GULBENKIAN CIÊNCIA

Aligments

Precision Oncology Course

cnio SPANISH NATIONAL CANCER RESEARCH CENTRE

BU CNIO **BIOINFORMATICS** UNIT

Coral Fustero Torre
Bioinformatics Unit,
Structural Biology Programme.
cfustero@cnio.es | bioinformatics.cnio.es

Alignments

What is an alignment?

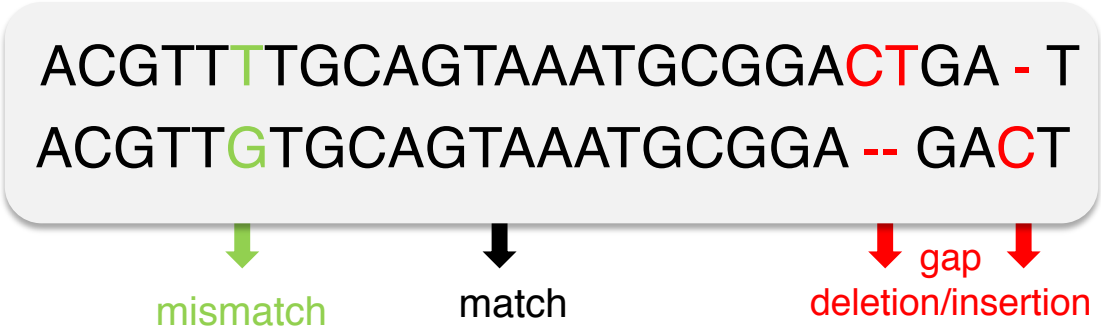
Sequence alignments are a way of arranging the sequences of DNA, RNA, or protein in order to **identify regions of similarity** that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

```
ACGTCTTGACTGG -TTAAAATAC
AC - TCTTGACTGGATTAACATAC
```

Alignments

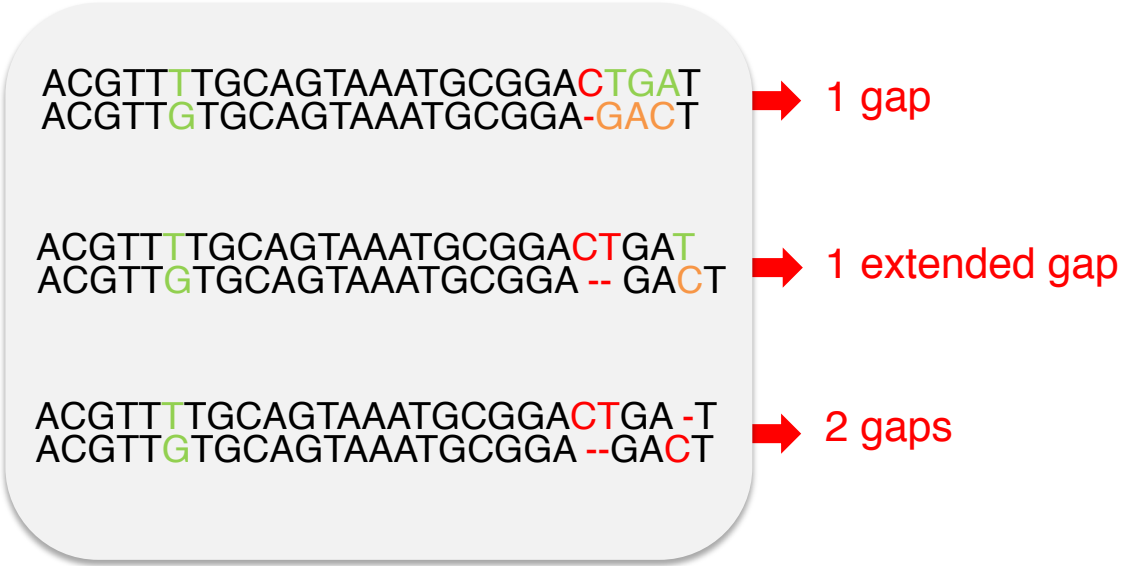
Elements of an aligment

Alignment seeks to **reduce gaps and mismatches** and **maximize matches**.
In the construction, each of these components has a penalty value associated. For gaps there is a penalty value for opening the gap and another for extending it.



Alignments

Elements of an aligment



Alignments

Elements of an aligment

1. Based on the number of sequences:

- **Pairwise** alignment: 2 sequences
- **Multiple** alignment: > 2 sequences

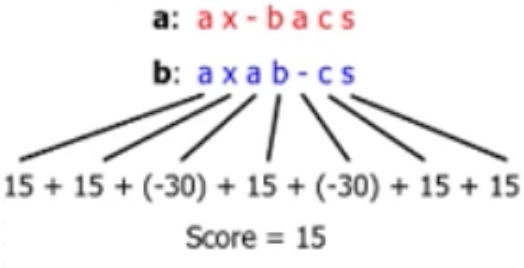
2. Based on the **region** to align:

- **Local**: sequence sub-region (Smith and Waterman, BLAST)
  Alignment is done only in the most similar regions
- **Global**: complete sequence (Needleman Wunsch)
  Alignment covers two sequences completely
  To align sequences that start and end in the same region (homologous genes of similar species)

Example:

P = xyaxbacsll, **T** = pqraxabcstvq

Answer:



Alignments

Objectives

The comparison between sequences in sequence alignment allows to:

1. Determine the **homology** degree
2. Identify **functional domains**
3. **Compare** the gene with its product
4. Find homologous **positions**
5. Identify **differences**

Alignments

Objectives

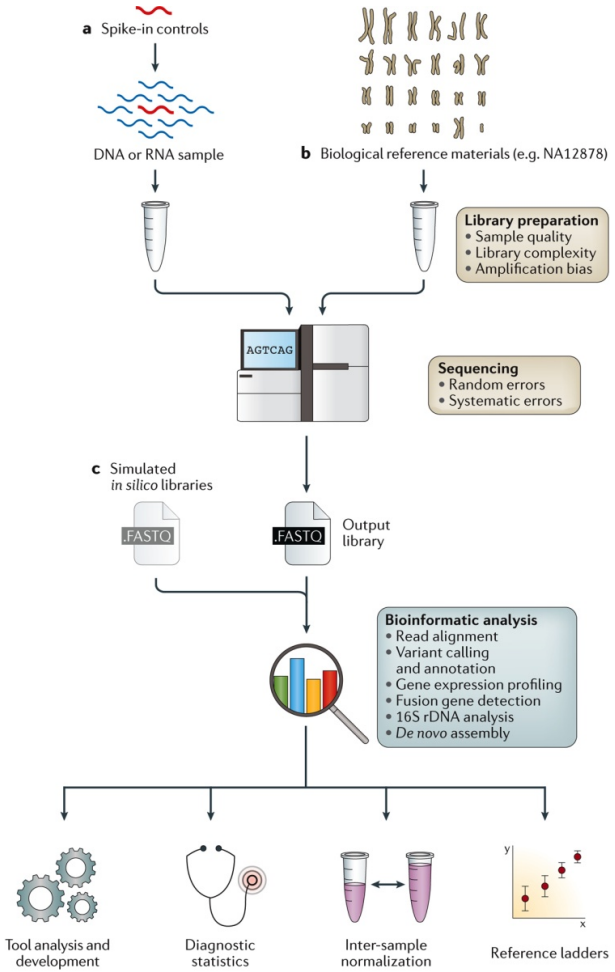
The comparison between sequences in sequence alignment allows to:

1. Determine the **homology** degree
2. Identify **functional domains**
3. Compare the gene with its product
4. Find homologous **positions**
5. Identify **differences**

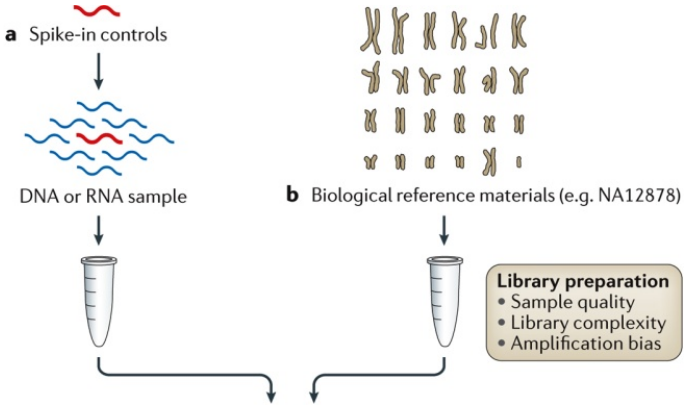
Differential Expression and Variant Detection in Next Generation Sequencing (NGS)

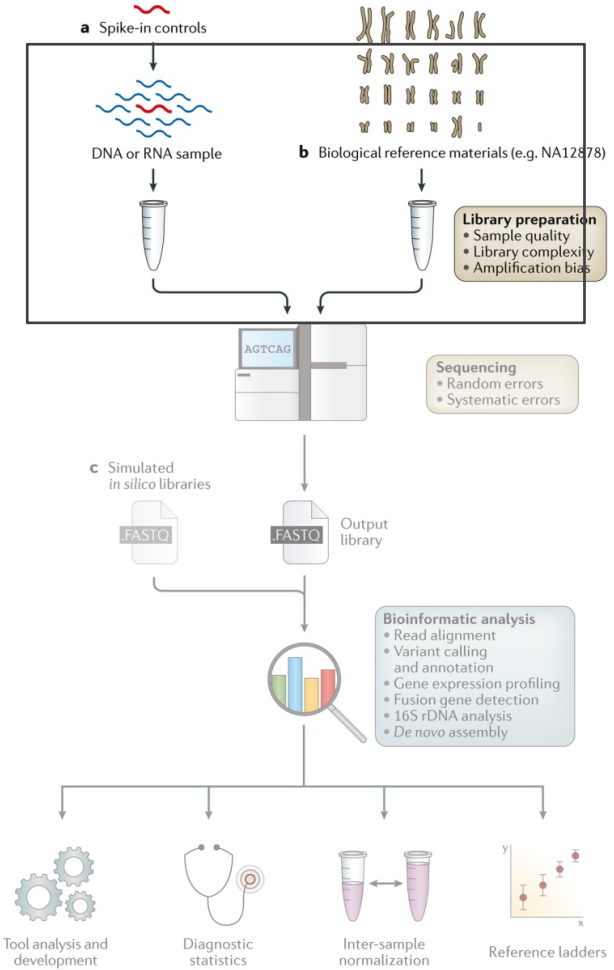
Next Generation Sequencing

One of the earliest and important steps in NGS analysis is the mapping of the reads to the original reference or **Read Alignment**



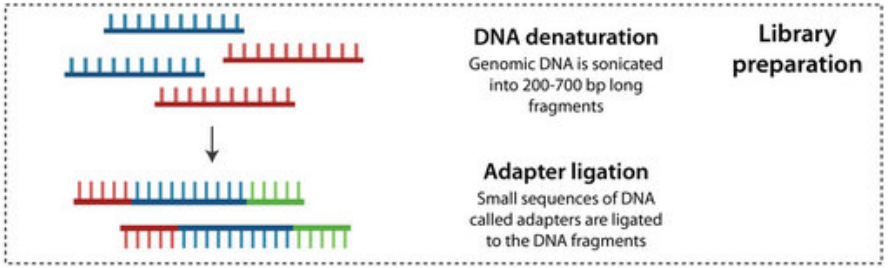
Next Generation Sequencing

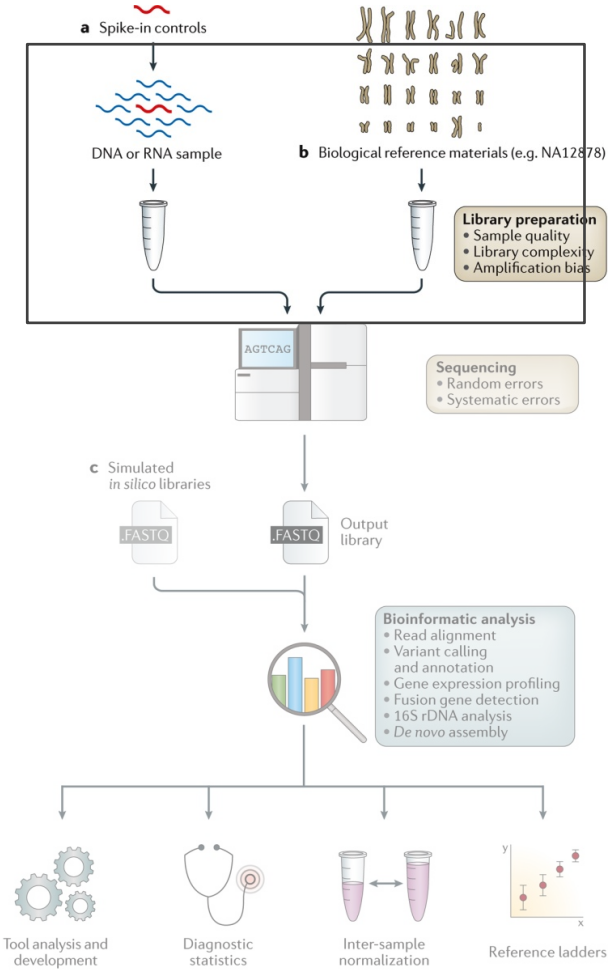




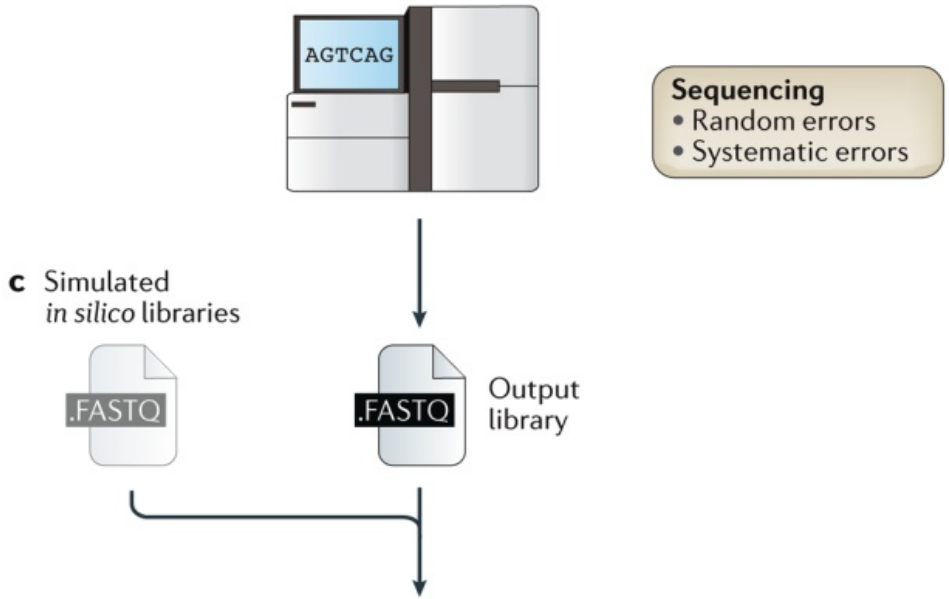
Next Generation Sequencing

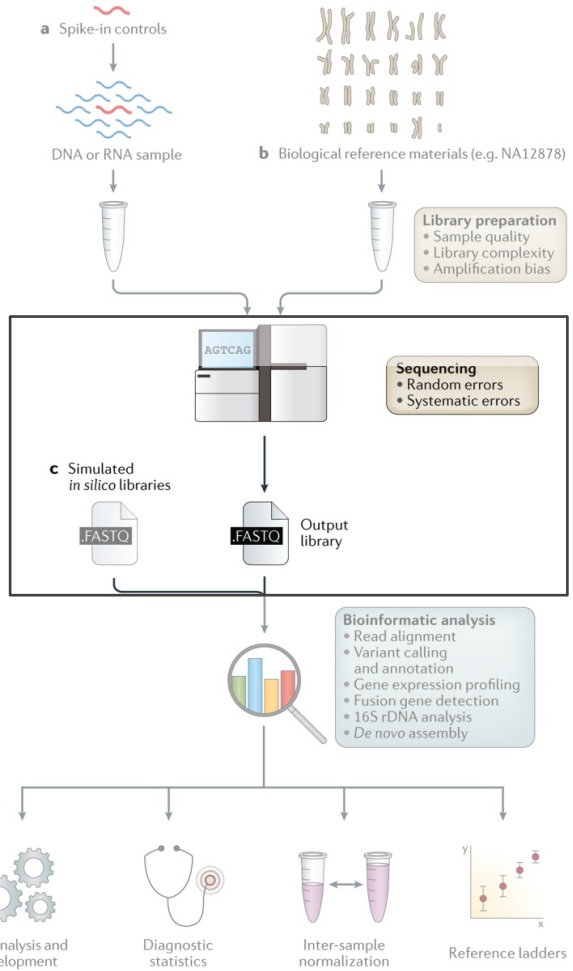
Video: https://www.youtube.com/watch?v=fCd6B5HRaZ8





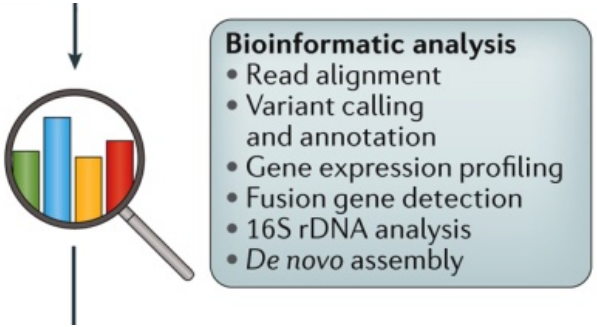
Next Generation Sequencing

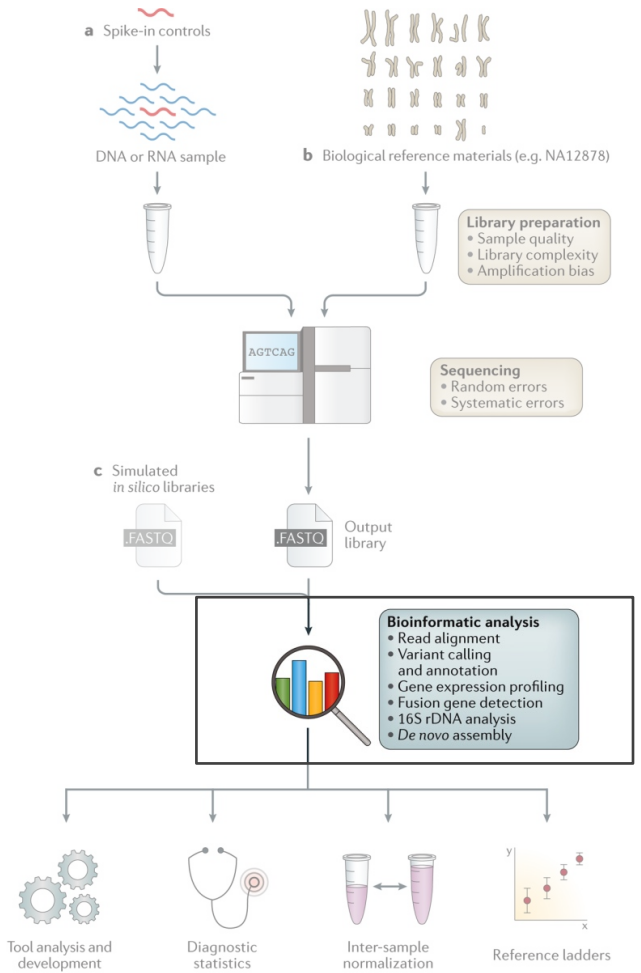




Nature Reviews | Genetics

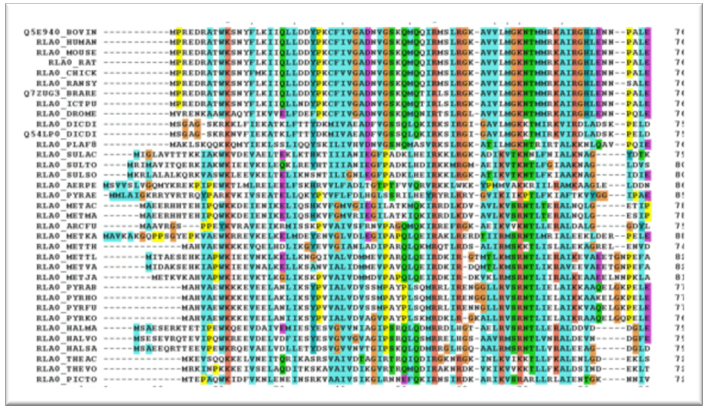
Next Generation Sequencing

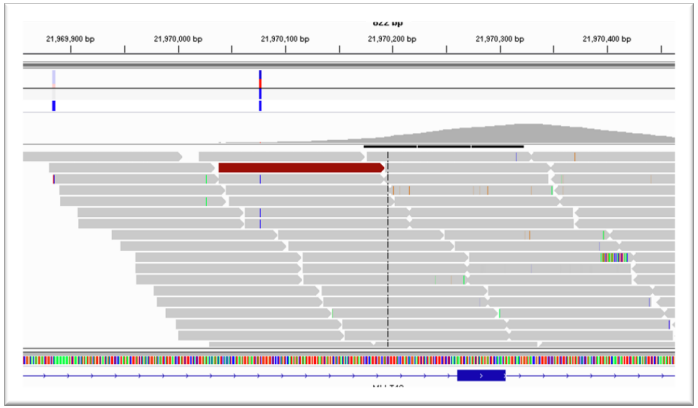




Classical vs NGS alignment

	Classical	NGS
Quantity	A few sequences (n < 30) between them	Billions of reads to a very large reference genome (n = 10^6 - 10^8)
Length	Long sequences (a whole gene, including introns, or a whole protein)	Reads have short sequences (I = 25-1000 bp)
Similarity	No very similar sequences	Highly similar sequences
Quality	High quality sequences coming from Sanger capillary sequencing	Lower quality sequences
Examples	ClustalW, T-Coffee	BWA, Bowtie





Short Read Aligners: Challenges

As we need to align billions of reads to a very large reference genome → SRA must be "**extraordinarily efficient algorithms**"

- Speed
- Memory use

As we need to align short reads, a read may align in multiple positions → SRA have to:

- Either report multiple positions
- Or pick heuristically one of them

Different NGS technologies have different error profiles to take into account:

- **454**: insertion or deletions in homopolymer runs
- **Illumina**: increasing likelihood of sequence errors towards the end of the read

Specific problems: splicing junctions in RNA-seq

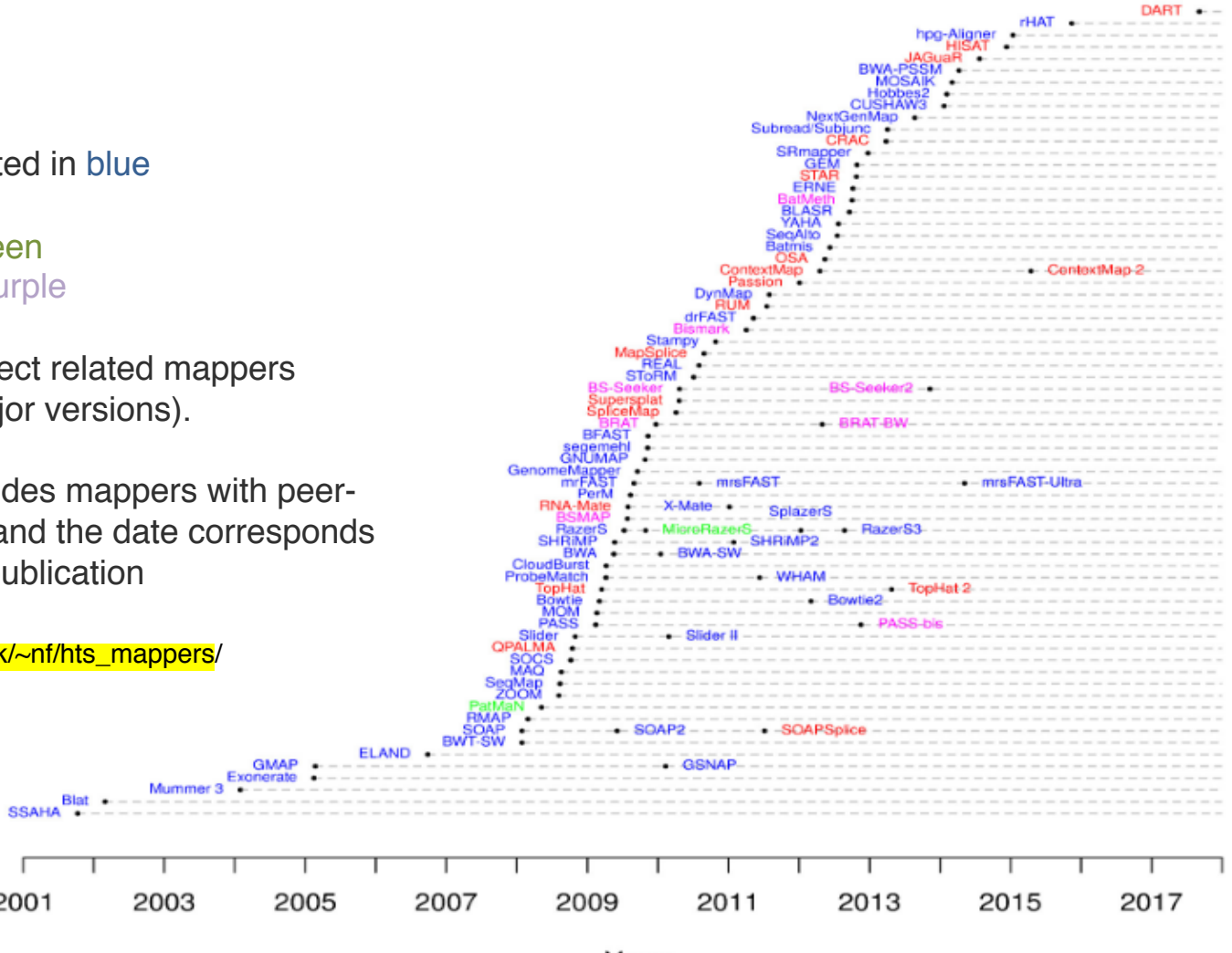
Timeline

DNA mappers are plotted in blue
RNA mappers in red
miRNA mappers in green
bisulfite mappers in purple

Gray dotted lines connect related mappers (extensions or new major versions).

The time line only includes mappers with peer-reviewed publications and the date corresponds to the earliest date of publication

Source: https://www.ebi.ac.uk/~nf/hts_mappers/

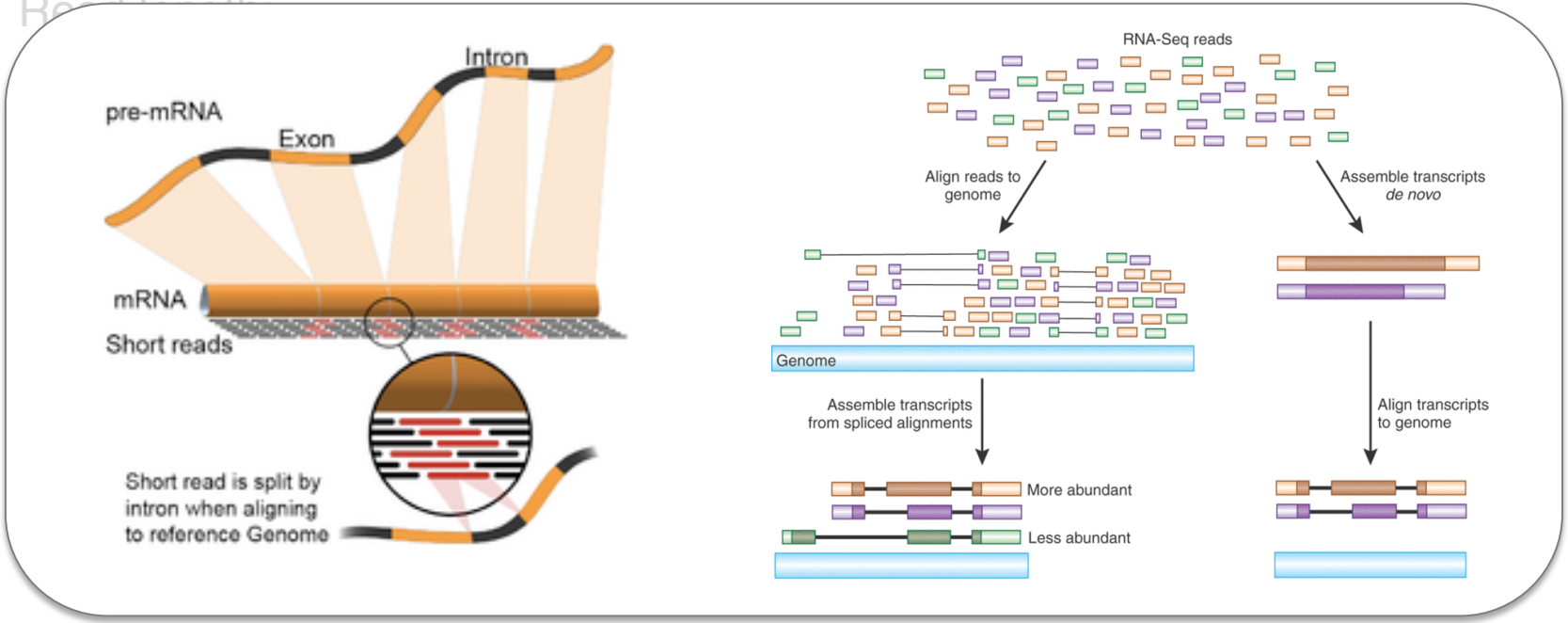


Elements to consider

- Read type: DNA, RNA, etc.
- Read length:
- Paired-end or not
- Computational requirements (number of processors, memory)
- Base quality (taken or not into account)
- Sequencing errors (can be platform dependent)
- Number of mismatches (limitation in allowed differences)

Elements to consider

- Read type: DNA, RNA, etc.
- Read length



Nature Biotechnology **28**, 421–423(2010)

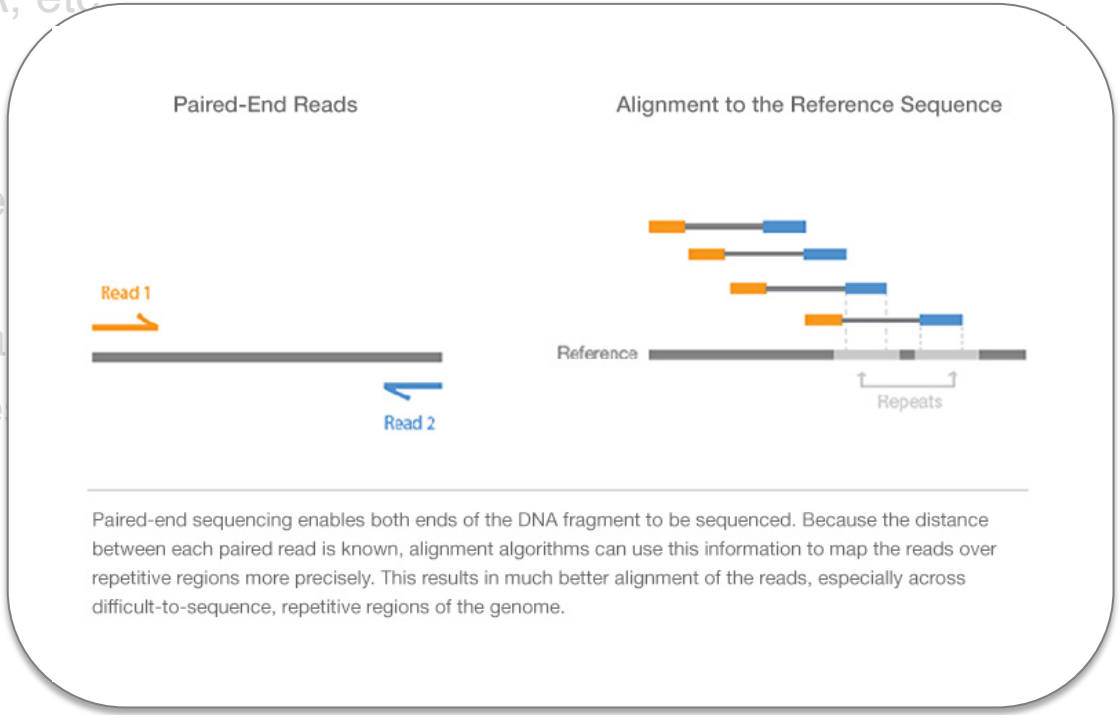
Elements to consider

- Read type: DNA, RNA, etc.
- Read length:
  - extremely short sequences (miRNA)
  - Increasing length of the reads (more probability of mismatches and gaps)

- Paired-end or not
- Computational requirements (number of processors, memory)
- Base quality (taken or not into account)
- Sequencing errors (can be platform dependent)
- Number of mismatches (limitation in allowed differences)

Elements to consider

- Read type: DNA, RNA, etc
- Read length:
- Paired-end or not
- Computational require
- Base quality (taken or
- Sequencing errors (ca
- Number of mismatche



https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

Elements to consider

- Read type: DNA, RNA, etc.
- Read length:
- Paired-end or not
- Computational requirements (number of processors, memory)
- Base quality (taken or not into account)
- Sequencing errors (can be platform dependent)
- Number of mismatches (limitation in allowed differences)

Elements to consider

- Read type: DNA, RNA, etc.
- Read length:
- Paired-end or not
- Computational requirements (number of processors, memory)
- Base quality (taken or not into account)
- Sequencing errors (can be platform dependent)
- Number of mismatches (limitation in allowed differences)

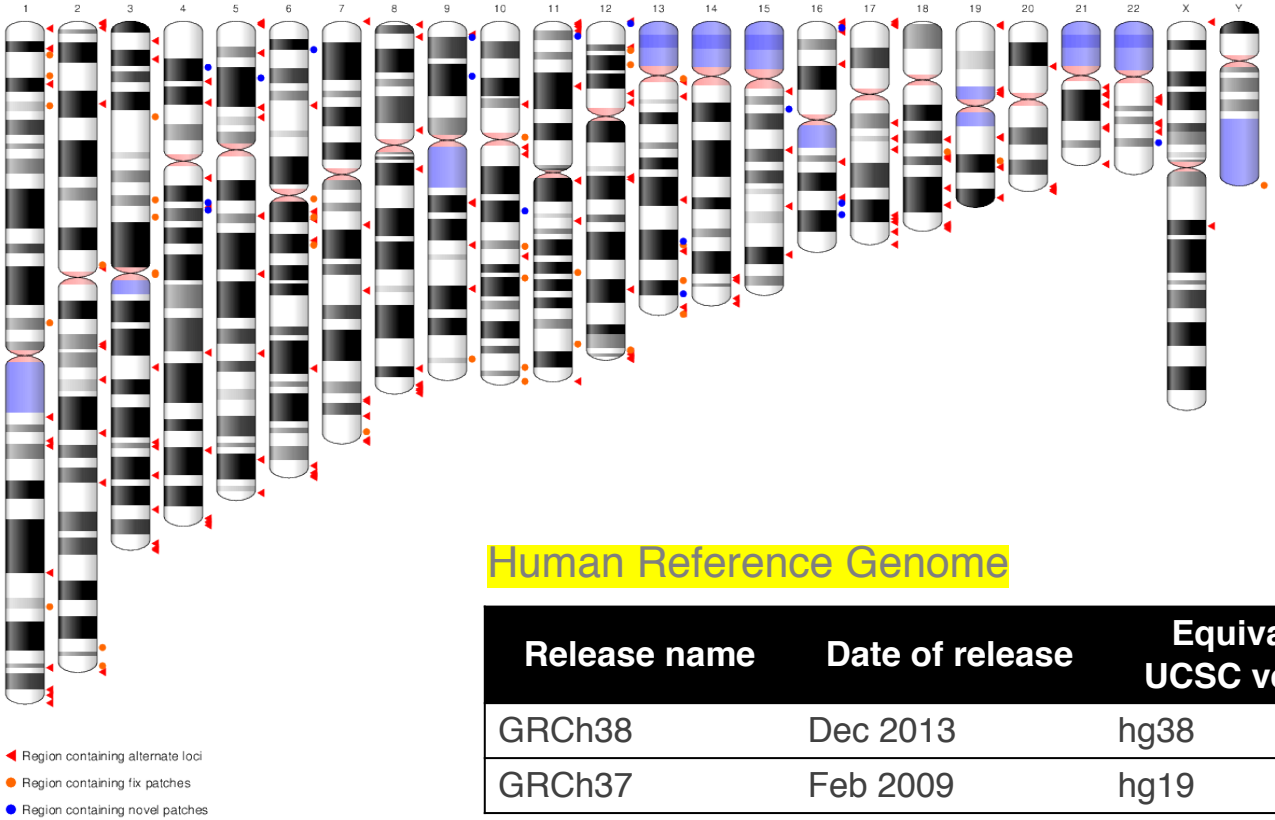
Elements to consider

- Read type: DNA, RNA, etc.
- Read length:
- Paired-end or not
- Computational requirements (number of processors, memory)
- Base quality (taken or not into account)
- Sequencing errors (can be platform dependent)
- Number of mismatches (limitation in allowed differences)

Elements to consider

- Read type: DNA, RNA, etc.
- Read length:
- Paired-end or not
- Computational requirements (number of processors, memory)
- Base quality (taken or not into account)
- Sequencing errors (can be platform dependent)
- Number of mismatches (limitation in allowed differences)

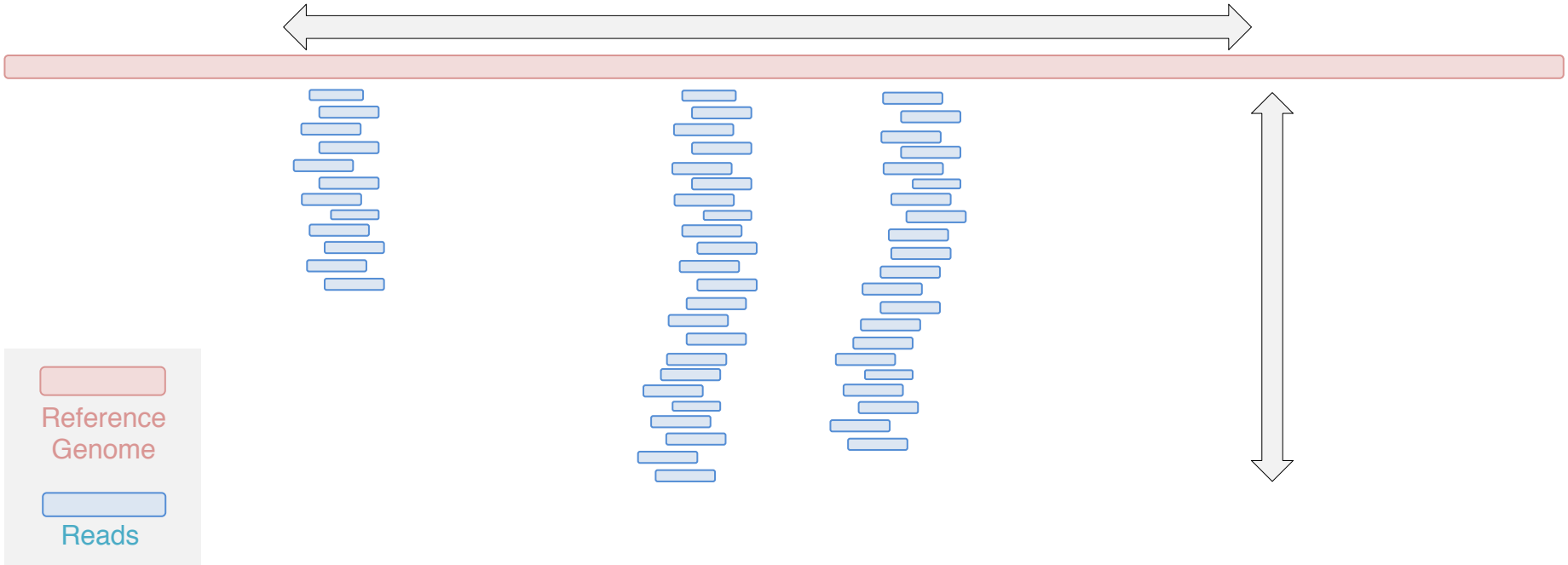
Reference Genome



Human Reference Genome

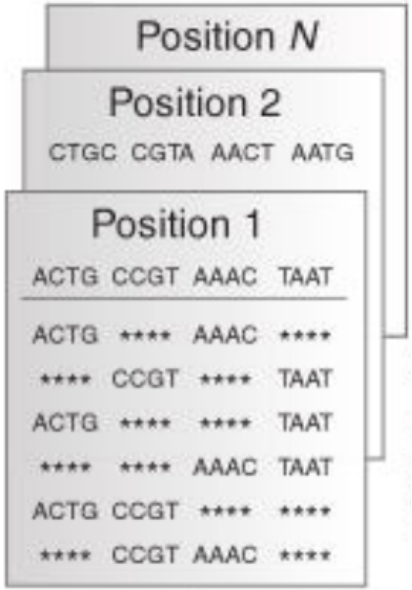
Release name	Date of release	Equivalent UCSC version	Base Pairs
GRCh38	Dec 2013	hg38	3,609,003,417
GRCh37	Feb 2009	hg19	3,326,743,047

The data dimensionality problem



Indexing

Index



Indexing allows to organize information in a more easier and faster way to search

Spaced seeds vs Burrows-Wheeler

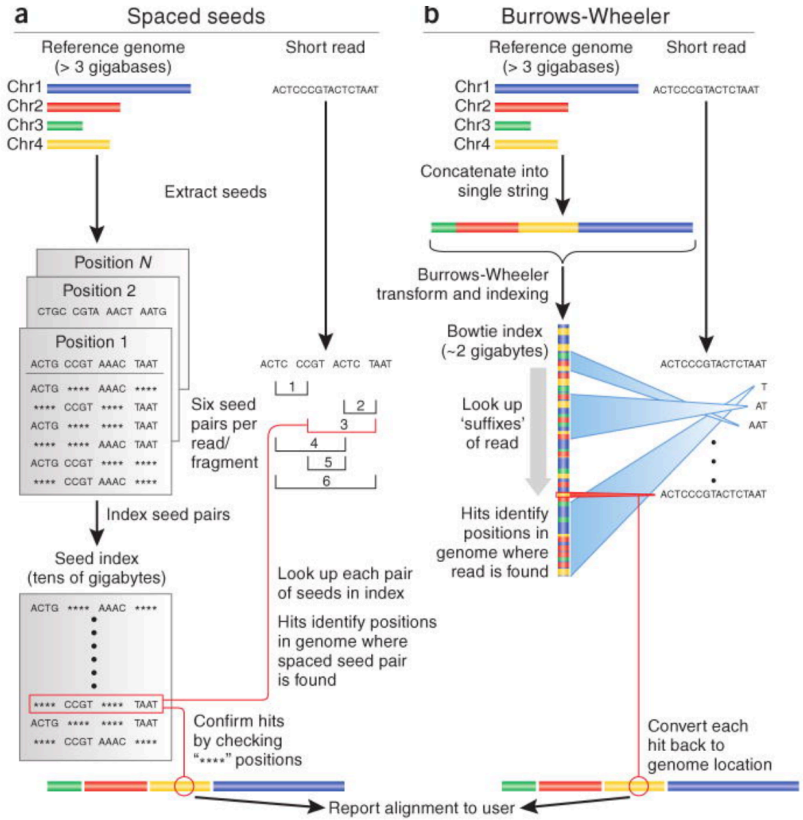
Spaced seeds
- Slower
- More mismatches allowed
- Indel detection
- Unspliced: MAQ, GSNAP
- Spliced: GMAP

Burrows-Wheeler Transform (BWT)
- Faster
- Few mismatches allowed
- Limited indel detection
- Unspliced: Bowtie, BWA
- Spliced: TopHat

More on the BW-transform on: https://youtu.be/4n7NPk5lwbI
More on FM index: https://youtu.be/kvVGj5V65io

Interesting cclass on alignments and BWT:
https://www.youtube.com/watch?v=P3ORBMon8aw&t=40s

Due to the increase in the quality of the reads and the increase in depth and coverage, **BWT aligners are more common**



Nat Biotechnol. 2009 May; 27(5): 455–457

Errors and biases

- Errors in reference sequence
- Sequencing errors:
  - Increases mismatches
  - Higher at the end of the reads
- Different regions in DNA sequence causes aligning biases:
  - Repetitive regions:
    - Similar regions in different locations
    - Place of sequencing errors
    - Place of real mutations and structural variants
  - Difficulties in the alignment of insertions/deletions (gaps)

Solutions: Quality Control Post-alignment, mapping quality scores, local realignment of indels

Thanks!

Credits for many class materials to:

Héctor Tejero: htejero@cnio.es
Elena Piñeiro: epineiro@cnio.es
Javier Perales-Patón: jperales@cnio.es

CNIO **BIOINFORMATICS** UNIT

Next Generation Sequencing

Video: https://www.youtube.com/watch?v=fCd6B5HRaZ8