Extended taxonomic profiling of diverse environments with mOTUs 2.6

Hans-Joachim Ruscheweyh[1], Alessio Milanese[1,2], Lucas Paoli[1], Quentin Clayssen[1], Daniel Mende[3], Georg Zeller[2*], Shinichi Sunagawa[1,*]

Affiliations:

[1] Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland

[2] Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[3] Department of Medical Microbiology, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

* Correspondence to: ssunagawa@ethz.ch, zeller@embl.de

Abstract

Summary:

Identifying species and estimating their relative abundance in metagenomic samples is a crucial task in microbiome research. However, for many environments, variable fractions of microbial species are not represented in reference genome databases, and thus, remain often unaccounted for in taxonomic composition analyses. Here, we present an updated version of the metagenomic profiling tool mOTUs. We extended its database of taxonomic marker genes by including information from ~600,000 prokaryotic draft genomes, >96% of which are metagenome assembled genomes from 23 environments. This extension enables researchers to profile a previously underrepresented diversity of microbes not only in some of the most intensively studied environments such as the human, soil and ocean microbiomes, but also in freshwater systems, the air, and the gastrointestinal tract of mice, cattle and other animals. This update doubles the number of detectable prokaryotic species to 33,570 and includes the release of 11,164 taxonomic profiles as a community resource.

Availability and implementation

mOTUs 2.6 is implemented in Python and licensed under GPLv3. Source code, profiles and documentation are available at: https://motu-tool.org/.

Supplementary information:

Supplementary information is available online.

Introduction:

Metagenomic sequencing has transformed the field of microbiology by enabling researchers to study the genomic composition of microbial communities in the environment (microbiomes) in a cultivation-independent manner. This advance has driven the development of computational methods to profile microbiome samples, *i.e.* to identify species and to quantify their relative abundance (Sczyrba *et al.*, 2017). Such methods often rely on the availability of reference genomes (Wood and Salzberg, 2014; Shah *et al.*, 2021), the generation of which usually requires successful cultivation. However, for many environments, the majority of microbial species is still uncultivated (Lloyd *et al.*, 2018; Konstantinidis et al., 2015). As a result, a significant proportion of microbial communities, especially at high taxonomic resolution, may be missed during profiling. Adding to this limitation, microbial community composition is typically measured as the relative abundance of its members, and hence, missing taxa can cause biased estimation of the abundance of the others (Milanese *et al.*, 2019).

To address these issues, we developed the taxonomic profiling tool, mOTUs, which leverages universal, protein coding, single-copy phylogenetic marker genes (MGs) that can not only be identified in reference genomes, but also retrieved from metagenomes (Milanese *et al.*, 2019; Sunagawa *et al.*, 2013). These MGs are clustered into species-level MG-based taxonomic units, referred to as 'ref-mOTUs' and 'meta-mOTUs' if they are represented by MGs from at least one reference genome or only from metagenomes, respectively. Recent algorithmic advances (e.g. Kang *et al.*, 2019) have made it possible to reconstruct metagenome-assembled genomes (MAGs) in studies of diverse host-associated and environmental microbiomes (Stewart *et al.*, 2019; Anantharaman *et al.*, 2016; Delmont *et al.*, 2018). As for reference genomes, MGs can be

extracted from MAGs and be readily used to extend the database of the mOTUs tool (Milanese *et al.*, 2019). This feature enables the mOTUs profiler to provide a more complete view of the microbial diversity across different environments and at the same time, to reduce bias in relative abundance estimation.

Here we present an updated version, mOTUs 2.6, with a database largely extended by *de novo*-generated MAGs, which, combined with the addition of publicly available draft genomes, cover the most intensely studied microbiomes to date. This greatly increases the scope of metagenomic profiling to environments that are of broad interest, but currently lack cultivation-based reference genome collections.

Results & Discussion

To increase the phylogenetic coverage of the mOTUs database for microbial species without any genome-sequenced representative available, we generated 150,880 MAGs from 4,531 publicly available metagenomes originating from 23 environments (Supplementary Table 1). These MAGs were complemented with 454,773 draft genomes (~96% MAGs; ~4% isolate and single cell genomes) gathered from previous work (Supplementary Table 1). After extracting MGs from these draft genomes, we extended the previous database (2.5.1) using existing functionality described at https://motu-tool.org/. A total of 79,833 of these genomes were not represented by any mOTU (neither meta-mOTU nor ref-mOTU) in version 2.5.1 and were thus used to extend the database of MGs by 19,358 new mOTUs ('ext-mOTUs', Figure 1a). Among these, more than half (10,401) were assigned to unknown families (Supplementary Table 4) based on taxonomic

annotations using the STAG classifier (v0.7, https://github.com/zellerlab/stag). MGs from assembled metagenomes, which were neither part of a MAG nor previously represented by a mOTU, were added to an unbinned group (shown as 'unassigned' in mOTUs abundance profiles). As a unique feature of mOTUs, quantifying the abundance of this group allows for estimating the relative abundance of unknown species, and thus improves the accuracy of relative taxonomic abundance estimates for all detected organisms (Milanese et al. 2019).

The updated database, version 2.6, increases the number of mOTUs by a factor of 2.3 (from 14,212 to 33,570). Notably, the proportion of newly represented species varies widely across different environments (Figure 1b). For example, in the well-studied mouse gut microbiome, less than 25% of the 5,345 MAGs were found to correspond to ref-mOTUs. The remaining 75% of the MAGs resulted in 587 ext-mOTUs (Supplementary Figure 2). Even for the human microbiome, 2,750 ext-mOTUs were generated, although 97% of the ~400,000 newly added MAGs had already been represented by ref- or meta-mOTUs. This finding illustrates that while the vast majority of human-associated microbial species are repeatedly sampled, novel species are still being recovered, possibly by the inclusion of underrepresented populations, dietary habits and/or disease states (Pasolli *et al.*, 2019; Wirbel *et al.*, 2019). The remaining 16,021 ext-mOTUs cover a large range of other animal-associated (e.g., cattle, fish, chicken, pig, bee, dog, cat) and environmental (e.g., soil, freshwater, wastewater, ocean, air) microbiomes for which reference genomes are generally still scarce (Figure 1b). The total number of species represented in the mOTUs 2.6 database varies across different environments from hundreds to several thousands of which 5 to 40% (Figure 1b) are typically detected in a single profiled metagenome. To facilitate comparative research, we profiled a total of 11,164 metagenomes and

metatranscriptomes from 23 environments using the mOTUs 2.6 database. These profiles are released alongside the updated mOTUs database as a basis for comparisons with newly generated ones.

Conclusions:

By including assembled metagenomes from previously underrepresented environments, the updated mOTUs tool (version 2.6) enables a greatly extended profiling of the microbial diversity on earth, a large fraction of which is missed if their detection and quantification relies exclusively on sequenced reference genomes.

Funding:

We acknowledge funding from the Swiss National Foundation through project grants 205321_184955, the NCCR Microbiomes (51NF40_180575) and the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (#2018-521) to S.S.. G.Z. was supported by the European Molecular Biology Laboratory and the German Federal Ministry of Education and Research (BMBF, the de.NBI network, grant no. 031A537B, and grant no. 031L0181A).

Conflict of Interest:

none declared

Acknowledgements:

The reconstruction of metagenome assembled genomes was performed on the EULER cluster managed by the HPC team at ETH Zurich.

References

Anantharaman,K. *et al.* (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.*, **7**, 1–11.

Delmont,T.O. *et al.* (2018) Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, **3**, 804–813.

Kang,D.D. *et al.* (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.

Konstantinidis,K.T. and Rosselló-Móra,R. (2015) Classifying the uncultivated microbial majority: A place for metagenomic data in the Candidatus proposal. *Syst. Appl. Microbiol.*, **38**, 223–230.

Lloyd,K.G. *et al.* (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems*, **3**.

Milanese,A. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.*, **10**, 1–11.

Pasolli,E. *et al.* (2019) Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, **176**, 649–662.e20.

Sczyrba,A. *et al.* (2017) Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.

Shah,N. *et al.* (2021) TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics*.

Stewart,R.D. *et al.* (2019) Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.*, **37**, 953–961.

Sunagawa,S. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods*, **10**, 1196–1199.

Wirbel,J. *et al.* (2019) Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.*, **25**, 679–689.

Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, 1–12.

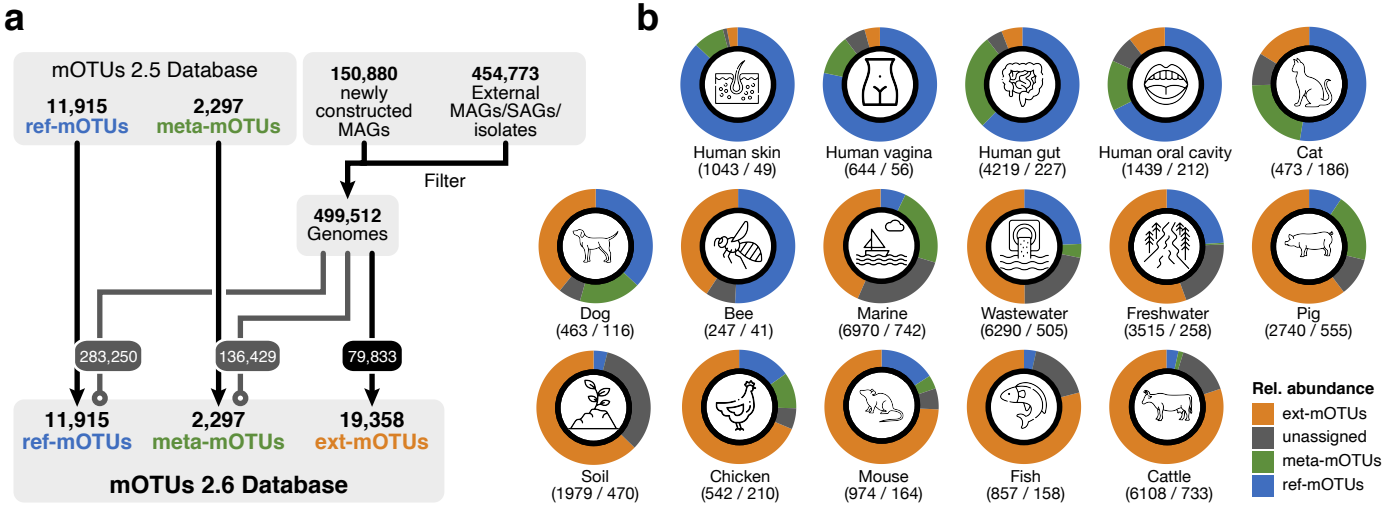
a
mOTUs 2.5 Database
11,915
ref-mOTUs
2,297
meta-mOTUs
150,880
newly
constructed
MAGs
454,773
External
MAGs/SAGs/
isolates
Filter
499,512
Genomes
283,250
136,429
79,833
11,915
ref-mOTUs
2,297
meta-mOTUs
19,358
ext-mOTUs
mOTUs 2.6 Database
b
Human skin
(1043 / 49)
Human vagina
(644 / 56)
Human gut
(4219 / 227)
Human oral cavity
(1439 / 212)
Cat
(473 / 186)
Dog
(463 / 116)
Bee
(247 / 41)
Marine
(6970 / 742)
Wastewater
(6290 / 505)
Freshwater
(3515 / 258)
Pig
(2740 / 555)
Soil
(1979 / 470)
Chicken
(542 / 210)
Mouse
(974 / 164)
Fish
(857 / 158)
Cattle
(6108 / 733)
Rel. abundance
ext-mOTUs
unassigned
meta-mOTUs
ref-mOTUs

Figure 1. An update of the mOTUs database enables comprehensive species-level profiling across diverse microbiomes. (a) The mOTUs 2.6 database was constructed based on version 2.5.1 of the database by adding phylogenetic marker genes from 605,653 MAGs (as well as a small fraction of isolate and single amplified genomes), of which 454,773 originated from external resources and 150,880 were reconstructed *de novo* from 4,531 metagenomes. This addition resulted in the generation of more than 19,358 new ext-mOTUs. Genomes already represented by ref- and meta-mOTUs were discarded (grey lines). **(b)** Breakdown of the relative abundance covered by the three types of mOTUs using the mOTUs 2.6 database. The numbers underneath the ring charts represent the total number of mOTUs that can be profiled per environment (left) considering only species with a prevalence of 0.1%, as well as the median number of mOTUs per sample that were detected after downsampling to 5,000 inserts (right).