TECHNICAL NOTE

CAMITAX: Taxon labels for microbial genomes

Andreas Bremges [1,2,*], Adrian Fritz[1] and Alice C. McHardy [1,*]

[1]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Inhoffenstraße 7, 38124 Braunschweig, Germany; and [2]German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Inhoffenstraße 7, 38124 Braunschweig, Germany

*Correspondence address. Andreas Bremges, E-mail: andreas.bremges@helmholtz-hzi.de http://orcid.org/0000-0001-6739-7899 and Alice C. McHardy, E-mail: alice.mchardy@helmholtz-hzi.de http://orcid.org/0000-0003-2370-3430

Abstract

Background: The number of microbial genome sequences is increasing exponentially, especially thanks to recent advances in recovering complete or near-complete genomes from metagenomes and single cells. Assigning reliable taxon labels to genomes is key and often a prerequisite for downstream analyses. **Findings:** We introduce CAMITAX, a scalable and reproducible workflow for the taxonomic labelling of microbial genomes recovered from isolates, single cells, and metagenomes. CAMITAX combines genome distance–, 16S ribosomal RNA gene–, and gene homology–based taxonomic assignments with phylogenetic placement. It uses Nextflow to orchestrate reference databases and software containers and thus combines ease of installation and use with computational reproducibility. We evaluated the method on several hundred metagenome-assembled genomes with high-quality taxonomic annotations from the TARA Oceans project, and we show that the ensemble classification method in CAMITAX improved on all individual methods across tested ranks. **Conclusions:** While we initially developed CAMITAX to aid the Critical Assessment of Metagenome Interpretation (CAMI) initiative, it evolved into a comprehensive software package to reliably assign taxon labels to microbial genomes. CAMITAX is available under Apache License 2.0 at https://github.com/CAMI-challenge/CAMITAX.

Keywords: Genome Taxonomy; Phylogenetic Placement; Reproducible Research; Docker; Nextflow; CAMI

Introduction

The direct costs for sequencing a microbial genome are at an all-time low: a high-quality draft now costs <$100, a "finished" genome sequence <$500. This has resulted in many culture-dependent genome studies, in which thousands of isolates—selected by, e.g., their distinct phylogeny [1,2], abundance in the human microbiome [3,4], or biotechnological relevance [5,6]—are sequenced.

Single-cell genome and shotgun metagenome studies further contribute to this expansion in genome numbers by enabling access to the genome sequences of (as-yet) uncultured microbes [7–9]. Notably, new bioinformatics methods can reconstruct complete or near-complete genomes even from complex environments [10,11] and easily scale to hundreds or even thousands of metagenome samples [12–16].

Typically, the sequencing and assembly of a new genome is merely a prerequisite for further bioinformatics analyses (and their experimental validation) to uncover novel biological insights by, e.g., functional annotation [17,18] or phenotype prediction [19,20], which often require the genome's taxonomy.

Historically, a bacterial or archaeal species was defined as a collection of strains that share 1 (or more) trait(s) and show DNA-DNA reassociation values of $\geq$70% [21]. However, with the advent of genomics and—more recently—culture-independent methods, this definition was found to be impractical and difficult to implement [22].

Today, 16S ribosomal RNA (rRNA) gene similarity, average nucleotide identity (ANI), genome phylogeny, or gene-centric voting schemes are used for taxonomic assignments [23–28]. These approaches all have their merits (see below), but, to the best of our knowledge, no unifying workflow implementation existed.

To jointly use these complementary approaches, we developed CAMITAX, a scalable and reproducible workflow that combines genome distance–, 16S rRNA gene–, and gene homology–based taxonomic assignments with phylogenetic placement onto a fixed reference tree to reliably infer genome taxonomy.

Methods

In the following, we describe CAMITAX's assignment strategies and its implementation (Fig. 1).

Genome distance–based assignment

An ANI value of 95% roughly corresponds to a 70% DNA-DNA reassociation value (the historical species definition) [24]. In other words, strains from the same species are expected to show >95% ANI [29]. This species boundary appears to be widely applicable and has been confirmed in a recent large-scale study, in which the analyses of 8 billion genome pairs revealed a clear genetic discontinuity among known genomes, with 99.8% of the pairs showing either >95% intraspecies ANI or <83% interspecies ANI values [30].

CAMITAX uses Mash [31] to rapidly estimate the input genomes' ANI to all bacterial or archaeal genomes in the RefSeq database [32] (114,176 strains as of 10 May 2018). CAMITAX's genome distance–based assignment is the lowest common ancestor (LCA) of all Mash hits with >95% ANI; a genome is placed at "root" if there is no RefSeq genome with >95% ANI.

This strategy works best if the query genome is >80% complete (Mash does not accurately estimate the genome-wide ANI of incomplete genomes [33]) and is represented in RefSeq. CAMITAX's other assignment strategies are complementary by design and better suited for incomplete genomes or underrepresented lineages. If a Mash hit is found, however, CAMITAX most likely assigns a taxonomy at the species or genus level.

16S rRNA gene–based assignment

The 16S rRNA gene is widely used for classification tasks because it is a universal marker gene likely present in all bacteria and archaea [34,35].

CAMITAX uses nhmmer [36] to identify 16S rRNA genes in the input genomes and Dada2 [37] to assign taxonomy. Dada2 uses the naive Bayesian classifier method [38] for kingdom to genus assignments, and exact sequence matching against a reference database for species assignments. CAMITAX supports 2 commonly used databases: SILVA [39] and Ribosomal Database Project (RDP) [40], which both were found to map back well to the NCBI Taxonomy [41].

Of course, this strategy only is applicable if the genome assembly contains a copy of the 16S rRNA gene—which is not always the case, particularly for genomes recovered from metagenomes or single cells.

Gene homology–based assignments

Metagenomics and single-cell genomics are complementary approaches providing access to the genomes of (as-yet) uncultured microbes, but both have strings attached: Single amplified genomes (SAGs) are hindered by amplification bias and, as a consequence, are often incomplete [42,43]. Metagenome-assembled genomes (MAGs) on the other hand rarely contain full-length 16S rRNA genes [44,45]. While there are notable exceptions to

this rule [46,47], the above assignment strategies are generally not expected to work well for today's SAGs and MAGs.

To overcome these problems, CAMITAX implements a gene-based voting scheme. It uses Prodigal [48] to predict protein-coding genes, and then Centrifuge [49] and Kaiju [50] for gene homology searches on the nucleotide and protein level, respectively. Both tools scale to large reference databases, such as NCBI's nr/nt [51], but (by default) CAMITAX resorts to the (much smaller) proGenomes genes and proteins datasets [52,53]. The proGenomes database was designed as a resource for consistent taxonomic annotations of bacteria and archaea.

Inferring genome taxonomy from a set of gene-level assignments is not trivial, and—inspired by procedures implemented in anvi'o [27] and dRep [33]—CAMITAX places the query genome on the lowest taxonomic node with ≥50% support in gene assignments (which corresponds to the interval-union LCA algorithm [28]) for nucleotide and protein searches.

Phylogenetic placement

CAMITAX uses CheckM [25] for a phylogeny-driven estimate of taxonomy. Relying on 43 phylogenetically informative marker genes (consisting primarily of ribosomal proteins and RNA polymerase domains), CheckM places the query genome onto a fixed reference tree with Pplacer [54] to infer taxonomy. We note that phylogenetic placement is often quite conservative and does not necessarily provide resolution at the species level [26,55].

Last, CAMITAX reports the query genome's completeness and contamination as estimated by CheckM using its lineage-specific marker genes [25].

Classification algorithm

CAMITAX considers the lowest consistent assignment as the longest unambiguous root-to-node path in the taxonomic tree spanned by the individual assignments; i.e., it retains the most specific, yet consistent taxonomic label among all tools. For example, CAMITAX would determine as "consistent" assignments for the individual assignments (derived with the different assignment strategies) the following:

- 3× *E. coli*, 2× Bacteria ↦ *E. coli*
- 3× *E. coli*, 2× *E. albertii* ↦ *Escherichia*
- 3× *E. coli*, 2× Archaea ↦ Root

This strategy is more robust than computing the LCA of individual assignments because outliers, e.g., missing predictions of conservative methods, do not affect the overall assignment.

At the same time, requiring a consistent assignment is less error prone than, e.g., selecting the maximal root-to-leaf path, which would introduce many false-positive assignments especially on lower ranks.

The trade-off is that incorrect individual assignments, e.g., due to potentially misassembled or misbinned 16S rRNA gene sequences in MAGs, can result in overly conservative assignments on high taxonomic ranks. CAMITAX therefore also reports the maximal root-to-leaf path as an alternative, and we suggest that the user investigate taxonomic discrepancies manually, taking individual assignments into account.

Implementation

CAMITAX incorporates many state-of-the-art pieces of software, and automatically resolves all software and database dependencies with Nextflow [56] in a containerized environment (Table 1).
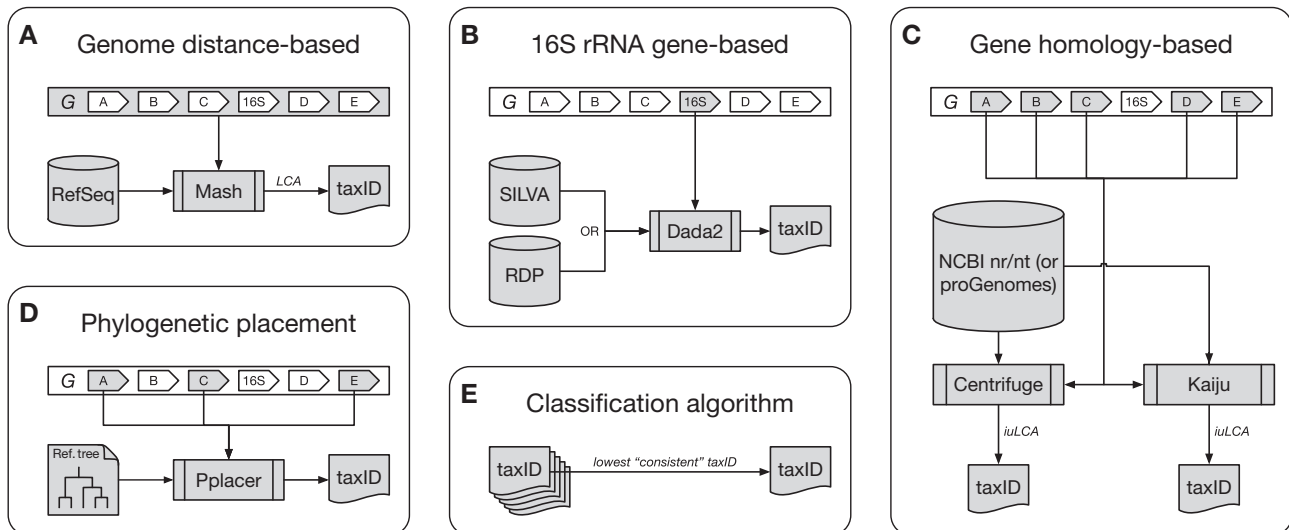
Figure 1: The CAMITAX taxonomic assignment workflow. CAMITAX assigns 1 NCBI Taxonomy ID (taxID) to an input genome G by combining genome distance–, 16S rRNA gene–, and gene homology–based taxonomic assignments with phylogenetic placement. (A) Genome distance–based assignment. CAMITAX uses Mash to estimate the average nucleotide identity (ANI) between G and >100,000 microbial genomes in RefSeq and assigns the lowest common ancestor (LCA) of genomes showing >95% ANI, which was found to be a clear species boundary. (B) 16S rRNA gene–based assignment. CAMITAX uses Dada2 to label G's 16S rRNA gene sequences using the naive Bayesian classifier method to assign taxonomy across multiple ranks (down to genus level) and exact sequence matching for species-level assignments against the SILVA or RDP database. (C) Gene homology–based assignments. CAMITAX uses Centrifuge and Kaiju to perform gene homology searches against nucleotide and amino acid sequences in NCBI's nr and nt (or proGenomes' genes and proteins datasets), respectively. CAMITAX determines the interval-union LCA (iuLCA) of gene-level assignments and places G on the lowest taxonomic node with ≥50% coverage. (D) Phylogenetic placement. CAMITAX uses Pplacer to place G onto a fixed reference tree, as implemented in CheckM, and estimates genome completeness and contamination using lineage-specific marker genes. (E) Classification algorithm. CAMITAX considers the lowest consistent assignment as the longest unambiguous root-to-node path in the taxonomic tree spanned by the 5 taxIDs derived in (A)–(D); i.e., it retains the most specific, yet consistent taxonomic label among all tools.

Table 1: Software used in the CAMITAX workflow

Software	Version	BioContainer
Centrifuge	1.0.3	centrifuge:1.0.3–py36pl5.22.0_2
CheckM	1.0.11	checkm-genome:1.0.11–0
Dada2	1.6.0	bioconductor-dada2:1.6.0–r3.4.1_0
Kaiju	1.6.2	kaiju:1.6.2–pl5.22.0_0
Mash	2.0	mash:2.0–gsl2.2_2
Nhmmer	3.1	
Pplacer	1.1	
Prodigal	2.6.3	prodigal:2.6.3–0

CAMITAX automatically resolves all software dependencies with Nextflow using BioContainers in a containerized environment. Nhmmer and Pplacer are bundled with CheckM.

This fosters reproducibility in bioinformatics research [57,58], and we strongly suggest running CAMITAX using BioContainers [59] (automated container builds for software in Bioconda [60]). CAMITAX can be run on a local machine or in a distributed fashion.

Results

We applied CAMITAX to real data not present in its databases, a recent collection of 885 bacterial and archaeal MAGs from Delmont et al. [15], who used state-of-the-art metagenomic assembly, binning, and curation strategies to create a non-redundant database of microbial population genomes from the Tara Oceans project [61].

Delmont et al. [15] used CheckM for an initial taxonomic inference of the MAGs. Thereafter, they used Centrifuge [49], RAST [62], and manual BLAST searches of single-copy core genes

against NCBI's nr/nt to manually refine their taxonomic inferences. Last, they trained a novel machine learning classifier to also identify MAGs affiliated to the Candidate Phyla Radiation (CPR) [8].

As expected, CAMITAX outperformed CheckM, which is rather conservative in its assignments, by adding low-ranking annotations based on high-quality predictions of other tools, such as Kaiju (Fig. 2). Notably, 95% of CAMITAX's predictions were consistent with Delmont et al. [15], i.e. the two assignments were on the same taxonomic lineage and their LCA is either of the two. CAMITAX assignments of 46 MAGs (5%) were in conflict with the manually curated taxonomy. Of these, CAMITAX made species assignments for 12 MAGs based on Mash hits against RefSeq genomes. These we consider trustworthy because >95% ANI was shown to be a clear species boundary [30], and we assume that Delmont et al. assigned them incorrectly. On the other hand, CAMITAX for instance misclassified MAGs affiliated to the CPR based on their 16S rRNA gene sequences to other phyla.

To quantify taxonomic assignment performance, we calculated precision, recall, and accuracy across all ranks with AMBER 2.0 [63] (Fig. 3). As the gold standard, we used the Delmont et al. [15] assignments up to genus rank. CAMITAX was very precise down to class level and reasonably (>80%) precise below. Overall, it was more accurate across all ranks than each of its assignment strategies individually. While the recall of CAMITAX dropped at the mid-range ranks, largely due to a more conservative assignment strategy compared with the expert curation of Delmont et al., it recovered for genus-level assignments.

We thus propose CAMITAX as a reliable and reproducible taxonomic assignment workflow, ideally followed by a manual refinement step—as always.
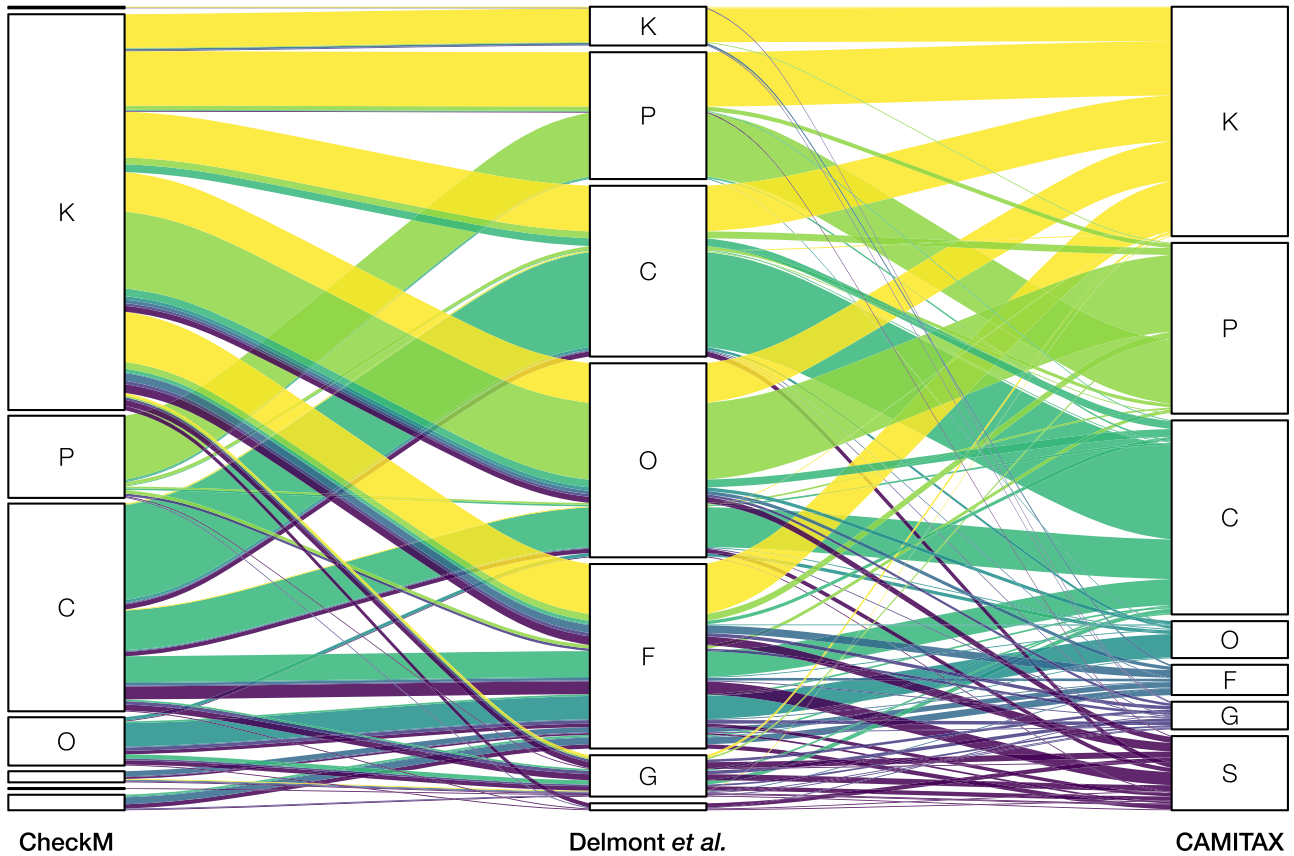
Figure 2: Comparison of high-quality taxonomic assignments for 885 MAGs. Using genome-resolved metagenomics, Delmont et al. [15] assembled 885 bacterial and archaeal genomes from the Tara Oceans metagenomes and used CheckM for an initial taxonomic inference. Subsequently, they manually refined the taxonomic assignments using additional analyses and expert knowledge. The alluvial diagram shows the assigned taxonomic ranks for CheckM (left), manual curation (middle), and CAMITAX (right) on kingdom (K), phylum (P), class (C), order (O), family (F), genus (G), and species (S) level. Colored links between these ranks represent the "flow," i.e. changes in the assignment depth, among the 3 methods.
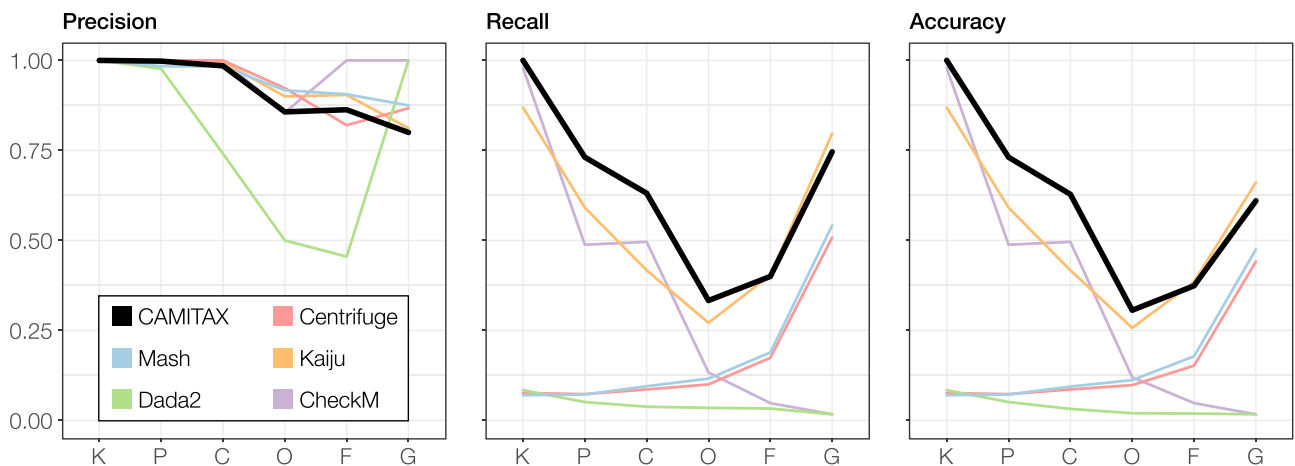

Figure 3: Taxonomic assignment performance metrics across ranks for 885 MAGs. Performance across ranks was assessed with the AMBER software using the manually assigned taxonomy by Delmont et al. [15] as the gold standard. Shown are precision, recall, and accuracy for CAMITAX (and the individual tools combined therein) on kingdom (K), phylum (P), class (C), order (O), family (F), and genus (G) level.

Discussion

CAMITAX was initially developed while preparing the second Critical Assessment of Metagenome Interpretation (CAMI) challenge [64]. The challenge datasets include new genomes from taxa (at different evolutionary distances) not found in public databases yet, which need high-quality taxon labels for the subsequent microbial community and metagenome data simulation [65]. Owing to this need, we created CAMITAX to system-

atically double-check, newly infer, or refine genome taxon label assignments in a fully reproducible way.

CAMITAX combines different taxonomic assignment strategies into one unifying workflow implementation. It uses Nextflow to orchestrate reference databases and software containers. Therefore, both databases and software can be easily substituted, providing the flexibility to cope with rapid change of standards oftentimes observed in the field. For instance, Parks et al. recently proposed a standardized bacterial taxonomy based on genome phylogeny, the so-called Genome Taxonomy Database (GTDB) [66]. While CAMITAX currently uses the NCBI Taxonomy [67], it is (at least in principle) agnostic to the underlying database and could thus be easily adapted to other taxonomy versions that will arise in future.

Software and Availability of Supporting Data and Materials

CAMITAX is implemented in Nextflow and Python 3 and is freely available under Apache License 2.0 at https://github.com/CAMI-challenge/CAMITAX.

Mash sketches for all bacterial and archaeal genomes in RefSeq, snapshots of the NCBI Taxonomy databases, and Centrifuge and Kaiju indices for the proGenomes genes and proteins datasets are collected in Zenodo [68], as are the snapshots used in this study, generated on 10 May 2018 [69].

Dada2-formatted training fasta files, derived from SILVA (release 132) and RDP (training set 16, release 11.5), are also available in Zenodo [70, 71].

The CheckM reference databases are available at https://data.ace.uq.edu.au/public/CheckM_databases.

Snapshots of our code and other data further supporting this work are available in the GigaScience respository, GigaDB [72].

Abbreviations

ANI: average nucleotide identity; BLAST: Basic Local Alignment Search Tool; CPR: Candidate Phyla Radiation; LCA: lowest common ancestor; MAG: metagenome-assembled genome; NCBI: National Center for Biotechnology Information; nr/nt: non-redundant nucleotide; RAST: Rapid Annotation using Subsystem Technology; RDP: Ribosomal Database Project; rRNA: ribosomal RNA; SAG: single amplified genome.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

A.B. implemented the software, performed experiments, and wrote the manuscript with comments from A.F. and A.C.M. A.F. thoroughly tested the software. A.B. and A.C.M. jointly conceived the project and evaluated results. All authors read and approved the final manuscript.

Acknowledgements

References

1. Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 2009;**462**(7276):1056–60.
2. Mukherjee S, Seshadri R, Varghese NJ, et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol 2017;**35**(7):676–83.
3. Browne HP, Forster SC, Anonye BO, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 2016;**533**(7604):543–6.
4. Lagier JC, Khelaifia S, Alou MT, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 2016;**1**:16203.
5. Maus I, Bremges A, Stolze Y, et al. Genomics and prevalence of bacterial and archaeal isolates from biogas-producing microbiomes. Biotechnol Biofuels 2017;**10**:264.
6. Seshadri R, Leahy SC, Attwood GT, et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol 2018;**36**(4):359–67.
7. Rinke C, Schwientek P, Sczyrba A, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 2013;**499**(7459):431–7.
8. Brown CT, Hug LA, Thomas BC, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 2015;**523**(7559):208–11.
9. Hug LA, Baker BJ, Anantharaman K, et al. A new view of the tree of life. Nat Microbiol 2016;**1**:16048.
10. Quince C, Walker AW, Simpson JT, et al. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 2017;**35**(9):833–44.
11. Sczyrba A, Hofmann P, Belmann P, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nat Methods 2017;**14**(11):1063–71.
12. Parks DH, Rinke C, Chuvochina M, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2017;**2**(11):1533–42.
13. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 2018;**5**:170203.
14. Stewart RD, Auffret MD, Warr A, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun 2018;**9**(1):870.
15. Delmont TO, Quince C, Shaiber A, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol 2018;**3**:804–13.
16. Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 2019;**176**(3):649–62.
17. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;**30**(14):2068–9.
18. Kunath BJ, Bremges A, Weimann A, et al. Metagenomics and CAZyme discovery. Methods Mol Biol 2017;**1588**:255–77.
19. Feldbauer R, Schulz F, Horn M, et al. Prediction of microbial phenotypes based on comparative genomics. BMC Bioinformatics 2015;**16**:S1, doi:10.1186/1471-2105-16-S14-S1.
20. Weimann A, Mooren K, Frank J, et al. From genomes to phenotypes: Traitar, the microbial trait analyzer. mSystems 2016;**1**(6), doi:10.1128/mSystems.00101-16.
21. Rosselló-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev 2001;**25**(1):39–67.

22. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 2005;**102**(7):2567–72.

23. Yarza P, Yilmaz P, Pruesse E, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 2014;**12**(9):635–45.

24. Varghese NJ, Mukherjee S, Ivanova N, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res 2015;**43**(14):6761–71.

25. Parks DH, Imelfort M, Skennerton CT, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;**25**(7):1043–55.

26. Stewart RD, Auffret M, Snelling TJ, et al. MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). Bioinformatics 2019;**35**(12):2150–2.

27. Eren AM, Esen OC, Quince C, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 2015;**3**:e1319.

28. Huson DH, Albrecht B, Bağcı C, et al. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biol Direct 2018;**13**(1):6.

29. Thompson CC, Chimetto L, Edwards RA, et al. Microbial genomic taxonomy. BMC Genomics 2013;**14**:913.

30. Jain C, Rodriguez-R LM, Phillippy AM, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;**9**(1):5114.

31. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;**17**(1):132.

32. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;**44**(D1):D733–45.

33. Olm MR, Brown CT, Brooks B, et al. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 2017;**11**(12):2864–8.

34. Pollock J, Glendinning L, Wisedchanwet T, et al. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. Appl Environ Microbiol 2018;**84**(7), doi:10.1128/AEM.02627-17.

35. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018;**16**(7):410–22.

36. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics 2013;**29**(19):2487–9.

37. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 2016;**13**(7):581–3.

38. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 2007;**73**(16):5261–7.

39. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013;**41**:D590–6.

40. Cole JR, Wang Q, Fish JA, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 2014;**42**:D633–42.

41. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?. BMC Genomics 2017;**18**(Suppl 2):114.

42. Clingenpeel S, Clum A, Schwientek P, et al. Reconstructing each cell's genome within complex microbial communities-dream or reality? Front Microbiol 2014;**5**:771.

43. Bremges A, Singer E, Woyke T, et al. MeCorS: metagenome-enabled error correction of single cell sequencing reads. Bioinformatics 2016;**32**(14):2199–201.

44. Hugenholtz P, Skarshewski A, Parks DH. Genome-based microbial taxonomy coming of age. Cold Spring Harb Perspect Biol 2016;**8**(6), doi:10.1101/cshperspect.a018085.

45. Bowers RM, Kyrpides NC, Stepanauskas R, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 2017;**35**(8):725–31.

46. Woyke T, Tighe D, Mavromatis K, et al. One bacterial cell, one complete genome. PLoS One 2010;**5**(4):e10314.

47. Krause S, Bremges A, Munch PC, et al. Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. Sci Rep 2017;**7**(1):3289.

48. Hyatt D, Chen GL, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;**11**:119.

49. Kim D, Song L, Breitwieser FP, et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 2016;**26**(12):1721–9.

50. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;**7**:11257.

51. Sayers EW, Agarwala R, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2019;**47**(D1):D23–8.

52. Mende DR, Sunagawa S, Zeller G, et al. Accurate and universal delineation of prokaryotic species. Nat Methods 2013;**10**(9):881–4.

53. Mende DR, Letunic I, Huerta-Cepas J, et al. proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. Nucleic Acids Res 2017;**45**(D1):D529–34.

54. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 2010;**11**:538.

55. Czech L, Barbera P, Stamatakis A. Methods for automatic reference trees and multilevel phylogenetic placement. Bioinformatics 2019;**35**(7):1151–8.

56. Di Tommaso P, Chatzou M, Floden EW, et al. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;**35**(4):316–9.

57. Bremges A, Maus I, Belmann P, et al. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. Gigascience 2015;**4**, doi:10.1186/s13742-015-0073-6.

58. Belmann P, Dröge J, Bremges A, et al. Bioboxes: standardised containers for interchangeable bioinformatics software. Gigascience 2015;**4**, doi:10.1186/s13742-015-0087-0.

59. da Veiga Leprevost F, Grüning BA, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics 2017;**33**(16):2580–2.

60. Grüning B, Dale R, Sjödin A, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 2018;**15**(7):475–6.

61. Sunagawa S, Coelho LP, Chaffron S, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science 2015;**348**(6237):1261359.

62. Aziz RK, Bartels D, Best AA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 2008;**9**:75.

63. Meyer F, Hofmann P, Belmann P, et al. AMBER: Assessment of Metagenome BinnERs. Gigascience 2018;**7**(6), doi:10.1093/gigascience/giy069.

64. Bremges A, McHardy AC. Critical assessment of metagenome interpretation enters the second round. mSystems 2018;**3**(4), doi:10.1128/mSystems.00103-18.

65. Fritz A, Hofmann P, Majda S, et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome 2019;**7**(1):17.

66. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018;**36**(10):996–1004.

67. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res 2012;**40**(Database issue):D136–43.

68. Bremges A. CAMITAX reference databases. Zenodo. 2019. https://doi.org/10.5281/zenodo.1250043.

69. Bremges A. CAMITAX reference databases (Version 1). Zenodo. 2018. https://doi.org/10.5281/zenodo.1250044.

70. CallahanB. Silva taxonomic training data formatted for DADA2 (Silva version 132). Zenodo. 2018. https://doi.org/10.5281/zenodo.1172782.

71. Callahan B. RDP taxonomic training data formatted for DADA2 (RDP trainset 16/release 11.5). Zenodo. 2017. https://doi.org/10.5281/zenodo.801827.

72. Bremges A, Fritz A, McHardy AC. Supporting data for "CAMITAX: Taxon labels for microbial genomes". GigaScience Database .2019. http://doi.org/10.5524/100680.