

NEW METHOD TO ENSURE REPRODUCIBILITY IN COMPUTATIONAL EXPERIMENTS

- **Scientists at the Centre for Genomic Regulation (CRG) in Barcelona, Spain, have developed a workflow management system that prevents irreproducibility when analyzing large genomics datasets with computers.**
- **Nextflow contributes to establishing good scientific practices and provides an important framework for those research projects where the analysis of large datasets are used to take decisions, for example, in precision medicine.**

Research reproducibility is crucial to move forward in science. Unfortunately, and according to recent studies and surveys*, the number of irreproducible experiments is increasing and research reproducibility is now recognized as one of the major challenges that scientists, institutions, funders and journals must address for science to remain credible and to keep progressing.

In order to make sense of genomic data, scientists are increasingly relying on a combination of computer software named pipelines. These pipelines process data and deliver analytical results such as genetic risks for instance. Unfortunately, the results of these pipelines are not always reproducible. In the era of precision medicine, this limited reproducibility can have important implications for our health.

Now, a team of researchers at the Centre for Genomic Regulation (CRG) in Barcelona, Spain, led by Cedric Notredame, have developed a workflow management system that ensures reproducibility in computational experiments. The system, named Nextflow, has been described in the current issue of *Nature Biotechnology*. “When doing computational analysis, tiny variations across computational platforms can induce numerical instability that result in irreproducibility. *Nextflow* allows scientists to avoid these variations and contributes to standardizing good practices in computational experiments” explains Cedric Notredame, lead author of the paper.

“A small variation may not seem to be a problem when using genomic data in a particular research project but, even the smallest variations may be crucial if we are using these conclusions to take a decision, for instance on a precision medicine treatment.” adds Paolo Di Tommaso, first author of the paper. “Irreproducibility will be a major issue in precision medicine” he concludes.

Containing irreproducibility

The main reason for irreproducibility is the complexity of modern computers. With all the libraries and software they contain, computers are like machines made of billions of moving parts. Even when using exactly the same pipeline and the same data, slight variations across computers can lead to irreproducibility. The solution to this problem is providing not only the data and the software, but also the complete pre-configured execution environment within a new generation of virtualization technology named *containers*. The CRG team implemented *Nextflow* as a tool to manage a computational workflow along with its dependencies by using these containers. “It is like freezing the experiment, so everyone aiming at reproducing it can do it the same way without having to manually re-introduce complex configurations. This way of doing things guarantees that the same dataset will produce the same results anywhere” explain the authors.

Nextflow helps integrate the most sophisticated resources for reproducibility: Zenodo for data, Github and Docker for software, and the cloud for computation. It provides a turning point for good practice in the computational processing of large datasets. The CRG is now committed to help promoting this important aspect of modern biology by making this new resource available for academic research but also for clinical and commercial production. It is also organizing a series of courses and workshops dedicated to the use of Nextflow and its uptake by the community.

NOTES DE THE EDITORS

Reference:

Di Tommaso et al. *Nextflow enables reproducible computational workflows*. *Nature Biotechnology* 35, 316-319 (2017). doi: [10.1038/nbt.3820](https://doi.org/10.1038/nbt.3820)

(*) Other articles cited in this news release:

Baker, M. *1,500 scientists lift the lid on reproducibility*. *Nature* **533**, 452–454 (2016).

Munafò MR et al. *A manifesto for reproducible science*. *Nature Human Behaviour* 1, 0021 (2017).

Funding information:

This project was supported by the Centre for Genomic Regulation, the Spanish Plan Nacional, and the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', The Elixir EXCELERATE project and the H2020 European framework project OpenRiskNet. Maria Chatzou and Evan W Floden were supported by the La Caixa Foundation. We acknowledge the support of the CERCA Programme/Generalitat de Catalunya.

Images available at:

<https://www.dropbox.com/sh/d4ow6i3tt dg8fte/AAAIAY7gRI9lp2vAK7cnMjWJa?dl=0>

Caption and credits:

Nextflow CRG team.jpg => CRG researchers developed Nextflow, a new method to ensure reproducibility in computational experiments. From left to right: Emilio Palumbo, Paolo Di Tommaso, Evan W Floden, Cedric Notredame, and Pablo Prieto © CRG.

Media Contact:

Laia Cendrós, Media Relations, Centre for Genomic Regulation (CRG)

Tel.+34 93 316 02 37 – Mobile +34 607 611 798 – E-mail: laia.cendros@crg.eu