

# Consensus genome generation and assembly

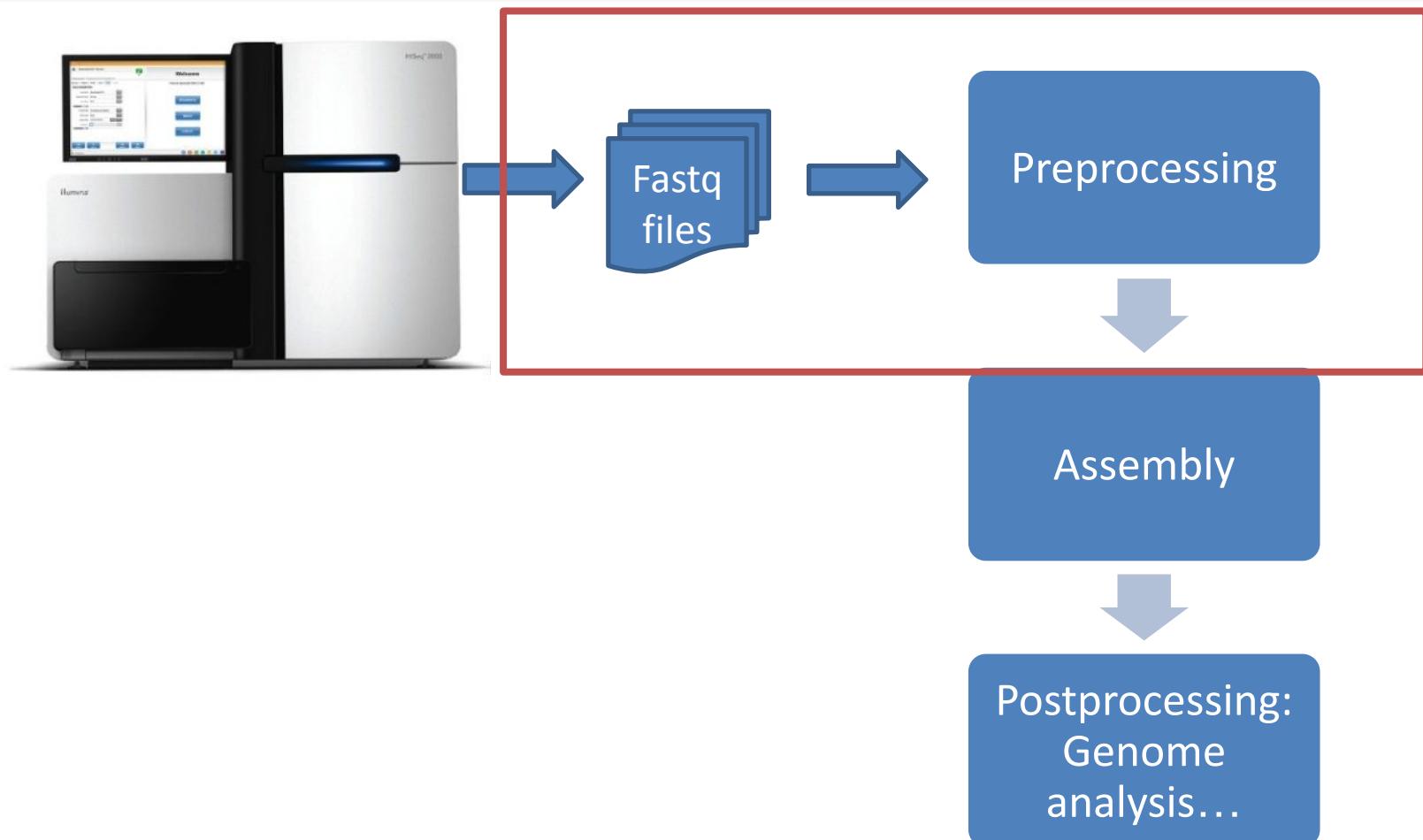
Sara Monzón Fernández

BU-ISCIII

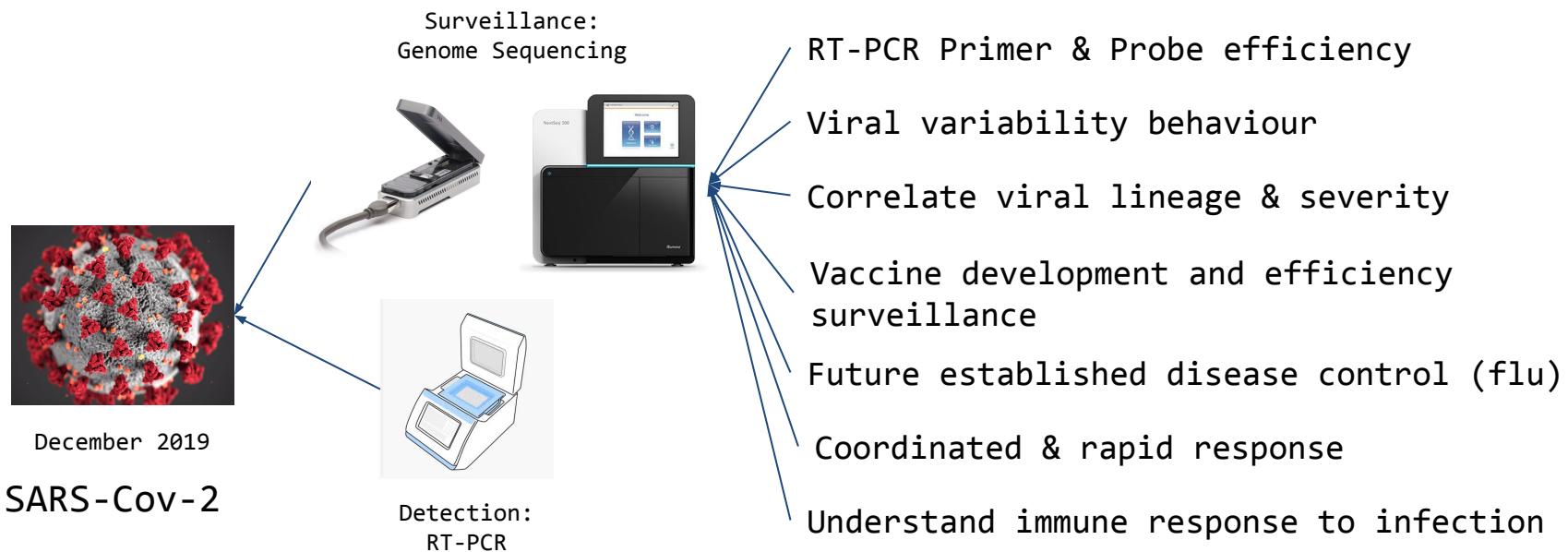
Unidades Comunes Científico Técnicas - SGSAFI-ISCIII

20 Diciembre 2023  
UAH - ISCIII

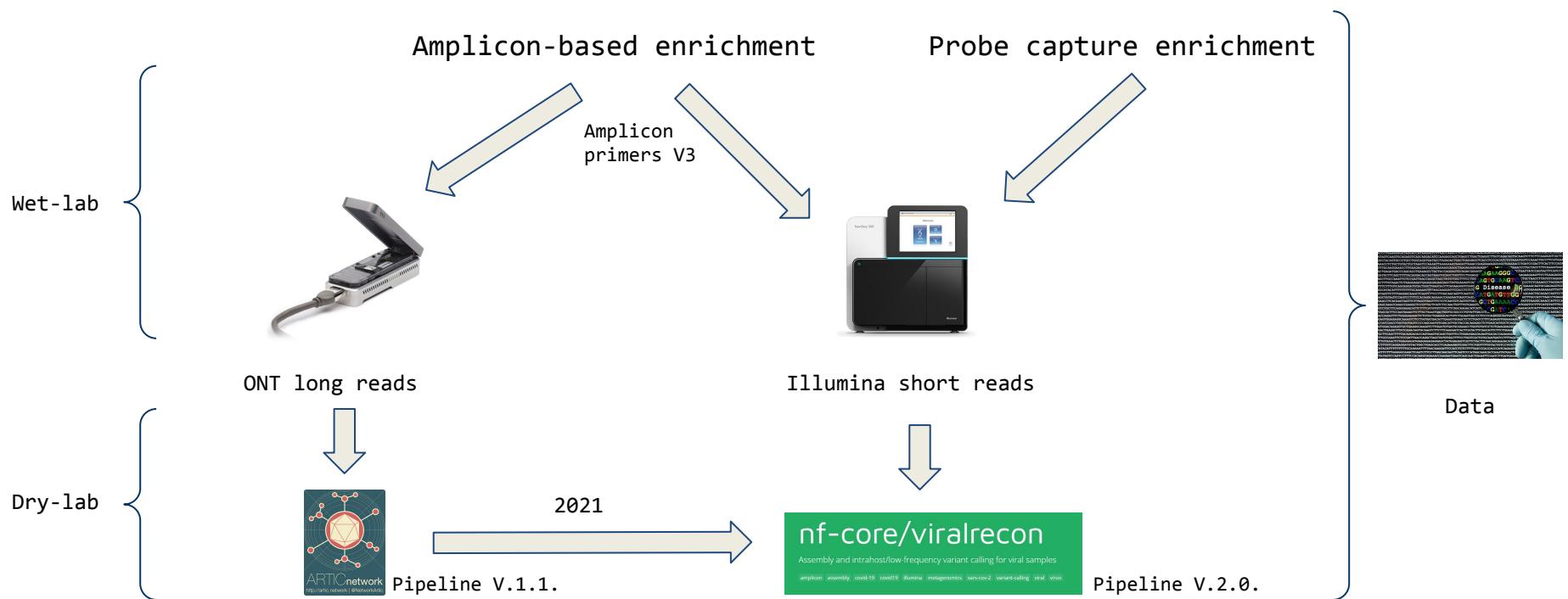
# Step in the process



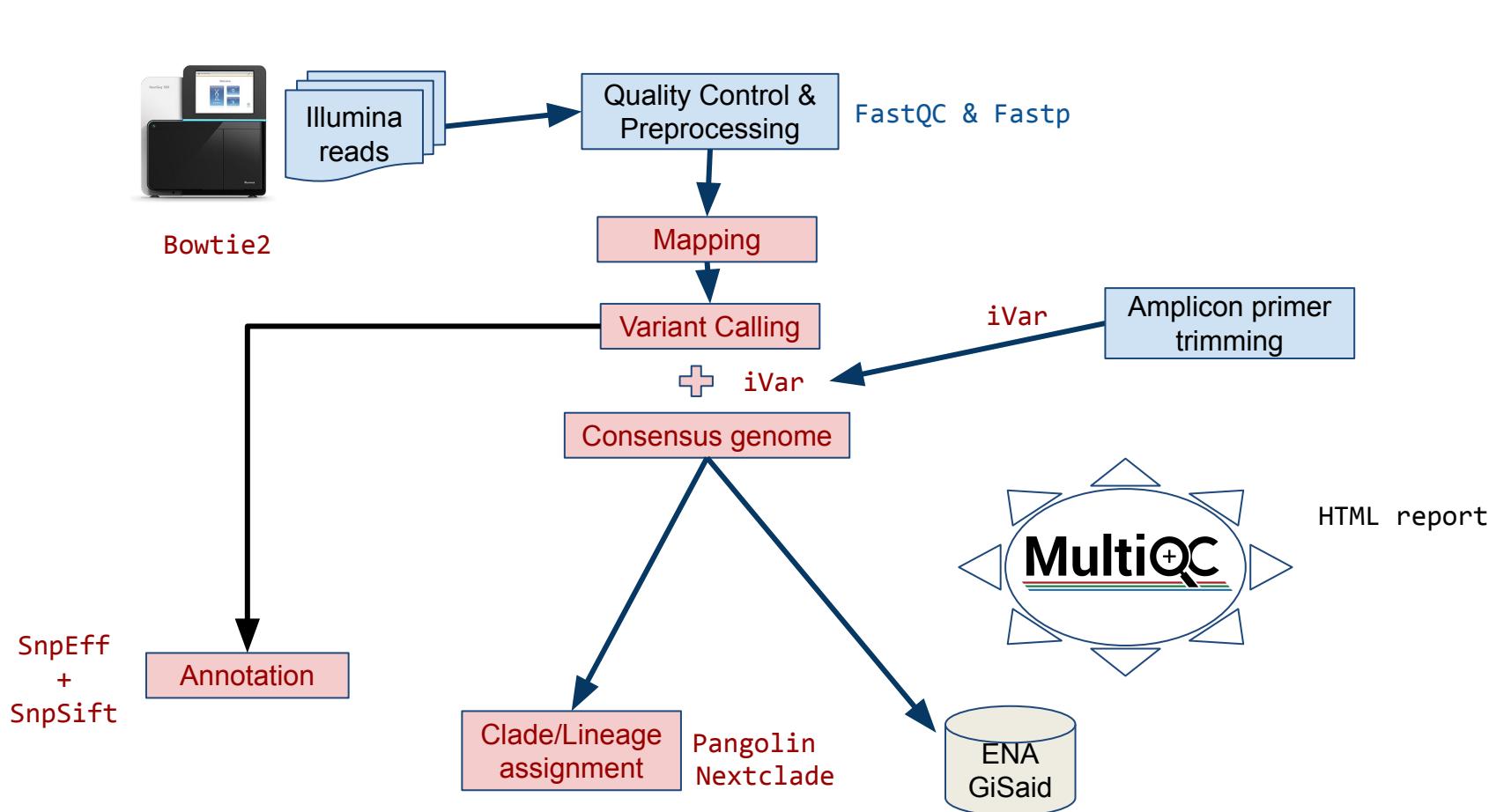
## 1. Background



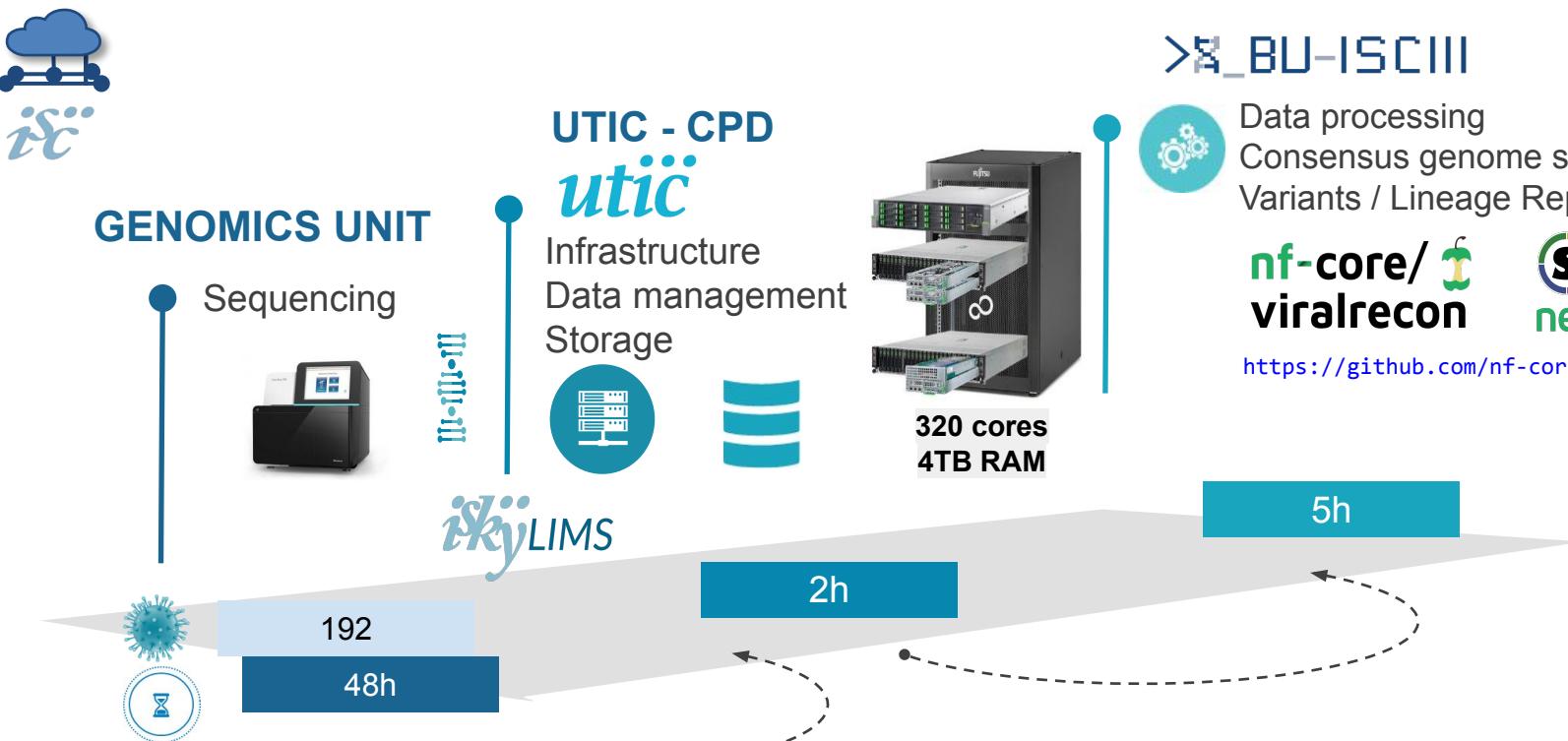
## 2. Sequencing Approaches



## 4. Viralrecon for Illumina reads



## 4. Viralrecon ISCIII nf-core





GOBIERNO  
DE ESPAÑA  
MINISTERIO  
DE CIENCIA, INNOVACIÓN  
Y UNIVERSIDADES

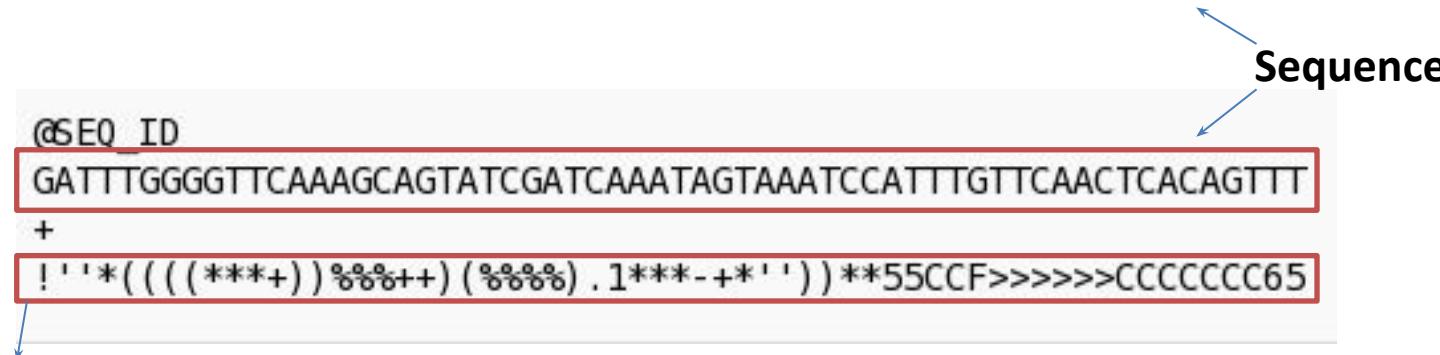
# Galaxy

# Introducción a Galaxy

# FASTQ format

- Is a FASTA file with quality information
- Within HTS, FASTA contain genomes y FASTQ reads

```
>SEQ_ID|  
AGCTTTCAATTCTGACTGCAACGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC  
TTCTGAAGTGGTTACCTGCCGTGAGTAAATTAAATTGACTTAGTCACAAATACTTAACCAA  
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC  
ATTACCACCAACCATTACCAACAGGTAACGGTGCGGGCTGACCGTACAGGAAACACAGAAAAAG
```



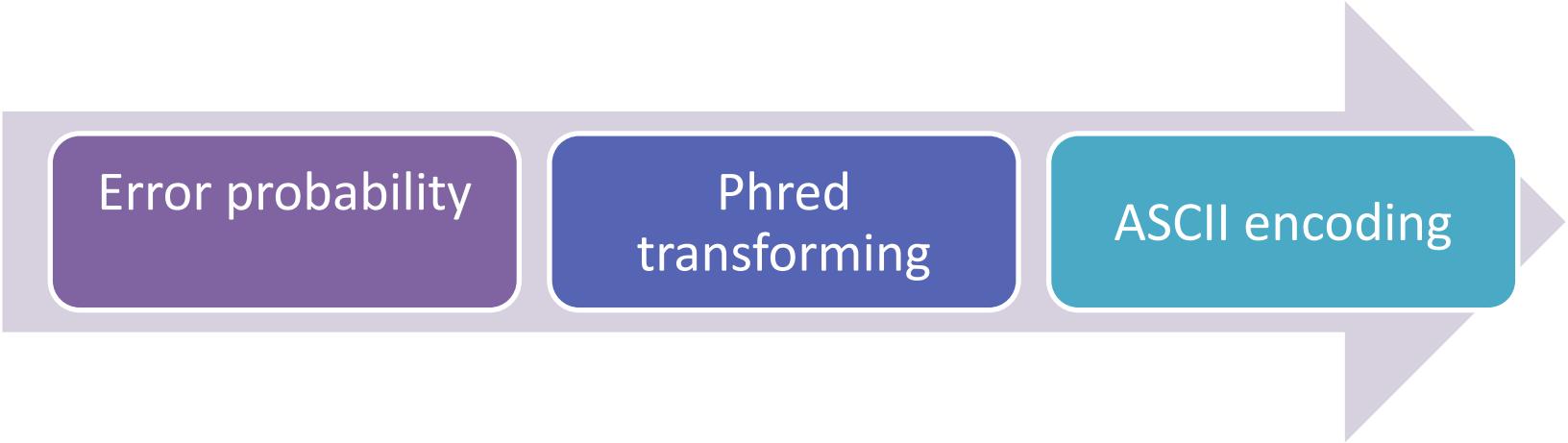
Sequence

```
@SEQ_ID  
GATTTGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT  
+  
! ' * ( ( ( ***+ ) % % + + ) ( % % % ) . 1 *** - + * ' ) ) **55CCF>>>>CCCCCCCC65
```

Quality: must be 1 bit

# FASTQ format

- Each base has an assigned quality score
  - Sequencing quality scores measure the probability that a base is called incorrectly
- How is it calculated?



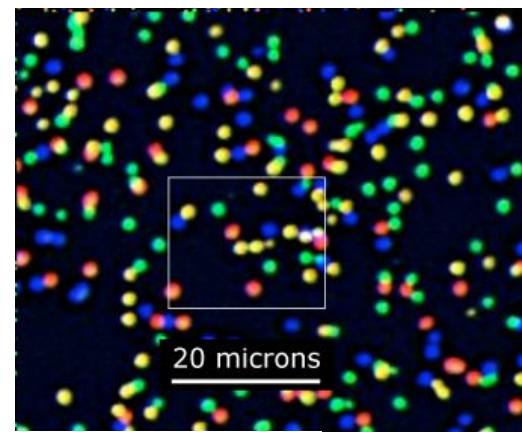
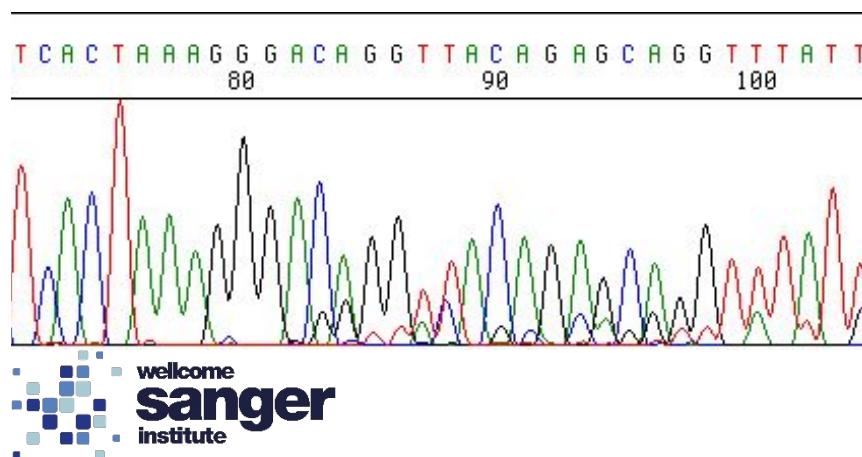
Error probability

Phred  
transforming

ASCII encoding

# Phred quality and error probability

- **Light intensity** is used to calculate the error probabilities
- Convert error probability into Phred score quality - Ewing B, Green P. (1998)
- Phred originated as an algorithmic approach that considered Sanger sequencing metrics, such as **peak resolution and shape**



illumina®

# Phred quality and error probability

- Convert error probability into Phred score quality - in real time on Illumina platforms
- Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)
- Phred quality range between 0-40 for Sanger and Illumina 1.8+

$$Q = -10 \log_{10} P$$

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%
50	1 in 100,000	99.999%

# Phred quality and error probability

- Convert Phred quality score into ASCII, a compact form, which uses only 1 byte per quality value

ASCII\_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII									
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41 )	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

- Phred+33 (Sanger and current Illumina). 0 Phred quality correspond to decimal 33, which is the symbol !

ASCII\_BASE=64 Old Illumina

Q	P_error	ASCII									
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93 ]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

- Phred+64 (Solexa and Illumina 1.3-1.5)

# Phred quality and error probability

- Phred 33 example

```
@HWI-ST731_6:1:1101:1322:1938#1@0/1
NTGACAAAGGGCTAATATCCAGAACATCTACA[AAGAACTTAAACAAATGTATAAGAATAAAAGTATAGTGCTAACAT
+
#1:BDDADDFDFDD@F>BGFIIIB@CFHIHICAGBC9CBCBGGIGCFF??>GGHFHIGGEGI<FECGDE=FHCHEG=
```

$$P=0.001 \rightarrow Q=-10 \cdot \log_{10}(0.001)= 30 \rightarrow \text{ASCII} 33+30 = 63 \rightarrow ?$$

$$P=0.0001 \rightarrow Q=-10 \cdot \log_{10}(0.0001)= 40 \rightarrow \text{ASCII} 33+40 = 73 \rightarrow ?$$

Quality encoding: !#\$%&'()\*+,-./0123456789:;=>?@ABCDEFGHI

| | | | |

Quality score: 0.....10.....20.....30.....40

# FASTQ format

## Illumina read header

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

<b>HWUSI-EAS100R</b>	the unique instrument name
<b>6</b>	flowcell lane
<b>73</b>	tile number within the flowcell lane
<b>941</b>	'x'-coordinate of the cluster within the tile
<b>1973</b>	'y'-coordinate of the cluster within the tile
<b>#0</b>	index number for a multiplexed sample (0 for no indexing)
<b>/1</b>	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

```

@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:2458:1027 1:N:0:ACAGTG
AGAAAAAAACCTGGANGGAAAAAAATCAGACATTCTAGAGGTGGAAGGCAAACGTGAAACAAAGAAATAATTCA
+
DGGGEDHHHHGGGFEBACBCA<?HHHHBHHHHHHHHHHHEHEFEGGGGG/GGDDDGHFHGFCHFHHHEH8
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3082:1029 1:N:0:ACAGTG
GGTAATACTAGACTGANATGATCAAAGGCATGCTGGAAACAAACCTATTAAAGATAAGCTGGATCAAGCTTCATT
+
B:B:B?BB:/=-55177#55877<775EDD>E=B?BBBBGGDDAG@G>GGGGGG@)EEEEBEG>GGGGGGAAA?<D
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3185:1033 1:N:0:ACAGTG
TCTGGGACATTGCTCNTGGCTGGAGTCACCTGCTGGACATTGCTCAGGGCTGGAGACACGTGTTGGAGGGAC
+
BC??A66;)74781<#??;452.27'64(8,851DDG8GB######
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3268:1033 1:N:0:ACAGTG
ATTCAAATTAGAAGANAGTTGATGTTCTCATGATGCCAAAATTCACTGAGAAAACCTTTTTAAGCCAC
+
IIIIIIIIIFFFF#ABACFEFFIFIIFIHE@IIIIIIIIHHIIIFIIF>HHIHFIDIIIIIGFHIEGH
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3400:1035 1:N:0:ACAGTG
TCCTGCTTAGGAGANTCCTCATGCTCTGACAGGATGCTCTATGTGAGTTGAGCTGGCTTCTCACTTTATAG
+
IIIIIIHHIIIGGE#AAC@?=?BHHIIIIHHIIHIIHIIHIIHGHGIGIHFEGGGGGH@EFGCEFAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3962:1033 1:N:0:ACAGTG
CCACCAACACAGTCTNCACCTTCTGGTGTGGATAGATTTGCACCTTCCATCCTCAGGTTCAAATAGC
+
HHFHDHDHH>C?CA#EEEE>?A?>HHDHGEBGBCEEHHF8HEEEHECH,=>>=EAEE>BEBBAEACAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:4491:1028 1:N:0:ACAGTG
AGAGAGAGAGAGAGAGAGAGACTCTGGAGATGCCAAGCACAAGCCTGCAAGAGTCCCAGCAAAGAAAATAAAA
+
GADGEGEGEGBBB?B#Q=@@72;64GGGFGB>GGGBDG<DBGB<DA??/?#####

```

ASCII-coded (0-40):

- “!#\$%” lowest quality
- “FGHI” highest quality

# Sequencing quality assessment

- To asses quality, software uses **Phred per-base quality score** is used
- Is the **first quality control step** after sequencing. There should be one after every step of the analysis
- After quality assessment user can know how **reliable** are their datasets
- QC will determine the next **filtering** step
- Filtering decisions will **impact** directly in **further analysis**
- Many other steps also use this quality as variable in their **algorithms**

# Sequencing quality assessment: Artifacts

HTS methods are bounded by their technical and theoretical limitations and sequencing errors cannot be completely eliminated (Hadigol M, Khiabanian H. 2018)

- **Artifacts in library preparation**

- Remaining adapters
- High rate of duplicates
- GC regions bias
- Polymerase error rate
- DNA damage during breakdown

- **Artifacts during sequencing**

- Low quality in sequence ends (Phasing: cluster loose sync)
- Complication in certain regions:
  - Repetitions
  - Homopolymers
  - High CG content

# FastQC: Basic Statistics

- Self defined overall stats
  - Encoding: Phred33 or Phred64



## Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47



## Basic Statistics

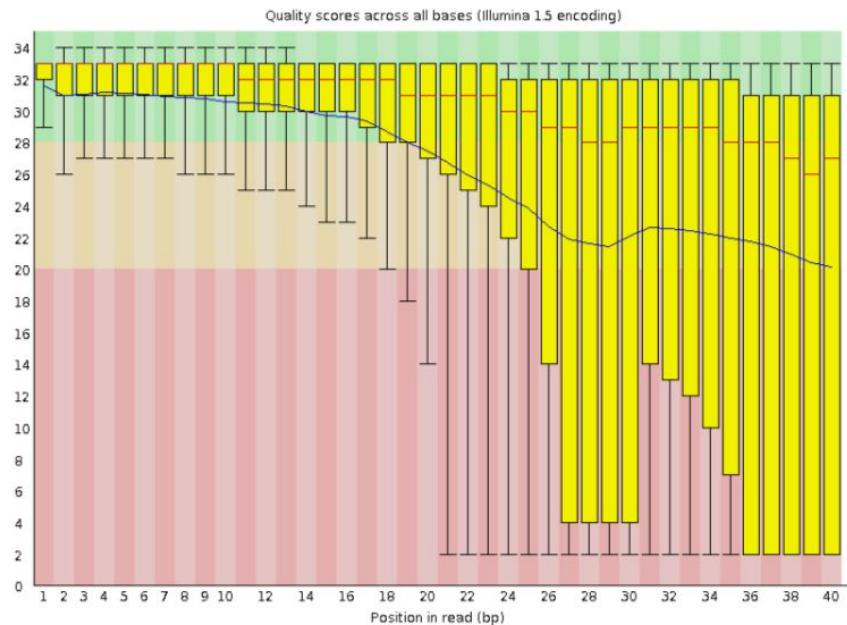
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

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

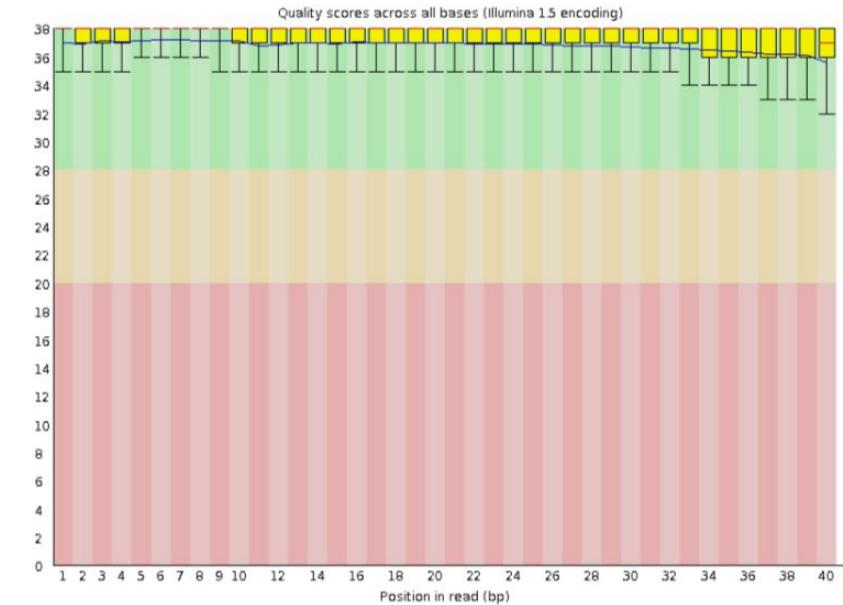
# FastQC: Per base sequence quality

- Overview of the range of quality values across all bases at each position in the FastQ file
- Median, inter-quartile range (25-75%), 10-90% points, mean quality

## ✖ Per base sequence quality



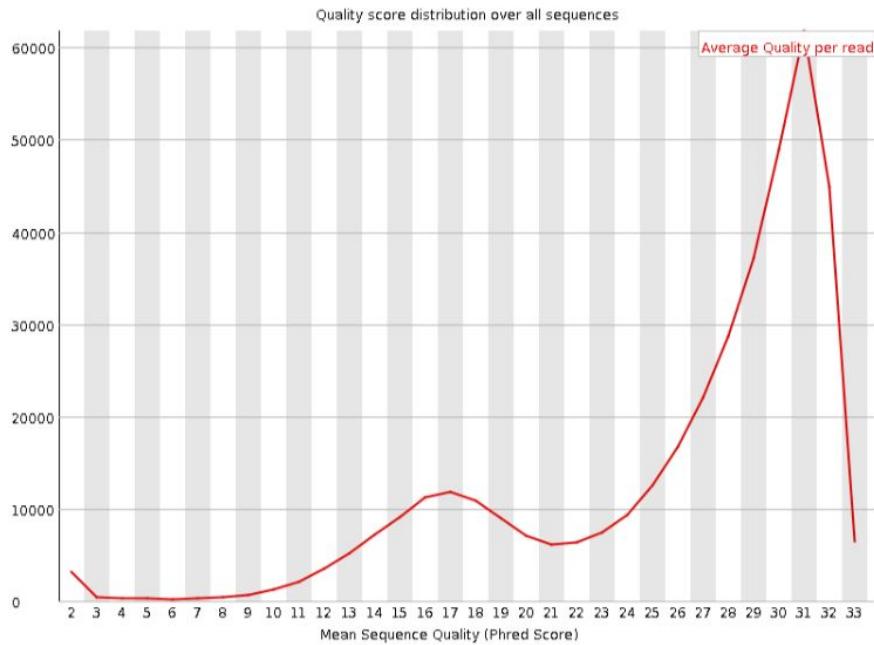
## ✓ Per base sequence quality



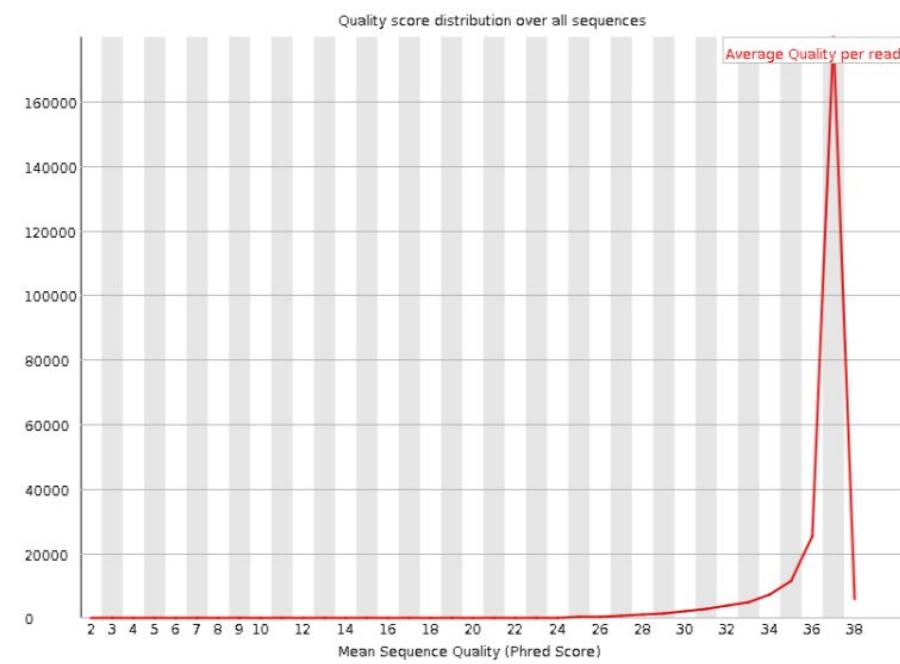
# FastQC: Per sequence quality score

- Number of sequences with the same mean quality

✓ Per sequence quality scores



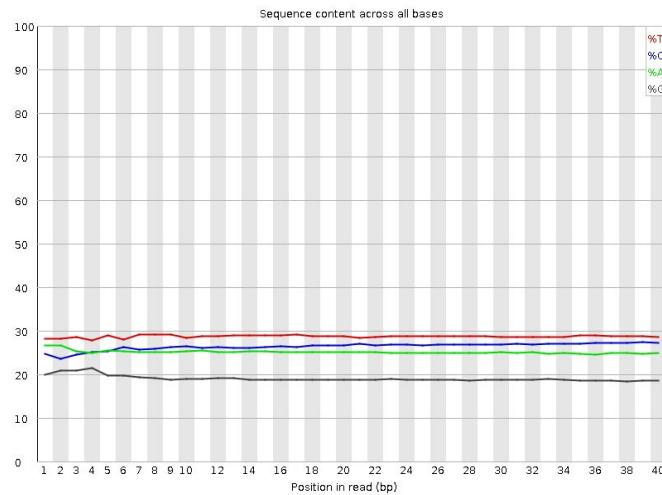
✓ Per sequence quality scores



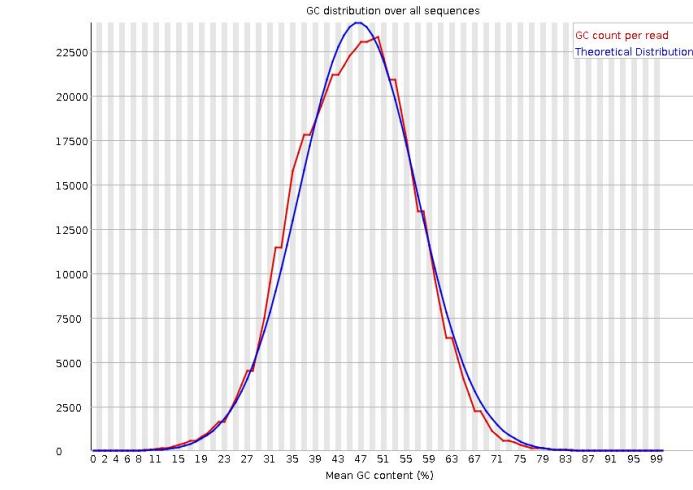
# FastQC: Nucleotide related errors

- How expected nucleotide distribution deviates from expected
  - Per base sequence content
  - Per base GC content
  - Per sequence GC content
  - Per base N content

⚠ Per base sequence content

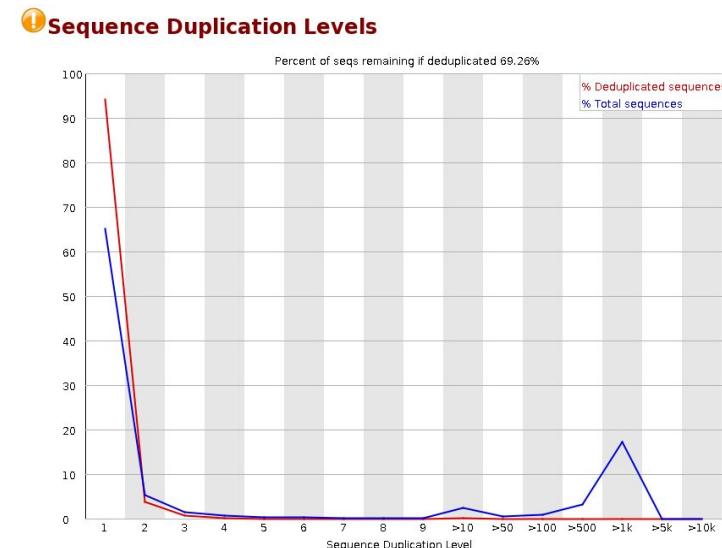
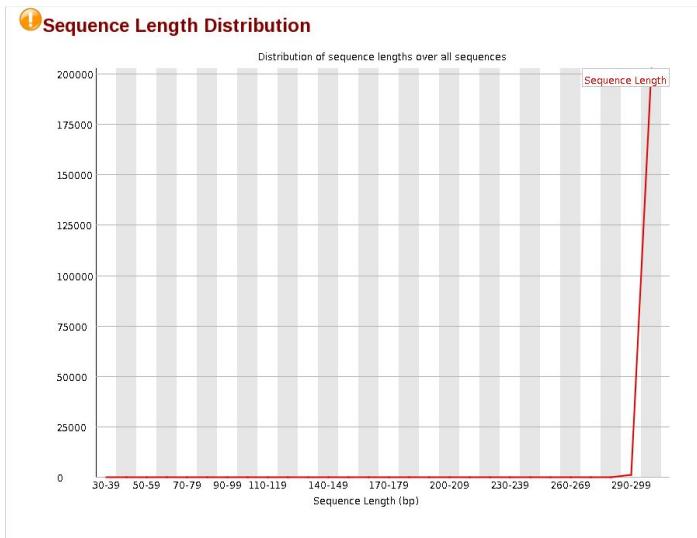


✅ Per sequence GC content



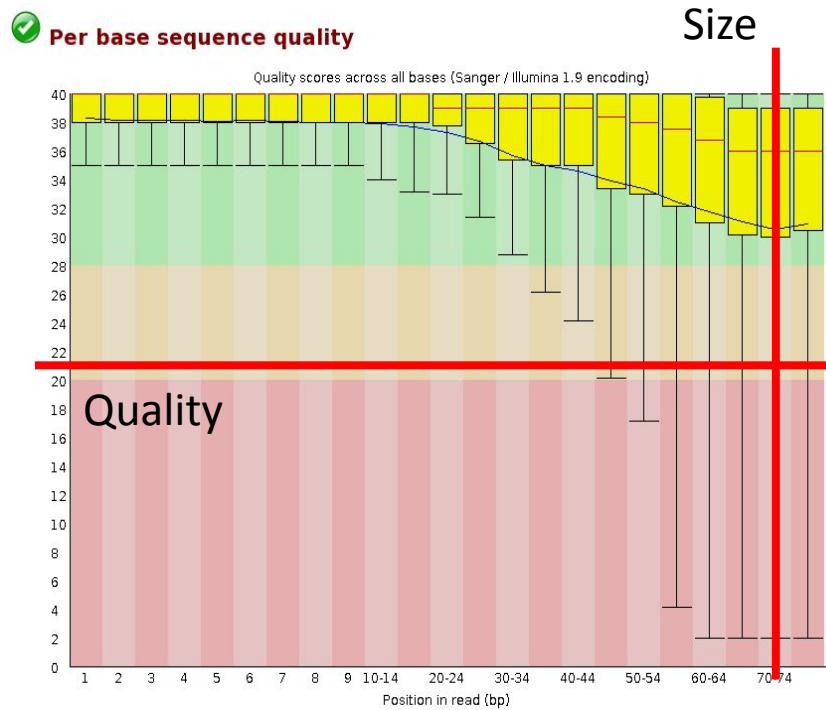
# FastQC: Sequence related errors

- How expected nucleotide distribution deviates from expected
  - Sequence Length Distribution - Fragments
  - Sequence Duplication Levels
  - Overrepresented sequences
  - Adapter Content



# Sequence filtering

- Remove residual adapters
  - Depending on used library
- Filtering parameters
  - Quality filtering
    - Overall mean quality
    - Local mean quality
      - Sequence end
      - Sliding window
  - Size filtering
    - Overall sequence size
    - Remaining sequence size after filtering





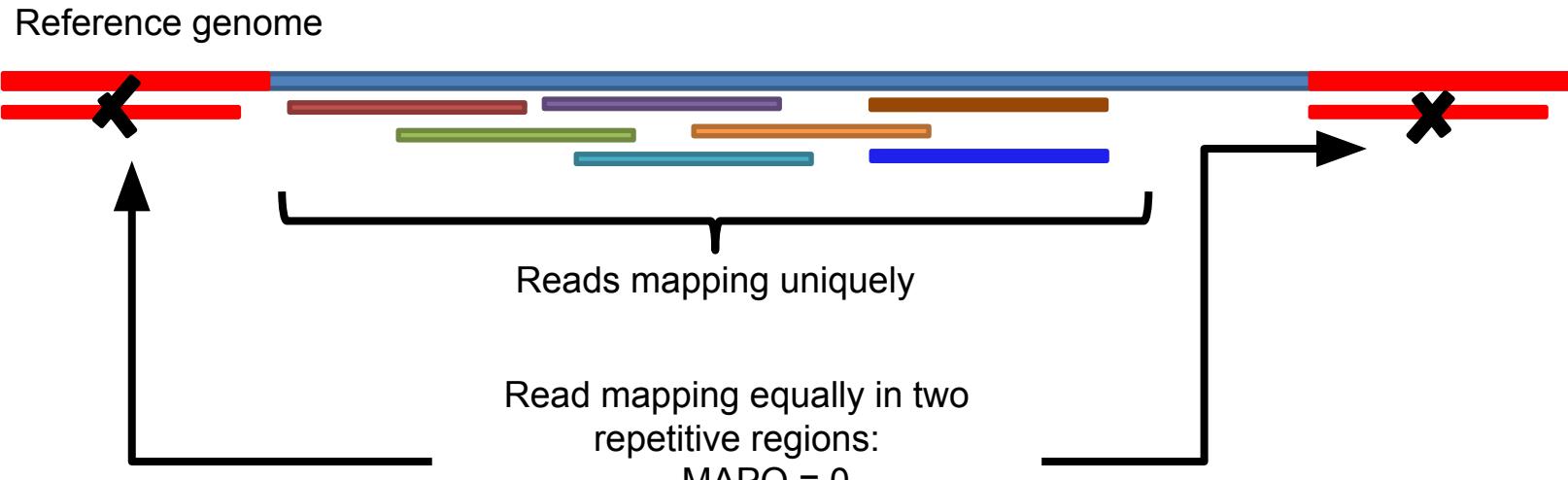
GOBIERNO  
DE ESPAÑA  
MINISTERIO  
DE CIENCIA, INNOVACIÓN  
Y UNIVERSIDADES

# CONTROL DE CALIDAD EN GALAXY

# Mapping

BOWTIE2

Mapping software looks for the best match for each read in the genome.  
Paired-end reads help the mapper to find the perfect spot!

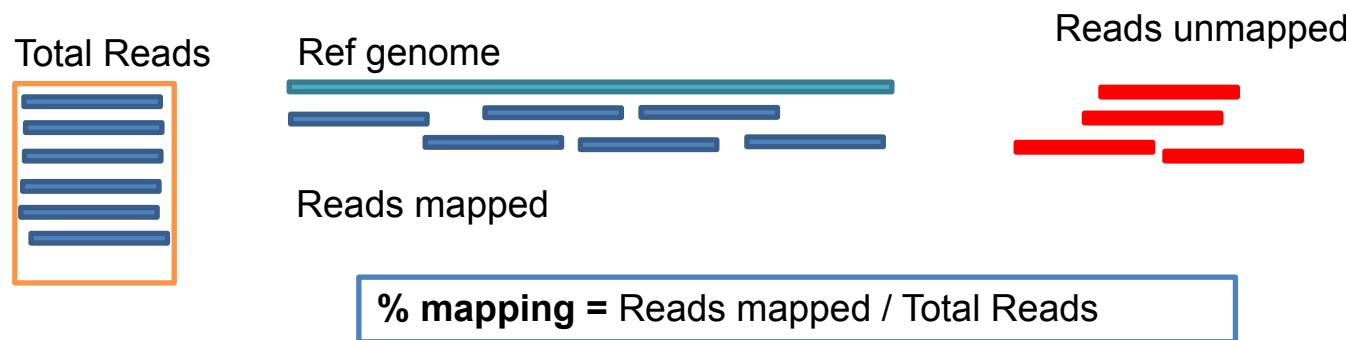


Depends on parameters!!

# Mapping quality control

- **% mapping:** number of reads mapping againts reference genome.

Picard  
Samtools



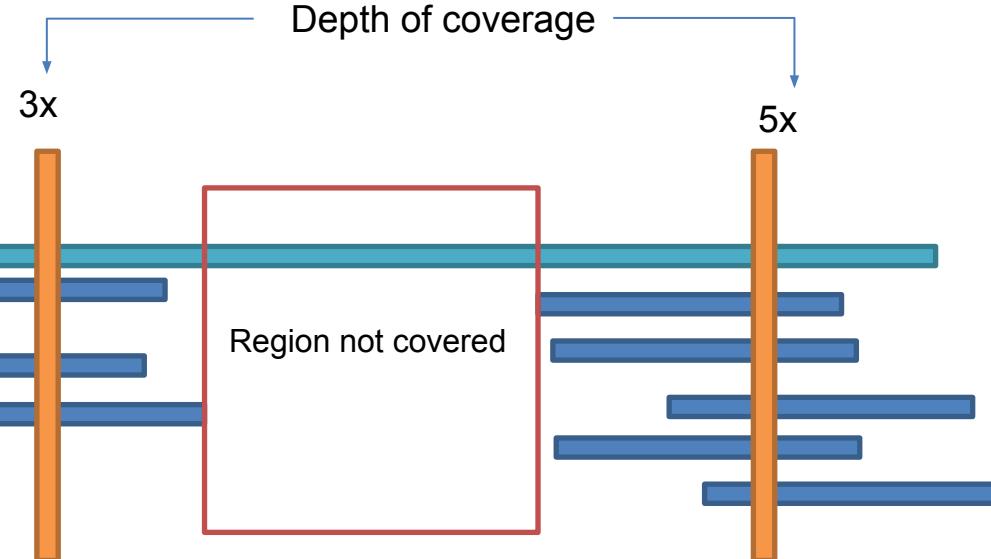
Mandatory parameter for microbial genomics!! It indicates us how many reads we have from our organism of interest. In human genomics this is almost always 99.99% unless something terrible happens. Not here!!!

# Mapping quality control

- **% genome > 10x:** percentage of genome covered with more than 10 reads.
- **Mean Depth of coverage:** mean of reads covering a genome position.

Picard  
Samtools

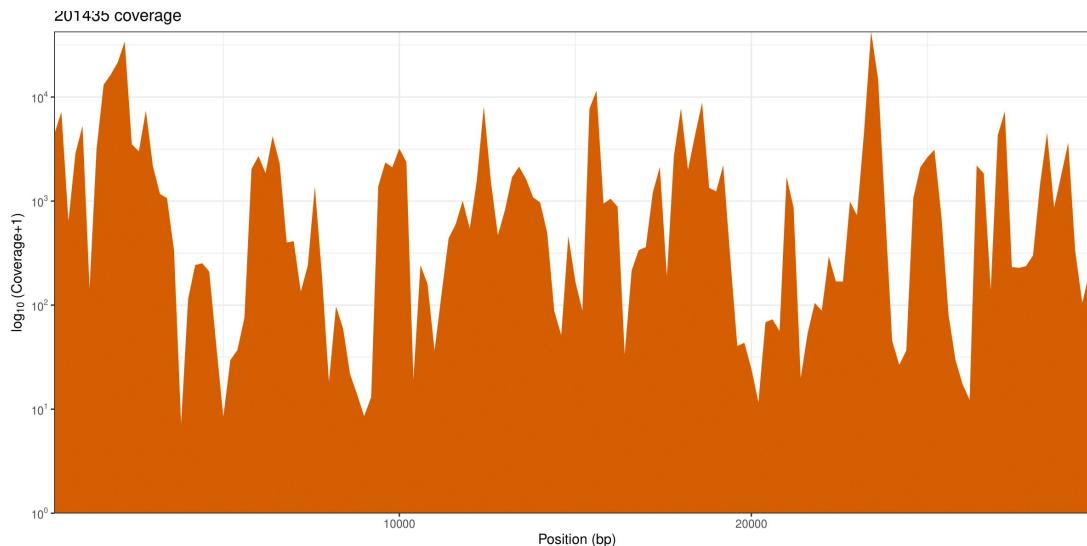
**% genome > 10x =**  
 $\text{Genome size} > 10x / \text{Genome size} * 100$



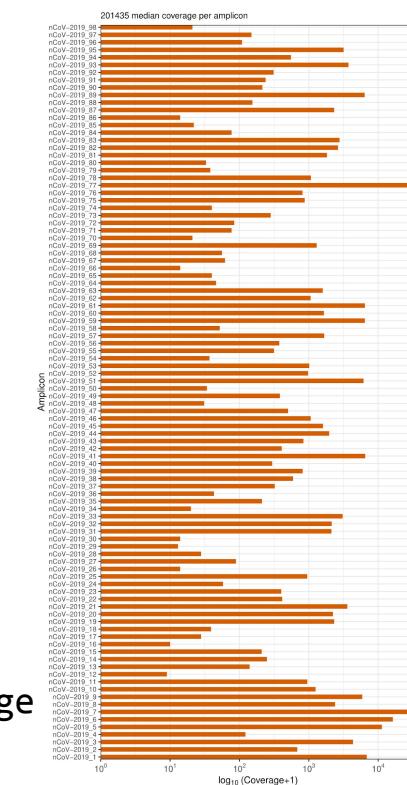
**Mean Depth of coverage =**  
 $\text{Depth of coverage in position 1, 2, 3 / Genome size}$

# Amplicon QC results

## Genome coverage



## Amplicon coverage

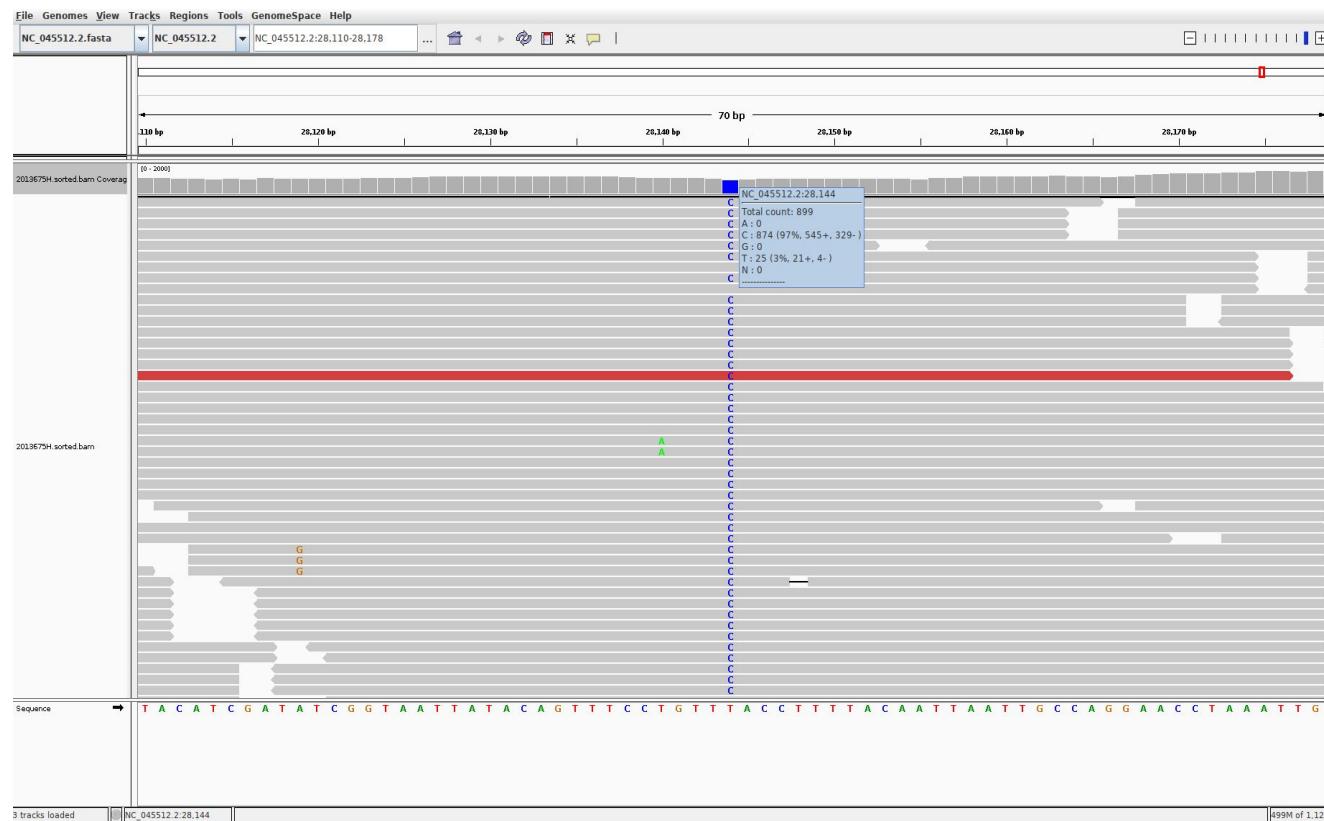




MINISTERIO  
DE CIENCIA, INNOVACIÓN  
Y UNIVERSIDADES

# MAPPING EN GALAXY

# Variant calling



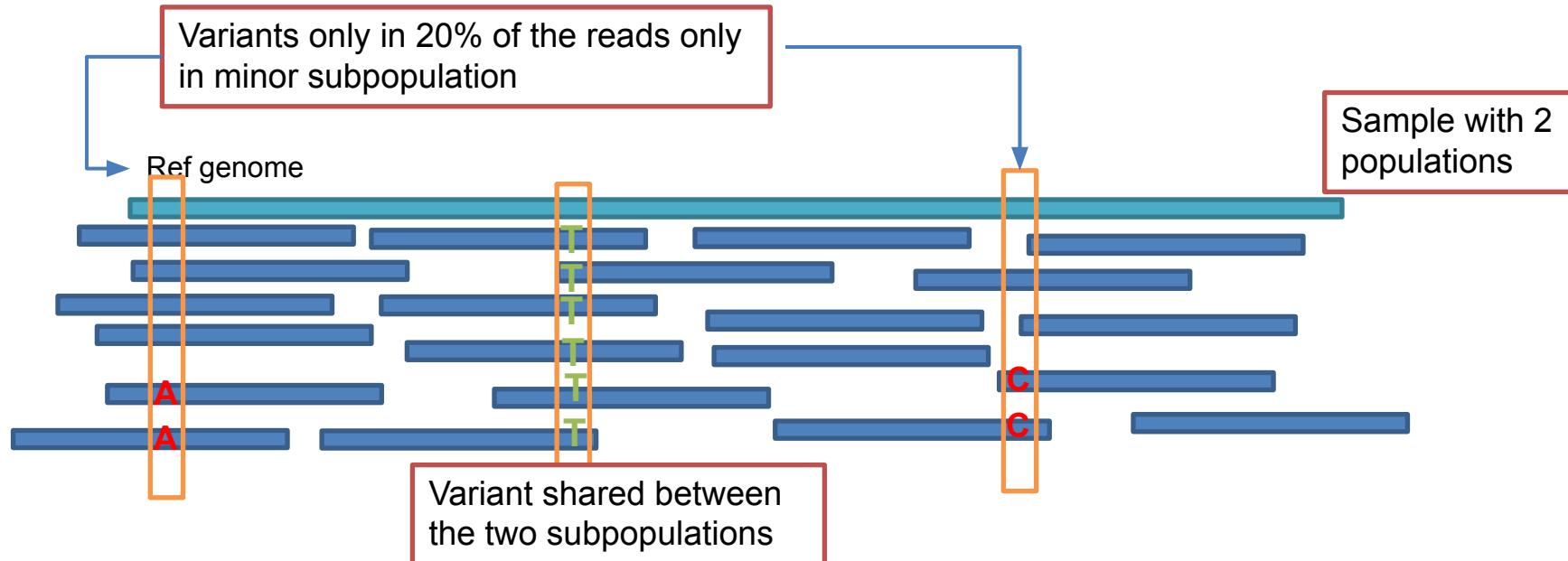
Variant callers walk through each alignment column independently and evaluate if there is a variant.

A variant is called in haploid genomes when it's supported by at least 90% of the reads.

VARSCAN2  
IVAR

# Viral subpopulation - Quasispecies

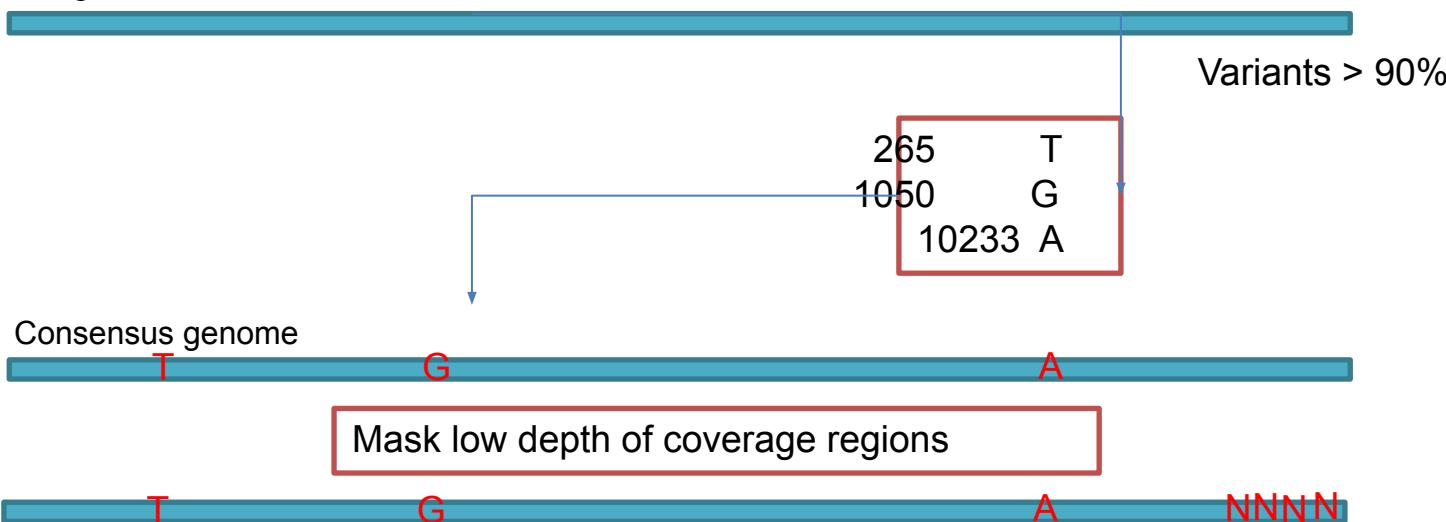
- Just as in clonal subpopulations in tumor samples, we can have viral subpopulations called quasispecies in viral samples.
- We detect them using the alternative allele frequency.



## Consensus genome

- Select variants: > 90% allele frequency
- Include variants in reference genome.
- Mask low frequency positions: <10x.

Ref genome

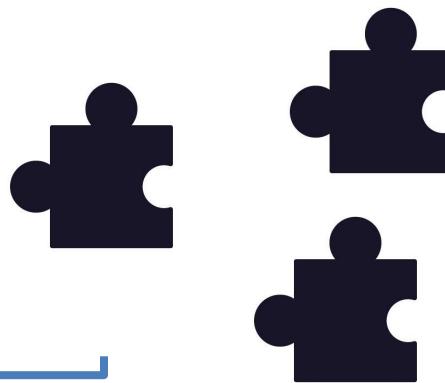
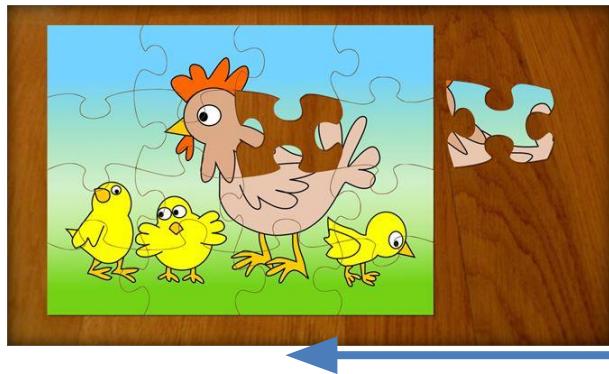




MINISTERIO  
DE CIENCIA, INNOVACIÓN  
Y UNIVERSIDADES

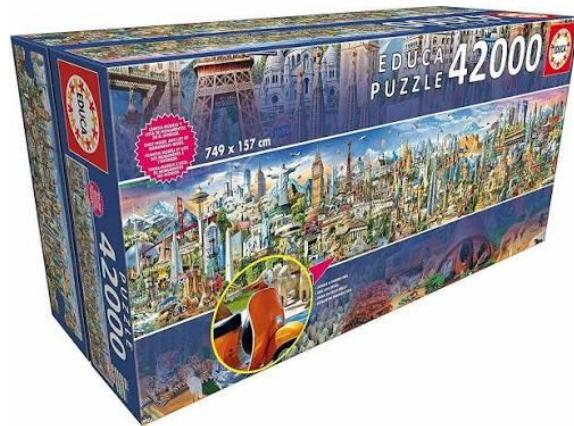
# VARIANT CALLING EN GALAXY

# Assembly



The bigger the pieces the easier to reconstruct the puzzle!

# Assembly



Obviously this a little bit  
more complicated....

We have LOADS of really  
little pieces and a big puzzle  
in proportion

# Assembly



Obviously this a little bit  
more complicated....

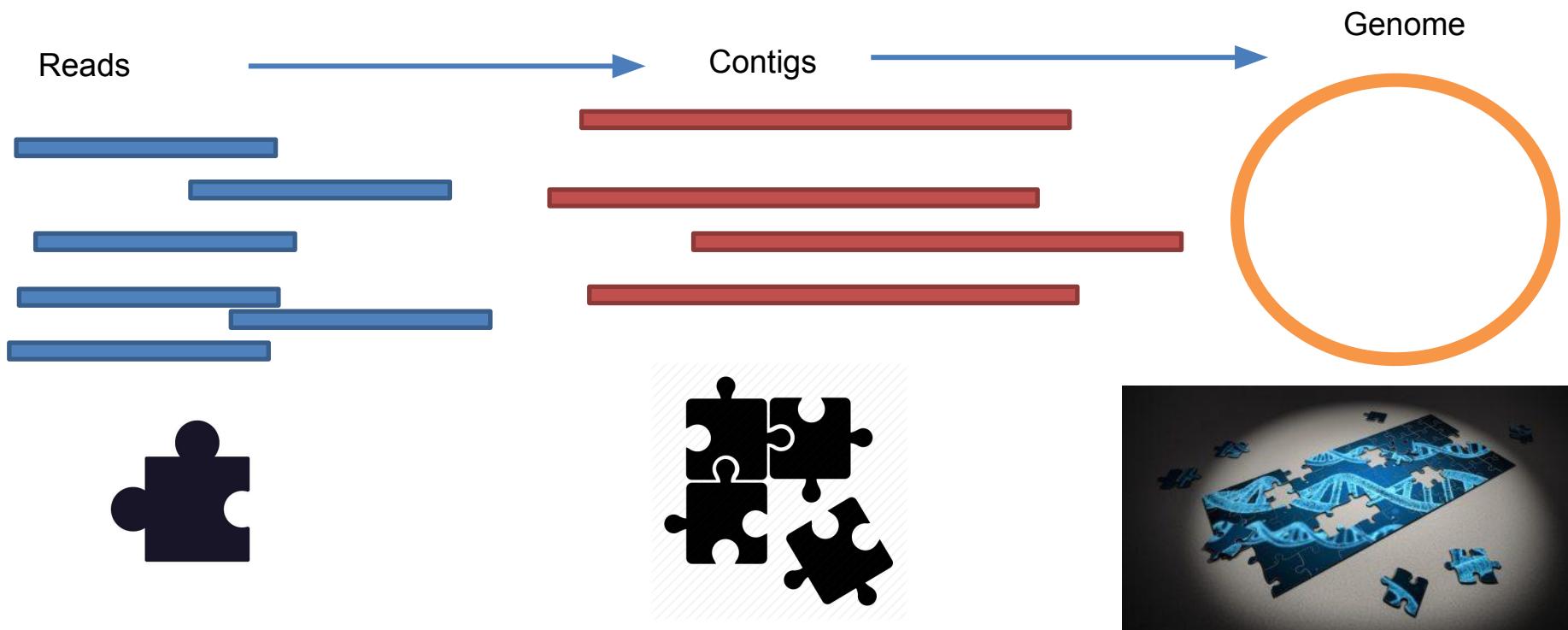
Actually we don't even have  
the box image most of the  
time

# Assembly

Reconstruct a representation of the original DNA from shorter DNA sequences or small fragments known as reads

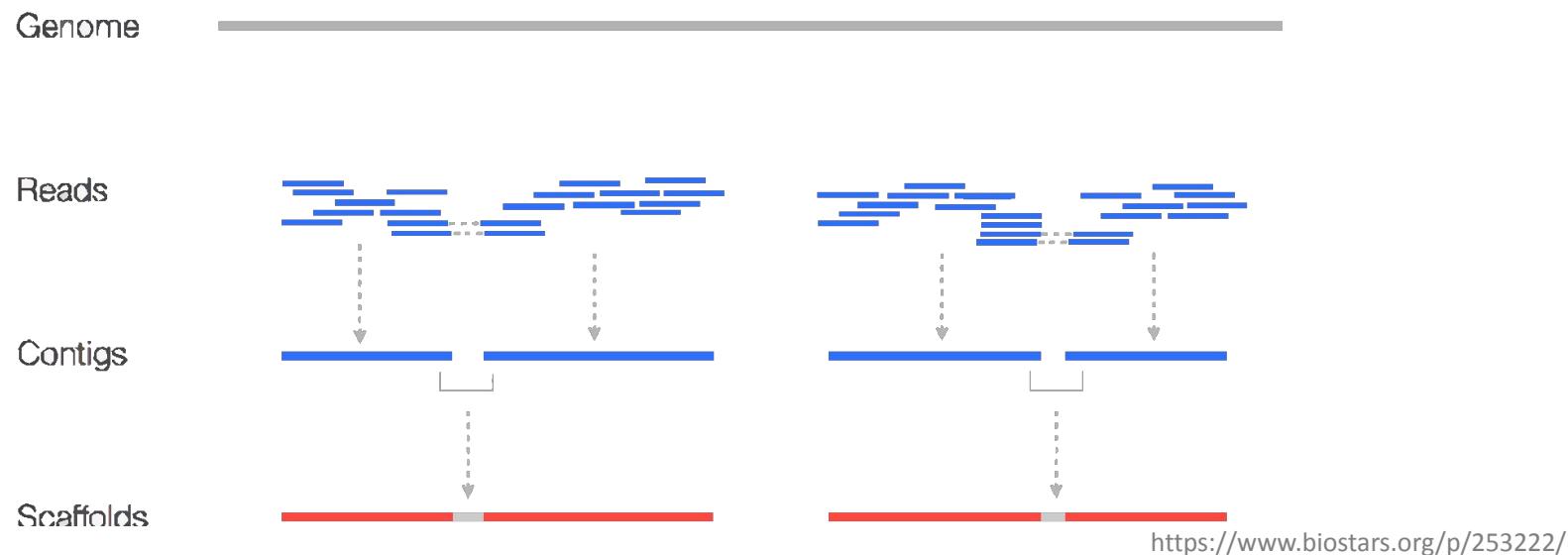
- ***De novo:*** with no previous knowledge of the genome to be assembled. It overlap the end of the end of each read in order to create a longer sequence.
- ***Assembly with reference:*** A similar but not identical genome guides the assembly process. Map reads over supplied genome.

# Assembly

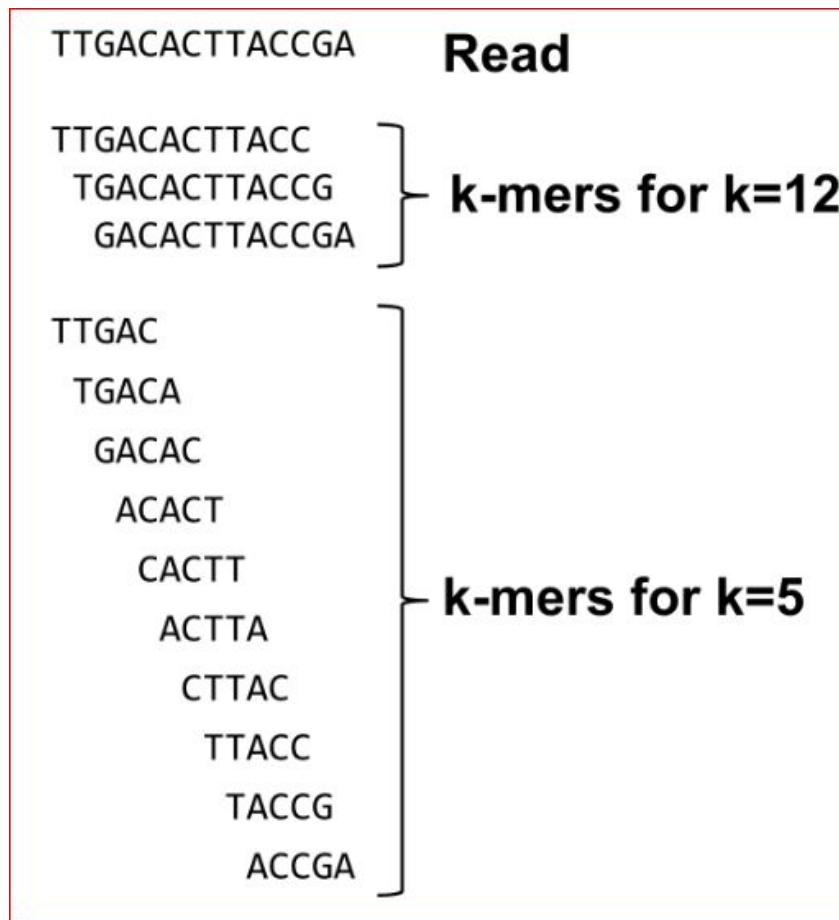


# Assembly: contig y scaffold

- **Contig:** continuous sequence made up of overlapping shorter sequences
- **Scaffold:** two or more contigs located and rearranged according to spatial information(pair-end, mate pair, reference)



# Algorithms: DBG



## Algorithms: DBG

Example #1:

HAPPI    PINE    INESS    APPIN

All 4-mers:

HAPP    PINE    INES    **APPI**  
**APPI**                    NESS    PPIN

*Unique* 4-mers:

HAPP **APPI** PINE PPIN INES NESS

## Algorithms: DBG

Example #1:

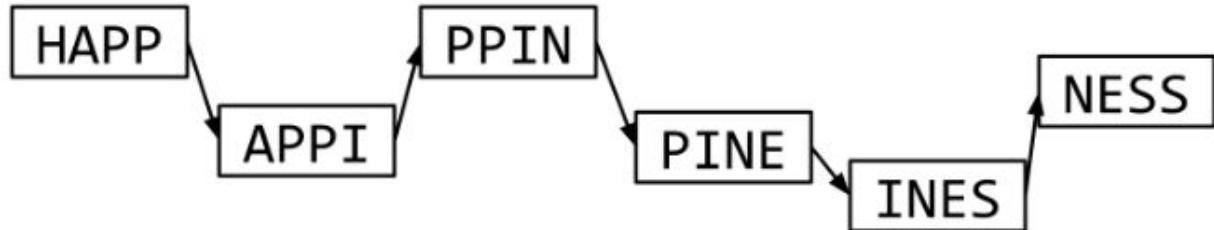
HAPPI   PINE   INESS   APPIN

k = 4 k-mers:

HAPP APPI

PINE PPIN

INES NESS



# Algorithms: DBG

Example #1:

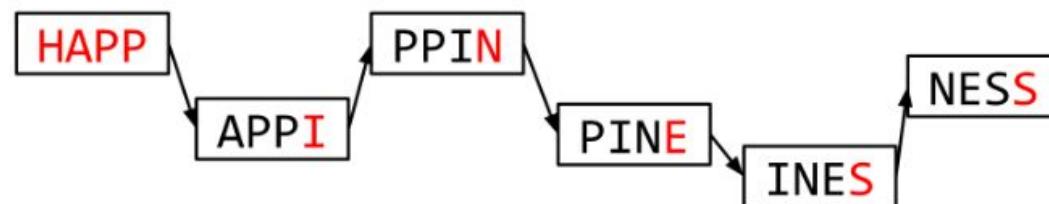
HAPPI    PINE    INESS    APPIN

k = 4 k-mers:

HAPP    APPI

PINE    PPIN

INES    NESS

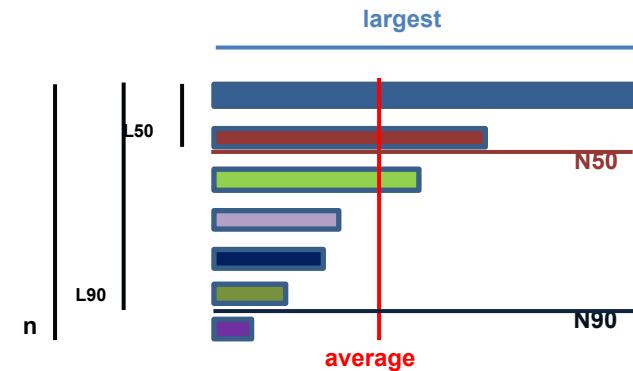
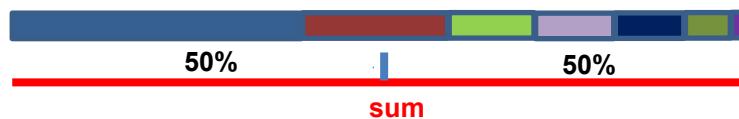


HAPPINESS

Easy!

# Assembly polishing and quality control

- sum = total bases number
- n = contigs number
- average = average contig length
- largest = largest contig
- N50 = length of the shortest contig where 50% of sum is held
- L50 = number of contigs which have 50% of the genome
- N90 = length of the shortest contig where 90% of sum is held.
- L90 = number of contigs which have 90% of the genome



Thanks for your attention!

> BU-ISCIII



(Find us in <https://github.com/BU-ISCIII>)