



HAL
open science

Empowering bioinformatics communities with Nextflow and nf-core

Björn E Langer, Andreia Amaral, Marie-Odile Baudement, Franziska Bonath,
Mathieu Charles, Praveen Krishna Chitneedi, Emily L Clark, Paolo Di
Tommaso, Sarah Djebali, Philip A Ewels, et al.

► **To cite this version:**

Björn E Langer, Andreia Amaral, Marie-Odile Baudement, Franziska Bonath, Mathieu Charles, et al.. Empowering bioinformatics communities with Nextflow and nf-core. 2024. hal-04770333

HAL Id: hal-04770333

<https://hal.science/hal-04770333v1>

Preprint submitted on 6 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Empowering bioinformatics communities with Nextflow and nf-core

Authors: Björn E. Langer¹, Andreia Amaral^{2,3}, Marie-Odile Baudement⁴, Franziska Bonath⁵, Mathieu Charles^{6,7}, Praveen Krishna Chitneedi⁸, Emily L. Clark⁹, Paolo Di Tommaso¹⁰, Sarah Djebali¹¹, Philip A. Ewels¹⁰, Sonia Eynard¹², James A. Fellows Yates^{13,14,15}, Daniel Fischer¹⁶, Evan W. Floden¹⁰, Sylvain Foissac¹², Gisela Gabernet¹⁷, Maxime U. Garcia¹⁰, Gareth Gillard⁴, Manu Kumar Gundappa⁹, Cervin Guyomar¹², Christopher Hakkaart¹⁰, Friederike Hanssen^{18,19,20}, Peter W. Harrison²¹, Matthias Hörtenhuber^{22,23}, Cyril Kurylo¹², Christa Kühn^{8,24}, Sandrine Lagarrigue²⁵, Delphine Lallias⁶, Daniel J. Macqueen⁹, Edmund Miller²⁶, Júlia Mir-Pedrol¹⁹, Gabriel Costa Monteiro Moreira²⁷, Sven Nahnsen^{18,19,20,28}, Harshil Patel¹⁰, Alexander Peltzer²⁹, Frederique Pitel¹², Yulixaxis Ramayo-Caldas³⁰, Marcel da Câmara Ribeiro-Dantas^{10,31}, Dominique Rocha⁶, Mazdak Salavati^{9,32}, Alexey Sokolov²¹, Jose Espinosa-Carrasco^{1*}, Cedric Notredame^{1,33*} and the nf-core community.

Affiliations:

- 1 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
- 2 MED. University of Évora, Polo da Mitra, 7002-554 Évora, Portugal
- 3 Centre for Interdisciplinary Research in Animal Health (CIISA), Faculty of Veterinary Medicine, University of Lisbon, 1300-477, Lisbon Portugal
- 4 Centre for Integrative Genetics (CIGENE), Faculty of Biosciences, Norwegian University of Life Sciences, 1430 Ås, Norway
- 5 Kungliga Techniska Högskolan, School of Engineering Sciences in Chemistry, Biotechnology and Health at the Department of Gene Technology
- 6 INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, F-78350, France
- 7 SIGENAE, INRAE, Jouy-en-Josas, F-78350, France
- 8 Research Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany
- 9 The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, UK.
- 10 Seqera, Carrer de Marià Aguiló, 28, Barcelona, 08005, Spain
- 11 IRSD, Université de Toulouse, INSERM, INRAE, ENVT, Univ Toulouse III - Paul Sabatier (UPS), Toulouse, France
- 12 GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet-Tolosan, France
- 13 Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Pl. 6, Leipzig, Saxony, 4103, Germany
- 14 Department of Paleobiotechnology, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Adolf-Reichwein-Straße 23, Jena, Thuringia, 07745, Germany (Current Address)
- 15 Associated Research Group of Archaeogenetics, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Adolf-Reichwein-Straße 23, Jena, Thuringia, 07745, Germany (Current Address)

- 16 Natural Resources Institute Finland (Luke), Applied Statistical Methods, Myllytie 1, 31600 Jokioinen, Finland
- 17 Department of Pathology, Yale School of Medicine. New Haven, CT 06511, USA
- 18 Biomedical Data Science, Department of Computer Science, University of Tübingen, Germany
- 19 Quantitative Biology Center (QBiC), University of Tübingen, Germany
- 20 M3 Research Center, Faculty of Medicine, University of Tübingen
- 21 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, Cambridge, UK.
- 22 Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
- 23 SciLifeLab, Uppsala University, Uppsala, Sweden
- 24 Friedrich-Loeffler-Institut (FLI), 17493 Greifswald-Riems, Germany
- 25 PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France
- 26 Department of Biological Sciences and Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75205, United States
- 27 Unit of Animal Genomics, GIGA & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium
- 28 Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen
- 29 Boehringer Ingelheim Pharma GmbH & Co KG, Birkendorfer Straße 65, 88400 Biberach/Riss, Germany
- 30 Animal Breeding and Genetics Program, Institute of Agrifood Research and Technology (IRTA), Caldes de Montbui, Barcelona, Spain
- 31 Postgraduate Program in Biotechnology at Universidade Potiguar (UnP) – Natal/RN, Brazil.
- 32 Dairy Research Innovation Centre, South and West Faculty, Scotland's Rural College, Dumfries, United Kingdom DG1 3NE
- 33 Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

*To whom correspondence should be addressed.

Abstract

Standardised analysis pipelines are an important part of FAIR bioinformatics research. Over the last decade, there has been a notable shift from point-and-click pipeline solutions such as Galaxy towards command-line solutions such as Nextflow and Snakemake. We report on recent developments in the nf-core and Nextflow frameworks that have led to widespread adoption across many scientific communities. We describe how adopting nf-core standards enables faster development, improved interoperability, and collaboration with the >8,000 members of the nf-core community. The recent development of Nextflow Domain-Specific Language 2 (DSL2) allows pipeline components to be shared and combined across projects. The nf-core community has harnessed this with a library of modules and subworkflows that can be integrated into any Nextflow pipeline, enabling research communities to progressively transition to nf-core best practices. We present a case study of nf-core adoption by six European research consortia, grouped under the EuroFAANG umbrella and dedicated to farmed animal genomics. We believe that the process outlined in this report can inspire many large consortia to seek harmonisation of their data analysis procedures.

Introduction

Advancements in large-scale molecular biology methods have driven an unprecedented increase in data generation¹⁻³. The volume of data generated pushes computational capacities to their limits, revealing an urgent need for robust and scalable data analyses⁴. Workflow management systems (WfMSs) are now the recommended solution when dealing with high-throughput data analysis pipelines⁵. While some frameworks such as the Galaxy framework⁶ focus on user-friendly graphical interfaces, tools such as Snakemake⁷ and Nextflow⁸ are designed for bioinformaticians familiar with programming. They combine the expressiveness⁹ of Bash with additional features to better support reproducibility, traceability, parallelisation, and portability across different infrastructures (HPC clusters, cloud computing, and workstations). When considering some of the most popular WfMSs⁹ we found that there were more than two and half times the number of citations in 2023 than in 2018 (Fig. 1, Supplementary Fig. 1, Supplementary Table 1, Supplementary Table 2). A notable trend over the last six years is a decrease in Galaxy citations in favour of Snakemake and Nextflow. In 2023, Nextflow was the resource with the highest usage growth and in that same year, Google Scholar data indicates that Galaxy and Nextflow each accounted for about a third of reported WfMS usage.

While WfMSs provide a powerful way to bundle available methods into pipelines, they do not provide standards for how this bundling should be carried out. This gap has led to the establishment of pipeline registries for Snakemake (Snakemake Workflow Catalog (SWC), <https://snakemake.github.io/snakemake-workflow-catalog/>; Snakemake-Workflows, <https://github.com/snakemake-workflows>) and Nextflow (nf-core, <https://nf-co.re/>). These workflow repositories enforce implementation guidelines, effectively providing best-practice standards for both WfMSs. The nf-core community has been a key driver behind Nextflow adoption, in part due to the excellent quality of its curated pipeline collection. A recent study found that 83% of nf-core pipelines could be deployed and run automatically without a crash, nearly 4x better than other catalogues¹⁰.

Nextflow and nf-core: recent developments

Expansion of the nf-core pipeline community. The nf-core community was created in 2018 as a curated collection of pipelines implemented according to agreed-upon best-practice standards¹¹. These off-the-shelf pipelines are characterised by reproducibility, standardisation, and rapid result generation. There are now nearly 100 pipelines available within nf-core,

covering the analysis of a broad range of data types, including high-throughput sequencing, mass spectrometry, protein structure prediction, microscopy, satellite imagery analysis, and economics. These curated pipelines are supported by over 2,000 contributors and developers on GitHub, over 100 institutions, and over 8,000 users on nf-core's primary communication platform Slack. Pipeline long-term maintenance is achieved by an efficient cooperation model allowing teams across sites to collectively and simultaneously contribute according to their needs and capacity (Fig. 2a).

Broadened nf-core outreach. nf-core organises programs and events to facilitate sustainable growth of the community. Since its inception, nf-core has run 13 virtual, hybrid, or in-person hackathon events that have attracted global attendance. Nearly 100 'nf-core/bytesize' webinars describing community advancements, coding approaches, and pipelines, have been shared online, attracting tens of thousands of views. A mentorship program pairs experienced Nextflow and nf-core developers with new community members from underrepresented groups. Free community Nextflow and nf-core training videos are recorded and shared online every 6 months, accounting for over 35,000 views on YouTube since October 2022 (Supplementary Table 4). nf-core strives for accessible community programs and events, irrespective of language, geography, time zones, and the ability to travel.

Evolution of the Nextflow workflow management system. The most notable development in Nextflow during the past five years has been the introduction of the DSL2 syntax. This introduced the ability to define complex workflows which can be organised into reusable components. This has dramatically improved the encapsulation and re-use of analytical steps shared across pipelines. Nextflow has also benefited from improved support for cloud-based computing platforms, support for on-demand container provisioning via the Wave service¹², and extended support for many HPC schedulers and software container engines (15 schedulers and 7 container engines, as of version 23.10.0). Community-driven projects such as nf-test¹³, a testing framework for Nextflow pipelines and components, have garnered rapid adoption. Notably, Nextflow plugins ease the extension of core Nextflow functionality, illustrative examples are nf-validation¹⁴ (JSON schema validation of pipeline parameters and sample sheets), nf-co2footprint¹⁵ (energy consumption and carbon intensity from pipeline compute), and nf-prov¹⁶ (provenance reports for pipeline runs). This wide range of improvements may explain why Nextflow is being adopted in fields beyond biological applications, with new pipelines in fields as diverse as astrophysics¹⁷, earth science¹⁸, and economics¹⁹.

Impact of module support. The main impact of Nextflow's DSL2 transition on nf-core has been to enable the establishment of a shared repository for standardised modules and subworkflows. The *nf-core/modules* repository currently features more than 1000 modules and 50 subworkflows, which are used by nf-core and non-nf-core pipelines. Each module is a wrapper around a single tool function and bundles the required software via Docker / Singularity images and a Conda environment, provided via the Bioconda²⁰ and Biocontainers communities²¹. In this way, modules encapsulate the correct environment for each embedded software, thus mitigating dependency conflicts between software packages. nf-core also supports self-hosted module repositories. This allows direct integration of internal development with community-maintained modules and subworkflows, thereby bridging closed- and open-source development.

Improved pipeline standardisation. Adopting a unified pipeline structure and incorporating shared features guarantees consistent functionality and high-quality standards across nf-core pipelines. For this, nf-core has developed a base template for pipelines. The template receives regular updates which are semi-automatically synchronised across all pipelines. The template establishes pipeline file structure, with code, documentation, and continuous integration (CI) tests. Pipelines are tested using automated linting tests for code formatting and syntax, and flags outdated nf-core modules. To guarantee deployability, each pipeline comes with minimal datasets used to test the pipeline for every pull request. CI testing also includes a "full-size" dataset with a realistic input to test portability to diverse platforms for each stable release, executed automatically on AWS and Microsoft Azure after every pipeline release. The output files are publicly accessible and can be browsed on the nf-core website.

Extended tooling. The *nf-core/tools* package provides helper tools to improve accessibility for users of any Nextflow pipeline. Functionality includes commands to download pipelines and associated software for offline use and methods for launching pipelines, including a command-line and a graphical web-browser-based pipeline-launch interface to enable users less comfortable with command-line interfaces. The CLI is a key part of nf-core development, with functionality to create new pipelines based on the nf-core template, create and modify JSON schema files for parameter validation/documentation. Subcommands allow developers to discover, install, and update modules and subworkflows, as well as create new ones from a template. Automation synchronises existing pipelines with new template versions, as well as

updating modules and subworkflows to the latest version of the tool. A lint command tests pipelines and components for adherence to nf-core guidelines.

Reproducibility and data provenance. Achieving full reproducibility when running a pipeline is an essential requirement of the FAIR paradigm. Nextflow and nf-core simplify this task by supporting the programmatic collection of execution parameters (eg. tool versions, command-line flags) through parameter files, the nf-prov plugin and MultiQC²² summary reports. Provenance is especially critical when comparing results across large consortiums, such as EuroFAANG (see next section). These features were essential to meet the FAANG metadata standards²³, facilitating comparability, reproducibility, and standardisation.

Uptake of nf-core by the farmed animals genomics community

Functional Annotation of Animal Genomes project (FAANG). FAANG²⁴ (Functional Annotation of Animal Genomes) is a coordinated international effort established in 2014, with the explicit purpose of harmonising genotype-to-phenotype research in farmed animals. EuroFAANG is a formalised initiative of FAANG partnering organisations across Europe, which currently encompasses six of EU's Horizon 2020 program consortia: AQUA-FAANG, BovReg, GENE-SWitCH, RUMIGEN, GEroNIMO, and HoloRuminant. These six consortia bring together more than 80 academic institutions and commercial companies spread across Europe, North America, Middle East, and Australia. While each project specialises in a distinct set of farmed animal species (multiple fish species, cattle, pig, and chicken), they share the common goal of improving our understanding of the genome-phenome relationship through the comprehensive annotation of genomic functional features. This collective effort systematically captures dynamic gene expression, chromatin accessibility, and epigenetic state under a range of biological conditions. Comprehensive functional annotation demands the generation and analysis of large sequencing datasets derived from distinct functional assays including RNA-seq, ATAC-seq, and CHIP-seq. Together, these projects are committed to implementing an analysis strategy adhering to the FAIR (findability, accessibility, interoperability, and reusability) principles of scientific reproducibility^{25,26}.

The projects AQUA-FAANG, BovReg, and GENE-SWitCH were the first to use the EuroFAANG name as a common umbrella. EuroFAANG is now developing as a future European Research Infrastructure²⁷ to bring together national facilities at the pan-European level, in the fields of animal genetic resources, phenotyping and breeding, and animal health. The EuroFAANG

Research Infrastructure project builds on the foundation provided by the six current H2020 EuroFAANG projects and connects with existing infrastructures for data management and animal agriculture and aquaculture in the European research infrastructure landscape. It also addresses the research priorities outlined in FAANG to Fork²⁸, which mirrors the European Commission's European Green Deal Farm to Fork strategy. The recommendation, adoption, adaptation, and reuse of nf-core workflows for future genotype-to-phenotype research remains a key aspect of the goals of the EuroFAANG Research Infrastructure, to ensure that analyses conducted across European institutes and industry are interoperable, standardised, and reusable.

Necessity of interoperability and standardisation. The use of nf-core provides an effective standardisation strategy, which avoids technical differences that can be introduced by using varied parameterisation and analysis tools. Such standardisation ensures the comparability of results. For instance, AQUA-FAANG relied on nf-core pipelines to exploit the evolutionary conservation of genomic regulatory elements and epigenetic states across six farmed fish species for multiple biological conditions and matched tissue panels²⁹. Data integration across projects is critical, as projects such as the farm animal GTEx³⁰⁻³² are expected to keep integrating a larger number of species and deeper sequencing over time. Interoperability, achieved through the use of a common analytic framework is therefore an essential component of long-term sustainability, well beyond the goal and lifetime of projects like EuroFAANG.

Main hurdles towards a common bioinformatics analysis framework. Establishing common standards is still one of the most challenging goals of bioinformatics³³. Research groups often have established methods and adopting new standards may require re-training staff, as well as re-implementing and re-validating methods already in production. The process of standards adoption can benefit from being spear-headed by a designated leader. When the AQUA-FAANG consortium decided to utilise nf-core, the University of Edinburgh and the Norwegian University of Life Sciences took the responsibility of deploying the framework and training the other AQUA-FAANG groups to use nf-core pipelines. Regardless of which framework is chosen, the transition roadmap is as important as the endpoint – a gradual transition is more likely to succeed in avoiding fragmentation across groups.

Choosing a bioinformatics analysis framework. Discussions and surveys about methods, languages, and standards were carried out among the EuroFAANG partners (Table 1). These consultations reflected the predominance of Nextflow as a favoured WfMS, and identified the needs of consortia that were not already met by the methods available in nf-core. We found that the 15 consortium partners involved in data analysis were carrying out 78 analyses, using four different programming languages: Nextflow (70), Snakemake (3), Bash (4), and Perl (1). A key feature of Nextflow is its lack of language dependency. Even though the Nextflow DSL is based on Groovy, the scripts and functions of a pipeline can be directly ported into Nextflow in their original language. This enables pipelines to be gradually ported into Nextflow without the need to rewrite scripts or do extensive revalidation. For example, the BovReg eQTL analysis workflow was initially written in Bash and was ported into a Nextflow pipeline. The wide range of analyses available in nf-core and its fast-growing community are important aspects of long-term pipeline maintenance.

EuroFAANG pipeline development within the nf-core ecosystem. All nf-core pipelines are open source and welcome new contributions. For instance, EuroFAANG partners have been heavily involved in the maintenance and development of several existing nf-core pipelines, including *nf-core/atacseq*, *nf-core/chipseq*, and *nf-core/smrnaseq* (Table 1). PacBio-based genome annotation was identified as a gap for the GENE-SWitCH project, leading one of its partners (Roslin Institute, University of Edinburgh, UK) to port their Iso-Seq Nextflow pipeline into nf-core (*nf-core/isoseq*³⁴). Similar collaborative efforts have been initiated within GEroNIMO and BovReg for eQTL analysis using GWAS and Allele-Specific Expression (ASE). This led to the development of *nf-pegASE* (INRAE, Table 1), a pipeline for ASE analysis that is based on the nf-core template which is intended to be contributed to nf-core. Some pipelines were also developed outside of the nf-core umbrella, like TAGADA³⁵ for the annotation and quantification of novel long non-coding and coding RNA transcripts and *BovReg/nf-cage*³⁶ for transcript annotation. Thanks to the standardisation provided by Nextflow, these non-nf-core pipelines can nonetheless provide new modules and elements lending themselves to nf-core integration. Overall, most pipeline methods were widely shared across the various consortium members and while nf-core dominates, other solutions were also part of a very active exchange process (Fig. 2b, Supplementary Table 3) with 4 of the 17 pipelines being used by at least three EuroFAANG consortiums.

Conclusion and perspective

The rapid growth of the nf-core community demonstrates the importance of strict best practice guidelines bundled with tooling for developers and users. The module and subworkflow infrastructure facilitates collaborative development both within nf-core and between external pipelines, leading to improved code quality and long-term maintenance. nf-core standards and community support enable faster development, improved interoperability, standardisation, and collaboration. As adoption spreads from traditional bioinformatics to other scientific fields these standards continue to enable other distributed communities to work together effectively and produce more reliable research.

In this report, we present the process of adopting a bioinformatics pipeline standard at the level of a group of consortia involving over 80 institutions and companies, with a primary focus on computational analysis. We argue here for soft standards, meaning a framework that accommodates a continuum between the current state of an implementation and the most desirable one. We found nf-core to be a good trade-off as it is based on Nextflow, which can wrap any existing code regardless of its implementation language. Strict adherence to nf-core best-practices is not mandatory for a Nextflow pipeline to be functional, allowing for progressive development towards nf-core standards. The extra development effort is offset by secondary benefits resulting from community input: nf-core pipelines are more visible, and have more accessible documentation. They benefit from community support, validation, bug fixes, and extension. These factors are especially important when considering long-term sustainability.

The EuroFAANG collaboration with nf-core has been a great success, resulting in mutual benefits and truly FAIR science. To foster more collaborations of this kind, nf-core has started a new initiative of “special interest groups”, where researchers with a common focus can collaborate on standardised usage of nf-core pipelines. We hope that these groups will work orthogonally to existing groups working on specific pipelines, encouraging improved collaboration and standardisation between scientists worldwide.

Figures and Tables

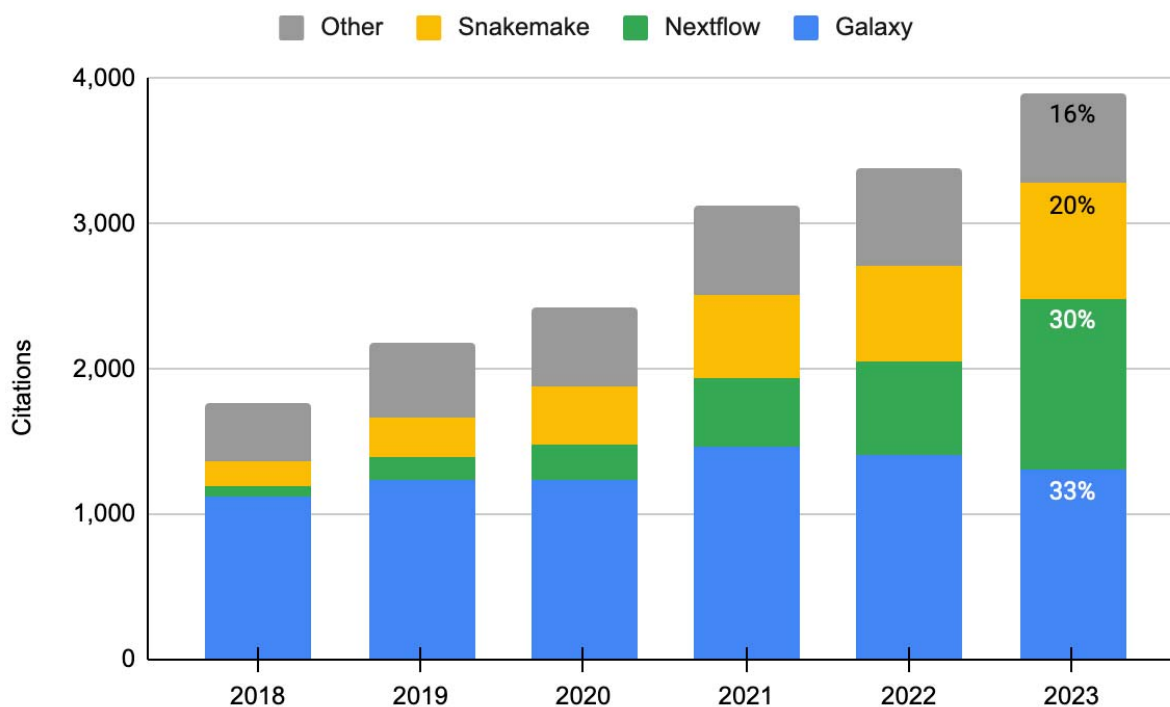


FIGURE 1: Google Scholar citation counts for bioinformatics workflow management systems. Sum of citations of the major publications of Galaxy, Nextflow, and Snakemake between 2018 and 2023 (Data in Supplementary Table 1).



FIGURE 2: Pipeline maintenance and usage. (A) Major contributions to the *nf-core/smrnaseq* pipeline over time by different academic institutions or private companies. Data for individual contributors is collapsed to their institution (SciLifeLab: 3; QBiC: 2; Boehringer Ingelheim: 3; Seqera: 4; all the others: 1) (B) Nextflow analysis pipelines used in the EuroFAANG consortia for the functional annotation of various species' genomes.

Funding

The EuroFAANG genomics communities have received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement Numbers 817923 (AQUA-FAANG), 815668 (BovReg), 817998 (GENE-SWitCH), 101000226 (RUMIGEN), 101000236 (GEroNIMO), and the Horizon Europe programme under Grant Agreement Number 101094718 (EuroFAANG Research Infrastructure). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

S.N. acknowledges support by the Carl Zeiss Foundation, project "Certification and Foundations of Safe Machine Learning Systems in Healthcare". SN is "gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder - EXC 2180 - 390900677 (iFIT)". SN is "gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder - EXC 2124 – 390838134 (CMFI)". This study was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the project NFDI 1/1 "GHGA - German Human Genome-Phenome Archive" (#441914366 to SN).

J.A.F.Y. was supported by Max Planck Society and the Werner Siemens Foundation grant 'Palaeobiotechnology' (awarded to Profs. Pierre Stallforth and Christina Warinner)

References

1. THE GTEx CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P. & Hayashizaki, Y. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci. Data* **4**, 170113 (2017).
4. Grüning, B. *et al.* Practical Computational Reproducibility in the Life Sciences. *Cell Syst.* **6**, 631–635 (2018).
5. Deelman, E. *et al.* The future of scientific workflows. *Int. J. High Perform. Comput. Appl.* **32**, 159–175 (2018).
6. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).
7. Mölder, F. *et al.* Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. *F1000Research* **10**, (2021).
8. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
9. Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat. Methods* **18**, 1161–1168 (2021).
10. Grayson, S., Marinov, D., Katz, D. S. & Milewicz, R. Automatic Reproduction of Workflows in

- the Snakemake Workflow Catalog and nf-core Registries. in *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability* 74–84 (ACM, Santa Cruz CA USA, 2023). doi:10.1145/3589806.3600037.
11. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).
 12. seqeralabs/wave. Seqera (2024). <https://github.com/seqeralabs/wave>
 13. askimed/nf-test. Askimed (2024). <https://github.com/askimed/nf-test>
 14. nextflow-io/nf-validation. nextflow-io (2024). <https://github.com/nextflow-io/nf-validation>
 15. nextflow-io/nf-co2footprint. nextflow-io (2024). <https://github.com/nextflow-io/nf-co2footprint>
 16. nextflow-io/nf-prov. nextflow-io (2024). <https://github.com/nextflow-io/nf-prof>
 17. Scott, D. R. *et al.* CELEBI: The CRAFT Effortless Localisation and Enhanced Burst Inspection pipeline. *Astron. Comput.* **44**, 100724 (2023).
 18. Lehmann, F., Frantz, D., Becker, S., Leser, U. & Hostert, P. FORCE on Nextflow: Scalable Analysis of Earth Observation Data on Commodity Clusters. in *Proceedings of the CIKM 2021 Workshops* (eds. Cong, G. & Ramanath, M.) vol. 3052 (CEUR-WS.org, Gold Coast, Queensland, Australia, 2021).
 19. Kanakaris, N. & Karacapilidis, N. Predicting prices of Airbnb listings via Graph Neural Networks and Document Embeddings: The case of the island of Santorini. *Procedia Comput. Sci.* **219**, 705–712 (2023).
 20. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).
 21. da Veiga Leprevost, F. *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **33**, 2580–2582 (2017).
 22. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
 23. Harrison, P. W. *et al.* The FAANG Data Portal: Global, Open-Access, “FAIR”, and Richly Validated Genotype to Phenotype Data for High-Quality Functional Annotation of Animal Genomes. *Front. Genet.* **12**, (2021).
 24. Andersson, L. *et al.* Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
 25. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
 26. Barker, M. *et al.* Introducing the FAIR Principles for research software. *Sci. Data* **9**, 622 (2022).
 27. A European infrastructure for farmed animal genotype to phenotype research. <https://doi.org/10.3030/101094718> (2022).
 28. Clark, E. L. *et al.* From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biol.* **21**, 285 (2020).
 29. Johnston, I. A. *et al.* Advancing fish breeding in aquaculture through genome functional annotation. *Aquaculture* **583**, 740589 (2024).
 30. Liu, S. *et al.* A multi-tissue atlas of regulatory variants in cattle. *Nat. Genet.* **54**, 1438–1447 (2022).
 31. Teng, J. *et al.* A compendium of genetic regulatory effects across pig tissues. *Nat. Genet.* **56**, 112–123 (2024).

32. Guan, D. *et al.* The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. 2023.06.27.546670 Preprint at <https://doi.org/10.1101/2023.06.27.546670> (2023).
33. Leng, R. *et al.* The Sequences and the Sequencers: What Can a Mixed-Methods Approach Reveal about the History of Genomics? *Hist. Stud. Nat. Sci.* **52**, 277–319 (2022).
34. Guizard, S. *et al.* nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. *Bioinformatics* **39**, btad150 (2023).
35. Kurylo, C., Guyomar, C., Foissac, S. & Djebali, S. TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data. *NAR Genomics Bioinforma.* **5**, lqad089 (2023).
36. Salavati, M. & Espinosa-Carrasco, J. MazdaX/nf-cage: nf-cage. Zenodo <https://doi.org/10.5281/zenodo.6855541> (2022).