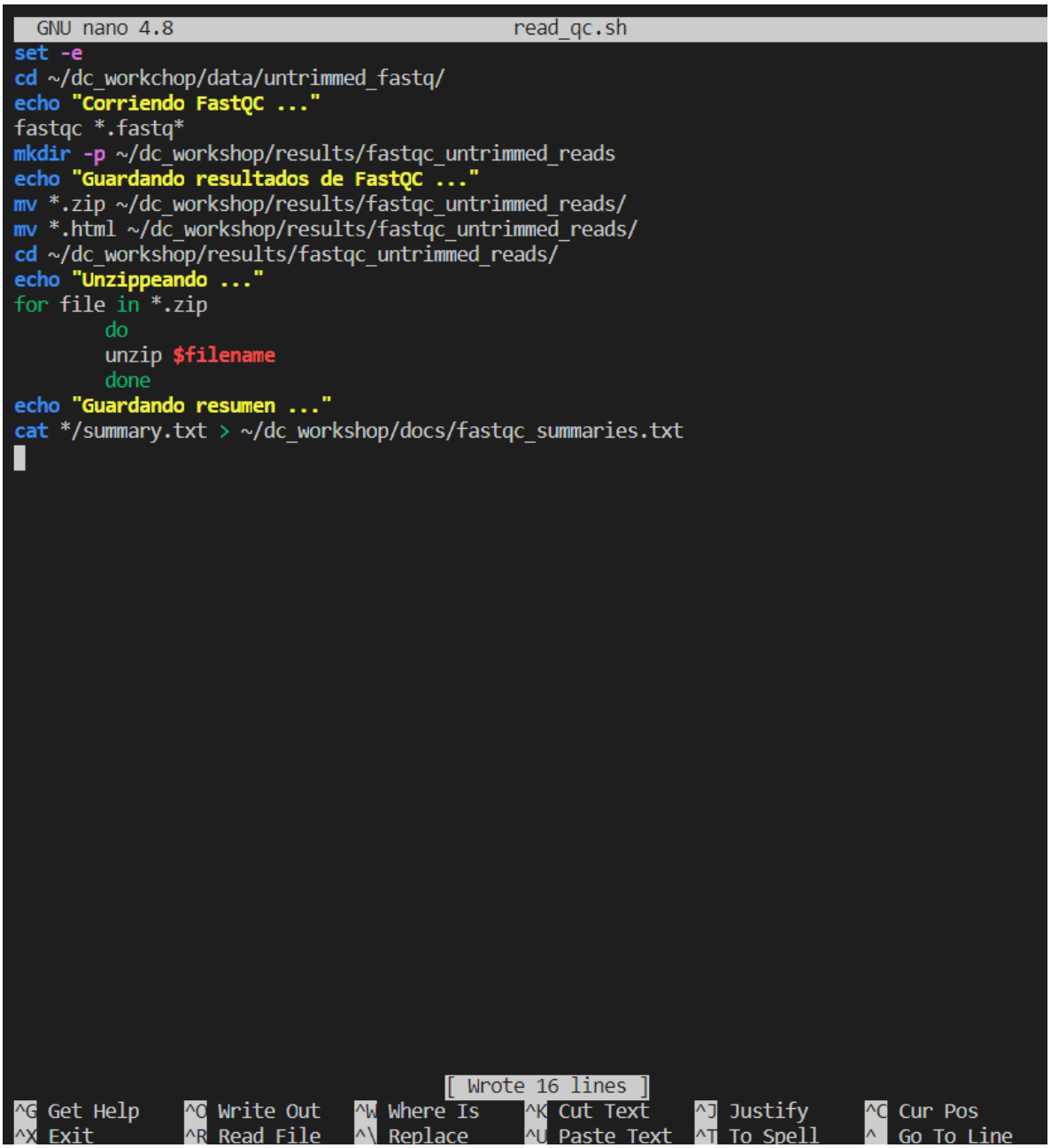
Data Wrangling and Processing for Genomics - Parte 5

Evidencia:

Script fastq:



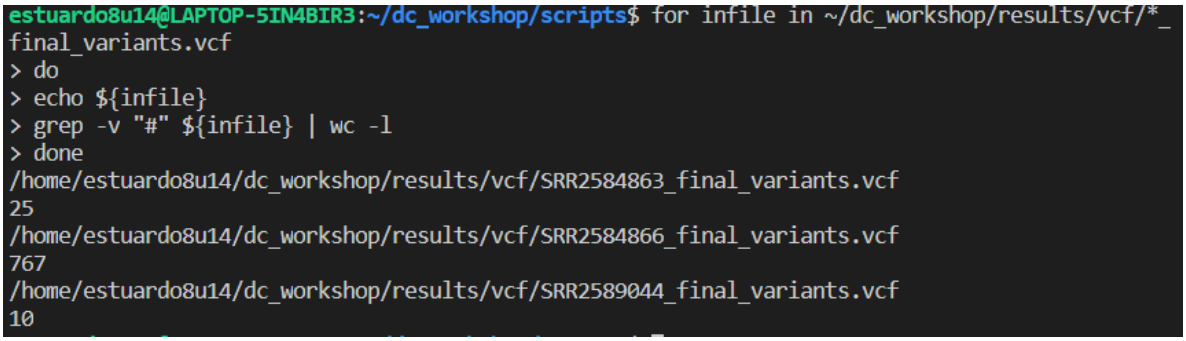
```
estuardo8u14@LAPTOP-5IN4BIR3:~/dc_workshop/scripts$ bash run_variant_calling.sh
[bwa_index] Pack FASTA... 0.08 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 1.63 seconds elapse.
[bwa_index] Update BWT... 0.03 sec
[bwa_index] Pack forward-only FASTA... 0.04 sec
[bwa_index] Construct SA from BWT and Occ... 0.75 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index /home/estuardo8u14/dc_workshop/data/ref_genome/ecoli_rel606.fasta
[main] Real time: 3.688 sec; CPU: 2.538 sec
working with file /home/estuardo8u14/dc_workshop/data/trimmed_fastq_small/SRR2584863_1.trim.su
b.fastq
base name is SRR2584863
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 78970 sequences (10000278 bp)...
```

Sprint samtools bfctools y bwa:

CON COMMENTS

```
  GNU nano 4.8                          run_cariant_calling.sh                          Modified
#indexar nuestro genoma de referencia para BWA
bwa index $genome

mkdir -p sam bam bcf vcf

#asigne el nombre del archivo FASTQ con el que estamos trabajando actualmente a una variable >
#decirle a la secuencia de comandos que nos devuelva el nombre del archivo paraque podamos ve>
for fq1 in ~/dc_workshop/data/trimmed_fastq_small/*_1.trim.sub.fastq
    do
    echo "working with file $fq1"

    base=$(basename $fq1 _1.trim.sub.fastq)
    echo "base name is $base"
#We can use the base variable to access both the base_1.fastq and base_2.fastq i
    fq1=~/dc_workshop/data/trimmed_fastq_small/${base}_1.trim.sub.fastq
    fq2=~/dc_workshop/data/trimmed_fastq_small/${base}_2.trim.sub.fastq
    sam=~/dc_workshop/results/sam/${base}.aligned.sam
    bam=~/dc_workshop/results/bam/${base}.aligned.bam
    sorted_bam=~/dc_workshop/results/bam/${base}.aligned.sorted.bam
    raw_bcf=~/dc_workshop/results/bcf/${base}_raw.bcf
    variants=~/dc_workshop/results/bcf/${base}_variants.vcf
    final_variants=~/dc_workshop/results/vcf/${base}_final_variants.vcf

    bwa mem $genome $fq1 $fq2 > $sam
    samtools view -S -b $sam > $bam
    samtools sort -o $sorted_bam $bam
    samtools index $sorted_bam
    bcftools mpileup -O b -o $raw_bcf -f $genome $sorted_bam
    bcftools call --ploidy 1 -m -v -o $variants $raw_bcf
    vcfutils.pl varFilter $variants > $final_variants

    done
#resumen pasos:
#alinee las lecturas con el genoma de referencia y genere un archivo .sam:
#convertir el archivo SAM a formato BAM
#ordenar el archivo BAM
#indexar el archivo BAM con fines de visualización
#calcular la cobertura de lectura de posiciones en el genoma
#llamar a SNP con bcftools
#filtrar e informar las variantes de SNP en formato de llamada de variante (VCF)


^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Paste Text ^T To Spell ^_ Go To Line
```

How did the number of mutations per sample change over time? Examine the metadata table. What is one reason the number of mutations may have changed the way they did?



Recordar:

Podemos combinar varios comandos en un script de shell para automatizar un flujo de trabajo.

Use declaraciones de echo dentro de sus scripts para obtener una actualización de progreso automatizada.