nf-core/sarek: an open-source pipeline for germline, tumor-only, and somatic analysis of NGS data

Friederike Hanssen[1], Maxime Garcia[2], Lasse Folkersen[3], Susanne Jodoin[1], Oskar Wacker[1], Anders Sune Pedersen[4], Edmund Miller[5], Francesco Lescai[6], Julius Joos[7], nf-core community, Gisela Gabernet[1], Sven Nahnsen[1]

[1]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen [2]SciLifeLab, Karolinska Institutet, Stockholm [3]Nucleus Genomics Ltd., New York [4] Danish National Genome Center, Copenhagen [5]University of Texas, Dallas [6]Department of Biology and Biotechnology, University of Pavia, Pavia [7]Internal Medicine, University Hospital, Tübingen

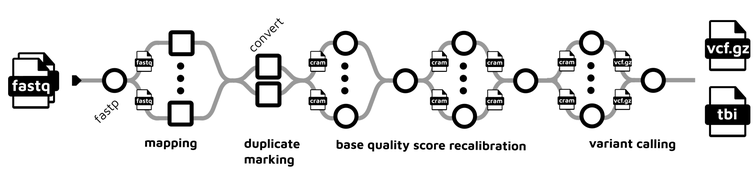
1. Introduction

Somatic variant calling studies often include many patients with dataset sizes varying widely between oncopanel, whole-exome, and whole-genome sequencing data. nf-core/sarek[1] is an established pipeline for exploring single-nucleotide variants, structural variation, microsatellite instability, and copy-number alterations of germline, tumor-only, and paired tumor-normal short-reads.

nf-core/sarek is part of nf-core[2], a community project which provides an infrastructure to create reproducible, scalable, and portable open-source Nextflow[3]-based pipelines.

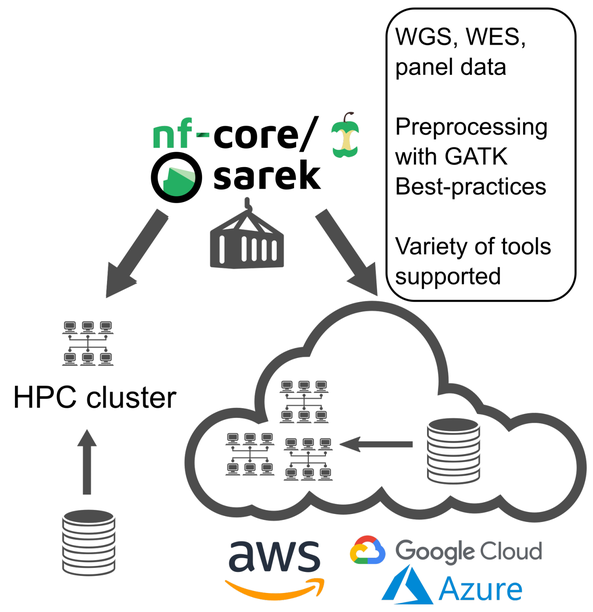
Here, we show the latest updates including improvements to the data flow and tool selection reducing time and compute resources, and modularization improving code maintainability.

2. Overview



- FASTQ or BAM inputs are split into files of equal size before alignment to speed up computation
- Resulting BAM files are then merged and duplicate marked in one step before they are converted into CRAM format.
- Subsequent steps are run on multiple genomic regions in parallel. By default an interval file with chromosomes cut at their centromeres is used for WGS, and a user-supplied target bed file is used for WES or panel data.
- For all data types, small regions are grouped resulting in approximately equal sizes being processed together.

3. Nextflow pipelines are portable



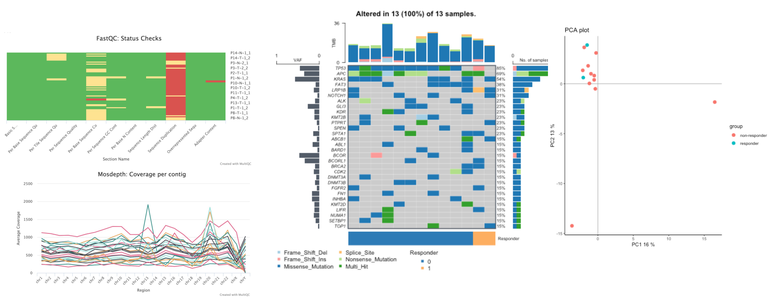
nf-core pipelines are containerized and versioned to ensure reproducible analyses. They are portable to different compute infrastructures mitigating time-consuming data transfer.

Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).

Pipeline metromap showing a high-level view of the different analysis steps. The pipeline can be started from six different entry points and run through all subsequent tasks. All optional tools can be selected in any combination. This allows to recompute and extend the results throughout a project's duration.

4. Application

Showcase of somatic SNP/Indel calling results (Strelka2) on 13 onco-panels datasets comparing treatment responses



Extensive QC report using the tool MultiQC

Oncoplot generated with the R package maftools to compare SNP variants of both groups

PCA plot to evaluate clustering of the mutational signatures

Tools	SNP recall	SNP precision	F1
Haplotypecaller	0.990652	0.992179	0.9915
Deepvariant	**0.992866**	0.99787	**0.99536**
Freebayes	0.992847	0.954791	0.97345
Strelka2	0.983946	**0.998364**	0.9911

Tools	INDEL recall	INDEL precision	F1
Haplotypecaller	0.971052	0.985107	0.97803
Deepvariant	**0.983122**	0.988865	**0.98599**
Freebayes	0.945148	0.94993	0.94753
Strelka2	0.959560	**0.990090**	0.97458

Benchmark of germline variant calling on the Genome in a Bottle sample HG0002 (26X coverage, Illumina NovaSeq, concatenated fastq files) using the pipeline's default settings. Variants in the high-confidence regions were evaluated.

Conclusion

- nf-core/sarek is a high-throughput reproducible pipeline ready to be used in high throughput variant calling projects.
- As a showcase project, 161 WGS germline samples were already analyzed with SNP, SV and CNV calling on a local HPC
- Cost and time evaluation on AWS cloud is currently under way.
- Continuous optimization & addition of community-requested tools
- Possible application: Reanalysis of ICGC /TCGA cohorts for comparative analyses with local cohorts

Join us

https://nf-co.re/sarek

References

1. Garcia et al. (2020), F1000Research 9:63
2. Ewels et al. (2020), Nature Biotechnology 38, 276–278
3. Di Tommaso et al. (2017), Nature Biotechnology, 35(4), 316–319

Acknowledgements

We would like to acknowledge funding from the Excellence cluster iFIT, and the SFB 209 & Amazon Web Services for cloud computing.

We are grateful to the nf-core and nextflow community for their support during the development.