







SciLifeLab



An introduction

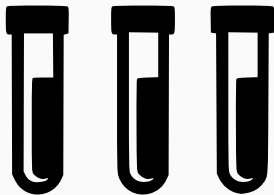


 Maxime U. Garcia
 maxulyse.github.io
 @MaxUlysse
 @gau



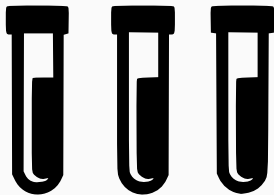
Barntumörbanken





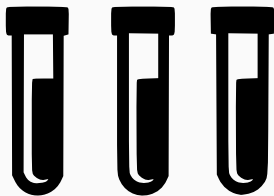
- Normal sample -> blood

Samples



- Normal sample -> blood
- Tumor sample

Samples



- Normal sample -> blood
- Tumor sample
- Eventual relapse or metastasis

WGS - WES - Targeted sequencing



Genome



Exome



Targeted



Illumina's HiSeq X

- Short reads

FASTQ: text-based format for storing both nucleotide sequence and corresponding quality scores.

FASTQ: text-based format for storing both nucleotide sequence and corresponding quality scores.

- At least 1 for each samples (single end)

FASTQ: text-based format for storing both nucleotide sequence and corresponding quality scores.

- At least 1 for each samples (single end)
- At least 1 pair for each samples (pair end)

FASTQ: text-based format for storing both nucleotide sequence and corresponding quality scores.

- At least 1 for each samples (single end)
- At least 1 pair for each samples (pair end)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++) (%%%) .1***-+*'') **55CCF>>>>>CCCCCCC65
```

- Map short reads to reference genome

- Map short reads to reference genome
- Cleanup

Sequence Alignment Map (SAM): text-based format for storing biological sequences aligned to a reference sequence.

SAM/BAM files

Sequence Alignment Map (SAM): text-based format for storing biological sequences aligned to a reference sequence.

Binary Alignment Map (BAM): compressed binary representation of SAM format.

Sequence Alignment Map (SAM): text-based format for storing biological sequences aligned to a reference sequence.

Binary Alignment Map (BAM): compressed binary representation of SAM format.

```
@HD VN:1.6 S0:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

- Germline

- Germline
 - Differences to Reference genome

- Germline
 - Differences to Reference genome
- Somatic

- Germline
 - Differences to Reference genome
- Somatic
 - Differences to Germline genome

VCF files

The Variant Call Format (VCF): text-based format for storing gene sequence variations.

VCF files

The Variant Call Format (VCF): text-based format for storing gene sequence variations.

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##phasing=partial
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA000001 NA000002 NA000003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ
    0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ
    0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP
    1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ
    0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP
    0/1:35:4 0/2:17:2 1/1:40:3
```

What do we want?

Do analysis!

Do analysis!

- Easy to use

Do analysis!

- Easy to use
- Easy to install



Do analysis!

- Easy to use
- Easy to install
- Reproducible

Do analysis!

- Easy to use
- Easy to install
- Reproducible
- Portable

Do analysis!

- Easy to use
- Easy to install
- Reproducible
- Portable
- Open Source

Do analysis!

- Easy to use
- Easy to install
- Reproducible
- Portable
- Open Source
- QC

What do we need?

- Tools

What do we need?

- Tools
 - Installation
 - Version

What do we need?

- Tools
 - Installation
 - Version
- Reference files

What do we need?

- Tools
 - Installation
 - Version
- Reference files
 - Download
 - Version

What do we need?

- Tools
 - Installation
 - Version
- Reference files
 - Download
 - Version
- Annotation files

What do we need?

- Tools
 - Installation
 - Version
- Reference files
 - Download
 - Version
- Annotation files
 - Download
 - Version

What do we need?

- Tools
 - Installation
 - Version
- Reference files
 - Download
 - Version
- Annotation files
 - Download
 - Version
- Works with cluster executor

What is Sarek?




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

- Analysis germline and somatic workflow

What is Sarek?



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

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing

What is 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



The logo for Nextflow, featuring the word "next" in a green, lowercase, sans-serif font, followed by "flow" in a black, lowercase, sans-serif font. The "x" in "next" is stylized with a green swoosh that extends over the "i" and "f".

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

- Docker-like container engine
 - Specific for HPC environment



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


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



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

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



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

- 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



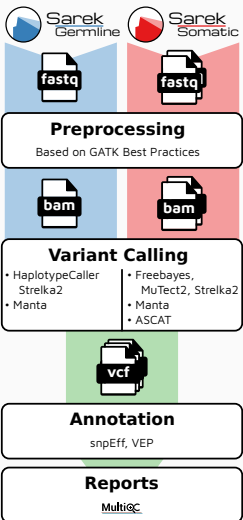
Sarek exists in multiple flavors




Sarek exists in multiple flavors



Data and files workflow






 <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

Somatic Variant Calling


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

- 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)

- 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
 - 1st tier: well known, high-impact variants
 - 2nd tier: variants in known cancer-related genes
 - 3rd tier: the remaining variants

Acknowledgments



Barntumörbanken	Elisa Basmaci Szilveszter Juhas 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 Remi-Andre Olsen Senthilkumar Panneerselvam Fanny Taborsak Chuan Wang	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	Clinical Genetics	Jesper Eisfeldt	Clinical Genomics	Kenny Billiau Hassan Foroughi Asl Valtteri Wirta
				Nextflow folks	Paolo Di Tommaso Sven Fillingner Alexander Peltzer



Any questions?

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

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

