

Nallo: a Nextflow pipeline for comprehensive human long-read genome analysis

Felix Lenner^{1,2,3}, Anders Jemt^{2,3}, Lucia Peña Pérez^{3,4,5}, Ramprasad Neethiraj³, Peter Pruisscher³, Daniel Schmitz⁶, Annick Renevey³, Pádraic Corcoran^{1,7}, Daniel Nilsson^{3,8}, Jesper Eisfeldt^{5,8}, Anna Lindstrand^{5,8}, Valtteri Wirta^{2,3,8}, Adam Ameer^{1,7,†,✉}, Lars Feuk^{1,7,†,✉}

¹Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, 751 08, Sweden

²Genomic Medicine Center Karolinska, Karolinska University Hospital, Stockholm, 171 77, Sweden

³Science for Life Laboratory, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, 171 77, Sweden

⁴Centre for Inherited Metabolic Diseases, Karolinska University Hospital, Stockholm, 171 77, Sweden

⁵Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, 171 77, Sweden

⁶Clinical Genomics Gothenburg, SciLifeLab, Sahlgrenska Academy, University of Gothenburg, Göteborg, 405 30, Sweden

⁷SciLifeLab, Uppsala University, Uppsala, 751 08, Sweden

⁸Department of Clinical Genetics and Genomics, Karolinska University Hospital, Stockholm, 171 77, Sweden

*Corresponding author: Department of Immunology, Genetics and Pathology, Box 815, Uppsala University, 751 08 Uppsala, Sweden.

E-mail: lars.feuk@igp.uu.se

[†]These authors contributed equally to this article.

Associate Editor: Peter Robinson

Abstract

Motivation: Long-read sequencing (LRS) is increasingly used for human medical research and clinical diagnostics due to its capacity to generate complete genome information. However, there is a lack of robust and easy-to-use pipelines for comprehensive LRS data analysis.

Results: Here we present Nallo, a Nextflow pipeline for analysis of PacBio and Oxford Nanopore data, with additional support for rare disease research projects. The pipeline detects a wide range of genetic variants, performs genome assembly, and reports CpG methylation. It also enables annotation and ranking of variants based on their predicted functional consequences.

Availability and implementation: Nallo is available from GitHub: <https://github.com/genomic-medicine-sweden/nallo>

1 Introduction

Long-read sequencing (LRS) technologies enable identification of a wide range of genetic variants in the human genome as well as DNA modifications. In recent years, the yield and accuracy of LRS have increased drastically, thereby paving the way for large-scale human LRS projects. Given the current trend toward higher throughput and lower cost per base, it is likely that LRS will replace short-read sequencing for many human genome studies. In fact, this transition has already started, with LRS being increasingly used in clinical diagnostics (Steyaert *et al.* 2024) and sequencing of population cohorts (Gustafson *et al.* 2024, Mahmoud *et al.* 2024).

To fully capitalize on the benefits of human LRS data, it needs to be analyzed with bioinformatic tools capable of extracting different types of genetic events, including single nucleotide variants (SNVs), structural variants (SVs), copy number variants (CNVs), tandem repeats (TRs), and methylation (5mC) signals.

Furthermore, the LRS data allow variants to be phased on the two haplotypes and enable *de novo* assembly of a diploid genome sequence (Jarvis *et al.* 2022). Since no single software is capable of providing all relevant analyses, there is a need to collect different tools into a joint workflow, or pipeline. Ideally, the pipeline should be easy to use while producing accurate and reproducible results. Moreover, for clinical research or diagnostics, annotation and ranking of the identified variants are needed in order to facilitate interpretation of the results.

Nallo is a comprehensive pipeline for human LRS analysis. It is developed in Nextflow (Di Tommaso *et al.* 2017), which is increasingly becoming the workflow of choice within the bioinformatics community (Langer *et al.* 2024). Nallo integrates commonly used LRS analysis tools including for alignment, variant calling, and genome assembly, with downstream tools for variant annotation and ranking. All tools are combined into a single pipeline that is under version control and easy to install

and run. Nallo is implemented using the nf-core template (Ewels *et al.* 2020), which ensures standardization of the code and facilitates collaborative development. The current version handles data from the two main LRS technology providers, i.e. Oxford Nanopore Technologies (ONT) and PacBio.

2 Materials and methods

2.1 Overview of the Nallo pipeline

Nallo takes (u)BAM or FASTQ files from any number of (related or unrelated) samples with long sequencing reads as input, i.e. PacBio HiFi reads or base-called ONT reads, and processes them in parallel. The analysis process for a single sample is outlined in Fig. 1, but Nallo also supports family-based analyses, such as trios. Single samples and families can be run simultaneously. Briefly, Nallo starts by aligning the LRS data to the human reference genome, followed by quality control and detection of genetic variants (SNVs, SVs, TRs, CNVs). Then annotation and ranking of potentially deleterious variants is performed, resulting in output files for each sample and combined per family. The alignment files are phased, and if the raw data contain 5mC CpG signals, then this information is preserved and allele-specific methylation pileup data are generated. Furthermore, a *de novo* assembly is created for each individual sample and is aligned to the reference genome. Since some of the analysis tools are preferable for, or restricted to, specific LRS technology, Nallo runs somewhat different tools for PacBio and ONT data, and sub-workflows can be turned on or off at the user's discretion. Nallo has mainly been evaluated on long-read whole genome sequencing (WGS) data but also works for targeted LRS. The complete list of tools in the current version is available online: <https://github.com/genomic-medicine-sweden/nallo>.

2.2 Implementation and run performance

Nallo is implemented in Nextflow using the nf-core template and guidelines. The implementation is modular, which facilitates

adding new tools and tailoring the pipeline for specific needs. It also provides flexibility to bypass analysis steps that are not relevant to a user's specific application, thereby saving compute time and resources. Moreover, the pipeline can be installed on various types of compute infrastructure. To evaluate the performance, we ran Nallo version 0.5.1 on a $\sim 32\times$ coverage PacBio trio (HG002, HG003, HG004). Our results show that a complete Nallo analysis, including alignment, variant calling, and *de novo* assembly, consumes about 344 core hours for each sample, completing in 5 hours 52 minutes by executing parallel jobs on a cluster with 68 compute nodes (dual Intel Xeon Gold 6248R, 24 cores @ 3.0 GHz). On the same compute infrastructure, we further executed Nallo on a larger dataset consisting of 192 in-house PacBio samples at $23\times$ average coverage. This analysis was completed in 34 hours. The Nextflow workflow report for this large Nallo run, including resource use for individual tools in the pipeline, is available as a [data File 1](#), available as [supplementary data](#) at *Bioinformatics* online.

2.3 Annotation, ranking, and filtering of variants

Nallo not only detects genetic variants but can also annotate SNVs, INDELs, SVs, and TRs and reports the results in VCF files. The annotation can be done using public and local databases with variant effect predictions and population frequencies. In addition, Nallo employs a customizable ranking algorithm that allows the user to prioritize variants based on their annotations, such as predicted effects on genes, transcripts, or protein sequences (Stranneheim *et al.* 2021). A filtering step can be applied using criteria such as population allele frequency and genes of interest, thereby generating a shortlist of variants with a high likelihood of being functionally relevant. Annotation, ranking, and filtering of variants are crucial steps in clinical genomics analyses, e.g. when searching for the genetic cause of a rare disease. Since the pipeline was primarily built to replace existing short-read analysis, Nallo currently performs no annotation of methylation signals or of *de novo* assembly results; however, there are external tools that can be used to process the

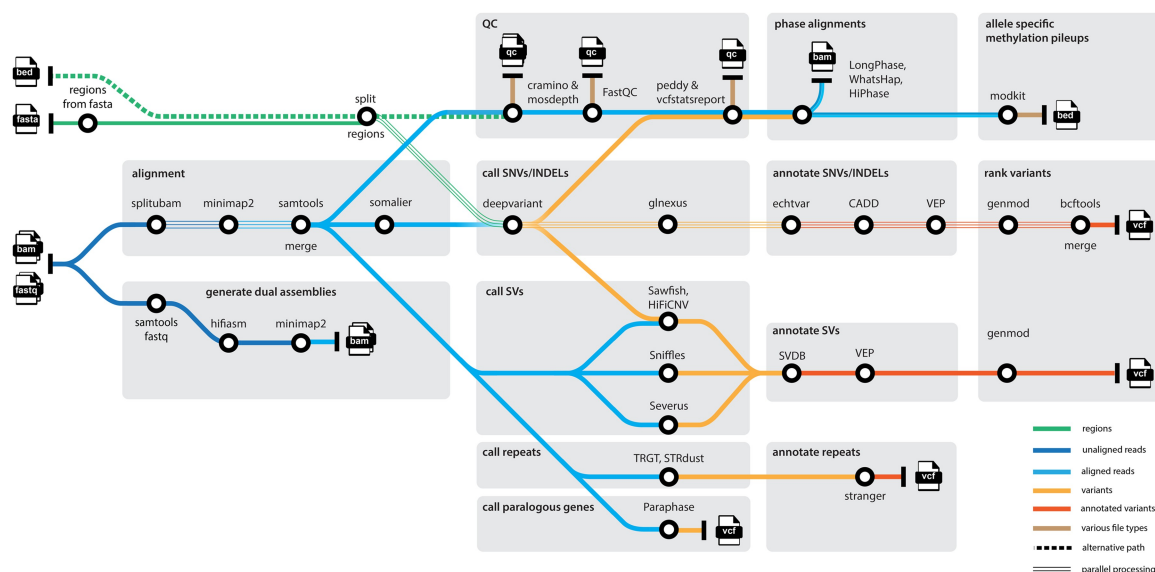


Figure 1 Graphical overview of the Nallo pipeline.

data further, e.g. PAV for assembly-based variant calling (Ebert *et al.* 2021). Furthermore, Nallo is continuously improved, and the addition of more tools is part of the future development plans.

3 Conclusions and future perspectives

In addition to Nallo, there exist other workflows for human LRS analysis, including those from the LRS providers ONT and PacBio. The epi2me platform, which is maintained by ONT, includes a Nextflow workflow for human variation (<https://github.com/epi2me-labs/wf-human-variation>). PacBio has developed its own solution using the workflow description language (WDL) (<https://github.com/PacificBiosciences/HiFi-human-WGS-WDL>). There have also been open-source efforts within the nf-core community, such as nf-core/nanoseq (<https://github.com/nf-core/nanoseq>) and nf-core/pacvar (Jain and Clelland 2025). Contrary to the above pipelines, Nallo is designed to handle both ONT and PacBio data. Furthermore, Nallo has additional features, such as the option to run multiple SV callers, perform single-sample or family-based analyses, and provide annotation and ranking to support the analysis of rare disease variants. However, all pipelines developed within or following the nf-core template can take advantage of the collaborative environment.

Nallo was originally developed for the identification of pathogenic genetic variants in rare disease patients (Eisfeldt *et al.* 2024), but the pipeline can also be used also for projects related to complex disease, pharmacogenomics, and population genomics. Specific application areas could benefit from additional analysis tools, e.g. for HLA and CYP genes, which we intend to include in future releases. Tools for additional quality control and cleanup, such as Breakinator to detect foldback artifacts in ONT data, would also improve Nallo further (Heinz *et al.* 2025). Another limitation with Nallo is that the current version is restricted to germline variation. Although Nallo could be expanded to somatic variation detection and paired tumor-normal analysis in cancer, such analyses are not yet supported.

Since bioinformatic analysis of LRS data is rapidly evolving, we anticipate that Nallo will be regularly updated as new tools and improved software versions become available. In this context, the ability of Nextflow to handle different execution platforms is a big advantage, thereby facilitating updates on local machines, clusters, and cloud environments. Nallo also lays a foundation for larger international collaborations. Having a common framework for LRS analysis facilitates data sharing and reduces batch effects between projects and institutes. In this way, our proposed workflow can play an important role in large-scale projects aiming to better understand the human genome and its role in health and disease.

Acknowledgements

We thank the authors of all tools included in the pipeline and urge all users of Nallo to include citations to the original publications for all individual tools that are used. The Nallo pipeline has been

evaluated using human LRS data from the SciLifeLab National Genomics Infrastructure (NGI) in Uppsala. The authors would like to acknowledge the expertise and support with NGS services provided by Clinical Genomics Stockholm facility at the Science for Life Laboratory (jointly hosted by Department of Microbiology, Tumor and Cell biology at Karolinska Institutet and Department of Gene Technology at School of Engineering Sciences in Chemistry, Biotechnology and Health at KTH Royal Institute of Technology) and the Genomic Medicine Center Karolinska at the Karolinska University Hospital. Computations were performed on resources provided by NAISS through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Author contributions

Felix Lenner (Conceptualization [supporting], Formal analysis [lead], Methodology [lead], Software [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Anders Jemt (Conceptualization [supporting], Data curation [supporting], Methodology [supporting], Supervision [supporting], Writing—review & editing [supporting]), Lucia Peña Pérez (Data curation [supporting], Formal analysis [supporting], Methodology [supporting], Writing—review & editing [supporting]), Ramprasad Neethiraj (Data curation [supporting], Formal analysis [supporting], Writing—review & editing [supporting]), Peter Pruisscher (Formal analysis [supporting], Methodology [supporting], Validation [supporting]), Daniel Schmitz (Formal analysis [supporting], Methodology [supporting], Software [supporting], Writing—review & editing [supporting]), Annick Renevey (Formal analysis [supporting], Methodology [supporting], Validation [supporting]), Pádraic Corcoran (Data curation [supporting], Formal analysis [supporting], Methodology [supporting], Writing—review & editing [supporting]), Daniel Nilsson (Formal analysis [supporting], Methodology [supporting], Writing—review & editing [supporting]), Jesper Eisfeldt (Formal analysis [supporting], Methodology [supporting], Writing—review & editing [supporting]), Anna Lindstrand (Conceptualization [supporting], Funding acquisition [supporting], Writing—review & editing [supporting]), Valtteri Wirta (Conceptualization [supporting], Funding acquisition [supporting], Methodology [supporting], Resources [supporting], Supervision [supporting], Writing—review & editing [supporting]), Adam Ameer (Conceptualization [supporting], Methodology [supporting], Supervision [supporting], Writing—original draft [lead], Writing—review & editing [equal]), and Lars Feuk (Conceptualization [lead], Funding acquisition [lead], Methodology [supporting], Project administration [lead], Resources [lead], Writing—original draft [supporting], Writing—review & editing [lead])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflicts of interest

None to declare.

Funding

This work has been supported by Genomic Medicine Sweden and funding from Hjärnfonden (project FO2022-0207) to L.F.

References

- Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017; **35**:316–9.
- Ebert P, Audano PA, Zhu Q *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021; **372**:eabf7117.
- Eisfeldt J, Ameer A, Lenner F *et al.* A national long-read sequencing study on chromosomal rearrangements uncovers hidden complexities. *Genome Res* 2024; **34**:1774–84.
- Ewels PA, Peltzer A, Fillinger S *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 2020; **38**:276–8.
- Gustafson JA, Gibson SB, Damaraju N *et al.* High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* 2024; **34**:2060–73.
- Heinz JM, Meyerson M, Li H. Detecting foldback artifacts in long-reads. *BMC Genomics* 2026; **27**:144.
- Jain T, Clelland C. nf-core/pacvar: a pipeline for analyzing long-read PacBio whole genome and repeat expansion sequencing data. *Bioinformatics* 2025; **41**:btaf116.
- Jarvis ED, Formenti G, Rhie A *et al.*; Human Pangenome Reference Consortium Semi-automated assembly of high-quality diploid human reference genomes. *Nature* 2022; **611**:519–31.
- Langer BE *et al.* Empowering bioinformatics communities with Nextflow and nf-core. *Genome Biol.* 2025; **26**:228.
- Mahmoud M, Huang Y, Garimella K *et al.* Utility of long-read sequencing for all of us. *Nat Commun* 2024; **15**:837.
- Steyaert W, Sagath L, Demidov G, Solve-RD DITF-EpiCARE *et al.* Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing. *Genome Res* 2025; **35**:755–68.
- Stranneheim H, Lagerstedt-Robinson K, Magnusson M *et al.* Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med* 2021; **13**:40.