



Sarek

Lab Meeting

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

2019-02-12



Karolinska
Institutet

SciLifeLab

NATIONAL CTAC
ATCAGENOMICS GT
INFRASTRUCTURE



Barntumörbanken

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



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

nextflow

🌐 <https://www.nextflow.io/>



🌐 <https://www.sylabs.io/singularity/>

Powered by

nextflow

🌐 <https://www.nextflow.io/>

Data-driven workflow language



🌐 <https://www.sylabs.io/singularity/>

HPC specific container engine

BIOCONDA[®]



<https://bioconda.github.io/>

- Virtual environment management system

Sarek exists in multiple flavors



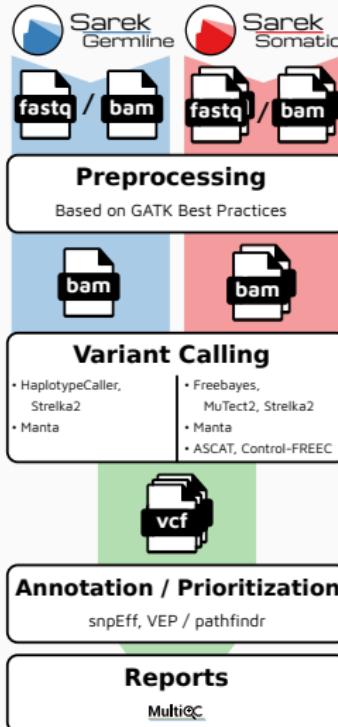
Sarek exists in multiple flavors



Sarek exists in multiple flavors



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

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

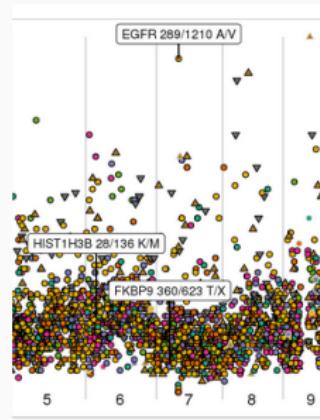
Prioritization with pathfindr

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

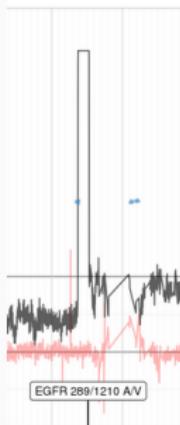
Gene	Mutation	Rank score	Rank Terms	Allele_ratio
EGFR	289/1210 A/V	10	T1_gene moderate_impact clinvar polyphen/SIFT hotspot cosmic_>50	0.94
TP53		9	T1_gene high_impact high+TSG clinvar cosmic_>50	0.86
HIST1H3B	28/136 K/M	7	T1_gene	0.23



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

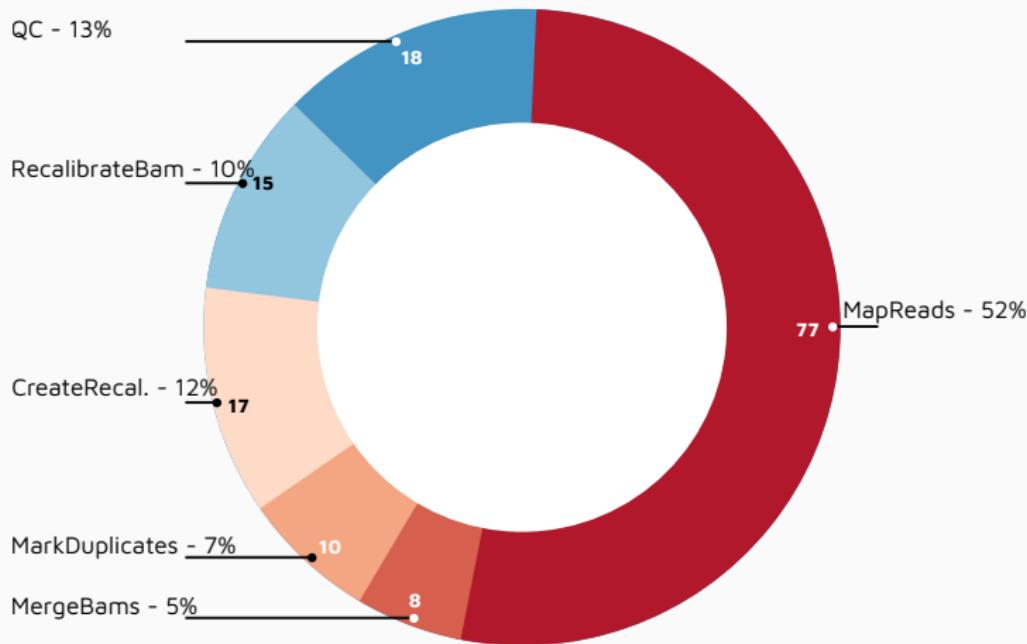
- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
 - 1 000 samples in germline settings

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

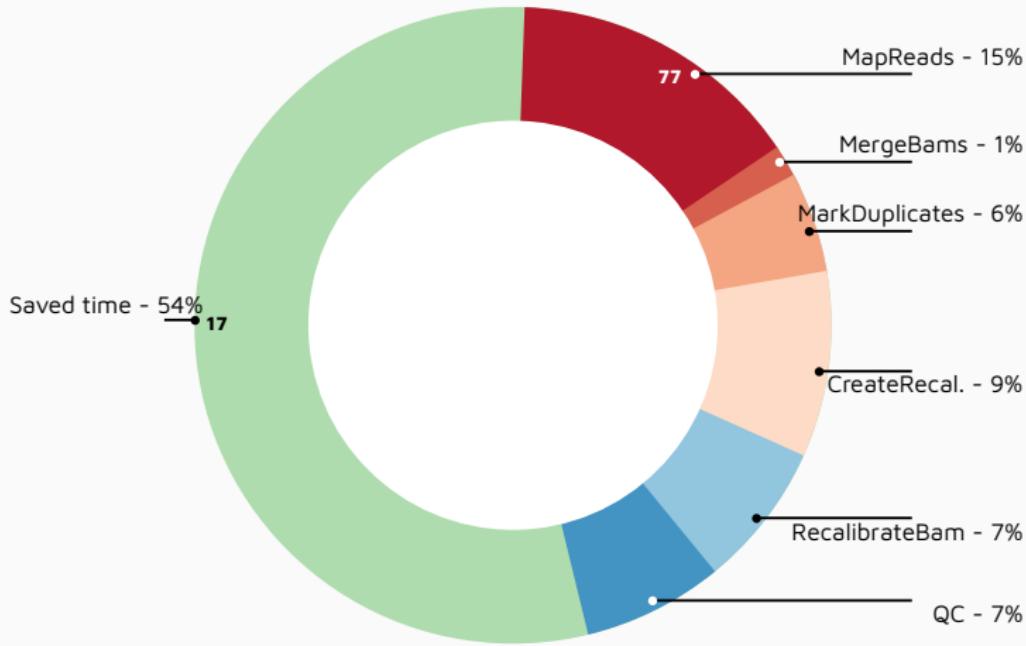
- 50 tumor/normal pairs with GRCh37 reference
- 90 tumor/normal pairs (with some relapse) with GRCh38 reference
- The whole SweGen dataset with GRCh38 reference
 - 1 000 samples in germline settings
- Clinical samples with Genomic Medicine Sweden initiative
- Used at NGI
 - 200 samples
 - testing it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



<https://aws.amazon.com/>

Johannes Alneberg



<https://aws.amazon.com/>

- Improved AWS usage

Johannes Alneberg

QC reports

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 Name: Maxime Garcia
Contact E-mail: max.u.garcia@gmail.com
Genome: smallGRCh37

⌚ Loading report...

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

General Statistics

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

The Sarek logo consists of a green triangle inside a white circle, followed by the word "Sarek" in a stylized font.

Toolbox

A

H

<http://multiqc.info/>

16/19



The CMM server room

munin



munin

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 Ringnér Pall I Olason Jonas Söderberg
Grupp Nistér	Saad Alqahtani Min Guo Daniel Hägerstrand Anna Hedrén Martin Proks Rong Yu Jian Zhao		Remi-Andre Olsen Senthilkumar Panneerselvam Fanny Taborsak Chuan Wang	Clinical Genomics	Kenny Billiau Hassan Foroughi Asl Valterti Wirta
		Clinical Genetics	Jesper Eisfeldt	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/>

