Impact of microbial genome completeness on functional metagenomics

Raphael Eisenhofer
https://orcid.org/0000-0002-3843-0749
Iñaki Odriozola
University of Copenhagen https://orcid.org/0000-0002-5289-7935
Antton Alberdi (antton.alberdi@sund.ku.dk)
University of Copenhagen https://orcid.org/0000-0002-2875-6446

Brief Communication

Keywords:

Posted Date: August 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1955526/v1

Abstract

Inferring the functional capabilities of bacteria from metagenome-assembled genomes (MAGs) is becoming a central process in microbiology. Here we show that the completeness of MAGs has a significant impact on the recovered functional signal, spanning all domains of metabolic functions. We identify factors that affect this relationship between genome completeness and function fullness, and provide baseline knowledge to guide efforts to correct for this new-found bias in functional metagenomic analyses.

Full Text

Genome-resolved metagenomics enables draft bacterial genomes to be reconstructed from DNA sequence data derived from complex microbial mixtures [1]. The metagenome-assembled genomes (MAG) derived from such a process can be annotated to predict their functional toolbox upon which microbiome-level functional analyses can be conducted [2,3]. One of the main issues of this methodology is that MAGs usually display different levels of genome completeness, i.e., the entirety of a microbe's DNA is not always captured in the reconstructed genome [4]. MAG completeness is primarily estimated through the presence of single-copy core genes (SCCGs), which are expected to be found in most bacteria [5]. It is common to use MAGs with completeness values as low as 70% for the functional analyses of microbial communities [6]. However, if a MAG is estimated to be 70% complete, it is probable that many of the actual functions encoded in the genome will not be captured in the MAG, and thus the functional capacity of the genome will be underestimated [3,7]. Not accounting for the level of completeness of MAGs could therefore lead researchers to incorrect interpretations of results, such as the artifactual deficit of functions being misinterpreted as real biological signal.

A major challenge of metagenomic research is correcting or accounting for biases in statistical analyses and modelling. However, we currently ignore how the loss of functional capacities is correlated with MAG completeness, and whether these relationships are constant or variable across microbial communities, metabolic domains, and microbial phylogeny. To address these issues, we investigated the relationship between estimated genome completeness and metabolic function fullness (defined as the proportion of biochemical reactions enabled by the genes present in a genome to accomplish a metabolic function) using 1680 MAGs derived from five animal host species; namely, *Apodemus sylvaticus* (448 MAGs), *Crocidura russula* (87), *Felis catus* (230), *Gallus gallus* (724) and *Mus musculus* (191) (Fig. 1, Supplementary Table S1). MAG completeness was estimated using CheckM [5], while functional fullness of KEGG modules was estimated using DRAM [2,8]. Only MAGs with CheckM contamination values under 10% and completeness values >70% were used in the analyses (Fig. S1).

We employed linear mixed models to understand the association of genome completeness and function fullness in a filtered set of 138 KEGG metabolic modules across 1554 MAGs (details in the Online Methods). We fitted two models with alternative host-level random effects: one with a random intercept

(random =1|Host), which assumed a constant fullness-completeness relationship across datasets; and the other with a random slope (random = Completeness|Host), which assumed host-specific responses. Modules showed a constant or near-constant fullness-completeness relationship across hosts, suggesting that the observed associations are not or are minimally host-specific. The random intercept model exhibited lower AIC than the random slope model in 56% of the modules, and for most of the modules supporting the model with random slope, the fullness-completeness relationships were not strikingly different between hosts (Fig. S2).

The models estimated a positive relationship between genome completeness and metabolic function fullness for 97% of the studied modules in Actinobacteriota, Bacteroidota and Proteobacteria and 92% of the modules in Firmicutes, spanning all functional domains and levels of complexity (i.e., number of enzymatic steps). Overall, the increase of completeness from 70% to 100% was associated with a 23±13% (mean±sd) increase in module fullness. This relationship remained constant across the completeness gradient, with a slight tendency for the slope of the relationship to increase with completeness (Fig. 2A). This indicates that while increasing the completeness threshold to exclude MAGs from functional analyses minimises the issue, the problem persists, even when only considering 'high quality' (>90%) MAGs. We also found evidence for significant differences between the fullness-completeness relationship across bacterial phyla. Considering all functional traits analysed, Proteobacteria showed the overall strongest fullness-completeness relationship followed by Actinobacteriota, Bacteroidota and Firmicutes (Fig. 2B).

Similarly to taxonomic differences, the fullness-completeness relationship did not change evenly across metabolic domains. The fullness of the modules belonging to the 'nucleotide metabolism' domain were the most affected by completeness, followed by 'carbohydrate metabolism', 'amino acid metabolism', and 'metabolism of cofactors and vitamins' (Fig. 2C). Modules of the domain 'biosynthesis of terpenoids and polyketides' followed by 'energy metabolism' showed the weakest fullness-completeness association. In addition, the complexity of the modules was negatively associated with the fullness-completeness relationship (Fig. 2D). This suggests that the fullness of the modules with the fewest steps are the ones that are more severely affected by genome incompleteness.

Our results highlight the need to consider MAG completeness when comparing the functional capacities between microbial genomes or metagenomes. We argue that completeness biases should be accounted for in functional diversity analyses, analogously to how DNA sequencing depth biases are considered in diversity modelling approaches [9]. At the very least, our observations urge scientists to revisit whether the functional differences observed across contrasting treatments, hosts, or populations are driven by differences in MAG completeness. Ideally, our results should serve as a baseline to account for completeness in statistical modelling, and when enough information is available, to reconstruct missing functions in MAGs through functional imputation. Only through the correction and mitigation of the functional biases introduced by uneven MAG completeness will researchers be able to robustly characterise, model, and assess the functional capabilities of microbial communities.

References

1. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**, 37–43 (2004).
2. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. **48**, 8883–8900 (2020).
3. Belcour, A. *et al.* Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species. Elife **9**, e61968 (2020).
4. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. Appl. Environ. Microbiol. **87**, (2021).
5. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **25**, 1043–1055 (2015).
6. Levin, D. *et al.* Diversity and functional landscapes in the microbiota of animals in the wild. Science (2021) doi:10.1126/science.abb5352.
7. Zhou, Z. *et al.* METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome **10**, 33 (2022).
8. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**, 27–30 (2000).
9. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput. Biol. **10**, e1003531 (2014).

Figures

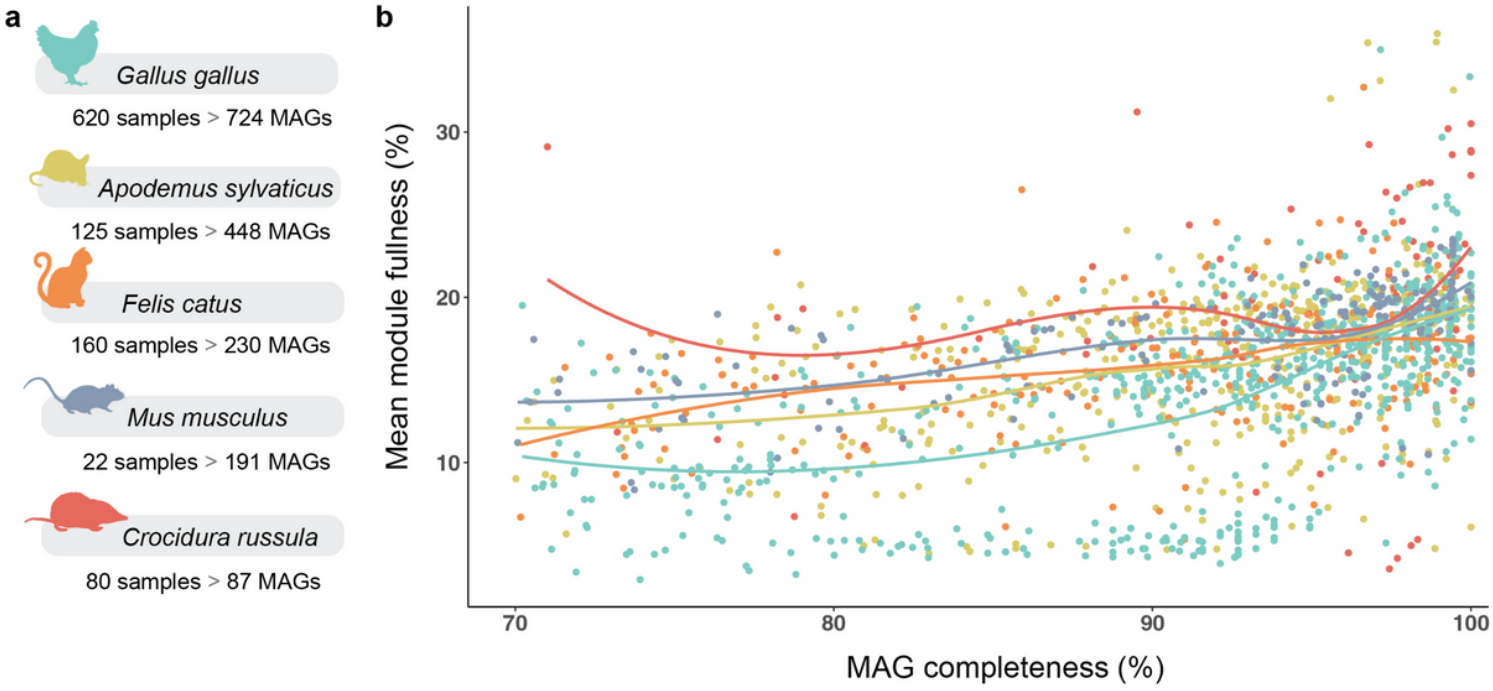

Figure 1

Origin and general statistics of the dataset.

(**a**) Overview of the number of samples and MAGs belonging to each host species used in this study. (**b**) Relationship between genome completeness and mean KEGG module fullness. Dots indicate individual metagenome-assembled genomes (MAGs) within each dataset (colour-coded). Solid lines indicate LOESS smoothed mean for each dataset.

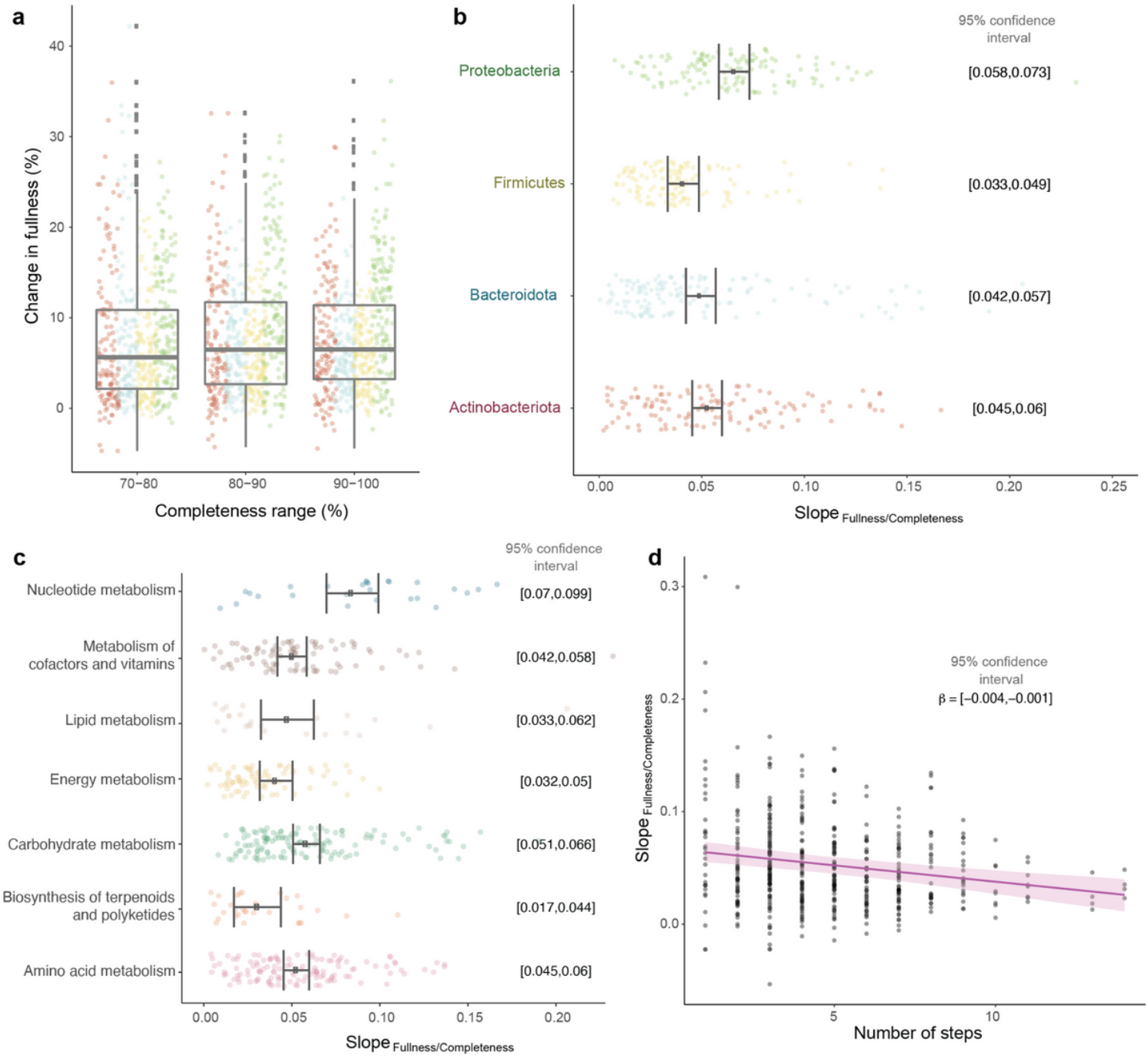

Figure 2

Relationship between function fullness and MAG completeness.

(**a**) Percentage of change in module fullness across different MAG completeness windows. (**b**) Mean slope variation of the fullness/completeness relationship across bacterial Phyla. (**c**) Mean slope variation of the fullness/completeness relationship across functional domains. (**d**) Relationship between the slope of the fullness/completeness relationship and the number of steps of the modules.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2022CompletenessfunctionSupplementaryInformation.docx