Published online 2 February 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqac007

nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning

Sabrina Krakau [1,2,*], Daniel Straub [1,3], Hadrien Gourlé [4], Gisela Gabernet [1] and Sven Nahnsen [1,2,5]

[1]Quantitative Biology Center (QBiC), University of Tübingen, 72076 Tübingen, Germany, [2]Cluster of Excellence – Controlling Microbes to Fight Infections, University of Tübingen, 72076 Tübingen, Germany, [3]Microbial Ecology, Center for Applied Geosciences, University of Tübingen, 72076 Tübingen, Germany, [4]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, S-75007 Uppsala, Sweden and [5]Biomedical Data Science, Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany

Received August 18, 2021; Revised November 19, 2021; Editorial Decision January 10, 2022; Accepted January 25, 2022

ABSTRACT

The analysis of shotgun metagenomic data provides valuable insights into microbial communities, while allowing resolution at individual genome level. In absence of complete reference genomes, this requires the reconstruction of metagenome assembled genomes (MAGs) from sequencing reads. We present the nf-core/mag pipeline for metagenome assembly, binning and taxonomic classification. It can optionally combine short and long reads to increase assembly continuity and utilize sample-wise group-information for co-assembly and genome binning. The pipeline is easy to install-all dependencies are provided within containers-portable and reproducible. It is written in Nextflow and developed as part of the nf-core initiative for best-practice pipeline development. All codes are hosted on GitHub under the nf-core organization https://github.com/nf-core/mag and released under the MIT license.

INTRODUCTION

Shotgun metagenomic approaches enable genomic analyses of all microbes within, for example, environmental or host-associated microbiome communities. Since most bacteria cannot be cultured, isolated and individually sequenced, reference databases for microbial genomes are often incomplete. Thus, one of the main tasks in metagenomic data analysis is to reconstruct the individual genomes directly from the given mixture of metagenomic reads.

A typical reference-independent metagenomic workflow consists of preprocessing raw reads, assembly and binning to generate so-called metagenome assembled genomes (MAGs), as well as taxonomic and functional annotation of MAGs. Assemblies based on short reads typically suffer from being highly fragmented. In contrast, long reads can be used to generate continuous assemblies but suffer from high error rates. Hybrid assembly approaches combine the advantages of both short and long reads and thus, can produce continuous and accurate assemblies at reasonable cost (1). Another important aspect is whether to combine information across samples for assembly. When analyzing multiple samples that contain the same microbes (e.g. cultures and time series), co-assembly increases sequencing depth and can thus improve assembly completeness, in particular with respect to low abundant genomes. The sample-wise sequencing depth information can be used to aid binning methods to decide what contigs should form a MAG. Co-assembly also allows to directly track MAGs through samples instead of computing links between MAGs of multiple assemblies. On the other hand, co-assembly can increase the metagenome complexity and result in more fragmentation or hybrid contigs of multiple similar genomes (2,3).

Several pipelines have been developed for the assembly and binning of metagenomes (4–6). However, only a few pipelines such as Muffin (7) and ATLAS (8) make use of workflow management systems, such as Snakemake (9) or Nextflow (10), which facilitate scalability, portability, reproducibility and ease of application. These pipelines have different strengths and weaknesses but only Muffin supports hybrid assembly and none of them supports co-assembly.

Here we introduce nf-core/mag, a Nextflow pipeline for hybrid assembly of metagenomes, binning and taxonomic classification of MAGs. The nf-core/mag pipeline is ideal for standardized, large-scale and high-throughput analysis. It is also versatile and can use mixed sequencing data (hybrid assembly) or single sequencing technology data, and allows the use of group information to perform co-assembly and the computation of co-abundances used for genome binning. The pipeline is part of the nf-core collection of community curated best-practice pipelines (11).

*To whom correspondence should be addressed. Tel: +49 7071 29 78596; Email: sabrina.krakau@qbic.uni-tuebingen.de

MATERIALS AND METHODS

Implementation and reproducibility

nf-core/mag is written in Nextflow, making use of the new DSL2 syntax (see Supplementary Data: Section S2). DSL2 enables a modularized pipeline structure, where each individual process, containing ideally only one tool, is provided as a 'module'. Additionally, sub-workflows can be integrated. The pipeline strongly benefits from the nf-core framework, which enforces a set of strict best-practice guidelines to ensure high-quality pipeline development and maintenance (11). For example, pipelines must provide comprehensive documentation as well as community support via GitHub Issues and dedicated Slack channels. Pipeline portability and reproducibility are enabled through (i) pipeline versioning (i.e. tagged releases on GitHub), (ii) building and archiving associated containers that contain the required software dependencies (i.e. the exact same compute environment can be used over time and across systems) and (iii) a detailed reporting of the used pipeline/software versions and applied parameters. The use of container technologies such as Docker and Singularity enable reproducibility and portability also across different compute systems, i.e. local computers, HPC clusters and cloud platforms. The nf-core/mag pipeline comes with a small test dataset that is used for continuous integration (CI) testing with GitHub Actions. In addition, 'full-size' pipeline tests are run on AWS for each pipeline release to ensure cloud compatibility and an error-free performance on real-world datasets. The full-size test results for each pipeline release are displayed on the nf-core website (https://nf-co.re/mag/results). Moreover, since the nf-core framework uses DSL2, commonly used processes can be shared across pipelines via nf-core/modules (https://github.com/nf-core/modules). This allows the efficient integration of new analysis tools into the pipeline in the future.

Although the nf-core framework facilitates reproducibility at several layers, it is further crucial to ensure that the individual tools that are part of the pipeline can be run in a deterministic and reproducible manner. For this purpose, nf-core/mag offers dedicated reproducibility settings, for example, to set a random seed parameter, to fix and report multi-threading parameters as well as to generate and/or save required databases, whose public versions do not always remain accessible (see Supplementary Data: Section S3).

Simulation of metagenomic data

To show exemplary results generated with the nf-core/mag pipeline, we simulated metagenomic time series data with CAMISIM (12). For this, CAMISIM was applied based on the genome sources from the 'CAMI II challenge toy mouse gut dataset' (13), containing 791 genomes, while set to generate Illumina and Nanopore reads. Two groups of samples were simulated, each comprising a time series of four samples (for details see Supplementary Data: Section S4). The simulated datasets as well as a sample sheet file, which can be used as input for the nf-core/mag pipeline, are available at https://doi.org/10.5281/zenodo.5155395.

RESULTS

Pipeline overview

An overview of the nf-core/mag pipeline is shown in Figure 1A. The input can be either directly provided FASTQ files containing the short reads or a sample sheet in CSV format containing the paths to short and, optionally, long read files as well as additional group information.

Pre-processing. The pipeline starts with preprocessing the raw reads. For short Illumina reads, fastp (14) is used for adapter and quality trimming, Bowtie2 (15) for identifying and removing host or PhiX reads, and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for quality control (QC) on the raw and preprocessed reads. For long Nanopore reads, porechop (https://github.com/rrwick/Porechop) is used for adapter trimming, NanoLyse (https://github.com/rrwick/Filtlong) to remove phage lambda control contamination, and Filtlong (16) for quality filtering (host read contamination is indirectly removed based on the filtered short reads). QC on the raw and processed long reads is performed using NanoPlot (https://github.com/rrwick/Filtlong).

Assembly and binning. The preprocessed reads are then de novo assembled using MEGAHIT (17) or SPAdes (18). If both short and long reads are provided, a hybrid assembly can be performed using hybridSPAdes (19). By default, nf-core/mag assembles the reads of each sample individually. However, it provides the option to compute co-assemblies according to user specified group information. MetaBAT2 (20) is then used to bin the contigs into individual MAGs based on nucleotide frequencies and co-abundance patterns across samples (within the same group, by default). The pipeline further estimates MAG abundances for the different samples from contig sequencing depths. QUAST (21) summarizes QC features of the generated assemblies and MAGs. MAG completeness and contamination is estimated by BUSCO (22), which makes use of near-universal single-copy orthologs.

Taxonomic classification. Finally, MAGs are taxonomically annotated using GTDB-TK (23) or CAT/BAT (24). While CAT/BAT is able to taxonomically classify any MAG, GTDB-TK requires a number of marker genes and is therefore only applied to MAGs passing quality thresholds regarding the completeness and contamination priorly estimated with BUSCO. Besides the results from the individual tools, nf-core/mag outputs a summary containing estimated abundances, as well as the main QUAST, BUSCO and GTDB-TK metrics for each MAG (see Figure 1C).

Quality assurance. Preprocessed short reads are classified using Kraken2 (25) or Centrifuge (26) and visualized in Krona charts (27) to assess potential contamination and the microbial community before the assembly. MultiQC (28) is used to generate a comprehensive quality report aggregating the QC results across all samples.

Table 1 shows a comparison of nf-core/mag's functionality to other existing pipelines for metagenome assembly and binning. The respective analysis tools used in nf-core/mag were chosen based on benchmarking results from

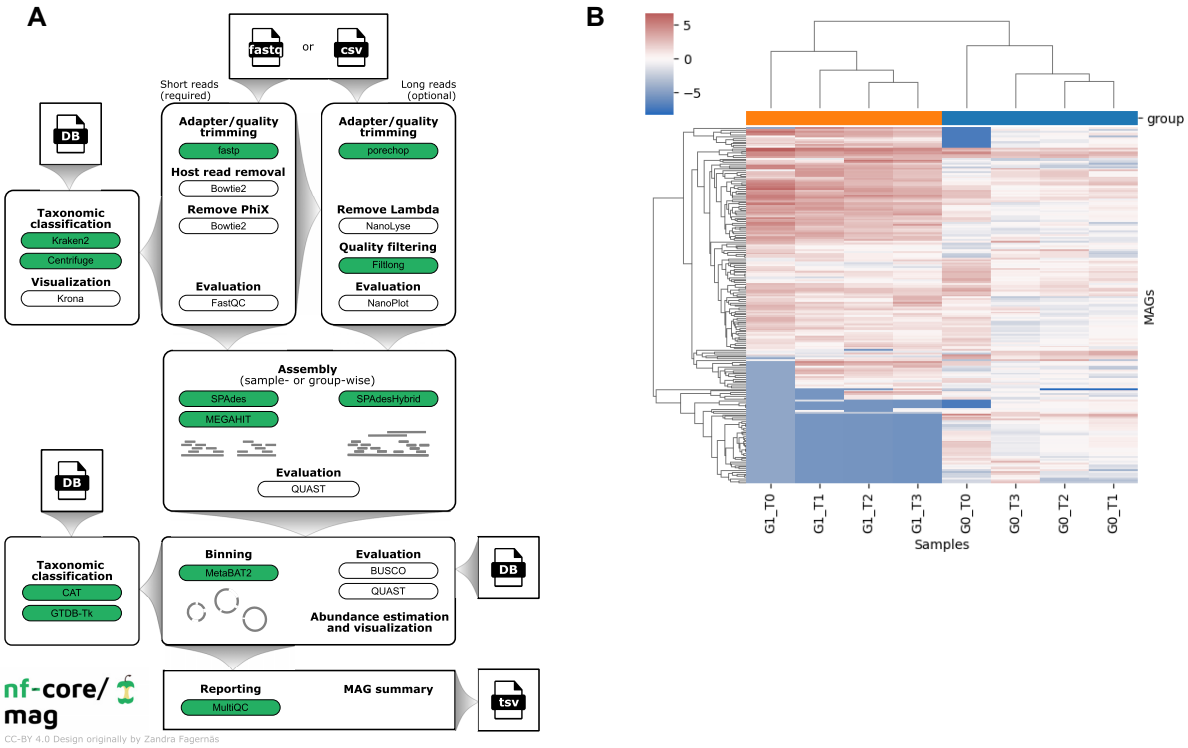

Figure 1. (**A**) Overview of the nf-core/mag pipeline (v2.1.0). (**B**) Clustered heatmap showing MAG abundances, i.e. centered log-ratio depths across samples. (**C**) Schematic representation of MAG summary output, containing abundance information, QC metrics and taxonomic classifications.

Table 1. Comparison of nf-core/mag's functionality with commonly used metagenome assembly and binning pipelines. A more detailed comparison is shown in Supplementary Table S1

	Functionality	Muffin v1.0.3	ATLAS v2.6a2	nf-core/mag v2.1.0
Assembly	Hybrid assembly	Yes	Partial	Yes
	Reassembly after binning	Yes	No	No
	Group-wise co-assembly	No	No	Yes
Genome binning	Group-wise co-abundances used for binning	No	Yes	Yes
	Bin refinement	Yes	Yes	No
	MAG abundance estimation	No	Yes	Yes
Annotation	Taxonomic classification	Yes	Yes	Yes
	Functional annotation	Yes	Yes	No
Usability	Reproducibility	No	No	Yes
	Adherence to set of strict best-practice guidelines for pipeline development	No	No	Yes

the CAMI challenge (13,29) and based on specific user requests from the scientific community. A more detailed pipeline comparison listing also the individual tools as well as a brief discussion about tool choices can be found in Supplementary Data: Section S1.

Exemplary results

We ran nf-core/mag v2.1.0 on the metagenomic time series data simulated with CAMISIM. Figure 1B shows an example heatmap representing MAG abundances across samples, obtained with nf-core/mag performing hybrid, group-wise co-assembly. To illustrate the possible impact of the assembly setting, we compared the results for four different nf-core/mag settings, i.e. (i) short read only, sample-wise assembly, (ii) hybrid, sample-wise assembly, (iii) short read only, group-wise co-assembly and (iv) hybrid, group-wise co-assembly. Figure 2 shows a comparison of the resulting assemblies with respect to commonly used assembly metrics. The results demonstrate that—for this particular time series data—both hybrid assembly as well as group-wise co-assembly increase the assembly's size, its N50 value and the

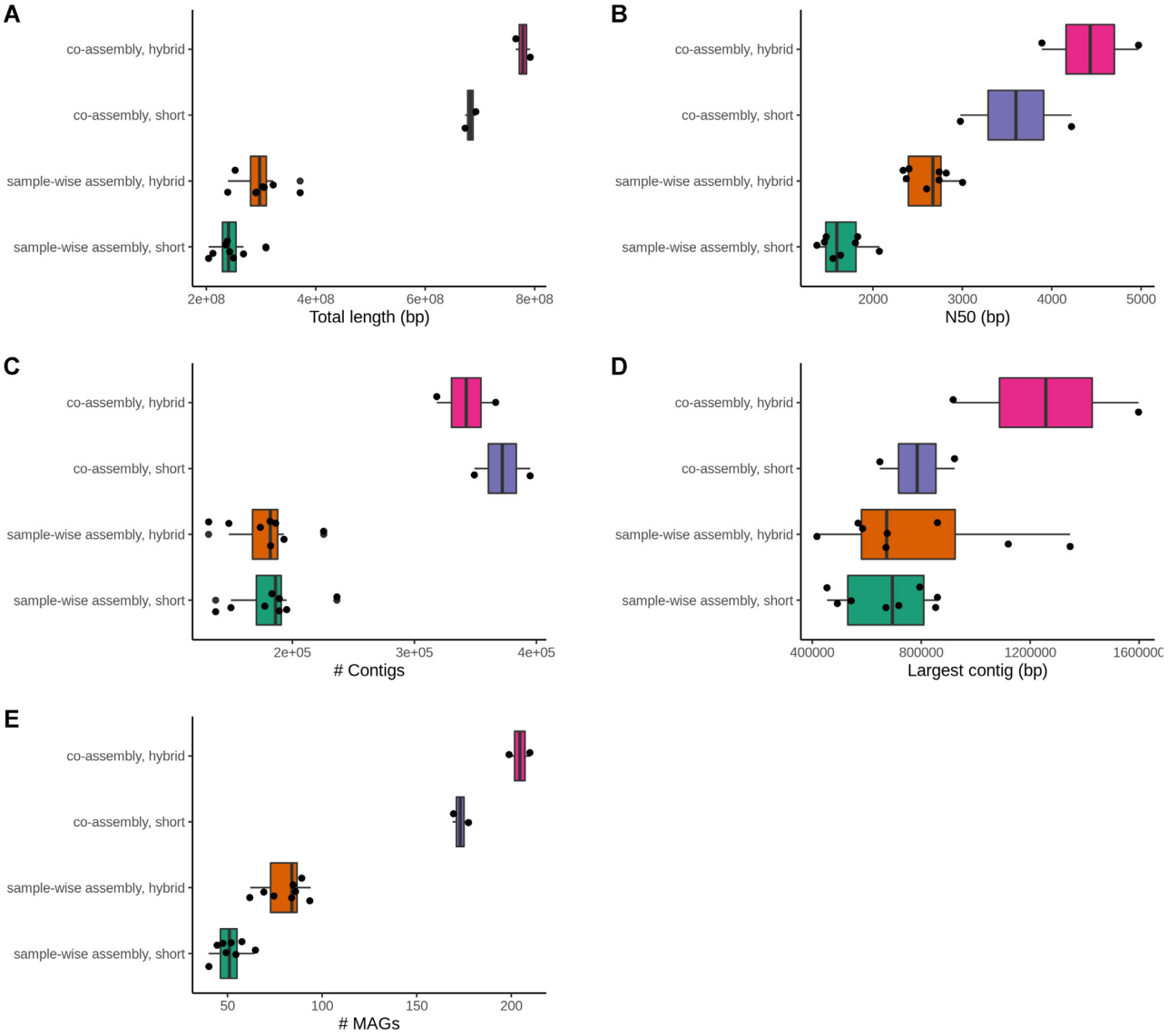

Figure 2. Assembly metrics obtained using different nf-core/mag assembly settings on the simulated data: sample-wise assembly, group-wise co-assembly, short read only assembly or hybrid assembly. Each point corresponds to one assembly, originating either from one sample or one group. Metrics displayed are (**A**) total length of the assembly in base pairs, (**B**) N50 value (i.e. the length of the shortest contig that needs to be included to cover least 50% of the genome), (**C**) number of contigs in the final assembly, (**D**) size of the largest contig in base pairs and (**E**) number of MAGs identified in the final assembly. The metrics (**A**)–(**D**) are part of the QUAST assembly summary.

number of reconstructed MAGs, and thus likely the overall assembly completeness. For further details and results see Supplementary Data: Section S5.

DISCUSSION

We implemented nf-core/mag, an easy to install, reproducible and portable pipeline for hybrid assembly, binning and taxonomic classification of metagenomes. It facilitates the analysis of large metagenomic datasets, while allowing the generation of results that can be reproduced by other scientists. It provides comprehensive usage documentation (https://nf-co.re/mag) and community support via a dedicated Slack channel (https://nfcore.slack.com/channels/mag). One important advantage over existing pipelines is that it can utilize sample-wise group information to perform co-assembly and/or to compute co-abundances used for the genome binning step. This can be particularly useful for the analysis of enrichment cultures, of longitudinal datasets—as often generated in clinical studies—or for studies with interest also in lower abundant genomes. Users are enabled to choose the approach most suitable to their specific research question and experimental setup, or even to compare different settings.

The pipeline was already successfully applied in microbial studies (30,31) and, as part of nf-core, will be constantly maintained and developed further to keep up with state-of-the-art analysis methods. We envision for the future that the here presented version 2.1.0 of nf-core/mag will be further improved, for example, by adding functional annota-

tion as well as assembly and bin refinement steps. The modular DSL2 structure efficiently allows future extensions with new tools and integrations with other (sub-)workflows, for example, for the analysis of metatranscriptomic data. As a community effort focusing on best-practices, an ongoing aim is to join forces with other tool or pipeline developers in the field of metagenomics. At the time of writing this article, for instance, work on a bin refinement step was already started by members of the nf-core community.

DATA AVAILABILITY

nf-core/mag code is hosted on GitHub under the nf-core organization https://github.com/nf-core/mag and released under the MIT license. The with CAMISIM simulated metagenomic data used to generate the exemplary results is available at https://doi.org/10.5281/zenodo.5155395.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the nf-core community for general support during the writing of the pipeline.

FUNDING

Deutsche Forschungsgemeinschaft [under Germany´s Excellence Strategy – EXC 2124 - 390838134 to S.K., D.S., G.G., S.N.; ZUK63 to D.S.; 398967434 - TRR 261 to S.N., im Rahmen der Exzellenzstrategie des Bundes und der Länder - EXC 2180 - 390900677 to S.N.]; German Federal Ministry of Education and Research [01ZX1301F to G.G.]; Chan Zuckerberg Initiative [EOSS2-0000000270 to G.G.]; Ministry of Science, Research and Art Baden-Württemberg [BioDATEN project to G.G., S.N.]. Funding for open access charge: Deutsche Forschungsgemeinschaft [under Germany´s Excellence Strategy – EXC 2124 – 390838134].
Conflict of interest statement. None Declared.

REFERENCES

1. Overholt,W.A., Hölzer,M., Geesink,P., Diezel,C., Marz,M. and Küsel,K. (2020) Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environ. Microbiol.*, **22**, 4000–4013.
2. Hofmeyr,S., Egan,R., Georganas,E., Copeland,A.C., Riley,R., Clum,A., Eloe-Fadrosh,E., Roux,S., Goltsman,E., Buluç,A. *et al.* (2020) Terabase-scale metagenome coassembly with MetaHipMer. *Sci. Rep.*, **10**, 10689.
3. Olm,M.R., Brown,C.T., Brooks,B. and Banfield,J.F. (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.*, **11**, 2864–2868.
4. Fourquet,J., Noirot,C., Klopp,C.C., Pinton,P., Combes,S., Hoede,C. and Pascal,G. (2020) Whole metagenome analysis with metagWGS [Poster]. https://hal.archives-ouvertes.fr/hal-03176836, (16 January 2022, date last accessed).
5. Tamames,J. and Puente-Sánchez,F. (2019) SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.*, **9**, 3349.
6. Uritskiy,G.V., DiRuggiero,J. and Taylor,J. (2018) MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, **6**, 158.
7. Van Damme,R., Hölzer,M., Viehweger,A., Müller,B., Bongcam-Rudloff,E. and Brandt,C. (2021) Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). *PLoS Comput. Biol.*, **17**, e1008716.
8. Kieser,S., Brown,J., Zdobnov,E.M., Trajkovski,M. and McCue,L.A. (2020) ATLAS: a snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinf.*, **21**, 257.
9. Köster,J. and Rahmann,S. (2018) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **34**, 3600–3600.
10. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
11. Ewels,P.A., Peltzer,A., Fillinger,S., Patel,H., Alneberg,J., Wilm,A., Garcia,M.U., Di Tommaso,P. and Nahnsen,S. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.*, **38**, 276–278.
12. Fritz,A., Hofmann,P., Majda,S., Dahms,E., Dröge,J., Fiedler,J., Lesker,T.R., Belmann,P., DeMaere,M.Z., Darling,A.E. *et al.* (2019) CAMISIM: simulating metagenomes and microbial communities. *Microbiome*, **7**, 17.
13. Meyer,F., Lesker,T.-R., Koslicki,D., Fritz,A., Gurevich,A., Darling,A.E., Sczyrba,A., Bremges,A. and McHardy,A.C. (2021) Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nat. Protoc.*, **16**, 1785–1801.
14. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
15. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
16. De Coster,W., D'Hert,S., Schultz,D.T., Cruts,M. and Van Broeckhoven,C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.
17. Li,D., Luo,R., Liu,C.-M., Leung,C.-M., Ting,H.-F., Sadakane,K., Yamashita,H. and Lam,T.-W. (2016) MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, **102**, 3–11.
18. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
19. Antipov,D., Korobeynikov,A., McLean,J.S. and Pevzner,P.A. (2016) hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, **32**, 1009–1015.
20. Kang,D.D., Li,F., Kirton,E., Thomas,A., Egan,R., An,H. and Wang,Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.
21. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
22. Manni,M., Berkeley,M.R., Seppey,M., Simao,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
23. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. and Parks,D.H. (2020) GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, **36**, 1925–1927.
24. von Meijenfeldt,F.A.B., Arkhipova,K., Cambuy,D.D., Coutinho,F.H. and Dutilh,B.E. (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.*, **20**, 217.
25. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.*, **20**, 257.
26. Kim,D., Song,L., Breitwieser,F.P. and Salzberg,S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, **26**, 1721–1729.
27. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a web browser. *BMC Bioinf.*, **12**, 385.

28. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
29. Sczyrba,A., Hofmann,P., Belmann,P., Koslicki,D., Janssen,S., Dröge,J., Gregor,I., Majda,S., Fiedler,J., Dahms,E. *et al.* (2017) Critical assessment of metagenome interpretation - a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.
30. Huang,Y.-M., Straub,D., Blackwell,N., Kappler,A. and Kleindienst,S. (2021) Meta-omics reveal Gallionellaceae and Rhodanobacter species as interdependent key players for Fe(II) oxidation and nitrate reduction in the autotrophic enrichment culture KS. *Appl. Environ. Microbiol.*, **87**, e0049621.
31. Huang,Y.-M., Straub,D., Kappler,A., Smith,N., Blackwell,N. and Kleindienst,S. (2021) A novel enrichment culture highlights core features of microbial networks contributing to autotrophic Fe(II) oxidation coupled to nitrate reduction. *Microb. Physiol.*, **31**, 280–295.