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

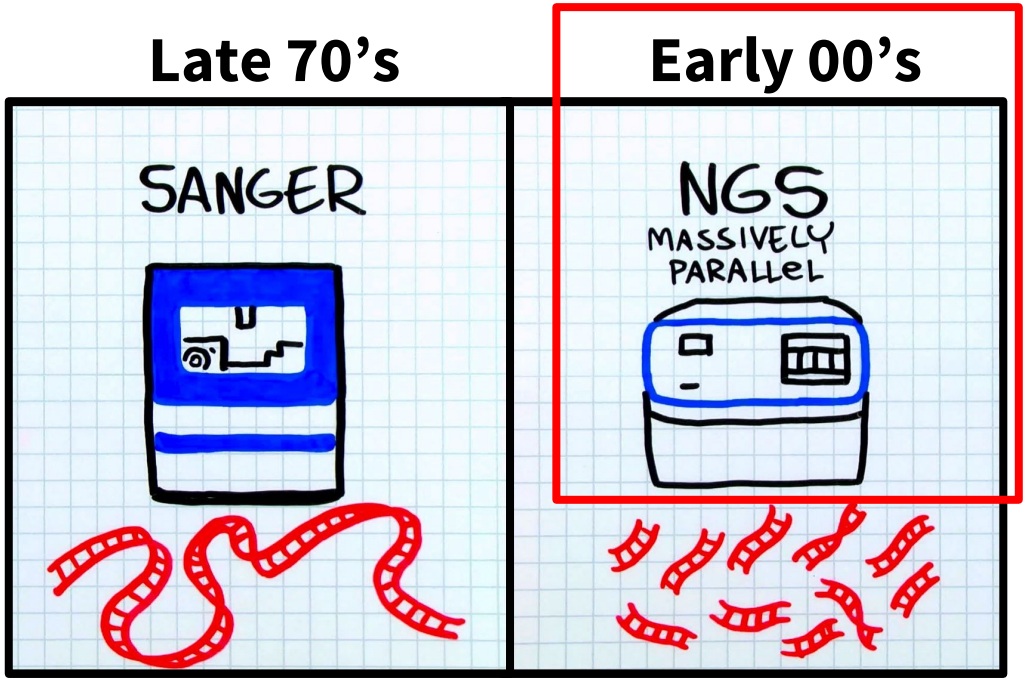
September 17-18th, 2018

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

University of Zurich[UZH]

What is sequencing?

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

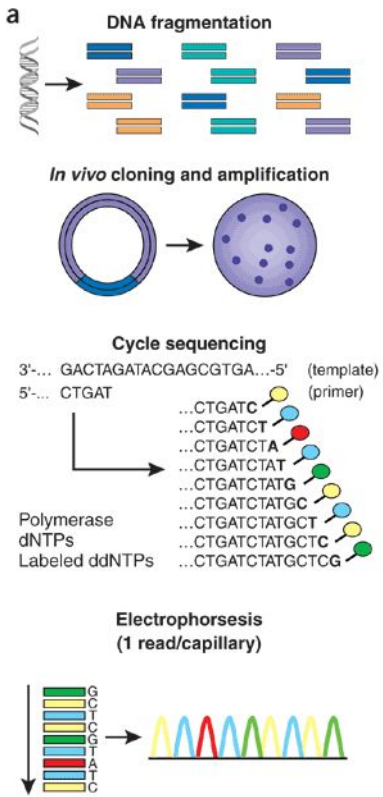


NGS is much more efficient

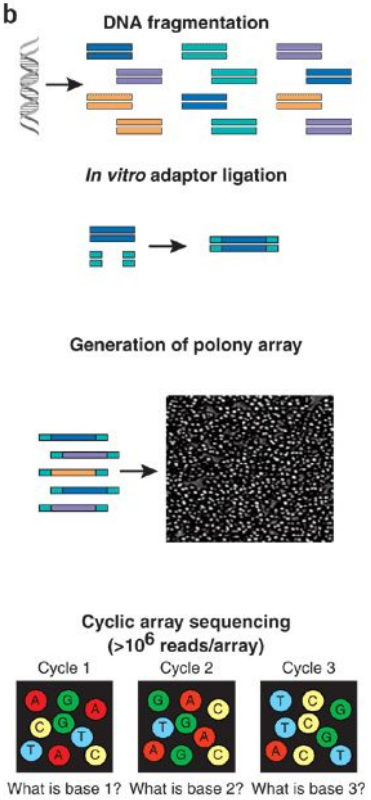
Shendure and Hanlee. 2008, Nat Biotechnol.

- **Few** targets
- Cost effective
- Familiar

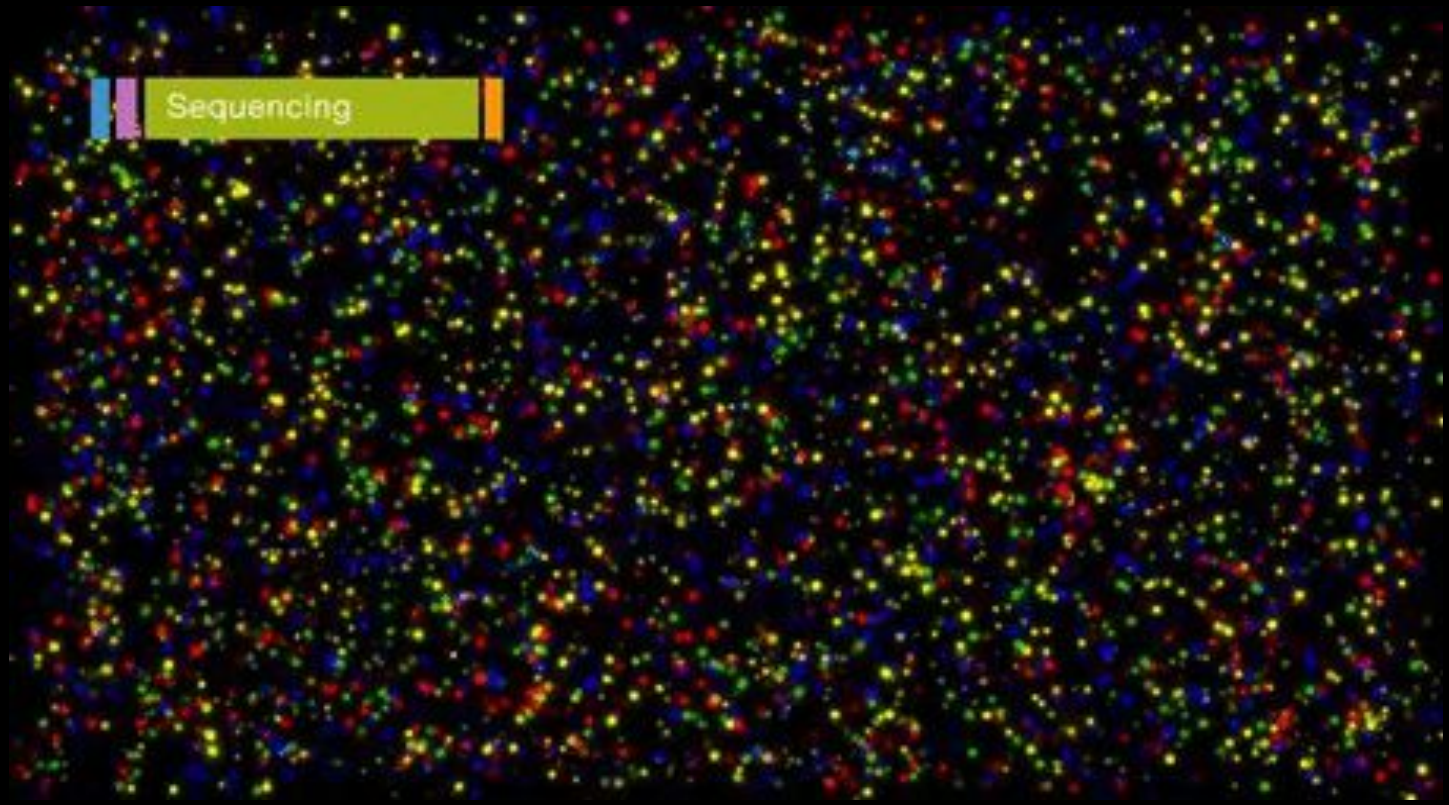
Sanger



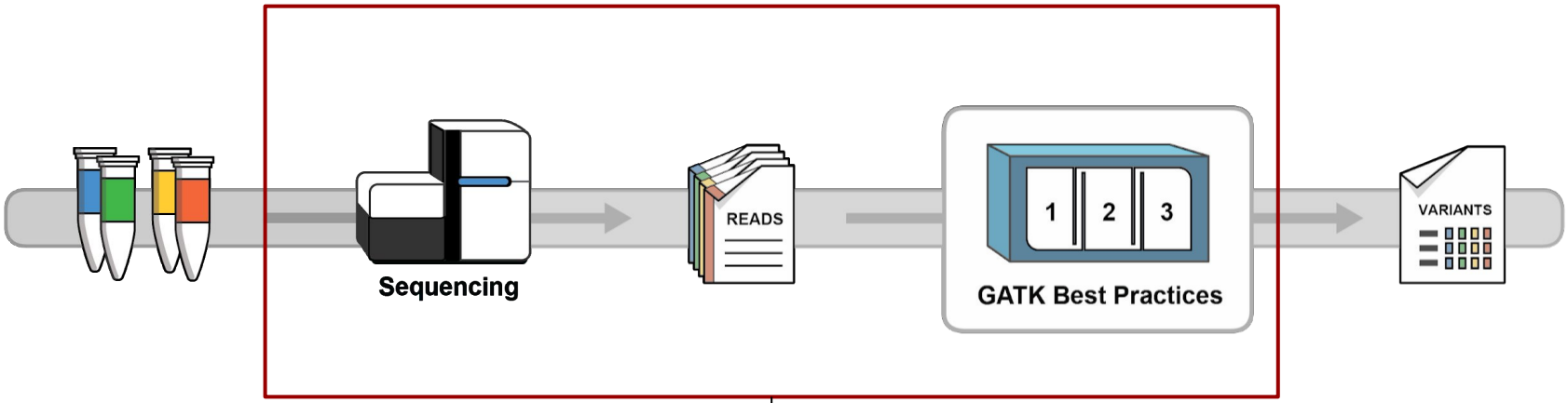
NGS (Illumina)



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


Sequencing

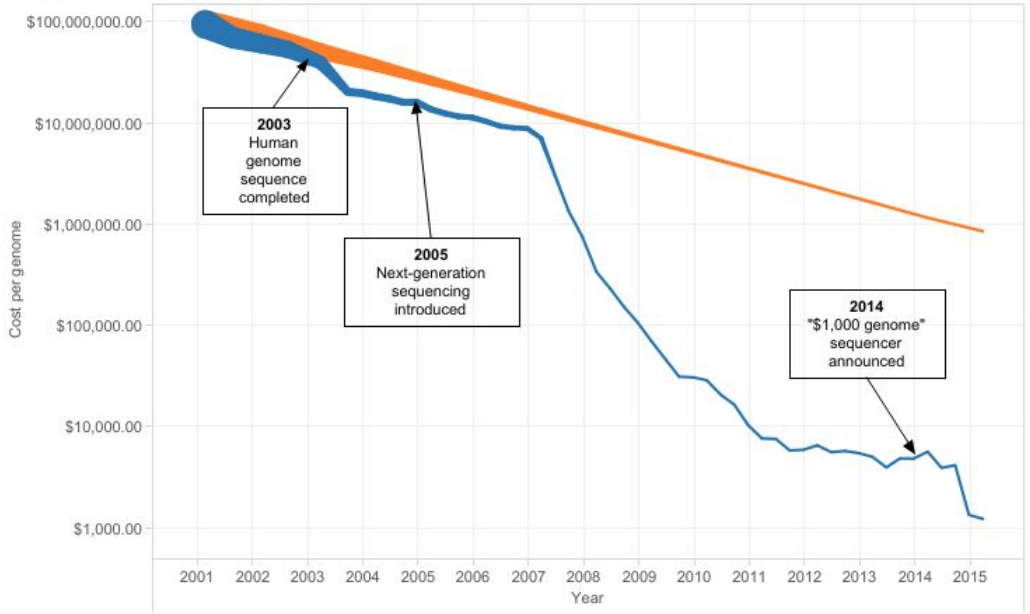
Sequencing and data analysis



1. **SEQUENCING**
2. **QUALITY CHECK**
3. **MAPPING**

Source: GATK webpage

Reduction of sequencing cost over time

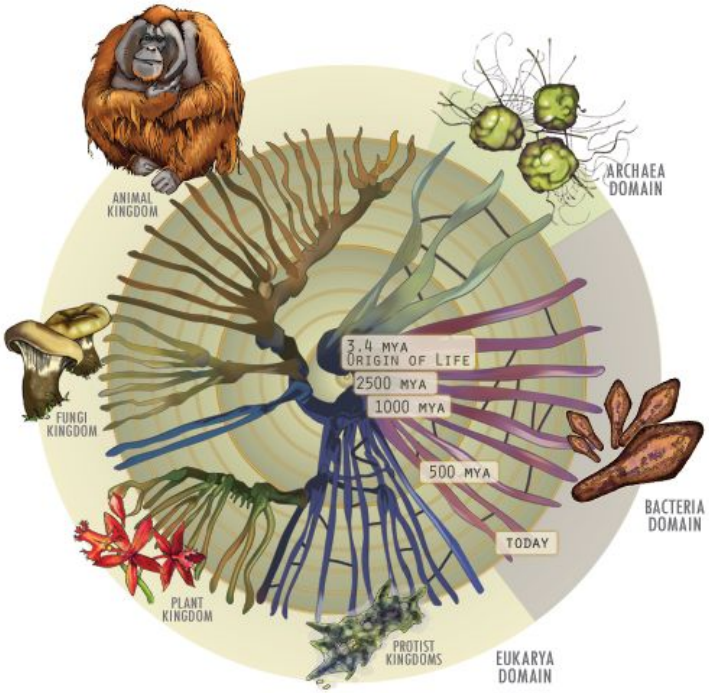


Decline in real costs compared to expected declines based on Moore's Law.
Trend line: Cost per human genome. Line width: Cost per megabase (Mb)
(*Data: NHGRI* https://www.genome.gov/27541954/dna-sequencing-costs-data/)

$3,000,000,000	**2003** Human Genome Project
$20,000,000	**2006** 1st individual genome
$2,000,000	**2007** 1st NGS Genome
$200,000	**2008** 1st 30x genome
$10,000	**2010** 1st sub-10K genome
$1,000	**2014** 1st $1,000 genome
$100	**2017** 1st $100 genome

¢1? | 2XXX 1st **¢1** genome?

Many species genomes are available



Consortia all over the world

1000 genomes

GENCODE

GENOMES 25 PROJECT

UK 10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE

VGP VERTEBRATE GENOMES PROJECT

DIGITAL NOAH'S ARK GENOME LIBRARY

Microbial Genomes

10,000 Salmonella GENOMES

International Cancer Genome Consortium

PCAWG PanCancer Analysis OF WHOLE GENOMES

PROJECT IWGSC

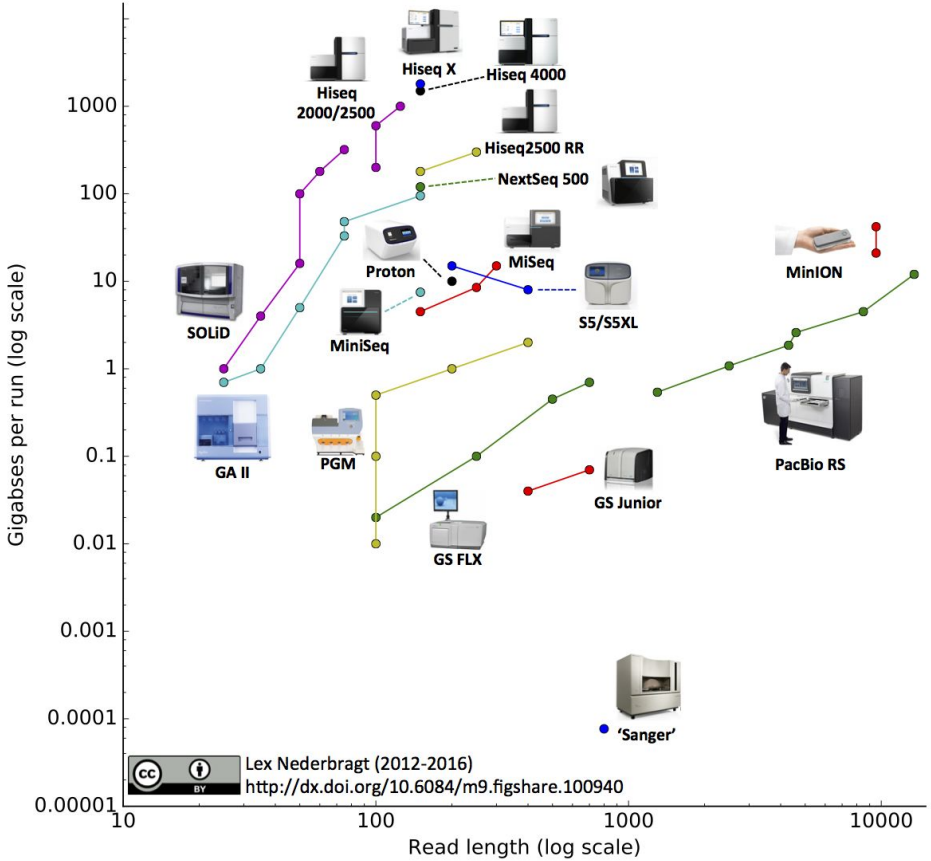
Most used sequencing platforms today

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

Choose wisely

http://dnatech.genomecenter.ucdavis.edu/next-generation-sequencing/

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**

Choosing between platforms

Table 1.

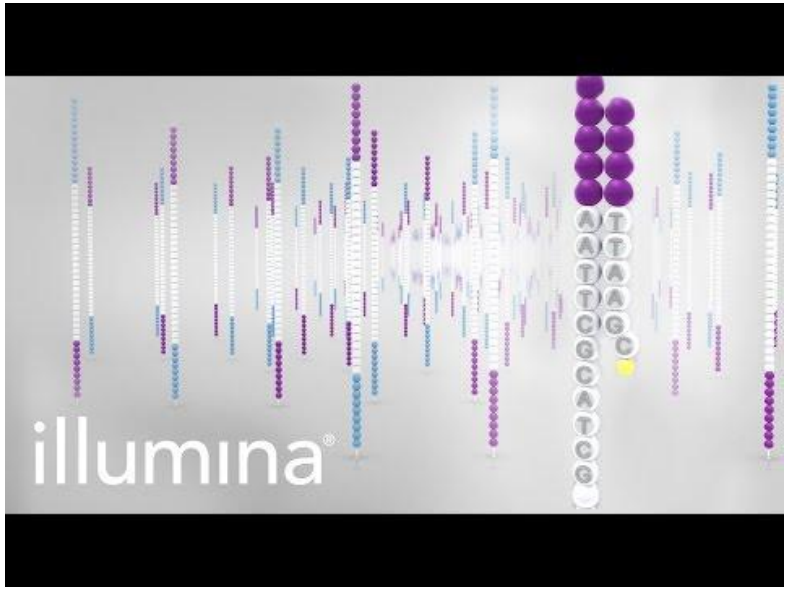
Comparison of PacBio sequencing platforms to two current industry standards

Platform	Read length	Number reads	Error rate	Run rime
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

company website (www.pacb.com and www.illumina.com) queries on 14 November 2017.

Current and upcoming technologies

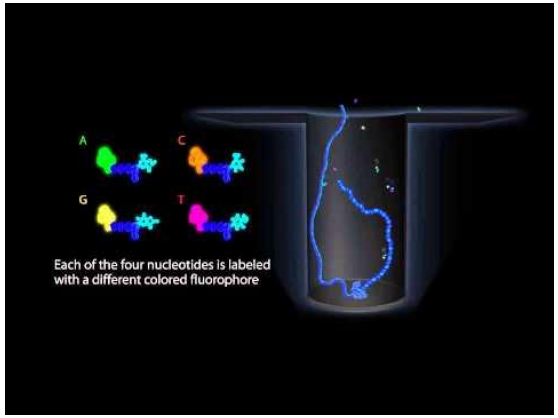
Illumina (HiSeq)



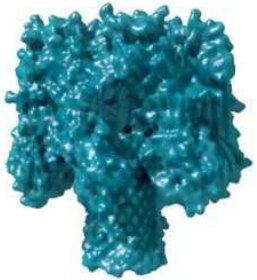
Links:

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

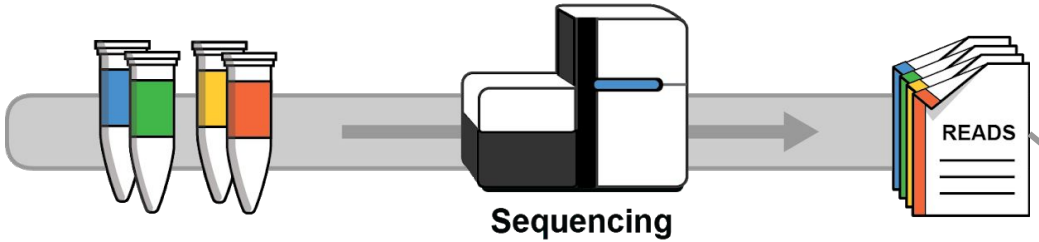
PacBio (SMRT seq)



Oxford Nanopore (ONT)



There are different file formats

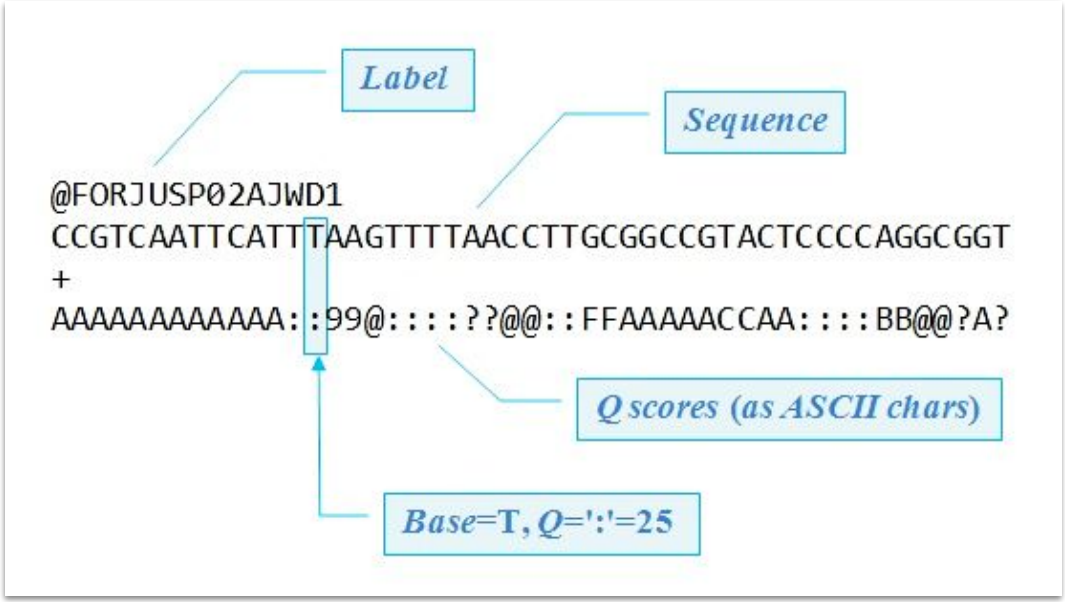


TEXT FILES IN DIFFERENT FORMATS

There are different file 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 is typically used

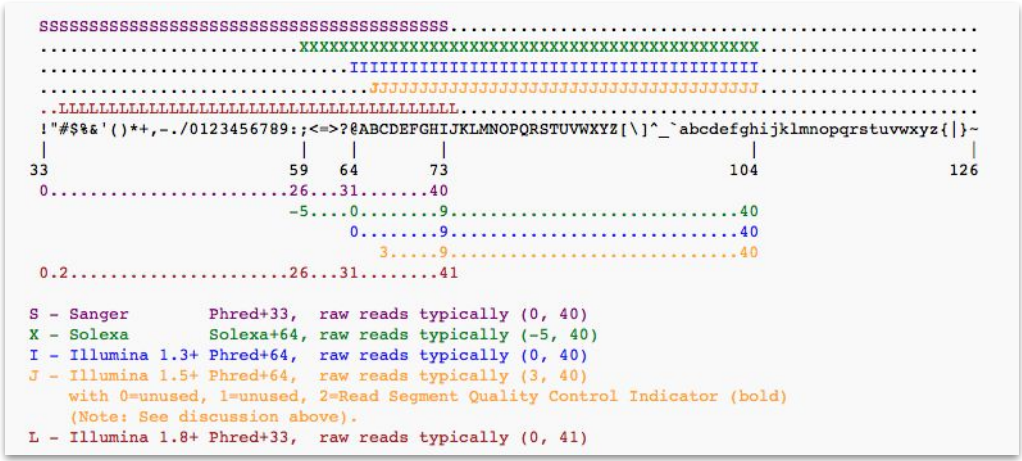


FASTQ: FASTA with quality scores (ASCII to Phred Q-scores)

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%

- $Q = -10 \log_{10} P$ <=> $P = 10^{-Q/10}$
  - 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) :

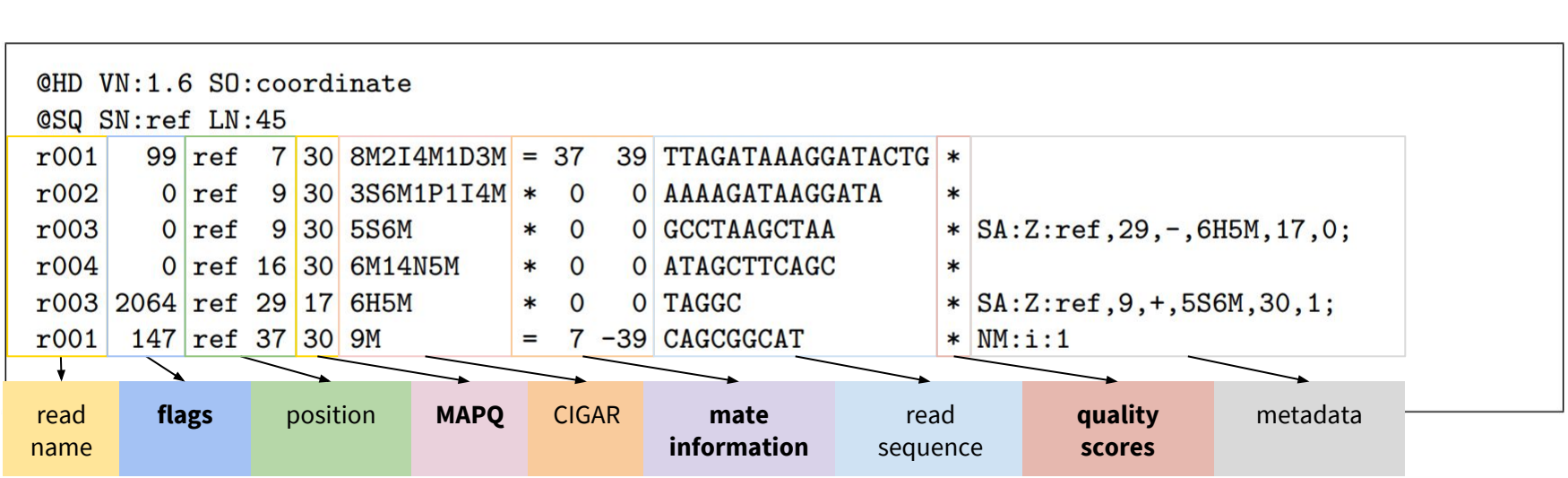
```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC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

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

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

File formats

SAM(Sequence Alignment/Map) :



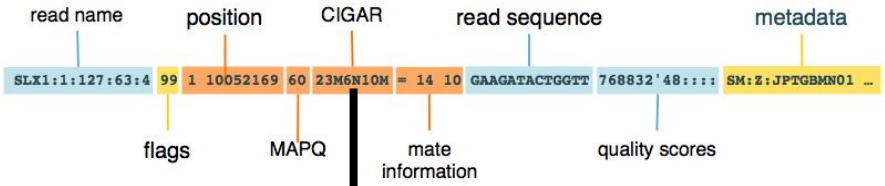
SAMTOOLS: http://www.htslib.org/doc/sam.html

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)

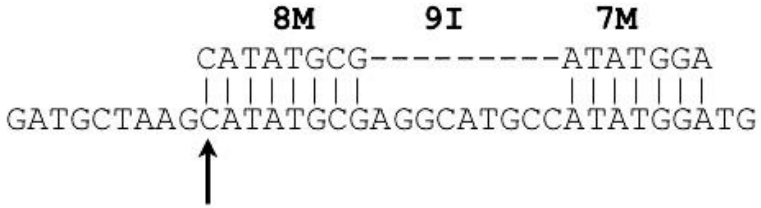


CIGAR

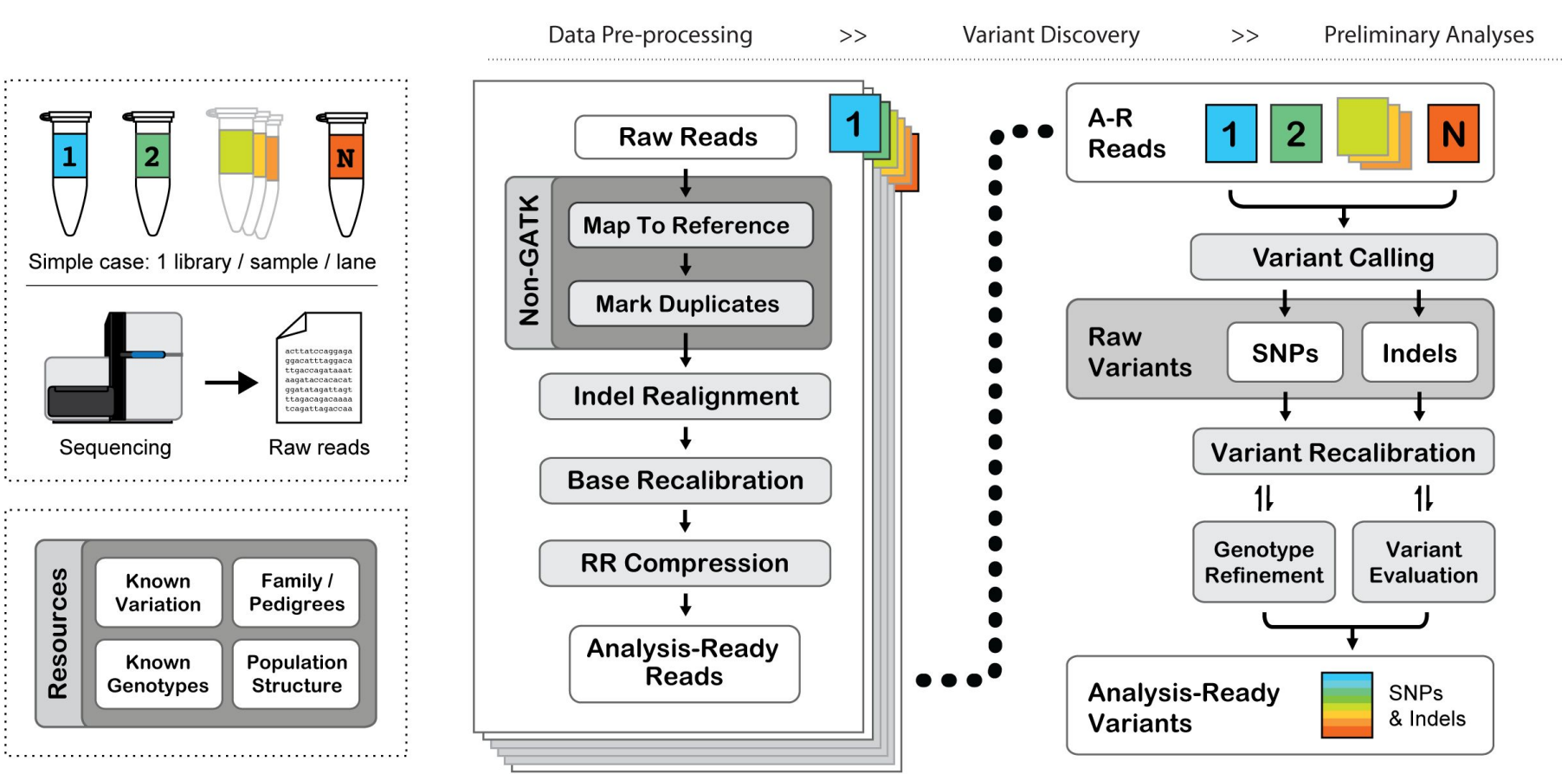
op	Description
M	Alignment match (can be a sequence match or mismatch)
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent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

Show alignment result simply
8M9I7M

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

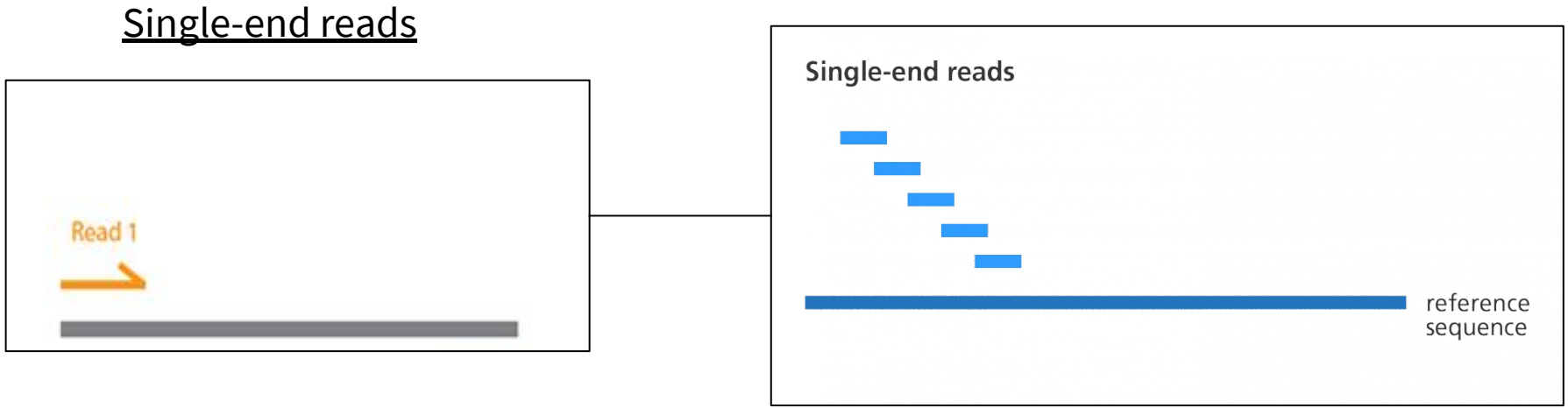


4th line “POS” indicates this position.



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

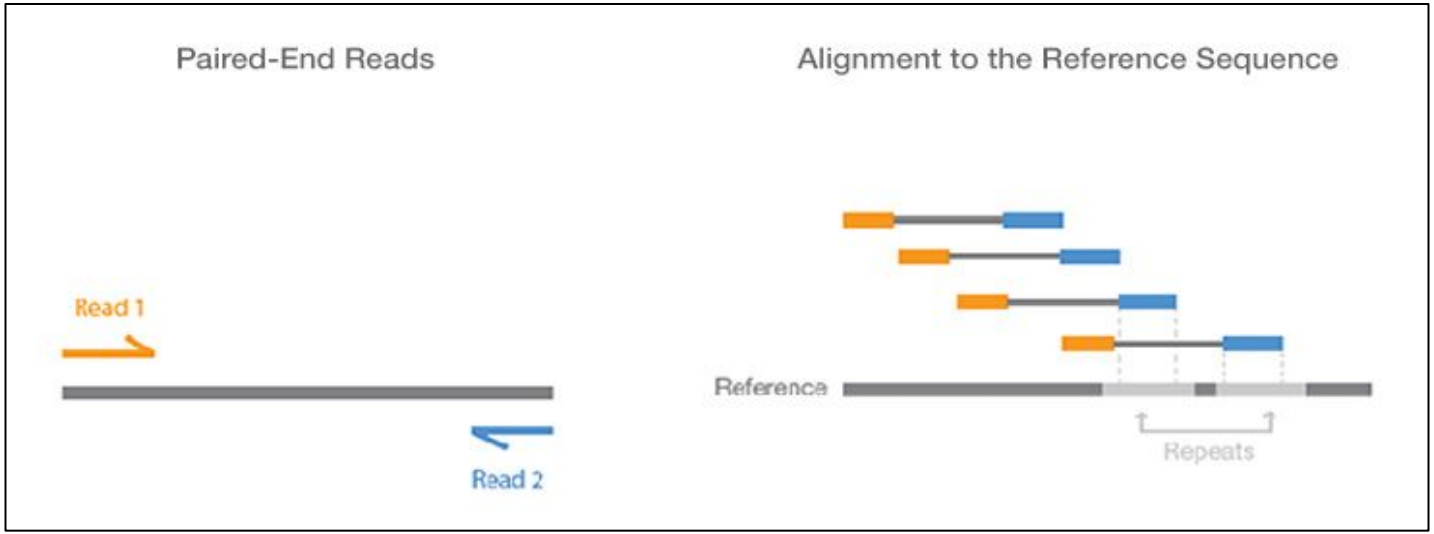
Single, Paired-End, Mate pair reads



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.**

Source: www.illumina.com

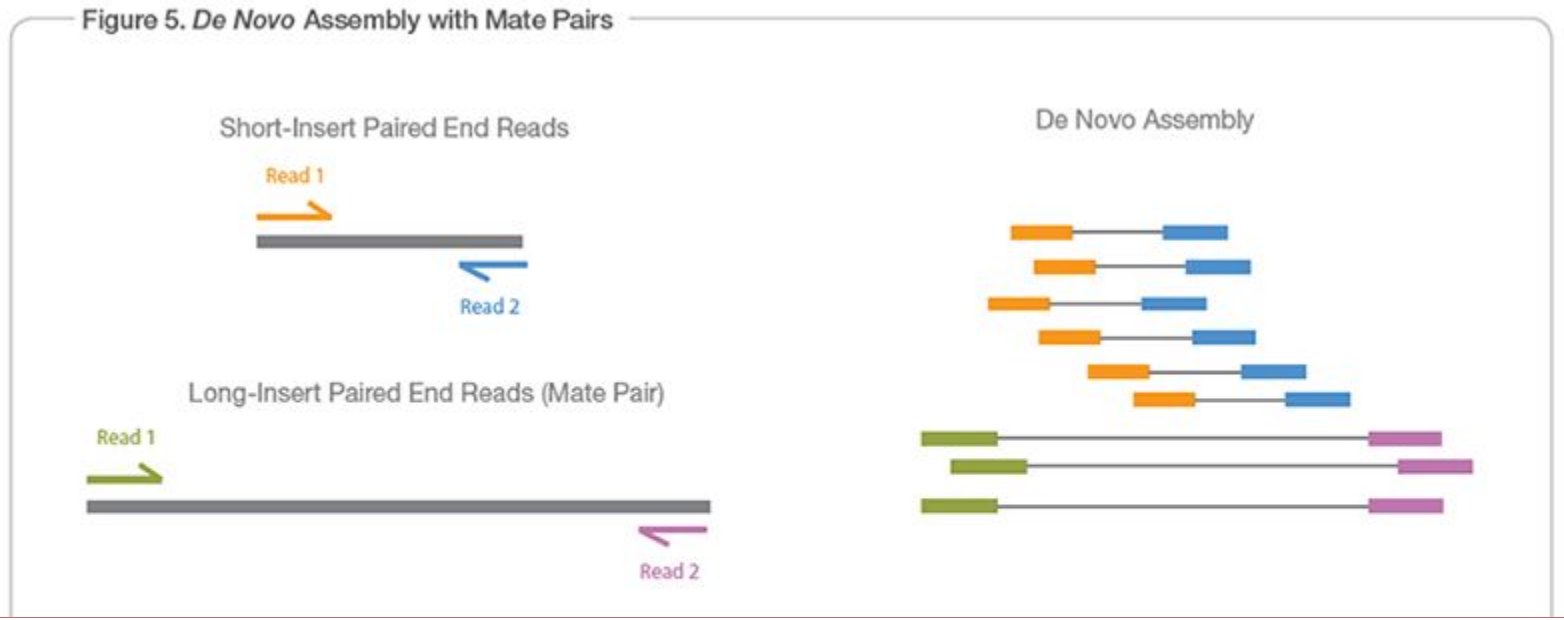
Single, Paired-End, Mate pair reads



Paired-end: 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.**

Source: www.illumina.com

Single, Paired-End, Mate pair reads



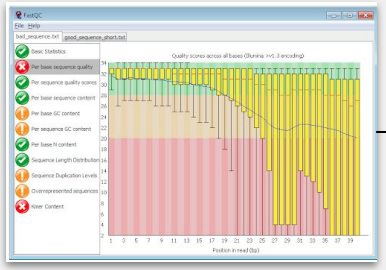
Mate-pair: Each read is **two sequences from each end of the same DNA fragment**, but the distance between the reads on the original genome sequence is much longer, e.g. **3000-10000 bp**

Source: www.illumina.com

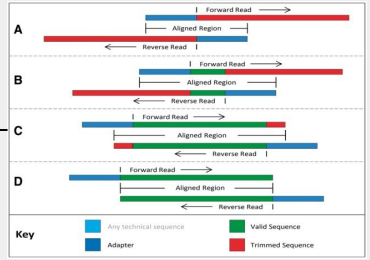
Toolbox for mapping NGS data

Quality check(QC) and cleaning:

FASTQC

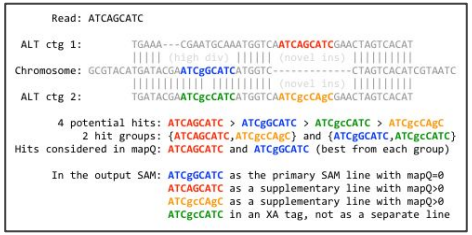


Trimmomatic/Cutadapt



Map and remove duplicates:

BWA/Tophat/STAR



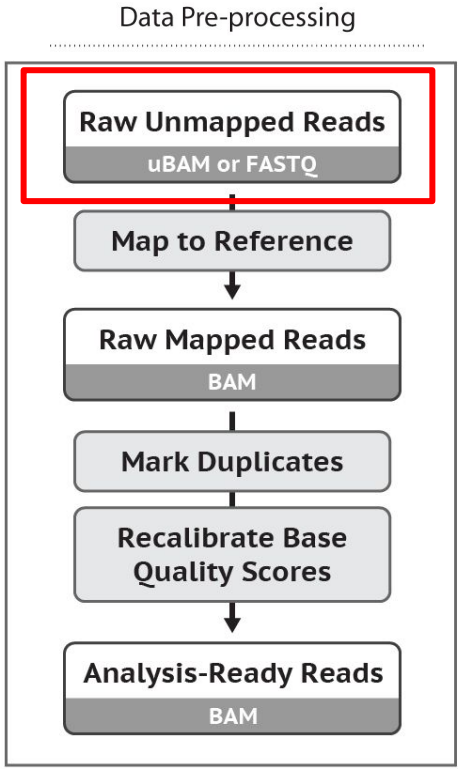
GATK and Picard Tools

Picard

build passing

gatk

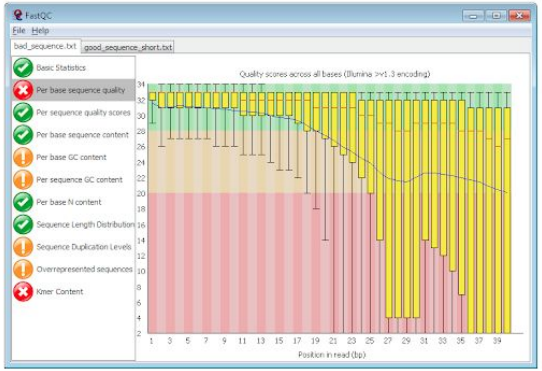
Quality check: FASTQC



FastQC

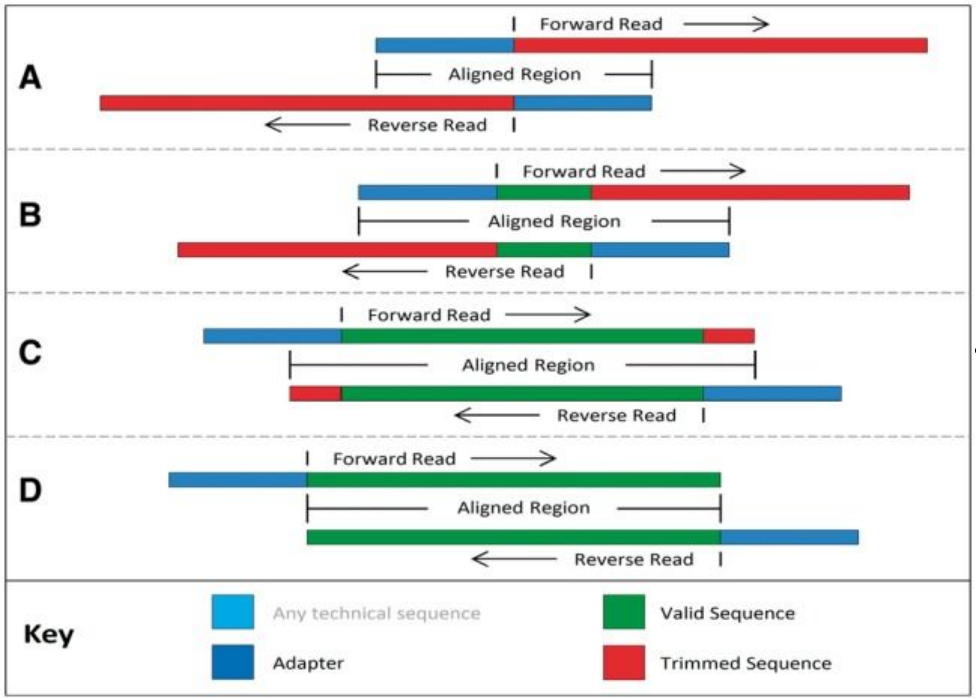
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now



FASTQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

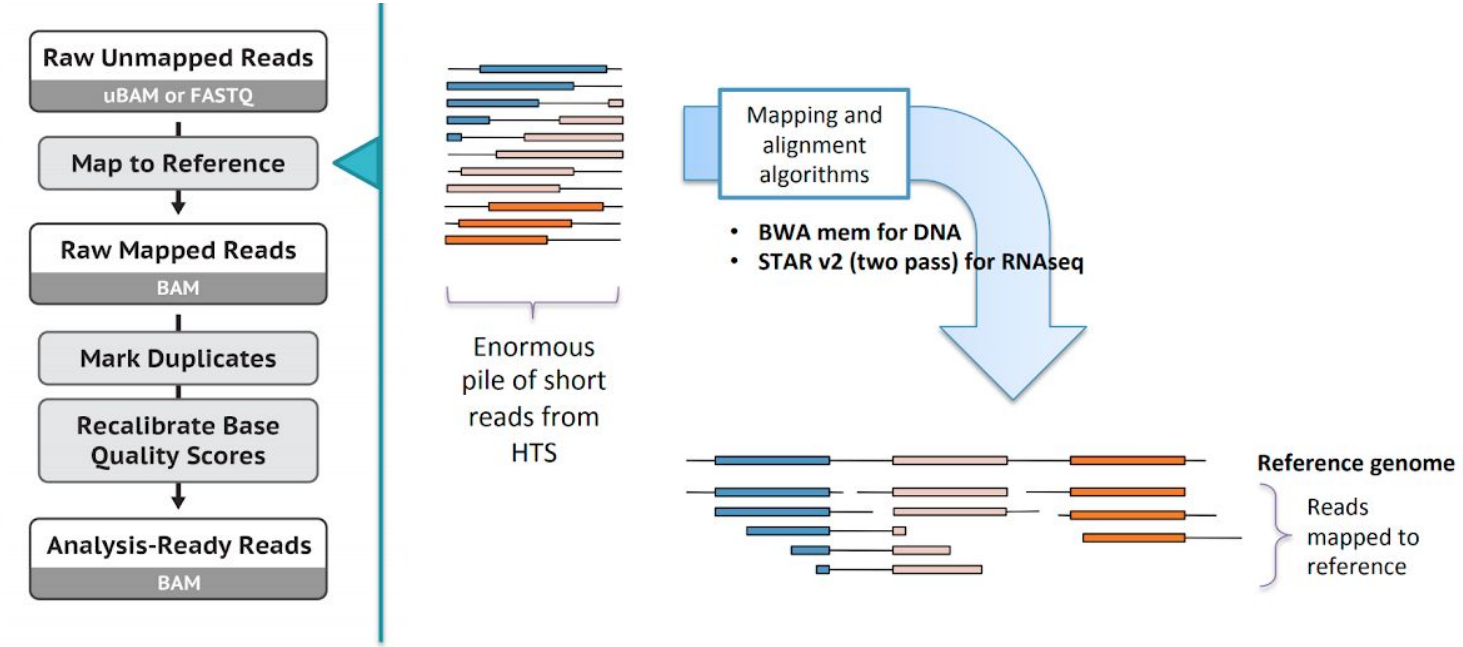
Clean-up: Trimmomatic



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

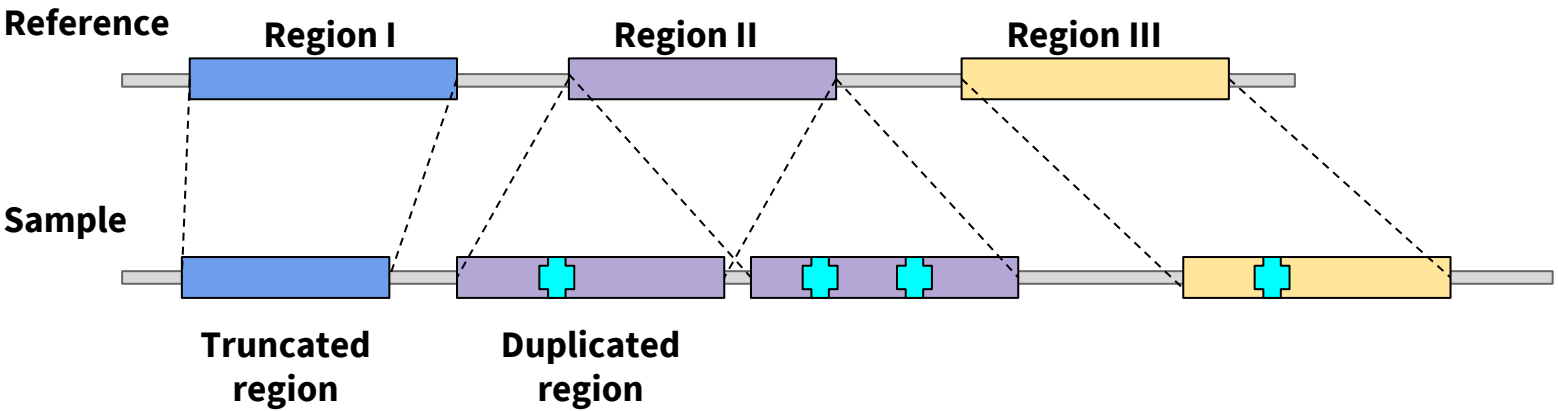
Trimmomatic: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Map DNA to reference genome: BWA



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

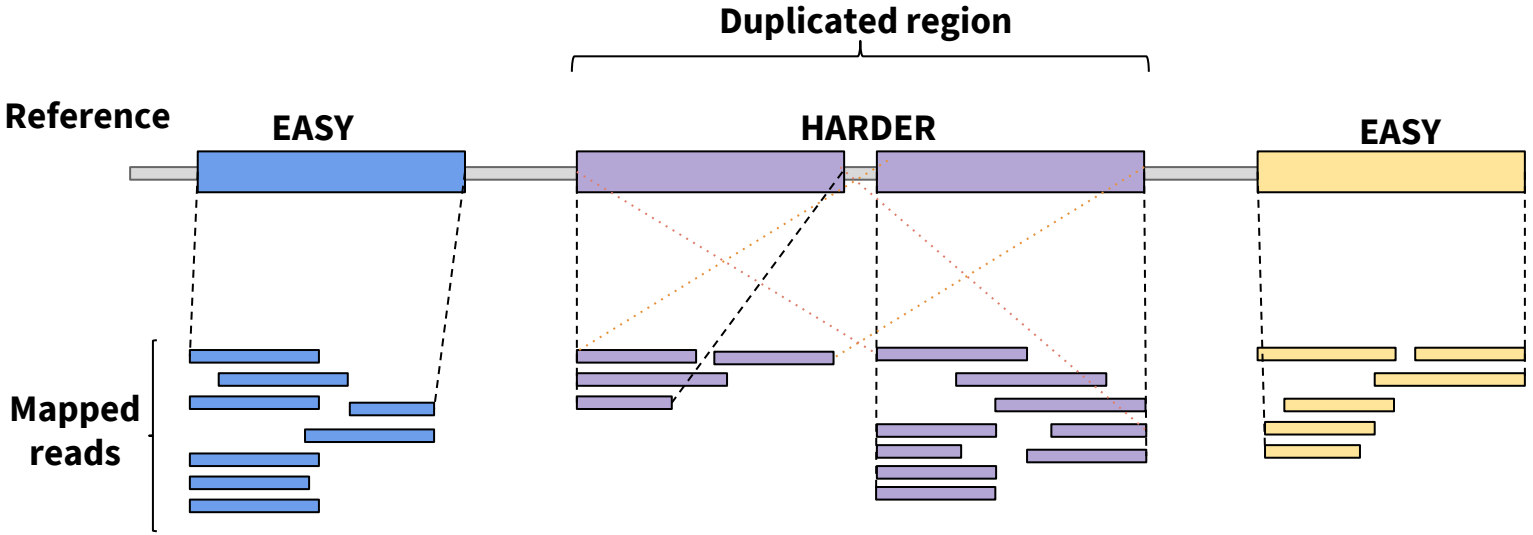
Map DNA to reference genome: BWA



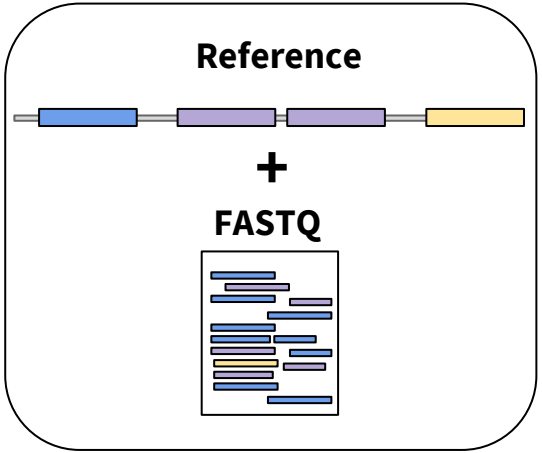
= local variant (SNP/INDEL)

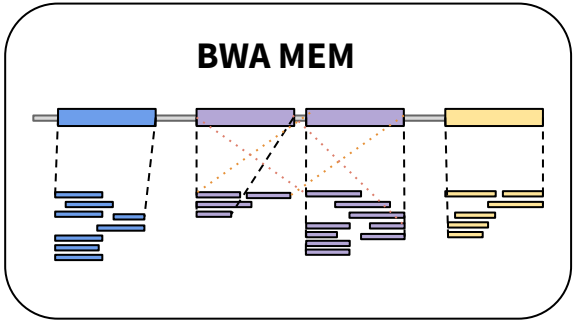
Our sample is **fragmented**, therefore **we need to put all the pieces together**!

Map DNA to reference genome: BWA



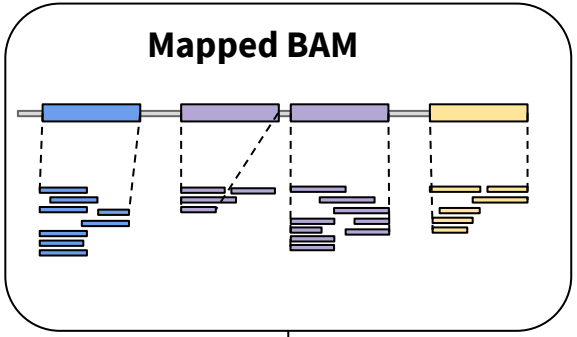
Map DNA to reference genome: BWA-MEM





Burrows-Wheeler Aligner (BWA):

Algorithm	Read length	Sensitivity
BWA-Backtrack	<= 100bp	Shorter reads
BWA-MEM	**70bp-1Mbp**	**Longer reads**
BWA-SW	70bp-1Mbp	Frequent gaps

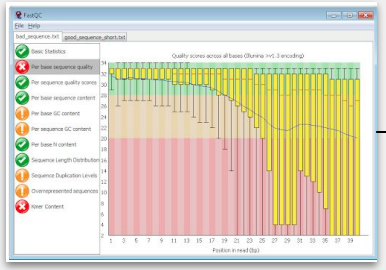


- Variant calling analysis
- Comparative genomics
- Targets
- ….

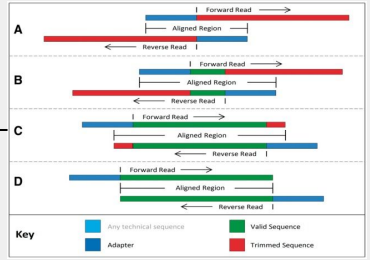
Toolbox for mapping NGS data

Quality check(QC):

FASTQC



Trimmomatic



Map and refine/clean alignments:

BWA

```
      Read: ATCAGCATC

 ALT ctg 1:         TGAAA---CGAATGCAAATGGTCAATCAGCATCGAACTAGTCACAT
                          ||||| (high div) |||||| (novel ins) ||||||||||
Chromosome: GCGTACATGATACGAATCgGCATCATGGTC------------CTAGTCACATCGTAATC
                          |||||||||||| ||||||||| (novel ins) ||||||||||
 ALT ctg 2:         TGATACGAATCgCATCATGGTCAATCgCCAgCGAACTAGTCACAT

       4 potential hits: ATCAGCATC > ATCgGCATC > ATCgCATC > ATCgCCAgC
           2 hit groups: {ATCAGCATC,ATCgCCAgC} and {ATCgGCATC,ATCgCATC}
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
                         ATCgCCAgC as a supplementary line with mapQ>0
                         ATCgCATC in an XA tag, not as a separate line
```

GATK

Picard

build passing

gatk

Hands-on session - Part I: FastQC and Mapping

Please, go here and follow the instructions:

https://github.com/carlalbc

https://github.com/carlalbc/BIO634_2018/