genomicrelatedness

bwRSE4HPC Project Report

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:	MIT
Link:	https://github.com/nf-core/genomicrelatedness/
Programming Language:	Nextflow, Python
Technologies:	nf-core
Target Cluster:	BinAC2

1. Executive Summary

This bwRSE4HPC-project [1,2] aimed to modernize and modularize an existing Nextflow [3] pipeline that was originally developed for relatedness analysis from low-coverage whole-genome sequencing (lcWGS) data [4], specifically in wild Guinea fowl.

Problem Statement: The original relatedness analysis workflow was monolithic, based on outdated Nextflow DSL1 code, and difficult to maintain or reuse beyond the original project context. This limited reproducibility, portability, and community adoption.

Solution: We implemented a new pipeline from scratch and aligned it with nf-core [5] standards. The resulting pipeline, `nf-core/genomicrelatedness`, modernizes the complete workflow using Nextflow DSL2, modular components, and standardized project structure. We also prepared user-facing usage instructions for BinAC2 to support practical HPC execution.

Impact: The proposal for the pipeline was accepted by nf-core and the project is now available there. At the time of writing, the pipeline is still marked as "under development", with only a few remaining tasks before a `v1.0` release. This work establishes a sustainable and community-facing basis for relatedness estimation from low-coverage sequencing data and substantially improves long-term maintainability and usability.

[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

Figure 1: Official logo of the `nf-core/genomicrelatedness` pipeline.

2. Description of Work

2.1. Initial Problem

The original codebase consisted of legacy DSL1 workflows that were tightly coupled and hard to extend. Missing modular boundaries and limited automation reduced developer productivity and made quality assurance difficult. In addition, the workflow was not integrated into nf-core, which created barriers for wider adoption and constrained reuse in standardized bioinformatics environments. The initial project work plan can be found in Ref. [6].

2.2. Solution Design

The core design decision was to reimplement the pipeline as a new DSL2-based nf-core pipeline rather than applying incremental patching to the legacy implementation. This enabled:

- Clear module boundaries and better maintainability
- Compatibility with nf-core conventions for structure, linting, and release processes
- Easier portability across local and HPC execution profiles
- Long-term sustainability through community standards

Where possible, nf-core conventions and reusable components were adopted. Custom implementation was retained where domain-specific behavior was required.

2.3. Contributions

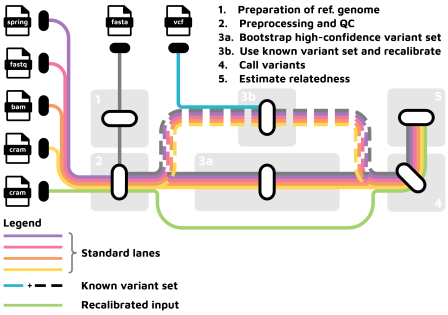
The following contributions were provided during the project:

- Full pipeline rewrite from scratch into the new `genomicrelatedness` pipeline
- Migration to Nextflow DSL2 and modular workflow structure
- Alignment with nf-core development and contribution workflow
- Successful proposal acceptance in nf-core and publication as `nf-core/genomicrelatedness`
- Preparation of user instructions for running the pipeline on BinAC2 bwHPC cluster
- Project documentation improvements to support onboarding and operational use

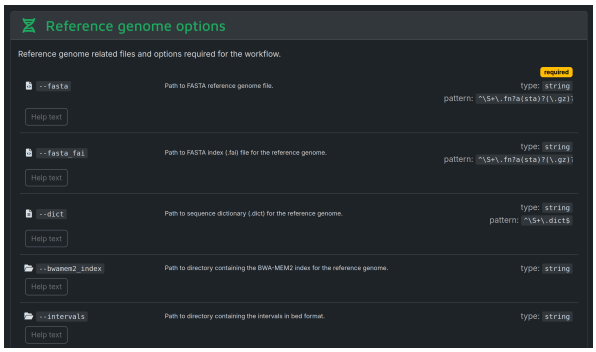
External Constraints for Additional Relatedness Tools The original plan included integration of additional relatedness estimation software, but in several cases this was blocked by external constraints such as unclear or missing licenses, or required [7,8] file format conversion software that was either unlicensed or not properly packaged. These constraints are outside project control, but the modular pipeline design now makes future integration straightforward once those tooling and licensing issues are resolved.

2.4. Showcase

The pipeline's nf-core website presents documentation, usage examples and parameters (see Figure 2) in such a way that onboarding of new users is as easy as it gets.



(a) Workflow schematic metro map.



(b) Parameters section at nf-core.

Figure 2: Pipeline documentation on it's GitHub page and https://nf-co.re/

With Nextflow and nf-core installed (standard), the pipeline can then simply be run via

```
nf-core pipelines launch nf-core/genomicrelatedness -params-file <path>
```

without dealing with repositories or any implementation details.

3. Possible Follow-Ups

The following steps are suggested to project partners and collaborators:

1. Finalize and publish `v1.0` release of `nf-core/genomicrelatedness` upon approval of the nf-core community.
2. Add additional relatedness estimation modules once licensing and packaging constraints of external software are resolved.

Bibliography

[1] T. Isensee, A. Baer, R. Caspart, J. Hörter, G.H. Hunter, D. Kempf, M. Koch, K. Riedmiller, T. Schrader, I. Ulusoy, bwRSE4HPC – Research Software Engineering Support for High-Performance Computing in Baden-Württemberg, (2026). https://doi.org/10.5281/zenodo.18770909.

[2] J. Hörter, D. Kempf, M. Koch, bwRSE4HPC – A blueprint for better research software in Baden-Württemberg, (2026). https://doi.org/10.5281/zenodo.18909685.

[3] 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.

[4] 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.

[5] 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.

[6] T. Isensee, GENOMICRELATEDNESS - Work Plan, (2026). https://doi.org/10.5281/zenodo.18770388.

[7] 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.

[8] 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.