GULBENKIAN CIÊNCIA

PO: Precision Oncology Course

Alignments

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer

Alignments

What is an alignment?

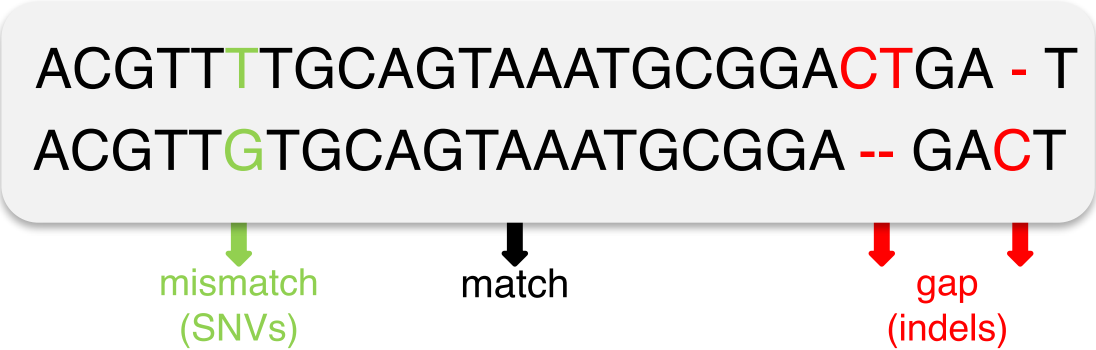
Sequence alignments are a way of arranging the sequences of DNA, RNA or proteins in order to **identify regions of similarity** that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

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

Alignments

Elements of an alignment

Alignment seeks to **reduce gaps and mismatches** and **maximize matches.**



```
ACGTTTTGCAGTAAATGCGGACTGA - T
ACGTTGTGCAGTAAATGCGGA -- GACT
```

mismatch (SNVs)

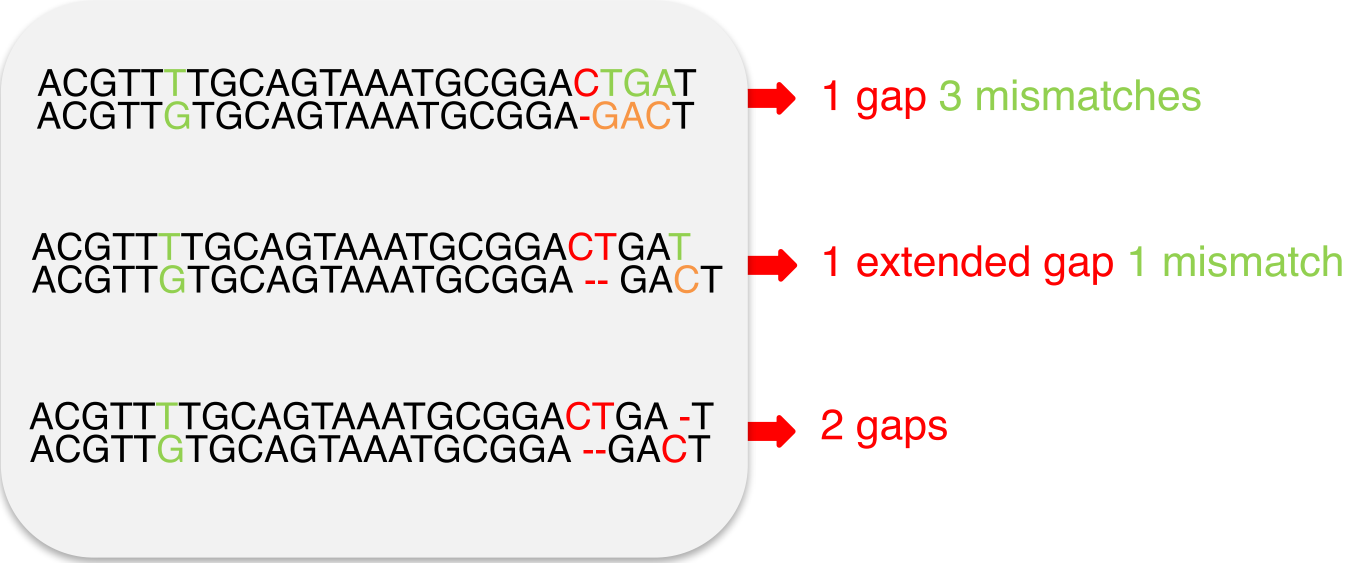
match

gap (indels)

Alignments

Elements of an alignment

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.



ACGTTTTGCAGTAAATGCGGACTGAT
ACGTTGTGCAGTAAATGCGGA-GACT

1 gap 3 mismatches

ACGTTTTGCAGTAAATGCGGACTGAT
ACGTTGTGCAGTAAATGCGGA -- GACT

1 extended gap 1 mismatch

ACGTTTTGCAGTAAATGCGGACTGA -T
ACGTTGTGCAGTAAATGCGGA --GACT

2 gaps

Alignments

Objectives

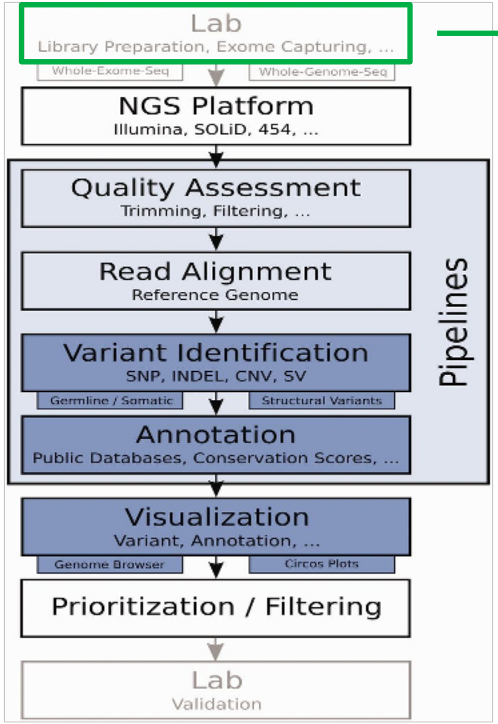
The comparison between sequences in sequence alignment allows to:

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

Objective of Variant Calling in Next Generation Sequencing (NGS)

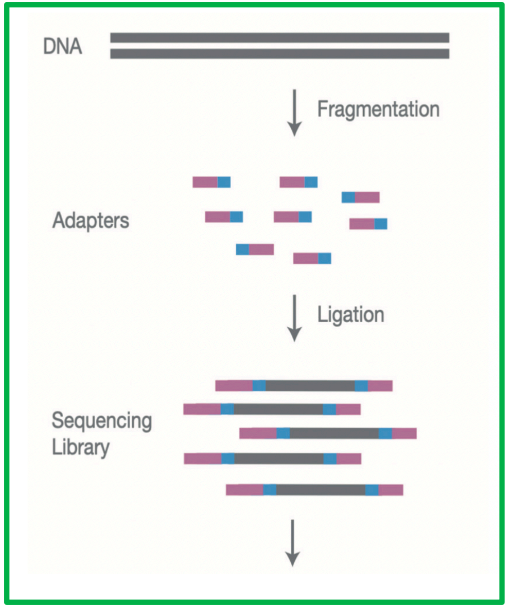
Next Generation Sequencing (NGS)

Workflow



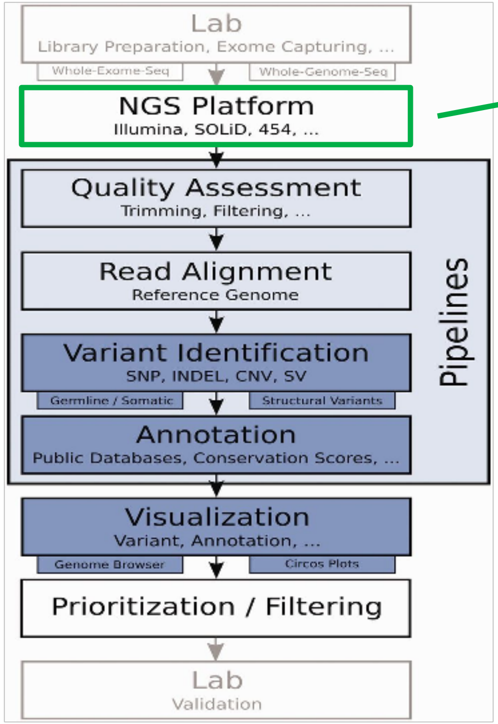
Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Library Preparation:
DNA is fragmented into small pieces to form a **library.**



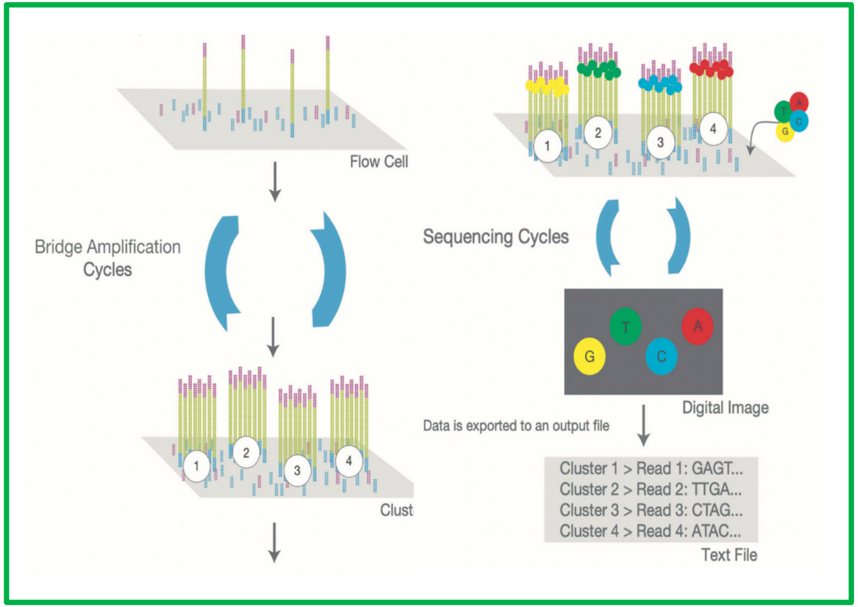
Next Generation Sequencing (NGS)

Workflow



Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

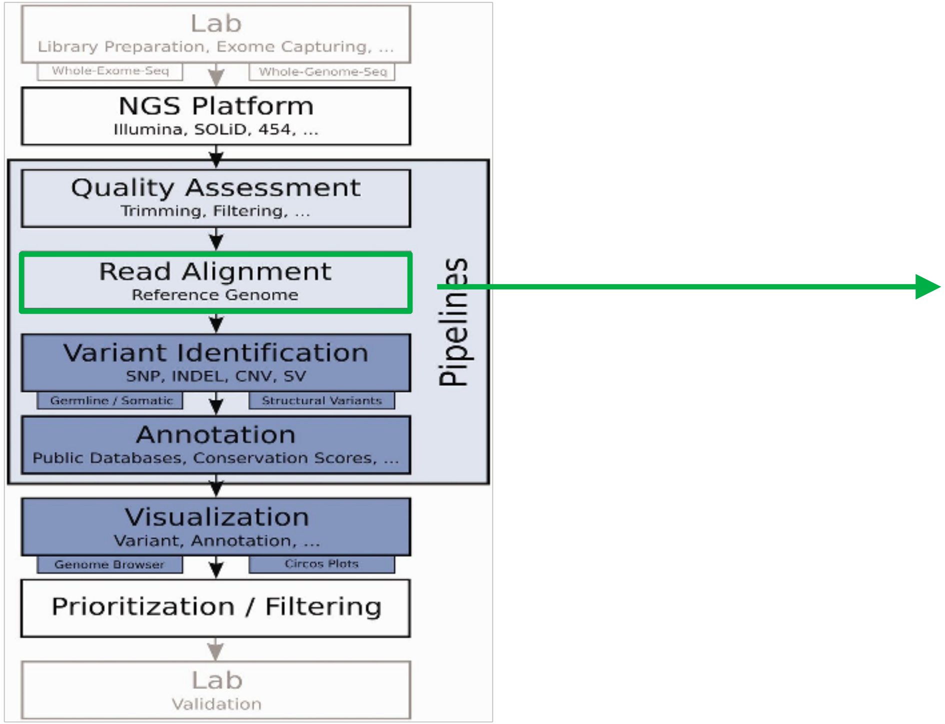
Sequencing:
Fragments are sequenced and stored as **reads** in text files.


Flow Cell
Bridge Amplification Cycles
Clust
1
2
3
4
Sequencing Cycles
G
T
C
A
Digital Image
Data is exported to an output file
Text File

```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

Next Generation Sequencing (NGS)

Workflow

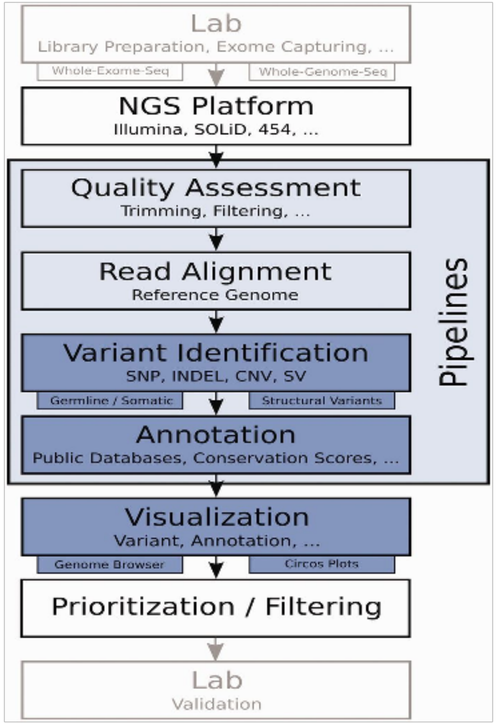


Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Alignment:
Align the reads against the reference genome in order to find out where do they come from.

Next Generation Sequencing (NGS)

Workflow



Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Video: https://youtu.be/fCd6B5HRaZ8

Short Read Aligners (SRA)

Challenges

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**

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

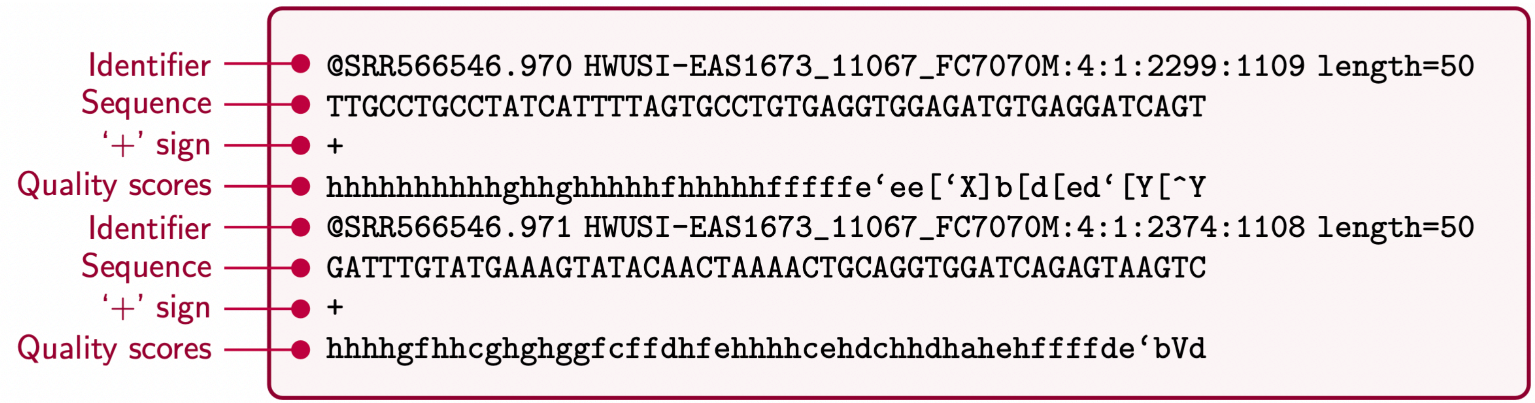
Different NGS technologies have different error profiles to take into account:

- **Roche 454:** Insertions or deletions in homopolymer runs
- **Illumina:** Increasing likelihood of sequence errors towards the end of the read

Alignment

Reads

Short reads are stored in **text files** with the extension **FASTQ.**



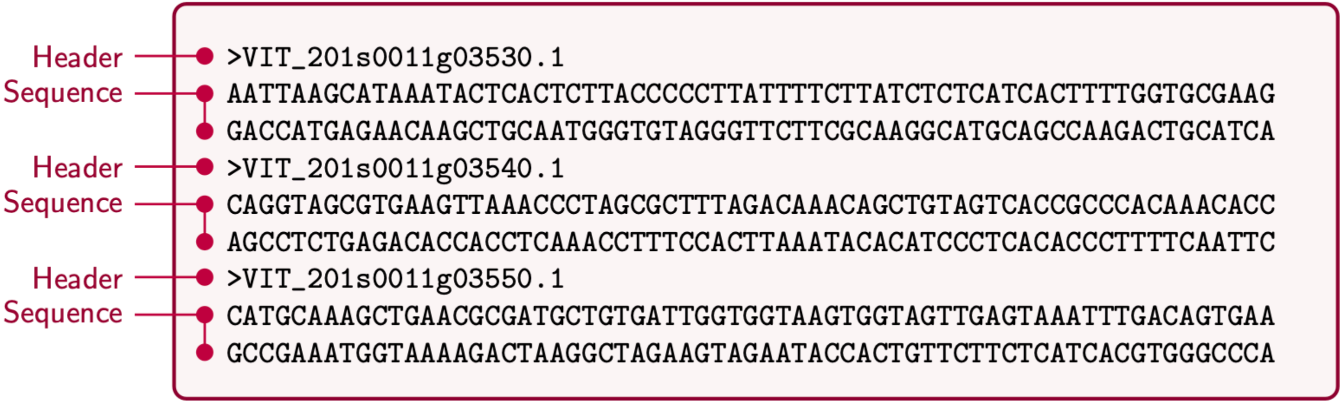
```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhghhghhhhhfhhhhhfffffe`ee[`X]b[d[ed`[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde`bVd
```

Hosseini, M.; Pratas, D.; Pinho, A.J. A Survey on Data Compression Methods for Biological Sequences. *Information* **2016**, *7*, 56. doi: 10.3390/info7040056

Alignment

Reference genome

The reference genome is stored in another **text file** with the extension **FASTA.**

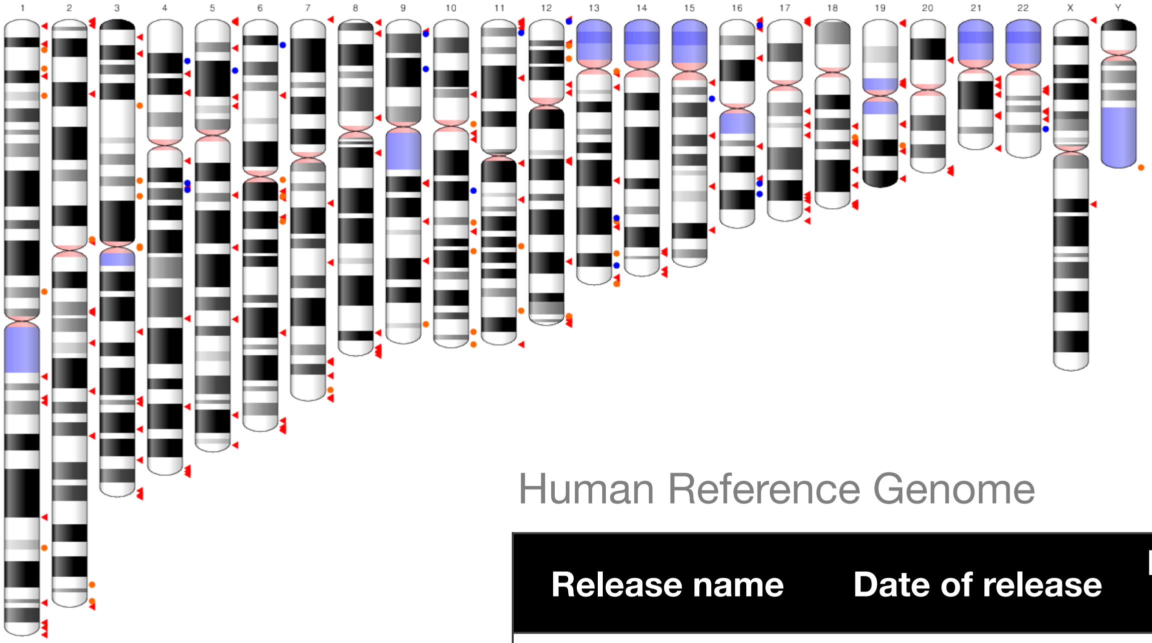


```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

Hosseini, M.; Pratas, D.; Pinho, A.J. A Survey on Data Compression Methods for Biological Sequences. *Information* **2016**, *7*, 56. doi: 10.3390/info7040056

Alignment

Reference genome

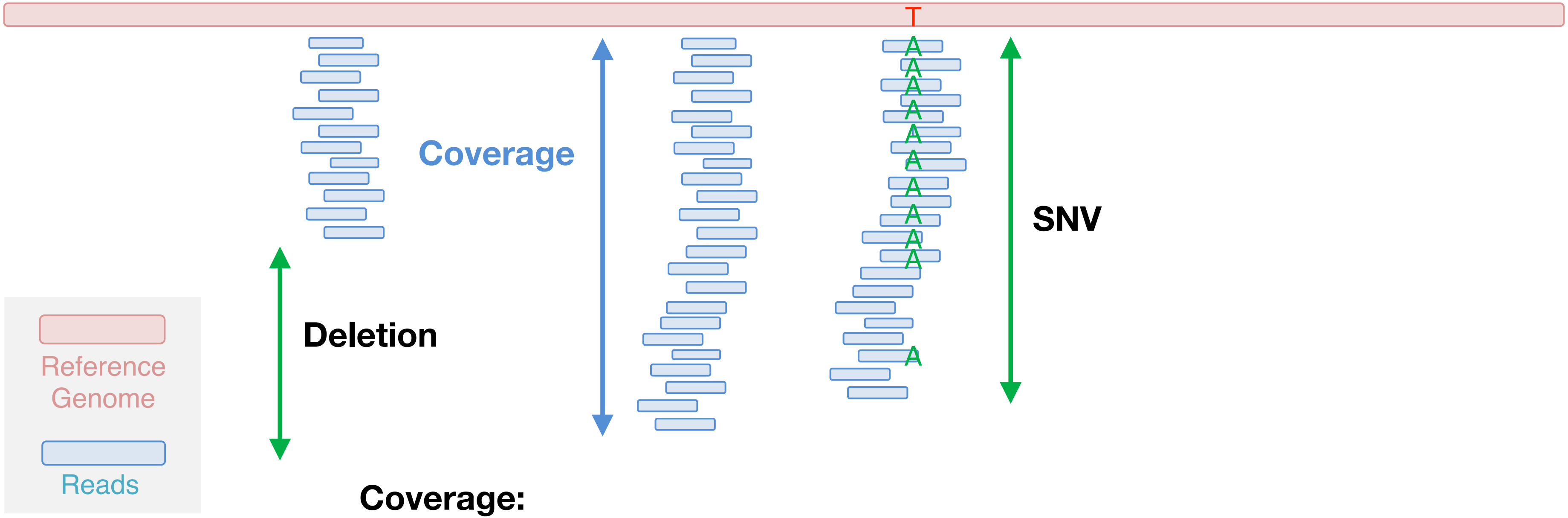

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
◀ Region containing alternate loci
● Region containing fix patches
● Region containing novel patches

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

Alignment

Reads are mapped against the reference genome in order to identify differences

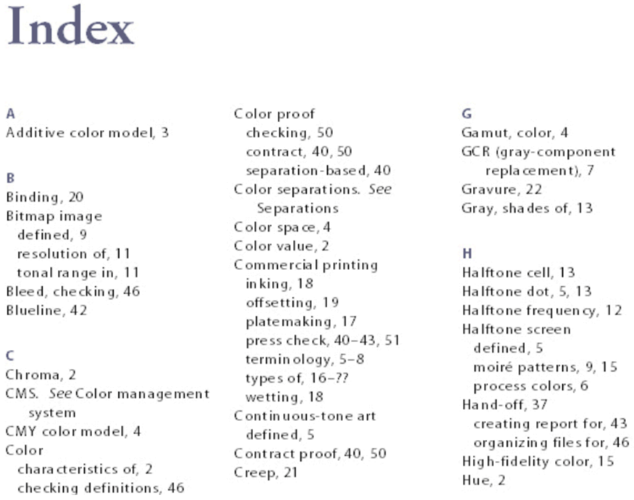


Coverage:
Number of reads that align to the same region (60x, 100x,…)

Alignment

Indexing

In order to optimize the alignment process, these algorithms use a strategy called **indexing. As in books,** indexing allows to organize information in a more easier and faster way to search.



Index

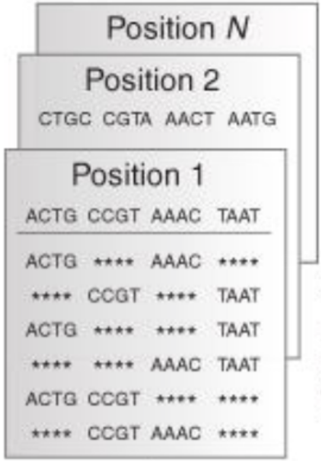
A
Additive color model, 3

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 ****

Alignment

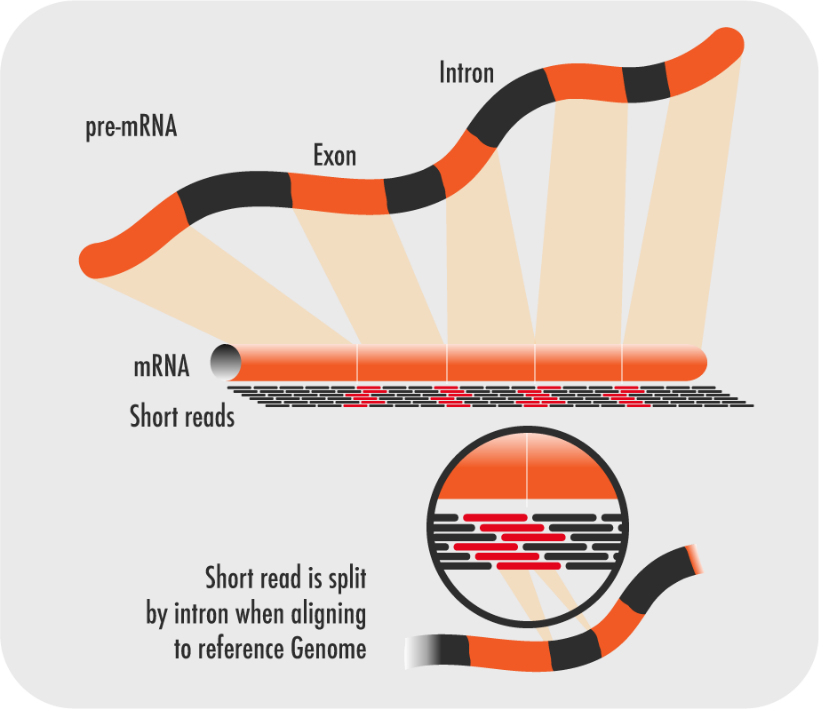
Elements to consider

- Read type: DNA, RNA, etc.
- Read length
- Paired-end or single-end
- Computational requirements (number of processors, memory)
- Number of mismatches (limitation in allowed differences)
- Sequencing errors (can be platform dependent)

Alignment

Elements to consider

- Read type: DNA, RNA, etc.



Mackenzie RJ. RNA-seq: Basics, Applications and Protocol. *Technology Networks Genomic Research*, **2018**. https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

Alignment

Elements to consider

- Read type: DNA, RNA, etc.
- Read length
- Paired-end or single-end
- Computational requirements (number of processors, memory)
- Number of mismatches (limitation in allowed differences)
- Sequencing errors (can be platform dependent)

Alignment

Elements to consider

- Read length

The longer the reads, the higher the likelihood of sequencing errors towards the end.

Alignment

Elements to consider

- Read type: DNA, RNA, etc.
- Read length
- Paired-end or single-end
- Computational requirements (number of processors, memory)
- Number of mismatches (limitation in allowed differences)
- Sequencing errors (can be platform dependent)

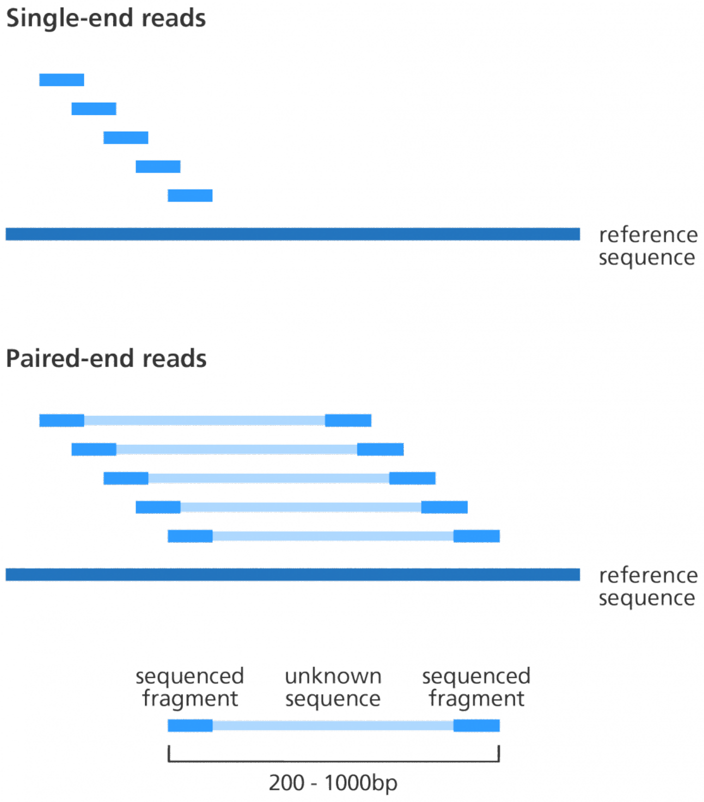
Alignment

Elements to consider

- Paired-end or single-end

In a **paired-end** experiment, **both ends** of the fragment are sequenced.

The **distance between** the each paired read **is known (insert size)**, thus the aligners can **map** the reads **more precisely.** Important for correctly mapping **repetitive regions** of the genome.



NGS overlapping reads. Biostars. Post 241139.

Alignment

Elements to consider

- Read type: DNA, RNA, etc.
- Read length
- Paired-end or single-end
- Computational requirements (number of processors, memory)
- Number of mismatches (limitation in allowed differences)
- Sequencing errors (can be platform dependent)

Alignment

Elements to consider

- Read type: DNA, RNA, etc.
- Read length
- Paired-end or single-end
- Computational requirements (number of processors, memory)
- Number of mismatches (limitation in allowed differences)
- Sequencing errors (can be platform dependent)

Alignment

Elements to consider

- Read type: DNA, RNA, etc.
- Read length
- Paired-end or single-end
- Computational requirements (number of processors, memory)
- Number of mismatches (limitation in allowed differences)
- Sequencing errors (can be platform dependent)

Alignment

Errors and biases

- Sequencing errors:
  - Increases mismatches
  - Higher at the end of the reads
  - Technology-dependent
- Different regions in the DNA sequence cause aligning biases:
  - Repetitive regions:
    - Found in different locations
    - Place of sequencing errors
    - Place of real mutations and structural variants
  - Difficulties in the alignment of indels (gaps)

Alignment

Solutions

Post-alignment Quality Control and local realignment of indels.

More about indel realignment: https://qcb.ucla.edu/wp-content/uploads/sites/14/2016/03/GATKwr12-3-IndelRealignment.pdf

GULBENKIAN CIÊNCIA

Thanks!

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer