

Módulo: Expresión diferencial

Bioinformática y Estadística 2

Dra. Evelia Coss

Dra. Alejandra Medina

21 al 24 de Febrero, 2023

Descarga de datos de SRA

Modificar las descargas de NCBI

Error 1

Para descargar archivos por cada USUARIO debes modificar lo siguiente:

```
module load sra/3.0.0  
vdb-config -i # disable storage of cache in ~
```

Vdb = virtual database

SRA configuration

[save] [exit] [discard] [default]

MAIN

CACHE

AWS

GCP

MET

TOOLS

 enable local file-caching

location of user-repository:

[choose]

[clear]

process-local location:

[choose]

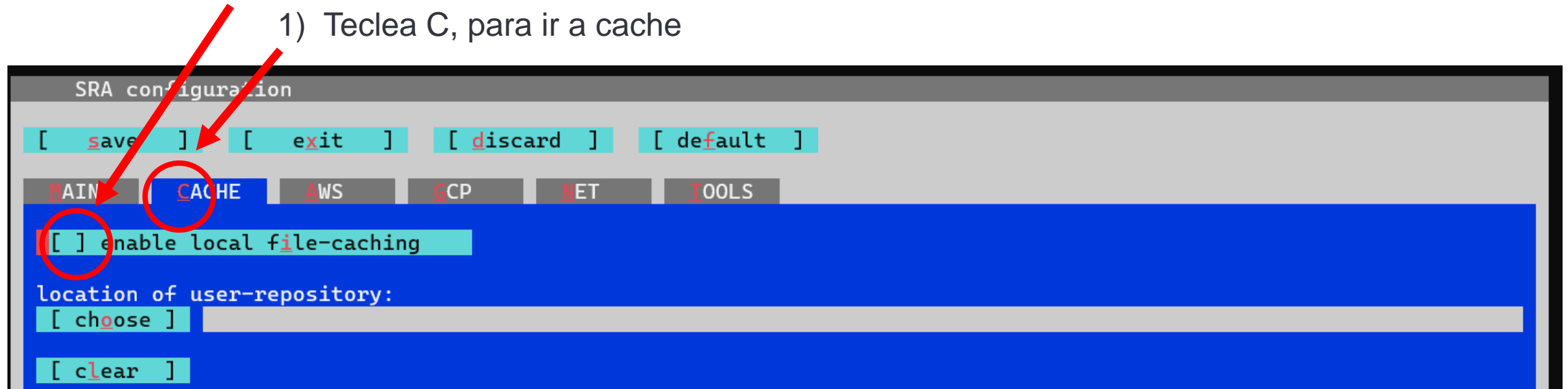
[clear]

RAM used: + 0 - MB

use local cache

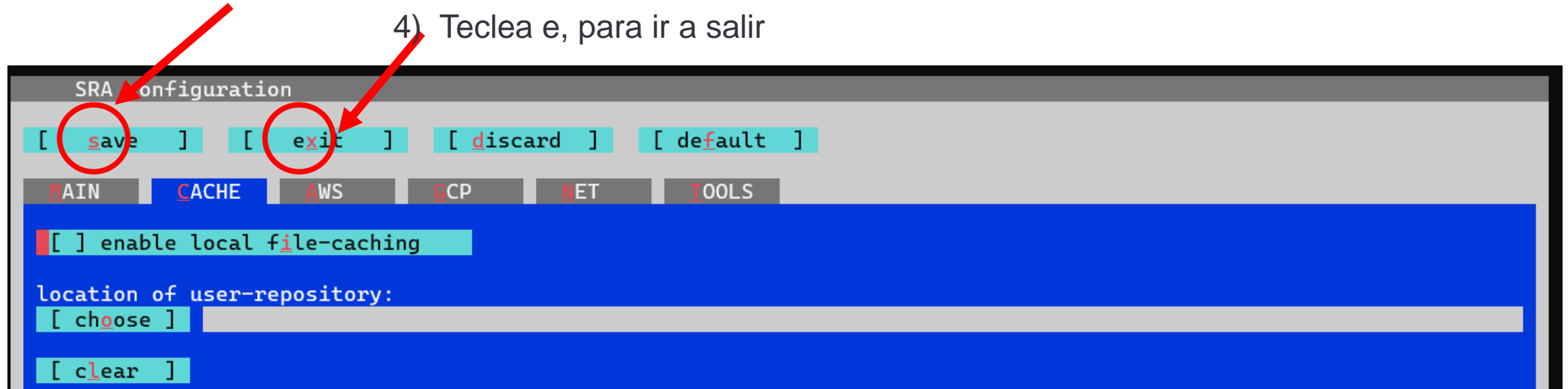
2) Teclea i, para ir a la primera opcion, despues con ENTER vas a deshabilitar la opcion.

1) Teclea C, para ir a cache



3) Teclea s, para guardar. Luego da O, de OK

4) Teclea e, para ir a salir



Error en permisos de los archivos

Error 2

Cada usuario tiene permisos diferentes cuando crea un archivo. Los permisos pueden modificarse con *chmod*.

Los caracteres atribuidos a los permisos son:

- *r* : escritura (Read)
- *w* : lectura (Write)
- *x* : ejecución (eXecute)

| permisos | pertenece |
|------------------------|-----------|
| <code>rwX-----</code> | usuario |
| <code>---r-X---</code> | grupo |
| <code>-----r-X</code> | otros |

```
chmod 777 Archivo
```

```
[ecoss@chromatin Homo_sapiens]$ ls -l data/*
-rw-r--r-- 1 ecoss amedina 4658351418 Feb 22 01:26 data/SRR18745762_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 4708166626 Feb 22 01:26 data/SRR18745762_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 2843661143 Feb 22 02:19 data/SRR18745763_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 2938220058 Feb 22 02:19 data/SRR18745763_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 3014775630 Feb 22 03:17 data/SRR18745764_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 3047537455 Feb 22 03:17 data/SRR18745764_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 2233640797 Feb 22 03:58 data/SRR18745765_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 2231114552 Feb 22 03:58 data/SRR18745765_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 621576574 Feb 22 04:10 data/SRR18745766_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 611866388 Feb 22 04:10 data/SRR18745766_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 2504474730 Feb 22 04:56 data/SRR18745767_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 2482010231 Feb 22 04:56 data/SRR18745767_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 305028530 Feb 22 05:01 data/SRR18745768_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 238906712 Feb 22 05:01 data/SRR18745768_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 453580963 Feb 22 05:10 data/SRR18745769_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 446989649 Feb 22 05:10 data/SRR18745769_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 2478200510 Feb 22 05:57 data/SRR18745770_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 2462545568 Feb 22 05:57 data/SRR18745770_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 286763123 Feb 22 06:02 data/SRR18745771_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 223025734 Feb 22 06:02 data/SRR18745771_2.fastq.gz
-rw-r--r-- 1 ecoss amedina 454432229 Feb 22 06:11 data/SRR18745772_1.fastq.gz
-rw-r--r-- 1 ecoss amedina 448228595 Feb 22 06:11 data/SRR18745772_2.fastq.gz
```

```
[ecoss@chromatin Homo_sapiens]$ ls -l data/*
-rwxrwxrwx 1 ecoss amedina 4658351418 Feb 22 01:26 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 4708166626 Feb 22 01:26 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2843661143 Feb 22 02:19 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2938220058 Feb 22 02:19 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 3014775630 Feb 22 03:17 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 3047537455 Feb 22 03:17 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2233640797 Feb 22 03:58 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2231114552 Feb 22 03:58 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 621576574 Feb 22 04:10 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 611866388 Feb 22 04:10 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2504474730 Feb 22 04:56 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2482010231 Feb 22 04:56 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 305028530 Feb 22 05:01 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 238906712 Feb 22 05:01 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 453580963 Feb 22 05:10 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 446989649 Feb 22 05:10 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2478200510 Feb 22 05:57 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 2462545568 Feb 22 05:57 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 286763123 Feb 22 06:02 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 223025734 Feb 22 06:02 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 454432229 Feb 22 06:11 data/SRR1874
-rwxrwxrwx 1 ecoss amedina 448228595 Feb 22 06:11 data/SRR1874
```

Actualizaciones de las descargas

```
[ecoss@compute-00-11 rawData]$ chmod 777 Athaliana_Phlo/data/*  
[ecoss@compute-00-11 rawData]$ ls -l Athaliana_Phlo/data/*  
-rwxrwxrwx 1 ecoss amedina 1681491954 Feb 21 23:16 Athaliana_Phlo/data/SRR18552040_1.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1720124867 Feb 21 23:16 Athaliana_Phlo/data/SRR18552040_2.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1689996192 Feb 22 00:04 Athaliana_Phlo/data/SRR18552041_1.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1710787262 Feb 22 00:04 Athaliana_Phlo/data/SRR18552041_2.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1654387039 Feb 22 00:52 Athaliana_Phlo/data/SRR18552043_1.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1679755978 Feb 22 00:52 Athaliana_Phlo/data/SRR18552043_2.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 2572005650 Feb 22 02:04 Athaliana_Phlo/data/SRR18552044_1.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 2622283328 Feb 22 02:04 Athaliana_Phlo/data/SRR18552044_2.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1644733567 Feb 22 02:49 Athaliana_Phlo/data/SRR18552045_1.fastq.gz  
-rwxrwxrwx 1 ecoss amedina 1673190738 Feb 22 02:49 Athaliana_Phlo/data/SRR18552045_2.fastq.gz
```

Athaliana_Phlo/data/SRR18552040:

total 2038312

-rw-r--r-- 1 ecoss amedina 2087225730 Feb 20 13:09 SRR18552040.sra

Athaliana_Phlo/data/SRR18552041:

total 2030448

-rw-r--r-- 1 ecoss amedina 2079173192 Feb 20 13:20 SRR18552041.sra

Este SRA bien
descargados

Athaliana_Phlo/data/SRR18552042:

total 255376

-rw-r--r-- 1 ecoss amedina 0 Feb 21 21:36 SRR18552042.sra.lock

-rw-r--r-- 1 ecoss amedina 0 Feb 21 21:36 SRR18552042.sra.prf

-rw-r--r-- 1 ecoss amedina 261501389 Feb 21 21:37 SRR18552042.sra.tmp

Este SRA no se
descargo
correctamente

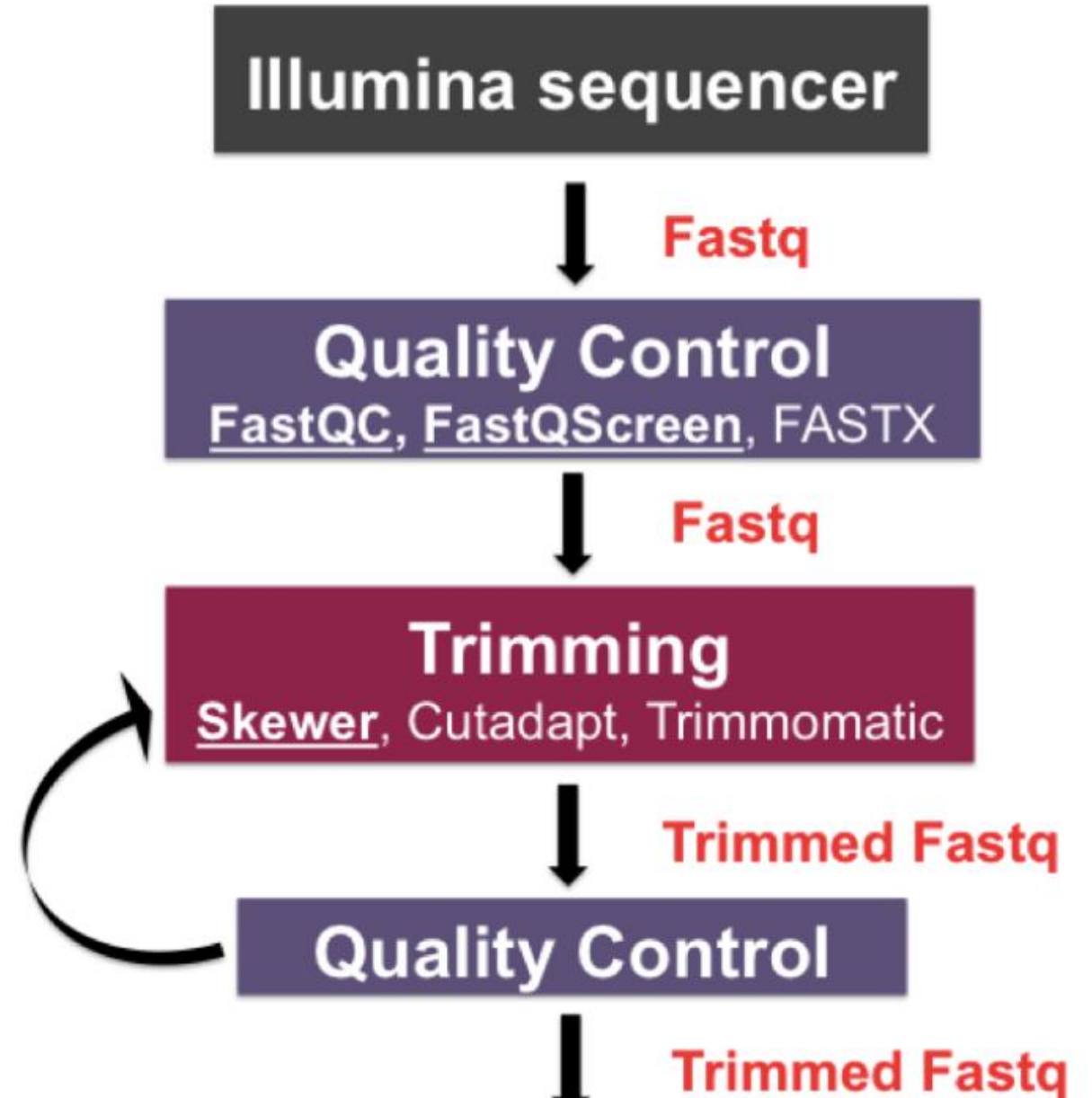
- Alguna duda...

Práctica 1: Calidad de secuencias y eliminación de adaptadores

Kallisto

Vamos a generar el script y dejar corriendo

- https://github.com/EveliaCoss/RNAseq_classFEB2023/blob/main/RNA_seq/README.md#practica1



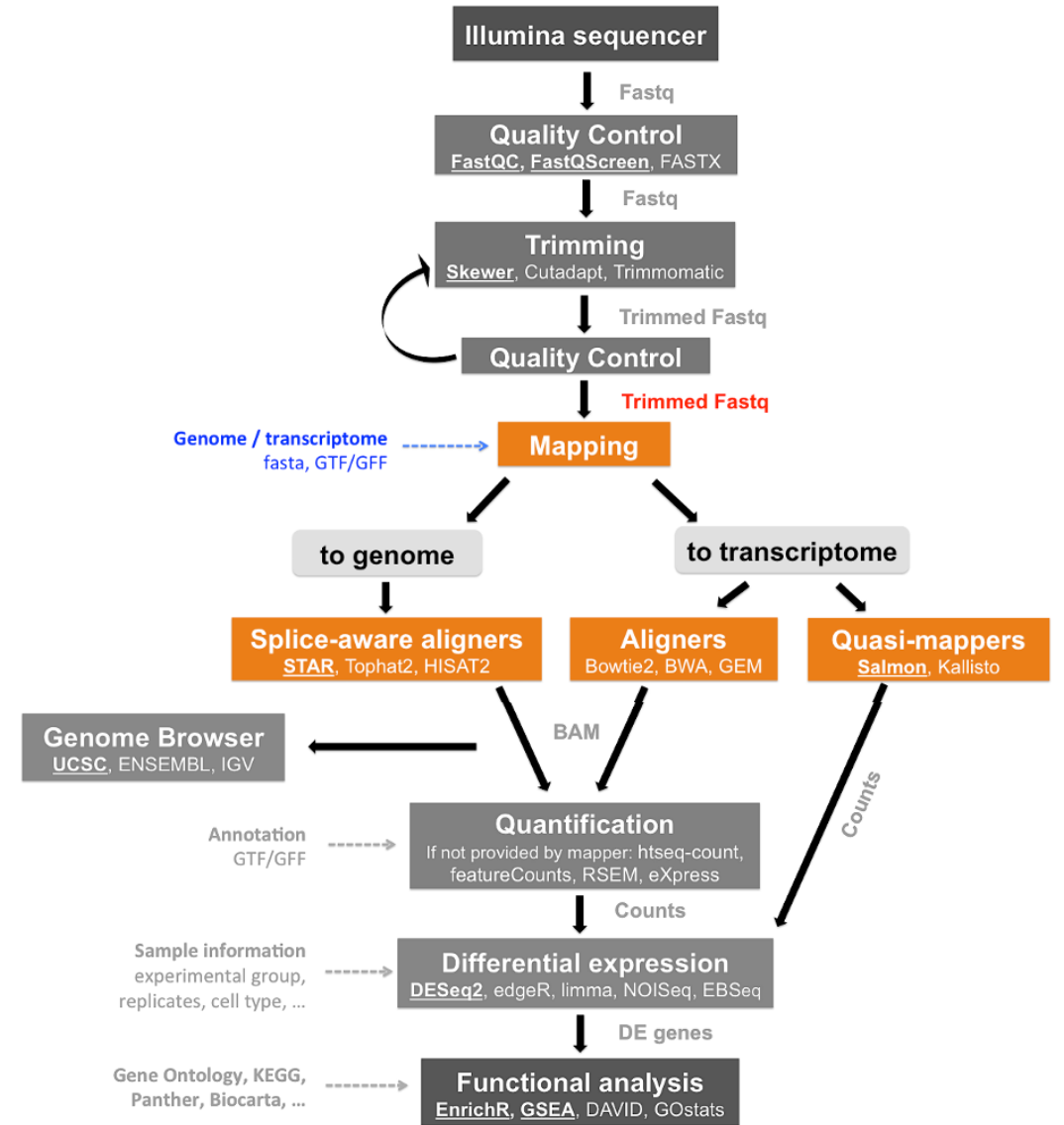
Día 2

- Diversos pipelines bioinformáticos:
 - Alineamiento al genoma de referencia
 - Ensamblaje con el transcriptoma de referencia
 - Ensamblaje *de novo*
- Ejercicio con Kallisto
 - Pseudoalineamiento de kallisto



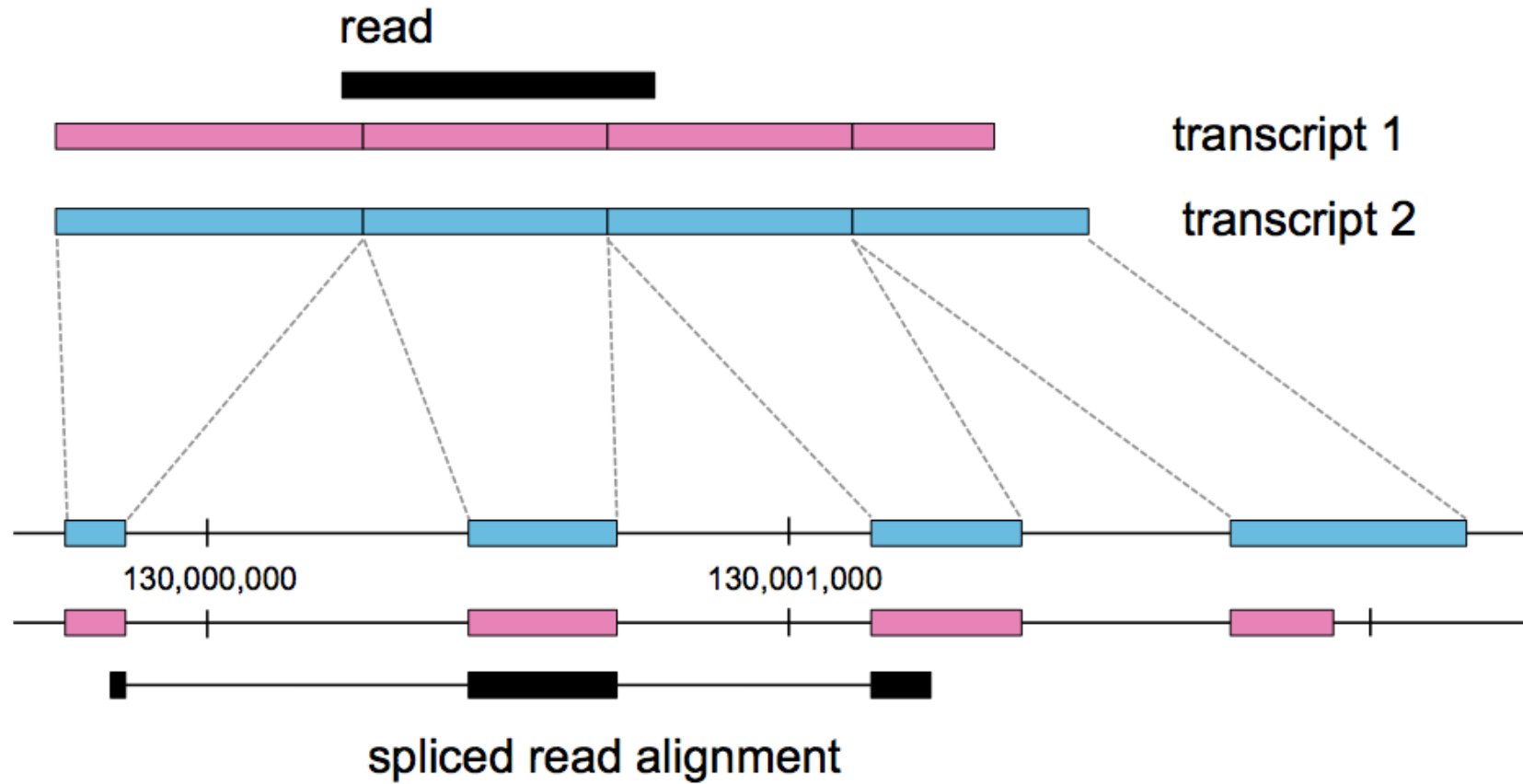
Pipeline bioinformática

Dónde estamos...



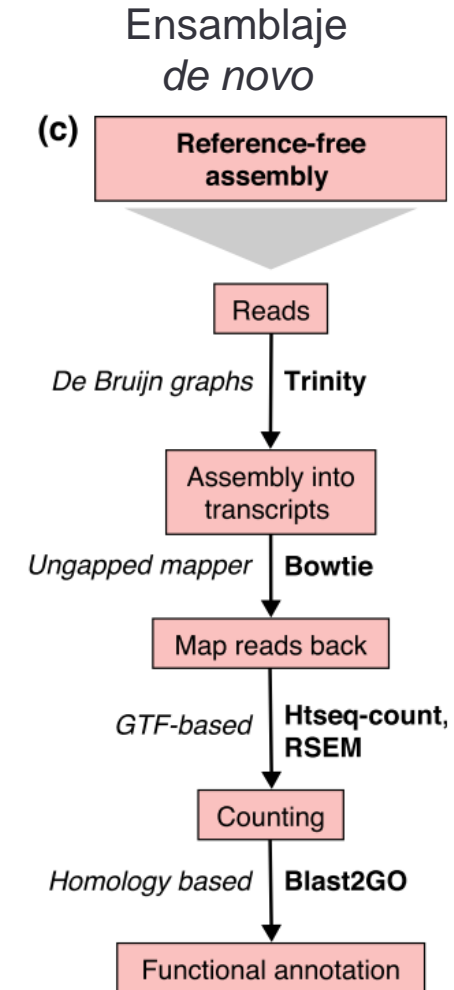
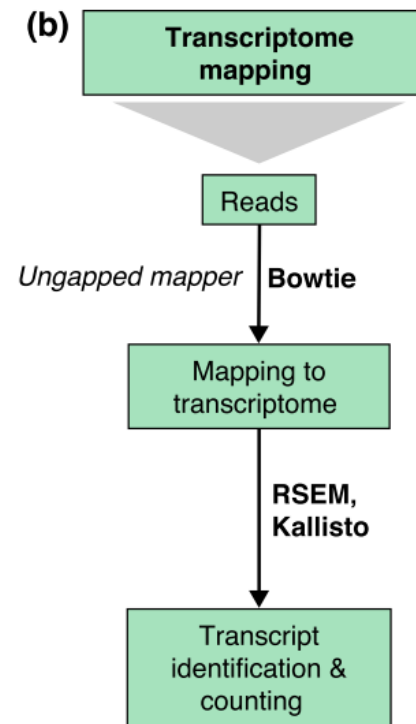
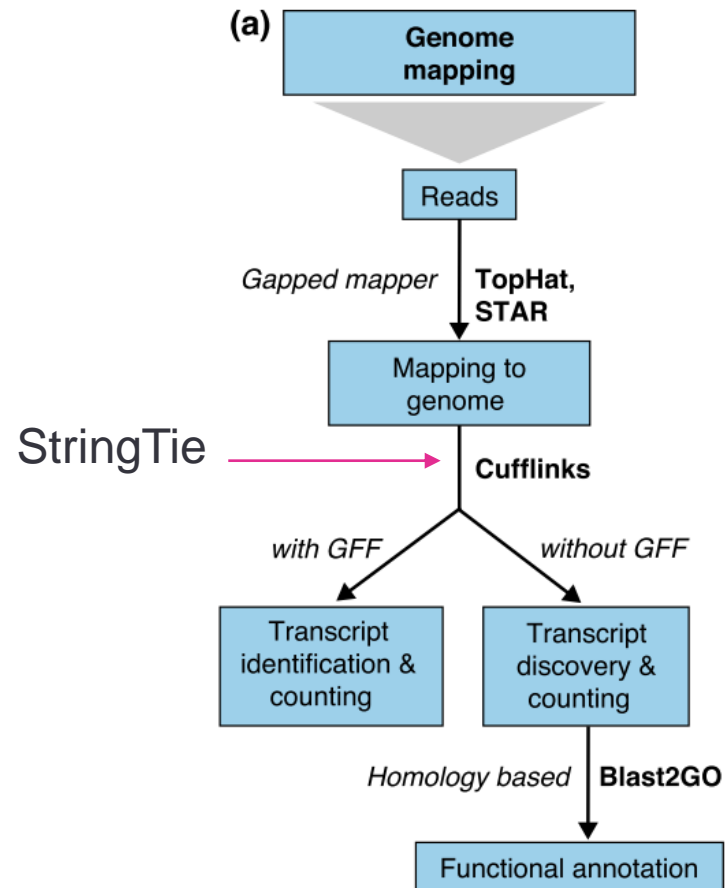
mRNA-Seq data analysis workflow
“https://biocorecrg.github.io/RNAseq_course_2019/workflow.html”

Alineamiento genómico

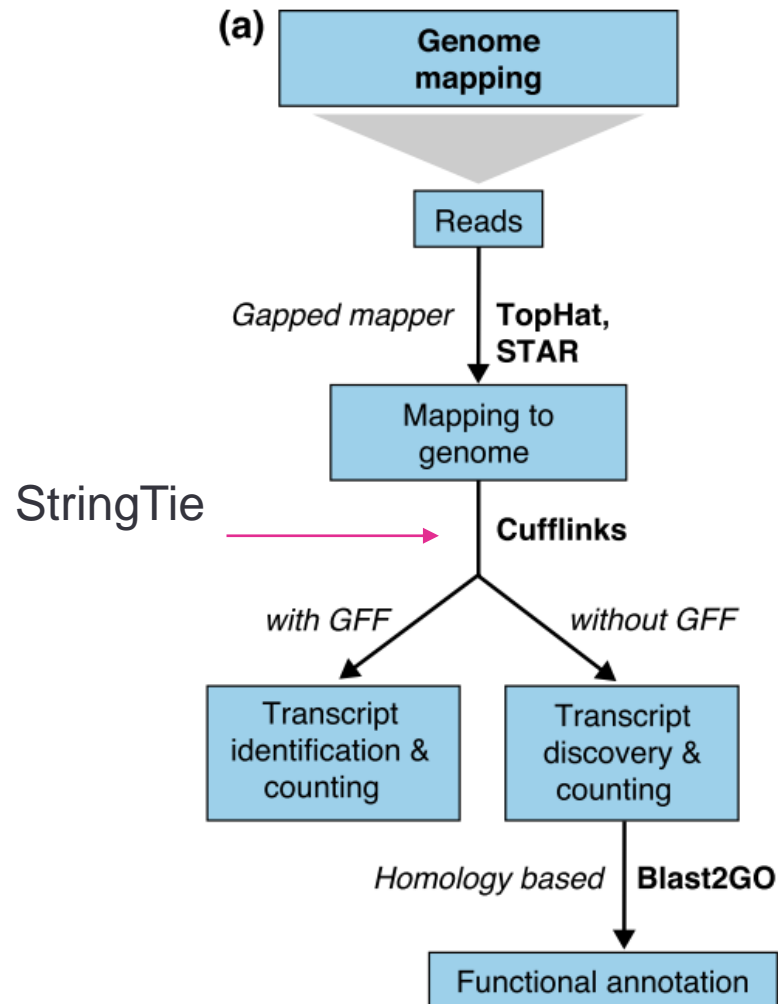


Diversos pipelines

¿Cómo saber qué tipo de algoritmo usar?

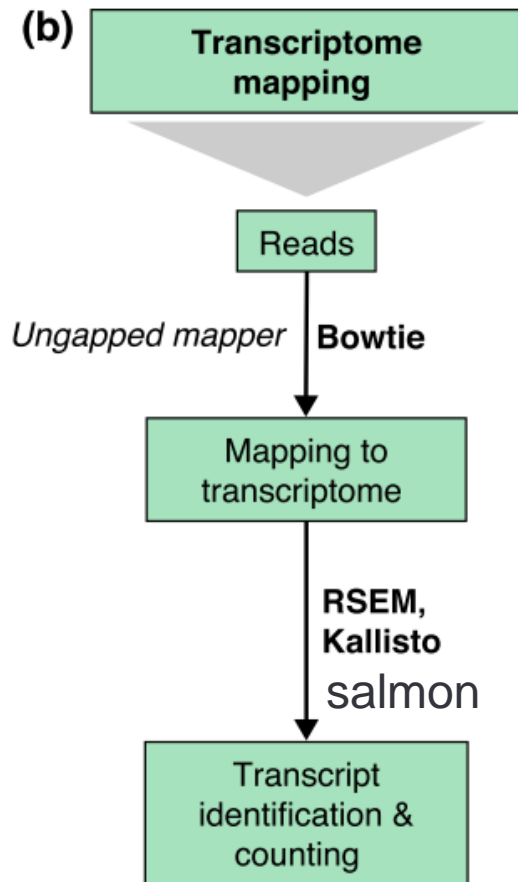


Alineamiento y ensamblaje de lecturas guiado por el genoma de referencia



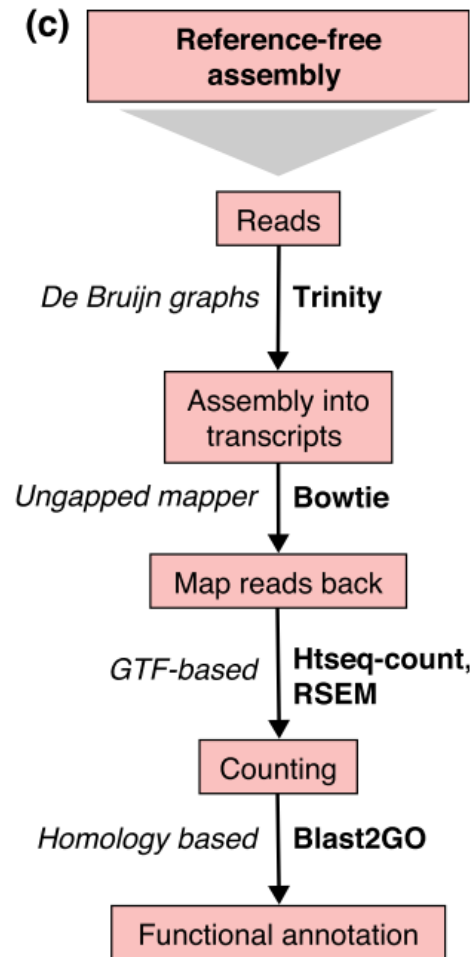
- Podemos anotar nuevos transcritos, así como cuantificarlos.
- Identificación de isoformas.
- Especie con genoma de buena calidad.
- De preferencia contar con un archivo de anotación.
- Empleado normalmente en organismo modelo.

Ensamblaje de transcriptoma guiado



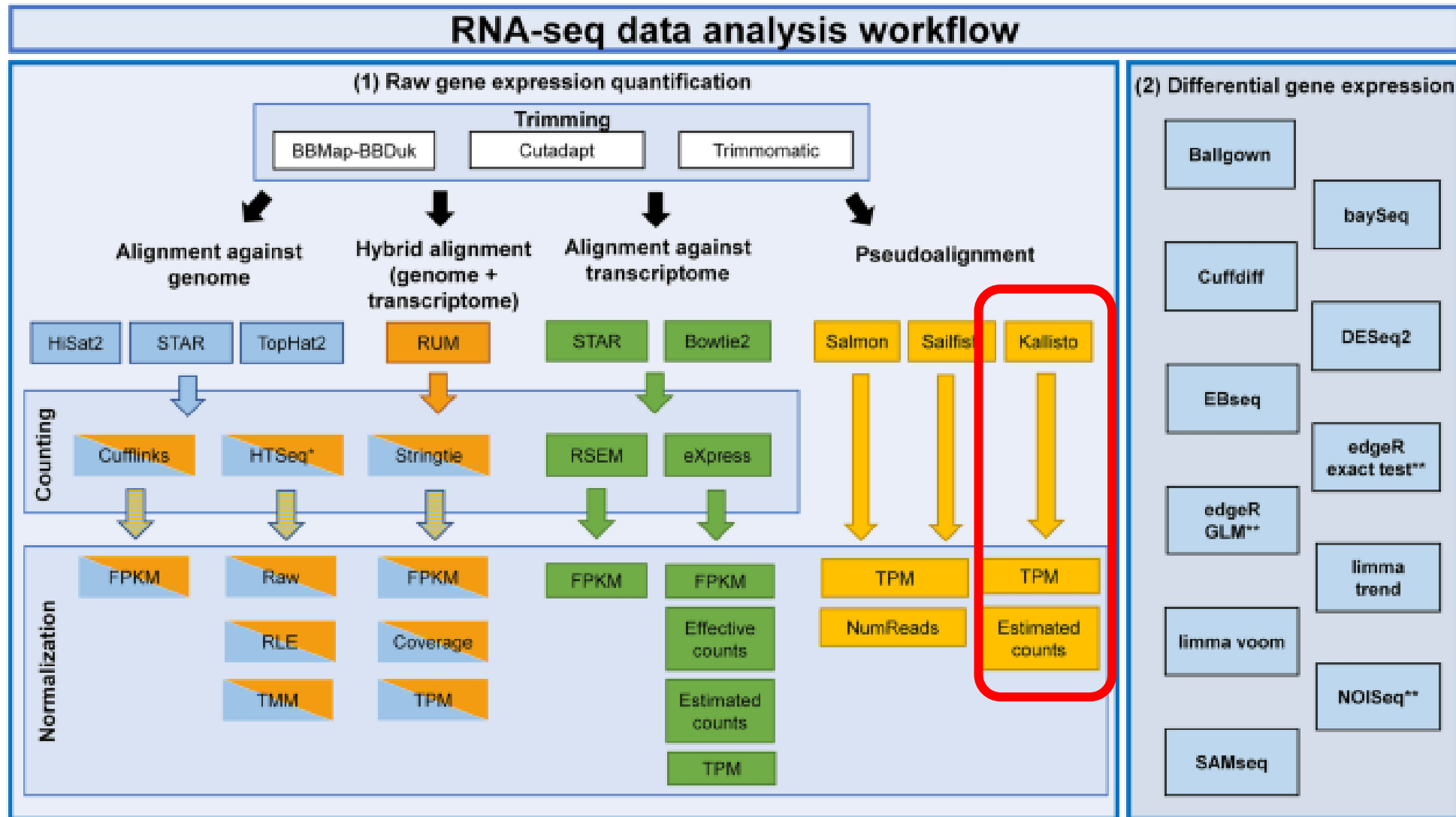
- Expresión por genes, por lo que no vemos isoformas.
- NO hay anotación de nuevos transcritos.
- Si no esta en el archivo de anotación (tx2gene / kallisto) no lo veremos.
- Es necesario un archivo de anotación con buena calidad

Ensamblaje *de novo*



- Especie con genoma de mala calidad.
- Organismo no modelo.
- No contamos con un archivo de anotación de buena calidad.
- Emplear reads tipo PE.

Multiples posibilidades



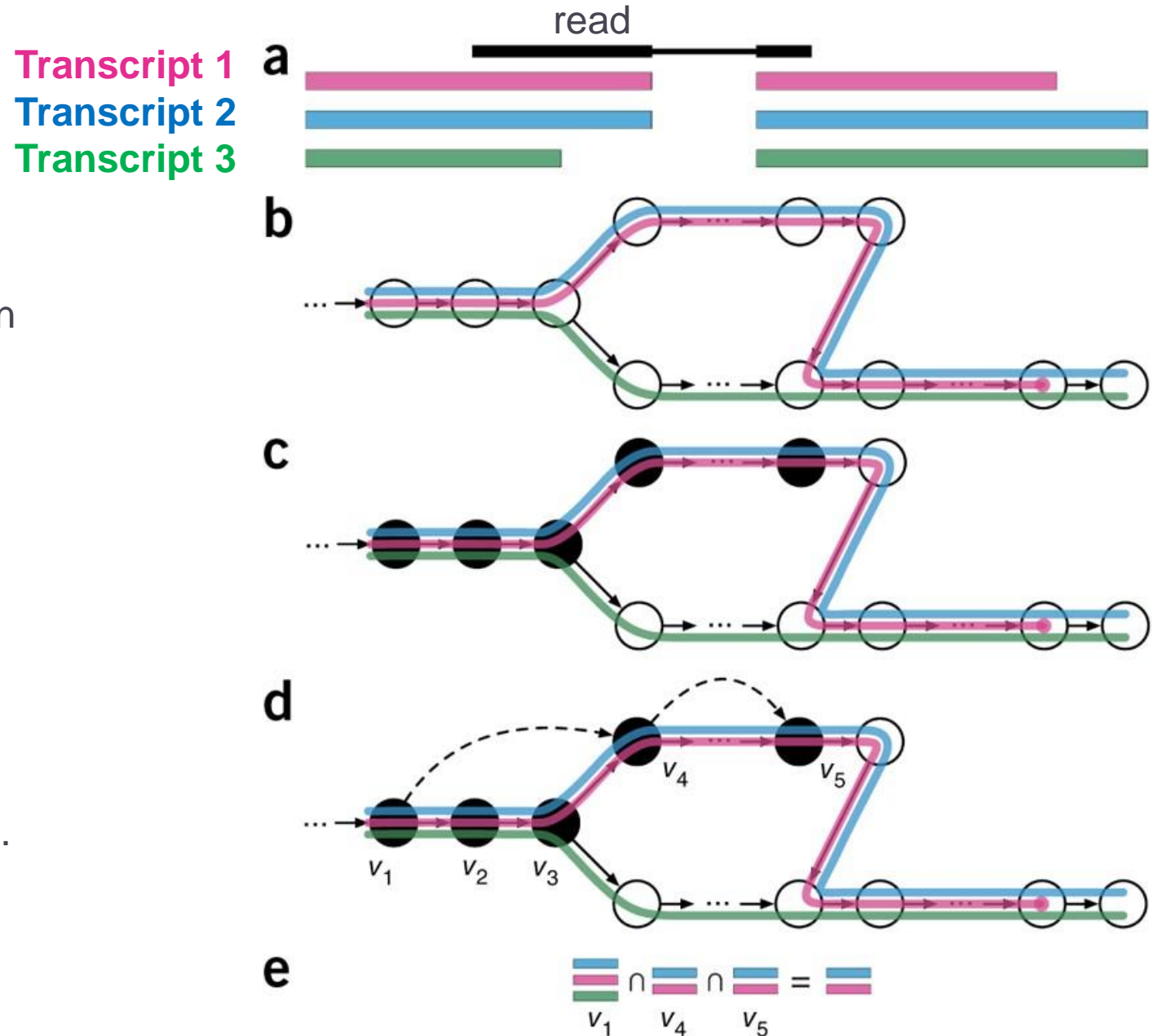
Práctica: Ensamblaje de transcriptoma guiado

Kallisto

Kallisto

- Se basa en la probabilidad de asignación correcta de las lecturas a un transcrito.
 - Pseudoalineamiento.
 - Es rápido.
 - Se puede ejecutar el programa desde tu computadora.
-
- Brujin Graph (T-DBG)
 - Los Nodos (v_1, v_2, v_3) son k -mers
 - Omite pasos redundantes en el T-DBG.

Bray, *et al.* 2016. Nature




Kallisto

Transcriptoma de referencia

2) Generar el index de Kallisto

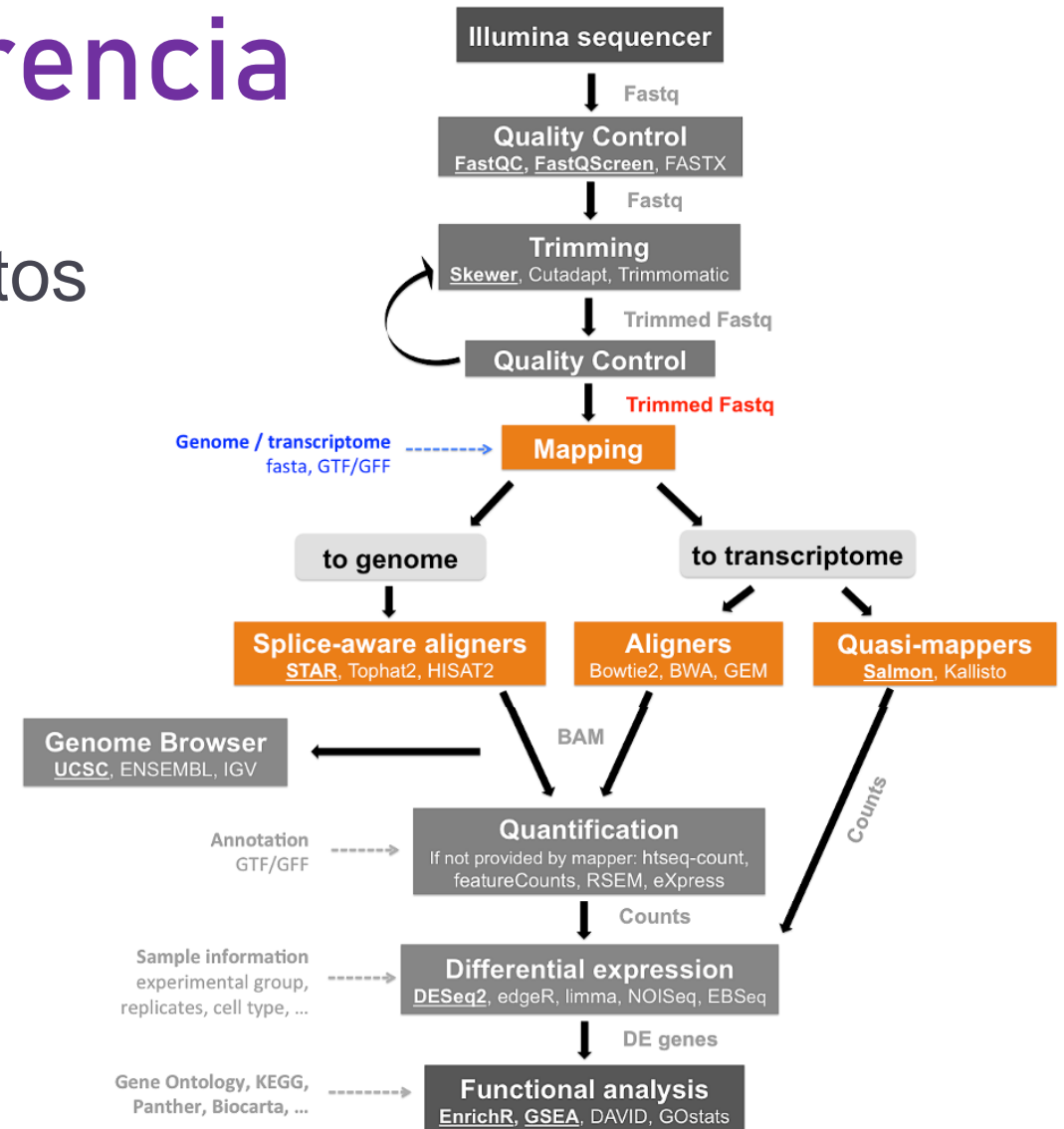
```
mkdir kallisto_quant

# Generar index de kallisto
module load kallisto/0.45.0 # cargar modulo de kallisto
kallisto index -i ./kallisto_quant/At_ref.kidx At_stringm_seq_v2.fasta
```

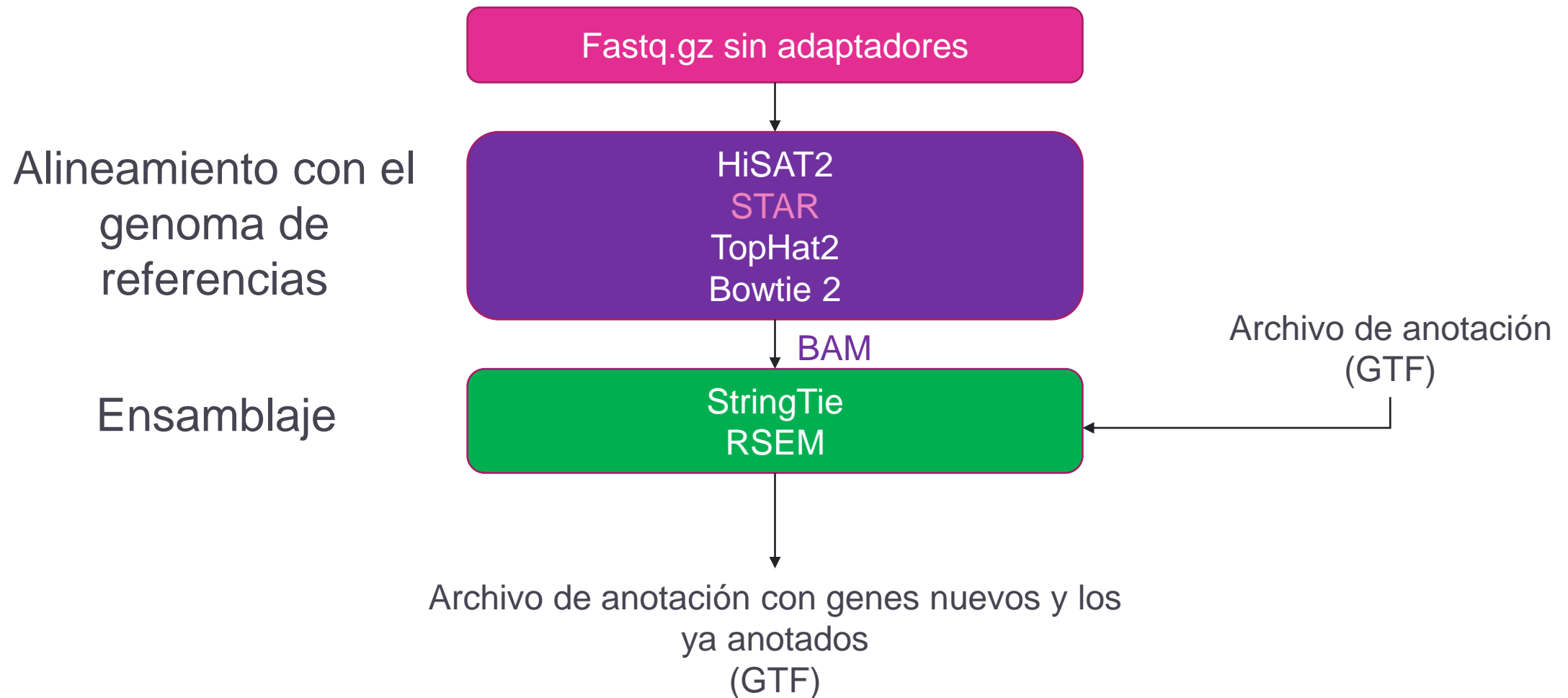


Transcriptoma de referencia

- A) Descargarlo de una base de datos
- B) Generarlo tu mismo



Generar transcriptome de referencia



Continuamos con Kallisto

3) Pseudoalineamiento con Kallisto

```
# Single-end
for file in ./data_trimmed/*.fastq.gz
do
    clean=$(echo $file | sed 's/\.fastq\.gz//')          # Nombre de la carpeta de salida, mismo nombre de SRA
    kallisto quant --index ./kallisto_quant/At_ref.kidx --output-dir $clean --threads 8 $file
done

# Paired-end
for file in ./data_trimmed/*_1.fastq.gz                  # Read1
do
    clean=$(echo $file | sed 's/_1\.fastq\.gz//')        # Nombre de la carpeta de salida, mismo nombre de SRA
    file_2=$(echo ${clean}_2.fastq.gz | sed 's/FP/RP/')  # Read2
    kallisto quant --index ./kallisto_quant/At_ref.kidx --output-dir $clean --threads 8 ${file} ${file_2}
done
```

Práctica 2

- Input
 - fastq.gz

- Github

https://github.com/EveliaCoss/RNAseq_classFEB2023/tree/main/RNA_seq#practica2