

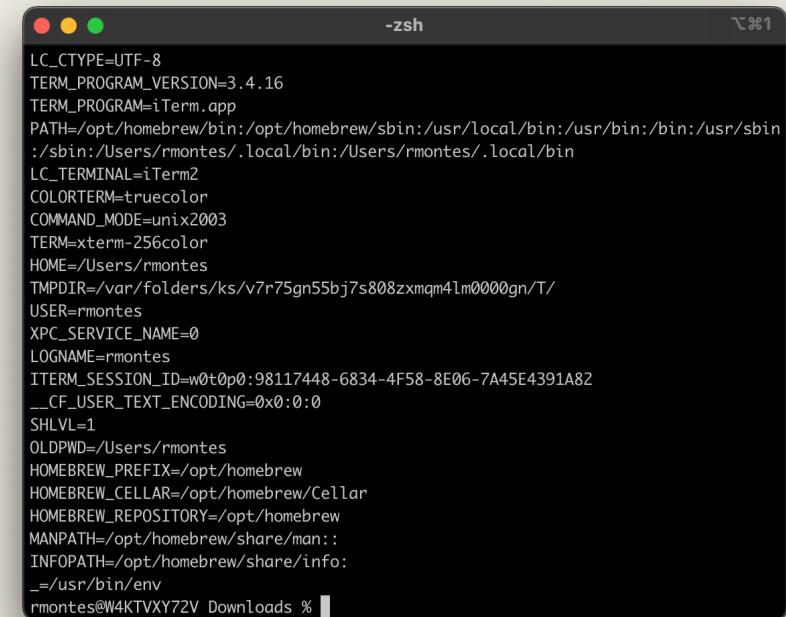
# INTRODUCCIÓN A LA SECUENCIACIÓN DE NUEVA GENERACIÓN



Robinson Montes Gómez  
Bioingeniero

# Contenido

- Módulo 1: Manejo y familiarización con la terminal
  - Qué es *la shell*?
  - Comandos básicos: *whoami, cal, date, pwd, echo, ls*
  - Edición de archivos y carpetas: *cd, touch, >, mkdir*
  - Copiar, mover, y eliminar ficheros: *rm, rmdir, cp, mv*
  - Comandos para búsqueda: *which, find, grep*
  - Comprimir/Descomprimir archivos: *tar, zip, unzip*
  - Otros comandos
- Módulo 2: Crear un pipeline de NGS
  - Qué es Secuenciación de nueva Generación (NGS)
  - Antecedentes
  - Obtener las secuencias
  - Procesamiento: *Alineamiento, Preprocesamiento, Llamado de variantes, Postprocesamiento, Anotación de variantes*

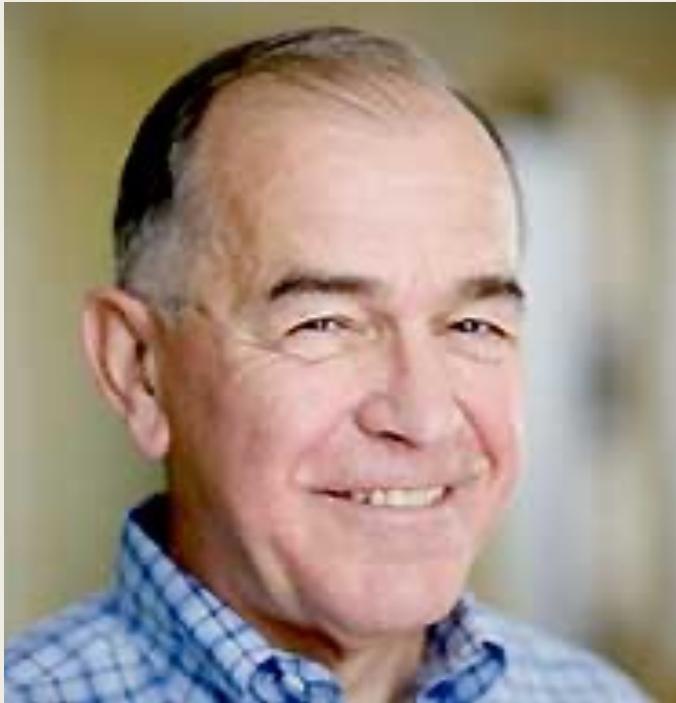


```
LC_CTYPE=UTF-8
TERM_PROGRAM_VERSION=3.4.16
TERM_PROGRAM=iTerm.app
PATH=/opt/homebrew/bin:/opt/homebrew/sbin:/usr/local/bin:/usr/bin:/bin:/usr/sbin
:/sbin:/Users/rmontes/.local/bin:/Users/rmontes/.local/bin
LC_TERMINAL=iTerm2
COLORTERM=truecolor
COMMAND_MODE=unix2003
TERM=xterm-256color
HOME=/Users/rmontes
TMPDIR=/var/folders/ks/v7r75gn55bj7s808zxmqm41m0000gn/T/
USER=rmontes
XPC_SERVICE_NAME=0
LOGNAME=rmontes
ITERM_SESSION_ID=w0t0p0:98117448-6834-4F58-8E06-7A45E4391A82
__CF_USER_TEXT_ENCODING=0x0:0:0
SHLVL=1
OLDPWD=/Users/rmontes
HOMEBREW_PREFIX=/opt/homebrew
HOMEBREW_CELLAR=/opt/homebrew/Cellar
HOMEBREW_REPOSITORY=/opt/homebrew
MANPATH=/opt/homebrew/share/man:::
INFOPATH=/opt/homebrew/share/info:
_=~/usr/bin/env
rmontes@W4KTVXY72V Downloads %
```

# Módulo 1:

# Manejo y familiarización con la terminal

# Qué es la Shell?



Steve Bourne

- Un shell es un intérprete de comandos que expone una interfaz para que el usuario trabaje con el sistema operativo subyacente.
- Permite ejecutar operaciones usando texto y comandos, y proporciona a los usuarios características avanzadas como la posibilidad de crear scripts.
- Esto es importante: los shells te permiten realizar cosas de una manera más optimizada de lo que una GUI (Graphical User Interface o Interfaz Gráfica de Usuario) podría permitirte hacer. Las herramientas de línea de comandos pueden ofrecer muchas opciones de configuración diferentes sin ser demasiado complejas de usar.

# CLI

## *Command Line Interface*

Interacción mediante texto.  
Se basa en el uso de un lenguaje codificado.

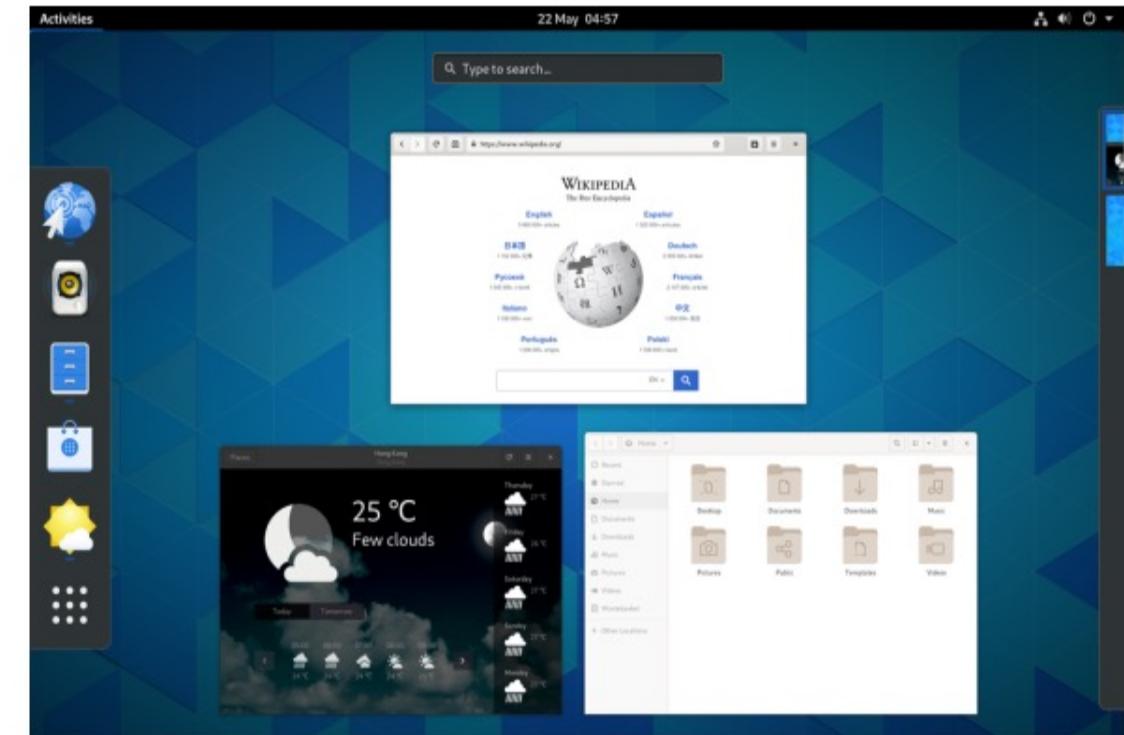
```
test@test-VirtualBox:~$ uname -a
Linux test-VirtualBox 4.15.0-58-generic #64-Ubuntu SMP Tue Aug 6 11:12:41
UTC 2019 x86_64 x86_64 x86_64 GNU/Linux
test@test-VirtualBox:~$ date
vie ago 30 20:09:10 CEST 2019
test@test-VirtualBox:~$ cal
    Agosto 2019
do lu ma mi ju vi sa
      1  2  3
 4  5  6  7  8  9 10
11 12 13 14 15 16 17
18 19 20 21 22 23 24
25 26 27 28 29 30 31

test@test-VirtualBox:~$ who
test    tty2        2019-08-30 19:26 (tty2)
test@test-VirtualBox:~$
```

# GUI

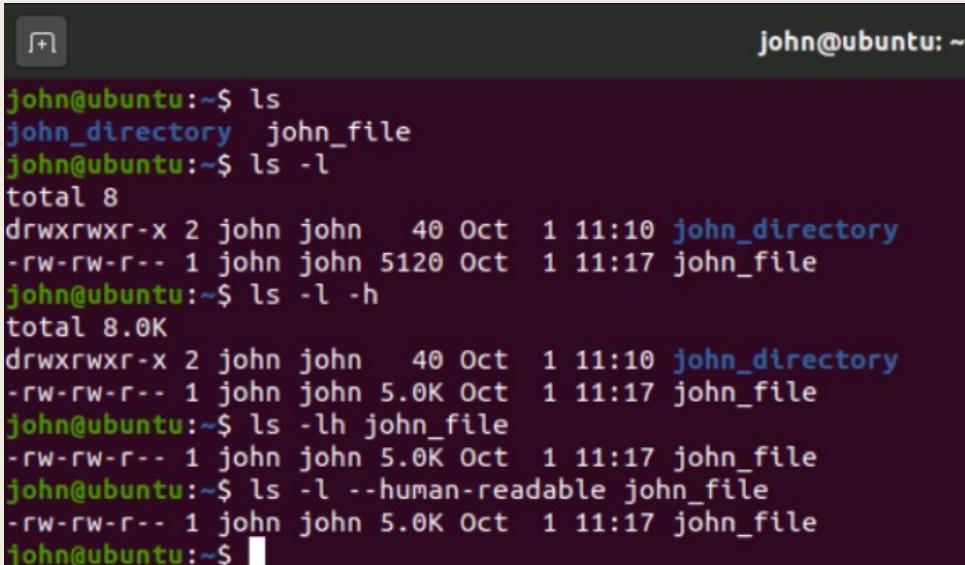
## *Graphical User Interface*

Interacción a través de elementos gráficos.  
Posibilita una interacción amigable e intuitiva.



# Comandos Básicos

- Prompt: Es el carácter o conjunto de caracteres que se muestran en una linea de comandos para indicar que está a la espera de órdenes. \$ >
- whoami: Muestra el nombre de usuario
- cal: Muestra el calendario
- date: Entrega la fecha
- pwd: “Print Working Directory”. Imprime la ruta del directorio actual
- echo: Imprimir texto en pantalla
- ls: Listar los archivos y carpetas
- clear: Limpiar la pantalla
- man: Muestra el manual
- -h / - -help



```
john@ubuntu:~$ ls
john_directory john_file
john@ubuntu:~$ ls -l
total 8
drwxrwxr-x 2 john john 40 Oct 1 11:10 john_directory
-rw-rw-r-- 1 john john 5120 Oct 1 11:17 john_file
john@ubuntu:~$ ls -l -h
total 8.0K
drwxrwxr-x 2 john john 40 Oct 1 11:10 john_directory
-rw-rw-r-- 1 john john 5.0K Oct 1 11:17 john_file
john@ubuntu:~$ ls -lh john_file
-rw-rw-r-- 1 john john 5.0K Oct 1 11:17 john_file
john@ubuntu:~$ ls -l --human-readable john_file
-rw-rw-r-- 1 john john 5.0K Oct 1 11:17 john_file
john@ubuntu:~$
```

# Edición de archivos y carpetas

- cd: “Change Directory” moverse entre carpetas. . .. - ~
- touch: Crear un archivo en blanco
- mkdir: Crear una carpeta
- >: Redireccionar texto
- cat: Mostrar el contenido de un archivo
- head: Muestra las primeras líneas de un archivo
- tail: Muestra las últimas líneas de un archivo
- wc: “Word Counter” muestra el número de líneas, palabras, y caracteres en un archivo

```
student01@server01:~$ wc laptop_inv.txt
14 40 352 laptop_inv.txt
```

Total lines

Total words

Total characters

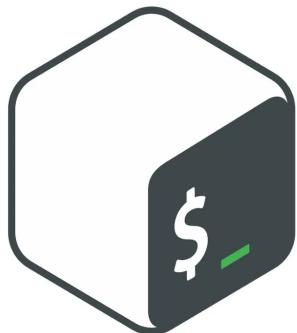
# Copiar, mover, y eliminar archivos

- [cp](#): Copiar archivos \$ cp <nombre del archivo> <destino/nombre>
- [mv](#): Mover archivos \$ mv <nombre del archivo> <destino/nombre>
- [rm](#): Remover un archivo
- [rmdir](#): Remover una carpeta

## Reto:

1. Crear un archivo vacío llamado mi\_texto.txt
2. Escribir el texto “Mi primer Hola Mundo desde la línea de comandos” en el archivo creado
3. Crear una carpeta llamada “mis\_archivos”
4. Copiar el archivo mi\_texto.txt a la carpeta mis\_archivos
5. Eliminar el archivo mi\_texto.txt
6. Moverse dentro de la carpeta mis\_archivos
7. Mostrar el contenido del archivo mi\_texto.txt
8. Mostrar el número de caracteres en el archivo mi\_texto.txt
9. Salir de la carpeta mis\_archivos
10. Eliminar la carpeta mis\_archivos

# Otros comandos:



- Comandos para búsqueda: which, find, grep
- Comprimir/Descomprimir archivos: tar, zip, unzip
- Revisar el historial: history
- sudo, su, alias, unalias
- wget, ping, awk, kill, xkill
- sed, ps, passwd, exit, git

Reto: <https://cmdchallenge.com>

# Módulo 2:

## Crear un pipeline de NGS

# Next Generation Sequencing - NGS

- La secuenciación de nueva generación (*Next Generation Sequencing [NGS]*) es un grupo de tecnologías diseñadas para secuenciar gran cantidad de segmentos de ADN de forma masiva y en paralelo, en menor cantidad de tiempo y a un menor costo por base.
- Su uso se dio inicialmente para detectar variantes de nucleótido único y cada vez se ha desarrollado para otro tipo de variantes, como inserciones, delecciones y grandes rearreglos.
- Gracias a los recientes desarrollos en las pruebas basadas en NGS, estas tecnologías se plantean como estrategias de gran utilidad para la prevención, el diagnóstico, el tratamiento y el seguimiento de un amplio espectro de enfermedades, incluidas condiciones genéticas, patologías crónicas y enfermedades infecciosas, y se prevé que en un futuro cercano su creciente aplicación clínica generará resultados favorables para lograr el diagnóstico molecular en un número mayor de pacientes y a un menor costo.
- En la práctica clínica es creciente el uso actual de las pruebas basadas en NGS; sin embargo, aún existe incertidumbre sobre aspectos importantes como las indicaciones adecuadas para su uso, limitaciones de la técnica (sensibilidad y especificidad), reporte de variantes, interpretación de resultados, relación costo-beneficio y cobertura

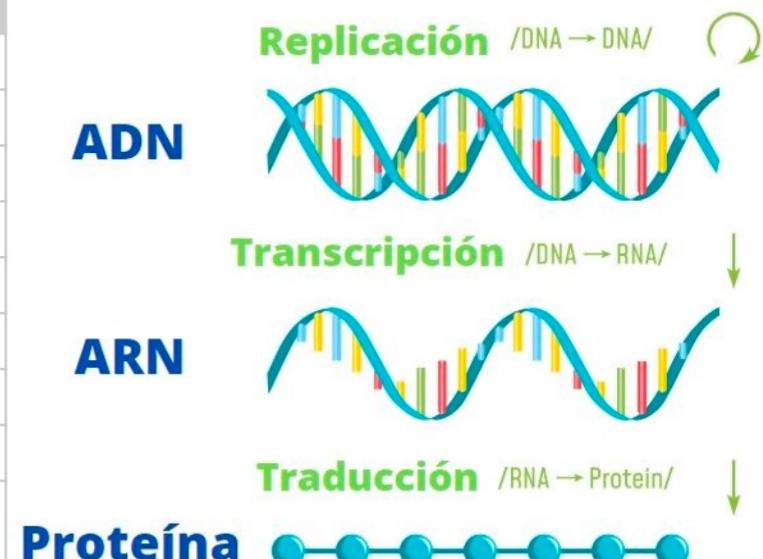
# Cómo está almacenada la información biológica?

## Proteínas: Aminoácidos

Aminoácido	Símbolo de una letra	Abreviatura común
Alanina	A	Ala
Arginina	R	Arg
Asparagina	N	Asn
Ácido aspártico	D	Asp
Cisteína	C	Cys
Glutamina	Q	Gln
Ácido glutámico	E	Glu
Glicina	G	Gly
Histidina	H	His
Isoleucina	I	Ile
Leucina	L	Leu
Lisina	K	Lys
Metionina	M	Met
Fenilalanina	F	Phe
Prolina	P	Pro
Serina	S	Ser
Treonina	T	Thr
Triptófano	W	Trp
Tirosina	Y	Tyr
Valina	V	Val

## DNA: Bases nitrogenadas

Símbolo	Nucleótido	Categoría
A	Adenina	Purina
C	Citosina	Pirimidina
G	Guanina	Purina
T	Timina	Pirimidina
N	Cualquier nucleótido	-----
R	A ó G	Purinas
Y	C ó T	Pirimidinas
S	C ó G	Enlace fuerte
-	Ninguno (agujero o gap)	-----

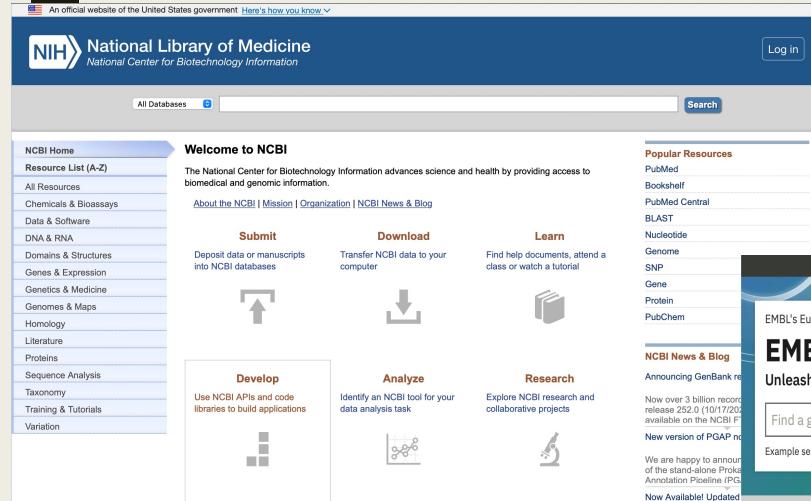


Para los científicos de datos son archivos de texto con un formato específico: fasta, fastq, sam, bam, vcf, pdb, xyz, ....

# Principales Bases de datos Biológicos

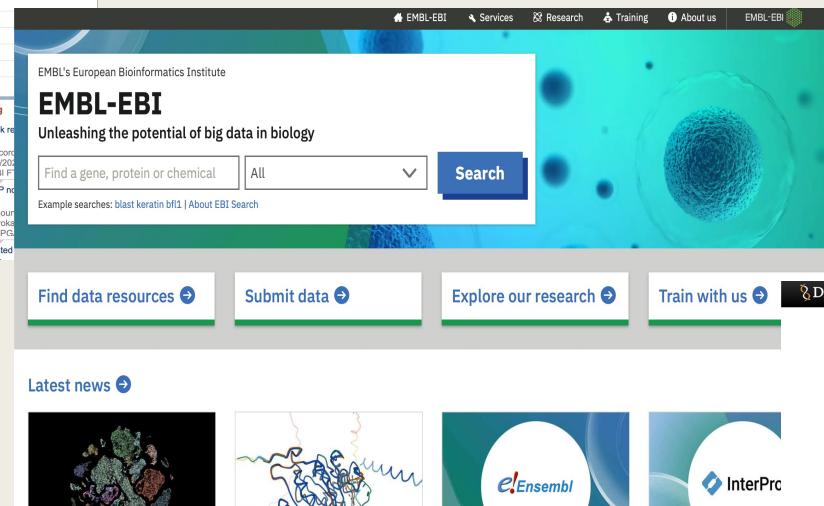
NCBI: National Center for Biotechnology Information

<https://www.ncbi.nlm.nih.gov>



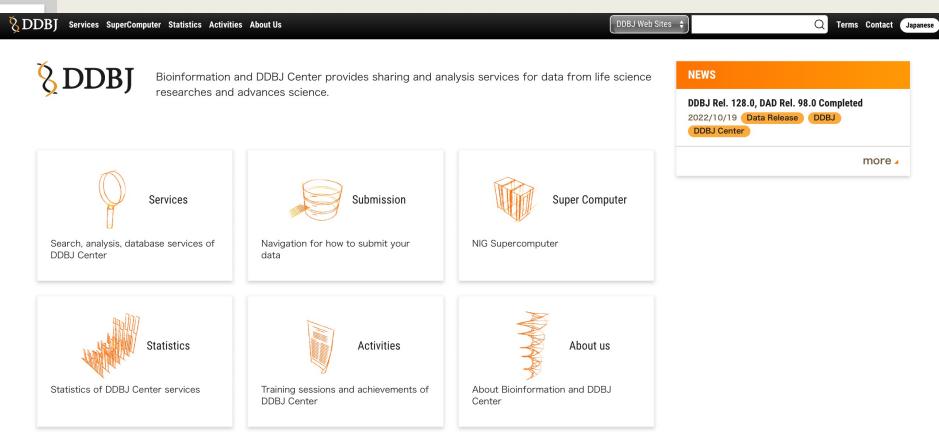
The screenshot shows the NCBI homepage with a dark blue header. It features a search bar, a 'Log in' button, and a 'Popular Resources' sidebar on the right. The main content area includes sections for 'Welcome to NCBI', 'Submit', 'Download', 'Learn', 'Develop', 'Analyze', and 'Research'. A central banner highlights 'Unleashing the potential of big data in biology'.

EMBL-EBI: European Nucleotide Archive  
- European Bioinformatics Institute  
<https://www.ebi.ac.uk>



The screenshot shows the EMBL-EBI homepage with a large image of a cell. It features a search bar at the top and several navigation links: 'Find data resources', 'Submit data', 'Explore our research', and 'Train with us'. Below these are sections for 'Latest news' and logos for 'eEnsembl' and 'InterPro'.

DDBJ: DNA Data Bank of Japan  
<https://www.ddbj.nig.ac.jp/index-e.html>



The screenshot shows the DDBJ homepage with a dark header. It features a 'NEWS' section with a link to 'DDBJ Rel. 128.0, DAD Rel. 98.0 Completed'. Below this are sections for 'Services', 'Submission', 'Super Computer', 'Statistics', 'Activities', and 'About us', each with a corresponding icon.

INSDC:  
International Nucleotide Sequence Database Collaboration  
<https://www.insdc.org>

# Formato fasta

Trabajo\_final retocado: Bloc de notas

Archivo Edición Formato Ver Ayuda

---

```
>ct179_r_b38
CCAAAAAAATAAAATTACAGACCGTCGTAAAGTGAATTAGTAATCTAAAACCTTATAAA
AATCGGTCGTAAAATTGAAGGACTTACACATCCTCCGGTGACGAATTGAGGGAGGAAGC
AAGCTACAGCGGCCGGTAAGTTGAGCAAGCATGTAGAAGGAGCCAGTGATAAAGGTAA
TATGTCCGCCGAAGGTTAACCCACAGGTGACCGCAGGGTTAACATGACCACCGGAGATGT
TAGCGGAAATCGAAACGGCTACGAATAGAGCAAATCCATGGCAAATGCAATAGCTACAA
GCCCAGCCGGATCAAGTGCAGCATTGTTCAACTTGCCTGTATCCAAACAATACCAAA
TTTAAAATCTCTACATCGATCAATAGTAAAAATACACAAAAATAGAATATATATTAC
TTGATCAGAAGCGGATCTAGTAAAAGGATATGATTTGGTATAACTTATAAACCTGAATC
AATCTATTTGATATGAATTCAAAATAAGCATTAACCATTATGCGGGGCATAATTCCGA
GTCAGAAGCAATAAATCAAACCTGGAACCGCAGAATTGTTAACATTAAAATGAAAAGA
GGGATCGGATCGGAGTTATTAACCGTAAGCAATGGCGGACCAACTCCGGCGAAGACAAAG
ATGAGTGTGGAGATGAATTGGCAAGATAGGACTTAAGAGAGACAACGCTGAATGAATCA
CTAACTC
```

- Cada secuencia inicia con el símbolo `>` seguido del identificador de la secuencia y todos los comentarios que se quiera sobre esa misma línea.
- La siguiente línea describe propiamente la secuencia, a la que se le pueden insertar tabuladores, cambios de línea y espacios, mismos que serán ignorados en cualquier análisis.
- La secuencia es leída hasta que es encontrado el fin de archivo ó una nueva secuencia inicia con `>`.

# Formato fastq

- Similar al formato fasta, pero en estos el punter para cada secuencia es un @.
  - También contiene la información para cada secuencia.
  - Incluye una tercera linea con la calidad PHRED para cada uno de los nucleótidos.

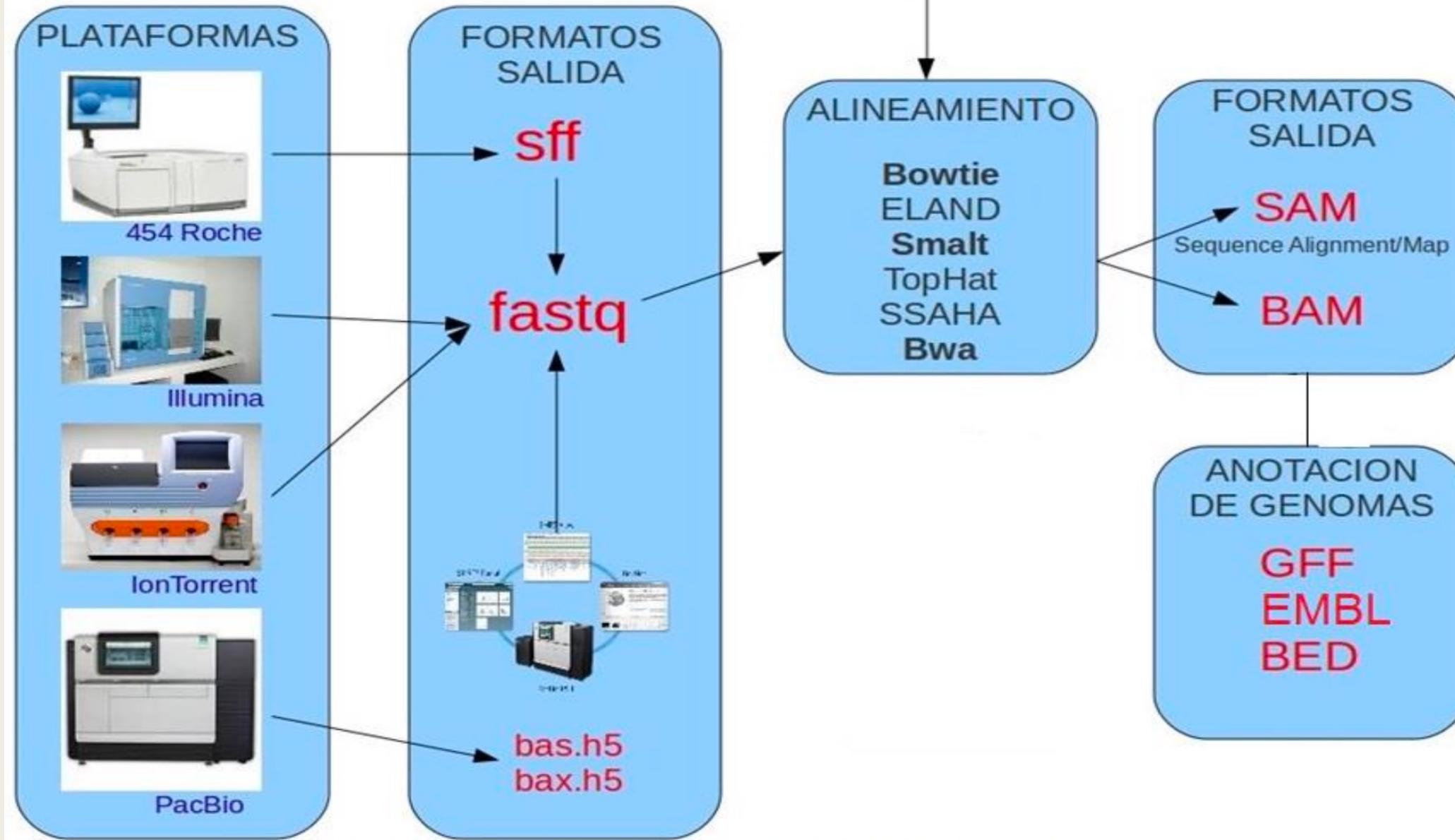
@HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:215:593  
GAGAAGTTAACACAGCTGGTATTATTTTGTAAACAT  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:215:593  
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE  
@HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:234:551  
TGGGACTTTATCTGGAGGGAGTGGAAAGCCATT  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:234:551  
hh  
@HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:338:194  
TGGTTTATGCAGAAATTCTAGAATAAGGGTAACCT  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:338:194  
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh  
@HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:363:717  
TCTCAGAAACTTGTTGTGATGTGTATTCAACTA  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:363:717  
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh  
@HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:208:209  
TTGATTTAACTCTGACAAAATAACAAAGTCTTAGG  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:208:209

Sequencing info

Nucleotide sequence

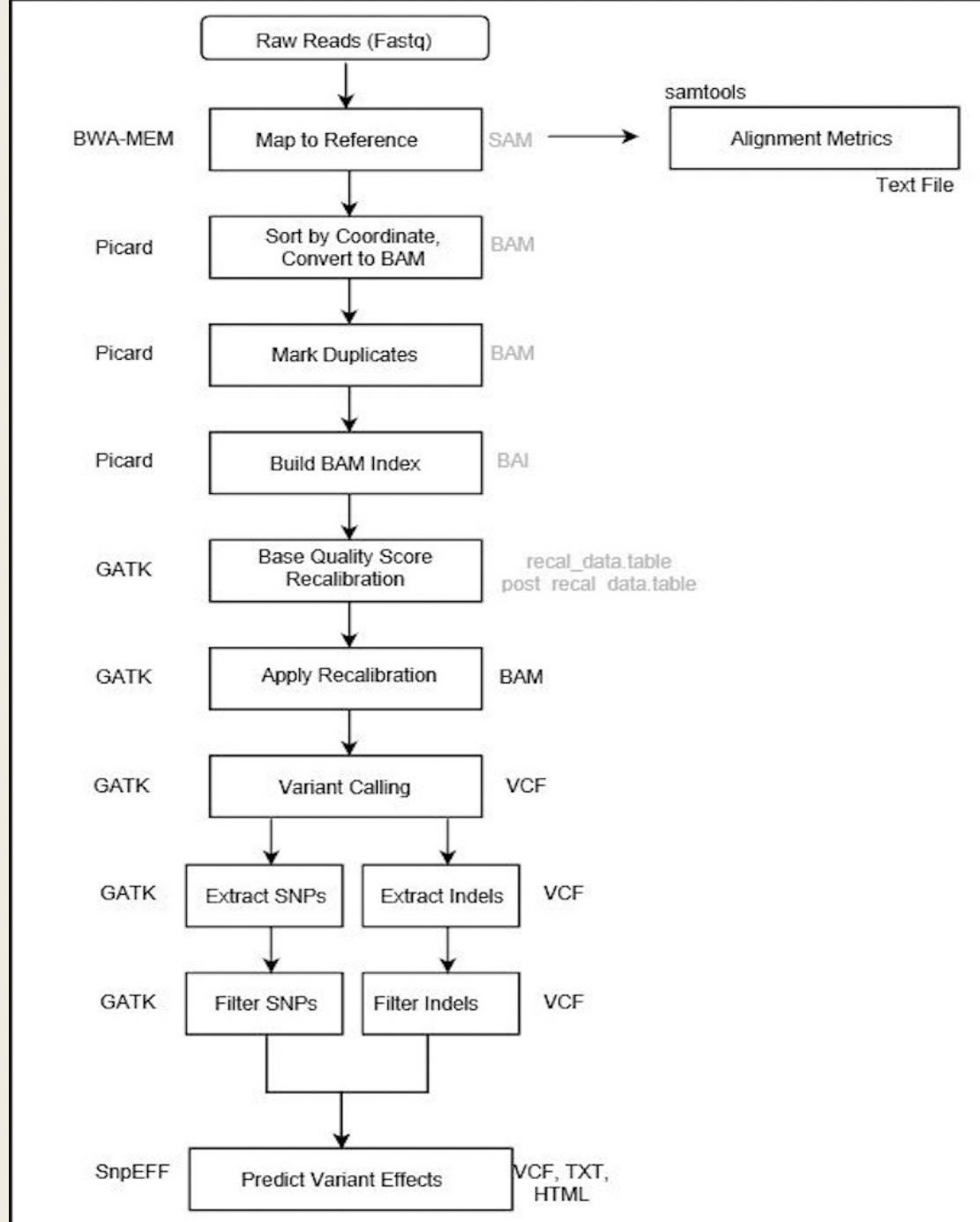
Quality score in ASCII

# Flujo de tipos de Archivos



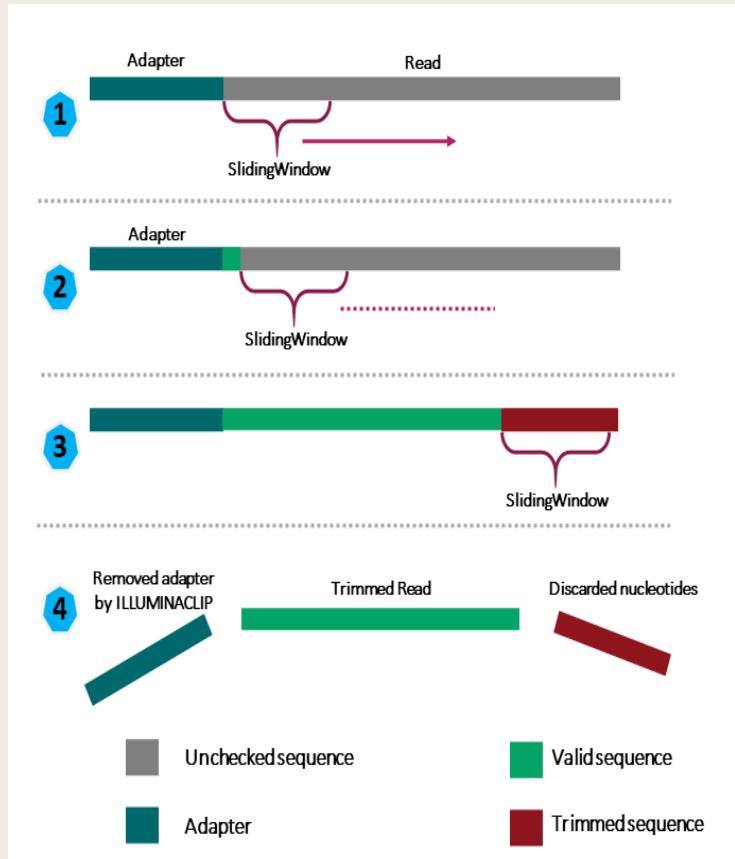
# Creando un pipeline para procesar secuencias de DNA.

1. Preparación de los datos
2. Alineamiento
3. Preprocesamiento
4. Llamado de Variantes
5. Postprocesamiento
6. Anotación de Variables



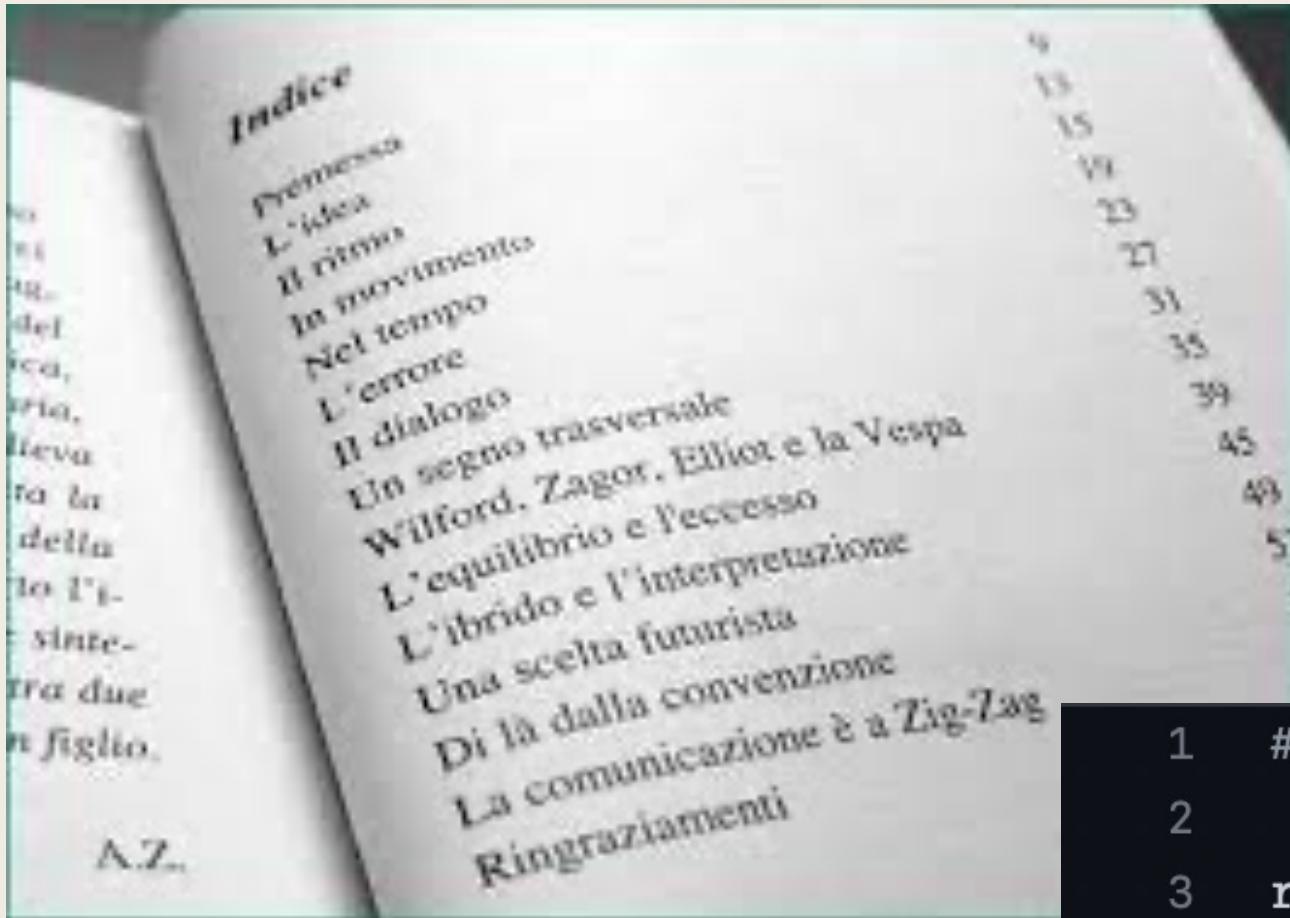
# 1. Preparación de los datos

- Trimming: Eliminar las secuencias usadas como adaptadores.
- Herramienta: Trimmomatic



```
1  #!/bin/bash
2
3  read_sample_R1=./input/sample_R1.fastq
4  read_sample_R2=./input/sample_R2.fastq
5
6  read_sample_R1_trimmed=./input/sample_R1_trimmed.fastq
7  read_sample_R2_trimmed=./input/sample_R2_trimmed.fastq
8
9  trimmomatic PE -phred33 \
10    -basein ${read_sample_R1} ${read_sample_R2} \
11    -baseout ${read_sample_R1_trimmed} ${read_sample_R2_trimmed} \
12    HEADCROP:15 \
13    TRAILING:3 \
14    SLIDINGWINDOW:4:15 \
15    LEADING:3 \
16
```

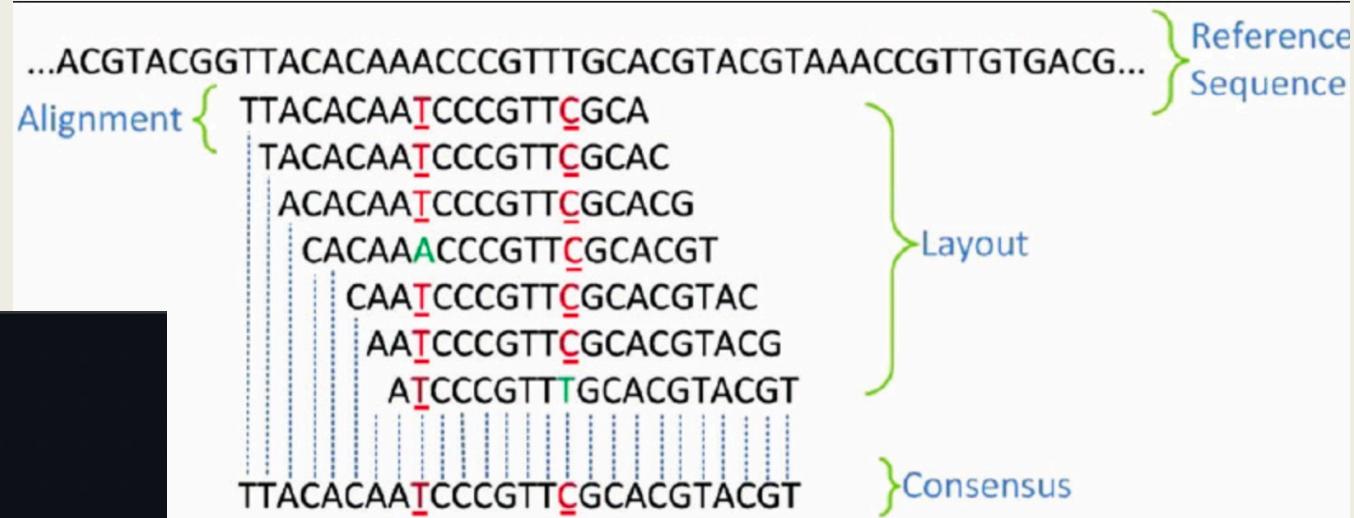
- **Indexing:** La indexación es un método usado para acceder fácilmente a los datos
- **Herramienta:** bwa index



```
1 #!/bin/bash
2
3 reference_genome=./input/hg38.fa
4
5 bwa index ${reference_genome}
6
```

# 2. Alineamiento

```
1 #!/bin/bash
2
3  read_sample_R1=./input/sample_R1.fastq
4  read_sample_R2=./input/sample_R2.fastq
5
6  # read_sample_R1=./input/sample_R1_trimmed.fastq
7  # read_sample_R2=./input/sample_R2_trimmed.fastq
8
9  reference_genome=./input/hg38.fa
10
11 output=./output/aligned_seqence.sam
12
13 bwa mem -t 10 -M \
14   ${reference_genome} \
15   ${read_sample_R1} ${read_sample_R2} \
16   > ${output}
17
```



- **BWA mem: Borrows-Wheeler**  
Aligner es un algoritmo usado para alinear múltiples secuencias contra una secuencia de referencia.
- **Herramienta: bwa mem**

# 3. Preprocesamiento

- Convertir SAM a BAM y ordenar el arreglo

The diagram illustrates the structure of a SAM file with several annotations:

- Obligatorio: Encabezado** (Red box): Annotations pointing to the header section, which includes entries like @HD, @SQ, and @RG.
- Importante: Contigs del genoma de referencia contra el que se alineó** (Blue box): Annotations pointing to the reference genome contig information.
- Importante: información de grupos de reads. Plataforma (PL) librería (LB) y muestra (SM)** (Green box): Annotations pointing to the read group information.
- Util: Herramientas de procesamiento aplicadas a las sec.** (Blue box): Annotations pointing to the processing tools section at the bottom of the file.

```
@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:247249719
@SQ SN:chr2 LN:242951149
[cut for clarity]
@SQ SN:chr9 LN:140273252
@SQ SN:chr10 LN:135374737
@SQ SN:chr11 LN:134452384
[cut for clarity]
@SQ SN:chr22 LN:49691432
@SQ SN:chrX LN:154913754
@SQ SN:chrY LN:57772954
@RG ID:20FUK.1 PL:illumina PU:20FUKAAXX100202.1 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.2 PL:illumina PU:20FUKAAXX100202.2 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.3 PL:illumina PU:20FUKAAXX100202.3 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.4 PL:illumina PU:20FUKAAXX100202.4 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.5 PL:illumina PU:20FUKAAXX100202.5 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.6 PL:illumina PU:20FUKAAXX100202.6 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.7 PL:illumina PU:20FUKAAXX100202.7 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.8 PL:illumina PU:20FUKAAXX100202.8 LB:Solexa-18484 SM:NA12878 CN:BI
@PG ID:BWA VN:0.5.7 CL:tk
@PG ID:GATK TableRecalibration VN:1.0.2864
20FUKAAXX100202:1:1:12730:189900 163 chrM 1 60 101M = 282 381
GATCACAGGTCTATCACCTTAAACCACTCACGGGAGCTCTCCATGCATTGGTA...[more bases]
?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCAB...[more quals]
RG:Z:20FUK.1 NM:i:1 SM:i:37 AM:i:37 MD:Z:72G28 MQ:i:60 PG:Z:BWA UQ:i:33
```

```
1#!/bin/bash
2
3reference_genome=./input/hg38.fa
4input=./output/aligned_sequence.sam
5output=./output/sorted_sequence.bam
6
7# java -Xmx10g -jar ./picard.jar \
8PicardCommandLine \
9    SortSam VALIDATION_STRINGENCY=SILENT \
10    I=${input} \
11    O=${output} \
12    SORT_ORDER=coordinate
13
```

```
1#!/bin/bash
2
3aligned_sequence=./output/aligned_sequence.sam
4metrics=./output/alignment_metrics.txt
5
6samtools flagstat ${aligned_sequence} > ${metrics}
7
```

## ■ Mark Duplicates:

```
1 #!/bin/bash  
2  
3 input=./output/markduplicates.bam  
4  
5 samtools index ${input}  
6
```

## ■ Crear el diccionario de referencia:

```
1 #!/bin/bash  
2  
3 reference_genome=./input/hg38  
4  
5 # java -Xmx10g -jar ./picard.jar \  
6 PicardCommandLine \  
7     CreateSequenceDictionary \  
8     R="$reference_genome".fa \  
9     O="$reference_genome".dict  
10
```

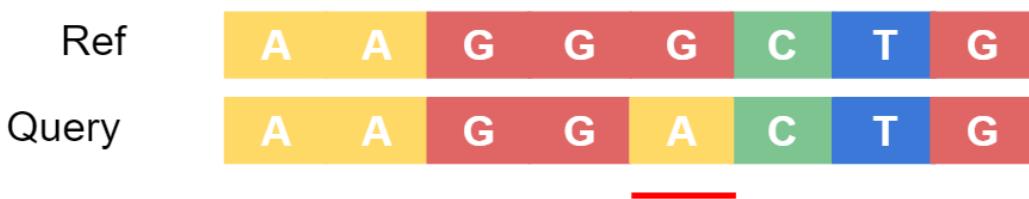
## ■ Indexar el BAM:

```
1 #!/bin/bash  
2  
3 reference_genome=./input/hg38.fa  
4 input=./output/sorted_seqence.bam  
5 output=./output/markduplicates.bam  
6  
7 # java -Xmx10g -jar ./picard.jar \  
8 PicardCommandLine \  
9     MarkDuplicates \  
10    I=${input} \  
11    O=${output} \  
12    M="$output".metrics  
13
```

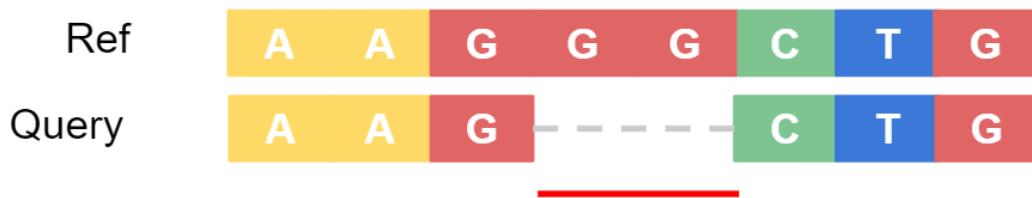
# 4. Llamado de Variantes

## Sequence Variants

### SNV (Single Nucleotide Variant)



### INDEL (Insertion or Deletion)



```
1 #!/bin/bash
2
3 reference_genome=./input/hg38.fa
4 input=./output/sequence_RG.bam
5 output=./output/variants.vcf.gz
6
7 ~/gatk-4.3.0.0/gatk --java-options "-Xmx10g" \
8     HaplotypeCaller \
9     -R ${reference_genome} \
10    -I ${input} \
11    -O ${output} \
12    -ERC GVCF
13
```

# FORMATO VCF

## Example

##fileformat=VCFv4.0  
##fileDate=20100707  
##source=VCFtools  
##reference=NCBI36  
##INFO<ID=AA,Number=1,Type=String>Description="Ancestral Allele">  
##INFO<ID=H2,Number=0,Type=Flag>Description="HapMap2 membership">  
##FORMAT<ID=GT,Number=1,Type=String>Description="Genotype">  
##FORMAT<ID=GQ,Number=1,Type=Integer>Description="Genotype Quality (phred score)">  
##FORMAT<ID=GL,Number=3,Type=Float>Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">  
##FORMAT<ID=DP,Number=1,Type=Integer>Description="Read Depth">  
##ALT<ID=DEL,Description="Deletion">  
##INFO<ID=SVTYPE,Number=1,Type=String>Description="Type of structural variant">  
##INFO<ID=END,Number=1,Type=Integer>Description="End position of the variant">

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**VCF header** {   
**Mandatory header lines** ←  
**Optional header lines** (meta-data about the annotations in the VCF body) ←

**Body** {   
**Reference alleles (GT=0)** →  
**Alternate alleles (GT>0 is an index to the ALT column)** →  
**Phased data** (G and C above are on the same chromosome) →

**Deletion** →  
**SNP** →  
**Large SV** ↑  
**Insertion** →  
**Other event** →

# 5. POSTPROCESAMIENTO

## Indexado del vcf

```
1 #!/bin/bash
2
3 input=./output/variants.vcf.gz
4
5 # tabix -p vcf ${input} # Not necessary for with GATK
6 vt peek ${input}
7
```

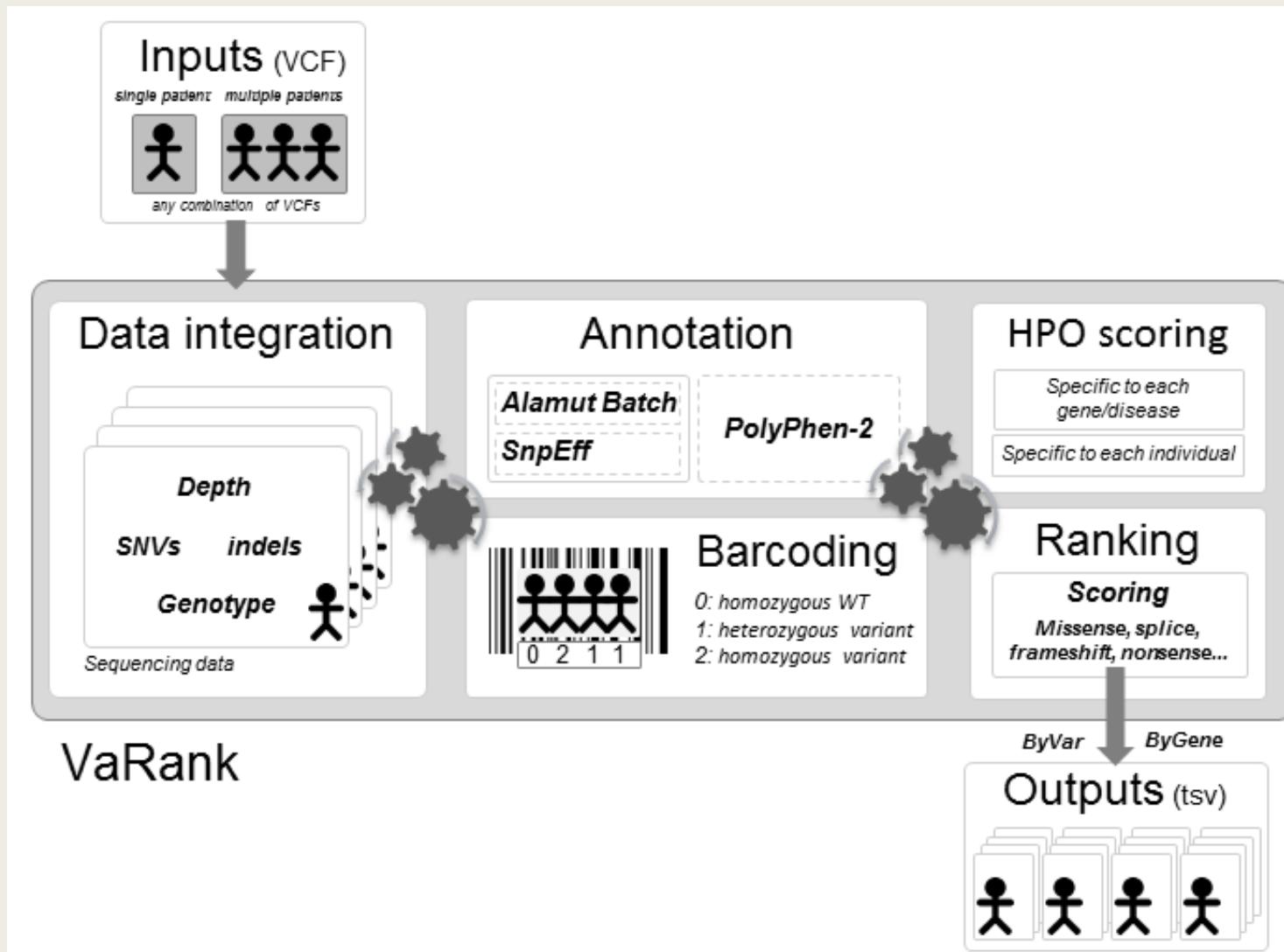
## Filtrar por calidad

```
1 #!/bin/bash
2
3 reference_genome=./input/hg38.fa
4 input=./output/variants.vcf.gz
5 output=./output/variants_InDels.vcf.gz
6
7 gatk --java-options "-Xmx10g" \
8     SelectVariants \
9     -R ${reference_genome} \
10    -V ${input} \
11    -select-type INDEL \
12    -O ${output} \
13
```

## Eliminar SNP / InDels

```
1 #!/bin/bash
2
3 reference_genome=./input/hg38.fa
4 input=./output/variants.vcf.gz
5 output=./output/variants_InDels.vcf.gz
6
7 gatk --java-options "-Xmx10g" \
8     SelectVariants \
9     -R ${reference_genome} \
10    -V ${input} \
11    -select-type INDEL \
12    -O ${output} \
13
```

# 6. ANOTACIÓN DE VARIABLES



## feature type

using Sequence Ontology  
transcript, motif, miRNA ...

## feature ID

dependent on annotation  
transcript ID, motif ID, ChipSwq peak ...

## gene name

common gene name (HGNC)

## biotype

Ensemble biotypes  
Coding, non-coding..

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	123456	.	C	A	.	.	ANN=A ...
chr1	234567	.	A	G,T	.	.	ANN=G ..., T ...

ANN = Annotation aka effect or consequence

## putative impact

description of consequence  
exon\_loss\_variant, stop\_lost,  
frameshift\_variant

## impact

estimation of level of impact  
HIGH, LOW, MODERATE

# SnpEff

Genetic variant annotation and effect prediction toolbox.

# Liponium

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	Gen	Gen-Position	Gen AA	Mutation typ	Probe	Position	Read	ference	Codutated Code	Counts	Frequencies	Aminated	Aminated	Resistan	Notes	ward_SON	Gen.1
2	ddn	ddn-144	ddn_L49P	sust	AGAAAGATT	217	AGGGACTG	CTG	TAC	1	0.024643	L	Y	Delamanid	2 ) Asso c w	L49P	
3	embB	embB-219	embB_L74F	sust	TGACCGCC	217	AGGGACTG	CTG	TAC	1	0.024643	L	Y	Ethambuto	2 ) Asso c w	L74R	
4	pncA	pncA-345	pncA_L116	sust	TCGACGAG	217	AGGGACTG	CTG	TAC	1	0.024643	L	Y	Pyrazinami	2 ) Asso c w	L116P	
5	pncA	pncA-474	pncA_L159	sust	CCACCAGG	217	AGGGACTG	CTG	TAC	1	0.024643	L	Y	Pyrazinami	1) Assoc w	L159R	
6	pykA	pykA-654	pykA-219	sust	GGGTGCCG	217	AGGGACTG	CTG	TAC	1	0.024643	L	Y			AAGCCGGA pykA-219	
7	rpoB	rpoB-1344	rpoB_L449I	sust	GGTTGACC	217	AGGGACTG	CTG	TAC	1	0.024643	L	Y	Rifampin	2) Assoc w	L449M	
8	ddn	ddn-144	ddn_L49P	sust	AGAAAGATT	240	ATCGTGGC	CTG	TTG	1	0.024643	L	L	Delamanid	2 ) Asso c w	L49P	
9	embB	embB-219	embB_L74F	sust	TGACCGCC	240	ATCGTGGC	CTG	TTG	1	0.024643	L	L	Ethambuto	2 ) Asso c w	L74R	
10	pncA	pncA-345	pncA_L116	sust	TCGACGAG	240	ATCGTGGC	CTG	TTG	1	0.024643	L	L	Pyrazinami	2 ) Asso c w	L116P	
11	pncA	pncA-474	pncA_L159	sust	CCACCAGG	240	ATCGTGGC	CTG	TTG	1	0.024643	L	L	Pyrazinami	1) Assoc w	L159R	
12	pykA	pykA-654	pykA-219	sust	GGGTGCCG	240	ATCGTGGC	CTG	TTG	1	0.024643	L	L			AAGCCGGA pykA-219	
13	rpoB	rpoB-1344	rpoB_L449I	sust	GGTTGACC	240	ATCGTGGC	CTG	TTG	1	0.024643	L	L	Rifampin	2) Assoc w	L449M	
14	ddn	ddn-144	ddn_L49P	sust	AGAAAGATT	244	CTCAATTG	CTG	GAT	49	1.207491	L	D	Delamanid	2 ) Asso c w	L49P	
15	embB	embB-219	embB_L74F	sust	TGACCGCC	244	CTCAATTG	CTG	GAT	49	1.207491	L	D	Ethambuto	2 ) Asso c w	L74R	
16	pncA	pncA-345	pncA_L116	sust	TCGACGAG	244	CTCAATTG	CTG	GAT	49	1.207491	L	D	Pyrazinami	2 ) Asso c w	L116P	
17	pncA	pncA-474	pncA_L159	sust	CCACCAGG	244	CTCAATTG	CTG	GAT	49	1.207491	L	D	Pyrazinami	1) Assoc w	L159R	
18	pykA	pykA-654	pykA-219	sust	GGGTGCCG	244	CTCAATTG	CTG	GAT	49	1.207491	L	D			AAGCCGGA pykA-219	
19	rpoB	rpoB-1344	rpoB_L449I	sust	GGTTGACC	244	CTCAATTG	CTG	GAT	49	1.207491	L	D	Rifampin	2) Assoc w	L449M	
20	ddn	ddn-144	ddn_L49P	sust	AGAAAGATT	203	CACCGGGC	CTG	TAG	1	0.024643	L	*	Delamanid	2 ) Asso c w	L49P	
21	embB	embB-219	embB_L74F	sust	TGACCGCC	203	CACCGGGC	CTG	TAG	1	0.024643	L	*	Ethambuto	2 ) Asso c w	L74R	
22	pncA	pncA-345	pncA_L116	sust	TCGACGAG	203	CACCGGGC	CTG	TAG	1	0.024643	L	*	Pyrazinami	2 ) Asso c w	L116P	
23	pncA	pncA-474	pncA_L159	sust	CCACCAGG	203	CACCGGGC	CTG	TAG	1	0.024643	L	*	Pyrazinami	1) Assoc w	L159R	
24	pykA	pykA-654	pykA-219	sust	GGGTGCCG	203	CACCGGGC	CTG	TAG	1	0.024643	L	*			AAGCCGGA pykA-219	
25	rpoB	rpoB-1344	rpoB_L449I	sust	GGTTGACC	203	CACCGGGC	CTG	TAG	1	0.024643	L	*	Rifampin	2) Assoc w	L449M	
26	ddn	ddn-144	ddn_L49P	sust	AGAAAGATT	47	TGGGCGGC	CTG	TCG	1	0.024643	L	S	Delamanid	2 ) Asso c w	L49P	
27	embB	embB-219	embB_L74F	sust	TGACCGCC	47	TGGGCGGC	CTG	TCG	1	0.024643	L	S	Ethambuto	2 ) Asso c w	L74R	
28	pncA	pncA-345	pncA_L116	sust	TCGACGAG	47	TGGGCGGC	CTG	TCG	1	0.024643	L	S	Pyrazinami	2 ) Asso c w	L116P	
29	pncA	pncA-474	pncA_L159	sust	CCACCAGG	47	TGGGCGGC	CTG	TCG	1	0.024643	L	S	Pyrazinami	1) Assoc w	L159R	
30	pykA	pykA-654	pykA-219	sust	GGGTGCCG	47	TGGGCGGC	CTG	TCG	1	0.024643	L	S			AAGCCGGA pykA-219	
31	rpoB	rpoB-1344	rpoB_L449I	sust	GGTTGACC	47	TGGGCGGC	CTG	TCG	1	0.024643	L	S	Rifampin	2) Assoc w	L449M	
32	ddn	ddn-144	ddn_L49P	sust	AGAAAGATT	94	TAAGAACAC	CTG	CGG	1	0.024643	L	R	Delamanid	2 ) Asso c w	L49P	
33	embB	embB-219	embB_L74F	sust	TGACCGCC	94	TAAGAACAC	CTG	CGG	1	0.024643	L	R	Ethambuto	2 ) Asso c w	L74R	
34	pncA	pncA-345	pncA_L116	sust	TCGACGAG	94	TAAGAACAC	CTG	CGG	1	0.024643	L	R	Pyrazinami	2 ) Asso c w	L116P	
35	pncA	pncA-474	pncA_L159	sust	CCACCAGG	94	TAAGAACAC	CTG	CGG	1	0.024643	L	R	Pyrazinami	1) Assoc w	L159R	
36	pykA	pykA-654	pykA-219	sust	GGGTGCCG	94	TAAGAACAC	CTG	CGG	1	0.024643	L	R			AAGCCGGA pykA-219	
37	rpoB	rpoB-1344	rpoB_L449I	sust	GGTTGACC	94	TAAGAACAC	CTG	CGG	1	0.024643	L	R	Rifampin	2) Assoc w	L449M	

# Pipelines Predefinidos

## Gratuitos:

- Snakemake
- Nextflow
- Bamtools
- EMBL
- GATK
- NCBI

## Pagos:

- Schödinger
- Globant Genomics
- Illumina

# Recursos para entrenamiento

- <https://www.embl.org/training/>
- [https://www.centogene.com/centocloud-features?creative=617206197286&keyword=whole%20genome%20analyses&matchtype=b&network=g&device=c&gclid=Cj0KCQjwteOaBhDuARIsADBqReh0bhhy3ETtkPPQvkxdrziSWQjbY3FilYeiApYYtEGM4jHu7NBYqlaAoD9EALw\\_wcB](https://www.centogene.com/centocloud-features?creative=617206197286&keyword=whole%20genome%20analyses&matchtype=b&network=g&device=c&gclid=Cj0KCQjwteOaBhDuARIsADBqReh0bhhy3ETtkPPQvkxdrziSWQjbY3FilYeiApYYtEGM4jHu7NBYqlaAoD9EALw_wcB)
- <https://www.coursera.org/learn/wgs-bacteria>
- <https://www.illumina.com/services/instrument-services-training/training.html>