(First) steps in NGS data analysis

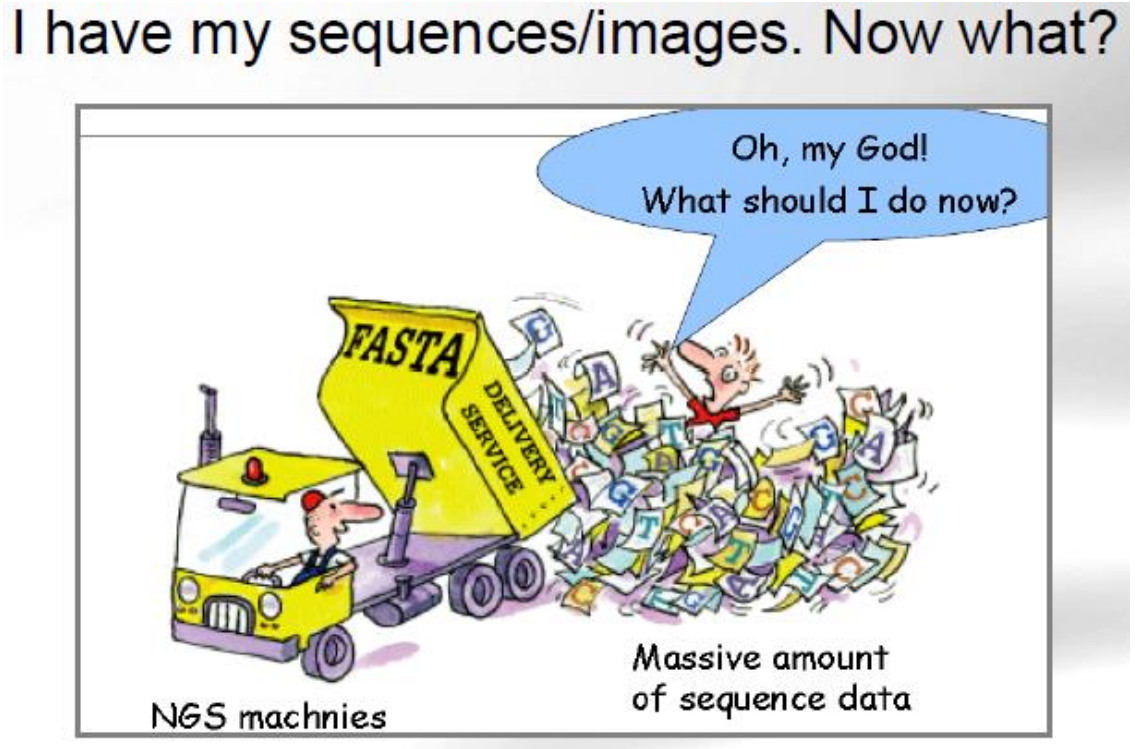
Bioinformatics Course UEB-VHIR
November 2023

Mireia Ferrer[1], Álex Sánchez[1,2], Esther Camacho[1], Berta Miró[1]

[1] Unitat d'Estadística i Bioinformàtica (UEB) VHIR
[2] Departament de Genètica Microbiologia i Estadística, UB

Steps in NGS analysis

Bioinformatics challenges of NGS



Steps in NGS analysis

Bioinformatics challenges of NGS

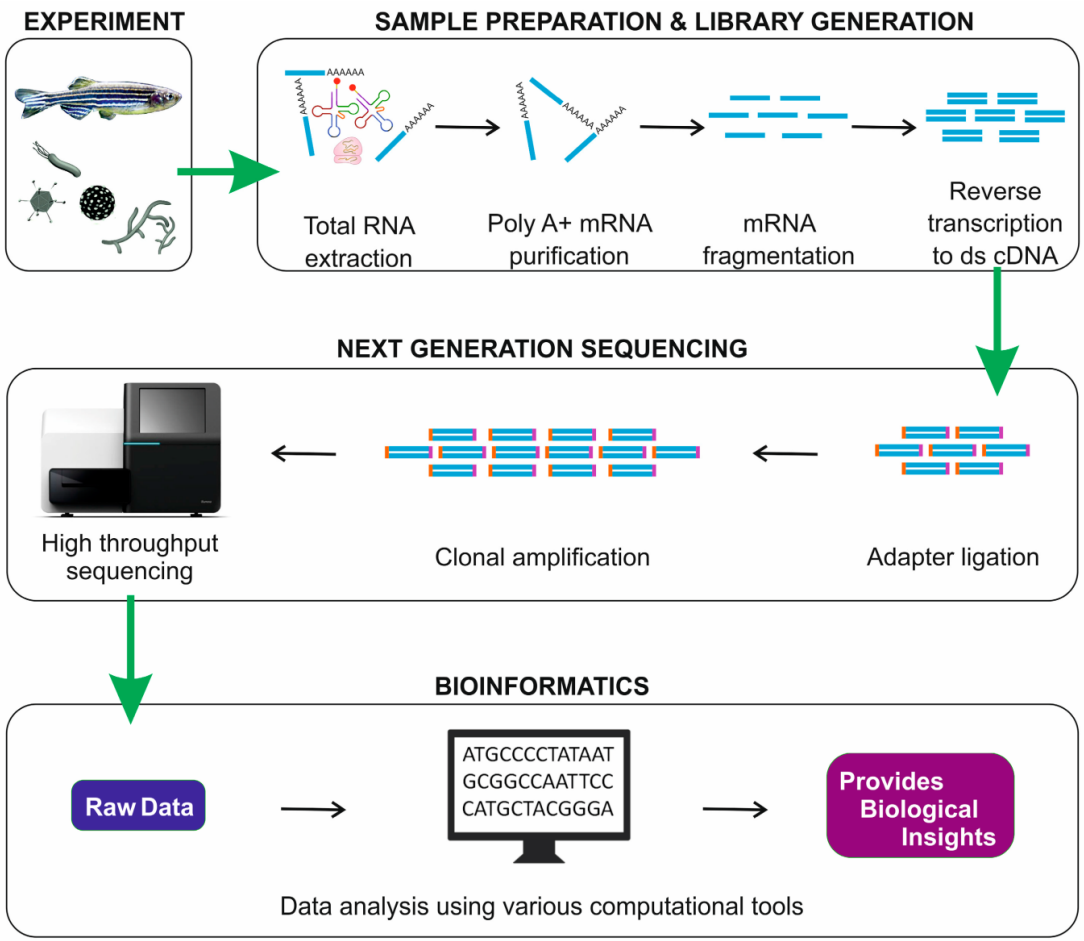
A single sequencing experiment can generate 100's of millions of reads, 10's to 100's gigabytes of data.

We need:

- Huge data storage and transfer technology
- Algorithms for managing, analyzing and visualizing data
- Reproducible workflows and standards for analysis
- Specialized tools for integrating various data types

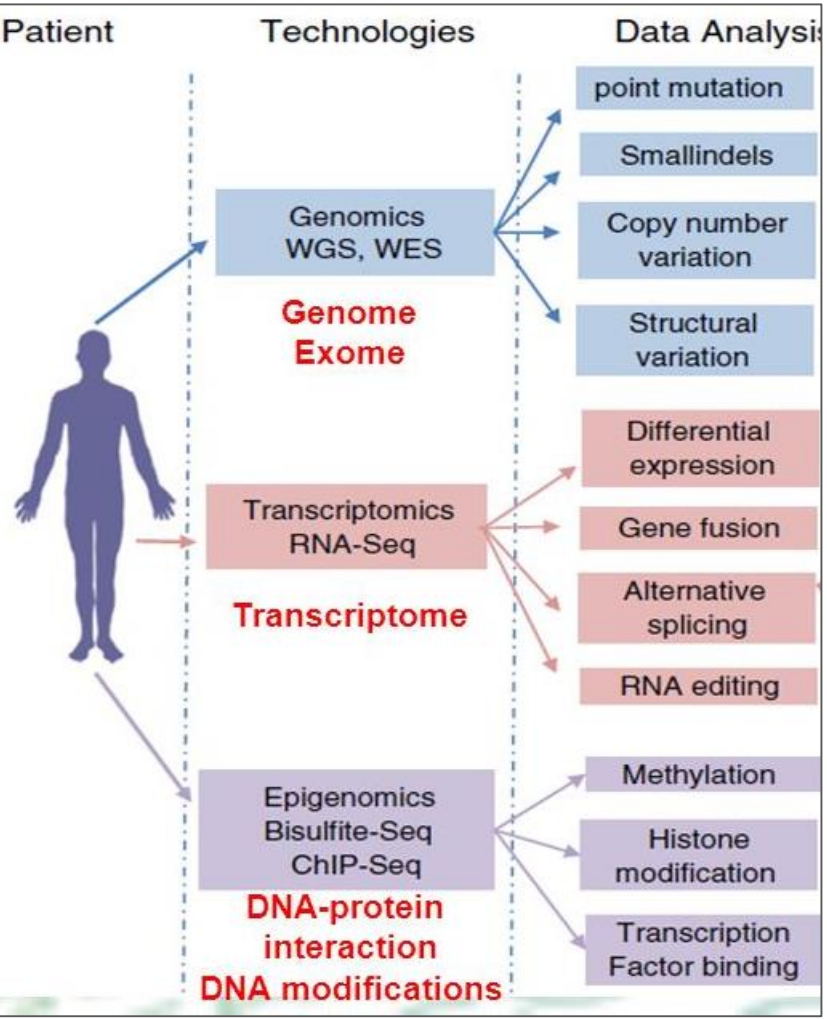
Steps in NGS analysis

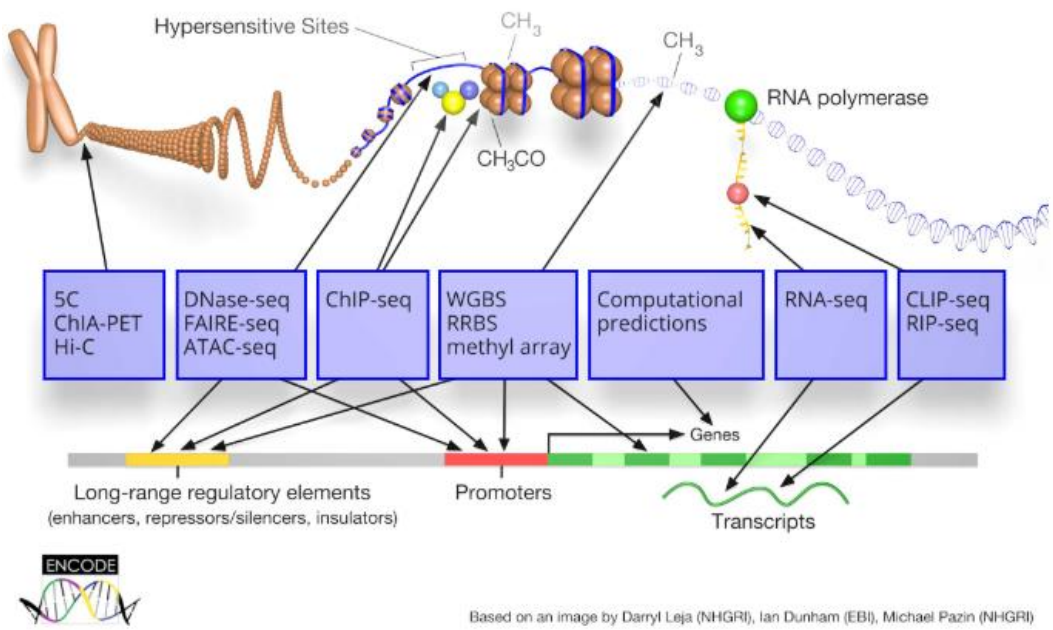
General workflow



Steps in NGS analysis

Applications



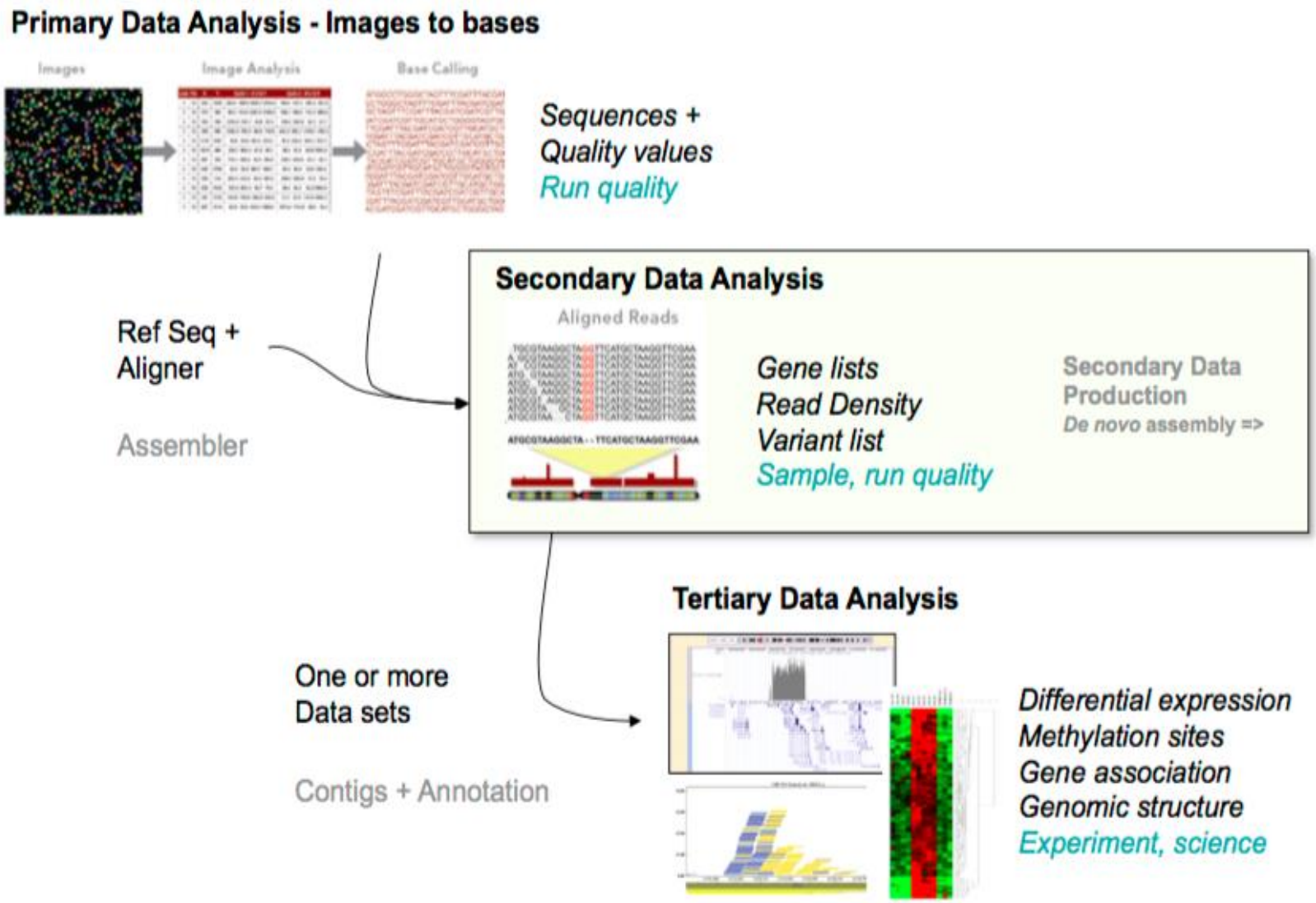


- But also...
  - Metagenomics
  - *De novo* genome assembly

More info: http://allseq.com/kb-category/applications/

Steps in NGS analysis

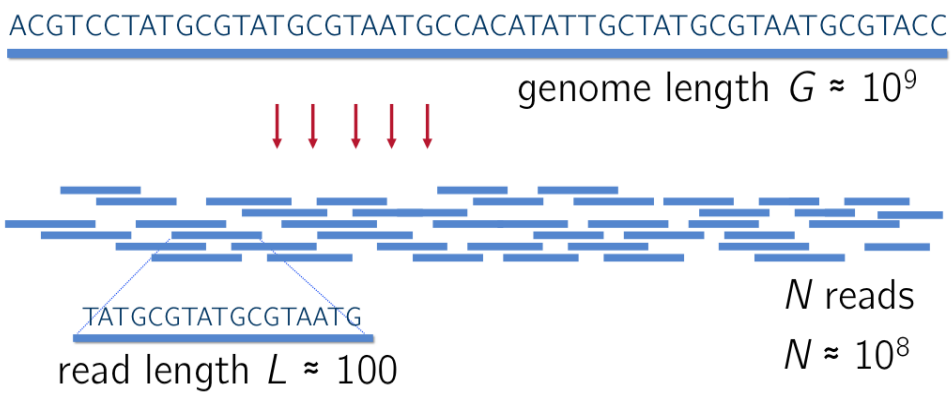
- NGS data is analyzed in three stages



Steps in NGS analysis

Terminology

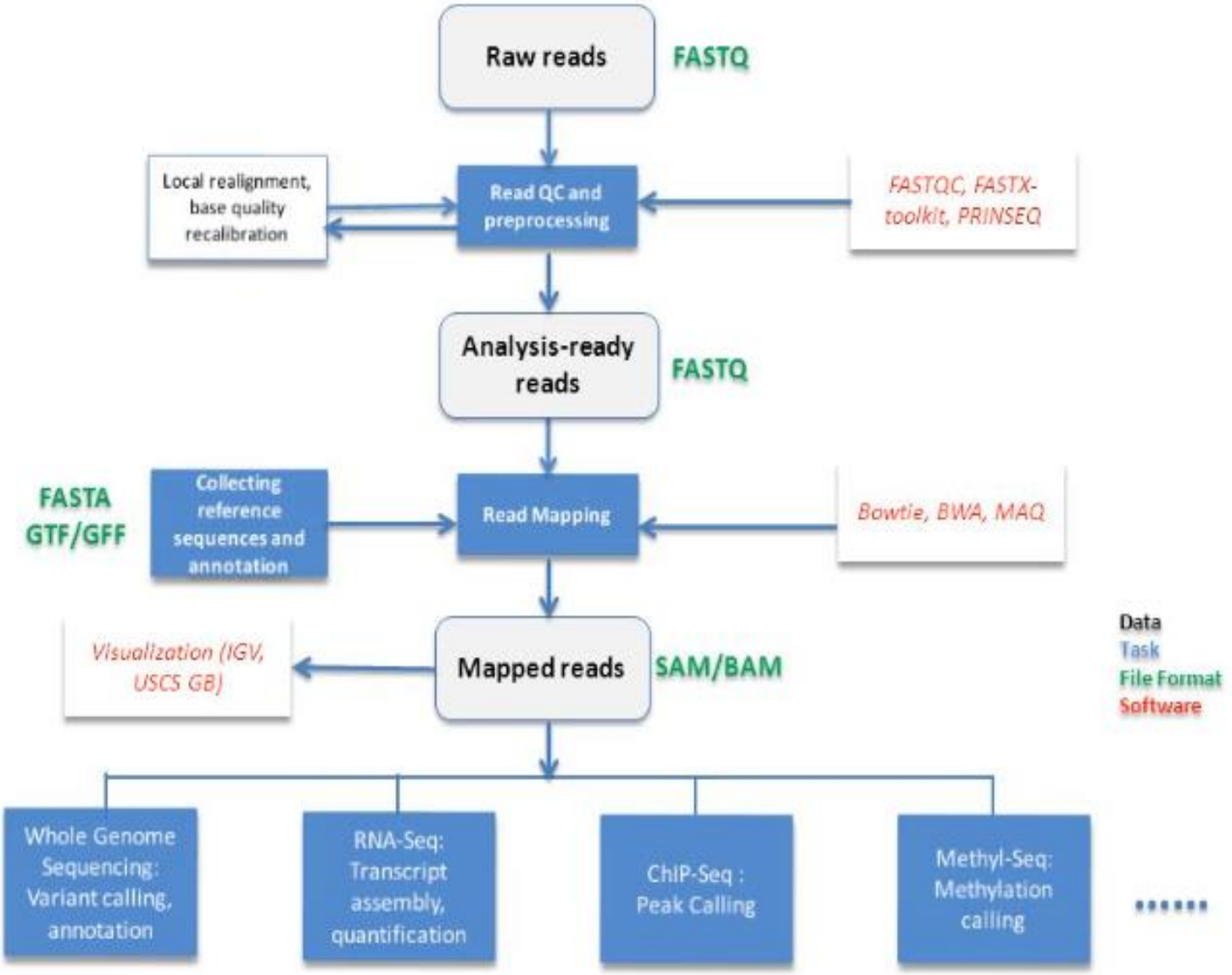
- **Library –** collection of DNA fragments for sequencing
- **Read –** a sequenced fragment
- **Read length** - the average number of contiguous nucleotide bases in a polynucleotide sequence that are produced by a particular sequencing instrument (14-400)



- **Contig –** set of overlapping reads
- **Sequencing depth/Library size –** total number of usable reads from the sequencing machine
- **Coverage –** Number of times a nucleotide base is read (# followed by X: 300X)
- **Single/Paired end** – in paired end sequencing each fragment is sequenced from the two ends and so generates two reads/fragment.
- **Call –** determination of a given base or base sequence by a sequencing instrument

Steps in NGS analysis

- We will have different data (file) formats and tools for each step

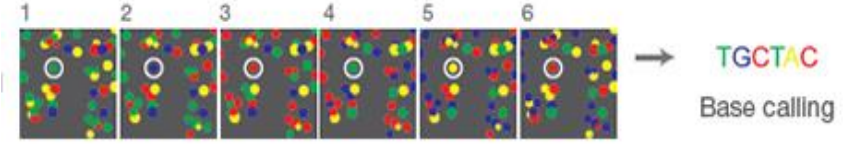


Steps in NGS analysis

Base calling: obtaining the raw read sequences (FASTQ files)

Sanger

Illumina (NGS)



- Base calling accuracy often measured by the Phred Quality Score (Q score) which assesses the accuracy of a sequencing platform.
- It indicates the probability that a given base is called incorrectly by the sequencer.

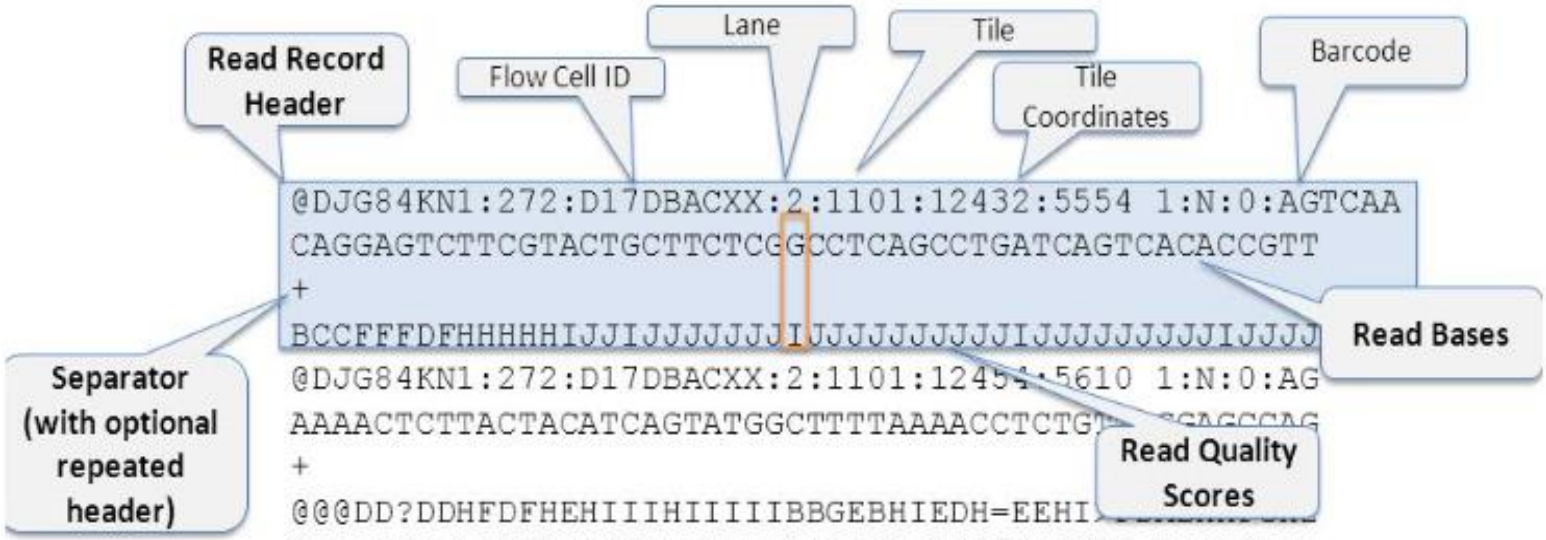
$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

- Ambiguous positions with Phred scores <= 20 are labeled with N.
- To assign each base a unique score identifier (instead of numbers of varying character length), Phred scores are typically represented as ASCII characters.

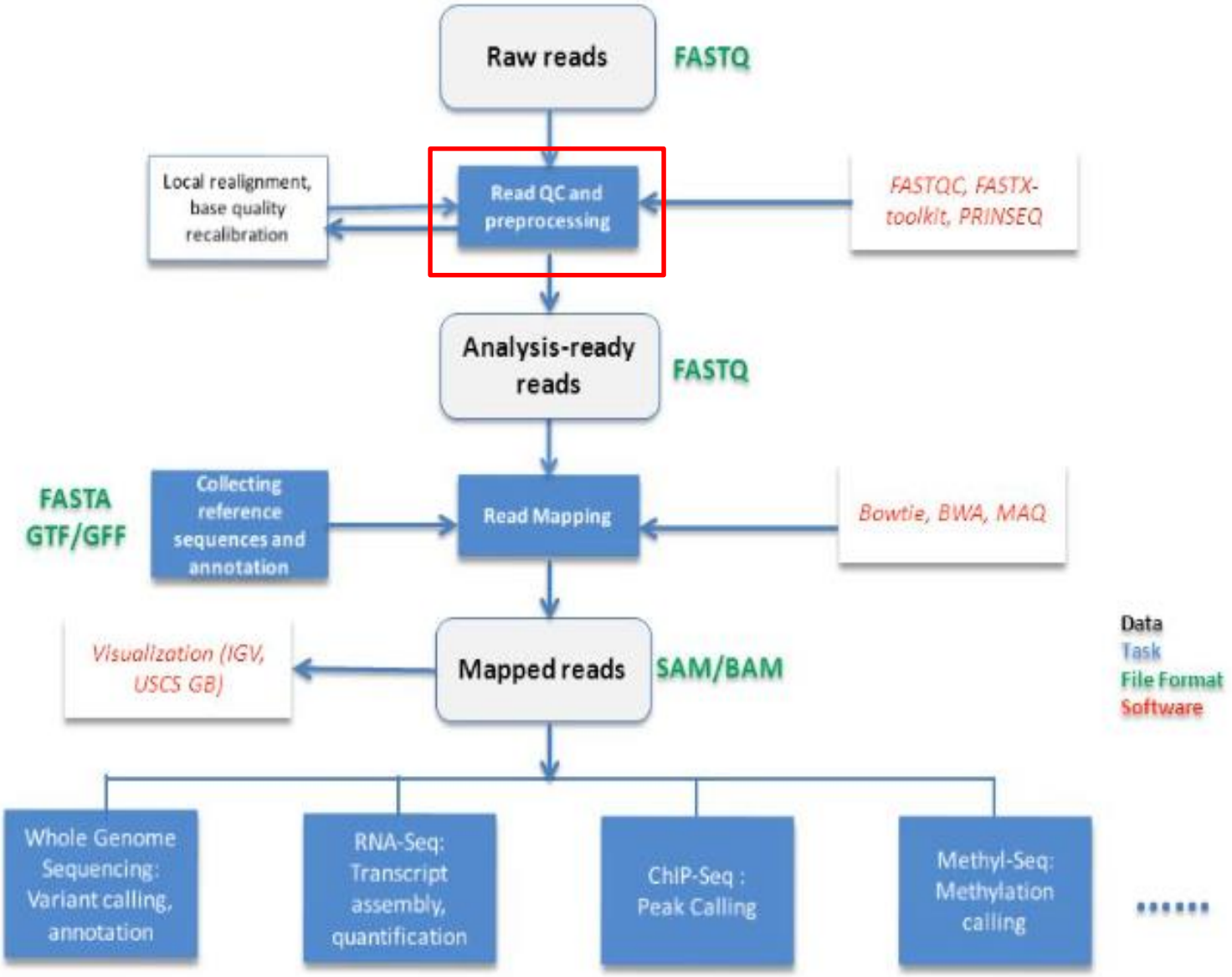
Steps in NGS analysis

FASTQ format = DNA sequence data + Phred quality scores of each base



NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

Steps in NGS analysis

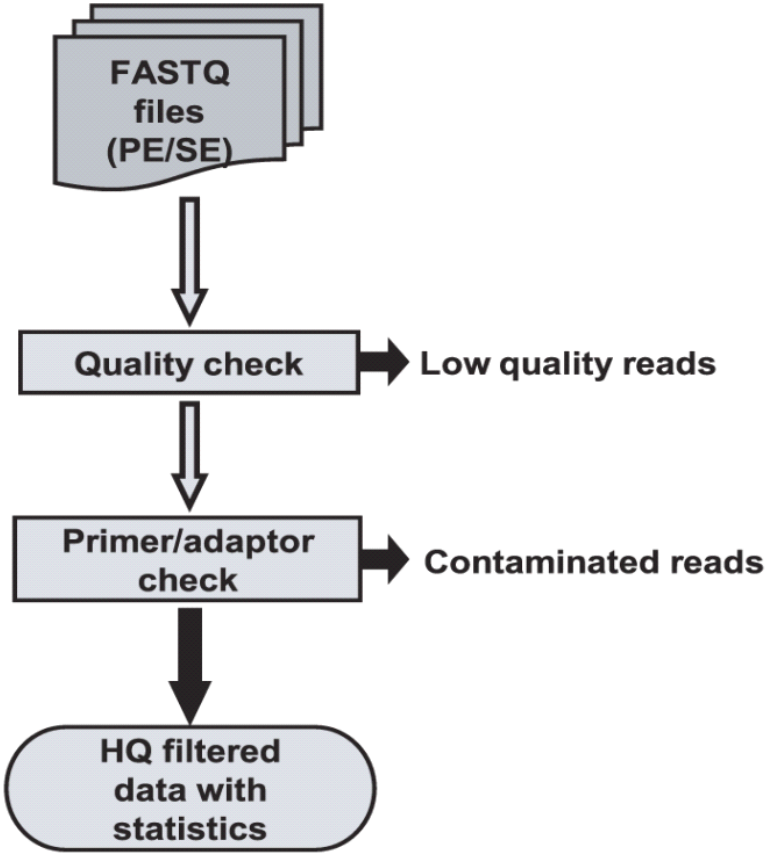


Steps in NGS analysis

Quality Control and Preprocessing

• Quality Control analysis of sequence data is extremely important for meaningful downstream analysis

- To analyze problems in quality scores/ statistics of sequencing data
- To check whether further analysis with sequence is possible
- To remove redundancy (filtering)
- To remove low quality reads from analysis
- To remove adapter contamination



Quality Control

FastQC tool

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

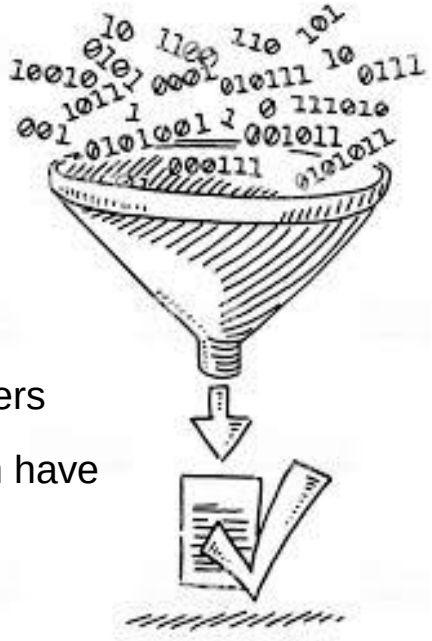
- Basic statistics
- Quality- Per base position
- Per Sequence Quality Distribution
- Nucleotide content per position
- Per sequence GC distribution
- Per base GC distribution
- Per base N content
- Length Distribution
- Overrepresented/ duplicated sequences
- K-mer content

Quality Control

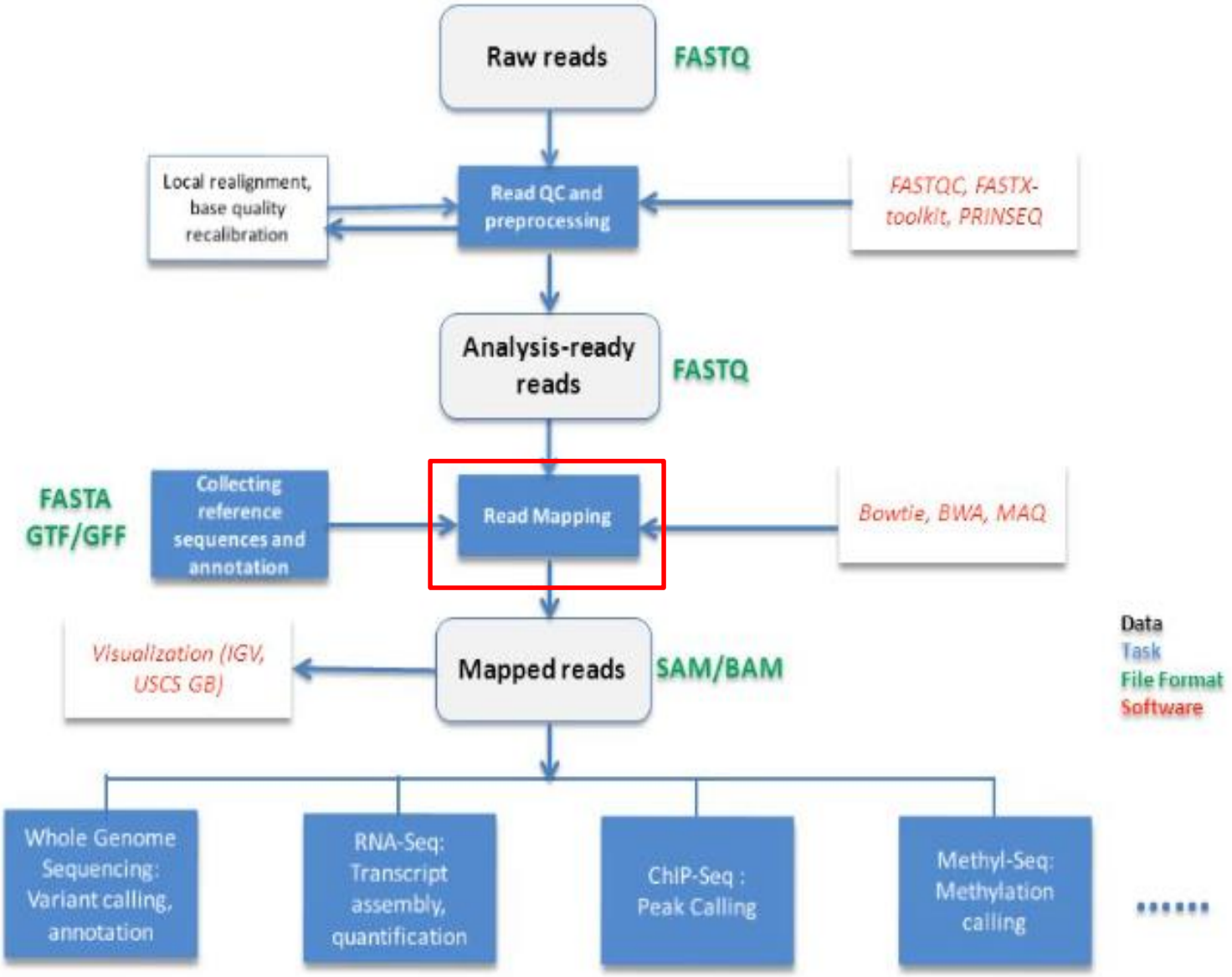
Preprocessing of raw data

Based on the information provided by the QC graphs, the sequences may be treated to reduce bias in downstream analysis:

- **Filtering sequences**
  - with low mean quality score
  - too short
  - with too many ambiguous (N) bases
  - based on their GC content
  - Biological contamination: polyA-tails, rRNA or mtDNA sequences,…
  - Technical contamination: PhiX internal control sequences, adapters/primers
  - Removing duplicate reads is not advised since high expressed genes can have genuine duplicate reads that are not due to the PCR amplification step.
- **Cutting/Trimming sequences**
  - from low quality score regions
  - beginning/end of sequence
  - removing adapters, primers

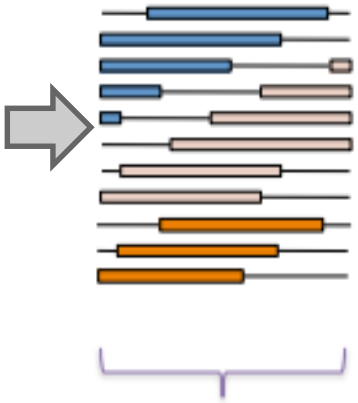


Steps in NGS analysis

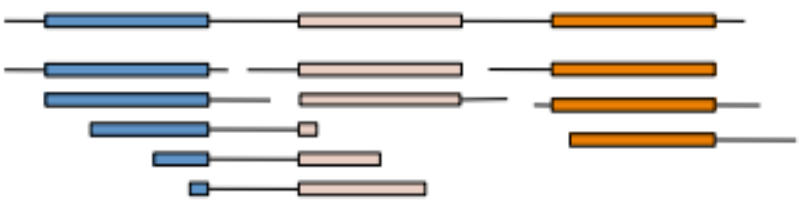


Steps in NGS analysis

Mapping reads to the genome



Mapping and alignment algorithms

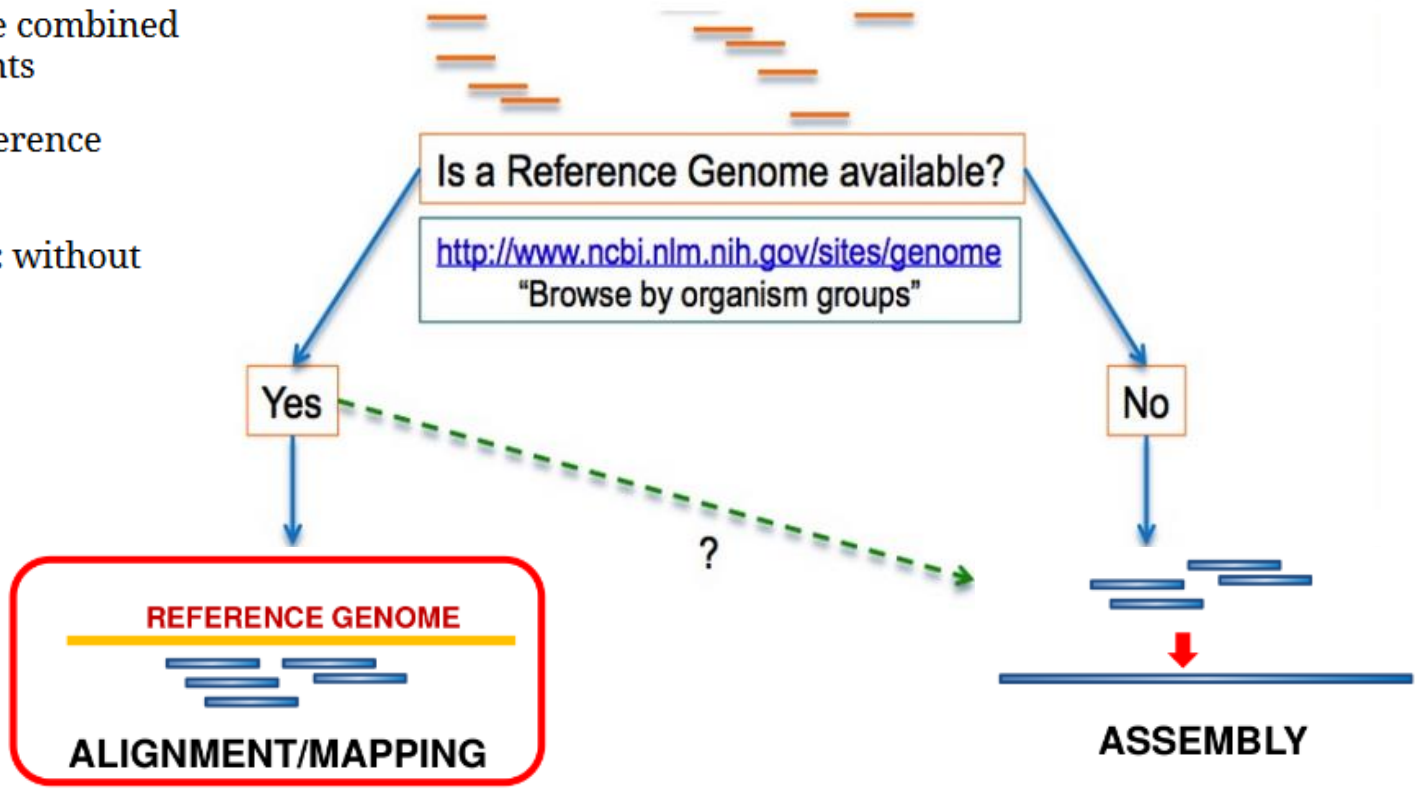


Steps in NGS analysis

Mapping reads to the genome

Mapping/Alignment vs Assembly

- Short reads must be combined into longer fragments
- **Mapping:** use a reference genome as a guide
- **De-novo assembly:** without reference genome



Steps in NGS analysis

Mapping reads to the genome

- Determine position of short read on the reference genome

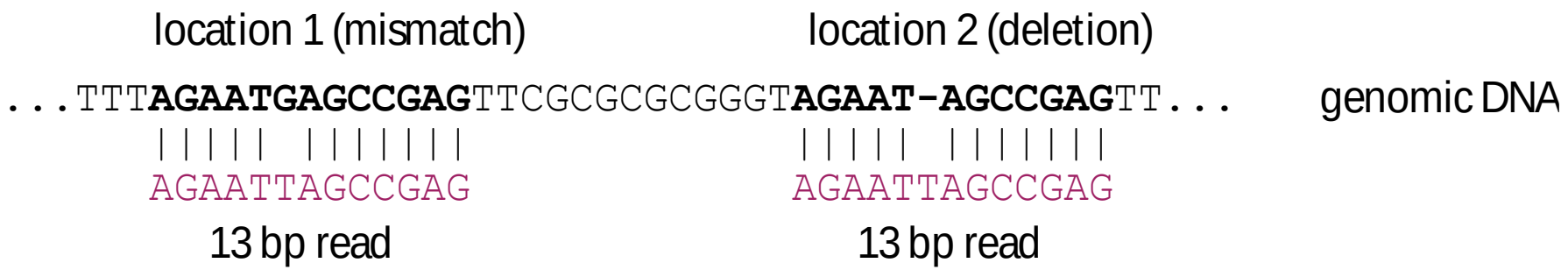
```
Reference: . . . A A - C G C C T T . . .          | = match
.                  | : - : | | | | |             : = mismatch
Read:              A G G G G C C T T             - = gap
```

Steps in NGS analysis

Mapping reads to the genome

Challenging!

- There is ambiguity mapping a read with a mismatch versus a deletion



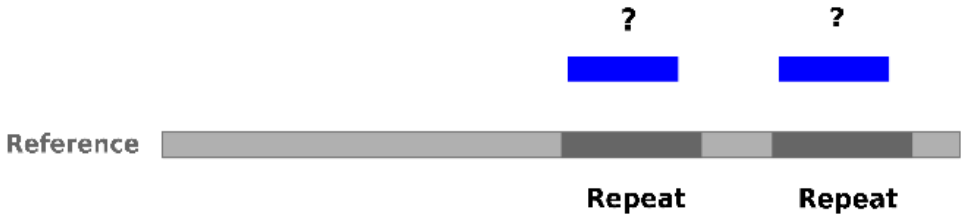
Steps in NGS analysis

Mapping reads to the genome

Challenging!

- There is ambiguity mapping a read with a mismatch versus a deletion

- A read could align to multiple places (repeats)

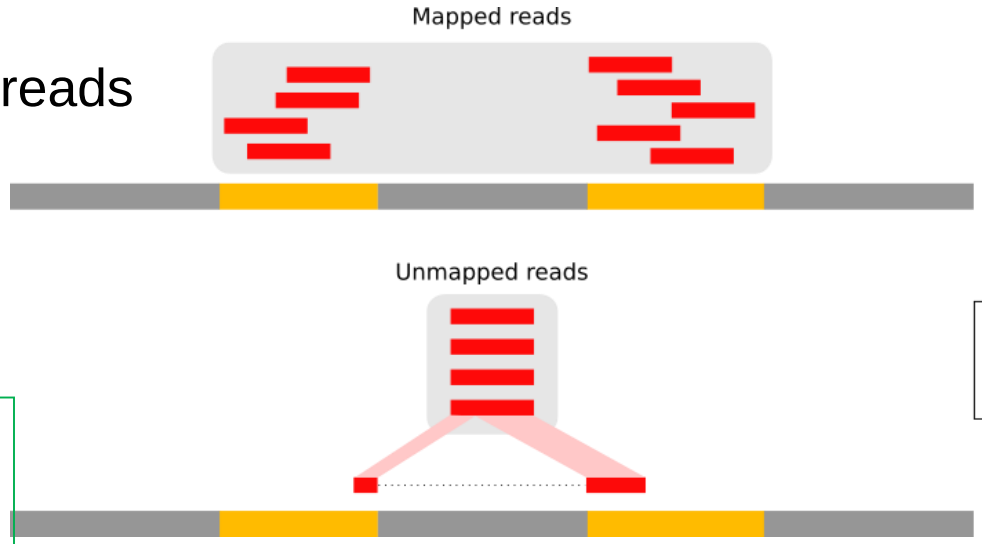


Steps in NGS analysis

Mapping reads to the genome

Challenging!

- There is ambiguity mapping a read with a mismatch versus a deletion
- A read could align to multiple places (repeats)
- In RNA-seq, splicing may split reads

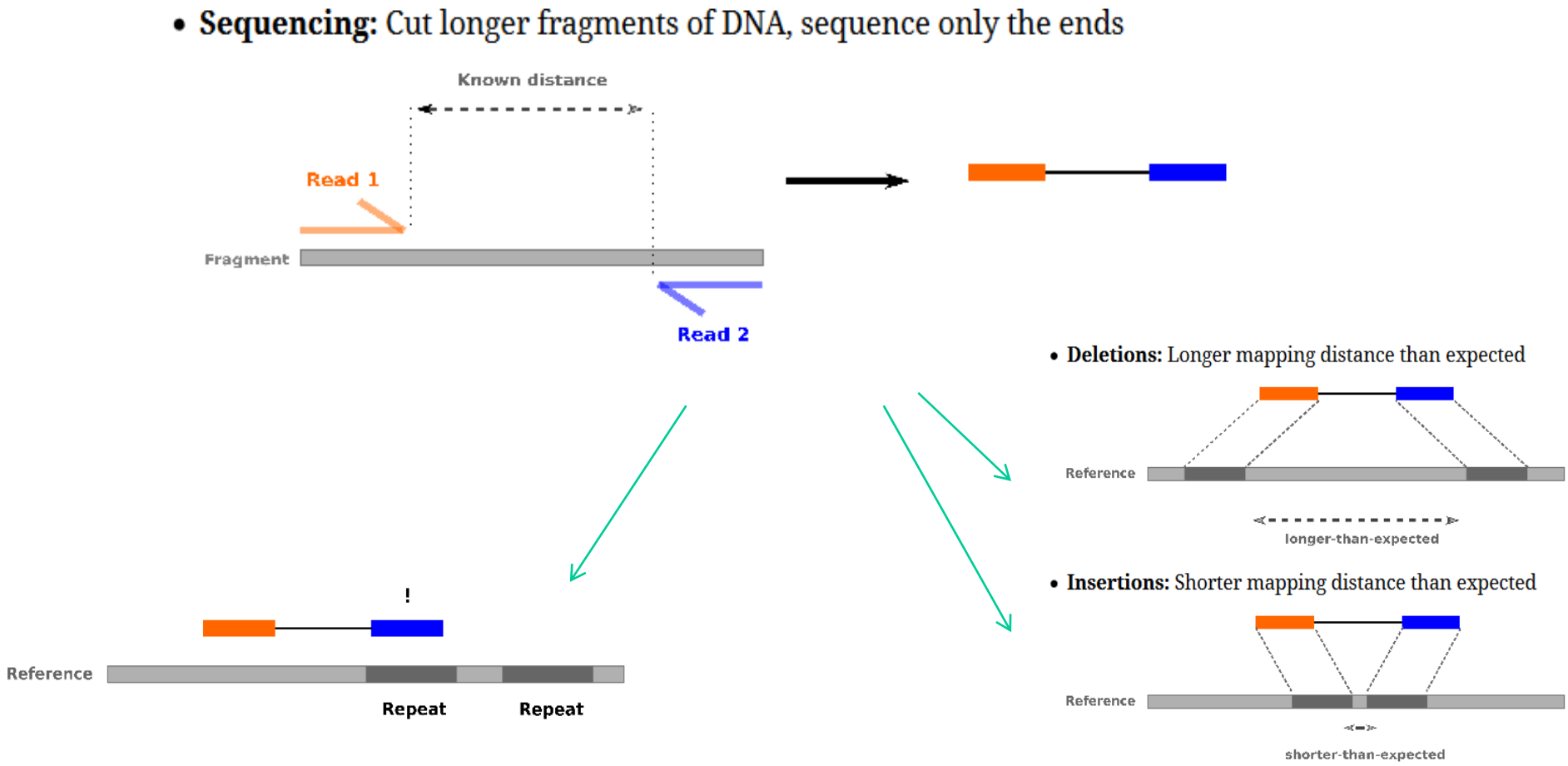


- Complex algorithms have been developped
- Choose appropriate tool/parameters

Steps in NGS analysis

Mapping reads to the genome

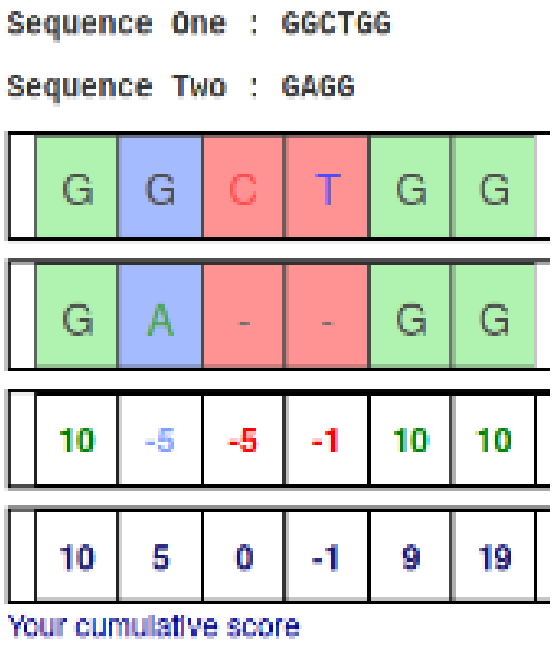
- Paired-end sequencing improves accuracy of mapping



Steps in NGS analysis

Mapping reads to the genome

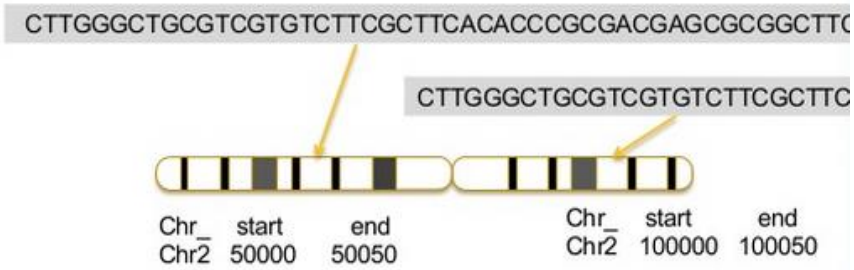
- Quality scores to assess mapping accuracy
  - quantify the probability that a read is misplaced.
  - Function of factors such as:
    - uniqueness (ie not a multi-mapper)
    - number of mismatches in read
    - number of insertions/deletions in read
    - quality of bases in read



Steps in NGS analysis

Mapping reads to the genome

SAM/BAM format = Aligned read sequence + Mapping info (position, quality score...)



- SAM files typically contain a short header section with information about the genomic loci of each read and a very long alignment section where each row represents a single read alignment. For each read, there are 11 mandatory fields that always appear in the same order:

header

```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
```

Read name | Alignment gap information | Quality scores (fastq format)

SLX1:1:127:63:4 ... 1 10052169 ... 23M6N10M ... GAAGATACTGGTTTTTTTCTTATGAGACGGAGT 768832'48:::::::;;:/78$88818099897 SM:Z:JPTGBMN01 ...

Locus | Read sequence | Meta data

Steps in NGS analysis