EBERHARD KARLS UNIVERSITÄT TÜBINGEN

QBiC

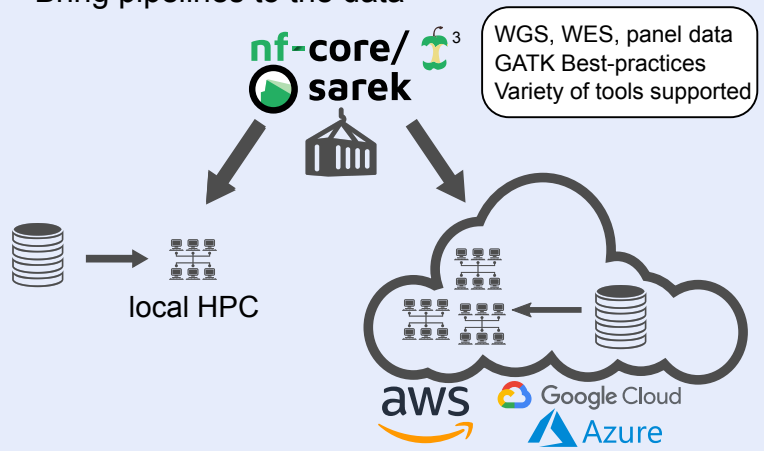
Optimization of nf-core/sarek for large-scale analysis of public cancer data in the cloud

Friederike Hanssen,* Maxime Garcia,+ Gisela Gabernet,* Sven Nahnsen*

*Quantitative Biology Center(QBiC), University of Tuebingen, +SciLifeLab, Karolinska Institutet, Stockholm

Introduction

- nf-core[1] provides portable, reproducible Nextflow[2] based pipelines
- (Re-)analyzing public data can support own data
- Many cancer DBs available in commercial clouds
- Datasets can be large:
  i.e. 300GB WGS/patient (tumor/normal)
- Bring pipelines to the data



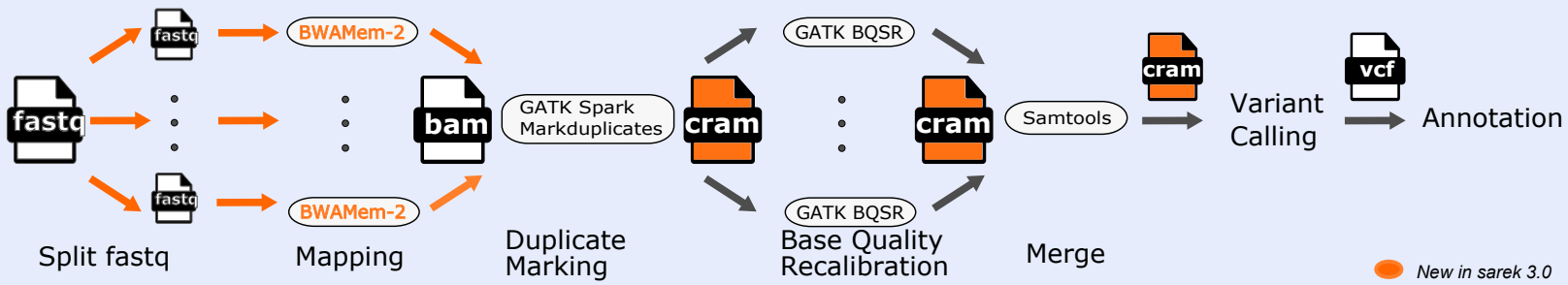
- **Limited compute resources**
- Data download time-consuming

- **Expensive**
- Data upload time-consuming
- Data security concerns

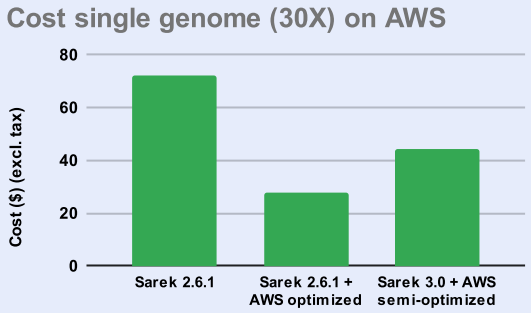
Methods

- CRAM: Storage reduction by 30-50%[4]
- Improve "Preprocessing":
- Change & update tools to reduce resource usage
- Improve cloud usage

nextflow tower + aws



Current Results & Outlook



- Further improve sarek workflow
  Split input into equal sizes to allow precise resource requests for mapping
- Tailor AWS setup & requested resources to new workflow
- Evaluate other commercial cloud providers
- Compare resource usage for the whole pipeline on the local HPC: Sarek 2.6.1 vs Sarek 3.0

Citations

1. Ewels et al. (2020), Nature Biotechnology 38, 276–278
2. Di Tommaso et al. (2017), Nature Biotechnology, 35(4), 316–319
3. Garcia et al. (2020), F1000Research 9:63
4. https://www.ga4gh.org/cram/

Funding

iFIT DFG LIVER CANCER SFB/TR 209 aws