A container-based pipeline for bacterial genome assembly and annotation

Felipe Marques de Almeida, Georgios Joannis Pappas Junior

Universidade de Brasília

Abstract

Advances in DNA sequencing technologies are reshaping bacterial genomics studies enabling chromosome level assemblies, at a fraction of cost and time, paving the way to population level genomic surveys. At the present, the computational analysis of sequencing data is the main hindrance to the field and withholds its move into mainstream clinical settings. To overcome this barrier we developed a complete container-based pipeline for bacterial genomics analysis, meaning that given raw sequencing data from multiple platforms (Illumina, Pacbio and Oxford Nanopore), it performs genome assembly and annotation, enabling identification and visualization of antibiotic resistance genes, virulence factors, prophages and integrative elements. In general terms the annotation phase of the pipeline can be executed in a few hours in a laptop. The assembly module, despite requiring a large amount of memory (>64 Gb RAM), can be executed in a day. The pipeline is designed to be modular, taking into account different analytical scenarios readily configured by the user. We also leverage the use of operating system virtualization meaning that there is no need for user installation of required pipeline components. When ready, all modules will be made available through GitHub. In conclusion, this pipeline offers a seamless exposition of computational tools to bridge the gap toward routine bacterial genomics.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Universidade de Brasília (UnB)