Sarek

SciLifeLab

Karolinska Institutet

GANATIONALCTAC ATCAGENOMICSGT INFRASTRUCTURE

NBIS

An introduction

Maxime U. Garcia

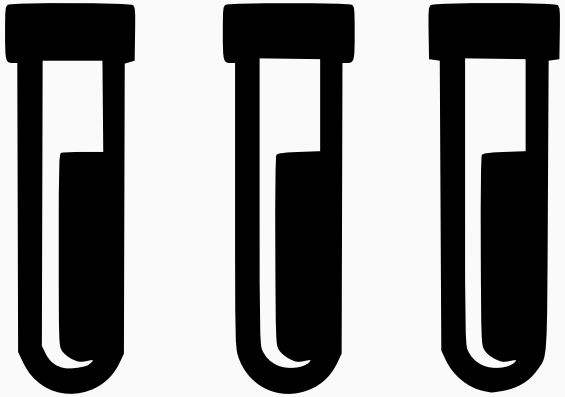
maxulysse.github.io

@MaxUlysse

@gau

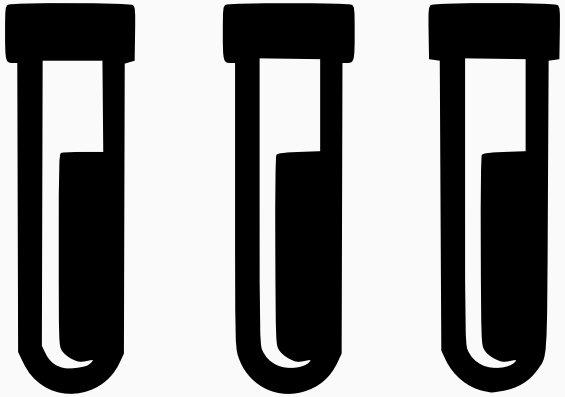
Barntumörbanken

Samples



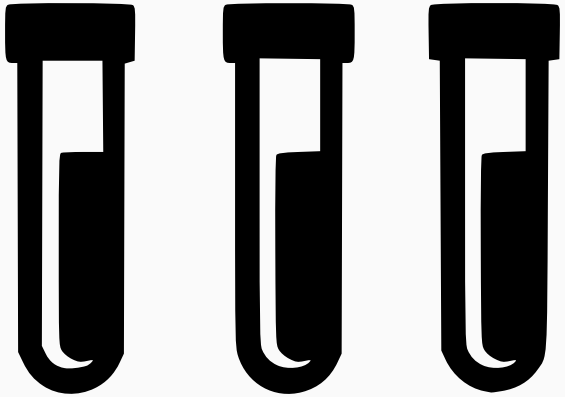
- Normal sample -> blood

Samples



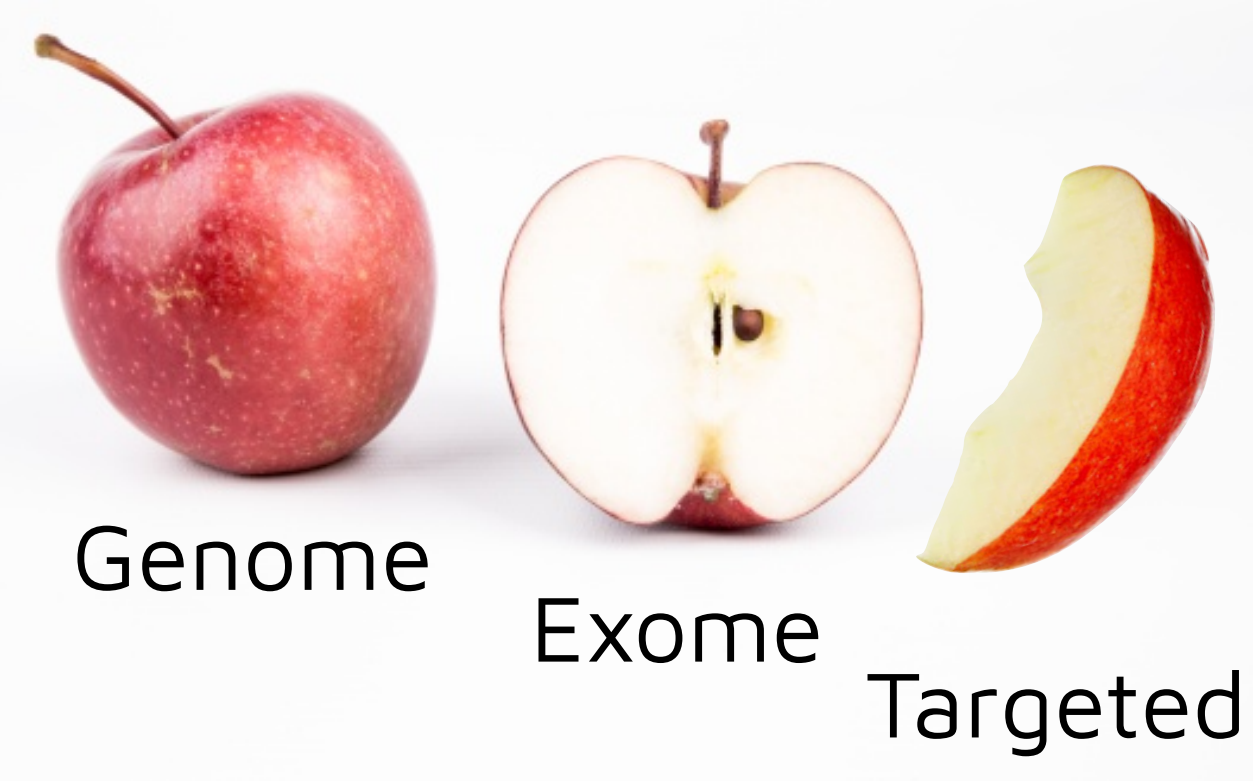
- Normal sample -> blood
- Tumor sample

Samples



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

WGS - WES - Targeted sequencing



Sequencing

Illumina's HiSeq X

- Short reads

FASTQ Files

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

FASTQ Files

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

- At least 1 for each samples (single end)

FASTQ Files

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 Files

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
```

Preprocessing

- Map short reads to reference genome

Preprocessing

- Map short reads to reference genome
- Cleanup

SAM/BAM files

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.

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.

```
@HD VN:1.6 SO: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
```

Variant Calling

- Germline

Variant Calling

- Germline
  - Differences to Reference genome

Variant Calling

- Germline
  - Differences to Reference genome
- Somatic

Variant Calling

- 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 NA00001 NA00002 NA00003
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!

What do we want?

Do analysis!

- Easy to use

What do we want?

Do analysis!

- Easy to use
- Easy to install

What do we want?

Do analysis!

- Easy to use
- Easy to install
- Reproducible

What do we want?

Do analysis!

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

What do we want?

Do analysis!

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

What do we want?

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
  - Dowload
  - Version

What do we need?

- Tools
  - Installation
  - Version
- Reference files
  - Dowload
  - Version
- Annotation files

What do we need?

- Tools
  - Installation
  - Version
- Reference files
  - Dowload
  - Version
- Annotation files
  - Dowload
  - Version

What do we need?

- Tools
  - Installation
  - Version
- Reference files
  - Dowload
  - Version
- Annotation files
  - Dowload
  - 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

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://www.sylabs.io/singularity/

- Docker-like container engine
  - Specific for HPC environnment

Singularity

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

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

Singularity

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

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

Singularity

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

- 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 multiple flavors

Sarek

Sarek exists in multiple flavors

Sarek

Sarek Germline

Sarek exists in multiple flavors

Sarek

Sarek Germline

Sarek Somatic

Sarek exists in multiple flavors

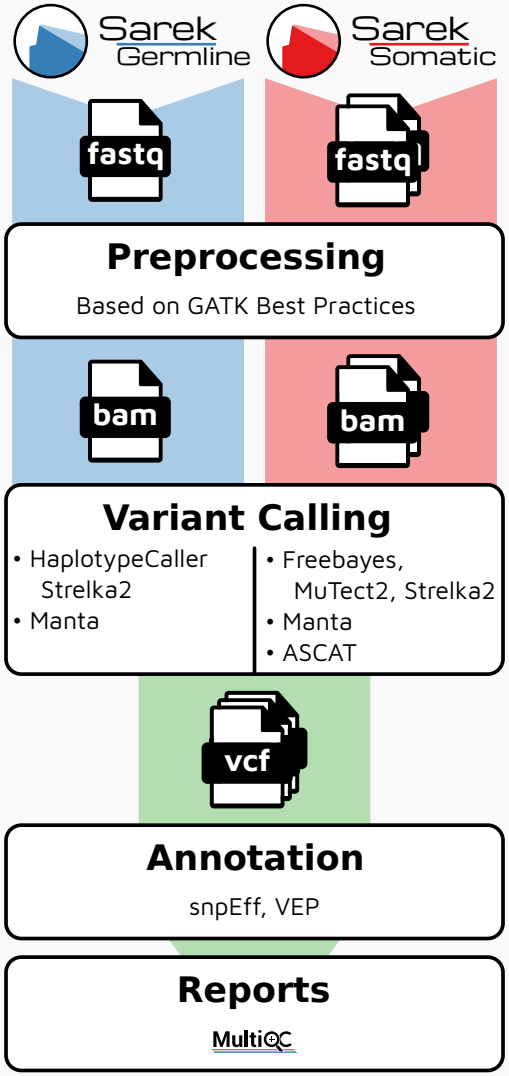
Sarek

Sarek Germline

Sarek Somatic

Sarek Exome

Data and files workflow



Preprocessing

gatk ☑ Best Practices

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

Based on GATK Best Practices (GATK 4.0)

Preprocessing

gatk ☑ Best Practices

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

Based on GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa`

Preprocessing

gatk ☑ Best Practices

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`

Preprocessing

gatk ☑ Best Practices

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`

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:
  - MuTect2
  - Freebayes
  - Strelka2

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
- 🗄 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
  - 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 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
Martin Proks
Rong Yu
Jian Zhao

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

Clinical Genetics
Jesper Eisfeldt

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

Nextflow folks
Paolo Di Tommaso
Sven Fillinger
Alexander Peltzer

Any questions?

http://sarek.scilifelab.se/

https://github.com/SciLifeLab/Sarek