Title

Quality MAGnified

Journal

Nature Reviews Microbiology, 21(12)

ISSN

1740-1526

Author

Malmstrom, Rex R

Publication Date

2023-12-01

DOI

10.1038/s41579-023-00981-4

Peer reviewed

Quality MAGnified

This Genome Watch highlights different tools and strategies used to enhance the quality of metagenome-assembled genomes (MAGs) generated in microbiome studies.

Decreasing sequencing costs and steady improvements to contig assembly and binning algorithms have led to an explosion of metagenome-assembled genomes (MAGs) derived from microbial communities. These MAGs provide insights into the metabolic potential of community members, which is particularly important for the uncultivated microorganisms whose ecological activities might otherwise remain unknown. However, reconstructing genomes is difficult, and MAGs are not perfect. Thus, scientists must use tools to determine MAG quality based on estimated completeness and contamination levels. For instance, 'high-quality' MAGs are defined as those that exhibit more than 90% completeness and less than 5% contamination (1).

CheckM has been the primary automated tool used for assessing MAG quality since its development in 2015 (2). Briefly, CheckM determines genome completeness by measuring the fraction of nearly universally distributed, single-copy genes found in the MAG. Multiple copies of these genes in a MAG signify potential contamination. This strategy works reasonably well for some phyla, such as those well-sampled phyla used to construct the gene lists, but not always for newly discovered phyla or those with smaller genomes. Like MAGs themselves, genome quality assessments are imperfect.

CheckM2 is an updated version of CheckM that uses machine learning models, rather than gene lists, to estimate genome quality (3). The new approach surpasses the original, particularly when analysing MAGs from novel or under-sampled lineages, or lower-quality MAGs that are less complete and more contaminated. CheckM2 can also be easily updated with new genomes, which enables continuous improvement of future versions. For these reasons, CheckM2 is poised to supplant CheckM as the primary tool for MAG quality assessment.

While improving estimates of genome quality is valuable, improving actual genome quality is even more desirable. Using differential coverage is a simple method to enhance contig binning accuracy. In this approach, multiple microbiome samples are analysed collectively, and the binning algorithm considers sequencing coverage patterns when deciding how to group contigs into MAGs. That is, the algorithm expects the coverage levels of contigs belonging to the same genome to rise and fall together based on changes in the genome's relative abundance in the different samples. Multiple studies have leveraged contig coverage patterns to generate MAGs, and a recent systematic comparison of single-sample versus multi-sample coverage binning further illustrates the impact on genome quality (4). At the same level of overall sequencing effort, MAGs binned with multi-sample coverage data had substantially lower contamination levels than those binned using single-sample coverage. These results re-emphasize that, whenever possible, differential coverage patterns should be used to generate MAGs.

While differential coverage improves automated binning, it is not a panacea, and algorithms will still produce some contaminated MAGs. uBin is a new manual curation tool with a graphical user interface that lets users interactively refine MAGs based on the guanine–cytosine content, coverage and taxonomy of individual contigs (5). Like CheckM, uBin also uses a set of single-copy genes to provide feedback on how completion and contamination levels change when

removing contigs. Manual curation of MAGs from multiple studies significantly improved CheckM quality scores, illustrating the utility of this new bioinformatic tool.

The continuous technological and methodological progress in the field of metagenomics, including the development of automated tools for evaluating the quality of MAGs, will provide deeper insights into the composition and functionality of microbial communities. Thus, ensuring data quality is essential for building a solid foundation for microbiome science and translating insights into actionable goals.

References

1. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
2. Parks, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
3. Chklovski, A. et al. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
4. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat. Methods* **20**, 1170–1173 (2023).
5. Bornemann, T. L. V. et al. uBin: a manual refining tool for genomes from metagenomes. *Environ. Microbiol.* **25**, 1077–1083 (2023).