

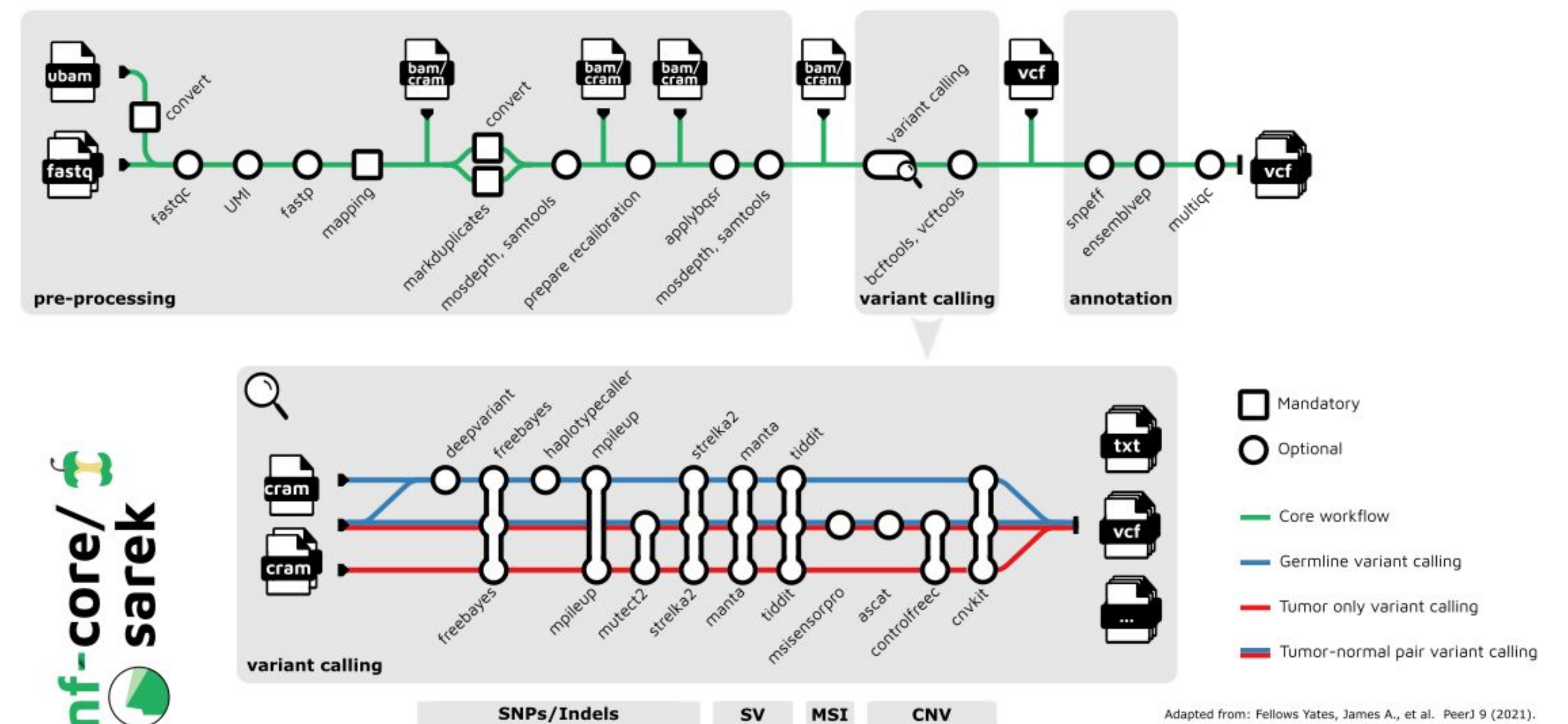
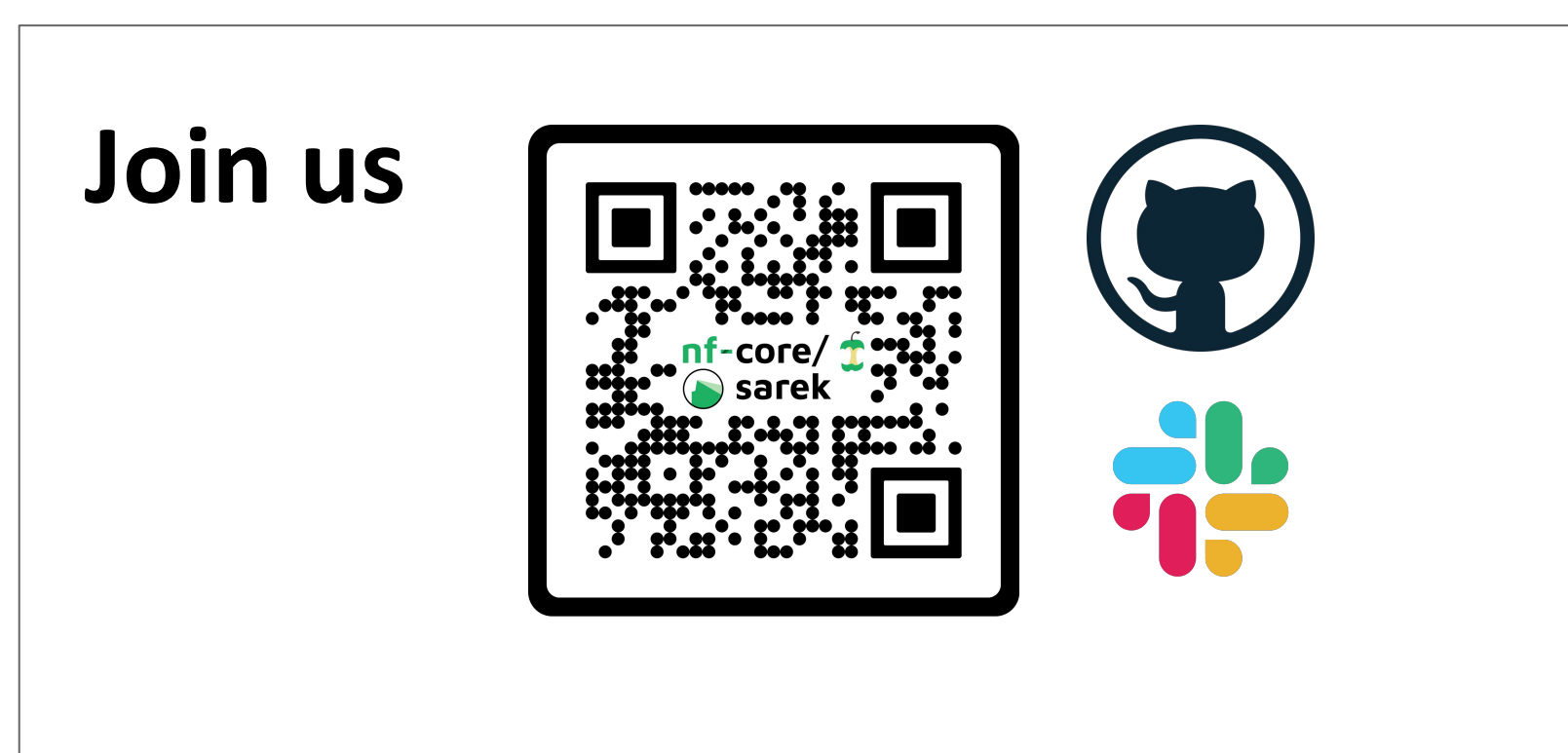
nf-core/sarek: a pipeline for efficient germline, tumor-only, and somatic analysis of NGS data on different compute infrastructures

Friederike Hanssen^{1,2}, Maxime U. Garcia³, Lasse Folkersen⁴, Susanne Jodoin¹, Oskar Wacker¹, Anders Sune Pedersen⁵, Edmund Miller⁶, Francesco Lescai⁷, Nick Smith⁸, nf-core community, Gisela Gabernet¹, Sven Nahnsen^{1,2}

¹Quantitative Biology Center, University of Tübingen, Tübingen ²Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen ³Seqera Labs, Barcelona ⁴Nucleus Genomics Ltd., New York ⁵Danish National Genome Center, Copenhagen ⁶University of Texas, Dallas ⁷Department of Biology and Biotechnology, University of Pavia ⁸German Human Genome-Phenome Archive

Overview

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, compute resources, and cloud computing costs, as well as modularization improving code maintainability. Preprint available at: [biorxiv.org/content/10.1101/2023.07.19.549462v1](https://www.biorxiv.org/content/10.1101/2023.07.19.549462v1)



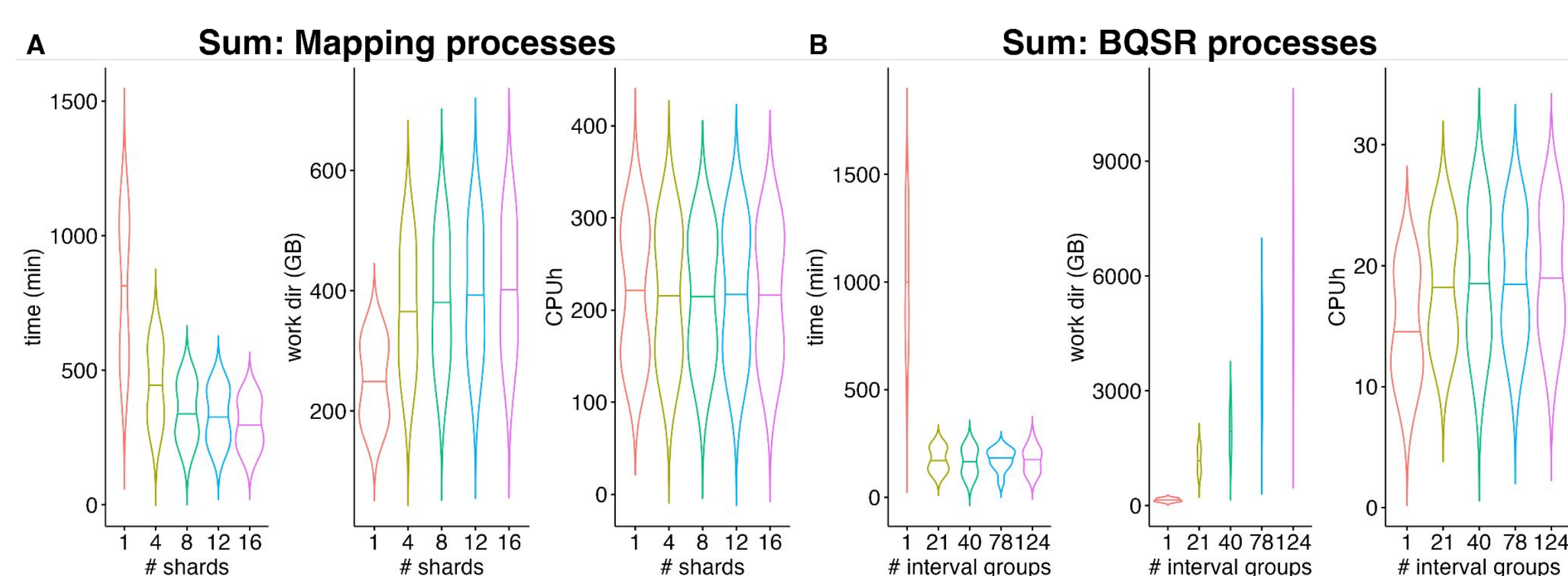
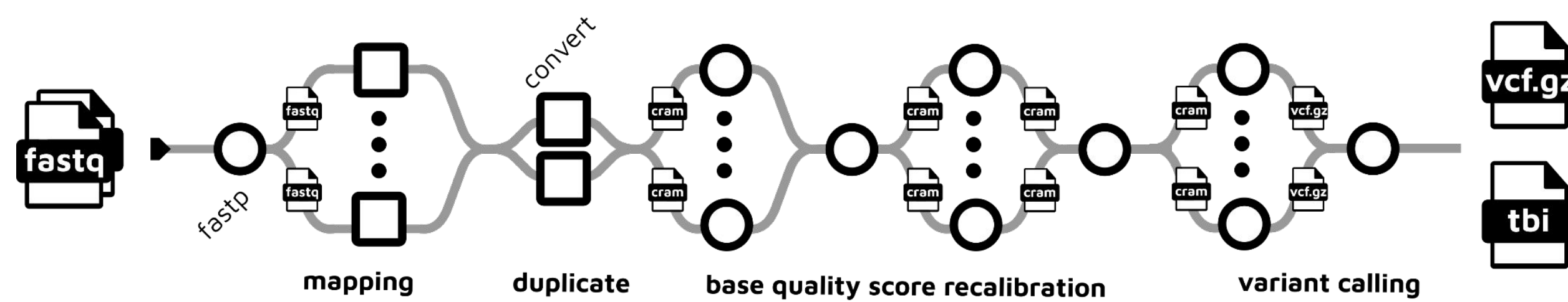
Pipeline metro map showing a high-level view of the different analysis steps. The pipeline can be started from six different entry points and run through all subsequent tasks. All optional tools can be selected in any combination. This allows to recompute and extend the results throughout a project's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

CRAM files reduce storage usage

Using CRAM format, allows us to **reduce work storage needs by 66%**. The additional memory needs are distributed over many tasks and usually comprised of a couple of GB at most, thus not posing a limiting factor in practice.

	BAM	CRAM
CPUh	3761.1	3252.4
Total memory (GB)	7738.5	10346.8
Work storage (TB)	170.4	59.7

Scatter/gathering speeds up analysis



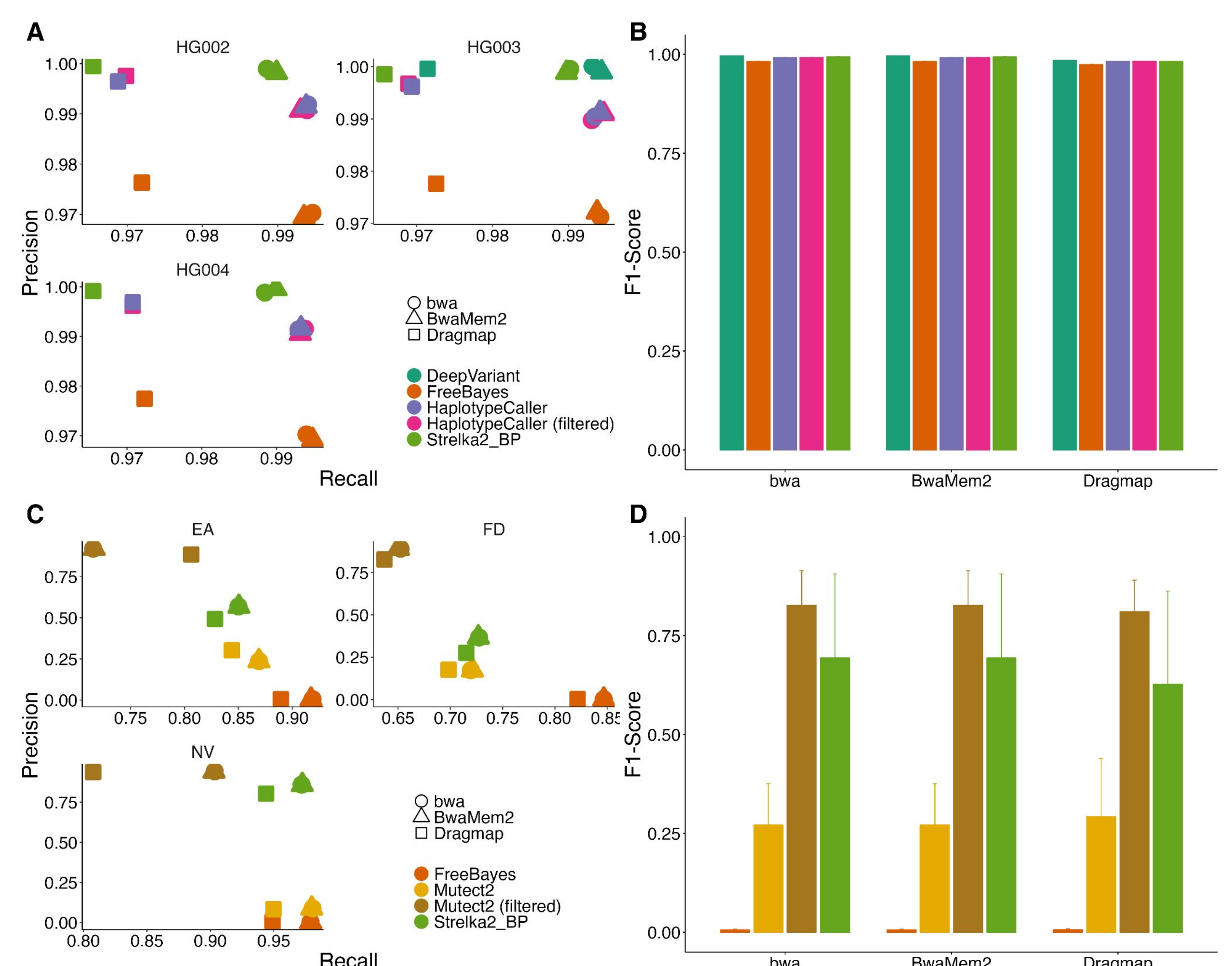
Inputs are split before alignment to speed up computation. Resulting BAM files are merged, duplicate marked, and converted into CRAM format. Speed up plateaus at 12 shards with a **median decrease of 37% in runtime**. BQSR and variant calling are run on multiple genomic regions in parallel. For WGS data interval files from the GATK resource bundle are provided. Users can supply their own, i.e. for WES or panel data. There were no further speed improvements beyond 21 interval groups on hg38. Storage needs increase with more interval groups for BQSR.

Cloud computation costs are reduced by two-third

	Costs
2.7.2	68.04\$
3.1.1	20.82\$

A 35X germline sample was run three times on AWS Batch (us-east-1, spot-instances). For nf-core/sarek 3.1.1 31 interval groups were used. Our optimisations **reduced the price per sample by two-thirds** enabling users to run the pipeline on commercial clouds at a reasonable price, a feature of growing importance for large scale analyses projects.

Benchmarking against truth datasets



We benchmarked nf-core/sarek with various datasets with respect to precision, recall, and F-score. The germline track was evaluated with 3 WGS Genome in a Bottle samples (A,B), the somatic track with 3 WES samples from SEQ2C (C,D). The results are in-line with reported values for these datasets.

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.

Literature

- 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