PhyloMagnet: Fast and accurate screening of short-read meta-omics data using gene-centric phylogenetics

Max E. Schön; Laura Eme; Thijs J. G. Ettema
max-emil.schon@icm.uu.se

UPPSALA UNIVERSITET

SINGEK training network
EUROPEAN UNION

Introduction & Main Results

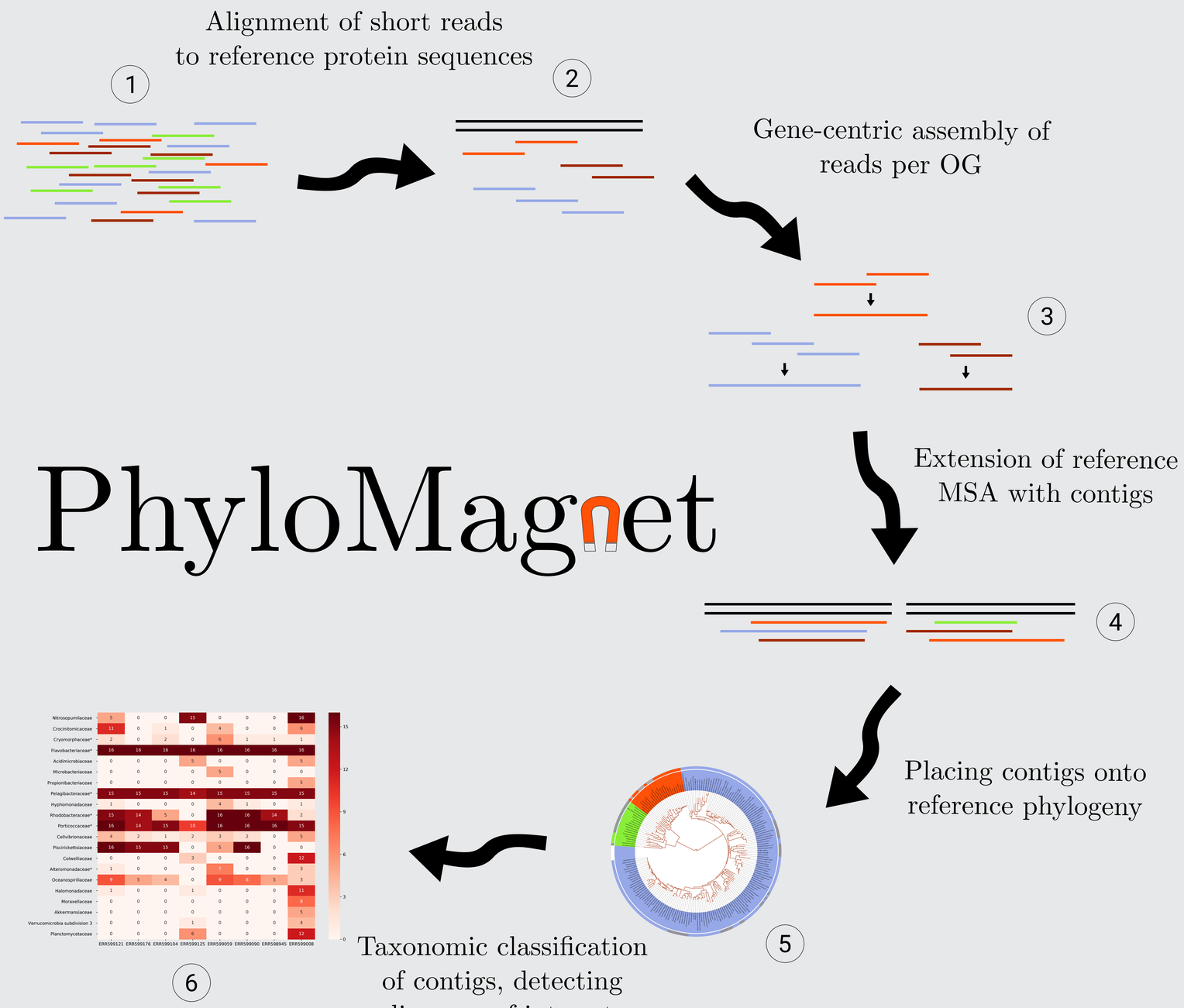
Metagenomic and metatranscriptomic sequencing analyses have become increasingly popular tools for producing massive amounts of short-read data, often used for the reconstruction of draft genomes (or MAGs) or the detection of (active) genes in microbial communities.

- Sequence assemblies of such datasets remain computationally challenging
- Often unnecessary if only a small group of organisms or genes is of interest
- Well known discrepancy between similarity-based and phylogenetic reconstruction

Here we present PhyloMagnet, a workflow to screen meta-omics datasets for taxa and genes of interest using gene-centric assembly and phylogenetic placement of sequences:

- PhyloMagnet could identify up to 87% of the genera in an *in vitro* mock community, while the false positive rate ranges from 0% to 23%.
- We could identify the same taxonomic labels as those given to MAGs reconstructed from the same samples
- The phylogenetic placement of contigs correspond to that of assembled transcript sequences

Workflow



Tree based decision

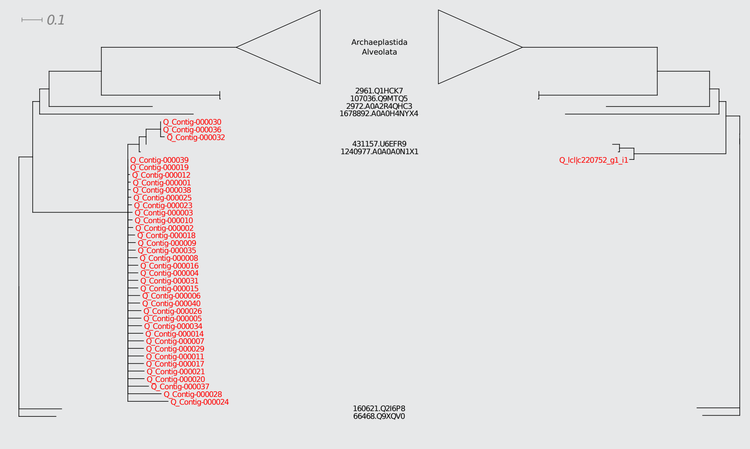


Figure 1: Phylogenetic placement of gene-centric contigs (left) and assembled transcripts (right) onto a references tree of the plastid gene psbB. Reference sequences were obtained from UniProt, transcripts and contigs were reconstructed from 26 pooled metatranscriptomes of corals and including their photosynthetic symbionts. The close branching reference sequences are from the genus *Symbiodinium*.

Performance

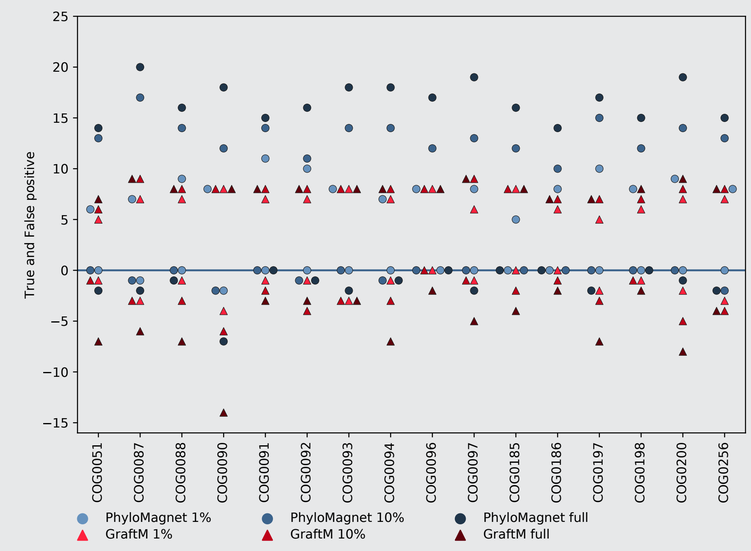


Figure 2: Classification results of PhyloMagnet and GraftM on the MBARC-26 dataset screened for 16 ribosomal proteins. True and false positive (on the negative y-axis including zero) classification results are shown for both PhyloMagnet (blue circles) and GraftM (red triangles). 3 different dataset sizes are shown by lighter (1% subsampled reads), middle (10%) and darker (full dataset) shades of the respective color.

Resources

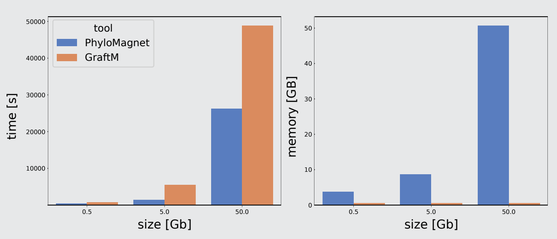


Figure 3: Runtimes (left) and memory (right) consumption of PhyloMagnet and GraftM at different dataset sizes. PhyloMagnet is faster, with the drawback that it consumes more memory (mostly in the overlap assembly step).

Conclusion

PhyloMagnet can accurately identify the organisms and genes present in metagenomic and metatranscriptomic data by assembling genes and placing them into a phylogenetic tree. By using PhyloMagnet to screen samples, researchers can make more efficient use of available resources and data.

References

[1] P. Barbera, *et al.*, *Systematic Biology* **68**, 365 (2019).
[2] S. A. Berger, A. Stamatakis, *Bioinformatics* **27**, 2068 (2011).
[3] J. A. Boyd, B. J. Woodcroft, G. W. Tyson, *Nucleic Acids Research* **46**, e59 (2018).
[4] B. Buchfink, C. Xie, D. H. Huson, *Nature methods* **12**, 59 (2015).
[5] L. Czech, A. Stamatakis, *bioRxiv* pp. 1–36 (2018).
[6] P. Di Tommaso, *et al.*, *Nature Biotechnology* **35**, 316 (2017).
[7] D. H. Huson, *et al.*, *Microbiome* **5**, 11 (2017).
[8] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, *Bioinformatics* pp. 1–3 (2019).

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675752.

github.com/maxemil/PhyloMagnet
phylomagnet.readthedocs.io/en/latest/

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY