Sarek

Lab Meeting

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

2019-02-12

Karolinska Institutet

Barntumörbanken

SciLifeLab

GANATIONALCTAC ATCAGENOMICSGT INFRASTRUCTURE

NBIS

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

NATIONAL GENOMICS INFRASTRUCTURE

NBIS

What is 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

Powered by

https://www.nextflow.io/

https://www.sylabs.io/singularity/

Powered by

https://www.nextflow.io/

Data-driven workflow language

https://www.sylabs.io/singularity/

HPC specific container engine

Conda

BIOCONDA®

https://bioconda.github.io/

- Virtual environment management system

Sarek exists in multiple flavors

Sarek exists in multiple flavors

Sarek

Sarek
Exome

Sarek exists in multiple flavors

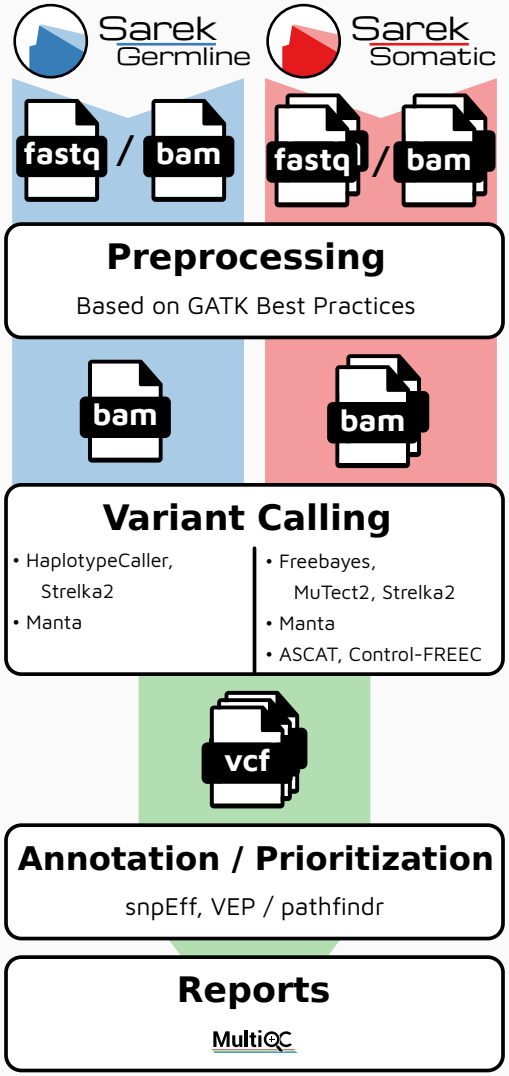
Sarek

Sarek Exome

Sarek Germline

Sarek Somatic

Data and files workflow



AWS iGenomes

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

- Human `GRCh37` from the GATK Resource Bundle
- Human `GRCh38` from the GATK Resource Bundle

AWS iGenomes

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

- Human `GRCh37` from the GATK Resource Bundle
- Human `GRCh38` from the GATK Resource Bundle
- 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

Prioritization

What we need

- Adapt settings where necessary, and ensure they give good results

Markus Mayrhofer

Prioritization

What we need

- Adapt settings where necessary, and ensure they give good results
- Coherent overview to allow critical assessment of sequence and variant quality

Markus Mayrhofer

Prioritization

What we need

- Adapt settings where necessary, and ensure they give good results
- Coherent overview to allow critical assessment of sequence and variant quality
- Tables of variants with probable relevance for the disease

Markus Mayrhofer

Prioritization with pathfindr

Our solution

- Parse all results into R environment

Prioritization with pathfindr

Our solution

- Parse all results into R environment
- Rank variants based on evidence for being a driver mutation

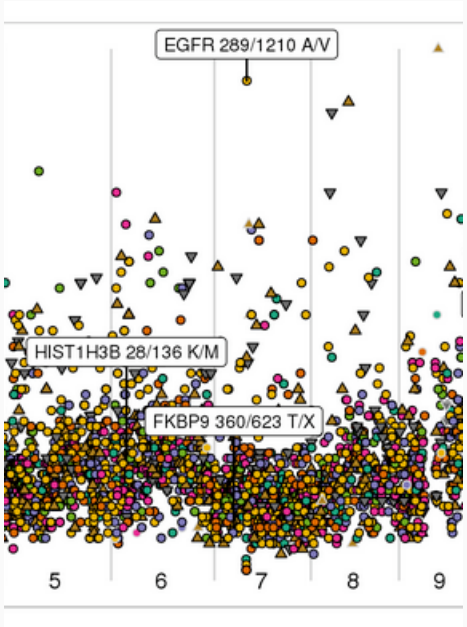
Prioritization with pathfindr

Our solution

- Parse all results into R environment
- Rank variants based on evidence for being a driver mutation
- Visualize in portable html report for easy browsing

Tables and visualization

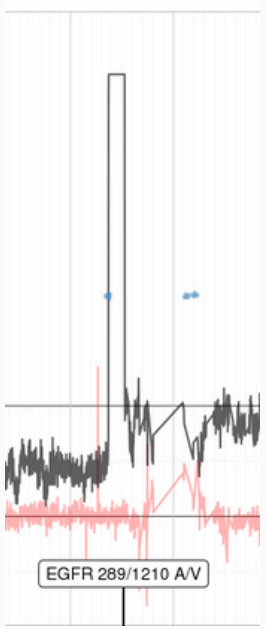
Gene	Mutation	Rank score	Rank Terms	Allele_ratio
EGFR	289/1210 A/V	10	T1_gene moderate_impact clinvar polyphen/SIFT hotspot cosmic_>50	0.94
TP53		9	T1_gene high_impact high+TSG clinvar cosmic_>50	0.86
HIST1H3B	28/136 K/M	7	T1_gene	0.23



SNVs, Indels

Tables and visualization

Gene	Mutation	Rank score	Rank Terms
EGFR	gain	5	T1_gene focal high_amp
ETNK1	gain	5	T1_gene focal high_amp
KRAS	gain	5	T1_gene focal high_amp



Copy number

Sarek at work

- 50 tumor/normal pairs with GRCh37 reference

Sarek at work

- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference

Sarek at work

- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
  - 1 000 samples in germline settings

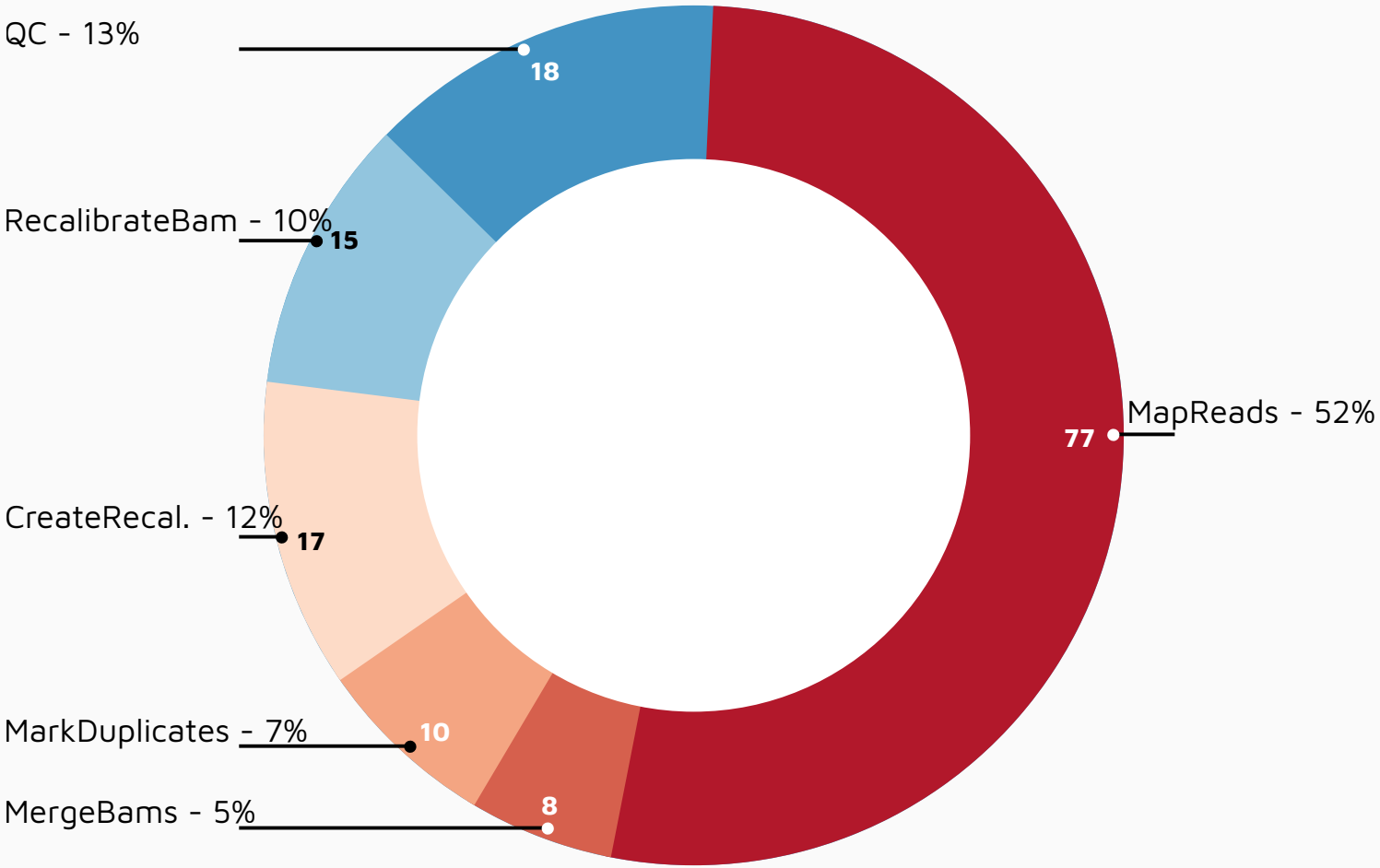
Sarek at work

- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
  - 1 000 samples in germline settings
- Clinical samples with Genomic Medicine Sweden initiative

Sarek at work

- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
  - 1 000 samples in germline settings
- Clinical samples with Genomic Medicine Sweden initiative
- Used at NGI
  - 200 samples
  - testing it in production
  - plans for validation

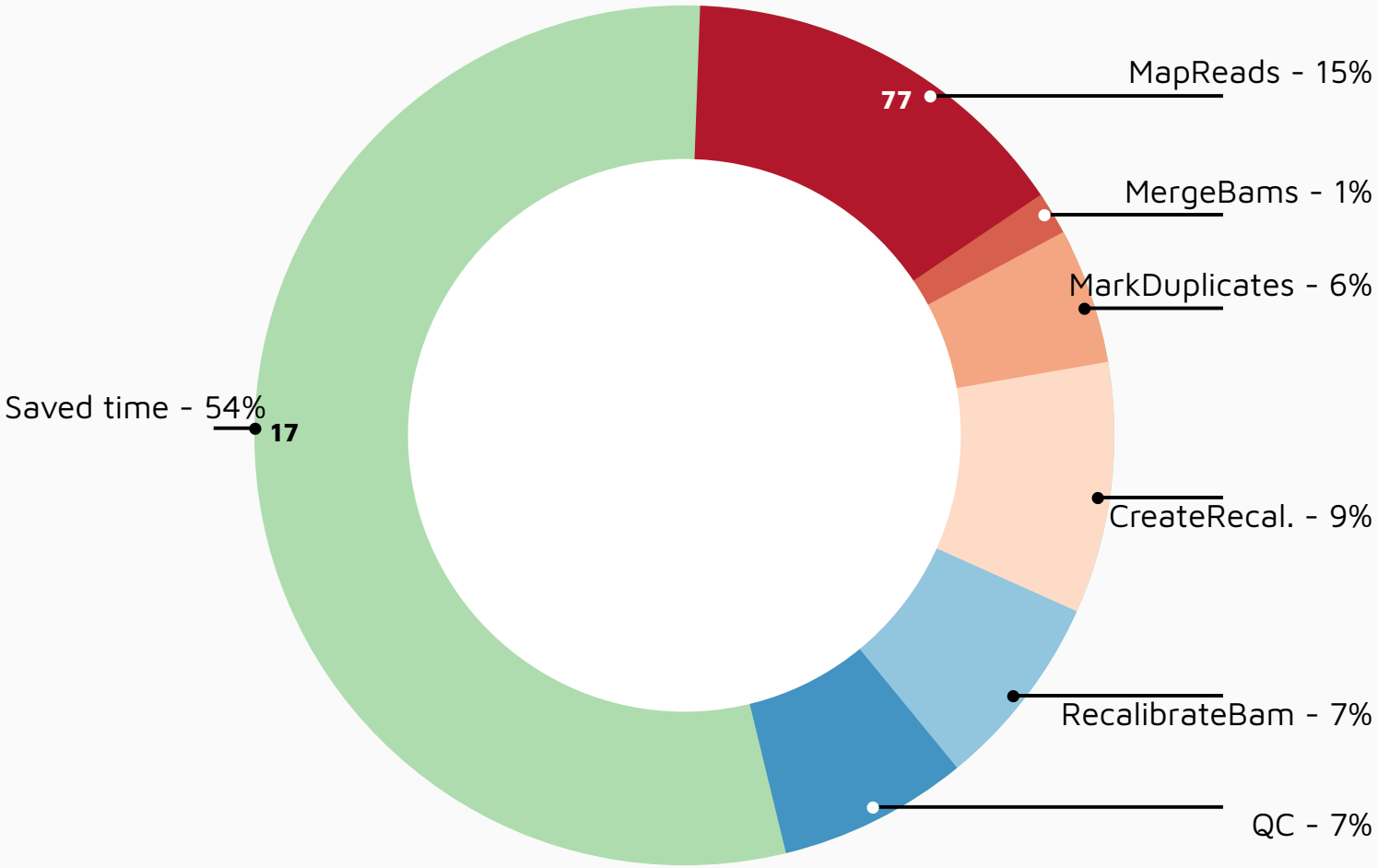
Preprocessing time



On a research settings

On UPPMAX secure cluster Bianca

Preprocessing time



On a clinical settings

On our own secure server munin

AWS

https://aws.amazon.com/

Johannes Alneberg

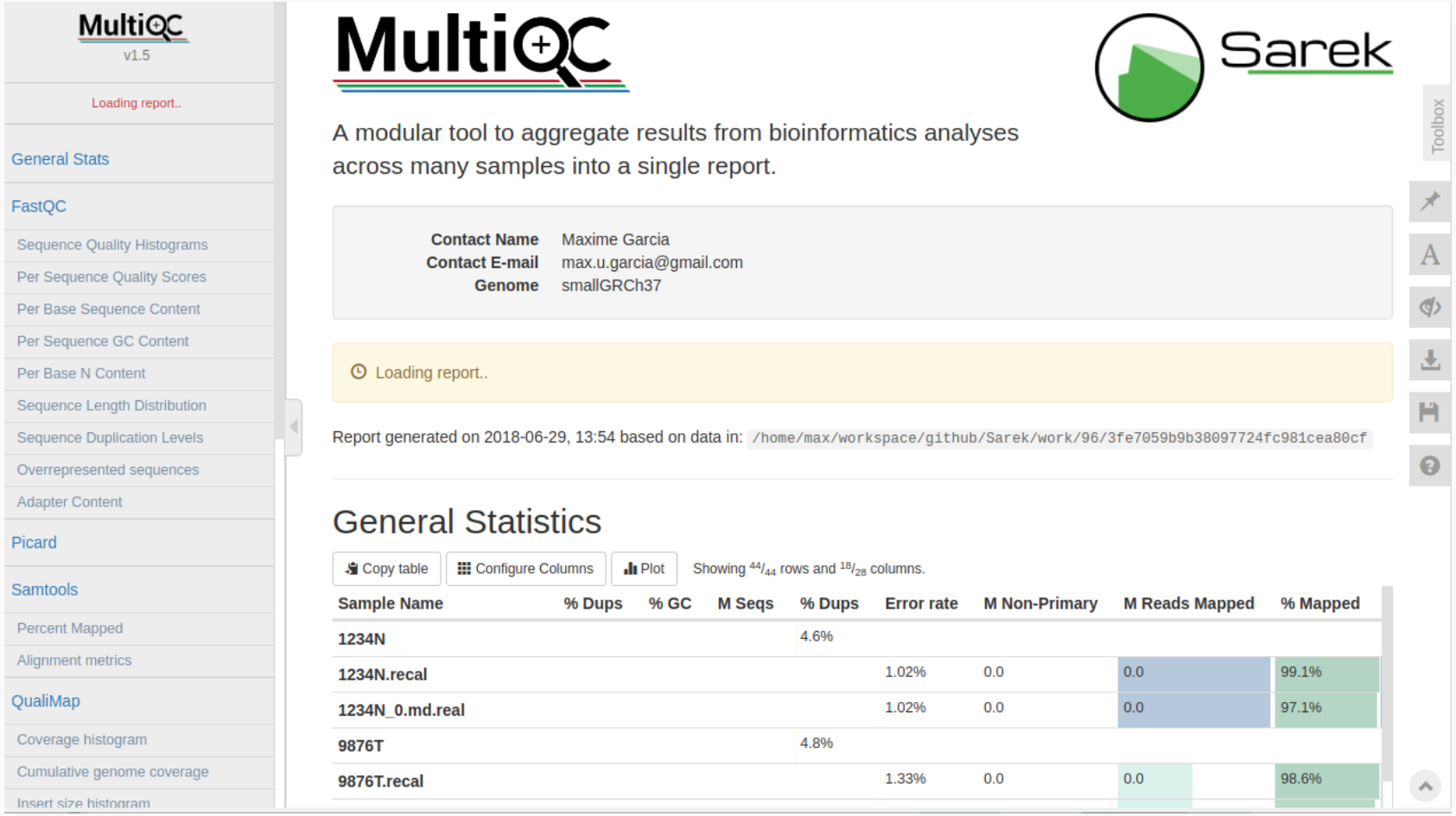
AWS

https://aws.amazon.com/

- Improved AWS usage

Johannes Alneberg

QC reports



http://multiqc.info/

munin

The CMM server room

munin

munin

Acknowledgments

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

Any questions?

https://github.com/SciLifeLab/Sarek/

https://gitter.im/SciLifeLab/Sarek/

http://sarek.scilifelab.se/