Taxonomic classification method for metagenomics based on core protein families with Core-Kaiju

Anna Tovo[1], Peter Menzel[2], Anders Krogh[3], Marco Cosentino Lagomarsino[4,5], Samir Suweis[1*]

1 Physics and Astronomy Department, LIPh Lab, University of Padova, Padova, Italy

2 Labor Berlin Charité Vivantes GmbH, Berlin, Germany

3 University of Copenhagen, Nørregade 10 Postboks 2177, DK-1017, Copenhagen K

4 IFOM, FIRC Institute of Molecular Oncology, Via Adamello 16, 20143, Milan, Italy

5 Physics Department, University of Milan, and I.N.F.N., Via Celoria 16, 20133, Milan, Italy

* suweis@pd.infn.it

Abstract

An increasing number of studies recognizes the importance of characterizing species diversity and composition of bacteria hosted by biota for systems that range from oceans to humans. This task is typically addressed by using environmental sequencing data ("metagenomics"). However, determining microbiomes diversity implies the classification of species composition within the sampled community, which is often done via the assignment of individual reads to taxa by comparison to a reference database. Although computational methods aimed at identifying the microbe(s) taxa are available, it is well known that the inference of the community from the same sample using different methods can vary widely depending on the various biases in each step of the analysis. In this study, we compare different bioinformatics methods based on amplicon sequencing of 16S ribosomal RNA and whole genome shotgun sequencing for taxonomic classification. We apply the methods to three mock communities of bacteria, of which the composition is known. We show that 16S data reliably allow to detect the number

of species, but not the abundances, while standard methods based on shotgun data give a reliable estimate of the most abundant species, but predict a large number of false-positive species. We thus propose a novel approach, that combines shotgun data with a classification based on core protein families (PFAM), hence similar in spirit to 16S. We show that this method reliably predicts both number of species and abundance of the bacterial mock communities.

Author summary

Characterizing species diversity and composition of bacteria hosted by biota is revolutionizing our understanding and conception of the role of symbiotic interactions in ecosystems. However, determining microbiomes diversity implies the classification of species composition within the sampled community. Although many computational methods aimed at identifying the microbe(s) taxa are available, it is well known that the inference of the community from the same sample using different methods can vary widely depending on the various biases in each step of the analysis. In most of the studies, when benchmarking protocols for taxonomic classification from biological samples, the "ground truth" of the contained species and their relative abundances is not known. Therefore, the use of mock communities or simulated datasets remains as basis for a robust comparative evaluation of a methods prediction accuracy. In this work, we first compare different bioinformatics methods for taxonomic classification. We apply the methods to three mock communities of bacteria, of which the composition is known. We show that no method is able to correctly predict both the number of species and their abundances. We then propose a novel approach based on core protein families, reliably inferring both number of species and abundance of the bacterial mock communities.

Introduction

Modern high-throughput genome sequencing techniques revolutionized ecological studies of microbial communities at an unprecedented range of taxa and scales [1–5]. It is now possible to massively sequence genomic DNA directly from incredibly diverse environmental samples [3, 6] and gain novel insights about structure and metabolic functions of microbial communities.

One major biological question is the inference of the composition of a microbial community, that is, the relative abundances of the sampled organisms. In particular, the impact of microbial diversity and composition for the maintenance of human health is increasingly recognized [7–10]. Indeed, several studies suggest that the disruption of the normal microbial community structure, known as dysbiosis, is associated with diseases ranging from localized gastroenterologic disorders [11] to neurologic illnesses [12]. However, it is impossible to define dysbiosis without first establishing what "normal microbial community structure" means within the healthy human microbiome. To this purpose, the Human Microbiome Project has analysed the largest cohort and set of distinct, clinically relevant body habitats [13], characterizing the ecology of healthy human-associated microbial communities. However there are several critical aspects. The study of the structure, function and diversity of the human microbiome has revealed that even healthy individuals differ remarkably in the contained species and their abundances. Much of this diversity remains unexplained, although diet, environment, host genetics and early microbial exposure have all been implicated. Characterizing a microbial community implies the classification of species/family composition within the sampled community, which in turn requires the assignment of sequencing reads to taxa, usually by comparison to a reference database. Although computational methods aimed at identifying the microbe(s) taxa have an increasingly long history within bioinformatics [14–16], it is well known that inference based on 16S or shotgun sequencing vary widely [17]. Moreover, even if data are obtained via the same experimental protocol, the usage of different computational methods or algorithm variants may lead to different results in the taxonomic classification. The two main experimental approaches for analyzing the microbiomes are based on 16S ribosomal RNA (rRNA) gene amplicon sequencing and whole genome

shotgun sequencing (metagenomics).

Sequencing of amplicons from a region of the 16S ribosomal RNA gene is a common approach used to characterize microbiomes [18, 19] and many analysis tools are available (see Materials and methods section). Besides the biases in the experimental protocol, a major issue with 16S amplicon-sequencing is the variance of copy numbers of the 16S genes between different species. Therefore, abundances inferred by read counts of the amplicons should be properly corrected by taking into account the copy number of the different species detected in the sample [3, 20, 21]. However, the average number of 16s copies is only known for a restricted selection of bacterial species. As a consequence, different algorithms have been proposed to infer from data the copy number of those taxa for which this information is not available [18, 22].

In contrast, whole genome shotgun sequencing of all the DNA present in a sample can inform about both diversity and abundance as well as metabolic functions of the species in the community [23]. The accuracy of shotgun metagenomics species classification methods varies widely [24]. In particular, these methods can typically result in a large number of false positive predictions, depending on the used sequence comparison method and its parameters. For example in k-mer based methods, the choice of k determines sensitivity and precision of the classification, such that sensitivity increases and precision decreases with increasing values for k, and vice versa. As we will show, false positive predictions often need to be corrected heuristically by removing all species/taxa with abundance below a given arbitrary threshold (see Materials and Methods section for an overview on different algorithms of taxonomy classification).

Several studies have performed comparisons of taxa inferred from 16s amplicon and shotgun sequencing data, with samples ranging from humans to studies of water and soil. Logares and collaborators [25] studied communities of bacteria marine plankton and found that shotgun approaches had an advantage over amplicons, as they rendered more truthful community richness and evenness estimates by avoiding PCR biases, and provided additional functional information. Chan et al. [26] analyzed thermophilic bacteria in hot spring water and found that amplicon and shotgun sequencing allowed for comparable phylum detection, but shotgun sequencing failed to detect three phyla. In another study [27] 16S and shotgun methods were compared in classifying community bacteria sampled from freshwater. Taxonomic composition of each 16S rRNA gene

library was generally similar to its corresponding metagenome at the phylum level. At the genus level, however, there was a large amount of variation between the 16S rRNA sequences and the metagenomic contigs, which had a ten-fold resolution and sensitivity for genus diversity. More recently Jovel et al. [28] compared bacteria communities from different microbiomes (human, mice) and also from mock communities. They found that shotgun metagenomics offered a greater potential for identification of strains, which however still remained unsatisfactory. It also allowed increased taxonomic and functional resolution, as well as the discovery of new genomes and genes.

While shotgun metagenomics has certain advantages over amplicon-sequencing, its higher price point is still prohibitive for many applications. Therefore amplicon sequencing remains the go-to established cost-effective tool to the taxonomic composition of microbial communities. In fact, the usage of the 16S-gene as a universal marker throughout the entire bacterial kingdom made it easy to collect sequence information from a wide distribution of species, which is yet unmatched by whole genome databases. Several curated databases exist to date, with SILVA [29, 30], GreenGenes [31, 32] and Ribosomal Database Project [33] being the most prominent. Additionally, NCBI also provides a curated collection of 16S reference sequences in its Targeted Loci project (https://www.ncbi.nlm.nih.gov/refseq/targetedloci/).

When benchmarking protocols for taxonomic classification from real samples of complex microbiomes, the "ground truth" of the contained species and their relative abundances is not known (see [28]). Therefore, the use of mock communities or simulated datasets remains as basis for a robust comparative evaluation of a methods prediction accuracy. In this work we apply two widely used taxonomic classifiers for metagenomics, Kaiju [34] and MetaPhlAn2 [35, 36], and two common methods for analyzing 16S-amplicon sequencing data, Dada2 [37] and Qiime2 [38] to three mock communities of bacteria, of which we know the exact composition. We show that 16S data efficiently allow to detect the number of species, but not their abundances, while shotgun metagenomics give a reliable estimate of the most abundant species, but the nature of the two algorithms makes them predict a very large number of false-positive species. Therefore, we propose an updated version of Kajiu, which combines the power of shotgun metagenomics data with a more focused marker gene classification method, similar to 16S, but based on core protein domain families [39–42] from the PFAM

database [43].

Our criterion for choosing the set of marker domain families is that we uncover the existence of a set of core families that are typically at most present in one or very few copies per genome, but together cover uniquely all 8116 bacteria species in the PFAM database with an overall quite short sequence. Using presence of these core PFAMs (mostly related to ribosomal proteins) as a filter criterion allows for detecting the correct number of species/taxa in the sample. We tested our approach in a protocol called "Core-Kaiju" and show that it has a higher accuracy than other classification methods on the mock communities considered here, overcoming in particular the problem of false-positive species.

Materials and methods

Taxonomic Classification: amplicon versus whole genome sequencing

Many computational tools are available for the analysis of amplicon and shotgun sequencing data. The main difference between them stands on how to take into account the next-generation sequencing error rate per nucleotide in the association of each sampled 16s sequence to a species label. Indeed, errors along the nucleotide sequence could lead to an inaccurate taxon identification and, consequently, to misleading diversity statistics.

The traditional approach to overcome this problem is to cluster amplicon sequences into the so-called operational taxonomic units (OTUs), which are based on an arbitrary shared similarity threshold usually set up equal to 97% for classification at the genus level or 98-99% at the species level. Of course, in this way, these approaches lead to a reduction of the phylogenetic resolution, since gene sequences below the fixed threshold cannot be distinguished one from the other.

That is why, sometimes, it may be preferable to work with exact amplicon sequence variants (ASVs), i.e. sequences recovered from a high-throughput marker gene analysis after the removal of spurious sequences generated during PCR amplification and/or sequencing techniques. The next step in these approaches is to compare the filtered

sequences with reference libraries as those cited above. In this work, we chose to conduct the analyses with the following two open-source platforms: Dada2 [37] and Qiime2 [38]. Dada2 is an R-package optimized to process large datasets (10s of millions to billions of reads) of amplicon sequencing data with the aim at inferring the ASVs from one or more samples. Once the spourious 16s rRNA gene sequences have been recovered, Dada2 allow for the comparison with both SILVA, GreenGenes and RDP libraries. We performed the analyses for all the three possible choices. Qiime2 is another widely used bioinformatic platforms for the exploration and analysis of microbial data which allows, for the sequence quality control step, to choose between different methods. For our comparisons, we performed this step by using Deblur [44], a novel sub-operational-taxonomic-unit approach which exploites information on error profiles to recover error-free 16s sequences from samples.

As shown in [28], where different amplicon sequencing methods are tested on both simulated and real data and the results are compared to those obtained with metagenomic pipelines, the whole genome approach resulted to outperform the previous ones in terms of both number of identified strains, taxonomic and functional resolution and reliability on estimates of microbial relative abundance distribution in samples.

Similar comparisons have also been performed with analogous results in [25, 26, 35, 45] (see [17] for a comprehensive summary of studies comparing different sequencing approaches and bioinformatic platforms). Standard widespread taxonomic classification algorithms for metagenomics (e.g. Kraken [46]) extract all contained $k-$mers (all the possible strings of length k that are contained in the whole metagenome) from the sequencing reads and compare them with index of a genome database. However, the choice of the length of k highly influences the classification, since, when k is too large, it is easy not to found a correspondence in reference database, whereas if k is too small, reads may be wrongly classified. Recently, a novel approach has been proposed for the classification of shotgun data based on sequence comparison to a reference database comprising protein sequences, which are much more conserved with respect to nucleotide sequences [34]. Kaiju indexes the reference database using the Borrows-Wheeler-Transform (BWT), and translated sequencing reads are searched in the BWT using maximum exact matches, optionally allowing for a certain number of mismatches via a greedy heuristic. It has been shown [34] that Kaiju is able to classify

more reads in real metagenomes than nucleotide-based $k-$mers methods. Therefore, previous studies on the community composition and structure of microbial communities in the human can be actually very biased by previous metagenomic analysis that were missing up to 90% of the reconstructed species (i.e. most of the species they found were not present in the gene catalog). We therefore chose to work with Kaiju (with MEM option) for our taxonomic analysis. Although it resulted to give better estimates of sample biodiversity composition with respect to amplicon sequencing techniques, we found that it generally overestimates the number of genera actually present in our community (see Sec. Results) of two magnitude orders. To overcome this, we implemented a new release of the program, Core-Kaiju, which contains an additional preliminary step where reads sequences are firstly mapped against a newly protein reference library we created containing the amino-acid sequence of proteomes' core PFAMs (see following section). We also compared Kaiju 1.0 and Core-Kaiju results with those obtained via another widely used program for shotgun data analysis, MetaPhlAn2 [35, 36].

Characterization of the core PFAM families

After downloading the PFAM database (version 32.0), we selected only bacterial proteomes and we tabulated the data into a $F \times P$ matrix, where each column represented a different proteome and each row a different protein domain. In particular, our database consisted of $P = 8116$ bacterial proteomes and $F = 11286$ protein families. In each matrix entry (f, p), we inserted the number of times the f family recurred in proteins of the p proteome, $n_{f,p}$. By summing up over the p column, one can get the proteome length, i.e. the total number of families of which it is constituted, which we will denote with l_p. Similarly, if we sum up over the f row, we get the family abundance, i.e. the number of times the f family appears in the PFAM database, which we call a_f. Figure 1 shows the frequency histogram of the proteome sizes (left panel) and of the family abundances (right panel). Our primary goal was to find the so-called **core families** [47], i.e. the protein domains which are present in the overwhelming majority of the bacterium proteomes but occurring just few times in each of them [40, 48]. In order to analyze the occurrences of PFAM in proteomes, we converted

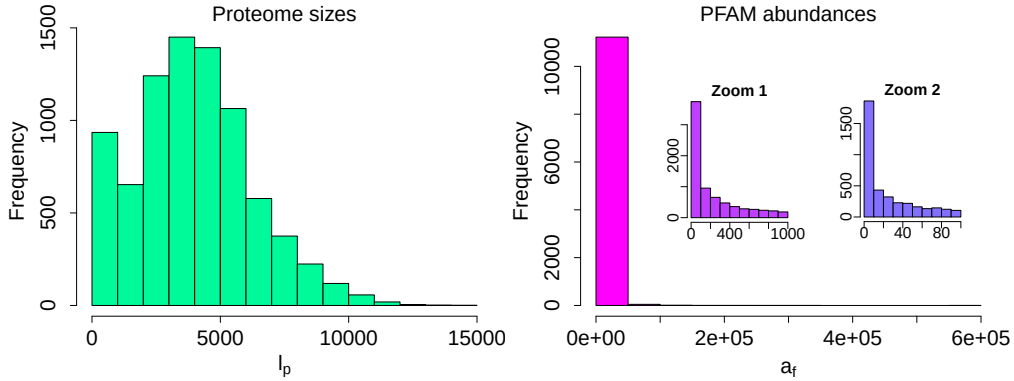

Fig 1. Proteome sizes and families abundances in PFAM database. On the left panel: frequency histogram of proteome lengths l_p (total number of families of which a proteome p is composed). On the right panel: frequency histogram of family abundances, a_f (number of times a PFAM f appear along a proteome).

the original $F \times P$ matrix into a binary one, giving information on whether each PFAM was present or not in each proteome. In the left panel of Figure 2 we inserted the histogram of the family occurrences, which displays the typical *u-shape*, already observed in literature [42, 49–51]: a huge number of families are present in only few proteomes (first pick in the histogram), whilst another smaller peak occurs at large values, meaning that there are also a percentage of domains occurring in almost all the proteomes. In the right panel, we show the plot of the number of rare PFAM (having abundance less or equal to four in each proteome) versus the percentage of proteomes in which they have been found. We thus selected the PFAMs found in more than 90% of

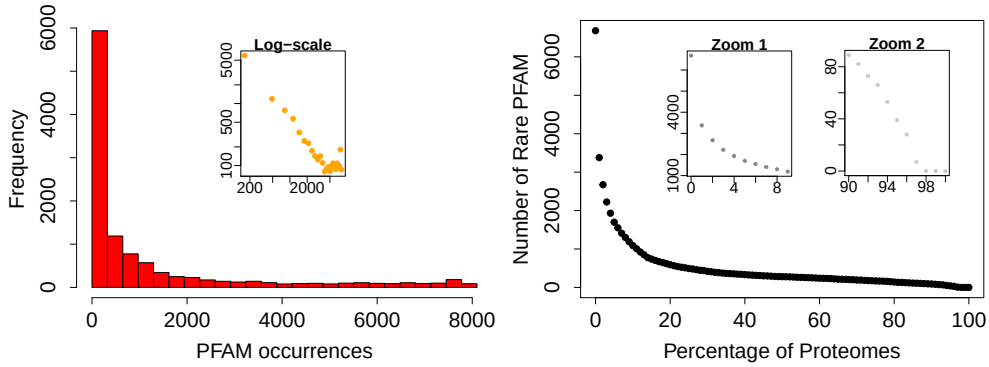

Fig 2. PFAM occurrences along proteomes. On the left panel: frequency histogram of family occurrences (number of proteomes in which a PFAM is contained). On the right panel: number of families with occurrence at most four versus the percentage of proteomes in which they are contained.

the proteomes and such that $\max_p n_{f,p} = 4$ (see Zoom 2 panel of Figure 2 and Table 1).

Table 1. Core PFAMs identity number and corresponding functions in proteomes.

PFAM ID	Function
PF00453	Ribosomal protein L20
PF00572	Ribosomal protein L13
PF01029	NusB family (involved in the regulation of rRNA biosynthesis by transcriptional antitermination)
PF01196	Ribosomal protein L17
PF01649	Ribosomal protein S20 (Bacterial ribosomal protein S20 interacts with 16S rRNA)
PF01795	MraW methylase family (SAM dependent methyltransferases)
PF03947	Ribosomal Proteins L2, C-terminal domain
PF08338	Domain of unknown function (DUF1731)
PF09285	EF-P (elongation factor P) translation factor required for efficient peptide bond synthesis on 70S ribosomes
PF17136	Ribosomal proteins 50S L24/mitochondrial 39S L24

Since we wish to have at least one representative core PFAM for each proteome in the database, we checked whether with these selected core families we could "cover" all bacteria. Unfortunately, none of them resulted to be present in proteomes 479430 and 1609106, corresponding to Actinospica robiniae DSM 44927 and Streptomyces sp. NRRL B-1568, respectively. We therefore added to our core PFAM list the most prevalent PFAMs present in such proteome, which resulted to be PFAM PF08338, occurring in 43% of the proteomes. In particular, we noticed that, if we selected, among our core protein domains, the PFAMs PF00453, PF00572, PF01029, PF01649, PF01795, PF03947, PF08338, PF09285 and PF17136, we were able to cover the whole list of proteomes of the databases. Therefore, in order to reduce the number of PFAM to work with, we restricted our following analyses to these ten domains that we consider the core PFAMs of bacteria proteomes.

Principal Coordinate Analysis

In order to explore whether the expression of the core PFAM protein domains are correlated with taxonomy, we did the following. First, we downloaded from the UniProt database the amino acid sequence of each PFAM along the different proteomes. In Figure 6, we show the frequency histogram of their sequence length, which resulted to be highly picked around specific values.

Second, for each family we computed the Damerau–Levenshtein (DL) distance

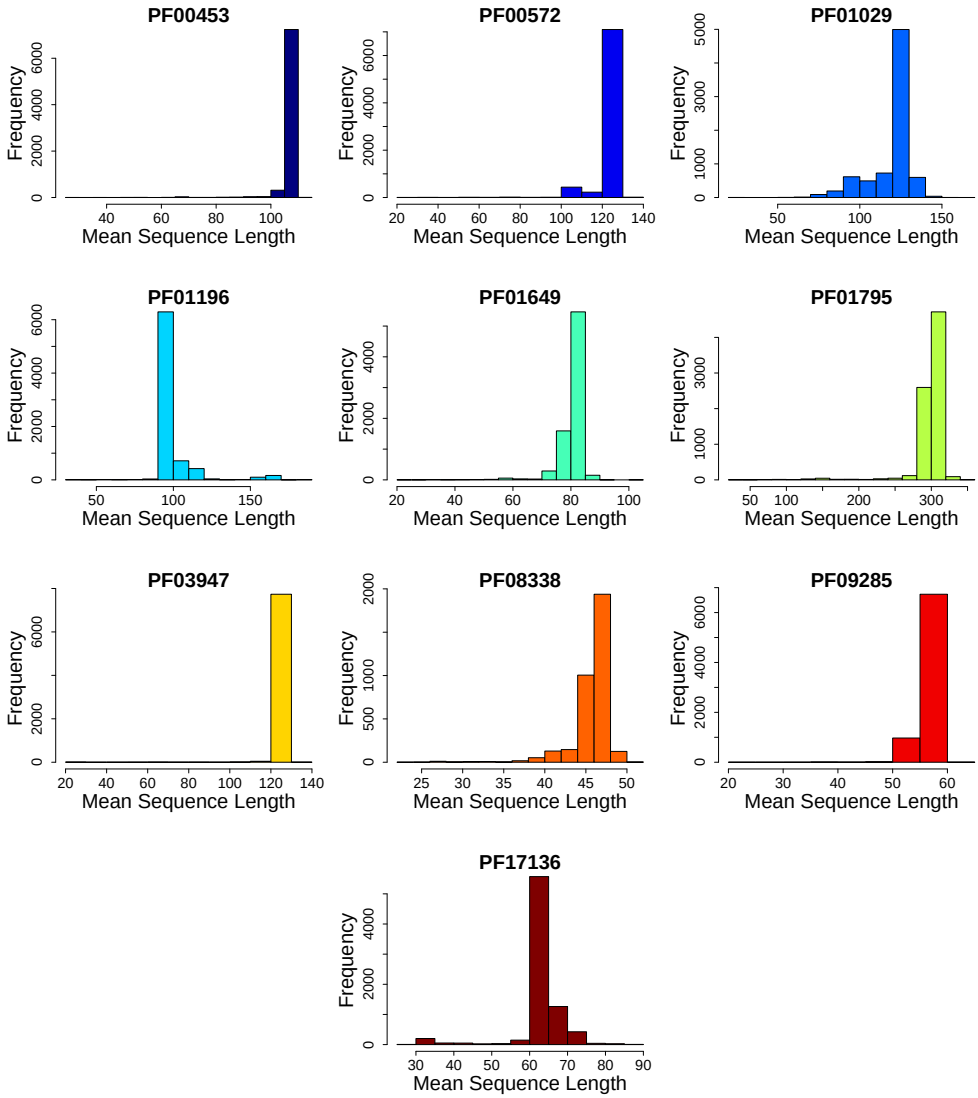

Fig 3. Frequency histogram of the core PFAMs sequence length. PF00453: mean=106.1, sd=5.2; PF00572: mean=122.8, sd=7.7; PF01029: mean=121.0, sd=13.3; PF01196: mean=100.4, sd=13.2; PF01649: mean=81.1, sd=4.8; PF01795: mean=296.8, sd=33.0; PF03947: mean=125.2, sd=6.0; PF08338: mean=46.0, sd=3.1; PF09285: mean=55.8, sd=1.5; PF17136: mean=63.6, sd=6.8.

between all its corresponding DNA sequences. DL measures the edit distance between two strings in terms of the minimum number of allowed operations needed to modify one string to match the other. Such operations include insertions, deletions/substitutions of single characters, transposition of two adjacent characters, which are common errors occurring during DNA polymerase. This analogy makes the DL distance a suitable metric for the variation between protein sequences. By simplicity and to have a more immediate insight, we conducted the analysis only for sequence points corresponding to the five most abundant phyla, i.e. Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes and Cyanobacteria.

After computing the DL distance matrices between all the amino-acid sequences of each PFAMs along proteomes, we performed the Multi Dimensional Scaling (MDS) or Principal Coordinate Analysis (PCoA) on the DL distance matrix. This step allow us to reduce the dimensionality of the space describing the distances between all pairs of core PFAMs of the different species and visualize it in a two dimensional space. In the last two columns of Table 2, we insert the percentage of the variance explained by the first two principal coordinates for the ten different core families, where the first one ranges from 3.3 to 12.1% and the second one from 2.4 to 7.7%. We then plotted the sequence points into the new principal coordinate space, colouring them by phyla. In general, we observed a two-case scenario. For some families as PF03883 (see Figure 4, left panel), Actinobacteria and Proteobacteria sequences are grouped in one or two highly visible clusters each, whereas the other three phyla do not form well distinguished structures, being their sequence points close one another, especially for Cyanobacteria and Firmicutes. For other families as PF01196 (see Figure 4, left panel), all five phyla result to be clustered, suggesting a higher correlation between taxonomy and amino-acid sequences (see Supporting Information, Figure S3, for the other core families graphics). These results suggest that some core-families (e.g. ribosomal ones) are phyla dependent, while other are not directly correlated with species taxa.

Mock Bacteria Communities

We tested shotgun versus 16S taxonomic pipelines on three artificial bacterial communities generated by Jovel et al. [28], whose raw data are publicly available at the

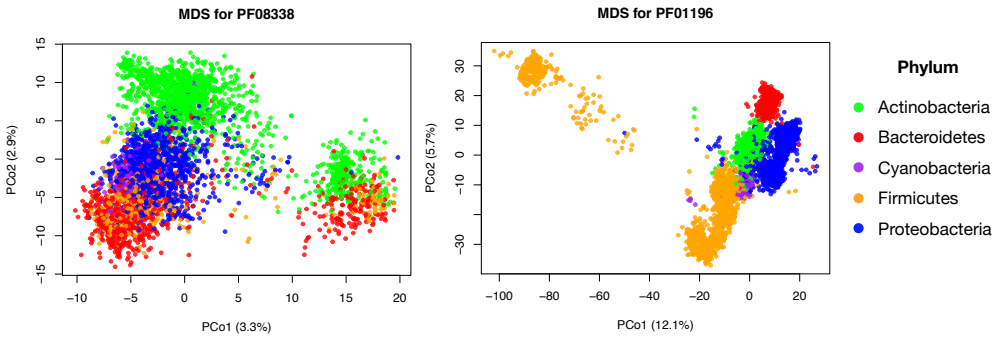

Fig 4. Phylum-based clustering for PF03883 and PF01196. For MDS analysis, only the sequences associated to the five most-abundant phyla (Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, Cyanobacteria) have been considered.

Table 2. Prevalence, Maximal/Total Occurences and Principal Coordinates of PFAM core families. We inserted, for each core family (PFAM ID, first column), the percentage of proteomes in which it appears (prevalence, second column), the maximum number of times it occurrs in one proteome (maximal occurrence, third column), the total number of times it is found among proteomes in the PFAM database (total occurrence, fourth column) and the percentage of variance explained by the firs two coordinates (PCo1 and PCo2, last two columns) when MDS is performed on sequences belonging to the five most abundant phyla (see Figure 4).

PFAM ID	Prevalence	Maximal Occurrence	Total Occurrence	PCo1	PCo2
PF00453	95%	3	7786	10.6%	6.6%
PF00572	97%	3	7897	5.4%	5.1%
PF01029	96%	4	12991	3.9%	2.4%
PF01196	97%	3	7888	12.1%	5.7%
PF01649	94%	3	7715	6.1%	4.6%
PF01795	96%	4	8113	5.2%	4.9%
PF03947	97%	4	7886	8.2%	7.7%
PF08338	43%	4	4267	3.3%	2.9%
PF09285	96%	4	8585	9.1%	4.9%
PF17136	97%	4	7896	5.4%	4.1%

Sequence Read Archive (SRA) portal of NCBI under accession number SRP059928. These mock populations contain DNA from eleven species belonging to seven genera: Salmonella enterica, Streptococcus pyogenes, Escherichia coli, Lactobacillus helveticus, Lactobacillus delbrueckii, Lactobacillus plantarum, Clostridium sordelli, Bacteroides thetaiotaomicron, Bacteroides vulgatus, Bifidobacterium breve, and Bifidobacterium animalis. For the taxonomic analysis at the genus level through 16S amplicon sequencing, we evaluated the performance of Dada2 [37] and QIIME2 pipelines [38]. In particular, as shown in [28], QIIME2 showed more reliable results in terms of relative

abundance of bacteria for all three mock communities when compared to Mothur [52] and the MiSeq Reporter v2.5 from Illumina.

As for shotgun libraries, we compared our Core-Kaiju protocol, to standard Kaiju [34] and to MetaPhlAn2 [35,36], which relies on unique clade-specific marker genes and which had been shown to have higher precision and speed over other programs [28].

Core-Kajiu

After defining the core PFAMs, we created two protein databases for Kaiju: The first database only contains the protein sequences from the core families, whereas the second database is the standard Kaiju database based on the bacterial subset of the NCBI NR database. The protocol then follows these steps (for the details, see Supporting Information):

1. Classify the reads with Kaiju using the database with the core protein domains
2. Classify the reads with Kaiju using the NR database to get the preliminary relative abundances for each genus
3. Discard from the list of genera detected in (2) those having absolute abundance of less than ten reads in the list obtained in point (1). This threshold represents our confidence level on the sequencing pipeline (see below).
4. Re-normalize the abundances of the genera obtained in point (3).

Results

Relative abundance vs absolute abundance thresholds

As stated in the introduction, metagenomic classification methods, such as Kaiju, often give a high number of false-positive predictions. In principle, one could set an arbitrary threshold on the detected relative abundances, for example 0.1% or 1%, to filter out low abundance taxa that are likely false positives. However, different choices of thresholds typically lead to very different results. The left panel of Fig 5 shows the empirical relative species abundance distribution of the 674 genera detected by Kaiju in our first mock community. Such biodiversity number would decrease to 34, 9 or 7 if one

considers only species accounting for more than 0.01%, 0.1% and 1% of the total number of sample reads, respectively. Moreover, looking at the empirical pattern, one can notice the main gap between genera of relative abundances of less than $5 \cdot 10^{-3}$ (black points) and those with relative population above 10^{-2} (green points), which correspond to the genera actually present in the artificial community. One could therefore guess that, whenever such a gap is detected in the RSA, this corresponds to the one between false-positive and truly present taxa. However, this is not always the case. We analysed a cohort of 26 healthy human fecal samples from the study [53] (metagenomic sequencing data are publicly available at the NCBI SRA under accession number SRP057027). We applied both Kaiju 1.0 and Core-Kaiju and found (when no threshold is set on either relative or absolute abundances) on average 2108 and 164 bacterial genera, respectively. The available amplicon-sequencing datasets from stool samples of healthy participants of the human microbiome project [1] suggest that there are on average 25 different bacterial genera found in each sample (based on 174 samples with at least >5k reads per sample using 97% OTU clustering [54]).

The right panel of Fig 5 shows the empirical RSA of one individual (sample ID: SRR2145359). As we can see, in this case the only apparent gap occurs between relative abundance of less than 10^{-1} and those above 0.5, with only one genus. It therefore results quite unrealistic that all the taxa but only should be considered false-positive. The same plot shows the vertical lines corresponding to threshold on relative population of 0.01%, 0.1% and 1% above which we have 97, 32 and 10 taxa, respectively. In contrast, by applying a threshold on the absolute abundance of reads on Core-Kaiju output, we end up with 25 genera (orange points in figure), which is compatible with the previous estimate.

Comparison between methods, using the same in silico test data-sets

After the preliminary exploratory analysis on the core protein domains, we tested Core-Kaiju pipeline in detecting the biodiversity composition of the three mock communities described above. In Figure 6 we plot the linear fit performed on predicted relative abundances via Core-Kaiju versus theoretical ones, known a priori. As we can

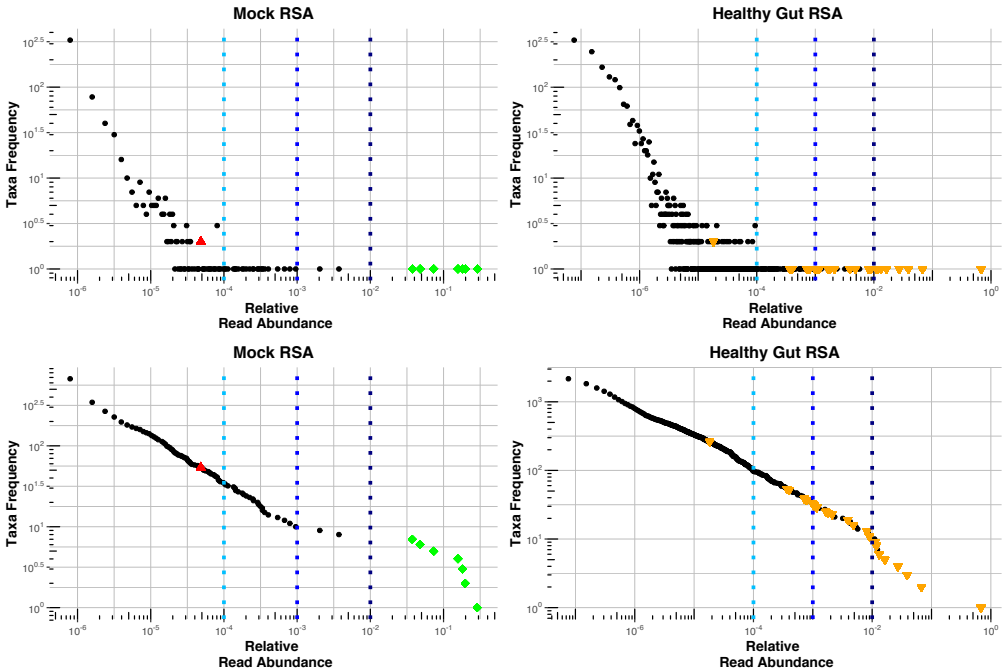

Fig 5. Relative vs absolute abundance thresholds for false-positive detection. On the left panels: Relative species abundance plots for the first mock community (see Materials and methods section). Green diamonds are the genera actually present in the artificial community and correctly detected by Core-Kaiju algorithm. The red triangle corresponds to the unique false-positive genus (*Moraxellaceae Acinetobacter*) detected with the newly proposed method. Dashed lines represent relative abundance thresholds on Kaiju 1.0 output of 0.01%, 0.1% and 1%, respectively, which would have led to a biodiversity estimate of 34, 9 and 7 genera, respectively. Imposing an absolute abundance threshold of ten reads on Kaiju 1.0 output directly, would instead lead to an estimation of 145 genera. On the right panels: the same analyses have been performed on a human gut sample of a healthy individual, where Kaiju 1.0 detects (without any threshold) 2165 genera. In this case the number (and label) of the actual present species is unknown. Nevertheless estimates from a reference cohort of stool microbiomes [54] from 174 healthy HMP participants (16S V3-V5 region, >5k reads per sample, 97% OTU clustering), the average number of genera per sample is 25 (max=46, min=9) [1]. Setting a threshold on the relative abundance of reads produced by Kaiju 1.0 gives a number of genera of 97 for the 0.01%, 32 for the 0.1% and 10 for the 1% threshold, respectively. In contrast, considering false-positive all genera with less or equal to ten reads in Kaiju 1.0 output, we end up with 864 genera. Orange triangles in plot correspond to the 25 genera detected with Core-Kaiju, a number compatible with the reported estimates. Bottom panels represent cumulative patterns.

see, in all three cases the predicted community composition was satisfactorily captured by our method, with an R^2 value higher than 0.75 (see Table 3). We also evaluated the performance of other shotgun and 16s pipelines for the taxonomic classification of the three mock communities. In particular, for 16s analysis, we applied Dada2 and Qiime2 with different reference libraries (GreenGenes [31, 32], the Ribosomal Database Project

(RDP) [33], and SILVA [29, 30]), whereas for shotgun analysis, we compared the obtained results with Kaiju 1.0 and MetaPhlAn 2. In Figure 7 we showed the true relative genus abundance composition of the three mock communities versus the ones predicted via the different taxonomic pipelines tested. R-squared values for the linear fit performed and number of genera predicted in the various cases are shown in Table 3. As we can see, both Core-Kaiju and MetaPhlAn 2 give a correct estimate of the number of genera in the communities (which is equal to seven), whereas all 16s methods slightly overestimate it and Kaiju 1.0 give a number which is around two orders of magnitude higher than the true one. Moreover, fit with Kaiju 1.0 and Core-Kaiju of the predicted abundances displayed a higher determination coefficient with respect to other pipelines. With Core-Kaiju, we were therefore able to produce a reliable estimate of both the number of species within the communities and their relative abundances.

Table 3. R-squared values and number of predicted genera. For all three analysed mock communities, we insert the R^2 value of the linear fit performed between estimated and true abundances together with the number of predicted genera, $\hat{G}$, with various taxonomic methods. The true number of genera is $G = 7$ for each of them.

		Mock 1		**Mock 2**		**Mock 3**	
		R^2	$\hat{G}$	R^2	$\hat{G}$	R^2	$\hat{G}$
Shotgun	**Core-Kaiju**	0.95	8	0.96	7	0.86	7
	Kaiju 1.0	0.97	674	0.98	501	0.94	738
	MetaPhlAn 2	0.46	7	0.60	7	0.08	7
16s	**Dada2 + SILVA**	0.59	18	0.73	22	0.41	13
	Dada2 + GG	0.45	17	0.60	21	0.35	12
	Dada2 + RDP	0.59	18	0.73	23	0.41	13
	Qiime2 + SILVA	0.50	41	0.59	41	0.43	41
	Qiime2 + GG	0.46	32	0.50	32	0.36	33

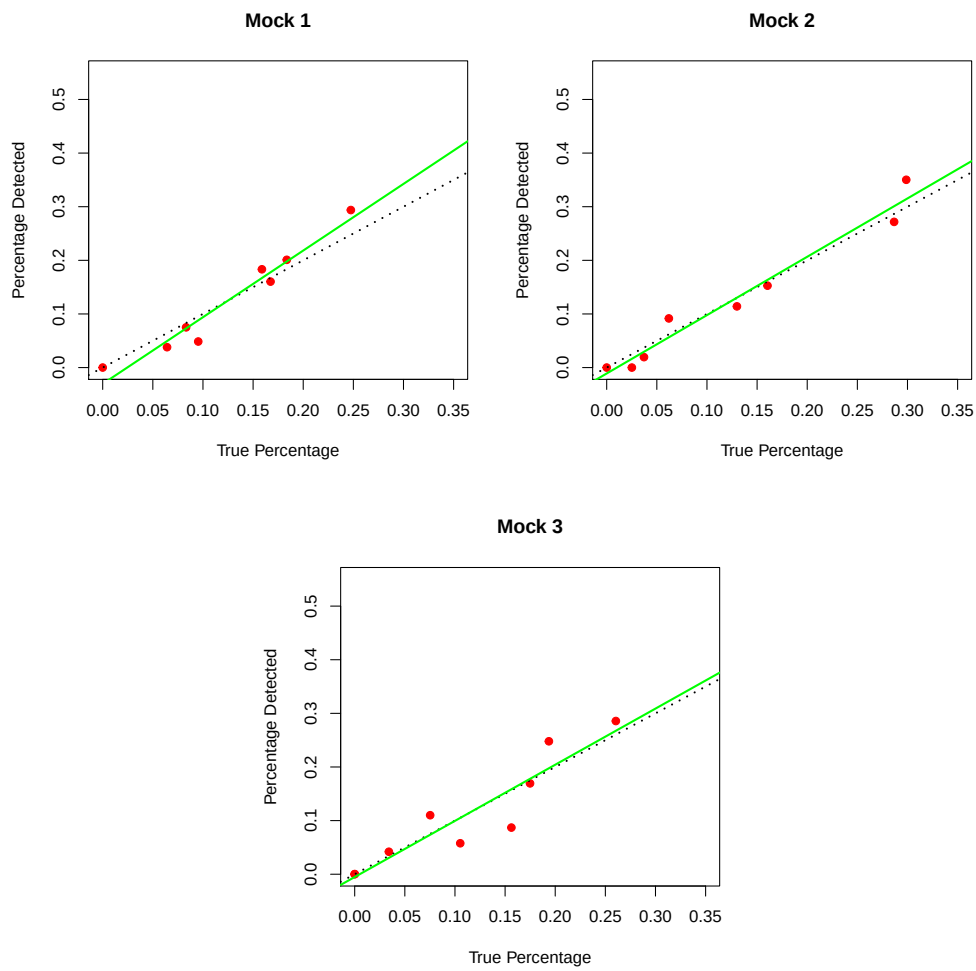

Fig 6. Linear fit between theoretical and predicted relative abundances. Red points represent data of the genus relative abundance predicted by Core-Kaiju on the three Mock communities versus the true ones, known a priori. The green line is the linear fit performed on obtained points which, in the best scenario, should coincide with the quadrant bisector (dotted black line). In all three cases the predicted community composition was satisfactorily captured by our method.

Discussion and Conclusions

An important source of errors in the performance of any algorithm working on shotgun data is the high level of plasticity of bacterial genomes, due to widespread horizontal transfer [40, 55–59]. Indeed, most highly abundant gene families are shared and exchanged across species, making them both a confounding factor and a computational burden for algorithms attempting to extract species presence and abundance information. Thus, while having access to the sequences from the whole metagenome is

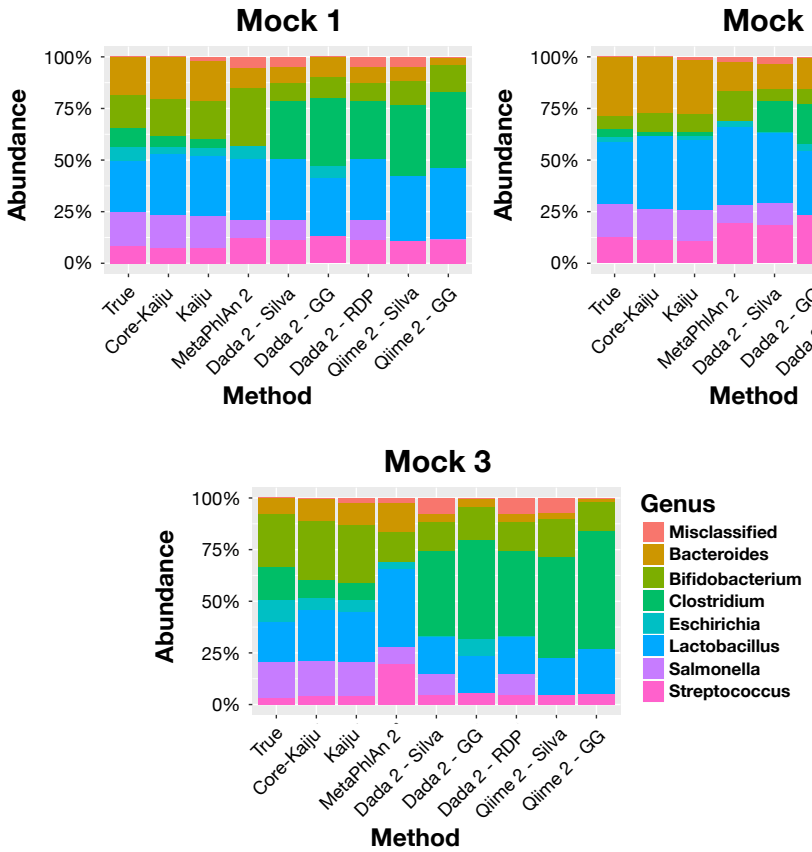

Fig 7. Predicted relative abundance composition of mock communities via different taxonomic classification methods.

very useful for functional characterization, restriction to a smaller set of families may be a very good idea when the goal is to identify the species and their abundance. Additionally, comparative genomics studies have shown that horizontal gene transfer (and plasticity) is gene-family dependent, and identified the degree that different families that contribute the least (and the most) to genome plasticity [40, 58]. As we have shown in this study, such knowledge can be exploited to restrict the set of sequences used to identify species.

To summarize, we have presented a novel method for the taxonomic classification of microbial communities which exploits the peculiar advantages of both whole-genome and 16s pipelines. Indeed, while the first approaches are recognised to better estimate the relative taxa composition of samples, the second are much more reliable in predicting the true biodiversity of a community, since the comparison between taxa-specific hyper-variable regions of bacterial 16s ribosomal gene and comprehensive

reference databases allow in general to avoid the phenomenon of false-positive taxa detection. Inspired by the role of 16S gene as a taxonomic fingerprint and by the knowledge that proteins are more conserved than DNA sequences, we proposed an updated version of Kaiju, an open-source program for the taxonomic classification of whole-genome high-throughput sequencing reads where sample metagenomic DNA sequences are firstly converted into amino-acid sequences and then compared to microbial protein reference databases. We identified a class of ten domains, here denoted by core PFAMs, which, analogously to 16s gene, on the one hand are present in the overwhelming majority of proteomes, therefore covering the whole domain of known bacteria, and which on the other hand occur just few times in each of them, thus allowing for the creation of a novel reference database where a fast research can be performed between sample reads and PFAMs amino-acid sequence. Tested against three in-silico generated mock communities, the proposed updated version of Kaiju, Core-Kaiju, outperformed popular 16s and shotgun methods for taxonomic classification in the estimation of both the total biodiversity and taxa relative abundance distribution. Our promising results pave the way for the application of the newly proposed pipeline in the field of microbiota–host interactions, a rich and open research field which has recently attracted the attention of the scientific world due to the hypothesised connection between human microbiome and healthy/disease [60, 61]. Having a trustable tool for the detection of both microbial species and their abundances could have a fundamental impact in our knowledge of human microbial communities and could therefore lay the foundations for the identification of the main ecological properties modulating the healthy or ill status of an individual, which, in turn, could be of great help in preventing and treating diseases on the basis of the observed patterns.

Supporting information

All codes, scripts and data used are freely available at the following links:

- http://kaiju.binf.ku.dk/
- https://github.com/liphlab/Kaiju-core

PFAM database. Proteins may be classified according to their function (metabolism, transcription, ...) or their structure (α-helix, β-sheets, ...). In this last case, it is possible to divide them into the so-called *protein domains* or *families*. In the **PFAM** database [43][1], one can find more than sixteen thousands protein families. Each PFAM file is organized as follows (see Figure S1). The first line gives information on

```
#Pfam-A regions from Pfam version 30.0 for ncbi taxid 253 'Chryseobacterium indologenes'
#Total number of proteins in proteome: 4334
#<seq id> <alignment start> <alignment end> <envelope start> <envelope end> <hmm acc> <hmm name> <type> <hmm start> <hmm end> <hmm length> <bit score> <E-value> <clan>
A0A0N1KUP3  55    313   50    317   PF16576    HlyD_D23    Domain      18    187   214   53.90 1.3e-11      CL0105
A0A0N1KUQ9  3     330   1     330   PF02277    DBI_PRT     Domain      9     329   329   394.80      2.9e-115    No_clan
A0A0N1KUS9  1     96    1     102   PF01149    Fapy_DNA_glyco    Domain      1     106   115   29.50 0.00093     No_clan
A0A0N1KUS9  116   186   114   201   PF06831    H2TH  Domain      4     73    92    42.00 6.6e-08     CL0303
Q5ICN6      17    49    17    49    PF00140    Sigma70_r1_2      Family      2     34    34    51.70 6.1e-11     No_clan
Q5ICN6      135   208   134   210   PF04539    Sigma70_r3  Family      3     76    78    67.00 1.2e-15     CL0123
Q5ICN6      54    123   54    124   PF04542    Sigma70_r2  Domain      1     70    71    61.90 3.6e-14     CL0123
Q5ICN6      222   275   222   275   PF04545    Sigma70_r4  Domain      1     50    50    63.20 1.1e-14     CL0123
```

Fig S1. Example of a PFAM file.

the PFAM database version we are using as well as the organism taxonomic name (*Chryseobacterium indologenes* in our example) and its NCBI identifier[2] (253), which equals the *.tsv* file label. In the second line we find the total number of proteins of which the proteome consists of (4334). Finally, the third lines gives information on the columns content of the file. Indeed, each of the following row corresponds to a different protein domains, for which different information are available. For example, the first column gives the protein code to which the PFAM belongs[3], the second and third column indicate the starting and ending position along the protein sequence of the domain, respectively, and in the sixth column the name of the protein domain. Note that protein domains are not specific of a particular protein, and one domain of the same family may be found in different proteins. In order to get the amino acid (each derived from the translation of three RNA nucleotides) sequence corresponding to a protein in the file one has to cross the information with the UniProtKB section of the

[1] https://pfam.xfam.org/

[2] https://www.ncbi.nlm.nih.gov/

[3] Unfortunately, for many of the proteins known to constitute the proteome, the information about their domains is not available, so that in the first column we may find a lower number of proteins than the one indicated at the second text line. In our example, for only 2936 of the 4334 proteins the corresponding families are indicated.

UniProt database [62]. There, for each protein, it is possible to download [4] a fasta file with its amino acid sequence. For example, for the Q5ICN6 protein in the PFAM data set, one obtains the fasta file shown in Figure S2, where the first two lines give information on the organism to which the protein belongs and the following lines give the complete amino acid sequence. Therefore, for example, if we are interested in the sequence of any protein family present in the PFAM file, it suffices to read the fasta file from the PFAM starting position to the ending one (in our example, for the PFAM PF04545, we have to read from the 222th amino acid to the 275th).

```
>tr|Q5ICN6|Q5ICN6_9FLAO RNA polymerase sigma factor RpoD/SigA
OS=Chryseobacterium indologenes OX=253 GN=rpoD PE=3 SV=1
MRQLKITKQVTNRETASLDKYLQEIGKVELITADEEVELAQRIRAGDRAALEKLIKANLR
FVVSVSKQYQNQGLSLPDLINEGNLGLMKAAKRYDETRGFKFISYAVWWIRQSILQALAE
QSRIVRLPLNKIGSINKINKAYAHLEQENERPPSPEELAEVLDMSEEDIKESMKNSGRHL
SMDAPLVEGEDSNLYDVLRSGESPSPDKDLMLESLQIEIERALNTLTPREADLVRLYFGL
NGKHPMTLEEIGETFDLTRERVRQIKEKAIKRLKHNTRSKILKSYLGK
```

Fig S2. Example of a UniProt protein sequence.

[4] https://www.uniprot.org/uniprot/

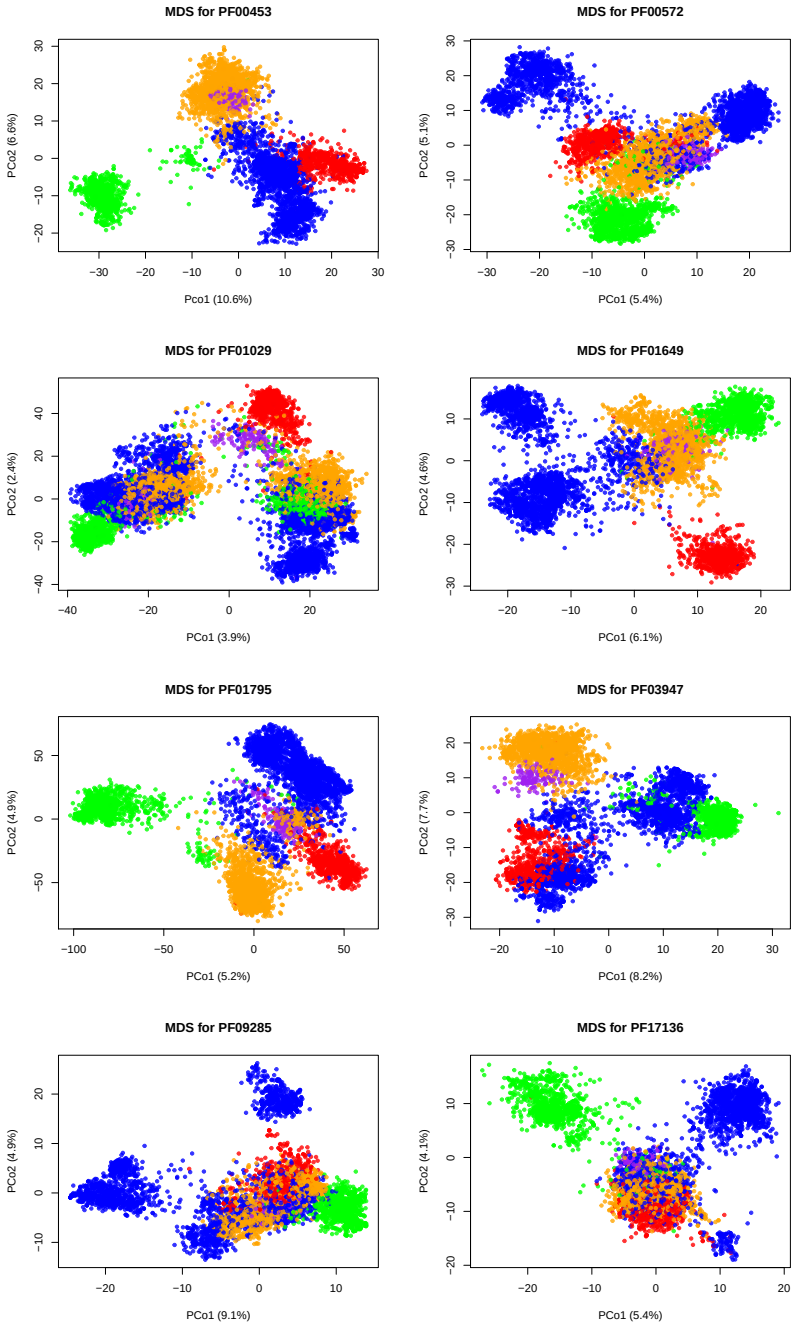

Fig S3. Phylum-based clustering for Core PFAMs. For MDS analysis, only the sequences associated to the five most-abundant phyla have been considered: Proteobacteria (blue), Firmicutes (yellow), Actinobacteria (green), Bacteroidetes (red), Cyanobacteria (purple).

Acknowledgments

SS and AT acknowledge funding from SID 2017 project "Quantitative Characterization of Biodiversity in Microbiomes: a Statistical Physics Approach" and the STARS 2018 grant "ReACT" from UNIPD. AK and MCL acknowledge Cariparo foundation Visiting Program.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804.

2. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, et al. The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS biology. 2012;10(8):e1001377.

3. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. Nature. 2017;551(7681):457.

4. Bork P, Bowler C, De Vargas C, Gorsky G, Karsenti E, Wincker P. Tara Oceans studies plankton at planetary scale; 2015.

5. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Scientific data. 2017;4:170093.

6. Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, et al. Emergent simplicity in microbial community assembly. Science. 2018;361(6401):469–474.

7. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, Relman DA. The application of ecological theory toward an understanding of the human microbiome. Science. 2012;336(6086):1255–1262.

8. Bashan A, Gibson TE, Friedman J, Carey VJ, Weiss ST, Hohmann EL, et al. Universality of human microbial dynamics. Nature. 2016;534(7606):259.

9. Gilbert JA, Lynch SV. Community ecology as a framework for human microbiome research. Nature medicine. 2019; p. 1.

10. Integrative H. The Integrative Human Microbiome Project. Nature. 2019;569(7758):641.

11. Lynch SV, Pedersen O. The human intestinal microbiome in health and disease. New England Journal of Medicine. 2016;375(24):2369–2379.

12. Wang Y, Kasper LH. The role of microbiome in central nervous system disorders. Brain, behavior, and immunity. 2014;38:1–12.

13. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. nature. 2012;486(7402):207.

14. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. Shotgun sequencing of the human genome; 1998.

15. Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome research. 2009;19(7):1141–1152.

16. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature communications. 2013;4:2304.

17. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Scientific reports. 2017;7(1):6589.

18. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology. 2013;31(9):814.

19. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and

next-generation sequencing-based diversity studies. Nucleic acids research. 2013;41(1):e1–e1.

20. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS computational biology. 2012;8(10):e1002743.

21. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 2017;551(7681):507.

22. Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome. 2014;2(1):11.

23. Hugenholtz P, Tyson GW. Microbiology: metagenomics. Nature. 2008;455(7212):481.

24. Peabody MA, Van Rossum T, Lo R, Brinkman FS. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC bioinformatics. 2015;16(1):362.

25. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, et al. Metagenomic 16S rDNA I llumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environmental microbiology. 2014;16(9):2659–2671.

26. Chan CS, Chan KG, Tay YL, Chua YH, Goh KM. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. Frontiers in microbiology. 2015;6:177.

27. Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PloS one. 2014;9(4):e93827.

28. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. Frontiers in microbiology. 2016;7:459.

29. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2013;41(D1):D590–D596.

30. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic acids research. 2013;42(D1):D643–D648.

31. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–5072.

32. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal. 2012;6(3):610.

33. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic acids research. 2013;42(D1):D633–D642.

34. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications. 2016;7:11257.

35. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods. 2012;9(8):811.

36. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature methods. 2015;12(10):902.

37. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581.

38. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology. 2019;37(8):852–857.

39. Grilli J, Bassetti B, Maslov S, Cosentino Lagomarsino M. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. Nucleic acids research. 2012;40:530–540. doi:10.1093/nar/gkr711.

40. Grilli J, Romano M, Bassetti F, Cosentino Lagomarsino M. Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers. Nucleic acids research. 2014;42:6850–6860. doi:10.1093/nar/gku378.

41. De Lazzari E, Grilli J, Maslov S, Cosentino Lagomarsino M. Family-specific scaling laws in bacterial genomes. Nucleic acids research. 2017;45:7615–7622. doi:10.1093/nar/gkx510.

42. Mazzolini A, Gherardi M, Caselle M, Lagomarsino MC, Osella M. Statistics of shared components in complex component systems. Physical Review X. 2018;8(2):021023.

43. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic acids research. 2019;47:D427–D432. doi:10.1093/nar/gky995.

44. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems. 2017;2(2):e00191–16.

45. Mitra S, Förster-Fromme K, Damms-Machado A, Scheurenbrand T, Biskup S, Huson DH, et al. Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. BMC genomics. 2013;14(5):S16.

46. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):R46.

47. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends in genetics : TIG. 2009;25:107–110. doi:10.1016/j.tig.2008.12.004.

48. Mazzolini A, Grilli J, De Lazzari E, Osella M, Lagomarsino MC, Gherardi M. Zipf and Heaps laws from dependency structures in component systems. Physical review E. 2018;98:012315. doi:10.1103/PhysRevE.98.012315.

49. Pang TY, Maslov S. Universal distribution of component frequencies in biological and technological systems. Proceedings of the National Academy of Sciences of the United States of America. 2013;110:6235–6239. doi:10.1073/pnas.1217795110.

50. Haegeman B, Weitz JS. A neutral theory of genome evolution and the frequency distribution of genes. BMC genomics. 2012;13:196. doi:10.1186/1471-2164-13-196.

51. Lobkovsky AE, Wolf YI, Koonin EV. Gene frequency distributions reject a neutral model of genome evolution. Genome biology and evolution. 2013;5:233–242. doi:10.1093/gbe/evt002.

52. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–7541.

53. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. Cell host & microbe. 2015;18(4):489–500.

54. NIH Human Microbiome Project;. Available from: "https://www.hmpdacc.org/hmp/HM16STR/".

55. Koonin EV, Wolf YI, Puigbo P. The phylogenetic forest and the quest for the elusive tree of life. Cold Spring Harbor symposia on quantitative biology. 2009;74:205–213. doi:10.1101/sqb.2009.74.006.

56. Puigbo P, Wolf YI, Koonin EV. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. Journal of biology. 2009;8:59. doi:10.1186/jbiol159.

57. Puigbo P, Wolf YI, Koonin EV. The tree and net components of prokaryote evolution. Genome biology and evolution. 2010;2:745–756. doi:10.1093/gbe/evq062.

58. Puigbo P, Wolf YI, Koonin EV. Genome-Wide Comparative Analysis of Phylogenetic Trees: The Prokaryotic Forest of Life. Methods in molecular biology (Clifton, NJ). 2019;1910:241–269. doi:10.1007/978-1-4939-9074-0_8.

59. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12:32.

60. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Current opinion in gastroenterology. 2015;31(1):69.

61. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. The evolution of the host microbiome as an ecosystem on a leash. Nature. 2017;548(7665):43–51.

62. Consortium U. UniProt: a hub for protein information. Nucleic acids research. 2014;43(D1):D204–D212.