NextflowWorkbench: Reproducible and Reusable Workflows for Beginners and Experts

Jason P. Kurs[1], Manuele Simi[1,2], Fabien Campagne[1,2,3,*]

[1]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, United States of America; [2]Clinical Translational Science Center, Weill Cornell Medicine, New York, NY, United States of America; [3]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, United States of America. *To whom correspondence should be addressed: fac2003@campagnelab.org

Keywords: Workflows, Pipelines, Reproducibility, Docker, Language Workbench Technology

Computational workflows and pipelines are often created to automate series of processing steps. For instance, workflows enable one to standardize analyses for large projects or core facilities, but are also useful for individual biologists who need to perform repetitive data processing.

Some workflow systems are designed for beginners: they offer a graphical user interface and have been very popular with biologists. In practice, these tools are infrequently used by more experienced bioinformaticians, who may require more flexibility or performance than afforded by the user interfaces, and seem to prefer developing workflows with scripting or command line tools.

The talk will introduce the NextflowWorkbench, a workflow system designed for both beginners and experts that blends the distinction between user interface and scripting language. This system extends and reuses the popular Nextflow workflow description language (http://nextflow.io) and shares its advantages. It is built using Language Workbench Technology, also used to develop the MetaR platform [1].

In contrast to Nextflow, NextflowWorkbench offers an integrated development environment that helps complete beginners get started with workflow development. Auto-completion helps users who do not know the syntax of the Nextflow language, and typesystem checks make it possible to validate a workflow as it is being developed, in order to provide immediate feedback to the developer. Expert bioinformaticians will also benefit from unique interactive features that help them work more productively with docker containers. Finally, reusable processes provide modular workflows, a feature useful to both beginners and experts.

We illustrate this tool with a workflow to estimate RNA-Seq counts using Kallisto. We found that beginners can be taught how to assemble this workflow in a two hours training session. The workflow can transparently run either on a laptop with docker, on a lab cluster, or in the cloud.

NextflowWorkbench simplifies the development of reproducible, implicitly parallel workflows [2]. Software is distributed under the Apache 2.0 license and available at https://github.com/CampagneLaboratory/NextflowWorkbench, http://workflow.campagnelab.org.

References: [1] Fabien Campagne, William ER Digan, Manuele Simi. MetaR: simple, high-level languages for data analysis with the R ecosystem. bioRxiv doi: http://dx.doi.org/10.1101/030254

[2] Jason P Kurs, Manuele Simi, Fabien Campagne. NextflowWorkbench: Reproducible and Reusable Workflows for Beginners and Experts. bioRxiv doi: http://dx.doi.org/10.1101/041236