Received: 12 September 2022 | Revised: 18 October 2023 | Accepted: 13 November 2023

DOI: 10.1111/1755-0998.13904

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

MuDoGeR: Multi-Domain Genome recovery from metagenomes made easy

Ulisses Rocha[1] | Jonas Coelho Kasmanas[1,2] | René Kallies[1] | Joao Pedro Saraiva[1] | Rodolfo Brizola Toscan[1] | Polonca Štefanič[3] | Marcos Fleming Bicalho[1] | Felipe Borim Correa[1] | Merve Nida Baştürk[1] | Efthymios Fousekis[1] | Luiz Miguel Viana Barbosa[1] | Julia Plewka[4] | Alexander J. Probst[4] | Petr Baldrian[5] | Peter F. Stadler[6,7,8,9] | CLUE-TERRA Consortium

[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

[2]Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil

[3]Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia

[4]Environmental Microbiology and Biotechnology, Department of Chemistry, University of Duisburg-Essen, Essen, Germany

[5]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Praha 4, Czech Republic

[6]Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany

[7]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

[8]Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria

[9]The Santa Fe Institute, Santa Fe, New Mexico, USA

Correspondence
Ulisses Rocha, Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany.
Email: ulisses.rocha@ufz.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 460129525; Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Number: 2019/03396-9 and 2022/03534-5; German Network for Bioinformatics Infrastructure; Helmholtz-Gemeinschaft, Grant/Award Number: VH-NG-1248; Lundin Energy Norway

Handling Editor: Sebastien Calvignac-Spencer

Abstract

Several computational frameworks and workflows that recover genomes from prokaryotes, eukaryotes and viruses from metagenomes exist. Yet, it is difficult for scientists with little bioinformatics experience to evaluate quality, annotate genes, dereplicate, assign taxonomy and calculate relative abundance and coverage of genomes belonging to different domains. MuDoGeR is a user-friendly tool tailored for those familiar with Unix command-line environment that makes it easy to recover genomes of prokaryotes, eukaryotes and viruses from metagenomes, either alone or in combination. We tested MuDoGeR using 24 individual-isolated genomes and 574 metagenomes, demonstrating the applicability for a few samples and high throughput. While MuDoGeR can recover eukaryotic viral sequences, its characterization is predominantly skewed towards bacterial and archaeal viruses, reflecting the field's current state. However, acting as a dynamic wrapper, the MuDoGeR is designed to constantly incorporate updates and integrate new tools, ensuring its ongoing relevance in the rapidly evolving field. MuDoGeR is open-source software available at

Ulisses Rocha and Jonas Coelho Kasmanas should be considered joint first author.

CLUE-TERRA Consortium group members are included in Acknowledgements.

https://github.com/mdsufz/MuDoGeR. Additionally, MuDoGeR is also available as a Singularity container.

KEYWORDS
genome reconstruction, metagenome-assembled genomes, metagenomics, multi-domain, uncultivated viral genomes

1 | INTRODUCTION

Metagenomics encompasses the concepts and techniques used to study genetic material recovered directly from mixed microbial communities (Dias et al., 2019; Keller-Costa et al., 2021; López-Mondéjar et al., 2020). Essentially, two main approaches are used in metagenomics research: read-based metagenomics and genome-centric metagenomics. While both approaches involve high-throughput reads sequencing all the DNA in a sample, they differ in how they analyse the generated data. Read-based metagenomics focuses on individual reads and can provide a broad overview of the genes and organisms in a sample but cannot link genes to a specific organism in a determined environment. In addition, read-based approaches heavily depend on the completeness and accuracy of reference databases, which may not include all microbial diversity. In contrast, genome-centric metagenomics focuses on assembling these reads into genomes. This approach allows for the characterization of novel and uncultured organisms not represented in databases, provides a more comprehensive understanding of the functional potential of individual organisms and helps link specific genes and metabolic pathways to their originating organisms (Breitwieser et al., 2018; Saraiva, Worrich, et al., 2021). Genomes reassembled from high-throughput sequencing reads have been done since 2004 (Tyson et al., 2004), although the exponential increase in genome recovery only occurred after 2011 with the development of many specialized metagenomic assemblers (Koren et al., 2011; Li et al., 2016; Namiki et al., 2012; Nurk et al., 2017; Peng et al., 2012). In 2013, differential coverage binning was used in the first open-source tool for genome-resolved metagenomics (Albertsen et al., 2013; Sharon et al., 2013). Since then, the bioinformatics community has developed different frameworks specific to genome recovery of prokaryotes (Corrêa et al., 2020; Sieber et al., 2018; Uritskiy et al., 2018), eukaryotes (Corrêa et al., 2020; Kasmanas et al., 2021; West et al., 2018) and viruses (Corrêa et al., 2020; Guo et al., 2021; Kallies et al., 2019; Kieft et al., 2020; Ren et al., 2017). Notice that different bioinformatic approaches are needed for the different domains because they have fundamentally different types of genomes. For instance, viral sequences are usually much smaller and have a higher relative diversity than prokaryotes and eukaryotes (Koonin et al., 2023). Eukaryotic sequences are generally more complex and contain more repetitive sequences, introns, exons and a higher proportion of non-coding DNA (Yandell & Ence, 2012).

Consequently, specific bioinformatic approaches must be developed to address its sequence particularities. For example, specialized machine learning models have been trained for identifying viral sequences in Virfinder, while the model trained in EukRep showed higher accuracy for identifying eukaryotic sequences (Ren et al., 2017; West et al., 2018). Genome-centric metagenomics allows the simultaneous exploration of individual populations' functional potential and phylogeny of uncultivated species (Evans et al., 2015; Liu et al., 2022; Melkonian et al., 2021; Saraiva et al., 2017; Saraiva, Bartholomäus, et al., 2021; Tláskal et al., 2021). Although software developers created several genome-centric tools, most require users with bioinformatics or computational biology expertise to understand the overall pipeline, instal the different dependencies, run a complex framework often written in various computer languages and understand the output (Oliveira Monteiro et al., 2022). Some efforts are worth mentioning, for instance, the Galaxy (Blankenberg et al., 2010), QIIME (Caporaso et al., 2010) and Anvi'o (Eren et al., 2015). While Galaxy's platform was designed to be user-friendly and flexible, the genome assembly in Galaxy can be limited by the computational resources of the specific instance and the availability of its installation in most research environments. In addition, the complexity of creating and managing workflows in Galaxy can be challenging for users without a strong bioinformatics background. On the other hand, QIIME was primarily designed for microbiome amplicon analysis. Consequently, while QIIME can handle some aspects of genome assembly, it may not have the same breadth of tools for genome assembly as platforms specifically designed for this task. Anvi'o has made significant progress in genome-centric analysis by providing a wide range of open-source and flexible tools. However, the platform requires a steep learning curve and lacks easier automation, a significant attribute for usability. Other MAGs recovery wrappers worth mentioning are MetaWrap (Uritskiy et al., 2018), ATLAS (Kieser et al., 2020), nf-core/mag (Krakau et al., 2022) and MAGNETO (Churcheward et al., 2022). The MAGNETO manuscript thoroughly compares these tools using the Human Microbiome Project (HMP) dataset. All four tools (MetaWrap, ATLAS, nf-core/mag and MAGNETO) utilize similar software for assembly and binning. However, they differ in their approaches to binning refinement and the number of MAGs they can reconstruct.

MAGNETO systematically reconstructed more MAGs than ATLAS. nf-core/mag and MAGNETO use the same tools for assembly and binning, and as expected, they yielded the same number of MAGs. MetaWrap includes a binning refinement module that conducts pairwise alignment of MAGs to identify redundant genomes, retaining only the highest quality MAGs among detected duplicates. In the MAGNETO experiment, MetaWrap produced more MAGs than other wrappers (Churcheward et al., 2022).

All mentioned wrappers are single-domain MAGs recovery systems and, therefore, do not simultaneously recover genomes from three domains: prokaryotes, viruses and eukaryotes. At the same time, the genome assembly from the three domains is biologically not separable, but the separation often observed is due to different technical approaches. Theoretically, a genome assembly wrapper that can recover multi-domain can provide a holistic view of all three domains simultaneously, offering a more integrated and complete perspective of the metagenomic environment. To extend the use of genome-centric metagenomics to those with entry-level expertise in bioinformatics basic knowledge of the Unix command-line environment, and ensure the data generated can be reused, the next steps in software development in the field need to guarantee that (i) the tools can be used, installed and maintained by non-experts in computational biology and bioinformatics (Mangul et al., 2019) and (ii) data generated by any tool follow the FAIR guiding principles (Corrêa et al., 2020; Kasmanas et al., 2021; Wilkinson et al., 2016). FAIR stands for findable, accessible, interoperable and reusable. Consequently, it is vital for the method and the researcher to provide data with unique identifiers and be easily discoverable (Findable), openly available (Accessible), structured using standardized formats (Interoperable), and Reusable well-documented (Reusable).

Here we present MuDoGeR (Multi-Domain Genome Recovery—/mjuːˈdoʊgər/), a framework that makes it easy to recover genomes from prokaryotes, eukaryotes and viruses from metagenomes. We developed MuDoGeR to be easily installed and used by those with entry-level bioinformatics expertise and basic Unix command-line environment knowledge. Our modular framework allows one to use the same input to recover genomes from prokaryotes, eukaryotes and viruses alone or in combination. At the same time, users may install and use the individual modules independently. We also constructed our framework to ensure the outputs could be used to simultaneously study the genetic potential and phylogeny of the individual species from which genomes were recovered.

2 | METHODS

We divided MuDoGeR into five modules (Figure 1a). Module 1 deals with the pre-processing of the raw sequences. After module 1, MuDoGeR branches into three Modules (2–4). Module 2 can be used to recover prokaryotic metagenome-assembled genomes (MAGs). Module 3 was assembled to generate uncultivated viral genomes (UViGs), and Module 4 can retrieve eukaryotic metagenome-assembled bins (MABs). We recover eukaryotic MABs as eukaryotes usually need much higher coverage to assemble than prokaryotic MAGs, allowing those interested in eukaryotes to study even partial eukaryotic genomes. We used the following definitions for the recovered sequences: MAGs are prokaryotic sequences with at least 50% completeness and less than 10% contamination based on CheckM results and a quality score higher or equal to 50, where quality score = completeness-5*contamination, as defined by Parks et al., 2017; UViG is a viral genome sequence that has been identified in metagenome or metatranscriptome datasets, its sequence is not derived from an isolate and follows the Minimum Information about an Uncultivated Virus Genome (MIUViG) (Roux et al., 2019); and eukaryotic MABs are eukaryotic bins with completeness values above 40% and contamination values below 5% based on the results from EukCC (Saary et al., 2020). Module 5 was designed to generate outputs to be used in genome-centric biodiversity analysis, and users can use it to calculate the coverage and relative abundance table of the recovered MAGs, UViGs, MABs and open reading frames (ORFs) in Modules 2–4.

Currently, MuDoGeR works on paired-end short-sequence reads generated by ILLUMINA machines, but future updates will include tools to work with data from long-read sequencing. Additionally, while MuDoGeR can recover eukaryotic viral sequences (via VirSorter2 (Guo et al., 2021) and VIBRANT (Kieft et al., 2020)), its characterization is predominantly skewed towards bacterial and archaeal viruses, reflecting the current state of the field. However, acting as a dynamic wrapper, the MuDoGeR is designed to constantly incorporate updates and integrate new tools, ensuring its ongoing relevance in the rapidly evolving field. The MuDoGeR framework is a wrapper of more than 20 tools and 50 custom scripts written using bash, Python and R. It was designed to be an easy-to-use tool that outputs ready-to-use comprehensive files (Figure 1b). After installation, the user can run the complete pipeline using five commands, one for each module. Each module of MuDoGeR generates several outputs that users may use genome-centric analysis focusing on genetic potential, phylogenetic biodiversity of MAGs, UViGs and MABs individually or merge both types of analysis in the same study (Figure 1b).

2.1 | MuDoGeR can be easily installed

Part of the effort to make MAGs recovery ‘easy’ involves making the complex bioinformatics tools easily accessible, requiring little user input. A MAG/UViG/MAB recovery pipeline requires several tools that have multiple dependencies. More often than not, these tools have conflicting dependencies. Consequently, making them cooperate in a single operating system is challenging even for bioinformatics and computational biology experts. To tackle this problem, MuDoGeR uses the Conda environment management system (Grüning et al., 2018) to create multiple environments automatically orchestrated during the pipeline's functioning. A Conda environment is an isolated workspace, or ‘environment’, where specific versions of software packages and their dependencies can be installed without interfering with each other. Conda environments create separate directories for each environment and adjust system paths accordingly so that only the desired versions of each software package are

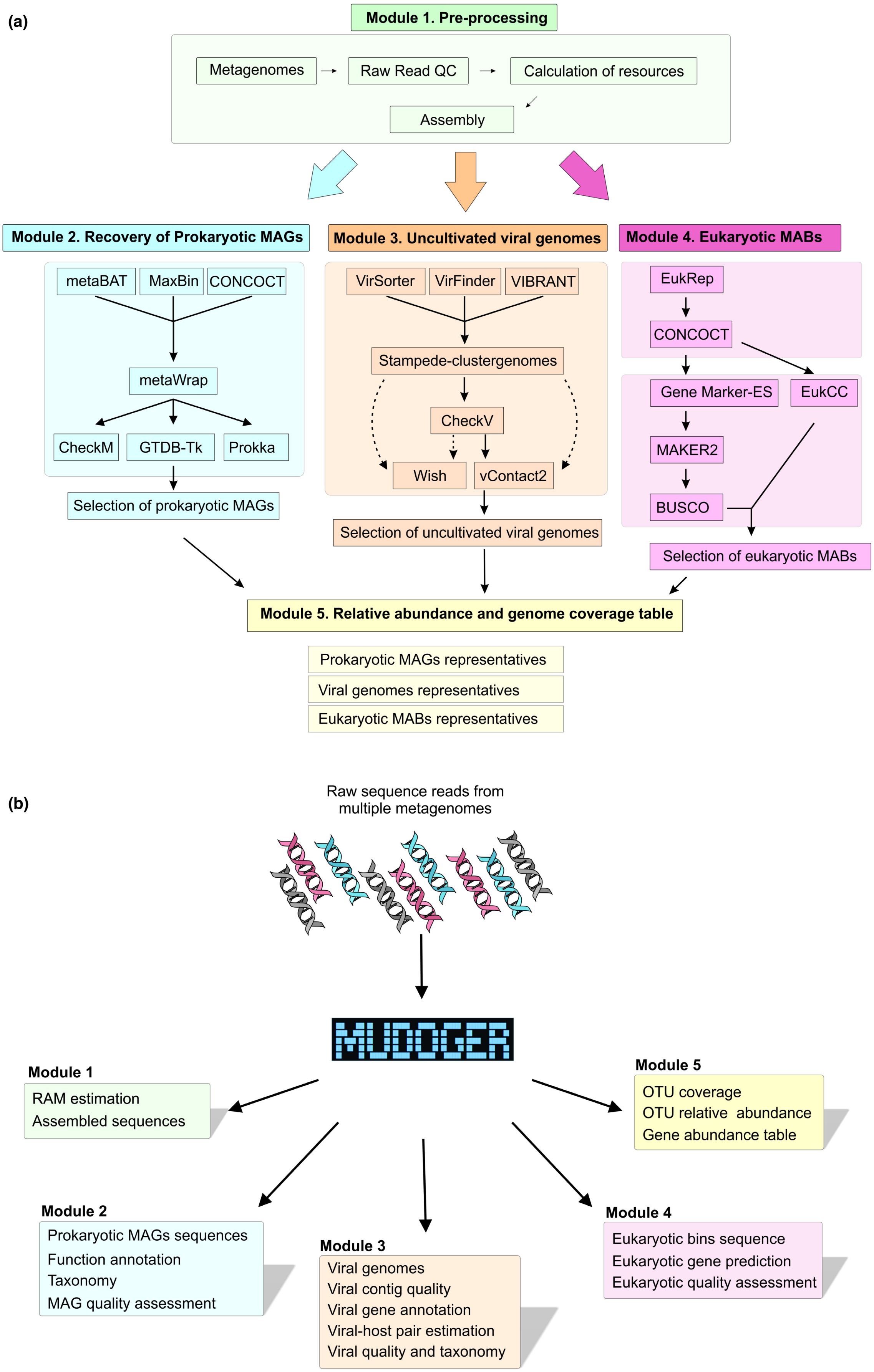
(a)
Module 1. Pre-processing
Metagenomes
Raw Read QC
Calculation of resources
Assembly
Module 2. Recovery of Prokaryotic MAGs
metaBAT
MaxBin
CONCOCT
metaWrap
CheckM
GTDB-Tk
Prokka
Selection of prokaryotic MAGs
Module 3. Uncultivated viral genomes
VirSorter
VirFinder
VIBRANT
Stampede-clustergenomes
CheckV
Wish
vContact2
Selection of uncultivated viral genomes
Module 4. Eukaryotic MABs
EukRep
CONCOCT
Gene Marker-ES
EukCC
MAKER2
BUSCO
Selection of eukaryotic MABs
Module 5. Relative abundance and genome coverage table
Prokaryotic MAGs representatives
Viral genomes representatives
Eukaryotic MABs representatives
(b)
Raw sequence reads from multiple metagenomes
MUDOGER
Module 1
RAM estimation
Assembled sequences
Module 2
Prokaryotic MAGs sequences
Function annotation
Taxonomy
MAG quality assessment
Module 3
Viral genomes
Viral contig quality
Viral gene annotation
Viral-host pair estimation
Viral quality and taxonomy
Module 4
Eukaryotic bins sequence
Eukaryotic gene prediction
Eukaryotic quality assessment
Module 5
OTU coverage
OTU relative abundance
Gene abundance table

FIGURE 1 MuDoGeR workflow and outputs. (a) MuDoGeR modular workflow. Module 1—Pre-processing; Module 2—Recovery of prokaryotic metagenome-assembled genomes (MAGs); Module 3—Recovery of Uncultivated viral contigs (UViGs); Module 4—Recovery of eukaryotic metagenome-assembled bins (MABs); Module 5—Calculation of coverage and relative abundances. (b) MuDoGeR modular outputs. In module 1, random access memory (RAM) is calculated (for optimization purposes), and sequences are assembled. In module 2, all prokaryotic MAG sequences are stored, assigned taxonomy, annotated and quality assessed. In module 3, all viral contigs are stored, their quality assessed, and taxonomy assigned and annotated. Additional outputs include the estimation of viral hosts. Outputs of module 4 include the storage of all eukaryotic bin sequences, quality assessment and gene prediction. Module 5 outputs consist of tables containing the operation taxonomic units (OTU) coverage, relative abundances and gene abundance.

accessible from within each environment. This separation allows users to switch between environments depending on the software requirements, ensuring compatibility and reproducibility of results. The installation protocol creates and automatically sets up 24 independent environments. Additionally, MuDoGeR is readily available as a singularity container (Kurtzer et al., 2017), increasing software distribution and reproducibility by making its installation easier.

In addition, several MAG/UViG/MAB recovery tools require specific databases. Knowing the correct databases and adequately installing them can also limit the genome recovery process by non-bioinformaticians. In MuDoGeR, we also developed a database 'download and set up'-tool that makes them ready to use. Currently, MuDoGeR makes 13 databases available and integrated into our pipeline. Briefly, the user needs to run the database-setup.sh script and specify the location to save them. Notice that we allow the user to download and set up all the required databases for all MuDoGeR modules or instal only the databases required by specific modules by choosing the value for the --dbs flag (all, prokaryotes, viruses, eukaryotes). We plan to update the automatic database installation script once a year or when major database updates are released. Finally, we understand that users might want to include their custom databases, although this would be true for more experienced users. In that case, we included a section in our GitHub documentation guiding the user on how to include their databases.

Consequently, it is worth mentioning that genome-centric metagenomic analysis is computationally intensive, and we recommend MuDoGeR usage under cloud environments or high-performance computers (HPC). The user can find more extensive documentation on the software dependencies and system requirements on MuDoGeR's GitHub page (https://github.com/mdsufz/MuDoGeR).

2.2 | MuDoGeR v1.0 at a glance

The only mandatory inputs for MuDoGeR are the paths to the samples' raw sequencing reads. Following this, the user can run each complete module with one command. Module 1 starts by using the procedure implemented in MetaWrap (Uritskiy et al., 2018) to quality control the raw sequences. We then created a regression model to estimate the Random Access Memory (RAM) required to assemble metagenomes with metaSPAdes (Nurk et al., 2017) using the selected 574 metagenomic libraries. From these, 558 (97.2%) fell in the predicted range (Figure S1). The total RAM correlates highly with k-mer frequencies (Figure S2).

By default, MuDoGeR uses metaSPAdes for assembly, but we also provide the option to use MegaHit (Li et al., 2016). While metaSPAdes results in better-assembled sequences in most samples, MegaHit is significantly more memory efficient and faster, and it may be used in case the user lacks the necessary computer resources (Meyer et al., 2021). If the user is interested in performing co-assembly, we provide a tutorial on MuDoGeR's Manual on GitHub. Module 2 integrates the prokaryotic binning procedure implemented in MetaWrap using Metabat2 (Kang et al., 2015), Maxbin2 (Wu et al., 2016), and CONCOCT (Alneberg et al., 2014) binning tools. MuDoGeR uses CheckM (Parks et al., 2015) and GTDB-tk (Chaumeil et al., 2020) to estimate quality and taxonomy. Moreover, Prokka (Seemann, 2014) annotates ORFs, and BBtools (sourceforge.net/projects/bbmap/) is employed to calculate sequence metrics from the prokaryotic MABs. Finally, MuDoGeR summarizes all the outputs and selects prokaryotic MAGs as defined by Parks and collaborators (Parks et al., 2017).

The recovery of UViGs, performed in module 3, starts by integrating the viral sequence recovery tools VirSorter2 (Guo et al., 2021), VirFinder (Ren et al., 2017) and VIBRANT (Kieft et al., 2020). Later, the potential viral sequences are dereplicated using Stampede-clustergenomes (https://bitbucket.org/MAVERICLab/stampede-clustergenomes/src/master/). Subsequently, putative viral sequences are annotated using vConTACT2 (Bin Jang et al., 2019), and quality assessment is performed using CheckV (Nayfach et al., 2021). MuDoGeR also uses WIsH (Galiez et al., 2017) to predict the potential prokaryotic hosts from the recovered viral sequences by integrating the results from modules 2 and 3. Finally, MuDoGeR compiles the tools' outputs and selects high-quality and complete UViGs, as indicated by CheckV developers (Nayfach et al., 2021).

In Module 4, MuDoGeR integrates EukRep (West et al., 2018) for selecting the eukaryotic contigs from the initial assembly, the CONCOCT binning tool, GeneMark (Besemer et al., 2001) for the prediction of eukaryotic genes, EukCC (Saary et al., 2020) for quality estimation from eukaryotic sequences, MAKER2 (Holt & Yandell, 2011) for gene annotation and BUSCO (Waterhouse et al., 2017) for detection of single-copy orthologous genes. Lastly, in module 5, MuDoGeR groups the recovered MAGs into operation taxonomic units (OTUs). Following, it maps the sequencing reads to the OTUs using two possible approaches: reduced or complete. The reduced method maps the recovered MAGs/UViGs/MABs on their respective libraries, while the complete method maps the recovered MAGs/UViGs/MABs on all available libraries. We kept the reduced mapping as default as, mostly, genomes recovered from different libraries may show coverage in a library even if, in reality, it was not present there. This is because a significant fraction of the genome encodes genes of the central metabolism that are highly conserved across most microbial species (Noor et al., 2010). Such genes are likely to map to a related species that may not be present in the sample. However, depending on the user interests, the complete mapping method might be useful as it can provide relevant insights into comparative genomics, detection of contaminants and identification of horizontal gene transfer, for instance. When performing the mapping, the users should, naturally, be aware that not all genomes present in a sample are successfully assembled. Later, it calculates the relative abundance and coverage tables from the mapped MAGs/UViGs/MABs and annotated prokaryotic genes within the assembled sequences. By the end, the user should have standardized results from a complete MAG pipeline.

We designed MuDoGeR as a wrapper for several complex tools. Further, we structured our tool to make all integrated software available for the user independently. This feature means that a more experienced user could integrate only pieces of MuDoGeR into their pipeline or even access a specific environment configured by MuDoGeR and use only the selected tool. Users can find a tutorial on activating these modules independently on the MuDoGeR GitHub page (https://github.com/mdsufz/MuDoGeR). Consequently, MuDoGeR modularity can give the researcher flexibility in their analysis and facilitate the investigator's software management necessities.

Finally, we have implemented a file-checking mechanism for automatic resuming. This mechanism checks and ensures that specific and necessary files are present and correctly formatted before the workflow begins. If the workflow is interrupted for any reason, this file check allows it to resume from the point of interruption, thus saving time and computational resources.

MuDoGeR is designed to support Linux x64 systems. Some of the used software requires a large amount of RAM (e.g., GDTB-Tk, metaSPAdes). However, specific resource requirements vary depending on the user's data and sequencing depth. For this reason, and to reduce the over/underuse of RAM, MuDoGeR attempts to calculate the memory necessary for metaSPAdes.

The complete software installation requires approximately 170 GB, but MAKER2, from Module 4, uses 99 GB. The entire database requirements, considering all tools, are currently around 439.9 GB. In addition, it is recommended that the user provides multiple processing units since analysing several metagenomes simultaneously may require significant time. Consequently, the MuDoGeR Conda installation procedure allows it to be installed on high-performance computers (HPC) or in cloud services such as Amazon Elastic Compute Cloud, Google Cloud, or, for researchers in Germany, the German Network for Bioinformatics Infrastructure, allowing users to work with larger metagenome datasets. It is worth mentioning that accessing and utilizing HPC has become significantly more user-friendly and accessible in recent years, with many platforms offering intuitive interfaces and comprehensive support. Furthermore, most universities and research institutes either have their own HPC clusters or access to one, and resources are typically readily available to students, postdocs and faculty members—often, all one needs to do is ask. Finally, the MuDoGeR singularity container should be directly usable in HPC environments.

3 | RESULTS AND DISCUSSION

We tested the MuDoGeR pipeline using 598 metagenome libraries (Illumina short reads) from public repositories (Table S1). We selected these libraries to encompass metagenomes (574) and individual-isolated genomes (24). Libraries from axenic genomes were explicitly chosen to test whether our pipeline can recover viruses in individual isolates (Table S2). We could not evaluate our pipeline using a mock community because, to the best of our knowledge, no standardized community containing prokaryotes, eukaryotes and viruses existed during our analysis. Additionally, we used MuDoGeR on samples from Taur et al., 2018 (SRR14092160, SRR14092310, SRR14143424), also used during the Kraken study (Lu et al., 2022). After, we compared the diversity recovered by Kraken, a read-based approach, with MuDoGeR, a genome assembly-based method.

Only 19 of the 574 metagenomic libraries showed no recovery of prokaryotic MAGs, putative viral contigs or eukaryotic MABs (Table S1). Several factors influence genome recovery from metagenomic samples, such as sample evenness, sequencing depth and taxonomic relatedness. Sequencing depth was found to influence the accurate recovery of genomes significantly. However, it is also worth mentioning that those samples have recovered bins but did not pass the requirements to be classified as MAGs (da Rocha et al., 2023). The following paragraphs detail the data for recovering prokaryotic MAGs, high-quality and complete UViGs and eukaryotic MABs. Nevertheless, MuDoGeR can also generate output files containing information on ORFs. The accession numbers of the generated prokaryotic MAGs, high-quality and complete UViGs and high-quality eukaryotic MABs can be found in Tables S3–S5.

3.1 | Recovery of prokaryotes MAGs

MuDoGeR recovered 5726 prokaryotic MAGs encompassing 3850 'species level' OTUs (Average Nucleotide Identity, ANI >0.95) (Table S3) from the 574 metagenomic libraries used in this study. These included 1969 high-quality genomes (>90% completeness and <5% contamination) (Table S3) (Figure 2a). These OTUs belonged to 3644 bacterial and 206 archaeal species, and GTDB-tk classified them over 77 Phyla, 141 Classes, 530 families and 1110 genera (Table S3). The number of prokaryotic MAGs per library ranged from 0 to 111 (average = 9.98, standard deviation = 15.81) (Table S1) (Figure 2d). MuDoGeR recovered no prokaryotic MAGs from approximately 27% (152) libraries. These libraries belonged to environments and materials with an extremely high diversity of microbes (e.g. soils) or samples where host DNA is present in higher yields than the microbial DNA before sequencing

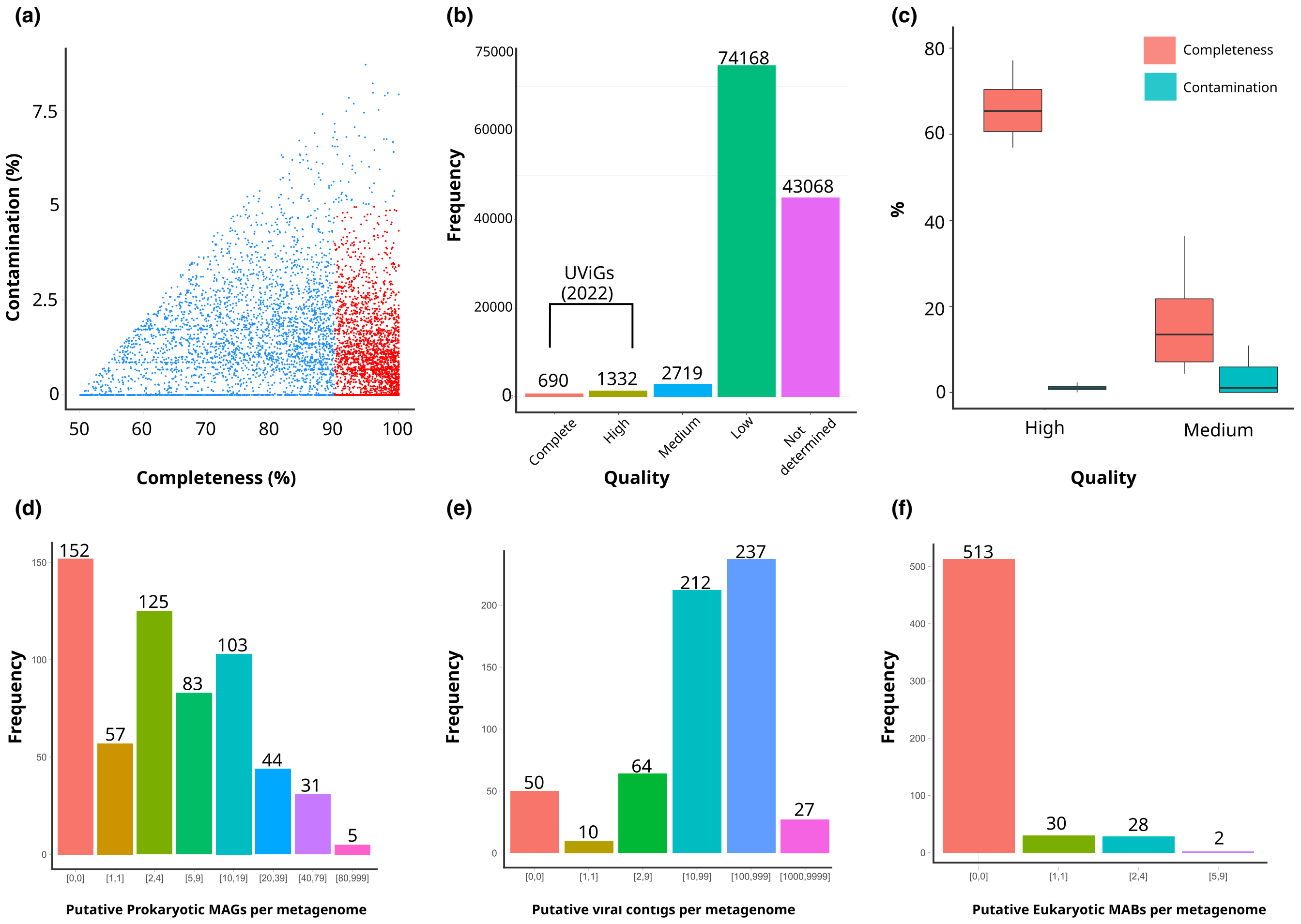

FIGURE 2 Results of prokaryotic and eukaryotic genome and viral contig recovery. (a) Completeness and contamination of prokaryotic metagenome-assembled genomes (MAGs) with a quality score above 50 recovered from all libraries. The completeness and contamination for prokaryotic MAGs were assigned by MuDoGeR using CheckM; (b) Number of complete, high-quality, medium-quality, low-quality and undetermined viral contigs recovered from all libraries; the number on top of the bars indicates the number of viral contigs per quality group. The viral contig quality was assigned by MuDoGeR using CheckV. Viral contigs classified as Complete or High quality are considered uncultivated viral genomes (UViGs); (c) Completeness and contamination boxplots for eukaryotic metagenome-assembled bins (MABs) recovered from all libraries; The eukaryotic MABs quality were assigned by MuDoGeR using EukCC. (d) Frequency in the number of metagenomes recovering different numbers of prokaryotic MAGs; (e) Frequency in the number of metagenomes recovering different numbers of putative viral contigs; (f) Frequency in the number of metagenomes recovering different numbers of MABs. (d–f) Numbers between brackets indicate the intervals of MAGs, putative viral contigs and MABs respectively; numbers on top of the bars indicate the number of metagenomes per interval.

(e.g. endosymbiotic communities). Seven libraries showed more than 70 prokaryotic MAGs. From these, six metagenomes belonged to anaerobic reactors and one to a cyanobacterial mat (Table S3). The relation between the sample's diversity and complexity with the quality of recovered MAGs was previously observed in the work of (Bornemann et al., 2020). They identified that bin quality of low-complexity datasets (e.g. bioreactors) was significantly higher than in medium or high-complexity datasets (e.g. soils).

3.2 | Recovery of UViGs

Our analysis uncovered 121,977 putative viral contigs (Table S4). After dereplication and quality check, 2719 were classified as medium quality and 2022 as high-quality or complete UViGs (Table S4) (Figure 2b). The number of dereplicated UViGs per library ranged from 0 to 166 (average = 8.25 and standard deviation = 15.40) (Table S1) (Figure 2e). We observed at least one high-quality or complete UViG in approximately 85% (493) of the libraries. The libraries that yielded no high-quality or complete UViGs belonged to environments and materials with high microbial diversity (e.g. soils) or samples where host DNA cannot be separated from microbial DNA before sequencing (e.g. endosymbiotic communities). Three libraries showed more than 100 high-quality or complete UViGs. All these belonged to aquatic environments (Table S4). The high-quality and complete dereplicated UViGs belonged to the Order Caudovirales (978 UViGs) and Ligamenvirales (4 UViGs). At the same time, most of them had their orders assigned as unclassified or undetermined (1092) by vConTACT2, indicating the high potential to identify new DNA viruses using MuDoGeR. Such a high number of still unknown or unclassified viruses is indeed not unusual when analysing large (viral) metagenomic datasets (Santos-Medellin et al., 2021; Tisza & Buck, 2021).

It is worth noticing that we used terrestrial metagenome libraries to benchmark MuDoGeR. We would expect prokaryotic viruses as the majority in these libraries since prokaryotes are very abundant in terrestrial environments. Tailed phages from the former order Caudovirales are the most abundant (known) group of prokaryotic viruses. Therefore, one reason for our findings is our dataset's composition of viral communities. Another reason is a database issue. Caudoviruses are over-represented in public databases because most metagenomic protocols exclusively target dsDNA (and >95% of known viral genomes come from metagenomes). In addition, Caudoviruses have been the best-described prokaryotic viruses. It is, therefore, much more likely to assign the identified sequences to this large group while other viral sequences remain unassigned. Furthermore, there is a technical reason. We have implemented Vcontact2 in MuDoGeR. This tool is especially suited to assign phage sequences to a taxonomic rank via similarity analyses of the individually translated ORFs. MuDoGeR could recover eukaryotic viral sequences with VirSorter2 (Guo et al., 2021) and VIBRANT (Kieft et al., 2020). However, given the existing research in the area, its current characterization focuses on bacterial and archaeal viruses. Additionally, the recovered genomes should be readily utilized in other characterization tools. Since MuDoGeR works as a wrapper, it is set up to regularly add updates and new tools, keeping it up-to-date in field.

3.3 | Recovery of Eukaryotic bins

We recovered a total of 52 eukaryotic MABs. Of these, seven showed completeness above 40%, of which five showed less than 5% contamination (Figure 2c, Table S5). As expected, due to the low number of MABs, we recovered no eukaryotes from 544 samples. The maximum number of MABs per sample was 6 (Figure 2f). The eukaryotic MABs showing more than 40% completeness were recovered from metagenomes from water and organic material retrieved from lakes, rivers, cropland and urban samples (Table S5). These bins are distributed over two phyla (Chlorophyta and Heterotrichea) and one species (*Micromonas commoda*) (Table S5). Chlorophyta (i.e. green algae) are commonly found in marine habitats (Leliaert et al., 2012), consistent with the biome from which the samples were retrieved. Challenges in the recovery of eukaryotic MABs stem from the presence of repeat regions (Delmont & Eren, 2016), assembly of genomes using short reads (Pearman et al., 2020), a limited number of reference genomes (Pawlowski et al., 2012) and lack of software for predicting eukaryotic genes in entire metagenomes (Saraiva et al., 2023). Once the scientific community advances in eukaryotic genome reconstruction, we will add them to the MuDoGeR.

3.4 | Biodiversity analysis with MuDoGeR

Although we did not select metagenomes to perform biodiversity analysis with a specific research question in mind, we prepared the output of MuDoGeR to allow biodiversity analysis from prokaryotes, eukaryotes and viruses alone or combined using standard biodiversity analyses with visualization pipelines (e.g. Phyloseq, McMurdie & Holmes, 2013). To indicate its potential, we selected the samples with at least one eukaryotic bin and one medium, high-quality or complete UViG. Using these 15 libraries, we performed Beta diversity analyses using phyloseq to demonstrate the potential for genome-centric biodiversity analysis using MuDoGeR (Figure S3).

3.5 | Comparing MuDoGeR genome-centric diversity with Kraken

To illustrate Kraken microbiome analysis, Lu et al. (2022) applied it to three metagenomic sequencing samples collected from Taur et al. (2018). The samples exemplify a normal state of the patient microbiome and its reduced diversity state after antibiotic treatment. MuDoGeR is an assembly-based wrapper designed to recover genomes from WGS samples, while Kraken is a read-based tool primarily used for assigning taxonomic labels to metagenomic samples. However, the comparison between MuDoGeR and Kraken on clinical metagenomes provided insightful data on how the assembly-based approach fares against the read-based approach, especially if it can capture the diversity of a sample while providing a deeper option for functional potential exploration.

After recovering the genomes from the clinical samples (MuDoGeR modules 1–4), we applied MuDoGeR module 5 to perform a complete read mapping against the recovered sequences (Table S6). We recovered 96 MAGs, 93 UViGs and no eukaryotic genome (the complete workflow results are accessible via the Data Availability section). The MAGs clustered into 55 gOTUs. The main results from Kraken are in their Table S1 (Lu et al. (2022)). Our analyses further highlighted the similarities and differences between the two methods across all samples. As expected, Kraken mapped the reads to a higher number of species per sample 550, 200 and 503, while MuDoGeR mapped to 55, 49 and 55 for the samples SRR14092160, SRR14092310 and SRR14143424 respectively. However, in terms of alpha diversity, the Shannon Index values for the samples SRR14092160, SRR14092310 and SRR14143424 were 3.55333, 0.391483 and 2.45485 for Kraken, and 3.04791, 0.341177 and 2.46034 for MuDoGeR, respectively, showing that MuDoGeR was able to capture the diversity of the samples. Additionally, an analysis of commonly mapped species showed an average relative abundance difference of 1.1%, indicating a core set of species consistently identified by both methods.

Naturally, read-based methods, such as Kraken, classify individual reads without assembling them into contiguous sequences. They are known for their speed and efficiency in taxonomic assignment. However, they are limited in resolution due to the short length of individual reads. Read-based methods may have their accuracy compromised based on the reference database. Comparatively, assembly-based methods like MuDoGeR aim to assemble reads into longer contiguous sequences (contigs) or complete genomes, facilitating a deeper exploration of the microbial community structure and function and enabling a more nuanced understanding of the metagenomic landscape.

3.6 | MuDoGeR as wrapper and its critical use

Our wrapper was designed to streamline the genome assembly process from metagenome samples across multiple domains. It integrates several state-of-the-art genome recovery tools and simplifies the installation and usage of several complex tools and databases. While using a wrapper like MuDoGeR can accelerate metagenomics analysis, it also makes it easier to ignore what is happening behind the scenes to the detriment of the analysis. Understanding the inherent limitations of any metagenomic approach, including ours, is essential. Every approach selected to integrate MuDoGeR currently was made based on benchmark and the original studies, and the default parameters used were selected based on the original tools' recommendations. However, it is always relevant for the user to understand its dataset and tools limitations to adapt the workflow to their hypothesis or research question (da Rocha et al., 2023).

MuDoGeR generates progress reports, intermediate files and error logs at various stages of the assembly process, all detailed in our GitHub documentation. These provide valuable insights into the tool's performance and the genome assembly and annotation progress and can be helpful for troubleshooting. It is important to check these regularly to ensure the tool works as expected for a specific dataset. It is also worth mentioning that, like any current metagenomic approach, our tool may not recover the complete microbial composition of the sample. This limitation should be considered when interpreting the results.

In addition, it is relevant to note that from a biological perspective, the domains are not inherently separable. The separation often observed is mainly due to the technical approaches employed in the analysis. In our study, we have strived to provide a holistic view of all three domains simultaneously, which offers a more comprehensive understanding of microbial composition. Our approach helps to reduce the cross-domain recovery bias as we initiate all genome recovery from the same assembly. This uniform starting point allows us to minimize the discrepancies that often arise when different pipelines are used for different domains. However, genome recovery from the different domains is in different stages of their technological progress and complexity of analysis. For instance, eukaryotic genomes are much more complex to recover than prokaryotic ones. Therefore, the user should be aware of potential recovery bias from a particular dataset.

For more experienced users, MuDoGeR allows the activation of each tool separately. This flexibility enables users to adapt the process to their specific needs, whether that involves dealing with particularly complex samples, limited computational resources, or other unique circumstances. To better understand and optimize the recovery process, we strongly recommend that users consult the MuDoGeR Manual regularly and check the direct links from the tools used within the wrapper, also provided in the MuDoGeR documentation. A better understanding of MuDoGeR structure and design will give users a deeper understanding of the underlying processes and help them optimize the parameters for their specific needs.

4 | CONCLUSIONS

MuDoGeR is a user-friendly tool encompassing state-of-the-art software and pipelines that recover prokaryotic, viral and eukaryotic genomes from metagenomes in combination or individually. It extends to any study using Illumina short-sequence reads. Users can easily install all 20 tools and 50 custom scripts. MuDoGeR generates around 48 comprehensive files and folders containing the summary and parsing of the taxonomic information, quality estimation, genome annotation, coverage and relative abundance calculation of the metagenome-assembled genomes. MuDoGeR requires only five simple commands to generate the output structure for the prokaryotic MAGs, UViGs and eukaryotic MABs. Further, users may use any of the tools found in MuDoGeR by loading their specific Conda environment, creating a flexible pipeline that expert users can adapt.

AUTHOR CONTRIBUTIONS

UR conceptualized, coordinated and supervised all work. RK joined the conceptualization of the viral genome recovery. JPS and PB joined the conceptualization of the eukaryotic genome recovery. PS joined the conceptualization for reaching non-bioinformaticians and the user-friendliness of the pipeline. JCK, RBT, MFB, FBC, LMVB and MNB worked on writing the code for the pipeline and its GitHub page. EF helped with the GitHub and standardization of the metadata from the metagenomes and genomes used in the study. JP and AP tested the tool for its user-friendly aspect and improved the GitHub and code. PFS discussed the conceptualization and the user-friendly aspect of the tool. All authors helped in the writing of the manuscript. The CLUE-TERRA Consortium was instrumental in discussing and designing the potential functionality of MuDoGeR and its user-friendly aspect.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 460129525. UNR, RBT, JPS, MFB and FBC were financed by the Helmholtz Young Investigator grant VH-NG-1248 Micro 'Big Data'. JCK was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [2019/03396-9] and [2022/03534-5]. JP was supported by Lundin Energy Norway AS within the framework of the GeneOil Project. This work also used the resources provided by the German Network for Bioinformatics Infrastructure (de.NBI). Open Access funding enabled and organized by Projekt DEAL.

Alexander Bartholomäus, Alexandre Soares Rosado, Ana-Maria Fiore-Donno, André Carlos Ponce de Leon Ferreira de Carvalho, Cecile Gubry-Rangin, Daniel Machado, Danilo S. Sanches, Dirk Wagner, Gabriele Berg, Ines Mundic Mulec, Marie Muehe, Michael Bonkowski, Newton Gomes, Raquel Peixoto, Rodrigo Costa, Sabine Kleinsteuber, Simonetta Gribaldo, Tina Keller Costa, Victor Pylro, Vivian Pellizari.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

MuDoGeR is open-source software available at https://github.com/mdsufz/MuDoGeR. All MAGs are available on NCBI SRA through https://www.ncbi.nlm.nih.gov/bioproject/PRJNA843551/. All high-quality and complete UViGs are available on the Helmholtz Center for Environmental Research - UFZ cloud services through the link https://nc.ufz.de/s/yFFBrceNoC6P3wY (password: pSNoAafzYB). All seven species-level MABs are available on NCBI SRA under the accessions SAMN26329102, SAMN26244053, SAMN26329113, SAMN26302841, SAMN26302842, SAMN26329290 and SAMN26302904. All the data resulting from MuDoGeR genome recovery from the clinical samples by Taur et al. (2018) can be downloaded at https://www.ufz.de/record/dmp/archive/14253.

ORCID

Ulisses Rocha https://orcid.org/0000-0001-6972-6692
Petr Baldrian https://orcid.org/0000-0002-8983-2721

REFERENCES

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533–538. https://doi.org/10.1038/nbt.2579

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, *29*(12), 2607–2618.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D., & Sullivan, M. B. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology*, *37*(6), 632–639. https://doi.org/10.1038/s41587-019-0100-8

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., & Taylor, J. (2010). Galaxy: A web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, *89*(1), 19.10.1–19.10.21. https://doi.org/10.1002/0471142727.mb1910s89

Bornemann, T. L. V., Esser, S. P., Stach, T. L., Burg, T., & Probst, A. J. (2020). uBin – A manual refining tool for metagenomic bins designed for educational purposes. *bioRxiv*, 2020.07.15.204776. https://doi.org/10.1101/2020.07.15.204776

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2018). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1139. https://doi.org/10.1093/bib/bbx120

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Churcheward, B., Millet, M., Bihouée, A., Fertin, G., & Chaffron, S. (2022). MAGNETO: An automated workflow for genome-resolved metagenomics. *mSystems*, *7*(4), e00432-22. https://doi.org/10.1128/msystems.00432-22

Corrêa, F. B., Saraiva, J. P., Stadler, P. F., & da Rocha, U. N. (2020). TerrestrialMetagenomeDB: A public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Research*, *48*(D1), D626–D632. https://doi.org/10.1093/nar/gkz994

da Rocha, U. N., Kasmanas, J. C., Toscan, R., Sanches, D. S., Magnusdottir, S., & Saraiva, J. P. (2023). Simulation of 69 microbial communities indicates sequencing depth and false positives are major drivers of bias in prokaryotic metagenome-assembled genome recovery. *bioRxiv*, p. 2023.05.02.539054. https://doi.org/10.1101/2023.05.02.539054

Delmont, T. O., & Eren, A. M. (2016). Identifying contamination with advanced visualization and analysis practices: Metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, *4*, e1839. https://doi.org/10.7717/peerj.1839

Dias, O., Saraiva, J., Faria, C., Ramirez, M., Pinto, F., & Rocha, I. (2019). iDS372, a phenotypically reconciled model for the metabolism of Streptococcus pneumoniae strain R6. *Frontiers in Microbiology*, *10*, 1283. https://doi.org/10.3389/fmicb.2019.01283

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ*, *3*, e1319. https://doi.org/10.7717/peerj.1319

Evans, P. N., Parks, D. H., Chadwick, G. L., Robbins, S. J., Orphan, V. J., Golding, S. D., & Tyson, G. W. (2015). Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*, *350*(6259), 434–438. https://doi.org/10.1126/science.aac7745

Galiez, C., Siebert, M., Enault, F., Vincent, J., & Söding, J. (2017). WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, *33*(19), 3113–3114. https://doi.org/10.1093/bioinformatics/btx383

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 476. https://doi.org/10.1038/s41592-018-0046-7

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, *9*(1), 37. https://doi.org/10.1186/s40168-020-00990-y

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. https://doi.org/10.1186/1471-2105-12-491

Kallies, R., Hölzer, M., Brizola Toscan, R., Nunes da Rocha, U., Anders, J., Marz, M., & Chatzinotas, A. (2019). Evaluation of sequencing library preparation protocols for viral metagenomic analysis from pristine aquifer groundwaters. *Viruses*, *11*(6), 484. https://doi.org/10.3390/v11060484

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *3*, e1165. https://doi.org/10.7717/peerj.1165

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., von Bergen, M., Stadler, P. F., Carvalho, A. C. P. L. F., & Nunes da Rocha, U. (2021). HumanMetagenomeDB: A public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research*, *49*(D1), D743–D750. https://doi.org/10.1093/nar/gkaa1031

Keller-Costa, T., Lago-Lestón, A., Saraiva, J. P., Toscan, R., Silva, S. G., Gonçalves, J., Cox, C. J., Kyrpides, N., Nunes da Rocha, U., & Costa, R. (2021). Metagenomic insights into the taxonomy, function, and dysbiosis of prokaryotic communities in octocorals. *Microbiome*, *9*(1), 72. https://doi.org/10.1186/s40168-021-01031-y

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, *8*(1), 90. https://doi.org/10.1186/s40168-020-00867-0

Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M., & McCue, L. A. (2020). ATLAS: A Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics*, *21*(1), 257. https://doi.org/10.1186/s12859-020-03585-4

Koonin, E. V., Krupovic, M., & Dolja, V. V. (2023). The global virome: How much diversity and how many independent origins? *Environmental Microbiology*, *25*(1), 40–44. https://doi.org/10.1111/1462-2920.16207

Koren, S., Treangen, T. J., & Pop, M. (2011). Bambus 2: Scaffolding metagenomes. *Bioinformatics*, *27*(21), 2964–2971. https://doi.org/10.1093/bioinformatics/btr520

Krakau, S., Straub, D., Gourlé, H., Gabernet, G., & Nahnsen, S. (2022). Nf-core/mag: A best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics and Bioinformatics*, *4*(1), lqac007. https://doi.org/10.1093/nargab/lqac007

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS One*, *12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., & De Clerck, O. (2012). Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences*, *31*(1), 1–46. https://doi.org/10.1080/07352689.2011.615705

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods (San Diego, Calif.)*, *102*, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020

Liu, B., Sträuber, H., Saraiva, J., Harms, H., Silva, S. G., Kasmanas, J. C., Kleinsteuber, S., & Nunes da Rocha, U. (2022). Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. *Microbiome*, *10*(1), 48. https://doi.org/10.1186/s40168-021-01219-2

López-Mondéjar, R., Tláskal, V., Větrovský, T., Štursová, M., Toscan, R., Nunes da Rocha, U., & Baldrian, P. (2020). Metagenomics and stable isotope probing reveal the complementary contribution of fungal and bacterial communities in the recycling of dead biomass in forest soil. *Soil Biology and Biochemistry*, *148*, 107875. https://doi.org/10.1016/j.soilbio.2020.107875

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L., & Steinegger, M. (2022). Metagenome analysis using the kraken software suite. *Nature Protocols*, *17*(12), 2839. https://doi.org/10.1038/s41596-022-00738-y

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., Hill, B., Brito, J., Littman, R. J., Statz, B., Lam, A. K.-M., Dayama, G., Grieneisen, L., Martin, L. S., Flint, J., Eskin, E., & Blekhman, R. (2019). Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS Biology*, *17*(6), e3000333. https://doi.org/10.1371/journal.pbio.3000333

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Melkonian, C., Fillinger, L., Atashgahi, S., da Rocha, U. N., Kuiper, E., Olivier, B., Braster, M., Gottstein, W., Helmus, R., Parsons, J. R., Smidt, H., van der Waals, M., Gerritse, J., Brandt, B. W., Röling, W. F. M., Molenaar, D., & van Spanning, R. J. M. (2021). High biodiversity in a benzene-degrading nitrate-reducing culture is sustained by a few primary consumers. *Communications Biology*, *4*(1), 530. https://doi.org/10.1038/s42003-021-01948-y

Meyer, F., Lesker, T.-R., Koslicki, D., Fritz, A., Gurevich, A., Darling, A. E., Sczyrba, A., Bremges, A., & McHardy, A. C. (2021). Tutorial: Assessing metagenomics software with the CAMI benchmarking toolkit. *Nature Protocols*, *16*(4), 1801. https://doi.org/10.1038/s41596-020-00480-3

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: An extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155. https://doi.org/10.1093/nar/gks678

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Noor, E., Eden, E., Milo, R., & Alon, U. (2010). Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Molecular Cell*, *39*(5), 809–820. https://doi.org/10.1016/j.molcel.2010.08.031

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Oliveira Monteiro, L. M., Saraiva, J. P., Brizola Toscan, R., Stadler, P. F., Silva-Rocha, R., & Nunes da Rocha, U. (2022). PredicTF: Prediction of bacterial transcription factors in complex microbial communities using deep learning. *Environmental Microbiomes*, *17*(1), 7. https://doi.org/10.1186/s40793-021-00394-x

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., … de Vargas, C. (2012). CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, *10*(11), e1001419. https://doi.org/10.1371/journal.pbio.1001419

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*, *21*(1), 220. https://doi.org/10.1186/s12859-020-3528-4

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*(1), 69. https://doi.org/10.1186/s40168-017-0283-5

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019). Minimum information about an uncultivated virus genome (MIUViG). *Nature Biotechnology*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Saary, P., Mitchell, A. L., & Finn, R. D. (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biology*, *21*(1), 244. https://doi.org/10.1186/s13059-020-02155-4

Santos-Medellin, C., Zinke, L. A., ter Horst, A. M., Gelardi, D. L., Parikh, S. J., & Emerson, J. B. (2021). Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *The ISME Journal*, *15*(7), 1956–1970. https://doi.org/10.1038/s41396-021-00897-y

Saraiva, J. P., Bartholomäus, A., Kallies, R., Gomes, M., Bicalho, M., Coelho Kasmanas, J., Vogt, C., Chatzinotas, A., Stadler, P., Dias, O., & Nunes da Rocha, U. (2021). OrtSuite: From genomes to prediction of microbial interactions within targeted ecosystem processes. *Life Science Alliance*, *4*(12), e202101167. https://doi.org/10.26508/lsa.202101167

Saraiva, J. P., Bartholomäus, A., Toscan, R. B., Baldrian, P., & Nunes da Rocha, U. (2023). Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes. *Molecular Ecology Resources*, *23*(5), 1066–1076. https://doi.org/10.1111/1755-0998.13776

Saraiva, J. P., Worrich, A., Karakoç, C., Kallies, R., Chatzinotas, A., Centler, F., & Nunes da Rocha, U. (2021). Mining synergistic microbial interactions: A roadmap on how to integrate multi-omics data. *Microorganisms*, *9*(4), 840. https://doi.org/10.3390/microorganisms9040840

Saraiva, J. P. L. F., Zubiria-Barrera, C., Klassert, T. E., Lautenbach, M. J., Blaess, M., Claus, R. A., Slevogt, H., & König, R. (2017). Combination of classifiers identifies fungal-specific activation of lysosome genes in human monocytes. *Frontiers in Microbiology*, *8*, 2366. https://doi.org/10.3389/fmicb.2017.02366

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., & Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, *23*(1), 111–120. https://doi.org/10.1101/gr.142315.112

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1

Taur, Y., Coyte, K., Schluter, J., Robilotti, E., Figueroa, C., Gjonbalaj, M., Littmann, E. R., Ling, L., Miller, L., Gyaltshen, Y., Fontana, E., Morjaria, S., Gyurkocza, B., Perales, M.-A., Castro-Malaspina, H., Tamari, R., Ponce, D., Koehne, G., Barker, J., … Xavier, J. B. (2018). Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant. *Science Translational Medicine*, *10*(460), eaap9489. https://doi.org/10.1126/scitranslmed.aap9489

Tisza, M. J., & Buck, C. B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(23), e2023202118. https://doi.org/10.1073/pnas.2023202118

Tláskal, V., Brabcová, V., Větrovský, T., Jomura, M., López-Mondéjar, R., Monteiro, L. M. O., Saraiva, J. P., Human, Z. R., Cajthaml, T., da Rocha, U. N., & Baldrian, P. (2021). Complementary roles of Wood-inhabiting fungi and bacteria facilitate deadwood decomposition. *mSystems*, *6*(1), e01078-20. https://doi.org/10.1128/mSystems.01078-20

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. https://doi.org/10.1038/nature02340

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—A flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*, 543–548. https://doi.org/10.1093/molbev/msx319

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*(4), 569–580. https://doi.org/10.1101/gr.228429.117

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics (Oxford, England)*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rocha, U., Coelho Kasmanas, J., Kallies, R., Saraiva, J. P., Toscan, R. B., Štefanič, P., Bicalho, M. F., Borim Correa, F., Baştürk, M. N., Fousekis, E., Viana Barbosa, L. M., Plewka, J., Probst, A. J., Baldrian, P., Stadler, P. F., & (2024). MuDoGeR: Multi-Domain Genome recovery from metagenomes made easy. *Molecular Ecology Resources*, *24*, e13904. https://doi.org/10.1111/1755-0998.13904