genomicrelatedness

bwRSE4HPC Project Proposal

Assigned RSEs:	Thomas Isensee[1]
Project Partners:	Gisela Kopp[2], Till Dorendorf[3]
Scientific field:	Biology
Subfield:	Bioinformatics
Start:	2025-10-01
Duration:	3 months
Type of Work:	Refactoring, packaging, documentation, and workflow standardization
License:	GNU GPL
Link:	https://github.com/didillysquat/kopp_2021_fowl
Programming Language:	Nextflow, Python
Technologies:	nf-core
Target Cluster:	BinAC2

1. Project Summary

This project aims to modernize and modularize an existing Nextflow pipeline that was originally developed for relatedness analysis from low-coverage whole-genome sequencing (lcWGS) data [1], specifically in wild Guineafowl. The current pipeline consists of four monolithic workflows written in Nextflow DSL1 and is no longer maintained. To ensure sustainability, reproducibility, and usability by the wider research community, the workflow will be refactored to follow the nf-core standards and best practices. This includes converting to Nextflow [2] DSL2, adopting modular components, and enabling automated testing and deployment via nf-core infrastructure.

2. Problem Statement

The existing Nextflow workflows were developed for genotype likelihood calling and relatedness analysis [1,3–6] in wild bird populations, particularly Guineafowl, and have proven useful in ecological genomics research. However, the workflows are monolithic, based on the outdated DSL1 syntax, and lack automated tests or modular reuse. This limits accessibility and maintainability, especially for researchers unfamiliar with software engineering practices. Furthermore, they are not integrated into nf-core [7], which prevents easy usage and sharing within the bioinformatics community. Updating and modularizing the pipeline

[1] Scientific Software Center (SSC), Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany; thomas.isensee@iwr.uni-heidelberg.de

[2] Zukunftskolleg and Department of Biology and Center for the Advanced Study of Collective Behaviour, University of Konstanz, 78457 Konstanz, Germany; Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, 78315 Radolfzell, Germany; gisela.kopp@uni-konstanz.de

[3] Department of Biology & Konstanz Research School Chemical Biology (KoRS-CB), University of Konstanz, 78457 Konstanz, Germany; till.dorendorf@uni-konstanz.de

would allow broader adoption and easier execution on both local systems and bwHPC clusters, fostering further scientific progress in population genomics and related fields.

3. Suggested Solution

- Refactor the existing pipelines using Nextflow DSL2 with modular structure. Where applicable, replace custom processes with existing nf-core modules to improve maintainability and transparency.
- Enforce nf-core coding standards throughout the codebase to enable long-term maintainability and to prepare the workflow for potential publication on nf-core.
- Create a minimal example dataset and use it to establish automated tests for continuous integration (CI) on both local systems and HPC clusters.

4. Milestones

Week 1-2

- Identify available nf-core modules corresponding to the tools used in the current repository.
- Identify missing features or requirements for nf-core compatibility.

Week 3 - Month 2

- Port the existing pipelines to a modular structure using the latest Nextflow DSL2 syntax.
- Ensure the pipeline is functional with a minimal dataset on local systems.

Month 3

- Test the updated pipeline with real data on an HPC cluster (e.g., BinAC2).
- Integrate minimal example setup into CI testing framework.
- Prepare the repository for publication on nf-core, including documentation, licensing, and metadata compliance.

5. Deliverables

- A fully refactored, modular Nextflow pipeline based on DSL2 according to nf-core standards.
- Integration of nf-core modules where appropriate, and custom modules otherwise.
- Comprehensive documentation and instructions for installation and usage, including support for Conda, Docker, and Apptainer/Singularity.
- Continuous integration setup for automatic test execution via GitHub Actions or similar CI platform.
- A GitHub repository that complies with nf-core structure and conventions, ready for publication (pending review and approval by the nf-core community).

6. Expected Contributions

- Designate domain experts as a point of contact for technical questions and validation of biological relevance.
- Provide access to example data and expected results to validate the pipeline logic.
- Grant access to the GitHub repository and approve any required changes (e.g., repository name or structure) needed for nf-core compatibility.

- Provide access to relevant HPC clusters for integration testing and performance validation.
- Test the refactored pipeline in practical use cases, including installation, execution, and result inspection.
- Acknowledge the contributions of Research Software Engineers involved, through co-authorship on relevant publications.

Bibliography

[1] N. Snyder-Mackler, W.H. Majoros, M.L. Yuan, A.O. Shaver, J.B. Gordon, G.H. Kopp, S.A. Schlebusch, J.D. Wall, S.C. Alberts, S. Mukherjee, X. Zhou, J. Tung, Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples, Genetics 203 (2016) 699–714. https://doi.org/ 10.1534/genetics.116.187492 .

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

[3] B.G. Milligan, Maximum-likelihood estimation of relatedness, Genetics 163 (2003) 1153–1167. https://doi.org/ 10.1093/genetics/163.3.1153 .

[4] T.S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data, BMC Bioinformatics 15 (2014) 356. https://doi.org/ 0.1186/s12859-014-0356-4 .

[5] E. Alaçamlı, T. Naidoo, M.N. Güler, M.N. Güler, Ş. Aktürk, I. Mapelli, K.B. Vural, M. Somel, H. Malmström, T. Günther, READv2: advanced and user-friendly detection of biological relatedness in archaeogenomics, Genome Biology 25 (2024) 216. https://doi.org/ 10.1186/s13059-024-03350-3 .

[6] A.B. Rohrlach, J. Tuke, D.R. Popli, W. Haak, BREADR: An R Package for the Bayesian Estimation of Genetic Relatedness from Low-coverage Genotype Data, Journal of Open Source Software 10 (2025) 7916. https://doi.org/ 10.21105/joss.07916 .

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