The Journal of Open Source Software

# Reproducible germline analysis with nf-core/sarek using Nix and DataLad

**Alexis Praga** [1¶], **Alexis Overs** [1], **and Gaétan Lepage** [2]

**1** Oncobiologie Génétique Bioinformatique, University Hospital of Besançon, France **2** Centre Inria de l'Université Grenoble Alpes, France ¶ Corresponding author

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:**

**Submitted:** 16 July 2025
**Published:** unpublished



## Summary

Analysis of the entire genome is now part of daily routine in clinical genomics thanks to technical breakthroughs in DNA sequencing. This results in large amounts of data that need to be processed by accurate, reproducible, and fast bioinformatics pipelines. However, achieving reproducibility across different computing platforms remains challenging. Recent tools have improved this situation with workflow management systems and consensus reference pipelines. This work advances the FAIR (Findable, Accessible, Interoperable, Reusable) principles in clinical bioinformatics by providing a complete reproducible environment for software dependencies and reference textual databases.

## Statement of need

In bioinformatics, there is a "reproducibility" crisis due to a wide variety of command-line utilities, possibly with non-deterministic output (Ziemann et al., 2023) and lack of good practices (Baykal et al., 2024). While workflow managers like Nextflow improve pipeline portability (Di Tommaso et al., 2017), and reference pipelines like nf-core/sarek provide standardized analysis workflows (Ewels et al., 2020; Garcia et al., 2020), reproducibility across different computing environments remains an issue. Traditional package managers may not be reproducible across different operating systems and architectures. Similarly, genomic databases are updated frequently, and current approaches may use outdated static databases or require manual management of database versions. We offer an alternative solution for reproducible package and database management for a reference pipeline in germline genetics.

Functional package managers like Nix or Guix allow for fully deterministic software builds, an approach more robust than containerization. (Courtès, 2013; Dolstra et al., 2004). Here, we packaged Sarek software dependencies in Nix for germline analysis and contributed all changes to Nix central package repository, nixpkgs. Instead of duplicating databases across servers, like Illumina iGenomes used by Sarek, we offer for the first time a decentralized approach for data management based on DataLad (Halchenko et al., 2021). All remote database locations are stored in a single configuration file, allowing for modular access and easier updates. In practice, our project is completely defined by several configurations, for Nix and Nextflow execution, and several minimal open-source GitHub repositories to track databases locations.

The project can be used for germline genetics, both in research and clinical care. This approach cleanly separates package management and database provisioning from workflow execution in a modular fashion to improve reusability and reproducibility.

The Journal of Open Source Software

## Acknowledgements

Computations have been performed on the supercomputer facilities of the Mésocentre de calcul de Franche-Comté (France).

## References

Baykal, P. I., Łabaj, P. P., Markowetz, F., Schriml, L. M., Stekhoven, D. J., Mangul, S., & Beerenwinkel, N. (2024). Genomic reproducibility in the bioinformatics era. *Genome Biology*, *25*(1), 213. https://doi.org/10.1186/s13059-024-03343-2

Courtès, L. (2013). *Functional Package Management with Guix*. arXiv. https://doi.org/10.48550/ARXIV.1305.4584

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dolstra, E., De Jonge, M., Visser, E., & others. (2004). Nix: A Safe and Policy-Free System for Software Deployment. *LISA*, *4*, 79–92.

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Garcia, M., Juhos, S., Larsson, M., Olason, P. I., Martin, M., Eisfeldt, J., DiLorenzo, S., Sandgren, J., Díaz De Ståhl, T., Ewels, P., Wirta, V., Nistér, M., Käller, M., & Nystedt, B. (2020). Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research*, *9*, 63. https://doi.org/10.12688/f1000research.16665.2

Halchenko, Y., Meyer, K., Poldrack, B., Solanky, D., Wagner, A., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S., Mönch, C., Markiewicz, C., Waite, L., Shlyakhter, I., De La Vega, A., Hayashi, S., Häusler, C., Poline, J.-B., Kadelka, T., … Hanke, M. (2021). DataLad: Distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software*, *6*(63), 3262. https://doi.org/10.21105/joss.03262

Ziemann, M., Poulain, P., & Bora, A. (2023). The five pillars of computational reproducibility: Bioinformatics and beyond. *Briefings in Bioinformatics*, *24*(6), bbad375. https://doi.org/10.1093/bib/bbad375