



# BIO634 - Day 1: Next generation sequencing (NGS) II

**Carla Bello**, carla.bello@ieu.uzh.ch

**June 3 - 4th, 2021**

Zürich, CH

# Overview

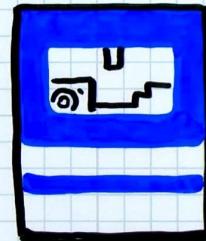
- **Sequencing technologies:** Which one to choose?
- **File formats:** FASTA, FASTQ, SAM/BAM, VCF, GTF, GFF
- **Raw reads:** Single-end, Paired-end, Mate pairs
- **Toolbox for mapping NGS data:** QC and Mapping
- **Hands-on session**

# What is sequencing?

DNA/RNA sequencing is the process of **determining the precise order** of nucleotides within a DNA/RNA molecule.

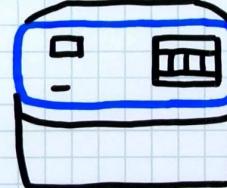
Late 70's

SANGER



Early 00's

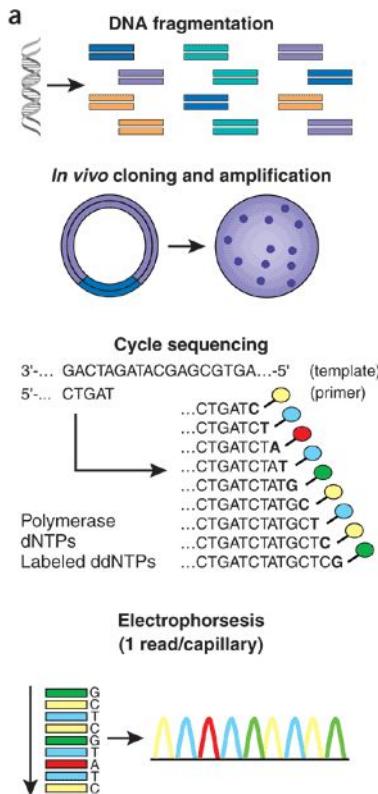
NGS  
MASSIVELY  
PARALLEL



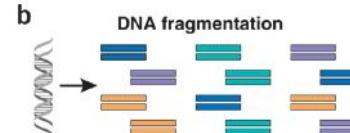
# NGS is much more efficient

## Sanger

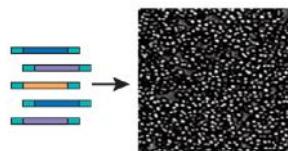
- Few targets
- Cloning
- Cost effective
- Familiar



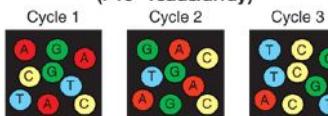
## NGS (Illumina)



Generation of polony array



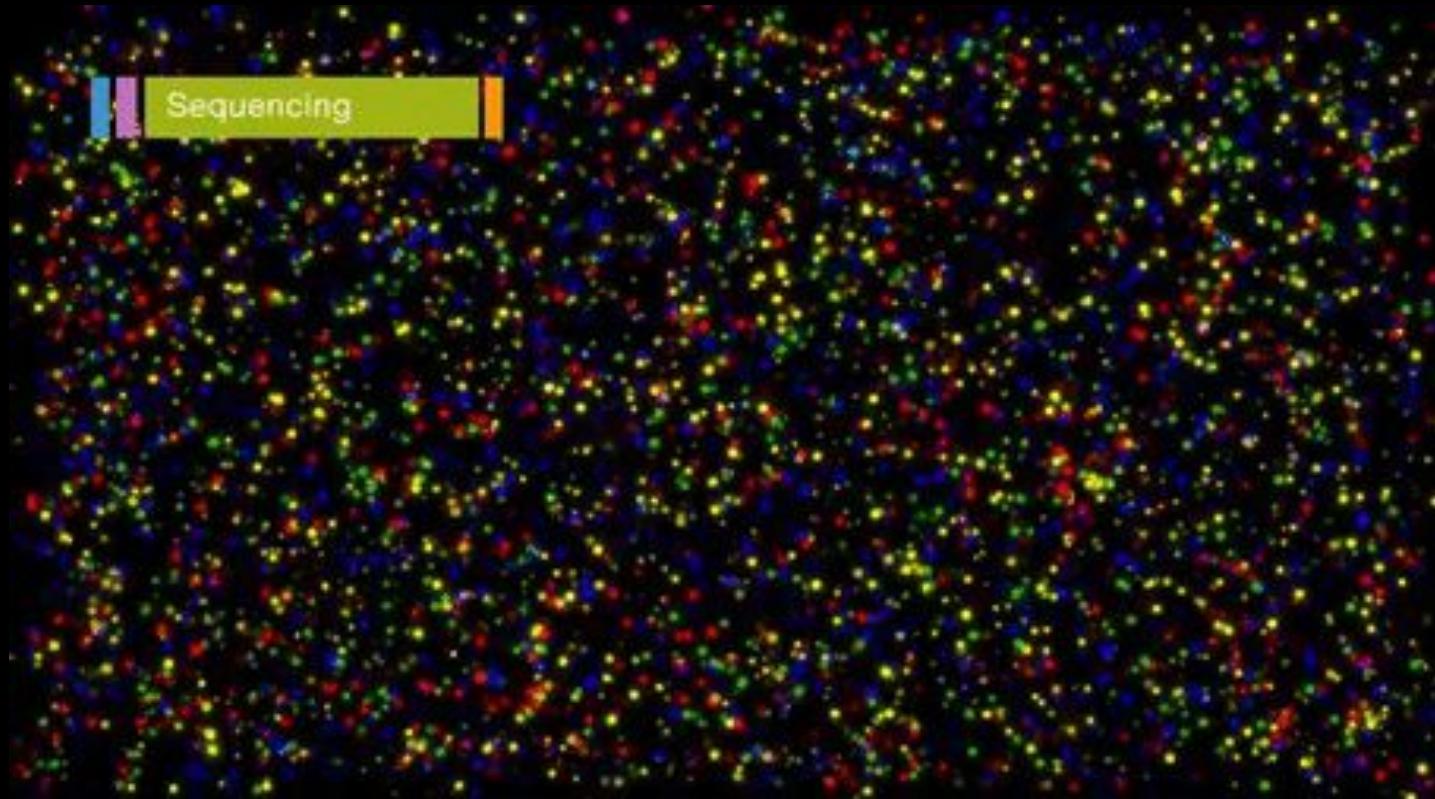
### Cyclic array sequencing (>10<sup>6</sup> reads/array)



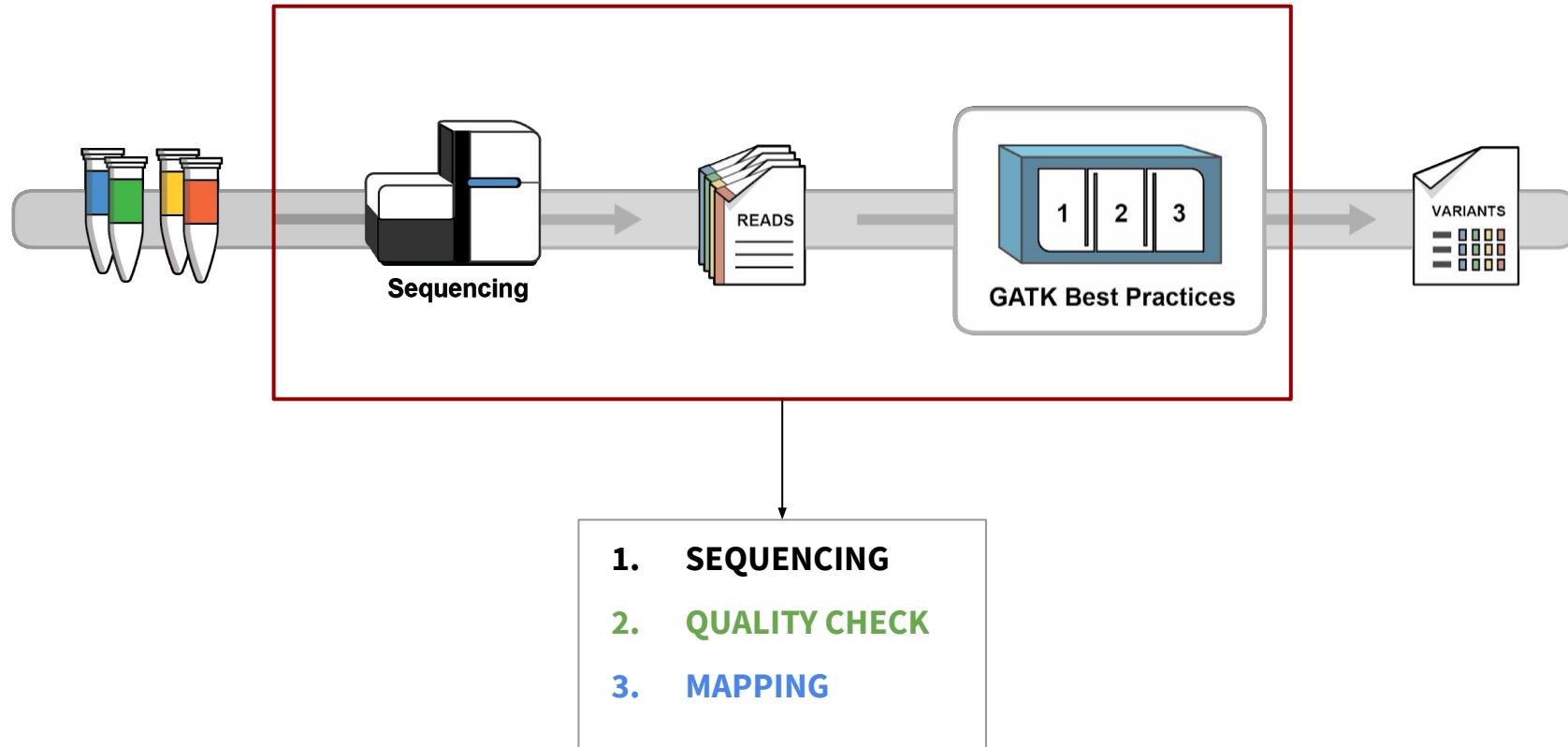
What is base 1? What is base 2? What is base 3?

- Many targets
- No cloning
- Detection of variants
- Quick

Sequencing

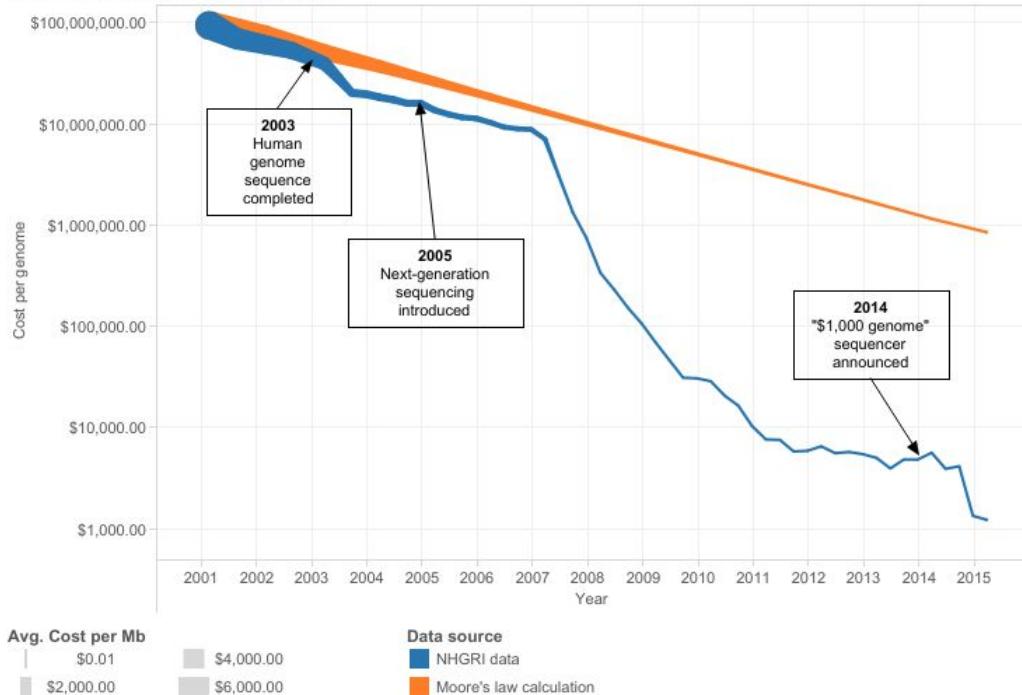


# Sequencing and data analysis

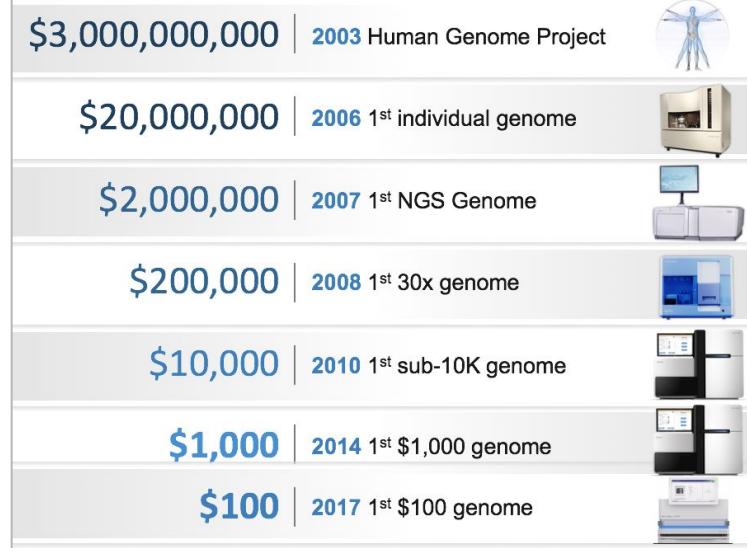


Source: GATK webpage

# Reduction of sequencing cost over time

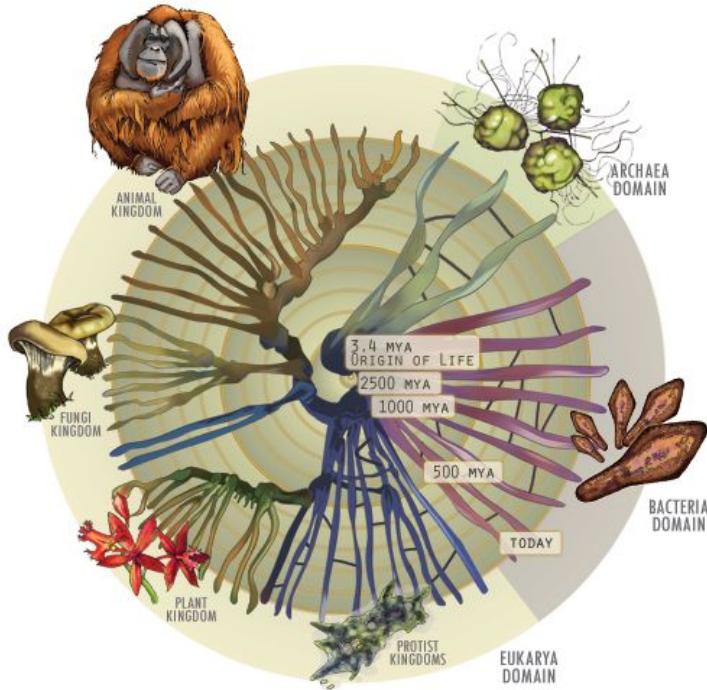


Decline in real costs compared to expected declines based on Moore's Law.  
Trend line: Cost per human genome. Line width: Cost per megabase (Mb)  
(Data: NHGRI <https://www.genome.gov/27541954/dna-sequencing-costs-data/>)

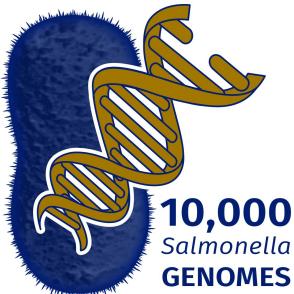
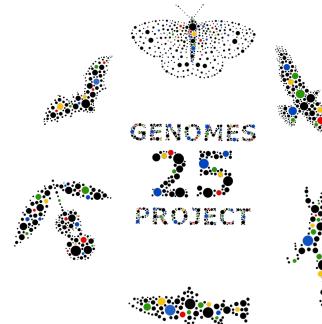
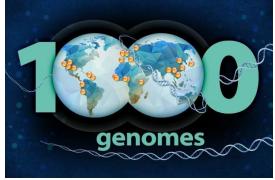


**¢1? | 2XXX 1st ¢1 genome?**

# Many species genomes are available



# Consortia all over the world



International  
Cancer Genome  
Consortium



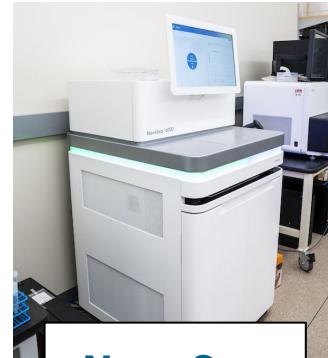
**PCAWG**  
PanCancer Analysis  
of WHOLE GENOMES



# Most used sequencing platforms today



Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	\$59 (PE100)	\$53 (PE150)	\$108 (PE300)	\$697

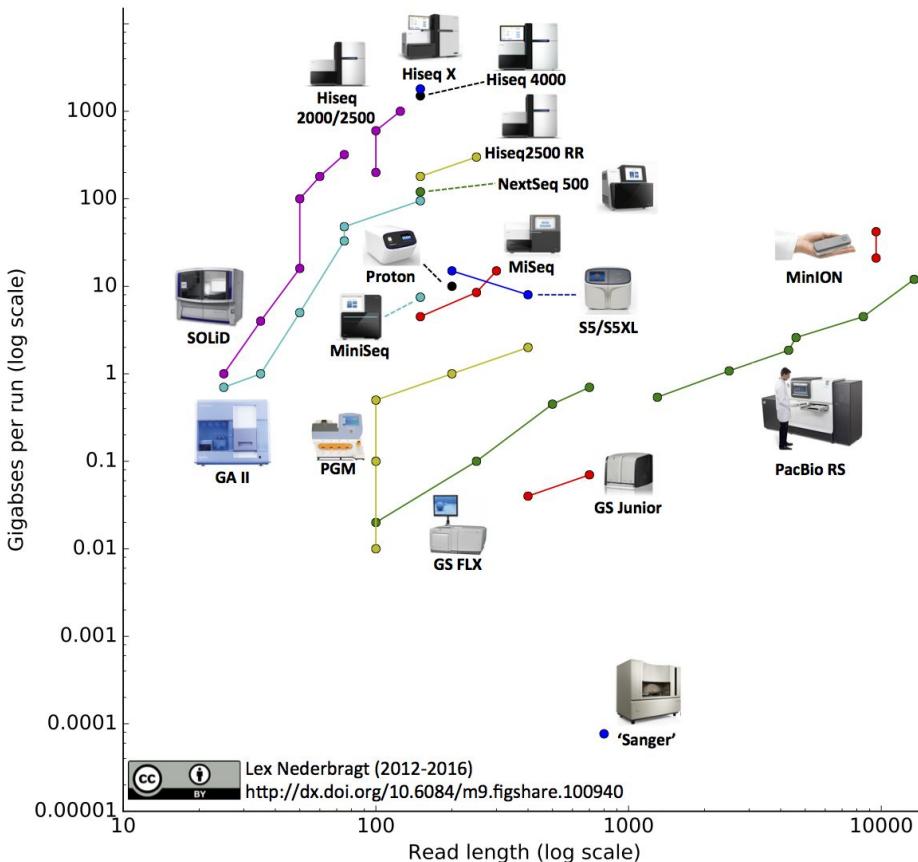


NovaSeq



**Choose wisely**

# Different sequencing outcomes



- **Paired end reads:** linear vs circularized fragments
- **Sequencing technologies:** DNA polymerase, DNA ligase, synthesis H+ detection, syntheses, and nanopore
- **Library amplification methods:** emPCR, bridge amplification, and its absence in some 3rd gen platforms
- **Run times**
- **Error rates**
- **Read lengths**



Lex Nederbragt (2012-2016)

<http://dx.doi.org/10.6084/m9.figshare.100940>

# Choosing between platforms

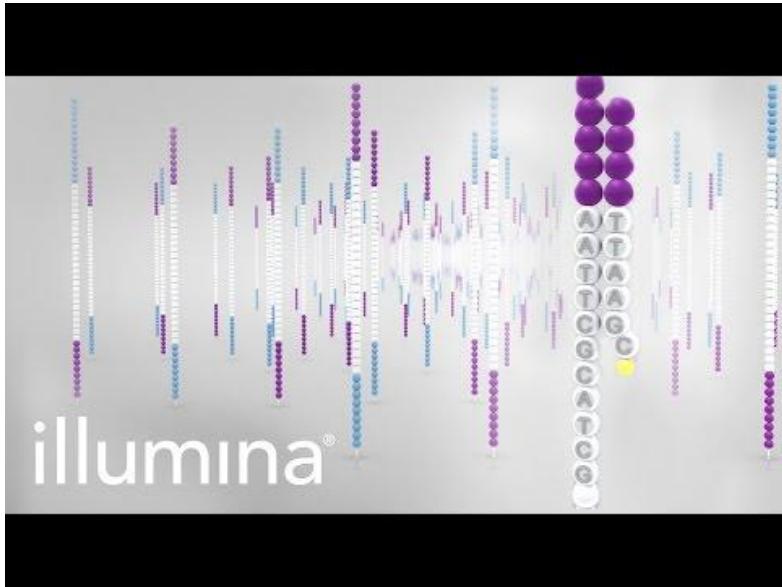
**Table 1.**

Comparison of PacBio sequencing platforms to two current industry standards

Platform	Read length	Number reads	Error rate	Run time
PacBio RSII (per SMRT cell)	Average 10–16 kb	~55 000	13–15%	0.5–6 hours
PacBio Sequel (per SMRT cell)	Average 10–14 kb	~365 000	13–15%	0.5–10 hours
Illumina HiSeq 4000	2 × 150 bp	5 billion	~0.1%	<1–3.5 days
Illumina MiSeq	2 × 300 bp	25 million	~0.1%	4–55 hours

# Current and upcoming technologies

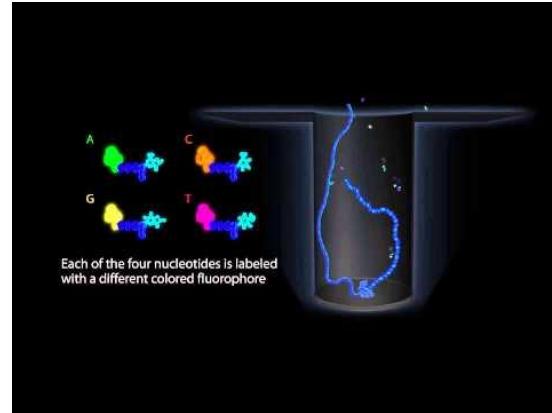
## Illumina (HiSeq)



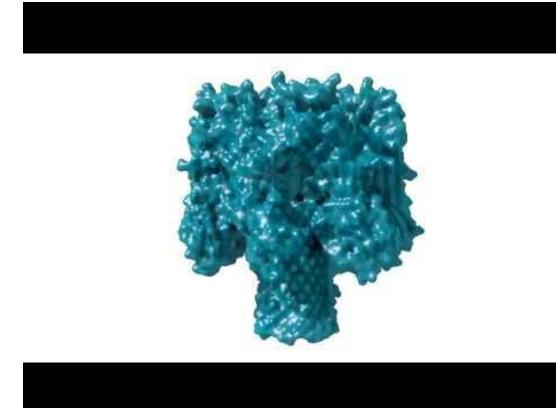
Links:

- <https://www.youtube.com/watch?v=fCd6B5HRaZ8>
- <https://youtu.be/WMZmG00uhwU>
- <https://www.youtube.com/watch?v=hs0FdiTHMbc>

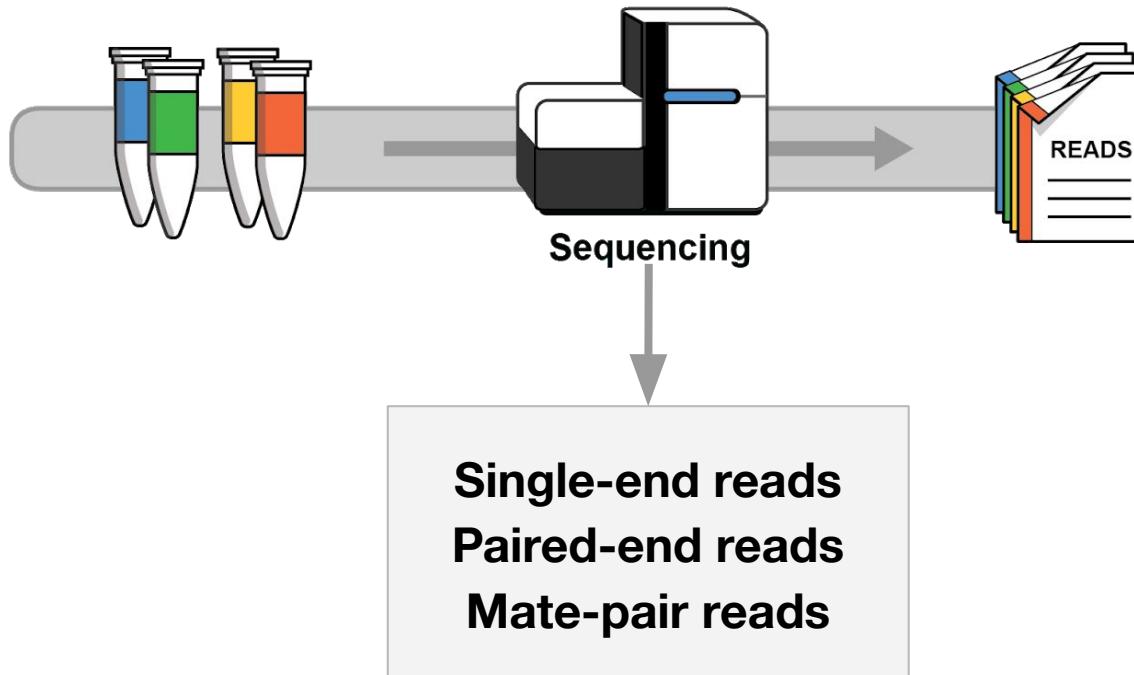
## PacBio (SMRT seq)



## Oxford Nanopore (ONT)

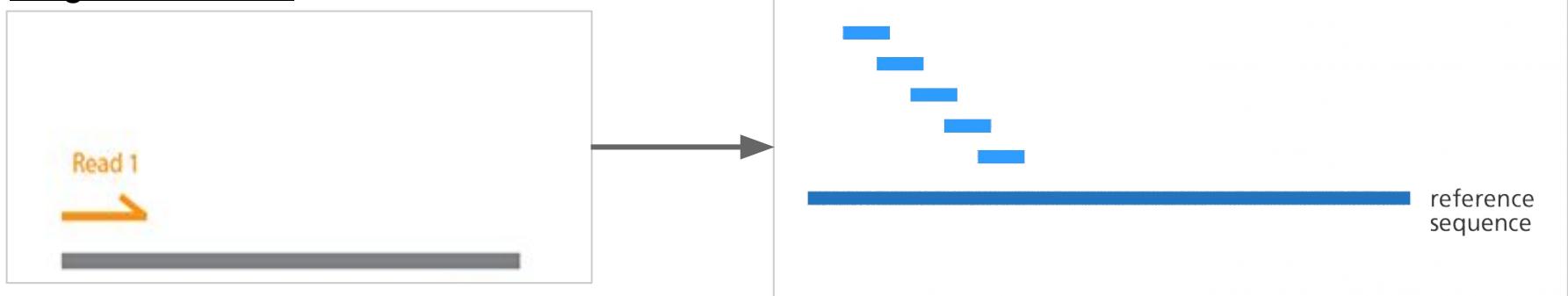


# There are different type of reads



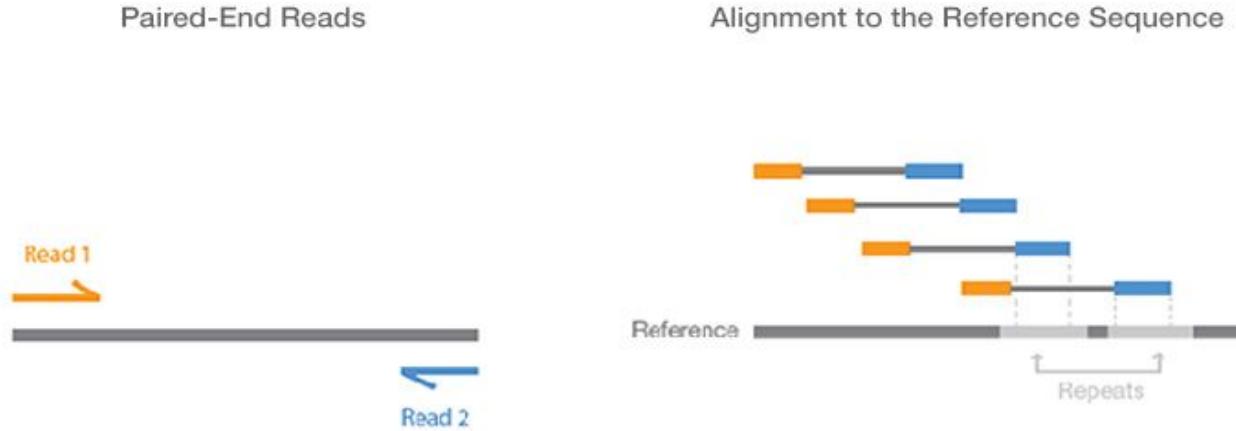
# Single-end reads are a single sequence

## Single-end reads



- Each read is a **single sequence** from one end of a DNA fragment (**single FASTQ file**).
- The fragment is usually 200-800bp long, with the amount being read can be chosen between **50 and 300 bp**.

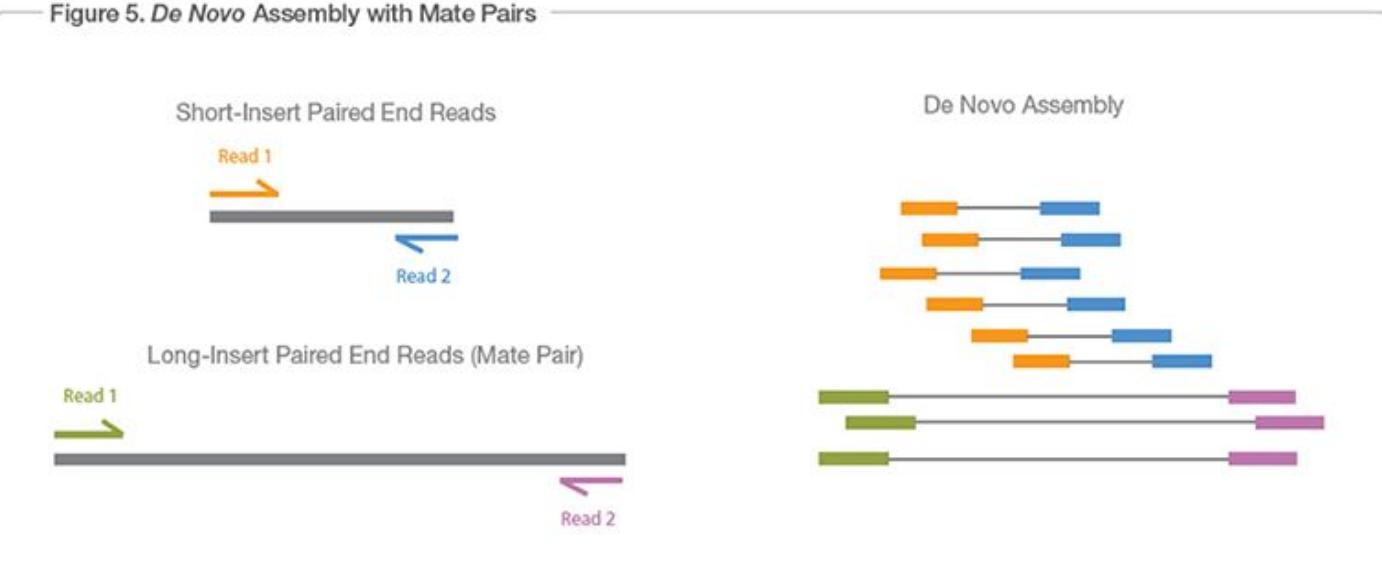
# Paired-end reads



- Each read is **two sequences** (a pair) from **each end** of the same DNA fragment (**two FASTQ files**).
- The distance between the reads on the original genome sequence is equal to the length of the DNA fragment that was sequenced, usually **200-800 bp**.

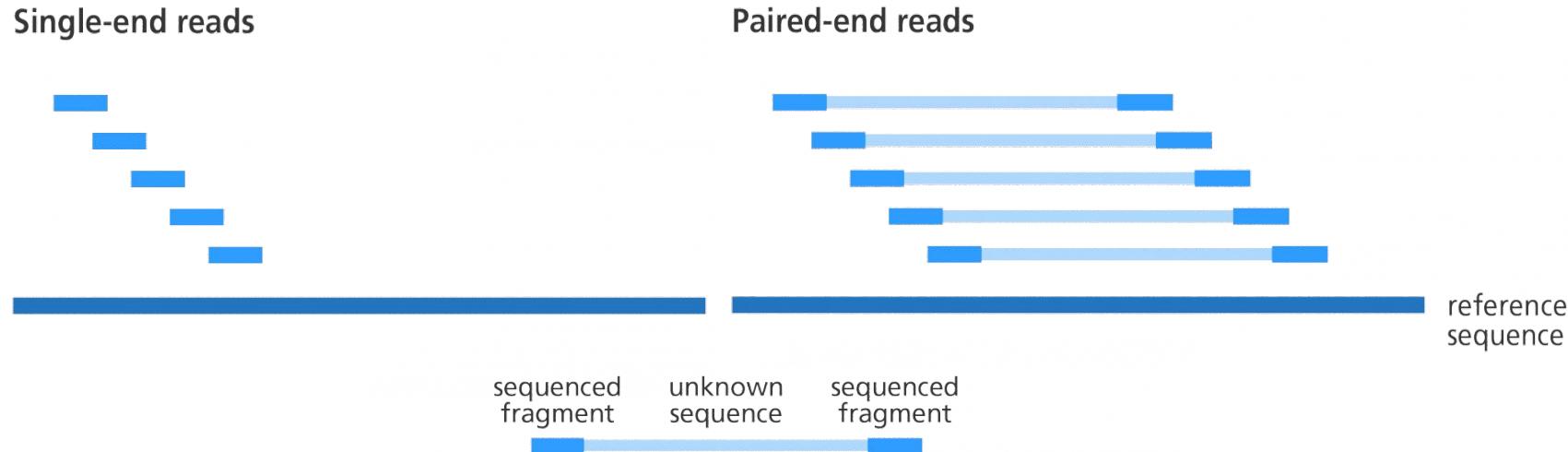
# Mate-pair reads

Figure 5. *De Novo Assembly with Mate Pairs*

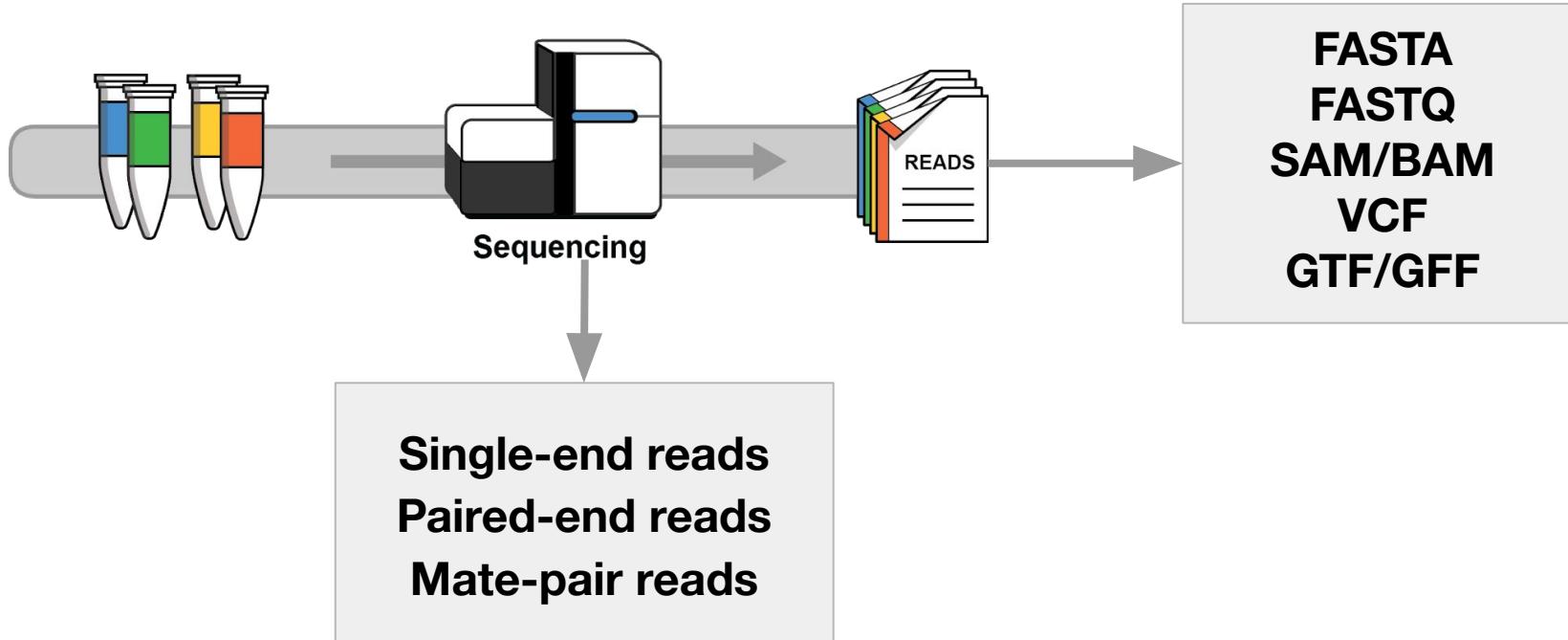


- Each read is **two sequences from each end of the same DNA fragment**, but the distance between the reads on the original genome sequence is much longer, e.g. **3000-10000bp**

# Single and paired-end reads are the most used



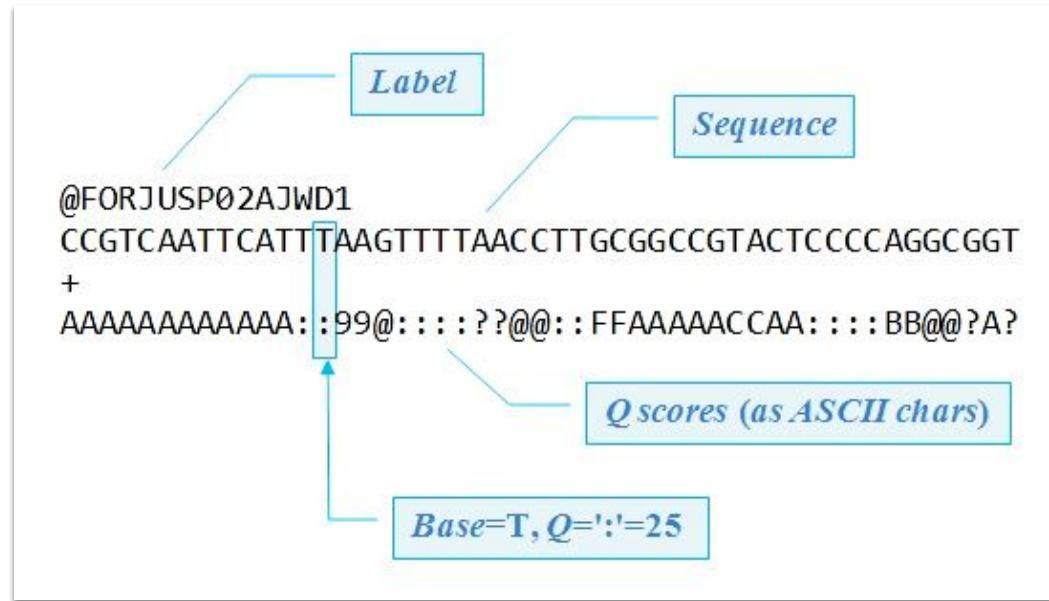
# Different files are obtained after sequencing



# Most commonly used formats

1. **FASTA**: Contains **raw nucleotide** sequence data
2. **FASTQ**: FASTA with **quality scores**.
3. **SAM/BAM**: Alignments with **genomic** information
4. **VCF**: Variant calling files (SNPs, CNVs, INDELs).
5. **GTF/GFF**: Genomic annotation information.

# FASTQ format contains NGS reads

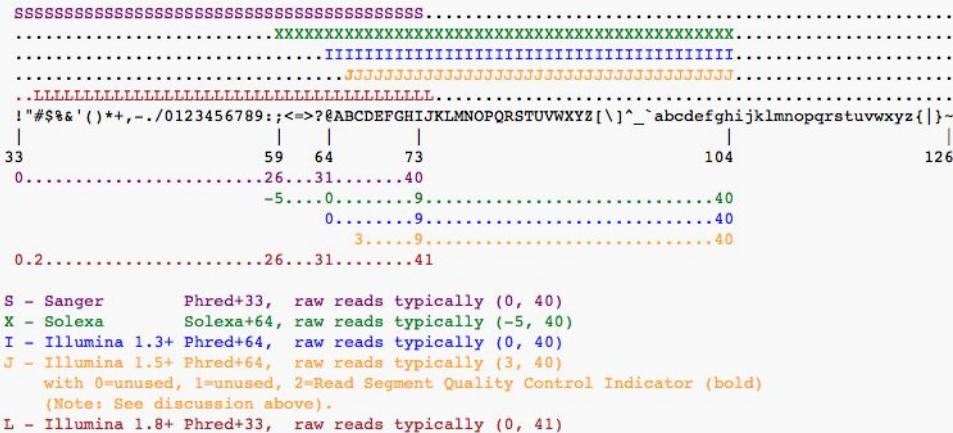


**FASTQ:** FASTA with quality scores (ASCII to Phred Q-scores)

# Phred scores / ASCII encoding

Quality value	Chance it is wrong	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

- $Q = -10 \log_{10} P \iff P = 10^{-Q/10}$ 
    - $Q$  = Phred quality score
    - $P$  = probability of base call being incorrect



**The Phred Quality scores Q is logarithmically related to base calling error probabilities**

# File formats

## SAM(Sequence Alignment/Map) :

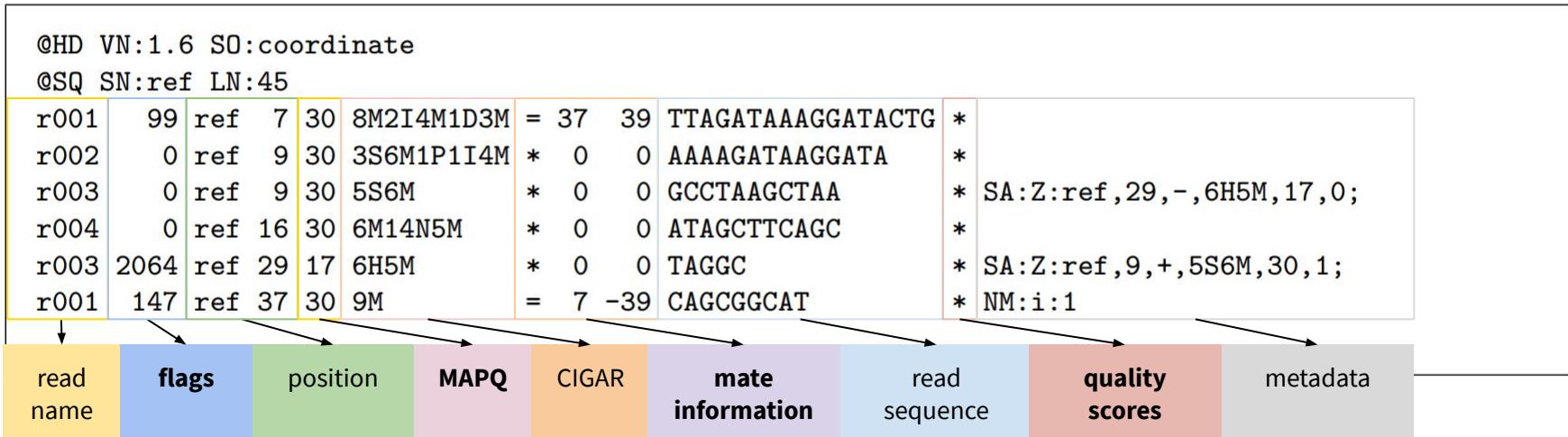
```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAACGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC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

SAMTOOLS: <http://www.htslib.org/doc/sam.html>

FLAGS EXPLAINED: <https://broadinstitute.github.io/picard/explain-flags.html>

# File formats

## SAM(Sequence Alignment/Map) :



SAMTOOLS: <http://www.htslib.org/doc/sam.html>

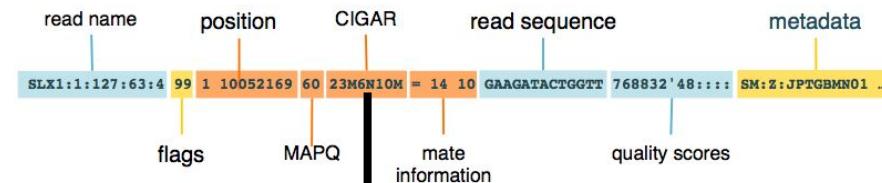
FLAGS EXPLAINED: <https://broadinstitute.github.io/picard/explain-flags.html>

# File formats

## SAM(Sequence Alignment/Map) :

HEADER containing metadata (sequence dictionary, read group definitions etc)

RECORDS containing structured read information (1 line per read record)

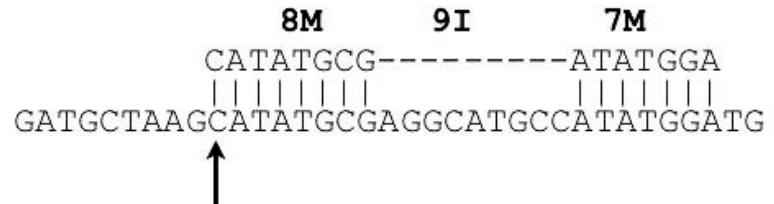


## CIGAR

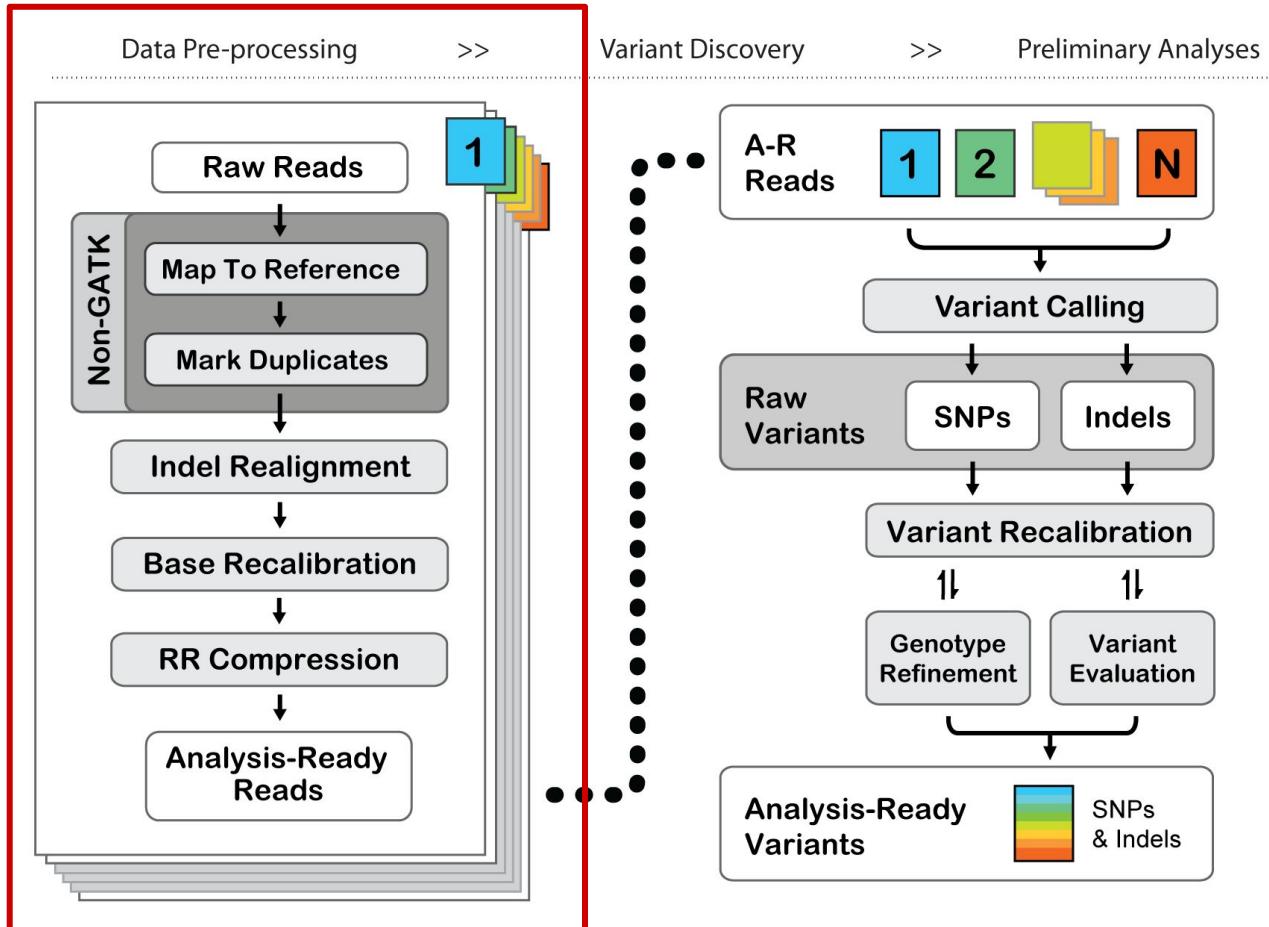
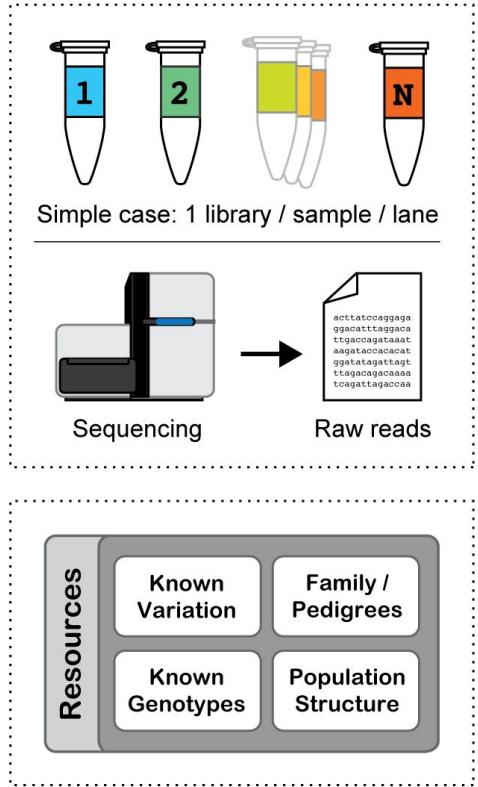
op	Description
M	Alignment match (can be a sequence match or mismatch)
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

Show alignment result simply  
8M9I7M

- 8bp match, 9bp insertion, and then 7bp match



4th line “POS” indicates this position.

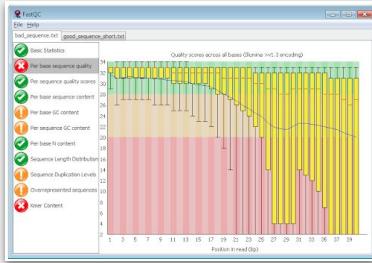


<https://software.broadinstitute.org/gatk/best-practices/>

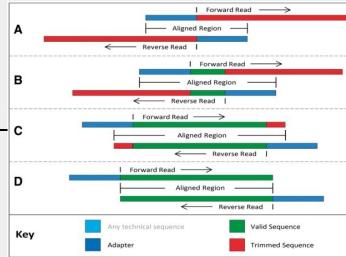
# Toolbox for mapping NGS data

Quality check(QC) and cleaning:

## FASTQC



## Trimmomatic/Cutadapt



Map and remove duplicates:

## BWA/Tophat/STAR

Read: ATCAGCATC

ALT ctg 1: CGAAATGCAATGCTC **ATCAGCATC** GAGCTACTGCACAT

Chromosome: GCGTACATGATAGA **ATCgGCATC**TGGTC-----CTAGTACATGTAAATC

ALT ctg 2: TGATACGA **ATCgcCATC**TGGCA **ATCgcCAG**GAACTAGTCACAT

4 potential hits: **ATCAGCATC** > **ATCgGCATC** > **ATCgcCATC** > **ATCgcCAG**  
2 hit groups: {**ATCAGCATC**, **ATCgcCAG**} and {**ATCgGCATC**, **ATCgcCATC**}

Hits considered in mapQ: **ATCAGCATC** and **ATCgGCATC** (best from each group)

In the output SAM: **ATCgGCATC** as the primary SAM line with mapQ=0  
**ATCAGCATC** as a supplementary line with mapQ>0  
**ATCgcCAG** as a supplementary line with mapQ>0  
**ATCgcCAG** in an XA tag, not as a separate line

## GATK and Picard Tools

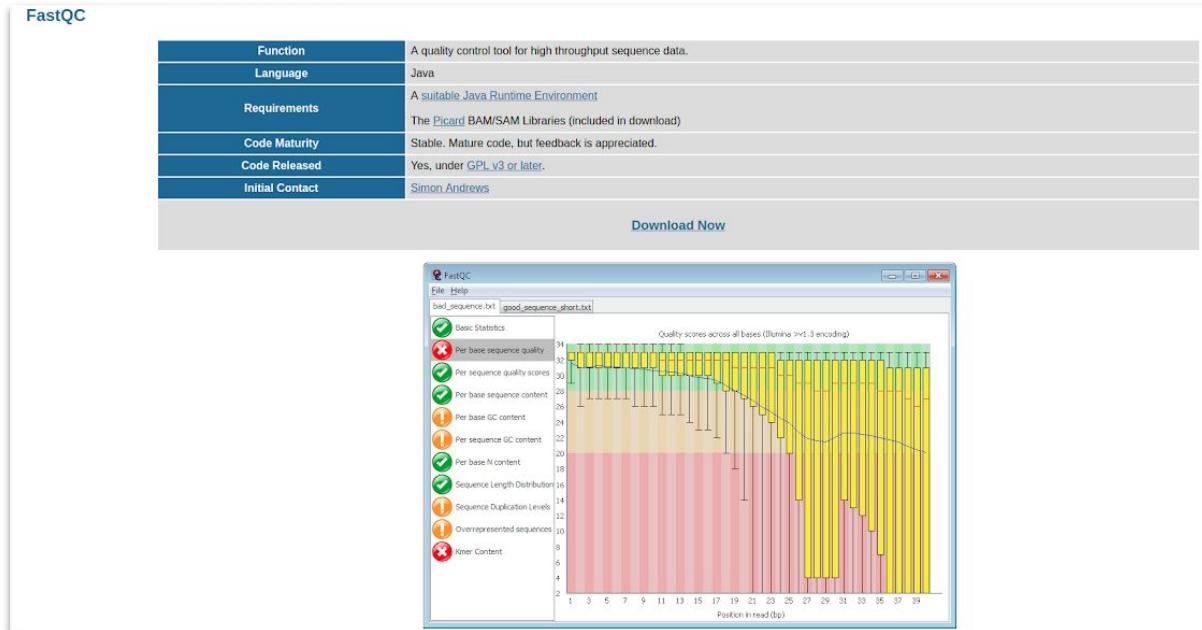
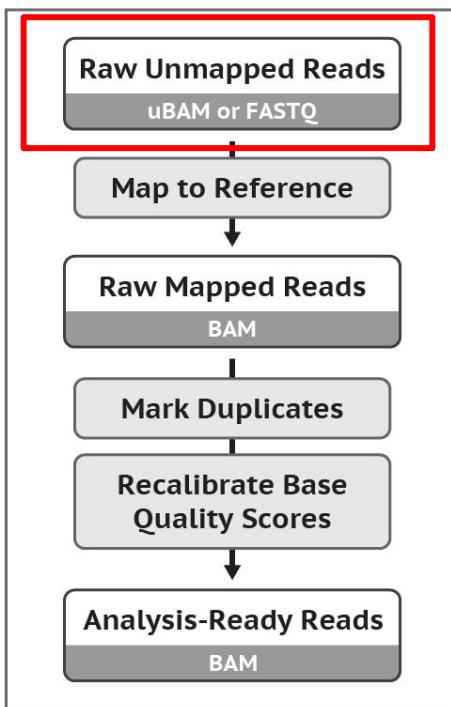
### Picard

build passing



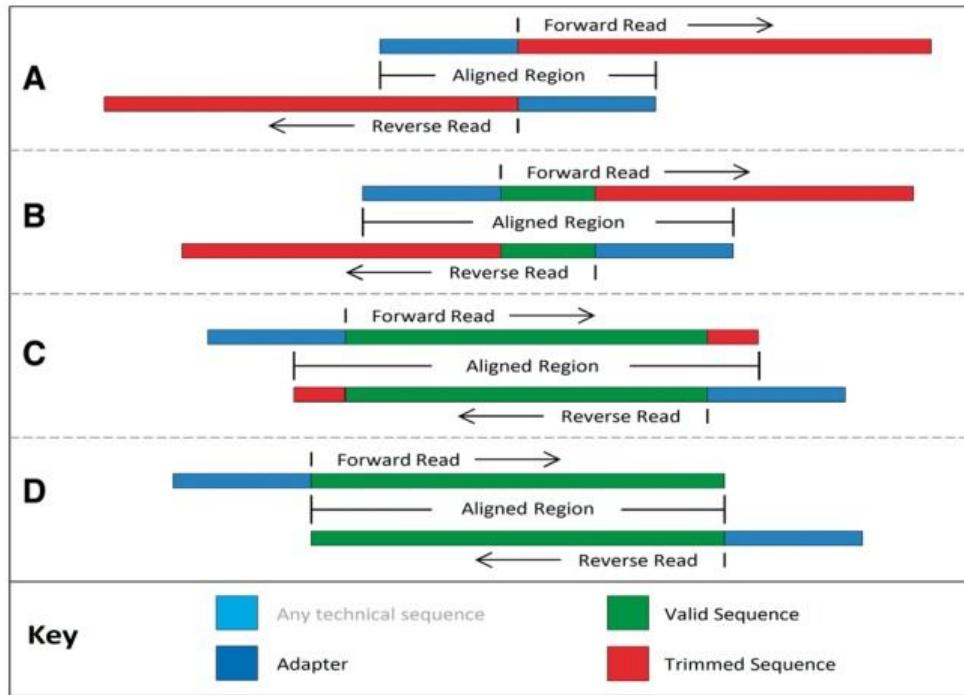
# Quality check: FASTQC

## Data Pre-processing



FastQC: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

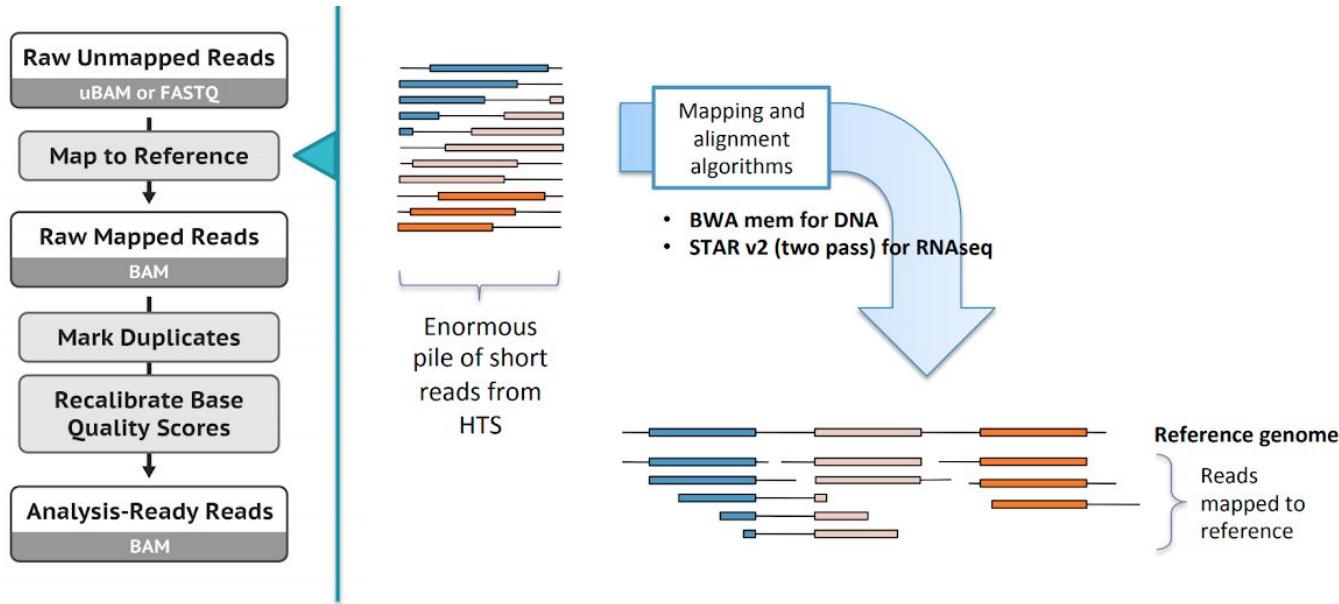
# Clean-up: Trimmomatic



Removing adapter sequences from the sequencing reads using adaptor templates from the NGS platform

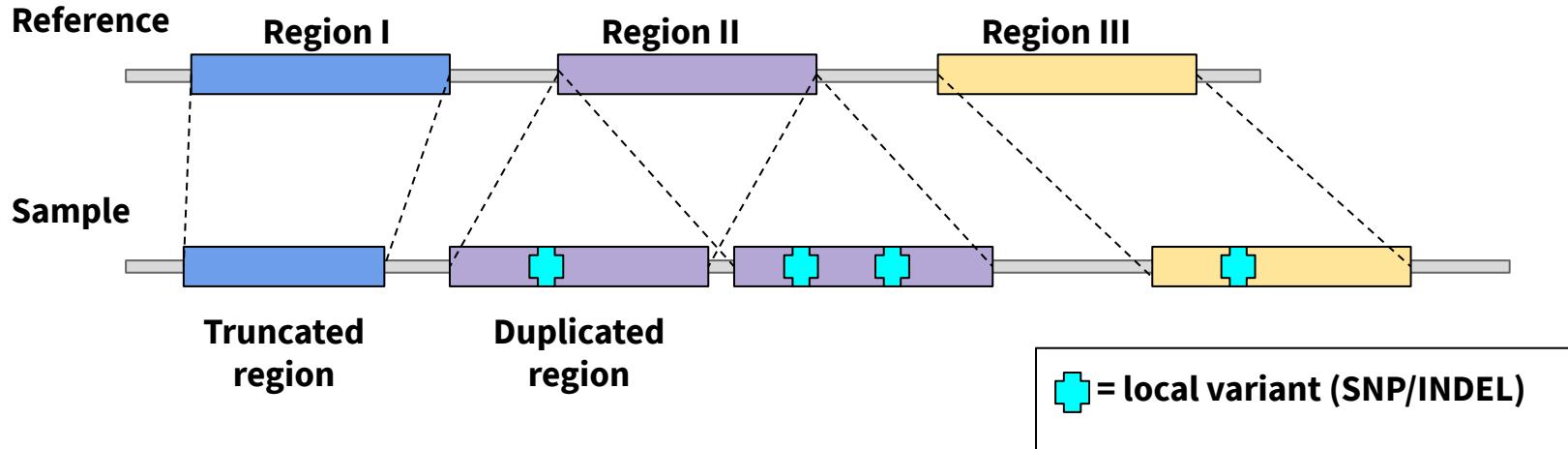
Trimmomatic: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

# Map DNA to reference genome: BWA



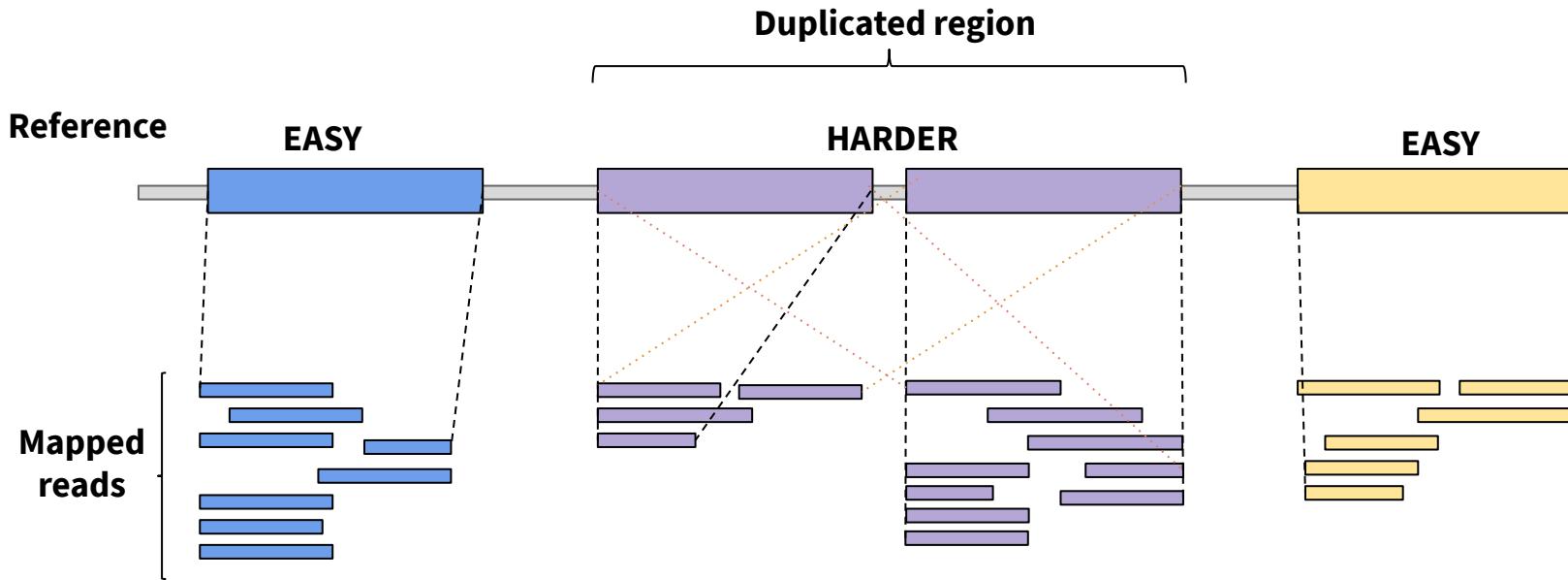
<https://software.broadinstitute.org/gatk/best-practices/workflow?id=11165>

# Map DNA to reference genome: BWA

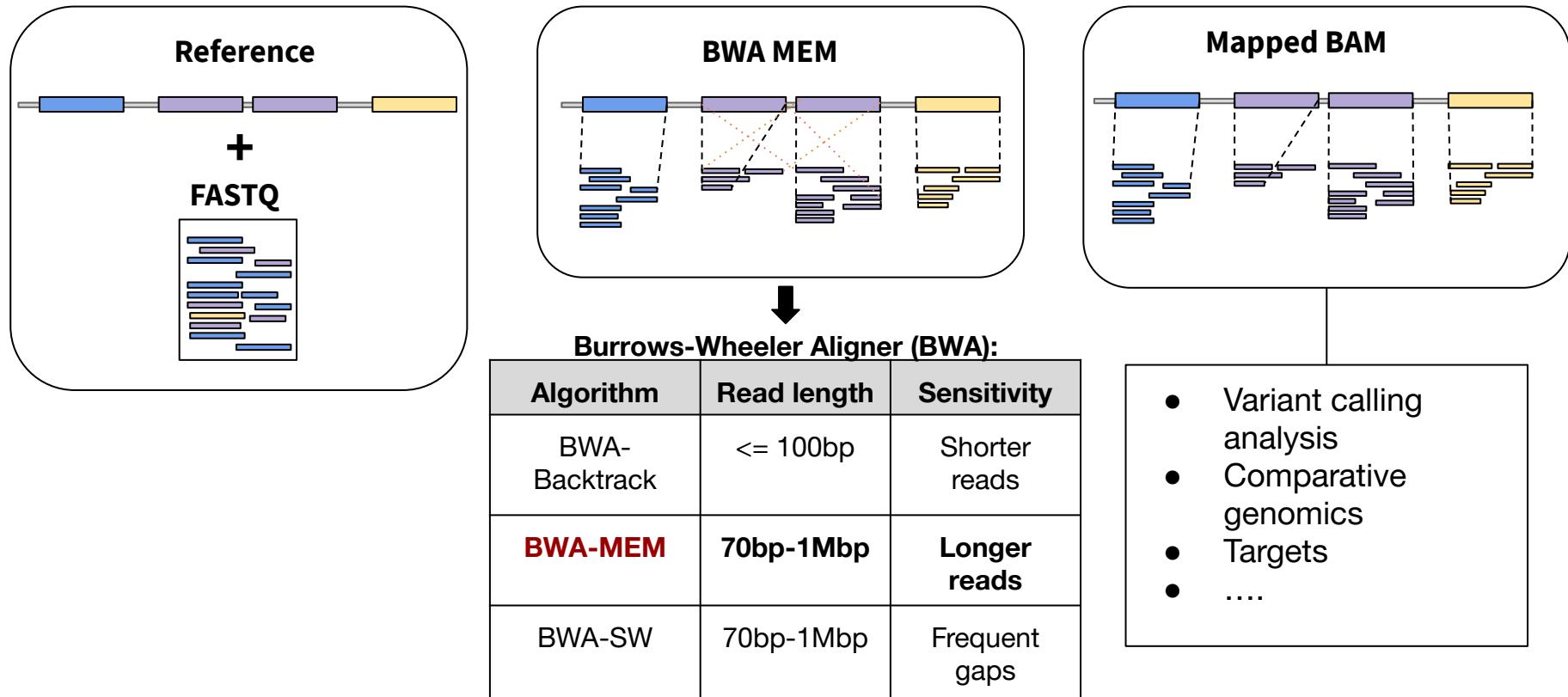


Our sample is **fragmented**, therefore **we need to put all the pieces together!**

# Map DNA to reference genome: BWA



# Map DNA to reference genome: BWA-MEM



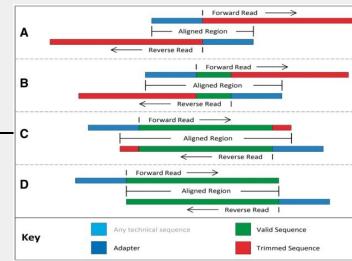
# Toolbox for mapping NGS data

## Quality check(QC):

## **FASTQC**



# Trimmomatic



Map and refine/clean alignments:

**BWA**

```

Read: ATCAGCATC

ALT ctg 1: TGAAA---CGAATGAAATGGTCATCAGCATCGACTAGTCACAT
            (novel int) (novel int) (novel int) (novel int)
Chromosome: CGCTACATGATAGCAATCAGCATCTGCTGCCTGATCTGATCCTGATC
ALT ctg 2: TGATAGCAATCAGCATCTGTCGAATCgcCAGcGACTGTCACAT

4 potential hits: ATCAGCATC > ATCgCAGTC > ATCgcCATC > ATCgcCAGC
2 hit groups: (ATCAGCATC, ATCgCAGTC) (ATCgcCATC, ATCgcCAGC)
Hits considered in mapQ: ATCAGCATC and ATCgCAGTC (best from each group)

In the output SAM: ATCgCAGTC as the primary SAM line with mapQ=0
ATCAGCATC as a supplementary line with mapQ>0
ATCgcCAGc as a supplementary line with mapQ>0
ATCgcCATC in an XA tag, not as a separate line

```

GATK

**Picard**

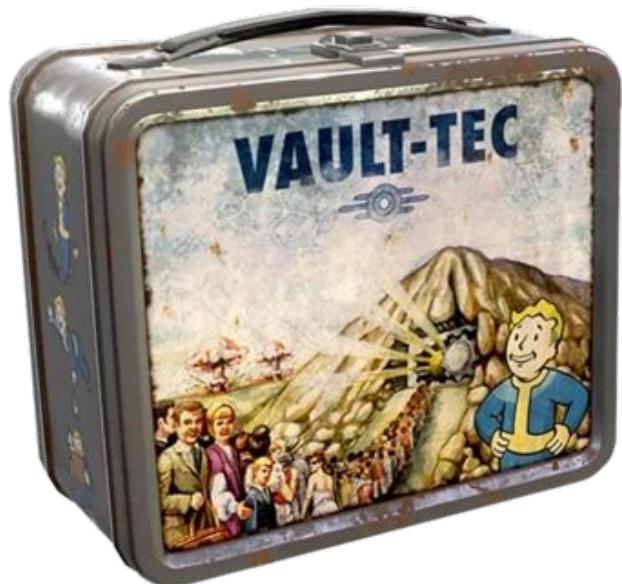
build passion



# Hands-on session

1. **Please, go here and follow the instructions:**  
<https://github.com/carlalbc>
2. [https://github.com/carlalbc/BIO634\\_2021/](https://github.com/carlalbc/BIO634_2021/)
3. **Day1 DataQC and mapping.md**

**~~ Lunch break from 12:00 to 13:30 ~~**



**Enjoy! See you in a bit! :)**