imed Research Institute for Medicines

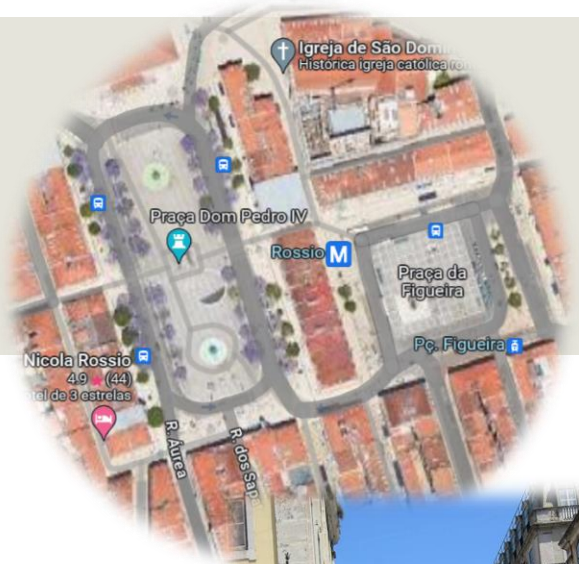
U LISBOA UNIVERSIDADE DE LISBOA

FACULDADE DE FARMÁCIA Universidade de Lisboa

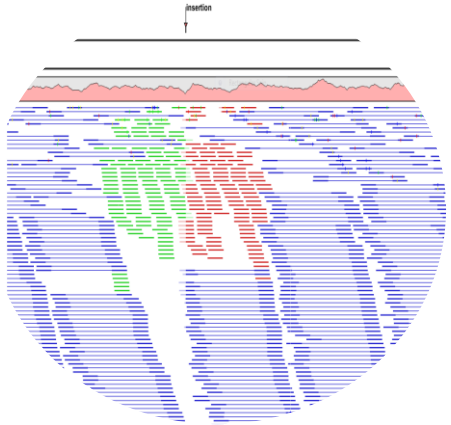
# PMB2025

## PATHOGEN MULTIOMICS AND BIOINFORMATICS

### Recife PE 2025



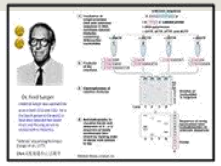

## Module 1: Mapping Sequence Data




João Perdigão

imed
Research Institute
for Medicines

# Sequencing through time…



Sanger method

1977


Human Genome Project

1990

Complete eukaryotic genome

1996

Second generation sequencer： 454 GS20

2005


Research Human Microbiome Project

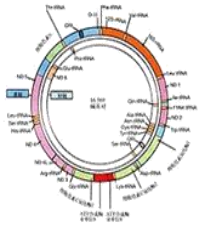
2008

Nanopore sequencing

2014

1981 1995 2001 2007 2011 2019

Human mitochondrial genome sequence



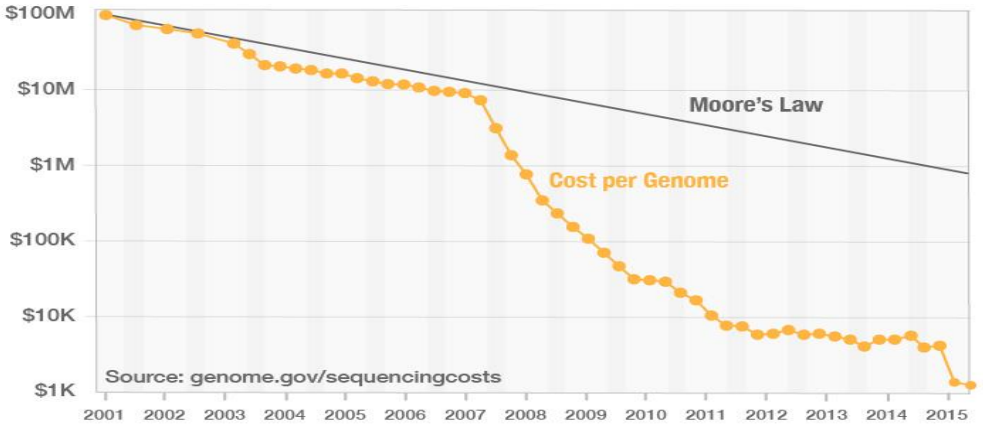
Complete cell genome

Complete the Human Genome Project

Second generation sequencer： Genetic Analyzer 2

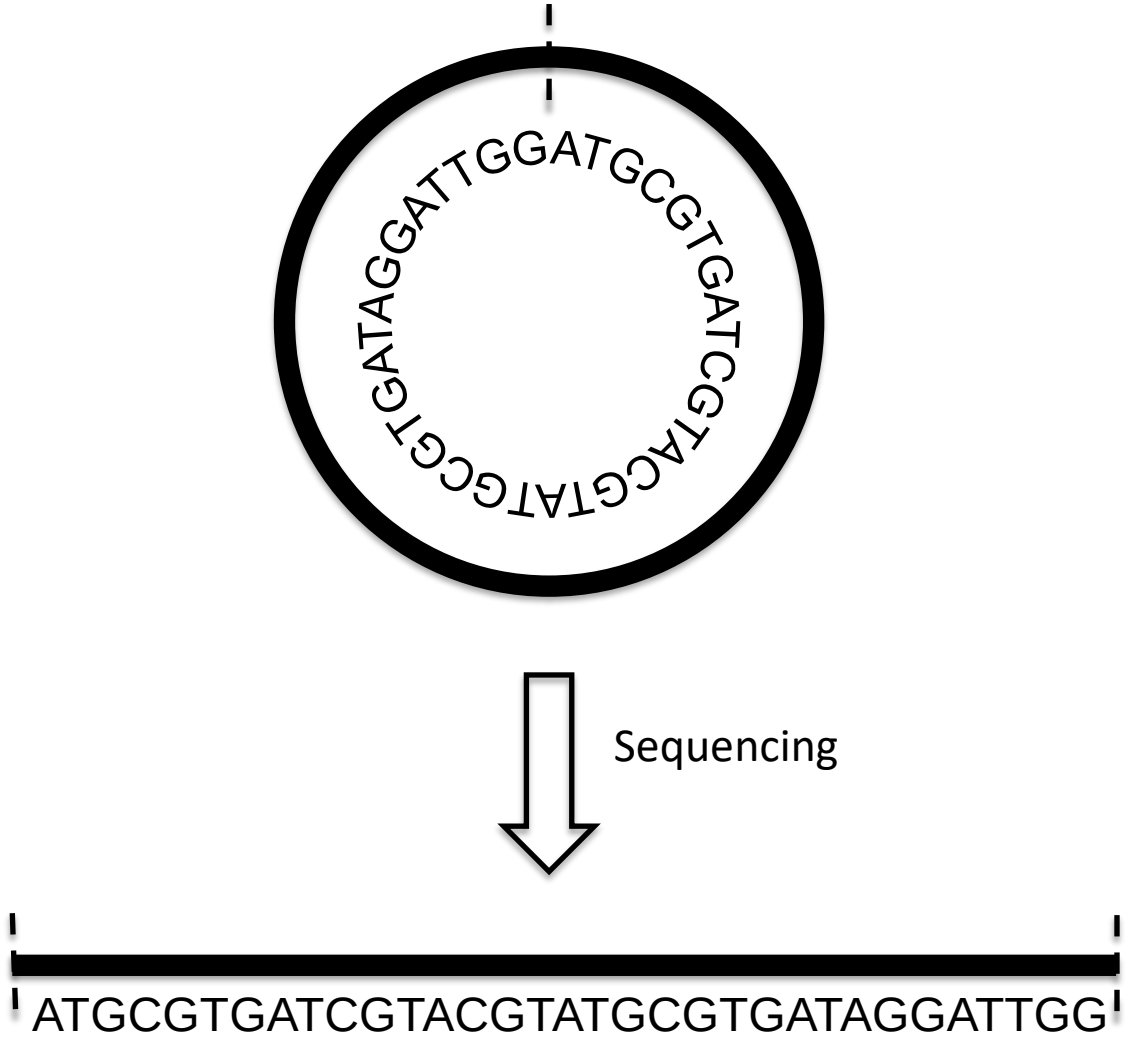
Third generation sequencer： PacBio RS

The third stage Human microbiome project



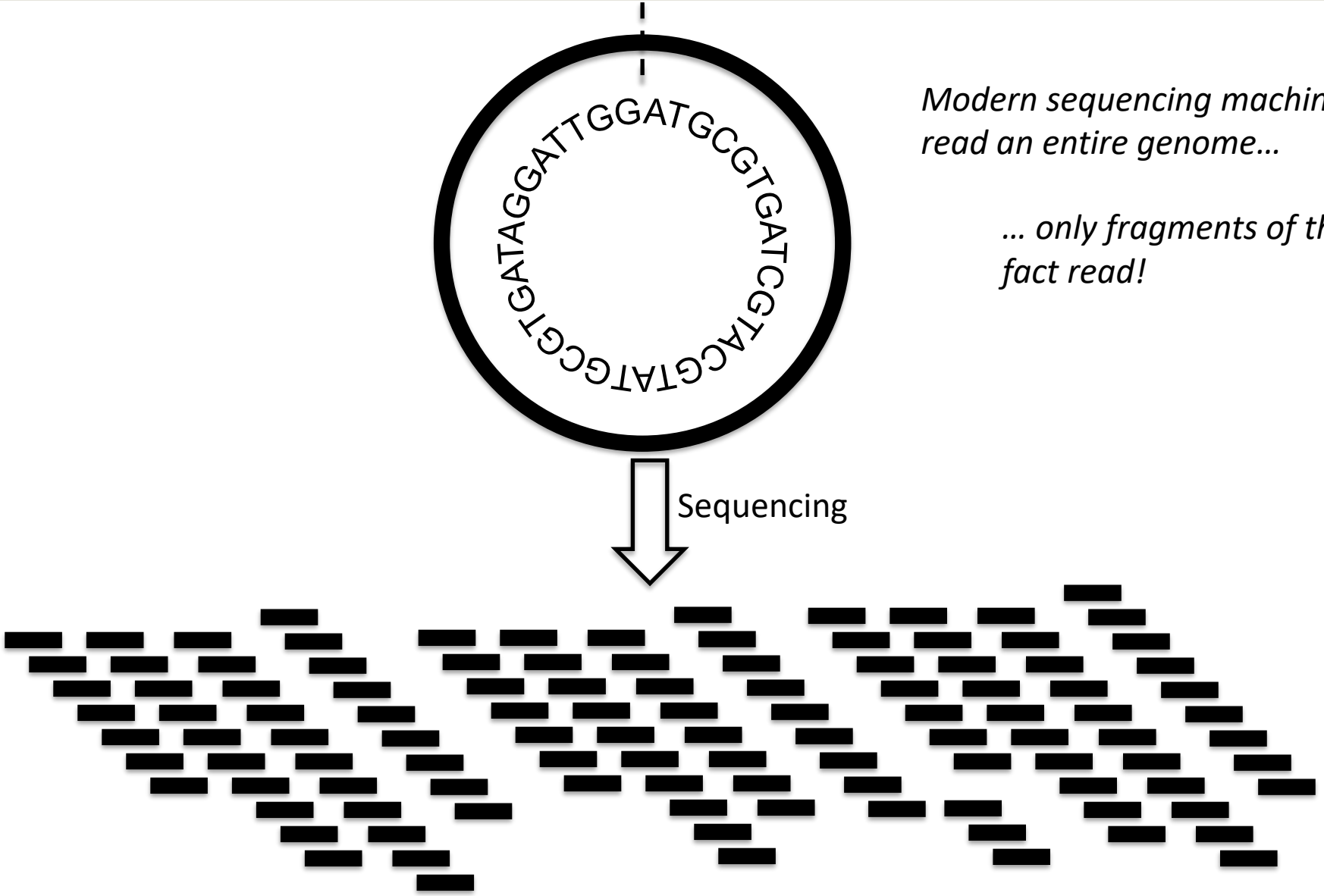
imed
Research Institute
for Medicines

# Genome Sequencing: Ideal situation…



imed
Research Institute
for Medicines

# Genome Sequencing: the hard reality…




*Modern sequencing machines cannot read an entire genome…*

*… only fragments of the genome are in fact read!*

imed Research Institute for Medicines

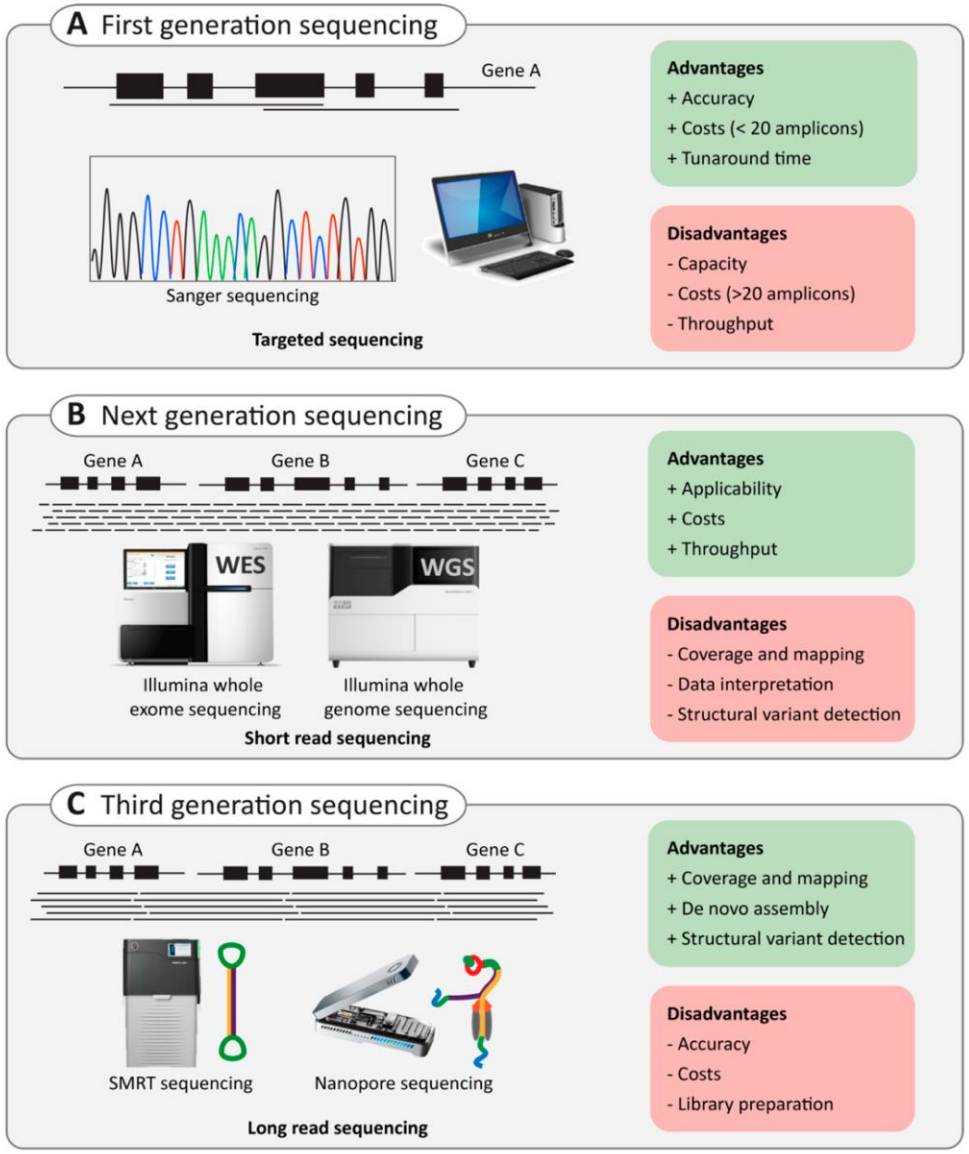
# NGS Platforms: An overview …

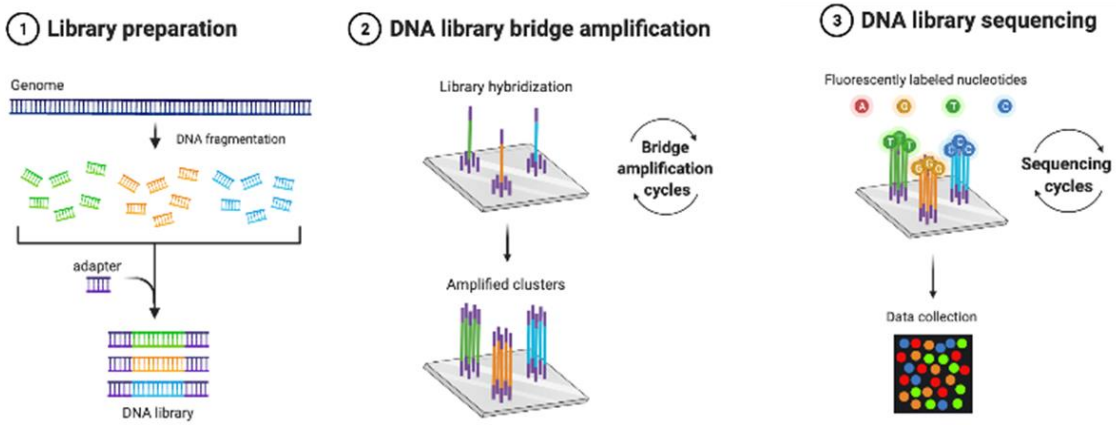
| | illumina® | ion torrent | PACIFIC BIOSCIENCES® | Oxford NANOPORE Technologies |
|---|---|---|---|---|
| Read Length (bp) | 50-300 | 200-400 | 10000-40000 | 1Mbp |
| Output (Gb) | 6000 | 0,05-1 | 0,5-1 | 5-40 |
| Cost / Million bp (USD) | 0,05-0,15 | 1 | 0,13-0,60 | variable |
| Accuracy | 99.9% | 99.6% | 87% | 92-97% |
| Time per run | 1-11d | 2h | 30min-20h | 1min-48h |

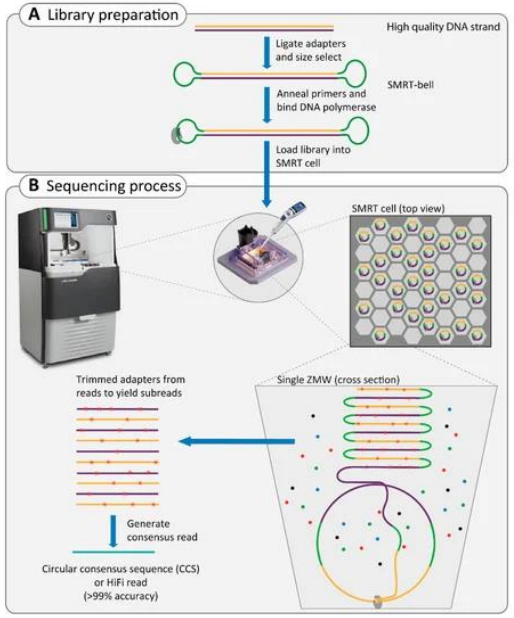
**Sanger Cost per MB: 2400USD**

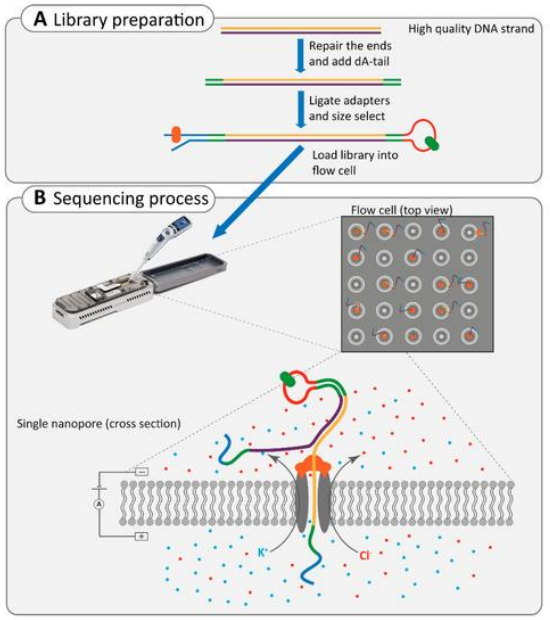
imed Research Institute for Medicines

# Next Generation Sequencing




*de Bruijn et al 2021*

# Illumina: Sequencing-by-synthesis

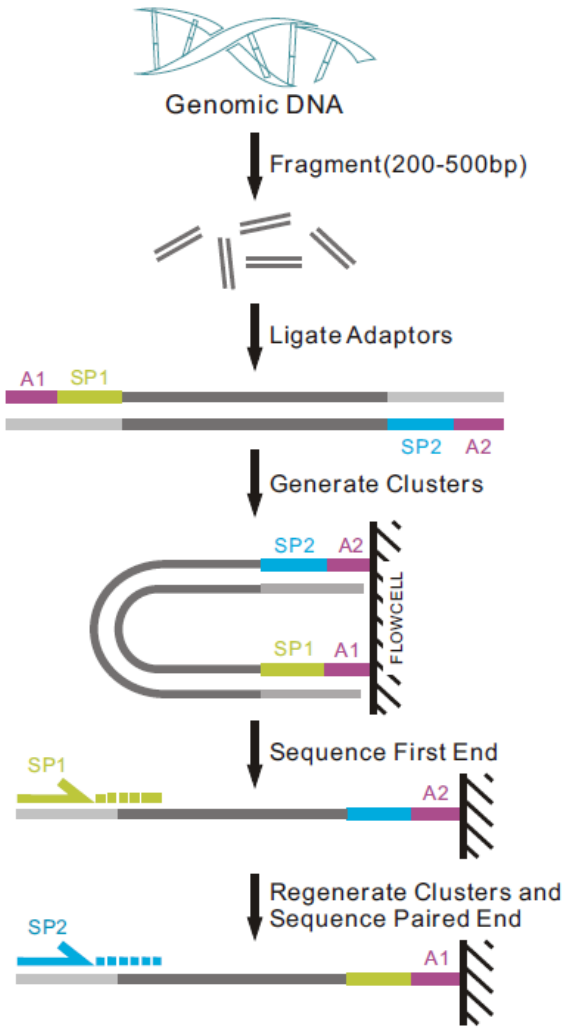
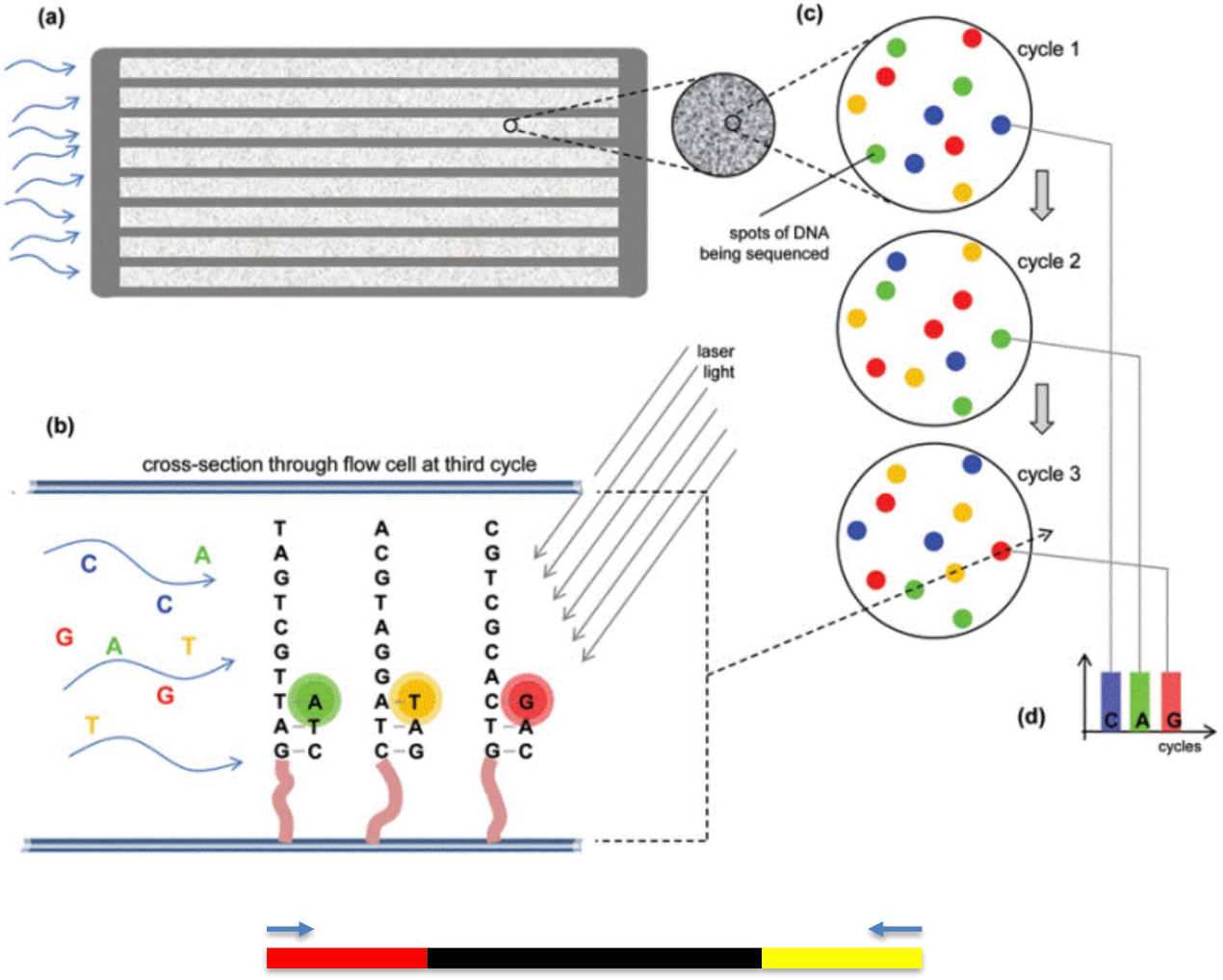
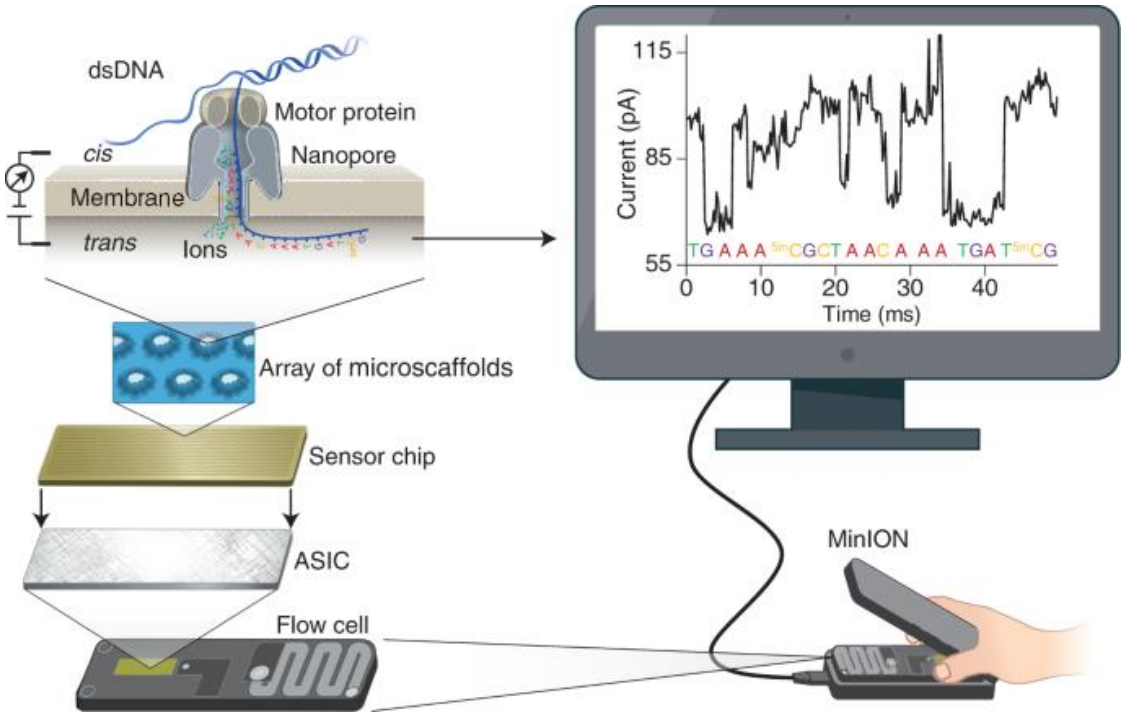


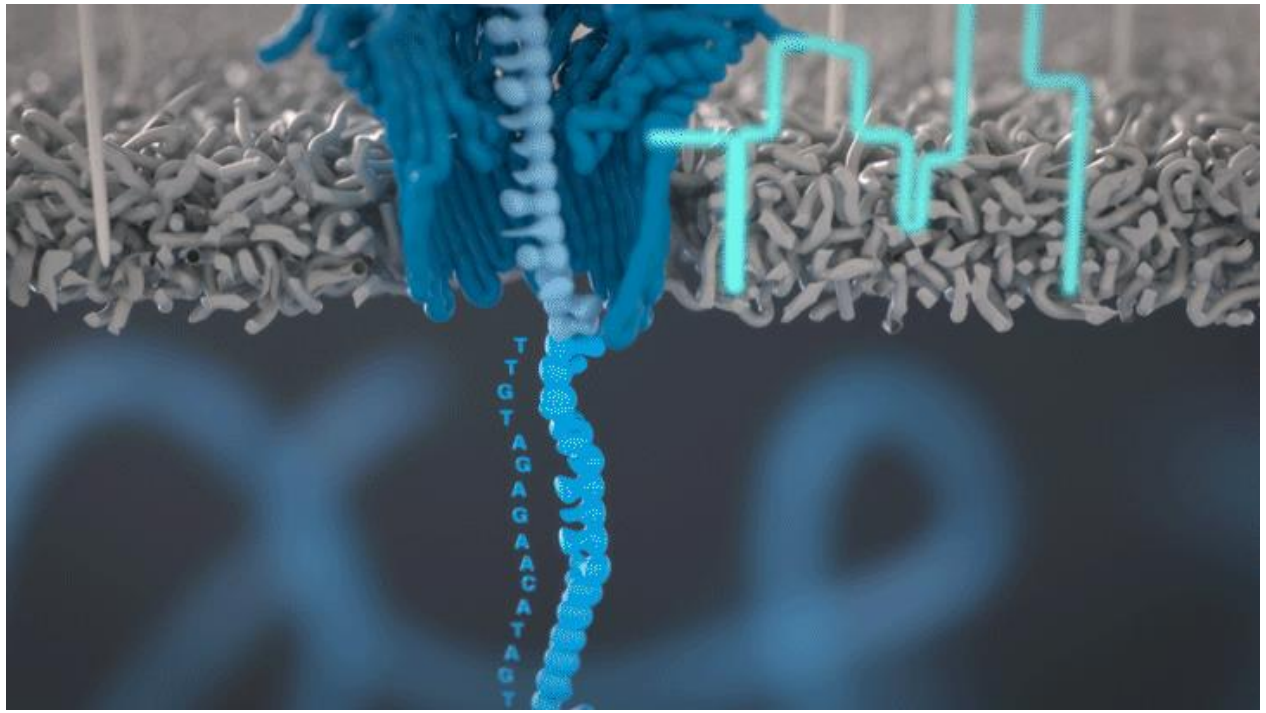

Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

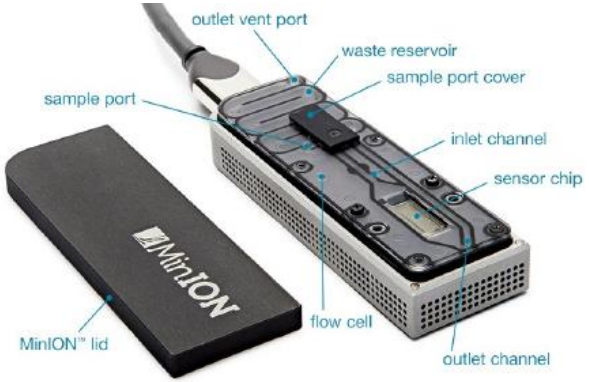



Read 1 Read 2

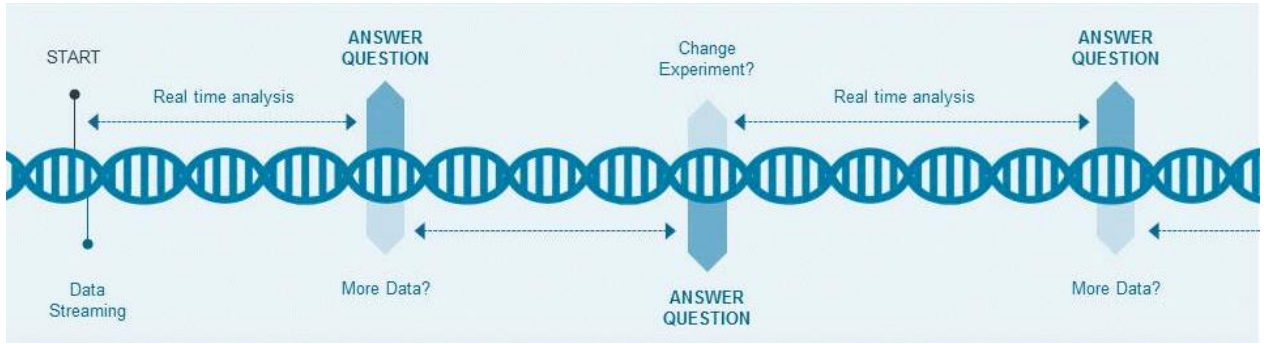
# Oxford Nanopore: nanopore sequencing with real-time analysis



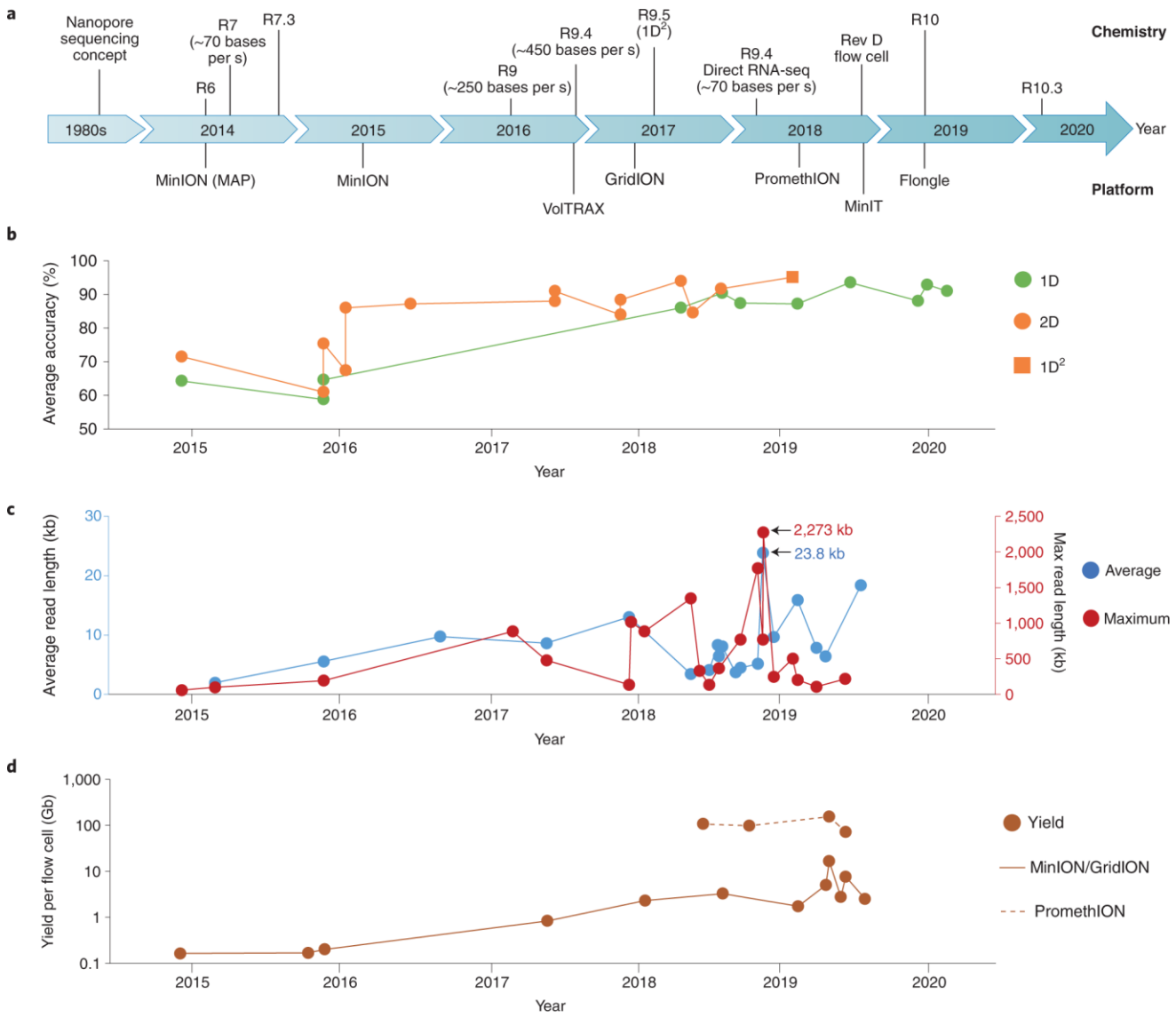






*Wang et al 2021*



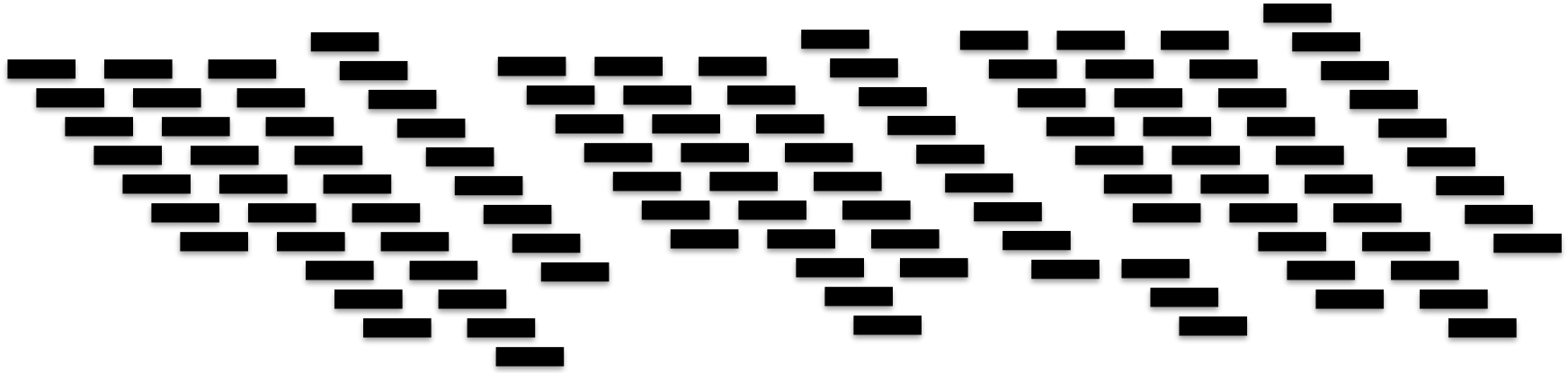
# Oxford Nanopore: nanopore sequencing with real-time analysis



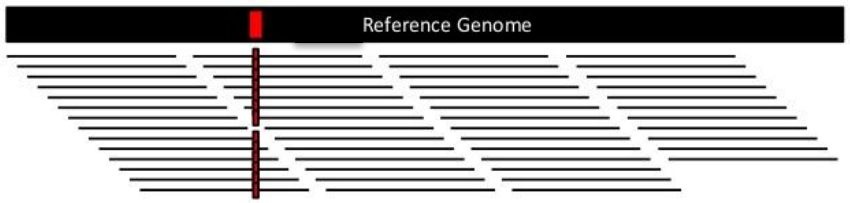

*Wang et al 2021*

imed
Research Institute
for Medicines

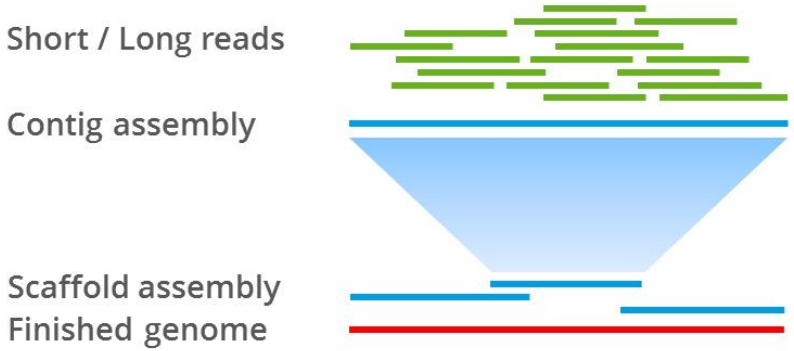
# Two main approaches for handling reads…



## Mapping or Reference Assembly

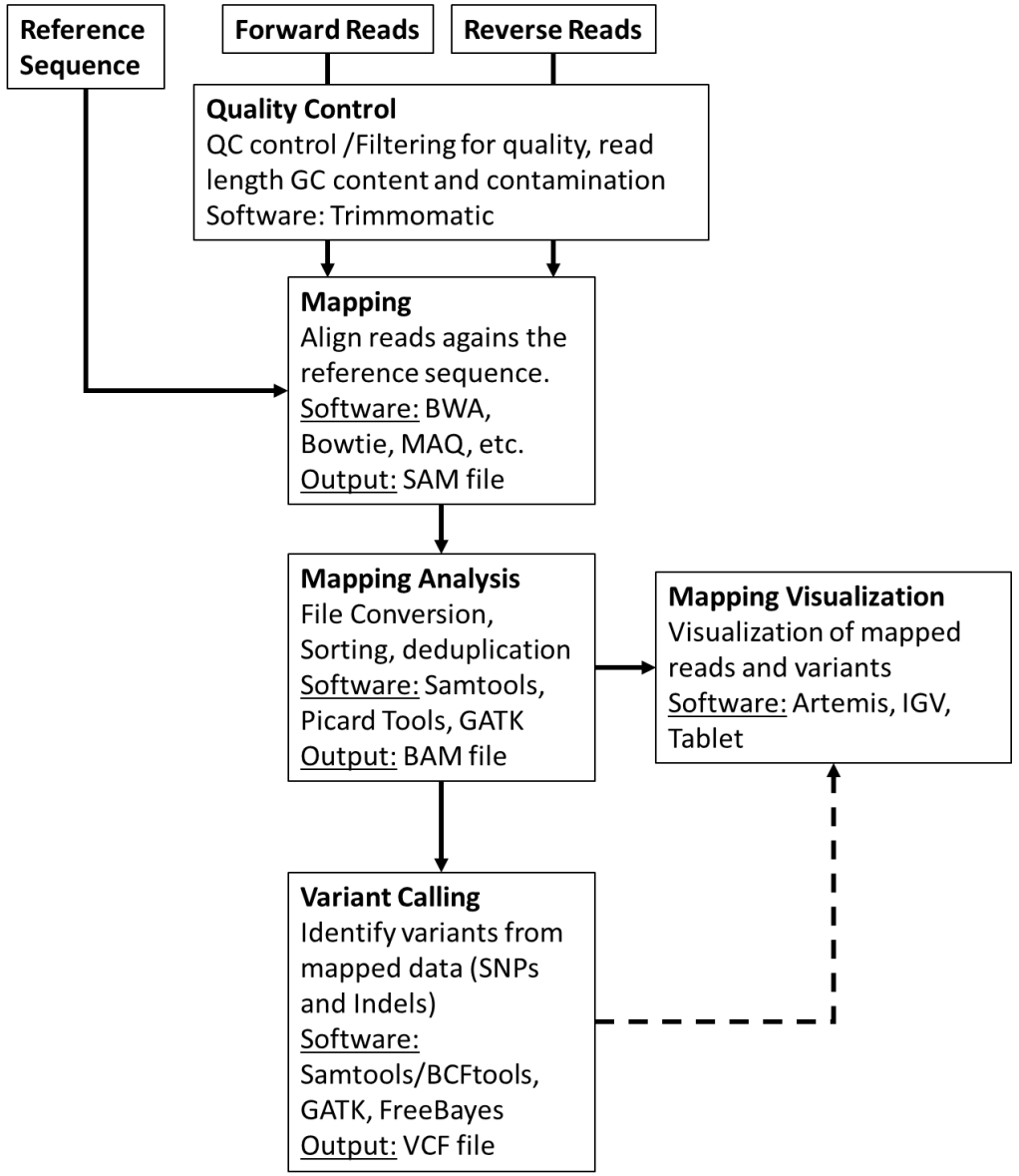



## *De novo* Assembly



imed
Research Institute
for Medicines

# Workflow

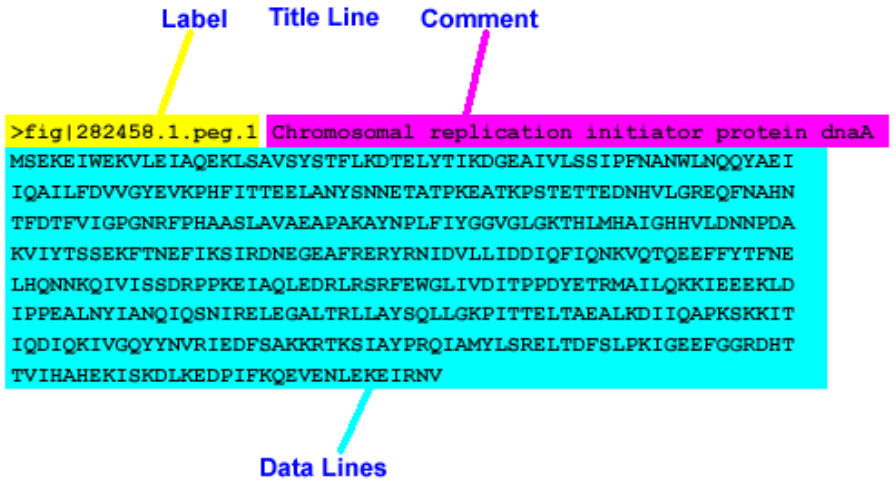



**Four main analytical stages:**

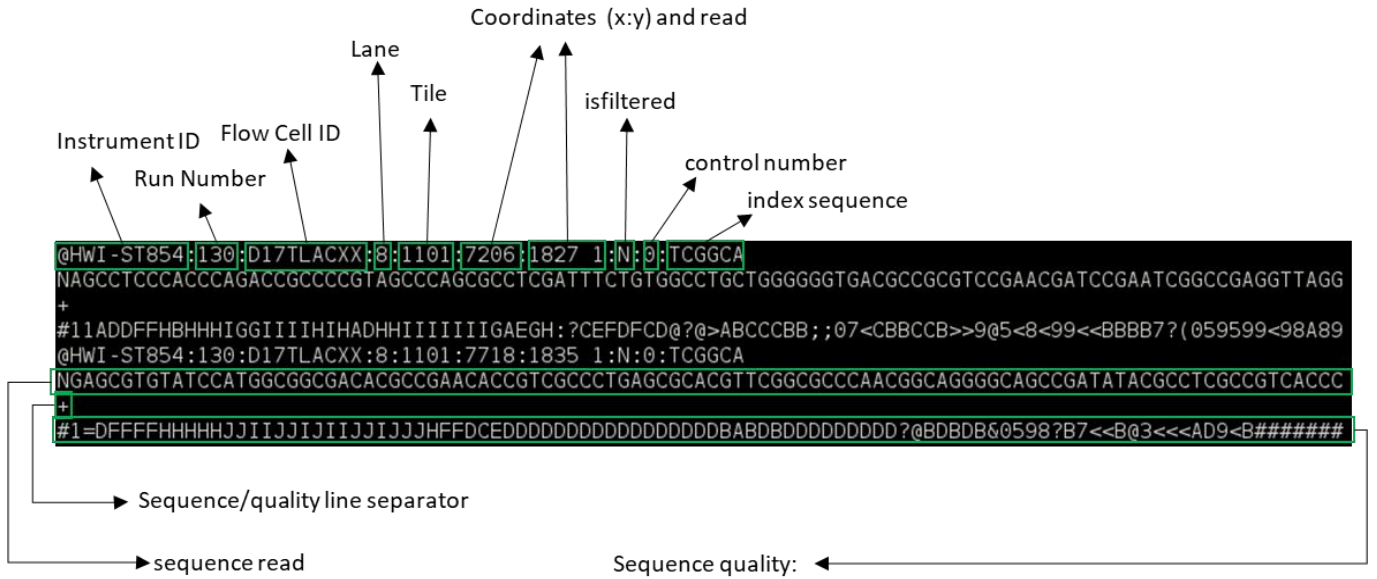
- *Quality-control* – filter out reads/bases associated with poor basecall quality;
- *Mapping* – map reads to a reference genome, obtain sample coverage at each position and read coordinates;
- *Variant Calling* – identify variants existing between sequenced and reference genome, either SNPs or INDELs;
- *Functional Annotation – determine the functional impact of each variant, e.g., which gene is affected? Is the mutation synonymous or non-synonymous? impact at the peptide primary structure?*

# What storage format for Sequencing Reads: FASTA vs FASTQ

## FASTA




## FASTQ



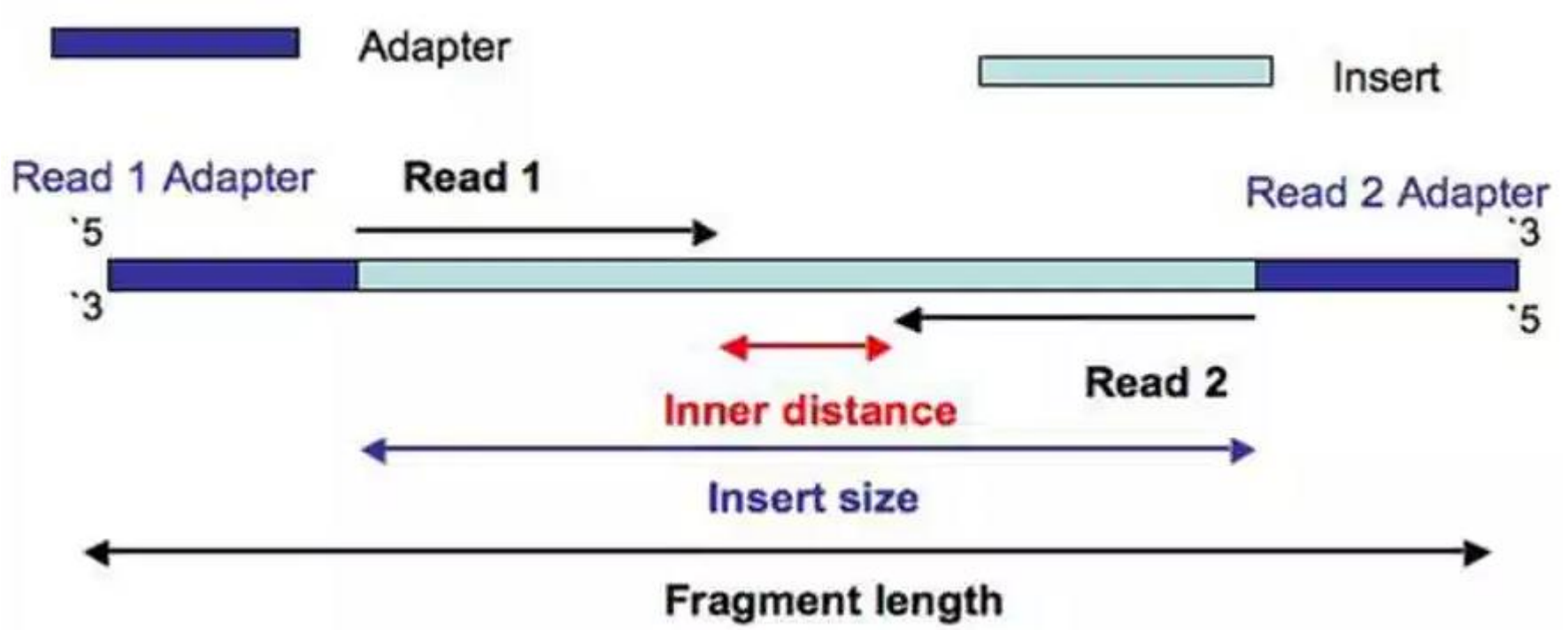

Sequence quality:
ASCII encoded Phred-33 quality score for each nucleotide of the sequence read above. Example: the 3rd nucleotide in the bottom read has a quality (Phred33 Q) of 28. Check the table below.
What about the 5th nucleotide of the same read? _____

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

| Q | P_error | ASCII | Q | P_error | ASCII | Q | P_error | ASCII | Q | P_error | ASCII |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.00000 | 33 ! | 11 | 0.07943 | 44 , | 22 | 0.00631 | 55 7 | 33 | 0.00050 | 66 B |
| 1 | 0.79433 | 34 " | 12 | 0.06310 | 45 - | 23 | 0.00501 | 56 8 | 34 | 0.00040 | 67 C |
| 2 | 0.63096 | 35 # | 13 | 0.05012 | 46 . | 24 | 0.00398 | 57 9 | 35 | 0.00032 | 68 D |
| 3 | 0.50119 | 36 $ | 14 | 0.03981 | 47 / | 25 | 0.00316 | 58 : | 36 | 0.00025 | 69 E |
| 4 | 0.39811 | 37 % | 15 | 0.03162 | 48 0 | 26 | 0.00251 | 59 ; | 37 | 0.00020 | 70 F |
| 5 | 0.31623 | 38 & | 16 | 0.02512 | 49 1 | 27 | 0.00200 | 60 < | 38 | 0.00016 | 71 G |
| 6 | 0.25119 | 39 ' | 17 | 0.01995 | 50 2 | 28 | 0.00158 | 61 = | 39 | 0.00013 | 72 H |
| 7 | 0.19953 | 40 ( | 18 | 0.01585 | 51 3 | 29 | 0.00126 | 62 > | 40 | 0.00010 | 73 I |
| 8 | 0.15849 | 41 ) | 19 | 0.01259 | 52 4 | 30 | 0.00100 | 63 ? | 41 | 0.00008 | 74 J |
| 9 | 0.12589 | 42 * | 20 | 0.01000 | 53 5 | 31 | 0.00079 | 64 @ | 42 | 0.00006 | 75 K |
| 10 | 0.10000 | 43 + | 21 | 0.00794 | 54 6 | 32 | 0.00063 | 65 A | | | |

# What storage format for Sequencing Reads: FASTA vs FASTQ

*Why do I get two FastQ files?*



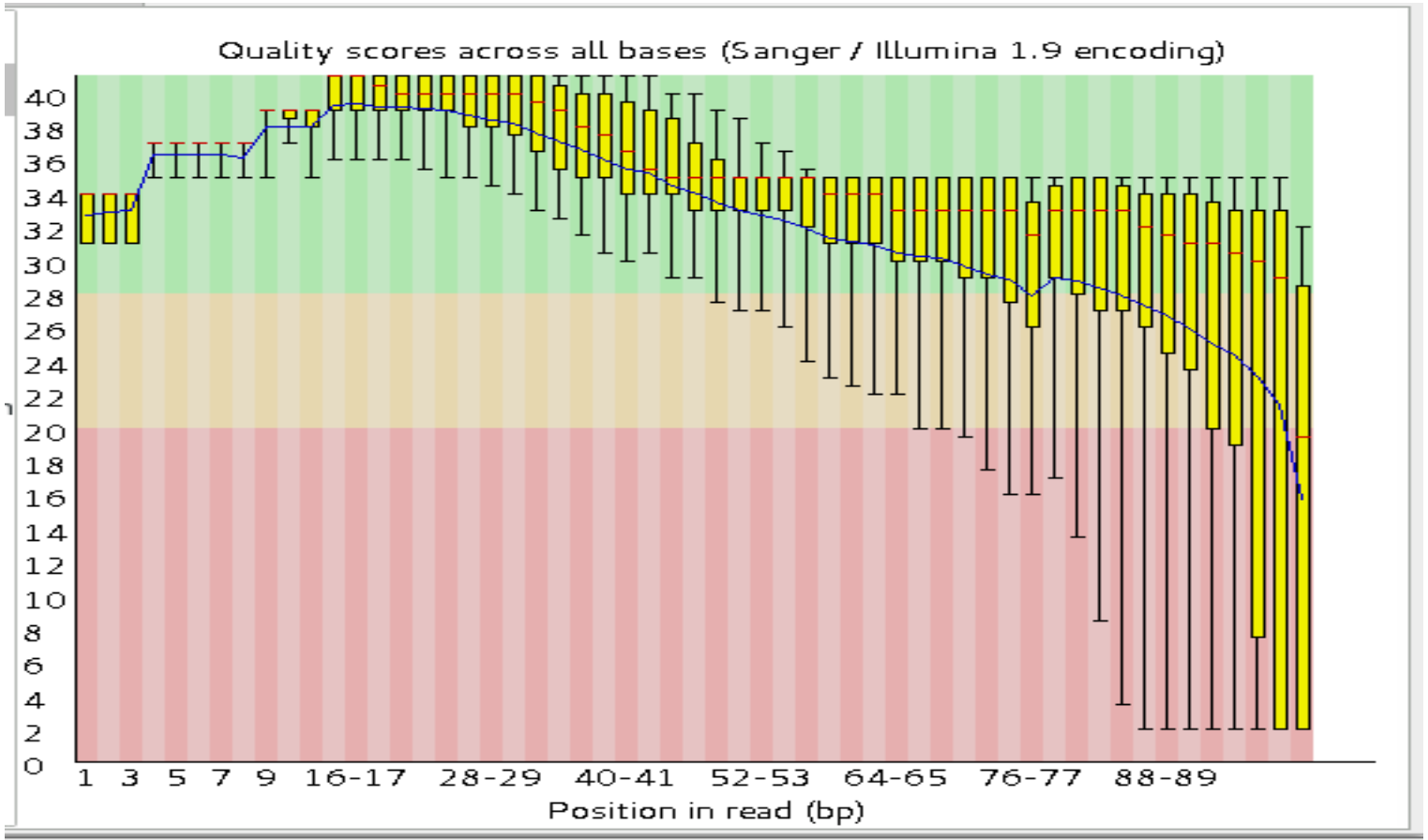

https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/

# Quality Control: Assessment

**Objective:**

- Assess overall sequencing quality and assess if sequencing metrics are within expected ranges;
- Remove base calls associated with low quality by removing or trimming sequencing reads;
- Taxonomical read QC – *did you sequence what you though you did*?

**Main/Most frequent problem**: base quality deterioration along the read length



imed Research Institute for Medicines

# Quality Control: Assessment

**How to assess sequencing metrics?**

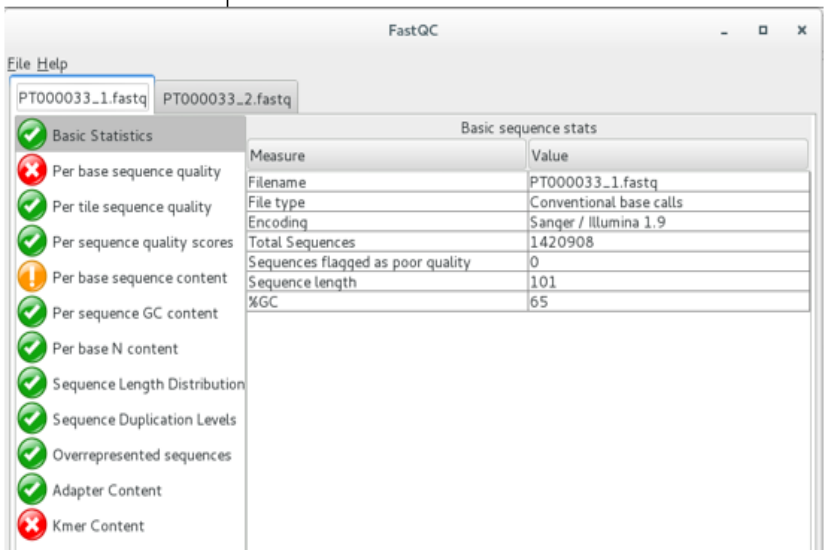
**Software:** FastQC, AfterQC, fastqp, HTSeq, etc.

FastQC – Java tool with both GUI and command-line as options.

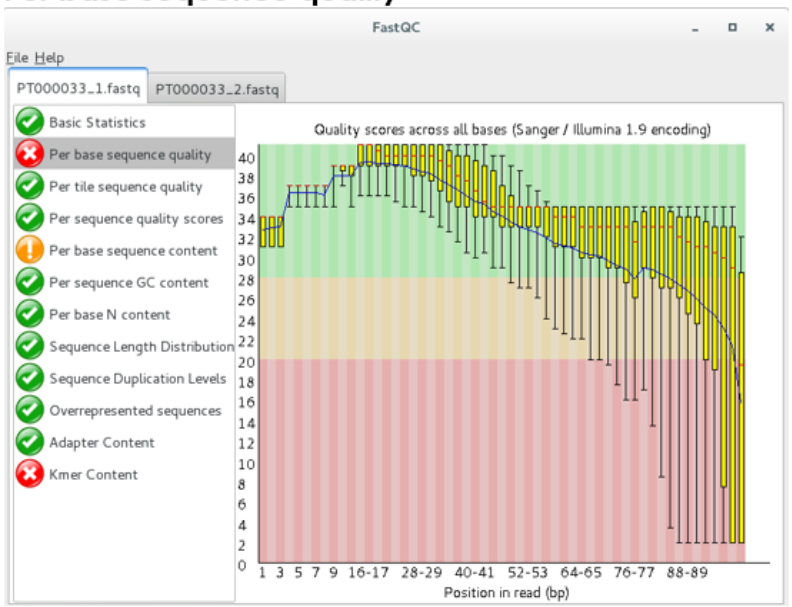
**Input:** FastQ files (or SAM/BAM files)

**Output:** sequencing metrics and plots

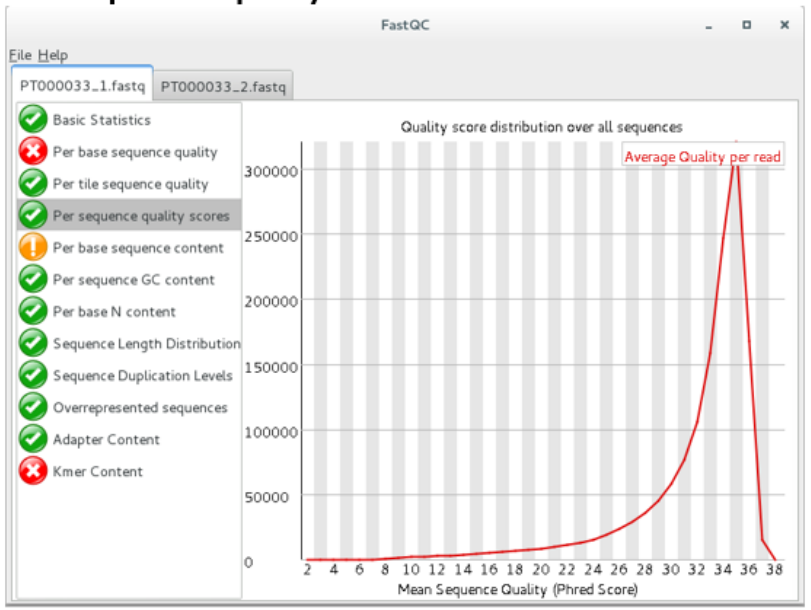
## Basic Statistics




## Per base sequence quality




## Per sequence quality score



# Quality Control: Assessment

**How to assess sequencing metrics?**

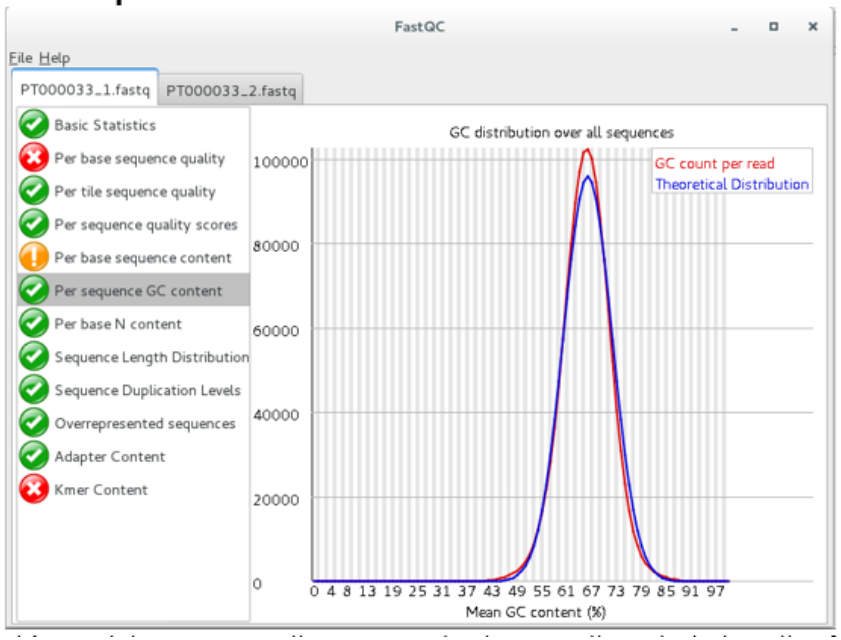
**Software:** FastQC, AfterQC, fastqp, HTSeq, etc.

FastQC – Java tool with both GUI and command-line as options.

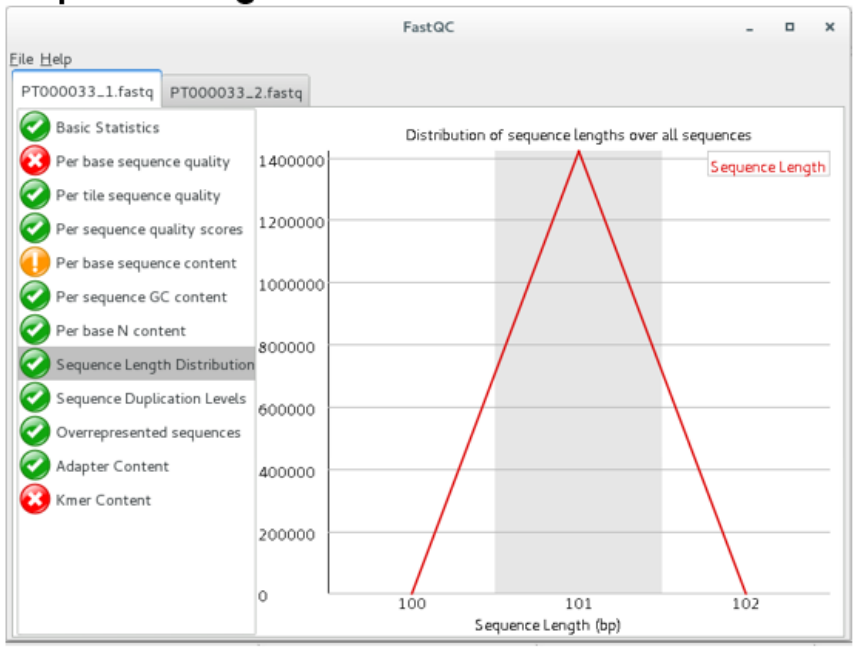
**Input:** FastQ files (or SAM/BAM files)

**Output:** sequencing metrics and plots

**Per Sequence GC content**




**Sequence Length Distribution**



# Quality Control: Correction

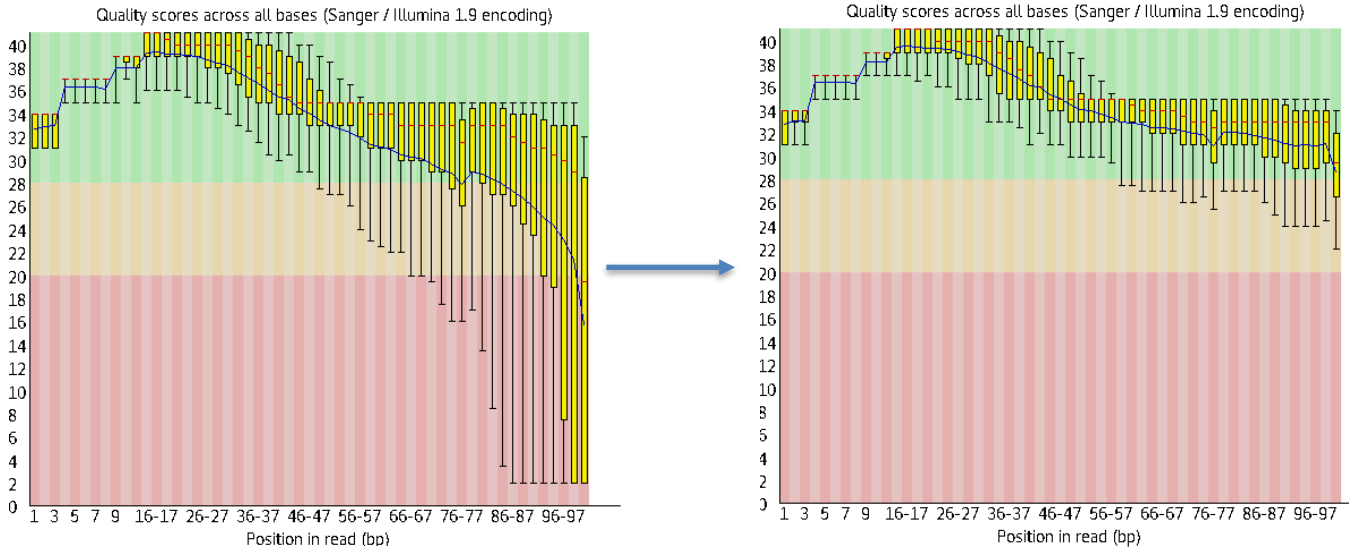
**How to correct, cut and filter out sequencing reads?**

**Software:** Trimmomatic, FASTX, etc.

Trimmomatic – command-line Java tool capable of handling SE and PE reads.

**Input:** FastQ files.

**Output:** FastQ files with trimmed/cut and surviving reads.




**Trimmomatic can:**

- Remove adapters;
- Remove leading and trailing low quality bases;
- Cut reads upon scanning at user defined sliding Windows when below a specified threshold;
- Remove reads when these don't meet a specified minimum length.

imed Research Institute for Medicines

# Quality Control: Taxonomical Read QC

***Did I sequence what I thought I did?*** or ***Why doesn't it map?*** or ***Why does the assembly look strange?***

**Software:** Kraken

Kraken – command-line tool that assigns reads to different taxonomical clades.

**Input:** FastQ files.

**Output:** Text Report.

```
0.89  28702   28702   U      0       unclassified
99.11  3210397 42095   -      1       root
97.81  3168297 1864    -      131567    cellular organisms
97.76  3166433 16142   D      2         Bacteria
97.25  3150142 4158    P      1224        Proteobacteria
97.12  3145930 19183   C      1236          Gammaproteobacteria
94.74  3068819 83719   O      91347           Enterobacterales
92.16  2985015 1245831 F      543               Enterobacteriaceae
53.46  1731479 124411  G      561                 Escherichia
49.60  1606565 1603919 S      562                   Escherichia coli
```

```
3.65  137472  137472  U      0       unclassified
96.35  3631172 7288    -      1       root
96.16  3623815 2151    -      131567    cellular organisms
96.10  3621664 10917   D      2         Bacteria
95.81  3610594 13835   P      1224        Proteobacteria
95.43  3596429 40211   C      1236          Gammaproteobacteria
92.48  3485213 108814  O      91347           Enterobacterales
89.53  3374249 342184  F      543               Enterobacteriaceae
79.20  2984764 1375209 G      570                 Klebsiella
38.86  1464425 1415727 S      573                   Klebsiella pneumoniae
```

imed
Research Institute
for Medicines

# Mapping or Reference Assembly

**Objective:** Find the origin of a sequencing read providing a reference genome is known

**Reference genome:** Should be a high quality genome, ideally finished, the close as possible to the sequenced genome.

**Software:** Burrows-Wheeler Aligner (BWA), Bowtie2, HISAT2

**Input:** FastQ files.

**Output:** Mapped/Alignment File
SAM/BAM file

Most mapping software implement the Burrows-Wheeler transformation algorithm which enables fast access to sequence data with an acceptable memory footprint.

| | Execution time | | | | Memory usage | | | | Accuracy | | | | % Prop. paired reads | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ins. (bp) | 350 | | 550 | | 350 | | 550 | | 350 | | 550 | | 350 | | 550 | |
| RL (bp) | 100 | 150 | 100 | 150 | 100 | 150 | 100 | 150 | 100 | 150 | 100 | 150 | 100 | 150 | 100 | 150 |
| BWA | + | + | + | + | + | + | + | + | +++ | +++ | +++ | +++ | +++ | +++ | +++ | +++ |
| Bowtie2 | ++ | ++ | ++ | ++ | +++ | +++ | ++ | ++ | +++ | +++ | +++ | +++ | ++ | ++ | +++ | +++ |
| HISAT2 | +++ | +++ | +++ | +++ | +++ | +++ | +++ | +++ | ++ | ++ | ++ | ++ | + | + | + | + |

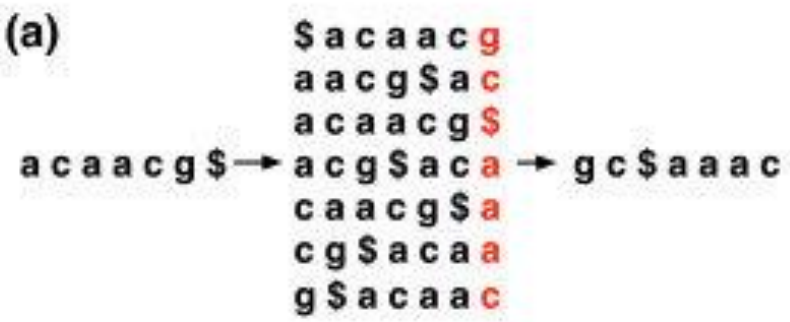
*"We conclude that there is not a single mapper that is ideal in all scenarios but rather the choice of alignment tool should be driven by the application and sequencing technology."*

*Keel et al 2018*

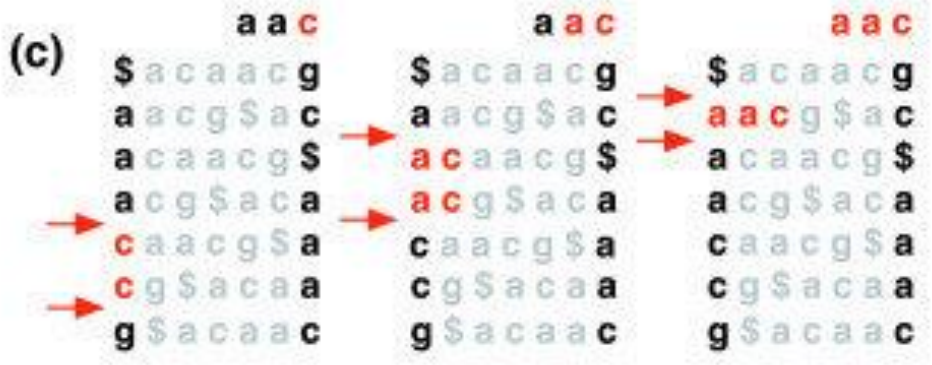
imed Research Institute for Medicines

# The Burrows-Wheeler Transform

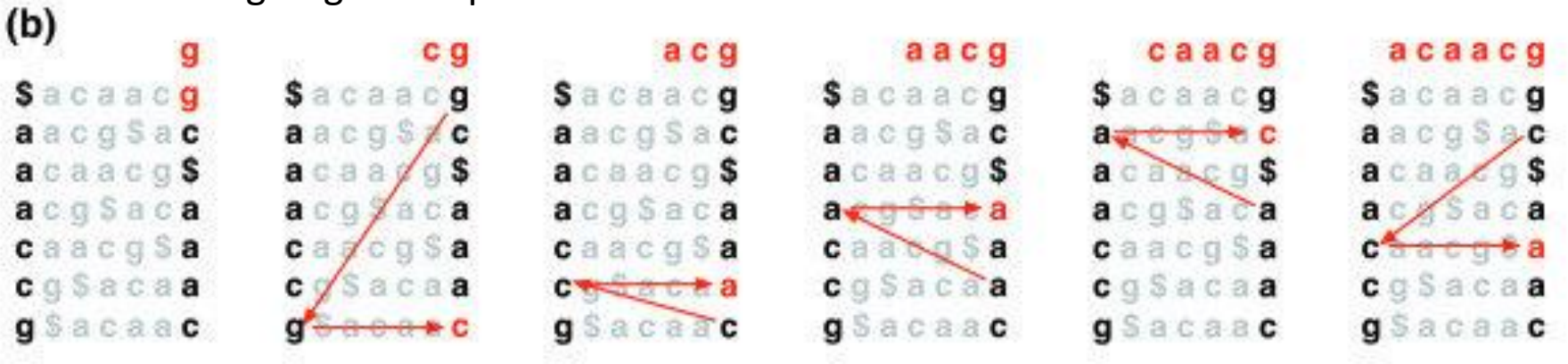
Reference compression:




Searching for *aac* string:

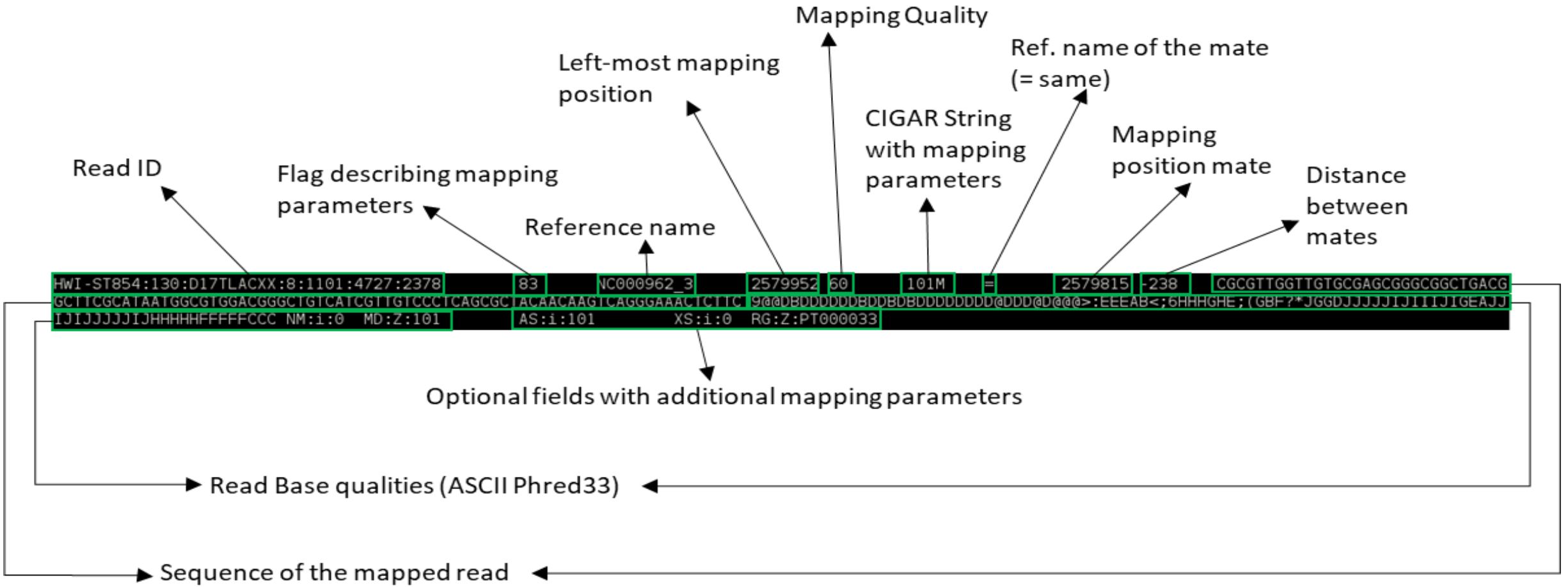



Reconstructing original sequence:




The *i*th x in the first column corresponds to the *i*th x in the last column.

# The SAM Format…




The BAM files are a binary version of the SAM files (text format)

Both BAM and SAM files can be manipulated and viewed with SAMtools or Picard Tools

imed
Research Institute
for Medicines

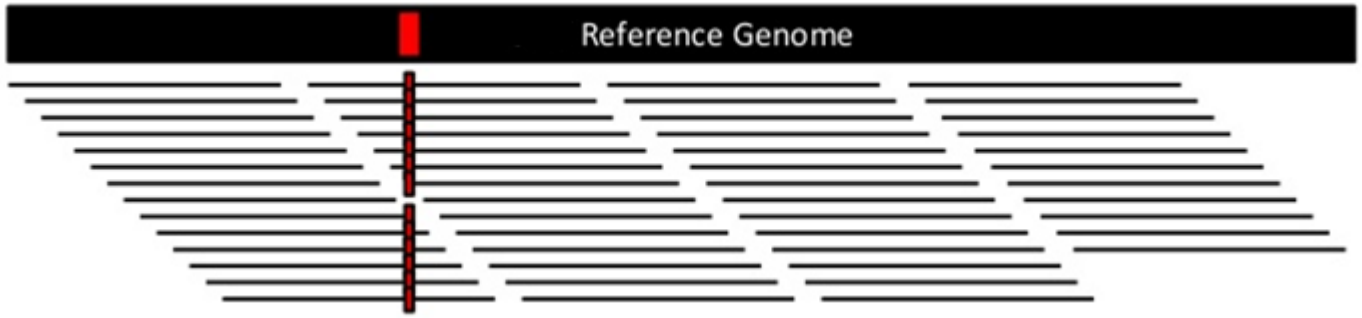
# Variant Calling

**Objective:** Identify, list and store genomic variants, either SNPs or INDELs.

**Software:** SAMtools/BCFtools; Genome Analysis Toolkit (GATK), FreeBayes, LoFreq

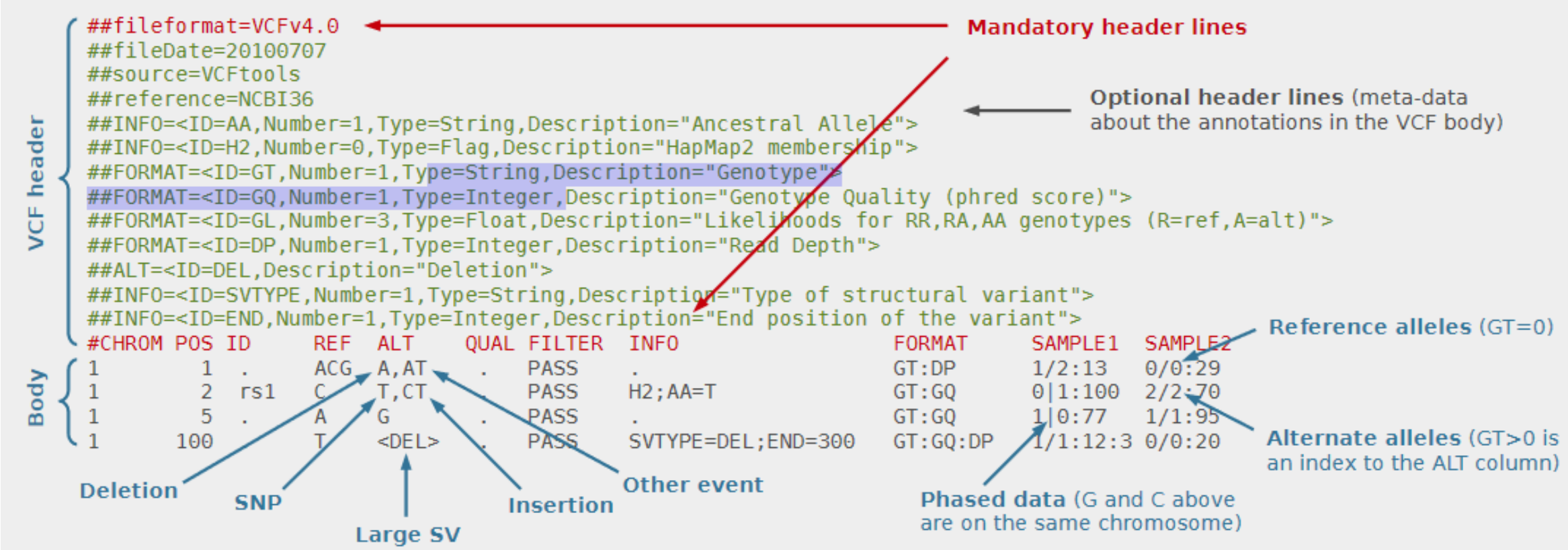
**Input:** BAM/SAM files.

**Output:** VCF files



imed
Research Institute
for Medicines

# VCF Format




http://vcftools.sourceforge.net/VCF-poster.pdf

See full specs:
https://samtools.github.io/hts-specs/VCFv4.2.pdf

imed Research Institute for Medicines

# VCF Format

## Format differences between variant callers - Examples

### Finding allelic depth and filtering

### SAMTools

NC000962_3 69871 . C T 225 . DP=23;VDB=0.641395;SGB=-0.69168;MQSB=0.537242;MQ0F=0;AC=2;AN=2;DP4=0,0,11,8;MQ=43 GT:PL 1/1:255,57,0

bcftools view --include 'QUAL>=20 && INFO/DP>=10 && **(INFO/DP4[2]+INFO/DP4[3])/(sum(INFO/DP4))>=0.9**'

### GATK

NC000962_3 69871 . C T 810 . AC=1;AF=1.00;AN=1;BaseQRankSum=1.988;DP=24;Dels=0.00;FS=0.000;HaplotypeScore=0.9469;MLEAC=1;MLEAF=1.00;MQ=60.00;MQ0=0;MQRankSum=0.000;QD=33.75;ReadPosRankSum=1.592;SOR=0.353 GT:AD:DP:GQ:PL 1:1,23:24:99:840,0

bcftools view --include 'QUAL>=20 && FORMAT/DP>=10 && **(FORMAT/AD[*:1])/(FORMAT/DP)>=0.9**'

### Freebayes

NC000962_3 69871 . C T 656.091 . AB=0;ABP=0;AC=1;AF=1;AN=1;AO=23;CIGAR=1X;DP=24;DPB=24;DPRA=0;EPP=3.10471;EPPR=5.18177;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=1;NUMAL
T=1;ODDS=151.07;PAIRED=0.956522;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=760;QR=14;RO=1;RPL=14;RPP=5.3706;RPPR=5.18177;RPR=9;RUN=1;SAF=12;SAP=3.10471;SAR=11;SRF=0;SRP=5.18177;SRR=1;TYPE=snp;technology.illumina=1 GT:DP:AD:RO:QR:AO:QA:GL 1:24:1,23:1:14:23:760:-67.3027,0

bcftools view --include 'QUAL>=20 && FORMAT/DP>=10 **&& (FORMAT/AO)/(FORMAT/DP)**>=0.9

### LoFreq

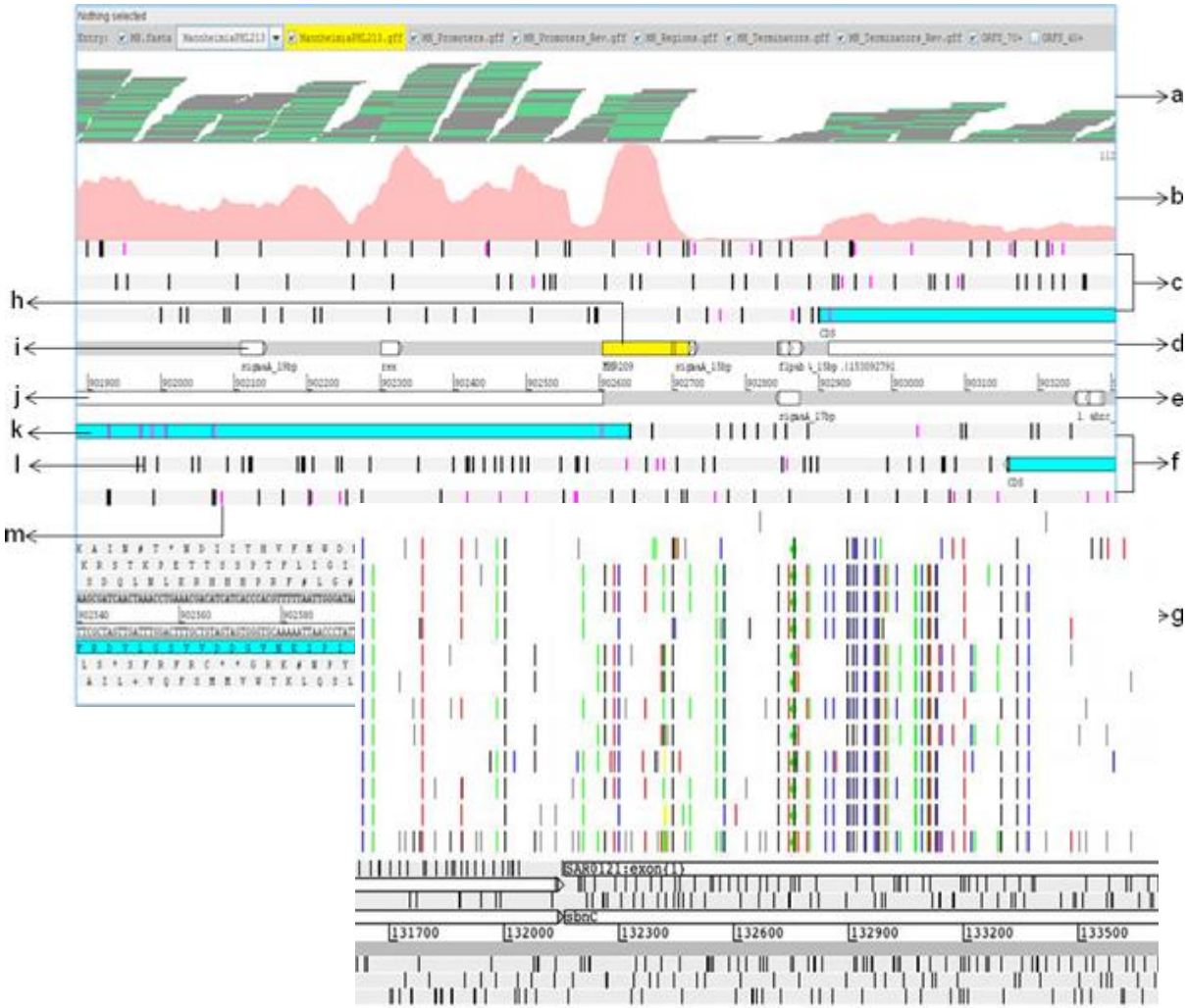
NC000962_3 69871 . C T 732 PASS DP=24;AF=0.958333;SB=0;DP4=0,1,12,11

bcftools view --include 'QUAL>=20 && INFO/DP>=10 && **(INFO/DP4[2]+INFO/DP4[3])/(sum(INFO/DP4))>=0.9**'c

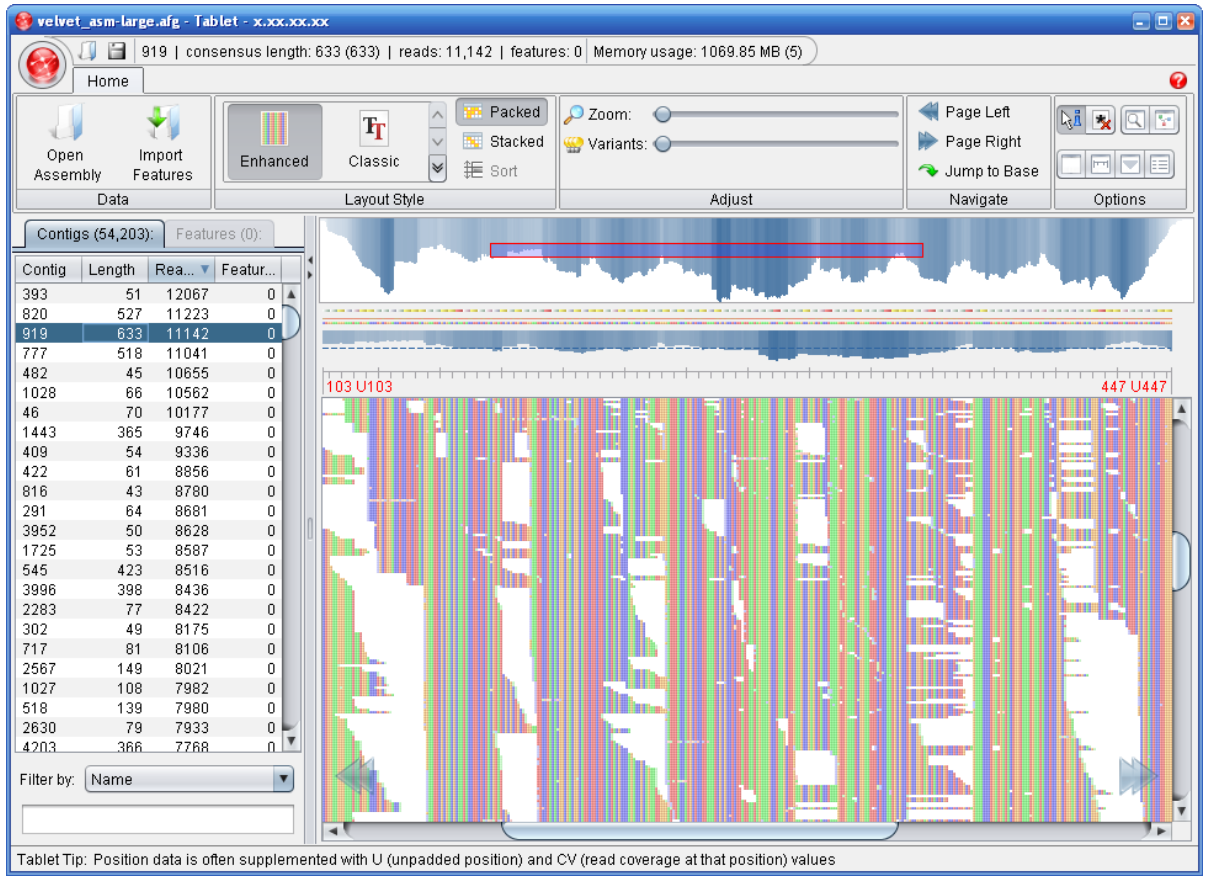
imed
Research Institute
for Medicines

# BAM and VCF Visualizatiom

## Artemis




## Tablet



imed Research Institute for Medicines

# Functional Annotation

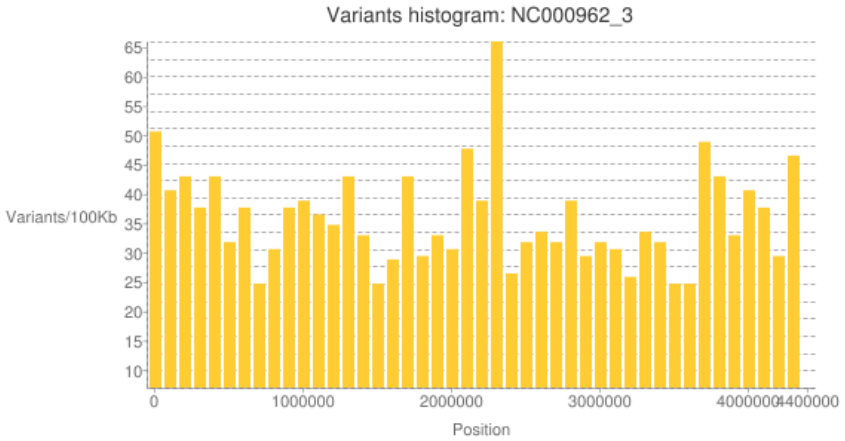
| | |
|---|---|
| **Genome** | NC000962_3 |
| **Date** | 2021-06-17 01:39 |
| **SnpEff version** | SnpEff 5.0e (build 2021-03-09 06:01), by Pablo Cingolani |
| **Command line arguments** | SnpEff -no-downstream -no-upstream NC000962_3 PT000033.filt.vcf |
| **Warnings** | 124 |
| **Errors** | 0 |
| **Number of lines (input file)** | 1,597 |
| **Number of variants (before filter)** | 1,598 |
| **Number of not variants (i.e. reference equals alternative)** | 0 |
| **Number of variants processed (i.e. after filter and non-variants)** | 1,598 |
| **Number of known variants (i.e. non-empty ID)** | 0 ( 0% ) |
| **Number of multi-allelic VCF entries (i.e. more than two alleles)** | 1 |
| **Number of effects** | 16,074 |
| **Genome total length** | 4,411,532 |
| **Genome effective length** | 4,411,532 |
| **Variant rate** | 1 variant every 2,760 bases |

**Objective:** Annotate each variant with tis functional impact/consequence.

**Software:** SnpEff, GATK etc

**Input:** VCF files.

**Output:** Annotated VCF files




## Number of effects by functional class

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| MISSENSE | 786 | 61.025% |
| NONSENSE | 13 | 1.009% |
| SILENT | 489 | 37.966% |

Missense / Silent ratio: 1.6074

'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGV S.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO' ">

ANN=G|missense_variant|MODERATE|katG|Rv1908c|transcript|Rv1908c|protein_coding|1/1|c.944G>C|p.Ser315Thr|944/2223|944/2223|315/740||