Improving nf-core/sarek reproductibility

07 July 2025

Summary

Analysis of the entirety of the 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 upon 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 functional databases.

Statement of need

In bioinformatics, there is a "reproducibility" crisis due to a a wide variety of command-line utilities, possibly with non-determinist output (Ziemann, Poulain, and Bora 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.problematic. 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 Similarly, genomic databases are updated frequently, and current approaches may use outdated static databases or require manual management of database versions.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 manager like Nix or Guix offers fully determistic build for software, an approach more robust than containerisation. (Dolstra et al. 2004; Courtès 2013). Here, we packaged in Nix Sarek software dependencies for germline analysis and all changes have been contributed to nixpkgs. Instead of duplicating database, like Illumina iGenomes used by Sarek, we offer for the first time a decentralized approach for data management based on Datalad.

All remote database locations are stored in a single configuration, allowing for modular accessand easier updates.

This approach builds upon the strengths of the Sarek pipeline with a modular approach to package management and database provisioning, making it suitable for deployment in both research and clinical environments.

Acknowledgements

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

References

Baykal, Pelin Icer, Paweł Piotr Łabaj, Florian Markowetz, Lynn M. Schriml, Daniel J. Stekhoven, Serghei Mangul, and Niko Beerenwinkel. 2024. "Genomic Reproducibility in the Bioinformatics Era." *Genome Biology* 25 (1): 213. https://doi.org/10.1186/s13059-024-03343-2.

Courtès, Ludovic. 2013. "Functional Package Management with Guix." arXiv. https://doi.org/10.48550/ARXIV.1305.4584.

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Dolstra, Eelco, Merijn De Jonge, Eelco Visser, et al. 2004. "Nix: A Safe and Policy-Free System for Software Deployment." In *LISA*, 4:79–92.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. "The Nf-Core Framework for Community-Curated Bioinformatics Pipelines." *Nature Biotechnology* 38 (3): 276–78. https://doi.org/10.1038/s41587-020-0439-x.

Garcia, Maxime, Szilveszter Juhos, Malin Larsson, Pall I. Olason, Marcel Martin, Jesper Eisfeldt, Sebastian DiLorenzo, et al. 2020. "Sarek: A Portable Workflow for Whole-Genome Sequencing Analysis of Germline and Somatic Variants." *F1000Research* 9 (September): 63. https://doi.org/10.12688/f1000research.16665.2.

Ziemann, Mark, Pierre Poulain, and Anusuiya Bora. 2023. "The Five Pillars of Computational Reproducibility: Bioinformatics and Beyond." *Briefings in Bioinformatics* 24 (6): bbad375. https://doi.org/10.1093/bib/bbad375.