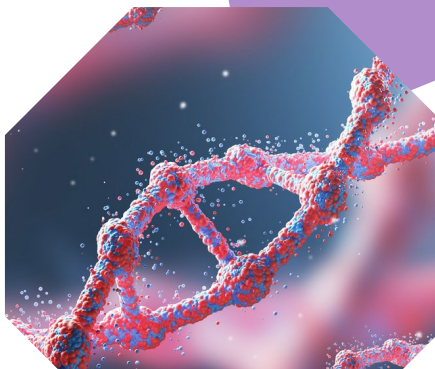
Taller de bioinformática básica y sus aplicaciones en la genómica de especies no modelos

Día 2

Verificación de calidad de genomas

MSc. Eduardo Pizarro G.



1

Uso del clúster

Conexión

2

Conda

Instalación paquetes
Gestión de ambientes

3

Ejecución y transferencia

Scripts
Transferencia datos

4

Verificación de Calidad

FastQC
MultiQC

1. Uso del clúster

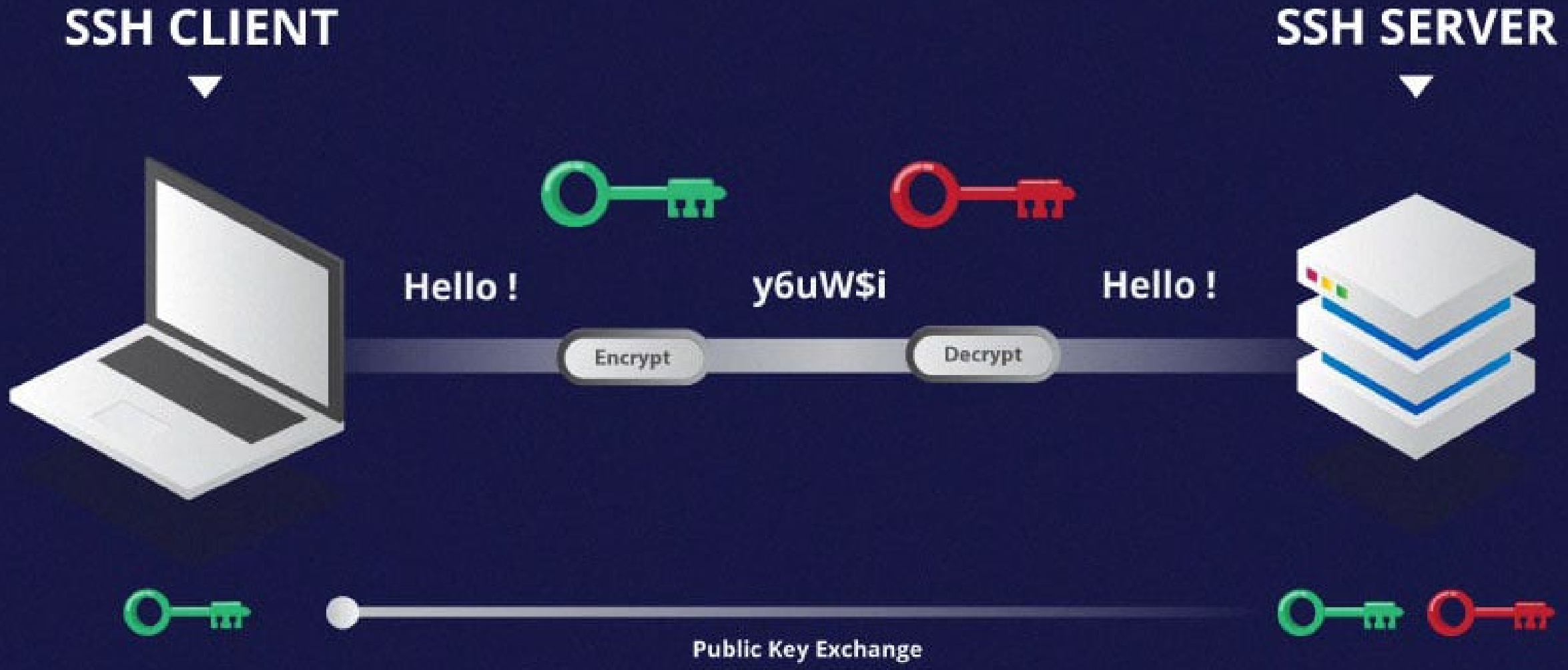
A) Qué es un HPC


High Performance Computing

1. Uso del clúster

A) Qué es un HPC
B) Conexión a un HPC

SSH


SSH CLIENT
SSH SERVER
Hello !
y6uW$i
Hello !
Encrypt
Decrypt
Public Key Exchange

HOSTINGER

Leftraru

- HPC – Computación de alto rendimiento
- Múltiples computadores conectados entre si
- Organización por *Schedulers*

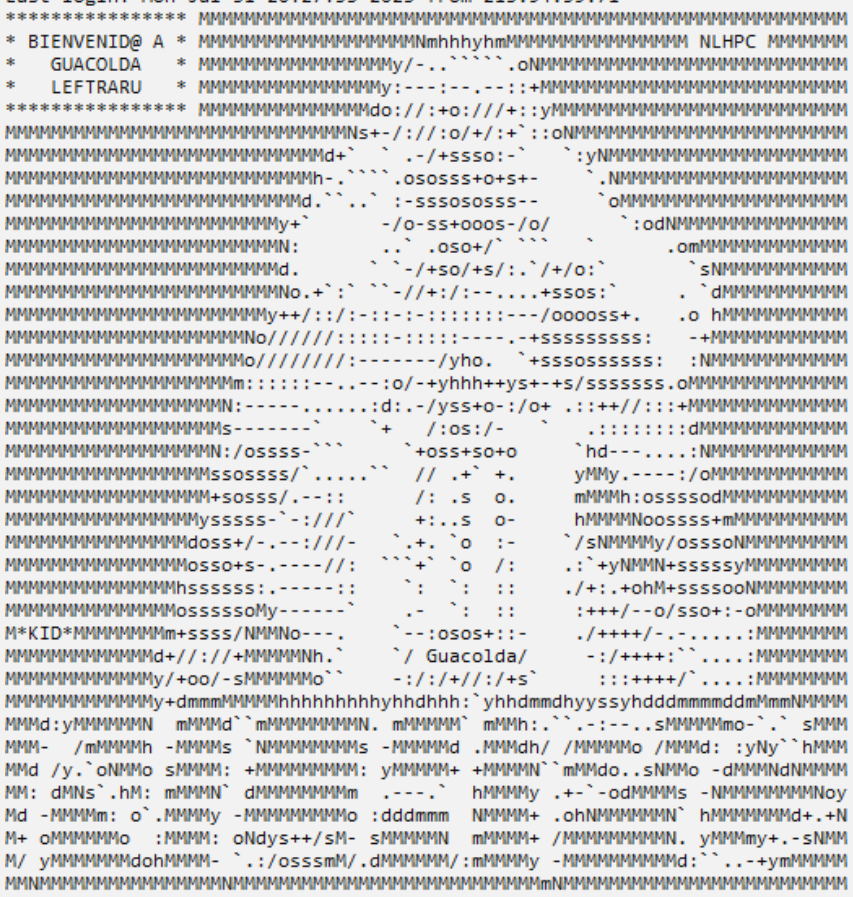
Conexión a clúster Leftraru

- Conexión se realiza al servidor (host) y a un usuario del servidor
- Usuarios disponibles: student88-99
- Comando para conexión:

  $ ssh usuario@host
- Dominio del host: leftraru.nlhpc.cl (podría ser IP)

  $ ssh student88@leftraru.nlhpc.cl
- Contraseña de usuarios: k7sm4wBz

```
(base) eduardo@LAPTOP-IDJV37SH:~$ ssh student88@leftraru.nlhpc.cl
student88@leftraru.nlhpc.cl's password:
Last login: Mon Jul 31 20:27:53 2023 from 213.94.59.71
```



```
Laboratorio Nacional de Computacion de Alto Rendimiento (NLHPC)
Centro de Modelamiento Matematico (CMM)
Universidad de Chile

IMPORTANTE: NO EJECUTAR PROCESOS EN ESTE NODO POR T > 30 min
PARA ESO DEBEN DE USARSE LAS COLAS DE EJECUCION
.................................................................

********************************************************************************
EN CASO DE TENER DUDAS CON SU SCRIPT, LO INVITAMOS A USAR NUESTRO GENERADOR
AUTOMÁTICO EN EL SIGUIENTE LINK: https://wiki.nlhpc.cl/Generador_Scripts
********************************************************************************

********************************************************************************
ESTIMADO USUARIO, A CONTINUACIÓN SE LISTAN LOS NODOS LIBRES PARA SER UTILIZADOS:
********************************************************************************

PARTICION  NODO   ESTADO
debug      3      idle
[student88@leftraru2 ~]$
```

2. Uso de Conda

2. Uso de Conda

- Gestor de paquetes y ambientes.
- Instalación por Miniconda
- Exploremos Conda!

2. Uso de Conda

➢Activar conda: source ~/miniconda3/bin/activate

conda list

conda info --envs

➢Crear un ambiente:

- Crear un ambiente con fastQC: conda create –n secuenciasQC –c bioconda fastqc
- Activar ambiente: conda activate <nombre_ambiente>
- Veamos las opciones que tiene fastqc y comparemos con lo que colocamos en nuestro comando: fastqc --help

2. Uso de Conda

- Gestor de paquetes y ambientes.
- Instalación por Miniconda
- Exploremos Conda!

- ¿Cómo instalo MultiQC en el ambiente de Conda?

3. Ejecución y transferencia

A) Ejecución de programas:
   i. En el prompt

3. Ejecución y transferencia

Ejecución de programas en el Prompt:

A) Revisar donde se encuentran nuestros genomas

$ ls

$ ls ~/genomes

B) Crear directorio para output de FastQC

mkdir ~/genomes/QC_estudianteX

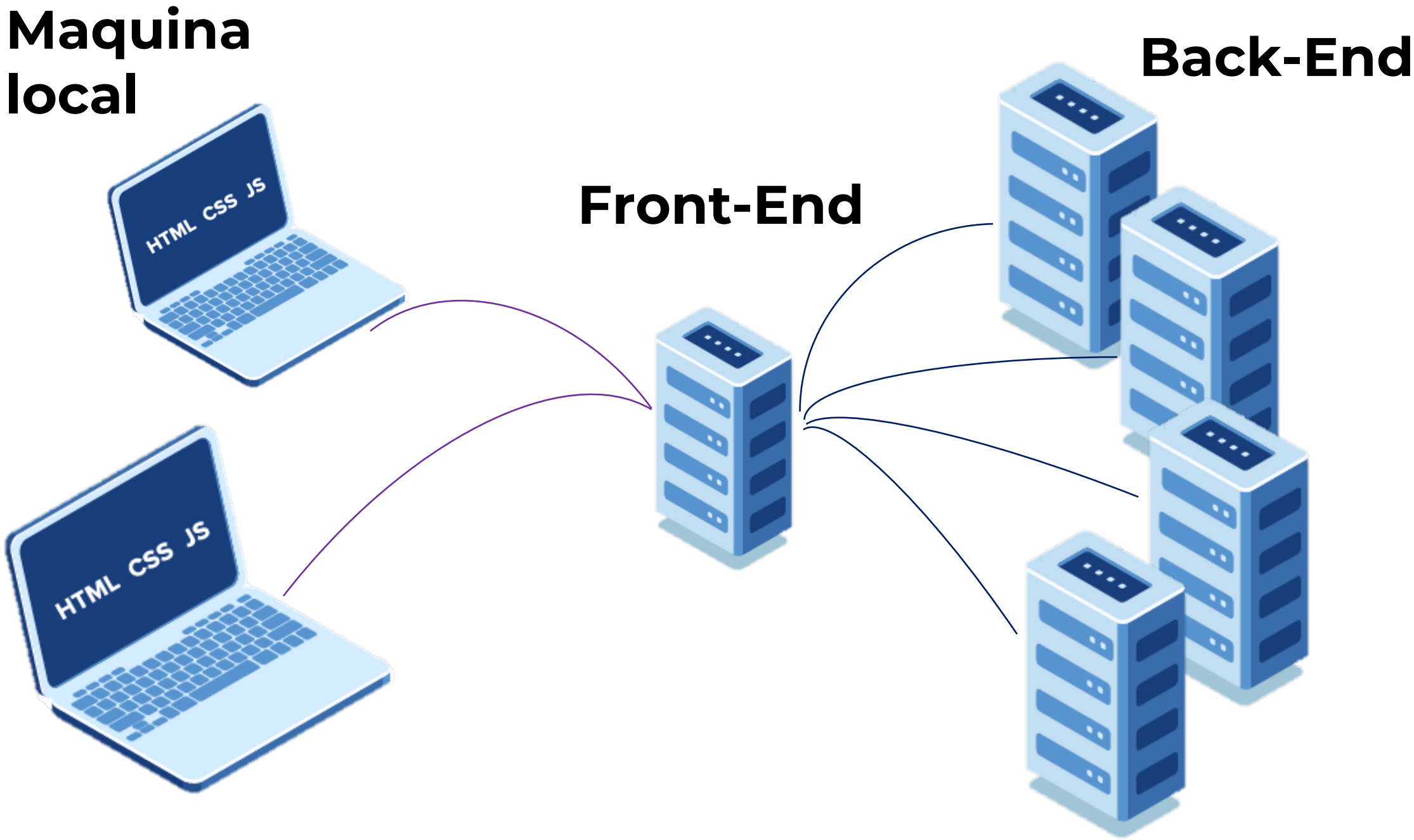
C) Activar ambiente, escribir comando y ejecutar

$ conda activate secuenciasQC

$ fastqc –o ~/genomes/QC_estudianteX –t 1 –f fastq ~/genomes/m2267sub2_R1.fastq.gz

3. Ejecución y transferencia

A) Ejecución de programas:
   i. En el prompt
   ii. Script: ¿qué son? ¿cómo elaborar uno?
      a. Back-End
      b. Front-End


Maquina local
Front-End
Back-End
HTML CSS JS
HTML CSS JS

Scripts para el Back-End

- Scheduler: sistema de gestión de tarea

SLURM

```
#!/bin/bash
#---------------Script SBATCH - NLHPC ----------------
#SBATCH -J FastQC-tunombre
#SBATCH -p slims
#SBATCH --reservation=bioagosto
#SBATCH -n 1
#SBATCH -c 1
#SBATCH --mem-per-cpu=2300
#SBATCH --mail-user=email
#SBATCH --mail-type=ALL
#SBATCH -t 2:2:5
#SBATCH -o FastQC_%j.out
#SBATCH -e FastQC_%j.err

GEN=/home/courses/studentXX/genomes
QC=/home/courses/studentXX/QC

source $HOME/miniconda3/bin/activate
conda activate assembly

fastqc -o $QC/ --noextract -t 1 -f fastq $GEN/*.gz
```

PBS

```
#!/bin/bash
#PBS -V
#PBS -N fastqcp
#PBS -k eo
#PBS -l nodes=1:ppn=40
#PBS -l walltime=40:00:00
################################

source activate preSNPcalling

RAW=/data6/testacc/Eduardo/PUDU/rawdata
QC=/data6/testacc/Eduardo/PUDU/QC
list=/data6/testacc/Eduardo/PUDU/rawdata/list


cd $RAW

while read sample
do
fastqc -o $RAW/fastqc-raw --noextract -t 4 -f fastq $RAW/${sample}.fq &
done < $list

wait
```

Scripts para el Front-End

- Editor de texto (nano):

```
$ nano script1.sh
------- 0 -------
#!/bin/bash
mkdir Prueba
------- 0 -------
$ chmod +x script1.sh

$ bash script1.sh
```

OJO: las personas que estén en un mismo usuario, deben crear un directorio de scripts para cada uno, ingresar al directorio, y crear ahí sus scripts

Ej:
mkdir Elisa
mkdir Fabian
mkdir Eduardo

Scripts para el Front-End

- Ejercicio: Crear un script llamado fqc_m2267_R2.sh para ejecutar en Front-End y que permita ejecutar el comando de FastQC con la muestra m2267sub2_R2.fastq.gz

Pasos:

1. Indicar el intérprete de comando
2. Activar gestor de paquetes y ambientes
3. Activar ambiente
4. Ejecutar programa

Scripts para el Front-End

```
#!/bin/bash

source ~/miniconda3/bin/activate
conda activate secuenciasQC

fastqc –o ~/genomes/QC_estudianteX –t 1 –f fastq ~/genomes/m2267sub2_R2.fastq.gz
```

Guardar el script, convertir en ejecutable, y correr en segundo plano

```
$ chmod +x fqc_m2267_R2.sh
$ ./fqc_m2267_R2.sh &
```

Scripts para el Front-End

- Tercera forma:

Copiaremos el script anterior en uno nuevo para ejecutar FastQC con la muestra m2293:

$ cp fqc_m2267_R2.sh fqc_m2293.sh

Luego modificar con nano para que quede de la siguiente forma:

Scripts para el Front-End

```
#!/bin/bash

source ~/miniconda3/bin/activate
conda activate secuenciasQC

fastqc –o ~/genomes/QC_estudianteX –t 1 –f fastq
~/genomes/m2293sub2_*.fastq.gz
```

Scripts para el Front-End

- Por último, ejecutar con nohup

$ nohup ./fqc_m2293.sh > fqc_m2293.out &

- Podemos monitorear con htop para ver que se está ejecutando
- Al finalizar los procesos, cambiar de ruta a ~/genomes/QC_estudianteX y ejecutar:

$ multiqc .

¿Cómo revisamos nuestro output? Revisar directorio ~/genomes/QC_estudianteX

3. Ejecución y transferencia

A) Ejecución de programas:
   i. En el prompt
   ii. Script: ¿qué son? ¿cómo elaborar uno?
      a. Back-End
      b. Front-End

B) Transferencia de archivos:
   i. Comando rsync

Transferir datos del servidor

- Abrir una nueva terminal en la computadora local, y ejecutar:

rsync -azvrP -e ssh student88@leftraru.nlhpc.cl:/home/courses/student88/genomes/QC_estudianteX .

- Abrir el html con navegador

FastQC

FastQC High Throughput Sequence QC Report
Version: 0.12.1
|www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewels 2011-22,
HTSJDK BAM/SAM reader v2.24.1 from Samtools 2022
BZip decompression ® Matthew J. Francis, 2011
Base64 encoding ® Robert Harder, 2012
Java HDF5 reader ® ETH, CISD and SIS, 2007-14

- FastQC webpage

Documentation

A copy of the FastQC documentation is availat

Example Reports

- Good Illumina Data
- Bad Illumina Data
- Adapter dimer contaminated run
- Small RNA with read-through adapter
- Reduced Representation BS-Seq
- PacBio
- 454

Babraham Institute

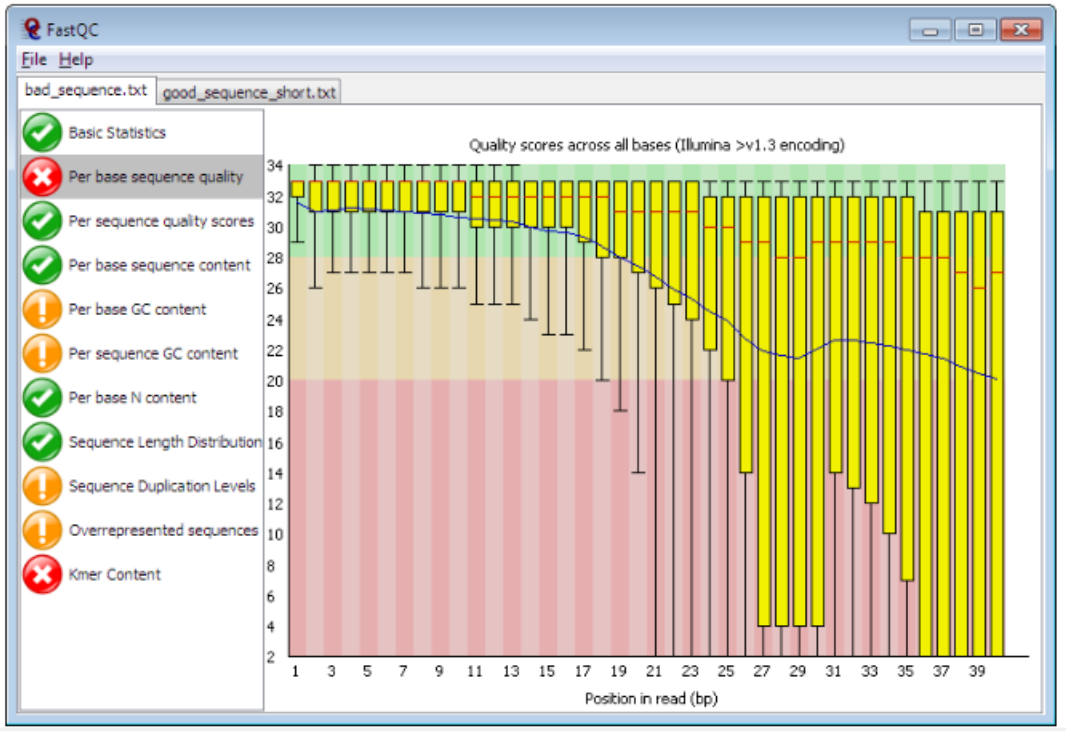
Babraham Bioinformatics

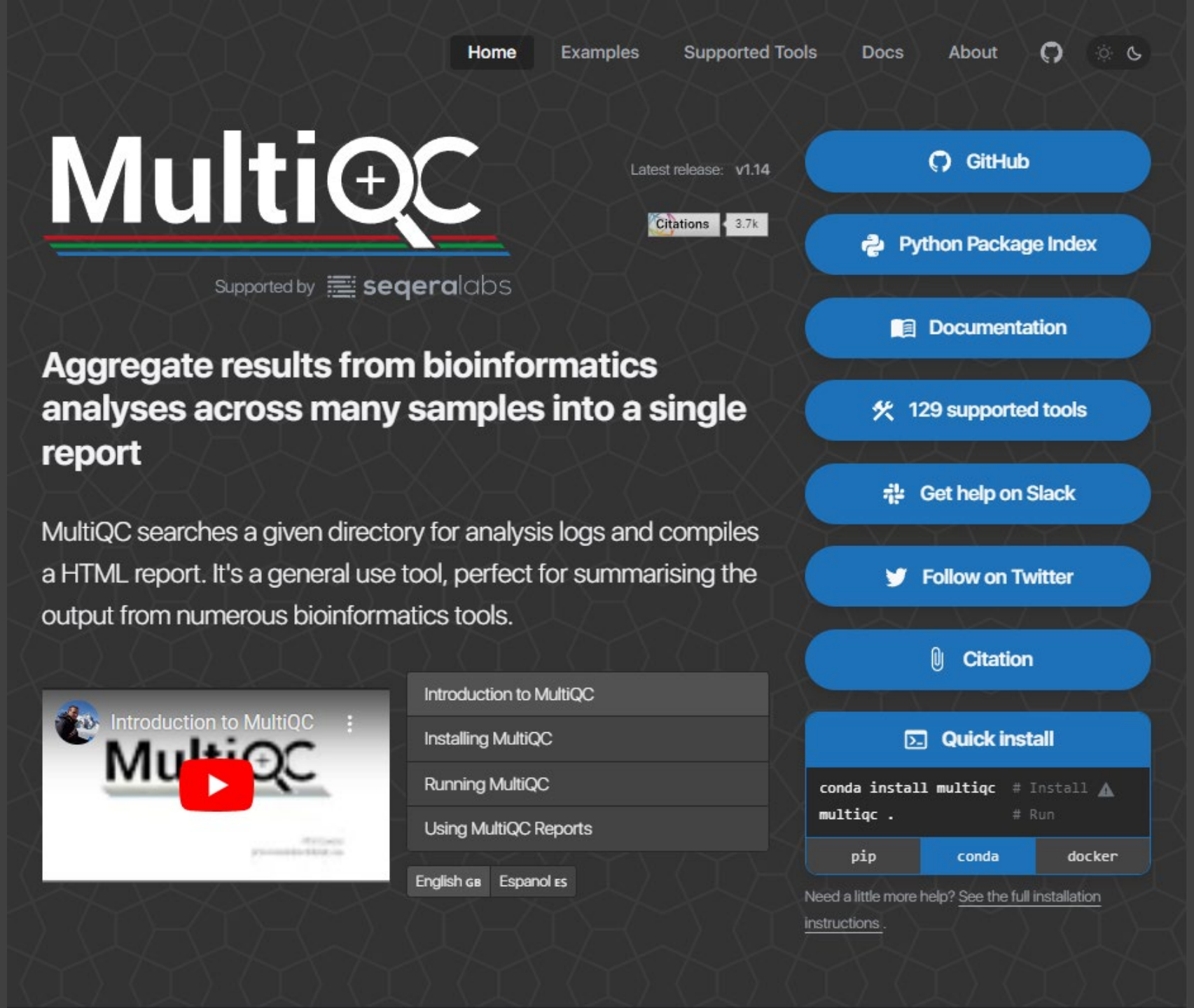
About | People | Services | Projects | Training | Publications

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now





MultiQC webpage

Preguntas a resolver con FastQC-MultiQC

- ¿Cuántos millones de reads tenemos?
- ¿Cuántas secuencias duplicadas?
- ¿Cuál es el tamaño promedio de los reads?
- ¿Poseen restos de adaptadores?
- ¿Qué significan los colores verdes, amarillos y rojos en cada análisis?
- ¿Cuál es la profundidad de secuenciación potencial que debiera obtener?

(Reads totales (R1 + R2) * largo de reads /1.000.000.000 (tamaño GigaBase))/2.4 (tamaño del genoma)

MultiQC

- multiqc_report.html Link

General Statistics

Copy table | Configure Columns | Plot | Showing 4/4 rows and 4/5 columns.

Sample Name	% Dups	% GC	Length	M Seqs
sample1-f	53.2%	44%	233 bp	0.0
sample1-r	55.1%	44%	233 bp	0.0
sample2-f	66.3%	44%	251 bp	0.1
sample2-r	65.7%	44%	251 bp	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 Quality Histograms

1 1 2

The mean quality value across each base position in the read. See the FastQC help.

Y-Limits: on

Phred quality score

LINK

```
+SEQ_ID
!''*(((((***+))%%%++)(%%%%).1***
```

A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

Taller de bioinformática básica y sus aplicaciones en la genómica de especies no modelos

Día 2

Verificación de calidad de genomas

MSc. Eduardo Pizarro G.