SnakeMAGs: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes

Nachida Tadrent, Franck Dedeine, Vincent Hervé

SOFTWARE TOOL ARTICLE

REVISED *SnakeMAGs*: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes [version 2; peer review: 2 approved]

Nachida Tadrent [1], Franck Dedeine [1], Vincent Hervé [1,2]

[1]Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS-Université de Tours, Tours, 37200, France
[2]Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, Palaiseau, 91120, France

v2 **First published:** 15 Dec 2022, **11**:1522
https://doi.org/10.12688/f1000research.128091.1
Latest published: 27 Feb 2023, **11**:1522
https://doi.org/10.12688/f1000research.128091.2

Abstract

Background: Over the last decade, we have observed in microbial ecology a transition from gene-centric to genome-centric analyses. Indeed, the advent of metagenomics combined with binning methods, single-cell genome sequencing as well as high-throughput cultivation methods have contributed to the continuing and exponential increase of available prokaryotic genomes, which in turn has favored the exploration of microbial metabolisms. In the case of metagenomics, data processing, from raw reads to genome reconstruction, involves various steps and software which can represent a major technical obstacle.

Methods: To overcome this challenge, we developed *SnakeMAGs*, a simple workflow that can process Illumina data, from raw reads to metagenome-assembled genomes (MAGs) classification and relative abundance estimate. It integrates state-of-the-art bioinformatic tools to sequentially perform: quality control of the reads (illumina-utils, Trimmomatic), host sequence removal (optional step, using Bowtie2), assembly (MEGAHIT), binning (MetaBAT2), quality filtering of the bins (CheckM, GUNC), classification of the MAGs (GTDB-Tk) and estimate of their relative abundance (CoverM). Developed with the popular Snakemake workflow management system, it can be deployed on various architectures, from single to multicore and from workstation to computer clusters and grids. It is also flexible since users can easily change parameters and/or add new rules.

Results: Using termite gut metagenomic datasets, we showed that *SnakeMAGs* is slower but allowed the recovery of more MAGs encompassing more diverse phyla compared to another similar workflow named ATLAS. Importantly, these additional MAGs showed no significant difference compared to the other ones in terms of completeness, contamination, genome size nor relative abundance.

Conclusions: Overall, it should make the reconstruction of MAGs more accessible to microbiologists. *SnakeMAGs* as well as test files and

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 27 Feb 2023	✓ view	
version 1 15 Dec 2022	? view	✓ view

1. **Célio Dias Santos Júnior**, Fudan University, Shanghai, China

2. **Aram Mikaelyan**, North Carolina State University, Raleigh, USA

Any reports and responses or comments on the article can be found at the end of the article.

an extended tutorial are available at
https://github.com/Nachida08/SnakeMAGs.

Keywords
Snakemake, metagenomics, microbiology, genomics, bioinformatics, microbial ecology

This article is included in the Bioinformatics gateway.

Corresponding author: Vincent Hervé (vincent.herve@inrae.fr)

Author roles: Tadrent N: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Review & Editing; **Dedeine F**: Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Hervé V**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the FEDER grant InFoBioS n°EX011185 (région Val de Loire, France), the Centre National de la Recherche Scientifique (CNRS) and the Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE). Ph.D. work of Nachida Tadrent (Ph.D. scholarship) was funded by the University of Tours.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Tadrent N, Dedeine F and Hervé V. ***SnakeMAGs*: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes [version 2; peer review: 2 approved]** F1000Research 2023, **11**:1522 https://doi.org/10.12688/f1000research.128091.2

First published: 15 Dec 2022, **11**:1522 https://doi.org/10.12688/f1000research.128091.1

REVISED **Amendments from Version 1**

Following the comments from the two reviewers, the authors have made the following update:

i) We released a new version of our workflow (v1.1.0) that now integrates GUNC, a software for detection of chimerism and contamination in prokaryotic genomes.
ii) We showed that the MAGs recovered by *SnakeMAGs* only were not significantly different from the MAGs recovered by both workflow, in terms of completeness, contamination, genome size and relative abundance.
iii) We demonstrated that regardless of the MAGs quality criteria, *SnakeMAGs* produces more MAGs and more diverse MAGs than ATLAS.

Any further responses from the reviewers can be found at the end of the article

Introduction

Over the last years, microbial ecology has progressively made the transition from gene-centric to genome-centric analyses,[1] allowing the clear assignment of (sometimes novel) microbial taxa to specific functions and metabolisms.[2–5] Indeed, technical and technological progresses such as binning methods applied to metagenomics,[6] single-cell genome sequencing[7] as well as high-throughput cultivation methods[8] have contributed to the continuing and exponential increase of available prokaryotic genomes.[9] This is particularly true for metagenomics that offers the possibility to reconstruct metagenome-assembled genomes (MAGs) on a large scale and from various environments, and thus has generated a huge amount of new prokaryotic genomes.[10,11]

Although the use of MAGs in microbial ecology is becoming a common practice nowadays, processing raw metagenomic reads up to genome reconstruction involves various steps and software which can represent a major technical obstacle, especially for non-specialists. To face this problem, several workflows such as MetaWRAP,[12] its Snakemake version called SnakeWRAP,[13] ATLAS[14] and more recently MAGNETO,[15] have been developed to automatically reconstruct genomes from metagenomes. However, these workflows contain various modules and perform more tasks than only generating MAGs. For instance, they will taxonomically assign the metagenomic reads, create gene catalogs or perform functional annotations. They rely on numerous dependencies, require significant computational resources and regenerate a lot of outputs which are not essential to most research projects. To simplify this procedure and make it more accessible while remaining efficient, reproducible and biologically relevant, we developed with the popular Snakemake workflow management system,[16] a configurable and easy-to-use workflow called *SnakeMAGs* to reconstruct MAGs in just a few steps. It integrates state-of-the-art bioinformatic tools to sequentially perform from Illumina raw reads: quality filtering of the reads, adapter trimming, an optional step of host sequence removal, assembly of the reads, binning of the contigs, quality assessment of the bins, taxonomic classification of the MAGs and estimation of the relative abundance of these MAGs.

Methods

Creation

Our tool was built by integrating a set of software needed to process metagenomic datasets, utilizing Snakemake. There are no additional equations/maths needed to recreate this tool.

Implementation

The workflow has been developed with the workflow management system Snakemake v7.0.0[16] based on the Python language. Snakemake enables reproducible and scalable data analyses as well as an independent management of the required software within a workflow. *SnakeMAGs* is composed of two main files:

The Snakefile, named "SnakeMAGs.smk", contains the workflow script. It is divided into successive rules which correspond to individual steps. Our workflow includes a total of 17 distinct rules. Each rule requires input files and relies on a single software installed independently when starting the workflow in a dedicated conda v4.12.0 environment. At the end of each rule, output files will be generated in a dedicated folder, as well as a log file (stored in the logs folder) summarizing the events of the software run and a benchmark file (stored in the benchmarks folder) containing the central processing unit (CPU) run time, the wall clock time and the maximum memory usage required to complete the rule. Thanks to Snakemake wildcards, our rules are generalized, so one can process multiple datasets in parallel without having to adjust the source code manually.

The configuration file,[43] named "config.yaml", is used to define some variable names (*e.g.* names of the input files), paths (*e.g.* working directory, location of the reference databases), software parameters and computational resource allocations (threads, memory) for each of the main steps.

To run the workflow, the user only requires Snakemake. It can be easily installed, for instance *via* Conda, as explained in the GitHub repository:

```
conda create -n snakemake_7.0.0 snakemake=7.0.0
```

After that, the user will only have to edit the config file (an example is provided on the GitHub repository) and then run *SnakeMAGs*:

```
#Example of command on a Slurm cluster
snakemake --snakefile SnakeMAGs.smk --cluster \
'sbatch -p <cluster_partition> --mem -c \
-o "cluster_logs/{wildcards}.{rule}.{jobid}.out" \
-e "cluster_logs/{wildcards}.{rule}.{jobid}.err" ' \
--jobs --use-conda --conda-frontend \
conda --conda-prefix/path/to/SnakeMAGs_conda_env/ \
--jobname "{rule}.{wildcards}.{jobid}" --configfile/path/to/config.yaml
```

During the first use of the workflow, a dedicated Conda environment will be installed for each of the bioinformatic tool to avoid conflict. Then the input files will be processed sequentially. Output files will be stored in eight dedicated folder: logs, benchmarks, QC_fq (containing FASTQ files), Assembly, Binning, Bins_quality (all three containing FASTA files), Classification (containing FASTA files and text files with the taxonomic information), and MAGs_abundances (text files).

The workflow has been successfully used on a workstation with Ubuntu 22.04 as well as on high-performance computer clusters with Slurm v18.08.7 and SGE v8.1.9.

Operation

The minimal system requirements to run the workflow will depend on the size of the metagenomic dataset. Small datasets (*e.g.* the test files provided on the GitHub repository) have been successfully analyzed on a workstation with an Intel Xeon Silver 4210, 2.20GHz (10 cores/20 threads) processor and 96GB of RAM. Larger datasets should be processed on cluster computing or within a high-performance infrastructure. For instance, performance evaluation of publicly available metagenomes (see below) was performed on a computer cluster under CentOS Linux release 7.4.1708 distribution with Slurm 18.08.7, on a node possessing an Intel Xeon CPU E7-8890 v4, 2.20GHz (96 cores/192 threads) and 512 GB RAM.

SnakeMAGs integrates a series of bioinformatic tools to sequentially perform from Illumina raw reads: quality filtering of the reads with illumina-utils v2.12,[17] adapter trimming with Trimmomatic v0.39[18] (RRID:SCR_011848), an optional step of host sequence removal (*e.g.* animal or plant sequences) with Bowtie2 v2.4.5[19] (RRID:SCR_016368), assembly of the reads with MEGAHIT v1.2.9[20] (RRID:SCR_018551), binning of the contigs with MetaBAT2 v2.15[21] (RRID:SCR_019134), quality assessment of the bins with CheckM v1.1.3[22] (RRID:SCR_016646) and optionally with GUNC v1.0.5,[23] classification of the MAGs with GTDB-Tk v2.1.0[24] (RRID:SCR_019136) and estimation of the relative abundance of these MAGs with CoverM v0.6.1. An overview of the workflow is presented in Figure 1.

Use cases

To demonstrate the benefits and potential of our workflow, we compared it to another Snakemake workflow named ATLAS v2.9.1.[14] To produce a fair comparison, ATLAS was run with the MEGAHIT assembler, without co-binning and dereplicating only 100% similar MAGs. To test these two workflows, we downloaded and analyzed ten publicly available termite gut metagenomes (accession numbers: SRR10402454; SRR14739927; SRR8296321; SRR8296327; SRR8296329; SRR8296337; SRR8296343; DRR097505; SRR7466794; SRR7466795) from five studies[25–29] and belonging to ten different termite species.

SnakeMAGs requires only a limited number of inputs files: the raw metagenomic reads in FASTQ format from the 10 above-mentioned metagenomes, a FASTA file containing the adapter sequences,[43] a YAML configuration file specifying the variable names, paths and computational resource allocations (available on the GitHub repository and on Zenodo), and here since we worked with host-associated metagenomes a FASTA file containing the termite genome sequences.[42] Regarding the outputs, *SnakeMAGs* produced quality-controlled FASTQ files without adapters nor termite sequences, in the QC_fq folder. Then the reads assembled into contigs and scaffolds (FASTA files) were saved in the Assembly folder. Products of the binning procedure were stored in the Binning folder. Bins with >50% completeness and <10% contamination (according to CheckM) were considered as medium-quality MAGs[30] and stored in the

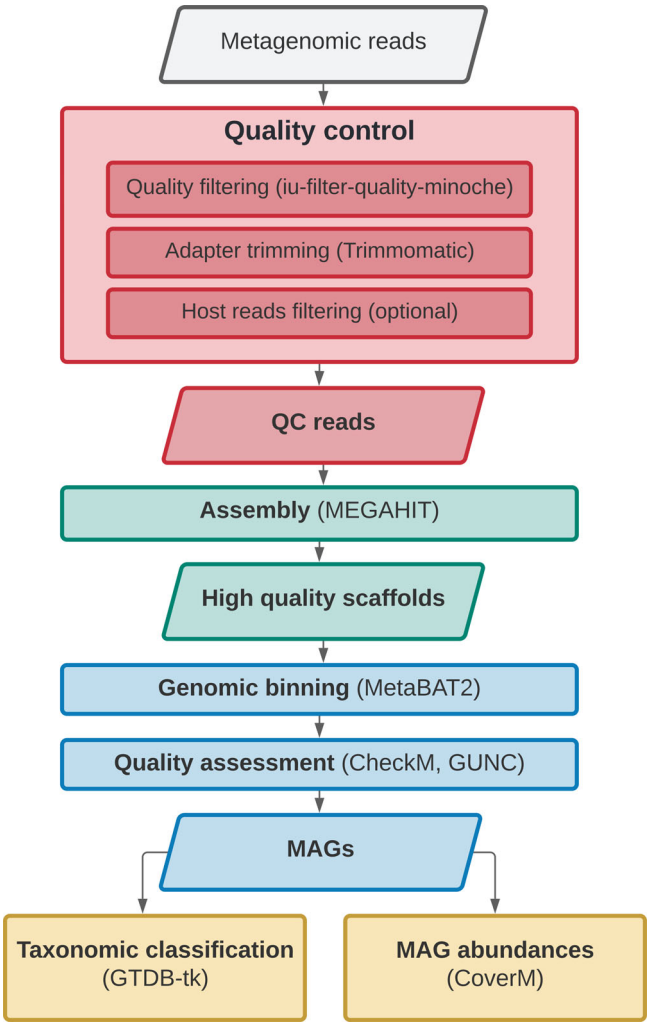

Figure 1. Directed acyclic graph describing the main steps performed by *SnakeMAGs* v1.1.0. The names of the software used for each step are showed in parentheses.

Bins_quality folder. At this step, it is also possible to use GUNC to remove potential chimeric and contaminated genomes. Subsequently, the results of the MAGs classification and relative abundance estimation were sent to the Classification and MAGs_abundances folders, respectively. ATLAS requires similar input files and produces, among others, similar outputs files.

ATLAS appeared to be faster than *SnakeMAGs* (Wilcoxon test, $P = 0.002$) to reconstruct MAGs from metagenomes (Figure 2A) with a similar memory usage (Wilcoxon test, $P = 0.393$). However, *SnakeMAGs* always recovered more MAGs (>50% completeness and <10% contamination according to CheckM) per metagenome or at least as much as ATLAS (Figure 2B). From the ten metagenomes, *SnakeMAGs* produced a total of 65 MAGs while ATLAS generated only 37 MAGs. Additionally, *SnakeMAGs* was able to recover MAGs encompassing a higher diversity of bacterial phyla ($n = 15$ phyla) compared to ATLAS ($n = 11$ phyla). Only one phylum, namely *Patescibacteria*, represented by a single MAG was recovered by ATLAS and not by *SnakeMAGs*. On the contrary, ATLAS failed to reconstruct MAGs belonging to *Verrucomicrobiota*, *Planctomycetota*, *Synergistota*, *Elusimicrobiota* and *Acidobacteriota* when *SnakeMAGs* succeeded (Figure 2C). We found no difference in MAG quality or genome size between the two workflows (Wilcoxon test, $P = 0.15$ for completeness; $P = 0.60$ for contamination and $P = 0.64$ for genome size). We also found that the additional MAGs recovered by *SnakeMAGs* did not differ from the others. MAGs belonging to phyla generated by *SnakeMAGs* only or recovered by both *SnakeMAGs* and ATLAS indeed did not significantly differ in terms of quality, relative abundance and genome size (Wilcoxon test, $P = 0.19$ for completeness; $P = 0.43$ for contamination; $P = 0.51$ for relative abundance and $P = 0.19$ for genome size).

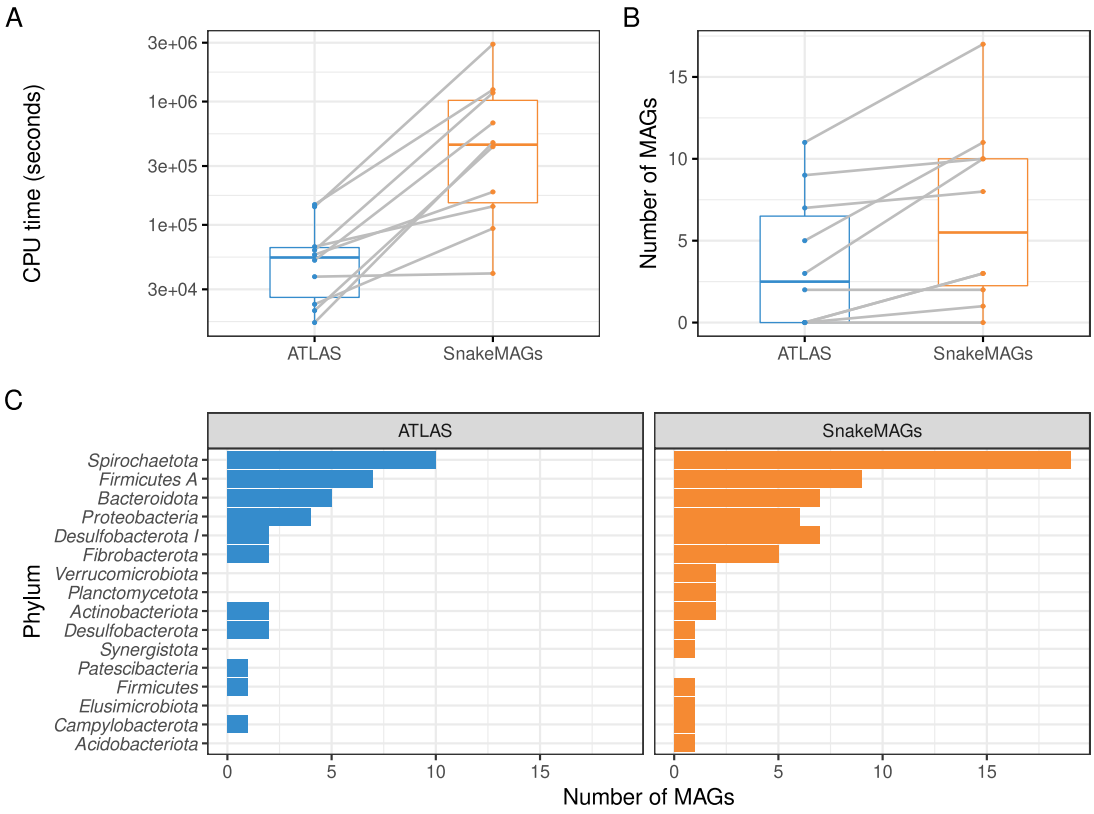

Figure 2. Comparison of the performance of *SnakeMAGs* v1.1.0 with another workflow, namely ATLAS v2.9.1[14] using 10 termite gut metagenomes. A. CPU time (in seconds) required to process each metagenome. B. Number of MAGs reconstructed from each metagenome. On both boxplots, gray lines link the result obtained with ATLAS and the one obtained with *SnakeMAGs* for each of the 10 analyzed termite metagenomes. C. Number of bacterial MAGs at the phylum level recovered from each workflow.

Then we evaluated the effect of MAG quality criteria on our workflow. Using an estimated quality threshold ≥50 (with quality defined as completeness − 5 × contamination),[31] *SnakeMAGs* still allowed the recovery of more MAGs ($n = 46$) than ATLAS ($n = 31$). In terms of diversity, *SnakeMAGs* also recovered more phyla ($n = 13$) than ATLAS ($n = 10$). Similarly, using GUNC combined with CheckM for genome quality assessment reduced the number of MAGs recovered by both workflows but a similar trend was observed: *SnakeMAGs* produced more MAGs and more diverse MAGs ($n = 59$ MAGs, encompassing 13 phyla) than ATLAS ($n = 29$ MAGs, encompassing 9 phyla). In summary, the advantages of our workflow are robust to the MAG quality criteria.

Discussion

Using metagenomic datasets from the gut of various termite species, our analyses revealed that while being slower, *SnakeMAGs* allowed the recovery of more MAGs encompassing more diverse phyla compared to ATLAS, another similar Snakemake workflow. More importantly our results showed that *SnakeMAGs* was able to recover MAGs encompassing the major bacterial phyla found in termite guts,[32,33] and that some of these phyla were not recovered by ATLAS. Indeed, taxa belonging to *Verrucomicrobiota,*[34] *Planctomycetota,*[33,35] *Synergistota,*[36] *Elusimicrobiota*[37] and *Acidobacteriota*[38,39] have been repeatedly found in the gut of various termite species. As such, they would represent relevant targets for genome-centric analyses of the termite gut microbiota. Although we found no significant difference between the relative abundance of the MAGs belonging to phyla recovered by *SnakeMAGs* only and the MAGs recovered by both workflows, it is worth mentioning that *Verrucomicrobiota*, *Planctomycetota*, *Synergistota*, *Elusimicrobiota* and *Acidobacteriota* are usually less abundant than *Spirochaetota*, *Firmicutes* and *Bacteroidota* which are dominant phyla in termite gut,[32,33] and that have been recovered by both workflows in our study. Therefore, it would suggest that *SnakeMAGs* is not restricted to the most abundant taxa. Altogether, we showed that *SnakeMAGs* has the potential to retrieve quantitatively more genomic information from metagenomes but also to extract genomic features of biological interest.

Thanks to the inherent flexibility of Snakemake, *SnakeMAGs* offers the possibility to the users to easily tune the parameters of the workflow (*e.g.* resource allocations for each rule, options of a specific tools) to adapt their analysis to the datasets and to the computational infrastructure. Additionally, advanced users will have the opportunity to edit or add new rules to the workflow. Regarding the future of *SnakeMAGs*, several avenues will be considered for the next versions of the workflow. Firstly, the workflow could give more freedom to the users by offering the choice of different tools to perform the same task (*e.g.* different trimming, assembly or binning software). Secondly, with the current emergence of metagenomic datasets generated with long-read DNA sequencing,[40] it might be relevant to adjust our workflow for long-read sequencing technology by including specific bioinformatic tools for this technology.[41] Meanwhile, since the majority of the metagenomic datasets have been and are still currently generated with Illumina short-read technology, *SnakeMAGs* can be widely used to explore the genomic content of various ecosystems *via* metagenomics.

Software availability

Source code available from: https://github.com/Nachida08/SnakeMAGs

Archived source code at time of publication: https://doi.org/10.5281/zenodo.7665149.[42]

License: CeCILL v2.1

Data availability

Source data

Termite genome references used for removing host sequences and their Bowtie2 index are available at: https://zenodo.org/record/6908287#.Y1JLANJBzUR

The termite gut metagenomes analyzed in the present study are available on NCBI with the following accession numbers: SRR10402454; SRR14739927; SRR8296321; SRR8296327; SRR8296329; SRR8296337; SRR8296343; DRR097505; SRR7466794; SRR7466795.

Underlying data

Zenodo. Reconstruction of prokaryotic genomes from ten termite gut metagenomes using two distinct workflows: SnakeMAGs and ATLAS: https://doi.org/10.5281/zenodo.7661004.[43]

- SnakeMAGs_config.yaml (The configuration file used to analyze the 10 termite gut metagenomes with *SnakeMAGs*)
- ATLAS_config.yaml (The configuration file used to analyze the 10 termite gut metagenomes with ATLAS)
- MAGs_SnakeMAGs.zip (A zipped folder containing the genomes of the 65 MAGs reconstructed with *SnakeMAGs*)
- MAGs_ATLAS.zip (A zipped folder containing the genomes of the 37 MAGs reconstructed with ATLAS)
- taxonomic_assignment_MAGs.csv (A text file containing the taxonomic assignment of all the MAGs reconstructed by both workflows)

Acknowledgments

The authors thank Emmanuelle Morin and Hélène Gardon for their valuable advice and feedback during the workflow development.

References

1. Prosser JI: **Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology.** *Nat. Rev. Microbiol.* 2015 Jun 8; **13**(7): 439–446.
PubMed Abstract | Publisher Full Text

2. Evans PN, Parks DH, Chadwick GL, *et al.*: **Methane metabolism in the archaeal phylum *Bathyarchaeota* revealed by genome-centric metagenomics.** *Science.* 2015 Oct 23; **350**(6259): 434–438.
PubMed Abstract | Publisher Full Text

3. Engelberts JP, Robbins SJ, de Goeij JM, *et al.*: **Characterization of a sponge microbiome using an integrative genome-centric approach.** *ISME J.* 2020 Jan 28; 1–11.
Publisher Full Text

4. Loh HQ, Hervé V, Brune A: **Metabolic potential for reductive acetogenesis and a novel energy-converting [NiFe] hydrogenase in *Bathyarchaeia* from termite guts – A genome-centric analysis.** *Front. Microbiol.* 2021 Feb 3; **11**: 3644.
Publisher Full Text

5. Bay SK, Dong X, Bradley JA, *et al.*: **Trace gas oxidizers are widespread and active members of soil microbial communities.** *Nat. Microbiol.* 2021 Feb 4; **6**(2): 246–256.
PubMed Abstract | Publisher Full Text

6. Sedlar K, Kupkova K, Provaznik I: **Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics.** *Comput. Struct. Biotechnol. J.* 2017 Jan 1; **15**: 48–55.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Woyke T, Doud DFR, Schulz F: **The trajectory of microbial single-cell sequencing.** *Nat. Methods.* 2017 Oct 31; **14**(11): 1045–1054.
PubMed Abstract | Publisher Full Text

8. Overmann J, Abt B, Sikorski J: **Present and future of culturing bacteria.** *Annu. Rev. Microbiol.* 2017 Sep 8; **71**(1): 711–730.
Publisher Full Text

9. Almeida A, Nayfach S, Boland M, *et al.*: **A unified catalog of 204,938 reference genomes from the human gut microbiome.** *Nat. Biotechnol.* 2021 Jan 20; **39**(1): 105–114.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Nayfach S, Roux S, Seshadri R, *et al.*: **A genomic catalog of Earth's microbiomes.** *Nat. Biotechnol.* 2021 Apr 9; **39**(4): 499–509.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Pasolli E, Asnicar F, Manara S, *et al.*: **Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle.** *Cell.* 2019 Jan; **176**(3): 649–662.e20.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Uritskiy GV, DiRuggiero J, Taylor J: **MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis.** *Microbiome.* 2018 Dec 15; **6**(1): 158.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Krapohl J, Pickett B: **SnakeWRAP: a Snakemake workflow to facilitate automated processing of metagenomic data through the metaWRAP pipeline [version 2; peer review: 1 approved].** *F1000Res.* 2022; **11**(265).
Publisher Full Text

14. Kieser S, Brown J, Zdobnov EM, *et al.*: **ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data.** *BMC Bioinformatics.* 2020 Dec 22; **21**(1): 257.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Churcheward B, Millet M, Bihouée A, *et al.*: **MAGNETO: An automated workflow for genome-resolved metagenomics.** *mSystems.* 2022; **7**(4): e00432–e00422.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Mölder F, Jablonski KP, Letcher B, *et al.*: **Sustainable data analysis with Snakemake.** *F1000Res.* 2021; Vol. **10**:
PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source

17. Eren AM, Vineis JH, Morrison HG, *et al.*: **A filtering method to generate high quality short reads using Illumina paired-end technology.** *PLoS One.* 2013; **8**(6).
Publisher Full Text

18. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014 Aug 1; **30**(15): 2114–2120.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat. Methods.* 2012 Apr 4; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Li D, Liu CM, Luo R, *et al.*: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly *via* succinct de Bruijn graph.** *Bioinformatics.* 2015 May 15; **31**(10): 1674–1676.
PubMed Abstract | Publisher Full Text

21. Kang DD, Li F, Kirton E, *et al.*: **MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.** *PeerJ.* 2019 Jul 26; **7**: e7359.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Parks DH, Imelfort M, Skennerton CT, *et al.*: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.** *Genome Res.* 2015 Jul; **25**(7): 1043–1055.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Orakov A, Fullam A, Coelho LP, *et al.*: **GUNC: detection of chimerism and contamination in prokaryotic genomes.** *Genome. Biol.* 2021; **22**: 178.
Publisher Full Text

24. Chaumeil PA, Mussig AJ, Hugenholtz P, *et al.*: **GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database.** *Bioinformatics.* 2022 Oct 11; btac672.
Publisher Full Text

25. Calusinska M, Marynowska M, Bertucci M, *et al.*: **Integrative omics analysis of the termite gut system adaptation to *Miscanthus* diet identifies lignocellulose degradation enzymes.** *Communications Biology.* 2020 Dec 1; **3**(1): 275.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Moreira EA, Persinoti GF, Menezes LR, *et al.*: **Complementary contribution of fungi and bacteria to lignocellulose digestion in the food stored by a neotropical higher termite.** *Front. Ecol. Evol.* 2021 Apr 26; **9**: 248.
Publisher Full Text

27. Romero Victorica M, Soria MA, Batista-García RA, *et al.*: **Neotropical termite microbiomes as sources of novel plant cell wall degrading enzymes.** *Sci. Rep.* 2020 Dec 2; **10**(1): 3864.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Tokuda G, Mikaelyan A, Fukui C, *et al.*: **Fiber-associated spirochetes are major agents of hemicellulose degradation in the hindgut of wood-feeding higher termites.** *Proc. Natl. Acad. Sci.* 2018 Dec 18; **115**(51): E11996–E12004.
Publisher Full Text

29. Waidele L, Korb J, Voolstra CR, *et al.*: **Ecological specificity of the metagenome in a set of lower termite species supports contribution of the microbiome to adaptation of the host.** *Animal Microbiome.* 2019 Dec 24; **1**(1): 13.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Bowers RM, Kyrpides NC, Stepanauskas R, *et al.*: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.** *Nat. Biotechnol.* 2017; **35**(8): 725–731.
Publisher Full Text

31. Parks DH, Rinke C, Chuvochina M, *et al.*: **Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life.** *Nat. Microbiol.* 2017; **2**: 1533–1542.
Publisher Full Text

32. Arora J, Kinjo Y, Šobotník J, *et al.*: **The functional evolution of termite gut microbiota.** *Microbiome.* 2022 Dec; **10**(1): 78.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Hervé V, Liu P, Dietrich C, *et al.*: **Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites.** *PeerJ.* 2020 Feb; **8**: e8614.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Wertz JT, Kim E, Breznak JA, *et al.*: **Genomic and physiological characterization of the *Verrucomicrobia* isolate *Diplosphaera colitermitum* gen. nov., sp. nov., reveals microaerophily and nitrogen fixation genes.** *Appl. Environ. Microbiol.* 2012 Mar 1; **78**(5): 1544–1555.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Köhler T, Stingl U, Meuser K, *et al.*: **Novel lineages of *Planctomycetes* densely colonize the alkaline gut of soil-feeding termites (*Cubitermes* spp.).** *Environ. Microbiol.* 2008 May; **10**(5): 1260–1270.
PubMed Abstract | Publisher Full Text

36. Ahmad F, Yang G, Zhu Y, *et al.*: **Tripartite symbiotic digestion of lignocellulose in the digestive system of a fungus-growing termite.** *Microbiology Spectrum.* 2022 Oct 17; e01234–e01222.
Publisher Full Text

37. Herlemann DPR, Geissinger O, Ikeda-Ohtsubo W, *et al.*: **Genomic analysis of "*Elusimicrobium minutum*," the first cultivated representative of the phylum "*Elusimicrobia*" (formerly termite group 1).** *Appl. Environ. Microbiol.* 2009 May 1; **75**(9): 2841–2849.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Hongoh Y, Deevong P, Inoue T, *et al.*: **Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host.** *Appl. Environ. Microbiol.* 2005 Nov 1; **71**(11): 6590–6599.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Bourguignon T, Lo N, Dietrich C, *et al.*: **Rampant host switching shaped the termite gut microbiome.** *Curr. Biol.* 2018 Feb; **28**(4): 649–654.e2.
PubMed Abstract | Publisher Full Text

40. Bickhart DM, Kolmogorov M, Tseng E, *et al.*: **Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities.** *Nat. Biotechnol.* 2022 Jan 3; **40**: 711–719.
PubMed Abstract | Publisher Full Text

41. Feng X, Cheng H, Portik D, *et al.*: **Metagenome assembly of high-fidelity long reads with hifiasm-meta.** *Nat. Methods.* 2022 Jun; **19**(6): 671–674.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Tadrent N, Dedeine F, Hervé V: **SnakeMAGs (v1.1.0).** *[Code] Zenodo.* 2022.
Publisher Full Text

43. Tadrent N, Dedeine F, Hervé V: Reconstruction of prokaryotic genomes from ten termite gut metagenomes using two distinct workflows: SnakeMAGs and ATLAS. [Data]. *Zenodo.* 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 01 March 2023

https://doi.org/10.5256/f1000research.144793.r164770

✔ **Célio Dias Santos Júnior**
Institute of Science and Technology for Brain-Inspired Intelligence - ISTBI, Fudan University, Shanghai, China

The new improvements implemented in the paper "*SnakeMAGs*: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes" not only rise its scientific value but also supply strong evidence of its applicability and importance. I feel convinced by the answers and additional analysis provided by Tradent *et al.* and praise the authors for their disposition in implementing some of my suggestions, especially about the release of a new version containing GUNC as a filter.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, microbiology, biochemistry, biotechnology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 15 February 2023

https://doi.org/10.5256/f1000research.140650.r159904

✔ **Aram Mikaelyan**

Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA

The authors describe SnakeMAGs as a dedicated pipeline for the compositional binning of contigs generated from Illumina sequencing reads obtained from microbial communities. It integrates the various steps in the process of comparative analyses in metagenomes, starting with the assembly of reads generated from community DNA, to the compositional binning of the assembled contigs into high-quality metagenome assembled contigs (MAGs). I believe this pipeline represents a valuable addition to the community. I found the manuscript was easy to read and have the following comments for the authors:

1. The main advantage of Snake is that it is highly task-specific to the generation of high-quality MAGs from complex communities. It therefore can save considerable time and resources by avoiding steps in processing data (e.g. extensive functional or taxonomic annotations) that are irrelevant to the researcher's objective.

   I would like the authors to expand on how their pipeline controls for chimeric sequences, since these may lead to an artefactual inflation of MAGs.

2. As a termite researcher, I am aware of the many quirks that termite gut microbiomes present, and the limitations of many data processing pipelines and databases when it comes to addressing those quirks. I was glad that the authors used this challenging community for their benchmarks, and impressed at the contrast between the results from SnakeMAGs and ATLAS. I will emphasize further that the apparent inability of ATLAS to report phyla such as Planctomycetota and Elusimicrobia, or reconstruct fewer MAGs from Fibrobacterota or Desulfobacterota, is not trivial. These bacterial members play major roles in the microbial ecology of the termite gut, and the SnakeMAGs' ability to detect them highlights its superiority.

   I would like the authors to list/explain some factors could potentially account for this difference.

3. I found the installation and usage on a Linux Mint server (with comparable processing power to the one used by the authors) to be straightforward. The dependencies are minimalistic and easy to install.

I congratulate the authors on sharing this tool with the community and hope they find my comments useful.

Aram Mikaelyan

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Entomology, Microbial Ecology, Microbial systematics, Symbiosis, Bioinformatics, Evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 22 Feb 2023

Vincent Hervé

The authors describe SnakeMAGs as a dedicated pipeline for the compositional binning of contigs generated from Illumina sequencing reads obtained from microbial communities. It integrates the various steps in the process of comparative analyses in metagenomes, starting with the assembly of reads generated from community DNA, to the compositional binning of the assembled contigs into high-quality metagenome assembled contigs (MAGs). I believe this pipeline represents a valuable addition to the community. I found the manuscript was easy to read and have the following comments for the authors:

1. The main advantage of Snake is that it is highly task-specific to the generation of high-quality MAGs from complex communities. It therefore can save considerable time and resources by avoiding steps in processing data (e.g. extensive functional or taxonomic annotations) that are irrelevant to the researcher's objective.
I would like the authors to expand on how their pipeline controls for chimeric sequences, since these may lead to an artefactual inflation of MAGs.

Reply – Indeed, this aspect was not considered in the first version of our workflow. Therefore, we decided to release a version 1.1 of *SnakeMAGs* that now includes GUNC, a software for detection of chimerism and contamination in prokaryotic genomes. We also evaluated the impact of this tool on the MAGs generated by both ATLAS and *SnakeMAGs.* For the *SnakeMAGs* genomes, we found that 59 out of 65 MAGs, encompassing 13 phyla, passed the GUNC filtering step. With ATLAS, we found that 29 out of 37 MAGs, encompassing 9 phyla, passed the GUNC filtering step. In summary, a few MAGs generated by both workflows did not pass the GUNC quality criteria. Importantly, this new analysis shows that including GUNC does not change the major outcome of our comparison with ATLAS: *SnakeMAGs* produces more MAGs and more diverse MAGs than ATLAS. This analysis has now been included in the revised version of the manuscript.

2. As a termite researcher, I am aware of the many quirks that termite gut microbiomes present, and the limitations of many data processing pipelines and databases when it comes to addressing those quirks. I was glad that the authors used this challenging community for their benchmarks, and impressed at the contrast between the results from SnakeMAGs and ATLAS. I will emphasize further that the apparent inability of ATLAS to report phyla such as Planctomycetota and Elusimicrobia, or reconstruct fewer MAGs from Fibrobacterota or Desulfobacterota, is not trivial. These bacterial members play major roles in the microbial ecology of the termite gut, and the SnakeMAGs' ability to detect them highlights its superiority.
I would like the authors to list/explain some factors could potentially account for this difference.

Reply – A potential explanation for this difference can be the relative abundance of the phyla recovered only with *SnakeMAGs* compared to the relative abundance of the phyla recovered by both workflows. We now further discuss this point in the revised version of the manuscript.
We found no significant difference in the relative abundance (Wilcoxon test, $P = 0.51$) of the MAGs belonging to phyla recovered only with *SnakeMAGs* compared to the phyla recovered by both workflows. However, it should be noted that although *Verrucomicrobiota*, *Planctomycetota*, *Synergistota*, *Elusimicrobiota* and *Acidobacteriota* have been reported among the major taxa in termite gut, they are usually less abundant than *Spirochaetota, Firmicutes and Bacteroidota* (Arora et al, 2022, *Microbiome*) that have been recovered by both workflows. This result suggests that *SnakeMAGs* is not restricted to the most abundant taxa unlike ATLAS, which could be relatively more prompted to recover only the most abundant microbial taxa.

3. I found the installation and usage on a Linux Mint server (with comparable processing power to the one used by the authors) to be straightforward. The dependencies are minimalistic and easy to install.

I congratulate the authors on sharing this tool with the community and hope they find my comments useful.

Reply – We thank Reviewer 2 for this positive feedback.

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 January 2023

https://doi.org/10.5256/f1000research.140650.r158646

Célio Dias Santos Júnior
Institute of Science and Technology for Brain-Inspired Intelligence - ISTBI, Fudan University, Shanghai, China

A method to bin MAGs from reads is described in the publication "SnakeMAGs: a simple, efficient, flexible and scalable approach to rebuild bacterial genomes from metagenomes" by Tradent *et al.* The use of this pipeline to standardize the work with MAGs and make it accessible to batch runs has an obvious benefit. Although the pipeline itself appears a little out of date in terms of conceptual work in the area, I think it has the potential to develop into a useful tool for bioinformatics in general with a few modifications. The critique I offer in this review is more directed at the pipeline's ongoing structural evolution than it is towards the actual article. Although the study reads extremely well, further work must be done to adequately support the use of SnakeMAGs.

Abstract

The authors here should have put greater emphasis on what may be anticipated in terms of the quality of the genomes and species retrieved. It is not favorable if you can recover more genomes but they are of inferior quality, smaller size, or skewed towards a certain species.

Methods

I want to congratulate the authors for the open scientific component. The databases, code, and software were all presented in an organized manner. The documentation reads quite well and appears to be a crucial component of your work. Both the tool's installation and usage are quite well-explained and simple.

As an adept of collaborative research, I saw that your GitHub lacked some unit tests that would have allowed other contributors to repair bugs that the authors had little interaction with or had previously overlooked without jeopardizing the distribution. That is not a question in this assessment, but it may represent a future enhancement.

I understand the authors' desire to keep the pipeline short to avoid several dependencies. However, it doesn't appear that the main pipeline is more effective or addresses that issue. For instance, ATLAS is faster, although relying on more dependencies. What benefit do SnakeMAGs have in this regard?

Despite being condensed, SnakeMAGs does appear to skip certain necessary stages to produce better MAGs, including:

1. Using software that controls for chimera, such as metaMIC[1], to fix contigs obtained after assembly;

2. Including multiple binning systems which may seem like a delay but ensures the best bin for each species and even offers a better resolution at the strain level. For instance, using a next-generation binner, such as semiBin or VAMB, then afterward clustering the MAGs on an ANI basis to assure a higher resolution of the species found. Several of these techniques even permit the use of long reads, which may ultimately be advantageous;

3. Since we now know that some of the species binned occasionally display under or overestimated completeness due to bias in the calculations using USCG, run a quality check of the bins using a more comprehensive approach than CheckM, for example, implementing software that addresses contamination, such as GUNC, or even more advanced updates of previous systems that now incorporate machine learning, such as CheckM2 , which does not present bias towards a taxonomy. It would be the best solution for this quality-checking step.

Is there a plan to integrate these actions in the upcoming versions? How may these omitted stages affect the outcomes in the current SnakeMAGs version?

Results

What was the reason for choosing termite gut metagenomes? Is there any indication of better sequencing, deeper sequencing, higher microbial diversity, or any other reason that led the authors to this choice?

The quality criteria used by authors to classify good genomes (>50% completeness and <10% contamination according to CheckM) is quite loose and not widely adopted. Parks *et al.* (2017)[2] - one of the first works using MAGs and also where checkM was presented - used an estimated quality ≥ 50 (defined as completeness $-$ 5 $\times$ contamination). In the present work, if a genome is in the bottom line, it would have an estimated quality of 0, which is far from the threshold Parks adopted. Other criteria are also quite important, e.g. the number of contigs, N50, and ambiguous base pairs. I recommend authors reassess these results using a more strict quality parameter and then report if their results are better or comparable to the other pipelines. I believe that if other binners were combined and similar bins merged as in Parks *et al.* (2017)[2], SnakeMAGs would represent a game changer.

The fact that "SnakeMAGs was able to recover MAGs encompassing a higher diversity of bacterial phyla" is impressive, and I wonder what the average quality of these MAGs recovered in those phyla that ATLAS was not able to bin. Is there any chance that ATLAS discarded bad-quality MAGs that SnakeMAGs is assuming as correct?

The analysis regarding the memory needed by both systems is missing. The authors mentioned that the peak of memory usage is registered in the SnakeMAGs' logs, and a comparison with ATLAS needs seems important. The longer time for processing, if accompanied by a reduction in memory needs, still seems advantageous in my POV.

Discussion

The fact that the phyla recovered by SnakeMAGs match well with reports of microbial diversity in the samples analyzed are quite appealing to me. However, do the abundances of these species also vary largely? How abundant the phyla recovered only by SnakeMAGs are? Authors should explore this factor to explain if SnakeMAGs' advantage lies in binning genomes from rare species.

It makes me happy to know that the authors plan versions of the pipeline along with improvements. This spirit of continuous development usually ends up in great bioinformatics tools.

References
1. Lai S, Pan S, Sun C, Coelho LP, et al.: metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies.*Genome Biol*. 2022; **23** (1): 242 PubMed Abstract | Publisher Full Text
2. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, et al.: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life.*Nat Microbiol*. 2017; **2** (11): 1533-1542 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, microbiology, biochemistry, biotechnology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 22 Feb 2023

Vincent Hervé

Abstract

The authors here should have put greater emphasis on what may be anticipated in terms of the quality of the genomes and species retrieved. It is not favorable if you can recover more genomes but they are of inferior quality, smaller size, or skewed towards a certain species.

Reply – We agree, it is indeed an important point. In the revised version of the manuscript, we show that the quality and genome size of the MAGs recovered by *SnakeMAGs* only did not differ from those belonging to phyla recovered by both *SnakeMAGs* and ATLAS. This is

now specified in the abstract of the manuscript.

Methods

I want to congratulate the authors for the open scientific component. The databases, code, and software were all presented in an organized manner. The documentation reads quite well and appears to be a crucial component of your work. Both the tool's installation and usage are quite well-explained and simple.

Reply – We thank Reviewer 1 for these positive comments.

As an adept of collaborative research, I saw that your GitHub lacked some unit tests that would have allowed other contributors to repair bugs that the authors had little interaction with or had previously overlooked without jeopardizing the distribution. That is not a question in this assessment, but it may represent a future enhancement.

Reply – We thank Reviewer 1 for this suggestion. We will fully consider it for future enhancement. Meanwhile, users have the opportunity to open Issues on the GitHub page to report any bug or request. Additionally, the test files provided on GitHub allow the user to test our workflow before running real and large datasets.

I understand the authors' desire to keep the pipeline short to avoid several dependencies. However, it doesn't appear that the main pipeline is more effective or addresses that issue. For instance, ATLAS is faster, although relying on more dependencies. What benefit do SnakeMAGs have in this regard?

Reply – Our main goal was to design a simple and minimalist workflow, so a non-specialist can easily bin MAGs without having to choose among (or test and compare) the myriads of software currently available to perform the different steps of our workflow. We believe that the strength of *SnakeMAGs* in this regard is its simplicity. Using *SnakeMAGs*, a non-specialist user just needs to follow general instructions, while using ATLAS, such a user need to make decisive choices of various tools at many steps of the workflow, an approach that can easily become quite confusing.

Despite being condensed, SnakeMAGs does appear to skip certain necessary stages to produce better MAGs, including:

1. Using software that controls for chimera, such as metaMIC[1], to fix contigs obtained after assembly;

2. Including multiple binning systems which may seem like a delay but ensures the best bin for each species and even offers a better resolution at the strain level. For instance, using a next-generation binner, such as semiBin or VAMB, then afterward clustering the MAGs on an ANI basis to assure a higher resolution of the species found. Several of these techniques even permit the use of long reads, which may ultimately be advantageous;

3. Since we now know that some of the species binned occasionally display under or

overestimated completeness due to bias in the calculations using USCG, run a quality check of the bins using a more comprehensive approach than CheckM, for example, implementing software that addresses contamination, such as GUNC, or even more advanced updates of previous systems that now incorporate machine learning, such as CheckM2 , which does not present bias towards a taxonomy. It would be the best solution for this quality-checking step.

Is there a plan to integrate these actions in the upcoming versions? How may these omitted stages affect the outcomes in the current SnakeMAGs version?

Reply – We thank Reviewer 1 for these constructive comments. We acknowledge that the processing of metagenomic reads as well as binning are currently very active fields of research and thus, new tools have been published since we started working on our workflow. As mentioned in our Discussion, in the upcoming versions we plan to include alternative software to perform certain tasks and the binning step will be one of them with the addition of SemiBin. CheckM2 is currently in a preprint stage but we also plan to include it as soon as it will be peer-reviewed.

Regarding the chimera, we felt that it was a significant gap in our workflow so we decided to release a version 1.1 of *SnakeMAGs* that now includes GUNC for genome quality evaluation. We also evaluated the impact of this tool on the MAGs generated by both ATLAS and *SnakeMAGs.* For the *SnakeMAGs* genomes, we found that 59 out of 65 MAGs, encompassing 13 phyla, passed the GUNC filtering step. With ATLAS, we found that 29 out of 37 MAGs, encompassing 9 phyla, passed the GUNC filtering step. In summary, a few MAGs generated by both workflows did not pass the GUNC quality criteria, but including GUNC does not change the major outcome of our comparison with ATLAS: *SnakeMAGs* produces more MAGs and more diverse MAGs than ATLAS. This analysis has now been included in the revised version of the manuscript.

Results

What was the reason for choosing termite gut metagenomes? Is there any indication of better sequencing, deeper sequencing, higher microbial diversity, or any other reason that led the authors to this choice?

Reply – It was indeed a deliberate choice to select termite gut metagenomes. We have been working on termite gut microbiota for many years so we are familiar with the microbial diversity present in such systems. Therefore, we have the expertise to properly evaluate the outputs of the workflows from a biological perspective. We would have been less confident with samples from other ecosystems (e.g. deep-sea sediments). Additionally, we would like to mention that termites are well-known to harbor a higher microbial diversity compared to other arthropods. We also selected samples encompassing hosts from different termite families and with different diets, two factors known to impact the gut microbial diversity. Therefore, we believe that termite gut metagenomes constitute relevant datasets to test our workflow.

The quality criteria used by authors to classify good genomes (>50% completeness and <10% contamination according to CheckM) is quite loose and not widely adopted. Parks *et*

al. (2017)[2] - one of the first works using MAGs and also where checkM was presented - used an estimated quality ≥50 (defined as completeness − 5 × contamination). In the present work, if a genome is in the bottom line, it would have an estimated quality of 0, which is far from the threshold Parks adopted. Other criteria are also quite important, e.g. the number of contigs, N50, and ambiguous base pairs. I recommend authors reassess these results using a more strict quality parameter and then report if their results are better or comparable to the other pipelines. I believe that if other binners were combined and similar bins merged as in Parks *et al.* (2017)[2], SnakeMAGs would represent a game changer.

Reply – Regarding the quality criteria, we used the criteria of "Medium quality MAG" as defined by the Genomic Standards Consortium for the Minimum Information about a Metagenome-Assembled Genome (see Bowers et al, 2017, Nature Biotechnology). This is now specified in the revised version of the manuscript. However, we acknowledge that other authors have used higher quality standard. Following the reviewer's suggestion, we reassessed our results using the estimated quality ≥50 (defined as completeness − 5 × contamination). As expected, using this higher quality threshold the number of recovered MAGs decreases compared to our initial criteria for both workflows: from 37 to 31 MAGs for ATLAS and from 65 to 46 MAGs for *SnakeMAGs*. Therefore, *SnakeMAGs* still allows the recovery of more MAGs than ATLAS. In terms of diversity, *SnakeMAGs* also recover more phyla (*n* = 13) than ATLAS (*n* = 10). In summary, the advantages of our workflow are robust to the MAG quality criteria. These results are now included in the revised version of the manuscript. It is noteworthy that in the version 1.1 of *SnakeMAGs* we have now implemented in the Quality assessment step an option to filter the MAGs according to this estimated quality criteria (completeness − 5 × contamination). This will allow users to freely select more stringent quality criteria if they want to.

The fact that "SnakeMAGs was able to recover MAGs encompassing a higher diversity of bacterial phyla" is impressive, and I wonder what the average quality of these MAGs recovered in those phyla that ATLAS was not able to bin. Is there any chance that ATLAS discarded bad-quality MAGs that SnakeMAGs is assuming as correct?

Reply – This is indeed an important point that is now reported in the revised version of the manuscript. Overall, we found no difference in MAG quality or genome size between the two workflows (Wilcoxon test, $P = 0.15$ for completeness; $P = 0.60$ for contamination and $P = 0.64$ for genome size). Regarding the MAGs generated by *SnakeMAGs*, we found no difference in MAG quality or genome size between the phyla also recovered by ATLAS and the phyla only recovered by *SnakeMAGs* (Wilcoxon test, $P = 0.19$ for completeness; $P = 0.43$ for contamination and $P = 0.19$ for genome size). Therefore, we found no evidence supporting the fact that ATLAS discarded bad-quality MAGs that *SnakeMAGs* is assuming as correct. These results are now included in the revised version of the manuscript.

The analysis regarding the memory needed by both systems is missing. The authors mentioned that the peak of memory usage is registered in the SnakeMAGs' logs, and a comparison with ATLAS needs seems important. The longer time for processing, if accompanied by a reduction in memory needs, still seems advantageous in my POV.

Reply – Following Reviewer 1 suggestion, we performed this comparison but found not

significant difference in memory usage between the two workflows (Wilcoxon test, *P* = 0.393). This result is now specified in the revised version of the manuscript.

Discussion

The fact that the phyla recovered by SnakeMAGs match well with reports of microbial diversity in the samples analyzed are quite appealing to me. However, do the abundances of these species also vary largely? How abundant the phyla recovered only by SnakeMAGs are? Authors should explore this factor to explain if SnakeMAGs' advantage lies in binning genomes from rare species.

Reply – We thank Reviewer 1 for this relevant comment. We now further discuss this point in the revised version of the manuscript. We found no significant difference in the relative abundance (Wilcoxon test, *P* = 0.51) of these MAGs compared to the other phyla. However, it should be noted that although *Verrucomicrobiota*, *Planctomycetota*, *Synergistota*, *Elusimicrobiota* and *Acidobacteriota* have been reported among the major taxa in termite gut, they are usually less abundant than *Spirochaetota, Firmicutes and Bacteroidota* (Arora et al, 2022, *Microbiome*) that have been recovered by both workflows. This result strongly suggests that *SnakeMAGs* is indeed not restricted to the most abundant taxa.

Competing Interests: No competing interests were disclosed.