Using the Nextflow framework for reproducible in-silico omics analyses across clouds and clusters

Paolo Di Tommaso[1], Evan Floden[1,2], Maria Chatzou[1,2] , Cedric Notredame[1]

[1] Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.
Email: paolo.ditommaso@crg.eu.

Project Website: http://www.nextflow.io
Source Code: https://github.com/nextflow-io/nextflow
License: GPLv3

Reproducibility has become one of biology's most pressing issues. This impasse has been fueled by the combined reliance on increasingly complex data analysis methods and the exponential growth of biological datasets. When considering the installation, deployment and maintenance of bioinformatic pipelines, an even more challenging picture emerges due to the lack of community standards. The effect of limited standards on reproducibility is amplified by the very diverse range of computational platforms and configurations on which these applications are expected to be applied (workstations, clusters, HPC, clouds, etc.). With no established standard at any level, diversity cannot be taken for granted.

Nextflow is a pipeline orchestration tool that has been designed to ease deployment and guarantee reproducibility across platforms. It is a computational environment, which provides a domain specific language (DSL) to simplify the writing of complex distributed computational workflows in a portable and replicable manner. It allows the seamless parallelization and deployment of any existing application with minimal development and maintenance overhead, irrespective of the original programming language.

The built-in support for container technologies such as Docker and Shifter, along with the native integration with the Git tool and popular code-sharing platforms like GitHub, make it possible to precisely prototype self-contained computational workflows, maintain all variations over time and rapidly reproduce any former configuration one may need to re-use. These capabilities guarantee consistent results over time and across different computing platforms.

Using a simple RNA-Seq based analysis we show how two seemingly irreproducible analyzes can be made stable across platforms when ported into Nextflow. Applying the transcript quantification tool Kallisto and the companion differential expression package Sleuth, we highlight how the same pipeline produces different results on Mac OSX and Linux. With a well studied dataset, 67 genes are only reported as significant on the Mac and 72 only reported in the Linux analysis out of a total of 6,138 differentially expressed genes. When the same pipeline was factored into Nextflow with Docker, the results were identical across platforms.

Nextflow was first introduced at BOSC 2015. Over the last year it has attracted an increasing amount of interest within the bioinformaticians community, including numerous users from leading institutions such as the Joint Genome Institute, the Pasteur Institute, the Sanger Institute, the International Agency for Research on Cancer and the Weill Cornell Medical College among others. An exhaustive list of third party curated pipelines is available from https://github.com/nextflow-io/awesome-nextflow