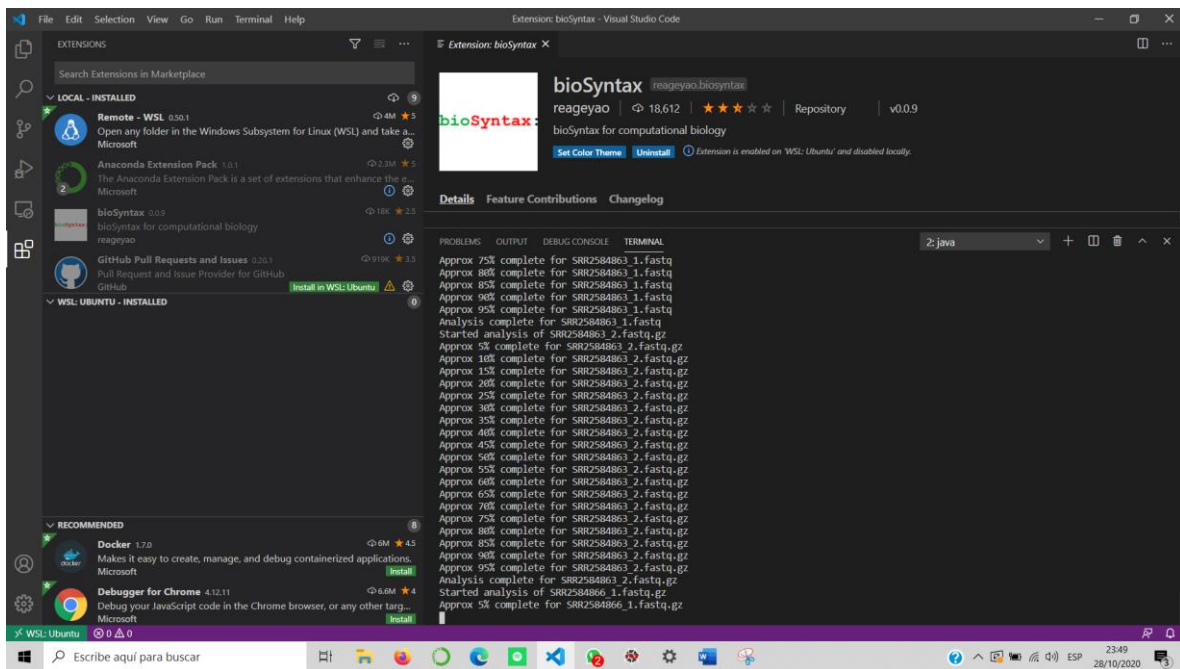# Data Wrangling and Processing for Genomics - Parte 2

Evidencia:


```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data/untrimmed_fastq$ cp ~/.backup/untrimmed_fastq/*fastq.gz .
cp: cannot stat '/home/estuardo8u14/.backup/untrimmed_fastq/*fastq.gz': No such file or directory
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data/untrimmed_fastq$ head -n 4 SRR2584863_1.fastq
@SRR2584863.1 HWI-ST957:244:H73TDADXX:1:1101:4712:2181/1
TTCACATCCTGACCATTCAGTTGAGCAAAATAGTTCTTCAGTGCCTGTTTAACCGAGTCACGCAGGGGTTTTTGGGTTACCTGATCCTGAGAGTTAACGGTAGAAACGGTCAGTAC
GTCAGAATTTACGCGTTGTTCGAACATAGTTCTG
+
CCCFFFFFGHHHHJIJJJJIJJJIITJJJJIIIJJGFIIIJEDDFEGGJIFHHJIJJDECCGGEGIIJFHFFFACD:BBBDDACCCCAA@@CA@C>C3>@5(8&>C:9?8+89<4(:
83825C(:A#######################
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data/untrimmed_fastq$ []
```

```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_woestuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data/untrimmed_
fastq$ tail -n
 4 SRR2584863_1.fastq
@SRR2584863.1553259 HWI-ST957:245:H73R4ADXX:2:2216:21048:100894/1
CTGCAATACCACGCTGATCTTTCACATGATGTAAGAAAAGTGGGATCAGCAAACCGGGTGCTGCTGTGGCTAGTTGCAGCAAACCATGCAGTGAACCCGCCTGTGCTTCGCTATAGCCGTGACTGATGAGGA
TCGCCGGAAGCCAGCCAA
+
CCCFFFFFHHHHGJJJJJJJJHGIJJJIJJJJIJJJJIIITJJJJJJJJJJJJJJIITJJJHHHHHFFFFFEEEEEDDDDDDDDDDDDDDDDDCDEDDBDBDDBDDDDDDDDDBDEEDDDD7@BDDDDDD>AA
>?B?<@BDD@BDC?BDA?
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data/untrimmed_fastq$ 
```

```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data/untrimmed_fastq$ ls -l -h
total 1.5G
-rw-r--r-- 1 estuardo8u14 estuardo8u14 545M Oct 28 23:13 SRR2584863_1.fastq
-rw-r--r-- 1 estuardo8u14 estuardo8u14 183M Oct 28 23:28 SRR2584863_2.fastq.gz
-rw-r--r-- 1 estuardo8u14 estuardo8u14 309M Oct 28 23:32 SRR2584866_1.fastq.gz
-rw-r--r-- 1 estuardo8u14 estuardo8u14 296M Oct 28 23:37 SRR2584866_2.fastq.gz
-rw-r--r-- 1 estuardo8u14 estuardo8u14 124M Oct 28 23:03 SRR2589044_1.fastq.gz
```

UVG
Genómica computacional
28/10/2020

Unzip

UVG
Genómica computacional
28/10/2020

```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop$ cd ~/dc_workshop/results/fastqc_untrimmed_reads/
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls
SRR2584863_1_fastqc.html   SRR2584866_1_fastqc.zip
SRR2584863_1_fastqc.zip    SRR2584866_2_fastqc.html
SRR2584863_2_fastqc.html   SRR2584866_2_fastqc.zip
SRR2584863_2_fastqc.zip    SRR2589044_1_fastqc.html
SRR2584866_1_fastqc.html   SRR2589044_1_fastqc.zip
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ unzip *.zip
Archive:  SRR2584863_1_fastqc.zip
caution: filename not matched:  SRR2584863_2_fastqc.zip
caution: filename not matched:  SRR2584866_1_fastqc.zip
caution: filename not matched:  SRR2584866_2_fastqc.zip
caution: filename not matched:  SRR2589044_1_fastqc.zip
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ for filename in *.zip
> do
> unzip $filename
> done
Archive:  SRR2584863_1_fastqc.zip
   creating: SRR2584863_1_fastqc/
   creating: SRR2584863_1_fastqc/Icons/
   creating: SRR2584863_1_fastqc/Images/
  inflating: SRR2584863_1_fastqc/Icons/fastqc_icon.png
  inflating: SRR2584863_1_fastqc/Icons/warning.png
  inflating: SRR2584863_1_fastqc/Icons/error.png
  inflating: SRR2584863_1_fastqc/Icons/tick.png
  inflating: SRR2584863_1_fastqc/summary.txt
  inflating: SRR2584863_1_fastqc/Images/per_base_quality.png
  inflating: SRR2584863_1_fastqc/Images/per_tile_quality.png
  inflating: SRR2584863_1_fastqc/Images/per_sequence_quality.png
  inflating: SRR2584863_1_fastqc/Images/per_base_sequence_content.png
  inflating: SRR2584863_1_fastqc/Images/per_sequence_gc_content.png
  inflating: SRR2584863_1_fastqc/Images/per_base_n_content.png
  inflating: SRR2584863_1_fastqc/Images/sequence_length_distribution.png
  inflating: SRR2584863_1_fastqc/Images/duplication_levels.png
  inflating: SRR2584863_1_fastqc/Images/adapter_content.png
  inflating: SRR2584863_1_fastqc/fastqc_report.html
  inflating: SRR2584863_1_fastqc/fastqc_data.txt
  inflating: SRR2584863_1_fastqc/fastqc.fo
Archive:  SRR2584863_2_fastqc.zip
   creating: SRR2584863_2_fastqc/
   creating: SRR2584863_2_fastqc/Icons/
   creating: SRR2584863_2_fastqc/Images/
  inflating: SRR2584863_2_fastqc/Icons/fastqc_icon.png
  inflating: SRR2584863_2_fastqc/Icons/warning.png
  inflating: SRR2584863_2_fastqc/Icons/error.png
```

UVG
Genómica computacional
28/10/2020

```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls -F
SRR2584863_1_fastqc/      SRR2584866_1_fastqc.zip
SRR2584863_1_fastqc.html  SRR2584866_2_fastqc/
SRR2584863_1_fastqc.zip   SRR2584866_2_fastqc.html
SRR2584863_2_fastqc/      SRR2584866_2_fastqc.zip
SRR2584863_2_fastqc.html  SRR2589044_1_fastqc/
SRR2584863_2_fastqc.zip   SRR2589044_1_fastqc.html
SRR2584866_1_fastqc/      SRR2589044_1_fastqc.zip
SRR2584866_1_fastqc.html
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls -F SRR2584863_1_fastqc/
Icons/    fastqc.fo          fastqc_report.html
Images/   fastqc_data.txt  summary.txt
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ less SRR2584863_1_fastqc/summary.txt
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ cat */summary.txt > ~/dc_workshop/docs/fastqc_summaries.txt
bash: /home/estuardo8u14/dc_workshop/docs/fastqc_summaries.txt: No such file or directory
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls
SRR2584863_1_fastqc      SRR2584866_1_fastqc.zip
SRR2584863_1_fastqc.html  SRR2584866_2_fastqc
SRR2584863_1_fastqc.zip   SRR2584866_2_fastqc.html
SRR2584863_2_fastqc      SRR2584866_2_fastqc.zip
SRR2584863_2_fastqc.html  SRR2589044_1_fastqc
SRR2584863_2_fastqc.zip   SRR2589044_1_fastqc.html
SRR2584866_1_fastqc      SRR2589044_1_fastqc.zip
SRR2584866_1_fastqc.html
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ cd ..
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results$ cd ..
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop$ ls
data   results
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop$ cd data
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data$ ls
untrimmed_fastq
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/data$ cd ..
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop$ cd results
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results$ ls
fastqc_untrimmed_reads
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results$ cd fastqc_untrimmed_reads
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls
SRR2584863_1_fastqc      SRR2584866_1_fastqc.zip
SRR2584863_1_fastqc.html  SRR2584866_2_fastqc
SRR2584863_1_fastqc.zip   SRR2584866_2_fastqc.html
SRR2584863_2_fastqc      SRR2584866_2_fastqc.zip
SRR2584863_2_fastqc.html  SRR2589044_1_fastqc
SRR2584863_2_fastqc.zip   SRR2589044_1_fastqc.html
SRR2584866_1_fastqc      SRR2589044_1_fastqc.zip
SRR2584866_1_fastqc.html
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls -F SRR2584863_1_fastqc/
```

Las muestras no superaron al menos una de las pruebas de calidad de FastQC:

```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ ls -F SRR2584863_1_fastqc/
Icons/    fastqc.fo          fastqc_report.html
Images/   fastqc_data.txt  summary.txt
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ cat */summary.txt > ~/dc_workshop/docs/fastqc_summaries.txt
bash: /home/estuardo8u14/dc_workshop/docs/fastqc_summaries.txt: No such file or directory
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ mkdir ~/dc_workshop/docs
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ cat */summary.txt > ~/dc_workshop/docs/fastqc_summaries.txt
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/results/fastqc_untrimmed_reads$ cd ~/dc_workshop/docs
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/docs$ grep FAIL fastqc_summaries.txt
FAIL    Per base sequence quality       SRR2584863_2.fastq.gz
FAIL    Per tile sequence quality       SRR2584863_2.fastq.gz
FAIL    Per base sequence content       SRR2584863_2.fastq.gz
FAIL    Per base sequence quality       SRR2584866_1.fastq.gz
FAIL    Per base sequence content       SRR2584866_1.fastq.gz
FAIL    Adapter Content SRR2584866_1.fastq.gz
FAIL    Adapter Content SRR2584866_2.fastq.gz
FAIL    Adapter Content SRR2589044_1.fastq.gz
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/docs$
```