

## **(First) steps in NGS data analysis**

Bioinformatics Course UEB-VHIR  
November 2021

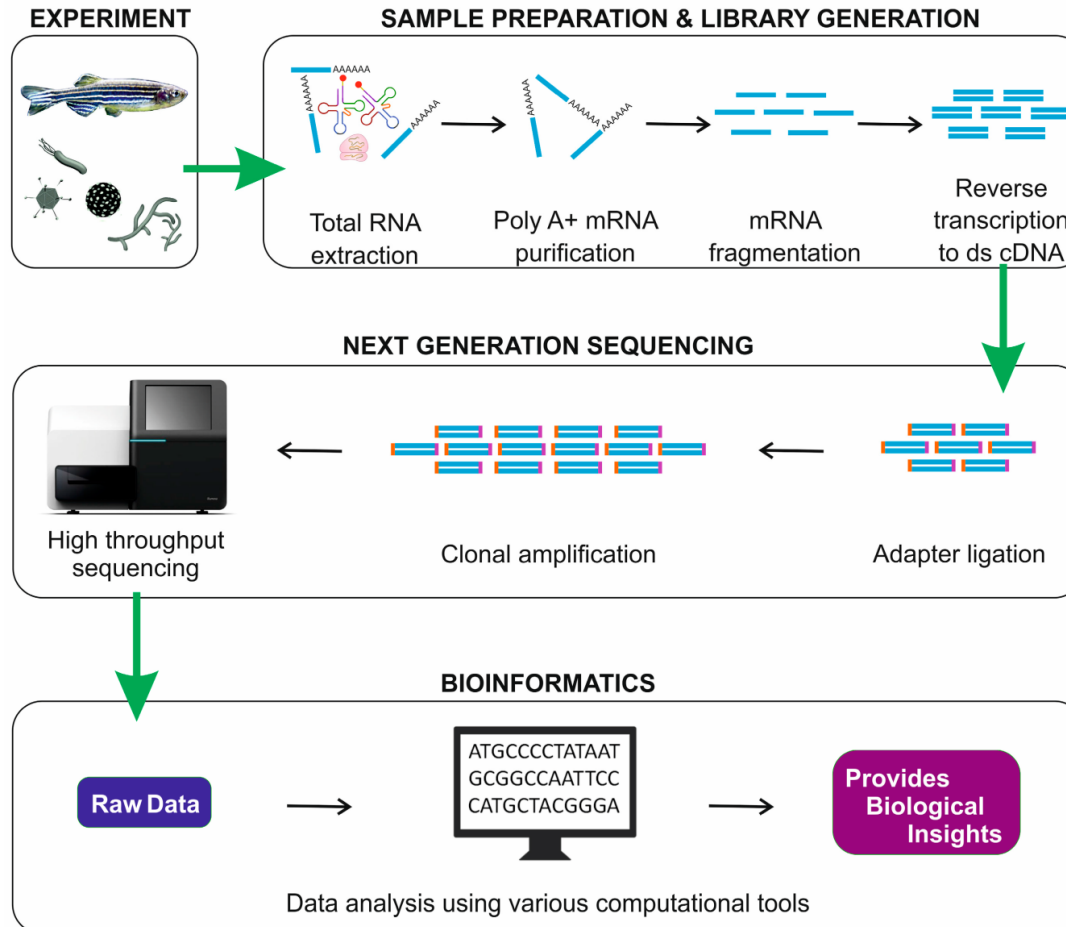
Mireia Ferrer<sup>1</sup>, Álex Sánchez<sup>1,2</sup>, Esther Camacho<sup>1</sup>, Angel Blanco<sup>1,2</sup>, Berta Miró<sup>1</sup>

<sup>1</sup> Unitat d'Estadística i Bioinformàtica (UEB) VHIR

<sup>2</sup> Departament de Genètica Microbiologia i Estadística, UB

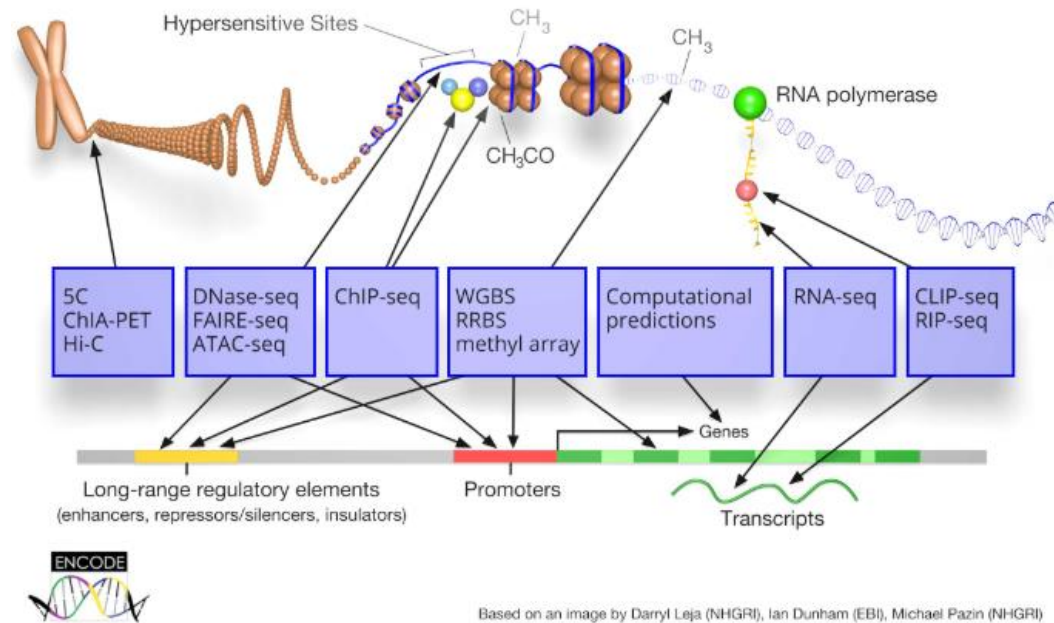
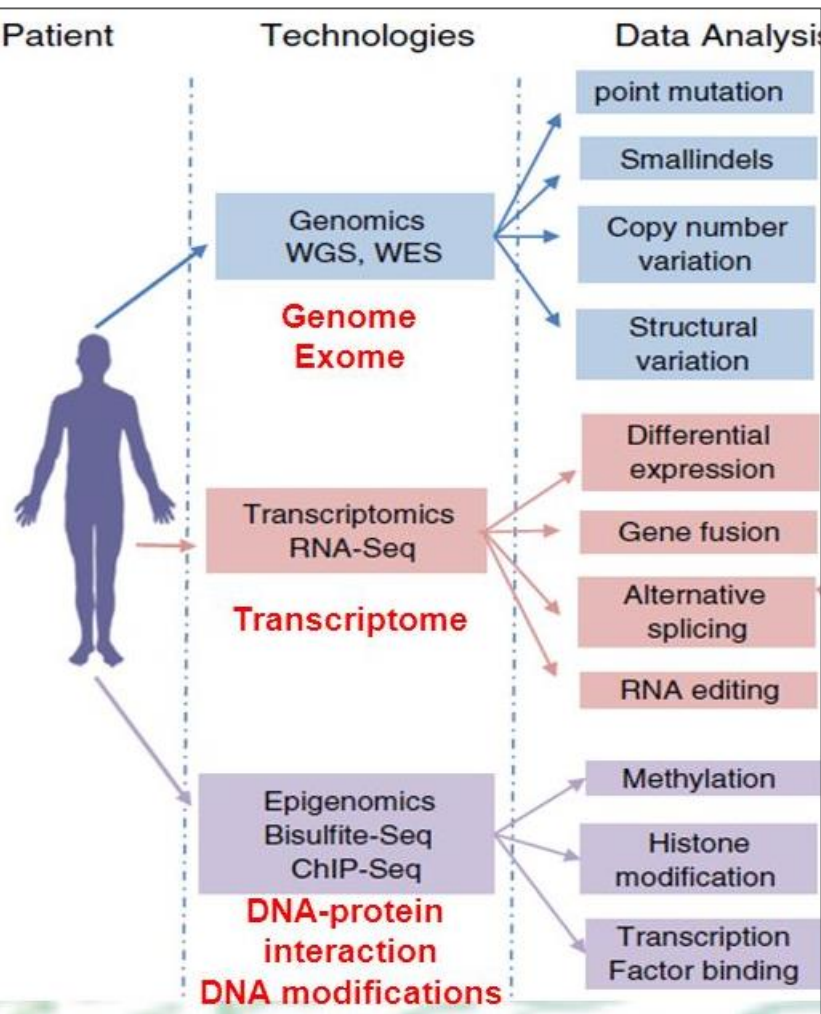
# Steps in NGS analysis

## General workflow



# Steps in NGS analysis

## Applications



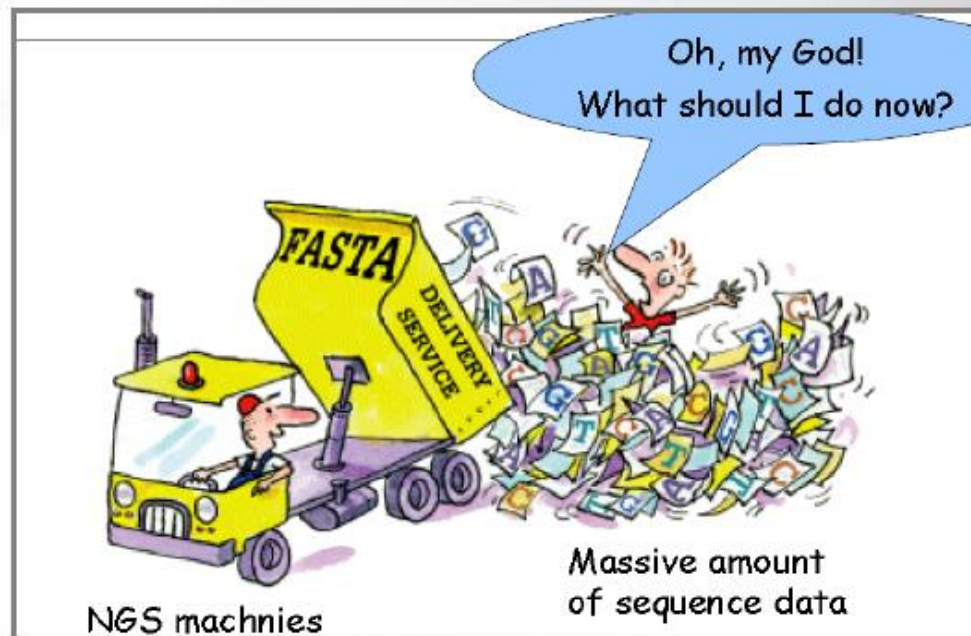
- But also...
  - Metagenomics
  - *De novo* genome assembly

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

# Steps in NGS analysis

## Bioinformatics challenges of NGS

I have my sequences/images. Now what?



# 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:

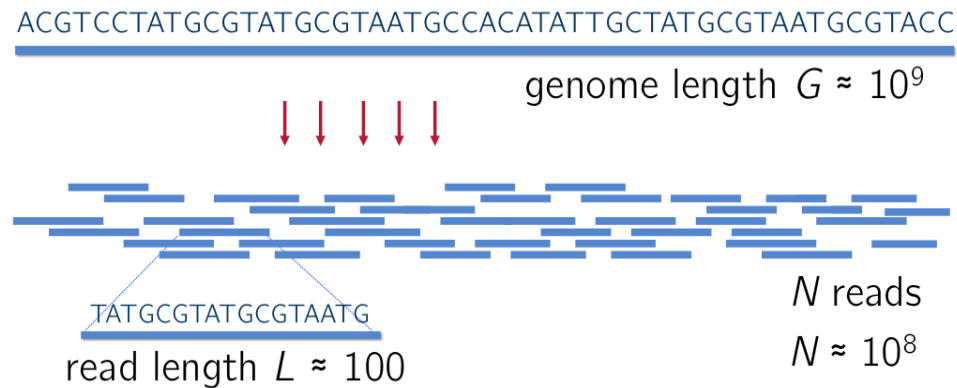
- Huge data storage and transfer technology
- Algorithms for managing, analyzing and visualizing data
- Reproducible workflows and standards for analysis
- Specialized tools for integrating various data types



# Steps in NGS analysis

## 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

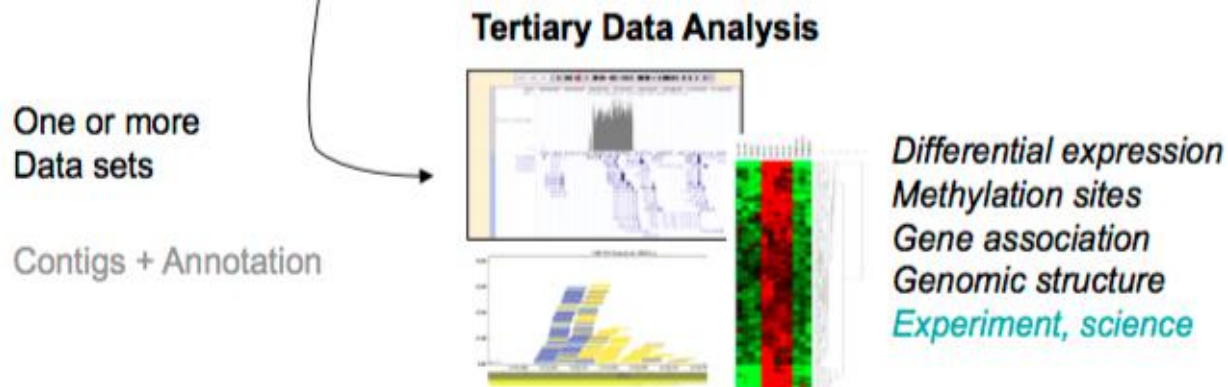
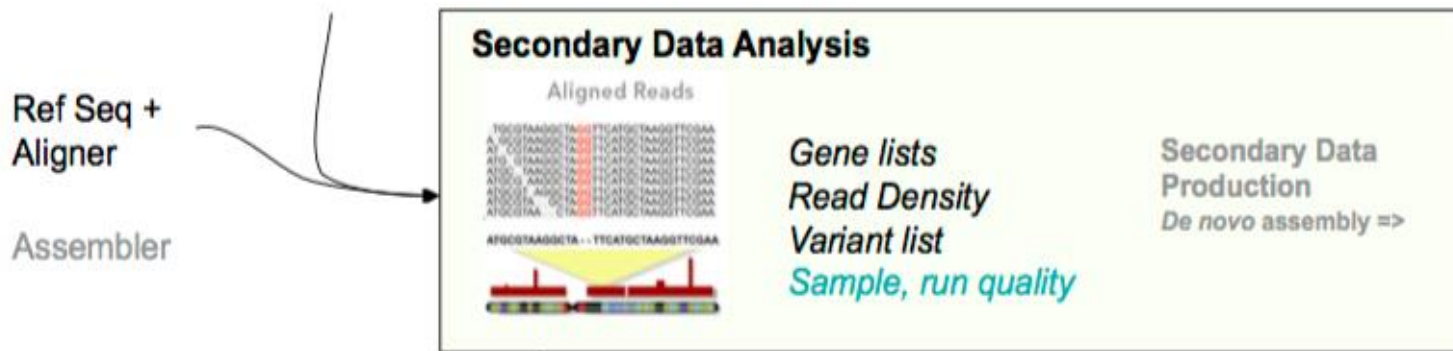




# Steps in NGS analysis

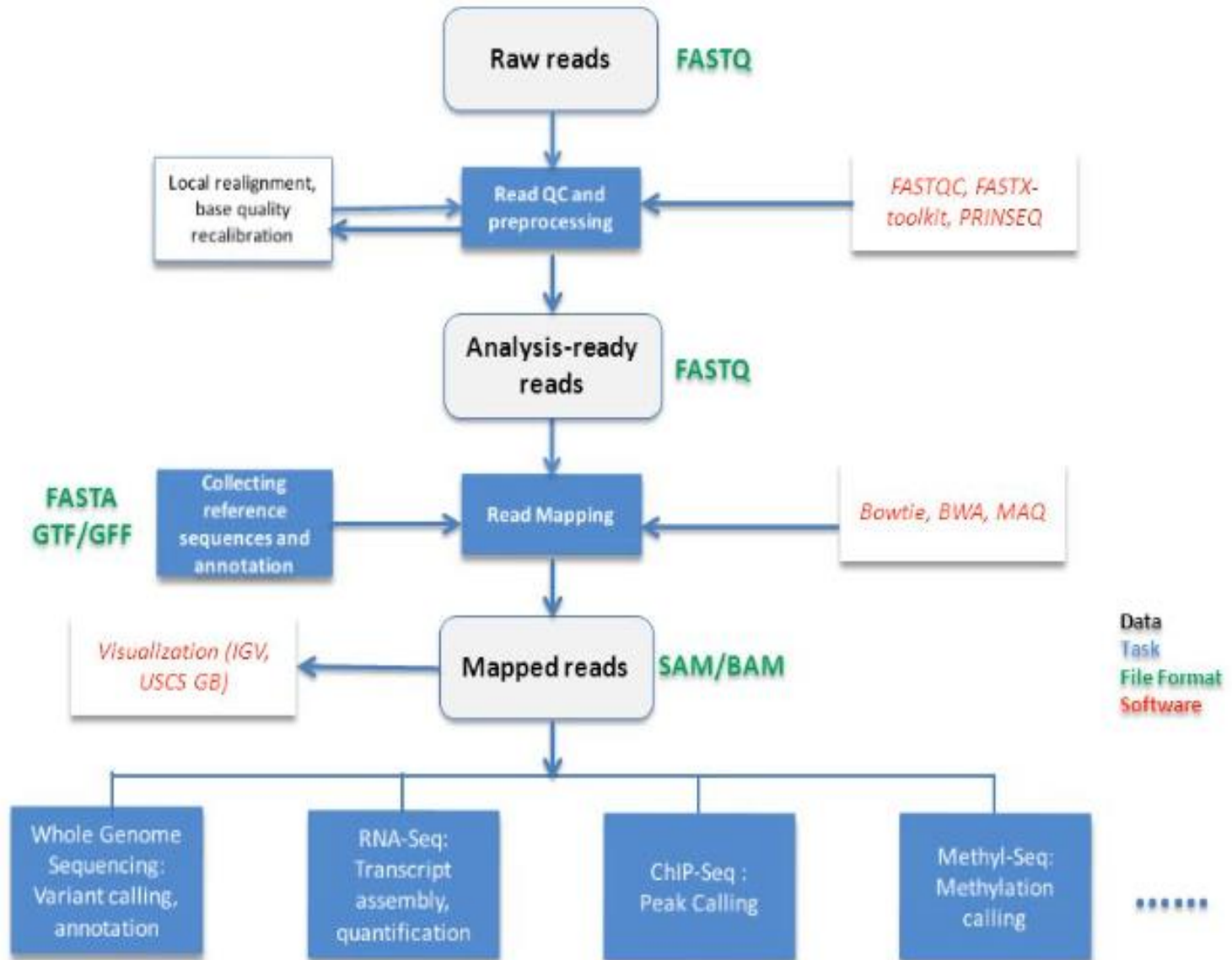
- NGS data is analyzed in three stages

## Primary Data Analysis - Images to bases



# Steps in NGS analysis

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

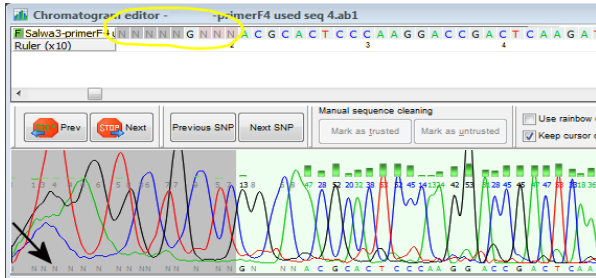




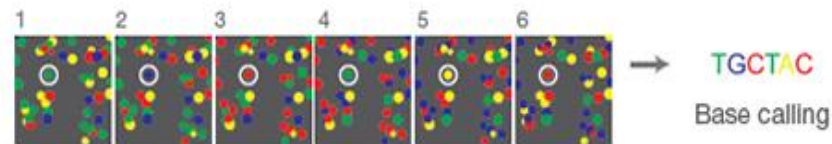
# Steps in NGS analysis

## 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.

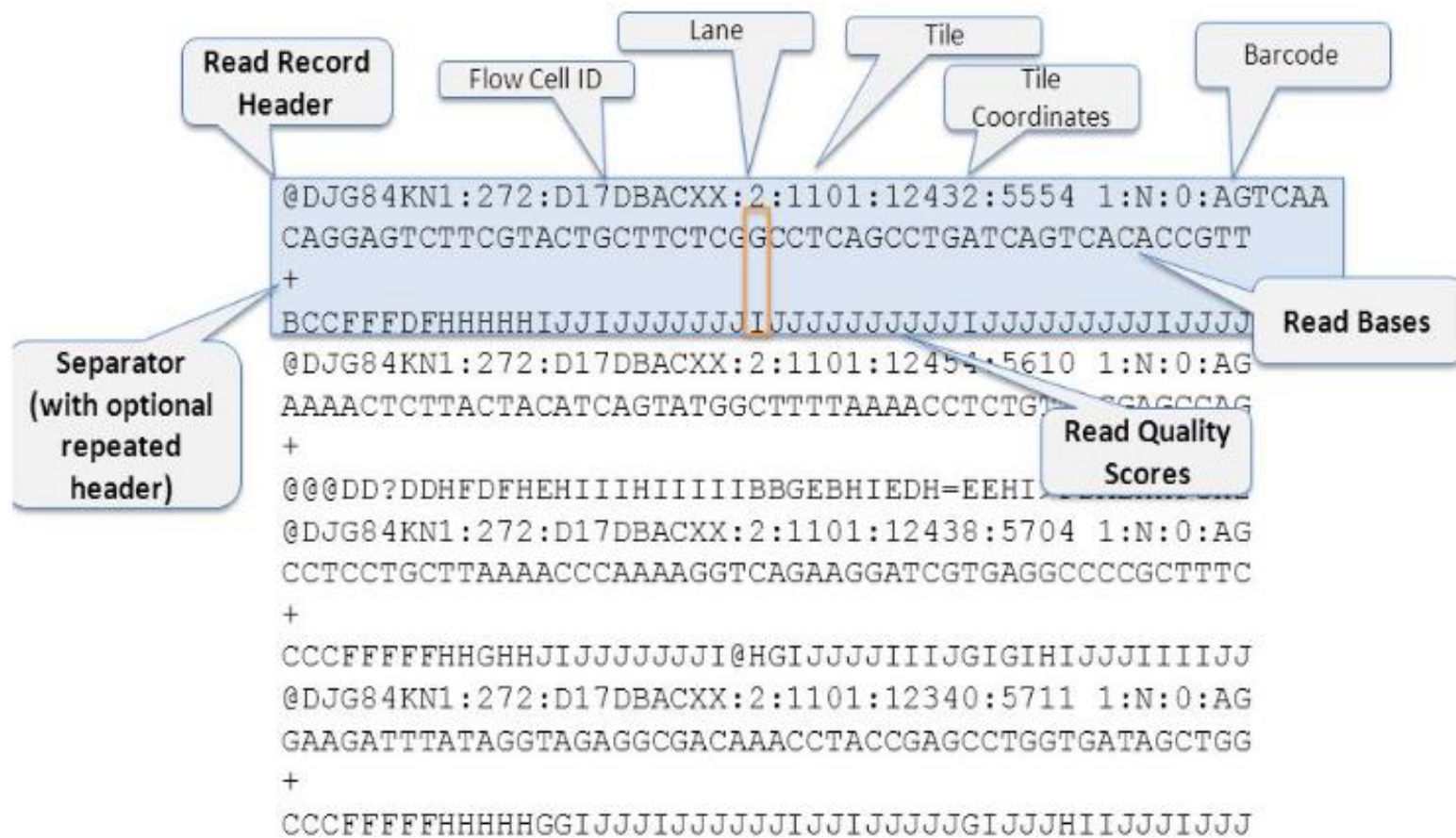
$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

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  $\leq 20$  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.

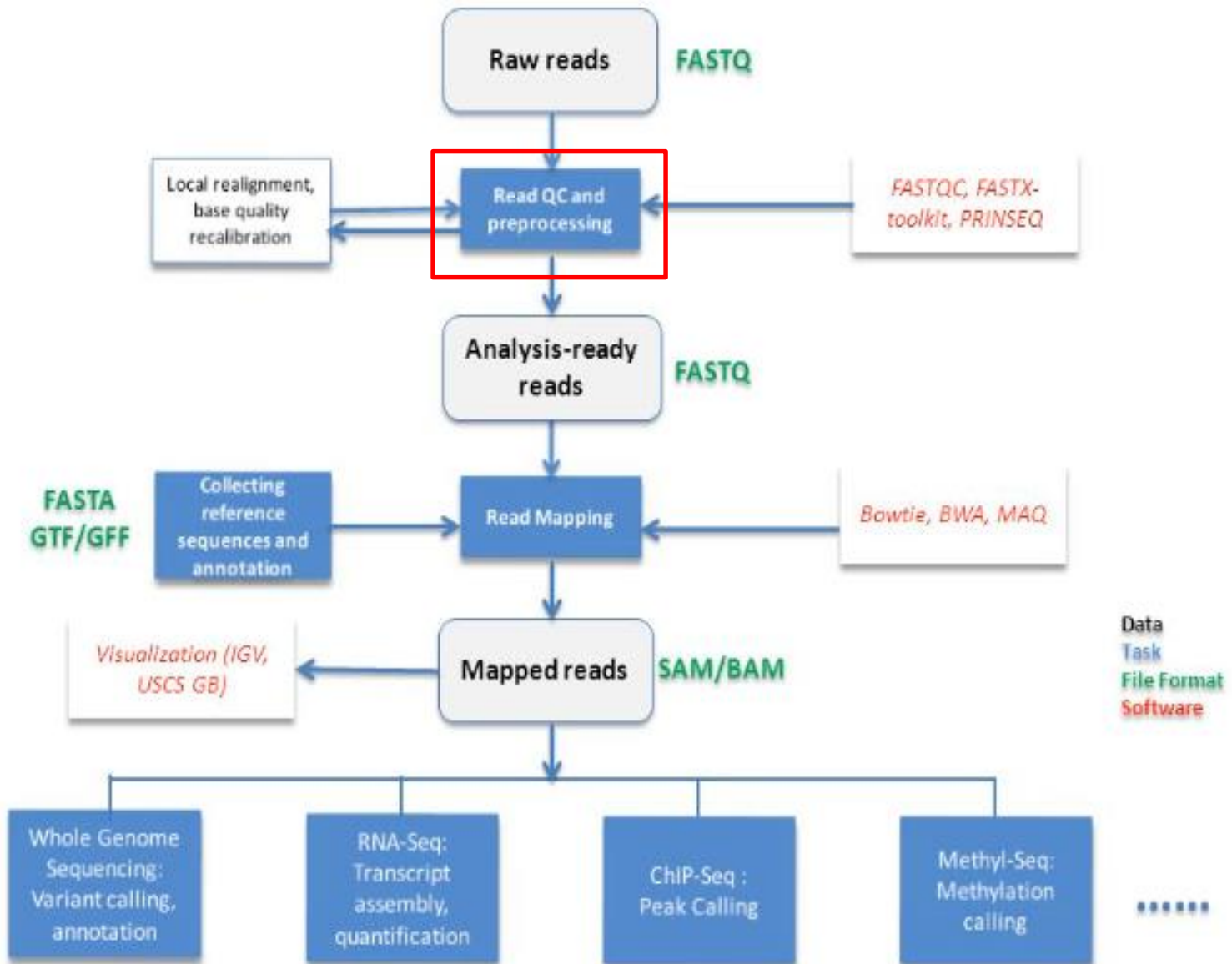
# Steps in NGS analysis

## 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

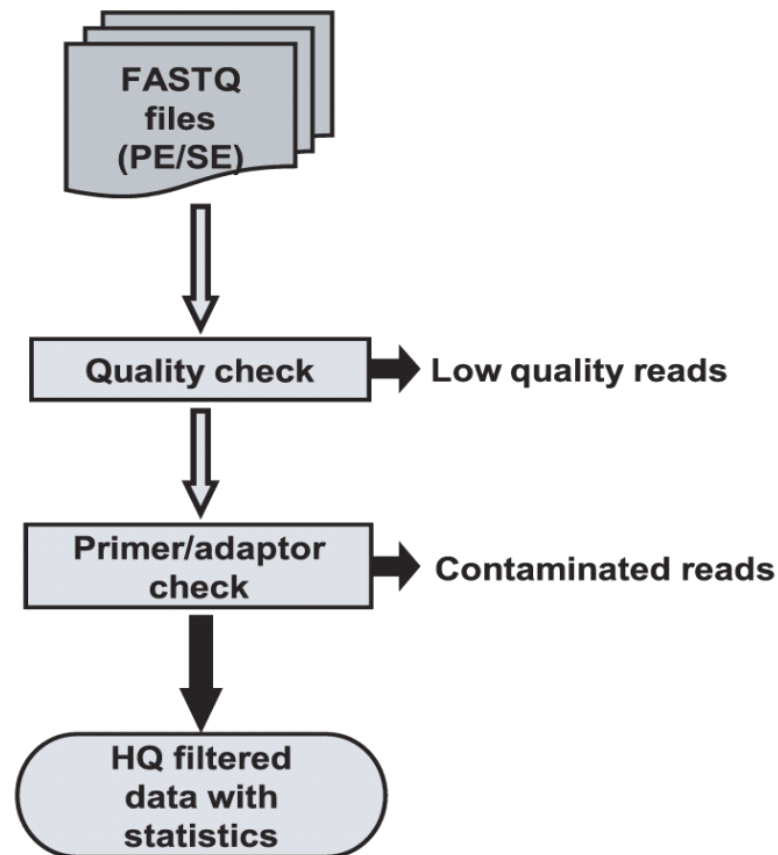


# Steps in NGS analysis

## 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

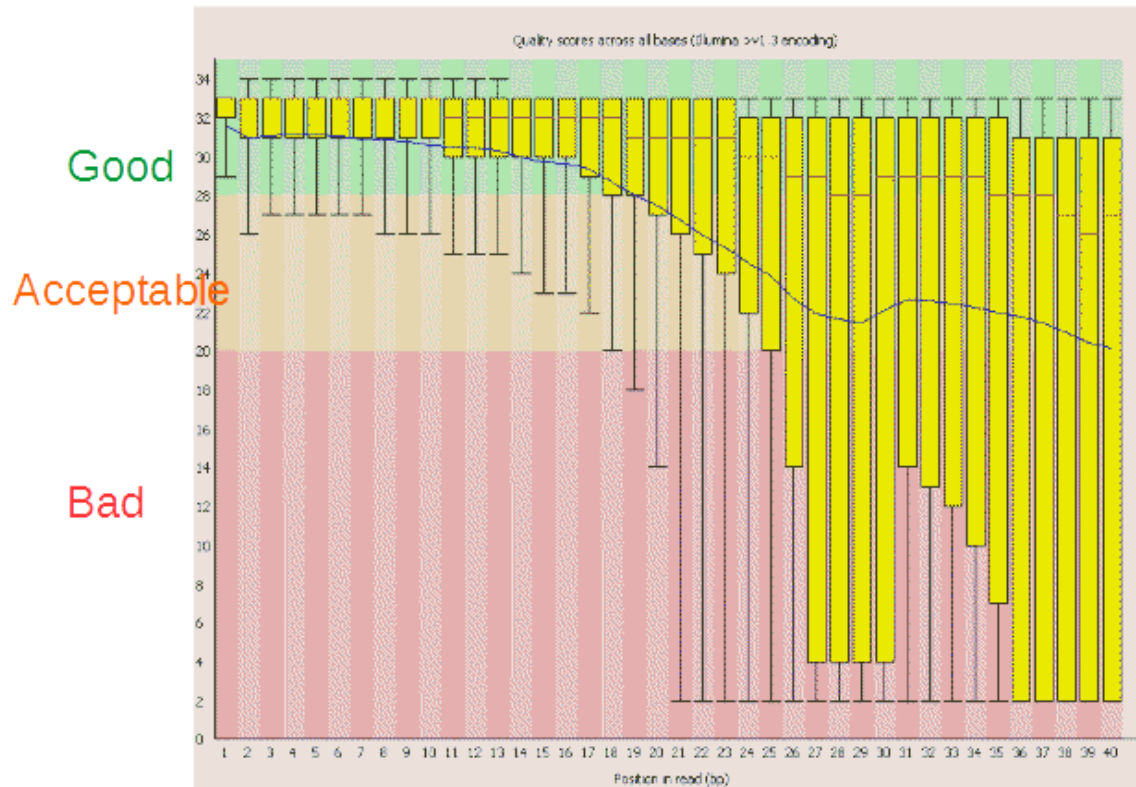
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- 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

## FastQC

### Per base sequence quality (Boxplot)



shows an overview of the range of quality values across all bases at each position in the FastQ file

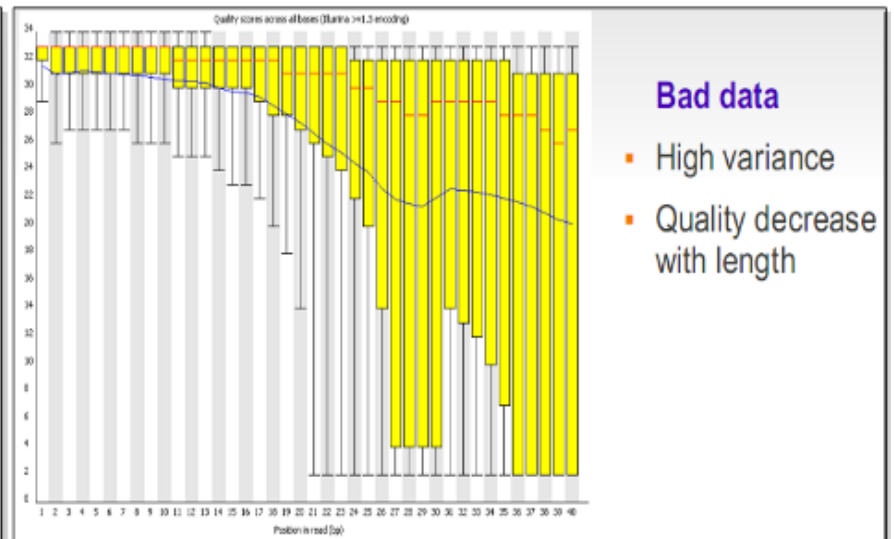
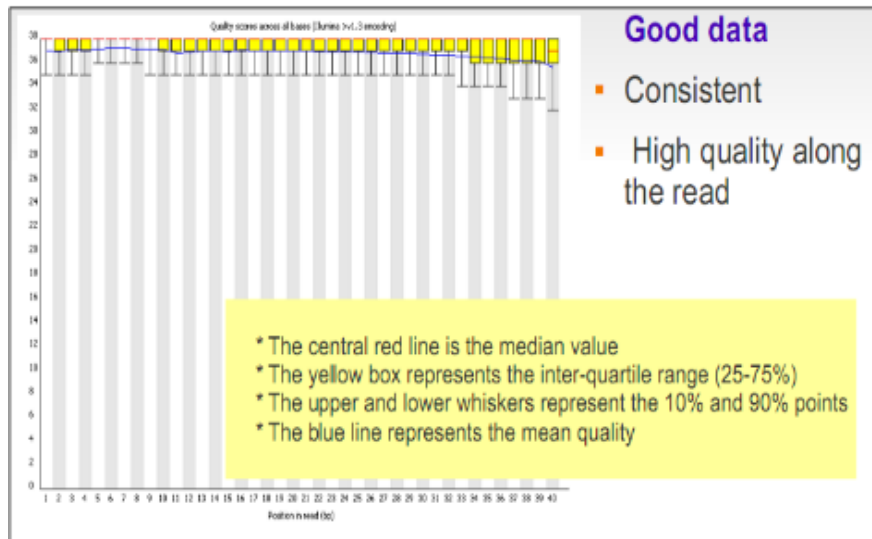
Y axis- Quality Score  
 X axis- Base position



# Quality Control

## FastQC

### Per base sequence quality (Boxplot)

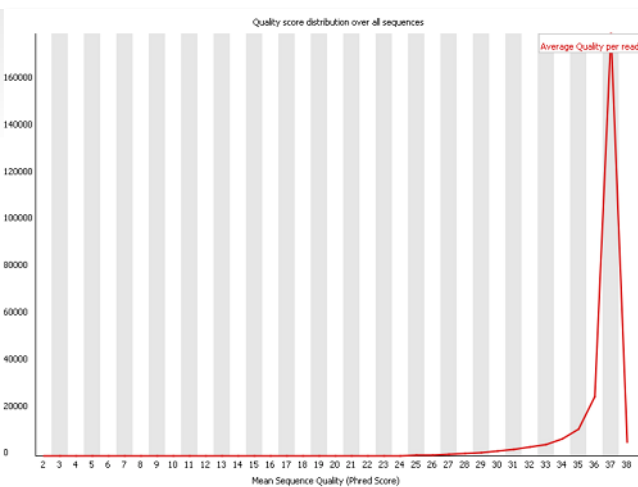


# Quality Control

## FastQC

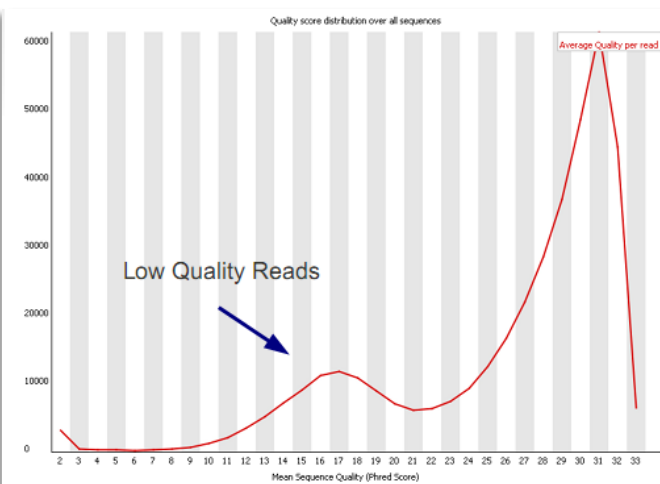
### Per sequence quality scores

allows you to see if a subset of your sequences have universally low quality values.



#### Good data

- Most are high-quality sequences



#### Bad data

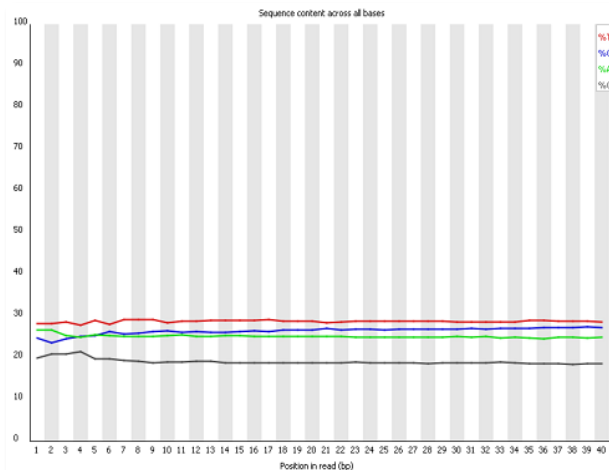
- Not uniform distribution

# Quality Control

## FastQC

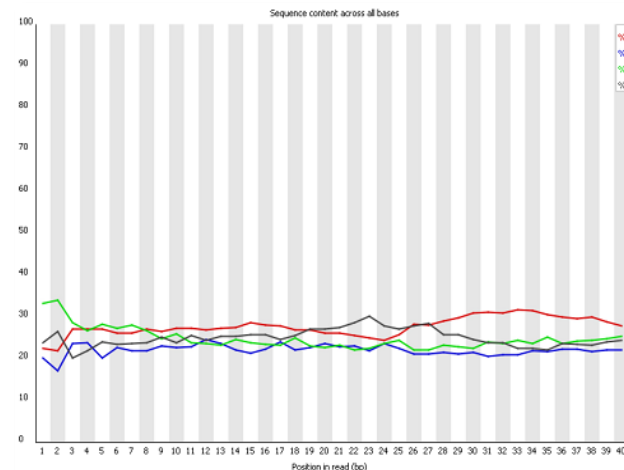
### Per base sequence content

proportion of each base position in a file for which each of the four normal DNA bases has been called



### Good data

- Smooth over length
- Organism dependent (GC)



### Bad data

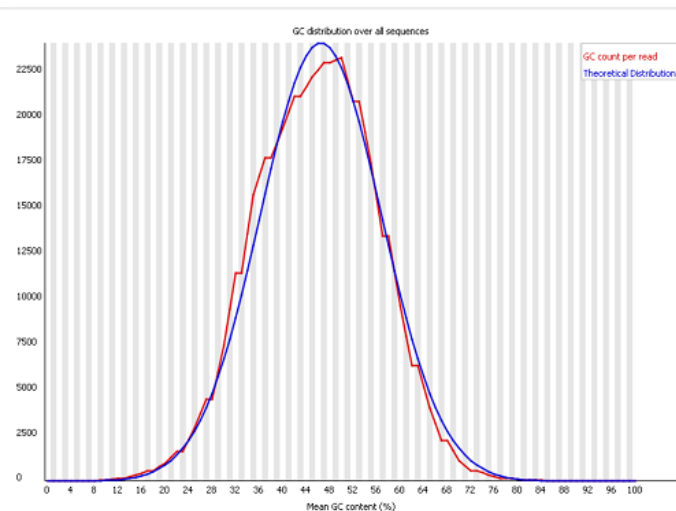
- Sequence position bias

# Quality Control

## FastQC

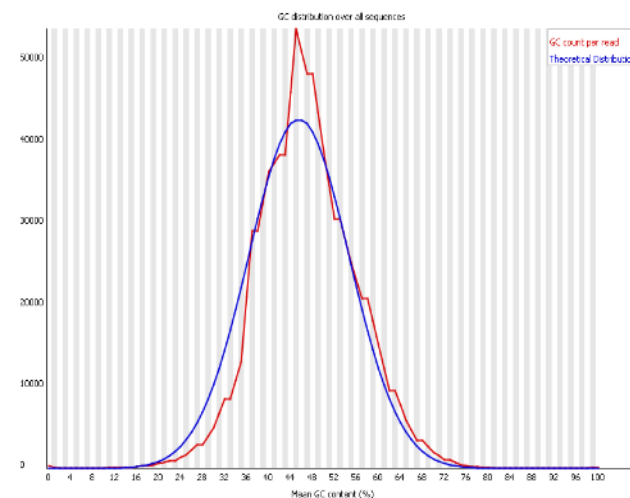
### Per sequence GC content

measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content



#### Good data

- Fits with the expected
- Organism dependent



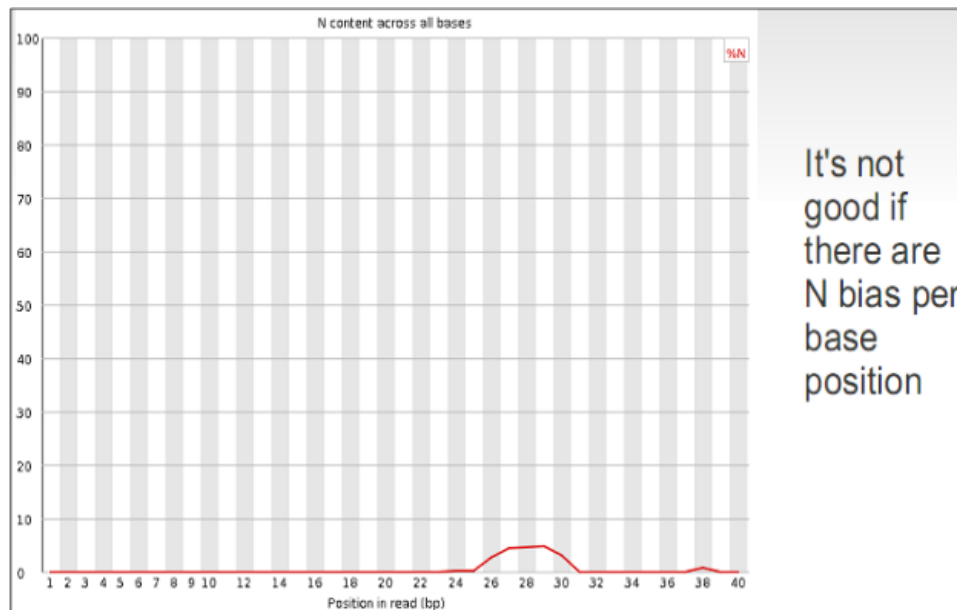
#### Bad data

- It does not fit with expected
  - Organism dependent
- Library contamination?

# Quality Control

## FastQC

### Per base N content

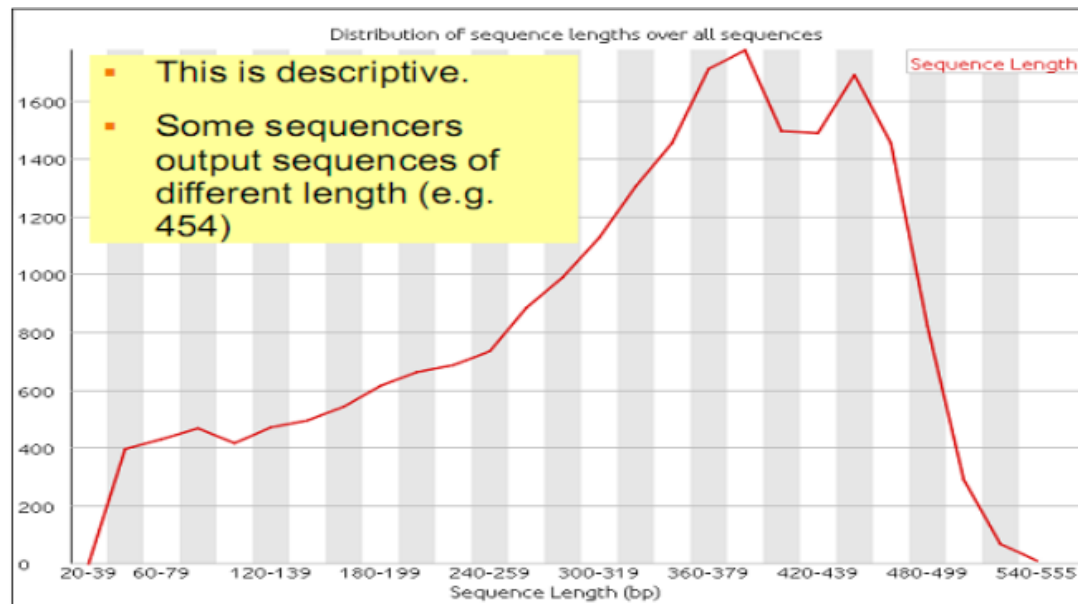


If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. It plots out the percentage of base calls at each position for which an N was called.

# Quality Control

## FastQC

### Sequence length distribution



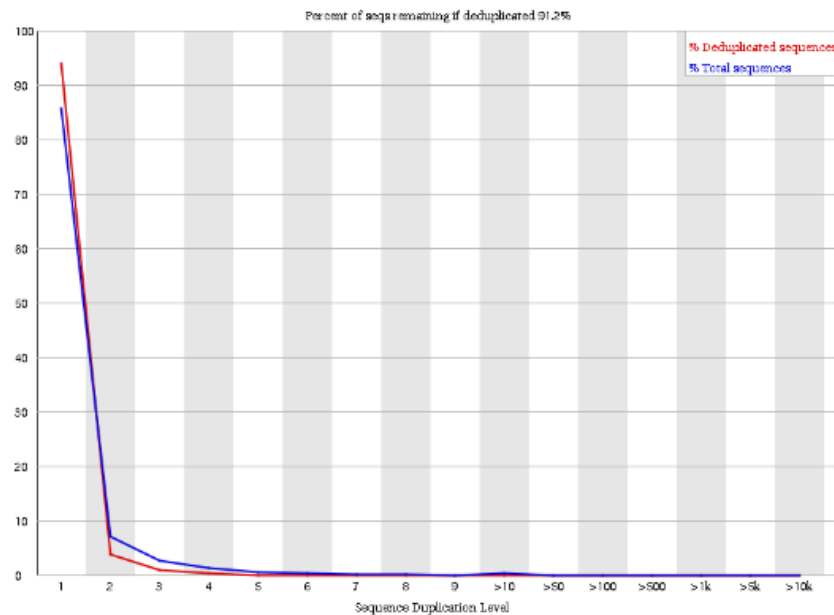
In many cases it will produce a simple graph showing a peak only at one size, but for variable length FASTQ files, it will show the relative amounts of each different size of sequence fragment.



# Quality Control

## FastQC

### Sequence duplication level



Counts the degree of duplication for every sequence. Too many duplicate regions in the sequence may indicate contamination or technical problems

## FastQC

### Overrepresented sequences

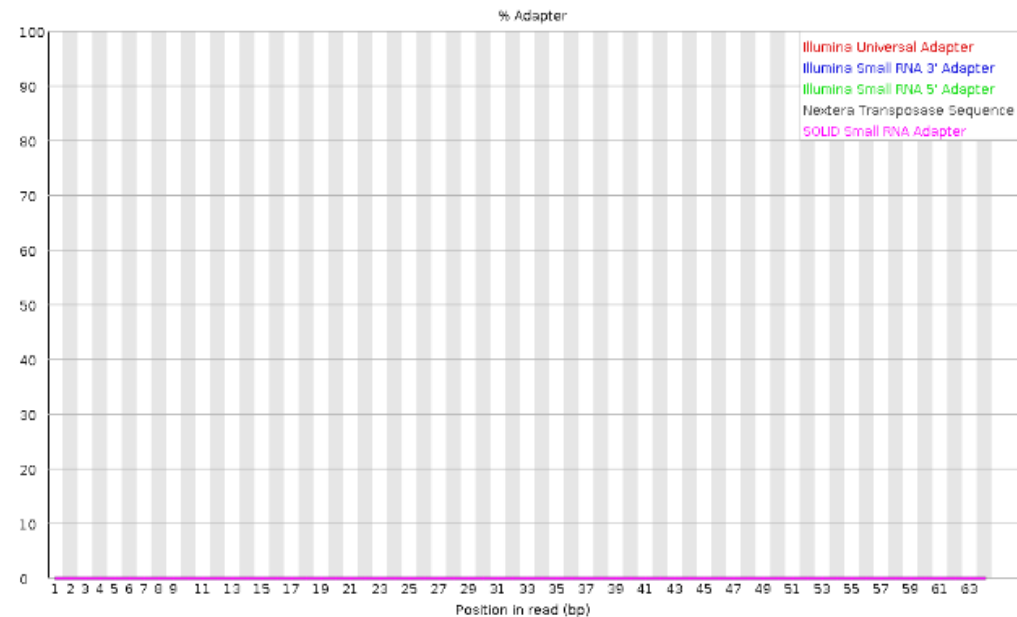
Sequence	Count	Percentage	Possible Source
AAGATCCGAGTCGTCCGAAATCCATTGCCCGTGTCTCACAGTTATTAA	432	0.43585733743631133	No Hit
AGATCCGAGTCGTCCGAAATCCATTGCCCGTGTCTCACAGTTATTAA	335	0.33799122231750994	No Hit
TGGCAGAAGTAGAGCAGAAGAAGAAGCGGACCTTCGCAAGTTCACCTAC	250	0.25223225546082834	No Hit
CAGAAGTAGAGCAGAAGAAGAAGCGGACCTTCGCAAGTTCACCTACCGC	237	0.23911617817686526	No Hit
GTAGAGCAGAAGAAGAAGCGGACCTTCGCAAGTTCACCTACCGCGCGT	223	0.22499117187105888	No Hit
AAGAAATCTGACCCGGTCTCGTACCGCGAGACGGTCAGTGAAGAGTC	204	0.2058215204560359	No Hit
AAGTAGAGCAGAAGAAGAAGCGGACCTTCGCAAGTTCACCTACCGCGGC	151	0.1523482822983403	No Hit
CACCTGGAGATCTGCCTGAAGGACCTGGAGGAGGACCACGCCTGCATCCC	147	0.14831256621096706	No Hit
TCTGCCTGAAGGACCTGGAGGAGGACCACGCCTGCATCCCATCAAGAAA	146	0.14730363718912376	No Hit

Lists all of the sequence which make up more than 0.1% of the total. Finding that a single sequence is very overrepresented in the set either means that is highly biologically significant, or that the library is contaminated. For each overrepresented sequence it will look for matches in a database of common contaminants.

# Quality Control

## FastQC

### Adapter content



Does a generic analysis of all the Kmers in the library to find those that don't have even coverage through the length of the reads.

## FastQC

- Good (Illumina) quality data:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good\\_sequence\\_short\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html)

- Bad (Illumina) quality data:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad\\_sequence\\_fastqc.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html)

# 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



# Quality Control

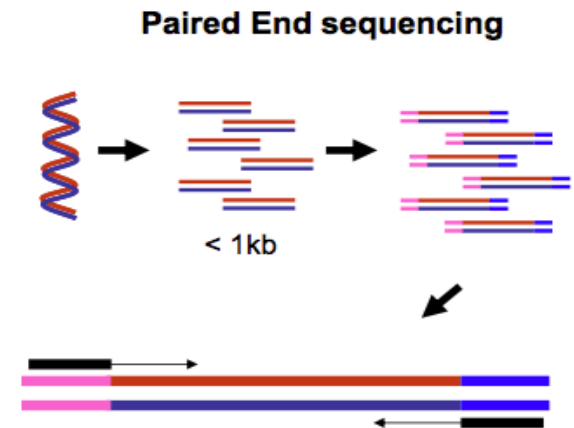
## Your turn!

- We will analyze exome sequencing data from a study that aimed to identify genetic variants associated to a disease. The data comes from paired-end sequencing, each file corresponding to the forward or reverse, respectively:

[https://zenodo.org/record/3243160/files/proband\\_R1.fq.gz](https://zenodo.org/record/3243160/files/proband_R1.fq.gz)

[https://zenodo.org/record/3243160/files/proband\\_R2.fq.gz](https://zenodo.org/record/3243160/files/proband_R2.fq.gz)

**Paired-end data:** a single physical piece of DNA/RNA is sequenced from two ends and so generates two reads. These can be represented as separate files (two fastq files with first and second reads) or a single file where reads for each end are interleaved.





# Introduction to Galaxy

## Training Infrastructure as a Service

We want to help you conduct your training seminars. You provide the training, we provide you training infrastructure *at no cost*.

Why use UseGalaxy.eu training infrastructure?

- Free
- Private queue, no wait times
- No Galaxy Maintenance
- No Galaxy Administration
- Official Galaxy Training Materials guaranteed to work



Simply fill out the infrastructure request form and we'll get back to you shortly.

[Find out more](#)

After registration in [European Galaxy server](#)



[https://usegalaxy.eu/join-training/ueb\\_bi2021](https://usegalaxy.eu/join-training/ueb_bi2021)

# Quality Control

## Your turn!

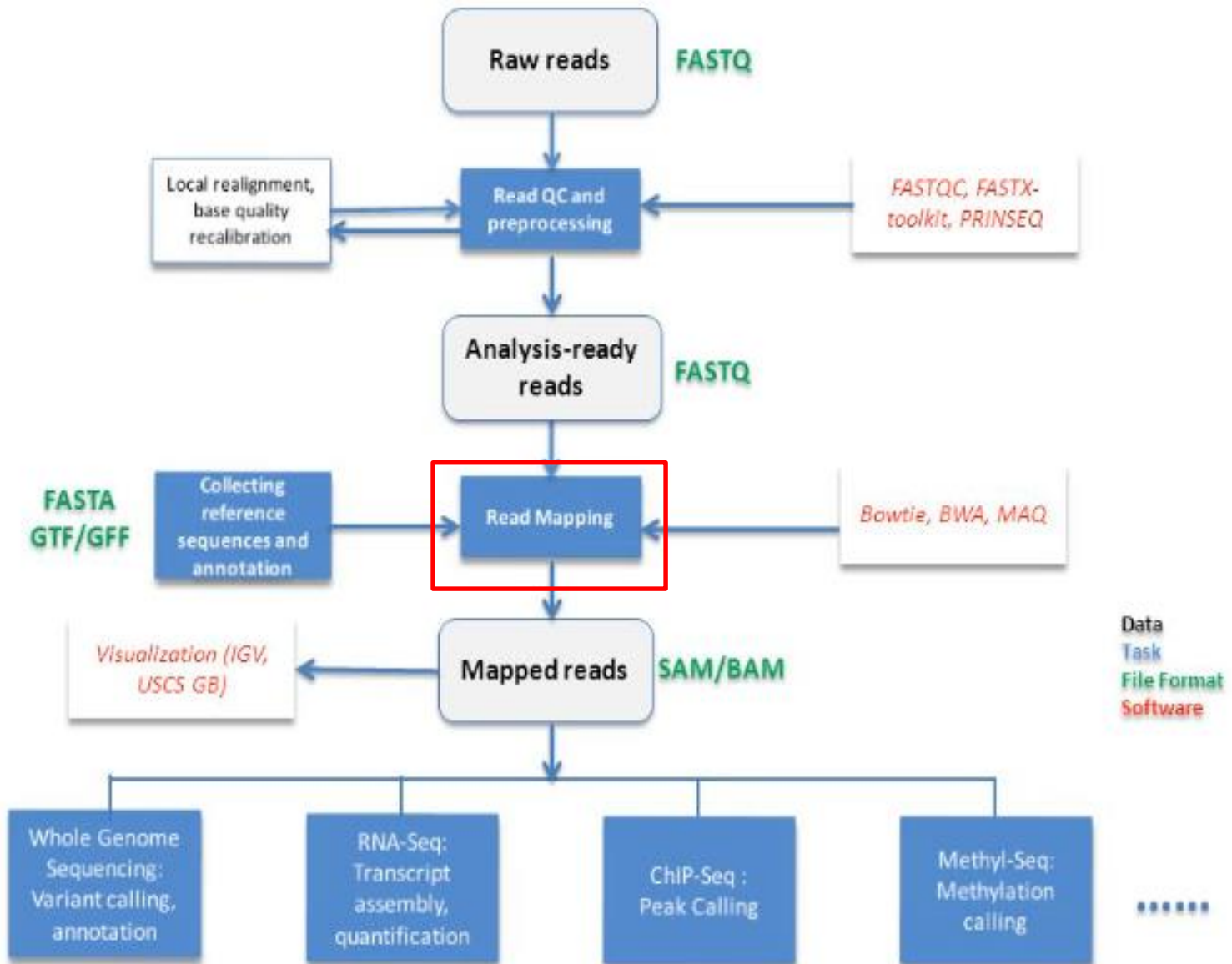
1. Create a new history and name it as you want (eg. Practica1)
2. Upload the fastq files into Galaxy from the urls copied above
3. Update the attributes of the two datasets (pencil icon):
  - a) Rename the datasets to “sample-f.fq.gz” and “sample-r.fq.gz”, respectively.
  - b) Check data type is set to “fastqsanger”
  - c) Associate the dataset with the human hg38 genome in the Database/Build field.
4. Run a quality control on each dataset using the FastQC tool.
  - a) What is the length of reads?
  - b) Are sequences of good quality? Any adapter that should be removed?
5. What would be the next step in the analysis workflow?

# Quality Control

## Your turn!

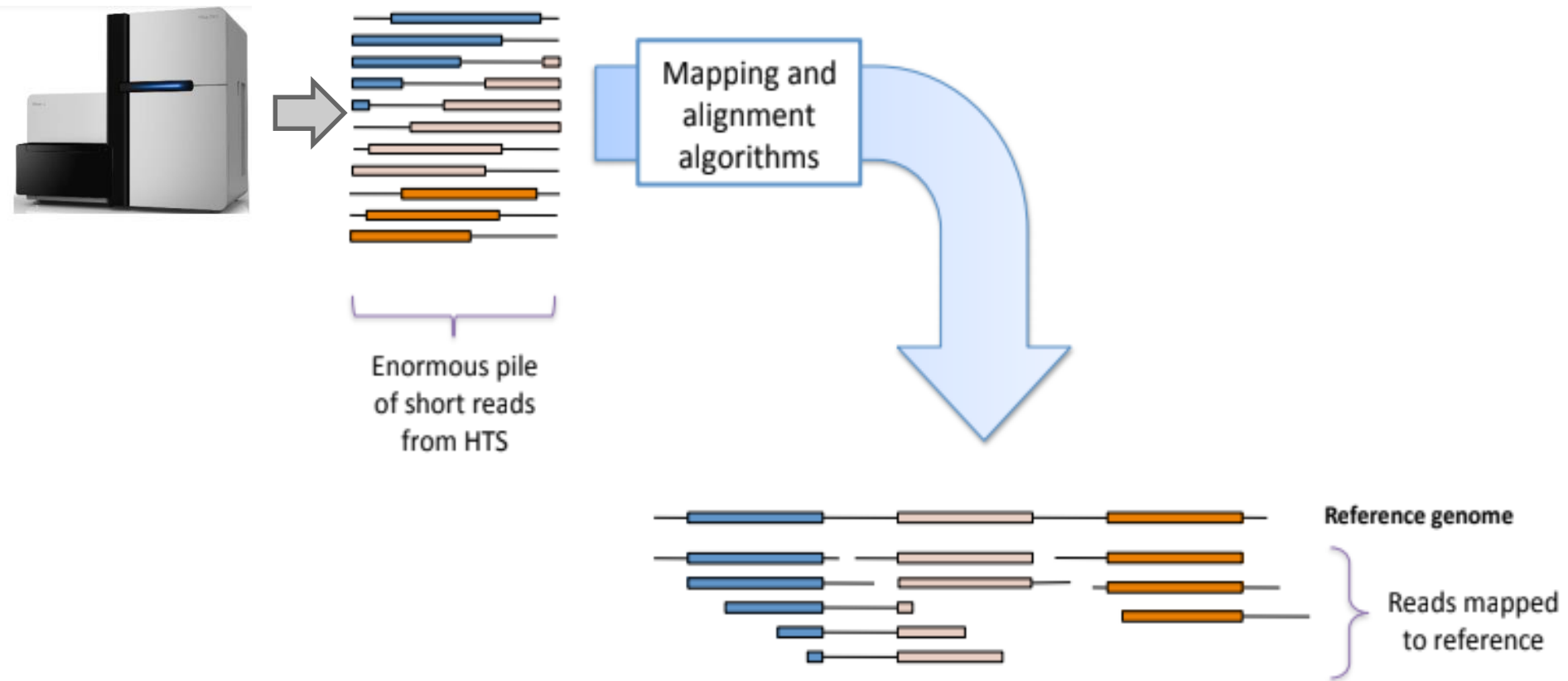
6. Trim the reads in each dataset using **Cutadapt** tool. Set the parameters:
  - a) Paired-end data
    - File 1: sample-f (forward)
    - File 2: sample-r (reverse)
  - b) Determine from the FastQC boxplot where the quality of the reads begins to drop off sharply. Calculate how many bases have to be trimmed from the end and use that number as the Offset from 3' end.
  - c) Output options: Report=yes
7. Inspect the results:
  - a) How many datasets do we get? Rename them to sample-f-trim / sample-r-trim, respectively. What is their format?
  - b) Do they have the same number of reads?
8. Re-run FastQC on the trimmed data, and inspect the new FastQC report. Has the sequence quality been improved?
9. Convert your analysis history into a workflow
10. What would be the next step in the analysis workflow?

# Steps in NGS analysis



# Steps in NGS analysis

## Mapping reads to the genome

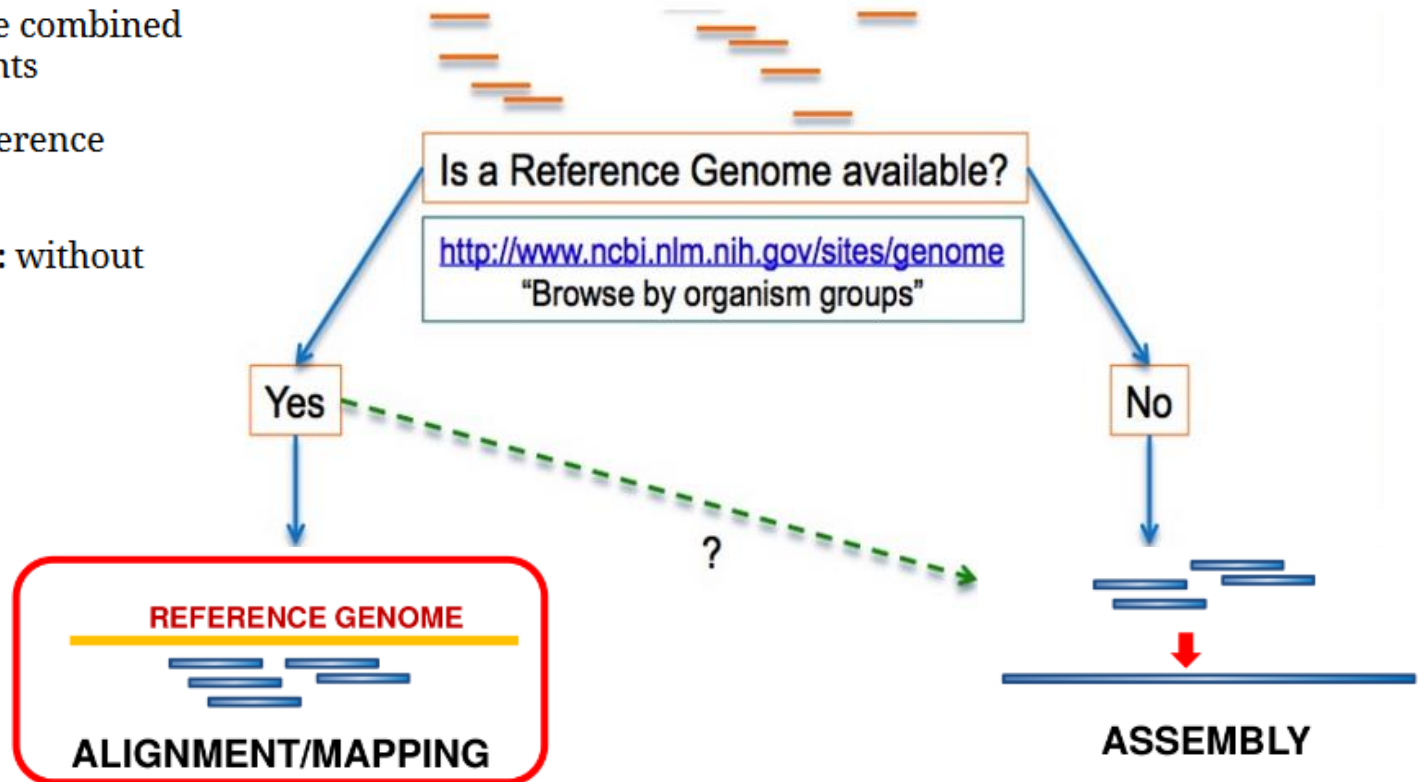


# Steps in NGS analysis

## 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

location 1 (mismatch)

. . . TTT**AGAATGAGCCGAG**TTCGCGCGCGGGT**AGAAT-AGCCGAG**TT . . .

||||| |||||  
AGAATTAGCCGAG

13 bp read

location 2 (deletion)

||||| |||||  
AGAATTAGCCGAG

13 bp read

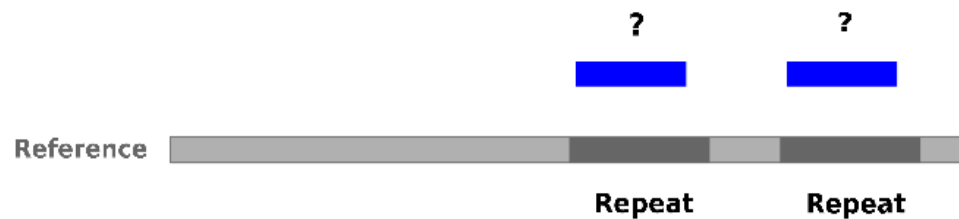
genomic DNA

# 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

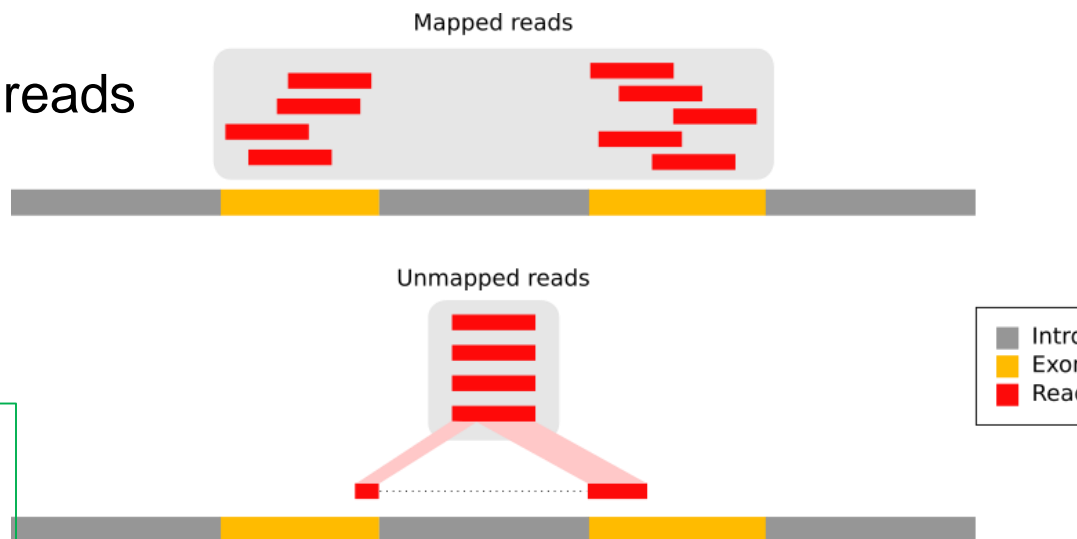
## Mapping reads to the genome

### 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



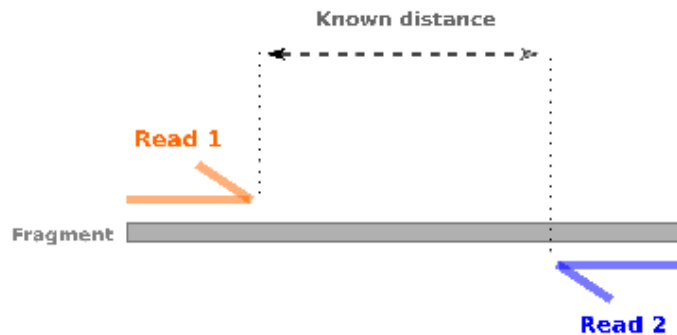
- Complex algorithms have been developed
- Choose appropriate tool/parameters



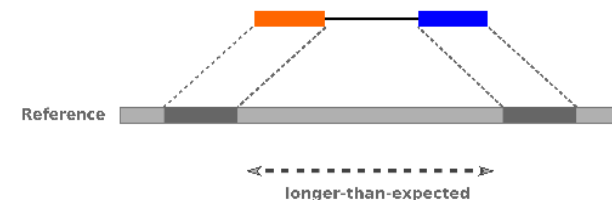
# Steps in NGS analysis

## Mapping reads to the genome

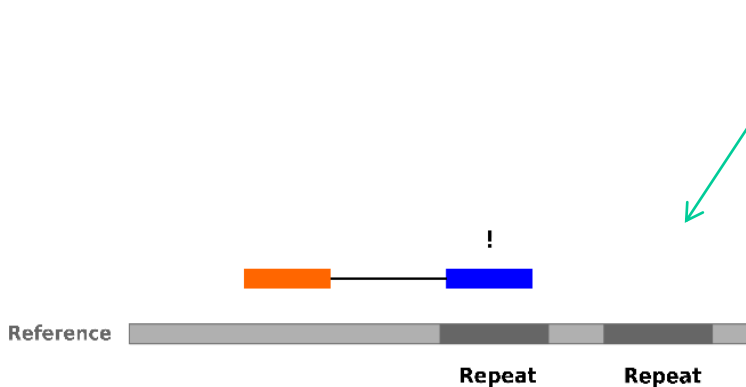
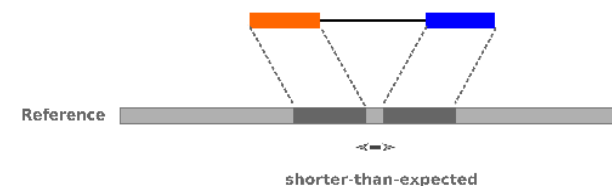
- Paired-end sequencing improves accuracy of mapping
- Sequencing:** Cut longer fragments of DNA, sequence only the ends



- Deletions:** Longer mapping distance than expected



- Insertions:** Shorter mapping distance than expected



# 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

Sequence One : GGCTGG

Sequence Two : GAGG

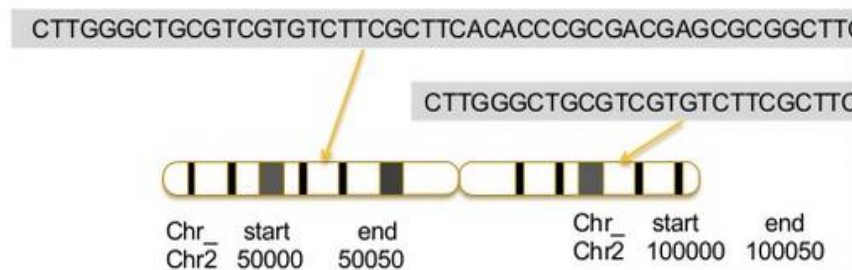
G	G	C	T	G	G
G	A	-	-	G	G
10	-5	-5	-1	10	10
10	5	0	-1	9	19

Your cumulative score

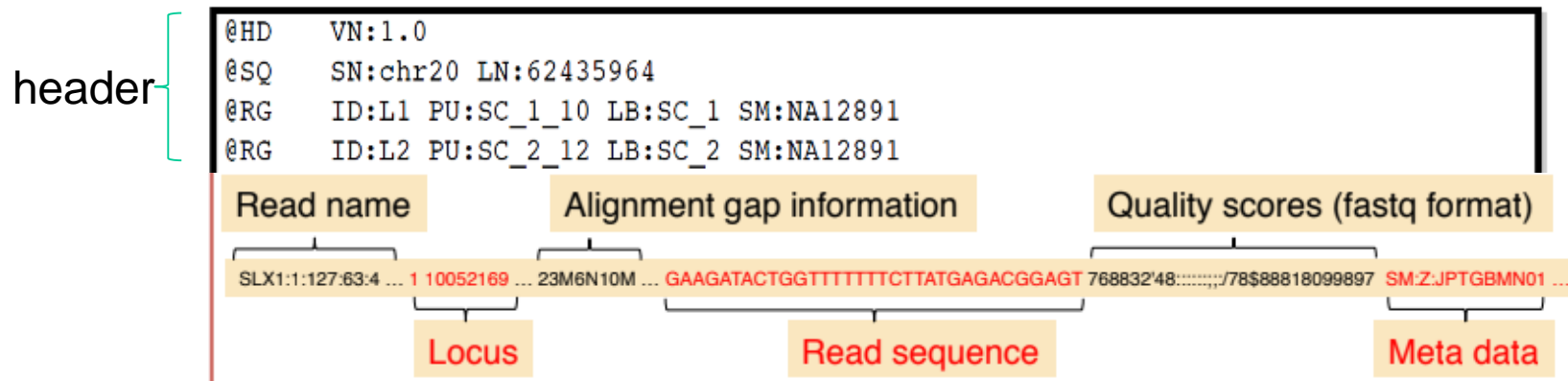
# Steps in NGS analysis

## 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:



# Steps in NGS analysis

