Large-scale analysis of genetic diversity in Patescibacteria across environments

Lodovico Sterzi[1,2], Diego Marco Minore[1], Clara Bonaiti[1], Simona Panelli[1], Francesco Comandatore[1]

[1] Department of Biomedical and Clinical Sciences, Pediatric Clinical Research Center "Romeo and Enrica Invernizzi", University of Milan, Milan, Italy
[2] Department of Evolutionary Biology, Ecology and Environmental Sciences, Biology Faculty, University of Barcelona, Barcelona, Spain

Patescibacteria, also referred to as the Candidate Phyla Radiation (CPR), represent a vast monophyletic division within the bacterial domain, comprising diverse lineages with reduced genomes, limited metabolic capabilities, and symbiotic lifestyles. Due to these features, CPR bacteria remain largely uncultivated and are often underdetected or misclassified in 16S rRNA gene surveys. Thus, these bacteria are commonly detected and analysed using shotgun metagenomics. Despite several reports of CPR occurrence across diverse natural and human-associated sources, their global distribution and habitat preferences remain incompletely understood.

Here, we performed a large-scale study on publicly available metagenomic datasets to investigate the environmental distribution and ecological associations of CPR lineages. We developed a machine learning–based classification approach leveraging the RecA protein as a marker for the detection and classification of CPR bacteria. We applied this approach to the MGnify protein database, which contains protein sequences from tens of thousands of metagenomic samples with associated biome information, to describe CPR diversity across environmental sources. Our results indicate that CPR bacteria are widespread in freshwater environments, with lineage-specific enrichments in wastewater and human microbiomes, alongside an expansion of human-associated lineages, consistent with potential adaptation to the human host. Overall, we provide a comprehensive overview of the global distribution patterns in CPR bacteria by applying a robust marker-based bioinformatic pipeline on an extensive metagenomic database.