EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Quantitative Biology Center

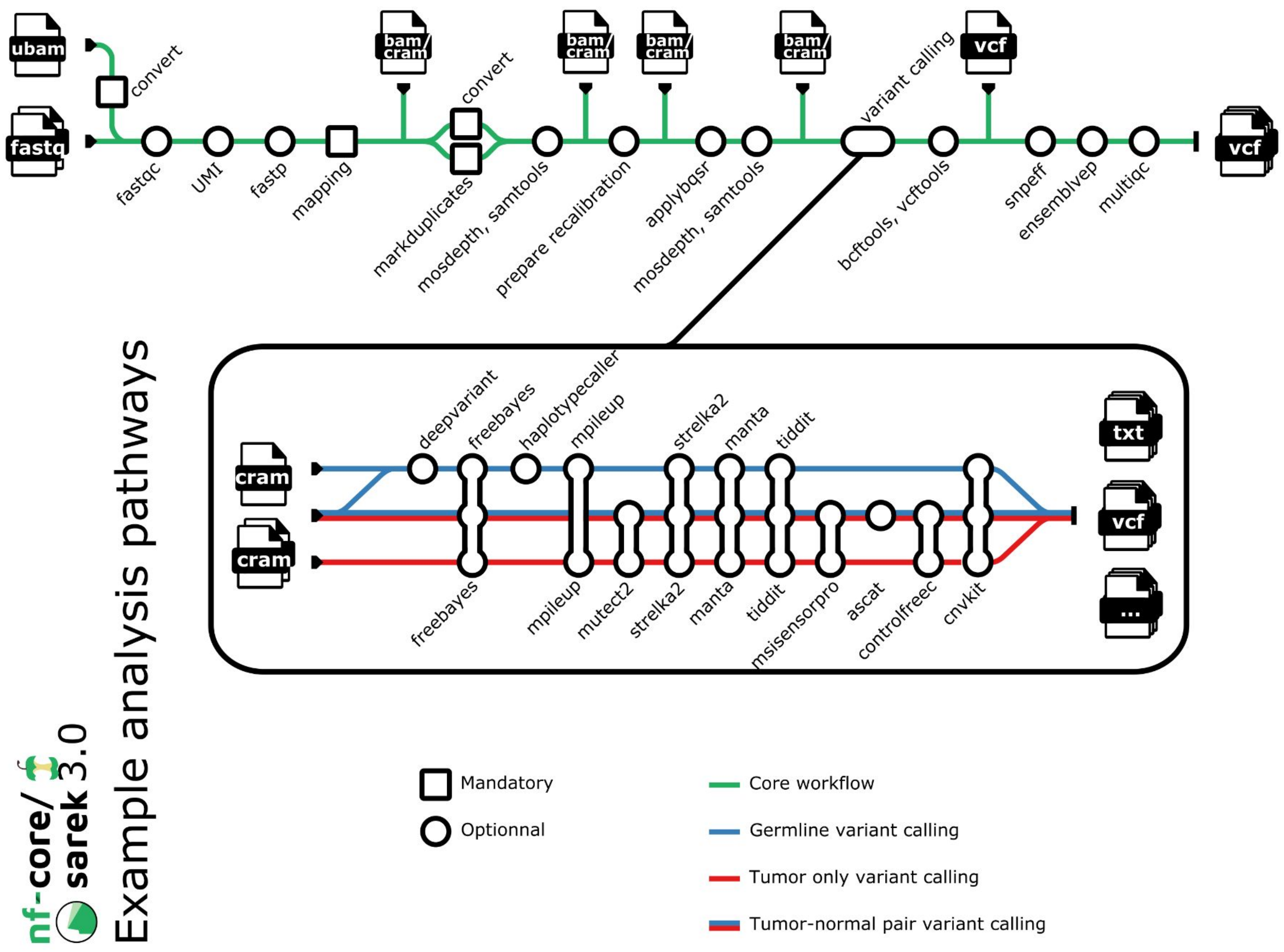
QBiC

# Optimization of nf-core/sarek

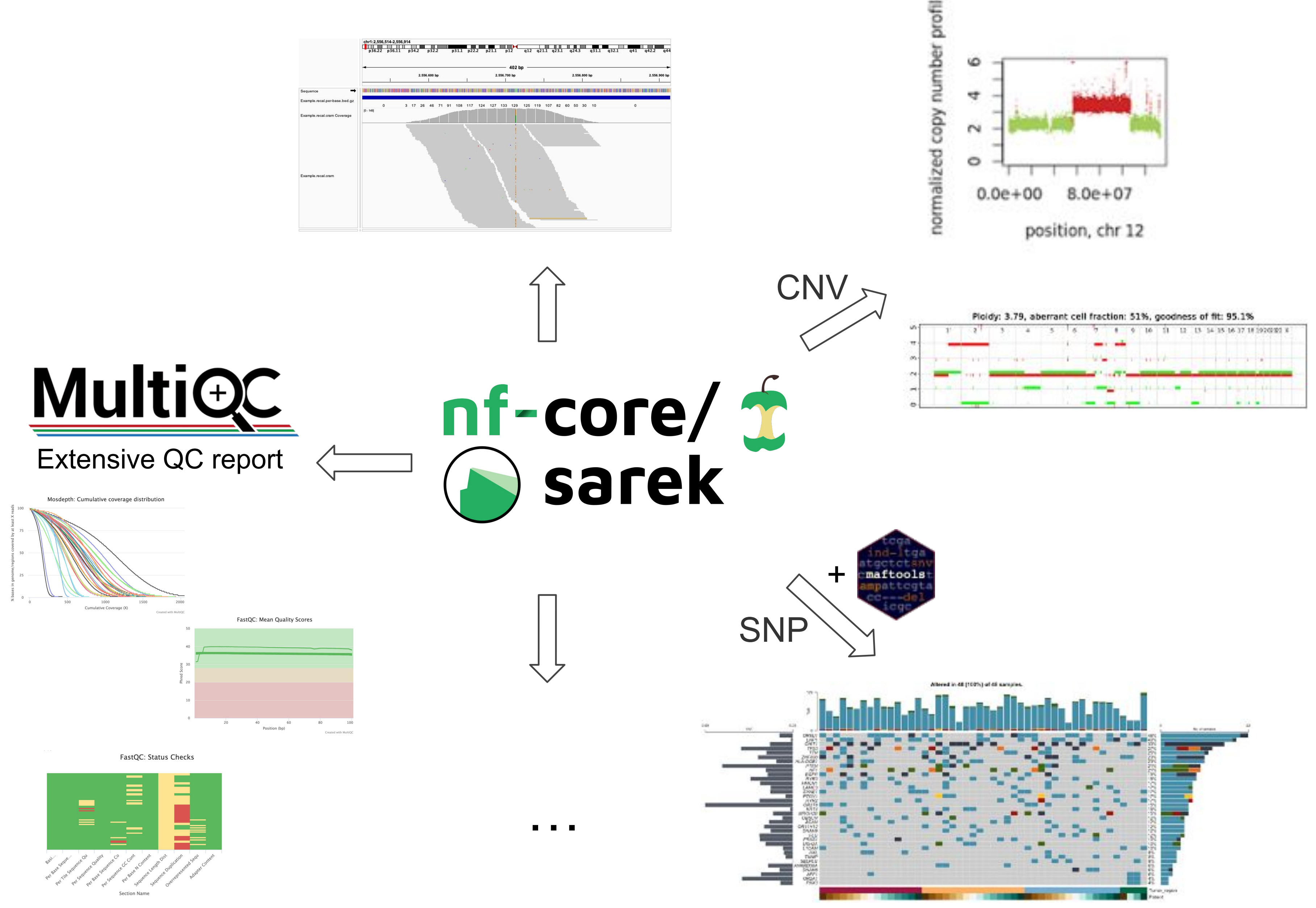
Friederike Hanssen[1], Maxime Garcia[2], Gisela Gabernet[1], Sven Nahnsen[1]
[1]Quantitative Biology Center(QBiC), University of Tuebingen, [2]SciLifeLab, Karolinska Institutet, Stockholm

Somatic variant calling studies often include many patients with dataset sizes varying widely between oncopanel, whole-exome, and whole-genome sequencing data. nf-core[1] provides reproducible, scalable, and portable open-source Nextflow[2]-based pipelines. nf-core/sarek[3] 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. Here, we show the latest updates to the pipeline including improvements to the data flow and tool selection reducing time and compute resources and, modularization improving maintainability.

## PIPELINE OVERVIEW

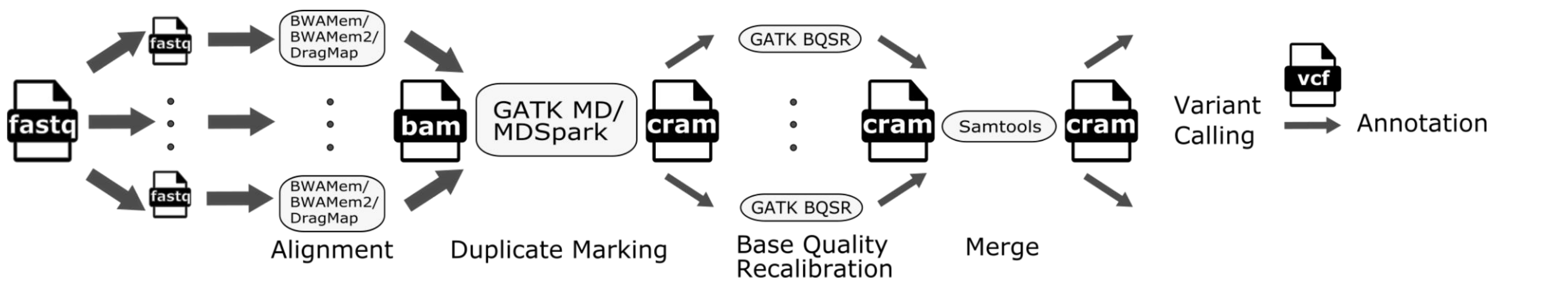


## EXAMPLE OUTPUTS



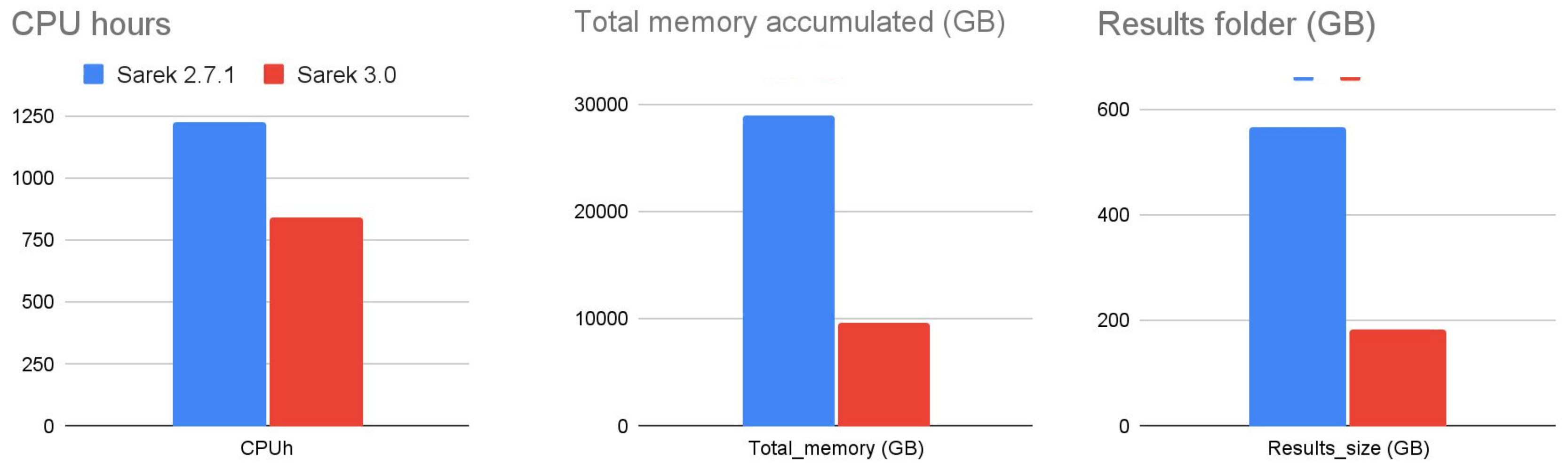
Oncoplot generated from the output vcf together with the R package maftools

## OPTIMIZING DATA FLOW



- FASTQ or BAM inputs are split into files of equal size before alignment.
- 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 for WGS a interval file with used with chromosomes cut at their centromers, for WES or panel data a user-supplied target bed file is used.
- For all data types, small regions are collected resulting in approximately equal sizes being processed together.

## RESULTS



- Somatic variant calling on 41 tumor/normal pairs, panel data, 708 genes
- Adapters trimming (trimgalore vs fastp), bwa, duplicate marking, BQSR, Strelka, Manta, VEP, all available QC steps respectively

## Literature

1. Ewels, P.A., Peltzer, A., Fillinger, S. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38,** 276–278 (2020).
2. Di Tommaso, P., Chatzou, M., Floden, E. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35,** 316–319 (2017)
3. Garcia, M., Juhos, S., Larsson, M. et al. "Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants." *F1000Research* vol. 9 63. 29 Jan. 2020

**Universität Tübingen**
**Quantitative Biology Center (QBiC)**
Auf der Morgenstelle 10, D-72076 Tübingen
Phone +49 7071 29-76499 · http://qbic.life

iFIT
Visualizing and Targeting Cancer Stress
LIVER CANCER
SFB/TR 209
aws

Join us
https://nf-co.re/sarek
GitHub
slack