





Sarek

A portable workflow for WGS analysis
of germline and somatic mutations

SciLifeLab

NATIONAL CTAC
ATCAGENOMICS
INFRASTRUCTURE

 Maxime U. Garcia
 [maxulysse.github.io](https://github.com/maxulysse)
 @MaxUlysse
 @gau

KAROLINSKA INSTITUTET
Karolinska
Institutet



DNA
club

Science for Life Laboratory

Karolinska Institutet Science Park



Sarek, the National Park in Northern Sweden

The most dramatic and grandiose of all

- Long, deep, narrow valleys and wild, turbulent water.
- **A tortuous delta landscape.**
- Completely lacking in comfortable accommodations.
- Sarek is one of Sweden's **most inaccessible national parks**
- **There are no roads leading up to the national park.**


Sarek National Park website

Where we're going we don't need roads




What is Sarek?



 <http://sarek.scilifelab.se/>

What is Sarek?



 <http://sarek.scilifelab.se/>

- Nextflow pipeline

What is Sarek?




 <http://sarek.scilifelab.se/>

- Nextflow pipeline
- Developed at NGI



What is Sarek?




 <http://sarek.scilifelab.se/>

- Nextflow pipeline
- Developed at NGI
- In collaboration with NBIS



What is Sarek?



 <http://sarek.scilifelab.se/>

- Nextflow pipeline
- Developed at NGI
- In collaboration with NBIS
- Support from The Swedish Childhood Tumor Biobank





 <https://www.nextflow.io/>

- Data-driven workflow language



 <https://www.nextflow.io/>

- Data-driven workflow language
- Portable (executable on multiple platforms)



 <https://www.nextflow.io/>

- Data-driven workflow language
- Portable (executable on multiple platforms)
- Shareable and reproducible (with containers)



 <https://singularity.lbl.gov/>

- Docker-like container engine
- Specific for HPC environment



🌐 <https://singularity.lbl.gov/>

- Docker-like container engine
- Specific for HPC environment
- Without the root user security problem



 <https://singularity.lbl.gov/>

- Docker-like container engine
- Specific for HPC environment
- Without the root user security problem
- Supported by Nextflow



 <https://singularity.lbl.gov/>

- Docker-like container engine
- Specific for HPC environment
- Without the root user security problem
- Supported by Nextflow
- Can pull containers from Docker-hub

Sarek exists in multiple flavors



Sarek exists in multiple flavors




Sarek exists in multiple flavors



Sarek exists in multiple flavors






 <https://software.broadinstitute.org/gatk/best-practices/>

Based on GATK Best Practices (GATK 4.0)



 <https://software.broadinstitute.org/gatk/best-practices/>

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa`



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

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa`
- Duplicates marked with `picard MarkDuplicates`



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

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa`
- Duplicates marked with `picard MarkDuplicates`
- Recalibrate with `GATK BaseRecalibrator`

- SNVs and small indels:

- SNVs and small indels:
 - HaplotypeCaller
 - Strelka2

- SNVs and small indels:
 - HaplotypeCaller
 - Strelka2
- Structural variants:

- SNVs and small indels:
 - HaplotypeCaller
 - Strelka2
- Structural variants:
 - Manta


- SNVs and small indels:

- SNVs and small indels:
 - MuTect2
 - Freebayes
 - Strelka2

- SNVs and small indels:
 - MuTect2
 - Freebayes
 - Strelka2
- Structural variants:

- SNVs and small indels:
 - MuTect2
 - Freebayes
 - Strelka2
- Structural variants:
 - Manta

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

- SNVs and small indels:
 - MuTect2
 - Freebayes
 - Strelka2
- Structural variants:
 - Manta
- Sample heterogeneity, ploidy and CNVs:
 - ASCAT
 - Control-FREEC ( adding)

- VEP and SnpEff

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

- First step towards clinical use

- First step towards clinical use
- Rank scores are computed for all variants
 - COSMIC, ClinVar, SweFreq and MSK-IMPACT (cancerhotspots.org)

- First step towards clinical use
- Rank scores are computed for all variants
 - COSMIC, ClinVar, SweFreq and MSK-IMPACT (cancerhotspots.org)
- Findings are ranked in three tiers

- First step towards clinical use
- Rank scores are computed for all variants
 - COSMIC, ClinVar, SweFreq and MSK-IMPACT (cancerhotspots.org)
- Findings are ranked in three tiers
 - 1st tier: well known, high-impact variants
 - 2nd tier: variants in known cancer-related genes
 - 3rd tier: the remaining variants

MultiQC
v1.5

Loading report...

General Stats

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

Picard

Samtools

Percent Mapped

Alignment metrics

QualiMap

Coverage histogram

Cumulative genome coverage

Insert size histogram

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Contact NameMaxime Garcia

Contact E-mailmax.u.garcia@gmail.com

GenomesmallGRCh37

Loading report...

Report generated on 2018-06-29, 13:54 based on data in: /home/max/workspace/github/Sarek/work/96/3fe7059b9b38097724fc981cea80cf

General Statistics

Copy table

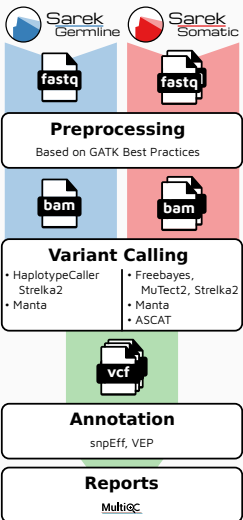
Configure Columns

Plot

Showing 44/44 rows and 10/28 columns.

Sample Name	% Dups	% GC	M Seqs	% Dups	Error rate	M Non-Primary	M Reads Mapped	% Mapped
1234N				4.6%				
1234N.rec.al					1.02%	0.0	0.0	99.1%
1234N_0.md.real					1.02%	0.0	0.0	97.1%
9876T				4.8%				
9876T.rec.al					1.33%	0.0	0.0	98.6%


 <http://multiqc.info/>



- GRCh37 and GRCh38

- GRCh37 and GRCh38
- Custom genome

Reference genomes

- GRCh37 and GRCh38
- Custom genome
-  Other organisms





- 50 tumor/normal pairs with GRCh37 reference

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

- 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

- 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
- 4 clinical samples
 - more coming with Genomic Medicine Sweden initiative

Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants

Maxime Garcia, Szilveszter Juhos, Malin Larsson, Pall I Olason, Marcel Martin, Jesper Eisfeldt, Sebastian DiLorenzo, Johanna Sandgren, Teresita Diaz de Ståhl, Valtteri Wirta, Monica Nistèr, Björn Nystedt, Max Käller

 <https://doi.org/10.1101/316976>

Get involved!

- Our code is hosted on Github
 - 🐙 <https://github.com/SciLifeLab/Sarek>
 - 🐙 <https://github.com/nf-core>

Get involved!

- Our code is hosted on Github
 - 🐙 <https://github.com/SciLifeLab/Sarek>
 - 🐙 <https://github.com/nf-core>
- We have gitter channels
 - 👥 <https://gitter.im/SciLifeLab/Sarek>
 - 👥 <https://gitter.im/nf-core/Lobby>

Acknowledgments



Barntumörbanken

Elisa Basmaci
Szilveszter Juhos
Gustaf Ljungman
Monica Nistér
Gabriela Prochazka
Johanna Sandgren
Teresita Díaz De Ståhl
Katarzyna Zielinska-Chomej

NGI

Johannes Alneberg
Anandashankar Anil
Franziska Bonath
Orlando Contreras-López
Phil Ewels
Sofia Haglund
Max Käller
Anna Konrad
Pär Lundin

NBIS

Sebastian DiLorenzo
Malin Larsson
Marcel Martin
Markus Mayrhofer
Björn Nystedt
Markus Ringné
Pall I Olason
Jonas Söderberg

Grupp Nistér

Saad Alqahtani
Min Guo
Daniel Hägerstrand
Anna Hedrén
Martin Proks
Rong Yu
Jian Zhao

NGI

Remi-Andre Olsen
Senthilkumar Panneerselvam
Fanny Taborsak
Chuan Wang

Clinical Genomics

Kenny Billiau
Hassan Foroughi Asl
Valtteri Wirta

Nextflow folks

Paolo Di Tommaso
Sven Fillingér
Alexander Peltzer

Clinical Genetics

Jesper Einfeldt



Any questions?

🌐 <http://sarek.scilifelab.se/>

🌐 <https://github.com/SciLifeLab/Sarek>

🌐 <https://maxulysse.github.io/dnaclub2018>

