Supplementary Information

Table of Contents

I. Supplementary Methods

1 Sample preparation

1.1 Sample collection

Sample collection for all European studies was approved by Ethics committee (different committees for different studies).

1.1.1 Danish individuals

Danish individuals were examined at Steno Diabetes Center, Gentofte. The participants were asked to provide a frozen, crude fecal sample. Samples were collected at home, and immediately frozen in their home freezer. The samples were delivered to Steno Diabetes Center using insulating polystyrene foam containers and stored at -80°C until analysis.

1.1.2 Spanish individuals

Patients with ulcerative colitis (UC) or Crohn's disease (CD) attending the outpatient clinic of Hospital Vall d'Hebron were asked to give written consent to take part in this study. Eligible patients were aged 18 to 75 years, had UC or CD previously diagnosed by endoscopy and histological examination of intestinal mucosal biopsies. They were in clinical remission for at least 3 months, and had stable maintenance therapy with mesalazine or azathioprine. Exclusion criteria included pregnancy or breast feeding, severe concomitant disease involving the liver, heart, lungs or kidneys, treatment with steroids, cyclosporine, anti-TNF drugs or topical anti-inflammatory preparations during the previous 3 months, and treatment with antibiotics during the previous 4 weeks. Healthy controls were recruited among family relatives of the UC and CD patients; antibiotic treatment for at least 4 weeks before fecal sample collection was excluded. The protocol was approved by the Ethics Committee of our institution (CEIC, Hospital Vall d'Hebron).

Patients and healthy controls were asked to provide a frozen stool sample. Fresh stool samples were obtained at home, and samples were immediately frozen by storing them in their home freezer. Frozen samples were delivered to the Hospital using insulating polystyrene foam containers, and then they were stored at -80°C until analysis. None of the patients or controls underwent bowel cleansing or endoscopic procedures before fecal sampling.

1.1.3 Italian individuals

Fecal samples were obtained from 6 elderly living in Camerino, Italy. The elderly volunteers consumed an unrestricted Western-type diet. They took neither antibiotics nor any drug known to influence the fecal microbiota composition for at least three months prior to sampling and were free of known metabolic or gastrointestinal diseases. Whole stools were collected in clean boxes and stored at 4°C under anaerobic conditions using an anaerocult® A (Merck, Nogent sur Marne, France) until sampling as 200 mg aliquots in 2 ml sterile screw-cap tubes which were frozen at -20°C for further analysis.

1.1.4 French individuals

Fecal samples were obtained from 4 obese and 4 healthy individuals, frozen immediately, and delivered at INRA.

1.2 DNA extraction

1.2.1 Danish, French, Italian and Spanish individuals

A frozen aliquot (200 mg) of each fecal sample was suspended in 250 μl of 4 M guanidine thiocyanate–0.1 M Tris (pH 7.5) and 40 μl of 10% N-lauroyl sarcosine. Then, DNA extraction was conducted using bead beating method as previously described[37]. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and agarose gel electrophoresis.

2 Sequence processing

2.1 European samples

Sanger sequencing was performed using standard protocols. Shotgun randomly shared DNA libraries were constructed using low copy plasmid (pCNS, 3 kb insert). Terminal clone end sequences were determined using BigDye terminator chemistry and capillary DNA sequencers (3730XL, Applied Biosystems) according to standard protocols established at Genoscope. Cloning vector and sequencing primer were removed from raw reads after aligning reads to the vector/primer sequences using BLASTN. Reads were quality trimmed by removing bases in either end with phred quality under 15. Lastly, reads shorter than 300bp were removed.

2.2 American samples

Sanger reads for two American adult human gut metagenomes[4] were downloaded from NCBI Trace Archive. The vector and sequence trimming coordinates from the trace information were used to remove the cloning vector and sequencing primer.

Titanium reads for two American female obese individuals[5] were downloaded from the NCBI Short Read Archive. These reads were not processed further than the trimming provided by the authors.

2.3 Japanese samples

We identified the following unclipped vector/linker sequences in the Japanese samples:

1. 5'- GAGAGCTCCTGCAGGCTAGCTTGCGCAAGGATCCTAGGCCTGAAGCTTGTC - 3'
2. 5'- GCATGGTACCACGCGTACGTAAGCAAGATCTTCCCGGGTGAATTCGTC - 3'

These sequences from the pTS1 cloning vector (K. K. and T. H., personal communication) were clipped from the 13 Japanese samples using the makeClip program from Forge assembler[41].

2.4 Further trimming

All Sanger reads were finally trimmed for low quality regions in the ends using makeClip.

2.5 Removal of potential human reads

Sequence reads were aligned against human genome assembly hg18 obtained from UCSC Genome Browser[42] using BLAT[43] (gfClient v 31, default parameters). Possible human DNA sequences were identified with a very low alignment threshold to maximize true positives and minimize false negatives ('pslFilter -minMatch=50' from the BLAT package), and were removed.

3 Assembly and gene prediction

Assembly and gene prediction were performed using the SMASH comparative metagenomics pipeline[38]. To obtain contigs and scaffolds from the reads, we employed SMASH's iterative assembly procedure using Arachne software[44-45]. This procedure iteratively assembles unassembled reads (singletons) from the previous iteration until no more assembly is possible. Protein coding genes were predicted using GeneMark[46] (v 2.6p) by the SMASH pipeline. SMASH uses the GC-content based heuristic models (provided with GeneMark software) to predict genes on scaffolds shorter than 200kb as well as unassembled reads, and a self-trained hybrid model using both GC-content and sequence content on scaffolds longer than 200kb.

4 Phylogenetic annotation

Phylogenetic annotation of each metagenome sample was performed using the SMASH pipeline[38].

4.1 Reference genome set

We obtained a set of 1511 reference microbial genomes from the National Center for Biotechnology Information (NCBI), Human Microbiome Project[10] and the MetaHIT Consortium[11]. We identified 16S rRNA gene sequences from each of these genomes using an HMM-based algorithm[47] and assigned a taxonomic rank to the genome based on the classification of the 16S rRNA gene using the RDP Classifier[39]. We used the taxonomic tree provided with the RDP Classifier, which is based on the bacterial taxonomy proposed by the Taxonomic Outline of Bacteria and Archaea[48], with further rearrangements proposed for Firmicutes and Cyanobacteria by the Bergey's Manual of Systematic Bacteriology[49-50].

4.2 Reference genome mapping

Sequence reads were aligned to the reference microbial genomes (listed in Supplementary Table 3) using BLASTN (WU-BLAST 2.0, default parameters except E=1e-20 Z=4000000000 B=5). Each read was assigned the taxonomy of the highest scoring hit(s) above the similarity threshold for the taxonomic rank (>65% for phylum and >85% for genus established by parameter exploration, see Supplementary Methods Section 4.3). Alignments were also required to span over 75bp covering >80% of the read length. Since paired-end reads are from two ends of a cloned DNA fragment, two reads from such a fragment represent only one physical DNA fragment. Hence taxonomy assignments of reads were

transferred to the corresponding fragments. The numbers of fragments assigned to each reference genome were counted. (A fragment assigned to *N* different reference genomes contributes 1/*N* to each genome). These counts were normalized by the sizes of these genomes to obtain the quantitative relative abundance (relative number of individuals) of each genome in the sample. Number of unassigned fragments was normalized by the average genome size in the reference set (3.54Mb) to calculate the approximate abundance of unknown genomes. Phylogenetic abundances at various phylogenetic ranks (species, genus, phylum etc) were calculated by adding the abundances of genomes under that rank.

4.3 Estimating sequence similarity barriers across phylogenetic ranks

Since there are no established sequence similarity barriers to differentiate genomes from different phylogenetic ranks, we estimated the sequence similarity cutoffs to safely assign a sequence to either a genus or a phylum. For this purpose, we retrieved 40 single copy marker genes[51] from a subset of 835 genomes (after removing some redundancy at species level) and generated 40 sets of pairwise alignments using BLASTN. These marker genes are highly representative of the reference genome set, and hence of at least the sequenced microbial species, since 801 of the 853 genomes (94.6%) contained at least 38 out of the 40 genes. Supplementary Figure 1a shows the distribution of sequence similarity levels between genomes from the same phylum (green) and different phyla (red). Supplementary Figure 1b shows the same distribution at genus level. We estimated the false positive rates at different similarity thresholds at both phylum and genus levels (Supplementary Figure 15). At the phylum level, a 65% threshold had 0.77% false positive rate; at the genus level, an 85% threshold had 1.84% false positive rate. Thus we chose 65% and 85% as the thresholds for the genus and phylum level assignments. This is a rather conservative cutoff, since the marker genes are among the genes under the highest levels of selective constraint[51].

4.4 HITChip analysis

10 ng from the fecal DNA extract was used to amplify the 16S rRNA genes with the *T7prom*-Bact-27-for and Uni-1492-rev primers. Subsequently, an *in vitro* transcription and labeling with Cy3 and Cy5 dyes, was performed. Fragmentation of Cy3/Cy5 labeled target mixes was followed by hybridization on the arrays at 62.5˚C for 16h in a rotation oven (Agilent Technologies, Amstelveen, The Netherlands). The slides were washed and dried before scanning. Signal intensity data was obtained from the microarray images using the Agilent Feature Extraction software, version 9.1 (http://www.agilent.com). Microarray data normalization and further analysis was performed using a set of R-based scripts (http://r-project.org) in combination with a custom designed relational database[20] which operates under the MySQL database management system (http://www.mysql.com).

5 Phylogenetic analysis of external datasets

5.1 Analyzing 16S rRNA sequences from 154 American individuals

Published 16S rRNA sequence data[5] derived from fecal samples of 154 individuals, including female monozygotic and dizygotic twin pairs and their mothers, were downloaded from

http://gordonlab.wustl.edu/NatureTwins_2008/V2.fasta.gz. This dataset containing 1119519 sequence reads from the V2 region of the 16S rRNA gene was processed using the SMASH pipeline[38] to classify the reads using the RDP Classifier[39] (using minimum read length of 200 and a minimum confidence score of 0.5). For each sample, the number of reads assigned to different genera by the RDP classifier were normalized by the average 16S gene copy number in genomes belonging to each genus obtained from rrnDB[52]. This resulting relative abundance profile at the genus level was used for further analysis.

5.2 Analyzing Illumina-based metagenome sequences from 85 Danish individuals

Illumina reads from 85 Danish individuals from a previously published dataset[8] (we focused on the Danish individuals as most of the Spanish ones came from patients with Crohn's disease with dysbiosis and a known reduction of species complexity, which introduces biases and caused technical difficulties in assignment and analysis) were quality trimmed and filtered using a customized pipeline based on the FASTX toolkit[53]. Briefly, (i) bases were trimmed from the beginning of reads unless the number of base calls for any base (A, T, G, C) was within the average across all cycles plus/minus two standard deviations, (ii) bases were trimmed from the end of reads if the quality score was <20, and (iii) reads shorter than 35 bp or reads with a median quality score < 20 were removed. The resulting high-quality reads were mapped to the reference microbial genomes (split into two files) using SOAP version 2.20[54] with the option –r 2 (keep all best hits). The mapping results were merged and filtered, i.e., only higher-scoring reads (or read-pairs) were kept if the number (or the sum) of mismatches of reads (or read-pairs) mapping to both reference files was not the same. Taxonomies were assigned by the same procedure as described above for the Sanger reads.

6 Functional annotation

Functional annotation of each sample was performed using the SMASH pipeline[38].

6.1 Estimating abundance of a gene/protein

Abundance of each predicted gene from a sample was estimated analogous to the contig coverage in sequence assembly. If $R = \{r\}$ is the set of assembled reads overlapping the locus of predicted gene g in a contig, abundance of g was calculated as

$$\text{abundance}(g) = \sum_{r \in R} \frac{\text{base\_overlap}(g, r)}{\text{base\_length}(g)} \quad (1)$$

Genes on a singleton read thus have an abundance of 1.

6.2 Assigning proteins to eggNOG orthologous groups

Each predicted protein was assigned to an orthologous group in the eggNOG v2 database[12]. We use the term "orthologous group" in the broadly accepted sense to mean group of genes "assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events"[55]. Readers are advised to refer to the original definition of clusters of orthologous groups (COGs) in ref. 55.

Predicted proteins were aligned to proteins from the eggNOG v2 database using BLASTP (WU-BLAST 2.0, default parameters except E=1e-5 B=10000) and were assigned to an orthologous group as described elsewhere[56]. From these alignments between the set of predicted proteins $G = \{g\}$ from a sample and the set of eggNOG reference proteins $K = \{k\}$, the abundance of each reference protein k in the sample was calculated as

$$\text{abundance}(k) = \sum_{g \in G} \frac{\text{aa\_overlap}(k, g) * \text{abundance}(g)}{\text{aa\_length}(k)} \quad (2)$$

Functional abundances at the OG level were calculated by adding abundances of reference proteins under each OG.

6.3 Assigning proteins to KEGG orthologous groups, modules and pathways

Predicted proteins were also aligned to proteins from Kyoto Encyclopedia of Genes and Genomes (KEGG) database[57] as before. Each protein was assigned to the KEGG orthologous group (KO) containing the highest scoring annotated hit(s) containing at least one HSP scoring over 60 bits. The abundance of each KEGG protein was calculated as in Equation (2). Functional abundances at KO, KEGG module and KEGG pathway levels were calculated by adding abundances of KEGG proteins under each KO, module and pathway, respectively. We use the term functional module to mean "smaller pieces of subpathways manually defined as consecutive reaction steps, operon or other regulatory units, phylogenetic units obtained by genome comparisons"[57] as defined by KEGG.

7 Clustering

7.1 Clustering algorithm

We used the Partitioning around medoids (PAM) clustering algorithm[58] to cluster the abundance profiles. PAM derives from the basic k-means algorithm, but has the advantage that it supports any arbitrary distance measure and is more robust than k-means[58].

7.2 Distance metric

Genus abundance profiles (phylogenetic) and OG abundance profiles (functional) were normalized to generate probability distributions (called abundance distributions hereafter). We used a probability distribution distance metric[59-60] related to Jensen-Shannon divergence (JSD) to cluster the samples. The distance $D(a, b)$ between samples a and b is defined as

$$D(a, b) = \sqrt{JSD(p_a, p_b)} \quad (3)$$

where p_a and p_b are the abundance distributions of samples a and b and $JSD(x, y)$ is the Jensen-Shannon divergence between two probability distributions x and y defined as

$$JSD(x,y) = \frac{1}{2}KLD(x,m) + \frac{1}{2}KLD(y,m) \quad (4)$$

where $m = \frac{x+y}{2}$ and $KLD(x,y)$ is the Kullback-Leibler divergence between x and y defined as

$$KLD(x,y) = \sum_i x_i \log \frac{x_i}{y_i} \quad (5)$$

We added a pseudocount of 0.000001 to the abundance distributions and renormalized them to avoid zero in the numerator and/or denominator of equation (5).

7.3 Optimal number of clusters

To assess the optimal number of clusters our dataset was most robustly partitioned into, we used the Calinski-Harabasz (CH) Index[61] that has shown good performance in recovering the number of clusters[62]. It is defined as:

$$CH_k = \frac{\frac{B_k}{k-1}}{\frac{W_k}{n-k}} \quad (6)$$

where B_k is the between-cluster sum of squares (i.e. the squared distances between all points i and j, for which *i* and *j* are not in the same cluster) and W_k is the within-clusters sum of squares (i.e. the squared distances between all points *i* and *j*, for which i and j are in the same cluster). This measure implements the idea that the clustering is more robust when between-cluster distances are substantially larger than within-cluster distances. Consequently, we chose the number of clusters *k* such that CH_k was maximal.

7.4 Cluster validation

Cluster validation methods are useful to assess the quality of a clustering with respect to the underlying data points. Here we use the silhouette validation technique[21]. The silhouette width *S(i)* of individual data points *i* is calculated using following formula:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (7)$$

where *a(i)* is the average dissimilarity (or distance) of sample *i* to all other samples in the same cluster, while *b(i)* is the average dissimilarity (or distance) to all objects in the closest other cluster.

The formula implies $-1 \leq S(i) \leq 1$. A sample which is much closer to its own cluster than to any other cluster has a high *S(i)* value, while *S(i)* close to 0 implies that the given sample lies somewhere between two clusters. Large negative *S(i)* values indicate that the sample was assigned to the wrong cluster. To

obtain a global assessment of the cluster quality, the average *S(i)* over all data points is a useful measure.

7.5 Cluster similarity measurement

To assess cluster similarity between the enterotype clusters and the new clusters, we determined (a) the number of pairs of related samples in both clustering, (b) the number of pairs of samples not related in both clustering, (c) the number of related samples in new clusters but not in the enterotype clusters, (d) the number of related samples in enterotype clusters but not in new clusters. These data were used to calculate the Rand index[63] R which give the cluster similarity between the enterotype clustering and the new clustering.

$$R = \frac{(a+b)}{(a+b+c+d)}$$

8 Simulating phylogenetic/functional compositions

For a given set of *N* feature vectors, each representing the phylogenetic (genus, phylum, etc) or the functional (gene, orthologous group, functional modules, etc) composition of a metagenomic sample, we simulated *N* hypothetical metagenomic samples containing the same number of features by sampling from a continuum as follows:

random-uniform: each generated feature is uniformly distributed in the interval [0,1)

template-uniform: each generated feature corresponds to a feature in the real dataset and its values across multiple samples is uniformly distributed between the minimum and maximum abundances of that feature in the real dataset

template-Gaussian: each generated feature corresponds to a feature in the real dataset and its values across multiple samples followed a Gaussian distribution with the same mean and standard deviation as observed in the corresponding real data

Generated values are then normalized so that abundances within a sample sum to 1.

9 Supervised learning

We built predictive models from the enterotype clusterings of all different datasets as well as from simulated data using decision tree learning[64]. We evaluated the ability of these models to accurately predict new data points in a leave-one-sample-out cross-validation scheme and compared the models using different statistics. We estimated the classification accuracy defined as the ratio between the number of correctly assigned samples and the number of all samples. Further, we estimated precision, which is the ratio between the number of true assignments to a given class and the number of all samples assigned to this class and averaged it over all classes. Additionally, average precision gain was estimated based on this as the ratio of the actual precision and the precision of random guessing averaged over the three classes. Here it is important to note that classification accuracy and average precision are measures which are influenced by class abundances, which may differ between real and

simulated data. In contrast, average precision gain is implicitly normalized for this and thus best suited for comparisons between different data sets.

10 Between-class analysis

Between-class analysis was performed to support the clustering and identify the drivers for the enterotypes. The analysis was done using R with the ade4 package[65]. Prior to the analysis the data was sample size normalized and very low abundant genera / orthologous groups were removed to decrease noise if their average abundance across all samples was below 0.01%. The between-class analysis is a particular case of principal component analysis with an instrumental variable: here the variable is a qualitative factor (i.e. enterotypes cluster). The between-class analysis enables us first to find the principal components based on the center of gravity of each group in a way to highlight differences between groups and then to link each sample with its group. According to the dataset, the between-class analysis was based on genera or OGs abundance using scaling and centering. In addition to this analysis concerning the microbial composition, best top species that mainly contribute to each principal component obtained from between-class analysis were highlighted in the graphical representation (such as Fig. 2).

11 Network correlation analysis

Spearman correlations were computed between the three main contributors (*Prevotella*, *Bacteroides* and *Ruminococcus*) and other genera. 5% of the correlations had an absolute Spearman correlation above 0.4, and these correlations were transformed into links between two genera in the genus network. The "network" package in R was then used to construct network figures with a spring-based algorithm.

12 Statistical treatment of over-/under-representation

Over- and underrepresented features (OGs, KEGG modules, KEGG maps, genera, and phyla) were identified using Fisher's exact test on pooled counts depending on the sample groups compared. Correction for multiple testing was done based on the Benjamini-Hochberg False Discovery Rate (corrected p-value <0.05). To avoid artifacts, we only took those features into account that were specifically overrepresented in only one metadata group (e.g., only in one enterotype). Case studies described in the main text were further manually scrutinized to avoid artifacts.

13 Correlations with host properties

Correlation analysis between host metadata (Supplementary Table 1) and feature (OG, module, pathway, genus, phylum) frequencies was done as described previously[40]. In short, Spearman pairwise correlations between continuous metadata variables (age, bmi) were calculated and p-values were corrected for multiple testing using Benjamini-Hochberg False Discovery Rate correction. Significant features were used as input for building linear models using stepwise regression (top-down and bottom-up feature selection) based on the Akaike Information Criterion. For categorical metadata, samples were pooled into bins (male/female, obese/lean, specific nationality/rest) and treated as in Supplementary

Methods Section 12. For nationality analysis, also the general variability of features across nations was investigated. For each nationality, we calculated the standard deviation (SD) of investigated features (relative abundance of OGs, genera) across samples, and compared this to the SD of the distribution of mean relative abundances of each nationality (to measure across-nationality variation). Examples with a across SD/within SD ratio >1 are discussed in the main text.

II. Supplementary Notes

1 Feasibility of comparative gut metagenomics

We compared the metagenomes derived from fecal samples from 22 European individuals and 17 others from two other continents. European individuals were drawn from four different nations (Denmark, France, Italy and Spain), included both healthy individuals and patients suffering from microbiota-associated disorders (6 obese individuals and 2 inflammatory bowel disease patients; see Supplementary Table 1) and were selected for a broad range of microbiota (8 of a larger group of 40 were found by HITChip[20] analysis to be particularly divergent and 6 were over 70 years old, as it was reported that the diversity of the microbiota increases with age[66-67]). Their fecal metagenomes were Sanger-sequenced at an average depth of 105 Mb each (Supplementary Table 1 and Supplementary Table 2). Of the non-European samples, 13 were from Japan[6] and 4 from America[4-5], sequenced at an average depth of 61Mb and 92 Mb, respectively; the latter include only 2 out of 18 metagenomes from one American study[5] determined using pyrosequencing[68] that had sufficient read length and sequencing depth.

We developed unified phylogenetic and functional annotation protocols to analyze data generated by different sequencing centers using different sample preparation protocols and sequencing technologies. We establish the feasibility of comparative metagenomic analysis on data with diverse origins using the following methods.

1.1 Functional repertoire of samples compared to bacterial genomes

We derived the relative abundance of COG functional categories in 575 microbial genomes in the STRING (v8) database[15], containing a subset of 53 gut-specific genomes (see Supplementary Notes Section 1.3 for details), after normalizing for genome size (red and blue dots in Supplementary Figure 16 respectively). We also derived their relative abundance in the metagenomes through the functional annotation of predicted genes as described in Supplementary Methods Section 6. The coverage of functional categories is similar between samples (grey box-plots). Functional profiles of metagenomes and 53 gut-specific microbial genomes are significantly different (overrepresented) only for category L (replication, recombination and repair) and V (defense mechanisms), even though other differences between the profiles of metagenomes and the 575 microbial genomes can be observed, such as enrichment of M (cell wall/membrane/envelope biogenesis) and G (carbohydrate transport and metabolism), and depletion of I(Lipid transport and metabolism) and Q(Secondary metabolites biosynthesis, transport and catabolism).

1.2 Comparing different sequencing technologies

We sequenced two of the Danish samples (DA-AD-1 and DA-AD-3) using Sanger and 454 Titanium methods and compared the species retrieval and function retrieval to compare the technologies (See Supplementary Table 20). We used the reference genome mapping approach (See Supplementary Methods Section 4.1) on all sequences and recorded the genus distribution. Although the greater depth of sequencing with 454 retrieves additional rare genera, the overall genus distributions are similar

(Pearson correlation coefficients 0.9852 and 0.9968, also see Supplementary Figure 17a). We recorded the functional abundance of eggNOG orthologous groups (OGs) in each sample (See Supplementary Methods Section 6). The OG abundance distributions remain similar between the technologies (correlation: 0.9482 and 0.9153, see Supplementary Figure 17b). Notably, deeper 454 sequences retrieve slightly fewer OGs (Supplementary Table 20), probably due to shorter gene fragments (resulting from shorter reads or contigs) that make orthology assignment more difficult. The OGs that differ most in abundance between the Sanger- and pyrosequencing-based metagenomes were mostly unknown functions (Supplementary Figure 18), showing that functional interpretation of comparative analysis is not affected by this difference. These results imply that future samples from different sequencing technologies can be integrated and compared, provided that the sequencing coverage is sufficient to discriminate between meaningful and random variation.

1.3 Functional rarefaction

We simulated the total number of orthologous groups (OGs) that could be functionally assigned in relation to the number of sequenced samples (Fig. 1a). As many genes might be 'bystanders' i.e. genes from transient, perhaps food-associated microbiota that just passage through the gut, we assigned habitat information to 1368 out of the 1511 reference genomes and distinguished between eggNOG orthologous groups from gut and non-gut species. As expected, OGs found in known gut species seem to be close to saturation while functions from 'non-gut' species still accumulate with each sample at our given coverage of 53-295Mb per individual (Fig. 1a). Thus, although the coverage at hand will miss rare gut species and genes from these, the coverage seems sufficient to cover major trends caused by resident gut species and to robustly identify genera and functionalities that are common and different between samples.

We computed the total number of eggNOG[12] orthologous groups present in random combinations of n individuals (with n=2 to 35 (all samples except infants), 100 replicates per bin). From all genomes present in STRING and our reference genome set, two sets of orthologous groups were defined: those present in genomes known to be found in the mammalian intestine ("gut", Supplementary Table 3) and all remaining organisms ("non-gut"). Gut genomes were identified using text mining of species descriptions on the sequencing centre websites[69-74], the NCBI genome overview page[75], the GOLD database[76], and by further manual curation. Of all 1511 genomes (1069 species) in the reference genome set, 1368 genomes (958 species) could be classified. 325 reference genomes (192 species) were classified as gut associated species. 54 of these 192 gut species are in STRING (see Supplementary Figure 19).

2 Global phylogenetic and functional variation of intestinal metagenomes

2.1 Non bacterial DNA content

Before we identified the phylogenetic profile of the metagenomes, we investigated the non-bacterial DNA content in the samples. For this part of the analysis only, we counted the number of reads per

samples, since this is not used in a quantitative manner for comparative analysis. This is different from the quantitative abundance estimation by counting the mate-paired reads as a single DNA fragment, as in Supplementary Notes Sections 2.2 and 2.3.

2.1.1 Eukaryotic contamination

Although the gut ecosystem consists mostly of microbes, host human genome can contaminate the gut metagenome samples, and so can DNA from ingested plants and meat. We screened for these two types of contamination and estimated the levels to be very low. These estimates are an upperbound for the eukaryotic contamination, since we use a more sensitive criterion.

2.1.1.1 Human DNA contamination

On an average 0.14% of the reads from the 22 European samples were classified as potential human reads using this method (see Supplementary Table 21). This is very low considering that we used very lenient criteria to capture as many human sequences as possible. Note that the published datasets may have been screened for human reads before being made publicly available and 0.02% of the reads from these datasets were classified as potential human reads.

2.1.1.2 Other eukaryotic contamination

The eukaryotic DNA fraction (possibly from food intake), estimated by identifying metagenome proteins whose best hit in STRING v8 database comes from a eukaryote, amounts to less than 1.3% of the DNA fragments from any sample (0.5% on an average, see Supplementary Table 21). Interestingly, 5 out of 7 eukaryotic kingdoms in STRING v8 were found in all 39 samples. We found extremely low human contamination (0.0068%) that our prescreening step (see Supplementary Notes Section 2.1.1.1) failed to remove, while other metazoan and fungal species contributed more than half of the eukaryotic fraction (0.3%, see Supplementary Table 22)

2.1.2 Prophage sequences

To assess the fraction of prophage sequences, we performed a BLAST search of all reads of our samples against the ACLAME mobile genetic elements database[77] as well as 2969 viral and 579 phage genomes from NCBI (from http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=5&name=Viruses and http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=6&name=Phages, respectively, as on 05 Feb 2010). On an average, 6.9% of the reads had a significant hit (more than 60 bits) to a sequence in these databases which is of the order of previous estimates of prophage sequences in bacterial genomes[78]. To estimate the lower bound of the prophage fraction, we estimated the number of these reads that had a significantly better hit to a viral sequence than to bacterial genome and found that at least 1.4% of our reads are of prophage origin (Supplementary Table 21).

2.2 Phylogenetic groups identified in gut metagenomes

We used the reference genome mapping procedure explained in Supplementary Methods Section 4.1 to estimate the phylogenetic abundance profile of each sample. Using an 85% DNA sequence similarity

threshold, which can reliably assign a sequence at the genus level (Supplementary Notes Section 1.2), we mapped the reads to the reference genome set and estimated the abundance of each genus in the reference set.

Firmicutes were identified as the most abundant phylum in our samples with a mean abundance of 38.8%, but they also show a high variability (standard deviation: 11.3%, abundance: 19.8 - 65.6%). The most abundant genus in this phylum is *Faecalibacterium* (mean abundance: 5.1%, standard deviation: 3.6%, 0.5%-15.0%). Unclassified Lachnospiraceae (mean abundance: 3.2%, standard deviation: 2.0%, 0.6%-9.5%) and *Roseburia* (mean abundance: 2.6%, standard deviation: 4.2%, 0.2 %-25.1%) are also abundant members of this phylum.

The second most abundant and most variable phylum is Bacteroidetes (mean abundance: 27.8%, standard deviation: 16.6%, abundance: 0.1%-64.9%). *Bacteroides* was the most abundant genus overall, but also showed the highest variability across all genera (mean abundance: 13.9%, standard deviation: 13.4%, 0.0%-54.7%). It also had the highest variation among cultivable genera in the HITChip analysis. *Prevotella* (mean abundance: 4.4%, standard deviation: 9.5%, 0.0%-35.5%) is also highly variable across all samples. *Alistipes* (mean abundance: 2.1%, standard deviation: 2.2%, 0.0%-9.1%) is another abundant genus from Bacteroidetes.

Actinobacteria represent the third most abundant phylum (mean abundance: 82%, standard deviation: 6.8%, abundance: 1.1%-32.5%). The most prominent Actinobacteria are *Bifidobacterium* (mean abundance: 4.5%, standard deviation: 4.7%, 0.0%-20.3%) and *Collinsella* (mean abundance: 1.8%, standard deviation: 2.2%, 0.0%-7.6%).

In addition to these phyla we also detect some lower abundant phyla such as Proteobacteria (mean abundance: 2.1%, standard deviation: 3.5%, abundance: 0.2 - 21.2%), Verrucomicrobia (mean abundance: 1.3%, standard deviation: 2.1%, abundance: 0.0 - 8.8%) and Euryarchaeota (mean abundance: 0.9%, standard deviation: 2.5%, abundance: 0.0 - 11.3%).

Faecalibacterium, *Bacteroides*, *Parabacteroides*, *Alistipes*, *Bifidobacterium* and *Collinsella* were among the five most abundant genera from their respective phyla in at least 32 out of 35 samples in agreement with array-based profiling in the subset of 22 European samples (using the HITChip[20]; Supplementary Table 4). This implies that species from these genera are predisposed and/or selected to be among the abundant species in the gut environment regardless of geographic location.

2.3 Functions identified in gut metagenomes

Some of the orthologous groups that the genes from fecal metagenomes are assigned to have no or only loose functional descriptions ('unknown' and 'general functions' account for 16.2% and 11%, respectively) comparable to other metagenomic samples from diverse habitats[79]. However, functionally uncharacterized genes usually form small OGs (Supplementary Figure 20) while large OGs with many genes are usually well-characterized[79] and their variation can be interpreted. The most frequent OG is formed by histidine kinases (COG0642), as reported previously for the Japanese dataset[6], which contributes on average 0.8% of all assigned genes in each sample, implying intensive signaling in this

community, for example triggered by environmental (nutritional or stress-related) compounds or in the context of specific quorum sensing communication[80].

The most variable OG in our gut metagenome samples is an ATPase (COG1132) component of ABC-type transporters (ranging from 0.3% to 1.1%, Fig. 1c). ABC type transport system is one of the most conserved molecular machines, which contributes not only for efflux but also for influx of compounds. These transporters participate in the persistence of bacteria in their ecological environment[81]. Their tremendous variety is also observed in the STRING database (Supplementary Figure 21), suggesting the contribution to the diversity of bacterial ability, such as drug resistance.

3 Highly abundant functions from low-abundance microbes

To identify functions that are predominantly from low-abundance microbes, we estimated the phylogenetic origin of each function, by combining phylogenetic assignment of reads to genera/phyla and functional annotation of genes to orthologous groups, and assigned orthologous groups to genera/phyla through the reads that constitute genes. We then looked for highly abundant functions (among the top 20% = above 80th percentile) that are primarily contributed by low-abundance genera (<2.5%), and found 122 such orthologous groups in all samples (Supplementary Figure 2 and Supplementary Table 6). Since we only chose functions that received more than 50% contribution from such genera, our observations will still be valid even if the unmapped portions of the genes are mapped to their rightful genera.

4 Robust clustering of samples across nations: Identification of enterotypes

4.1 Deriving enterotypes

Before identifying enterotype clusters in the samples, we removed the two American samples[4] because they had very low Bacteroidetes potentially due to a technical artifact[19]. The remaining 33 Sanger-based metagenomic samples were clustered using genus abundance profiles in Fig. 1b as explained in Supplementary Methods Section 7. The Calinski-Harabasz (CH) index showed a clear global maximum at 3 clusters in the Sanger dataset (Supplementary Figure 3a). When including the two American samples they formed a fourth cluster of their own (data not shown). The number of clusters in the other data sets was also 3 and thus confirmed our findings for the Sanger dataset (Supplementary Figure 3). Sanger- and pyrosequencing-based metagenomes from the same samples cluster together, reinforcing the feasibility of comparisons across sequencing platforms (Supplementary Figure 22).

4.2 Drivers of enterotypes in three different datasets

In the Sanger dataset, *Bacteroides*, *Prevotella* and *Ruminococcus* were the drivers of the three enterotypes. When we derived enterotypes from the 85 Illumina based metagenomes and the 154 16S rDNA datasets, we also identified three enterotypes where *Bacteroides* and *Prevotella* were still driving two of the three clusters. However, the third cluster was driven by different groups in these two

datasets: *Blautia* in the 16S rDNA dataset and unclassified Lachnospiraceae in the Illumina dataset. These three related groups fall under the Clostridiales order.

There is still uncertainty in the Clostridiales order of the phylogenetic tree of bacteria, especially in the Lachnospiraceae/Ruminococcaceae families, although microbiologists are working hard to resolve this issue. The placement of the reference genomes to this section of the tree (we use the tree from the Bergey manual) thus inherits these issues. Most of the strains with species identification in our reference genome database originally named by microbiologists as *Ruminococcus* have now been moved to Lachnospiraceae in the Bergey Manual (4 out of 6 genomes from named *Ruminococcus* species in our database). Furthermore, a number of genomes with almost full length 16S rRNA genes can only be classified at the family level as "unclassified Lachnospiraceae". Some of these are closely positioned with respect to *Blautia* in a phylogenetic tree based on 40 single copy marker genes (Supplementary Figure 23a). The uncertainty in the tree is further exemplified by *Ruminococcus lactaris* ATCC 29176:

1) a full length 16S rRNA gene sequence was classified as *Ruminococcus*
2) a partial 16S rRNA gene sequence was classified as Lachnospiraceae
3) the marker gene based tree clearly places it among Lachnospiraceae.

Given these uncertainties in this section of the tree we believe that that Enterotype 3 is driven by the same phylogenetic group but it is hard to ascertain who it is. In contrast, sections of the tree corresponding to *Bacteroides* and *Prevotella* show remarkable consistency (Supplementary Figure 23b) explaining why we do not face this issue in enterotypes 1 and 2.

The difference between 16S dataset and metagenomic datasets can also be explained by the different reference databases used for the 16S based and reference genome alignment based phylogenetic analysis procedures. For example, in the 154 samples from the 16S dataset, 38 genera have more than 0.1% abundance in more than 25% of the samples (≥39) as identified by the RDP Classifier. However, 13 out of these 38 genera do not have a representation in our reference genome database containing 1511 genomes, since there is no publicly available genome sequence for any member of these genera.

The difference between the Sanger-based metagenomes and the Illumina-based metagenomes can also be explained by the resolution at which phylogenetic compositions are measured. We used the SOAP aligner with the same parameters as in the original study[8], and this allows for only two differences between the Illumina read and the reference genome. This maps metagenomic reads that are closely related to the reference genome and hence is not at the same level of resolution as the BLAST-based mapping of Sanger-based metagenomes where we allow down to 85% sequence similarity to map at the genus level.

4.3 Robustness of enterotype clusters

To assess whether the samples are continuously distributed in the phylogenetic composition space, or whether they predominantly congregate around a few cluster centers, we simulated phylogenetic composition of hypothetical samples based on three models (namely random-uniform, template-uniform and template-Gaussian) that capture properties of real abundance data to different extents as

explained in Supplementary Methods Section 8. Random uniform simulations are expected to behave very differently than real data, while we expect the template-based uniform simulations to be more similar to the real data. Template-based Gaussian simulations should be very similar to the real data because they capture many characteristics of the real data except for non-normal distributions and interactions between features. We used the silhouette validation technique to compare real data with simulated data (Supplementary Methods Section 7.4). The real data had a higher silhouette width than 99% of the simulations for all datasets (Supplementary Figure 5). This implies that feature-feature interactions and non-normality of the distribution of features across samples contribute to the distinctiveness of clusters. We quantified the effect of removing samples from a cluster on the overall clustering behavior. Supplementary Figure 24 shows that the enterotypes generally stay intact (even when half of the samples in a cluster are removed, less than 6% of the samples are wrongly assigned).

We also quantified the effect of removing the major driver of each enterotype in forming clusters. We selectively removed the major driver before clustering the samples and recalculated the clusters. In two out of the three enterotypes, it is very clear that even after the main driver is removed, the network of co-occurring genera recovers the enterotype (see Supplementary Figure 25 and Supplementary Methods Section 7.5).

We also measured the effect of the number of samples on the optimal number of clusters by randomly sub-sampling half of each enterotype and estimating the optimal number of clusters using the CH-index. Comparison of Supplementary Figure 3 and Supplementary Figure 26 shows that more samples do not make the clustering less robust and thus the quality of the datasets determine the clustering quality and not the sample size.

4.4 Generalization and predictive power of enterotype clusters

The results show that accuracy, average precision and average precision gain are significantly higher for the classifier trained on real data than for any of the ones based on simulations for all datasets used in this study. Moreover, we observed that for the different simulations, classification improved with increasing similarity to real data -- “random uniform” simulations were most difficult to classify, while “template-based uniform” and “template-based Gaussian” were much easier to classify. Even though we employed a very simple classification algorithm, we obtained a leave-one-sample-out cross-validation accuracy of 90.1% (Sanger data set), 86% (16S data set) and 97.6% (Illumina data set). An average precision gain of 3.2, 4, and 5.26, respectively, demonstrates that the enterotype clustering indeed lends itself to building models that are much more predictive than is expected for a random clustering. As we analyzed the generalization behavior by means of cross-validation, the model trained on the actual samples is expected to classify future samples (generated similarly) with similar accuracy.

4.5 Independent experimental verification of enterotypes using HITChip

As this is a far reaching concept we further validated the enterotypes identified in our Sanger-based metagenomes by an independent experimental approach: we analyzed the 22 European samples using HITChip to validate the stratification of the enterotypes and the drivers.

A phylogenetic analysis of the DNA extracts of the 22 European samples was performed with the Human Intestinal Tract Chip (HITChip)[20]. This phylogenetic microarray has over 4,800 oligonucleotide probes, which target the 16S rRNA genes of more than 1,100 intestinal bacterial phylotypes. Hybridization and analysis were performed as described before[20] (brief summary in Supplementary Methods Section 4.4). Between class analysis on the genus profiles of these 22 samples using the enterotype classes from Fig. 2a shows the same three drivers: *Bacteroides*, *Prevotella* and *Ruminococcus* (see Supplementary Figure 4). This excludes possible confounding effects due to cloning bias of the *E. coli* host used.

5 Phylogenetic and functional variation between enterotypes

5.1 Phylogenetic composition of enterotypes

The abundance of genera in each enterotype is listed in detail in Supplementary Table 23.

5.1.1 Enterotype 1

Enterotype 1 is dominated by the genus *Bacteroides*, which represents between 21.3% and 54.3% of all genera found in the samples. The second most abundant genus is *Faecalibacterium* at abundance between 0.5% and 8.7%. *Bifidobacterium* makes up between 0.6% and 12.1%.

Unclassified Lachnospiraceae (1.5% - 5.0%) and *Parabacteroides* (0.2%-3.4%) also make up a reasonable portion of the samples from this enterotype. The assignment rate on genus level in this enterotype is between 50.0% and 71.0%.

5.1.2 Enterotype 2

Enterotype 2 is dominated by genera *Prevotella* (5.8%-35.9%) and *Bacteroides* (1.8%-15.3%). Other more abundant contributors are *Faecalibacterium* (1.6%-3.8%), unclassified Lachnospiraceae (0.8%-5.2%) and *Collinsella* (0.0%-4.9%). The assignment rate in this enterotype is between 21.0% and 60.2%.

5.1.3 Enterotype 3

In enterotype 3 *Bacteroides* (1.8%-17.6%) is the most abundant genus. *Bifidobacterium* (0.0%-20.3%) is the second most abundant genus. Other abundant genera in this enterotype are *Faecalibacterium* (0.6%-10.3%), unclassified Lachnospiraceae (0.6%-8.2), *Alistipes* (0.1%-6.3%), *Ruminococcus* (0.3%-7.1%), *Collinsella* (0.0%-7.5%) and *Akkermansia* (0.0%-8.9%). The assignment rate was between 20.1% and 50.1%.

6 Phylogenetic and functional biomarkers for host properties

While correcting for multiple testing, we lose a very large number of significant associations which is a direct result of the low number of samples. In addition, we only take associations that are exclusively linked to one metadata factor into account, to compensate for confounding variables. This conservative approach leaves us with only few correlations (and e.g. no association between IBD and an enterotype, due to the fact that there are only two IBD patients), but allows us to report likely true findings (within the given probability margins).

6.1 Age bias in the dataset

Enterotype 1 is enriched in Japanese individuals. The oldest Japanese subject is 45 years old and 20 out of the 22 European subjects are older than 45 years, which makes enterotype 1 younger than the rest of the dataset. Italian subjects are from a cohort of elderly individuals, so Italian subjects are older than the rest of the subjects.

6.2 Identification of an unknown Clostridiales genus correlating with host-age

When we classified our reference genome set of 1511 microbial genomes using the RDP Classifier that implements the recent reclassification of Firmicutes based on Bergey's Manual of Systematic Bacteriology[49], three genomes (from *Clostridium bolteae* ATCC BAA-613, *Clostridium asparagiforme* DSM 15981 and *Clostridiales bacterium* 1_7_47FAA) were classified as "Unclassified Clostridiales". We compared the almost full length (>1500bp long) 16S sequences of these three genomes and found that they were all more than 95% identical to each other. Therefore we conclude that the reads that are classified as "Unclassified Clostridiales" through their mappings (at >85% similarity) to these three genomes can be considered as a single genus that we call "unknown Clostridiales genus" in the manuscript. Abundance of this genus has a significant negative correlation with host-age ($p<0.02$). The increased chance for micro-aerobic regions in the gut with age could cause a decrease in their abundance[82-83].

6.3 Functions correlating with host-nationality

With regard to nationality, the abundance of 10 orthologous groups (mismatch repair ATPases, DNA methylases and DNA polymerases; Supplementary Table 14) varies more between than within nationalities (Supplementary Figure 12b) and several functionalities are specifically overrepresented in different ethnic groups (e.g., a polar amino acid transporter module in Japanese individuals; see Supplementary Table 15 and Supplementary Table 16), possibly linked to nutrition (e.g., the strong presence of glutamate in Japanese diet[84]). Despite those potential molecular markers, overall, the functional composition of the metagenomes of individuals from different nations was similar at the given sequencing depth. For example, the core metagenome (the set of functions present in all individuals) has a similar size in each nation, suggesting that the core functioning of the human intestine is similar in different ethnicities (Supplementary Figure 13). This also confirms the similarities of gut metagenomes from a large cohort of deeply sequenced Danish and Spanish individuals[8].

6.4 Verifying host-phenotypic classifications based on hydrogenotrophic microorganisms

With the current sequencing depth in our data set, additional phenotypic classification attempts such as those based on hydrogenotrophic microorganisms (methanogens, reductive acetogens or sulphate reducers) could not be verified using the functional marker approach, as the respective marker genes (e.g., coenzyme-M reductase mcrA, formyltetrahydrofolate synthetase, or dissimilatory sulphite reductase dsrA/dsrB) from these less abundant microbes could barely be identified. For example, mcrA

was only found in 3 out of the 22 European samples, although 30-50% of the western population are estimated to have dominant methanogenic bacteria in their feces[17]. Yet, the three distinct pathways for hydrogen disposal could trigger the three different enterotypes and indeed *Methanobrevibacter* (a methanogen) and *Desulfovibrio* (a known sulfate-reducer) are enriched in enterotypes 3 and 1, respectively. However, as the enterotypes seem to be driven by a complex mixture of functional properties, they could also have been shaped by hitherto unexplored physiological conditions such as transit time or pH of luminal contents.

6.5 Effect of genome size in functional over-representation in enterotypes

We checked the genome size distributions for 6 COGs which are over-represented in each enterotype (Supplementary Figure 27). These COGs have a relatively stable average genome size, suggesting that genome size does not have a major effect on functional abundance.

References

41 DiGuistini, S. *et al.* De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* **10**, R94 (2009).

42 Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006, doi:10.1101/gr.229102 (2002).

43 Kent, W. J. BLAT--The BLAST-Like Alignment Tool. *Genome Research* **12**, 656-664, doi:10.1101/gr.229202 (2002).

44 Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**, 177-189 (2002).

45 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**, 91-96 (2003).

46 Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-1115 (1998).

47 Huang, Y., Gilna, P. & Li, W. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**, 1338-1340, doi:10.1093/bioinformatics/btp161 (2009).

48 Garrity, G. M. *et al.* (Michigan State University Board of Trustees, MI, USA, 2007).

49 Ludwig, W., Schleifer, K.-H. & Whitman, W. B. in *Bergey's Manual of Systematic Bacteriology* Vol. 3 (Springer-Verlag, 2008).

50 Wilmotte, A. & Herdman, M. in *Bergey's Manual of Systematic Bacteriology* Vol. 1 487-493 (Springer-Verlag, New York, 2001).

51 Sorek, R. *et al.* Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. *Science* **318**, 1449-1452, doi:10.1126/science.1147112 (2007).

52 Lee, Z. M.-P., Bussema, C., III & Schmidt, T. M. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucl. Acids Res.* **37**, D489-493, doi:10.1093/nar/gkn689 (2009).

53 Goecks, J., Nekrutenko, A., Taylor, J. & Team, T. G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**, R86 (2010).

54 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967, doi:btp336 [pii]

10.1093/bioinformatics/btp336 (2009).
55 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).
56 Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).
57 Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-484 (2008).
58 Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. (Wiley, 1990).
59 Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *Information Theory, IEEE Transactions on* **49**, 1860 (2003).
60 Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences* **106**, 2677-2682, doi:10.1073/pnas.0813249106 (2009).
61 Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1 - 27 (1974).
62 Milligan, G. W. & Cooper, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159-179 (1985).
63 Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846-850 (1971).
64 Breiman, L., Friedman, J., Olshen, R. & Stone, C. *Classification and regression trees*. (Wadsworth and Brooks, 1984).
65 Dufour, A.-B. & Dray, S. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* **22**, 1-20 (2007).
66 Hayashi, H., Sakamoto, M., Kitahara, M. & Benno, Y. Molecular analysis of fecal microbiota in elderly individuals using 16S rDNA library and T-RFLP. *Microbiol Immunol* **47**, 557-570 (2003).
67 Saunier, K. & Dore, J. Gastrointestinal tract and the elderly: functional foods, gut microflora and healthy ageing. *Dig Liver Dis* **34 Suppl 2**, S19-24 (2002).
68 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376 (2005).
69 Human Genome Sequencing Center at Baylor College of Medicine. *Microbial Genome Projects at BCM HGSC*, <http://www.hgsc.bcm.tmc.edu/projects/microbial/microbial-index.xsp> (2009).
70 J. Craig Venter Institute. *JCVI: HMP / Human Microbiome Project*, <http://hmp.jcvi.org/status.shtml> (2009).
71 MetaHIT Consortium. *MetaHIT draft bacterial genomes at the Sanger Institute*, <http://www.sanger.ac.uk/pathogens/metahit/> (2009).
72 Microbial Sequencing Center at the Broad Institute of MIT and Harvard. <http://www.broadinstitute.org/seq/msc> (2009).
73 The Genome Center at Washington University. *Microbial Genomes*, <http://genome.wustl.edu/genomes/list/microbes> (2009).
74 Human Microbiome Project DACC. *Reference Genomes of the Human Microbiome Project*, <http://www.hmpdacc.org/reference_genomes.php> (2009).
75 National Center for Biotechnology Information. *Complete Microbial Genomes*, <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi> (2009).
76 Liolios, K., Mavromatis, K., Tavernarakis, N. & Kyrpides, N. C. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucl. Acids Res.* **36**, D475-479, doi:10.1093/nar/gkm884 (2008).

77 Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements. *Nucl. Acids Res.* **32**, D45-49, doi:10.1093/nar/gkh084 (2004).
78 Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nat Rev Micro* **3**, 504 (2005).
79 Harrington, E. D. *et al.* Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A* **104**, 13913-13918 (2007).
80 Kleerebezem, M., Quadri, L. E., Kuipers, O. P. & de Vos, W. M. Quorum sensing by peptide pheromones and two-component signal-transduction systems in Gram-positive bacteria. *Mol Microbiol* **24**, 895-904 (1997).
81 Davidson, A. L., Dassa, E., Orelle, C. & Chen, J. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol Mol Biol Rev* **72**, 317-364, table of contents (2008).
82 Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One* **5**, e10667, doi:10.1371/journal.pone.0010667 (2010).
83 Guigoz, Y., Doré, J. & Schiffrin, E. J. The inflammatory status of old age can be nurtured from the intestinal environment. *Current Opinion in Clinical Nutrition & Metabolic Care* **11**, 13-20, doi:10.1097/MCO.0b013e3282f2bfdf (2008).
84 Loliger, J. Function and Importance of Glutamate for Savory Foods. *J. Nutr.* **130**, 915- (2000).
85 Ciccarelli, F. D. *et al.* Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* **311**, 1283-1287, doi:10.1126/science.1123061 (2006).
86 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:btl446 [pii]
10.1093/bioinformatics/btl446 (2006).