Instituto Nacional de Salud

Taller virtual:

Análisis bioinformático de SARS-CoV-2

Septiembre 1 al 3 2021

Análisis de Calidad

Universidad Nacional de Colombia

Andrés M. Pinzón Ph. D.

ampinzonv@unal.edu.co

Instituto de Genética

Universidad Nacional de Colombia

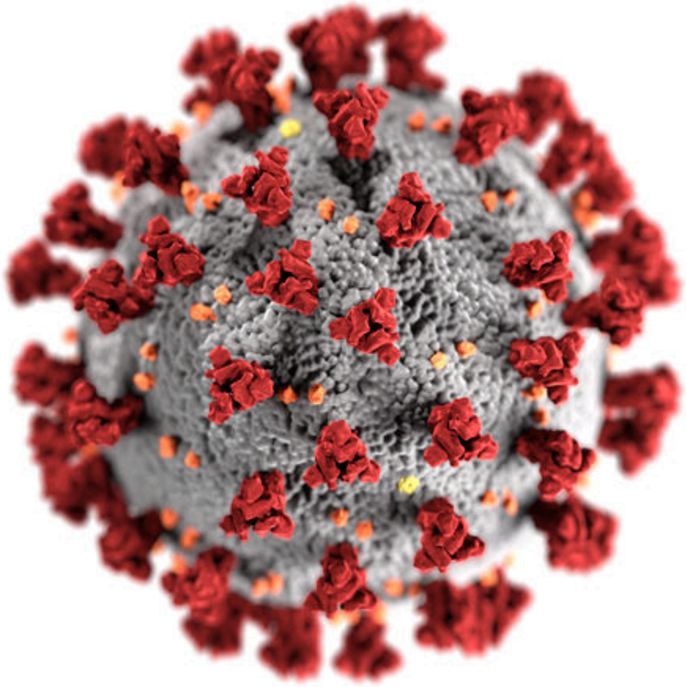
GiββS

BIOINFORMATICS AND SYSTEMS BIOLOGY GROUP

https://gibbslab.github.io/

@gibbsclab gibbslab

La Secuenciación Genómica nos permite obtener la secuencia completa del Genoma de Sars-CoV-2



>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

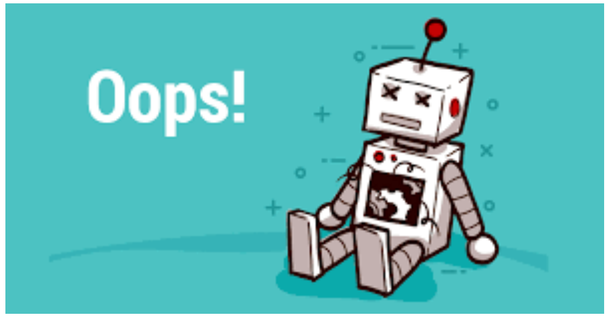
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
...

Pero… ¿Qué tan seguros estamos de que la secuencia obtenida es totalmente correcta?

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
…

Finalmente los secuenciadores también cometen errores!



ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

ATTAAAGGTTTATACGTTCCCAGGTAACAAACCAACCAACNNNCGATCTCTTGTAGATCTGTTCTCTAAA

https://blogs.unsw.edu.au/nowideas/blog/2019/03/innovacion-empresarial-error-no-es-fracaso/

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA

CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC

TAATTACTGTCG

TTGCAGCCGAT

CCTGGTTTCAA

GTGGCTTTGGA

CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGA

GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTA

GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAA

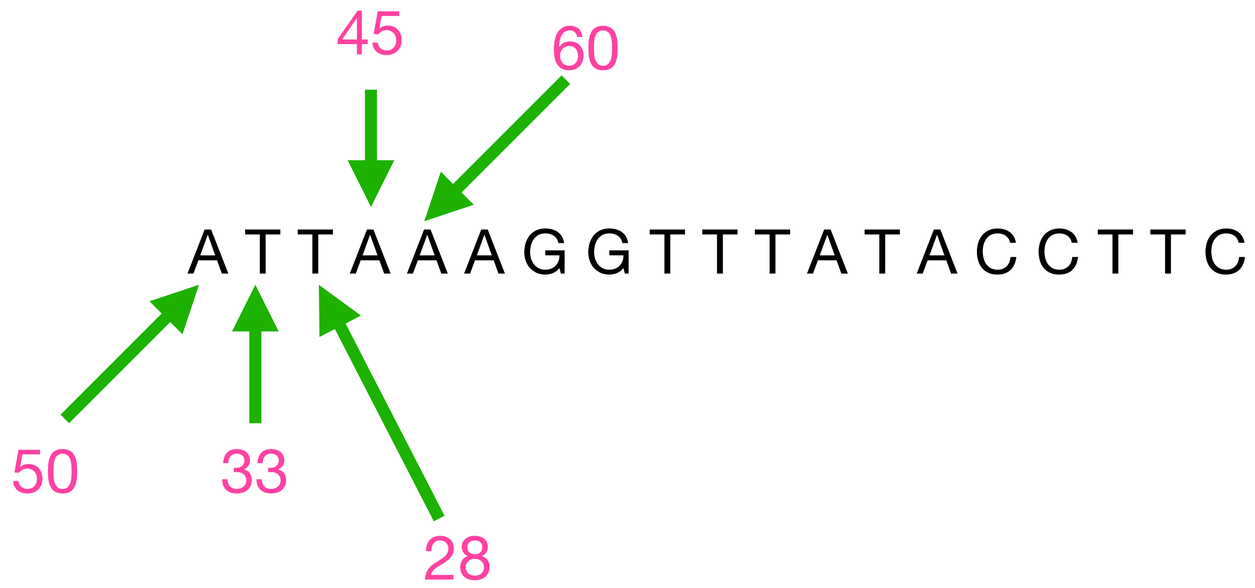
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA

GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG

...

El ANÁLISIS DE CALIDAD nos permite evaluar qué tan “errada” fue nuestra secuenciación.

Para eso se le asigna a cada base secuenciada **un valor numérico** que nos dice qué tan buena fue la detección de esa base!



Este valor numérico se conoce como:
PHRED QUALITY SCORE (Q-SCORE)

Phred quality score (Q)

Originally developed by the program **Phred** to help in the automation of DNA sequencing in the Human Genome Project. Phred quality scores are assigned to each nucleotide **base call** in automated sequencer traces.

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

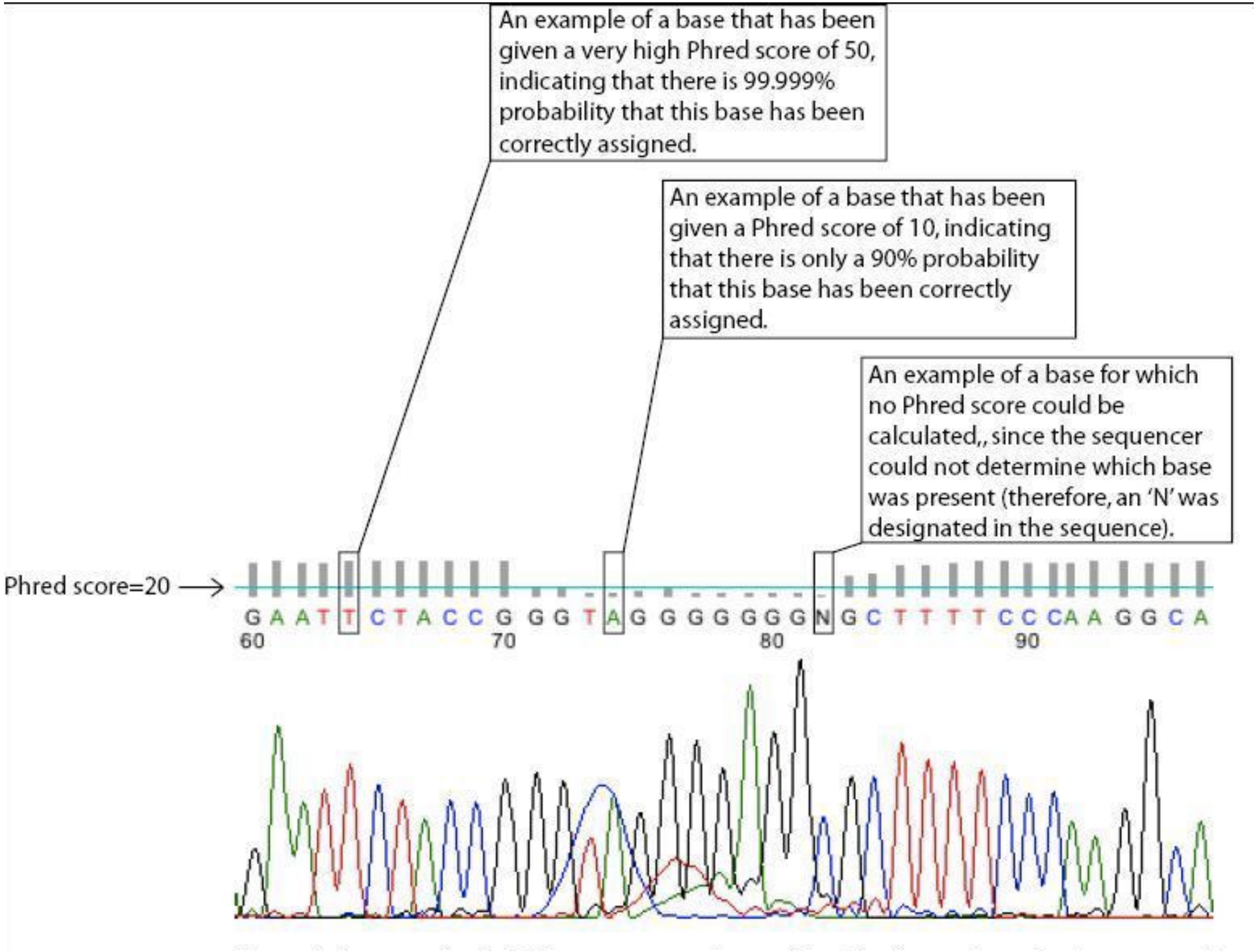
or

$$P = 10^{\frac{-Q}{10}}$$

The quality is logarithmically related to the base-calling error probabilities.

Y, en general, valores Q-SCORE por encima de 30 se consideran muy buenos!

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



http://upload.wikimedia.org/wikipedia/en/archive/e/e7/20070222215757%21Phred_Figure_1.jpg

Para guardar esta información, los bioinformáticos crearon un formato de archivo especial que contiene las bases nucleotídicas secuenciadas, conjuntamente con los valores de Q-SCORE

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

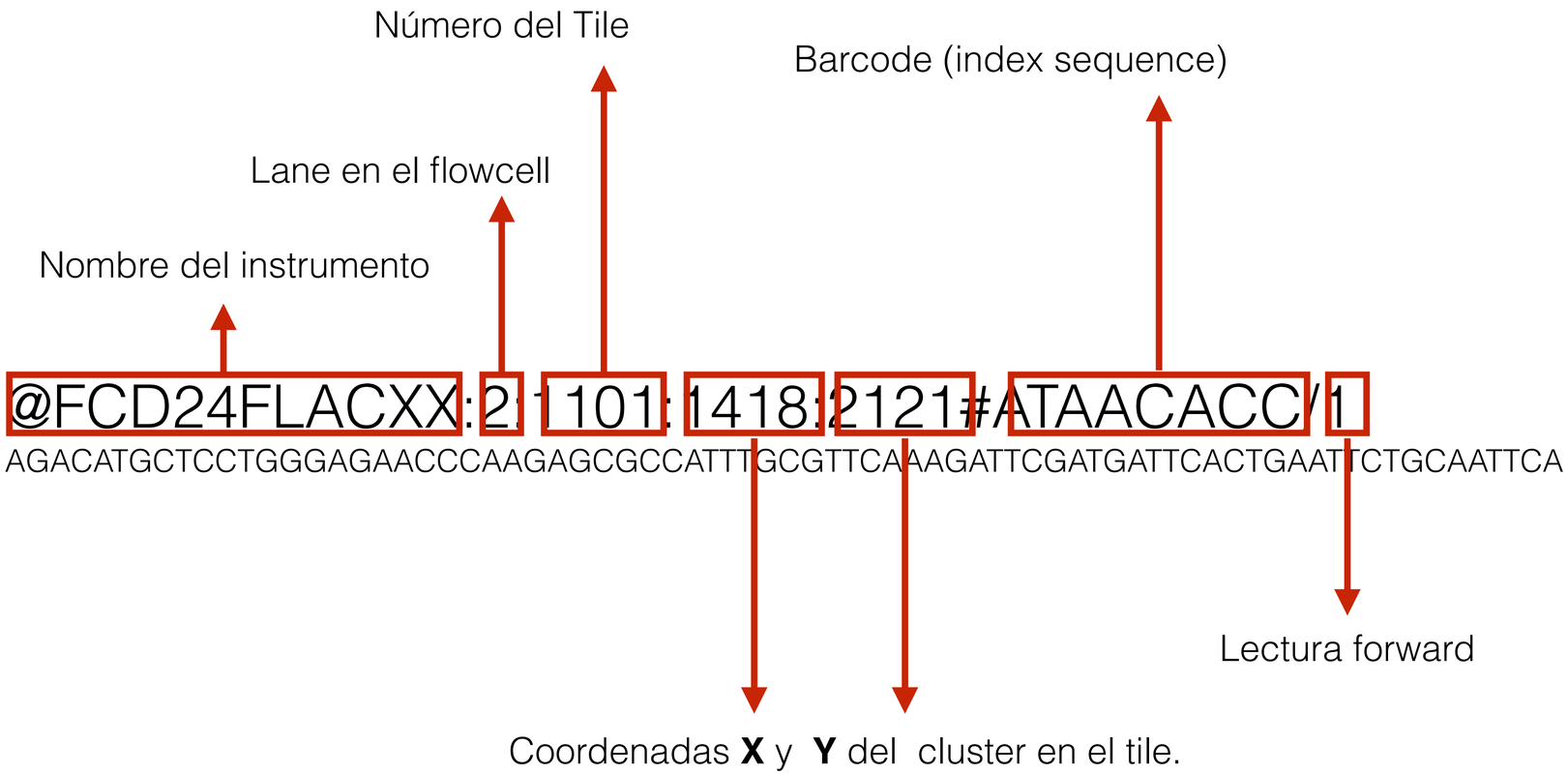
Archivo FASTQ

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Lowest → Highest

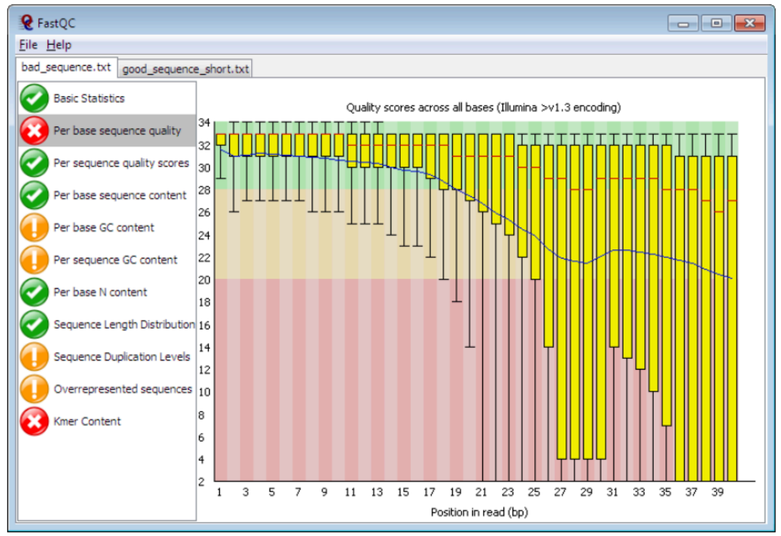
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Significado del contenido del header del fastq



¿Pero cómo analizar la calidad de todo un genoma o grupo de genomas de manera rápida y eficiente?

pycoQC



Para eso existen varias herramientas bioinformáticas. Como FASTQC (Illumina) o PycoQC (Nanopore)

Veamos como funciona

Corre el siguiente comando desde tu terminal:

```
(base) gibbs@cursoVigilancia:~/pycoqc$ pycoQC -f /data/setTutorial/sequencing_summary.txt -o analisisCalidad.html
```

Nota que estamos usando el archivo *sequencing_summary*, no los FASTQ directamente.

Este será nuestro archivo con los resultados

Veras mensajes en la consola como estos:

```
Checking arguments values
Check input data files
Parse data files
Merge data
Cleaning data
        Discarding lines containing NA values
Note: NumExpr detected 50 cores but "NUMEXPR_MAX_THREADS" not set, so enforcing safe limit of 8.
                0 reads discarded
        Filtering out zero length reads
                0 reads discarded
        Sorting run IDs by decreasing throughput
                Run-id order ['96a5435eb11667db349d8df1cada7aa60ef7d897', '86d5d3cee5b59984516683259bdca88c30a81ede']
        Reordering runids
                Processing reads with Run_ID 96a5435eb11667db349d8df1cada7aa60ef7d897 / time offset: 0
                Processing reads with Run_ID 86d5d3cee5b59984516683259bdca88c30a81ede / time offset: 235904.21025
        Cleaning up low frequency barcodes
                0 reads with low frequency barcode unset
        Cast value to appropriate type
        Reindexing dataframe by read_ids
                1,000,574 Final valid reads
Loading plotting interface
        Found 1,000,574 total reads
        Found 949,531 pass reads (qual >= 7.0 and length >= 0)
Generating HTML report
        Parsing html config file
        Running method run_summary
                Computing plot
        Running method basecall_summary
                Computing plot
        Running method alignment_summary
                No Alignment information available
        Running method read_len_1D
                Computing plot
        Running method align_len_1D
```

Después de unos segundos tendrás el resultado en tu carpeta:

```
(base) gibbs@cursoVigilancia:~/pycoqc$ ls -l
total 6663
-rw-rw-r-- 1 gibbs gibbs 6740936 Sep  2 14:10 analisisCalidad.html
```

NOTA: dado que estas trabajando en un sistema sin entorno gráfico no podrás ver los resultados directamente en el terminal, para eso descarga el archivo "analisisCalidad.html" a tu equipo local.

(ahora en tu equipo local) abre el archivo analisisCalidad.html con cualquier navegador web!

PycoQC report

Generated on 02/09/21 with pycoQC 2.5.2

General run summary

Status	Run Duration (h)	Active Channels	Number of Runids	Number of Barcodes
All Reads	138.39	508	2	8
Pass Reads	138.39	508	2	8

Basecall summary

Status	Reads	Bases	MSO	Median Read Length	Median PHRED score
All Reads	1.080874e+6	4.722867e+8	486	466	11.390
Pass Reads	9.695310e+5	4.482853e+8	485	465	11.387

Basecalled reads length

Basecalled reads PHRED quality

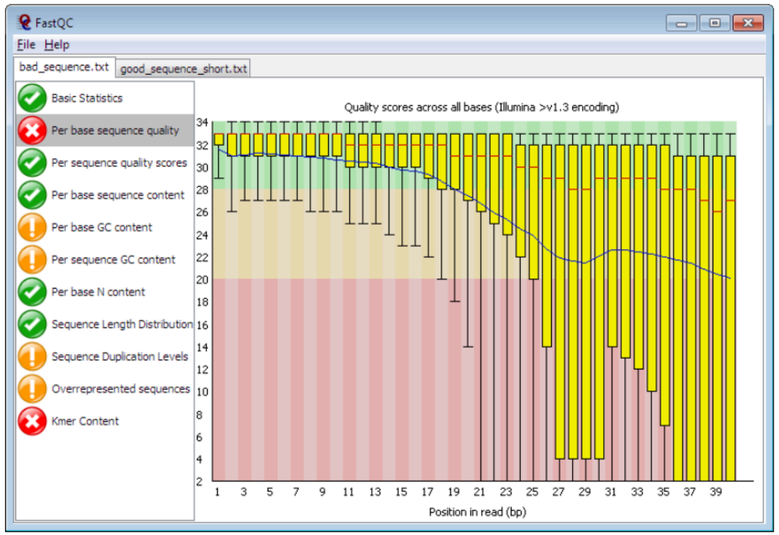
Basecalled reads length vs reads PHRED quality

Que emoción…

Analicemos los resultados!

AYUDA: https://timkahlke.github.io/LongRead_tutorials/QC_P.html

Ahora veamos como funciona



fastQC

Corre el siguiente comando desde tu terminal:

```
(base) gibbs@cursoVigilancia:~/fastqc$ fastqc /data/setTutorial/fastq_pass/barcode09/*.gz -o ./
```

Nota estamos analizando todos los archivos terminados en .gz (.fastq.gz).
Estamos analizando solamente los FASTQ del barcode09

Este será nuestra carpeta con los resultados (es decir la carpeta actual).

Veras mensajes en la consola como estos:

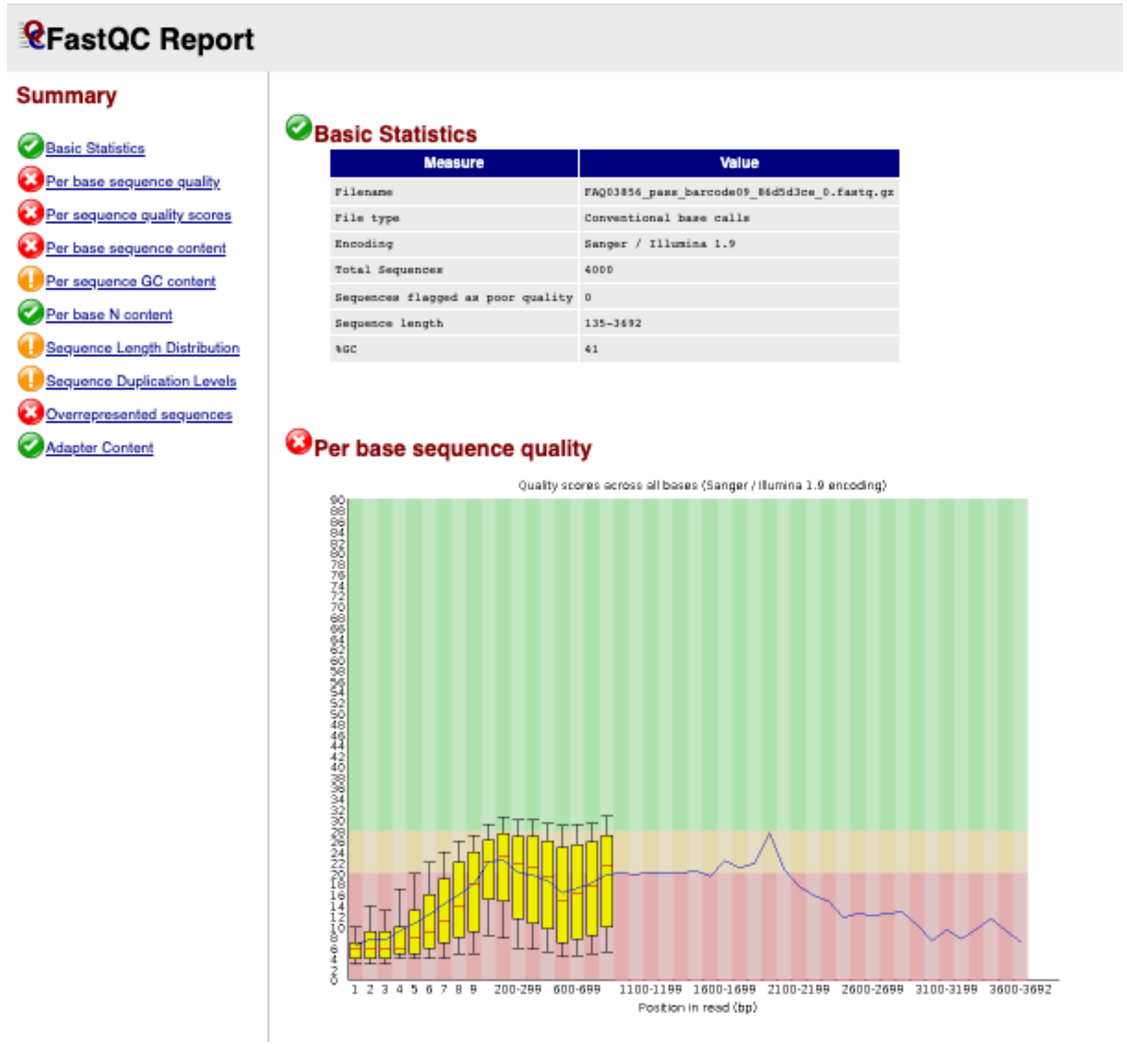
```
(base) gibbs@cursoVigilancia:~/fastqc$ fastqc /data/setTutorial/fastq_pass/barcode09/*.gz -o ./
Started analysis of FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
Approx 25% complete for FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
Approx 50% complete for FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
Approx 75% complete for FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
Approx 100% complete for FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
Analysis complete for FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
Started analysis of FAQ03856_pass_barcode09_86d5d3ce_1.fastq.gz
Approx 25% complete for FAQ03856_pass_barcode09_86d5d3ce_1.fastq.gz
Approx 50% complete for FAQ03856_pass_barcode09_86d5d3ce_1.fastq.gz
Approx 75% complete for FAQ03856_pass_barcode09_86d5d3ce_1.fastq.gz
Approx 100% complete for FAQ03856_pass_barcode09_86d5d3ce_1.fastq.gz
Analysis complete for FAQ03856_pass_barcode09_86d5d3ce_1.fastq.gz
Started analysis of FAQ03856_pass_barcode09_86d5d3ce_10.fastq.gz
Approx 20% complete for FAQ03856_pass_barcode09_86d5d3ce_10.fastq.gz
Approx 50% complete for FAQ03856_pass_barcode09_86d5d3ce_10.fastq.gz
Approx 75% complete for FAQ03856_pass_barcode09_86d5d3ce_10.fastq.gz
Approx 100% complete for FAQ03856_pass_barcode09_86d5d3ce_10.fastq.gz
Analysis complete for FAQ03856_pass_barcode09_86d5d3ce_10.fastq.gz
Started analysis of FAQ03856_pass_barcode09_86d5d3ce_11.fastq.gz
Approx 25% complete for FAQ03856_pass_barcode09_86d5d3ce_11.fastq.gz
Approx 50% complete for FAQ03856_pass_barcode09_86d5d3ce_11.fastq.gz
Approx 70% complete for FAQ03856_pass_barcode09_86d5d3ce_11.fastq.gz
Approx 100% complete for FAQ03856_pass_barcode09_86d5d3ce_11.fastq.gz
Analysis complete for FAQ03856_pass_barcode09_86d5d3ce_11.fastq.gz
Started analysis of FAQ03856_pass_barcode09_86d5d3ce_12.fastq.gz
Approx 20% complete for FAQ03856_pass_barcode09_86d5d3ce_12.fastq.gz
Approx 45% complete for FAQ03856_pass_barcode09_86d5d3ce_12.fastq.gz
Approx 70% complete for FAQ03856_pass_barcode09_86d5d3ce_12.fastq.gz
Approx 100% complete for FAQ03856_pass_barcode09_86d5d3ce_12.fastq.gz
Analysis complete for FAQ03856_pass_barcode09_86d5d3ce_12.fastq.gz
Started analysis of FAQ03856_pass_barcode09_86d5d3ce_13.fastq.gz
Approx 25% complete for FAQ03856_pass_barcode09_86d5d3ce_13.fastq.gz
```

Después de unos segundos tendrás los resultados en tu carpeta:

```
(base) gibbs@cursoVigilancia:~/fastqc$ ls -l
total 28270
-rw-rw-r-- 1 gibbs gibbs 685217 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_0_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 428600 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_0_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 700579 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_10_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 432112 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_10_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 678135 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_11_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 423294 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_11_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 709384 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_12_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 438401 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_12_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 683942 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_13_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 414395 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_13_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 747471 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_14_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 459444 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_14_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 781630 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_15_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 471322 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_15_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 759386 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_16_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 445032 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_16_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 672869 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_17_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 409571 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_17_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 678753 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_18_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 399731 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_18_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 754535 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_19_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 461211 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_19_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 783750 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_1_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 464570 Sep  2 14:59 FAQ03856_pass_barcode09_86d5d3ce_1_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 783887 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_20_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 465074 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_20_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 701233 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_21_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 431628 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_21_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 681753 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_2_fastqc.html
-rw-rw-r-- 1 gibbs gibbs 421701 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_2_fastqc.zip
-rw-rw-r-- 1 gibbs gibbs 745938 Sep  2 15:00 FAQ03856_pass_barcode09_86d5d3ce_3_fastqc.html
```

NOTA: dado que estas trabajando en un sistema sin entorno gráfico no podrás ver los resultados directamente en el terminal, para eso descarga los archivos terminados en "html" a tu equipo local.

(ahora en tu equipo local) abre el archivo "FAQ03856…html" con cualquier navegador web!



FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	FAQ03856_pass_barcode09_86d5d3ce_0.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4000
Sequences flagged as poor quality	0
Sequence length	135-3692
%GC	41

Per base sequence quality

Que emoción…

Analicemos los resultados!

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

Instituto Nacional de Salud

Taller virtual:

Análisis bioinformático de SARS-CoV-2

Septiembre 1 al 3 2021

Análisis de Calidad

Universidad Nacional de Colombia

Andrés M. Pinzón Ph. D.

ampinzonv@unal.edu.co

Instituto de Genética

Universidad Nacional de Colombia

GiββS

BIOINFORMATICS AND SYSTEMS BIOLOGY GROUP

https://gibbslab.github.io/

@gibbsclab gibbslab