Universidad de los Andes

Pontificia Universidad Católica del Ecuador

CALIDAD DE LECTURAS Y LIMPIEZA

Laura Natalia González, MSc

Romain Guyot, PhD

SGR
Sistema General de Regalías

FORMATOS DE DATOS

FASTA

FASTQ

FAST5 (binarios)

FASTA

```
>gi|5524211|gb|AAD44166.1| cytochrome b
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

https://en.wikipedia.org/wiki/FASTQ_format

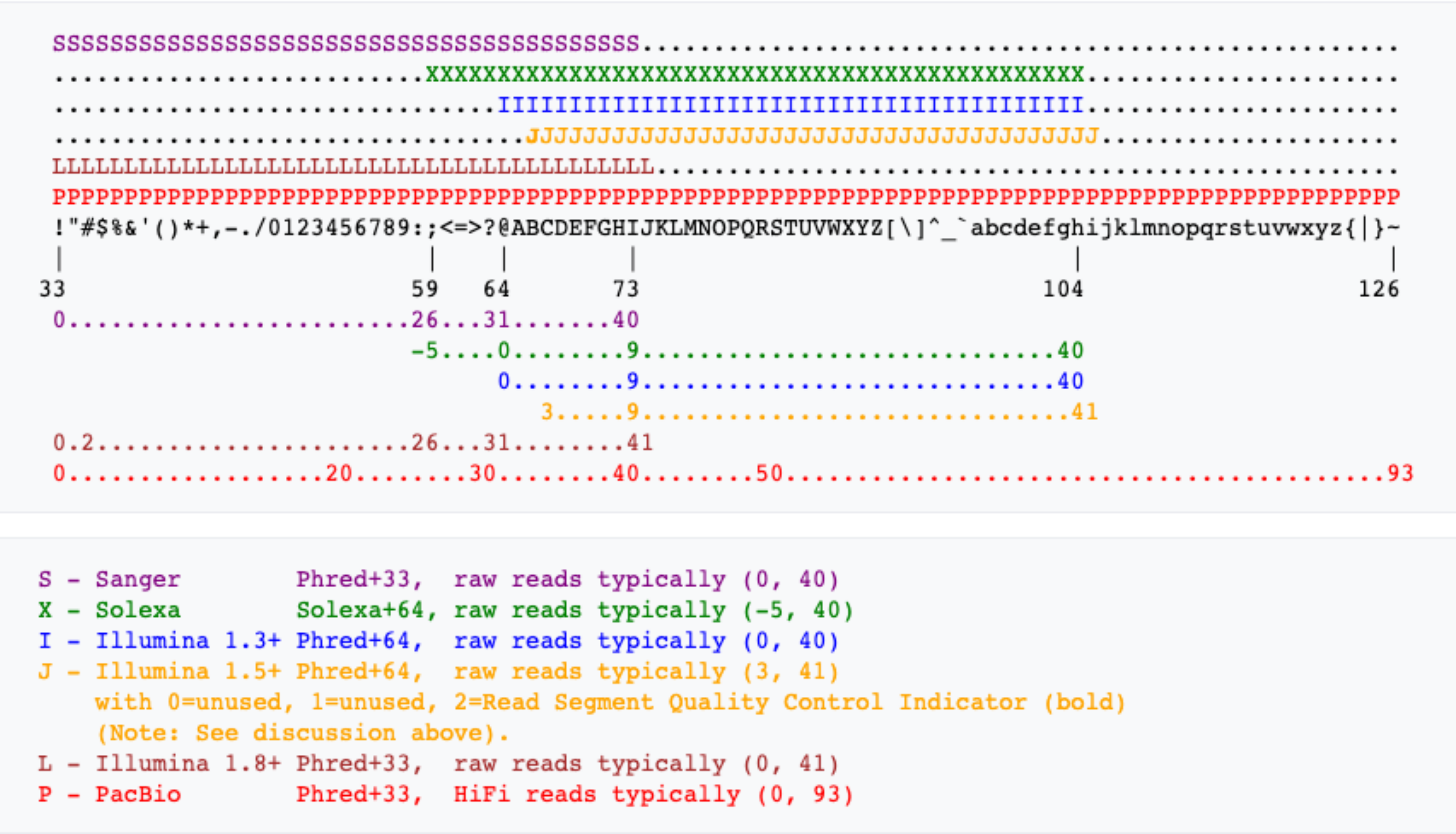
FASTQ

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.1 071113_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATACGGACAAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIH9IG9IB
```

- Línea 1 (comienza con @)
- Línea 2
- Línea 3 (comienza con +)
- Línea 4 —ASCII char

https://en.wikipedia.org/wiki/FASTQ_format

FASTQ



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                  126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9..............................41
  0.2......................26...31........41
  0..................20........30........40........50.........................................93
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

https://en.wikipedia.org/wiki/FASTQ_format

ANTES DE EMPEZAR

- ¿Cuántas lecturas tenemos en cada archivo?
- ¿Qué diferencia hay entre los formatos de salida de cada secuenciador?

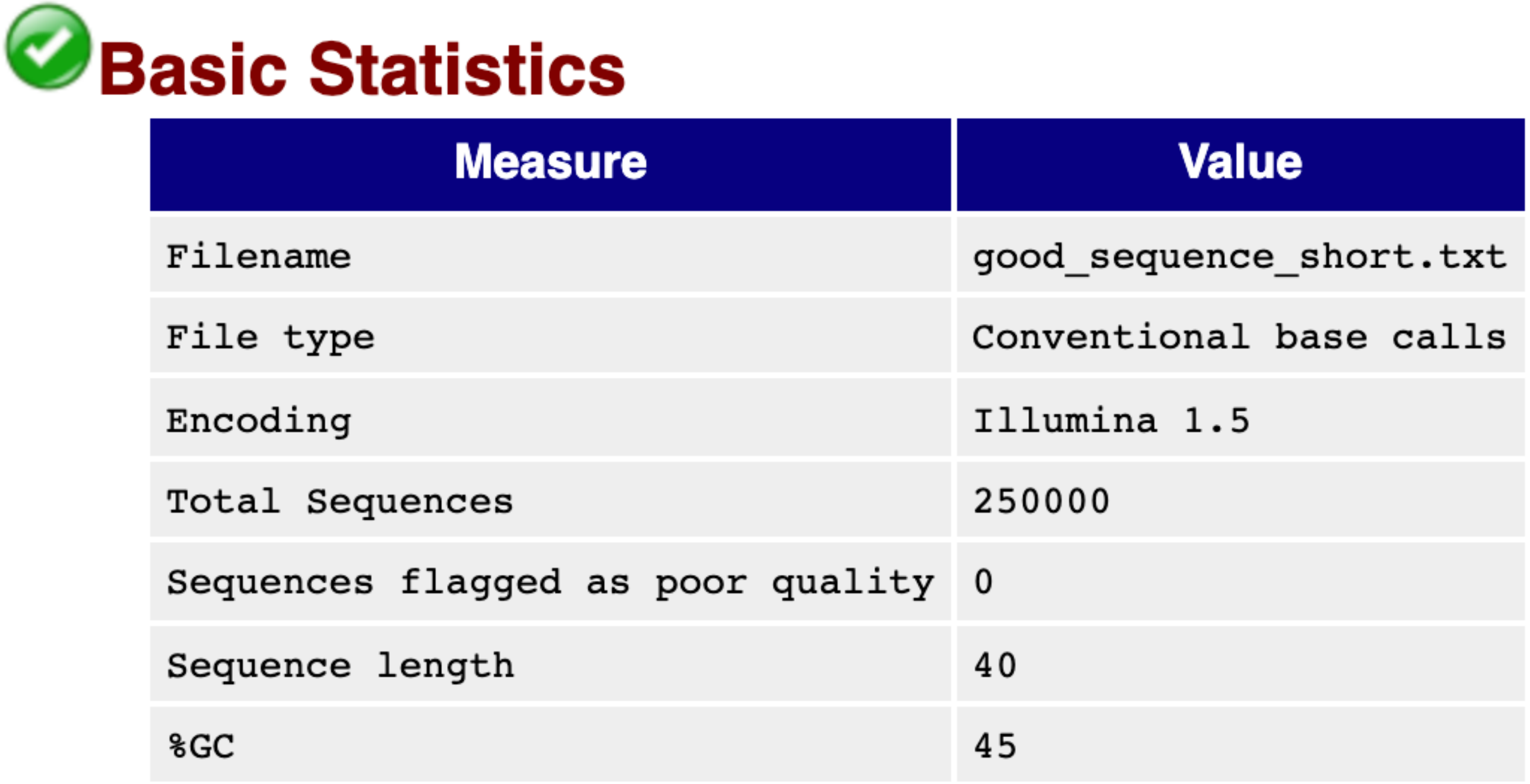
CALIDAD

FASTQC

Babraham Institute

Babraham Bioinformatics

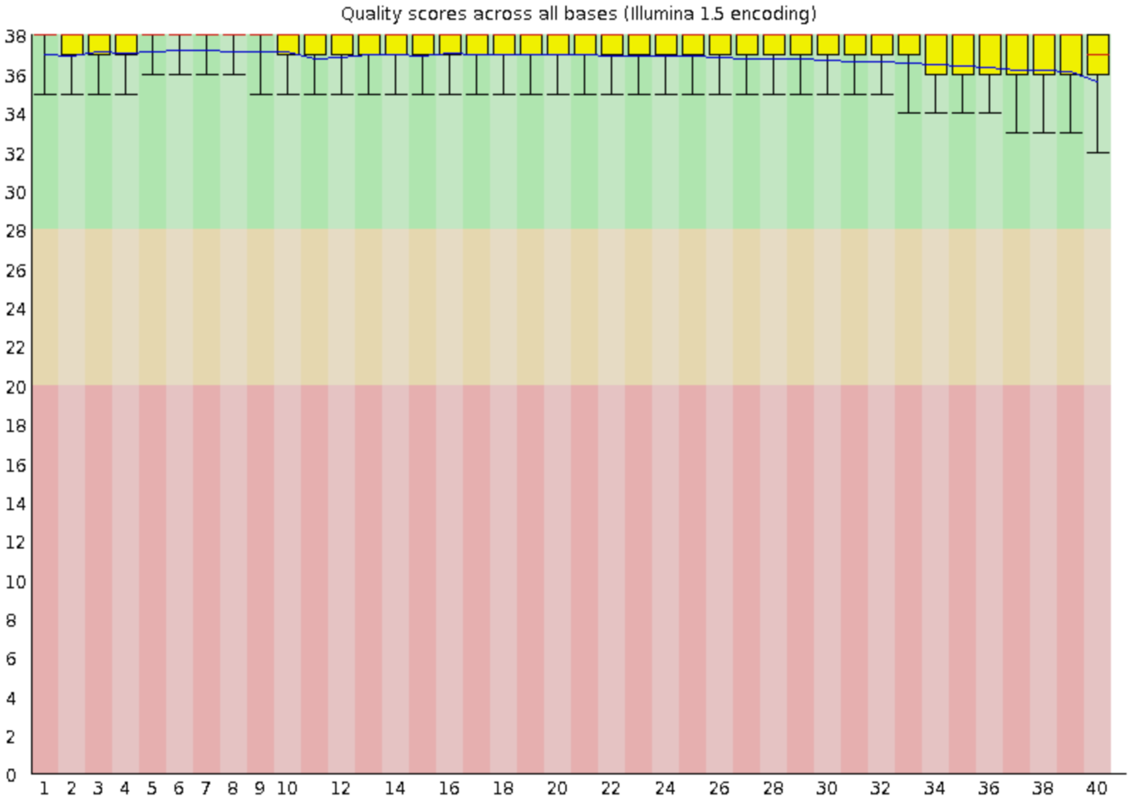
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



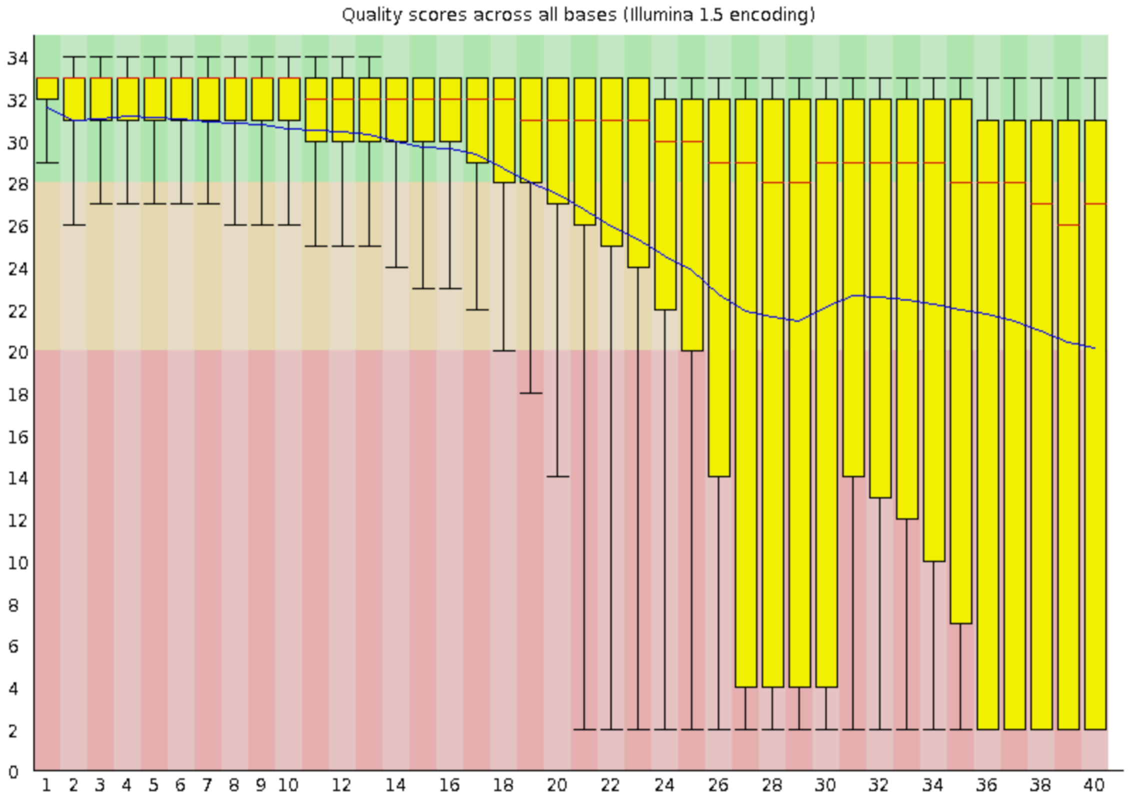
✓ Basic Statistics

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

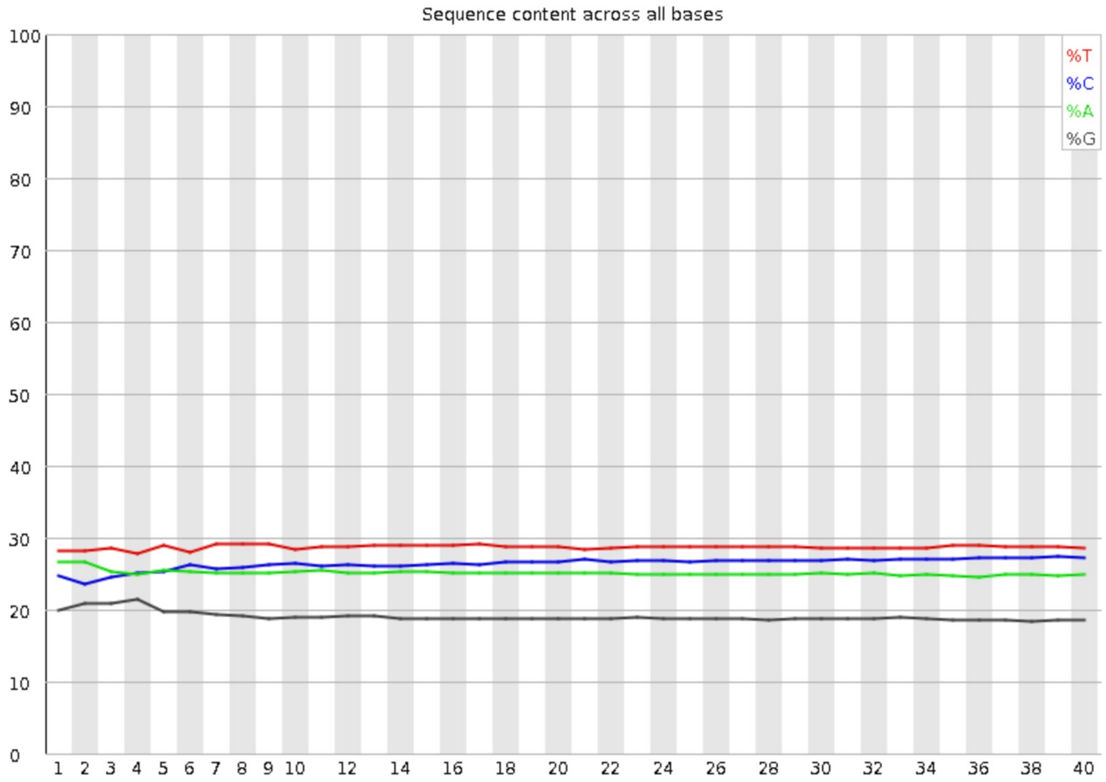
✓ Per base sequence quality



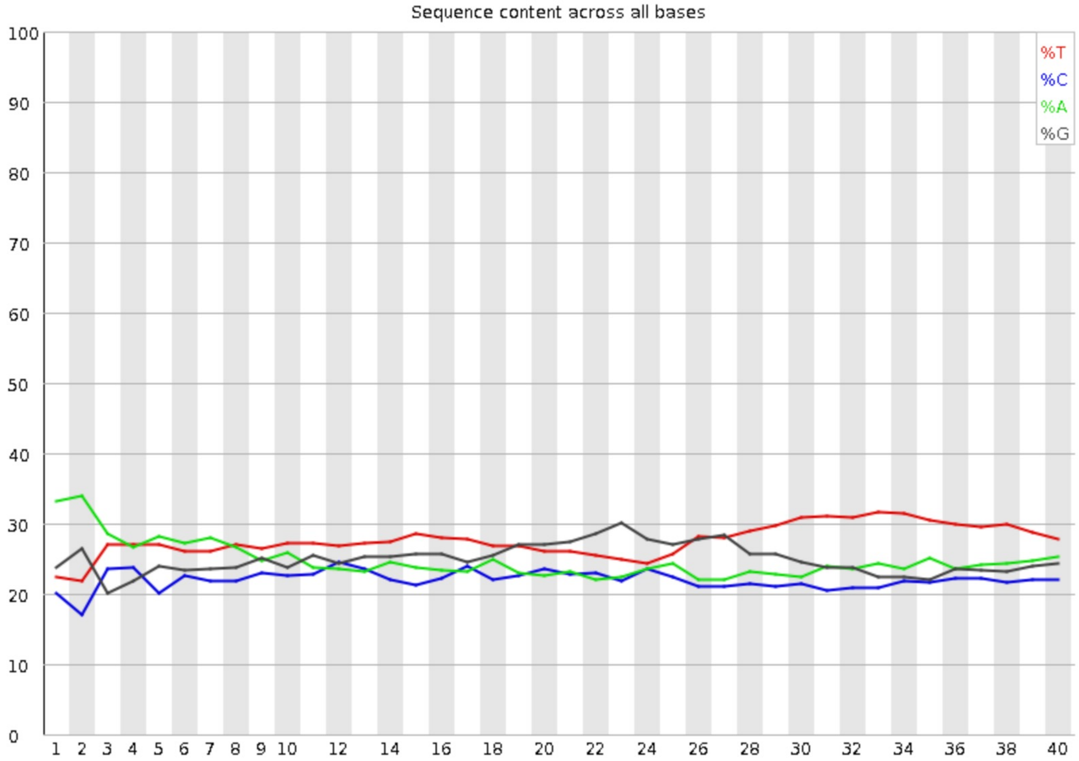
✗ Per base sequence quality



✓ Per base sequence content



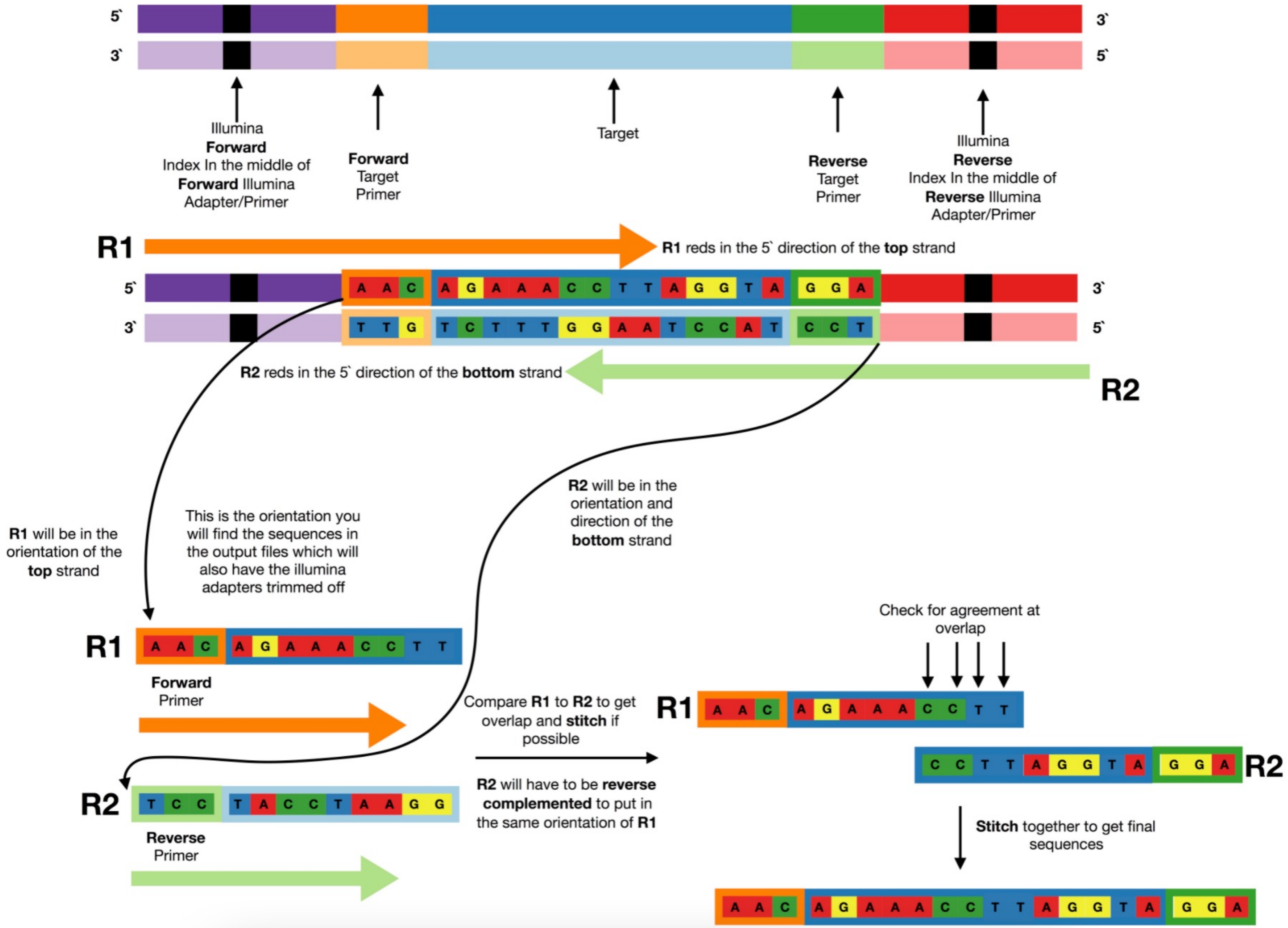
! Per base sequence content



LIMPIEZA

ILLUMINA – ION TORRENT

NANOPORE


5`
3`
3`
5`
A A C A G A A A C C T T A G G T A G G A
T T G T C T T T G G A A T C C A T C C T
R1
A A C A G A A A C C T T
R2
T C C T A C C T A A G G
R1
A A C A G A A A C C T T
C C T T A G G T A G G A
R2
A A C A G A A A C C T T A G G T A G G A

Illumina **Forward** Index In the middle of **Forward** Illumina Adapter/Primer

Forward Target Primer

Target

Reverse Target Primer

Illumina **Reverse** Index In the middle of **Reverse** Illumina Adapter/Primer

R1 reds in the 5` direction of the **top** strand

R2 reds in the 5` direction of the **bottom** strand

R1 will be in the orientation of the **top** strand

This is the orientation you will find the sequences in the output files which will also have the illumina adapters trimmed off

R2 will be in the orientation and direction of the **bottom** strand

Forward Primer

Reverse Primer

Compare **R1** to **R2** to get overlap and stitch if possible

R2 will have to be **reverse complemented** to put in the same orientation of **R1**

Check for agreement at overlap

Stitch together to get final sequences

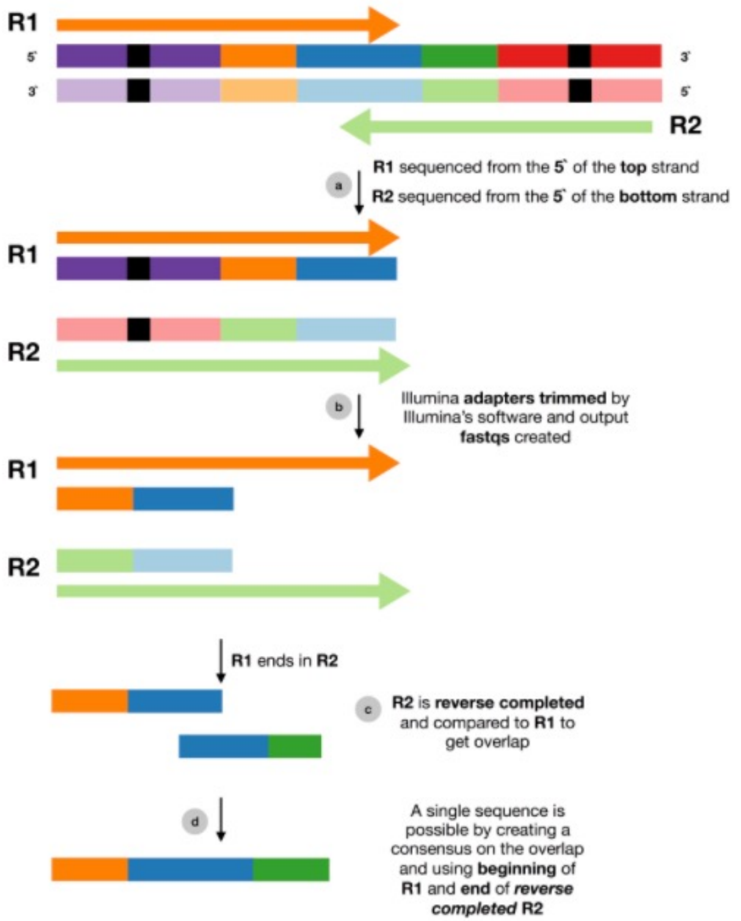
LECTURAS ILLUMINA O ION TORRENT

https://seekdeep.brown.edu/illumina_paired_info.html

1

R1EndsInR2

This situation happens when **target** sequence **plus primers** are **more than** read length but **shorter than** 2 times the read (e.g. **target** size of **200** on **2x150** sequencing run)



a: **R1** sequenced from the **5`** of the **top** strand
R2 sequenced from the **5`** of the **bottom** strand

b: Illumina **adapters trimmed** by Illumina's software and output **fastqs** created

R1 ends in **R2**

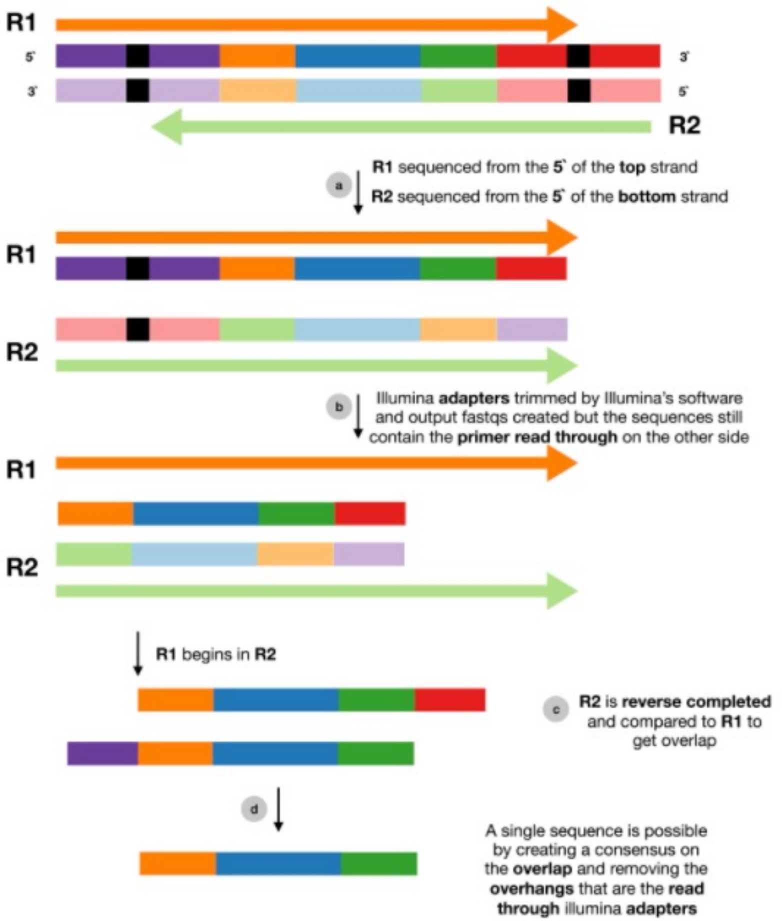
c: **R2** is **reverse completed** and compared to **R1** to get overlap

d: A single sequence is possible by creating a consensus on the overlap and using **beginning** of **R1** and end of ***reverse completed*** **R2**

2

R1BeginsInR2

This situation happens when **target** sequence **plus primers** are are **less than** read length (e.g. **target** size of **100** on **2x150** sequencing run)



a: **R1** sequenced from the **5`** of the **top** strand
R2 sequenced from the **5`** of the **bottom** strand

b: Illumina **adapters** trimmed by Illumina's software and output fastqs created but the sequences still contain the **primer read through** on the other side

R1 begins in **R2**

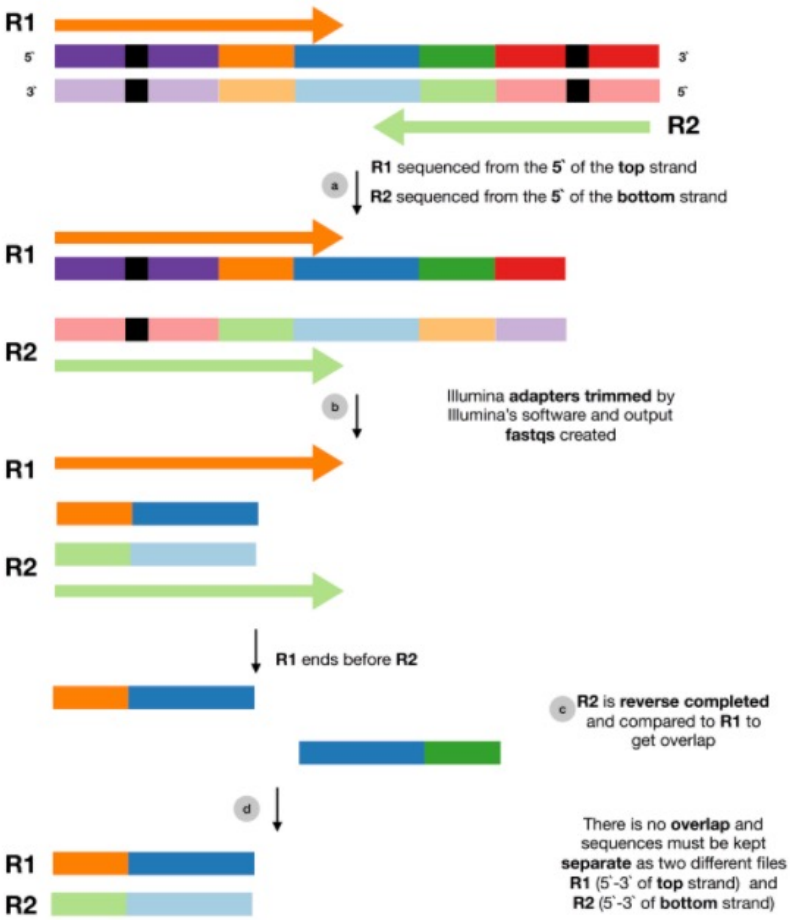
c: **R2** is **reverse completed** and compared to **R1** to get overlap

d: A single sequence is possible by creating a consensus on the **overlap** and removing the **overhangs** that are the read **through** illumina **adapters**

3

NoOverlap

This situation happens when **target** sequence **plus primers** are are **more than** 2 times read length (e.g. **target** size of **400** on **2x150** sequencing run)



a: **R1** sequenced from the **5`** of the **top** strand
R2 sequenced from the **5`** of the **bottom** strand

b: Illumina **adapters trimmed** by Illumina's software and output **fastqs** created

R1 ends before **R2**

c: **R2** is **reverse completed** and compared to **R1** to get overlap

d: There is no **overlap** and sequences must be kept **separate** as two different files **R1** (5`-3` of **top** strand) and **R2** (5`-3` of **bottom** strand)

https://seekdeep.brown.edu/illumina_paired_info.html

¿QUÉ LIMPIAMOS?

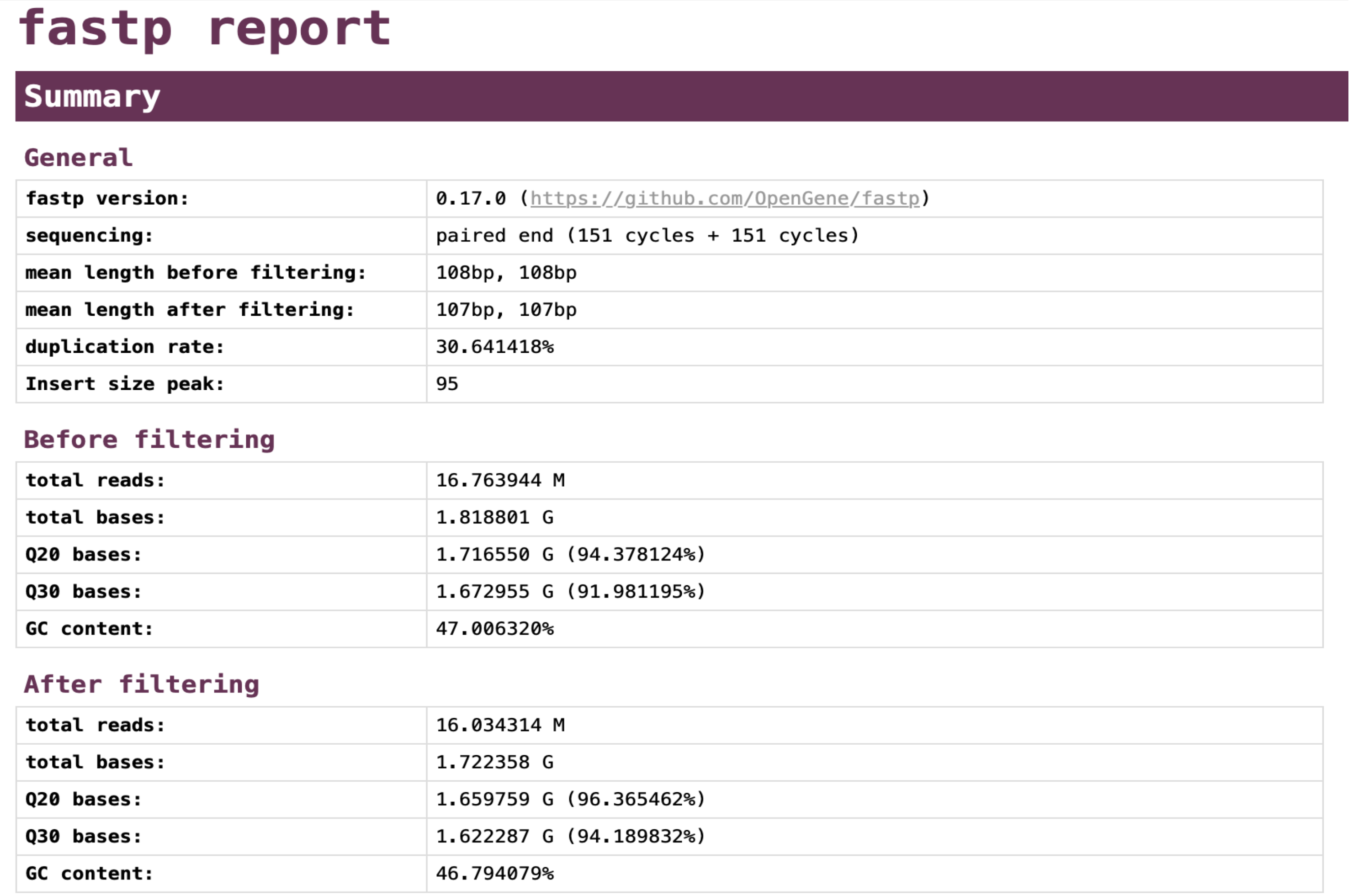
- Secuencias de adaptadores de PCR o secuenciación
- Regiones de baja calidad
- Lecturas de baja calidad (total o ventanas)
- Lecturas muy cortas

TRIMMOMATIC

- ILLUMINACLIP
- SLIDINGWINDOW
- LEADING
- TRAILING
- CROP
- HEADCROP
- MINLEN

FASTP

- https://github.com/OpenGene/fastp



fastp report

Summary

General

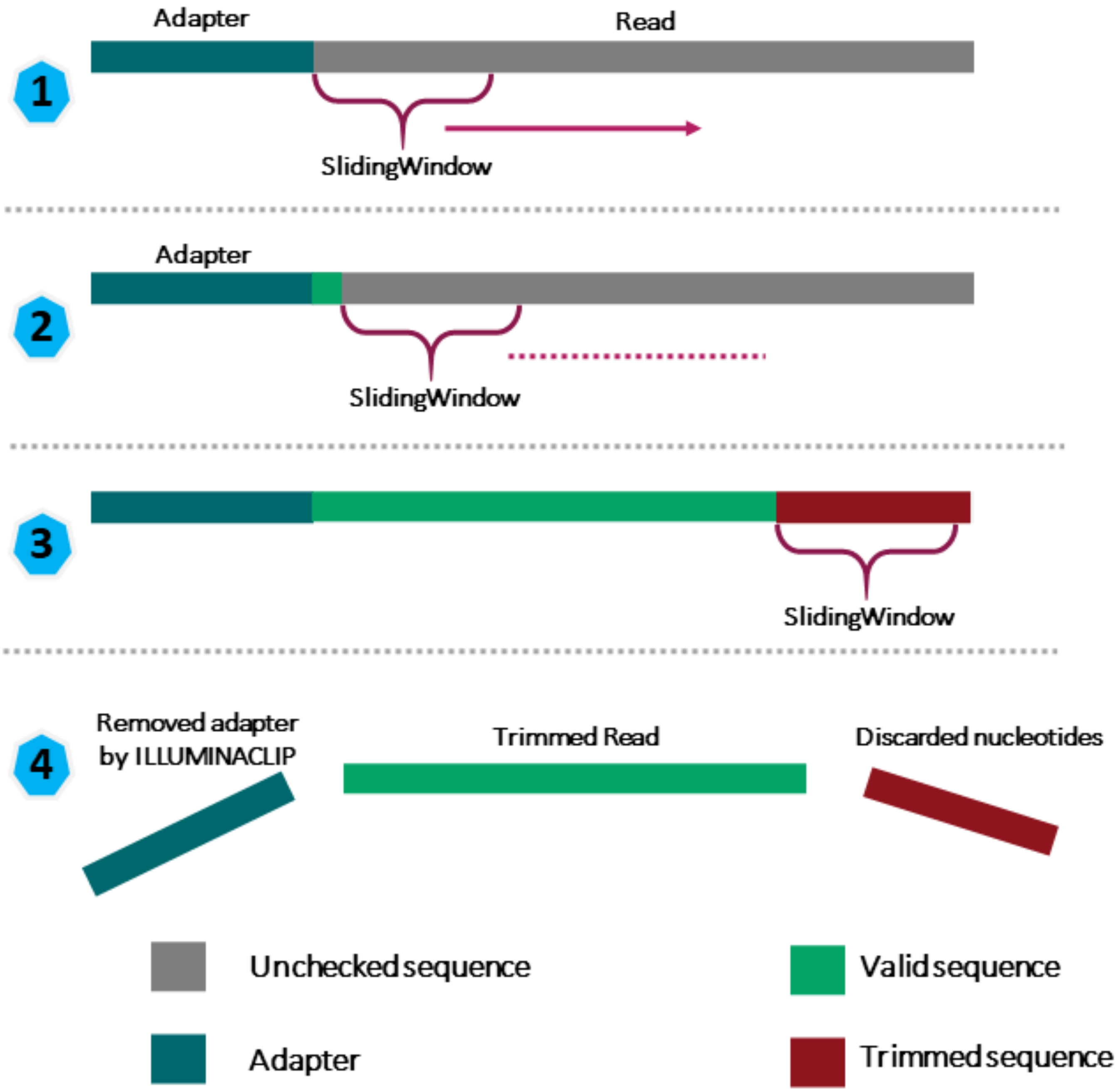
fastp version:	0.17.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (151 cycles + 151 cycles)
mean length before filtering:	108bp, 108bp
mean length after filtering:	107bp, 107bp
duplication rate:	30.641418%
Insert size peak:	95

Before filtering

total reads:	16.763944 M
total bases:	1.818801 G
Q20 bases:	1.716550 G (94.378124%)
Q30 bases:	1.672955 G (91.981195%)
GC content:	47.006320%

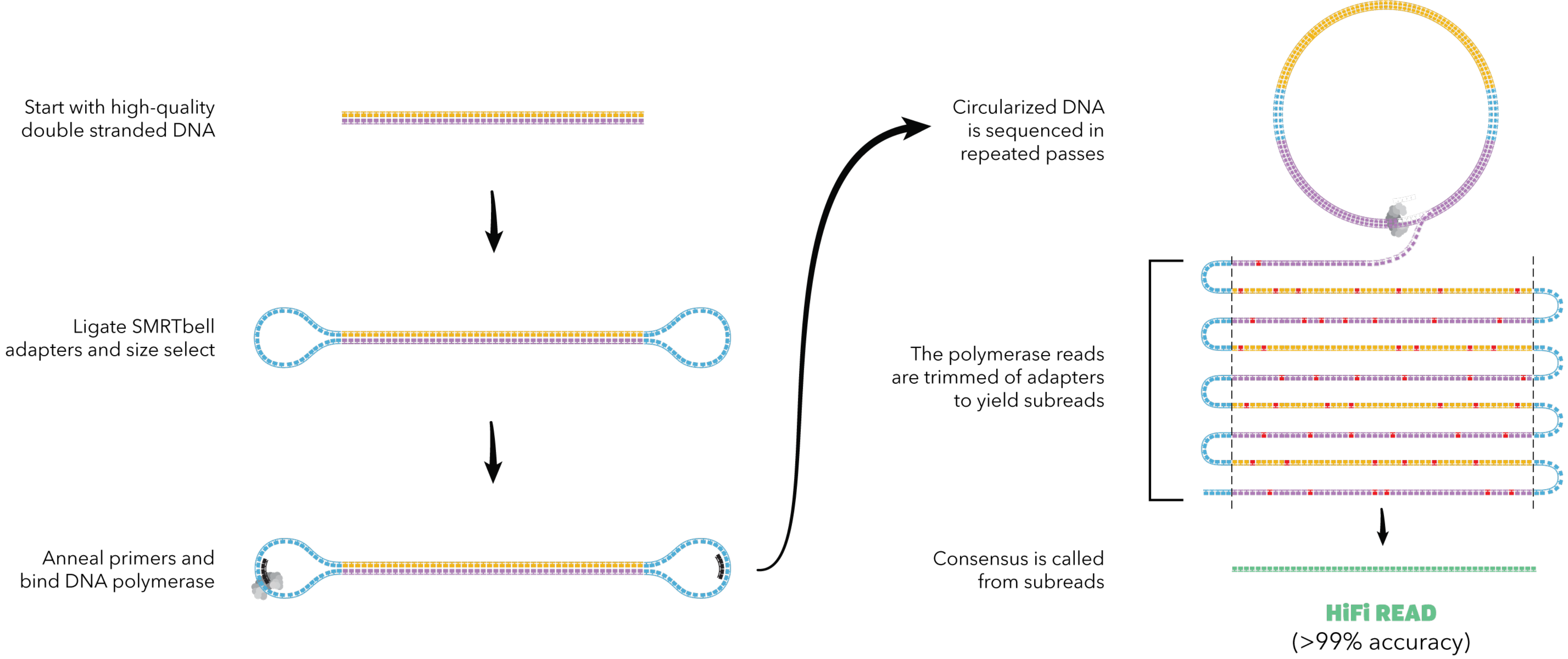
After filtering

total reads:	16.034314 M
total bases:	1.722358 G
Q20 bases:	1.659759 G (96.365462%)
Q30 bases:	1.622287 G (94.189832%)
GC content:	46.794079%



https://carpentries-incubator.github.io/metagenomics/03-trimming-filtering/index.html

HiFi READS



Start with high-quality double stranded DNA

Ligate SMRTbell adapters and size select

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

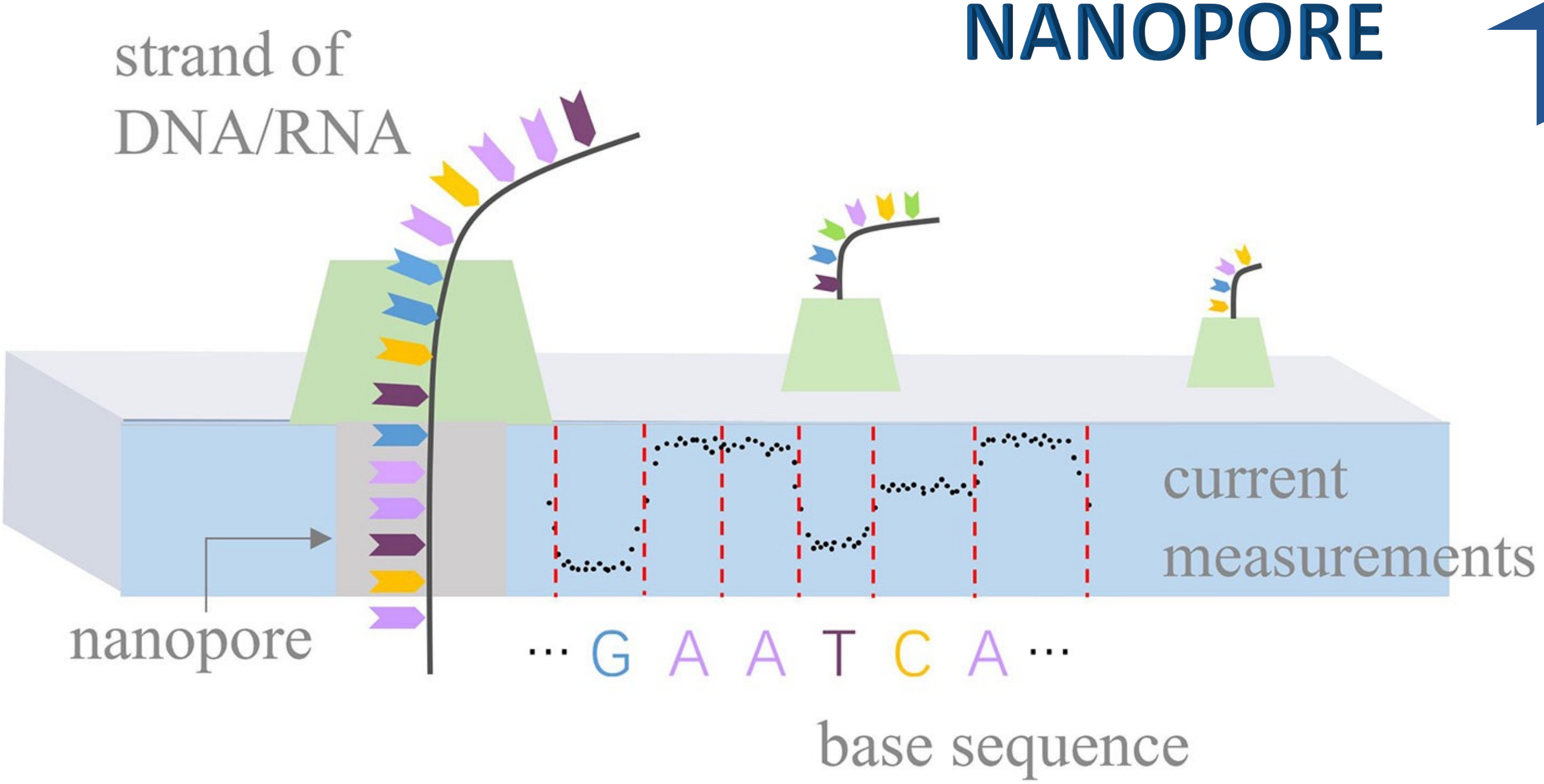
The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads

HiFi READ

(>99% accuracy)

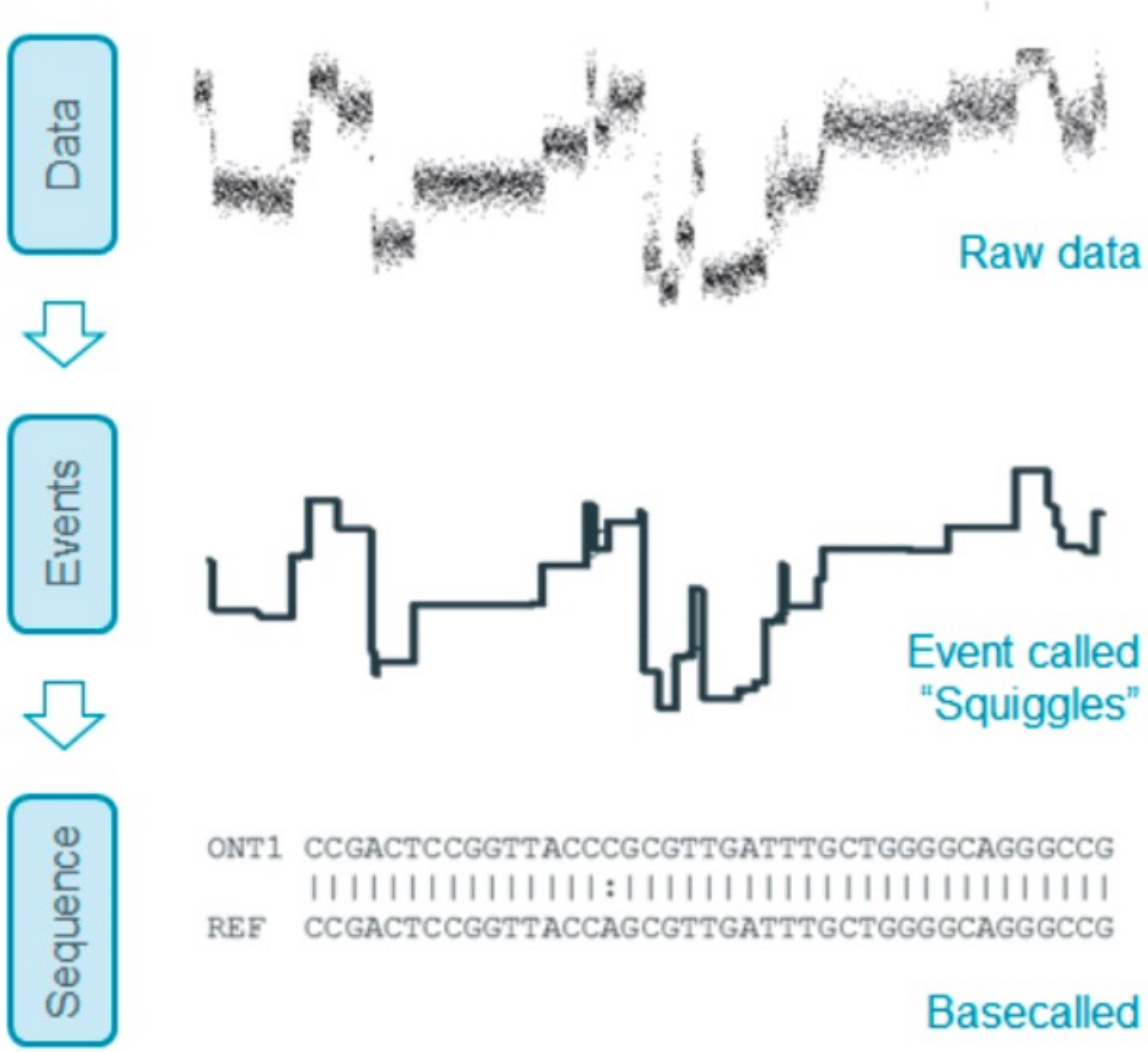
NANOPORE



https://www.frontiersin.org/articles/10.3389/fgene.2019.01332/full

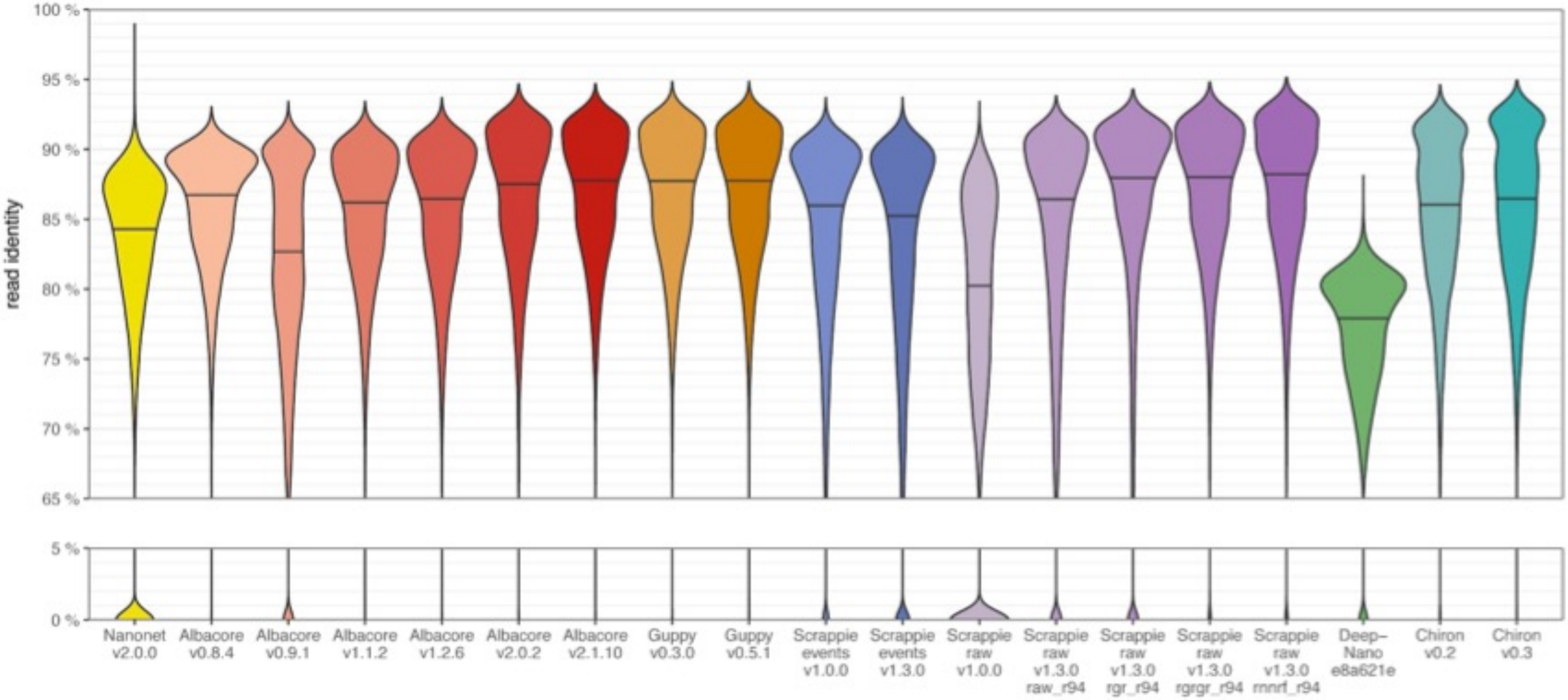
LLAMADO DE BASES

FAST5 (HDF5) a FASTQ



```
ONT1 CCGACTCCGGTTACCCGCGTTGATTTGCTGGGGCAGGGCCG
     |||||||||||||||:|||||||||||||||||||||||||
REF  CCGACTCCGGTTACCAGCGTTGATTTGCTGGGGCAGGGCCG
```

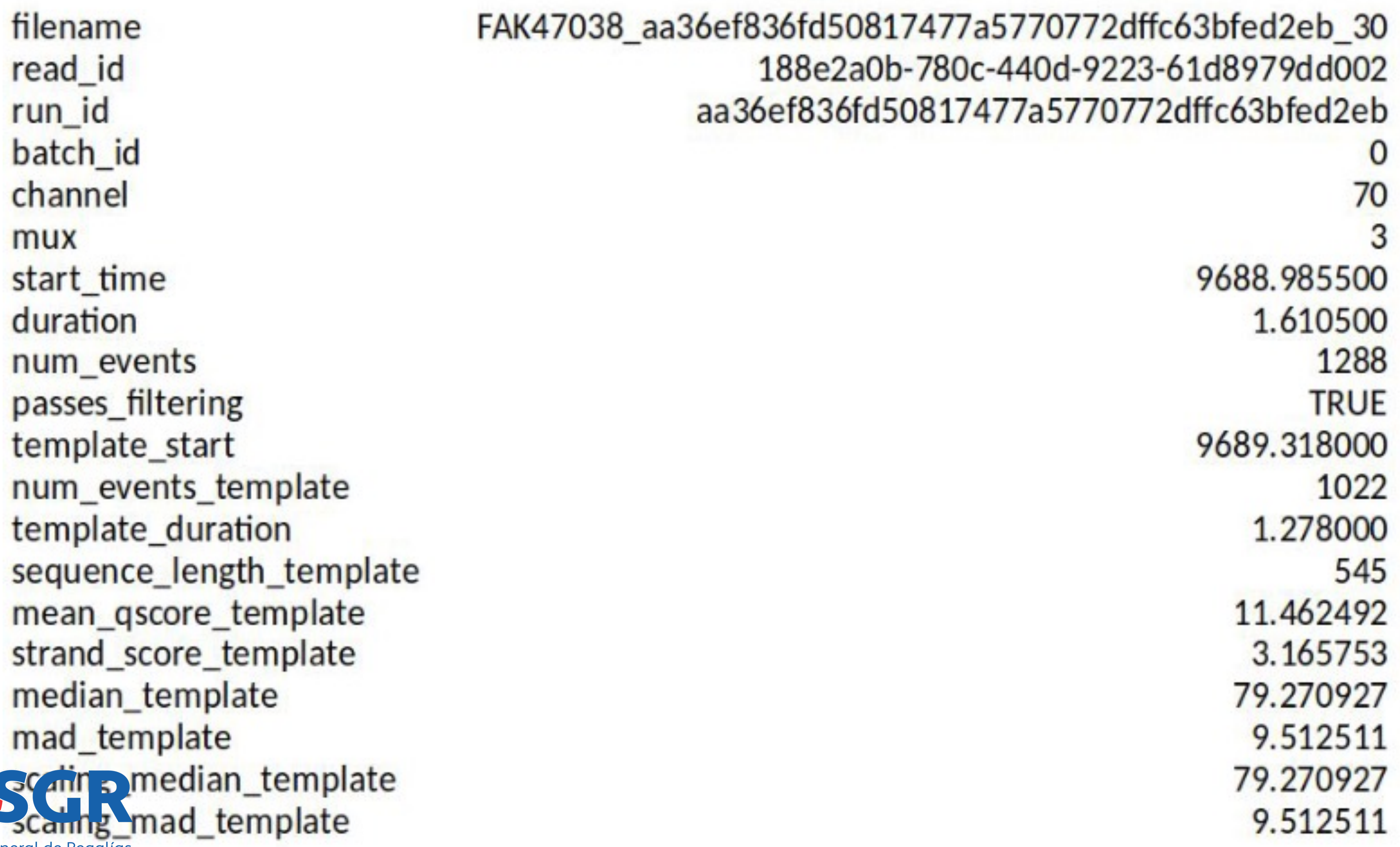
ONT Read calling



Sistema General de Regalías

summary_file.txt

Universidad de los Andes



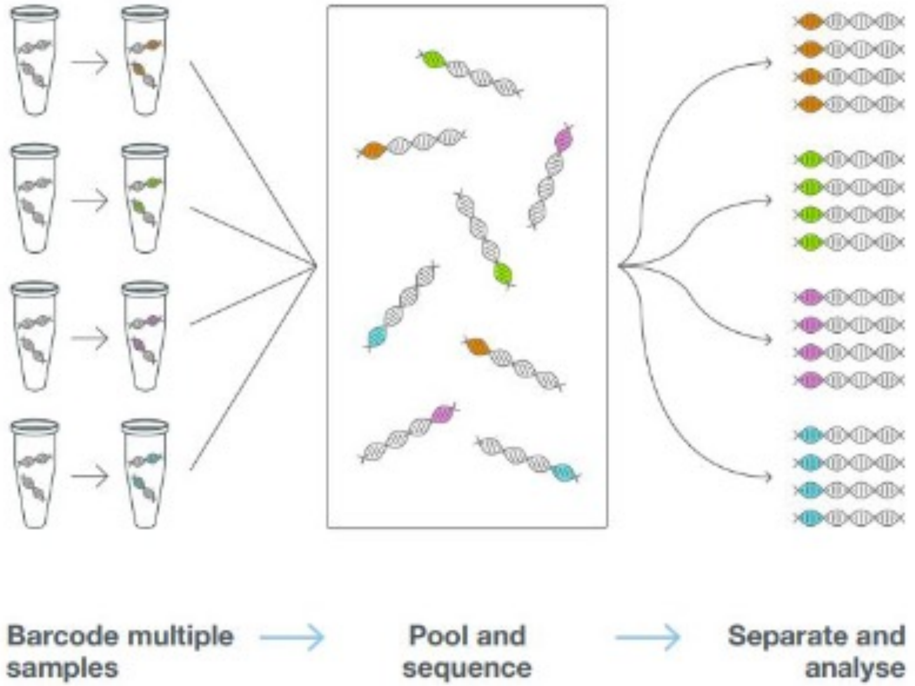
filename	FAK47038_aa36ef836fd50817477a5770772dffc63bfed2eb_30
read_id	188e2a0b-780c-440d-9223-61d8979dd002
run_id	aa36ef836fd50817477a5770772dffc63bfed2eb
batch_id	0
channel	70
mux	3
start_time	9688.985500
duration	1.610500
num_events	1288
passes_filtering	TRUE
template_start	9689.318000
num_events_template	1022
template_duration	1.278000
sequence_length_template	545
mean_qscore_template	11.462492
strand_score_template	3.165753
median_template	79.270927
mad_template	9.512511
scaling_median_template	79.270927
scaling_mad_template	9.512511

SGR
Sistema General de Regalías

ONT demultiplexing

Deepbinner: Demultiplexing barcoded ONT reads with deep convolutional neural networks (CNN). The network is trained to classify barcodes based on the raw nanopore signal.

Guppy
In contrast to Deepbinner, guppy barcoding requires basecalling of all reads and detects barcodes in the sequence



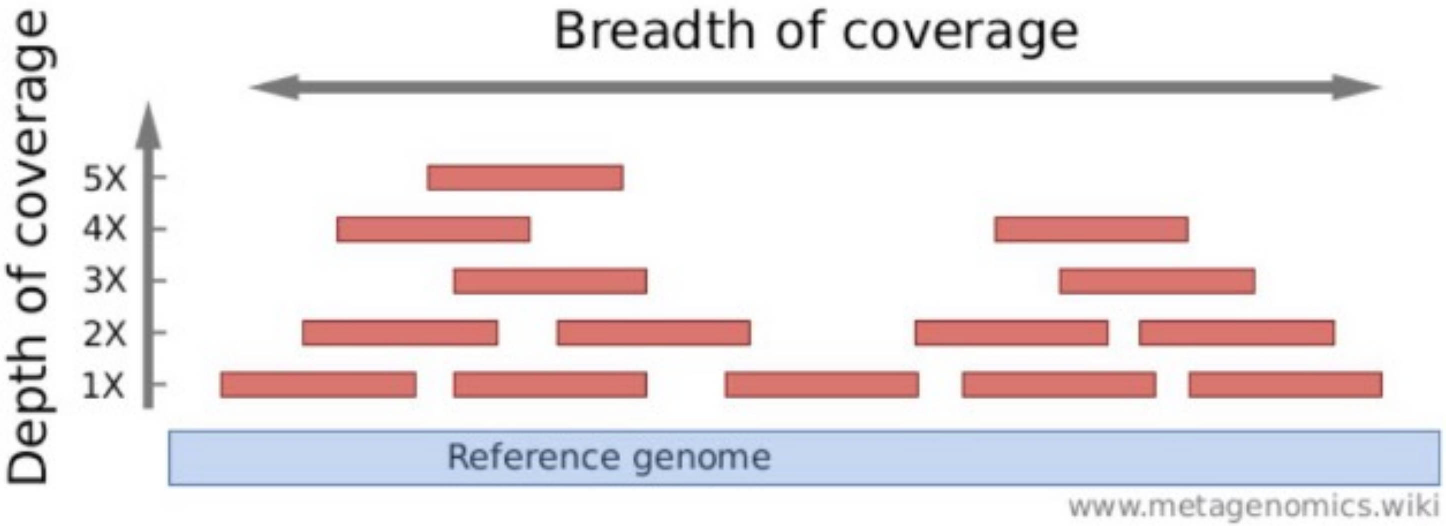
ONT Read calling, cleaning and filtering

Sequencer ONT : raw fast5 files

- Transform fast5 signal in fastq standard format *Guppy, Bonito*
- Optional Demultiplexing and removing adapters *Guppy options*
- Optional Find and remove adapters from reads *Porechop*
- Optional Quality filtering using the *sequencing_summary.txt* information : *Guppy options, filtlong, nanofilt*

Guppy is a neural network based basecaller that in addition to basecalling also performs filtering of low quality reads, clipping of Oxford Nanopore adapters and estimation of methylation probabilities per base

Calculate depth of coverage

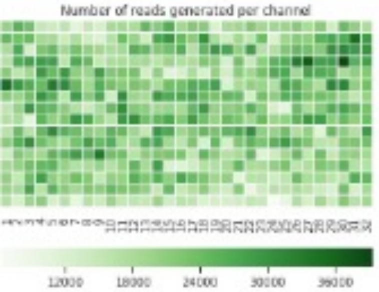


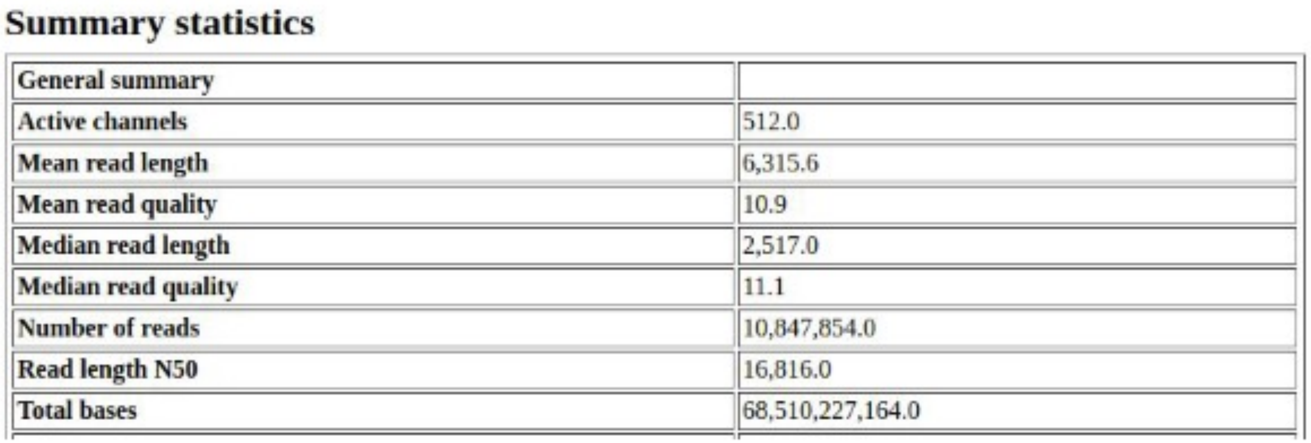
depth of coverage estimation :

- Count how much base pairs in all sequenced reads? *total_pb*
- What is the expected genome size? genome_size

depth_of_coverage = total_pb/genome_size

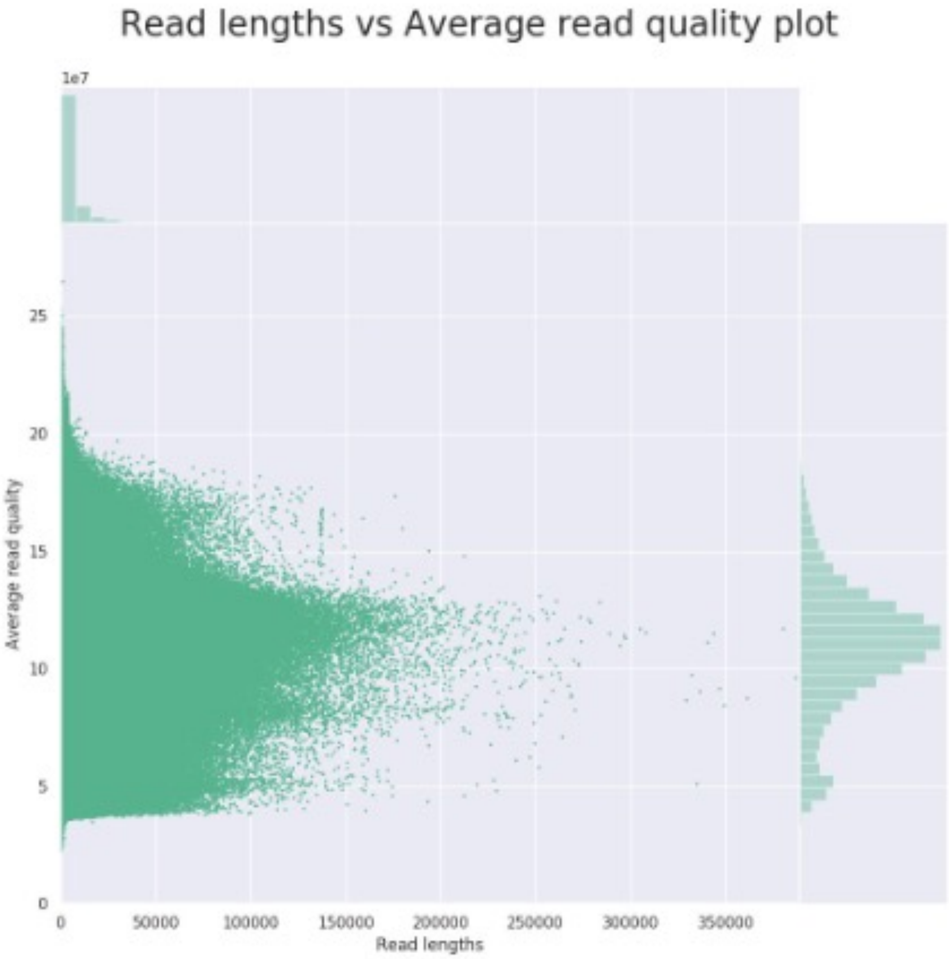
Reads Quality control : *NanoPlot*

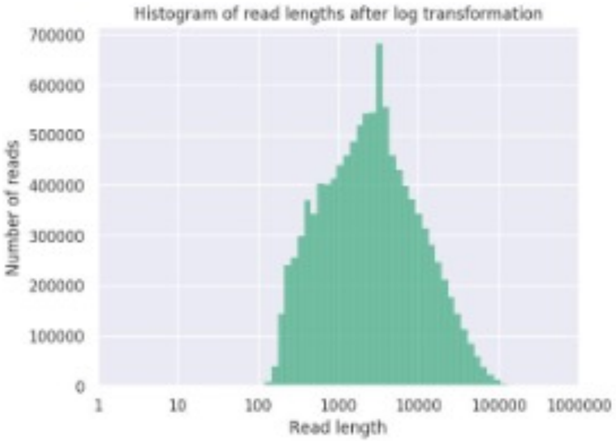

Number of reads generated per channel

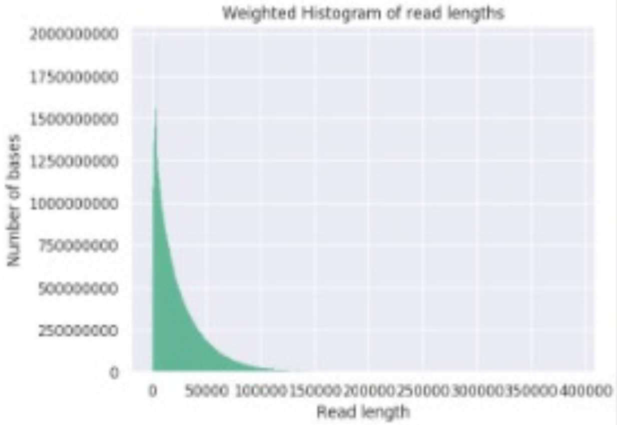


Summary statistics

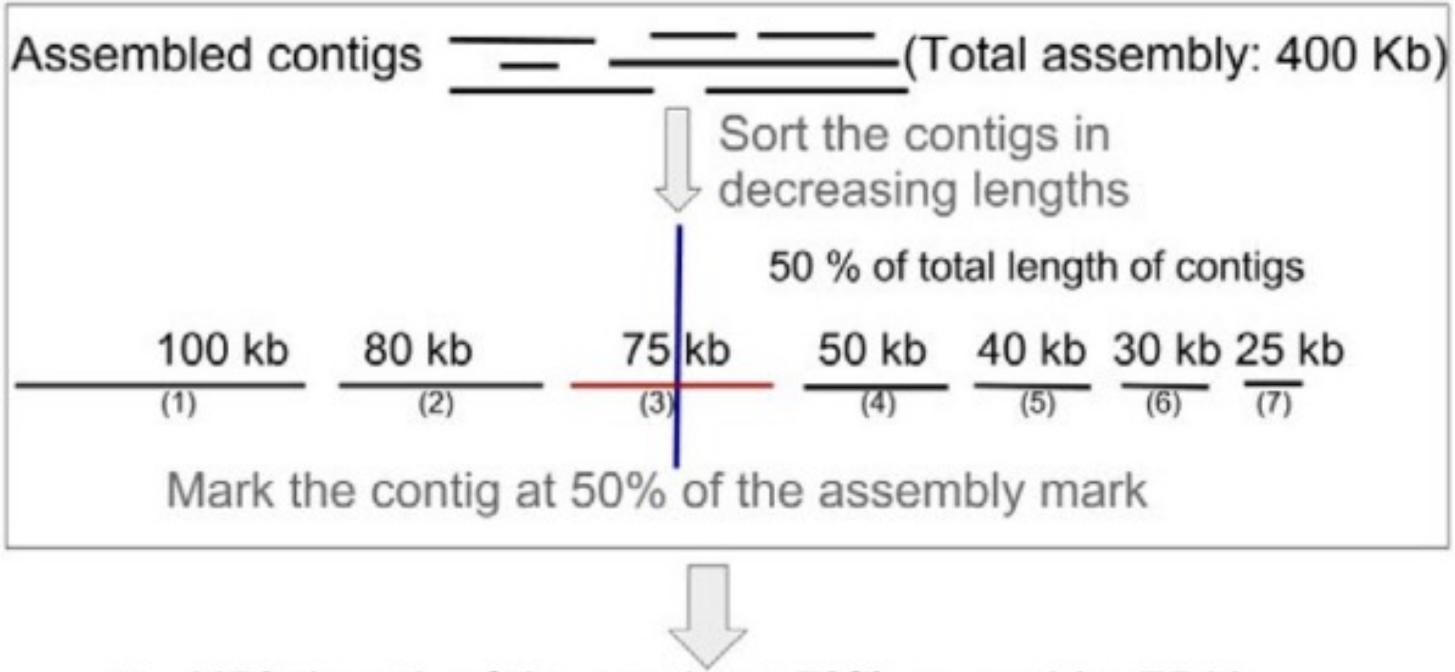
General summary	
Active channels	512.0
Mean read length	6,315.6
Mean read quality	10.9
Median read length	2,517.0
Median read quality	11.1
Number of reads	10,847,854.0
Read length N50	16,816.0
Total bases	68,510,227,164.0


Read lengths vs Average read quality plot


Histogram of read lengths after log transformation


Weighted Histogram of read lengths

What is N50 and L50?



Assembled contigs (Total assembly: 400 Kb)

Sort the contigs in decreasing lengths

50 % of total length of contigs

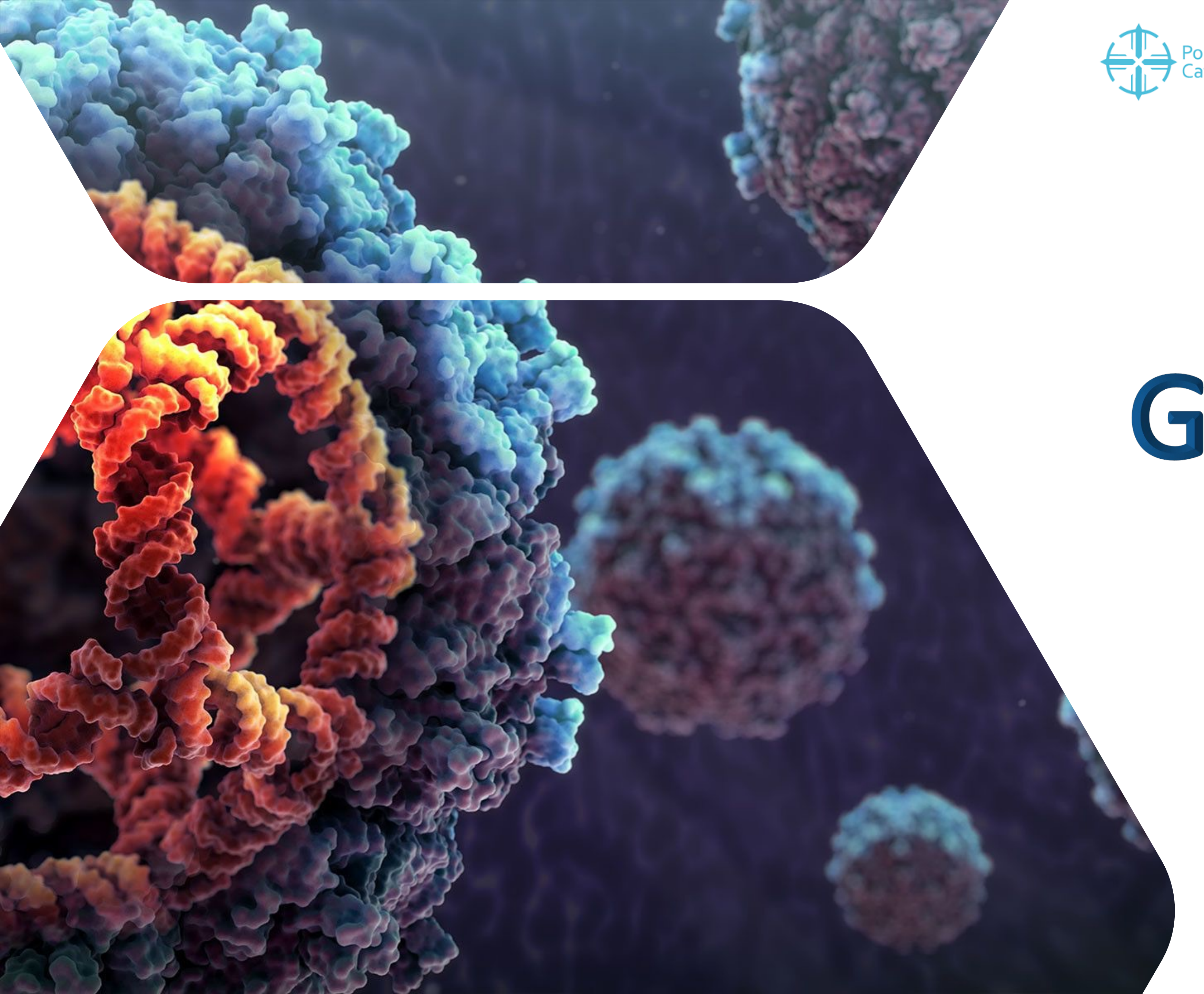
Mark the contig at 50% of the assembly mark

➔ N50, length of the contig at 50% assembly: 75 kb

➔ L50, number of contigs until 50% assembly: 3

Pontificia Universidad Católica del Ecuador

Universidad de los Andes



GRACIAS

SGR
Sistema General de Regalías