Sarek

A portable workflow for WGS analysis of germline and somatic mutations

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

JOBIM 2018 MARSEILLE

Science for Life Laboratory

SciLifeLab

https://scilifelab.se/

SciLifeLab is a national centre for molecular biosciences with focus on health and environmental research

Karolinska Institutet

KTH VETENSKAP OCH KONST

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab

https://scilifelab.se/

Infrastructure Services

Genomics	Proteomics	Metabolomics
Single Cell Biology	Bioimaging and Molecular Structure	Chemical Biology and Genome Engineering
Drug Discovery	Diagnostics	Bioinformatics

National Genomics Infrastructure

https://ngisweden.scilifelab.se/

- National resource

National Genomics Infrastructure

https://ngisweden.scilifelab.se/

- National resource
- State-of-the-art infrastructure
  - massively parallel DNA sequencing and SNP genotyping

National Genomics Infrastructure

https://ngisweden.scilifelab.se/

- National resource
- State-of-the-art infrastructure
  - massively parallel DNA sequencing and SNP genotyping
- Guidelines and support
  - sample collection, study design, protocol selection
  - bioinformatics analysis

National Bioinformatics Infrastructure Sweden

https://www.nbis.se/

- Swedish ELIXIR node

National Bioinformatics Infrastructure Sweden

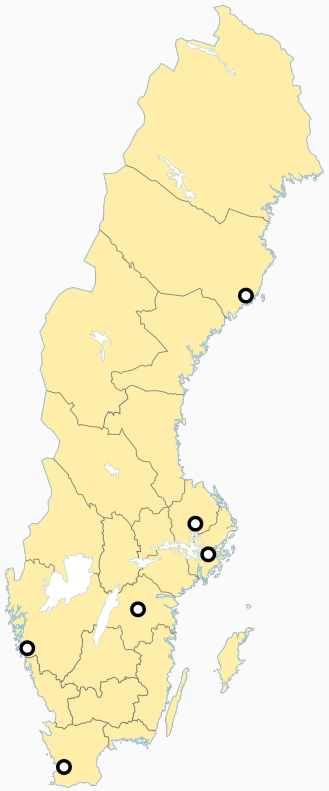
https://www.nbis.se/

- Swedish ELIXIR node
- Bioinformatics support for Swedish researchers

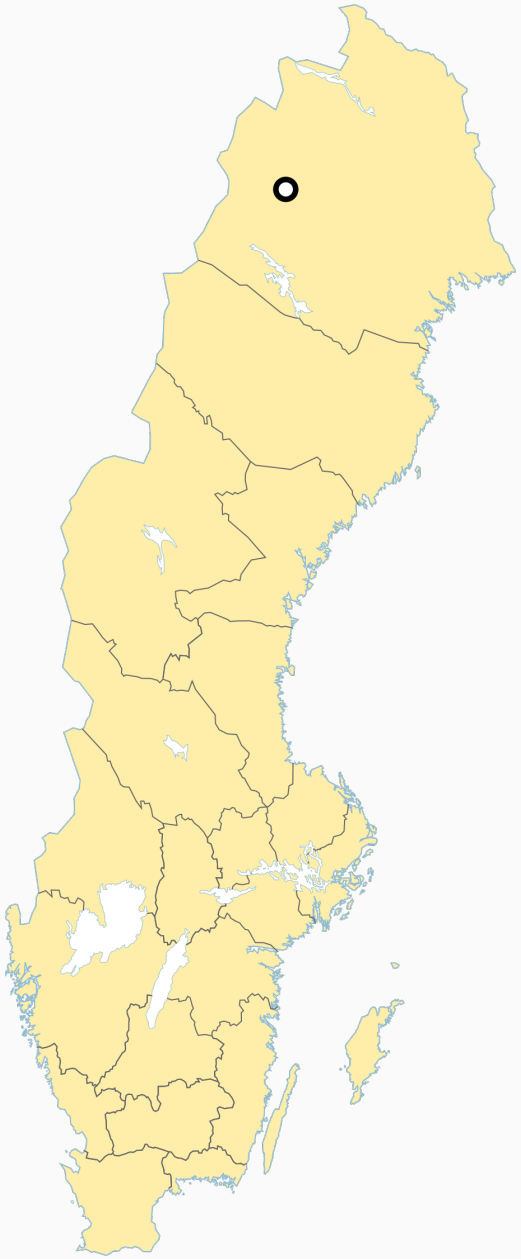
National Bioinformatics Infrastructure Sweden

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

https://www.nbis.se/



Sarek



Sarek

Sarek, the National Park in Northen 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://opensource.scilifelab.se/projects/sarek/

What is Sarek?

http://opensource.scilifelab.se/projects/sarek/

- Nextflow pipeline

What is Sarek?

http://opensource.scilifelab.se/projects/sarek/

- Nextflow pipeline
- Developed at NGI

NATIONAL GENOMICS INFRASTRUCTURE

What is Sarek?

http://opensource.scilifelab.se/projects/sarek/

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

NATIONAL GENOMICS INFRASTRUCTURE

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

What is Sarek?

http://opensource.scilifelab.se/projects/sarek/

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

Nextflow

https://www.nextflow.io/

- Data-driven workflow language

Nextflow

https://www.nextflow.io/

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

Nextflow

https://www.nextflow.io/

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

Singularity

https://singularity.lbl.gov/

- Docker-like container engine
- Specific for HPC environnment

Singularity

https://singularity.lbl.gov/

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

Singularity

https://singularity.lbl.gov/

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

Singularity

https://singularity.lbl.gov/

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

Sarek exists in two flavors

Sarek exists in two flavors

Sarek exists in two flavors

Preprocessing

🌍 `https://software.broadinstitute.org/gatk/best-practices/`

Based on GATK Best Practices (GATK 3.8)

🔧 Switching to GATK 4.0

Preprocessing

🌍 `https://software.broadinstitute.org/gatk/best-practices/`

Based on GATK Best Practices (GATK 3.8)

🔧 Switching to GATK 4.0

- Reads mapped to reference genome with `bwa`

Preprocessing

🌍 `https://software.broadinstitute.org/gatk/best-practices/`

Based on GATK Best Practices (GATK 3.8)

🔧 Switching to GATK 4.0

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

Preprocessing

🌍 `https://software.broadinstitute.org/gatk/best-practices/`

Based on GATK Best Practices (GATK 3.8)

🔧 Switching to GATK 4.0

- Reads mapped to reference genome with `bwa`
- Duplicates marked with `picard MarkDuplicates`
- Realign indels with `GATK IndelRealigner`

Preprocessing

gatk ☑ Best Practices

🌍 `https://software.broadinstitute.org/gatk/best-practices/`

Based on GATK Best Practices (GATK 3.8)

🔧 Switching to GATK 4.0

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

Germline Variant Calling

- SNVs and small indels:

Germline Variant Calling

- SNVs and small indels:
  - HaplotypeCaller
  - Strelka2

Germline Variant Calling

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

Germline Variant Calling

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

Somatic Variant Calling

- SNVs and small indels:

Somatic Variant Calling

- SNVs and small indels:
  - MuTect1 (🔧 removing)
  - MuTect2
  - Freebayes
  - Strelka2

Somatic Variant Calling

- SNVs and small indels:
  - MuTect1 (🔧 removing)
  - MuTect2
  - Freebayes
  - Strelka2
- Structural variants:

Somatic Variant Calling

- SNVs and small indels:
  - MuTect1 (🔧 removing)
  - MuTect2
  - Freebayes
  - Strelka2
- Structural variants:
  - Manta

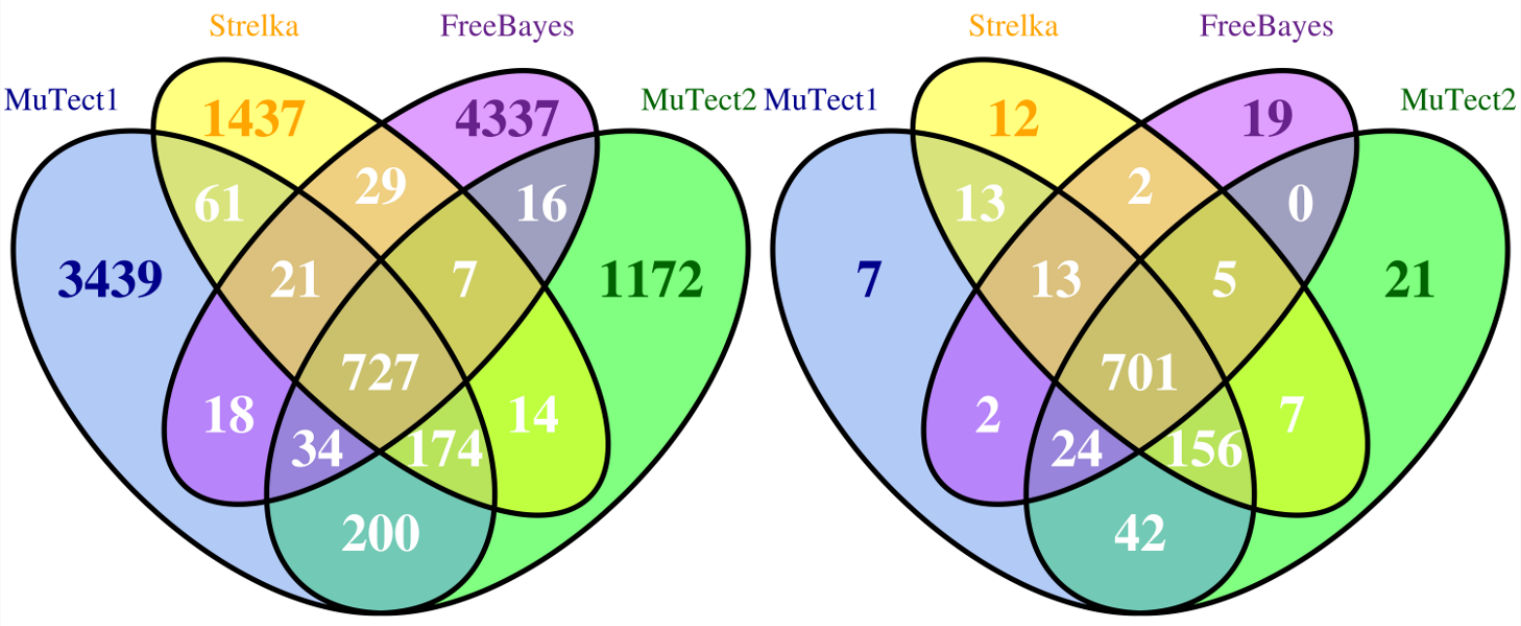
Somatic Variant Calling

- SNVs and small indels:
  - MuTect1 (🔧 removing)
  - MuTect2
  - Freebayes
  - Strelka2
- Structural variants:
  - Manta
- Sample heterogeneity, ploidy and CNVs:

Somatic Variant Calling

- SNVs and small indels:
  - MuTect1 (🔧 removing)
  - MuTect2
  - Freebayes
  - Strelka2
- Structural variants:
  - Manta
- Sample heterogeneity, ploidy and CNVs:
  - ASCAT
  - Control-FREEC (🔧 adding)

SNV Calling overlap



All calls True positives only

Number and overlap of somatic SNV calls from a WGS medulloblastoma dataset

Alioto TS et al. (2015) doi https://doi.org/10.1038/ncomms10001

Annotation

- VEP and SnpEff

Annotation

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

Prioritization

- First step towards clinical use

Prioritization

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

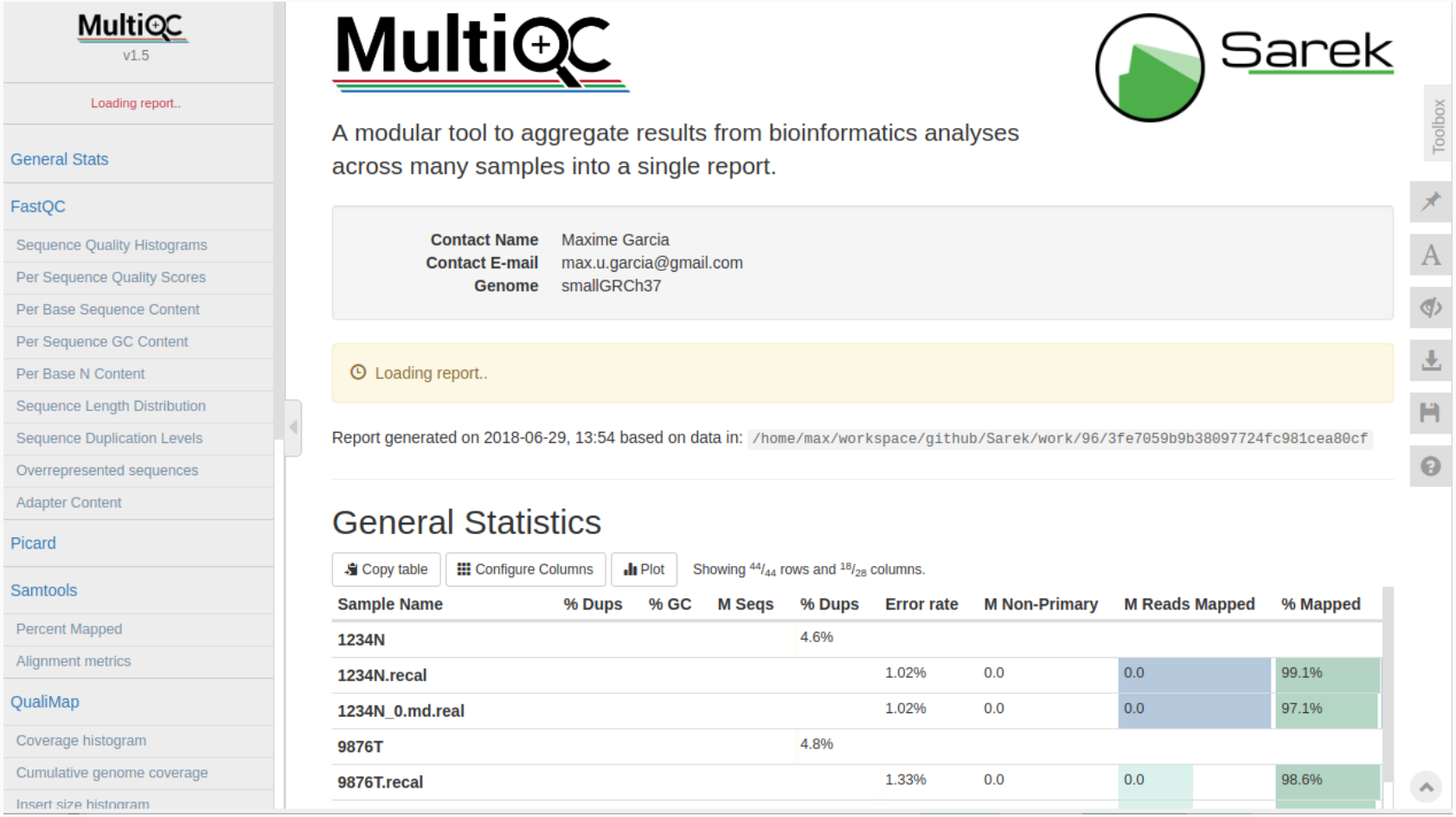
Prioritization

- 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

Prioritization

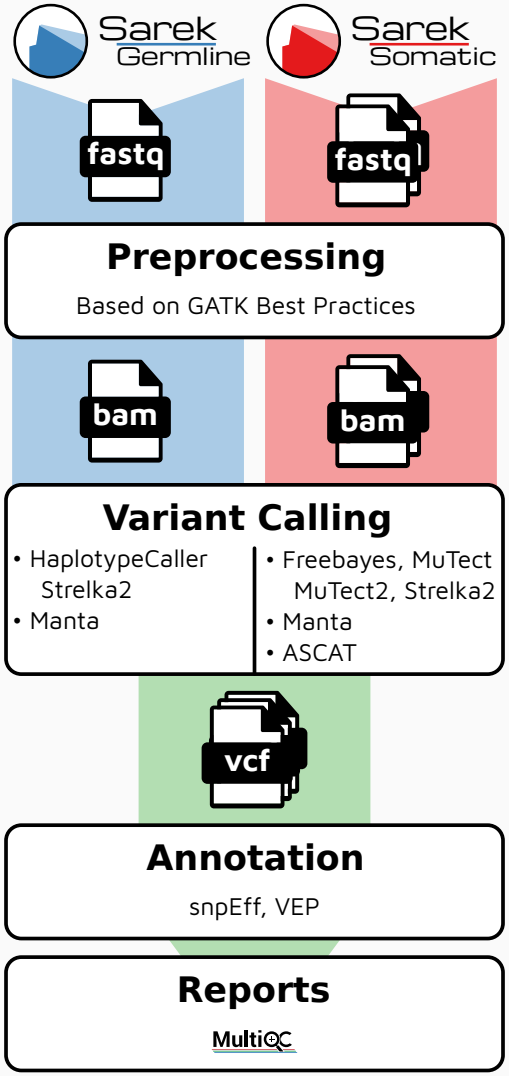
- First step towards clinical use
- Rank scores are computed for all variants
  - COSMIC, ClinVar, SweFreq and MSK-IMPACT (cancerhotspots.org)
- Findings are ranked in three tiers
  - 1^{st} tier: well known, high-impact variants
  - 2^{nd} tier: variants in known cancer-related genes
  - 3^{rd} tier: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

Reports



http://multiqc.info/

Workflow


Sarek Germline
Sarek Somatic
fastq
fastq
Preprocessing
Based on GATK Best Practices
bam
bam
Variant Calling
• HaplotypeCaller
Strelka2
• Manta
• Freebayes, MuTect
MuTect2, Strelka2
• Manta
• ASCAT
vcf
Annotation
snpEff, VEP
Reports
MultiQC

WES and gene panels

- Preprocessing is done with the whole genome

WES and gene panels

- Preprocessing is done with the whole genome
- Variant call only on the target regions

Reference genomes

- GRCh37 and GRCh38

Reference genomes

- GRCh37 and GRCh38
- Custom genome

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

Preprint available at BioRxiv

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

http://nf-co.re/

A community effort to collect a curated set of Nextflow analysis pipelines

- GitHub organisation to collect pipelines in one place
- No institute-specific branding
- Strict set of guideline requirements
- Automated testing for code style and function
- Conda environnment, Docker and Singularity container

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 Nistèr
Gabriela Prochazka
Johanna Sandgren
Teresita Díaz De Ståhl
Katarzyna Zielinska-Chomej

Grupp Nistèr
Saad Alqahtani
Min Guo
Daniel Hägerstrand
Anna Hedrén
Rong Yu
Jian Zhao

NGI
Anandashankar Anil
Orlando Contreras-López
Phil Ewels
Sofia Haglund
Max Käller
Anna Konrad
Pär Lundin
Remi-Andre Olsen
Senthilkumar Panneerselvam
Fanny Taborsak
Chuan Wang

Nextflow folks
Paolo Di Tommaso
Alexander Peltzer
Sven Fillinger

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

Clinical Genetics
Jesper Eisfeldt

Any questions?

- http://opensource.scilifelab.se/projects/sarek/
- https://github.com/SciLifeLab/Sarek
- https://maxulysse.github.io/jobim2018