





MagCluster: a Tool for Identification, Annotation, and Visualization of Magnetosome Gene Clusters

 Runjia Ji,^{a,b} Wensi Zhang,^{a,b} Yongxin Pan,^{a,b,c}  Wei Lin^{a,b}

^aKey Laboratory of Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China

^bFrance–China Joint Laboratory for Evolution and Development of Magnetotactic Multicellular Organisms, Chinese Academy of Sciences, Beijing, China

^cCollege of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China

ABSTRACT Magnetosome gene clusters (MGCs), which are responsible for magnetosome biosynthesis and organization in magnetotactic bacteria (MTB), are the key to deciphering the mechanisms and evolutionary origin of magnetoreception, organelle biogenesis, and intracellular biomineralization in bacteria. Here, we report the development of MagCluster, a Python stand-alone tool for efficient exploration of MGCs from large-scale (meta)genomic data.

The discovery of magnetotactic bacteria (MTB) has transformed our understanding of magnetoreception, organelle biogenesis, and biomineralization in the domain *Bacteria* (1–4). MTB biomineralize intracellular, membrane-bound, nano-sized magnetic crystals called magnetosomes and are characterized by their ability to sense and swim along the geomagnetic field lines (5). Genes responsible for magnetosome biosynthesis and organization are clustered together in MTB genomes, referred to as magnetosome gene clusters (MGCs) (6). Recent advances in omics-based and cultivation approaches have led to the recovery of unprecedented amounts of (meta)genomic data, sparking a need for rapid and accurate identification and comparison of various MGCs in newly reconstructed genomes. FeGenie is a hidden Markov model (HMM)-based tool that was recently developed to identify iron-related genes, including a small group of magnetosome genes, in genomes and metagenomic assemblies (7). However, the library of FeGenie lacks accessory magnetosome genes such as *mms*, *mad*, and *man* genes, and the comparison and visualization of MGCs are not supported by FeGenie.

Here, we present MagCluster, a tool for identification, annotation, comparison, and visualization of MGCs from large-scale (meta)genomic data. MagCluster comprises three modules (Fig. 1), (i) genome annotation with Prokka (8), (ii) MGC screening with MGC_Screen developed here, and (iii) MGC comparison and visualization with clinker (9).

For genome annotation, MagCluster provides a mandatory reference file containing a total of 192 magnetosome protein sequences, including both Fe₃O₄- and Fe₃S₄-producing proteins and both core and accessory magnetosome proteins (Mam, Mms, Mad, and Man), from seven representative MTB genomes (see <https://doi.org/10.6084/m9.figshare.16863646.v3>). This magnetosome protein reference file is applied with the --proteins parameter as default during genome annotation using Prokka v1.13.4 (8).

The MGC_Screen module retrieves putative MGCs or MGC-containing contigs from GenBank files generated by the genome annotation module. MGC_Screen applies a text-mining strategy for product names containing “magnetosome” to identify putative magnetosome proteins. Because magnetosome genes are clustered together in the genome, it is a useful and robust criterion to identify MGCs based on the existence of multiple magnetosome genes adjacent to each other. MGC_Screen identifies putative MGCs based on the existence of multiple magnetosome genes (–threshold, 3 by default) in a given contig (–contiglength, 2,000 bp by default) and a given size of the

Editor Irene L. G. Newton, Indiana University, Bloomington

Copyright © 2022 Ji et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Wei Lin, weilin@mail.iggcas.ac.cn.

The authors declare no conflict of interest.

Received 24 October 2021

Accepted 15 December 2021

Published 13 January 2022

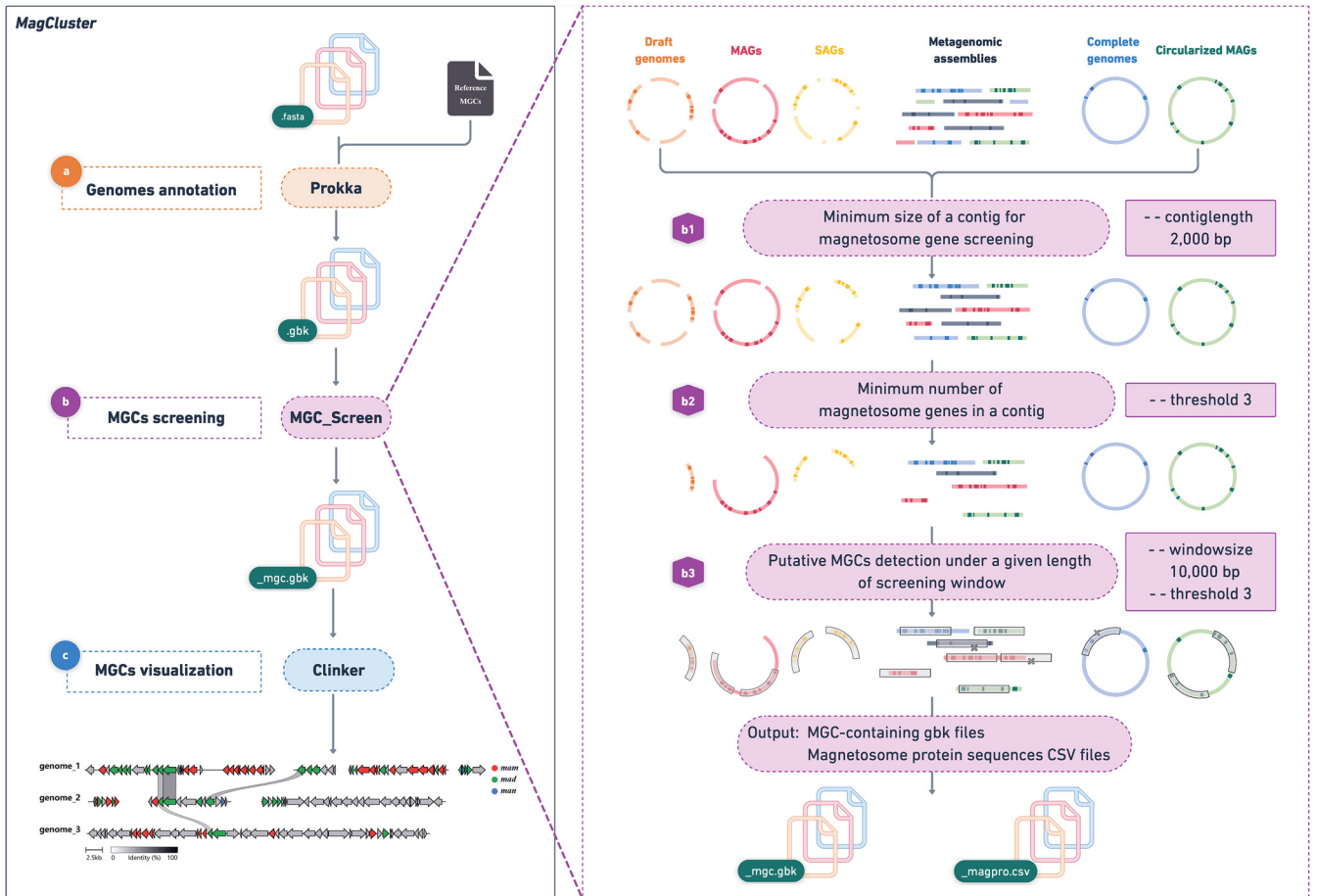


FIG 1 Overview of the MagCluster workflow. (a) Genomes are annotated using Prokka with a mandatory reference file of magnetosome proteins via `--proteins`. (b) Putative MGCs or MGC-containing contigs are retrieved by the MGC_Screen module from GenBank files generated by the annotation module. First, contigs are filtered by the contig length (`--contiglength`) and the minimum number of magnetosome genes in a contig (`--threshold`). Then, the length of a genomic region containing no less than the given number of magnetosome genes is checked to meet the value of `--windowsize`. Finally, contigs that pass all restrictions are regarded as putative MGC-containing contigs. (b1) Contigs shorter than 2,000 bp (by default) are discarded. (b2) Magnetosome genes are identified through a text-mining strategy using the keyword “magnetosome” in protein names, and contigs containing fewer than 3 (by default) magnetosome genes are discarded. (b3) Putative MGCs are screened under a 10,000-bp (by default) window, and the minimum number of magnetosome genes (3 by default) in each window size is rechecked. (c) Putative MGCs are compared and visualized using clinker. MAGs, metagenome-assembled genomes; SAGs, single amplified genomes.

sequence screening window (`--windowsize`, 10,000 bp by default) (Fig. 1b1 to b3). Users are advised to explore the different values of `--threshold` and `--windowsize` to achieve the best result. Note that, although MGC_Screen could efficiently identify putative MGCs, further manual review is necessary, considering the high level of genomic diversity of MGCs across different lineages.

MagCluster incorporates clinker v0.0.23 (9) to conduct the comparison and visualization of identified MGCs. An interactive HTML webpage is generated, where users can modify the MGC figure. Automatic modifications are conducted, including coloring the magnetosome genes and revising gene labels and legends.

Four MTB genomes (see <https://doi.org/10.6084/m9.figshare.16864372.v2>) from different taxonomic lineages were chosen to validate the effectiveness of MagCluster. MagCluster processed all four genomes and generated the MGC figure (see <https://doi.org/10.6084/m9.figshare.16831012.v2>) on a personal laptop using 6 cores and 8 GB of RAM, with a total runtime of 11 min 54.3 s.

In summary, MagCluster leverages the colocalization of magnetosome genes on the chromosome to identify MGCs, which are otherwise difficult to accurately identify based solely on sequences. MagCluster will facilitate future surveys of MGCs and MTB from large-scale (meta)genomic data.

Data availability. MagCluster can be downloaded from Python Package Index (PyPI) and Bioconda under the GNU General Public License v3.0. MagCluster is available on GitHub (<https://github.com/runjiaji/magcluster>) and Gitee (<https://gitee.com/runjiaji/magcluster>).

ACKNOWLEDGMENTS

We thank Yinzhao Wang and Jia Liu for their feedback.

This work was supported by National Natural Science Foundation of China (NSFC) grants 41822704 and 41621004 and the Youth Innovation Promotion Association of the Chinese Academy of Sciences.

REFERENCES

1. Blakemore R. 1975. Magnetotactic bacteria. *Science* 190:377–379. <https://doi.org/10.1126/science.170679>.
2. Frankel RB. 2009. The discovery of magnetotactic/magnetosensitive bacteria. *Chin J Ocean Limnol* 27:1–2. <https://doi.org/10.1007/s00343-009-0001-7>.
3. Grant CR, Wan J, Komeili A. 2018. Organelle formation in bacteria and archaea. *Annu Rev Cell Dev Biol* 34:217–238. <https://doi.org/10.1146/annurev-cellbio-100616-060908>.
4. Lin W, Kirschvink JL, Paterson GA, Bazylinski DA, Pan Y. 2020. On the origin of microbial magnetoreception. *Natl Sci Rev* 7:472–479. <https://doi.org/10.1093/nsr/nwz065>.
5. Uebe R, Schüler D. 2016. Magnetosome biogenesis in magnetotactic bacteria. *Nat Rev Microbiol* 14:621–637. <https://doi.org/10.1038/nrmicro.2016.99>.
6. Lin W, Pan Y, Bazylinski DA. 2017. Diversity and ecology of and biomineralization by magnetotactic bacteria. *Environ Microbiol Rep* 9:345–356. <https://doi.org/10.1111/1758-2229.12550>.
7. Garber AI, Neelson KH, Okamoto A, McAllister SM, Chan CS, Barco RA, Merino N. 2020. FeGenie: a comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies. *Front Microbiol* 11:37. <https://doi.org/10.3389/fmicb.2020.00037>.
8. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
9. Gilchrist CLM, Chooi Y-H. 2021. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* 37:2473–2475. <https://doi.org/10.1093/bioinformatics/btab007>.