EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Quantitative Biology Center
(QBiC)

GHGA

dkfz.

QBiC

Homogeneous data processing with public data

GHGA flex funds

Project Goal (as described in application)

Loose coupling of external data to GHGA to increase the number of phenotype-genotype pairs and if possible integration of (clinical) metadata.

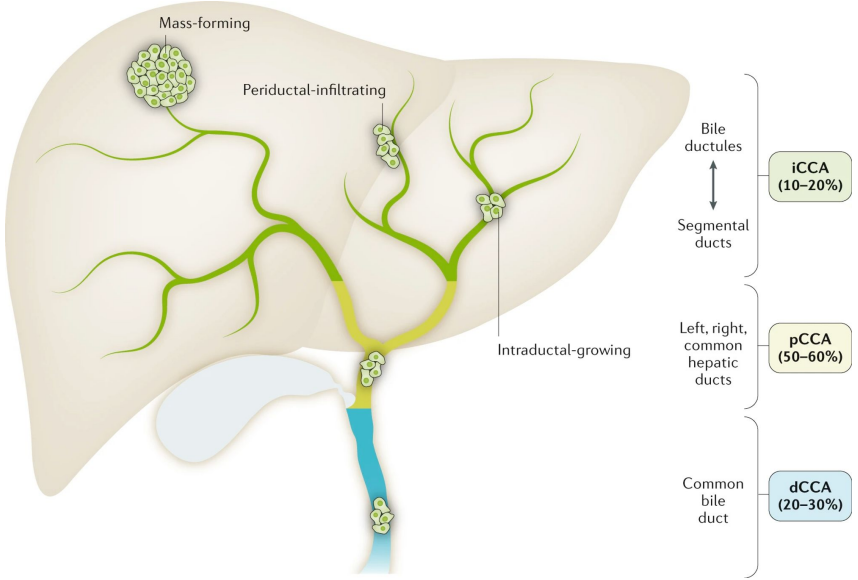
Using GHGA pipelines (DKFZ OTP and nf-core/sarek), we integrate GHGA with external data into ad-hoc variant stores for ML and genomics applications.

Investigation of batch effects through comparison of homogeneously processed data and inhomogeneously processed data.

Biological Background

Cholangiocarcinoma (CCA) is a malignant tumor of the bile ducts which can usually only be diagnosed by tissue examination.

There is medical expertise and interest in Tübingen and Heidelberg towards CCA.



Banales et al., Nat Rev Gastroenterol Hepatol, 2020

Task 1: Test out GHGA process for data access

- We want to go through the GHGA process to get access to data from the MASTER cohort in HD (until data freeze 2018, available in GHGA metadata catalog) at the QBiC in Tü
  - needs DACO

Task 2: Homogeneous data processing / data

- TCGA-CHOL: 51 WXS / 49 WGS patients + called variants on WXS (Pindel, Mutect2, VarScan, MuSE, Ascat, Arriba)
- MASTER: ~100 patients
  - 9 WGS / 20 WXS patients available via GHGA metadata catalog
  - others only accessible for people working in HD
- Tü CHOL: ~35 patients WXS
  - access probably possible in Tü
- ICGC: BTCA_SG
  - ongoing work to download data
  - maybe processing on AWS is possible
- UK Biobank CCA cohort: 300 intrahepatic + 117 extrahepatic patients

Analysis:

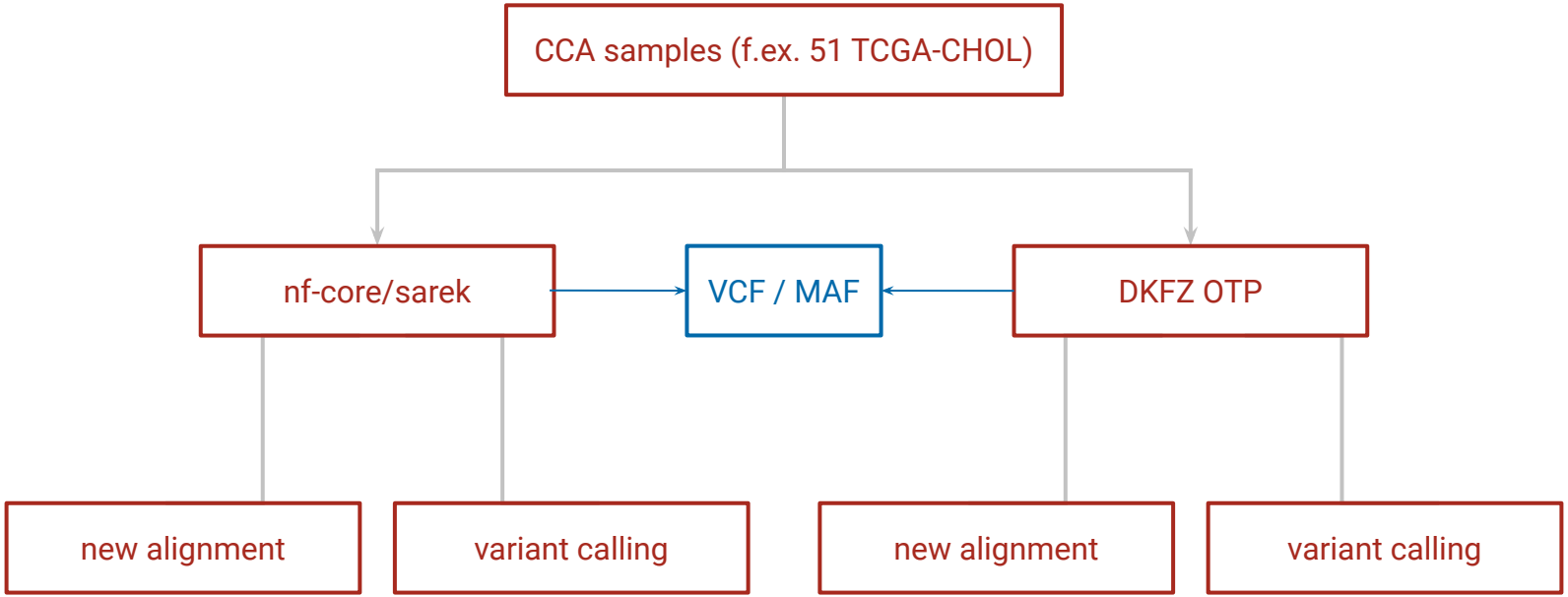
- nf-core/sarek
- DKFZ OTP

Reference genome:

- GRCh38 / hg38

Data overview table

Task 2: Homogeneous data processing / process



Questions we are trying to answer

- Regarding data access: Does the GHGA process work?
- *Hypothesis 1* “Homogenous (using a single pipeline) processing reduces the variation for integrated variant analysis” → reject / accept
- *Hypothesis 2* “The choice of the pipeline (OTP/SAREK) is less important than the choice of the processing strategy (homogeneous/heterogeneous)” → reject / accept
- Can we quantify the difference between start from alignment and start from variant calling? Is a re-alignment necessary?

Involved People

- Famke Bäuerle: works @QBiC, data analysis nf-core/sarek
- Kübra Narci: works @DKFZ, data analysis with OTP
- Sabrina Krakau: Team leader of RDDS @QBiC, project supervision
- Sven Nahnsen: Director of QBiC, DACO application for GHGA data
- Daniel Hübschmann: Head of the research group Computational Oncology @DKFZ
- Ivo Buchhalter: Head of Omics IT and Data Management Core Facility @DKFZ