MuDoGeR: Multi-Domain Genome Recovery from metagenomes made easy

Ulisses Nunes da 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 Probst[4], Petr Baldrian[5], Peter Stadler[6], 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, 45141, Essen, Germany.
[5]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Videnska 1083, Praha 4, 14220, Czech Republic.
[6]Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany.

*corresponding author: Ulisses Nunes da Rocha

Ulisses Nunes da Rocha:	ulisses.rocha@ufz.de
Jonas Coelho Kasmanas:	jonas.kasmanas@ufz.de
René Kallies:	rene.kallies@ufz.de
Joao Pedro Saraiva:	joao.saraiva@ufz.de
Rodolfo Brizola Toscan:	ta.bioinf@gmail.com
Polonca Štefanič:	polonca.stefanic@bf.uni-lj.si
Marcos Fleming Bicalho:	marcos9bicalho@gmail.com
Felipe Borim Correa:	felipe.correa@ufz.de
Merve Nida Baştürk:	mnidabasturk@gmail.com
Efthymios Fousekis:	efousekis@gmail.com
Luiz Miguel Viana Barbosa:	lmv.barbosa@gmail.com
Julia Plewka:	julia.plewka@uni-due.de
Petr Baldrian:	baldrian@biomed.cas.cz
Peter Stadler:	studla@bioinf.uni-leipzig.de

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

Abstract

Several frameworks that recover genomes from Prokaryotes, Eukaryotes, and viruses from metagenomes exist. For those with little bioinformatics experience, it is difficult to evaluate quality, annotate genes, dereplicate, assign taxonomy and calculate relative abundance and coverage from genomes belonging to different domains. MuDoGeR is a user-friendly tool accessible for non-bioinformaticians that make genome recovery from metagenomes of Prokaryotes, Eukaryotes, and viruses alone or in combination easy. By testing MuDoGeR using 574 metagenomes and 24 genomes, we demonstrated users could run it in a few samples or high-throughput. MuDoGeR is an open-source software available at https://github.com/mdsufz/MuDoGeR.

Background

Metagenomics encompasses the concepts and techniques used to study genetic material recovered directly from microbial communities [1–8]. Even if only partially successful, high-throughput sequencing reads have been used to reassemble genomes since 2004 [9]. In 2013, differential coverage binning was used in the first open-source tool for genome-resolved metagenomics [10]. Since then, the bioinformatics community has developed different frameworks specific for genome recovery in Prokaryotes [11–13], Eukaryotes [13–15], and viruses [13,16–19]. Genome-centric metagenomics allows the simultaneous exploration of individual species' functional potential and phylogeny in a microbial community [20–23]. Although software developers created several genome-centric tools, most require users with expertise in bioinformatics or computational biology to understand the overall pipeline, install the different dependencies, and run a complex framework that is often written in various computer languages, and understand the output. To extend the use of genome-centric metagenomics to those with entry level expertise in bioinformatics and assure the data generated can be reused, the next steps in software development in the field need to guarantee: (i) the tools can be used, installed, and maintained by non-experts in computational biology and bioinformatics [24]; and, (ii) the data generated by any tool follows the FAIR guiding principles [25].

Here we present MuDoGeR, 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 non-experts. 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.

Results and discussion

We tested the MuDoGeR pipeline using 598 Illumina short reads found in public repositories (Supplementary Table 1). We selected these libraries to encompass metagenomes (574) and genomes (24). We selected the libraries from axenic genomes to test if our pipeline could recover viruses in individual isolates (Supplementary Table 2). We did not test our pipeline in standard microbial communities as, to the best of our knowledge, a mock-community standard containing Prokaryotes, Eukaryotes, and viruses did not exist when we performed our analysis. Only 8 of the 574 metagenomic libraries showed no recovery of Prokaryotic metagenome-assembled genomes (MAGs), putative viral contigs, or Eukaryotic metagenome-assembled bins (MABs) (Supplementary Table 1). In the following paragraphs, we detailed the data for recovering Prokaryotic MAGs, high-quality and complete uncultivated viral genomes (UViGs), and Eukaryotic MABs. Nevertheless, MuDoGeR can also generate output files for Open Reading Frames (ORFs). The accession numbers of the generated Prokaryotic MAGs, high-quality and complete UViGs, and high-quality Eukaryotic MABs can be found in Supplementary Tables 3, 4, and 5.

Recovery of Prokaryotes MAGs. MuDoGeR recovered 5726 Prokaryotic MAGs divided into 3850 'species level' OTUs (ANI >0.95) (Supplementary Table 3) from the 574 metagenomic libraries used in this study, from which 1969 were high-quality genomes (>90% completeness and <5% contamination) (Supplementary Table 3) (Fig. 1A). These OTUs belonged to 3644 Bacteria and 206 Archaea species, and GTDB-tk distributed them over 77 Phyla, 141 Classes, 530 families, and 1110 genera (Supplementary Table 3). The number of Prokaryotic MAGs per library ranged from 0 to 111 (average=9.98, standard deviation=15.81) (Supplementary Table 1) (Fig. 1D). MuDoGeR recovered no MAGs from approximately 23% (132) 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 (e.g., endosymbiotic communities). Only seven libraries showed more than 70 Prokaryotic MAGs. From these, six metagenomes belonged to anaerobic reactors and one to a cyanobacterial mat (Supplementary Table 3).

Recovery of UViGs. Our analysis uncovered 121 977 putative viral contigs (Supplementary Table 4). After dereplication and quality check, 2719 were classified as medium quality and 2074 as high-quality or complete UViGs (Supplementary Table 4) (Fig. 1B). The number of dereplicated UViGs per library ranged from 0 to 166 (average=8.25 and standard deviation=15.40) (Supplementary Table 1) (Fig. 1E). We observed no high-quality or complete UViG in approximately 15% (87) of the libraries. These libraries 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). Only three libraries showed more than 100 high-quality or complete UVIGs. All of these belonged to aquatic environments (Supplementary Table 4). The high-quality and complete dereplicated UViGs belonged to the Order Caudovirales 978) and Ligamenvirales (4). 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.

Recovery of Eukaryotic bins. We recovered a total of 52 Eukaryotic MABs. Of these, only 7 showed completeness above 40%, of which only 5 showed less than 5% contamination (Fig. 1C, Supplementary Table 5). As expected, due to the low number of MABs, we recovered no Eukaryotes from most samples (544), and the maximum number of MABs per sample was 6 (Fig. 1F). The Eukaryotic MABs showing more than 40% completeness were recovered from metagenomes from water and organic material retrieved from lake, river, cropland, and urban samples (Supplementary Table 5). These bins are distributed over 2 Phyla (Chlorophyta and Heterotrichea) and one species (Micromonas commoda) (Supplementary Table 5). Chlorophyta (i.e., green algae) are commonly found in marine habitats [26], which is consistent with the biome in which the samples were isolated. Challenges in the recovery of Eukaryotic MABs stem from the presence of repeat regions [27], assembly of genomes using short reads [28], and a limited number of reference genomes [29]. Once the scientific community makes advances in Eukaryotic genome reconstruction, we will add them to the MuDoGeR.

Biodiversity analysis with the outputs of 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 analysis and visualization pipeline (e.g., Phyloseq [30]). To indicate its potential, we selected the samples with at least one Eukaryotic bin and one medium, high-quality or complete UViG. Using these 18 libraries, we performed Beta diversity using phyloseq to demonstrate the potential for genome-centric biodiversity analysis using MuDoGeR (Supplementary Figure 1).

Conclusions

MuDoGeR is a user-friendly tool encompassing state-of-the-art software and pipelines able to recover Prokaryotic, viruses, 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 taxonomical information, quality estimation, gene annotation, coverage, and relative abundance calculation from the recovered Prokaryotic MAGs, UViGs, and Eukaryotic MABs using five simple commands. 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.

Methods

MuDoGeR was divided into five modules (Figure 2A). Module 1 deals with the pre-processing of the raw sequences. After module 1, MuDoGeR is divided into three branches (Modules 2-4). Module 2 can be used to recover Prokaryotic MAGs. Module 3 was assembled to generate UViGs, and Module 4 can retrieve Eukaryotic MABs. We recover Eukaryotic MABs as Eukaryotes usually need a much higher coverage to assemble MAGs, allowing those interested in Eukaryotes to study even partial Eukaryotic genomes. 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 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. The MuDoGeR framework is a wrapper of more than 20 tools and 50 custom scripts written using Shell, Python, and R. It was designed to be an easy-to-use tool that outputs ready-to-use comprehensive files (Fig. 2B). 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 2B).

MuDoGeR can be easily installed. Part of the effort to make MAGs recovery "easy" involves making the complex bioinformatic tools easily accessible, requiring little user input. A MAG/UViG/MAB recovery pipeline requires several tools that have multiple dependencies. More often than not, those tools have conflicting dependencies. Consequently, making those tools work in tandem 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 [31] to create multiple environments automatically orchestrated during the pipeline's functioning. The installation protocol creates and automatically sets up 24 independent environments.

In addition, several MAG/UViG/MAB recovery tools require specific databases to work. Knowing the correct database 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.

MuDoGeR v1.0 at a glance. The only mandatory inputs for MuDoGeR are the paths to the samples' raw sequencing reads. Following, the user can use one command to run each complete module. Module 1 starts by using the procedure implemented in metaWRAP [12] to quality control the raw sequences. Following this, we created a regression model to estimate the Random Access Memory (RAM) required to assemble metagenomes with metaSPAdes [32] using the selected 574 metagenomic

libraries. From these, 558 (97.2%) fell in the predicted range (Supplementary Figure 2). The total RAM is highly correlated with chimer frequencies (Supplementary Figure 3).

After, MuDoGeR uses metaSPAdes for assembly. Module 2 integrates the prokaryotic binning procedure implemented in metaWrap using Metabat2 [33], Maxbin2 [34], and CONCOCT [35] binning tools. To estimate quality and taxonomy, MuDoGeR uses CheckM [36] and GTDB-tk [37]. Further, Prokka [38] is used to annotate ORFs, and BBtools (sourceforge.net/projects/bbmap/) to enrich sequence metrics calculation from the Prokaryotic MABs. Finally, MuDoGeR summarizes all the outputs and selects Prokaryotic MAGs as defined by Parks (2018) [39].

The recovery of UViGs, performed in module 3, starts by integrating the viral sequence recovery tools VirSorter2 [16], VirFinder [17], and VIBRANT [18]. Later, the potential viral sequences are dereplicated using Stampede-clustergenomes (https://bitbucket.org/MAVERICLab/stampede-clustergenomes/src/master/). After, the potential viral sequences are annotated using Vcontact2 [40], and quality is performed using CheckV [41]. MuDoGeR also uses WiSH [42] to predict the potential prokaryotic hosts from the recovered viral sequences by integrating the results from modules 2 and 3. Finally, MuDoGeR compiles the outputs from the tools used and selects high-quality and complete UViGs as indicated by CheckV developers [41].

In Module 4, MuDoGeR integrates EukRep [14] for selecting the eukaryotic contigs from the initial assembly, the CONCOCT binning tool, GeneMark [43] for prediction of eukaryotic genes, EukCC [44] for quality estimation from eukaryotic sequences, MAKER2 [45] for gene annotation, and BUSCO [46] for detection of single-copy orthologous genes. Lastly, in module 5, MuDoGeR groups the recovered MAGs into operational 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 may happen as large fractions of the genome are composed of central metabolism and

conserved across most microbial species [47]. That means that even when a species might not be found in a sample, it may be possible to map genes present in that genome from other species in that library. Later, it calculates the relative abundance and coverage tables from the mapped MAGs/UViGs/MABs and annotated prokaryotic genes within the assembled sequences. The user should have standardized results from a complete MAG pipeline by the end.

Finally, we designed MuDoGeR as a wrapper for several complex tools. It also makes all integrated software available for the user independently. This means that a more experienced user could integrate only pieces of MuDoGeR to their pipeline or even access a specific tool environment configured by MuDoGeR, and use only the selected tool. Users can find a tutorial on how to activate 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.

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 your data and its sequencing depth. For this reason, and to reduce the over/underuse of RAM, MuDoGeR attempts to calculate the amount of memory necessary for metaSPades.

The complete software installation requires approximately 170 GB, but MAKER2, from Module 4 uses 99 GB. The complete database requirements, considering all tools, are currently around 439.9 GB. In addition, it is recommended that the user provides multiple processing units since the analysis of several metagenomes simultaneously may require significant time. Consequently, the MuDoGeR Conda installation procedure allows it to be installed on high-performance computers 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.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

MuDoGeR is an 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.

Competing interests

The authors declare that they have no competing interests.

Funding

U.N.R, R.B.T., J.P.S, M.F.B. and F.B.C. were financed by the Helmholtz Young Investigator grant VH-NG-1248 Micro 'Big Data'. J.C.K was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [2019/03396-9]. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460129525.

Authors' contributions

UNR 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 of reaching non-bioinformaticians and user-friendliness of the pipeline. JCK, RBT, MFB, FBC, LMVB and MNB worked on writing the code for the pipeline and the GitHub. EF helped with the GitHub and standardization of the metadata from the metagenomes and genomes used in the study. JP tested the tool for its user-friendly aspect and corrected bugs in 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.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460129525 and the Helmholtz Young Investigator grant VH-NG-1248 Micro 'Big Data'. CLUETERRA consortium: Alexander Bartholomäus, Alexandre Soares Rosado, Alexander Probst, Ana-Maria Fiore-Donno, 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, Simonetta Gribaldo, Tina Keller Costa

References

1. Wooley JC, Ye Y. Metagenomics: Facts and Artifacts, and Computational Challenges*. J Comput Sci Technol. 2009;25:71–81.

2. Keller-Costa T, Lago-Lestón A, Saraiva JP, Toscan R, Silva SG, Gonçalves J, et al. Metagenomic insights into the taxonomy, function, and dysbiosis of prokaryotic communities in octocorals. Microbiome. 2021;9:72.

3. Tláskal V, Brabcová V, Větrovský T, López-Mondéjar R, Monteiro LMO, Saraiva JP, et al. Metagenomes, metatranscriptomes and microbiomes of naturally decomposing deadwood. Scientific Data [Internet]. 2021;8. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85111986556&doi=10.1038%2fs41597-021-00987-8&partnerID=40&md5=a94efb189b59b299bbce51edac093447

4. Oliveira Monteiro LM, Saraiva JP, Brizola Toscan R, Stadler PF, Silva-Rocha R, Nunes da Rocha U. PredicTF: prediction of bacterial transcription factors in complex microbial communities using deep learning. Environmental Microbiome. 2022;17:7.

5. Saraiva JP, Worrich A, Karakoç C, Kallies R, Chatzinotas A, Centler F, et al. Mining Synergistic Microbial Interactions: A Roadmap on How to Integrate Multi-Omics Data. Microorganisms. Multidisciplinary Digital Publishing Institute; 2021;9:840.

6. López-Mondéjar R, Tláskal V, Větrovský T, Štursová M, Toscan R, Nunes da Rocha U, et al. 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. 2020;148:107875.

7. Dias O, Saraiva J, Faria C, Ramirez M, Pinto F, Rocha I. iDS372, a Phenotypically Reconciled Model for the Metabolism of Streptococcus pneumoniae Strain R6. Frontiers in Microbiology. 2019;10:1283.

8. Saraiva JP, Oswald M, Biering A, Röll D, Assmann C, Klassert T, et al. Fungal biomarker discovery by integration of classifiers. BMC genomics. 2017;18:601.

9. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428:37–43.

10. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.

11. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/02/11/107789.abstract

12. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6:158.

13. Corrêa FB, Saraiva JP, Stadler PF, da Rocha UN. TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. Nucleic Acids Research. 2020;48:D626–32.

14. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Res. 2018;28:569–80.

15. Kasmanas JC, Bartholomäus A, Corrêa FB, Tal T, Jehmlich N, Herberth G, et al. HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Res. 2021;49:D743–50.

16. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9:37.

17. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.

18. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8:90.

19. Kallies R, Hölzer M, Brizola Toscan R, Nunes da Rocha U, Anders J, Marz M, et al. Evaluation of Sequencing Library Preparation Protocols for Viral Metagenomic Analysis from Pristine Aquifer Groundwaters. Viruses. 2019;11:484.

20. Evans Paul N., Parks Donovan H., Chadwick Grayson L., Robbins Steven J., Orphan Victoria J., Golding Suzanne D., et al. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. Science. American Association for the Advancement of Science; 2015;350:434–8.

21. Saraiva JP, Bartholomäus A, Kallies R, Gomes M, Bicalho M, Coelho Kasmanas J, et al. OrtSuite: from genomes to prediction of microbial interactions within targeted ecosystem processes. Life Sci Alliance. 2021;4:e202101167.

22. Tláskal V, Brabcová V, Větrovský T, Jomura M, López-Mondéjar R, Monteiro LMO, et al. Complementary Roles of Wood-Inhabiting Fungi and Bacteria Facilitate Deadwood Decomposition. mSystems [Internet]. American Society for Microbiology Journals; 2021 [cited 2021 Jan 23];6. Available from: https://msystems.asm.org/content/6/1/e01078-20

23. Saraiva JPLF, Zubiria-Barrera C, Klassert TE, Lautenbach MJ, Blaess M, Claus RA, et al. Combination of classifiers identifies fungal-specific activation of lysosome genes in human monocytes. Frontiers in Microbiology. 2017;8.

24. Mangul S, Mosqueiro T, Abdill RJ, Duong D, Mitchell K, Sarwal V, et al. Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLOS Biology. Public Library of Science; 2019;17:e3000333.

25. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. Nature Publishing Group; 2016;3:160018.

26. Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, et al. Phylogeny and Molecular Evolution of the Green Algae. Critical Reviews in Plant Sciences. Taylor & Francis; 2012;31:1–46.

27. Delmont TO, Eren AM. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. PeerJ. 2016;4:e1839.

28. Pearman WS, Freed NE, Silander OK. Testing the advantages and disadvantages of short- and long-read eukaryotic metagenomics using simulated reads. BMC Bioinformatics. 2020;21:220.

29. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, et al. CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLOS Biology. Public Library of Science; 2012;10:e1001419.

30. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE. 2013;8:e61217.

31. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. Nature Publishing Group; 2018;15:475–6.

32. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.

33. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.

34. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605–7.

35. Alneberg J, Bjarnason BS, Bruijn I de, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nature Methods. 2014;11:1144–6.

36. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.

37. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020;36:1925–7.

38. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

39. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.

40. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. Nature Publishing Group; 2019;37:632–9.

41. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. Nature Publishing Group; 2021;39:578–85.

42. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 2017;33:3113–4.

43. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29:2607–18.

44. Saary P, Mitchell AL, Finn RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. Genome Biology. 2020;21:244.

45. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.

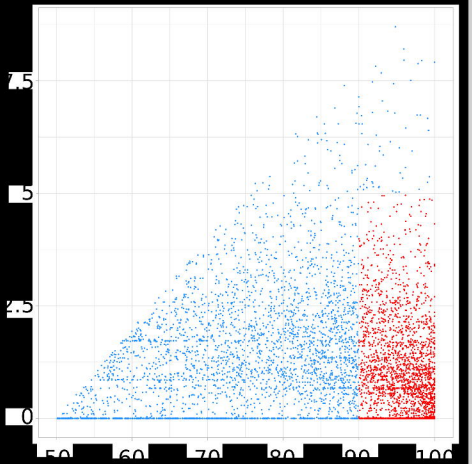
46. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017;

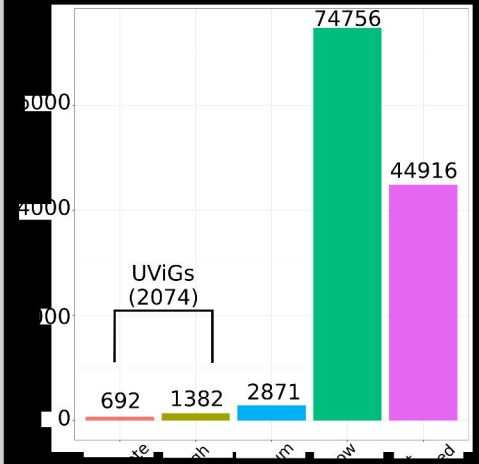
47. Noor E, Eden E, Milo R, Alon U. Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. Molecular Cell. 2010;39:809–20.

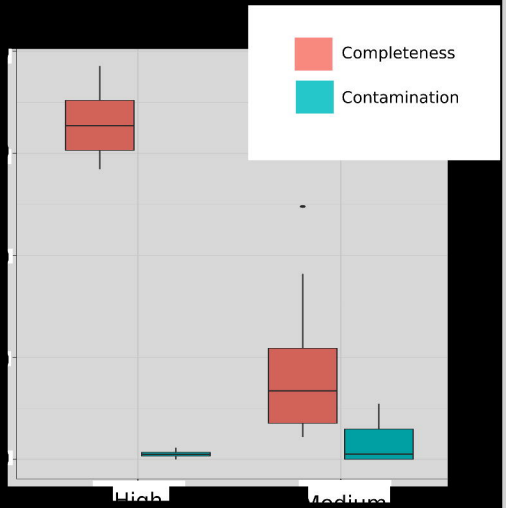
Figure Legends

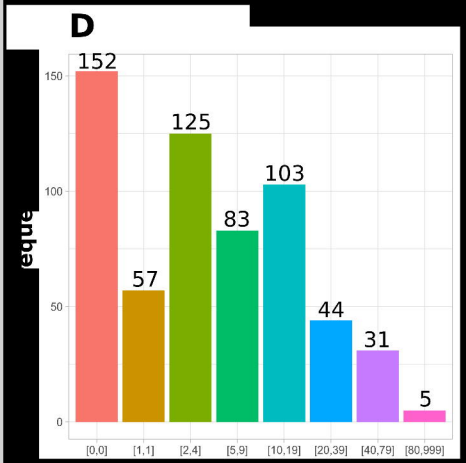
Figure 1. Results of prokaryotic and eukaryotic genome and viral contig recovery. A – Completeness and contamination of prokaryotic metagenome-assembled genomes (MAGs) with a quality score above 50 recovered from all libraries; **B –** Number of complete, high-quality, medium-quality, low-quality and undetermined viral contigs (UViGs) recovered from all libraries; number on top of the bars indicate the number of UViGs per quality group.; **C –** Completeness and contamination boxplots for eukaryotic metagenome-assembled bins (MABs) recovered from all libraries; **D –** Frequency in the number of metagenomes recovering different number of MAGs; **E -** Frequency in the number of metagenomes recovering different number of UViGs; **F -** Frequency in the number of metagenomes recovering different number of MABs. **D**, **E**, **F** numbers between brackets indicate the intervals of MAGs, UViGs and MABs, respectively; numbers on top of the bars indicate the number of metagenomes per interval.

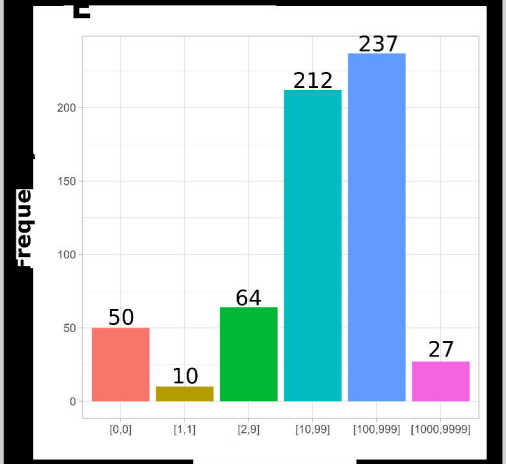
Figure 2. MuDoGeR workflow and outputs. A - MuDoGeR modular workflow. Module 1 – Pre-processing; Module 2 – Recovery of prokaryotic metagenome-assembled genomes (MAGs); Module 3 – Recovery of Uncultivated viral contigs (UViGs); Module 4 – Recovery of Eukaryotic metagenome-assembled bins (MABs); Module 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, 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 (OUT) coverage, relative abundances, and gene abundance.

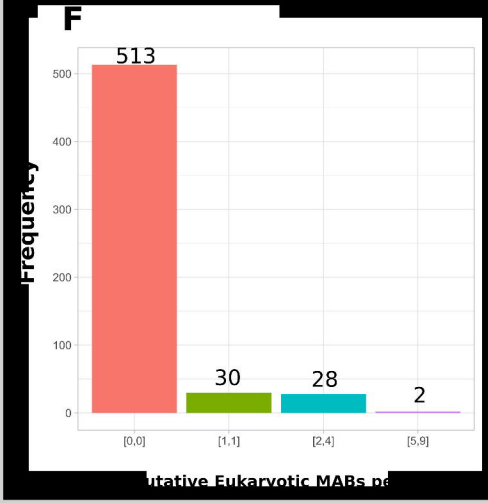



74756
44916
UViGs
(2074)
692
1382
2871


Completeness
Contamination


D
152
57
125
83
103
44
31
5
[0,0]
[1,1]
[2,4]
[5,9]
[10,19]
[20,39]
[40,79]
[80,999]


50
10
64
212
237
27
[0,0]
[1,1]
[2,9]
[10,99]
[100,999]
[1000,9999]


F
513
30
28
2
[0,0]
[1,1]
[2,4]
[5,9]

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

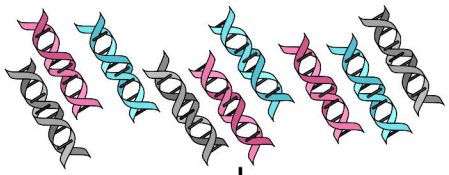
Module 5. Relative abundance and genome coverage table

Prokaryotic MAGs representatives

Viral genomes representatives

Eukaryotic MABs representatives

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