Sarek

SciLifeLab

KAROLINSKA INSTITUTET ANNO 1810
Karolinska
Institutet

DNA club

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

NBIS

Maxime U. Garcia
maxulysse.github.io
@MaxUlysse
@gau

Barntumörbanken

Science for Life Laboratory
Karolinska Institutet Science Par

What is Sarek?

Sarek

🌐 http://sarek.scilifelab.se/

- Analysis germline and somatic workflow

What is Sarek?

Sarek

http://sarek.scilifelab.se/

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing

What is Sarek?

Sarek

🌐 http://sarek.scilifelab.se/

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing
- Developed with NGI and NBIS

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

NBIS

What is Sarek?

Sarek

🌐 http://sarek.scilifelab.se/

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing
- Developed with NGI and NBIS
- Support from The Swedish Childhood Tumor Biobank

NATIONAL GENOMICS INFRASTRUCTURE   Barntumörbanken   NBIS

Using

nextflow

🌐 https://www.nextflow.io/

🌐 https://www.sylabs.io/singularity/

Using

nextflow

🌐 https://www.nextflow.io/

Data-driven workflow language

🌐 https://www.sylabs.io/singularity/

HPC specific container engine

Sarek exists in multiple flavors

Sarek

Sarek exists in multiple flavors

Sarek
Germline

Sarek
Somatic

Sarek exists in multiple flavors

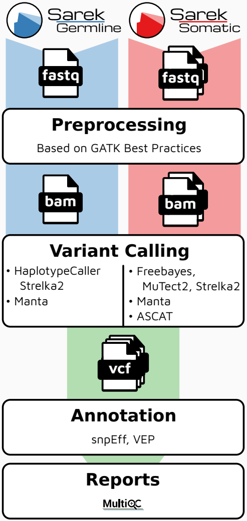
Sarek

Sarek Germline

Sarek Somatic

Sarek Exome

Data and files workflow



Preprocessing

Based on GATK Best Practices

Variant Calling

- HaplotypeCaller Strelka2
- Manta

- Freebayes, MuTect2, Strelka2
- Manta
- ASCAT

Annotation

snpEff, VEP

Reports

References Genomes

AWS iGenomes

🌐 https://ewels.github.io/AWS-iGenomes/

- Human `GRCh37`
- Human `GRCh38`

References Genomes

AWS iGenomes

🌐 https://ewels.github.io/AWS-iGenomes/

- Human `GRCh37`
- Human `GRCh38`
- Dog `CanFam3.1` 🔧
- Mouse `GRCm38` 🔧

Preprocessing

🌐 https://software.broadinstitute.org/gatk/best-practices/

Based on GATK Best Practices (GATK 4.0)

Preprocessing

🌐 https://software.broadinstitute.org/gatk/best-practices/

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`

Preprocessing

🌐 https://software.broadinstitute.org/gatk/best-practices/

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`
  - FASTQs or BAMs 🔧

Preprocessing

🌐 https://software.broadinstitute.org/gatk/best-practices/

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`
  - FASTQs or BAMs 🔧
- Duplicates marked with `picard MarkDuplicates`

Preprocessing

🌐 https://software.broadinstitute.org/gatk/best-practices/

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa mem`
  - FASTQs or BAMs 🔧
- Duplicates marked with `picard MarkDuplicates`
- Recalibrate with `GATK BaseRecalibrator`

Variant Calling

- SNVs and small indels:

Variant Calling

- SNVs and small indels:
  - Freebayes 🔴
  - HaplotypeCaller 🔵
  - MuTect2 🔴
  - Strelka2 🔵/🔴

Variant Calling

- SNVs and small indels:
  - Freebayes 🔴
  - HaplotypeCaller 🔵
  - MuTect2 🔴
  - Strelka2 🔵/🔴
- Structural variants:

Variant Calling

- SNVs and small indels:
  - Freebayes 🔴
  - HaplotypeCaller 🔵
  - MuTect2 🔴
  - Strelka2 🔵/🔴
- Structural variants:
  - Manta 🔵/🔴
- Sample heterogeneity, ploidy and CNVs:

Variant Calling

- SNVs and small indels:
  - Freebayes 🔴
  - HaplotypeCaller 🔵
  - MuTect2 🔴
  - Strelka2 🔵/🔴
- Structural variants:
  - Manta 🔵/🔴
- Sample heterogeneity, ploidy and CNVs:
  - ASCAT 🔴
  - Control-FREEC 🔧 🔴

Annotation

- VEP and SnpEff
- ClinVar, COSMIC, dbSNP, GENCODE, gnomAD, polyphen, sift, etc.

Annotation

- VEP and SnpEff
- ClinVar, COSMIC, dbSNP, GENCODE, gnomAD, polyphen, sift, etc.
- Possibility to use cache directories

Annotation

- VEP and SnpEff
- ClinVar, COSMIC, dbSNP, GENCODE, gnomAD, polyphen, sift, etc.
- Possibility to use cache directories 🔧
- Prioritization 🔧
  - Rank scores are computed for all variants, and can be explored

Singularity containers

🌐 https://www.sylabs.io/singularity/

Singularity containers

🌐 https://www.sylabs.io/singularity/

- Available on `rackham` and/or `bianca`
- `/sw/data/uppnex/ToolBox/sarek`

Singularity containers

🌐 https://www.sylabs.io/singularity/

- Available on `rackham` and/or `bianca`
- `/sw/data/uppnex/ToolBox/sarek`
  - Updated by myself at each new Sarek release

Singularity containers

🌐 https://www.sylabs.io/singularity/

- Available on `rackham` and/or `bianca`
- `/sw/data/uppnex/ToolBox/sarek`
  - Updated by myself at each new Sarek release
- Next step `Sarek` module

Conda 🔧

BIOCONDA®

🌐 https://bioconda.github.io/

Conda

BIOCONDA®

https://bioconda.github.io/

- Execute Sarek within a conda environment

AWS

🌐 https://aws.amazon.com/

AWS

https://aws.amazon.com/

- Improving AWS usage

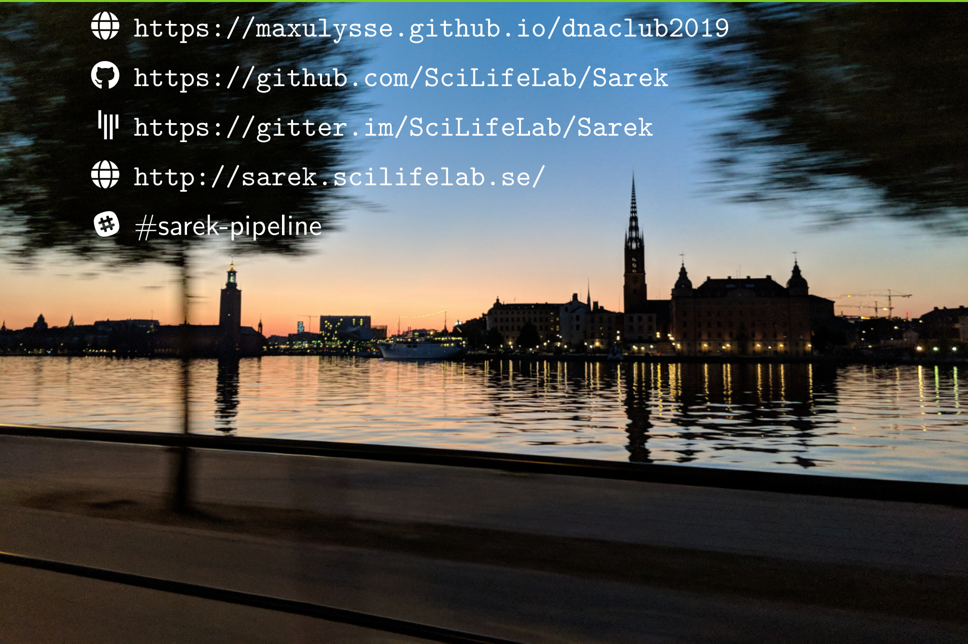
Acknowledgments

BARNCANCER FONDEN · Karolinska Institutet · Barntumörbanken · SciLifeLab

Barntumörbanken	Elisa Basmaci
	Szilveszter Juhos
	Gustaf Ljungman
	Monica Nistèr
	Gabriela Prochazka
	Johanna Sandgren
	Teresita Díaz De Ståhl
	Katarzyna Zielinska-Chomej
Grupp Nistèr	Saad Alqahtani
	Min Guo
	Daniel Hägerstrand
	Anna Hedrén
	Martin Proks
	Rong Yu
	Jian Zhao
NGI	Johannes Alneberg
	Anandashankar Anil
	Franziska Bonath
	Orlando Contreras-López
	Phil Ewels
	Sofia Haglund
	Max Käller
	Anna Konrad
	Pär Lundin
	Remi-Andre Olsen
	Senthilkumar Panneerselvam
	Fanny Taborsak
	Chuan Wang
Clinical Genetics	Jesper Eisfeldt
NBIS	Sebastian DiLorenzo
	Malin Larsson
	Marcel Martin
	Markus Mayrhofer
	Björn Nystedt
	Markus Ringnér
	Pall I Olason
	Jonas Söderberg
Clinical Genomics	Kenny Billiau
	Hassan Foroughi Asl
	Valtteri Wirta
Nextflow folks	Paolo Di Tommaso
	Sven Fillinger
	Alexander Peltzer

NATIONAL GENOMICS INFRASTRUCTURE · NBIS · nextflow · nf-core · UPPMAX

Any questions?



- https://maxulysse.github.io/dnaclub2019
- https://github.com/SciLifeLab/Sarek
- https://gitter.im/SciLifeLab/Sarek
- http://sarek.scilifelab.se/
- #sarek-pipeline