Sarek

A portable workflow for WGS analysis of germline and somatic mutations

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

DNA club
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

ONE DOES NOT SIMPLY WALK INTO SAREK
What is Sarek?

http://sarek.scilifelab.se/
What is Sarek?

- Nextflow pipeline

http://sarek.scilifelab.se/
What is Sarek?

- Nextflow pipeline
- Developed at NGI

http://sarek.scilifelab.se/
What is Sarek?

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

http://sarek.scilifelab.se/
What is Sarek?

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

http://sarek.scilifelab.se/
Nextflow

[Image]

- Data-driven workflow language

https://www.nextflow.io/
Nextflow

- 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)
Singularity

- Docker-like container engine
- Specific for HPC environment

https://singularity.lbl.gov/
Singularity

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

https://singularity.lbl.gov/
Singularity

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

https://singularity.lbl.gov/
Singularity

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

https://singularity.lbl.gov/
Sarek exists in multiple flavors
Sarek exists in multiple flavors

Sarek

Somatic

Germline

Exome
Sarek exists in multiple flavors

- Sarek
- Germline
- Somatic
- Exome
Sarek exists in multiple flavors

- Sarek
- Sarek Germline
- Sarek Somatic
- Sarek Exome
Preprocessing

Based on GATK Best Practices (GATK 4.0)

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

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with **bwa**

[https://software.broadinstitute.org/gatk/best-practices/](https://software.broadinstitute.org/gatk/best-practices/)
Preprocessing

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/
Preprocessing

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with bwa
- Duplicates marked with picard MarkDuplicates
- 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:
  - MuTect2
  - Freebayes
  - Strelka2
- Structural variants:
  - Manta
- Sample heterogeneity, ploidy and CNVs:
  - ASCAT
  - Control-FREEC
Somatic Variant Calling

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

- Structural variants:
  - Manta

- Sample heterogeneity, ploidy, and CNVs:
  - ASCAT
  - Control-FREEC
Somatic Variant Calling

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

- Structural variants:
Somatic Variant Calling

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

- Structural variants:
  - Manta
Somatic Variant Calling

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

- Structural variants:
  - Manta

- Sample heterogeneity, ploidy and CNVs:
Somatic Variant Calling

- SNVs and small indels:
  - MuTect2
  - Freebayes
  - Strelka2
- Structural variants:
  - Manta
- Sample heterogeneity, ploidy and CNVs:
  - ASCAT
  - Control-FREEC (adding)
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\textsuperscript{st} tier: well known, high-impact variants
  - 2\textsuperscript{nd} tier: variants in known cancer-related genes
  - 3\textsuperscript{rd} tier: the remaining variants
Reports

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: smalGRCh37

Report generated on 2018-06-29, 13:54 based on data in: /home/max/workspace/github/Sarek/work/96/3fe7050b9b38897724fc981ce88cf

General Statistics

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>% Dups</th>
<th>% GC</th>
<th>M Seqs</th>
<th>% Dups</th>
<th>Error rate</th>
<th>M Non-Primary</th>
<th>M Reads Mapped</th>
<th>% Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234N</td>
<td></td>
<td>4.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1234N.recal</td>
<td>1.02%</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>99.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1234N_md.real</td>
<td>1.02%</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>97.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9876T</td>
<td>4.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9876T.recal</td>
<td>1.33%</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>98.6%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://multiqc.info/
Workflow

Preprocessing
Based on GATK Best Practices

Variant Calling
• HaplotypeCaller
  • Strelka2
• Manta
• MuTect2, Strelka2
• Manta
• ASCAT

Annotation
snpEff, VEP

Reports
MultiQC
Reference genomes

- GRCh37 and GRCh38
Reference genomes

- GRCh37 and GRCh38
- Custom genome
Reference genomes

- GRCh37 and GRCh38
- Custom genome
- 🛠️ Other organisms
More AWS testing
Production ready
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
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

doi 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 Nistèr
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
Remi-Andre Olsen
Senthilkumar Panneerselvam
Fanny Taborsak
Chuan Wang

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

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

Clinical Genetics
Jesper Eisfeldt

Clinical Genomics
Kenny Billiau
Hassan Foroughi Asl
Valtteri Wirta

Nextflow folks
Paolo Di Tommaso
Sven Fillinger
Alexander Peltzer
Any questions?

http://sarek.scilifelab.se/
https://github.com/SciLifeLab/Sarek
https://maxulysse.github.io/dnaclub2018