ARCHIVOS Y PROGRAMAS BIOINFORMÁTICOS

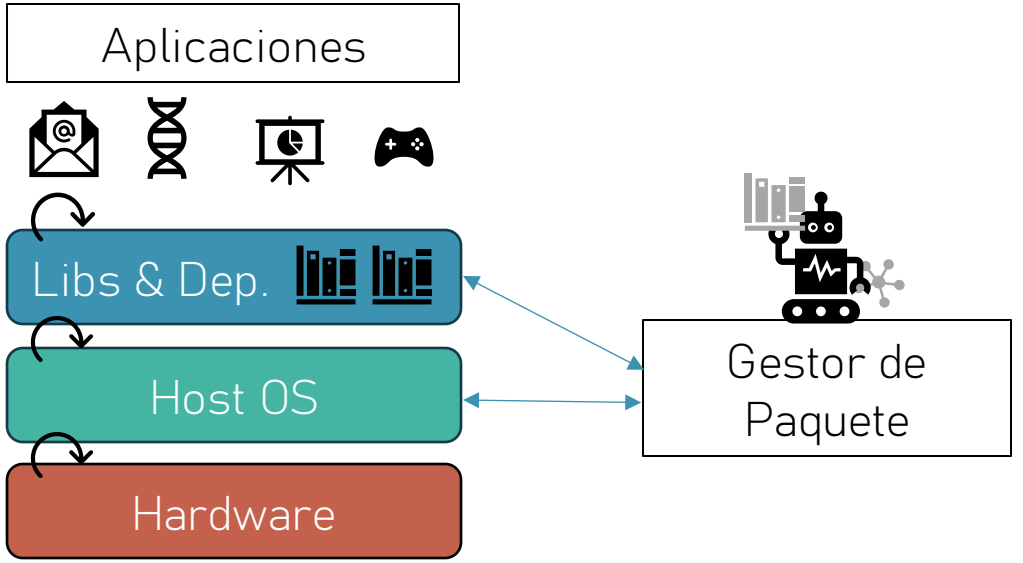
Murilo Cassiano, M.Sc.

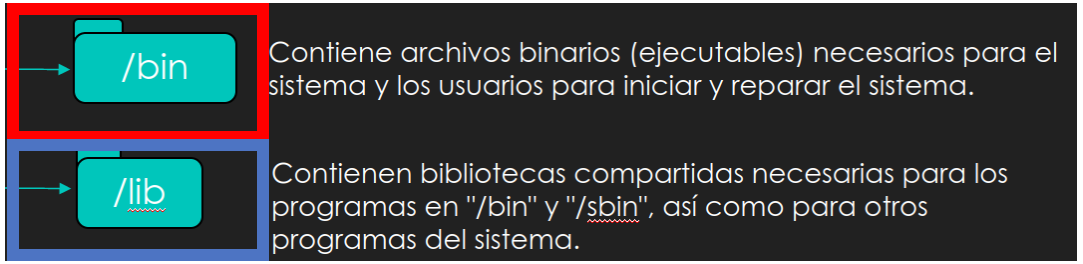
PARTE 1

Instalando programas en Linux y manejando entornos con Conda

GESTORES DE PAQUETES

- Es una herramienta que ayuda a administrar software en un sistema operativo.
- Facilita la instalación, actualización, configuración y eliminación de programas en una computadora.
- Mantienen un registro de las dependencias del software, asegurando que las aplicaciones funcionen correctamente y que todas las bibliotecas necesarias estén presentes en el sistema.



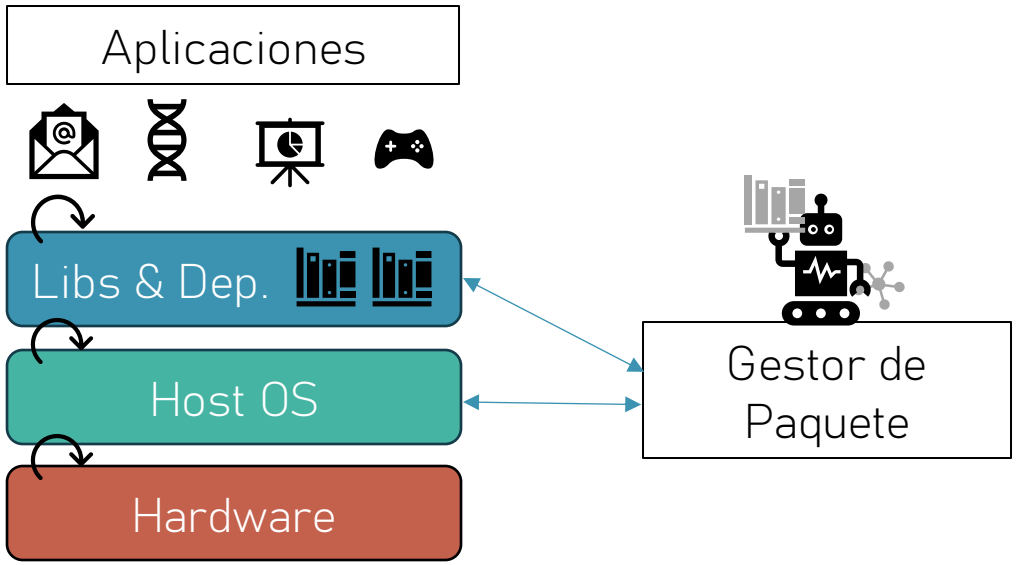


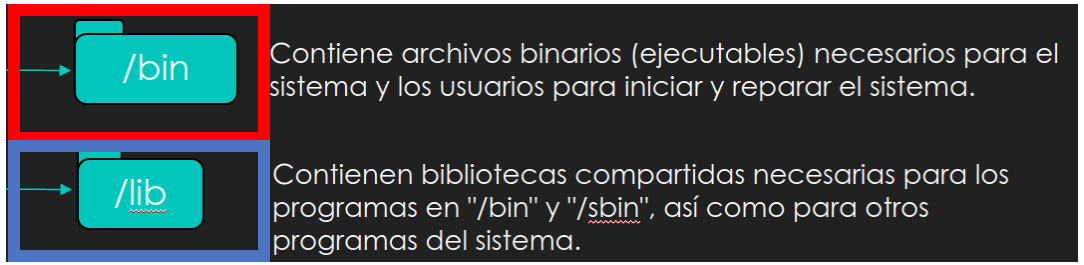
GESTOR DE PAQUETE: APT

- **APT (Advanced Package Tool)** es el sistema de gestión de paquetes utilizado en sistemas basados en Debian
  - Ubuntu
  - Mint
  - PopOs
  - Debian Linux

- APT simplifica la instalación y actualización de software, automatizando muchos procesos.

- A través de la línea de comandos, los usuarios pueden utilizar varios comandos para interactuar con APT.





GESTOR DE PAQUETE: APT

- Buscar un paquete:
  apt search término_de_búsqueda
  Este comando busca paquetes relacionados con el término de búsqueda especificado.

```
murilo@muca10-t14:~$ apt search htop
Sorting... Done
Full Text Search... Done
aha/stable,stable,now 0.5.1-3 amd64 [installed,automatic]
  ANSI color to HTML converter

bpytop/stable,stable 1.0.68-1 all
  Resource monitor that shows usage and stats

btop/stable,stable 1.2.13-1 amd64
  Modern and colorful command line resource monitor that shows usage and stats

htop/stable,stable 3.2.2-2 amd64
  interactive processes viewer

libauthen-oath-perl/stable,stable 2.0.1-2 all
  Perl module for OATH One Time Passwords

pftools/stable,stable 3.2.12-1 amd64
  build and search protein and DNA generalized profiles
```

- Mostrar información sobre un paquete:
  apt show nombre_del_paquete
  Este comando muestra información detallada sobre un paquete específico, incluyendo su descripción y dependencias.

```
murilo@muca10-t14:~$ apt show htop
Package: htop
Version: 3.2.2-2
Priority: optional
Section: utils
Maintainer: Daniel Lange <DLange@debian.org>
Installed-Size: 387 kB
Depends: libc6 (>= 2.34), libncursesw6 (>= 6), libnl-3-200 (>= 3.2.7), libnl-genl-3-200 (>= 3.2.7), libtinfo6 (>= 6)
Suggests: lm-sensors, lsof, strace
Homepage: https://htop.dev/
Tag: admin::monitoring, implemented-in::c, interface::text-mode,
 role::program, scope::utility, uitoolkit::ncurses, use::monitor,
 works-with::software:running
Download-Size: 152 kB
APT-Sources: http://ftp.au.debian.org/debian bookworm/main amd64 Packages
Description: interactive processes viewer
 Htop is an ncursed-based process viewer similar to top, but it
 allows one to scroll the list vertically and horizontally to see
 all processes and their full command lines.
 .
 Tasks related to processes (killing, renicing) can be done without
 entering their PIDs.
```

GESTOR DE PAQUETE: APT

```
murilo@muca10-t14:~$ apt search mafft
Sorting... Done
Full Text Search... Done
mafft/stable,stable 7.505-1 amd64
  Multiple alignment program for amino acid or nucleotide sequences

muscle/stable,stable 1:5.1.0-1 amd64
  Multiple alignment program of protein sequences

probalign/stable,stable 1.4-10 amd64
  multiple sequence alignment using partition function posterior probabilities

t-coffee/stable,stable 13.45.0.4846264+really13.41.0.28bdc39+dfsg-1 amd64
  Multiple Sequence Alignment
```

Intenta instalar con apt isntall htop y apt install muscle

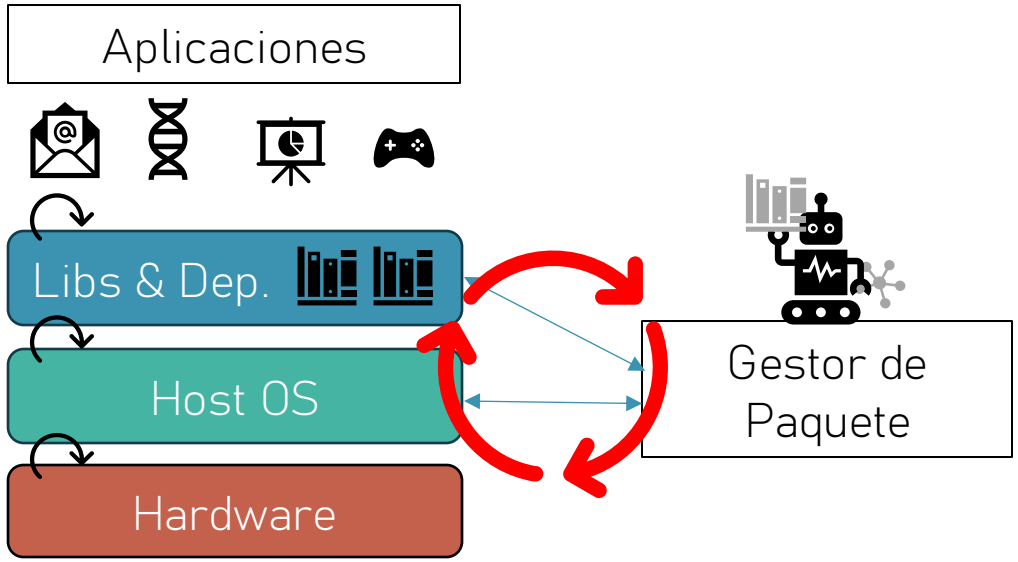
```
murilo@muca10-t14:~$ apt show muscle
Package: muscle
Version: 1:5.1.0-1
Priority: optional
Section: science
Maintainer: Debian Med Packaging Team <debian-med-packaging@lists.alioth.debian.org>
Installed-Size: 863 kB
Provides: muscle-doc
Depends: libc6 (>= 2.34), libgcc-s1 (>= 3.0), libgomp1 (>= 4.9), libstdc++6 (>= 11)
Conflicts: muscle-doc
Replaces: muscle-doc
Enhances: seaview, t-coffee
Homepage: https://www.drive5.com/muscle/
Tag: biology::format:aln, biology::peptidic, field::biology,
 field::biology:bioinformatics, implemented-in::c++,
 interface::commandline, role::program, scope::utility, use::comparing,
 works-with-format::plaintext, works-with::TODO
Download-Size: 297 kB
APT-Sources: http://ftp.au.debian.org/debian bookworm/main amd64 Packages
Description: Multiple alignment program of protein sequences
 MUSCLE is a multiple alignment program for protein sequences. MUSCLE
 stands for multiple sequence comparison by log-expectation. In the
 authors tests, MUSCLE achieved the highest scores of all tested
 programs on several alignment accuracy benchmarks, and is also one of
 the fastest programs out there.
 .
 Muscle v5 is a major re-write of MUSCLE based on new algorithms.
 .
 Users should be aware that command line arguments compared to version
 3.x of MUSCLE have changed!
 .
 Highest accuracy, scalable to thousands of sequences
 .
 Compared to previous versions, Muscle v5 is much more accurate, is often
 faster, and scales to much larger datasets. At the time of writing (late
 2021), Muscle v5 has the highest scores on multiple alignment benchmarks
 including Balibase, Bralibase, Prefab and Balifam. It can align tens of
 thousands of sequences with high accuracy on a low-cost commodity computer
 (say, an 8-core Intel CPU with 32 Gb RAM). On large datasets, Muscle v5
 is 20-30% more accurate than MAFFT and Clustal-Omega.
 .
 Alignment ensembles
 .
 Muscle v5 can generate ensembles of high-accuracy alternative alignments.
 All replicates have equal average accuracy on benchmark test, including
 the MSA made with default parameters. By comparing results of downstream
 analysis (trees, structure prediction...) on different replicates, you can
 assess the effects of alignment errors on your study.
```

GESTOR DE PAQUETE: APT

- La instalación de un paquete requiere que se modifiquen las dependencias y otras bibliotecas que están instaladas o que faltan en el sistema operativo.

- ¡El usuario debe ser parte de los administradores! ¡SUDO!

    - Siempre es arriesgado porque los programas instalados ya utilizan librerías que están configuradas



Instalar un nuevo paquete:

`sudo apt install nombre_del_paquete`

Este comando instala un nuevo paquete en el sistema.

`sudo apt install muscle mafft`

Eliminar un paquete:

`sudo apt remove nombre_del_paquete`

Este comando elimina un paquete del sistema, pero mantiene sus configuraciones.
Para eliminar también las configuraciones, se puede usar `sudo apt purge nombre_del_paquete`.

`sudo apt purge mafft`

GESTOR DE PAQUETE: APT

•Actualizar la lista de paquetes:

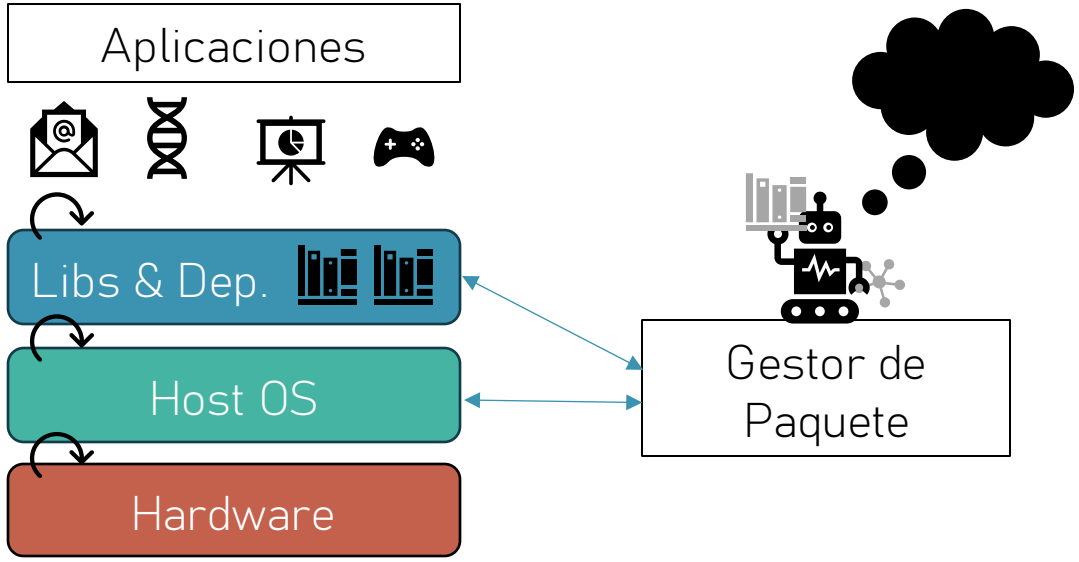
sudo apt update

Este comando actualiza la lista de paquetes disponibles en los repositorios.

•Actualizar los paquetes instalados:

sudo apt upgrade

Este comando actualiza todos los paquetes instalados en el sistema a sus últimas versiones.




```
murilo@muca10-t14:~$ sudo apt update
Hit:1 https://brave-browser-apt-release.s3.brave.com stable InRelease
Get:2 http://deb.debian.org/debian-security bookworm-security InRelease [48.0 kB]
Hit:3 http://repository.spotify.com stable InRelease
Hit:4 http://deb.debian.org/debian bookworm InRelease
Get:5 http://deb.debian.org/debian-security bookworm-security/main amd64 Packages [87.2 kB]
Get:6 http://deb.debian.org/debian-security bookworm-security/main Translation-en [49.2 kB]
Hit:7 http://cloud.r-project.org/bin/linux/debian bookworm-cran40/ InRelease
Hit:8 http://ftp.au.debian.org/debian bookworm InRelease
Get:9 http://ftp.au.debian.org/debian bookworm-updates InRelease [52.1 kB]
Fetched 236 kB in 3s (75.6 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
203 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

```
murilo@muca10-t14:~$ sudo apt upgrade
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Calculating upgrade... Done
The following packages were automatically installed and are no longer required:
  coinor-libcbc3 coinor-libcgl1 coinor-libclp1 coinor-libcoinmp1v5 coinor-libcoinutils3v5 coinor-libosi1v5
  libabw-0.1-1 libbox2d2 libcdr-0.1-1 libe-book-0.1-1 libepubgen-0.1-1 libetonyek-0.1-1 libfreehand-0.1-1
  libmspub-0.1-1 libmwaw-0.3-3 libodfgen-0.1-1 libpagemaker-0.0-0 libqxp-0.0-0 libreoffice-base-core
  libstaroffice-0.0-0 libvisio-0.1-1 libwpd-0.10-10 libwpg-0.3-3 libwps-0.4-4 libzmf-0.0-0 lp-solve
Use 'sudo apt autoremove' to remove them.
The following NEW packages will be installed:
  libayatana-appindicator3-1 libayatana-ido3-0.4-0 libayatana-indicator3-7 libdbusmenu-glib4 libdbusmenu-gtk3-4
  linux-headers-6.1.0-13-amd64 linux-headers-6.1.0-13-common linux-image-6.1.0-13-amd64
The following packages will be upgraded:
  amd64-microcode base-files brave-browser cups cups-bsd cups-client cups-common cups-core-drivers cups-daemon
  cups-ipp-utils cups-ppdc cups-server-common curl dbus dbus-bin dbus-daemon dbus-session-bus-common
  dbus-system-bus-common dbus-user-session dbus-x11 debian-archive-keyring debianutils exim4-base exim4-config
  exim4-daemon-light firefox-esr firefox-esr-l10n-ar firefox-esr-l10n-ast firefox-esr-l10n-be firefox-esr-l10n-bg
  firefox-esr-l10n-bn firefox-esr-l10n-bs firefox-esr-l10n-ca firefox-esr-l10n-cs firefox-esr-l10n-cy
  firefox-esr-l10n-da firefox-esr-l10n-de firefox-esr-l10n-el firefox-esr-l10n-en-gb firefox-esr-l10n-eo
  firefox-esr-l10n-es-ar firefox-esr-l10n-es-cl firefox-esr-l10n-es-es firefox-esr-l10n-es-mx firefox-esr-l10n-et
  firefox-esr-l10n-eu firefox-esr-l10n-fa firefox-esr-l10n-fi firefox-esr-l10n-fr firefox-esr-l10n-ga-ie
  firefox-esr-l10n-gl firefox-esr-l10n-gu-in firefox-esr-l10n-he firefox-esr-l10n-hi-in firefox-esr-l10n-hr
  firefox-esr-l10n-hu firefox-esr-l10n-id firefox-esr-l10n-is firefox-esr-l10n-it firefox-esr-l10n-ja
  firefox-esr-l10n-kk firefox-esr-l10n-km firefox-esr-l10n-kn firefox-esr-l10n-ko firefox-esr-l10n-lt
  firefox-esr-l10n-lv firefox-esr-l10n-mk firefox-esr-l10n-mr firefox-esr-l10n-nb-no firefox-esr-l10n-ne-np
  firefox-esr-l10n-nl firefox-esr-l10n-nn-no firefox-esr-l10n-pa-in firefox-esr-l10n-pl firefox-esr-l10n-pt-br
  firefox-esr-l10n-pt-pt firefox-esr-l10n-ro firefox-esr-l10n-ru firefox-esr-l10n-si firefox-esr-l10n-sk
  firefox-esr-l10n-sl firefox-esr-l10n-sq firefox-esr-l10n-sr firefox-esr-l10n-sv-se firefox-esr-l10n-ta
  firefox-esr-l10n-te firefox-esr-l10n-th firefox-esr-l10n-tr firefox-esr-l10n-uk firefox-esr-l10n-vi
  firefox-esr-l10n-zh-cn firefox-esr-l10n-zh-tw firmware-ath9k-htc ghostscript gir1.2-gtk-3.0 grub-common
  grub-efi-amd64 grub-efi-amd64-bin grub-efi-amd64-signed grub2-common gtk-update-icon-cache libbrlapi0.8 libc-bin
  libc-dev-bin libc-devtools libc-l10n libc6 libc6-dev libc6-i386 libclang-dev libcolorcorrect5 libcups2
  libcurl3-gnutls libcurl4 libcurl4-openssl-dev libdbus-1-3 libgs-common libgs10 libgs10-common libgssapi-krb5-2
  libgtk-3-0 libgtk-3-bin libgtk-3-common libgtk-4-1 libgtk-4-bin libgtk-4-common libhwy1 libk5crypto3
  libkf5plasma5 libkf5plasmaquick5 libkfontinst5 libkfontinstui5 libkrb5-3 libkrb5support0 libkworkspace5-5
  libldb2 libmariadb-dev libmariadb-dev-compat libmariadb3 libnftables1 libnotificationmanager1 libnss-myhostname
  libpam-modules libpam-modules-bin libpam-runtime libpam-systemd libpam0g libplasma-geolocation-interface5
  libqt5positioning5 libsmbclient libssl-dev libssl3 libsystemd-shared libsystemd0 libtaskmanager6abi1 libudev1
  libunbound8 libvpx7 libwbclient0 libweather-ion7 libwebp-dev libwebp7 libwebpdemux2 libwebpmux3 libx11-6
  libx11-data libx11-dev libx11-xcb1 libxpm4 linux-compiler-gcc-12-x86 linux-headers-6.1.0-10-amd64
  linux-headers-6.1.0-10-common linux-headers-amd64 linux-image-amd64 linux-kbuild-6.1 linux-libc-dev locales
  mariadb-client-core mariadb-common mariadb-server-core nftables openssh-client openssl plasma-framework
  plasma-workspace plasma-workspace-data plasma-workspace-wayland python3-brlapi python3-ldb python3-samba
  samba-common samba-common-bin samba-dsdb-modules samba-libs sddm-theme-breeze sddm-theme-debian-breeze smbclient
  spotify-client systemd systemd-coredump systemd-sysv udev xbrlapi
203 upgraded, 8 newly installed, 0 to remove and 0 not upgraded.
Need to get 532 MB of archives.
After this operation, 492 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
```

GESTORES DE PAQUETES: PROBLEMAS

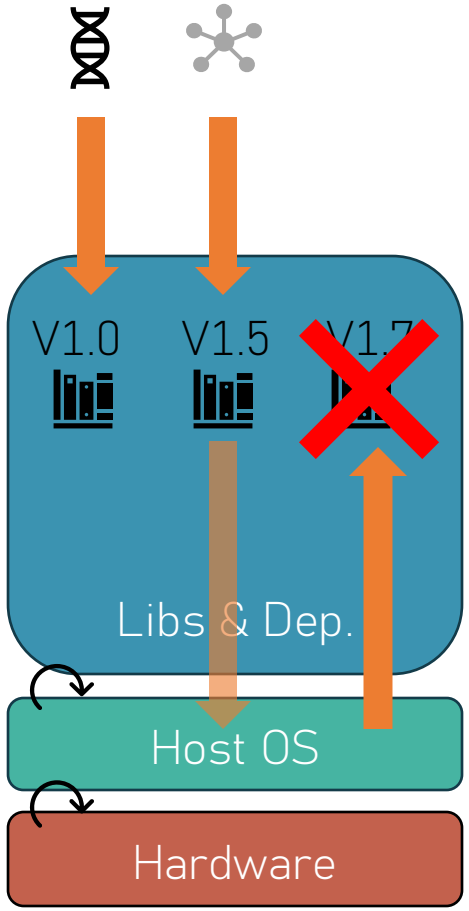
- ¡El usuario debe ser parte de los administradores! ¡SUDO!

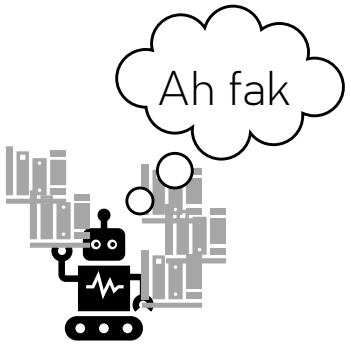
- Siempre es arriesgado porque los programas instalados ya utilizan librerías que están configuradas

- En algunas situaciones, como cuando se utilizan HPC, no hay forma de utilizar "sudo"

- Gestionar las dependencias del paquete: a menudo, diferentes proyectos requerirán diferentes versiones del mismo software.

    - Raramente, pero no imposible, un mismo proyecto requiere diferentes versiones de un mismo programa





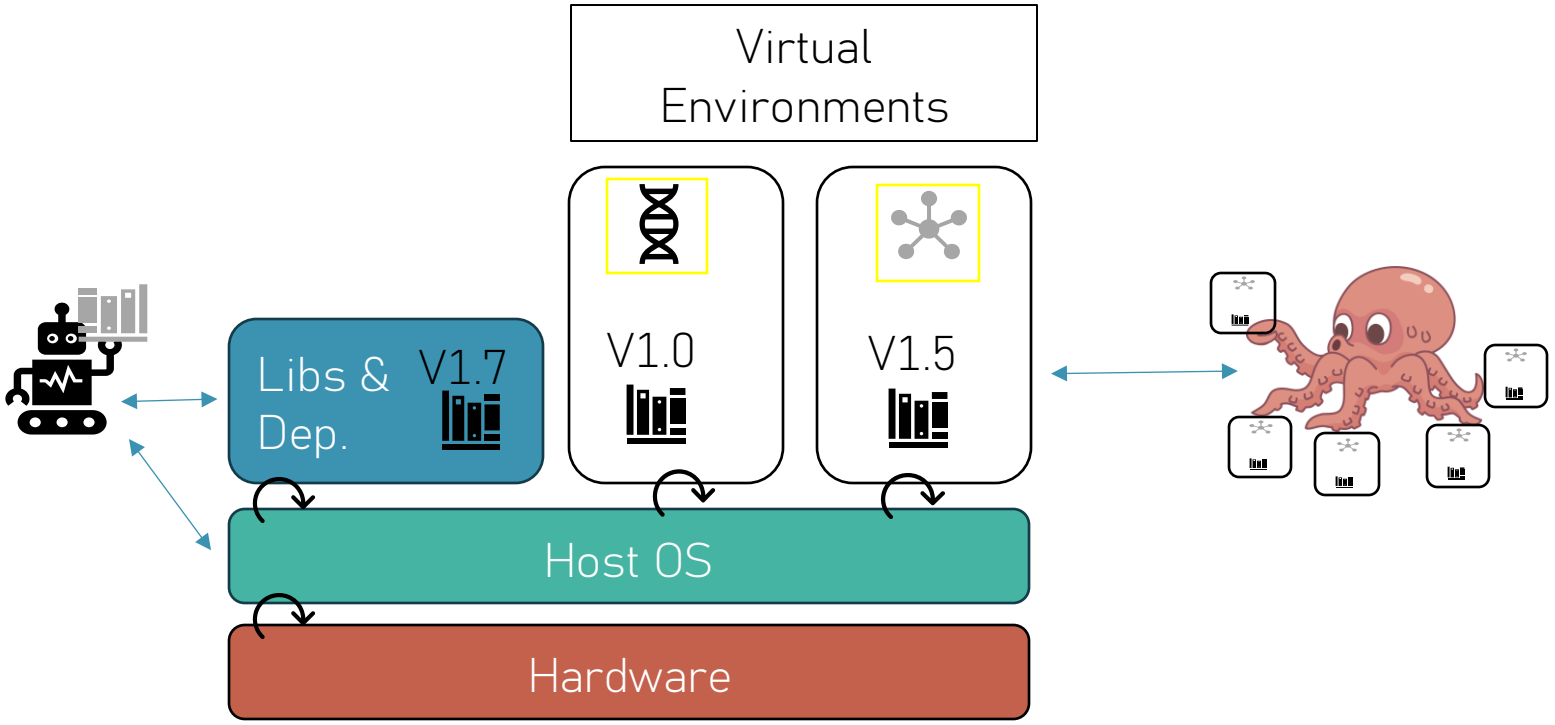
UTILIZAR ENTORNOS VIRTUALES TE PERMITE INSTALAR LA VERSIÓN QUE NECESITAS PARA TU PROYECTO.

- **Reproducibilidad mejorada**: Usar un entorno virtual permite un seguimiento preciso del software y sus versiones, facilitando la recreación del entorno.

- **Sistema más limpio y ligero**: Evitar múltiples instalaciones previene el caos y mantiene el sistema operativo ágil.

- ¡El usuario no necesita ser parte de los administradores! ¡SUDO!

Software que solo utilizaremos adentro el entorno virtual



Conda: Sistema de gestión de paquetes y entornos para lenguajes variados.

- **Compatibilidad**: Windows, macOS, y Linux.
- **Funcionalidad**: Instalación, ejecución y actualización ágil de paquetes y sus dependencias.
- **Versatilidad**: Originalmente para Python, pero extensible a lenguajes como R, Ruby, Java y más.
- **Gestión de entornos**: Facilita la creación, guardado y cambio entre diferentes entornos locales.
- **Flexibilidad**: Permite manejar diversas versiones de Python sin cambiar de gestor de entornos.

Mamba: Reimplementación del gestor de paquetes conda en C++ (para máxima eficiencia).

- Descarga paralela de datos de repositorio y archivos de paquetes con multi-threading.
- Uso de **libsolv** para resolución de dependencias mucho más rápida
- Al mismo tiempo, mamba utiliza el mismo analizador de línea de comandos, código de instalación y desinstalación de paquetes y rutinas de verificación de transacciones que conda para mantener la máxima compatibilidad.

https://docs.conda.io/en/latest/

https://github.com/mamba-org/mamba

INSTALACIÓN RÁPIDA DESDE LA LÍNEA DE COMANDOS

Ingrese al sitio web https://docs.conda.io/projects/miniconda/en/latest/index.html

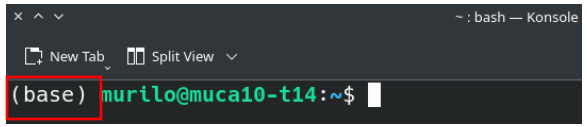
Windows macOS Linux

These four commands quickly and quietly install the latest 64-bit version of the installer and then clean up after themselves. To install a different version or architecture of Miniconda for Linux, change the name of the `.sh` installer in the `wget` command.

```
mkdir -p ~/miniconda3
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O ~/miniconda3/miniconda.sh
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
rm -rf ~/miniconda3/miniconda.sh
```

After installing, initialize your newly-installed Miniconda. The following commands initialize for bash and zsh shells:

```
~/miniconda3/bin/conda init bash
~/miniconda3/bin/conda init zsh
```



En general no es recomendable instalar nada en nuestro entorno "base"

Excepto la biblioteca mamba

De forma predeterminada, cuando abres la terminal, el entorno "base" siempre está activo. Podemos cambiar esto con:

```
conda config --set auto_activate_base false
```

USANDO A CONDA

- Para agregar un "channels" a conda:

```
conda config –add channels conda-forge
conda config –add channels bioconda
```

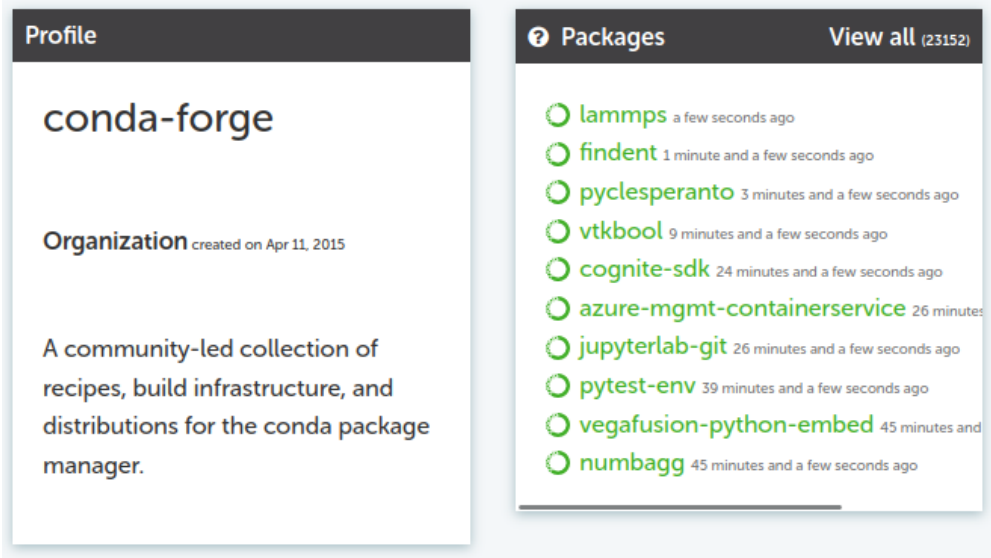
- Para agregar el paquete mamba que usaremos para los demás casos:

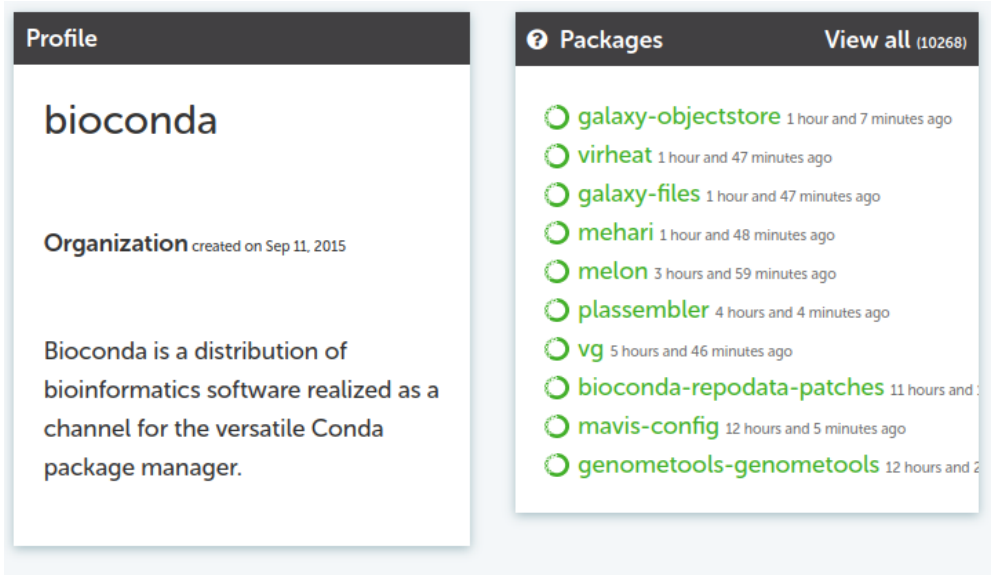
```
conda install -c conda-forge mamba
```

  - -c para es un parámetro con el nombre del channel que usaremos

    `-c conda-forge`

- conda se instala desde "channels"





Para buscar más paquetes: https://anaconda.org/

USANDO A CONDA

- Para crear un entorno, utilice mamba create
  `mamba create -n {your-env-name}`

- Para activar/desactivar un entorno, utilice conda activate/deactivate
  `conda activate {your-env-name}`
  `conda deactivate`

- para instalar paquetes en un entorno activo, use mamba install
  `mamba install {package-name}`
  o especificando el channel
  `mamba install -c conda-forge mamba`

- Para listar los entornos disponibles
  `conda env list`

- Para remover a un entorno
  `conda env remove {your-env-name}`

- Para instalar paquetes en un entorno específico no activo, agregue el indicador "-n", seguido del nombre del entorno.
  `mamba install -n {YOUR-ENV-NAME} -c {CHANNEL-NAME} {PACKAGE-NAME}`

- Para desinstalar paquetes en un entorno activo
  `mamba uninstall {PACKAGE-NAME}`

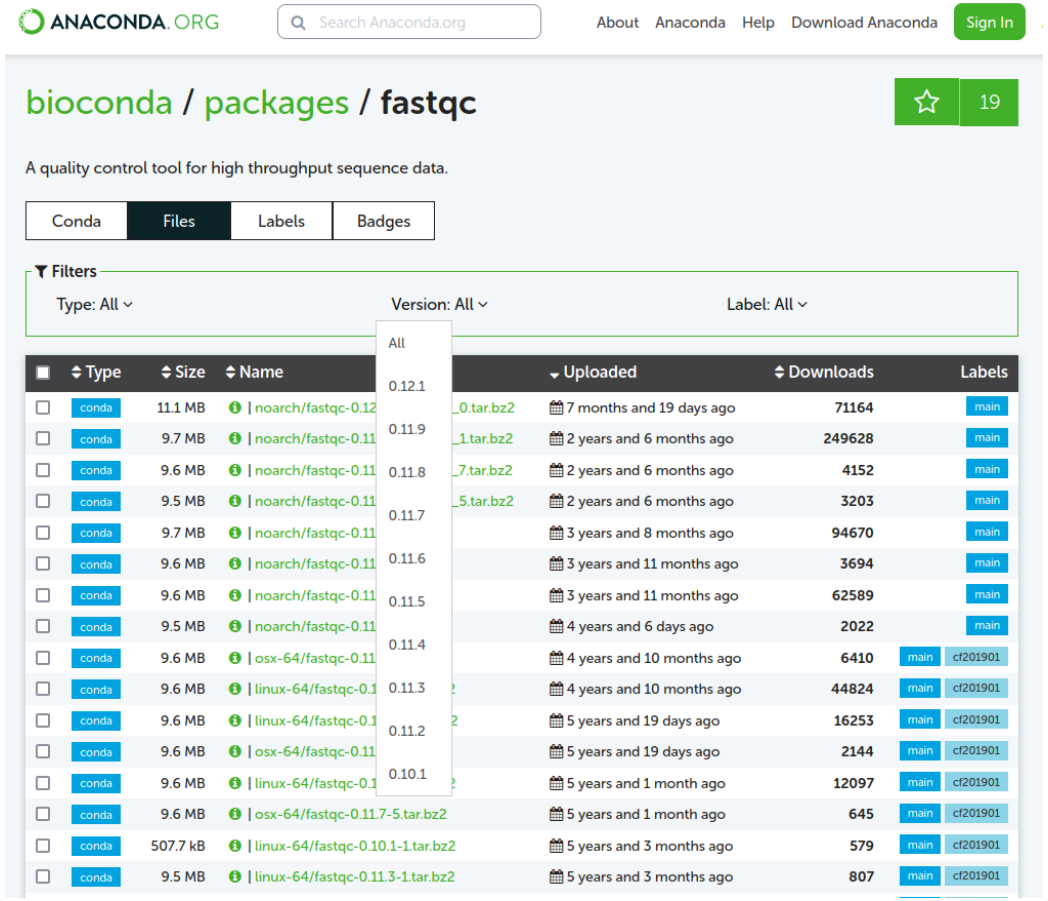
{YOUR-ENV-NAME} = ENTRENO, **ENTRENO2**
{CHANNEL-NAME} = BIOCONDA
{PACKAGE-NAME} = FASTQC, MAFFT, BWA

Practica: crear dos entornos, ver la lista de entornos disponibles, activarlos, remover el entorno 2, remover los paquetes del entorno uno menos a fastqc

USANDO A CONDA

- También podemos controlar la versión de los paquetes que instalamos

```
mamba create –n {YOUR-ENV-NAME}  -c {CHANNEL-NAME} {PACKAGE-NAME}={VERSION}
```



You can **install packages** when creating a new environment by specifying the package name(s) afterwards

```
mamba create -n {your-env-name} {package-name-1} {package-name-2}
```

You can install a **specific version** by adding the '=' sign, followed by the version (no spaces)

```
mamba create -n {your-env-name} {package-name}={version}
mamba install {package-name}={version}
```

You can install from a **specific channel** by using the '-c' flag, followed by the channel

```
mamba create -n {your-env-name} -c {a-channel} {package-name}
mamba install -c {a-channel} {package-name}
```

Here are a couple of examples:

```
mamba create -n  my-proj  -c conda-forge  python=3.8  pandas
mamba install -c conda-forge  python=3.8  pandas
```

¿QUÉ ES EXPORTAR ENTORNOS CONDA Y POR QUÉ HACERLO?

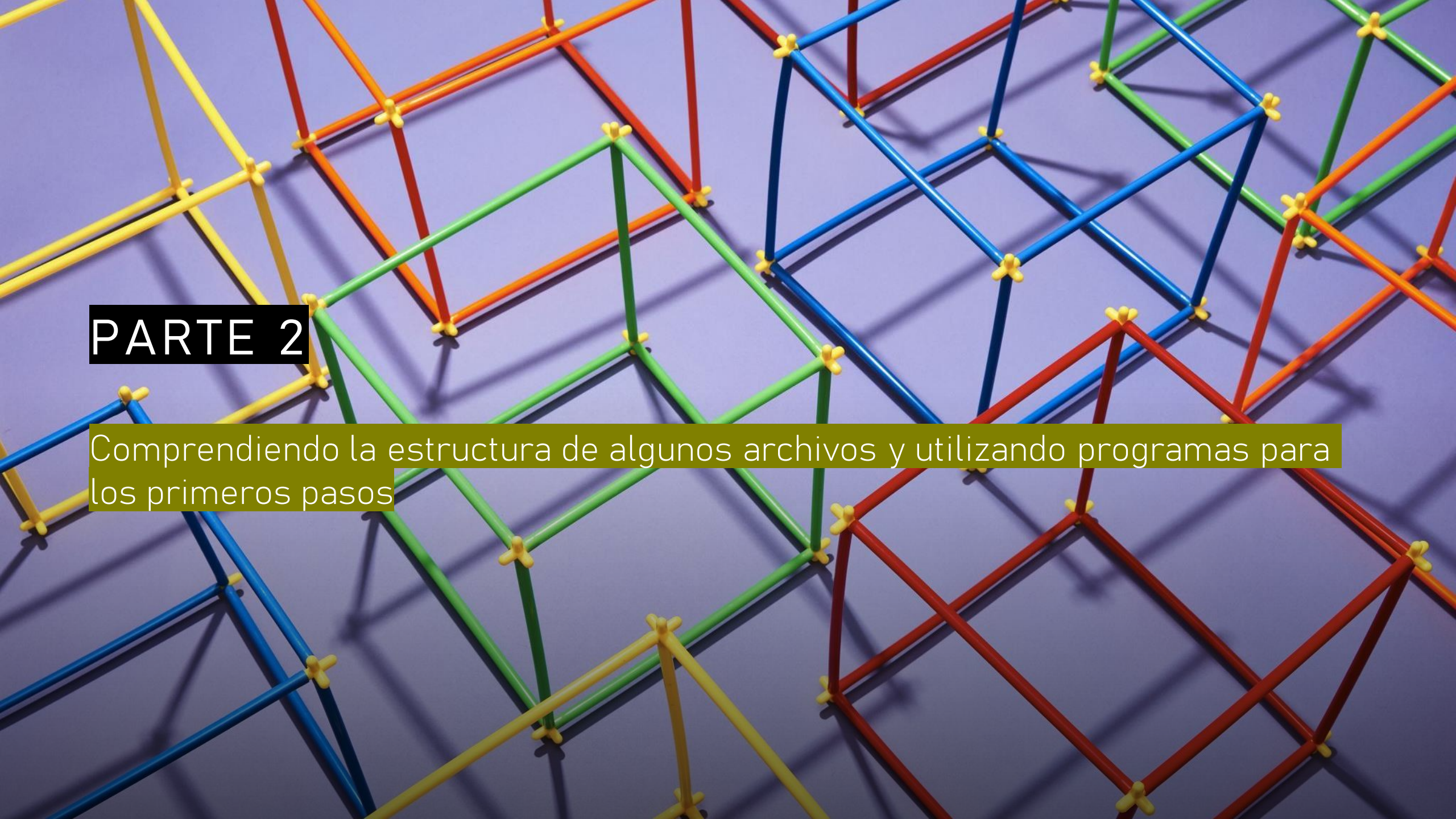
- Exportar un entorno crea un archivo (yaml) que lista todos los paquetes y versiones que están instalados en ese entorno.
- Este archivo se puede guardar en un repositorio de tu código de proyecto para que otros puedan crear fácilmente un entorno que ejecutará exitosamente tu código.
- También puede ser utilizado por ti para recrear el entorno (en caso de que no hayas trabajado en el proyecto durante meses).

{YOUR-ENV-NAME} = ENTRENO3
{CHANNEL-NAME} = BIOCONDA
{PACKAGE-NAME} = FASTQC, MAFFT, BWA

Practica: crear un entornoA, instalar paquetes, exportar este entorno y crear un entornoB con el yaml

- Para exportar un entorno activo, utilice
`mamba env export > {mi-entorno}.yaml`
- Para recrear un entorno de un archivo yaml, utilice
`mamba env create -f {mi-entorno}.yaml`

```
name: snakemake
channels:
  - bioconda
  - conda-forge
  - defaults
dependencies:
  - _libgcc_mutex=0.1=conda_forge
  - _openmp_mutex=4.5=2_gnu
  - aioeasywebdav=2.4.0=pyha770c72_0
  - aiohttp=3.8.5=py311h459d7ec_0
  - aiosignal=1.3.1=pyhd8ed1ab_0
  - amply=0.1.6=pyhd8ed1ab_0
  - appdirs=1.4.4=pyh9f0ad1d_0
  - async-timeout=4.0.3=pyhd8ed1ab_0
  - attmap=0.13.2=pyhd8ed1ab_0
  - attrs=23.1.0=pyh71513ae_1
  - backports=1.0=pyhd8ed1ab_3
  - backports.functools_lru_cache=1.6.5=pyhd8ed1ab_0
  - bcrypt=4.0.1=py311h46250e7_1
  - boto3=1.28.55=pyhd8ed1ab_0
  - botocore=1.31.55=pyhd8ed1ab_0
```



PARTE 2

Comprendiendo la estructura de algunos archivos y utilizando programas para los primeros pasos

MODELO DE ARCHIVO: FASTA

- Una secuencia comienza con un carácter mayor que (">") seguido o no de una descripción de la secuencia (todo en una sola línea).
- Las siguientes líneas inmediatamente después de la línea de descripción son la representación de la secuencia.
- Una letra por aminoácido o ácido nucleico y, por lo general, no tienen más de 80 caracteres de longitud.

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
```

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

https://blast.ncbi.nlm.nih.gov/doc/blast-topics/ y https://en.wikipedia.org/wiki/FASTA_format

MODELO DE ARCHIVO: (MULTI)FASTA

- Multifasta: un único archivo que contiene varias secuencias biológicas.
- Tenga en cuenta la falta de patrón en los nombres y descripciones de cada secuencia.
- Lo ideal sería no utilizar espacios y elegir algún carácter como separador.
    - facilita la construcción de scripts en lenguaje de programación para manipular estos datos

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

MODELO DE ARCHIVO: FASTA

código de letras

convenciones para facilitar la identificación

```
A  adenosine          C  cytidine             G  guanine
T  thymidine          N  A/G/C/T (any)        U  uridine
K  G/T (keto)         S  G/C (strong)         Y  T/C (pyrimidine)
M  A/C (amino)        W  A/T (weak)           R  G/A (purine)
B  G/T/C              D  G/A/T                H  A/C/T
V  G/C/A              -  gap of indeterminate length
```

```
A  alanine              P  proline
B  aspartate/asparagine Q  glutamine
C  cystine              R  arginine
D  aspartate            S  serine
E  glutamate            T  threonine
F  phenylalanine        U  selenocysteine
G  glycine              V  valine
H  histidine            W  tryptophan
I  isoleucine           Y  tyrosine
K  lysine               Z  glutamate/glutamine
L  leucine              X  any
M  methionine           *  translation stop
N  asparagine           -  gap of indeterminate length
```

Extension	Meaning	Notes
fasta, fa	generic FASTA	Any generic fasta file. See below for other common FASTA file extensions
fna	FASTA nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	FASTA amino acid	Contains amino acid sequences. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

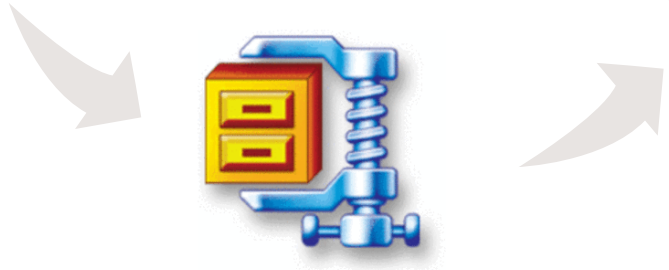
MODELO DE ARCHIVO: FASTQ

Un archivo FASTQ tiene cuatro líneas con cuatro campos separados por secuencia:

- El campo 1 comienza con un carácter '@' y va seguido de un identificador de secuencia.
- El campo 2 son las letras de secuencia.
- El campo 3 contiene un carácter '+'.
- El campo 4 codifica los valores de calidad para la secuencia en el campo 2 y debe contener la misma cantidad de símbolos que letras en la secuencia.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

En general, los archivos FASTQ pueden contener desde miles hasta millones de lecturas.



*fastq.gz

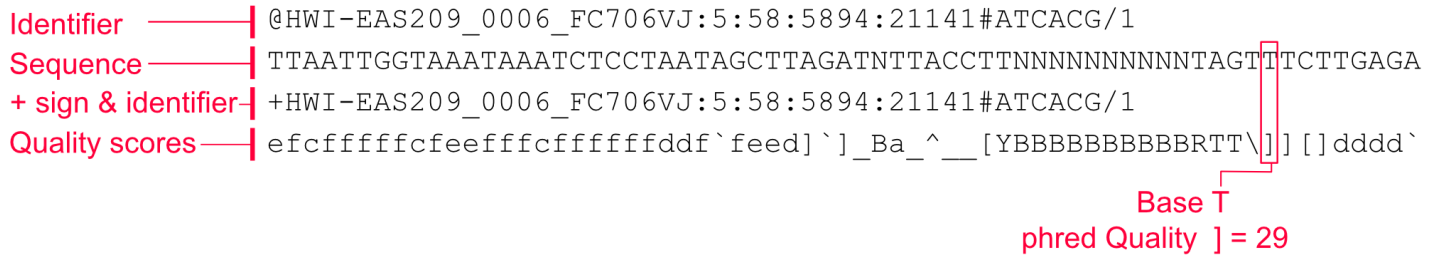
https://compgenomr.github.io/book/fasta-and-fastq-formats.html y https://en.wikipedia.org/wiki/FASTQ_format

MODELO DE ARCHIVO: FASTQ

Element	Requirements	Description
`@`	@	Each sequence identifier line starts with @
`<instrument>`	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID
`<run number>`	Numerical	Run number on instrument
`<flowcell ID>`	Characters allowed: a–z, A–Z, 0–9	Flowcell ID
`<lane>`	Numerical	Lane number
`<tile>`	Numerical	Tile number
`<x_pos>`	Numerical	X coordinate of cluster
`<y_pos>`	Numerical	Y coordinate of cluster
`<read>`	Numerical	Read number. 1 can be single read or Read 2 of paired-end
`<is filtered>`	Y or N	Y if the read is filtered (did not pass), N otherwise
`<control number>`	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X systems, control specification is not performed and this number is always 0.
`<sample number>`	Numerical	Sample number from sample sheet

identificador de una lectura Illumina

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<sample number>



¡Esta información es utilizada posteriormente por otros programas!

https://help.basespace.illumina.com/files-used-by-basespace/fastq-files

DESCARGAR LOS ARCHIVOS DE SECUENCIACIÓN EJEMPLO

- wget -q -O - "https://trace.ncbi.nlm.nih.gov/Traces/sra-reads-be/fastq?acc=SRR24053947" > SRR24053947.fastq.gz
- wget -q -O - "https://trace.ncbi.nlm.nih.gov/Traces/sra-reads-be/fastq?acc=SRR24054051" > SRR24054051.fastq.gz
- wget -q -O - "https://trace.ncbi.nlm.nih.gov/Traces/sra-reads-be/fastq?acc=SRR24053926" > SRR24053926.fastq.gz

USAREMOS EN DOS CLASES

ALGUNAS COMBINACIONES DE COMANDOS INTERESANTES PARA EVALUAR RÁPIDAMENTE ARCHIVOS FASTA Y FASTQ

Visualizar contenido de archivos

a. Usar cat para visualizar el contenido de ejemplo.fasta

```
cat ejemplo.fasta
```

b. Usar zcat para visualizar el contenido de ejemplo.fastq.gz

```
zcat ejemplo.fastq.gz
```

Visualizar las primeras y últimas líneas de un archivo

a. Usar head para ver las primeras 4 líneas de ejemplo.fastq

```
head -n 4 ejemplo.fasta
```

b. Usar tail para ver las últimas 4 líneas de ejemplo.fastq

```
tail -n 4 ejemplo.fasta
```

¿Qué pasa si intentamos lo siguiente?

```
head -n 4 ejemplo.fastq.gz
```

Intenta:

```
zcat ejemplo.fastq.gz | head -n 4
```

ALGUNAS COMBINACIONES DE COMANDOS INTERESANTES PARA EVALUAR RÁPIDAMENTE ARCHIVOS FASTA Y FASTQ

Contar cantidad de secuencias en un archivo fasta
Podemos contar cuántas hay:

```
grep -c '^>' ejemplo.fasta
```

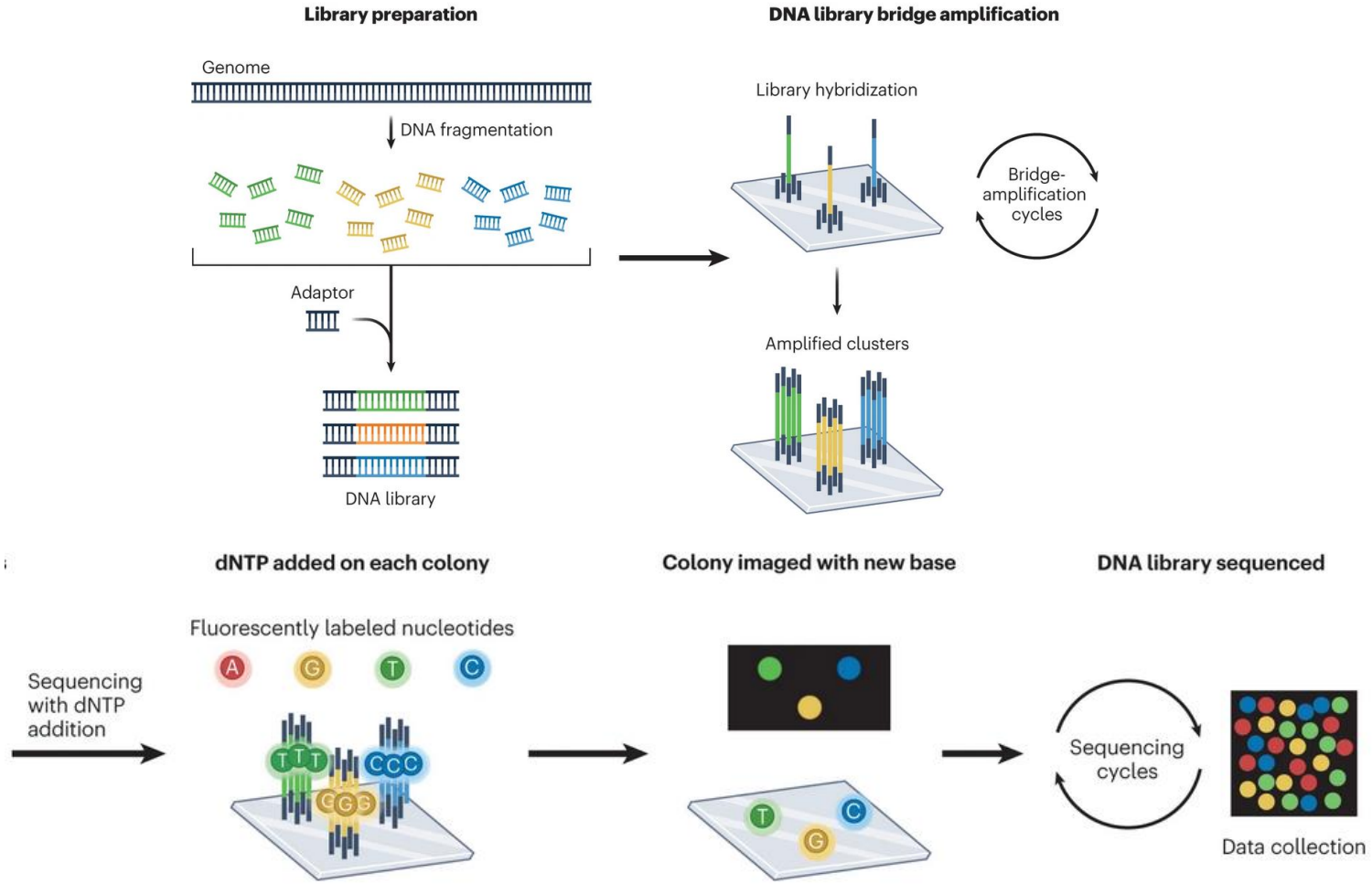
O podemos visualizar tolas las líneas de encabezamiento

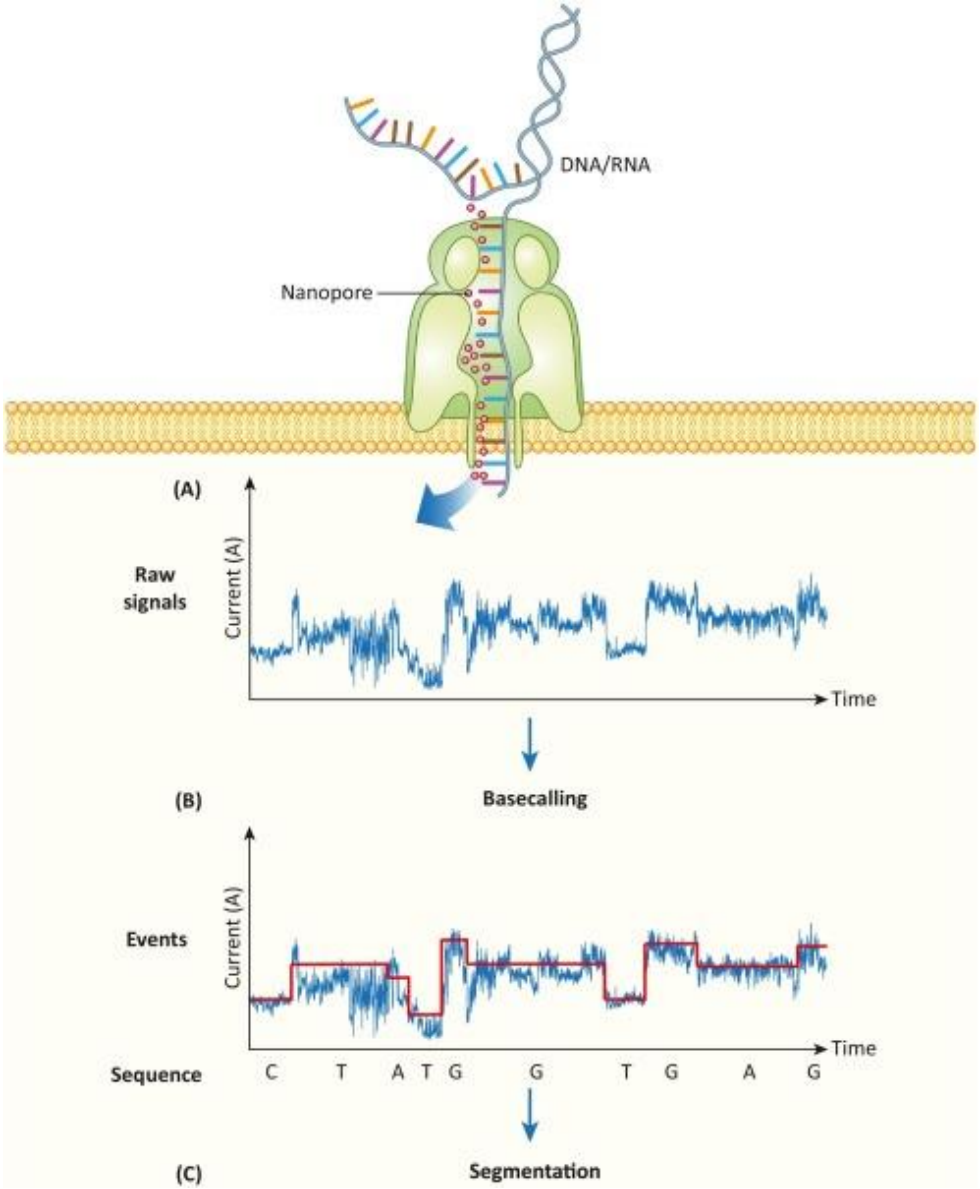
```
grep '^>' ejemplo.fasta
```

De manera análoga
a. Usar head para ver las primeras 4 líneas de ejemplo.fastq

```
zcat ejemplo.fastq.gz | grep -c '^@'
```

¿DE DÓNDE PROVIENE LA INFORMACIÓN DE CALIDAD DE LAS 'LECTURAS'?





https://doi.org/10.1038/s41587-023-01986-3

https://doi.org/10.1016/j.tig.2021.09.001

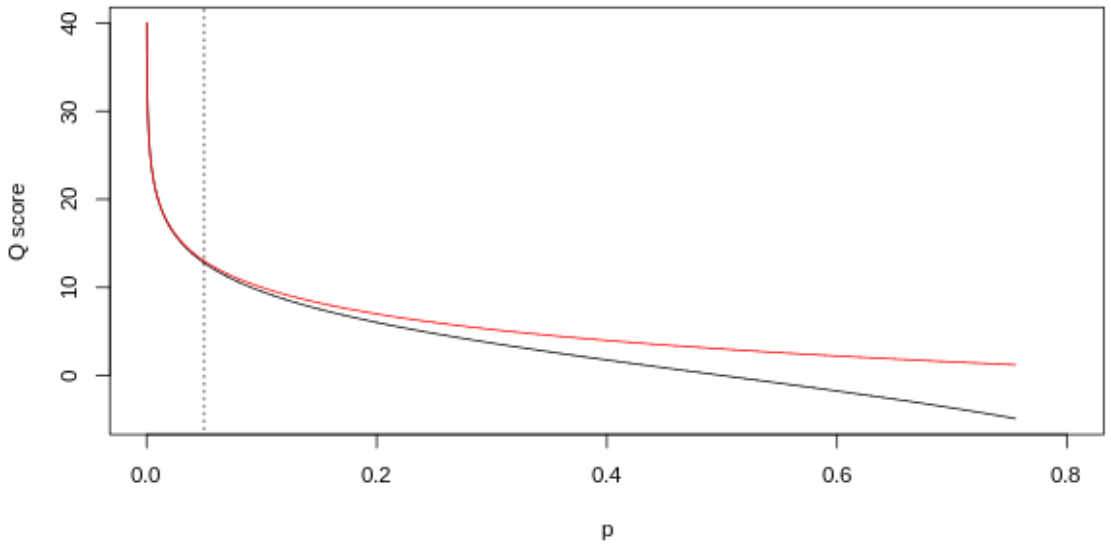
LIDIAR CON LA INCERTIDUMBRE

Dada una afirmación, A, el puntaje de calidad, Q(A), expresa la probabilidad de que A no sea cierta, P(~A), según la relación:

$$Q(A) = -10 \log_{10}(P(\sim A))$$

donde P(~A) es la probabilidad estimada de que una afirmación A sea incorrecta.

Quality score, Q(A)	Error probability, P(~A)
10	0.1
20	0.01
30	0.001



Symbol	ASCII Code	Q-Score			
!	33	0	?	63	30
"	34	1	@	64	31
#	35	2	A	65	32
$	36	3	B	66	33
%	37	4	C	67	34
&	38	5	D	68	35
'	39	6	E	69	36
(	40	7	F	70	37
)	41	8	G	71	38
*	42	9	H	72	39
+	43	10	I	73	40

https://help.basespace.illumina.com/files-used-by-basespace/quality-scores

FASTQC

Babraham Institute

Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

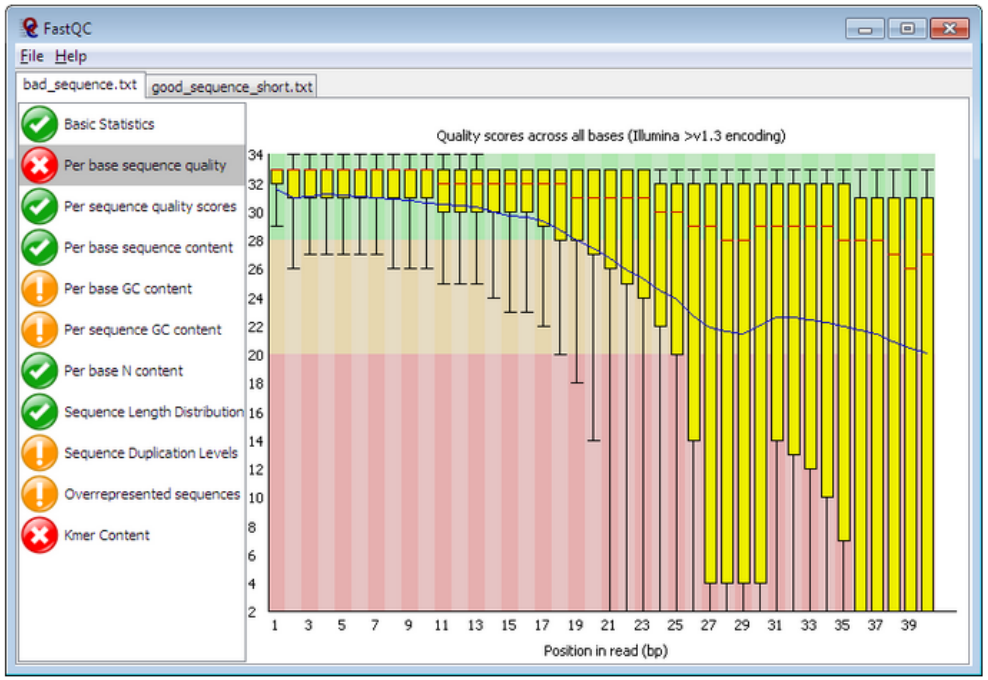
FastQC

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

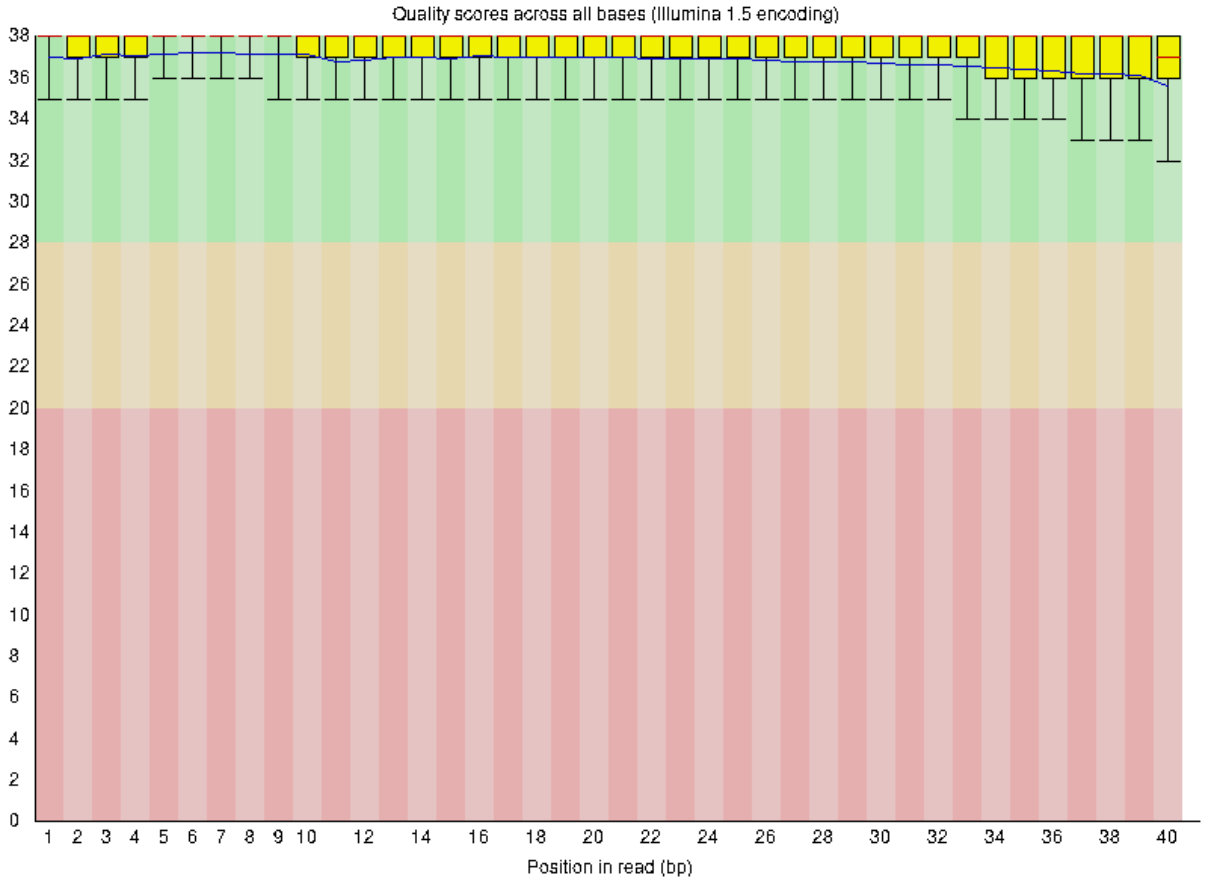
- FastQC tiene como objetivo proporcionar una manera sencilla de realizar algunas verificaciones de control de calidad en datos de secuencia crudos provenientes de pipelines de secuenciación de alto rendimiento.
- Ofrece un conjunto modular de análisis que puedes usar para obtener una rápida impresión de si tus datos tienen algún problema del cual deberías estar al tanto antes de realizar cualquier análisis adicional.

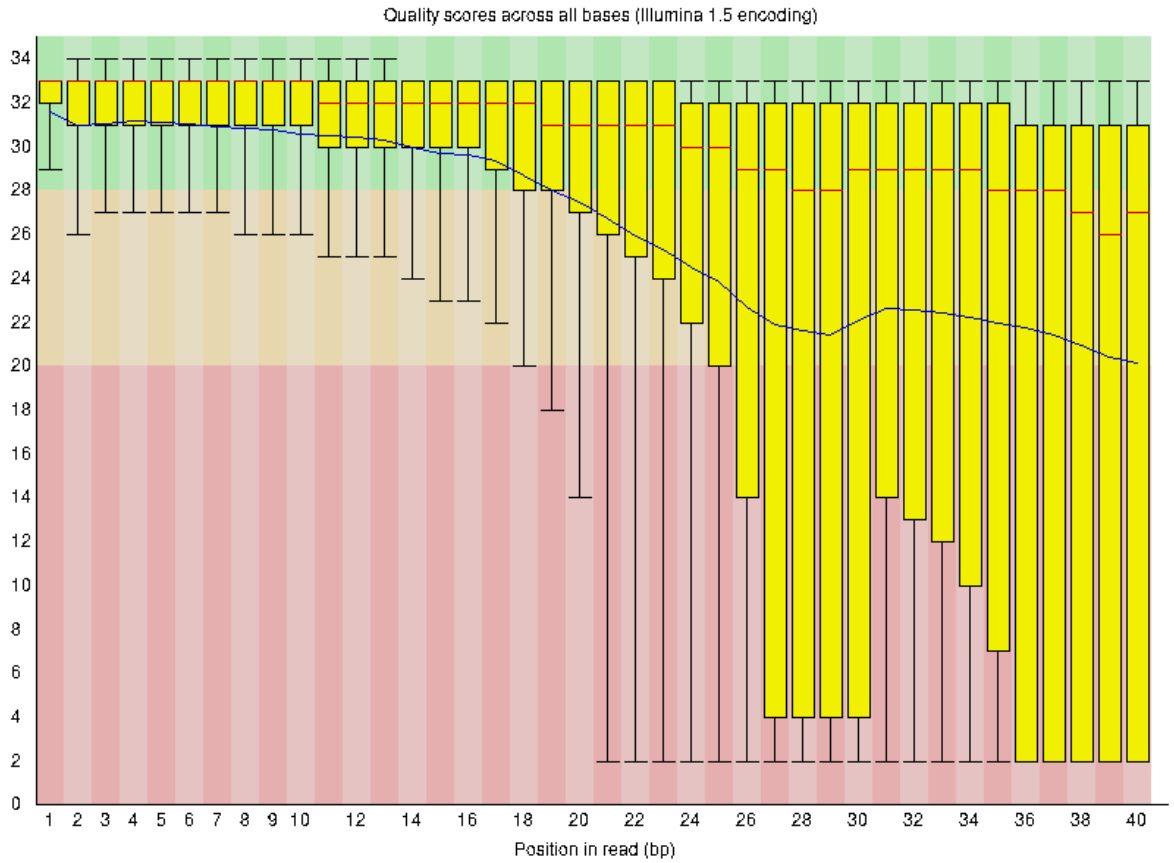
¡12 conjuntos de estadísticas en informes gráficos!



FASTQC

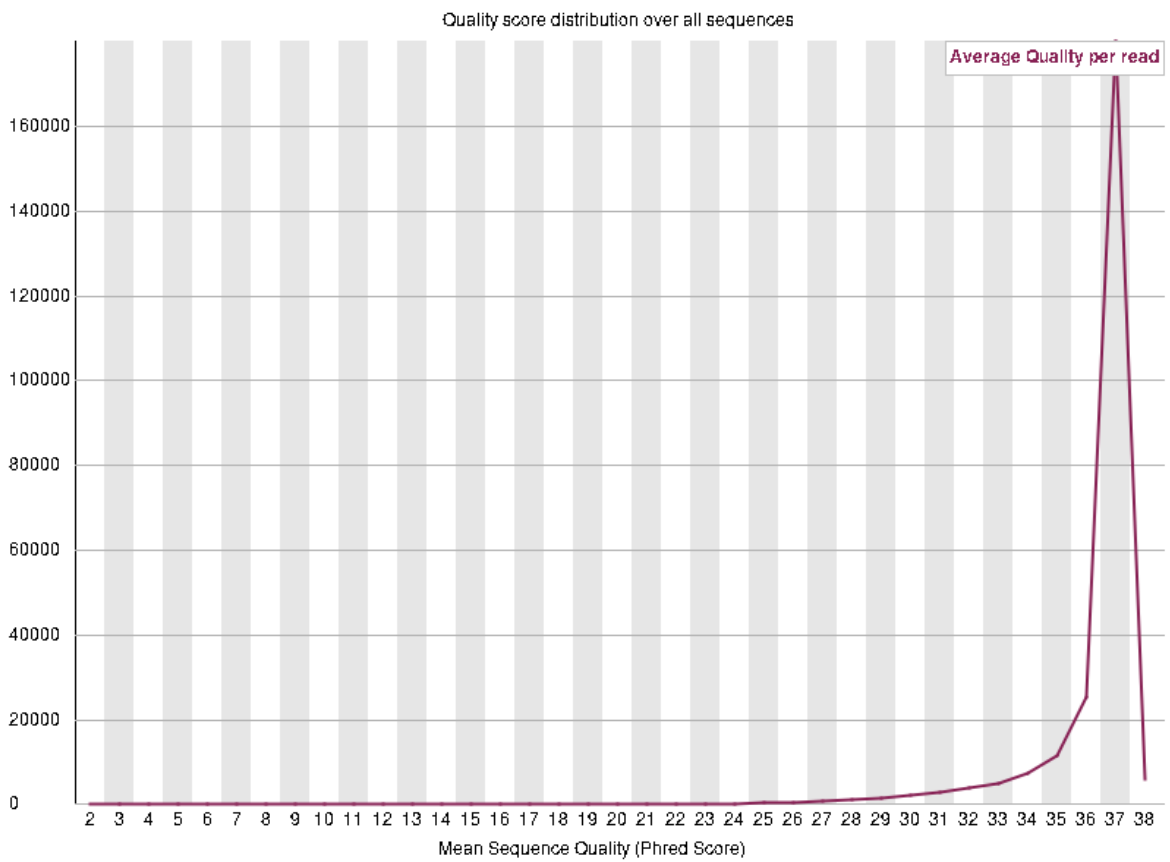
Calidad de la secuencia por base: Muestra la distribución de scores de calidad a través de todas las posiciones en las lecturas. Se espera que la calidad disminuya hacia el final de las lecturas, pero grandes desviaciones pueden indicar problemas.

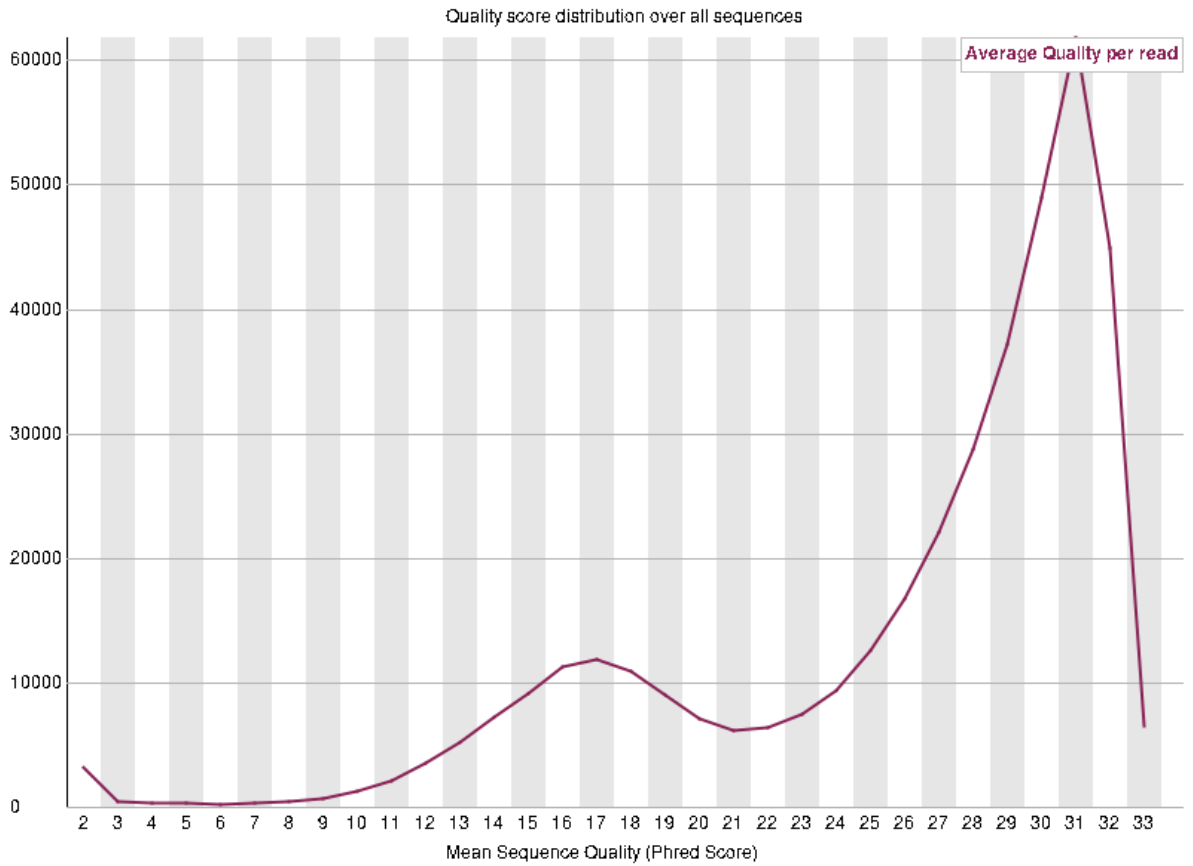




FASTQC

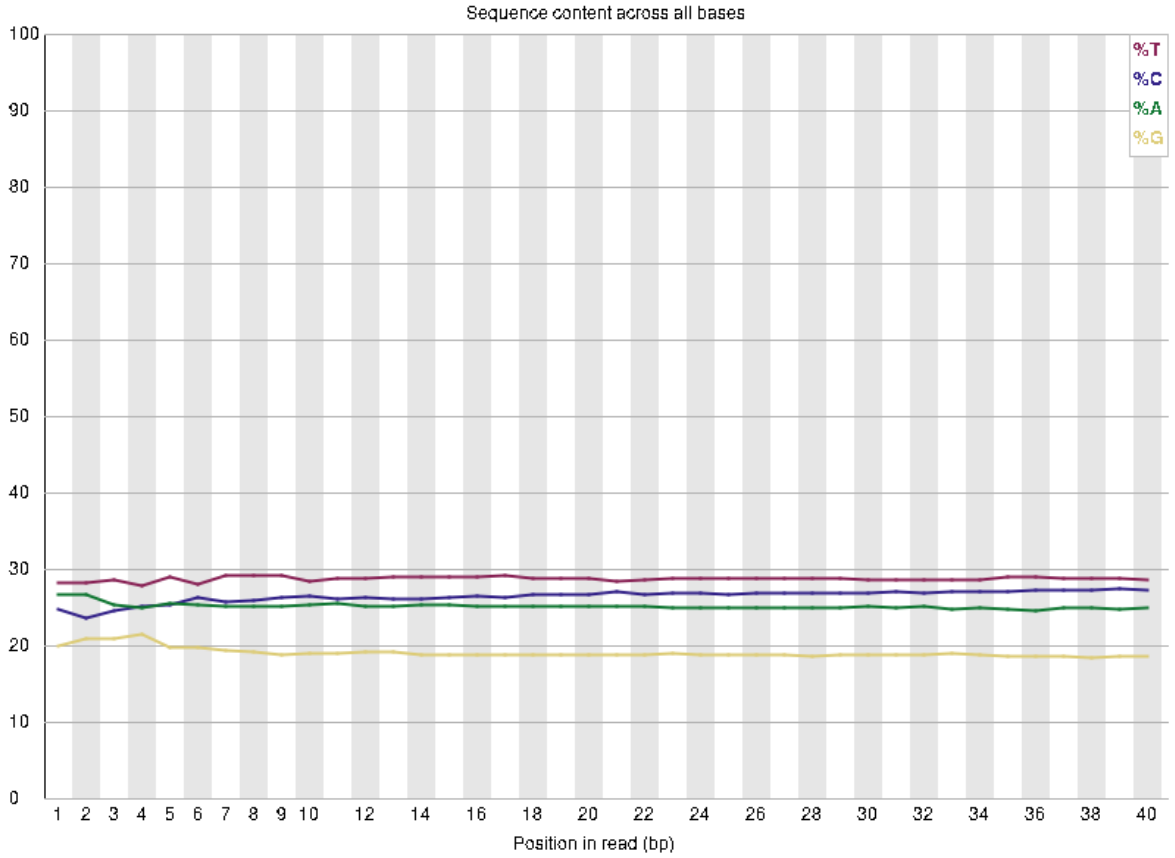
Distribución de calidad de la secuencia: Proporciona una visión general de la calidad de todo el archivo, no solo por base. Se esperaría ver la mayoría de las lecturas con una calidad alta.

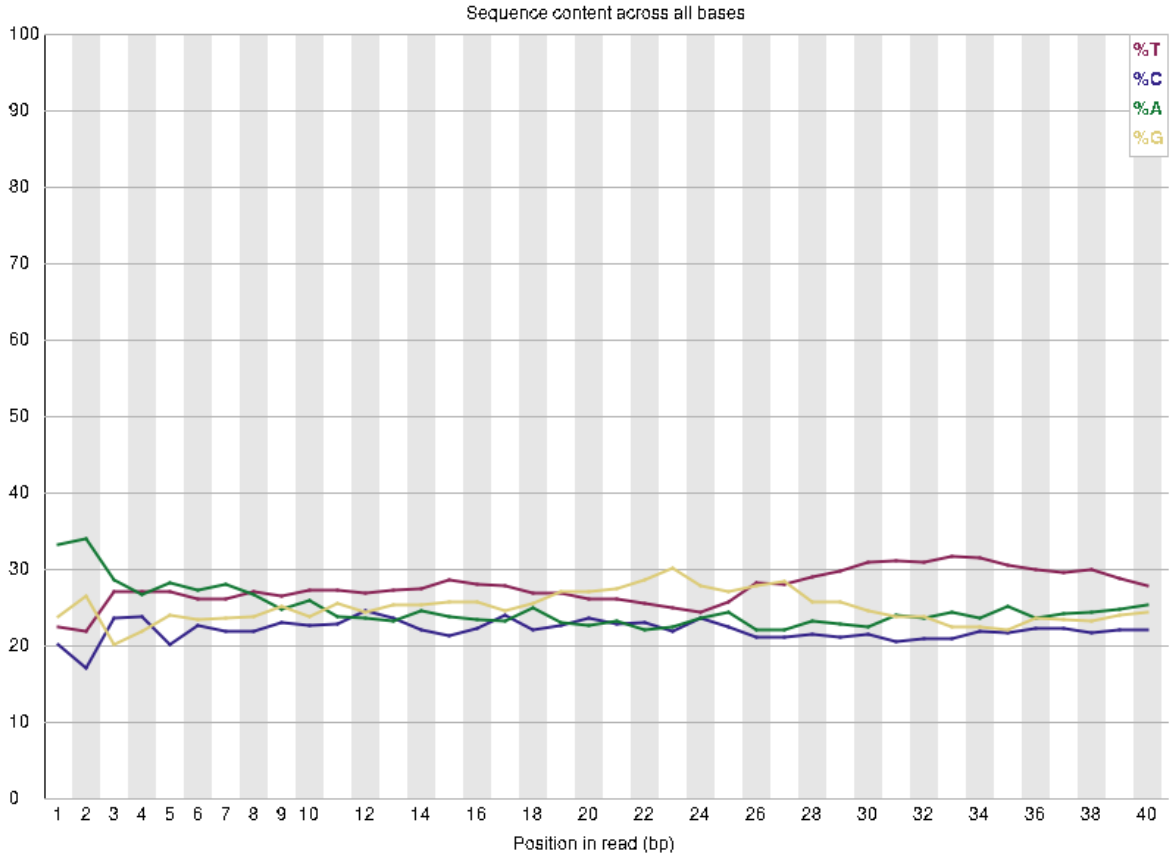




FASTQC

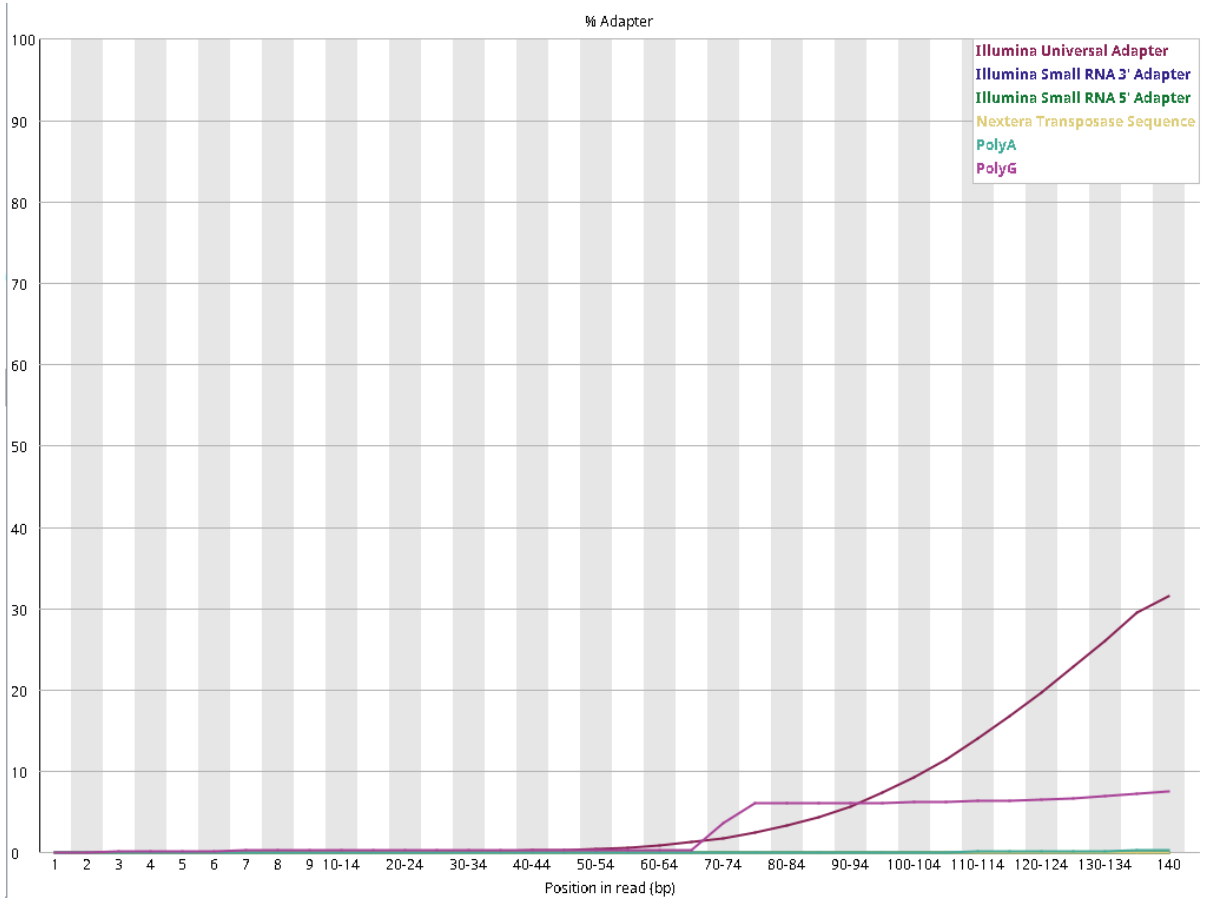
Contenido de base por base: En una biblioteca sin sesgo, el contenido de base debería ser constante. Variaciones pueden deberse a contaminación con adaptadores o a otros tipos de sesgo.





FASTQC

Contaminación por adaptadores: Si no se han eliminado los adaptadores, FastQC identificará las secuencias de adaptadores más comunes.



Contaminación por sobrerrepresentación de secuencias: Las secuencias que aparecen con mucha frecuencia pueden ser indicativas de contaminación o secuencias altamente conservadas.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGC	532	0.1345854162028698	No Hit
CTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCGC	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTTG	411	0.10397482341988626	No Hit

FASTQC

- Basado en GUI
  Abra la terminal, active el entorno virtual que creamos con fastqc y escriba fastqc para iniciar el programa. En "archivo" -> "abrir..."

- Basado en línea de comando
  Vaya al directorio donde se encuentra un archivo fastqc (use cd, ls y pwd para navegar si es necesario).
  Escriba fastqc --help y vea las opciones para ejecutar el programa a través de la línea de comando.

  Intentar:

  fastqc --threads 8 "archivo.fastq.gz"

MODELO DE ARCHIVO: SAM/BAM

- El Sequence Alignment/Map (SAM) es un formato de archivo para guardar información de alineación de lecturas cortas mapeadas contra secuencias de referencia.
- Normalmente comienza con una sección de encabezado seguida de información de alineación en líneas separadas por tabuladores para cada lectura.

Header section

```
@HD    VN:1.3    SO:coordinate
@SQ    SN:conticA    LN:443
@SQ    SN:contigB    LN:1493
@SQ    SN:contigC    LN:328
```

Tab-delimited read alignment information lines

https://www.metagenomics.wiki/tools/samtools/bam-sam-file-format

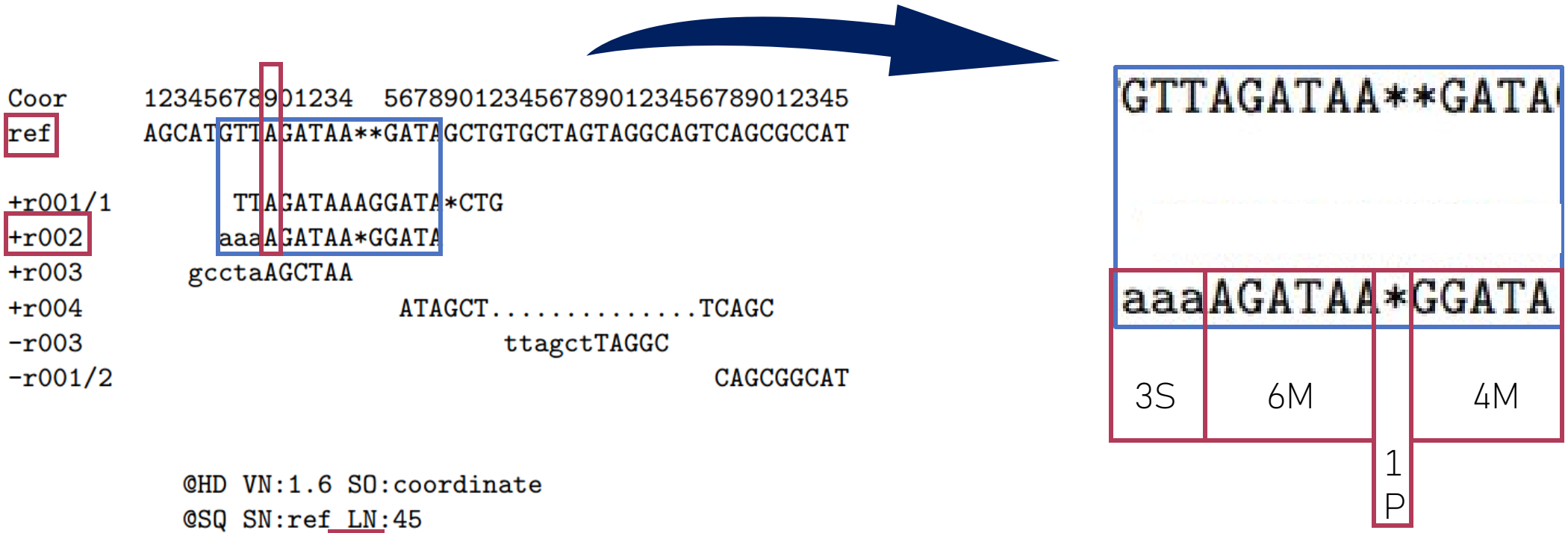
MODELO DE ARCHIVO: SAM

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGGCAT
```

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

http://samtools.github.io/hts-specs/SAMv1.pdf

MODELO DE ARCHIVO: SAM

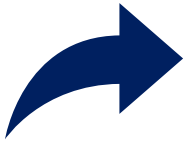



```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003     0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

http://samtools.github.io/hts-specs/SAMv1.pdf

MODELO DE ARCHIVO: SAM

Las líneas de encabezado siempre comenzarán con un símbolo "@" seguido de un identificador que indica el tipo y subtipo de la línea de encabezado. Algunos de los ejemplos más comunes pueden verse de la siguiente manera:



@SQ lleva cada una de las secuencias del conjunto de referencia, en el que mapeamos las lecturas

```
@SQ SN:chr14 LN:107349540
@PG ID:bwa PN:bwa VN:0.7.7-r441 CL:bwa mem ref/seq.fa r1.fastq r2.fastq
```

@PG describe el programa utilizado para generar el archivo SAM (y por tanto, realizar la alineación). Si se combinan varios archivos SAM, es posible que aparezcan varias líneas @PG, aunque es común tener solo una.

https://www.zymoresearch.de/blogs/blog/what-are-sam-and-bam-files

MODELO DE ARCHIVO: SAM

Información de alineación en líneas separadas por tabuladores para cada lectura.

1	2	3	4	5	6	7	8	9	10	11
SRR067577.2766	99	chr14	73240003	60	101M	=	73240004	102	GCTA…	FHG@…

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

https://www.zymoresearch.de/blogs/blog/what-are-sam-and-bam-files, http://samtools.github.io/hts-specs/SAMv1.pdf

MODELO DE ARCHIVO: NEXUS

```
#NEXUS
begin < blockname >;
    < command > < argument > [additional argument];
    [ < another command with args >; ]
end;
[ < another block with commands > ]
```

- Un archivo NEXUS está compuesto por un encabezado fijo #NEXUS seguido de múltiples bloques.
- Cada bloque comienza con BEGIN block_name; y termina con END;.
- Las palabras clave no distinguen entre mayúsculas y minúsculas.
- Los comentarios están encerrados entre corchetes [...]

- Bloque TAXA
  Contiene información sobre taxones.
- Bloque DATA
  Contiene la matriz de datos (por ejemplo, alineamiento de secuencias).
- Bloque TREES
  Contiene árboles filogenéticos descritos usando el formato Newick, por ejemplo, ((A,B),C);:

```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf;
End;

Begin data;
  Dimensions nchar=15;
  Format datatype=dna missing=? gap=- matchchar=.;
  Matrix
    [ "matchchar" means that it is the same as the first entry at the same position. ]
    SpaceDog   atgctagctagctcg
    SpaceCat   ......??...-.a.
    SpaceOrc   ...t.......-.g. [ same as atgttagctag-tgg ]
    SpaceElf   ...t.......-.a.
  ;
End;

BEGIN TREES;
  Tree tree1 = (((SpaceDog,SpaceCat),SpaceOrc,SpaceElf));
END;
```

http://wiki.christophchamp.com/index.php?title=NEXUS_file_format; *doi:10.1093/sysbio/46.4.590;* https://plewis.github.io/nexus/

MODELO DE ARCHIVO: NEXUS

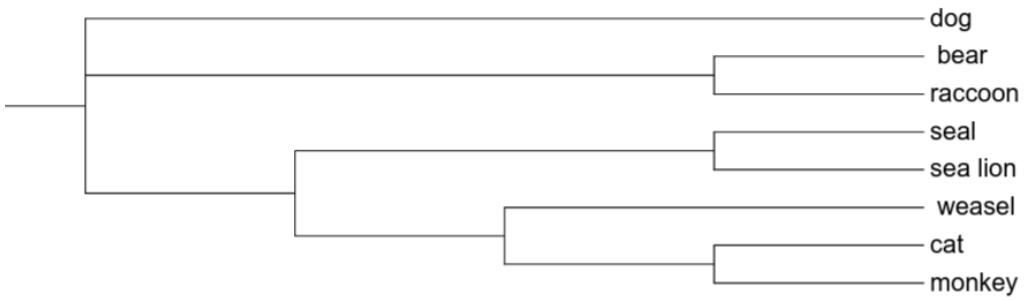
Puedes tener más o menos bloques, con tipos de datos diferentes, dependiendo de cómo una determinada aplicación requiera la entrada de datos.

```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf;
End;

Begin data;
  Dimensions nchar=15;
  Format datatype=dna missing=? gap=- matchchar=.;
  Matrix
    [ "matchchar" means that it is the same as the first entry at the same position. ]
    SpaceDog   atgctagctagctcg
    SpaceCat   ......??...-.a.
    SpaceOrc   ...t.......-.g. [ same as atgttagctag-tgg ]
    SpaceElf   ...t.......-.a.
  ;
End;

BEGIN TREES;
  Tree tree1 = (((SpaceDog,SpaceCat),SpaceOrc,SpaceElf));
END;
```

MODELO DE ARCHIVO: NEWICK

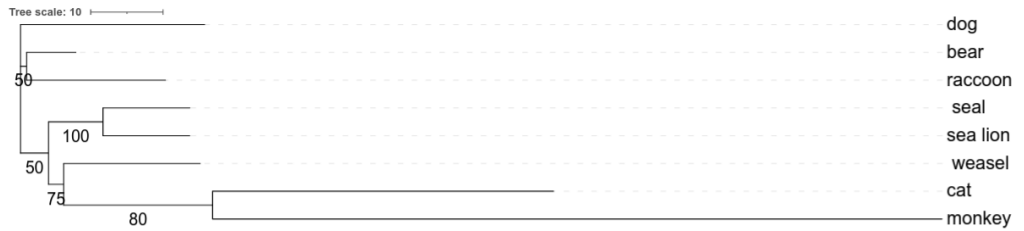
- NEWICK es un formato simple que se utiliza para escribir árboles en un archivo de texto.
- Este es un formato difícil de leer para los humanos, pero muy útil para intercambiar árboles entre diferentes tipos de software.
- Siempre se necesita punto y coma para finalizar el árbol).

```
((raccoon, bear),((sea_lion,seal),((monkey,cat),
weasel)),dog);
```




```
((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99700,
seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):20.59
201, weasel:18.87953):2.09460):3.87382,dog:25.46154);
```


```
((raccoon:19.19959,bear:6.80041)50:0.84600,((sea_lion:11.9970
0,
seal:12.00300)100:7.52973,((monkey:100.85930,cat:47.14069)80:
20.59201,
weasel:18.87953)75:2.09460)50:3.87382,dog:25.46154);
```



https://www.megasoftware.net/mega4/WebHelp/glossary/rh_newick_format.htm

¡VAMOS A PRACTICAR!

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

cooperación alemana
DEUTSCHE ZUSAMMENARBEIT

giz