Reproducible analysis of metatranscriptomics either through nf-core/metatdenovo or nf-core/magmap

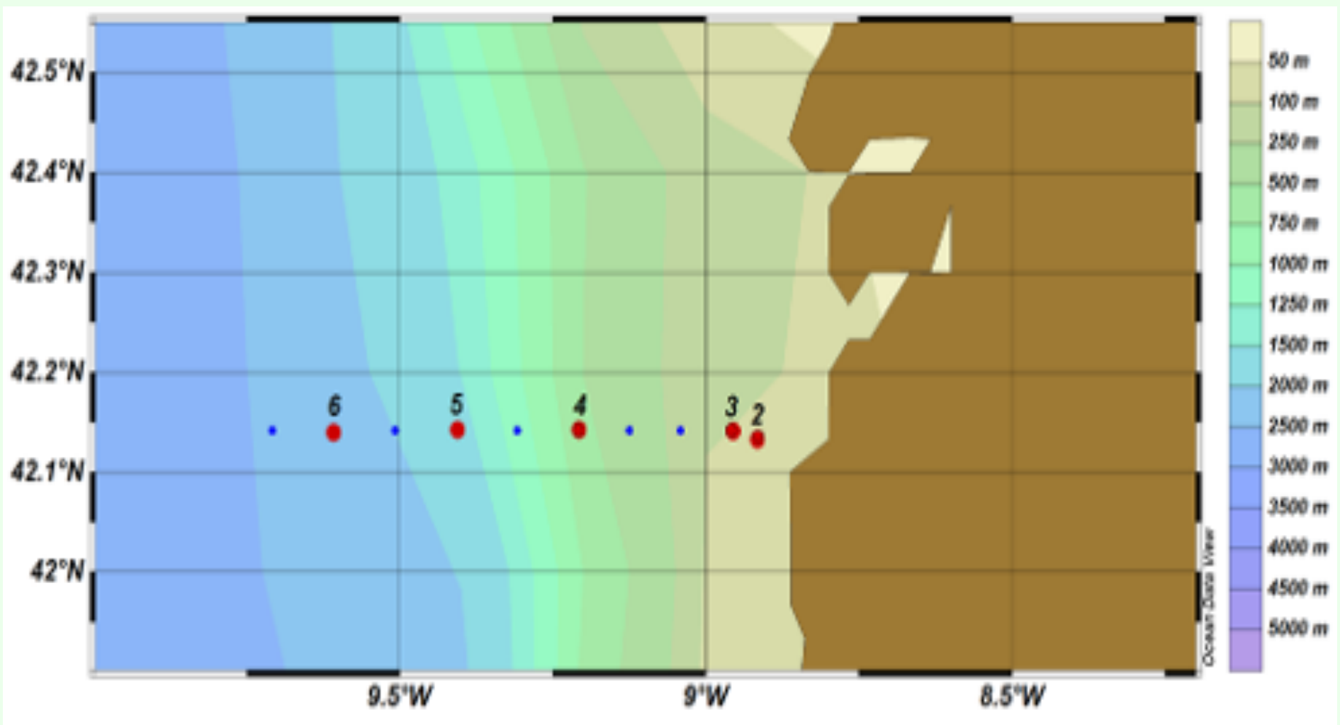
Danilo Di Leo & Emelie Nilsson, Jarone Pinhassi, Daniel Lundin
Linnaeus universty, Kalmar, Sweden

Danilo Di Leo, MSc
PhD Student in Marine Microbiology
Linnaeus University
Dept. of Biology & Environmental Science
Universitetsplatsen 1, 392 31 Kalmar, Sweden
Mobile: +46 (0) 72 369 61 08
danilo.dileo@lnu.se
www.lnu.se

Introduction

In the last decade, the study of microbial communities through RNA sequencing has significantly increased. Metatranscriptomics offers insights into metabolic processes within microbial communities, providing a snapshot of gene expression based on in situ environmental conditions. To support biologists in this endeavor and to promote reproducibility and standardization in data analysis, we developed two complementary pipelines with the help of the nf-core community: nf-core/metatdenovo and nf-core/magmap. These pipelines, designed to be user-friendly and reproducible, aim to investigate the activity of microbial communities with varying levels of genomic knowledge.

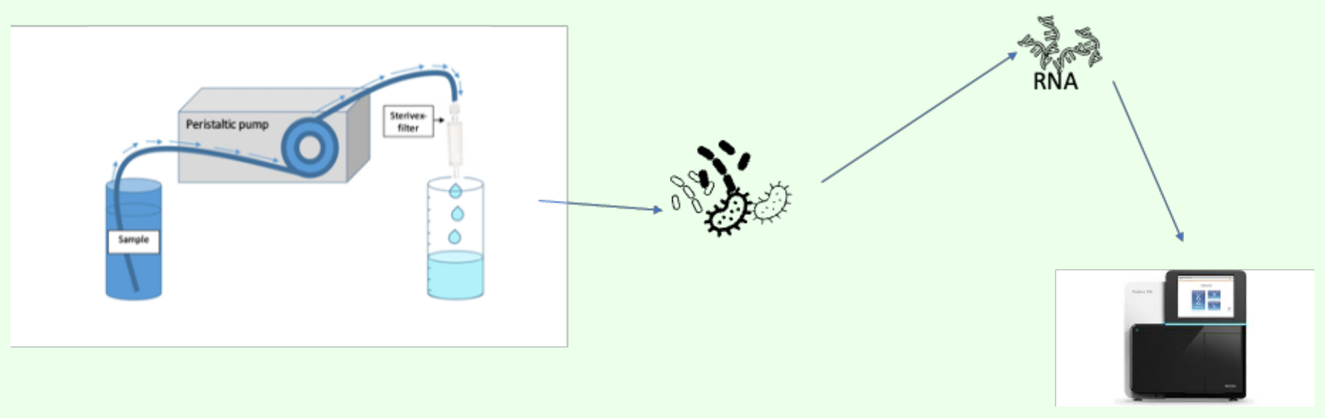
Study site



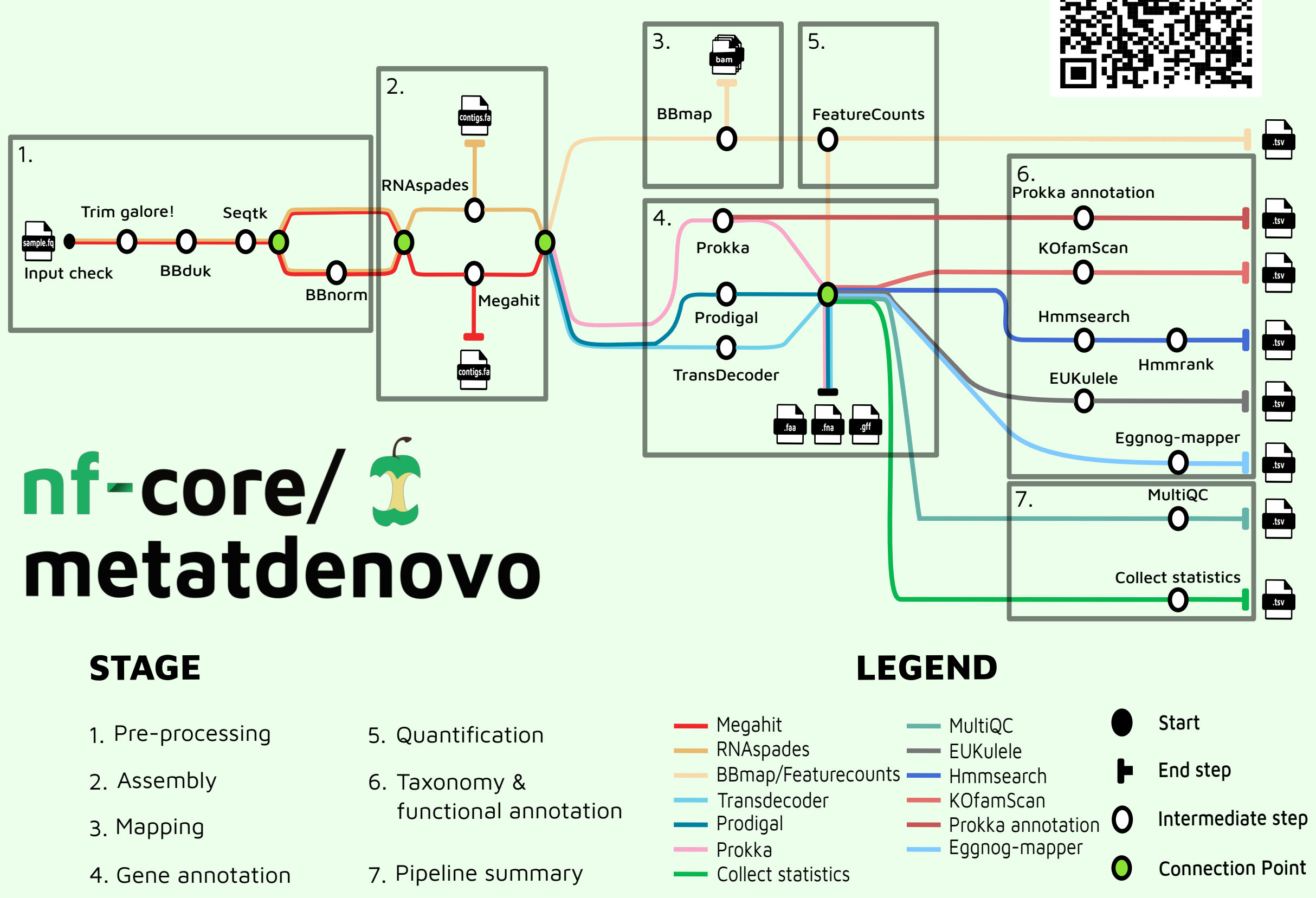
To compare the two pipelines, we chose a metatranscriptomic dataset from the north-west of Spain. The samples come from two different cruises in winter and spring at two different sites (3 and 6) and one mesocosm experiment M.

What is metaT?

Metatranscriptomics is the study of RNA transcripts present in a microbial community at a given time. It provides insights into which genes are being actively expressed by different organisms in the environment, allowing us to understand the functional activities of microbes. In our field, we sample water from the environment and, after filtering through an 0.22 um filter, we extract mRNA and sequence it.

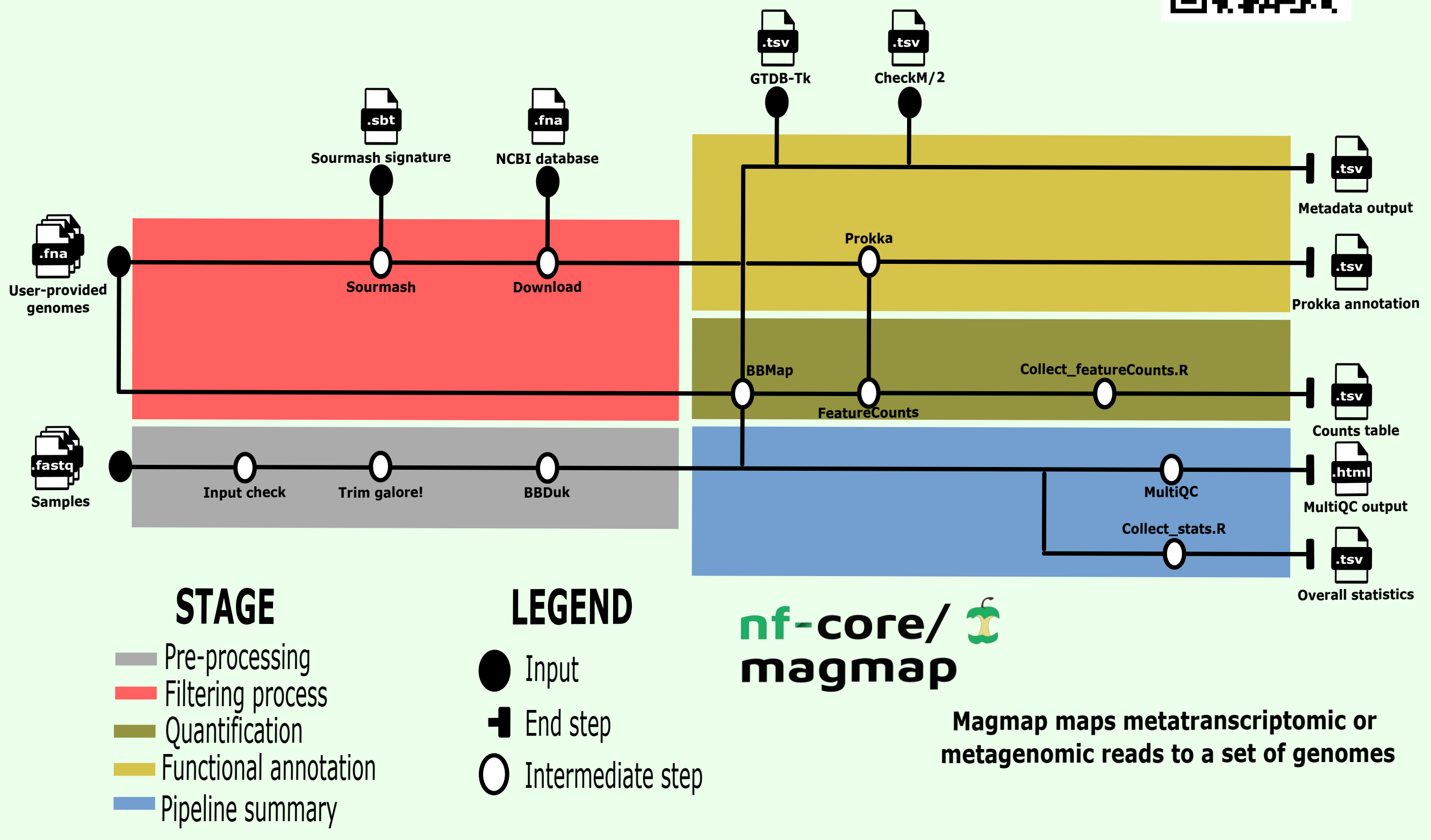


Nf-core/metatdenovo



Nf-core/metatdenovo creates an assembly *de novo*. ORFs are called, annotated taxonomically and functionally, and quantified by mapping the reads back. This pipeline suits samples from environments for which genomes databases are not sufficiently covered by reference genomes such as deep sea, or soil sediments.

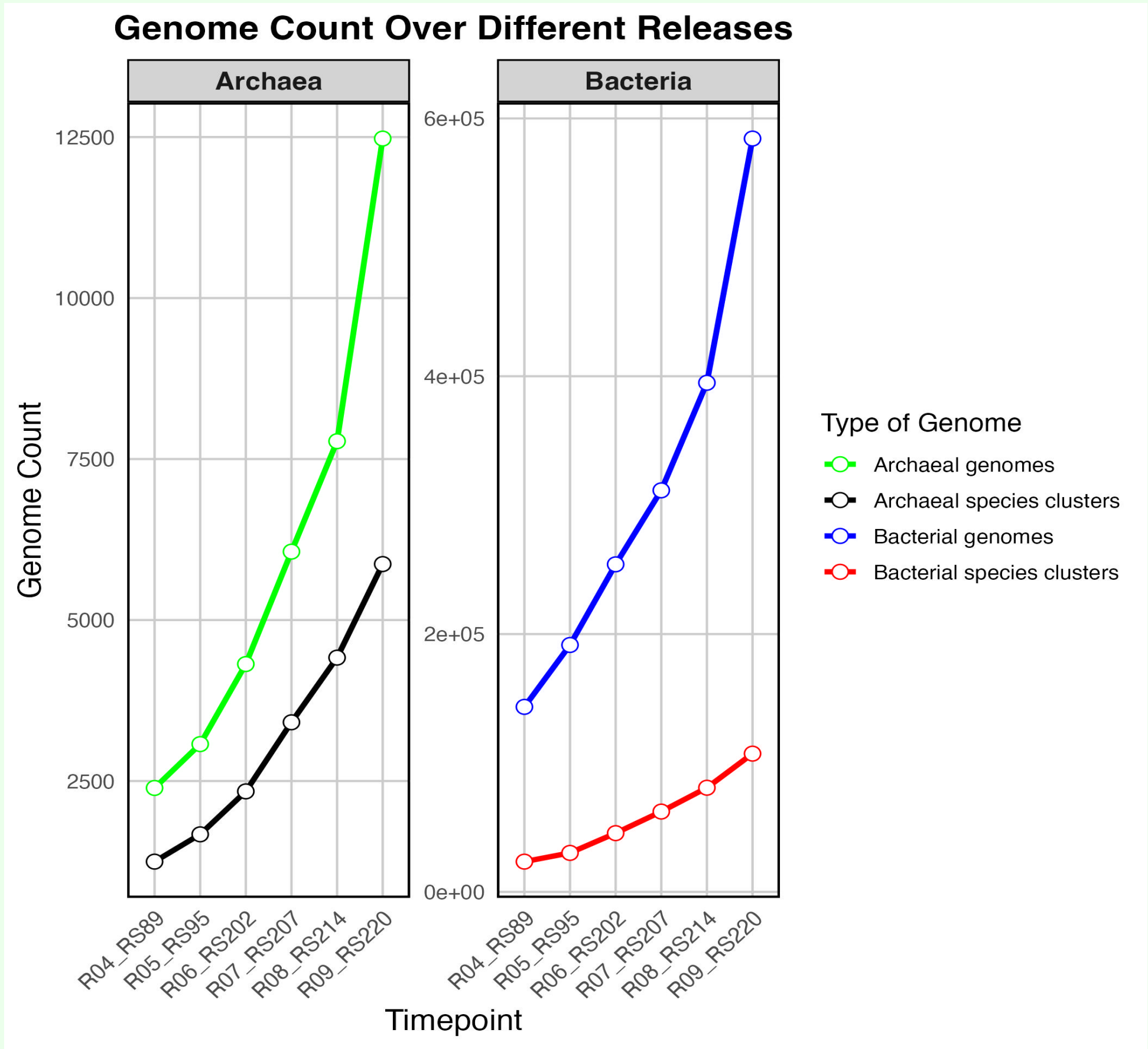
Nf-core/magmap



Magmap maps metatranscriptomic or metagenomic reads to a set of genomes

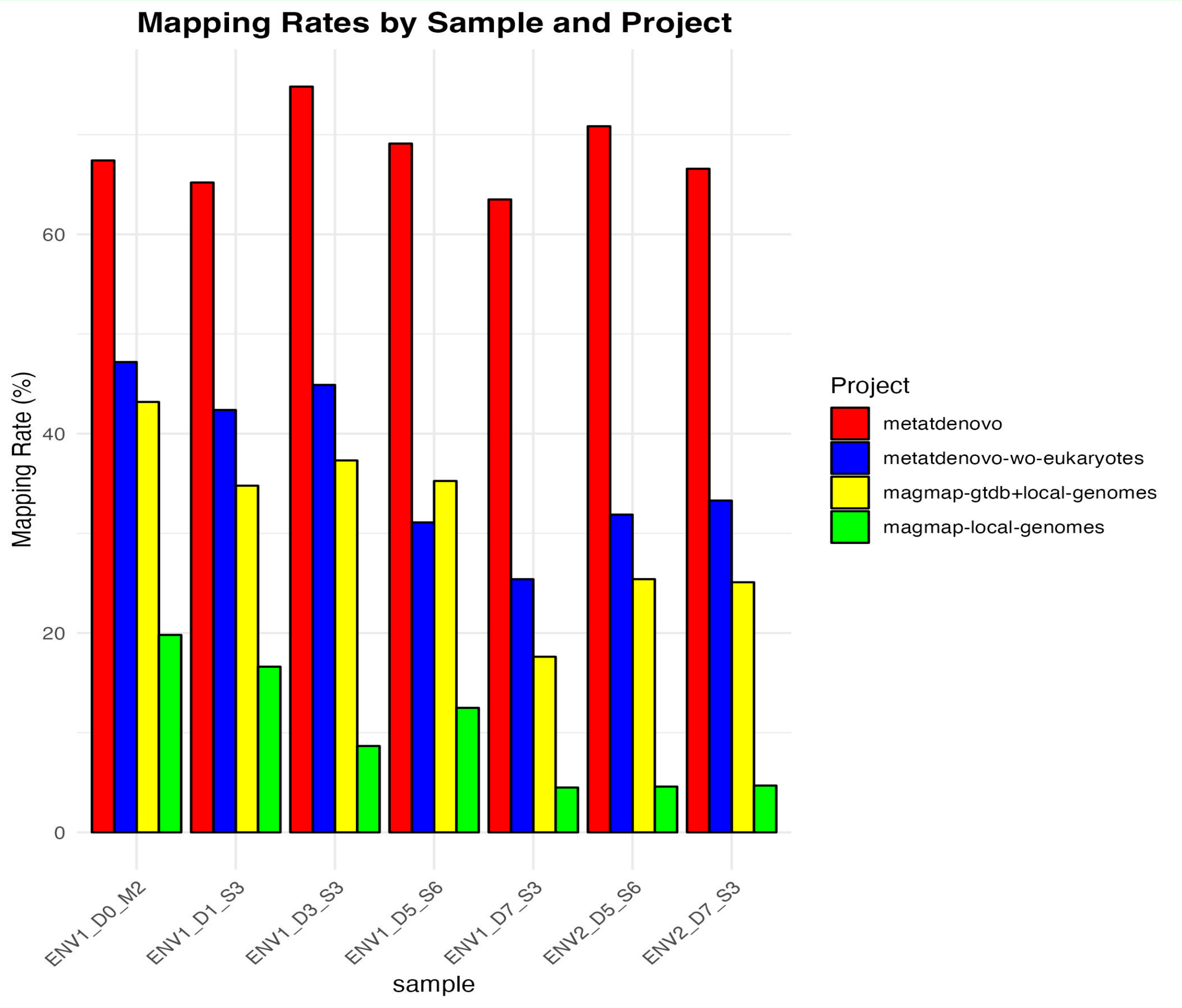
Nf-core/magmap maps the metatranscriptomics reads to reference genomes. It can use either metagenomes from the study site or external genomes. Target genomes can be selected with the Sourmash tool. It is recommended to use for datasets from environments well represented with genomes, such as gut microbiomes or surface water.

Growth of genomes



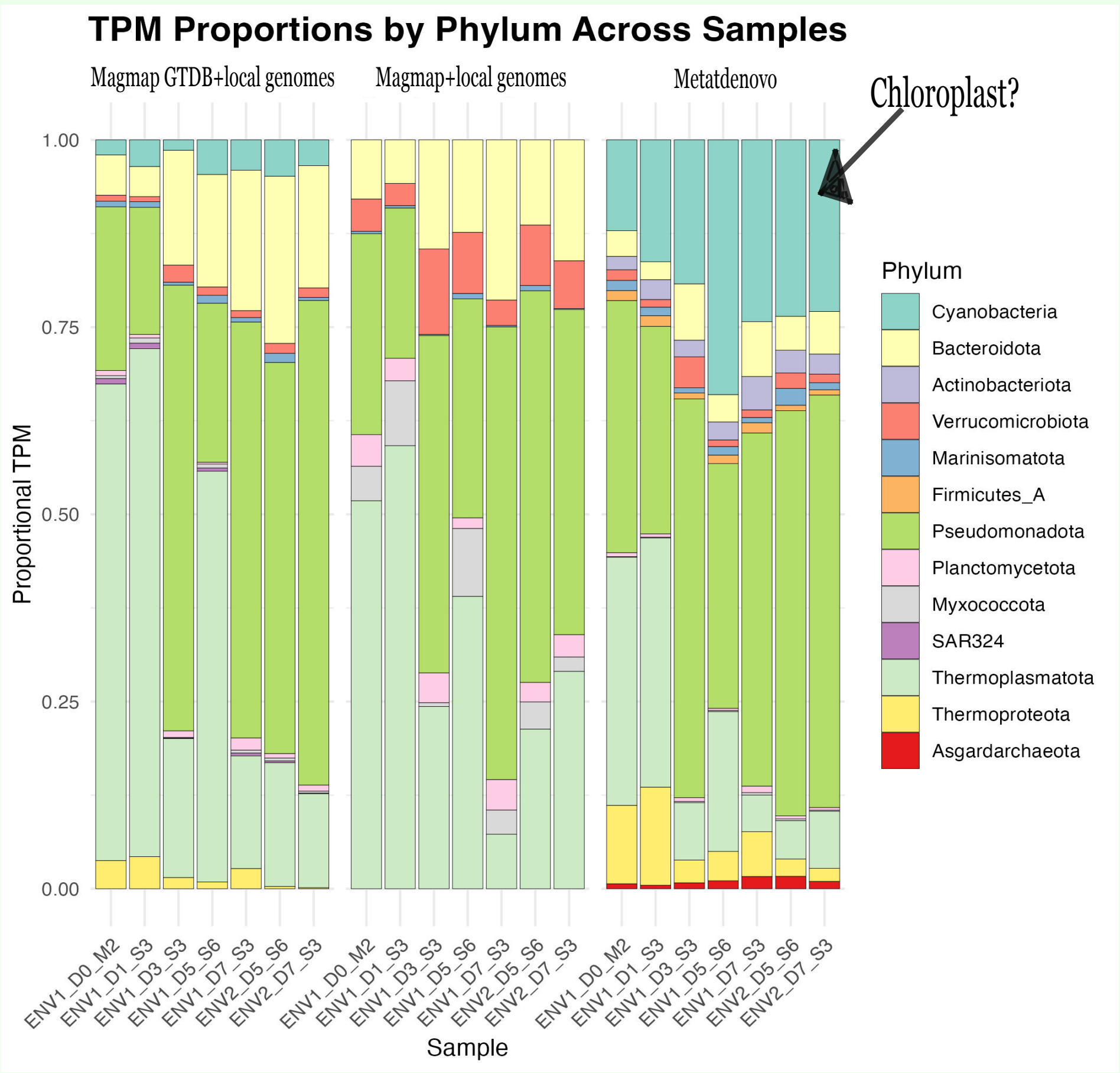
Genomes from Archaea and Bacteria grow exponentially, opening more and more environments to mapping with nf-core/magmap.

Mapping rate



The mapping rate from nf-core/metatdenovo was higher than nf-core/magmap. Most of this was, however, explained by the presence of eukaryotes. For prokaryotes only, mapping to a combination of metagenomes from the same cruises and public genomes was very similar to nf-core/metatdenovo, whereas mapping to only metagenomes was much lower.

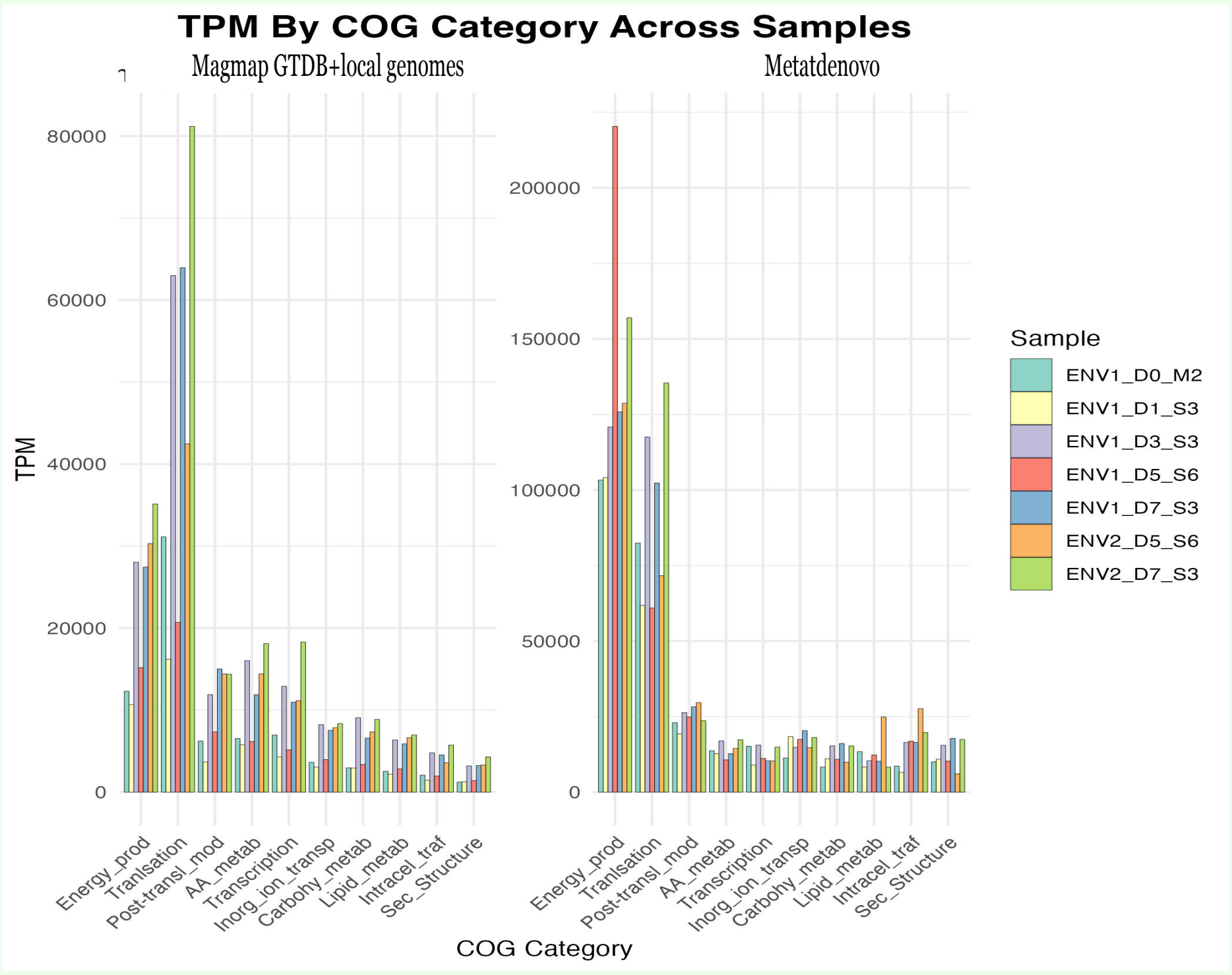
Taxonomy



In the nf-core/metatdenovo output, cyanobacteria are much higher than in magmap. This is possibly chloroplast sequences. Furthermore, using only local metagenomes resulted in lower diversity.

Function

Energy production and Translation are the most abundant functional categories. The balance between the two are different in the nf-core/metatdenovo and nf-core/magmap analyses, potentially due to the presence of photosynthetic eukaryotes in the nf-core/magmap output.



Conclusion

Our test study shows how in a surface marine water environment both nf-core/metatdenovo and nf-core/magmap performed well, but with important differences. The user has three strategies to use depending on the availability of reference genomes: 1) Are the communities well covered by metagenomes from the same environments: use nf-core/magmap with them, 2) Are they well covered by public genomes: use nf-core/magmap with them (plus any local), 3) Otherwise, use nf-core/metatdenovo. Mixed communities with prokaryotes and eukaryotes are likely better with nf-core/metatdenovo, while nf-core/magmap gives important genomic context.