with 95% confidence ellipses. **f**) Boxplot of $W_1^{(AFT)}$, $W_2^{(AFT)}$ and $W_3^{(AFT)}$ levels structured by statin intake. $W_1^{(AFT)}$ is significantly lower for statin intake, whereas $W_2^{(AFT)}$ is significantly higher. No significant shift is observed for $W_3^{(AFT)}$.

Additional file 6 — CD-related profile characterization.

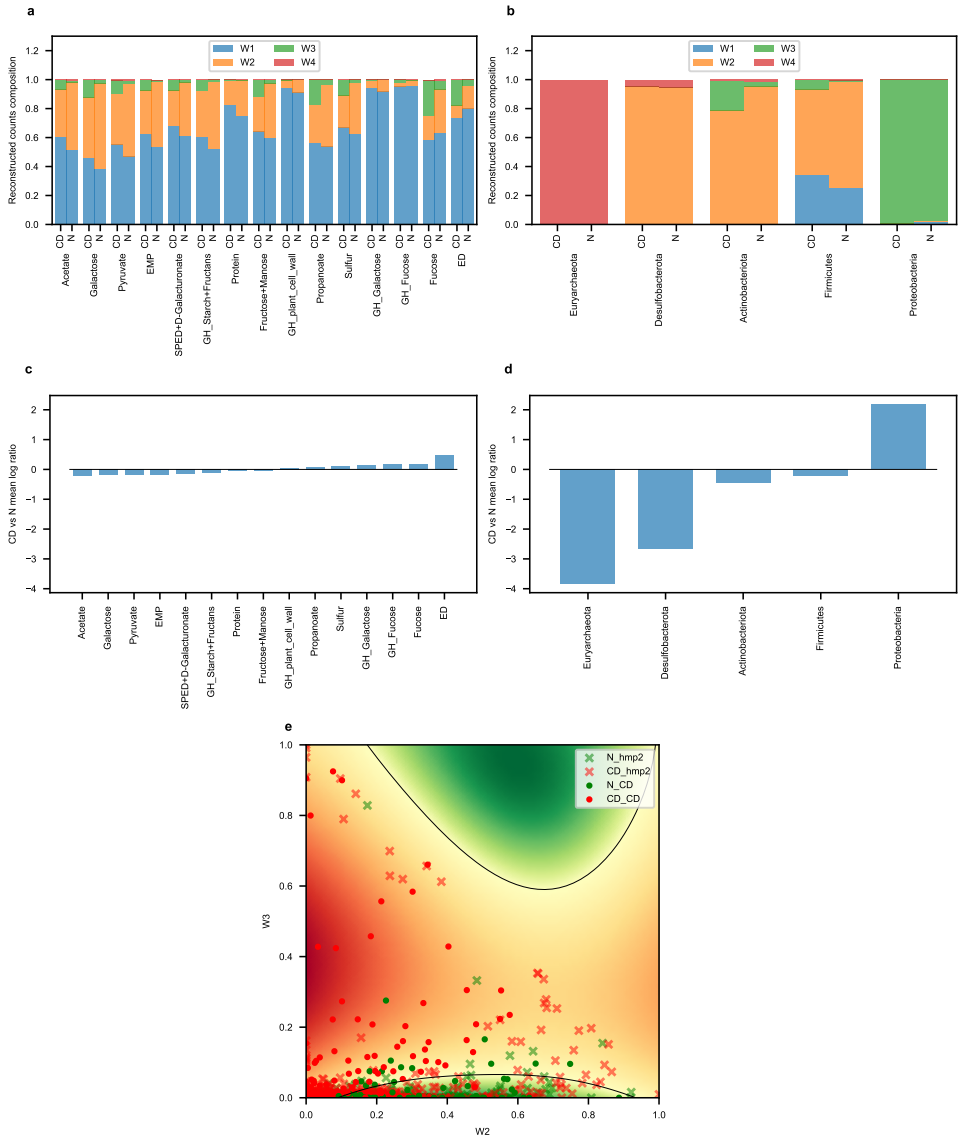


Figure S6: **CD-related profile characterization.** **a**) Functional differential analysis between CD and healthy samples (N). Average profile contribution in the significantly different functional module frequencies for CD and N groups. Functional modules are defined in fig. 1. a. We averaged the L_1 normalized $W^{(AFT)}$ (resp. $X^{(AFT)}$) for the CD and N groups of the train dataset, noted $\bar{W}^{(AFT)}_{train,L_1,g}$ (resp. $\bar{X}^{(AFT)}_{train,L_1,g}$) for $g = CD$ or N, and computed $\bar{W}^{(AFT)}_{train,L_1,g}H^{(AFT)}$. We then gathered the columns of $\bar{X}^{(AFT)}_{train,L_1,g}$ by functional modules and filtered functions with significant changes (t-test, 0.05 fdr Benjamini-hochberg correction) between N and CD groups. For selected modules, we computed $\bar{W}^{(AFT)}_{train,L_1,g,I}H_{Ij}/\sum_i \bar{W}^{(AFT)}_{train,L_1,g,i}H_{ij}$, for profile I, group $g = CD$ or N, and functional module j, displayed in barplots, in order to display profile contribution for each functional module. **b**) Taxonomic differential analysis between CD and healthy samples. The same procedure is repeated on phyla counts. After computing $\bar{X}^{(PG)}_{train,L_1,g}$ and pooled representative genome counts by phyla, the significantly varying phyla (t-test, 0.05 fdr Benjamini-hochberg correction) between N and CD groups are filtered. Then, $\bar{W}^{(PG)}_{train,L_1,g,I}H^{(PG)}_{Ij}/\sum_i \bar{W}^{(PG)}_{train,L_1,g,i}HPG_{ij}$, for profile I, group $g = CD$ or N, and functional module j, is displayed in barplots, in order to display the profile contribution to the reconstruction of each phyla. **c**) Log2 ratio between CD and N groups of filtered functional groups are displayed. **d**) Log2 ratio between CD and N groups of filtered phyla are displayed. Whereas functional variations are limited, taxonomic variations are more acute. **e**) Classification of CD samples. The SVM classifier for CD/N, trained on the hmp2 cohort, is displayed in the $W_2 - W_3$ space (normalised with min-max scaling). The black line separates the negative (green) and the positive region (red). The samples of the 'hmp2' (train, crosses) and 'CD' (test, circles) cohorts are displayed, coloured by disease status. We observe that $W_2 - W_3$ variations for CD samples are strong enough to capture this signal with a classifier (recall: 0.94, precision: 0.81, AUC: 0.92 for the CD unseen

Additional file 7 — Prevalent genomes functional profiling.

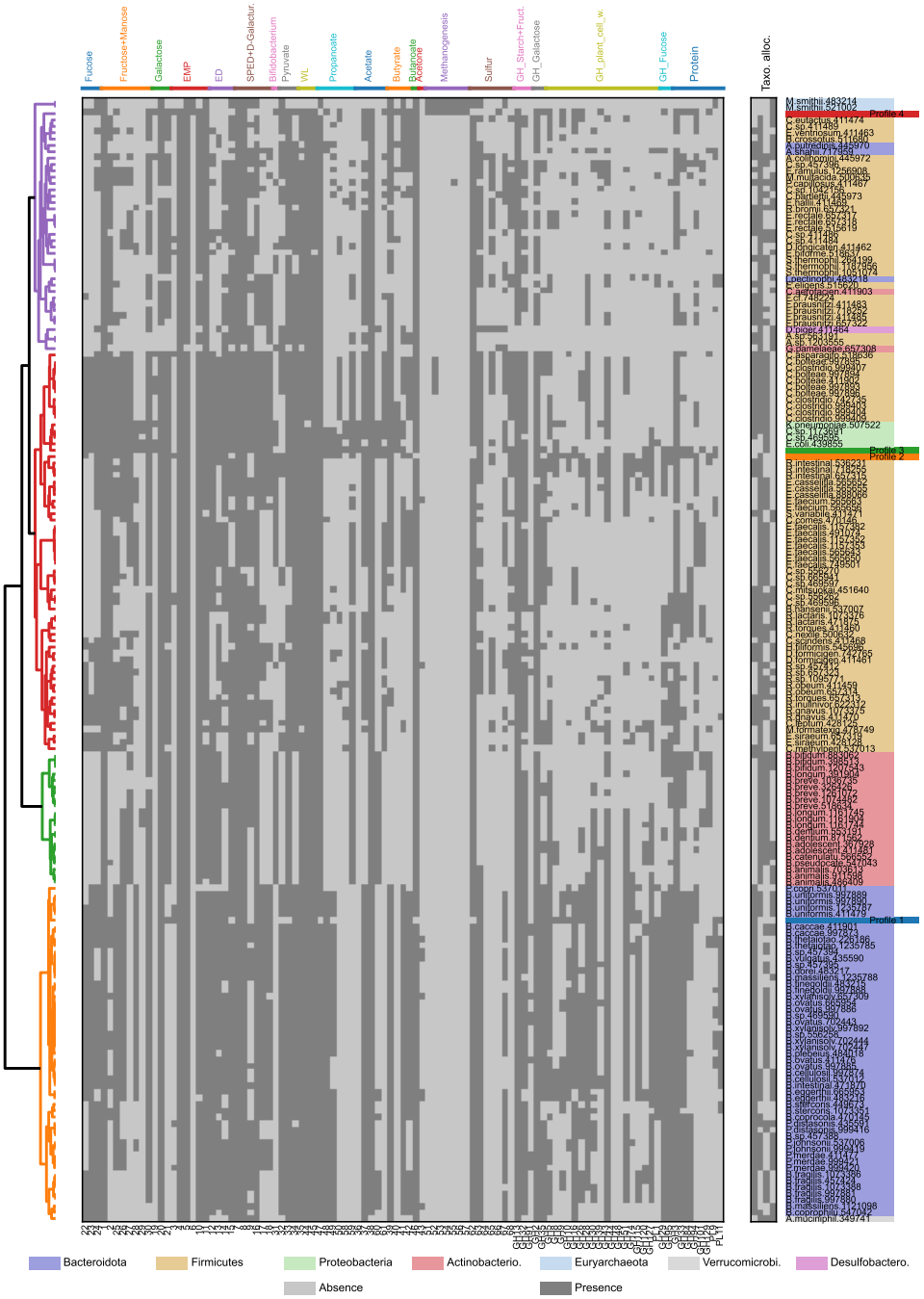


Figure S7: **Prevalent genomes functional profiling.** Selected AFTs are annotated in the prevalent genomes and presence/absence is displayed (middle panel), sorted by functional blocks (top). The genome names are indicated (short name and NCBI ID, right panel), colorcoded by phylum, and the the taxonomic allocation of the genomes in profiles is indicated by the presence/absence matrix in $H^{(PG)}$ (right panel, taxo. alloc. Profile i is the i-th column of this matrix). The 4 profiles are added to the genome list and displayed with presence/absence tags (a KO is assumed present in the profile if its frequence is higher than $1e-3$). Hierarchical clustering is performed ($k=4$ clusters), based on pairwise-Jaccard distance computed on AFT presence/absence matrix (corresponding dendogram in the left panel), and genomes are sorted accordingly in the middle and right panels. We note that the 4 profiles are clustered at the same time than the genomes. *Bacteroidetes* and *Actinobacteria* are gathered into their own cluster (orange and green clusters), whereas *Firmicutes* are splitted in two clusters: the main part is clustered with *Proteobacteria* (red), while the others are clustered with less prevalent phyla such as *Desulfobacterota* or *Euryarchaeota* (purple), the separation being based on difference in presence of ED and SPED-related AFTs. Profile 1 clusters with *Bacteroidetes*, consistantly with its taxonomic profiling (Fig. 4). This cluster is marked by higher presence of GH and PL, consistantly with its functional profiling (Fig. 3). Profile 2 and 3 cluster with *Firmicutes* (red cluster), Profile 3 being included in a sub-cluster involving *Proteobacteria*, again consistantly with their respective taxonomic profiling (Fig. 4). Profile 4

Additional file 8 — Profile 4 association with Bristol score and Parkinson's disease.

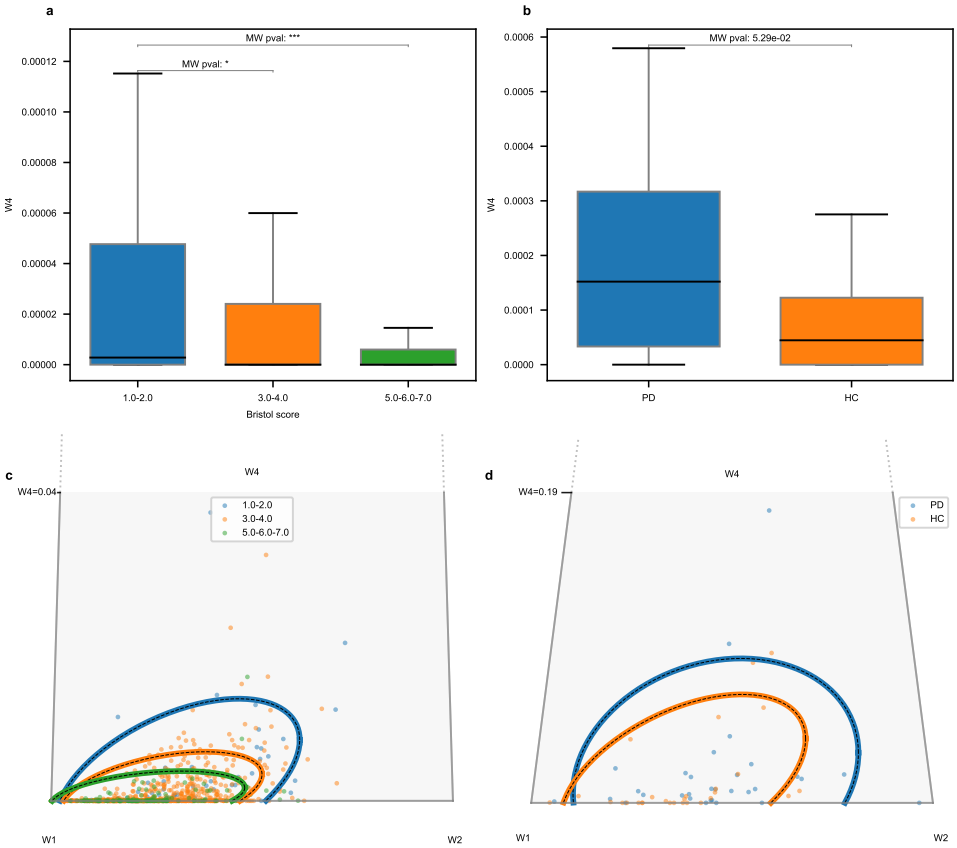


Figure S8: **Profile 4 association with Bristol score and Parkinson's disease.** **a**) Bristol score. Boxplot of $W_4^{(AFT)}$ levels, structured by Bristol stool score. **b**) Parkinson's disease. Boxplot of $W_4^{(AFT)}$ levels in PD and healthy control samples. We can observe that the significance of the difference between groups is slight ($p = 5.3e-2$, MW test) **c**) Ternary plot in the $W_1 - W_2 - W_4$ space, coloured by Bristol stool score. 95% confidence ellipses of each class are displayed. **d**) Ternary plot in the $W_1 - W_2 - W_4$ space of PD and healthy control samples. 95% confidence ellipses of each class are displayed.

Additional file 9 — Dataset count matrices, profile decomposition and metadata.

The input data needed for the analysis are provided.

- *Metadata.xlsx*: excel file containing the metadata of all the datasets used in the analysis. Field definition:
  - Sample_ID : internal ID.
  - ProjectID : study accession number
  - SRA : SRA sample accession ID
  - Patient_ID : internal patient ID
  - Nationality, sex, age, BMI : patient nationality, sex, age and BMI.
  - Diagnosis : internal diagnosis code. N = healthy, CD = Crohn's disease, UC = Ulcerative colitis, Control = Control sample, ObCIII = class 3 obesity, ObCII = class 2 obesity, ObCI= class 1 obesity, OW = overweight (but not obese), UW = underweight, PD = Parkinson disease, HC = healthy control, Diab = diabetis, ankylosing_spondylitis = ankylosing spondylitis
  - Dysbiosis_index : dysbiosis index computed from HMP2 samples (see material and methods), Is_dysbiotic = above or under dysbiotic threshold (see material and methods).
  - Study : internal study ID.
  - reference : doi of the related publication.
  - alias : internal alias of the sample (HMP2 dataset only).
  - enterotype : sample enterotypes (metacardis dataset only)
  - Statin : statin intake (metacardis dataset only).
  - Bristol : Bristol score (metacardis dataset only).
  - Diet : diet taken by the patient (control or Mediterranean diet, med_diet dataset only).
  - Timepoint : baseline, 4_weeks, 8_weeks (med_diet dataset only).
- *W.xlsx*: weight matrix for the different datasets.
- *X_AFT.xlsx*: AFT count matrix for the different datasets. The header indicates the AFT names as in Table 2.
- *X_mgs.xlsx*: MGS count matrix at the genus level (train dataset only).
- *X_pg.xlsx*: Prevalent genomes count matrix for the different datasets. The first sheet indicates the NCBI taxonomy ID and the name of the 203 prevalent genomes included in the study.
- *Genome_list.xlsx*: list of the 203 genomes included in the study.
- *H.xlsx*: matrices $H^{(AFT)}$, $H^{(PG)}$ and $H^{(mgs)}$.
- *List_of_Reactions.xlsx*: List of reactions as indicated in Fig. 1 and Table 2, with complete aggregated biochemical reaction, and reactant names.
- *F.xlsx*: matrix of metabolic constraints used during NMF.

Additional file 10 — Supplementary materials.

This document recapitulates additional precisions on the material and methods involved in this study.

Additional methodological details

We further detail here methodological steps of the method.

KO selection

We recall here the method presented in [24]. From the pathways selected, a list of putative reactions was compiled. For each reaction, a KEGG database reaction entry was retrieved. Since each reaction is catalyzed by enzymes linked to KO, a list of putative KO was obtained. The manual curation of the KO candidates followed the rules: (1) a KO not found in the IGC annotation was ignored, (2) a reaction that was not linked to a KO was ignored, (3) when an enzyme from a KO could catalyze more than one reaction, because we could not accurately link a KO frequency to a target reaction, all the KO associated to the target reaction were ignored. Exceptions were made regarding key-reactions. Multiple KO associated to a unique reaction were kept since they correspond to different enzymes catalyzing the same chemical reaction in different species (L-ribulose to G-Gly 3 Phosphate, Acetyl-CoA to Acetyl-Phosphate, Lactate to propionate) or different subunits of the same enzyme, such as K01034 and K01035. Reactions associated with microorganisms unlikely present in the gut microbiomes, such as aerobes from soil, halophilic extremophiles, were excluded. Reactions associated to micro-aerophilic or facultative anaerobes were kept. KO from very low abundant microorganism from the gut microbiome were kept.

Rationale and assemblage for the constraint matrix F

The metabolites in the model were parted between those that were known to be extracellular (gathered in a set noted E, and displayed with bold box in Fig. 1.A) and the others (not known to be extracellular, gathered in a set NE, and gathered with gray box in Fig.1.A). For a metabolite $m \in NE$, we considered all the reactions in our list that could produce m and gathered all the associated traits in a set called P_m. In the same way, all the traits involved in reactions that could consume m were gathered in a set called C_m. Each profile (line of H) was constrained so that the total sum of producing and consuming can not be simultaneously null. Namely, for profile number l, we state that

$$\sum_{j \in C_m} H_{lj} \leq \alpha_m^+ \left(\sum_{j \in P_m} H_{lj} \right) \tag{4}$$

$$\sum_{j \in P_m} H_{lj} \leq \alpha_m^- \left(\sum_{j \in C_m} H_{lj} \right) \tag{5}$$

for given coefficients α_m^+ and α_m^- to be defined.

The rationale of these constraints is to prevent the accumulation of intracellular compounds, and therefore if there is a potential in the profile for producing a metabolite, the same profile should also carry a functional potential for using it, and conversely. The bounds α_m^+ and α_m^- were derived from the analysis of the 190 prevalent genomes. In a nutshell, metabolites were constrained only if bounds α_m^+ or α_m^- could be found such that Eq. 4 or 5 are satisfied for more than 95% of the 190 genomes when replacing the H_{lj} by the corresponding KO frequencies in each genomes. Moreover, a security margin was taken on the values of α_m^+ and α_m^- to account for a possible discrepancy between the 190 representative genomes and the full microbiota. See [24] for extended justification of these constraints and a study of their impact on a toy model and a real case study.

Hyper-parameter selection procedure

Following [24], we follow a three-step hyperparameter selection procedure based on (1) a reconstruction error criteria, (2) a bi-cross validation and (3) a criteria based on the stability of the inferred H.

Reconstruction error. The relative reconstruction error criteria measures the proportion of information recovered by the NMF decomposition

$$C_{rec.err.} = \frac{\|X_{train}^{(AFT)} - W_{train}H^{(AFT)}\|_F}{\|X_{train}^{(AFT)}\|_F}.$$

This criteria mechanically decreases with the number of profiles k, thus a slope discontinuity on the graph is searched for, indicating that additional profiles carry less information and mainly approximate noise.

Bi-cross validation criteria. Starting from $X_{train}^{(AFT)}$ of size $n_s = 1126$ samples times $n_{AFT} = 101$ AFTs, we set a 5-fold random splitting l_s and l_{AFT} of respectively the set of samples and AFTs. Taking $I \in l_s$ and $J \in l_{AFT}$ two index subsets taken from the 5-fold splitting, we note

$$X_{11}^{IJ} = ({X_{train}^{(AFT)}}_{kl})_{k\in I, l\in J}, \quad X_{12}^{IJ} = ({X_{train}^{(AFT)}}_{kl})_{k\in I, l\notin J},$$
$$X_{21}^{IJ} = ({X_{train}^{(AFT)}}_{kl})_{k\notin I, l\in J}, \quad X_{22}^{IJ} = ({X_{train}^{(AFT)}}_{kl})_{k\notin I, l\notin J},$$

We then note W_1^{IJ}, H_1^{IJ} the solution of the NMF decomposition of X_{11}^{IJ}

$$(W_1^{IJ}, H_1^{IJ}) = \underset{\substack{W\geq 0\\ H\geq 0}}{\arg\min} \|(X_{11}^{IJ} - WH)D^{-1}\|_F^2 + \alpha\left(\|W\|_F^2 + \|HD^{-1}\|_{1,2}^2\right)$$

We note that this NMF is unconstrained since the undersampling breaks up the structure of the constraints.

We note W_2^{IJ} and H_2^{IJ} the respective non-negative least-square regression of X_{21}^{IJ} and X_{12}^{IJ} of the (unconstrained) problems

$$W_2^{IJ} = \underset{W\geq 0}{\arg\min} \|(X_{21}^{IJ} - WH_1)D^{-1}\|_F^2 + \alpha\left(\|W\|_F^2\right)$$

and

$$H_2^{IJ} = \underset{H\geq 0}{\arg\min} \|(X_{12}^{IJ} - W_1H)D^{-1}\|_F^2 + \alpha\left(\|HD^{-1}\|_{1,2}^2\right)$$

The criteria is the average over I and J of the relative reconstruction error of X_{22}^{IJ}

$$C_{bi-cross} = \frac{1}{|l_s| \times |l_{AFT}|} \sum_{I\in l_s} \sum_{J\in l_{AFT}} \frac{\|X_{22}^{IJ} - W_2^{IJ}H_2^{IJ}\|_F}{\|X_{22}^{IJ}\|_F}$$

Stability. After splitting the training set in two balanced random subsets X_1^J and X_2^J of $X_{train}^{(AFT)}$, $1 \leq J \leq 20$ being the index of the splitting in 20 splitting repetition, a constrained NMF is performed on X_1^J and X_2^J to get (W_1^J, H_1^J) and (W_2^J, H_2^J). To assess the similarity between profiles, we compute the similarity matrix for $I = 1, 2$ of dimension $n_{AFT} \times n_{AFT}$

$$S_{I\,lm}^J = \frac{\sum_{i=1}^k H_{I\,il}^J H_{I\,im}^J}{\left(\sum_{i=1}^k H_{I\,il}^{J\,2}\right)^{1/2}\left(\sum_{i=1}^k H_{I\,im}^{J\,2}\right)^{1/2}}, \quad \text{for } 1 \leq l, m \leq N_{AFT}$$

The criteria is finally

$$C_{stability} = 1 - \frac{1}{20\sqrt{n_{AFT}(n_{AFT}-1)}} \sum_{J=1}^{20} \|S_1^J - S_2^J\|_F$$

Next, we compute the different criteria in a grid with $\alpha \in \{0.001, 0.01, 0.03162, 0.1, 1\}$ and $k \in \{2, 4, 6, 8, 11\}$ (see S9) and selected $\alpha = 0.03162$, providing the minimal value

for the bi-cross validation and equivalent values for the other criteria. Next, in order to have a deeper accuracy on the selection ot the number of profiles, we computed the criteria for $\alpha = 0.03162$ and $k = 2, \cdots, 12$. We selected $k = 4$ due to the clear slope break for the stability criteria and un-degraded accuracy for the other criteria.

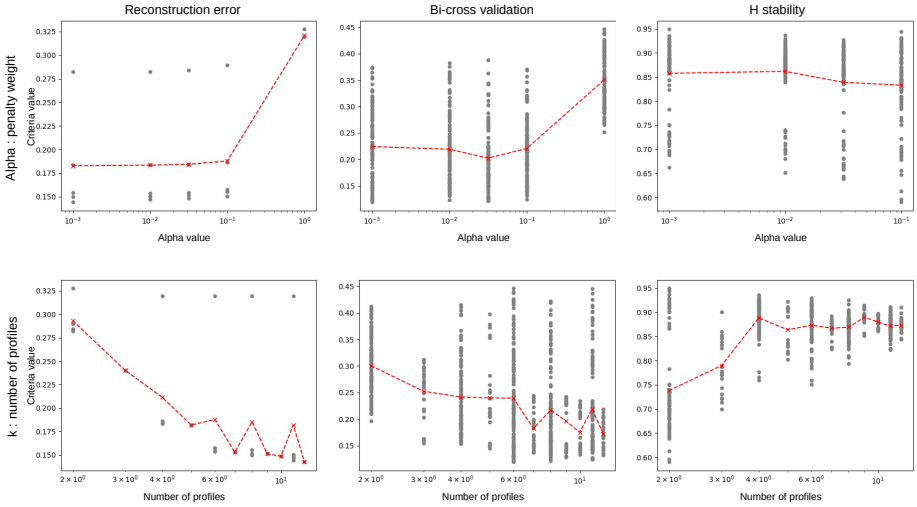


Figure S9: **Hyperparameter selection.** The results of the hyperparameter selection procedure are displayed for α selection (upper panel) and the number k of profiles (lower panel) for the reconstruction error, the bi-cross validation and the H stability criteria. This procedure leads to select $k = 4$ profiles and a value of $\alpha = 0.03162$