From reproducible to reusable bioinformatics workflows

GEORGE MARCHMENT, Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, France
BRYAN BRANCOTTE, Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, France
MARIE SCHMIT, Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, France
FRÉDÉRIC LEMOINE, Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, France and Institut Pasteur, Université Paris Cité, CNR Virus Des Infections Respiratoires, France
SARAH COHEN-BOULAKIA, Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, France

Additional Key Words and Phrases: workflows, reusability, structure, nextflow, bioinformatics

ACM Reference Format:
George Marchment, Bryan Brancotte, Marie Schmit, Frédéric Lemoine, and Sarah Cohen-Boulakia. 2018. From reproducible to reusable bioinformatics workflows. 37, 4, Article 111 (August 2018), 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 ABSTRACT

Data intensive science has ushered in a new era of bioinformatics analyses, resulting in a substantial growth in the scope, complexity and variety of bioinformatics analyses. Over the past decade, workflows management systems, such as Nextflow [3], have become pivotal in the development of these analyses. They are widely used and serve as indispensable tools in the creation, execution, and sharing of complex analyses, in the form of workflows, all while enhancing reproducibility and scalability, among other benefits [4]. Workflows consist of multiple data processing steps chained together by data flow: the input of one step is connected to the output of the previous one, determining their execution order. Hence, a workflow can be represented as a directed graph (see Fig. 1), called the *workflow structure*.

If workflow management systems have solved many issues related to the reproducibility of bioinformatics analyses, their usefulness in terms of a workflow's reusability is still limited. Indeed despite the growing number of bioinformatics workflows developed, there exists a major lack of their reuse among different users. [4] showed that for all Nextflow workflows, the top-30 most reused steps are only reused in 2.42% of the total steps.

This lack of reuse may stem from the inherent complexity of workflow code, particularly challenging for biologists and bioinformaticians who lack expertise in the field [1, 5].

Authors' Contact Information: George Marchment, george.marchment@universite-paris-saclay.fr, Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France; Bryan Brancotte, Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France; Marie Schmit, Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France; Frédéric Lemoine, frederic.lemoine@pasteur.fr, Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France and Institut Pasteur, Université Paris Cité, CNR Virus Des Infections Respiratoires, Paris, France; Sarah Cohen-Boulakia, sarah.cohen-boulakia@universite-paris-saclay.fr, Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France.

ACM 2476-1249/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

As underlined by the Workflows Community Summit [2], providing access to the workflow's structure (its directed graph representation depicting its sequential steps) enables workflow developers to quickly grasp the main workflow steps without the necessity of delving into its code. Structure thus plays a major role in allowing users to understand the workflow's overall functionality and thus makes users more inclined to reuse workflows.

In response to this, we developed **BioFlow-Insight**, a tool that analyses the code of Nextflow workflows and automatically extracts and rebuilds the workflow structure and provides multiple simplified representations of it. We believe that **BioFlow-Insight**, by allowing developers to easily extract a global view of their and others workflows, constitutes a first step towards a better reuse of workflow code.

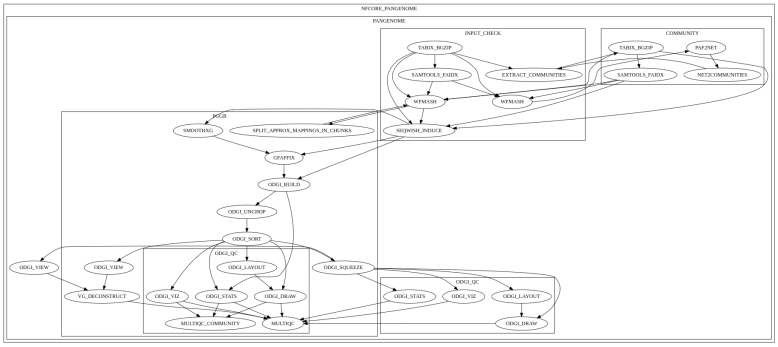

Fig. 1. Graph structure generated by BioFlow-Insight, showing a simplified representation of the pangenome nf-core workflow.

2 TOOL AVAILABILITY

BioFlow-Insight is an open-source tool, available as both a command-line interface (https://pypi.org/project/bioflow-insight/) and a web service (https://bioflow-insight.pasteur.cloud/).

3 FUNDING

This work received support from the National Research Agency under the France 2030 program, with reference to ANR-22-PESN-0007.

REFERENCES

[1] Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, and Christophe Blanchet. 2017. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems* 75 (Oct. 2017), 284–298. https://doi.org/10.1016/j.future.2017.01.012

[2] Rafael Ferreira et al. da Silva. 2023. *Workflows Community Summit 2022: A Roadmap Revolution.* Technical Report. https://doi.org/10.5281/zenodo.7750670 arXiv:2304.00019 [cs].

[3] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* 35, 4 (April 2017), 316–319. https://doi.org/10.1038/nbt.3820

[4] Marine Djaffardjy, George Marchment, Clémence Sebe, Raphael Blanchet, Khalid Bellajhame, Alban Gaignard, Frédéric Lemoine, and Sarah Cohen-Boulakia. 2023. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. *Computational and Structural Biotechnology Journal* 21 (Jan. 2023), 2075–2085. https://doi.org/10.1016/j.csbj.2023.03.003

[5] Onur Yukselen, Osman Turkyilmaz, Ahmet Rasit Ozturk, Manuel Garber, and Alper Kucukural. 2020. DolphinNext: a distributed data processing platform for high throughput genomics. *BMC Genomics* 21, 1 (April 2020), 310. https://doi.org/10.1186/s12864-020-6714-x

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009