



## SOFTWARE TOOL ARTICLE

# Large-scale quality assessment of prokaryotic genomes with metashot/prok-quality [version 1; peer review: 2 approved]

Davide Albanese , Claudio Donati

Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, TN, 38098, Italy

**V1** First published: 17 Aug 2021, 10:822  
<https://doi.org/10.12688/f1000research.54418.1>  
Latest published: 17 Aug 2021, 10:822  
<https://doi.org/10.12688/f1000research.54418.1>

## Abstract

Metagenomic sequencing allows large-scale identification and genomic characterization. Binning is the process of recovering genomes from complex mixtures of sequence fragments (metagenome contigs) of unknown bacteria and archaeal species. Assessing the quality of genomes recovered from metagenomes requires the use of complex pipelines involving many independent steps, often difficult to reproduce and maintain. A comprehensive, automated and easy-to-use computational workflow for the quality assessment of draft prokaryotic genomes, based on container technology, would greatly improve reproducibility and reusability of published results. We present metashot/prok-quality, a container-enabled Nextflow pipeline for quality assessment and genome dereplication. The metashot/prok-quality tool produces genome quality reports that are compliant with the Minimum Information about a Metagenome-Assembled Genome (MIMAG) standard, and can run out-of-the-box on any platform that supports Nextflow, Docker or Singularity, including computing clusters or batch infrastructures in the cloud. metashot/prok-quality is part of the metashot [collection of analysis pipelines](#). Workflow and documentation are available under GPL3 licence on [GitHub](#).

## Keywords

metagenome-assembled genome, MAG, genome quality, MIMAG, dereplication, completeness, contamination, nextflow, docker

## Open Peer Review

Reviewer Status  

Invited Reviewers

1

2

version 1




17 Aug 2021



report



report

1. **Thomas S. B. Schmidt** , European Molecular Biology Laboratory, Heidelberg, Germany  
**Anthony Fullam**, European Molecular Biology Laboratory, Heidelberg, Germany  
**Askarbek Orakov** , European Molecular Biology Laboratory, Heidelberg, Germany
2. **Denis Baurain** , University of Liège, Liège, Belgium

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

**Corresponding author:** Davide Albanese ([davide.albanese@fmach.it](mailto:davide.albanese@fmach.it))

**Author roles:** **Albanese D:** Conceptualization, Formal Analysis, Methodology, Project Administration, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Donati C:** Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by the Autonomous Province of Trento (Accordo di Programma).  
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2021 Albanese D and Donati C. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Albanese D and Donati C. **Large-scale quality assessment of prokaryotic genomes with metashot/prok-quality [version 1; peer review: 2 approved]** F1000Research 2021, **10**:822 <https://doi.org/10.12688/f1000research.54418.1>

**First published:** 17 Aug 2021, **10**:822 <https://doi.org/10.12688/f1000research.54418.1>

## Introduction

Genome-resolved metagenomics is one of the most promising approaches to identify and characterize novel microbial species. Large-scale environmental and host-associated studies demonstrated how metagenomics can expand our knowledge of uncultivated prokaryotes, recovering thousands of metagenome-assembled genomes (MAGs) of new archaeal and bacterial species.<sup>1,2</sup> For this reason, automated and reproducible methods for assessing the quality of MAGs play a critical role.

To recover MAGs, metagenomic sequence reads are first assembled into contigs using specific algorithms.<sup>3</sup> Contigs are then processed by tools like MetaBAT 2<sup>4</sup> or VAMB<sup>5</sup> that use tetra-nucleotide frequency (TNF) profiles and abundance patterns to group sequences that are likely to belong to the same organism (binning). Binning improves the interpretability of metagenomic data, but at the same time represents (together with assembly) a significant source of error.<sup>6</sup> Manual refinement<sup>7</sup> can increase the quality of resulting MAGs, but undermines the reproducibility of the analysis and is unfeasible for large-scale studies.

The recently introduced Minimum Information about a Metagenome-Assembled Genome (MIMAG) standard<sup>8</sup> recommends a set of measures for assessing the quality of MAGs. This comprises basic assembly statistics (e.g. N50), genome *completeness*, *contamination* and the presence of ribosomal RNA (rRNA) and transfer RNA (tRNA) genes.

Recovering this information involves computational pipelines composed of a series of specialized tools that are often difficult to use and install. Moreover, each task can require parameters and custom scripts that are often poorly documented, making reproducibility of results challenging. Tools and standards such as Galaxy,<sup>9</sup> Nextflow<sup>10</sup> and the Common Workflow Language,<sup>11</sup> coupled with container technologies like **Docker**, allows researchers to circumvent these issues, providing a way to build, run and share reproducible computational workflows.<sup>12</sup>

We present metashot/prok-quality, a comprehensive and easy-to-use Nextflow pipeline for assessing the quality of draft prokaryotic genomes. Metashot/prok-quality reports the quality statistics and estimates recommended by the MIMAG standard. Basic assembly statistics, completeness, both redundant and non-redundant contamination, rRNA and tRNA genes are reported in a single, comprehensive table.

## Methods

### Implementation

Metashot/prok-quality is written using the Nextflow domain-specific language. Nextflow is a framework for building scalable scientific workflows using containers, allowing implicit parallelism on a wide range of computing systems. Reproducibility is guaranteed by versioned Docker images, which enclose software applications together with their dependencies, allowing isolation from the host environment and portability across platforms. metashot/prok-quality v1.2.0 is composed of five main modules (**Figure 1**) and includes several custom scripts, designed to manipulate the output of the different tasks.

Software included in version 1.2.0:

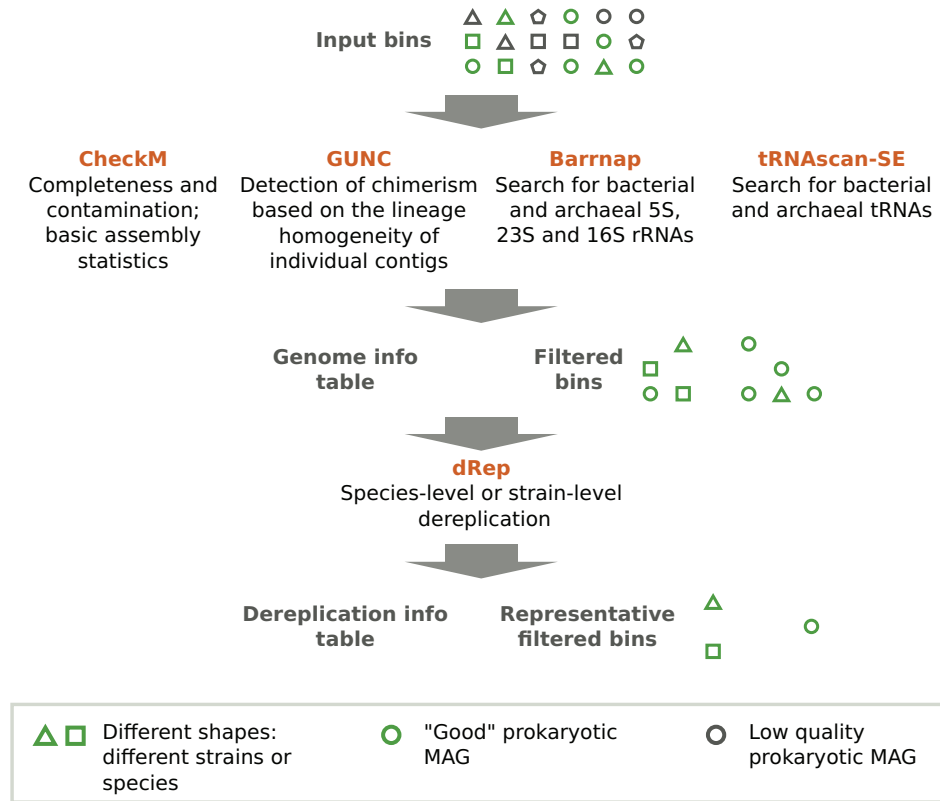
*CheckM v1.1.2.* Several tools have been developed for the assessment of completeness and contamination of MAGs. The proposed workflow includes the widely used CheckM<sup>13</sup> which estimate these metrics using ubiquitous and lineage-specific, single-copy core genes (SCGs) catalogs. CheckM is also used to recover the basic assembly statistics.

*GUNC v1.0.1.* SCG-based tools like CheckM can have very low sensitivity towards contamination by fragments from unrelated organisms (non-redundant contamination).<sup>6</sup> In order to circumvent this problem, the recent GUNC<sup>14</sup> tool was added to the pipeline. GUNC quantifies the lineage homogeneity of contigs with respect to the full gene complement, accurately detecting chimerism induced by both redundant and non-redundant contamination.

*Barrnap v0.9.* The presence of 5S, 23S and 16S rRNA genes is predicted by the BAsic Rapid Ribosomal RNA Predictor (**Barrnap**) using Hidden Markov models (HMM). Both bacteria and archaea databases are used.

*tRNAscan-SE v2.0.6.* tRNA genes are searched using tRNAscan-SE,<sup>15</sup> using bacteria and archaea covariance models. The number of tRNAs and tRNA isotypes found is reported.

*dRep v2.6.2.* Dereplication is a procedure that groups the input genomes according to their whole-genome similarity, using metrics such as the Average Nucleotide Identity<sup>16</sup> (ANI). Dereplication dramatically simplifies downstream analysis when the input genomes come from different sources.<sup>17</sup> In the proposed workflow, filtered genomes (genomes



**Figure 1. Metashot/prok-quality workflow.** The workflow takes a series of genomes (input bins) in FASTA format and returns: i) a tab-separated values (TSV) file including, for each input genome, the quality information recommended by the MIMAG standard (genome info table); ii) a directory containing the bins filtered according to the completeness and contamination thresholds; iii) a TSV file listing the cluster membership of each genome after the dereplication (optional) and iv) a directory containing the cluster representatives. The original outputs of each task (e.g. Barrnap’s GFF output) are also reported in dedicated folders.

that pass completeness, contamination and GUNC filters) are optionally dereplicated using dRep.<sup>18</sup> For each cluster, dRep reports, as the cluster representative, the best-scoring MAG using the CheckM’s quality estimates. The score is computed using the following formula:

$$\text{score} = \text{completeness} - 5 \times \text{contamination} + 0.5 \times \log(\text{N50})$$

*Python3 custom scripts.* The workflow includes three Python3 custom scripts, designed to manipulate the output of the different steps. The scripts make use of NumPy,<sup>17</sup> Pandas and scikit-learn libraries.

### Operation

metashot/prok-quality v1.2.0 requires Docker and Nextflow (tested on v20.07.1). Alternatively, the Singularity container engine can be used in place of Docker. At least 70 GB of RAM is required, a limit imposed by CheckM (v1.1.2). The workflow can run in a workstation with 16 GB of RAM using the options `--reduced_tree` and `--max_memory 16.GB`.

### Use case

As mentioned above, metagenome assembly tools combine the sequence reads into larger regions called contigs. Recently, many metagenomic assembly tools have been proposed. Amongst these, metaSPAdes<sup>3</sup> and MEGAHIT<sup>19</sup> have been shown to be able to efficiently handle large-scale short read sequencing data, producing high-quality contigs. Metagenomics contigs are then processed by tools like MetaBAT 2<sup>4</sup> in order to group sequences that are likely to belong to the same organism (binning). After binning, it is essential to assess the quality of the resulting candidate draft genomes.

In this section, we will show how to assess the quality of draft prokaryotic genomes using metashot/prok-quality. Given a series of candidate genomes in FASTA format stored in the “bins” directory, the version 1.2.0 of the workflow can be run with the following command line:

```
nextflow run metashot/prok-quality -r 1.2.0
\--genomes 'bins/*.fa'
\--outdir results
```

A series of files and directories are created in the output directory results. The main output file is “genome\_info.tsv”. This TSV file contains, for each input genome, a set of quality statistics, including completeness, contamination, GUNC filter, N50, rRNA genes found, number of tRNA and tRNA types. The columns included in this file are:

- **Genome:** the genome filename;
- **Completeness, Contamination, Strain heterogeneity:** CheckM estimates;
- **GUNC pass:** if a genome does not pass GUNC analysis it means it is likely to be chimeric;
- **Genome size (bp), ... , # predicted genes:** basic genome statistics (see <https://github.com/Ecogenomics/CheckM/wiki/Genome-Quality-Commands#qa>);
- **5S rRNA, 23S rRNA, 16S rRNA:** “Yes” if the rRNA gene was found;
- **# tRNA, # tRNA types:** the number of tRNA and tRNA types found, respectively.

The directory “filtered” contains the genomes (in FASTA format) filtered according to `--min_completeness`, `--max_contamination` and `--gunc_filter` options (see below). The TSV file “genome\_info\_filtered.tsv” includes the same information as “genome\_info.tsv”, but for the filtered genomes only. Representative (dereplicated) genomes (default ANI threshold 0.95) are reported in the “filtered\_repr” folder. The companion file “drep\_info.tsv” contains the summary of the dereplication procedure, including the genome filename, the cluster ID and the representativeness. A set of secondary directories contains the original output of each tool included in the pipeline:

- **checkm:** contains the original CheckM’s “qc” file;
- **gunc:** contains the original GUNC output file;
- **barrnap:** includes the predicted rRNA sequences for bacteria (.bac) and archaea (.arc) models in GFF and FASTA formats;
- **trnscan\_se:** includes the predicted tRNA sequences for bacteria (.bac) and archaea (.arc) models in TSV and FASTA formats;
- **drep:** dRep original data tables, figures and log file.

The command options are:

#### Input and output

- **--genomes:** input genomes/bins in FASTA format (default “data/\*.fa”);
- **--ext:** FASTA files extension, files with different extensions will be ignored (default “fa”);
- **--outdir:** output directory (default “results”);
- **--gunc\_db:** the GUNC database. If “none” the database will be automatically downloaded and will be placed the output folder (gunc\_db directory) (default “none”);

## CheckM

- `--reduced_tree`: reduce the memory requirements to approximately 14 GB, set `--max_memory` to 16.GB (default false);
- `--checkm_batch_size`: run CheckM on “checkm\_batch\_size” genomes at once in order to avoid memory issues, see <https://github.com/GenomeTools/CheckM/issues/118> (default 1000);

## GUNC

- `--gunc_batch_size`: run GUNC on “gunc\_batch\_size” genomes at once (default 100);

## Filtering

- `--min_completeness`: discard sequences with less than “min\_completeness” % completeness (default 50);
- `--max_contamination`: discard sequences with more than “max\_contamination” % contamination (default 10);
- `--gunc_filter`: if true, discard genomes that do not pass the GUNC filter (default false);

## Dereplication

- `--skip_dereplication`: skip the dereplication step (default false);
- `--ani_thr`: ANI threshold for dereplication (> 0.90) (default 0.95);
- `--min_overlap`: minimum required overlap in the alignment between genomes to compute ANI (default 0.30);

## Resource limits

- `--max_cpus`: maximum number of CPUs for each process (default 8);
- `--max_memory`: maximum memory for each process (default 70.GB);
- `--max_time`: maximum time for each process (default 96.h).

## Software availability

Source code available from: <https://github.com/metashot/prok-quality>

Archived source code at time of publication: <http://doi.org/10.5281/zenodo.4475355>.<sup>20</sup>

License: [GPL-3.0](https://www.gnu.org/licenses/gpl-3.0.html)

Docker image definitions available from: <https://github.com/metashot/docker>

## Data availability

### Underlying data

Zenodo: metashot/prok-quality v1.2.0 with test data, v1.2.0, <http://doi.org/10.5281/zenodo.4475355>.<sup>3</sup>

This project contains test data and workflow documentation.

Data are available under the terms of [GNU General Public License version 3 \(GPL-3\)](https://www.gnu.org/licenses/gpl-3.0.html).

## Extended data

Docker Hub: metashot docker images, <https://hub.docker.com/u/metashot>

This registry contains the pre-built Docker images

GitHub: metashot/docker, <https://github.com/metashot/docker>

This project contains Docker image definitions.

## Acknowledgements

The authors wish to thank Giuseppe Cossu and the Information Technology team of the Fondazione Edmund Mach for technical support.

## References

- 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 24; **176**(3): 649–62.e20.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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 Nov; **2**(11): 1533–1542.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Nurk S, Meleshko D, Korobeynikov A, *et al.*: **metaSPAdes: a new versatile metagenomic assembler.** *Genome Res*. 2017 May; **27**(5): 824–834.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- Nissen JN, Johansen J, Allesøe RL, *et al.*: **Improved metagenome binning and assembly using deep variational autoencoders.** *Nat Biotechnol*. 2021 Jan 4;  
[Publisher Full Text](#)
- Chen L-X, Anantharaman K, Shaiber A, *et al.*: **Accurate and complete genomes from metagenomes.** *Genome Res*. 2020 Mar; **30**(3): 315–333.  
[Publisher Full Text](#)
- Shaiber A, Eren AM: **Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories.** *MBio*. 2019 Jun 4; **10**(3).  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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 Aug 8; **35**(8): 725–731.  
[Publisher Full Text](#)
- Afgan E, Baker D, Batut B, *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update.** *Nucleic Acids Res*. 2018 Jul 2; **46**(W1): W537–44.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol*. 2017 Apr 11; **35**(4): 316–319.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Strozzi F, Janssen R, Wurmus R, *et al.*: **Scalable Workflows and Reproducible Data Analysis for Genomics.** *Methods Mol Biol*. 2019; **1910**: 723–745.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol*. 2020 Mar; **38**(3): 276–278.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- 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](#)
- Orakov A, Fullam A, Coelho LP, *et al.*: **GUNC: Detection of Chimerism and Contamination in Prokaryotic Genomes.** *bioRxiv*. 2020.  
[Reference Source](#)
- Chan PP, Lowe TM: **tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences.** *Methods Mol Biol*. 2019; **1962**: 1–14.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goris J, Konstantinidis KT, Klappenbach JA, *et al.*: **DNA-DNA hybridization values and their relationship to whole-genome sequence similarities.** *Int J Syst Evol Microbiol*. 2007 Jan; **57**(Pt 1): 81–91.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Evans JT, Denev VJ: **To Dereplicate or Not To Dereplicate?** *mSphere*. 2020 May 20; **5**(3).  
[Publisher Full Text](#)
- Olm MR, Brown CT, Brooks B, *et al.*: **dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.** *ISME J*. 2017 Dec; **11**(12): 2864–2868.  
[Publisher Full Text](#)
- Li D, Liu C-M, Luo R, *et al.*: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.** 2015.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Albanese D, Donati C: **metashot/prok-quality v1.2.0 with test data (Version 1.2.0).** *Zenodo*. 2021, January 28.  
[Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:  

---

Version 1

Reviewer Report 31 January 2022

<https://doi.org/10.5256/f1000research.57902.r100040>

© 2022 Baurain D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Denis Baurain** 

InBioS - PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium

This tool is useful and works as advertised by the authors. The manuscript reads well and I only have very minor comments.

## Functionality:

I tested metashot/prok-quality on a HPC facility running the SLURM grid engine. The authors' instructions were nearly enough to make it work on my first attempt. However, getting right the Singularity options was a little trickier than expected. In this regard, Docker is easier to use but generally not available in HPC environments. I contacted the authors by email (posing as a regular user) and they replied immediately, confirming that the reported issue was on my side and not a bug in their package. Unfortunately, I had to attend more pressing matters and only got back to the issue recently, hence this delayed review report. I apologize for this. For the record, here are the lines required to run the tool successfully on my system:

```
# in prok-quality/nextflow.config
singularity.enabled = true
singularity.cacheDir = "$PWD"
singularity.autoMounts = false
singularity.runOptions = "-B /path-to-user-home/prok-quality-1.2.0_with_test_data/ -B /tmpscratch/username"
```

The interface of metashot/prok-quality is straightforward and well designed, with attention to details, such as the possibility to deal with limited RAM (`--reduced_tree` option for CheckM) and to economize resources (`--gunc_db` option to avoid downloading the GUNC database multiple times). I only have a minor complaint: the `--genomes` option expects input filenames but if these do not match the `--ext` option (e.g., .fasta instead of .fa), infiles are silently ignored. In my view, the `--ext` option would make more sense with a `--genomes` option expecting an input directory. Indeed, when infiles are specified, there is no real need for filtering.



The tool's outputs are easy to understand and well documented. Testing it with default thresholds on a chimerical bacterial genome (Cornet and Baurain 2022, *Genome Biology*, in press) indeed flags it as contaminated and excludes it from the downstream dereplication step. In contrast, the two corresponding clean genomes undergo through all the steps of the pipeline, as expected.

### Manuscript:

- **Abstract:**
  - Eukaryotic sequences can be found in metagenomic samples. These are more difficult to deal with, but should be mentioned in the second sentence.
  - abstract: "... platform that supports Nextflow, Docker or Singularity" should read "... platform that supports Nextflow and either Docker or Singularity".
- **Introduction:** I would suggest adding a small paragraph about the need to use multiple tools when assessing genomic contaminations. This paragraph could include (and slightly expand) the rationale for GUNC (presently located in the Methods). For some ideas, see Lupo *et al.* (2021<sup>1</sup>).
- **Methods:**
  - Some sentences are either too much or too little, for example those about the rationale for dRep. I think that one more sentence would help and it would be better to move the whole idea of dereplication at the end of the Introduction. Indeed, Methods should not explain concepts.
  - Figure 1: The figure is nice. If I had to quibble, I would say that "filtered bins" can be misleading: here, the authors mean "bins satisfying their criteria of completeness and contamination" whereas one could imagine that they mean "bins cleaned up from contaminating sequences" (i.e., bins are passed "as is" and are not modified by the pipeline). Moreover, showing on the figure the parameters controlling the thresholds would help the reader to realize that they can be user-specified. Finally, a word is missing in the legend: "according the completeness.." should read "according to the completeness..."
  - The score formula should specify if completeness and contamination metrics are computed in percentage or not. Moreover, it might be interesting to make it user-tweakable (only a suggestion).
- **Use case:** I am not sure about the position of the backslash chars in the code snippet. I would have put them at end of lines.

### References

1. Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, et al.: Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Front Microbiol.* 2021; **12**: 755101 [PubMed Abstract](#) | [Publisher Full Text](#)

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

Yes

**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?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** bioinformatics, including genomic contaminations

**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.**

Reviewer Report 24 November 2021

<https://doi.org/10.5256/f1000research.57902.r100039>

© 2021 Schmidt T et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Thomas S. B. Schmidt** 

Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

**Anthony Fullam**

Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

**Askarbek Orakov** 

Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

The authors present 'prok-quality', a one-stop-shop Nextflow workflow for prokaryotic genome quality assessment, wrapping several state-of-the-art tools into a user-friendly pipeline. prok-quality fills two relevant niches: it provides an easy to install, easy to use interface to relevant

bioinformatics tools for non-expert users; and a portable, scalable workflow to process datasets of increasing sizes.

The code is available on github. We were able to download, install and use the workflow as advertised. We have only very minor comments, to be addressed at the authors' discretion:

- The authors restrict their discussion to metagenome-assembled genomes. However, there is no inherent reason why the prok-quality workflow shouldn't be used for reference genomes or isolate sequencing as well, and this may be worth emphasizing. In practice, many users work on large integrated datasets of MAGs and reference genomes, where fast, consistent quality control and clustering using a workflow such as prok-quality is highly relevant.
- In the bigger picture of an entire genome-resolved metagenomics workflow, from raw reads to biological analyses, prok-quality covers a reasonable chunk: the quality control and 'dereplication' (clustering) of genomes. That way, it can be used as a module, independently of upstream (assembly & binning) and downstream (annotation, analysis) tool choices. However, in our view, prok-quality would benefit from the addition of a taxonomic classifier, e.g. GTDB-tk. Taxonomic information would fit in very well with the reported quality metrics; but in particular, the workflow could (optionally) provide consensus taxonomies for dRep 95% ANI clusters which would greatly add value for non-expert users.
- Minor: the GUNC preprint has in the meantime been peer reviewed and published (Orakov *et al.* (2021<sup>1</sup>)).
- Minor: An option to switch from running with docker to running with singularity (e.g. '-profile singularity' with config profiles) would greatly enhance the ease of using this workflow as some clusters prohibit use of docker.

## References

1. Orakov A, Fullam A, Coelho L, Khedkar S, et al.: GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology*. 2021; **22** (1). [Publisher Full Text](#)

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

Yes

### **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?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** metagenomics; microbiome

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

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**