Supplementary Information

Supplementary Methods.

Total DNA and RNA extraction and sequencing

Fecal samples were stored in RNA later (Life Technologies, Carlsbad, CA) at -80 ºC until use. Total DNA was extracted as previously described (Serrano-Villar et al., 2017). Briefly, the samples, diluted in phosphate-buffered saline, were centrifuged at 2,000 r.p.m. at 4 ºC for 2 min to remove fecal debris. Total DNA was extracted from the pelleted cells (roughly 50 mg) with QIAamp DNA Stool Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The metagenomes were sequencing using both MiSeq platform (Illumina, Eindhoven, Netherlands) and Genome FLX Sequencer (Roche Diagnostics, Branford, CT, USA).

Total RNA from the pelleted cells (roughly 100 mg) was extracted using the RiboPure™ Bacteria kit (Ambion, Austin, TX, USA) as described in (Pérez-Cobas et al., 2013) and then treated with Baseline-ZEROTM DNAse (Illumina, Eindhoven, Netherlands). The rRNA removal was performed using Ribo-Zero Magnetic Kit (Bacteria) (Illumina, Eindhoven, Netherlands). Then, the rRNA-depleted RNA was processed with ScriptSeq V2RNA-Seq Library Preparation kit (Illumina, Eindhoven, Netherlands) to generate the double strand cDNA and the sequencing libraries according to the manufacturer's instructions. The metatranscriptomes were sequenced using the Kit V3 (2x230 cycles) in MiSeq platform (Illumina, Eindhoven, Netherlands).

For the metagenome analysis, we sequenced an average 98 Mb per sample (total 9.5 Gb) using a combination of Illumina and 454 sequencing technologies. The coverage of each metagenome was estimated using the Nonpareil software (Rodriguez-R and Konstantinidis, 2014) using the options (-T alignment -f fasta), resulting in a coverage mean of 42.04% (sd=17.09). The metatranscriptome was obtained in a subset of 41 samples (8 VU, 13 IR, 6 INR and 14 HIV-). The total mRNA was purified and the cDNA sequenced, obtaining an average 16 Mb per sample (a total of 715 Mb).

Initial processing, assembly and functional annotation

Illumina paired-end sequences retrieved from metagenomic (DNAseq) and metatranscriptomic (RNAseq) sequences were trimmed for adapters and the transposase sequence using the cutadapt software (v1.6). Sequence artifacts and low quality reads

from the raw sequences were removed using the fastx_artifacts_filter software and the fastq_quality_trimmer (parameters: -t 20 -l 50) available into the fastxtoolkit. Then, we used the prinseq-lite.pl script (parameters: -min_len 50 -min_qual_mean 20 -ns_max_n 4 -trim_qual_left 20 -trim_qual_right 25 -trim_qual_type min -trim_qual_rule lt 5-trim_qual_window 5 -trim_qual_step 1) to trim the 5' end and the 3' ends of each sequence, removing the trimmed reads that did not achieve one of the following criteria: read length above the 170 nucleotides, mean quality score above 20 or contain less than the 1% of ambiguous base.

The possible reads derived from the host genome was removed using the deconseq.pl script (v0.4.3) using as reference the human genome (Human Reference GRCh38).

Taxonomic assignment

First, we constructed our own non-redundant genome database. The sequences of the reference genomes from the species found in the gastrointestinal tract were downloaded from the HMP web site (http://downloads.hmpdacc.org/data/reference_genomes/body_sites/Gastrointestinal_tract.nuc.fsa). Similarly, all the reference genomes and the latest assemblies of the genomes of the species related to the HMP genome database were downloaded from NCBI RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/genomes/all) using the assembly_summary.txt file as a reference (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt). The resulting database contains 1,113,533 sequence entries in 39, 537, 222, 120 bp. The sequences that share the same taxa-id were grouped in a fasta file. Each file was then compared against itself using megablast (parameters: blastn -task megablast -perc_identity 95). Those sequences that share a 95% of sequence identity and an alignment that cover at least the 90% of the smallest of the sequences were clustered, taking as the reference the longest one. For this task, we excluded all the plasmid sequences from the database. The taxonomy of the clustered database was obtained using the acc_taxid.zip, taxdump.zip files available on the NCBI website. The final genome database (Genomeclustdb) contains 2463 species, 76,478 sequence entries in 10,545,961,079 bp.

The taxonomic assignation of the metagenomic and the metatranscriptomic sequences was carried out by mapping the unassembled sequences in the Genomeclustdb using

blastn tool (v2.2.31 parameters: -task megablast -perc_identity 75). Then, all reads whose alignment length was below 100 nucleotides, or the percentage of identity was below 75%, or e-value was above 0.00001, were discharged. The assignation criterion was based by the read best-hit, prioritizing first the e-value, then the bit-score, the percentage of identity and finally the alignment's length. In the case of two or more database genomes contain the same best-hit values; the read coverage was divided by the number of hits with the same score. We mapped the RNAseqs and the DNAseqs against Genomeclustdb using the megablast software to calculate their relative abundance.

Ecological, clustering, ordination and correlation analyses

All the analyses were conducted with custom Perl scripts and R statistical software version 3.2.0 (R Development Core Team, 2011). Given the high-dimensional data and to control the false discovery rate, we validated the statistical tests adjusting all p-values using the Benjamini-Hochberg correction (library "stats", function "p.adjust") (BH adjusted P-value).

The abundance matrix retrieved from the taxonomic and functional annotations of the DNAseq and RNAseq data were standardized by the total number of sequences. Then, to avoid biases given species composition we normalize data using the Hellinger transformation (library "vegan" function "decostand"). The Hellinger distance (library "vegan" function "vegdist") and the Bray-Curtis index (library "vegan" function "vegdist") were calculated from the abundance matrix from the functional and taxonomic composition, respectively, to quantify the compositional dissimilarity between two different communities. Similarly, the evenness within groups was measured using the Pielou's index.

The non-metric multidimensional scaling (NMDS) was performed to reduce dimensionality in the taxonomic and functional distance matrix (library "vegan" function "metaMDS"). The clustering analysis for the species and functional abundance profiles was performed using the Partitioning Around Medoids (PAM) algorithm (library "cluster", function "pam"). The optimal cluster configuration was the one that maximizes the silhouette index (library "cluster", function "silhouette"). The hierarchical clustering was done using the ward method (library "stats", function "hclust", method "ward").

To statistically assess the effect of the environmental factors on the bacterial and functional composition, a multivariate analysis of variance based on dissimilarity tests (ADONIS) was applied as implemented in the vegan library in R package (library "Vegan" function "adonis").

All the correlation analyses were performed using the Spearman's rank correlation coefficient (library ''stats'', function ''cor.test'').

To evaluate differences between groups in continuous variables, we used the Kruskal–Wallis test. The pairwise comparisons of continuous variables were analyzed using Wilcoxon rank-sum test or T-student test.

Association between taxonomic and functional biomarkers

The putative associations between the pathways and species biomarkers were estimated by means of Generalized Linear Models (GLM). The GLM was estimated setting as the response variable the pathway abundance and as the predictors the species matrix. In order to avoid over-fitting, each model was cross-validated 54 times and the least absolute shrinkage and selection operator (lasso) was used to perform variable selection and regularization (library "glment" function "cv.glmnet" alpha = 1). Additionally, the Spearman's rank correlation coefficient (library ''stats'', function ''cor.test'') was calculated for each of pairwise taxonomic-pathway variables selecting only those associations whose adjusted p-value were below 0.05.

Once the statistical associations were determined, we look whether the genomes of those biomarker species contained the genes of the biomarker metabolic pathways from which they had been associated. In order to achieve this, we downloaded the ORFs fasta file from the NCBI RefSeq genomes database (ftp://ftp.ncbi.nlm.nih.gov/genomes/all) from all the selected species. Then, the program makeblastdb (v2.2.31) produces BLAST databases for each of the species ORF translated to amino acid sequences fasta files. Finally, the blastp tool (v2.2.31) was used to search all the genes related to those metabolic pathways in each genome ORF translated database. We took as a good hit if the alinement's length were above 60 amino acids, the e-value were below to 0.000001 and if the percentage of identity were above the 60%.

Sensitive metabolic pathway detection

The remote homolog detection was carried out by Hidden Markov Models (HMM) for each of the KEGG maps of the tryptophan metabolism (M00038) and the enzymes: tryptophanase (K01667), catalase-peroxidase (K03781) and the o-aminophenol oxidase (K20219). Briefly, we search a Pfam family (a collection of related protein regions) for each of the enzymes mentioned above in the Pfam database (http://pfam.xfam.org/). The Pfam database is a large collection of protein families. Each family consists of a curated "seed" alignment, an alignment small set of representative sequences of the family, and an HMM profile based on the seed alignment. We download the corresponding seed alignment for each of the enzymes, mention before, to construct its respective HMM. Those seed alignments are: PF06052 (KO: K00452), PF03301 (KO: K00453), PF01231 (KO: K00463), PF01494 (KO: K00486), K01432 (KO: PF07859), PF00266 (KO: K01556), PF04909 (KO: K03392), PF04199 (KO: K07130), PF00171 (KO: K10217) and PF07859 (KO: K14263).

Each of the seed alignments was used to create an HMM using the hmmbuild (v3.1b2) software. Finally, all the HMM from the tryptophan metabolism were used to detect remote homologs in the ORFaanr database using the hmmsearch tool (v3.1b2). The homology detection was given if the full sequence e-value were below 0.001.

In a similar way, the enzymes related to the formation of trimethylamine (TMA) from choline and L-carnitine; the oxygenase component YeaW/cntA (ACCESSION APQ93983 and APP31394 respectively) and the reductase component YeaX/cntB (ACCESSION WP_001287026 and APQ93983 respectively) were searched into the ORFaanr database using Blastp (v2.2.31+) (alignment coverage >= 0.7 and e-value < 230.00001).

Ecological and metabolic networks

To determine if both co-occurrence and metabolic networks met the properties of biological networks, we assessed the topological properties using the package igraph in the R environment. The connectivity distribution that was calculated applying Kolmogorov-Smirnov test (function "power.law.fit" and "degree.distribution") (Kolmogorov Smirnov p-value > 0. 1, represents the distribution of edge per node and in a biological network fits a power law distribution. This feature means that few nodes presented a high number of interactions while most nodes possessed sparse connections. The small world effect was calculated using function "shortest.paths" and it indicated

that the network presented a high number of clusters with shortcuts between them. In order to test the consistency of the results of the ecological and the functional networks, we simulated 10000 random networks using the Erdős–Rényi model (library igraph function “erdos.renyi.game”). The probability of drawing an edge between two arbitrary nodes was sampled from a random distribution (mean = 0.5, sd = 0.3) for each of 10000 networks (Supplementary Table 1).

Modularity is another important characteristic of the biological networks. The modules or clusters represent link-dense areas separated by regions of low connectivity. We used the “walktrap community” algorithm from the igraph R package to define modules in the co-occurrence network and to obtain the modularity coefficient. The taxonomic composition of such modules was plotted as bar plots (R function “barplot”).

The network statistics, betweenness, degree and eigenvector centrality, were calculated for both ecological and metabolic networks. The degree centrality is the average of the numbers of edges per species and it was calculated using degree algorithm from igraph library. While betweenness centrality is a measure of centrality based on shortest paths, eigenvector centrality measures the relevance of a node based on the concept that connections to highly connected nodes contribute more to the score of the node in question than equal connections to low connecting nodes. We used the “betweenness" and “evcent” algorithms from the igraph R package to determine betweenness and eigenvector centrality, respectively.

The network fragmentation is defined as the relative fraction of disconnected compartments in a co-occurrence network. The fragmentation was calculated using the number of clusters derived from the function “clusters” and the walktrap.community”.

Multiomic Bayesian network

The multiomic BN was estimated using the log normalized relative abundance of the metagenome, metatranscriptome and metabolome data from the HIV-associated microbiota. We reanalyzed the metabolic data described in Serrano-Villar et al. (2016a) and Serrano-Villar et al. (2016b). The metabolite identity was confirmed comparing the fragments that were obtained with the structure of the proposed compound in the MS/MS spectra against a public database (METLIN:https://metlin.scripps.edu/metabolites_list.php) or commercially available standards. If the metabolite contains more than one annotation, we prioritized the

KEGG metabolite annotation, otherwise, we select one from the most abundant annotation of the METLIN database or we took one from the most abundant subclass annotation of the Lipidomics Gateway database (http://www.lipidmaps.org/data/structure/). The abundance matrix that we retrieved from the reanalysis of the metabolomic data was standardized by the total number of metabolites.

The network was estimated using the hill-climbing score-based learning algorithm (R library "bnlearn" function "hc"). The algorithm states the optimal network and in consequence the ''father'' to ''child'' node relationships which maximizes the Bayes information criterion (BIC). The whitelist option was used to define the set of arcs that will be included in the model. For this task we performed a GLM using the Lasso variable selection and regularization (library "glment" function "cv.glmnet" alpha = 1) As response variables, we selected those RNAseq and DNAseq pathways, species and metabolites that were overrepresented or underrepresented in HIV condition and as predictor the clinical variables related to the markers of innate and T cell activation variables, thymic function, and bacterial translocation as done in the multivariate statistical framework MaAsLin pipeline. To avoid over-fitting problems, all the models were validated by means of n-fold cross validation (library "glment" function "cv.glmnet"). Additionally, we included all the Spearman correlation adjusted p-value was above 0.01. We used the blacklist option (R library ''bnlearn'' function ''hc'') to exclude all the edges whose Spearman correlation adjusted p-value were above 0.1. The quantile 95 of the BN node statistics was estimated using the R function quantile (R library ''stats function ''quantile'').

The function mb (R Package ''bnlearn'' function ''mb'') was used to dissect all the Markov Blankets to perform the statistical analysis and to plot the butanoate metabolism Markov blanket. The size of each Markov Blanket was estimated counting the number of nodes per Markov Blanket.

References.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing. *http://www.R-project.org/.*

Pérez-Cobas AE, Gosalbes MJ, Friedrichs A, Knecht H, Artacho A, Eismann K, et al. (2013). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* **62**: 1591–1601.

Rodriguez-R LM and Konstantinidis KT. (2014). Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**: 629-635.

Serrano-Villar S, Rojo D, Martinez M, Deusch S, Vazquez-Castellanos J, Bargiela R, et al. (2016a). Gut Bacteria Metabolism Impacts Immune Recovery in HIV-infected Individuals. *EBioMedicine* **8**: 203–216.

Serrano-Villar S, Rojo D, Martinez M, Deusch S, Vazquez-Castellanos J, Bargiela R, et al. (2016b). HIV infection results in metabolic alterations in the gut microbiota different from those induced by other diseases. *Scientific Reports **6**: 26192.*

Serrano-Villar S, Vázquez-Castellanos JF, Vallejo A, Latorre A, Sainz T, Ferrando-Martínez S, et al. (2017). The Effects of Prebiotics on Microbial Dysbiosis, Butyrate Production and Immunity in HIV-Infected Subjects. *Mucosal Immunol* **10**:1279-1293.

Supplementary figure legends

Supplementary Figure 1. Comparison of the microbiota gene composition across HIV+ groups and uninfected subjects from metagenomics. (**a**) NMDS analysis of the gene (KO) composition for VU (red), IR (green), INR (orange) and HIV- (blue) individuals. The ellipses represent the optimal clustering configuration retrieved from the PAM algorithm and their size encompasses the 70% of variance of the cluster. The cluster configuration was validated using the ADONIS test with the four groups of the cohort (P-value=0.001). (**b**) Linear discriminative (LDA) effect size LEfSe analysis of genes (KO) between HIV+ (red) and HIV- (blue) subjects, (**c**) VU (red) versus HIV- (blue), (**d**) IR (green) versus HIV- (blue) and (**e**) INR (orange) versus HIV- (blue). The cladogram represented the biomarkers of the upper hierarchical classes within the KEGG database. VU, viremic untreated; IR, immunological responder; INR, immunological non-responder.

Supplementary Figure 2. Species composition from metagenomic dataset. (**a**) NMDS analysis of the taxonomic species composition. VU (red), IR (green), INR (orange) and HIV- (blue) subjects are represented by ellipses corresponding to 70% of samples; cluster configuration was validated using the ADONIS test for the four groups of the cohort (P-value=0.002). (**b**) Linear discriminative analysis (LDA) effect size LEfSe analysis between HIV+ (red) and HIV- (blue) subjects. LDA scores (log 10) for the most discriminative species in controls are represented on the positive scale (blue bars); LDA-negative scores indicate enriched species in HIV+ subjects (red bars). VU, viremic untreated; IR, immunological responder; INR, immunological non-responder.

Supplementary Figure 3. Comparison of microbiota gene composition across HIV+ groups and uninfected subjects from metatranscriptomic dataset. (**a**) NMDS analysis of the KO composition for VU (red), IR (green), INR (orange) and HIV- (blue) individuals. The ellipses represent the optimal clustering configuration retrieved from the PAM algorithm and their size encompasses the 70% of variance of the cluster. Cluster configuration was validated using the ADONIS test for the four groups of the cohort (P-value=0.001). (**b**) Linear discriminative analysis effect size LEfSe analysis between the HIV+ (in red) and HIV- (in blue) subjects. LDA scores (log 10) for the most discriminative KO in HIV- are represented on the negative scale, whereas LDA-

positive scores indicate enriched pathways in HIV+ subjects. VU, viremic untreated; IR, immunological responder; INR, immunological non-responder.

Supplementary Figure 4. Taxonomic abundance. (**a**) Relative abundance of the most abundant genera from metagenomics (mean relative abundance > 0.05 and present in at least 70% of samples) and (b) from the metatranscriptomics. VU (red), IR (green), INR (orange) and HIV- (blue) subjects. Grey represents Euryarchaeota; pink, Actinobacteria; brown gradient, Bacteroidetes; blue gradient, Firmicutes; and green gradient, Proteobacteria. VU, viremic untreated; IR, immunological responder; INR, immunological non-responder.

Supplementary Figure 5. Species composition from metatranscriptomic dataset. Linear discriminative analysis (LDA) effect size LEfSe analysis between HIV+ (red) and HIV- (blue) subjects. LDA scores (log 10) for most discriminative species in controls are represented on positive scale (blue bars); LDA-negative scores indicated enriched species in HIV+ subjects (red bars).

Supplementary Figure 6. Comparisons of gene expression level in HIV-infected and uninfected individuals. (**a**) Tryptophanase. (**b**) Choline TMA-lyase activating protein. (**c**) Choline TMA-lyase. Differences between groups were assessed by Kruskal-Wallis test, while group pairwise comparison was set by the Wilcoxon signed-rank test.

Supplementary Figure 7. Metabolic network. Metabolic network retrieved from pathways inferred by ORF KEGG-annotation enzymes. Each enzyme of the network can be associated with several reactions, and each reaction can be involved with different enzymes. Nodes represent KEGG orthologous groups (KO); direct edges indicate that product metabolite of one enzyme is substrate for enzyme to which the edge is pointing. Blue: all KO biomarkers related to HIV- subjects; red, KO biomarkers related to the HIV condition.

Supplementary Figure 8. HIV+ species centrality. (**a**) Betweenness centrality, which quantifies the number of times a node acts as a bridge along shortest path between two other nodes. (**b**) Eigenvector centrality measures the influence of a node in a network by

assigning a weighted score that considers its own degree of centrality and the centrality of nodes from which it is connected. Nodes represent species belonging to Firmicutes (cyan), Bacteroidetes (orange), Actinobacteria (pink) and Proteobacteria (green) phyla. Only outlier species with quantile 95 are labeled in the boxplot.

Supplementary Figure 9. Spearman correlation between LDA scores (log 10) of species biomarkers and corresponding centrality score. (**a**) Betweenness centrality. (**b**) Eigenvector centrality. The regression line is represented in red. Correlations were set for HIV+ and HIV- species biomarkers independently. Species belong to Firmicutes (cyan), Bacteroidetes (orange), Actinobacteria (pink) and Proteobacteria (green) phyla.

Supplementary Figure 10. Centrality of the KO biomarkers. Betweenness, degree and eigenvector centrality for the HIV- (blue), HIV+ (red) KO biomarkers and the rest of the enzymes in the network (grey). The centrality values are expressed on a log scale.

Supplementary Figure 11. HIV+ Bayesian network. "Multiomic" Bayesian networks composed of the metagenomic data (blue nodes), metatranscriptomic data (green nodes), metabolomic data (pink circle nodes) and the clinical variables (gold square nodes) from the HIV+ subjects. The data from the metagenomic and metatranscriptomic included the information of the species relative abundance (circles) and pathway relative abundance (squares). The labels of the nodes represent over representation in HIV- subjects (blue label) or in HIV+ subjects (red label). Arrows indicate the conditional dependencies between variables. The Spearman correlation coefficient is represented by the arrow's color, blue if it is significantly positive (BH adjusted P-value < 0.1), red if it has a significantly negative correlation (BH adjusted P-value < 0.1) or grey if it is non-significant.

Supplementary Figure 12. Bayesian network statistics. (**a**) Degree centrality, (**b**) number of nodes per Markov Blanket (MB) and (**c**) number of clinical variables within each MB for each component of the BN. The metagenomic and metatranscriptomic species (Sp_DNA and Sp_RNA respectively), the metagenomic and metatranscriptomic pathways (ko_DNA and ko_RNA respectively), the metabolites (Metabol) and the clinical variables (Clinic). Differences between groups were assessed by the Kruskal-

1Wallis test, while the pairwise comparison was set by the Wilcoxon signed-rank test.

2The centralities values are expressed on a log scale.

3

Supplementary tables

Supplementary Table 1. Network bootstrap P-value.

	Mean Degree	Mean Betweenness	Mean Closeness	Mean Eigenvector centrality	[a]%Pos	[b]%Neg	[c]Ratio	[d]Modularity	[e]Number clusters>1	[f] Transitivity	[g]Diameter	[h]Mean-short pathways	K-Smirnov P-value
Co-ocuurence network	*	*	*	*	0	0	0	0	0	0	0	0	*
Enzimatic network	*	*	*	*	*	*	*	*	*	*	*	*	*

* P-value <0.001

(a) Percentage of edges with a statistically significant positive correlation

(b) Percentage of edges with a statistically significant negative correlation

(c) Ratio between the positive and negative correlations.

(d) Modularity index retrieved from the modules estimate by the Walktrap community finding algorithm

(e) Number of modules, estimate by the Walktrap community finding algorithm, that posses at least two nodes

(f) Clustering coefficient of the network. measure of the degree to which nodes in a graph tend to cluster together.

1(g) Longest path between two vertices within the network

2(h) Average path length between any two nodes

3

4

5

6

2

1**Supplementary Table 2**. Relative abundance of the species with genes related to the tryptophan catabolism via the kynurenine pathway.

2

DNA mapping species	VU (mean /sd)	IR (mean /sd)	INR (mean /sd)	HIV- (mean /sd)	K/T ratio Spearman correlation index	P-value	BH adjusted P-value *
Xanthomonas arboricola pruni	9.41e-05/6.5e-05	7.66e-05/ 3.57e-05	0.0001108/8.45e-05	7.61e-05 / 6.1e-05	0.280	0.041	0.222
Trichodesmium erythraeum IMS101	8.97e-05/6.94e-05	6.91e-05 /4.56e-05	9.77e-05 /6.18e-05	9.11e-05/6.69e-05	0.167	0.227	0.265
Streptomyces scabiei	0.0001091/9.94e-05	8.52e-05 / 4.64e-05	0.000125 / 8.8e-05	0.0001094/8.81e-05	0.213	0.122	0.250
Pseudomonas putida	0.0001933/0.0001906	0.0001985/0.0001281	0.0002391/0.000237	0.0001844/0.00015	0.173	0.211	0.265
Pseudomonas fluorescens	0.0001734 / 0.0001575	0.0001706/0.0001096	0.0001964/0.0001917	0.0001549/0.00012	0.204	0.140	0.250
Pseudomonas aeruginosa	0.0001707 / 0.0001523	0.0001658/0.0001028	0.0002052/0.0001739	0.0001493/0.000117	0.199	0.149	0.250
Nitrosococcus oceani ATCC19707	0.0002297 / 0.000164	0.0002088/0.0001064	0.0002464/0.0001768	0.0002198/0.00014	0.194	0.161	0.250
Nitrosococcus halophilus Nc4	0.0001412 / 0.0001042	0.0001196/6.85e-05	0.0001532/0.0001356	0.0001245/0.00010	0.271	0.048	0.222
Neptuniibacter caesariensis MED92	5.38e-05 / 5.39e-05	5.6e-05 / 4.18e-05	5.4e-05 / 5.33e-05	3.99e-05 / 3.99e-05	0.211	0.126	0.250
Methylocystis sp ATCC 49242	0.0001391 / 0.0001011	8.43e-05 / 4.64e-05	0.0001506 /7.82e-05	0.0001229/8.09e-05	0.171	0.216	0.265
Gemmatimonas aurantiaca	8.17e-05 / 6.32e-05	5.42e-05 / 3.16e-05	8.89e-05 / 4.56e-05	7.15e-05 / 6e-05	0.213	0.121	0.250
Erythrobacter sp SD 2	2.47e-05 / 2.95e-05	6e-06 / 1.19e-05	2.23e-05 / 2.76e-05	2.12e-05 / 2.01e-05	0.287	0.035	0.222
Erythrobacter litoralis	0.0001235/9.76e-05	7.31e-05 / 4.4e-05	0.0001265/6.67e-05	0.0001322/8.27e-05	0.134	0.334	0.338
Citromicrobium bathyomarinum	0.0001136/9.57e-05	6.87e-05 /4.21e-05	0.0001276/6.37e-05	0.0001259/8.17e-05	0.133	0.338	0.338

1

RNA mapping species	VU (mean/sd)	IR (mean/sd)	INR (mean/sd)	HIV- (mean/sd)	K/T ratio Spearman correlation index	P-value	BH adjusted P-value *
Xanthomonas arboricola pruni	9.5e-06 / 1.75e-05	7.4e-06 / 1.36e-05	1.7e-06/4.1e-06	1.32e-05/1.69e-05	-0.030	0.868	0.935
Trichodesmium erythraeum IMS101	0 / 0	0 / 0	0 / 0	2.1e-06 / 7.7e-06	0.000	1.000	1.000
Streptomyces scabiei	1.2e-06 / 3.6e-06	2.1e-06 / 3e-06	0 / 0	6.1e-06/1.53e-05	-0.162	0.367	0.728
Pseudomonas putida	5e-06 / 6e-06	1.08e-05/1.55e-05	1.4e-06/3.4e-06	9.6e-06/1.93e-05	-0.075	0.679	0.792
Pseudomonas fluorescens	4.6e-06 / 7.4e-06	7.4e-06 / 1.24e-05	1.2e-06 / 3e-06	6.6e-06/1.27e-05	-0.155	0.389	0.728
Pseudomonas aeruginosa	3.8e-06 / 5.7e-06	9.3e-06 / 1.38e-05	3e-06 / 4.7e-06	6.5e-06/1.06e-05	-0.147	0.415	0.728
Nitrosococcus oceani ATCC19707	4.4e-06 / 8.7e-06	1.22e-05 / 2.21e-05	2.4e-06 / 6e-06	1e-05 / 2e-05	-0.119	0.509	0.728
Nitrosococcus halophilus Nc4	3.8e-06 / 7.5e-06	1.04e-05 / 1.88e-05	2.1e-06 / 5.1e-06	7.6e-06 1.43e-05	-0.118	0.512	0.728
Neptuniibacter caesariensis MED92	0 / 0	1.04e-05 / 1.66e-05	0 / 0	3e-06 / 8.2e-06	-0.219	0.220	0.728
Methylocystis sp ATCC 49242	9.7e-06 / 1.01e-05	5.9e-06 / 7.1e-06	0 / 0	2.7e-06 / 5.4e-06	0.110	0.543	0.728
Gemmatimonas aurantiaca	3.3e-06 / 6.6e-06	4.2e-06 / 7e-06	1.8e-06 / 4.5e-06	6.5e-06 / 1.1e-05	-0.172	0.339	0.728
Erythrobacter sp SD 2	7.7e-06 / 1.63e-05	1.02e-05 / 2.04e-05	0 / 0	7.8e-06/1.68e-05	-0.161	0.372	0.728
Erythrobacter litoralis	6.1e-06 / 9.3e-06	8.6e-06 / 1.57e-05	0 / 0	8.6e-06/1.05e-05	-0.151	0.400	0.728
Citromicrobium bathyomarinum	6e-06 / 9.1e-06	8.3e-06 / 1.84e-05	0 / 0	9.6e-06/1.24e-05	-0.102	0.572	0.728

2

3*P-value adjusted using the Benjamini and Hochberg method

4

1**Supplementary Table 3.** GLM between the taxonomic and functional biomarkers.

2

ko_pathway	Species	GLM Coefficient	Spearman Correlation index	Spearman P-value	BH adjusted P-value
Protein_export_ko03060	Prevotella_salivae	0.630327914	0.708557271	5.93E-009	3.17E-008
Protein_export_ko03060	Bacteroides_sp D20	-0.000344	-0.54473032	2.80E-005	0.0000749
Protein_export_ko03060	Alistipes_finegoldii	-0.00292209	-0.364665523	0.00700326	0.0095950
RNA_degradation_ko03018	Prevotella_salivae	0.489633859	0.748580141	0	0
RNA_degradation_ko03018	Prevotella_copri	0.000982003	0.737145035	0	0
RNA_degradation_ko03018	Bacteroides_pectinophilus	0.00491378	0.312826377	0.02166191	0.0276129
RNA_degradation_ko03018	Bibersteinia_trehalosi	-0.33164526	0.293629397	0.03116270	0.0383202
Epithelial_cell_signaling_in_Helicobacter_pylori_infection_ ko05120	Fermentimonas_caenicola	1.023448054	0.448348472	0.00067385	0.0011784
Epithelial_cell_signaling_in_Helicobacter_pylori_infection_ko05120	Streptococcus_pseudopneumoniae	2.067914861	0.396214039	0.00301834	0.0045471
Epithelial_cell_signaling_in_Helicobacter_pylori_infection_ko05120	Odoribacter_splanchnicus	-0.00151189	-0.254507338	0.06350798	0.0734245
Vibrio_cholerae_pathogenic_cycle_ko05111	Prevotella_copri	0.000224926	0.686601868	3.56E-008	1.72E-007
Vibrio_cholerae_pathogenic_cycle_ko05111	Prevotella_salivae	0.057339171	0.662206975	1.36E-007	6.13E-007
Vibrio_cholerae_pathogenic_cycle_ko05111	Prevotella_ruminicola	2.115990035	0.640627435	1.81E-007	7.89E-007
Vibrio_cholerae_pathogenic_cycle_ko05111	Prevotella_melaninogenica	0.276605282	0.6355523	2.43E-007	1.00E-006
Vibrio_cholerae_pathogenic_cycle_ko05111	Acidaminococcus_intestini	0.012038269	0.560434534	1.49E-005	4.19E-005
Vibrio_cholerae_pathogenic_cycle_ko05111	Dialister_succinatiphilus	0.098358216	0.554640747	1.89E-005	5.13E-005
Vibrio_cholerae_pathogenic_cycle_ko05111	Bilophila_wadsworthia	-0.019762194	-0.539944814	0.00002511	6.77E-005
Vibrio_cholerae_pathogenic_cycle_ko05111	Bacteroides_eggerthii	-0.004636945	-0.50409758	0.00012585	0.0002772
Vibrio_cholerae_pathogenic_cycle_ko05111	Bifidobacterium_longum	-0.002182973	-0.490909091	0.00019772	0.0004000
Vibrio_cholerae_pathogenic_cycle_ko05111	Bacteroides_finegoldii	-0.000509064	-0.413607776	0.00203953	0.0031685
Vibrio_cholerae_pathogenic_cycle_ko05111	Ruminococcus_bicirculans	-0.022716714	-0.375314332	0.00516630	0.0073382
Vibrio_cholerae_pathogenic_cycle_ko05111	Coprobacillus_sp 8138FAA	-0.063941187	-0.304921283	0.02496391	0.0312497

Vibrio_cholerae_pathogenic_cycle_ko05111	Blautia_obeum	-0.005948943	-0.262511912	0.05541975	0.0647183
Vibrio_cholerae_pathogenic_cycle_ko05111	Bacteroides_sp 325	0.008743212	0.252144082	0.06607022	0.0761339
Alanine_aspartate_and_glutamate_metabolism_ko00250	Prevotella_salivae	0.597740174	0.806899181	0	0
Alanine_aspartate_and_glutamate_metabolism_ko00250	Prevotella_copri	0.006633203	0.890909091	0	0
Alanine_aspartate_and_glutamate_metabolism_ko00250	Prevotella_fusca	0.827176859	0.870195445	1.30E-017	1.41E-016
Alanine_aspartate_and_glutamate_metabolism_ko00250	Prevotella_sp oral taxon 299	0.787709191	0.845978791	8.13E-016	7.45E-015
Alanine_aspartate_and_glutamate_metabolism_ko00250	Prevotella_intermedia	0.882819982	0.779715516	3.73E-012	2.60E-011
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Prevotella_salivae	0.122106449	0.752391843	0	0
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Prevotella_copri	0.002417975	0.828092243	0	0
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Prevotella_fusca	0.358790813	0.841638161	1.58E-015	1.41E-014
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Prevotella_dentalis	0.155941444	0.836258167	3.52E-015	3.06E-014
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Prevotella_ruminicola	2.445823921	0.827009201	1.30E-014	1.03E-013
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Prevotella_intermedia	2.023986526	0.77558943	5.72E-012	3.90E-011
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Dialister_succinatiphilus	0.029707247	0.61982085	1.08E-006	3.97E-006
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Bacteroides_eggerthii	-0.001618533	-0.566152087	1.18E-005	3.45E-005
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Bacteroides_sp HPS0048	-0.271486108	-0.565770917	1.20E-005	3.47E-005
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Bacteroides_salyersiae	-0.002680836	-0.538860301	3.52E-005	8.93E-005
Phenylalanine_tyrosine_and_tryptophan_biosynthes	Odoribacter_splanchnicus	-0.021115026	-0.502496665	0.00013307	0.0002894

is_ko00400					
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Bifidobacterium_sp 12147BFAA	-0.006566333	-0.485248511	0.00020002	0.0004000
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Bilophila_wadsworthia	0.005086793	-0.370744814	0.00578439	0.0081496
Phenylalanine_tyrosine_and_tryptophan_biosynthesis_ko00400	Ruminococcus_bicirculans	-0.029883963	-0.229935053	0.09440045	0.1073573
Citrate_cycle_TCA_cycle__ko00020	Prevotella_salivae	0.642743553	0.756279779	0	0
Citrate_cycle_TCA_cycle__ko00020	Acidaminococcus_intestini	4.92E-005	0.686983038	3.48E-008	1.71E-007
Fructose_and_mannose_metabolism_ko00051	Prevotella_denticola	0.536892355	0.482560962	0.00021955	0.0004316
Fructose_and_mannose_metabolism_ko00051	Prevotella_dentalis	0.085864011	0.47820565	0.00025492	0.0004821
Fructose_and_mannose_metabolism_ko00051	Prevotella_copri	0.000296255	0.437926434	0.00103337	0.0017206
Carbon_fixation_in_photosynthetic_organisms_ko00710	Prevotella_salivae	0.836564665	0.770688012	0	0
Carbon_fixation_in_photosynthetic_organisms_ko00710	Prevotella_copri	0.003241563	0.765885268	0	0
Carbon_fixation_in_photosynthetic_organisms_ko00710	Prevotella_melaninogenica	0.066786524	0.793648178	8.21E-013	5.96E-012
Carbon_fixation_in_photosynthetic_organisms_ko00710	Bacteroides_stercoris	-0.009513908	-0.506155899	0.00011710	0.0002612
Carbon_fixation_in_photosynthetic_organisms_ko00710	Streptococcus_pneumoniae	2.91801112	0.495672105	0.00013834	0.0002955
Carbon_fixation_in_photosynthetic_organisms_ko00710	Clostridiales_bacterium_VE20214	1.224250705	0.493119878	0.00018352	0.0003791
Carbon_fixation_in_photosynthetic_organisms_ko00710	Alistipes_finegoldii	-0.006455168	-0.437392796	0.00104944	0.0017308
Carbon_fixation_in_photosynthetic_organisms_ko00710	Bibersteinia_trehalosi	-1.310449328	0.270662686	0.04775561	0.0565880
Oxidative_phosphorylation_ko00190	Prevotella_copri	0.000435808	0.542443301	3.06E-005	8.07E-005
Oxidative_phosphorylation_ko00190	Prevotella_dentalis	0.01313649	0.531386305	3.56E-005	8.97E-005
Oxidative_phosphorylation_ko00190	Bacteroides_fluxus	0.00358923	0.483666857	0.00025160	0.0004810
Lipopolysaccharide_biosynthesis_ko00540	Prevotella_fusca	0.150236935	0.673539598	2.39E-008	1.22E-007
Lipopolysaccharide_biosynthesis_ko00540	Prevotella_ruminicola	9.183981911	0.66860578	3.29E-008	0.0000001

Lipopolysaccharide_biosynthesis_ko00540	Prevotella_copri	0.004909974	0.687059272	3.46E-008	1.71E-007
Lipopolysaccharide_biosynthesis_ko00540	Prevotella_salivae	0.364890392	0.671964932	8.15E-008	3.73E-007
Lipopolysaccharide_biosynthesis_ko00540	Prevotella_melaninogenica	0.851269653	0.641698989	1.71E-007	7.51E-007
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_fluxus	0.005905104	0.624623594	0.00000086	3.26E-006
Lipopolysaccharide_biosynthesis_ko00540	Prevotella_oralis	-1.385958208	0.606900116	1.14E-006	4.15E-006
Lipopolysaccharide_biosynthesis_ko00540	Acidaminococcus_intestini	0.040889744	0.61661902	1.26E-006	4.52E-006
Lipopolysaccharide_biosynthesis_ko00540	Bifidobacterium_longum	-0.043668026	-0.581551363	6.14E-006	1.89E-005
Lipopolysaccharide_biosynthesis_ko00540	Riemerella_anatipestifer	0.158894854	0.533289653	3.30E-005	8.50E-005
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_stercoris	0.022152402	-0.521250238	6.81E-005	0.0001613
Lipopolysaccharide_biosynthesis_ko00540	Bibersteinia_trehalosi	6.667541297	0.50484658	9.90E-005	0.0002266
Lipopolysaccharide_biosynthesis_ko00540	Fermentimonas_caenicola	2.671034976	0.49719101	0.00013097	0.0002866
Lipopolysaccharide_biosynthesis_ko00540	Parabacteroides_sp__HGS0025	-1.983556429	-0.493045463	0.00015198	0.0003225
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_plebeius	-0.006023894	-0.494034687	0.00017792	0.0003707
Lipopolysaccharide_biosynthesis_ko00540	Odoribacter_splanchnicus	0.006868554	-0.490680389	0.00019925	0.0004000
Lipopolysaccharide_biosynthesis_ko00540	Dialister_succinatiphilus	0.34678993	0.483666857	0.00025160	0.0004810
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_eggerthii	0.012927647	-0.475967219	0.00032332	0.0006049
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_sp__14_A_	-0.008273623	-0.465878773	0.00038492	0.0007050
Lipopolysaccharide_biosynthesis_ko00540	Coprobacillus_sp__8_1_38FAA	-0.044760255	-0.45458011	0.00055411	0.0009788
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_caccae	0.010611613	-0.450962455	0.00070374	0.0012123
Lipopolysaccharide_biosynthesis_ko00540	Parabacteroides_sp__ASF519	-0.114727764	-0.437712139	0.00093308	0.0015762
Lipopolysaccharide_biosynthesis_ko00540	Alistipes_finegoldii	0.025365913	-0.420163903	0.00170576	0.0027229
Lipopolysaccharide_biosynthesis_ko00540	Blautia_obeum	-0.032163756	-0.414751286	0.00197741	0.0030858
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_sp__HPS0048	-0.552418218	-0.36611397	0.00676863	0.0093102
Lipopolysaccharide_biosynthesis_ko00540	Coprococcus_sp__HPP0048	-2.033517547	-0.331400197	0.01436865	0.0187980
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_sp__3_2_5	0.156426259	0.303602058	0.02601956	0.0324545
Lipopolysaccharide_biosynthesis_ko00540	Bacteroides_pectinophilus	-0.015602444	-0.260377359	0.05748977	0.0669111
N_Glycan_biosynthesis_ko00510	Prevotella_copri	0.000175869	0.516142558	0.00008206	0.0001909
N_Glycan_biosynthesis_ko00510	Fermentimonas_caenicola	0.781727328	0.491985222	0.00015782	0.0003328
N_Glycan_biosynthesis_ko00510	Prevotella_dentalis	0.008432226	0.41325801	0.00189718	0.0030010
N_Glycan_biosynthesis_ko00510	Bacteroides_eggerthii	-0.000986601	-0.399504479	0.00296259	0.0045107
N_Glycan_biosynthesis_ko00510	Parabacteroides_sp__HGS0025	-0.314962578	-0.373900806	0.00535100	0.0075697
N_Glycan_biosynthesis_ko00510	Bacteroides_stercoris	-0.001063969	-0.314503526	0.02094032	0.0267913
N_Glycan_biosynthesis_ko00510	Coprococcus_sp__HPP0048	-0.072767395	-0.265921471	0.05195059	0.0608713

N_Glycan_biosynthesis_ko00510	Faecalibacterium_prausnitzii	-0.000823985	-0.227901658	0.09744031	0.1100949
Peptidoglycan_biosynthesis_ko00550	Bacteroides_sp__D20	-0.002507641	-0.728606823	0	0
Peptidoglycan_biosynthesis_ko00550	Prevotella_salivae	0.771067843	0.812388031	0	0
Peptidoglycan_biosynthesis_ko00550	Prevotella_copri	0.00519608	0.837164094	0	0
Peptidoglycan_biosynthesis_ko00550	Prevotella_dentalis	1.382983757	0.872323209	8.70E-018	1.01E-016
Peptidoglycan_biosynthesis_ko00550	Prevotella_sp__oral_taxon_299	3.20351697	0.846514681	7.48E-016	7.04E-015
Peptidoglycan_biosynthesis_ko00550	Faecalibacterium_prausnitzii	0.016487105	0.371755289	0.00591925	0.0082726
Nicotinate_and_nicotinamide_metabolism_ko00760	Prevotella_copri	0.002565147	0.765199161	0	0
Nicotinate_and_nicotinamide_metabolism_ko00760	Prevotella_sp__oral_taxon_299	1.188732661	0.727317452	4.70E-010	2.82E-009
Nicotinate_and_nicotinamide_metabolism_ko00760	Prevotella_denticola	0.81085705	0.725235907	5.56E-010	3.28E-009
Nicotinate_and_nicotinamide_metabolism_ko00760	Prevotella_fusca	0.120275338	0.716642772	1.10E-009	6.38E-009
Nicotinate_and_nicotinamide_metabolism_ko00760	Bacteroides_sp__D20	-0.000525527	-0.598170383	2.94E-006	9.66E-006
Nicotinate_and_nicotinamide_metabolism_ko00760	Bacteroides_sp__HPS0048	-0.072662376	-0.462321327	0.00049765	0.0008973
One_carbon_pool_by_folate_ko00670	Prevotella_salivae	0.81149853	0.591614256	3.95E-006	0.0000126
One_carbon_pool_by_folate_ko00670	Bacteroides_sp__3_1_23	-0.033884814	-0.487173623	0.00022403	0.0004379
One_carbon_pool_by_folate_ko00670	Prevotella_oralis	0.785144755	0.468924719	0.00034816	0.0006444
One_carbon_pool_by_folate_ko00670	Bacteroides_caccae	-0.001178167	-0.39310082	0.00349279	0.0051437
|Metabolism_of_Other_Amino_Acids| D_Alanine_metabolism_ko00473	Prevotella_salivae	0.361960852	0.755441205	0	0
Metabolism|Metabolism_of_Other_Amino_Acids| D_Alanine_metabolism_ko00473	Prevotella_copri	0.012571451	0.839146179	0	0
Metabolism|Metabolism_of_Other_Amino_Acids| D_Alanine_metabolism_ko00473	Prevotella_dentalis	0.493603635	0.827394724	1.23E-014	9.99E-014
Metabolism|Metabolism_of_Other_Amino_Acids| D_Alanine_metabolism_ko00473	Bacteroides_fluxus	0.047915676	0.673413379	7.54E-008	3.50E-007
Metabolism|Metabolism_of_Other_Amino_Acids| D_Alanine_metabolism_ko00473	Parabacteroides_sp__ASF519	-0.166124553	-0.433898983	0.00104591	0.0017308
Terpenoid_backbone_biosynthesis_ko00900	Bacteroides_sp__D20	-0.000990142	-0.760853821	0	0
Terpenoid_backbone_biosynthesis_ko00900	Prevotella_copri	0.004773795	0.897846388	0	0
Terpenoid_backbone_biosynthesis_ko00900	Prevotella_dentalis	1.547405825	0.90693648	3.63E-021	4.35E-020
Terpenoid_backbone_biosynthesis_ko00900	Prevotella_sp__oral_taxon_299	0.895389824	0.863663158	4.29E-017	4.52E-016
Terpenoid_backbone_biosynthesis_ko00900	Bacteroides_sp__4_1_36	-0.000347414	-0.708557271	5.93E-009	3.17E-008
Terpenoid_backbone_biosynthesis_ko00900	Fermentimonas_caenicola	6.137082373	0.631086934	3.12E-007	1.25E-006

Terpenoid_backbone_biosynthesis_ko00900	Bacteroides_clarus	-0.04616053	-0.61623785	1.28E-006	4.55E-006
Terpenoid_backbone_biosynthesis_ko00900	Bacteroides_sp__14_A_	-0.007914154	-0.560541375	0.00001042	3.10E-005
Terpenoid_backbone_biosynthesis_ko00900	Bacteroides_salyersiae	-0.011945584	-0.509738898	0.00010319	0.0002347
Zeatin_biosynthesis_ko00908	Prevotella_copri	0.001080912	0.77442348	0	0
Zeatin_biosynthesis_ko00908	Prevotella_dentalis	2.314246958	0.739447979	1.70E-010	1.10E-009
Zeatin_biosynthesis_ko00908	Fermentimonas_caenicola	0.744041474	0.559966686	1.07E-005	3.15E-005
Zeatin_biosynthesis_ko00908	Parabacteroides_sp__HGS0025	-3.340567	-0.519453814	5.70E-005	0.0001367
Zeatin_biosynthesis_ko00908	Bacteroides_sp__HPS0048	-0.033429809	-0.452029731	0.00068152	0.0011853
Zeatin_biosynthesis_ko00908	Megasphaera_elsdenii	-0.036514968	0.347131694	0.01046170	0.0139489
Zeatin_biosynthesis_ko00908	Acidaminococcus_sp__D21	0.013281067	0.325404993	0.01673093	0.0215643
Zeatin_biosynthesis_ko00908	Coprobacillus_sp__8_1_38FAA	-0.174983397	-0.32215149	0.01751841	0.0224959
Toluene_degradation_ko00623	Prevotella_copri	0.001917409	0.745683248	0	0
Toluene_degradation_ko00623	Prevotella_dentalis	0.820839323	0.752437507	5.37E-011	3.53E-010
Toluene_degradation_ko00623	Prevotella_salivae	0.195595019	0.695826186	1.94E-008	1.01E-007
Toluene_degradation_ko00623	Bacteroides_fluxus	0.000418738	0.644215742	3.34E-007	1.31E-006
Toluene_degradation_ko00623	Bacteroides_sp__4_1_36	-0.001583292	-0.537869259	0.00003652	9.14E-005
Toluene_degradation_ko00623	Bacteroides_eggerthii	0.009439637	-0.308862207	0.02345195	0.0296773
Toluene_degradation_ko00623	Ruminococcus_bicirculans	-0.014396793	-0.272576579	0.04614249	0.0551807
Toluene_degradation_ko00623	Blautia_obeum	-0.010927638	-0.227215552	0.09847345	0.1109021

1
2
3

1**Supplementary Table 4.** Biomarker species gene mapping.

2

Species [a]	ko [b]	KO in ko [c]	KO [d]	Percentage %	KO list [e]
Acidaminococcus_intestini	ko00020	56	11	19.64	K00030 K00031 K00174 K00175 K00176 K00177 K01610 K01679 K01681 K01958 K01960
Prevotella_salivae	ko00020	56	8	14.29	K00024 K00174 K00175 K00176 K00177 K01610 K01676 K01960
Prevotella_copri	ko00051	81	19	23.46	K00011 K00100 K00754 K00847 K00848 K00850 K00895 K00966 K00971 K01624 K01629 K01711 K01803 K01805 K01808 K01813 K01840 K02377 K04041
Prevotella_dentalis	ko00051	81	20	24.69	K00100 K00754 K00847 K00848 K00850 K00895 K00966 K00971 K01624 K01629 K01711 K01803 K01805 K01808 K01809 K01813 K01840 K02377 K04041 K05305
Prevotella_denticola	ko00051	81	16	19.75	K00100 K00754 K00847 K00850 K00895 K00966 K00971 K01218 K01624 K01711 K01803 K01808 K01809 K01840 K02377 K04041
Bacteroides_fluxus	ko00190	212	26	12.26	K00330 K00331 K00335

					K00337 K00338 K00339 K00340 K00341 K00342 K00343 K02108 K02109 K02111 K02112 K02113 K02114 K02115 K02117 K02118 K02120 K02121 K02123 K02124 K03885 K13378 K15987
Prevotella_copri	ko00190	212	20	9.43	K00330 K00331 K00337 K00338 K00339 K00340 K00341 K00342 K00343 K02108 K02109 K02110 K02111 K02112 K02113 K02114 K02115 K02120 K03885 K13378
Prevotella_dentalis	ko00190	212	21	9.91	K00330 K00331 K00337 K00338 K00339 K00340 K00341 K00342 K00343 K02108 K02109 K02110 K02111 K02112 K02113 K02114 K02115 K02120 K03885 K13378 K15987
Prevotella_copri	ko00250	60	21	35	K00259 K00262 K00264 K00265 K00266 K00278 K00609 K00610 K00764 K00820 K01424 K01755 K01756 K01914 K01939 K01940 K01953 K01955 K01956 K11540 K11541
Prevotella_fusca	ko00250	60	16	26.67	K00262 K00266 K00278 K00609 K00610 K00764 K00820 K01424 K01744 K01755 K01756 K01914

					K01939 K01955 K01956 K11541
Prevotella_intermedia	ko00250	60	16	26.67	K00262 K00266 K00278 K00609 K00610 K00764 K00820 K01424 K01744 K01755 K01756 K01939 K01953 K01955 K01956 K11541
Prevotella_salivae	ko00250	60	16	26.67	K00259 K00262 K00266 K00609 K00610 K00764 K00820 K01424 K01744 K01755 K01756 K01914 K01939 K01955 K01956 K11541
Prevotella_sp__oral_taxon_299	ko00250	60	14	23.33	K00262 K00609 K00610 K00764 K00820 K01424 K01744 K01755 K01756 K01914 K01939 K01955 K01956 K11541
Dialister_succinatiphilus	ko00400	70	21	30	K00014 K00766 K00800 K00812 K00817 K00832 K00891 K01609 K01626 K01657 K01658 K01695 K01696 K01713 K01735 K01736 K03856 K04517 K06001 K13497 K14170
Prevotella_copri	ko00400	70	19	27.14	K00014 K00210 K00766 K00800 K00812 K00817 K00891 K01609 K01657 K01658 K01695 K01696 K01735 K01736 K01817 K03786 K04516 K04518

Prevotella_dentalis	ko00400	70	20	28.57	K06001 K00014 K00210 K00766 K00800 K00812 K00817 K00891 K01609 K01626 K01657 K01658 K01695 K01696 K01735 K01736 K01817 K03786 K04516 K04518
Prevotella_fusca	ko00400	70	16	22.86	K06001 K00014 K00210 K00766 K00800 K00812 K00817 K00891 K01657 K01658 K01735 K01736 K01817 K03786 K04516 K04518
Prevotella_intermedia	ko00400	70	12	17.14	K06001 K00014 K00210 K00800 K00812 K00891 K01609 K01657 K01658 K01735 K01736 K03786 K04516
Prevotella_ruminicola	ko00400	70	19	27.14	K00014 K00210 K00766 K00800 K00812 K00817 K00891 K01609 K01657 K01658 K01695 K01696 K01735 K01736 K01817 K03786 K04516 K04518
Prevotella_salivae	ko00400	70	14	20	K06001 K00014 K00210 K00766 K00800 K00812 K00891 K01609 K01626 K01735 K01736 K03786 K04516 K04518
Bacteroides_fluxus	ko00473	5	1	20	K06001 K01775

Prevotella_copri	ko00473	5	1	20	K01775
Prevotella_dentalis	ko00473	5	1	20	K01775
Prevotella_salivae	ko00473	5	1	20	K01775
Fermentimonas_caenicola	ko00510	44	1	2.27	K00721
Prevotella_copri	ko00510	44	1	2.27	K00721
Prevotella_dentalis	ko00510	44	1	2.27	K00721
Acidaminococcus_intestini	ko00540	34	15	44.12	K00677 K00748 K00979 K01627 K02517 K02527 K02535 K02536 K02843 K03270 K03271 K03272 K03273 K03274 K16363
Bacteroides_fluxus	ko00540	34	11	32.35	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K16363
Bacteroides_sp__3_2_5	ko00540	34	13	38.24	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K03271 K03273 K16363
Bibersteinia_trehalosi	ko00540	34	15	44.12	K00677 K00748 K00912 K00979 K01627 K02527 K02535 K02536 K02560 K02843 K03270 K03271 K03272 K03273 K03274
Dialister_succinatiphilus	ko00540	34	12	35.29	K00677 K00748 K00912 K00979 K01627 K02527 K02535 K02536 K03271 K03272 K07031 K16363
Fermentimonas_caenicola	ko00540	34	10	29.41	K00677 K00748 K00912 K00979 K01627 K02527 K02536 K03269 K03270

Prevotella_copri	ko00540	34	14	41.18	K16363 K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K03271 K03273 K07031 K16363
Prevotella_fusca	ko00540	34	11	32.35	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K16363
Prevotella_melaninogenica	ko00540	34	11	32.35	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K16363
Prevotella_oralis	ko00540	34	13	38.24	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K03271 K03273 K16363
Prevotella_ruminicola	ko00540	34	11	32.35	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K16363
Prevotella_salivae	ko00540	34	11	32.35	K00677 K00748 K00912 K00979 K01627 K02517 K02527 K02536 K03269 K03270 K16363
Riemerella_anatipestifer	ko00540	34	6	17.65	K00677 K00748 K01627 K02517 K02536 K16363
Prevotella_copri	ko00550	38	15	39.47	K00075 K00790 K01000 K01921 K01924 K01925 K01928 K01929 K02563

Prevotella_dentalis	ko00550	38	15	39.47	K03587 K03814 K05366 K05515 K06153 K07259 K00075 K00790 K01000 K01921 K01924 K01925 K01928 K01929 K02563
Prevotella_salivae	ko00550	38	15	39.47	K03587 K03814 K05366 K05515 K06153 K07259 K00075 K00790 K01000 K01921 K01924 K01925 K01928 K01929 K02563
Prevotella_sp__oral_taxon_299	ko00550	38	15	39.47	K03587 K03814 K05366 K05515 K06153 K07259 K00075 K00790 K01000 K01921 K01924 K01925 K01928 K01929 K02563
Bacteroides_fluxus	ko00623	51	3	5.88	K03587 K03814 K05366 K05515 K06153 K07259 K00239 K00240 K00241
Prevotella_copri	ko00623	51	3	5.88	K00239 K00240 K00241
Prevotella_dentalis	ko00623	51	3	5.88	K00239 K00240 K00241
Prevotella_salivae	ko00623	51	3	5.88	K00239 K00240 K00241
Prevotella_oralis	ko00670	27	11	40.74	K00239 K00240 K00241 K00287 K00297 K00548 K00560 K00602 K00604 K00605 K01491 K01934 K01938 K08289
Prevotella_salivae	ko00670	27	10	37.04	K00287 K00297 K00548 K00560 K00602 K00604 K01491 K01938 K03465 K08289
Bibersteinia_trehalosi	ko00710	37	7	18.92	K00029 K00134 K00615 K00927 K01595 K01783 K01807
Clostridiales_bacterium_VE202_14	ko00710	37	6	16.22	K00029 K00134 K00615

					K00927 K01006 K01783
Prevotella_copri	ko00710	37	6	16.22	K00029 K00134 K00615 K00927 K01006 K01783
Prevotella_melaninogenica	ko00710	37	6	16.22	K00029 K00134 K00615 K00927 K01006 K01783
Prevotella_salivae	ko00710	37	6	16.22	K00029 K00134 K00615 K00927 K01006 K01783
Streptococcus_pneumoniae	ko00710	37	7	18.92	K00134 K00150 K00615 K00927 K01595 K01783 K01807
Prevotella_copri	ko00760	44	9	20.45	K00767 K00858 K00969 K01081 K01950 K03426 K03517 K03783 K03787
Prevotella_denticola	ko00760	44	10	22.73	K00763 K00767 K00858 K00969 K01081 K01950 K03426 K03517 K03783 K03787
Prevotella_fusca	ko00760	44	10	22.73	K00763 K00767 K00858 K00969 K01081 K01950 K03426 K03517 K03783 K03787
Prevotella_sp__oral_taxon_299	ko00760	44	9	20.45	K00763 K00767 K00858 K00969 K01081 K01950 K03426 K03783 K03787
Fermentimonas_caenicola	ko00900	50	10	20	K00099 K00806 K00919 K00991 K01662 K01770 K02523 K03526 K03527 K13789
Prevotella_copri	ko00900	50	11	22	K00099 K00806 K00919

					K00991 K01662 K01770 K02523 K03526 K03527 K12506 K13789
Prevotella_dentalis	ko00900	50	11	22	K00099 K00806 K00919 K00991 K01662 K01770 K02523 K03526 K03527 K12506 K13789
Prevotella_sp__oral_taxon_299	ko00900	50	11	22	K00099 K00806 K00919 K00991 K01662 K01770 K02523 K03526 K03527 K12506 K13789
Acidaminococcus_sp__D21	ko00908	8	1	12.5	K00791
Fermentimonas_caenicola	ko00908	8	1	12.5	K00791
Megasphaera_elsdenii	ko00908	8	1	12.5	K00791
Prevotella_copri	ko00908	8	1	12.5	K00791
Prevotella_dentalis	ko00908	8	1	12.5	K00791
Bibersteinia_trehalosi	ko03018	75	12	16	K00962 K01689 K03628 K03654 K03666 K03732 K04043 K04077 K05592 K08300 K08311 K12573
Prevotella_copri	ko03018	75	10	13.33	K00962 K00970 K03628 K03654 K04043 K04077 K05592 K11927 K12573 K12574
Prevotella_salivae	ko03018	75	10	13.33	K00962 K00970 K01689 K03628 K03654 K04043 K04077 K05592 K11927 K12573
Prevotella_salivae	ko03060	39	10	25.64	K03070 K03075 K03076 K03100 K03101 K03106 K03110 K03210 K03217 K12257

Acidaminococcus_intestini	ko05111	43	1	2.33	K07173
Prevotella_copri	ko05111	43	1	2.33	K07173
Prevotella_melaninogenica	ko05111	43	1	2.33	K07173
Fermentimonas_caenicola	ko05120	90	1	1.11	K08303

1

2a Biomarker specie.

3b Biomarker pathways.

4c Number of genes in pathways.

5d Number of genes in pathways that map in the biomarker specie.

6e List of genes in pathways that map in the biomarker specie.

7

8

9

10

11

12

13

14

15

16

17

18

19

Supplementary Table 5. Network statistics.

	Mean Degree	Mean betweenness centrality	Mean closeness centrality	Mean Eigenvector centrality	%Pos	%Neg	Ratio	Fragmentation	Modularity	Transitivity	Diameter	Mean short-pathways	K-Smirnov p-value	Node
Co-corrence network	3.88	262.38	0.001	0.106	0.431	0.569	0.758	0.134	0.428	0.155	10	4.079	0.603	176
Enzimatic network	17.15	10364.49	1.97e-07	0.018				0.433	0.780	0.369	26	6.600	0.702	3700

(a) Percentage of edges with a statistically significant positive correlation

(b) Percentage of edges with a statistically significant negative correlation

(c) Ratio between the positive and negative correlations.

(d) Modularity index retrieved from the modules estimate by the Walktrap community finding algorithm

(e) Clustering coefficient of the network. measure of the degree to which nodes in a graph tend to cluster together.

(f) Longest path between two vertices within the network

(g) Average path length between any two nodes

Supplementary Table 6. Bayesian network quantile 95 nodes. The quantile 95 based on the three different statistics: the degree centrality, the number of direct connections with a clinical variable and the MB size. DNAseq referred to metagenomics and RNAseq to metatranscriptomics.

Node name	Number of direct connections (degree centrality)
DNA seq Terpenoid backbone biosynthesis ko00900	17
BPI (ng/mL)	17
DNA seq Valine leucine and isoleucine degradation ko00280	18
RNA seq Butanoate metabolism ko00650	18
%CD4+HLA−DR+CD38+ T−cells	18
DNA seq Alanine aspartate and glutamate metabolism ko00250	19
DNA seq Aminobenzoate degradation ko00627	19
sj/β−TREC ratio	22
CD4+ T−cell counts (cells/uL)	24
DNA seq Nicotinate and nicotinamide metabolism ko00760	25

Node name	Number of direct connections with a clinical variable
DNA seq Limonene and pinene degradation ko00903	7
Metabolite Diradylglycerols [GL02]|Glycerolipids [GL]|LMGL02010346	7
Metabolite Diradylglycerols [GL02]|Glycerolipids [GL]|582.5218@10.206741	7
DNA seq Glyoxylate and dicarboxylate metabolism ko00630	9
Metabolite Steroid conjugates [ST05]|Sterol Lipids [ST]|LMST05010043	9
DNA seq Valine leucine and isoleucine degradation ko00280	10
RNA seq ABC transporters ko02010	11
DNA seq Nicotinate and nicotinamide metabolism ko00760	12
RNA seq Butanoate metabolism ko00650	16

Node name	Markov blanket size in nodes
DNA seq Aminobenzoate degradation ko00627	67
Metabolite Diradylglycerols [GL02]|Glycerolipids [GL]|582.5218@10.206741	67
Metabolite Steroid conjugates [ST05]|Sterol Lipids [ST]|LMST05010043	68
RNA seq ABC transporters ko02010	69
DNA seq Glyoxylate and dicarboxylate metabolism ko00630	74
DNA seq D Alanine metabolism ko00473	75