Dugong: a Docker image, inspired on Ubuntu Linux, designed to enhance reproducibility and replicability during computational analyses of biological data

Fabiano Bezerra Menegidio [1], Luiz R. Nunes [2]

[1]Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes, Brasil, [2] Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, Santo André, Brasil

The increasing use of computational methods for the analysis of biological data and the constant expansion verified in the fields of Bioinformatics and Computational Biology have revolutionized the study of Biology during the past few decades. However, grasping the complex nature of some softwares employed for such analyses and adapting to the rapid changes observed in computational ecosystems has become a major challenge for biologists, hampering the full use of such resources by the scientific community. This problem becomes more serious due to the fact that many computational methods often rely on pipelines composed by multiple analytical steps, involving different scripts, softwares and/or algorithms with unique requirements and/or dependencies. As a result, utilization of computational resources during bioinformatics analyses has become increasingly heterogeneous across laboratories, compromising reproducibility and replicability of results obtained from a given experiment or dataset. Although the Bioinformatics community has heavily relied on the production of Open Source softwares as a way to minimize these problems, mandatory installation of libraries for the proper functioning of scripts, lack of proper documentation and incompatibility with different operating systems and/or hardware still represent major obstacles to ensure replicability and reproducibility of data analysis in different computational environments. Fortunately, emergence of the Docker project is providing a promising new strategy to tackle these problems, by allowing the configuration of a complete computing environment, in which all libraries, codes and additional data required for a particular application may be implemented in a single container, which can be consistently exchanged and launched in different platforms, regardless the specificities of their hardware and/or operating systems. Thus, to explore and demonstrate the usefulness of Docker-based systems as a strategy to enhance replicabability and reproducibility of bioinformatics analyses in multiple computing environments, we developed the application Dugong, a Docker image based on Ubuntu 15.10, specifically designed for the analysis of large-scale biological data. Using a graphic interface generated by Xfce4, Dugong provides the managers Linuxbrew (Homebrew Science) and Conda (Bioconda), which allow distribution and installation of over 3000 bioinformatics-related packages and libraries, with automated installation of their respective dependencies. Simulations performed in virtual machines demonstrate that Dugong allows effective creation of reusable containers for different bioinformatics analyses in a uniform computational environment, allowing acquisition of consistent and reproducible results by the scientific community, thus assisting in the development of Open Science projects.