

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

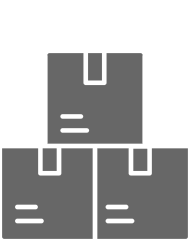
Friederike Hanssen¹, Maxime Garcia², Lasse Folkersen³, Susanne Jodoin¹, Oskar Wacker¹, Anders Sune Pedersen⁴, Edmund Miller⁵, Francesco Lescai⁶, Nick Smith⁷, nf-core community, Gisela Gabernet¹, Sven Nahnsen^{1,8}

¹Quantitative Biology Center, University of Tübingen, Tübingen ²SciLifeLab, Karolinska Institutet, Stockholm ³Nucleus Genomics Ltd., New York ⁴Danish National Genome Center, Copenhagen ⁵University of Texas, Dallas ⁶Department of Biology and Biotechnology, University of Pavia, Pavia ⁷German Human Genome-Phenome Archive ⁸Biomedical Data Science, Department of Computer Science, University of Tübingen, 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/sarek¹ 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-core², a community project which provides an infrastructure to create reproducible, scalable, and portable open-source Nextflow³-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. Implementation Details



Ported to DSL2:

- Using nf-core/modules for 80 of 82
- Related steps are bundled into subworkflows



Extensive CI using pytest:

- Testing for docker, conda, and singularity
- Md5sum or file content checks
- Tests for a variety of use cases



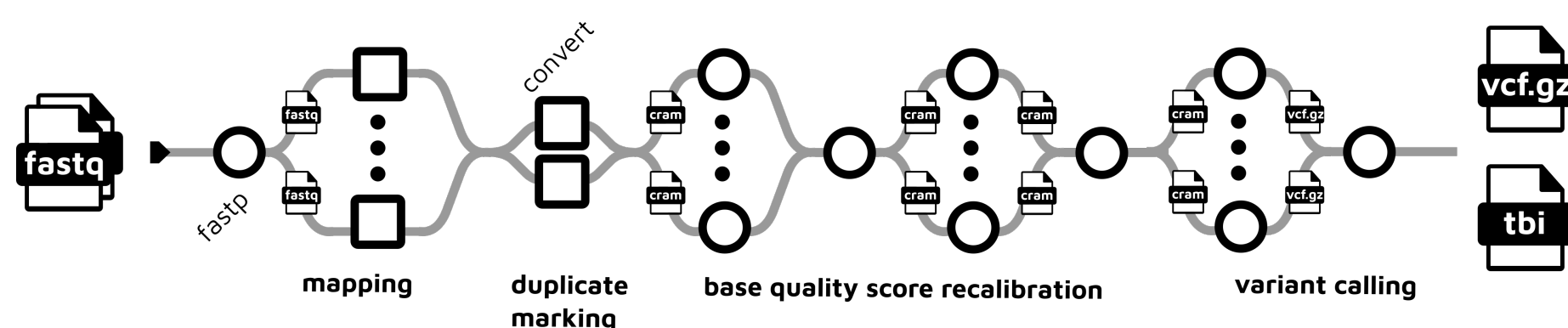
Usage of CRAM files



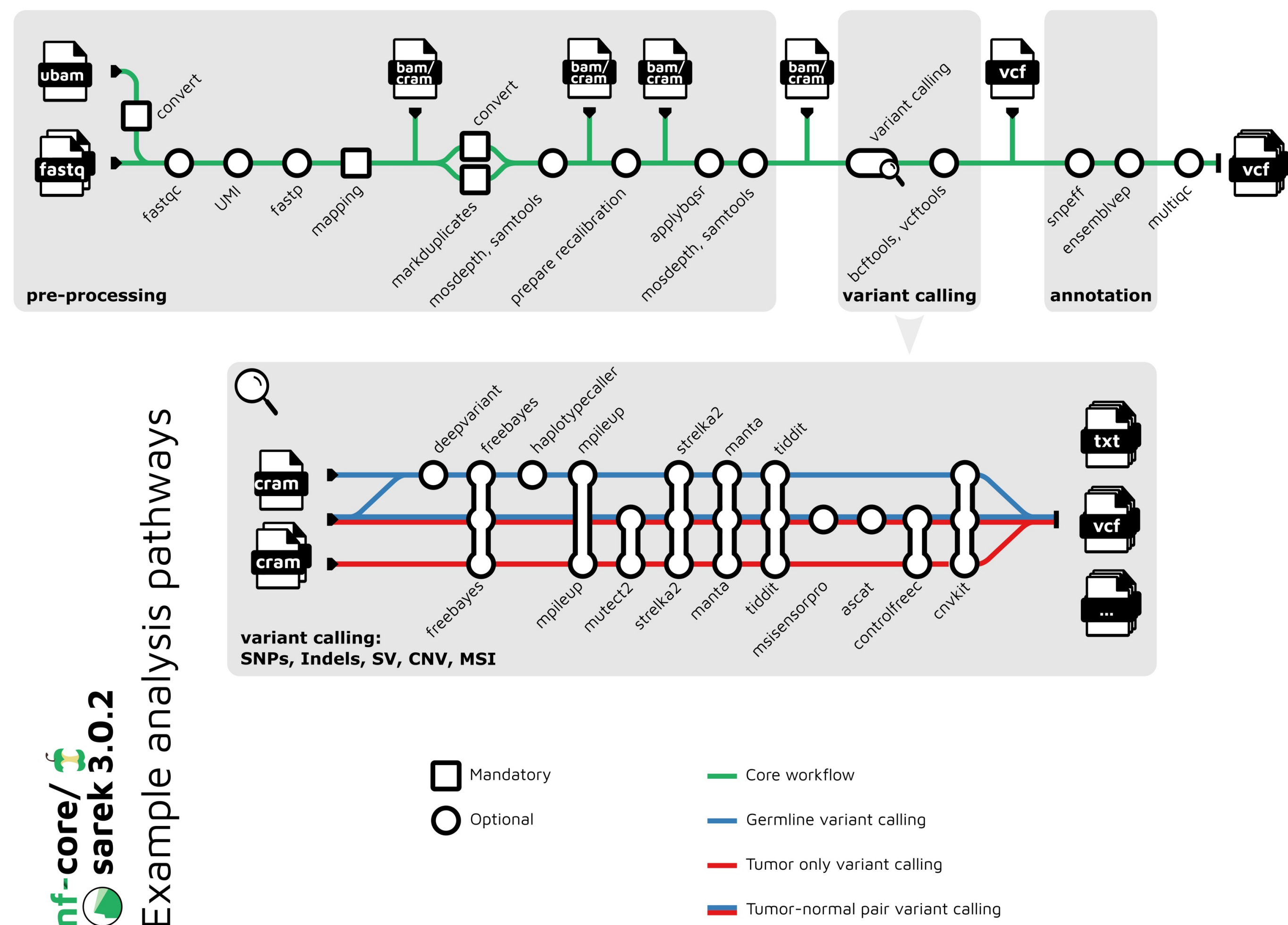
Automatic testing on AWS on release with:

- Germline: GIAB HG0001
- Somatic: HCC1395
- Both have a ground truth set available and can be used for validation

3. Overview



- FASTQ or BAM inputs are split into files of equal size before alignment to speed up computation. Resulting BAM files are 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.
- Larger interval groups reduce storage space consumption but increase runtime



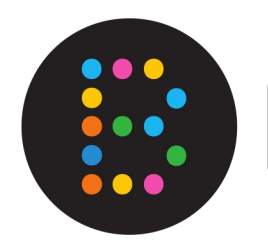
Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).

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. Supported by



SciLifeLab



Barntumörbanken



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

- Garcia et al. (2020), F1000Research 9:63
- Ewels et al. (2020), Nature Biotechnology 38, 276–278
- 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.