



Alignments

Precision Oncology Course



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.



The diagram shows two DNA sequences side-by-side. The top sequence is ACGTCTTGACTGG - TTAAAATAC. The bottom sequence is AC - TCTTGACTGGATT AACATAC. In the alignment, the bases 'T' at position 4 and 'G' at position 5 are highlighted in yellow, indicating they are identical. The bases 'C' at position 1, 'T' at position 2, 'A' at position 3, 'A' at position 6, 'A' at position 7, 'A' at position 8, and 'T' at position 9 are highlighted in red, indicating they are also identical.

Alignments

Elements of an alignment

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.

The diagram shows two sequences aligned vertically:

ACGTTTTGCAGTAAATGC~~GG~~**ACT**GA - T
ACGTT**G**TGCAGTAAATGC~~GG~~GA -- **G**~~A~~**C**T

Below the sequences, three types of alignment elements are indicated:

- A green arrow pointing down between the first two bases of the top sequence is labeled "mismatch".
- A black arrow pointing down between the 10th and 11th bases of the top sequence is labeled "match".
- Two red arrows pointing down between the 12th and 13th positions of the top sequence are labeled "gap deletion/insertion".

Alignments

Elements of an alignment

ACGTTTGCAGTAAATGCGGA~~CTGAT~~
ACGTTGTCAGTAAATGCGGA-~~GACT~~

→ 1 gap

ACGTTTGCAGTAAATGCGGA~~CTGAT~~
ACGTTGTCAGTAAATGCGGA -- GACT

→ 1 extended gap

ACGTTTGCAGTAAATGCGGA~~CTGA~~-T
ACGTTGTCAGTAAATGCGGA --~~GACT~~

→ 2 gaps

Alignments

Elements of an alignment

1. Based on the number of sequences:

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

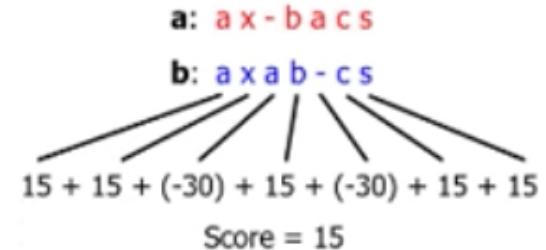
2. Based on the **region** to align:

- **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 = xy**a**xbacsl, **T** = pqr**a**bcs**t**vq

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:

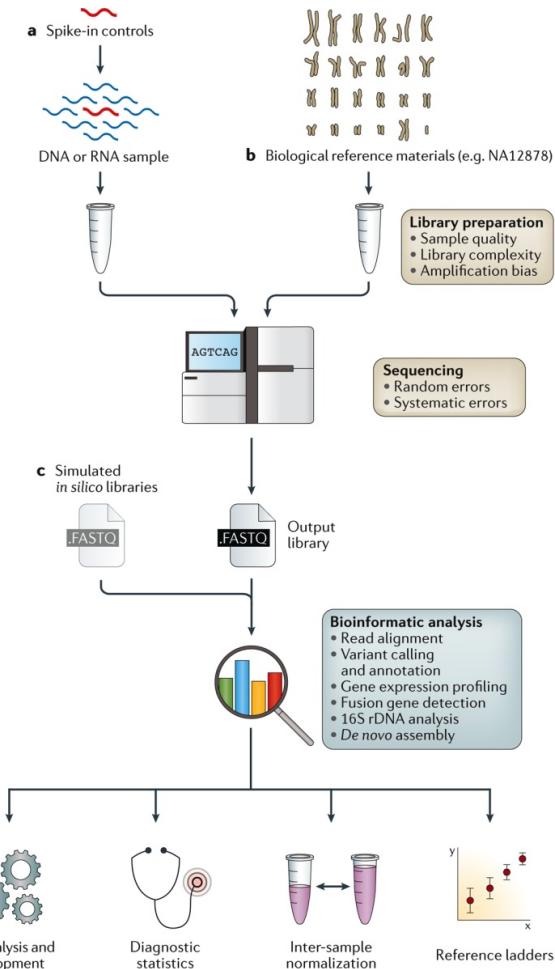
1. Determine the **homology** degree
2. Identify **functional domains**
3. Compare the gene with its product
4. Find homologous **positions**
5. Identify **differences**



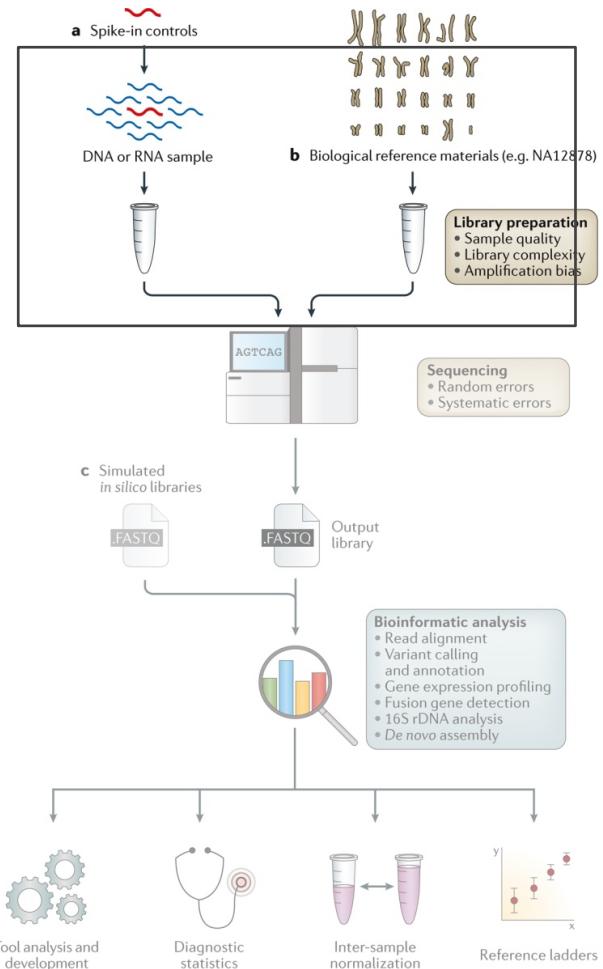
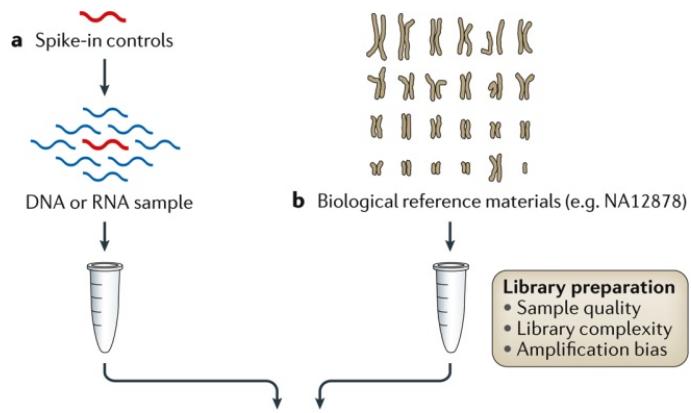
Differential Expression and Variant Detection in Next Generation Sequencing (NGS)

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**

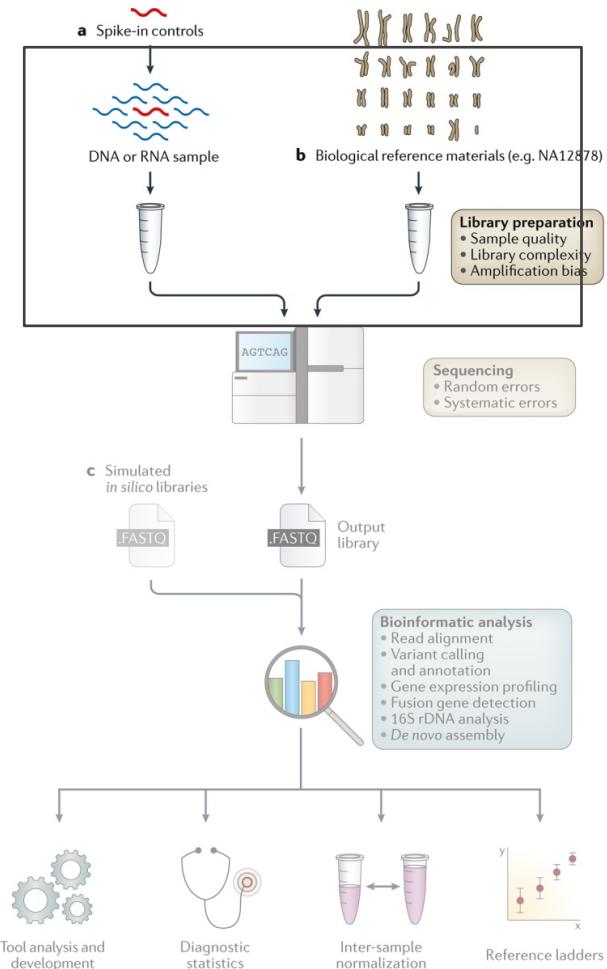
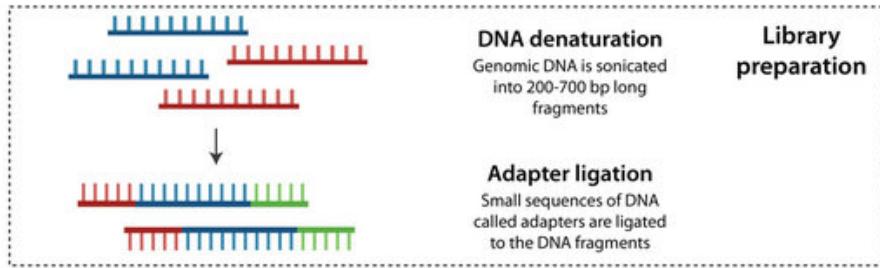


Next Generation Sequencing

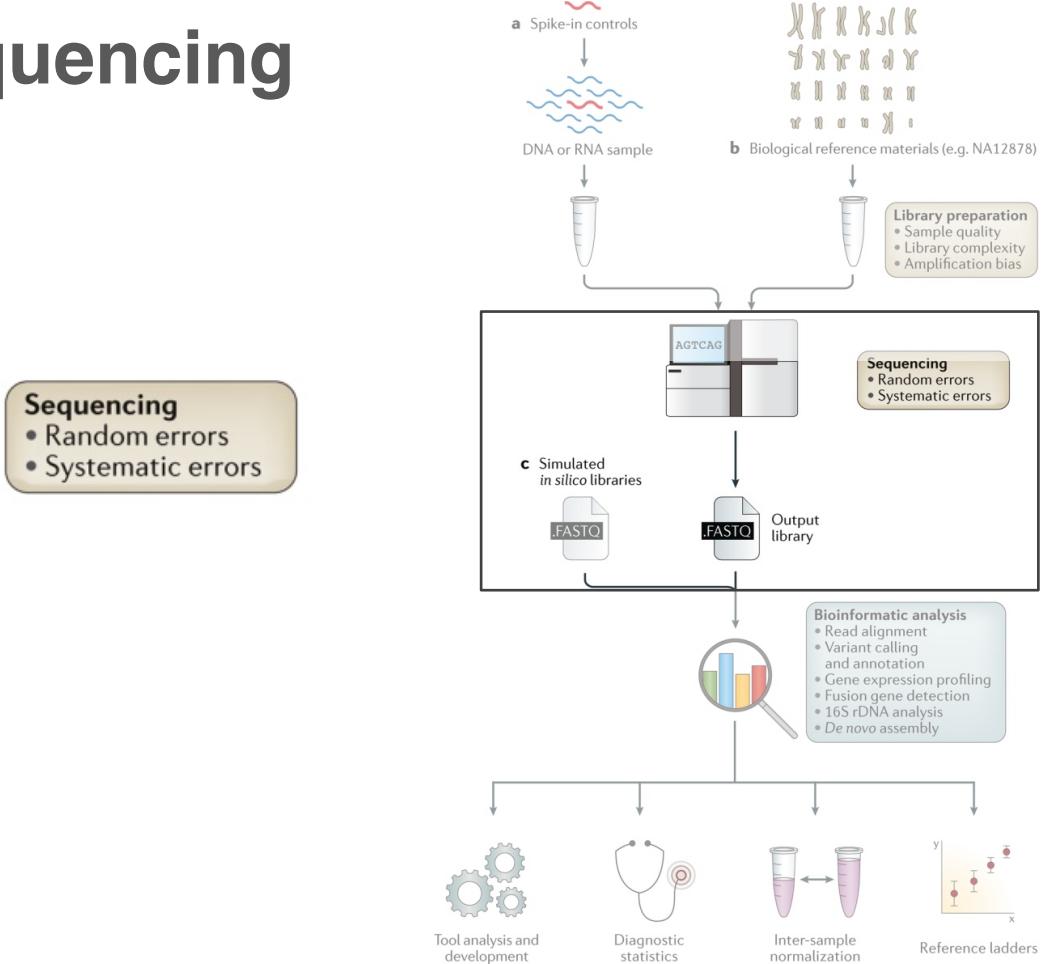
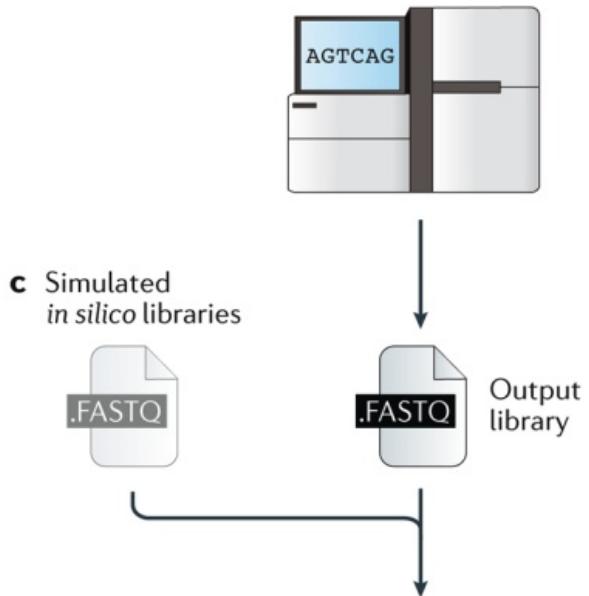


Next Generation Sequencing

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



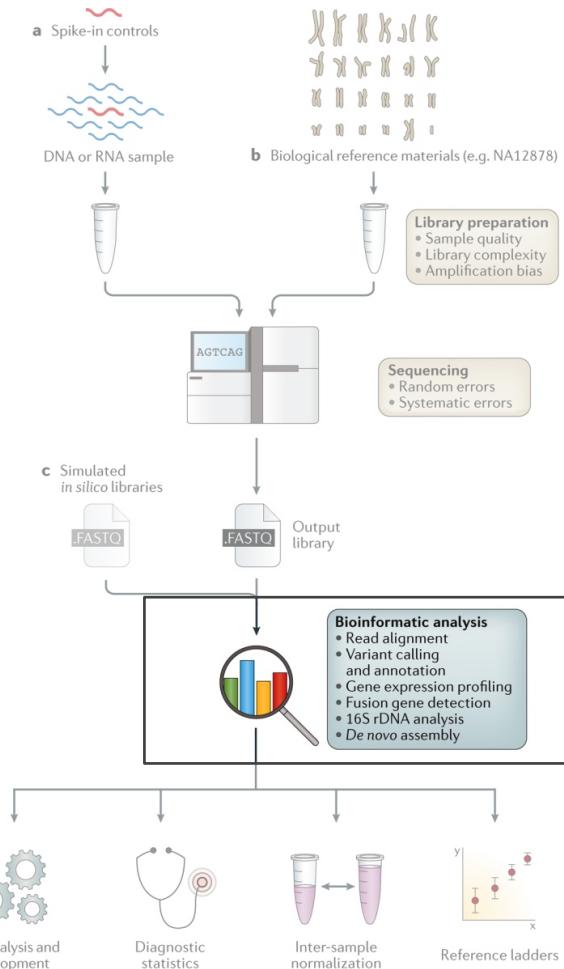
Next Generation Sequencing



Next Generation Sequencing

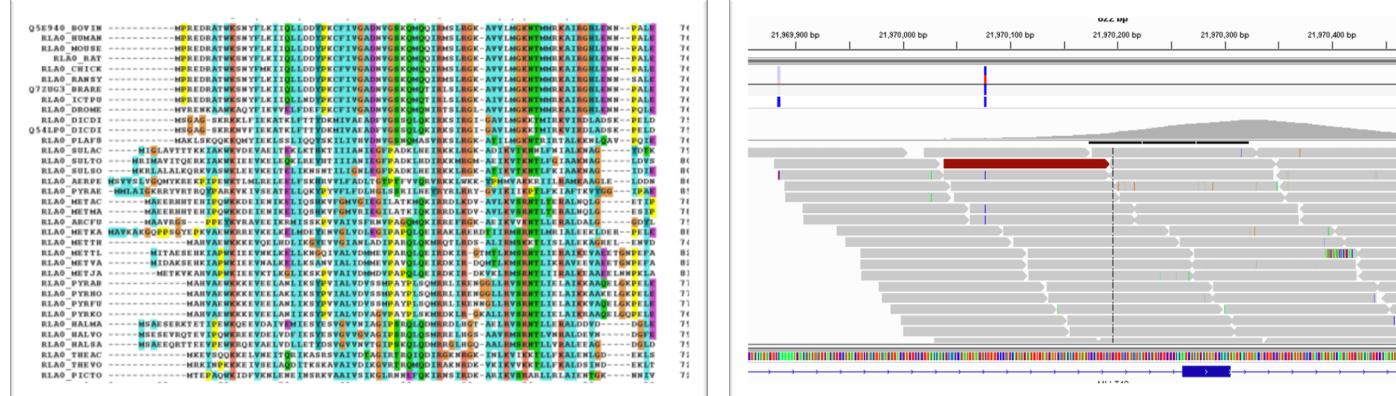


- Bioinformatic analysis**
- Read alignment
 - Variant calling and annotation
 - Gene expression profiling
 - Fusion gene detection
 - 16S rDNA analysis
 - *De novo* assembly



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 (l = 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

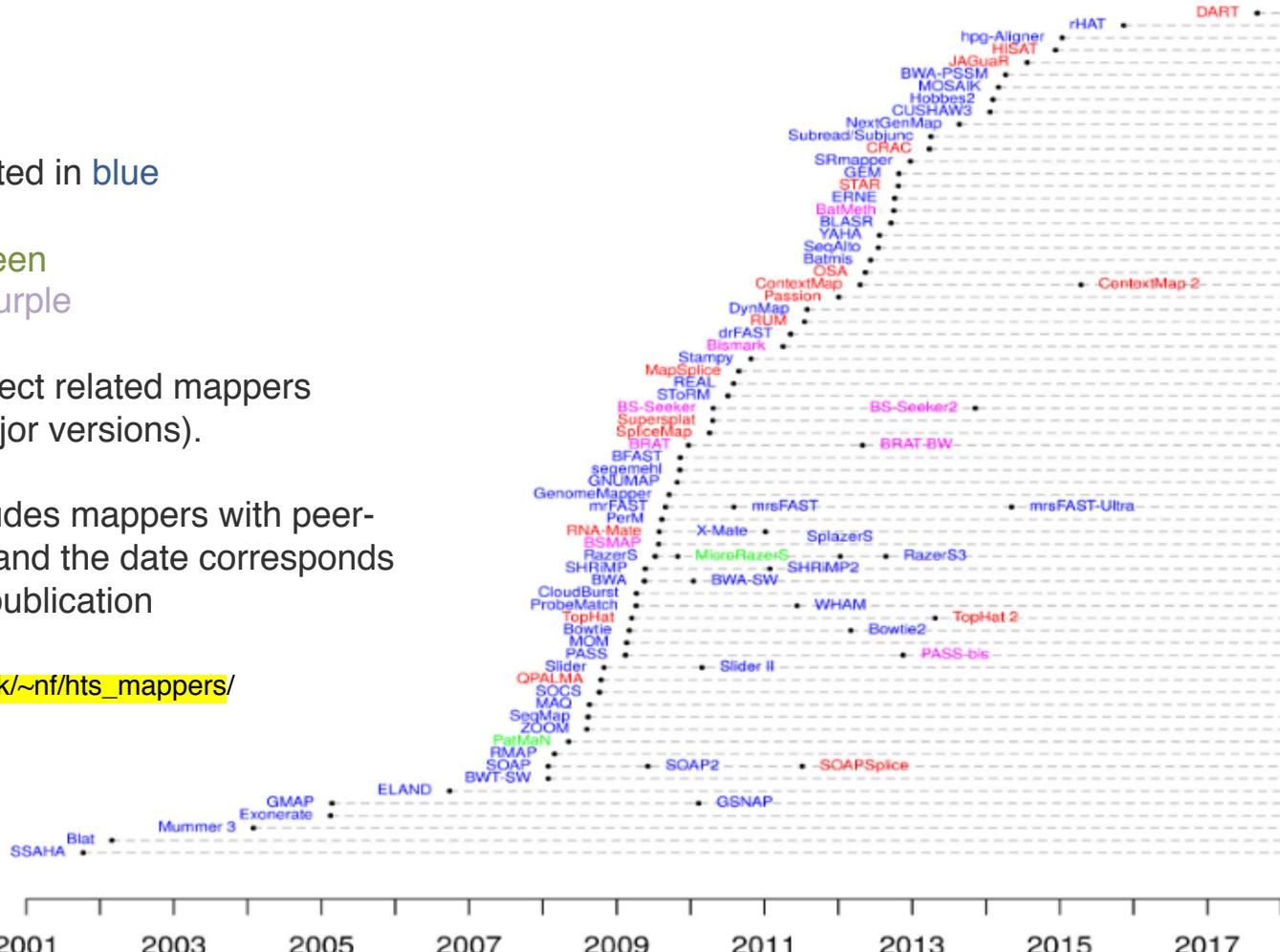
miRNA mappers in green

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/



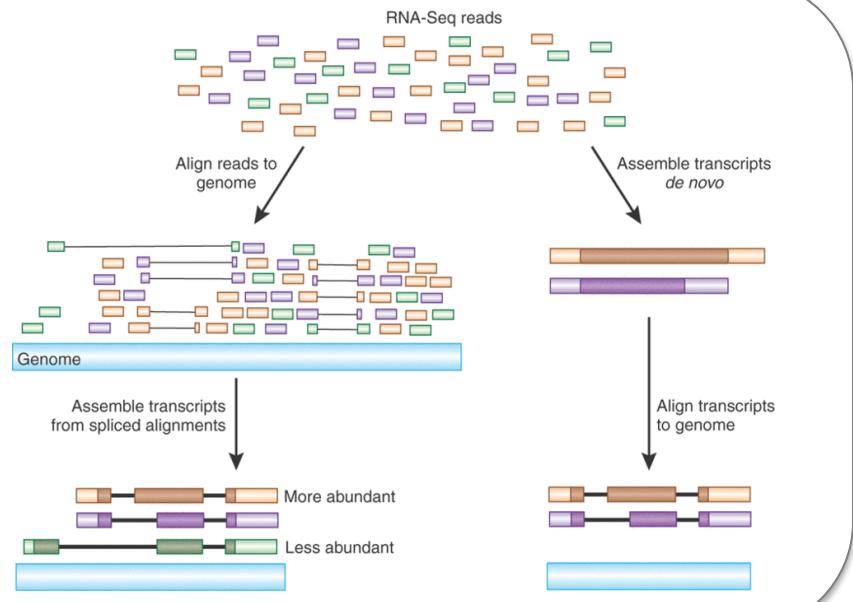
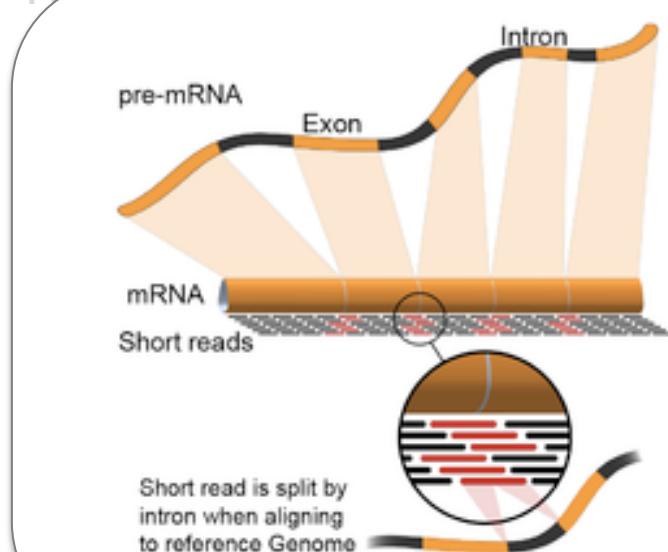
Elements to consider

- 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



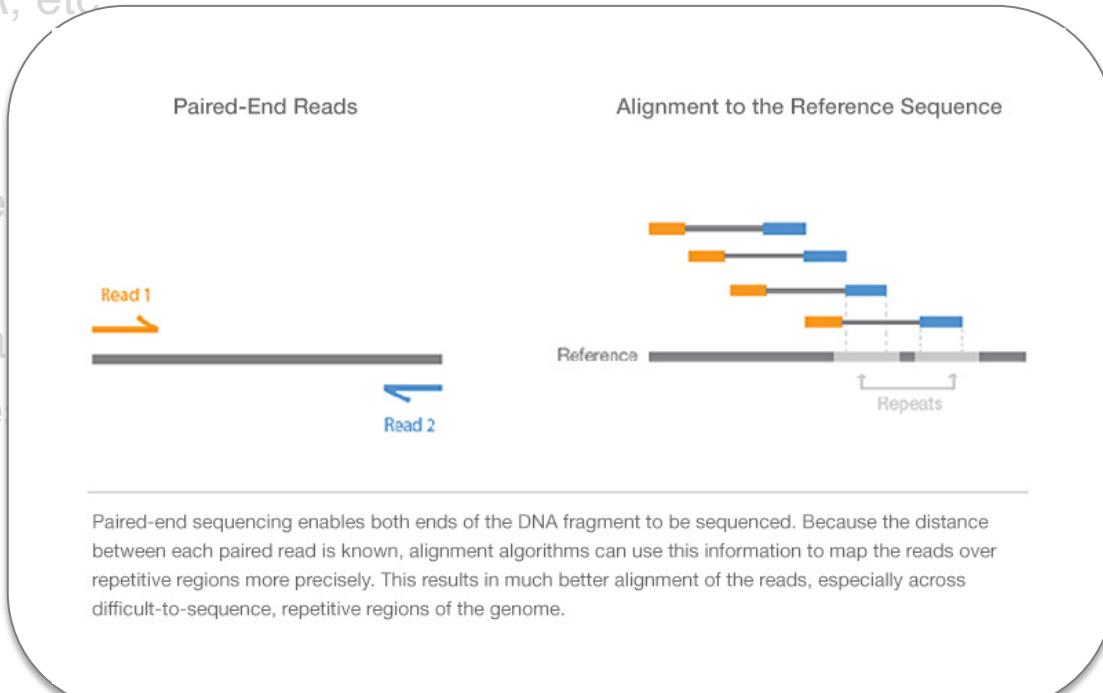
Elements to consider

- 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 requirements
- Base quality (taken or not)
- Sequencing errors (calling)
- Number of mismatches



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)

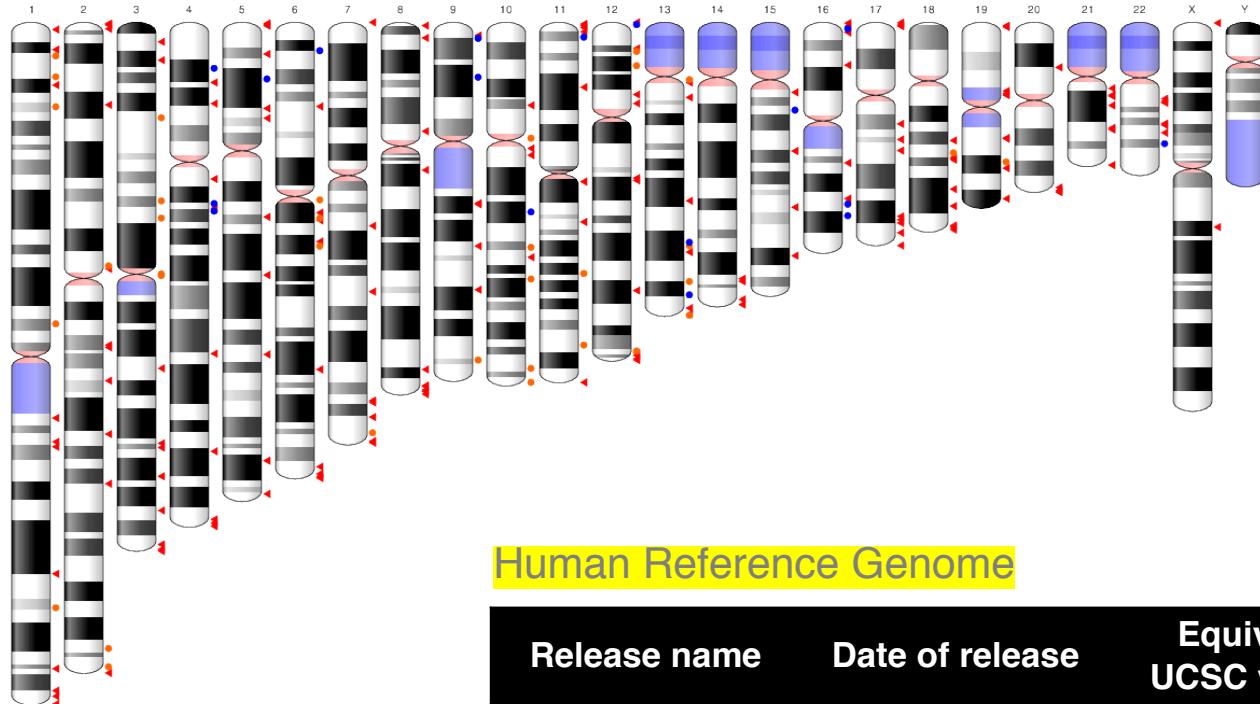
Elements to consider

- 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

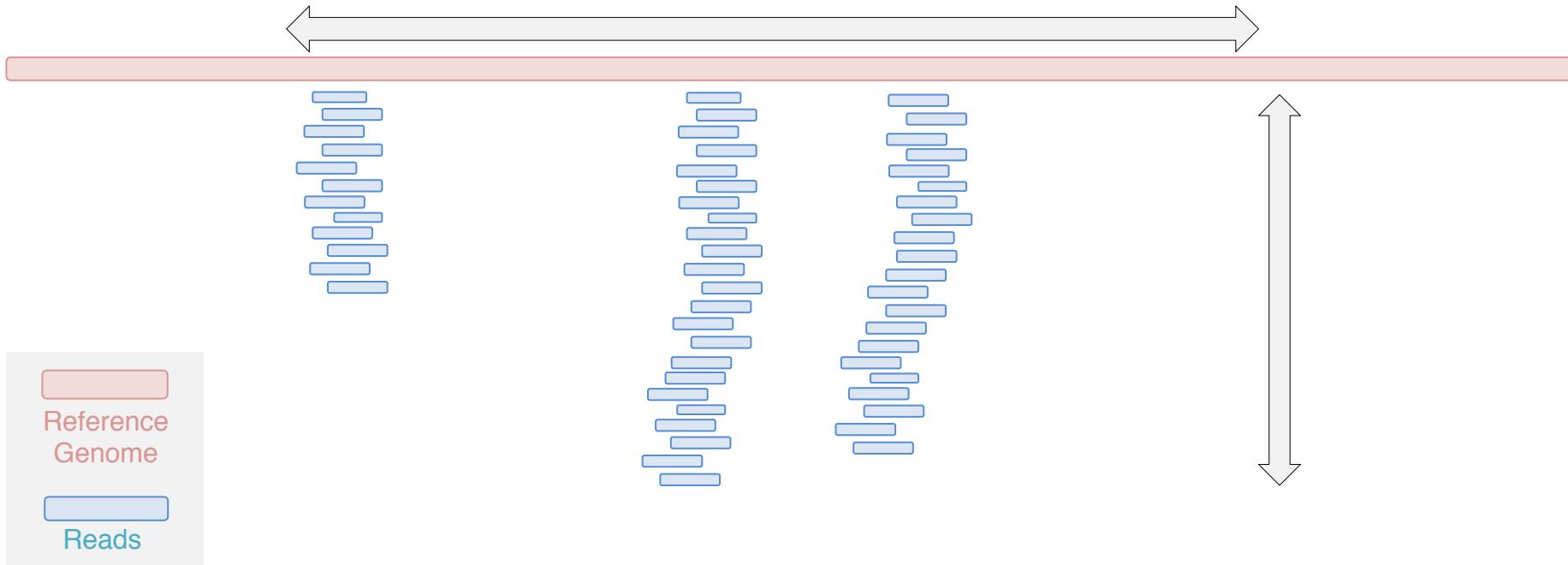


Human Reference Genome

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

- ◀ Region containing alternate loci
- Region containing fix patches
- Region containing novel patches

The data dimensionality **problem**



Indexing

Index

A

Additive color model, 3

B

Binding, 20

Bitmap image

defined, 9

resolution of, 11

tonal range in, 11

Bleed, checking, 46

Blueline, 42

C

Chroma, 2

CMS. *See* Color management

system

CMY color model, 4

Color

characteristics of, 2

checking definitions, 46

Color proof

checking, 50

contract, 40, 50

separation-based, 40

Color separations. *See* Separations

Color space, 4

Color value, 2

Commercial printing

inking, 18

offsetting, 19

platemaking, 17

press check, 40–43, 51

terminology, 5–8

types of, 16–??

wetting, 18

Continuous-tone art

defined, 5

Contract proof, 40, 50

Creep, 21

G

Gamut, color, 4

GCR (gray-component replacement), 7

Gravure, 22

Gray, shades of, 13

H

Halftone cell, 13

Halftone dot, 5, 13

Halftone frequency, 12

Halftone screen

defined, 5

moiré patterns, 9, 15

process colors, 6

Hand-off, 37

creating report for, 43

organizing files for, 46

High-fidelity color, 15

Hue, 2

Position N

Position 2

CTGC CGTA AACT AATG

Position 1

ACTG CCGT AAAC TAAT

ACTG ***** AAAC *****

***** CCGT ***** TAAT

ACTG ***** ***** TAAT

***** ***** AAAC TAAT

ACTG CCGT ***** *****

***** CCGT AAAC *****

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

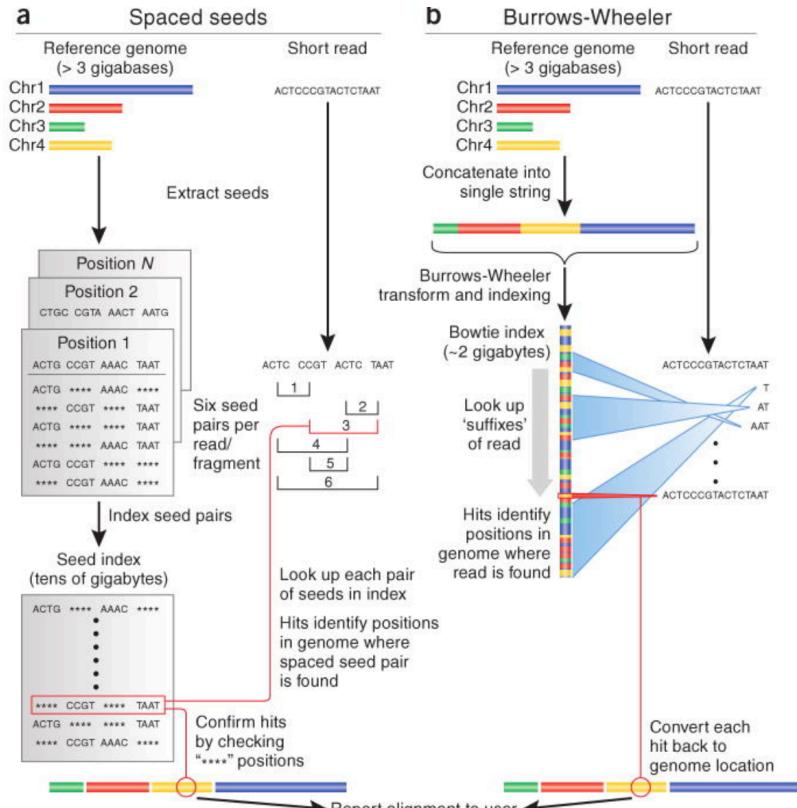
More on the BW-transform on: <https://youtu.be/4n7NPk5lwbl>

More on FM index: <https://youtu.be/kvVGj5V65io>

Interesting class 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

Next Generation Sequencing

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

