Bioinformatics 101

Analysing WGS data

SciLifeLab

Karolinska Institutet

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

NBIS

Maxime U. Garcia

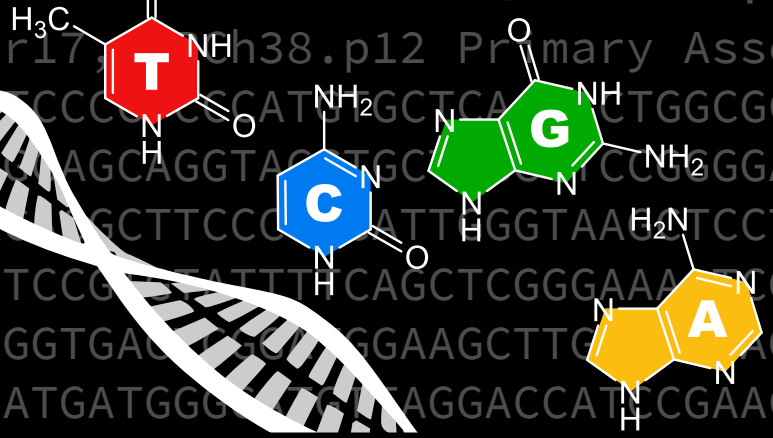
maxulysse.github.io

@MaxUlysse

@gau

Barntumörbanken

```
bwa mem -R \"${readGroup}\" ${extra} -t ${task.cpus
${genomeFile} ${fastqFile1} ${fastqFile2} | \
samtools sort --threads ${task.cpus} -m 2G - > ${id
>Homo sapiens chr17, GRCh38.p12 Primary Assembly
```



Plan

- DNA: From the sequencing to files

Plan

- DNA: From the sequencing to files
- Preprocessing: What to do with these files?

Plan

- DNA: From the sequencing to files
- Preprocessing: What to do with these files?
- Variant Calling and Annotation: Finally getting some results

DNA: From the sequencing to files

DNA: From the sequencing to files

- Sequencing
- Formats

How to store nucleotide sequence?

How to store nucleotide sequence?

```
AGCATCATACGGGGCTTTGG
CTGTACTGTACAGTTACTGT
AGGGGCAGTGACGCCGC
```

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence
- Some meta data

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence
- Some meta data

```
AGCATCATACGGGGCTTTGG
CTGTACTGTACAGTTACTGT
AGGGGCAGTGACGCCGC
```

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence
- Some meta data

```
> My sequence
AGCATCATACGGGGCTTTGG
CTGTACTGTACAGTTACTGT
AGGGGCAGTGACGCCGC
```

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence
- Some meta data

```
> My sequence|P3X-974
AGCATCATACGGGGCTTTGG
CTGTACTGTACAGTTACTGT
AGGGGCAGTGACGCCGC
```

FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence
- Some meta data

```
> My sequence|P3X-974|Homo Sapiens
AGCATCATACGGGGCTTTGG
CTGTACTGTACAGTTACTGT
AGGGGCAGTGACGCCGC
```

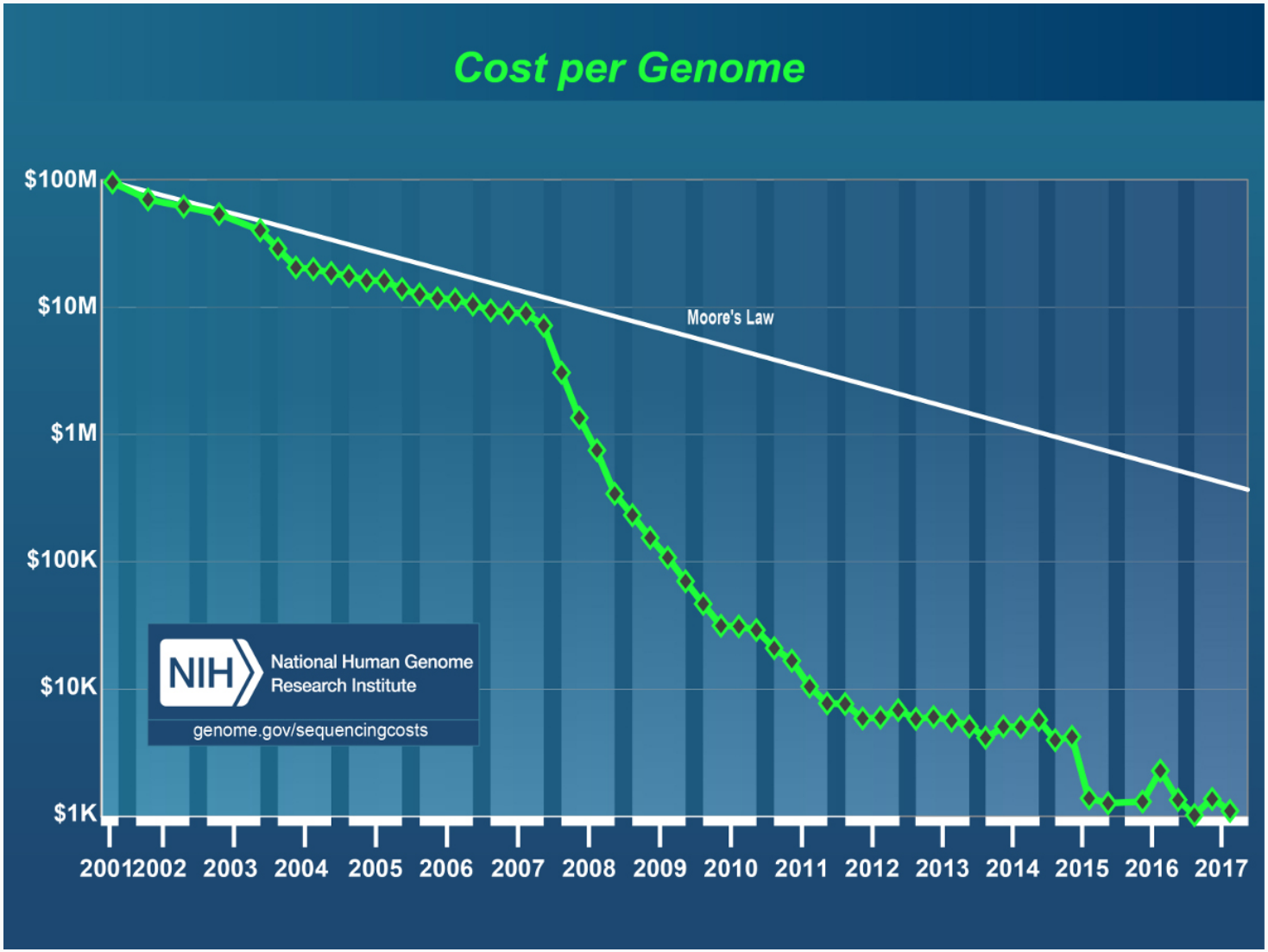
FASTA

FASTA: text-based format for storing either nucleotide or peptide sequences.

- Plainly store sequence
- Some meta data

```
> My sequence|P3X-974|Homo Sapiens|GRCh38
AGCATCATACGGGGCTTTGG
CTGTACTGTACAGTTACTGT
AGGGGCAGTGACGCCGC
```

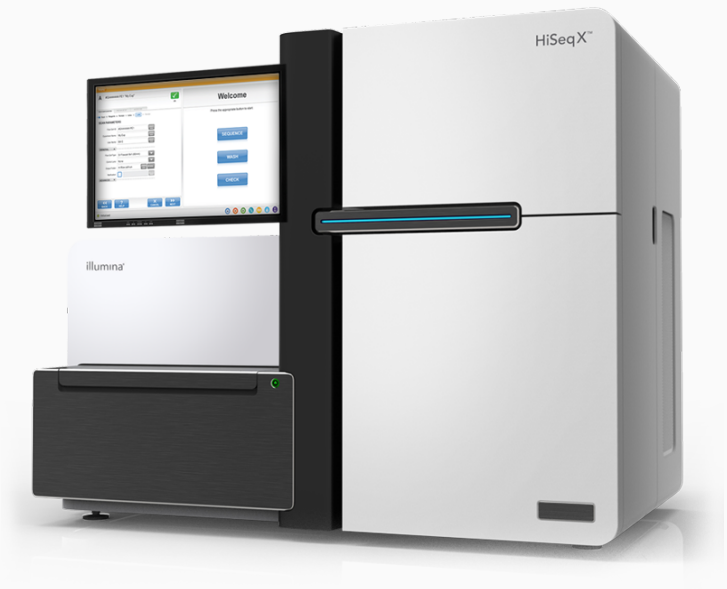
Moore's law in Bioinformatics



Sequencing with Illumina

Illumina's HiSeq X

Sequencing with Illumina



Illumina's HiSeq X

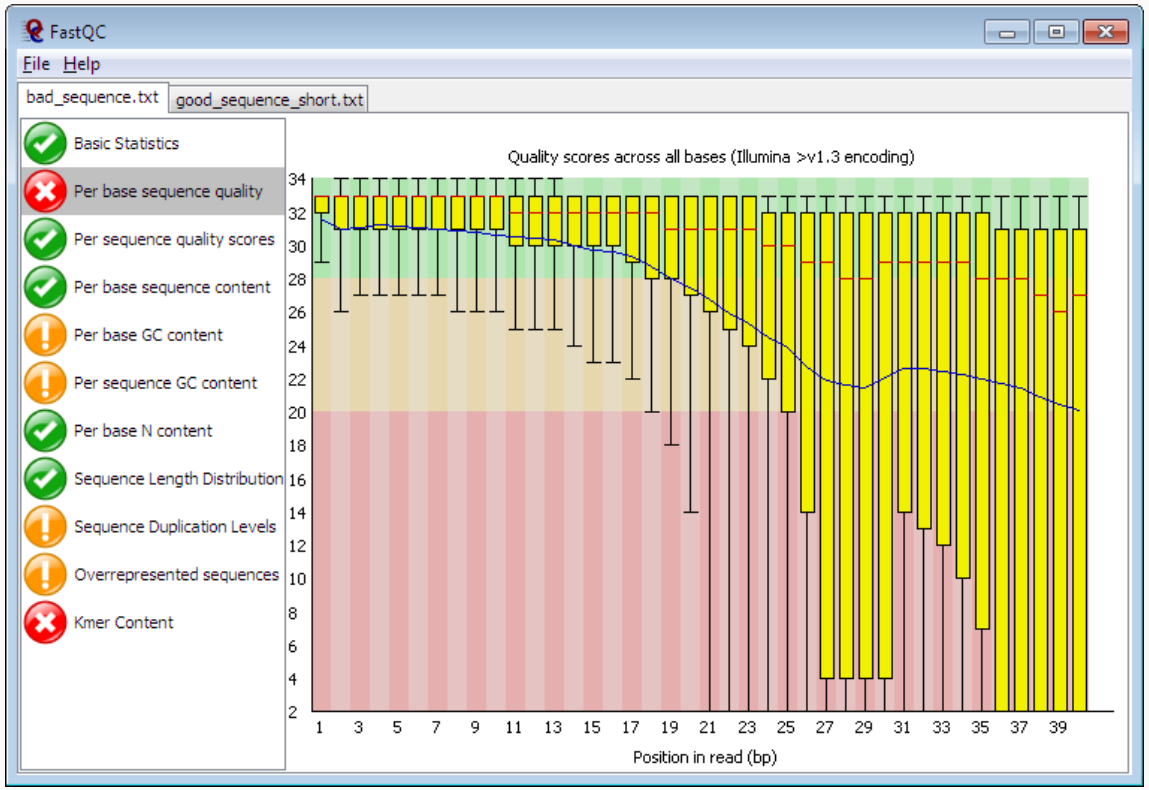
- Short reads ($\sim$ 120 -> 150 bp)

Sequencing with Illumina

Illumina's NovaSeq

- Short reads ($\sim$ 120 -> 150 bp)

Back to sequencing



Each base in a read is assigned a quality score probability of error

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.

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

FASTQ Files

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

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

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

First conclusion

- FASTA: Used to store sequences
  - You might use or even open such file

First conclusion

- FASTA: Used to store sequences
  - You might use or even open such file
- FASTQ: Used to store sequences and quality
  - You will see that
  - You won't use that directly
  - You will never open such file
  - You will transform it

Preprocessing: What to do with these files?

- Assembly

Preprocessing: What to do with these files?

- Assembly
- Cleanup

Assembly - the reads

Difficult question is coming

- Human Genome: $\sim$ 3,234.83 Mb

Assembly - the reads

Difficult question is coming

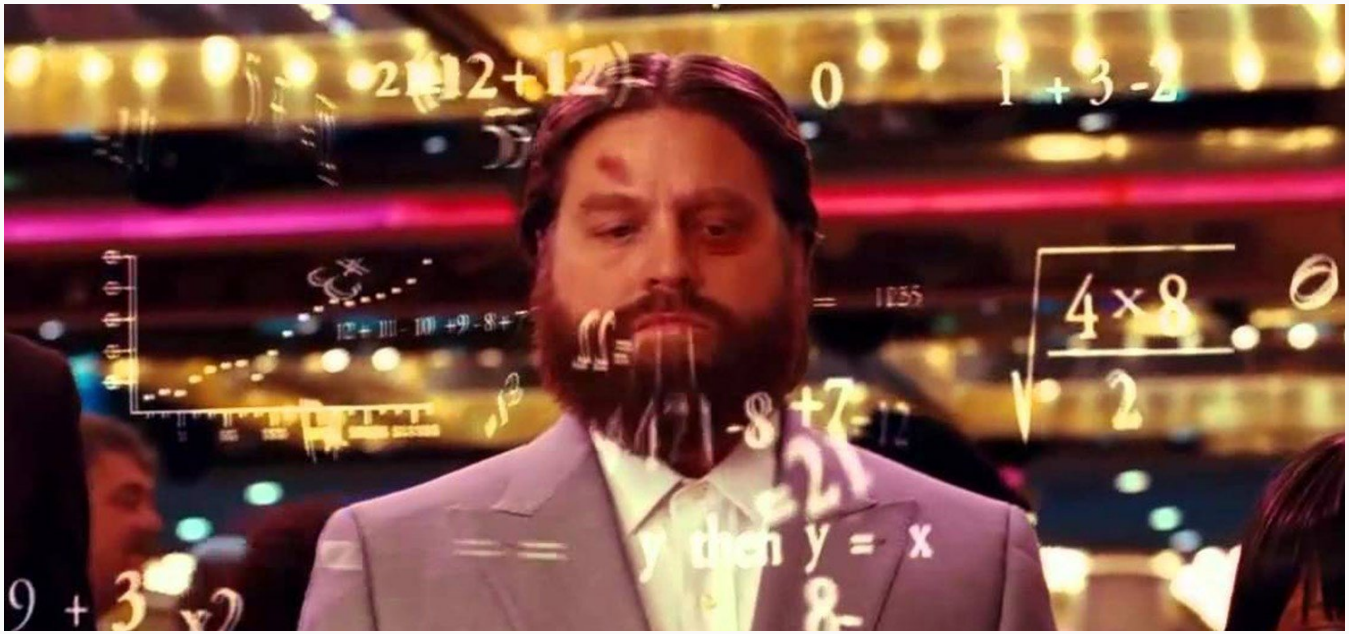
- Human Genome: $\sim$ 3,234.83 Mb
- Depth of sequencing: 30X

Assembly - the reads

Difficult question is coming

- Human Genome: $\sim$ 3,234.83 Mb
- Depth of sequencing: 30X
- Reads: $\sim$ 120 -> 150 bp

Assembly - the reads



How many reads in the end?

- Human Genome: $\sim$ 3,234.83 Mb
- Depth of sequencing: 30X
- Reads: $\sim$ 120 -> 150 bp

Assembly - the reads

600 M reads

- Human Genome: $\sim$ 3,234.83 Mb
- Depth of sequencing: 30X
- Reads: $\sim$ 120 -> 150 bp

Assembly with short reads - I

```
                TACGGGG                                GGCTTTG
                            AGCATCA
ATCATAC                                     ACGCCGC
```

Some short reads

Assembly with short reads - II

```
            GGCTTTG
      TACGGGG
   ATCATAC
AGCATCA
```

Some can be assembled

Assembly with short reads - III

```
            GGCTTTG
        TACGGGG
   ATCATAC
AGCATCA                                                  ACGCCGC
AGCATCATACGGGGCTTTGGCTGTACTGTACAGTTACTGTAGGGGCAGTGACGCCGC
```

Easier with a Reference

Quality guidelines for processing data

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

From the Broad Institute: GATK Best Practices (GATK 4.0)

Quality guidelines for processing data

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

From the Broad Institute: GATK Best Practices (GATK 4.0)

- Reads mapped to reference genome with `bwa`

Quality guidelines for processing data

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

From the Broad Institute: GATK Best Practices (GATK 4.0)

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

Quality guidelines for processing data

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

From the Broad Institute: GATK Best Practices (GATK 4.0)

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

Tools

Burrows-Wheeler Aligner

http://bio-bwa.sourceforge.net/

Software package for mapping low-divergent sequences against a large reference genome.

Tools

Burrows-Wheeler Aligner

http://bio-bwa.sourceforge.net/

Software package for mapping low-divergent sequences against a large reference genome.

GATK/Picard

https://software.broadinstitute.org/gatk/
https://broadinstitute.github.io/picard/

Sets of bioinformatic tools for analyzing/manipulating high-throughput sequencing (HTS) data.

Command lines

```
bwa mem -R \"@RG\tID:group1\tSM:file1\tPL:illumina\tLB:lib1\tPU:unit1\" -M \
Reference.fasta file1.fastq file2.fastq | \
samtools sort - > file.bam

gatk MarkDuplicates \
--INPUT file.bam \
--METRICS_FILE file.bam.metrics \
--ASSUME_SORT_ORDER coordinate \
--CREATE_INDEX true \
--OUTPUT file.md.bam

gatk BaseRecalibrator \
--input file.md.bam \
--output file.recal.table \
-R Reference.fasta
```

BAM files

Binary Alignment Map (BAM): compressed binary representation for storing biological sequences aligned to a reference sequence.

BAM files

Binary Alignment Map (BAM): compressed binary representation for storing biological sequences aligned to a reference sequence.

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 AGCATCATACGGGGCTTTG *
```

BAM files

Binary Alignment Map (BAM): compressed binary representation for storing biological sequences aligned to a reference sequence.

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 AGCATCATACGGGGCTTTG *
```

1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1-based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

CRAM files

Basically even more compressed BAM.

CRAM files

Basically even more compressed BAM.

Not yet widely adopted, but good to know about.

Second conclusion

- Follow Best Practices, unless you know what you're doing

Second conclusion

- Follow Best Practices, unless you know what you're doing
- Try to save space
  - Keep only your latest BAMs and your FASTQs
  - Look at BAM files with visualization tools (IGV…)

Variant Calling

- Differences to Reference genome

How many variants?

- According to 1 000 Genomes Project:

How many variants?

- According to 1 000 Genomes Project:
  - $\sim$ 1 000 deletions

How many variants?

- According to 1 000 Genomes Project:
  - $\sim$ 1 000 deletions
  - $\sim$ 1 000 insertions

How many variants?

- According to 1 000 Genomes Project:
  - $\sim$ 1 000 deletions
  - $\sim$ 1 000 insertions
  - $\sim$ 160 copy number variation

How many variants?

- According to 1 000 Genomes Project:
  - $\sim$ 1 000 deletions
  - $\sim$ 1 000 insertions
  - $\sim$ 160 copy number variation
  - $\sim$ 10 inversions

Some Variant Calling tools

- SNVs[1] and small indels[2]

[1] Single Nucleotide Variant
[2] insertion or deletion

Some Variant Calling tools

- SNVs[1] and small indels[2]
  - HaplotypeCaller (GATK)
  - Strelka2 (Illumina)

[1] Single Nucleotide Variant
[2] insertion or deletion

Some Variant Calling tools

- SNVs[1] and small indels[2]
  - HaplotypeCaller (GATK)
  - Strelka2 (Illumina)
- Structural variants:

[1] Single Nucleotide Variant
[2] insertion or deletion

Some Variant Calling tools

- SNVs[1] and small indels[2]
  - HaplotypeCaller (GATK)
  - Strelka2 (Illumina)
- Structural variants:
  - Manta (Illumina)

[1] Single Nucleotide Variant
[2] insertion or deletion

Annotation

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

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:H
  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
```

Third conclusion

- Lots of different variant callers
  - Lots of variants found

Third conclusion

- Lots of different variant callers
  - Lots of variants found
  - Need to filter
  - Need to annotate
  - Can even improve that with prioritisation

What do we want?

Do analysis!

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 need?

- Tools

What do we need?

- Tools
  - Installed
  - Specific version

What do we need?

- Tools
  - Installed
  - Specific version
- Reference files

What do we need?

- Tools
  - Installed
  - Specific version
- Reference files
  - Dowloaded
  - Specific version

What do we need?

- Tools
  - Installed
  - Specific version
- Reference files
  - Dowloaded
  - Specific version
- Annotation files / databases

What do we need?

- Tools
  - Installed
  - Specific version
- Reference files
  - Dowloaded
  - Specific version
- Annotation files / databases
  - Dowloaded
  - Specific version

What do we need?

- Tools
  - Installed
  - Specific version
- Reference files
  - Dowloaded
  - Specific version
- Annotation files / databases
  - Dowloaded
  - Specific version
- Works with cluster executor

What is Sarek?

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?

Sarek

http://sarek.scilifelab.se/

- Analysis germline and somatic workflow
- Whole genome or targeted sequencing
- Developed with NGI and NBIS

NATIONAL GENOMICS INFRASTRUCTURE

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

Build with

nextflow

https://www.nextflow.io/

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

Build with

https://www.nextflow.io/

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

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

- Docker-like container engine
  - Specific for HPC environnment

Sarek exists in multiple flavors

Sarek exists in multiple flavors

Sarek

Sarek Germline

Sarek exists in multiple flavors

Sarek

Sarek Germline

Sarek Somatic

Sarek exists in multiple flavors

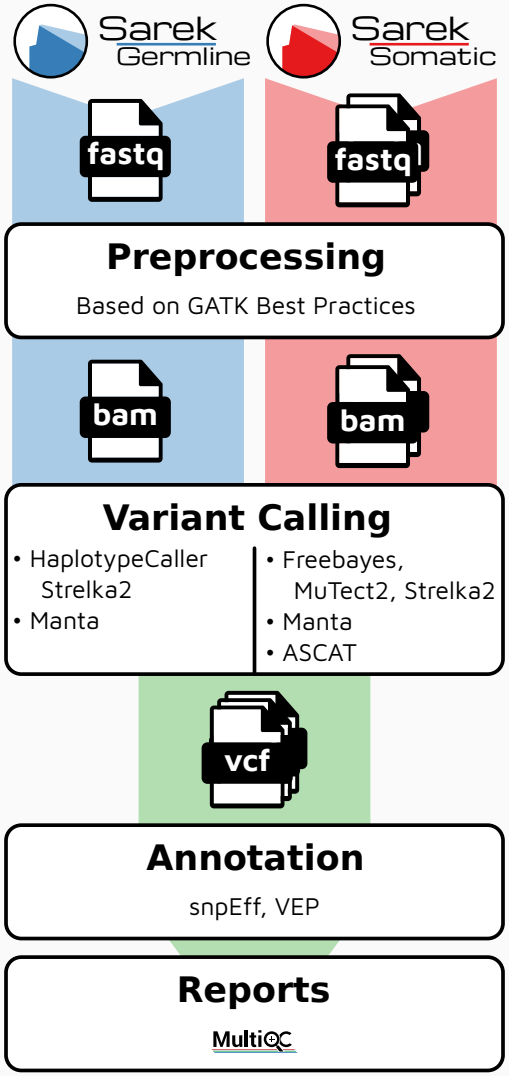
Sarek

Sarek Germline

Sarek Somatic

Sarek Exome

Data and files workflow



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
- https://maxulysse.github.io/2018-Analysing-WGS-data/
- http://rsg-sweden.iscbsc.org/
- https://www.biostars.org/