Microbial abundance, activity and population genomic profiling with mOTUs2

To cite this version:
Alessio Milanese, Daniel Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, et al.. Microbial abundance, activity and population genomic profiling with mOTUs2. Nature Communications, Nature Publishing Group, 2019, 10 (1), 10.1038/s41467-019-08844-4 . hal-02281129

HAL Id: hal-02281129
https://hal.archives-ouvertes.fr/hal-02281129
Submitted on 7 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Metagenomic sequencing has greatly improved our ability to profile the composition of environmental and host-associated microbial communities. However, the dependency of most methods on reference genomes, which are currently unavailable for a substantial fraction of microbial species, introduces estimation biases. We present an updated and functionally extended tool based on universal (i.e., reference-independent), phylogenetic marker gene (MG)-based operational taxonomic units (mOTUs) enabling the profiling of >7700 microbial species. As more than 30% of them could not previously be quantified at this taxonomic resolution, relative abundance estimates based on mOTUs are more accurate compared to other methods. As a new feature, we show that mOTUs, which are based on essential housekeeping genes, are demonstrably well-suited for quantification of basal transcriptional activity of community members. Furthermore, single nucleotide variation profiles estimated using mOTUs reflect those from whole genomes, which allows for comparing microbial strain populations (e.g., across different human body sites).
Microorganisms live in complex communities of interacting species that impact life on earth and geochemical processes in the environment. It is thus of fundamental interest to accurately profile and compare the composition of the communities they form. The most common approach for microbial community profiling is by classification of PCR amplicon sequences from the small subunit ribosomal RNA gene (i.e., the 16S rRNA gene of bacteria and archaea). While powerful, this approach is also known to introduce biases in composition estimates due to, for instance, variations in 16S rRNA gene copy numbers per genome (Supplementary Figure 1), unequal efficiencies of PCR-primers in different species\(^1\).\(^2\) as well as the use of different sub-regions of this gene\(^3\). In addition, the high level of its sequence conservation limits the power for resolving closely related organisms\(^4\).

More recent methods sample environmental DNA directly by shotgun sequencing (metagenomics), which resolves some of these biases. Different strategies have been introduced to determine microbial community compositions from metagenomic data. One approach is based on classifying sequencing reads using publicly available and taxonomically annotated reference genome sequences of ‘known’ species. The resulting read abundance distributions require subsequent normalization by genome length\(^5\).\(^6\) to derive relative abundances of individual species (Supplementary Figure 1). Rather than using whole genomes, an alternative approach is to quantify read coverage of genes that are found to be clade-specific based on analyzing current reference genome databases\(^7\). If such marker genes occur only once per genome, then the resulting read coverages do not need to be normalized by copy number or genome length. However, a downside to any method depending on prior knowledge of genome sequences is that genomically uncharacterized taxa remain unaccounted for, which can lead to inaccurate relative abundance estimates at species-level resolution (Supplementary Figure 1).

Taxa that are missed by such reference-dependent methods can collectively be referred to as biological ‘dark matter’\(^8\). These include organisms—hereon referred to as ‘unknown’ species—that may be detected, but remain difficult to quantify using standard methods and up-to-date genome databases. To overcome this issue, we previously introduced a profiling tool that uses universally occurring, protein coding, single copy phylogenetic marker gene (MG)-based operational taxonomic units (mOTUs) as an approach to capture and quantify microbial taxa at species-level resolution in metagenomic samples\(^9\). mOTUs are built on the basis of MGs from both known and unknown species, the latter of which are extracted from existing metagenomes, enabling higher taxonomic resolution and more accurate quantification of species when profiling new microbial communities\(^9\).

Here, we present an updated and functionally extended profiling tool, the mOTU profiler version 2 (mOTUs2), which consolidates data from >3100 metagenomic samples into an updated mOTU database to substantially improve the representation of human-associated and ocean microbial species. Evaluations of mOTUs2 relative to state-of-the-art methods demonstrate improved sensitivity and quantification accuracy for both known and unknown species. We illustrate how species missed by other approaches can skew relative abundance estimates from compositional metagenomic data. Moreover, mOTUs enable quantifying baseline transcriptional activity of microbial community members from metatranscriptomic data, while avoiding quantification artefacts due to the use of non-housekeeping genes. Finally, heterogeneous populations of microbial strains have been reported in metagenomic studies to co-exist in a given microbial community, differ between individuals and environmental sampling sites, and remain stable over time\(^10\)–\(^12\). We show that differences between such strain populations can be estimated using the MGs of mOTUs as an efficient alternative to using whole genome sequences for metagenomic single-nucleotide-variation profiling.

**Results**

Reference-extended microbial community profiling with mOTUs2. We first identified 40 previously selected and benchmarked MGs in a total set of >25,000 sequenced genomes\(^13\). To obtain species-level taxonomic groups of (possibly redundant) sequences, we clustered these genomes based on a calibrated cutoff of 96.5% sequence identity\(^4\) into 5232 non-redundant, reference MG-based operational taxonomic units (ref-mOTUs) that contained more than half of a subset of ten MGs that were found suitable for metagenomic analyses\(^9\). Next, we assembled >3100 metagenomes from studies that included, as a requirement, a large number of systematically processed samples per site (Supplementary Data 1). These comprised 1210 samples from major human body sites (oral, skin, gut and vaginal\(^14\).\(^15\)), an additional 1693 samples from various human gut metagenomic studies including different disease cohorts\(^16\)–\(^21\) and 243 ocean water samples\(^22\). MGs predicted in these assemblies were clustered into marker gene clusters (MGCs). Finally, we devised an improved method for co-abundance-based binning of the MGCs into metagenomic mOTUs (meta-mOTUs) applying the same inclusion criterion (>5 MGs per mOTU) as for ref-mOTUs (Fig. 1a, Methods). To evaluate the binning accuracy of meta-mOTUs, we assessed individual MGCs in terms of taxonomic consistencies (Methods), variations in abundance, prevalence and GC-content of individual MGCs in comparison to ref-mOTUs (Supplementary Figure 2, Methods). Overall, we found high agreement in all categories. For example, at the species level, >97% (s.d.: ±1.5%) of meta-mOTUs are expected to be completely consistent in their taxonomic annotation (Supplementary Figure 2a), despite known incongruencies between species name assignments and MG-based sequence divergence\(^4\).

After quality control, the resulting 2494 meta-mOTUs, together with the 5232 ref-mOTUs, comprise the updated mOTU database. Compared to the previous version, these numbers correspond to a 3-fold and 7-fold increase in known and unknown species, respectively, that can now be profiled using mOTUs2. Taxonomic ranks for each mOTU were assigned by a last common ancestor-based consensus assignment (Supplementary Figure 3, Methods). Also, phylogenetic reconstruction shows that meta-mOTUs were sampled from a broad taxonomic distribution (Supplementary Figure 4), including from taxa that were hypothesized to represent novel phyla\(^23\). Across all included biomes (four major human body sites and the ocean), the number and fraction of unknown species (85%) were highest in ocean water samples (Fig. 1b), which is in congruence with previous results\(^22\). Notably, even in presumably well-explored human gut samples, we found that more than half of the species still lacked sequenced representatives in our reference genome database\(^13\) (Fig. 1b, c). A breakdown of mOTUs by biome showed that ref-mOTUs are often detected in multiple biomes, while meta-mOTUs tend to be more biome-specific (Supplementary Figures 5a, b). As shown by rank-abundance analyses, we find meta-mOTUs to be well distributed across the range from dominant to rare species (Supplementary Figure 6). Finally, the MGCs that could not be binned were used to quantify the cumulative abundance of organisms that are known to be present, but not quantified as mOTUs (Methods). This fraction was higher for the ocean than for samples from human body sites (Fig. 1c), which may be improved by increasing the number of profiled ocean metagenomes in the future.
We next evaluated the sensitivity of mOTUs2 for unknown species and assessed the resulting impact on relative abundance estimations compared to other approaches. To accomplish this, we analyzed the correspondence between mOTUs and metagenome-assembled genomes (MAGs). MAGs involve binning assembled metagenomic contigs by sequence composition and/or read abundance variation as a strategy to detect and genomically characterize organisms found in environmental samples. Thus, similar to meta-mOTUs, MAGs may include taxa that are not yet represented in genomic databases, and thus provide a way to test if and how many environmental microbes would be captured by mOTUs. More specifically, we reconstructed MAGs from 4880 published human gut metagenomes (Supplementary Data 2) and used 1845 MAGs identified in ocean water samples as a subset of 8000 recently published MAGs. Using these MAGs, we determined how many of them could be assigned to previously known (ref-mOTUs) vs. unknown species (meta-mOTUs) and evaluated the impact on relative abundance estimations. We found that >97% of MAGs from human gut samples were represented by mOTUs (Fig. 1d). Among these, 76% could be matched to ref-mOTUs and the remainder to meta-mOTUs. In addition, although the majority of the MAGs could be assigned to mOTUs, they represented only 42% of all human gut meta-mOTUs. For ocean water MAGs, 55% were represented by mOTUs (19% of these matching ref-mOTUs), while MAGs represented only 25% of ocean meta-mOTUs (Supplementary Figure 7). Our results indicate that the most abundant organisms in the human gut are already represented in public genome databases, whereas a substantial additional fraction becomes accessible through metagenomic data analysis. While assembly opens possibilities for many additional analyses, higher sequence coverage is required for the reconstruction of high-quality MAGs than for mOTUs, explaining why meta-mOTUs capture many more species. In the ocean, even some of the most abundant species still appear to lack representative genomic information (Supplementary Figure 7).

Next, we assessed the advantage of using a reference-independent method for species quantification in microbial communities. To this end, we compared mOTUs2 with two popular reference-dependent approaches, as well as its original version (mOTUs1), using: (i) simulated metagenomes from human gut-associated MAGs (Supplementary Figures 8, 9 and Methods), (ii) the Critical Assessment of Metagenome Interpretation (CAMI) dataset (Supplementary Figures 10, 11), and (iii) the simulated metagenomes used to evaluate MetaPhlan2 for benchmarking (Fig. 2; Supplementary Data 3, 4; Supplementary Table 1). Our results based on simulated MAGs show that in terms of precision, mOTUs2 and MetaPhlan2 outperformed mOTUs1 and Kraken (Fig. 2e). The fact that the reference-dependent methods MetaPhlan2 and Kraken can only detect genomes that are closely related to those present in current reference databases was well reflected in a reduced sensitivity, higher mean absolute error and deviations from expected taxonomic diversity estimates (Fig. 2e–g). Additional simulations showed that relative abundance estimates may be highly inaccurate when solely relying on reference genomes if unknown species are present in medium to high abundance (Supplementary Figures 9, 11). For the CAMI dataset, our results show that the mOTUs2 profiler outperformed many other tools (Fig. 2h–o; Supplementary Figures 10, 11). More specifically, mOTUs2 not only outperformed mOTUs1 at all taxonomic ranks, but also other tools, including MetaPhlan2 above the genus level for medium complexity simulations and above the species level for high complexity samples (Fig. 2k, o). We should note that in the CAMI benchmark (and the OPAL evaluation tool) profiled abundance data are re-normalized based on the detected taxa (see

Fig. 1 Construction of marker gene-based OTUs (mOTUs) for metagenomic profiling. a Schematic illustration of the mOTUs concept (Methods). b The observed richness of ref-mOTUs (containing exclusively MG sequences from reference genomes; blue) and meta-mOTUs (containing only MG sequences from metagenomes; green) per biome, and c mean cumulative relative abundance of species profiled across 2481 metagenomic samples. d Correspondence between mOTUs and 19,302 metagenome assembled genomes (MAGs) from the human gut. While less than 3% of MAGs are not represented (dark grey bar), mOTUs allow for profiling of 900 species not captured by MAGs. Source data are provided as a Source Data file.
**Fig. 2** Evaluation of mOTU profiling on simulated samples. Benchmarks of quantification accuracy (a–g) on ten simulated metagenomic samples (Methods) containing MAGs with \( n = 50 \) and MAGs without \( n = 50 \) a representative reference genome sequence, (h–o) and the CAMI challenge datasets\(^{25} \). a–d A representative simulated metagenome (out of ten; Supplementary Figures 8, 9) analysed with four profilers. e Precision-recall plot, where each data point corresponds to one of the ten simulated samples. Mean absolute error (MAE, also referred to as L1 norm) (f) and differences of the Shannon diversity index (g) from the expected values (error bars in f and g show standard deviation). h–j Average precision-recall values over the two medium complexity samples and (l–n) average precision-recall values over the five high complexity samples of the CAMI dataset (see also Supplementary Figure 10). Each precision-recall plot contains five values for mOTUs2, which correspond to different sets of parameters: high precision (\(-l\ 140 \ -g\ 6\), default (\(-l\ 100 \ -g\ 3\), high recall (\(-l\ 75 \ -g\ 3\), high recall (\(-l\ 50 \ -g\ 2\) and maximum recall (\(-l\ 30 \ -g\ 1\)), indicating the versatility of mOTUs2 in optimising precision or recall. In (k) and (o), mean absolute errors (MAE; referred to as L1-norm in CAMI) at different taxonomic ranks are shown for several tools. For mOTUs2, results for two options of calculating relative abundances are shown: one with relative abundances re-normalized based on detected taxa, which is enforced in the CAMI evaluation (but artificially deteriorates quantification accuracy), and one without this additional re-normalization (see main text and Supplementary Figure 11 for details). Data are provided in Supplementary Data 3, 4. Other taxonomic profilers (MetaPhyler, TIPP, Taxy-Pro, FOCUS, CLARK, Quickr) evaluated in CAMI\(^{25}\) are denoted by grey dots. Source data are provided as a Source Data file.
Supplementary Figure 11b). This re-normalisation procedure penalises tools, such as mOTUs2, that can account for the relative abundance of unknown taxa (Supplementary Figures 1, 11b). This feature leads to improved quantification (hence, a further reduction of the mean absolute error), in particular at the species level (Fig. 2k; Supplementary Figure 11a). Finally, since Kraken was not included in the CAMI benchmark25 dataset, we compared the performance of mOTUs2 to the results reported for the evaluation of MetaPhlan27, which included Kraken2. We find that mOTUs2 and MetaPhlan2 performed similarly, while both (and mOTUs1) outcompeted Kraken (Supplementary Table 1).

Given that profiling unknown species in addition to those represented in genome databases significantly improves relative abundance estimates, we sought to assess potential impacts on describing community structural properties. The total number of detected species and their relative abundance distribution determines the alpha diversity of a microbial community. This parameter is of fundamental interest in microbial ecology because it quantifies the breakdown of unknown species into mOTUs provides more accurate estimates of relative species abundances, measures for alpha diversity, such as the Shannon index (H'), were expected to be more accurate for mOTU-based profiles compared to reference-dependent approaches (based on simulations, Fig. 2g). To test this further using real microbial community data, we compared mOTUs2 to reference-dependent methods against 16S rRNA gene-based approaches. In two example data sets, one from a colorectal cancer study21 (n = 129) and one from an ocean ecosystem survey22 (n = 139), we found mOTUs2 profiles to have higher correlations with 16S rRNA gene-based estimates of alpha diversity (Spearman R = 0.71, P < 0.0001 and R = 0.78, P < 0.0001, respectively) than the reference-dependent methods (Fig. 3 and Supplementary Figure 12).

We also assessed the performance of methods at estimating how similar taxonomic compositions are between samples (beta diversity). For this, we used data from healthy individuals who donated samples from four different major body sites on multiple sampling occasions14, so that composition similarities could be compared within and between individuals. Given that compositional differences are expected to be smaller within than between individuals15, we tested in how many cases a sample from one subject would be most similar to another sample from the same individual (and body site) than from any other sample in the set of >1200 samples tested. As a result, we found that mOTUs2 performed similarly to the reference-dependent, clade-specific gene-based method2, while both outperformed the whole genome-based method used by Kraken6 (Supplementary Figure 13).

Unbiased metatranscriptomic profiling using marker genes. Although metagenomics data can be used for taxonomic profiling of microbial communities, it does not allow determining whether community members are physiologically active or not. Analogous to DNA for metagenomics, metatranscriptomics refers to the sequencing of reverse-transcribed RNA present in a microbial community. Depending on environmental conditions, the number of transcripts per cell varies for most genes. An exception to this are housekeeping genes that are expressed constitutively and with low variability under different conditions. Thus, the abundance of transcripts from such genes should strongly correlate with the abundance of active cells in a community. As all ten MGs are universal and involved in the highly conserved process of translating mRNA to proteins, we hypothesized that metatranscriptomic abundances would serve as particularly good proxies for relative cell abundances. To test this, we compared mOTUs2 to reference-dependent methods that have been used in recent metatranscriptomic studies28, 29 or analysis workflows30 relating metatranscriptomic profiles to microbial abundance and/or activity. More specifically, we correlated matching metagenome and metatranscriptome profiles from human stool samples31. At the species level (Fig. 4a, Supplementary Figure 14), mOTUs2-based correlations were considerably higher (median Spearman’s R = 0.76) than for reference-dependent methods (R = 0.37 and 0.45). Furthermore, we summarized mOTU abundances at the class level and computed all pairwise distances for all metagenomic and metatranscriptomic profiles to test for each metagenomic profile whether the most similar metatranscriptomic profile matched the same sample. For mOTUs2, this was the case for 92% of the samples compared to 78% and 64% for reference-dependent methods (Fig. 4b, Supplementary Figure 15).

MG-based SNV profiling for microbial population analyses. Originally, the ten MGs were identified as a subset of candidate phylogenetic marker genes deemed suitable for reconstructing the tree of life32 due to their universal occurrence and low rate of horizontal gene transfer33. These properties provided us with the opportunity to test how well single nucleotide variation (SNV) profiles of microbial populations could be recapitulated by the MGs comprising mOTUs as a compute-efficient alternative to using whole reference genome sequences. To this end, we generated metagenomic SNV profiles34 for sets of samples from different human body sites and ocean water using reference OTUs and representative genome sequences as reference databases. Despite some differences between biomes (Fig. 5a) and a few species, we found overall that the distances of SNV profiles using MGs were highly correlated (R > 0.8; Pearson) with those obtained using whole genomes. For example, we find almost perfect correlations for ocean microbial species (median R = 0.96), and for most gut microbial species (median R = 0.84) including those for which sub-species population structure was recently identified12, 15, 35 (Supplementary Figure 16).

Having established the possibility of resolving mOTUs below the species level, we addressed the question of how variable
microbial strains populations were over time in different human body sites. Previously, microbial strain populations were shown to display a high degree of individuality e.g., in the gut, skin, and oral sites. However, a comparative analysis of the degree of individuality of strain populations across different human body sites. Previously, microbial strain populations were shown to have a representative genome sequenced. Future efforts could aim at extracting MGs from high-quality MAGs and single amplified genome sequences to incorporate these into the mOTU database.

**Discussion**

The original development of the mOTU profiler was driven by the motivation to extend reference-dependent profiling of human gut microbial species to uncharacterized taxa. As more environments are subjected to metagenomic profiling, more data sets are becoming available that can be used for approaches based on binning genes by co-abundance analysis. With the inclusion of new microbiomes, we found that some human body sites are very well represented by available reference genomes (in particular skin and vagina). In contrast, more than 50% of gut microbial species still lack representative reference genomes (see also ref. 38), which may seem unexpected, but this estimate is in the same range as reported for an independent approach. 39 This may in part be due to methodological improvements in the binning of MGs into meta-mOTUs (Methods), increasing the number of potentially uncharacterized species that can be profiled. In addition, we included not only more samples, but also data from a number of disease-related studies (e.g., CRC, liver cirrhosis, type 2 diabetes) with large geographic distribution both contributing to an extended diversity of species that were not profiled previously. These may include species of particular relevance for differentiating healthy from diseased states. Furthermore, our results highlight the critical need to generate more reference genomes for the ocean environment where we find only 15% of species to have a representative genome sequenced. Future efforts could aim at extracting MGs from high-quality MAGs and single amplified genome sequences to incorporate these into the mOTU database.

Although metagenomics data can be used to profile the abundance of microbial taxa in a given community, they do not inform us as to whether they are also (transcriptionally) active. To discern genomic potential from activity, the combined use of metagenomics with metatranscriptomics profiling is becoming increasingly popular. Here, we found that metatranscriptomic abundances of mOTUs are highly correlated with metagenomic abundances, which highlights the property of MGs as constitutively expressed housekeeping genes across different conditions. This suggests that mOTUs should be useful for normalizing metatranscriptomics data for differential gene expression analyses. Other methods depending on genes that are conditionally or variably expressed are demonstrably less suitable for this approach and may also give the misleading impression that many taxa are rare, but highly active, or abundant, but inactive or dead (Supplementary Figure 15).
The computation of metagenomic SNV profiles to study microbial strain population differences is both resource- and time-consuming when using methods based on whole reference genome sequences.\(^4,\ 34\) We show that the use of mOTUs provides a fast and efficient alternative for profiling abundant species in microbial communities. In addition to the improved efficiency, mOTUs enable studying differences in strain populations of species that currently lack a representative genome sequence. This may be particularly relevant for disease-associated species and biomes for which only few reference genomes are available. A breakdown of intra-individual strain population similarity by species also allows for distinguishing those with high specificity, potentially under the control of the immune system, from those that only transiently populate their host. Promising applications of this approach could include testing the efficacy of strain-retention after faecal microbiota transplantation\(^10\) or studying of this approach could include testing the efficacy of strain-retention after faecal microbiota transplantation\(^10\). Promising applications of this approach could include testing the efficacy of strain-retention after faecal microbiota transplantation\(^10\).

**Methods**

The mOTUs2 profiler. The mOTU profiler version 2 (mOTUs2) is a stand-alone, open-source, computational tool that estimates the relative abundance of known as well as characterized microbial community genomes. The mOTU profiler was designed to be used as a pre-processing step for downstream analysis, and enables studying differences in strain populations of species at the genus level using metagenomic shotgun sequencing data. The taxonomy profiling method is based on ten universally occurring, protein-coding, single-copy phylogenetic marker genes (MGs), which were extracted from more than 25,000 reference genomes\(^1\) and more than 3100 metagenomic samples (Supplementary Data 1). The total ca. 367,000 non-redundant MG sequences). The MGs were grouped into \(\geq 7700\) MG-based operational taxonomic units (mOTUs) that represent microbial species, many of which (ca. 30%) still lack sequenced reference genomes. In addition to (i) taxonomy profiling, the tool allows for (ii) basal transcriptional activity profiling of community members using metatranscriptomic data as well as (iii) determining proxies for strain population genomic distances based on single-nucleotide variations (SNVs) within the phylogenetic markers that comprise mOTUs.

**Generation and annotation of the mOTUs database.** The mOTU2 profiler relies on a custom-built database of MG sequences extracted from reference genomes (ref-MGs) and from metagenomic samples (meta-MGs). The reference genomes were grouped into species-level clusters (specI clusters) and MG sequences from these reference genomes were grouped based on their species affiliation into reference marker gene clusters (ref-MGCs). These ref-MGs were augmented by meta-MGs and the remaining meta-MGs were clustered into meta-MG clusters. Meta-MGs of different MGs were subsequently grouped based on their species affiliation or binning based on co-abundance analysis into reference genome-based mOTUs (ref-mOTUs), or metagenomic mOTUs (meta-mOTUs). The resulting mOTUs were quality-controlled, compiled into a sequence database for short-read mapping and taxonomically annotated. Updates of the mOTU database will be made available at: http://motu-tool.org.

**Collection of MGs from reference genomes and metagenomes.** The 25,038 reference genomes used for the mOTU2 database were downloaded from the pro-Genomes database\(^11\). Metagenomic data were downloaded from the Genbank Sequence Read Archive (https://www.ncbi.nlm.nih.gov/oa) and the European Nucleotide Archive (https://www.ebi.ac.uk/ena) (accession numbers are listed in Supplementary Data 1). Most samples were obtained from human microbiome studies, including 1210 samples from different major human body sites (oral, skin, gut and vagina)\(^1\), and 1693 further samples from various human gut microbiomes\(^2\). We used 243 metagenomic samples from an ocean biome sample (50 samples, Pearson correlation of log-transformed relative abundance; ocean - prevalence filter: five samples, Pearson correlation of relative abundance; ocean oral cavity - prevalence filter: 30 samples, Pearson correlation of relative abundance; ocean anaerobic - prevalence filter: five samples from meta-mOtu). The MGs used in the mOTU2 pipeline were downloaded from the NCBI using HMM models built with HMMER3 (http://hmmer.org) applying a set of optimized cutoffs\(^9\), and extracts corresponding nucleotide sequences with the SeqOlq tool. With this workflow we extracted a set of 40 MGs (COG0012, COG0016, COG0018, COG0048, COG0049, COG0052, COG0080, COG0081, COG0085, COG0087, COG0088, COG0090, COG0091, COG0092, COG0093, COG0094, COG0096, COG0097, COG0099, COG0099, COG1000, COG1002, COG1003, COG1012, COG1014, COG1018, COG1019, COG1020, COG1021, COG1022, COG1025, COG0495, COG0522, COG0525, COG0533, COG0541, COG0552)\(^3\) from all 25,038 reference genomes. Not all of these genes are currently suitable for metagenomic applications due to high rates of ambiguous mapping of short reads owing to highly conserved regions within MG sequences as well as lower assembly rates observed for some MGs\(^9\). Hence, a selected subset of ten MGs (COG0012, COG0016, COG0018, COG0072, COG0025, COG00495, COG0052, COG00533, COG00541, COG00552) was extracted from genes that were predicted in metagenomes as described above.

**Grouping of MGs into ref-MGCs and meta-MGCs.** Reference genomes were processed and clustered into spec clusters to build ref-MGCs\(^4\). To this end, we calculated pairwise global nucleotide identities for all genome for each of the 40 MGs using vsearch (version v1.9.3)\(^15\). Genome-to-genome distances were calculated as the gene length-weighted arithmetic mean of the individual MG sequence length weighted by the resulting distance matrix was used as input for computing sequence similarity. This mOTU database will be made available at:http://motu-tool.org.

**Binning of MGs into mOTUs.** As the clustering of meta-MGs into meta-MG clusters was performed independently for each of the ten MGs, it resulted in unbinned meta-MGs (as opposed to the ref-MGs, which were grouped into mOTUs based on their spec cluster affiliation). In order to bin MGs into mOTUs (i.e., to link MGs originating from the same species), we utilized the property that (and therefore, MGs) from the same species are expected to co-vari in abundance across metagenomic samples\(^14\). Accordingly, we used the correlation between pairwise MG abundances across all samples for each biome. We optimized the correlation measure and prevalence filtering (as a means against the spurious correlation between low-abundance MGs\(^16\)), for each biome separately based on the AU-ROC determined by cross-validating the grouping of ref-MGs for which membership in the same spec cluster served as a ground truth. As a result, we defined the following biome-specific parameters: human gut - prevalence filter: five samples, Pearson correlation of log-transformed relative abundance; ocean - prevalence filter: five samples, Pearson correlation of relative abundance; human oral cavity - prevalence filter: 30 samples, Pearson correlation of relative abundance; human anaerobic - prevalence filter: five samples, mean correlation (meta-mOTUs), Spearman correlation of log-transformed relative abundance. In order to combine the biome-specific correlations, we transformed each of these into an FDR-calibrated association measure in such a way that for a given FDR value, the same association value was assigned. To obtain an association for each pair of MGs, we computed the maximum of the FDR-calibrated association values across biomes. For the actual binning, we used a slightly modified version of the greedy algorithm described in ref.\(^9\). As an initialization step, the ref-MGs were grouped according to their spec cluster affiliations. Then, meta-MGs were iteratively binned starting from the highest FDR-calibrated association values and decreasing until a cutoff value of 0.8 was reached. In this procedure, an MGC was added (binned) to an existing group (or another MGC to form a bin of size two) if this MG (among the ten possible ones) was not already present. Only groups with at least 6 members were retained as MGCs, whereas meta-MGs were designated as meta-MOTUs (consisting only of meta-MGAs) and 5232 ref-MOTUs (containing at least one ref-MG and possibly additional meta-MGAs). MGs that remained unbinned were grouped into a single unbinned group. Note that although spec clusters and ref-MOTUs are conceptually similar, there are two major differences: first, ref-MOTUs are composed of MGs of at least six out of the ten different MGs used for metagenomics, while spec clusters represent genomes that are grouped based on distances calculated from up to 40 MGs; second, ref-MOTUs can, as described above, contain MGs and MGs that were assembled from metagenomic samples. To assess the expected taxonomic consistency of the binning strategy of meta-MGAs, a fraction of the ref-MOTUs were treated in the same ways as meta-MGAs and their taxonomic affiliation (known from ref-MOTU membership) was only used afterwards to ascertain the error rate of the binning algorithm (Supplementary Figure 2a). Across all metagenomic samples used to construct the mOTUs, 1223 ref-MOTUs were detected and could be used for 100-fold resampled 5-fold cross-validation. We also assessed the agreement of the MGs for each mOTU in terms of relative abundance and prevalence across metagenomic samples (Supplementary
Figures 2b,c). Relative abundance and prevalence showed higher agreement for meta-mOTUs than for ref-mOTUs. This was expected since the binning algorithm is directly influenced by two parameters. We additionally evinced the homogeneity of GC content among the MG sequences within each mOTU (Supplementary Figure 2d). meta-MGCs showed very homogeneous GC content, as expected for genes that originate from the same genome, but not for erroneously binned MG sequences.

**Construction of the mOTUs2 mapping database.** We compiled a sequence database against which short metagenomic reads can be aligned to quantify the abundance of MGCs and mOTUs. To construct a non-redundant mOTUs mapping database, we removed identical MG sequences. MG sequences in the database were extended at the start and end of the gene by up to 100 nt, based on their genome or metagenomic assembly of origin, to reduce known mapping artifacts at gene boundaries. The resulting non-redundant database consists of the sequence files in FASTA format along with MGc and mOTU annotations, as well as the coordinates of the coding segments of the MG sequences. The sequence files were further indexed for searches with BWA.

**Taxonomic annotation of meta-mOTUs.** To assign taxonomic affiliations to meta-mOTUs, we first annotated each MG using Uniprot’s UniRef90 (https://www.uniprot.org/uniref, release 2017_08) as a reference protein sequence database.税务, release 2017_08) as a reference protein sequence database. The abundance of this group is calculated as the median of unbinned MGCs summed by COG.

**Description of taxonomic profiling outputs.** The mOTUs2 profiler returns multiple taxonomic profiles, since abundances based on read mappings can be calculated in different ways. One major distinction is the unit of counts. Either fragments such as inserts (or reads for single-pair sequencing) or mapped base-pairs can be counted. Counting the mapped base-pairs has the advantage that the mean base coverage can easily be computed by dividing the number of bases aligned to a certain gene by its corresponding length (mOTUs2 output -y option: “base.coverage”). Count based statistics are powerful for differential abundance testing (output -y option: “insert.raw_counts”). As the counts could in principle be non-integer numbers due to inserts mapping to multiple genes (see section 3.1), all counts are rounded to integers. For relative abundance-based estimates, gene-length normalizations are required to account for varying lengths of MG sequences and varying numbers of MGcs present in each mOTU. To this end, we previously introduced “scaled counts” that retains most of the characteristics of insert counts. In this approach, coverages are calculated as described above and are then normalized to sum up to the number of inserts that align to mOTUs (output -y option: “insert.scaled_counts”).

**Phylogenetic analysis of mOTUs.** To explore the phylogeny of mOTUs (ref-MGcs and meta-mOTUs), a reference tree was reconstructed by combining the phylogenetic signal of the ten sets of marker genes selected (Supplementary Figure 4). For this, all marker genes were translated into amino acid sequences and analyzed using ETE Toolkit v3.1.15. In particular, the program ete-build was used to run the following phylogenetic workflow: First, each set of marker proteins was independently aligned using ClustalOmega25. Next, alignment columns with less than three aligned residues were removed. Finally, the individual MG alignments were concatenated and used to infer a maximum likelihood phylogenetic tree using IQ TREE26 and the LG model.

**The mOTUs2 profiling workflow.** The mOTUs2 workflow for taxonomic profiling consists of three steps: alignment of metagenomic sequencing reads to MGcs, estimation of read abundances for every marker gene cluster (MGC), and calculation of mOTU abundances. As input, mOTUs2 expects the user to provide quality controlled sequencing reads. These are aligned to the MGcs of the mOTU database using BWA (mem algorithm, default parameters)22. The resulting alignments are filtered and only those with at least 97% nucleotide identity are retained. Further, alignments are filtered according to their lengths (default: 75 bp minimum alignment length; can be adjusted using the -1 option).

Next, we compute the best alignment(s) for every insert (read pair) to the MGcs using BWA alignment scores. Inserts with a single highest scoring alignment are flagged as “unique alignments”, whereas inserts with multiple highest scoring alignments are flagged as “multiple alignments”. Subsequently, abundances for each MGC are calculated by summing up the number of all inserts flagged as unique alignments resulting in a unique alignment profile. Inserts flagged as multiple alignments are distributed among their best-scoring MGcs in accord with their respective abundances estimated based on the unique alignment profile. Thus, the final abundances are calculated as the sum of the unique abundance profiles and the distributed contributions of the inserts flagged as multiple alignments. In addition to these MGc insert counts, MGC base coverages are calculated by first summing up the total number of bases aligning to each MGC and then dividing by the respective gene lengths. Finally, the abundances of the MGcs are calculated as the median of their respective MGC abundances (insert counts and base coverages). In order to reduce false positive results, we require a certain number of MGcs to be detected, that is to have metagenome-assembled reads mapping to them (default 10, -s option).

**Benchmarking mOTUs2 against other tools.** To evaluate its accuracy and robustness, we benchmarked mOTUs2 against two established tools for taxonomic profiling of metagenomic samples: MetaPhAn27, which is based on clade-specific marker genes, and Kraken50, which is based on exact alignments of genomic k-mers. MetaPhAN (version 2.6.0) was executed with default parameters. For Kraken-labelled analyses, we executed Kraken for read classification and called relative abundance tables using Kraken. Kraken and Bracken were installed as version 1.0.0 using conda. The Minikraken database (version mini-kraken_20171101_8GB_dustmasked) was downloaded from https://ccb.jhu.edu/software/bracken/. The Minikraken database was downloaded from https://ccb.jhu.edu/software/bracken/ on 1 February 2018. We executed kraken using paired-end reads and metadata data using default parameters. Abundance estimation with kraken was performed with the following parameters: -k minikraken_8GB_75mers_distrib. -t 1 -S o -result-abundance.bracken.
Comparison of mOTUs with metagenome-assembled genomes. We further validated the mOTUs using metagenome-assembled genomes (MAGs) reconstructed from Illumina-sequenced metagenomes. For this purpose, we first extracted 4880 metagenomic sequencing runs from human gut samples available from the European Nucleotide Archive (accession numbers are listed in Supplementary Data 2). Raw reads from each run were assembled using metaSPAdes v3.10.018 and subsequently binned with MetaBAT2 v2.12.119 with a minimum contig length threshold of 2000 bp. Sequencing coverage required for binning was inferred by mapping the raw reads back to the assemblies using BWA v0.7.1620 and then retrieving the percentage of mapped read bases with samtools v1.818 and the jgi_summarybam_contig_depths function from MetaBAT2. Quality scores (QS) of each contig were calculated using MetaBAT (v1.0.7)19, calculated as the level of completeness - 5 x contamination, as previously described19. Good-quality MAGs (QS > 50) were kept for subsequent downstream analyses. MAGs from marine samples (Ocean MAGs) were obtained as a subset of about 8000 MAGs, which are described in a recent publication23. In order to identify ocean-associated MAGs, we first searched for the keywords: ocean, marine, baltic sea and north sea to extract entries in Supplementary Table 1 and found 400 samples matching these keywords. From these samples, we selected 1845 MAGs (from Supplementary Data 2) that were reconstructed from these metagenomes.

Dependency between MAGs and mOTUs was established using the following procedure: first, we extracted the ten MGs from the MAGs using fetchMGs (see above), obtaining a set of MG-MAGs. Second, we aligned the MG-MAGs to the MG database of the mOTUs using vsearch -usearch_global (parameters: --id 0.96 –minqt 0.7). Finally, we evaluated the congruency of the MG-MAGs to mOTU matches. For this, we first checked if at least three MG-MAGs were assigned to a mOTU (by vsearch -id 0.97) resulting in a mOTU. If this was not the case the MAG was annotated as "unassigned". If not, we removed all alignments to MGCS not assigned to mOTUs and assigned a MAG to a mOTU if > 50% of the MG-MAGs were consistently matched to the same mOTU. Otherwise (if no majority mOTU was found) the MAG is annotated as "inconsistent".

Benchmarking mOTUs2 using simulated metagenomes. To be able to assess taxonomic quantification accuracy, ten human gut metagenomic samples were simulated using 15,102 Human gut MAGs: a subset of the 19,302 MAGs described before, excluding the MAGs created from samples used to construct the mOTU database (Supplementary Figure 8). MAGs with an ANI > 96.5% were de-replicated to have a reference MAG per species, of which according to the ANI the MAG was calculated with the fastANI tool21 [https://github.com/PatBLISS/FastANI]. The corresponding fastq files (as well as the simulated abundance data) are available at: http://moto-tool.org/download.html. Metagenomic read data were simulated using BEAR22: first, we generated 100 M insertions (2 x 100 M paired-end reads of 150 nt length) with 350 nt insert distance (standard deviation: 30) using generate_reads.py. Second, trim_reads.pl with default parameters was used to add the quality scores, introduce errors and shorten the reads. Every sample was simulated based on mOTU profiles relative abundances from ten real samples. For each simulated sample, we randomly selected 50 MAGs with a representative reference genome, as the supercontigs of the Kraken, MetaPhlAn or ref-mOTU databases. The OTU abundance tables were downsampled to the minimum number of reads per sample (CRC: 40,805 reads, TARA: 1494 reads) to normalize for uneven sequencing depths using the R function diversity of the vegan package. In order to obtain a 95% confidence interval we used bootstrapping (n = 100,000) by resampling pairs of Shannon index values. The confidence intervals reflect the 2.5 and 97.5 percentile of the bootstrapped samples.

Analysis of community structure. We assessed correlations of the Shannon index calculated based on 16S rRNA gene-based analyses and three metagenomic profiling tools (mOTUs2, MetaPhlAn2 and Kraken). For this we used data from two different biomes: metagenomes generated from stool samples of a colorectal cancer (CRC) study23 and metagenomes from seawater samples of the Tara Oceans expedition24. For the CRC study, amplicon sequencing data of the V4 region of the 16S rRNA were downloaded from the European Nucleotide Archive (ENA) database (http://www.ebi.ac.uk/ena/) accession number ER805534. For the ocean water samples, 16S rRNA gene containing fragments were extracted from metagenome sequencing reads (mTAs)24. To ensure comparability between the data sets, we first extracted the 100 bp from each miTAG sequence starting from the V4 primer sequence.

Ribosomal RNA data were initially processed using USEARCH25 (version 9.2.64) as follows: quality-filtered reads were merged and quality-filtered reads were used to create the fastq_mergepairs command with default settings. Merged reads were filtered using the fastq_filter command (fastq_maxee 0.1). Sequences were de-replicated using the fastx_uniques command, singlets were excluded and the remaining unique sequences were clustered into operational taxonomic units (OTUs) at 97% with chimera checking using the cluster_otus command. Finally, OTU abundances for each sample were determined using the usearch_global command (-strand both; -id 0.97). The OTU abundance tables were downsampling to the minimum number of reads per sample (CRC: 40,805 reads, TARA: 1494 reads) to normalize for uneven sequencing depths using the R function diversity of the vegan package. In order to obtain a 95% confidence interval we used bootstrapping (n = 100,000) by resampling pairs of Shannon index values. The confidence intervals reflect the 2.5 and 97.5 percentile of the bootstrapped samples.

Analysis of metatranscriptomes. To demonstrate the use of mOTUs2 to assess basal transcriptional activity of microbial community members, we used a dataset from 36 samples for which metagenomic and metatranscriptomic sequencing data are available26. Each sample (36 metagenomes and 36 metatranscriptomes) was subjected to profiling using mOTUs2, Kraken/Bracken and MetaPhlAn2. All differentially expressed genes were transformed to relative abundances and log-transformed using the R function diversity of the vegan package. The Euclidean distance was computed as the Euclidean distance of the log-transformed relative abundances after the addition of a pseudocount smaller than the smallest non-zero value. For each of the three distances and each sample, we identified the most similar sample (i.e. the one with the minimum distance value) and determined the proportion of cases in which both samples belonged to the same individual. More specifically, for each body site, we compared community compositional distances between samples from the same individual (intra-individual) to distances between this and other individuals (inter-individual). Canberra and Bray-Curtis distances were computed with the vegan package and the function vegdist using the Euclidean distance as the distance measure. Euclidean distance was computed as the Euclidean distance of the log-transformed relative abundances after the addition of a pseudocount smaller than the smallest non-zero value. For each of the three distances and each sample, we identified the most similar sample (i.e. the one with the minimum distance value) and determined the proportion of cases in which both samples belonged to the same individual.
and in addition to a set of 5306 reference genomes. Genomic distances of strain populations between samples were estimated based on SNV profiles computed both on mOTUs and the whole genomes using the motus snv_call command. The filtering parameters used within the snv_call command were adapted to the specificity of datasets and references. The allele frequency tables were filtered using a horizontal coverage (-fb) equal to 40% for whole genome-mapped reads and 80% for mOTU-mapped reads, a vertical coverage (-fd) of 10, a per position coverage (-fc) of 5 and a position prevalence (-fp) of 0.90. The minimum number of samples per reference (-fm) was 20 for the human samples and 5 for the Ocean samples. Whole-genome-based distances were compared to those from mOTUs using Pearson’s correlation (Fig. 5a). We selected the ref-mOTUs/genomes that passed the filtering thresholds for both methods and correlated between sample distances between the two methods (n.b. there were no species from the vaginal superspecies passing the filtering requirements for both methods).

**Individuality of microbial populations across body sites.** We tested for the individuality of microbial strain populations on the subset of the human microbiome samples described above (5.4.1), for which at least two time point data were available. For each body site, we compared SNV profiles distances (see Supplementary Figure 17b) to distances from the same individual (intra-individual, intra-body-site distances) to distances between this and other individuals (inter-individual, intra-body-site distances). To determine whether intra-individual distances were smaller than inter-individual distances (see Supplementary Figure 17b)—indicating individuality of strain populations—we used ROC analysis. ROC curves (see Supplementary Figure 17a) ascertain how accurately small distances predict whether a pair of samples originated from the same individual (with similarly small inter-individual distances being considered false positives) when systematically varying the distance cutoff. ROC curves can be summarized by the area under the curve (AU-ROC) with higher values corresponding to clearer separation between intra- and inter-individual distances (Fig. 5b and Supplementary Figure 17a). Confidence intervals on the AU-ROC (Fig. 5b) were obtained by bootstrapping using the pROC package.

**Code availability.** The mOTU profiler version 2 and additional information are available at: https://motu-tool.org. Its source code is accessible at: https://github.com/motu-tool/mOTUs_v2.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**
To generate the mOTU database, we used reference genome sequence data from the proGenomes database (http://progenomes.embl.de) as well as metagenomic sequence data from the Genbank Sequence Read Archive (https://www.ncbi.nlm.nih.gov/era) and the EMBL European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) with access numbers listed in Supplementary Data 1. Human gut metagenomic data and metagenome assembled genomes are available at the ENA (accession numbers are listed in Supplementary Data 2; MAGs can be downloaded from: http://ftp.ebi.ac.uk/pub/databases/metagenomics/mags-gut_19.tar.gz). The 10 human gut metagenomic samples simulated from metagenome-assembled genomes are available on Zenodo (https://doi.org/10.5281/zenodo.1473645). All other relevant data is available upon request.

Received: 7 June 2018 Accepted: 2 February 2019
Published online: 04 March 2019

**References**


Acknowledgements

We thank members of the IT core facility and Y. Yuan for managing computing resources at EMBL, and the IT support team of the Institute of Microbiology and the HPC team at ETH Zürich for managing computing resources and the Euler/Leonhard clusters of ETH Zürich. This work was supported by EMBL, ETH Zürich [PHRT-521], the Helmut Horten Foundation, and the Novartis Foundation for Medical-Biological Research [17B077], the BMBF through the German Network for Bioinformatics Infrastructure (de.NBI) [031A537B], and the European Molecular Biology Organization [ALTF 721-2015], LTFCOFUND2013 [PCOFUND-GA-2013-609409].

Author contributions


Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-08844-4.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019