The Journal of Open Source Software

Magphi: Sequence extraction tool from FASTA and GFF3 files using seed pairs

Magnus G. Jespersen[1], **Andrew Hayes**[1], **and Mark R. Davies**[1]

1 Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia

DOI: 10.21105/joss.04369

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @gtonkinhill
- @DavidNickle

Submitted: 17 December 2021
Published: 12 July 2022

Summary

Researchers working with genomes originating from microorganisms often work with multiple genomes in a single analysis. The number of genomes in datasets can pose challenges when it comes to extracting specific regions of interest from multiple genomes. Manual extraction of regions becomes impractical and time consuming when datasets exceed 10-20 genomes. The complexity of this task increases when working within complex regions of genomes that may not assemble into a single contiguous sequence using some existing technologies such as short read-based sequencing technologies. Therefore, automation is required as datasets of microbial genomes routinely consist of tens or hundreds of genomes. Here we present Magphi, a BLAST (Altschul et al., 1990; Mount, 2007) based contig aware genome extraction tool utilising seed sequences to identify and extract regions of interest.

Statement of need

Magphi extracts genomic regions of interest from FASTA and Gene Feature Format 3 (GFF3) files, both being common file types in bioinformatics. Packages such as Seqkit (Shen, 2016) allow for extraction and manipulation of FASTA and FASTQ files; However, such tools do not work with GFF3, or when regions of interest may span across contigs. Handling of GFF3 files are often necessary when researchers examine annotated genomes, as these are not included in FASTA formatted files.

Magphi is a command-line tool written in Python 3. It uses the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990; Mount, 2007), BEDtools (Quinlan & Hall, 2010) and implements logic to identify possible connections between given seed sequences to return the optimal solution in terms of genetic sequence and possible annotations between a set of seed sequences. Magphi can handle FASTA or GFF 3 files with included genomes, as given by the microbial annotation tool Prokka (Seemann, 2014). Magphi is contig aware, and will return a file containing the confidence level for each pair of seed sequences and genomes, providing the researcher with feedback on their run. The file containing confidence levels, distances between seed sequences, and number of annotations can be imported into Phandango (Hadfield et al., 2017), along with a phylogenetic tree of genomes for quick and visual inference of patterns or potential problems. Magphi also produces an output folder for each seed sequence pair, containing FASTA and GFF3 files when possible.

Magphi is scalable and can take multiple genomes and pairs of seed sequences. Outputs are divided by the input seed sequences for easier file management.

Acknowledgements

MGJ is supported by The Melbourne Research Scholarship from The University of Melbourne. MRD is supported by a University of Melbourne CR Roper Fellowship.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M., & Harris, S. R. (2017). Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, *34*(2), 292–293. https://doi.org/10.1093/bioinformatics/btx610

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harbor Protocols*, *2007*(7), pdb–top17. https://doi.org/10.1101/pdb.top17

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shen, S. A. L., Wei AND Le. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/q file manipulation. *PLOS ONE*, *11*(10), 1–10. https://doi.org/10.1371/journal.pone.0163962