

The nf-core/variantbenchmarking Pipeline within the GHGA Framework

Kübra Narci^{1,2}, GHGA consortium

¹ German Human Genome-Phenome Archive (GHGA), DKFZ, Heidelberg, Germany

² Computational Oncology Group (CO), Molecular Precision Oncology Program, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany

Background: The German Human Genome-Phenome Archive (GHGA) provides a secure, scalable infrastructure for processing and benchmarking human omics data. As part of this effort, the [nf-core/variantbenchmarking pipeline](https://github.com/nf-core/variantbenchmarking) (<https://github.com/nf-core/variantbenchmarking>) was developed to benchmark small variants, indels, and structural variants in germline and somatic datasets, enabling reproducible and standardized analysis.

Materials and Method: Built with Nextflow, the pipeline ensures scalability and reproducibility across diverse environments, including local systems, HPCs, and cloud platforms. Users can benchmark using truth datasets (e.g., Genome in a Bottle, SEQC2) or custom VCF files. The pipeline supports normalization steps such as deduplication, variant splitting, filtration, and alignment correction, as well as benchmarking tools like hap.py, rtgtools, and truvari. It computes precision, recall, and F1 scores, offering detailed performance metrics.

Results: The pipeline facilitates accurate benchmarking and harmonization of variant-calling workflows, generating actionable performance insights like precision, recall, and F1 scores. By integrating with GHGA's secure infrastructure, it supports the preprocessing and benchmarking of human omics data while ensuring FAIR compliance. It is in nf-core github and It adheres to nf-core community guidelines, ensuring high-quality, reviewed code, modularity, and extensibility.

Conclusion: The nf-core/variantbenchmarking pipeline is specifically designed to accommodate varied research requirements, offering a modular approach that allows researchers to apply benchmarking methods most suitable for their analyses. GHGA's architecture is built on cloud computing infrastructures and includes an ethico-legal framework to ensure data protection compliance. GHGA enables researchers to conduct reproducible, rigorous, and secure research by standardizing bioinformatics workflows and governing reusability through harmonized metadata schemas.

Keywords: Variant benchmarking, FAIR-workflows, NGS analysis

Grants: German Research Foundation (DFG) 441914366 (NFDI 1/1).