

Instalación de FastQC usando conda

Conda es un programa que maneja las instalaciones de paquetes, programas y librerías, que además nos da la opción de “encapsular” las instalaciones en “ambientes”, de forma que no intervengan entre ellos.

Conda está instalado en JupyterHub.

Para instalar FastQC utilizando conda ejecutamos:

```
conda install -c bioconda fastqc
```

Para correr FastQC ejecutamos directamente:

```
fastqc archivo.fastq
```

Análisis de calidad de datos crudos de secuenciación

Ejercicio:

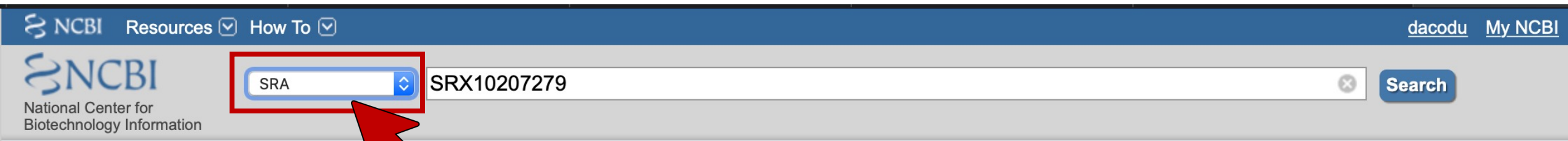
Objetivo:

Análisis de calidad de datos de secuenciación masiva bajados del NCBI

1. Bajar los datos de secuenciación masiva correspondientes al experimento [SRX10207279](#)
2. Descomprimirlo en la línea de comandos, de forma que queden los reads forward y reverse en diferentes archivos
3. realizar el análisis de calidad de los reads con FastQC

1. Bajar los datos de secuenciación masiva del experimento [SRX10207279](#)

Buscamos el accession SRX10207279 en el NCBI



The screenshot shows the NCBI search bar. On the left is the NCBI logo and the text "National Center for Biotechnology Information". To the right of the logo are links for "Resources" and "How To". Further right are links for "dacodu" and "My NCBI". The search bar itself contains the text "SRX10207279". To the left of the search bar is a dropdown menu with "SRA" selected. A red rectangle highlights the dropdown menu, and a red arrow points to it from below. To the right of the search bar is a "Search" button.

Seleccionar **SRA**

1. Bajar los datos de secuenciación masiva del experimento [SRX10207279](#)

SRX10207279: DNA-seq of K.pneumoniae phage

1 ILLUMINA (Illumina HiSeq 2500) run: 55,990 spots, 24.3M bases, 14.5Mb download

Design: Paired-end libraries were prepared from total DNA for each sample according to the Illumina DNA Library Prep Kit (New England Biolabs, USA). Libraries were indexed using NEB Next DNA Library Prep Kit (New England Biolabs, USA) kits. Shotgun genomic sequencing was performed on HiSeq 2500 (Illumina, USA) recommendations using the following reagent kits: HiSeq Rapid PE Cluster Kit v2, HiSeq 2500 Rapid PE Reagent Kit v2 (Illumina, USA).

Submitted by: Federal Research and Clinical Center of Physical-Chemical Medicine

Study: Klebsiella pneumoniae bacteriophages

[PRJNA705078](#) • [SRP308894](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample:

[SAMN18062703](#) • [SRS8353132](#) • [All experiments](#) • [All runs](#)

Organism: Klebsiella virus KpS8

Library:

Name: L001

Instrument: Illumina HiSeq 2500

Strategy: WGS

Source: GENOMIC

Selection: RANDOM

Layout: PAIRED

Runs: 1 run, 55,990 spots, 24.3M bases, 14.5Mb

Run	# of Spots	# of Bases	Size	Published
SRR13827880	55,990	24.3M	14.5Mb	2021-03-02

Identificamos:

- Plataforma de secuenciación
- Tipo de librería (PE, SE)

Copiamos el accession
correspondiente a la corrida

1. Bajar los datos de secuenciación masiva del experimento [SRX10207279](#)

En la terminal de JupyterHub:

```
jovyan@jupyter-236944:~$ mkdir Bioinfo2
jovyan@jupyter-236944:~$ cd Bioinfo2/
jovyan@jupyter-236944:~/Bioinfo2$ prefetch SRR13827880

2021-03-23T17:42:39 prefetch.2.10.8: 1) Downloading 'SRR13827880'...
2021-03-23T17:42:39 prefetch.2.10.8:   Downloading via HTTPS...
2021-03-23T17:43:11 prefetch.2.10.8:   HTTPS download succeed
2021-03-23T17:43:11 prefetch.2.10.8:   'SRR13827880' is valid
2021-03-23T17:43:11 prefetch.2.10.8: 1) 'SRR13827880' was downloaded successfully
2021-03-23T17:43:11 prefetch.2.10.8: 'SRR13827880' has 0 unresolved dependencies
```

```
cd # ir al directorio /home/joyvan
mkdir Bioinfo2 # crear el directorio Bioinfo2
cd Bioinfo2 # entrar al directorio Bioinfo2
prefetch SRR13827880 # descargar los datos desde el NCBI
```

2. Descomprimir

Compruebo que se bajó el archivo y su tamaño

```
jovyan@jupyter-236944:~/Bioinfo2$ ls
SRR13827880
jovyan@jupyter-236944:~/Bioinfo2$ ll SRR13827880/
total 14852
drwxr-sr-x 2 jovyan users      4096 Mar 23 17:43 ./
drwxr-sr-x 3 jovyan users      4096 Mar 23 17:42 ../
-rw-r--r-- 1 jovyan users 15199369 Mar 23 17:43 SRR13827880.sra
```

2. Descomprimir

Descomprimo el archivo .sra:

```
jovyan@jupyter-236944:~/Bioinfo2$ cd SRR13827880/
jovyan@jupyter-236944:~/Bioinfo2/SRR13827880$ fastq-dump --split-files SRR13827880.sra
Read 55990 spots for SRR13827880.sra
Written 55990 spots for SRR13827880.sra
jovyan@jupyter-236944:~/Bioinfo2/SRR13827880$ ll
total 70240
drwxr-sr-x 2 jovyan users      4096 Mar 23 17:50 ./
drwxr-sr-x 3 jovyan users      4096 Mar 23 17:42 ../
-rw-r--r-- 1 jovyan users 28229618 Mar 23 17:50 SRR13827880_1.fastq
-rw-r--r-- 1 jovyan users 28486070 Mar 23 17:50 SRR13827880_2.fastq
-rw-r--r-- 1 jovyan users 15199369 Mar 23 17:43 SRR13827880.sra
```

```
cd SRR13827880/ # ir al directorio SRR13827880
fastq-dump --split-files SRR13827880.sra # Descomprime separando los reads
forward y los reverse en diferentes archivos
```


3. Realizar el análisis de calidad de los reads con FastQC

Ejecutamos FastQC para los dos archivos fastq

```
jovyan@jupyter-236944:~/Bioinfo2/SRR13827880$ ~/FastQC/fastqc SRR13827880_1.fastq
Started analysis of SRR13827880_1.fastq
Approx 5% complete for SRR13827880_1.fastq
Approx 10% complete for SRR13827880_1.fastq
Approx 15% complete for SRR13827880_1.fastq
Approx 20% complete for SRR13827880_1.fastq
Approx 25% complete for SRR13827880_1.fastq
Approx 30% complete for SRR13827880_1.fastq
Approx 35% complete for SRR13827880_1.fastq
Approx 40% complete for SRR13827880_1.fastq
Approx 45% complete for SRR13827880_1.fastq
Approx 50% complete for SRR13827880_1.fastq
```

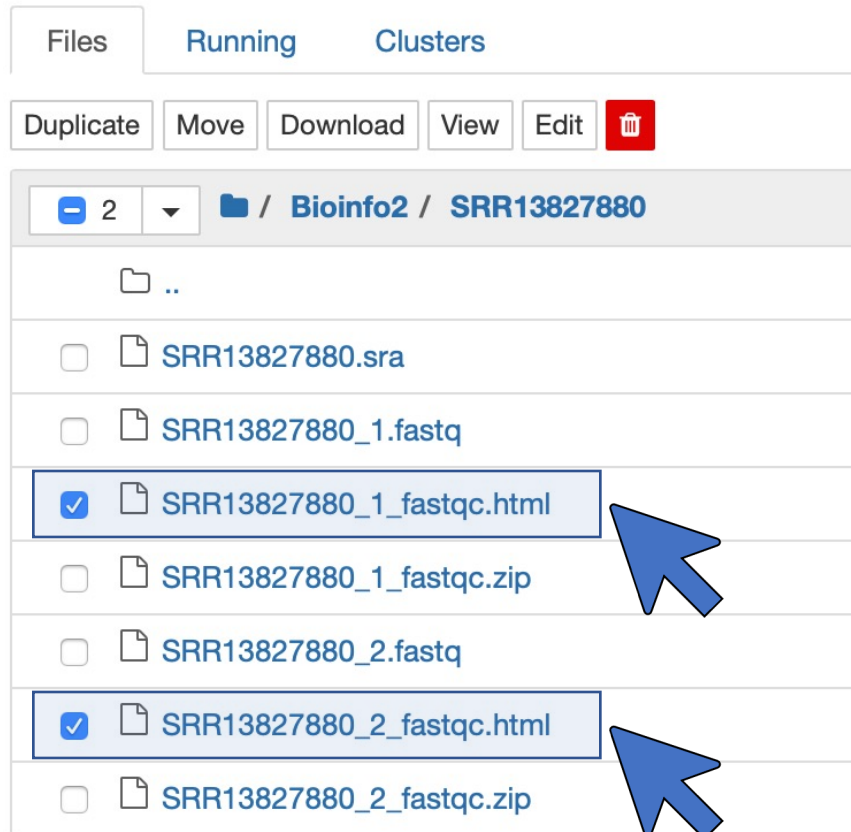
```
fastqc SRR13827880_1.fastq # ejecuta Fastq sobre los F
fastqc SRR13827880_2.fastq # ejecuta Fastq sobre los R
```

3. Realizar el análisis de calidad de los reads con FastQC

Vemos los archivos creados por fastqc:

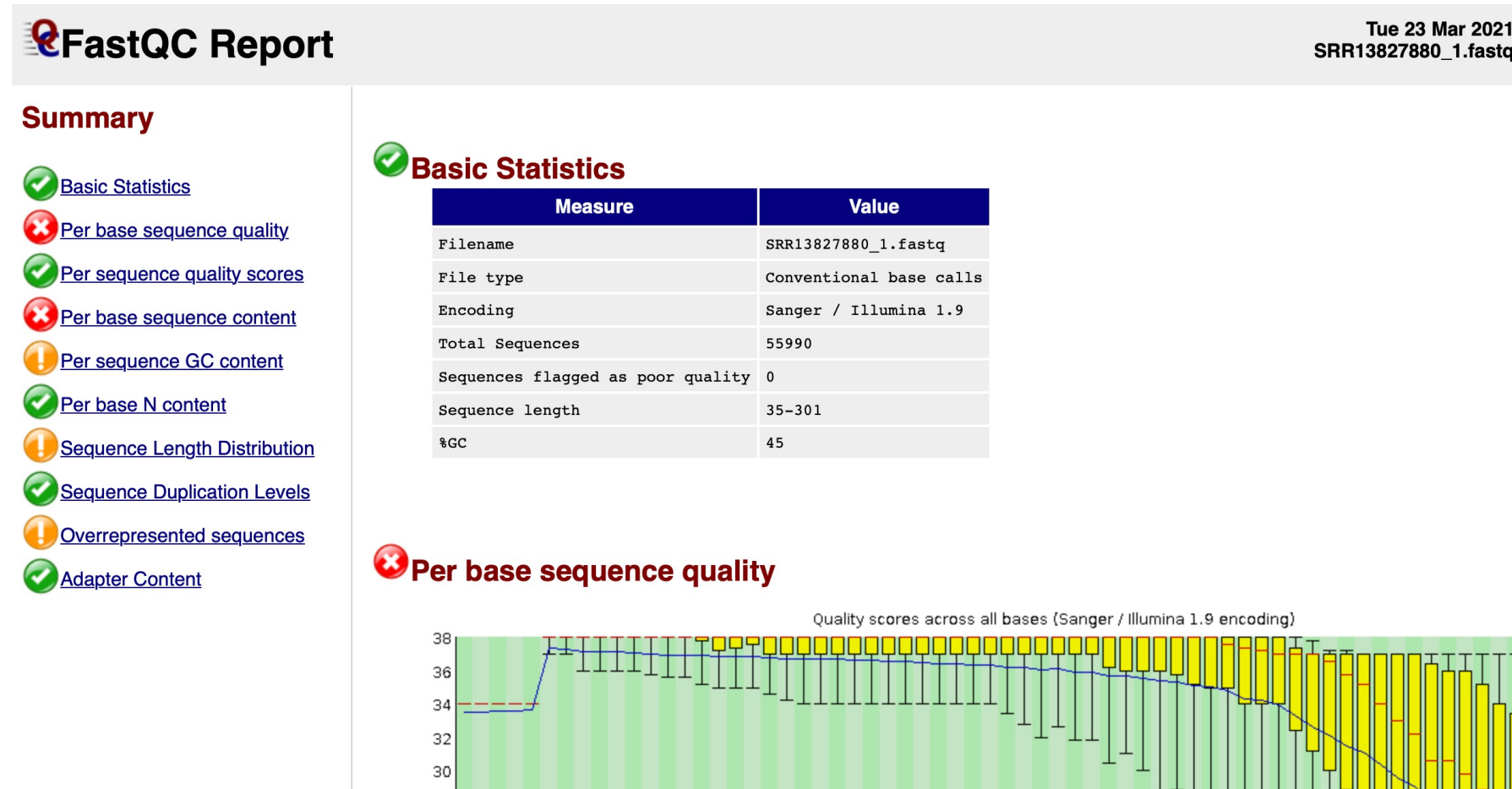
```
jovyan@jupyter-236944:~/Bioinfo2/SRR13827880$ ll
total 72292
drwxr-sr-x 2 jovyan users      4096 Mar 23 17:57 ./
drwxr-sr-x 3 jovyan users      4096 Mar 23 17:42 ../
-rw-r--r-- 1 jovyan users 28229618 Mar 23 17:50 SRR13827880_1.fastq
-rw-r--r-- 1 jovyan users   693944 Mar 23 17:55 SRR13827880_1_fastqc.html
-rw-r--r-- 1 jovyan users   349088 Mar 23 17:55 SRR13827880_1_fastqc.zip
-rw-r--r-- 1 jovyan users 28486070 Mar 23 17:50 SRR13827880_2.fastq
-rw-r--r-- 1 jovyan users   689828 Mar 23 17:57 SRR13827880_2_fastqc.html
-rw-r--r-- 1 jovyan users   358533 Mar 23 17:57 SRR13827880_2_fastqc.zip
-rw-r--r-- 1 jovyan users 15199369 Mar 23 17:43 SRR13827880.sra
```

3. Realizar el análisis de calidad de los reads con FastQC



Hacemos clic sobre los archivos .html para abrirlos en el Browser

3. Realizar el análisis de calidad de los reads con FastQC



Resultados de FastQC



Basic Statistics

Measure	Value
Filename	SRR13827880_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	55990
Sequences flagged as poor quality	0
Sequence length	35-301
%GC	45

Nombre del archivo

Codificación de la calidad

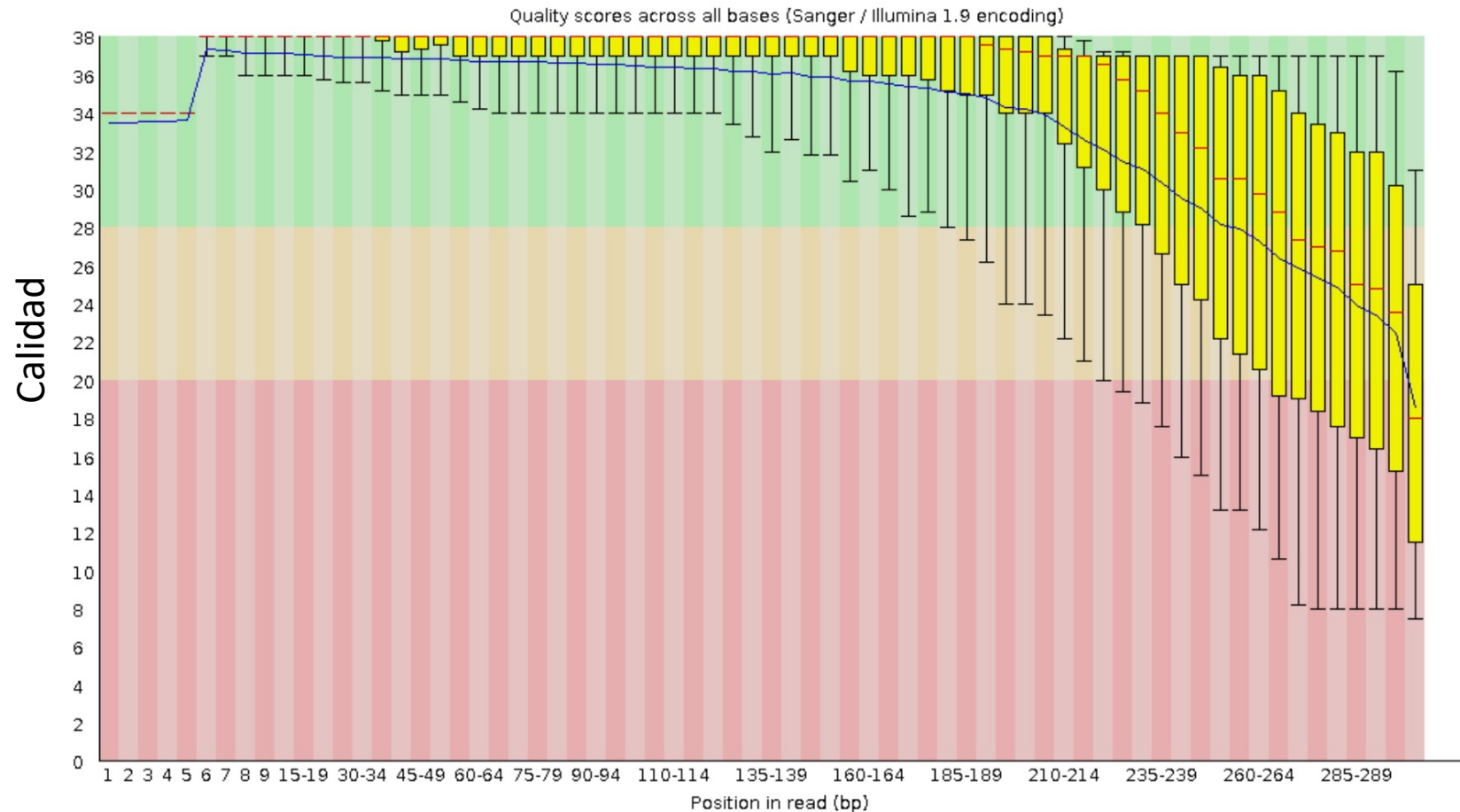
Número de reads

Largo de los reads (rango)

%GC

Resultados de FastQC

✖ Per base sequence quality



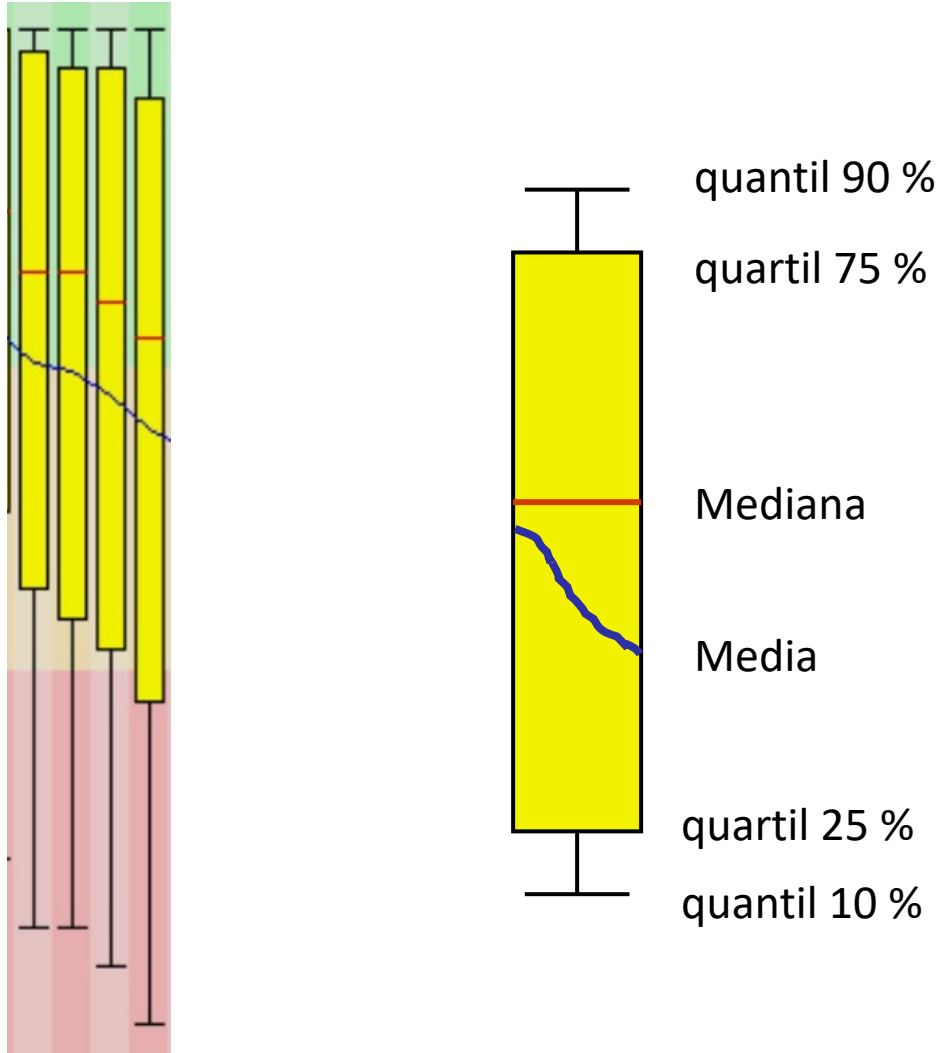
Calidad por base

Posiciones de buena calidad

Posiciones de calidad media

Posiciones de mala calidad

Resultados de FastQC



❌ Per base sequence quality

- quartil 25% en cualquier base es menor a 5
- mediana en cualquier posición es menor a 20

✅ Per base sequence quality

- quartil 25% en todas las bases es mayor a 10
- mediana en todas las bases es mayor a 25

Visualización de los resultados de calidad con multiqc

MultiQC es un programa que genera reportes mucho más amigables que los de FastQC.

Para instalarlo en JupyterHub Notebook:

```
pip install multiqc
```


Visualización de los resultados de calidad con multiqc

Para correr MultiQC sobre los resultados de FastQC contenidos en una carpeta:

```
multiqc -d <Carpeta>
```





modificar Carpeta por el nombre de la carpeta que contiene los resultados de FastQC







Visualización de los resultados de calidad con multiqc



Files Running Clusters

Select items to perform actions on them.

☐ 0   / Bioinfo2 / SRR13827880

-  ..
- ☐  multiqc_data
- ☐  multiqc_report.html
- ☐  SRR13827880.sra
- ☐  SRR13827880_1.fastq
- ☐  SRR13827880_1_fastqc.html

Para visualizar el reporte creado por MultiQC hacemos clic sobre los archivos en la página inicial de JupyterHub:



Visualización de los resultados de calidad con multiqc

MultiQC
v1.10

General Stats

FastQC

Sequence Counts

Sequence Quality Histograms

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

Status Checks



A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2021-03-23, 19:40 based on data in: `/home/jovyan/Bioinfo2/SRR13827880`

General Statistics

Copy table

Configure Columns

Plot

Showing 2/2 rows and 3/5 columns.

Sample Name	% Dups	% GC	M Seqs
SRR13827880_1	24.7%	45%	0.1
SRR13827880_2	22.0%	45%	0.1

FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.

Number of reads

Percentages

Help

FastQC: Sequence Counts

Export Plot

El reporte creado por multiqc se abre en una nueva pestaña del Browser

Explora el reporte creado por **multiqc** y los creados con **FastQC**

- ¿Cuál te parece mejor?
- ¿A qué corresponden los diferentes módulos del reporte?
- ¿Consideras que los reads son de buena calidad?
- ¿Consideras necesario filtrar los reads? En caso afirmativo: ¿qué les harías?