Establishing a metagenomic workflow for eukaryotic analysis in drinking water system

Marco Gabrielli[1, 2], Solize Vosloo[2], Manuela Antonelli[1], Ameet Pinto[2]

[1] Department of Civil and Environmental Engineering (DICA) - Environmental Section, Politecnico di Milano, Milan, Italy
[2] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

The analysis of metagenomic short-reads derived from drinking water systems has traditionally focused on the study of prokaryotes (i.e., bacteria and archaea) neglecting the presence of eukaryotes, even though such (micro)-organisms can have an impact on water quality. This limitation stems from the fact that traditional prokaryotic-centric bioinformatic tools are not suited for eukaryotic genomes due to their higher complexity and typically lower abundance in comparison with prokaryotes. Noticeably, while established bioinformatic pipelines exist for the analysis of prokaryotic-derived genomes, only recently researchers have started to develop tools suitable for the identification and subsequent analysis of eukaryotic populations in mixed communities; for this reason, there is currently no validated and widely accepted bioinformatic workflow for analyses of eukaryotes in metagenomes. In this study we propose a workflow tailored for the identification and analysis of eukaryotic genomes derived from metagenomic drinking water system surveys by comparing the performances of several tools for the identification of eukaryotic contigs, contigs binning, and gene prediction on both synthetically generated metagenomes and a representative drinking water system case study, consisting of samples collected for over 6 months. Our findings indicate that the performance of eukaryotic-centric tools differs between the analysis of real-world and synthetic data. However, an ensemble approach involving the combination of reference-dependent and reference-independent tools was found to be beneficial for the eukaryotic identification and analysis and was included in the workflow. For example, several *k*-mer-based (i.e., EukRep, Tiara, Whokaryote) and a reference-based (i.e., CAT) contigs classifiers were selected for the identification of eukaryotic contigs, leveraging the benefits of both approaches. Overall, the workflow proposed in this study will enable a more systematic metagenomic characterization of eukaryotic populations in the drinking water microbiome and thus help to better understand their ecological role in drinking water systems.