Sarek, a workflow for WGS analysis of germline and somatic mutations

Maxime Garcia[1], Szilveszter Juhos[1], Malin Larsson[2], Teresita Diaz de Ståhl[3], Johanna Sandgren[3], Jesper Eisfeldt[4], Sebastian DiLorenzo[5], Marcel Martin[6], Pall Olason[7], Phil Ewels[8], Björn Nystedt[7], Monica Nistér[3], Max Käller[9]

1 - BarnTumörBanken, Dept. of Oncology Pathology, Science for Life Laboratory, Karolinska Institutet
2 - Dept. of Physics, Chemistry and Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Linköping University
3 - BarnTumörBanken, Dept. of Oncology Pathology, Karolinska Institutet
4 - Clinical Genetics, Dept. of Molecular Medicine and Surgery, Karolinska Institutet
5 - Dept. of Medical Sciences, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University
6 - Dept. of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University
7 - Dept. of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University
8 - Dept. of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University
9 - Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Royal Institute of Technology

Summary

Sarek

Portable WGS germline and normal/tumor pairs analysis **workflow** written in

nextflow

Easily deployable with **containers**

docker

Preprocessing based on GATK best practices

Sarek Germline

Variant Calling with:
- HaplotypeCaller
- Manta
- Strelka

Sarek Somatic

Variant Calling with:
- ASCAT
- Freebayes
- HaplotypeCaller
- Manta
- MuTect1
- MuTect2
- Strelka

Annotation with:
- snpEff
- VEP

Reports aggregated by MultiQC

Can be used on amazon web services

Open source, contribute on GitHub

Join the chat on Gitter

Acknowledgements

BARNCANCER FONDEN

UPPMAX

We present Sarek, a **portable** Open Source pipeline to resolve **germline** and **somatic** variants from WGS data: it is written in **Nextflow**[1], a domain-specific language for workflow building.

It **processes normal samples** or **normal/tumor pairs** (with the option to include matched relapses).

Sarek is **based on GATK best practices** to prepare short-read data, which is done in parallel for a tumor/normal pair sample.

After these preprocessing steps several variant callers scan the resulting BAM files:

- **Manta** for structural variants

- **Strelka** and **GATK HaplotypeCaller** for germline variants

- **Freebayes**, **MuTect1**, **MuTect2** and **Strelka** for somatic variants

- **ASCAT** to estimate sample heterogeneity, ploidy and CNVs

At the end of the analysis the resulting VCF files can be annotated to facilitate further downstream processing.

Fig1: CPU usage for 90x tumor/normal pair sample (hours)

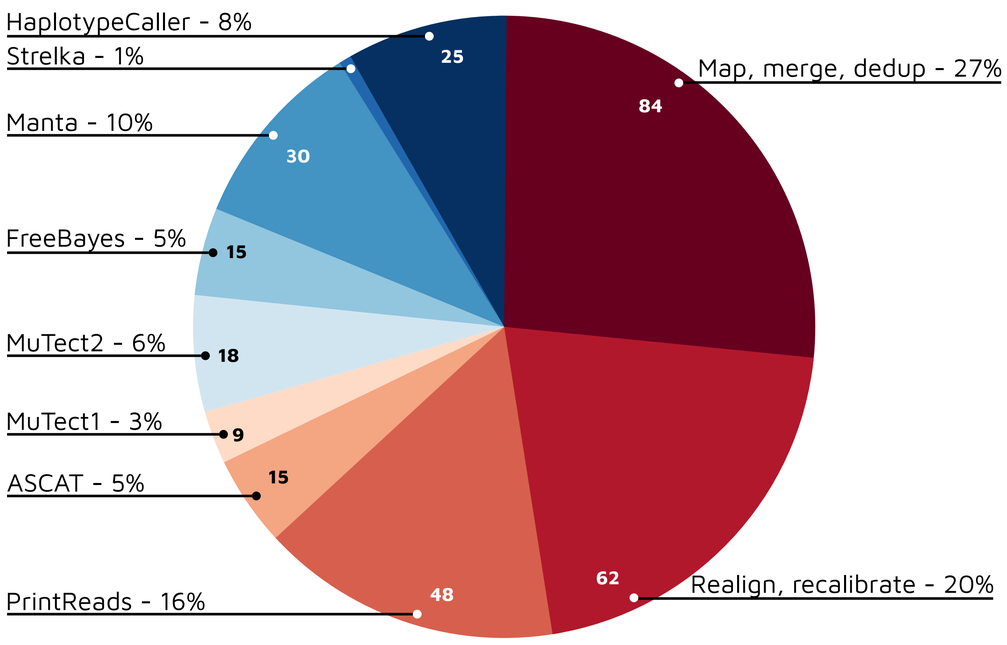
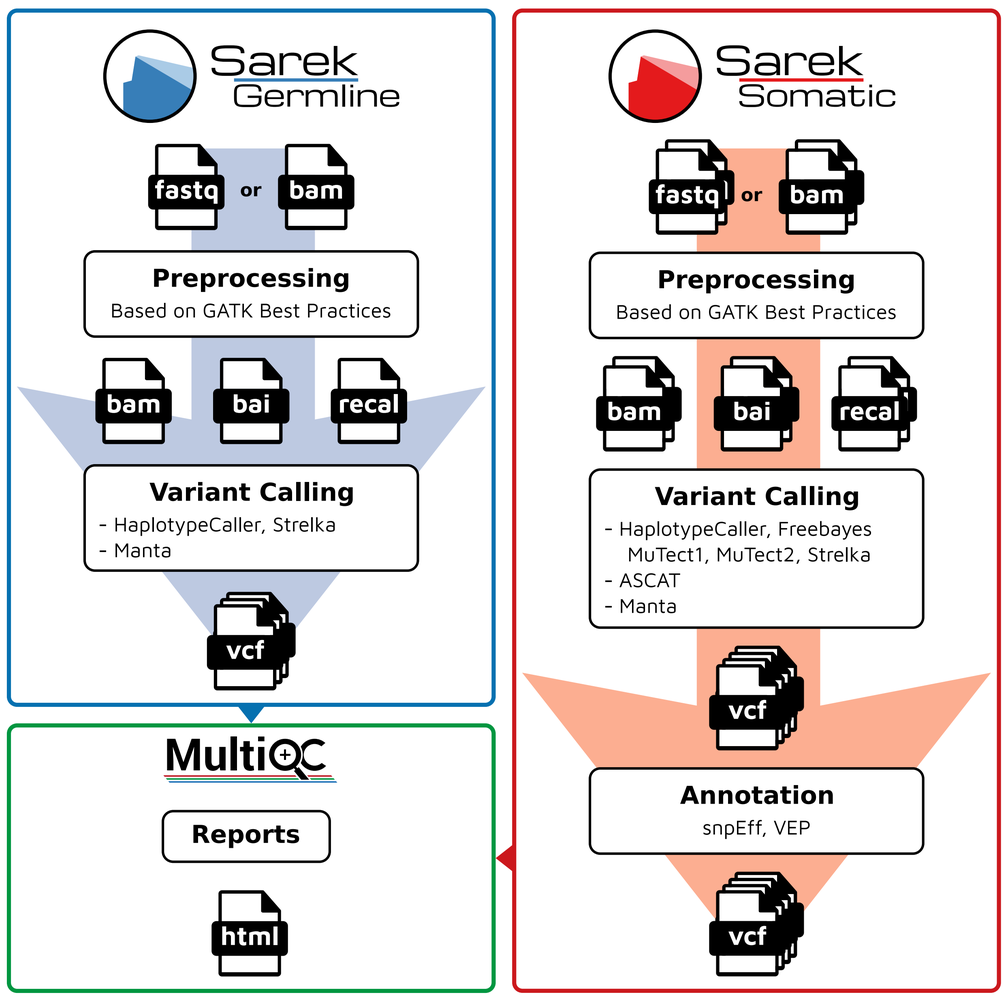


Fig2: Workflow organization



Sarek is based on **Docker** and **Singularity**[2] **containers**, enabling version tracking, reproducibility and handling sensitive data.

The workflow is capable of accommodating further variant callers.

Besides variant calls, the workflow provides quality controls presented by **MultiQC**[3].

Checkpoints allow the software to be started from FastQ, BAM or VCF.

The pipeline currently use **GRCh37** or **GRCh38** as a reference genome, it is also possible to add **custom genomes**.

The **MIT licensed** Open Source code can be downloaded from GitHub.

References

1: doi.org/10.1038/nbt.3820
2: doi.org/10.1371/journal.pone.0177459
3: doi.org/10.1093/bioinformatics/btw354

Links

opensource.scilifelab.se/projects/sarek
github.com/SciLifeLab/Sarek
gitter.im/SciLifeLab/Sarek
ngisweden.scilifelab.se

Acknowledgements

The authors thank the Swedish Childhood Cancer Foundation for the funding of Barntumörbanken. We would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure.