GNU/Linux on Bioinformatics

Jordi Durban

Institut de Biomedicina de València (CSIC)

17 Junio 2019

WODA

WEB-BASED OMICS DATA ANALYSIS

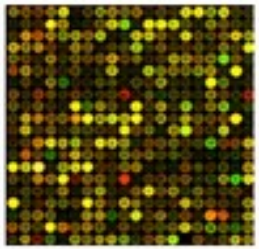
INSTITUTO DE BIOMEDICINA DE VALENCIA CSIC

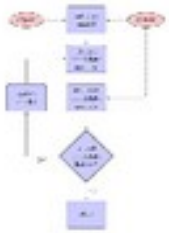
Unidad de Bioinformática y Bioestadística

PRINCIPE FELIPE CENTRO DE INVESTIGACION

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

GNU/Linux on Bioinformatics





Biological Data

GUI usage

CLI usage

Ad hoc script

programming

Algorithm

Math and IT staff

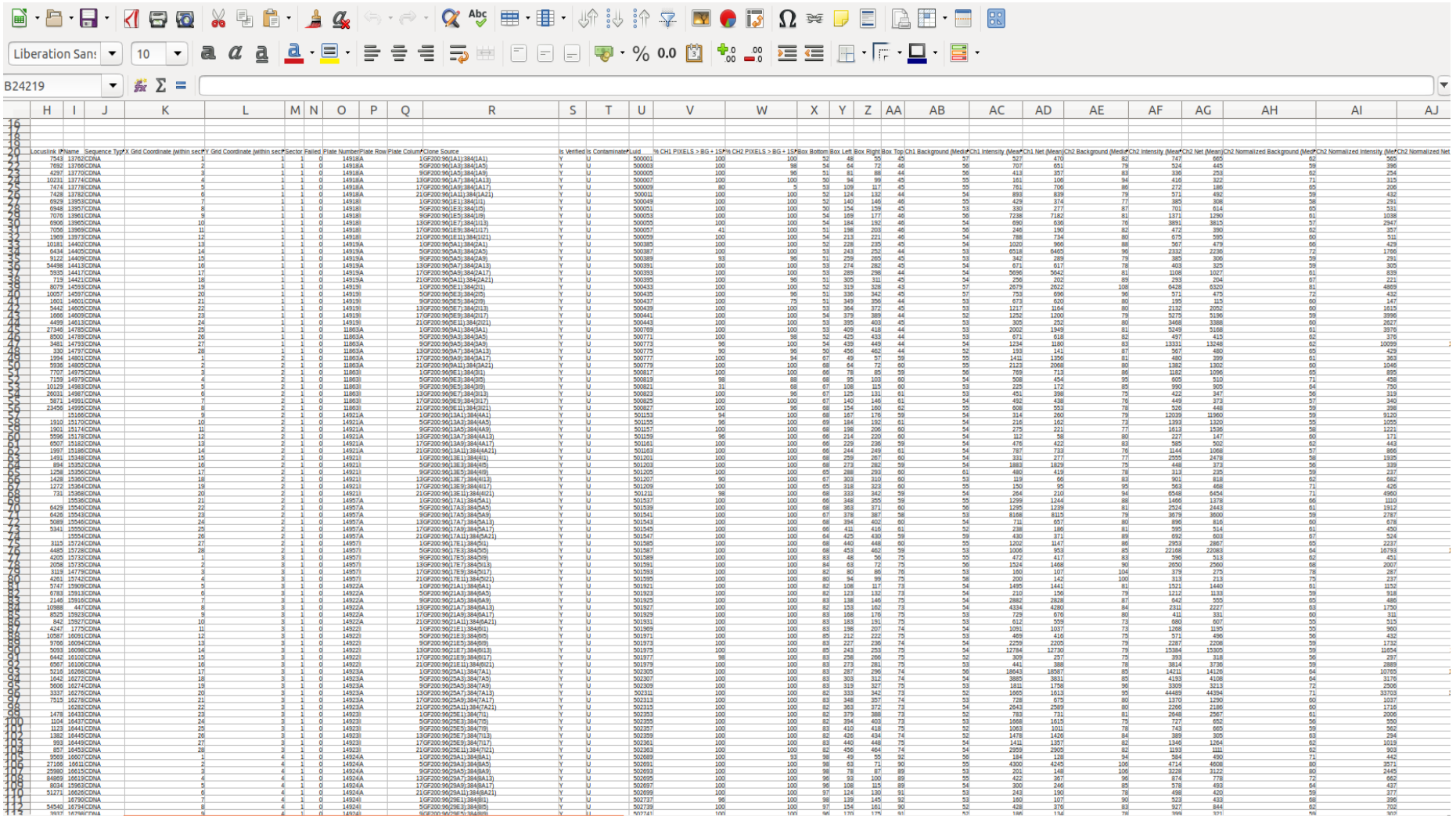
Bioinformaticians

Biologists


Configuring update for Windows 10
35% complete
Do not turn off your computer

GNU/Linux on Bioinformatics

EJEMPLO: MICROARRAY (Excel)



Excel specifications and limits

Excel for Office 365, Excel 2019, Excel 2016, Excel 2013, Excel 2010, Excel 2007

Newer versions Office 2010 Office 2007

Worksheet and workbook specifications and limits

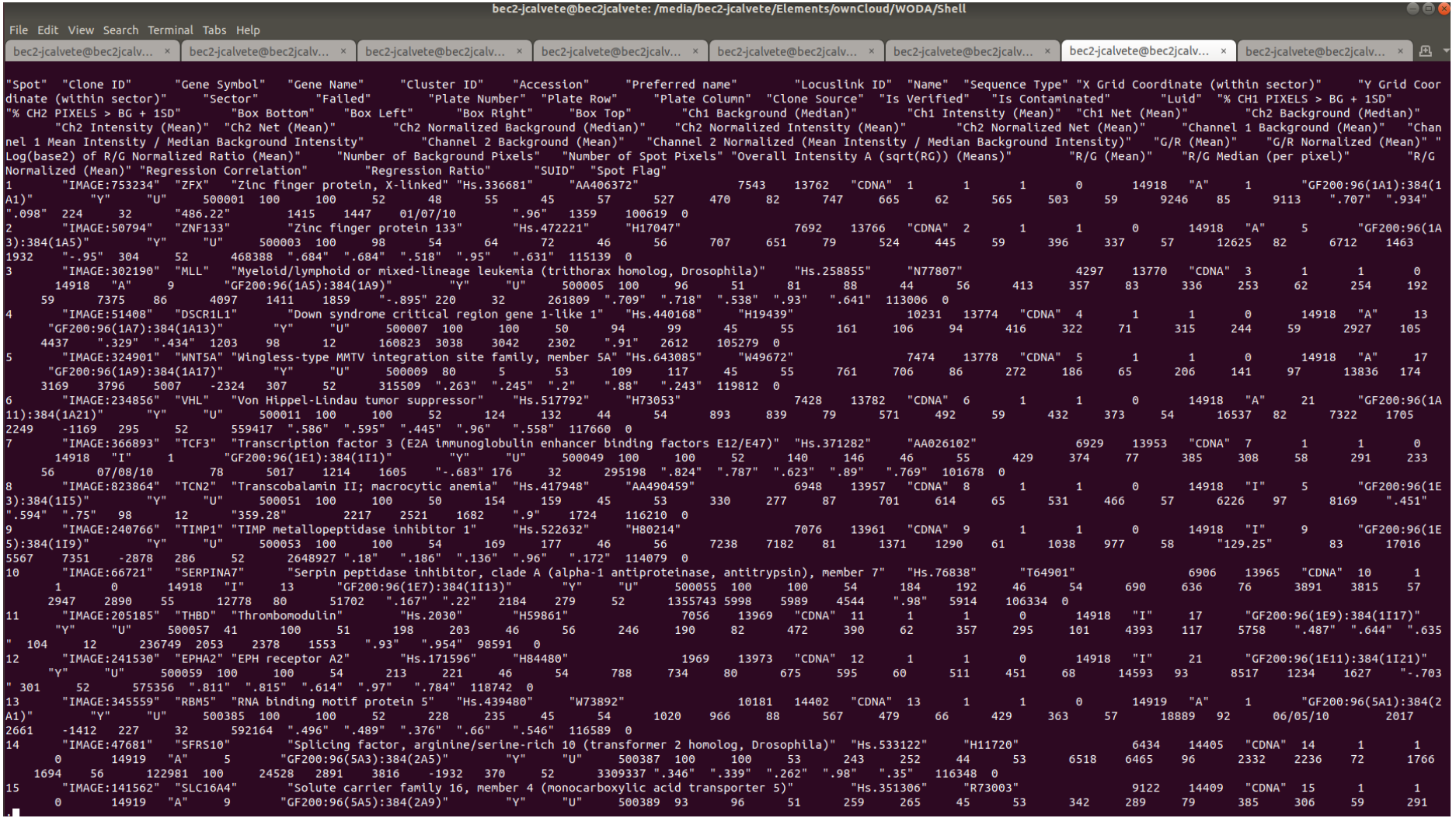
Feature	Maximum limit
Open workbooks	Limited by available memory and system resources
Total number of rows and columns on a worksheet	1,048,576 rows by 16,384 columns
Column width	255 characters
Row height	409 points
Page breaks	1,026 horizontal and vertical
Total number of characters that a cell can contain	32,767 characters

https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3

Uno de los puntos fundamentales de la filosofía Unix, es la utilización de **ficheros de texto**. Mientras otros sistemas operativos favorecen la utilización de ficheros binarios, que deben ser acompañados de herramientas especiales para poder manipularlos, en Unix se optó por crear un conjunto de herramientas para manipulación de ficheros de texto. Estas herramientas de manejo de ficheros de texto nos permiten realizar complejas manipulaciones de un modo muy sencillo y son uno de los principales atractivos de los sistemas Unix para el manejo de grandes cantidades de información.

GNU/Linux on Bioinformatics

EJEMPLO: MICROARRAY (Excel)



EJEMPLO: MICROARRAY (Excel)

¿Cómo podemos obtener un fichero nuevo con las columnas “Clone ID” y “Gene Name”?

¿Cuántos clones cuyo Gene name es “ZXD family zinc finger C” aparecen?

¿Cuál es la expresión de los genes relacionados con la leucemia?

EJEMPLO: MICROARRAY (Excel)

¿Cómo podemos obtener un fichero nuevo con las columnas “Clone ID” y “Gene Name”?

¿Cuántos clones cuyo Gene name es “ZXD family zinc finger C” aparecen?

¿Cuál es la expresión de los genes relacionados con la leucemia?

Filtros? Crtl + C , Crtl + V?

EJEMPLO: MICROARRAY (Excel)

¿Cómo podemos obtener un fichero nuevo con las columnas “Clone ID” y “Gene Name”?

cut -f 2,4 microarray_adenoma_hk69.csv |grep "^\""

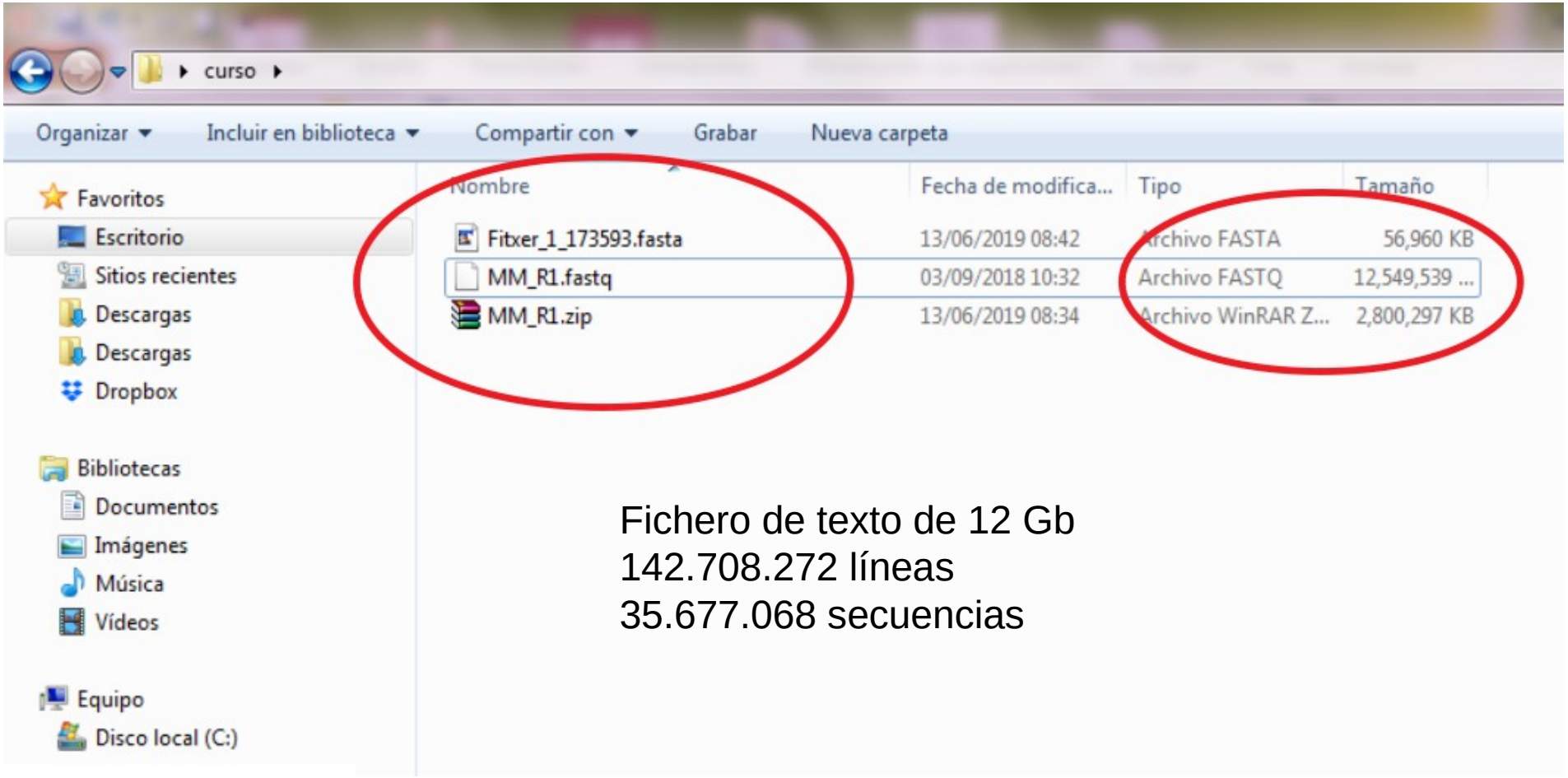
¿Cuántos clones cuyo Gene name es “ZXD family zinc finger C” aparecen?

awk -F "\t" '{if ($4 ~ /ZXD family zinc finger C/) print}' microarray_adenoma_hk69.csv

¿Cuál es la expressión de los genes relacionados con la leucemia?

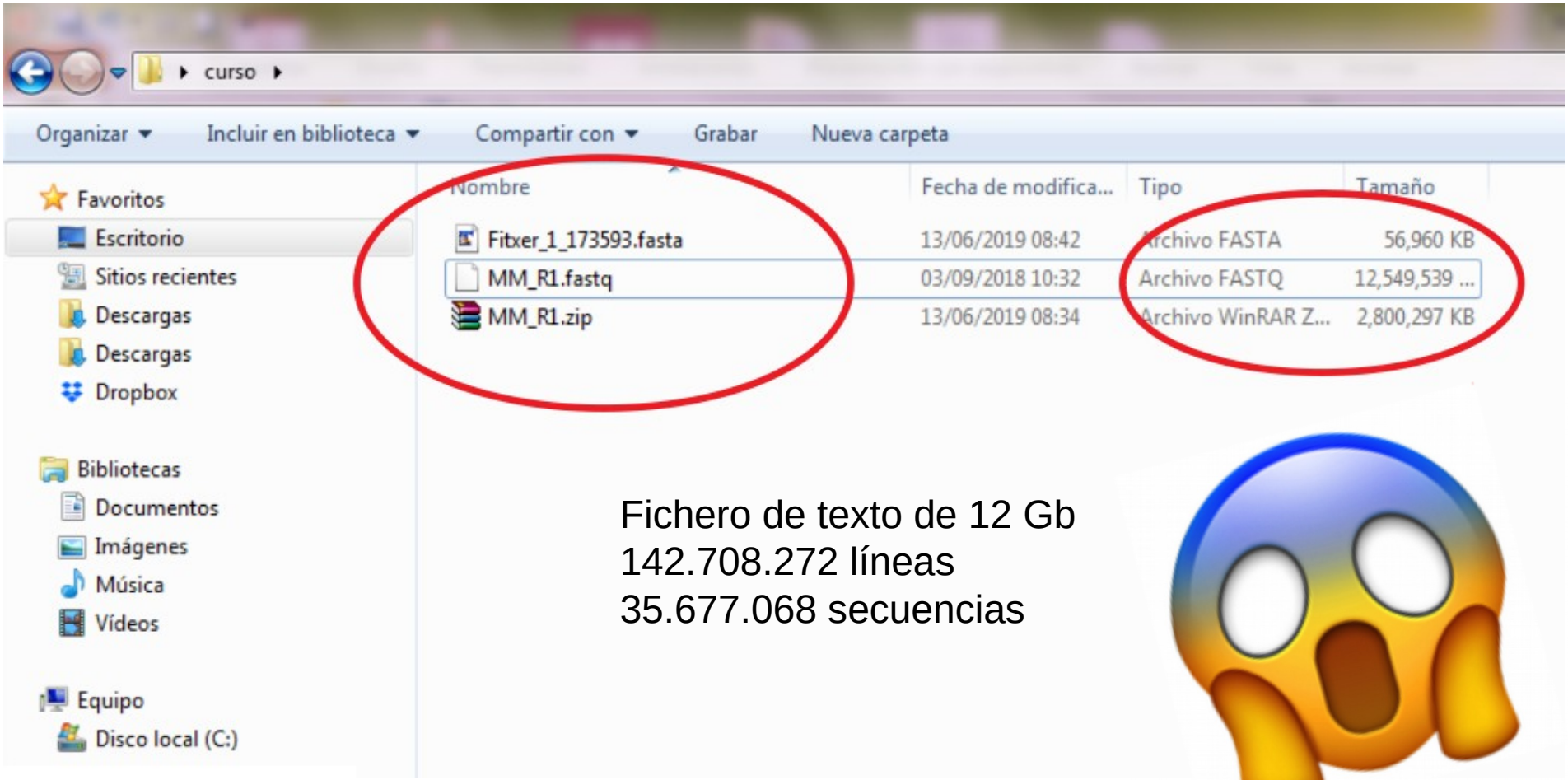
grep -i leukemia microarray_adenoma_hk69.csv

EJEMPLO: FASTQ



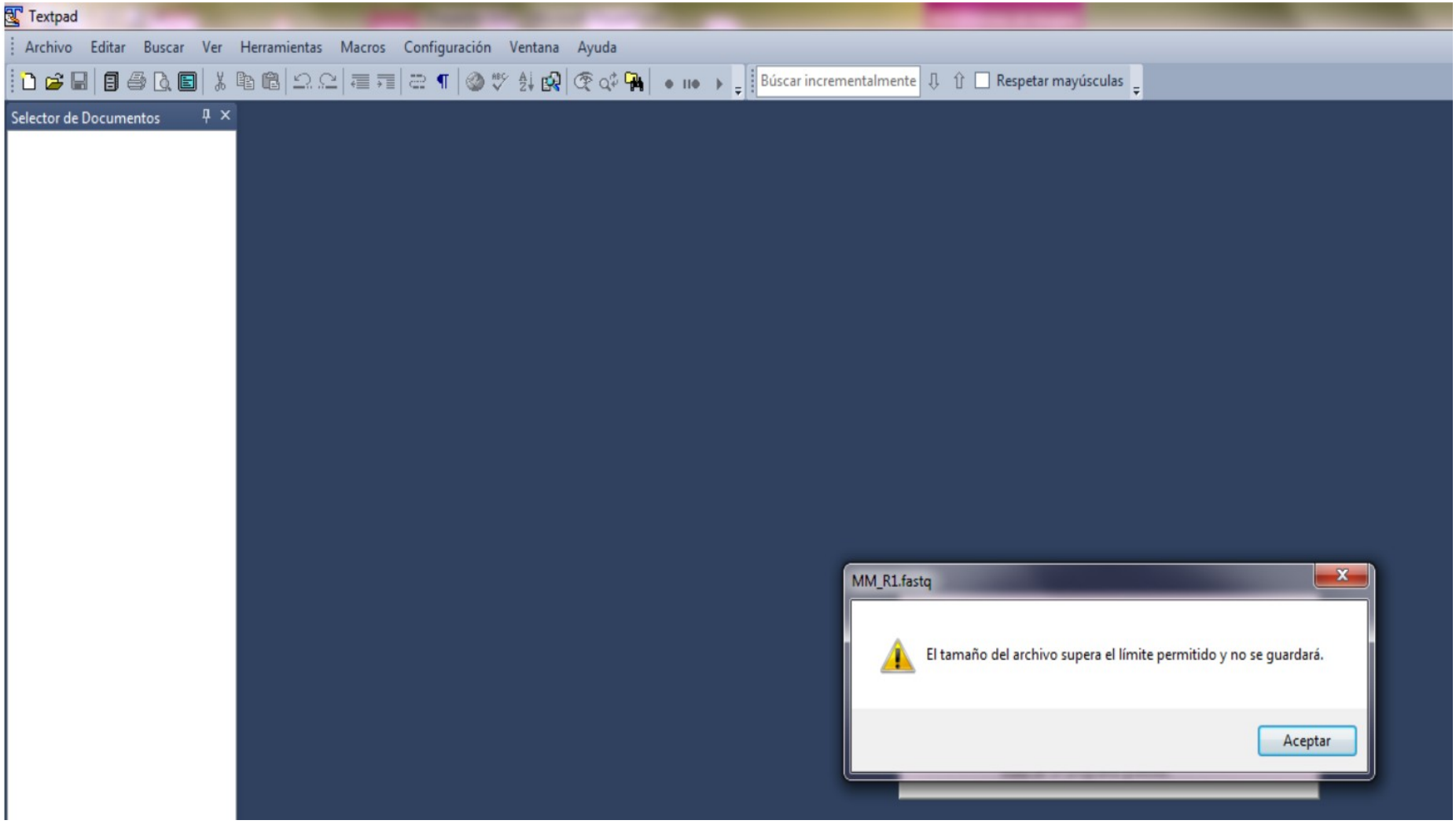
Fichero de texto de 12 Gb
142.708.272 líneas
35.677.068 secuencias

EJEMPLO: FASTQ

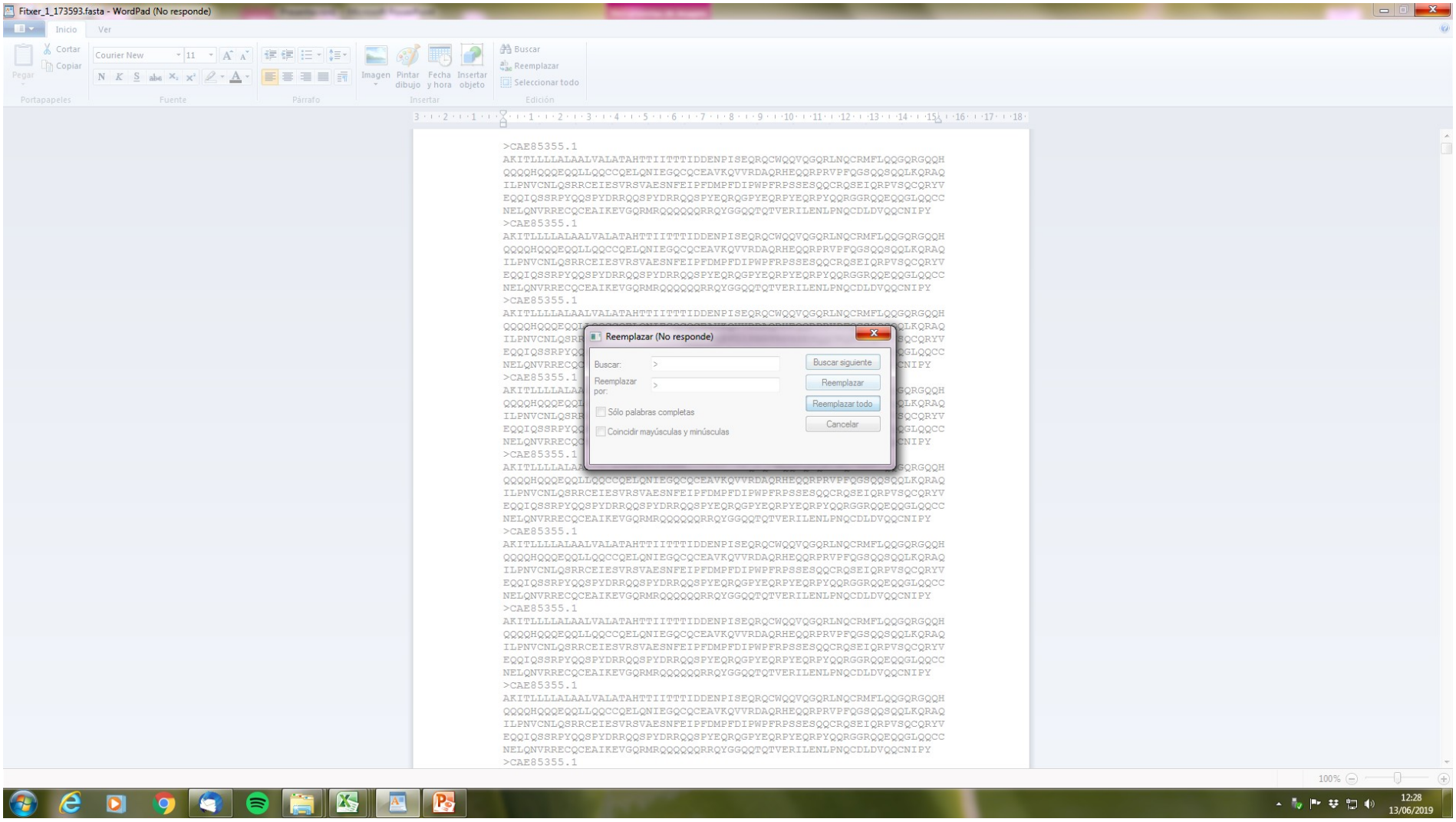


Fichero de texto de 12 Gb
142.708.272 líneas
35.677.068 secuencias

EJEMPLO: FASTQ

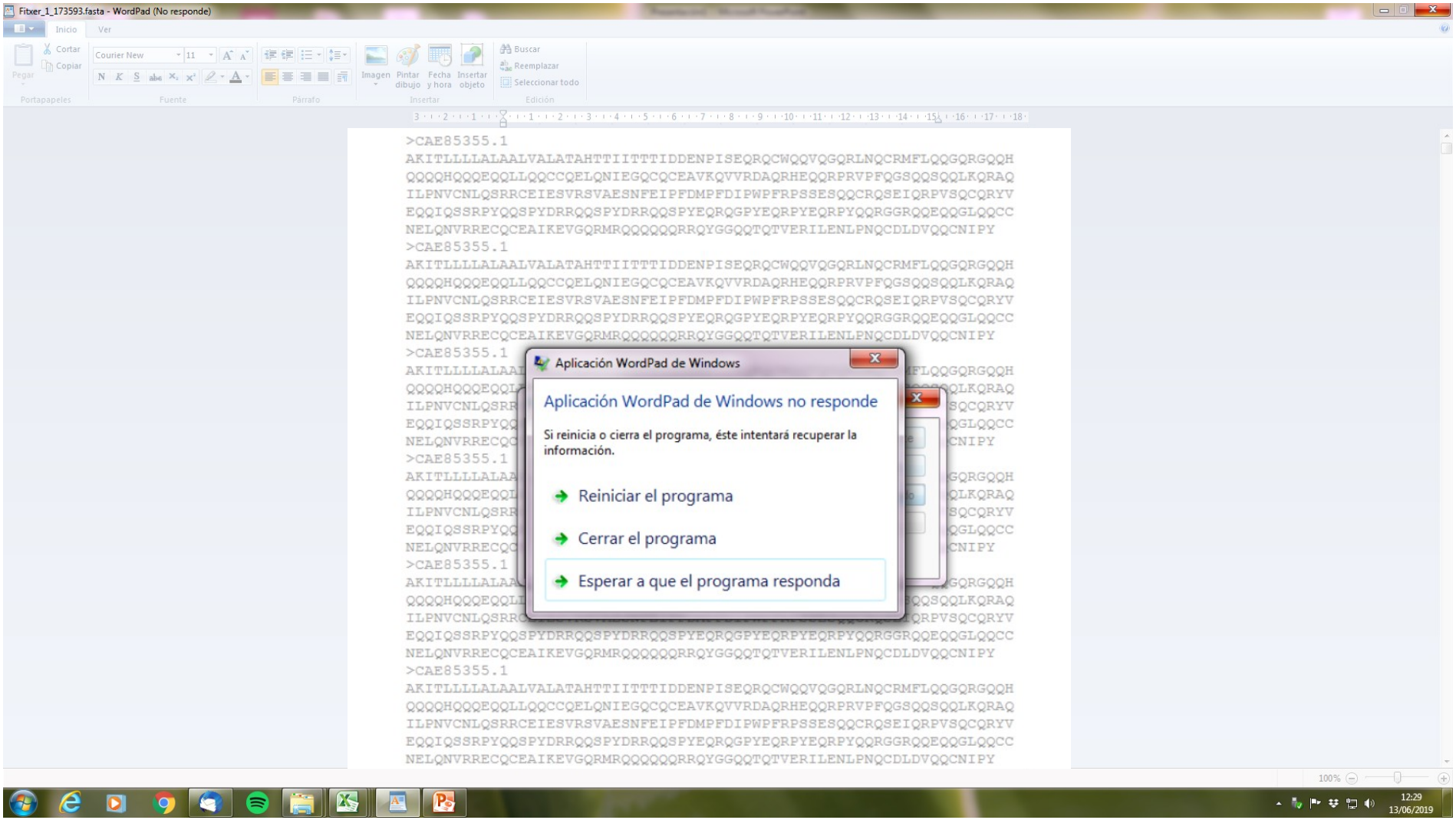


EJEMPLO: FASTA de 173593 secuencias



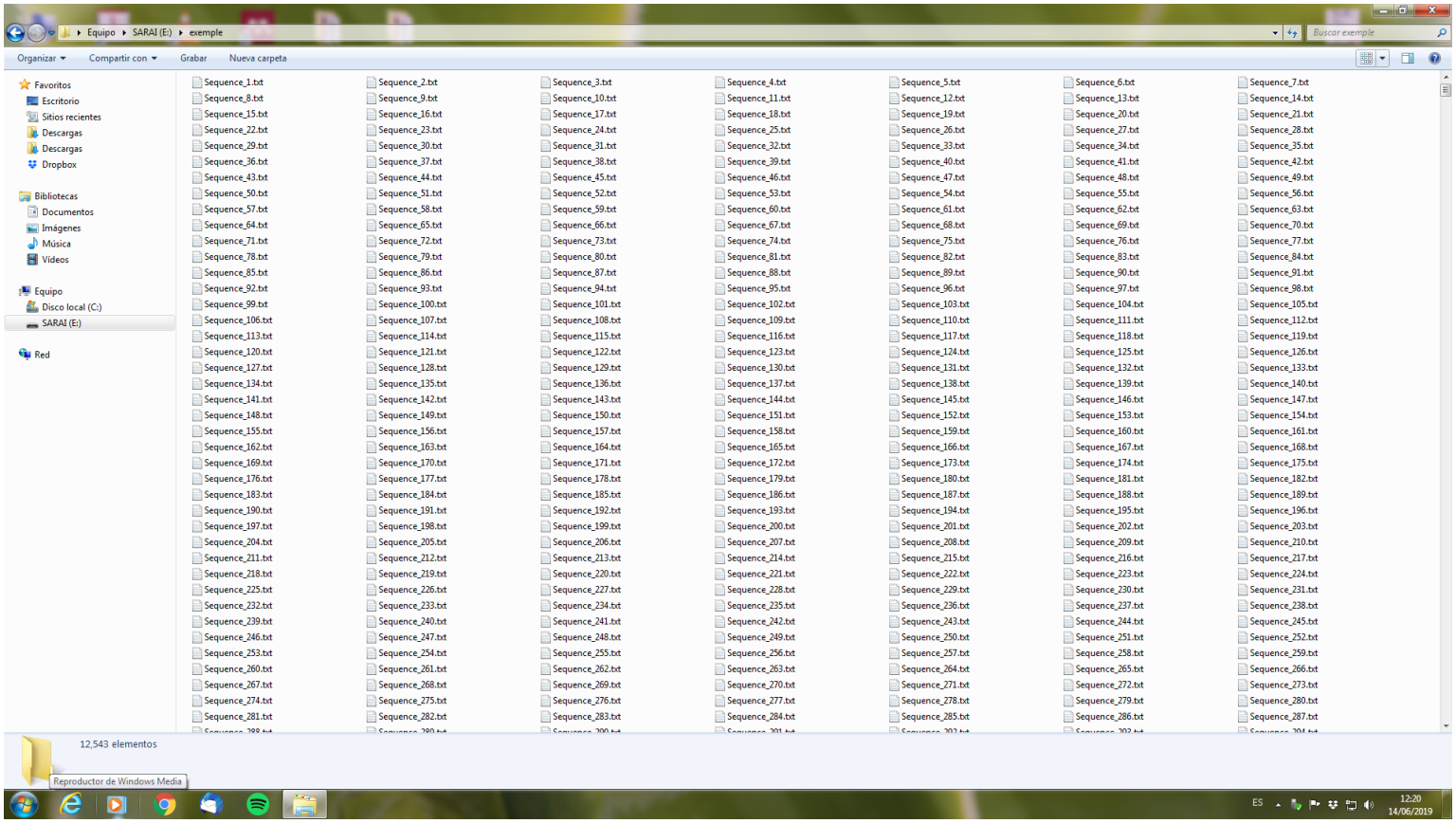
GNU/Linux on Bioinformatics

EJEMPLO: FASTA de 173593 secuencias



EJEMPLO: RENOMBRAR

Tenemos 12.543 ficheros con la siguiente nomenclatura: Sequence_1.txt, Sequence_2.txt, Sequence_3.txt, Sequence_4.txt, Sequence_5.txt, Sequence_6, txt… Es decir, este escenario:



EJEMPLO: RENOMBRAR

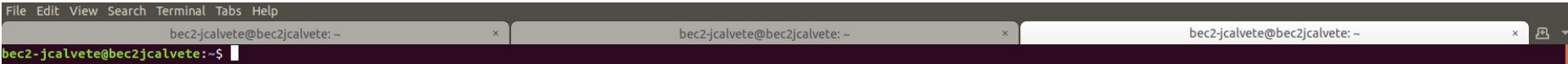
Queremos renombrar **TODOS** los archivos poniéndole la fecha de secuenciación: 23_04_2018

EJEMPLO: RENOMBRAR

Queremos renombrar **TODOS** los archivos poniéndole la fecha de secuenciación: 23_04_2018

rename 's/Sequence/23_04_2018_Sequence/' *.txt

LA SHELL DE UNIX



Flexibilidad. Los programas gráficos suelen ser muy adecuados para realizar la tarea para la que han sido creados, pero son difíciles de adaptar para otras tareas. La línea de comandos de Unix es por el contrario muy flexible puesto que está formada por pequeñas herramientas que podemos combinar según nuestras necesidades.

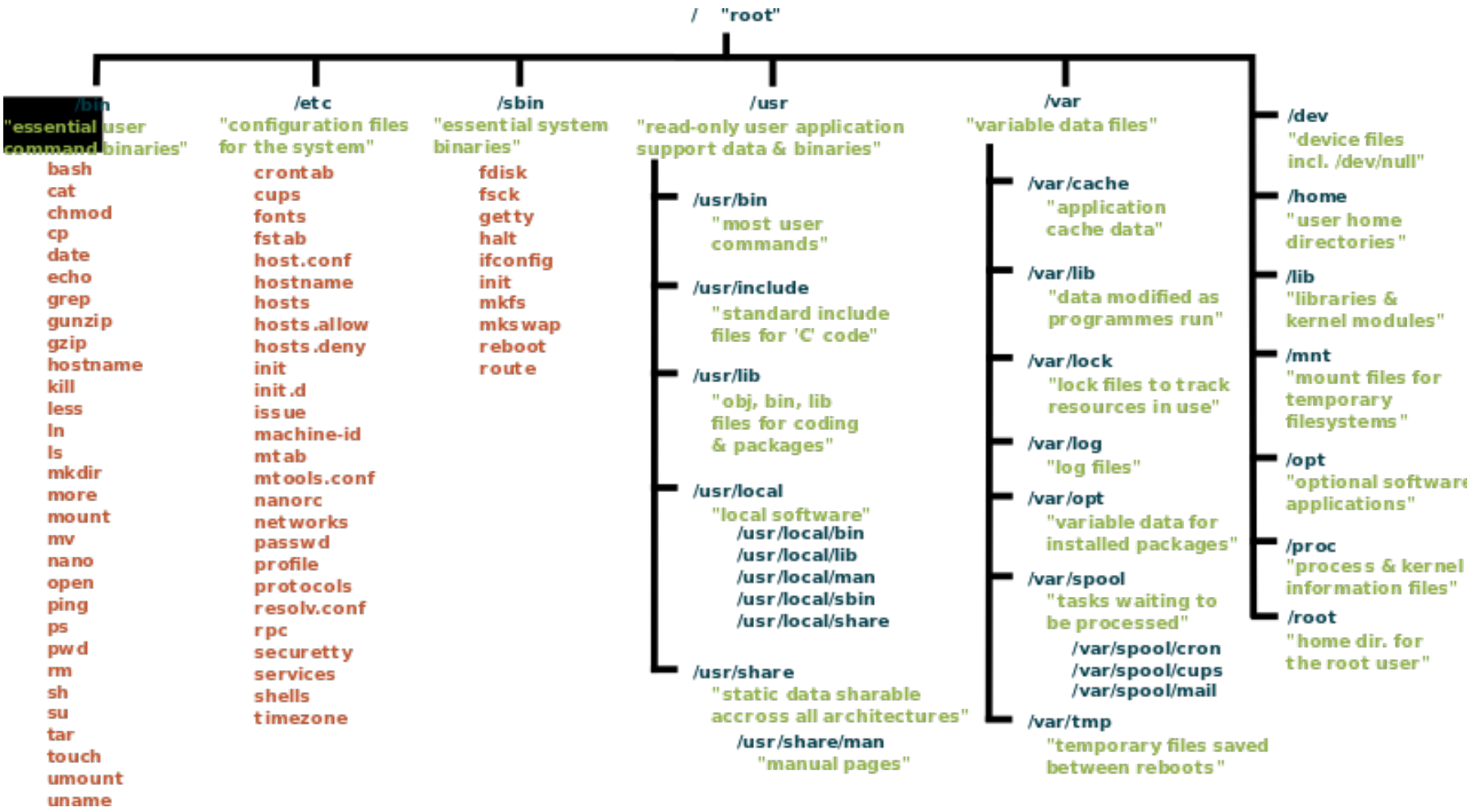
Reproducibilidad. Documentar y repetir el proceso seguido para realizar un análisis con un programa gráfico es muy costoso puesto que es difícil describir la secuencia de clicks y doble clicks que hemos realizado. Por el contrario, los procesos realizados mediante la línea de comandos son muy fáciles de documentar puesto que tan sólo debemos guardar el texto que hemos introducido en la pantalla.

Fiabilidad. Los programas básicos de Unix fueron creados en los años 70 y han sido probados por innumerables usuarios por lo que se han convertido en piezas de código extraordinariamente confiables.

Necesidad. Hay aplicaciones que sólo pueden utilizarse mediante la línea de comandos.

Velocidad. Las interfaces gráficas suelen consumir muchos recursos mientras que los programas que funcionan en línea de comandos suelen ser extraordinariamente livianos y rápidos.

ESTRUCTURA DE FICHEROS



ESTRUCTURA DE FICHEROS

/ directorio raíz	**/bin**	Contiene programas ejecutables básicos para el sistema.
	/boot	Contiene los ficheros necesarios para el arranque del sistema.
	/dev	Contiene los ficheros correspondientes a los dispositivos: sonido, impresora, disco duro, lector de cd/dvd, video, etc.
	/etc	Contiene ficheros y directorios de configuración.
	/home	Contiene los directorios de trabajo de los usuarios. Cada usuario tiene su propio directorio en el sistema dentro de /home/.
	/lib	Contiene las librerías compartidas y los módulos del kernel
	/media	Dentro de este directorio se montan los dispositivos como el CD-ROM, memorias USB, discos duros portátiles, etc
	/opt	Directorio reservado para instalar aplicaciones.
	/sbin	Contiene los ficheros binarios ejecutables del sistema operativo.
	/srv	Contiene datos de los servicios proporcionado por el sistema.
	/tmp	Directorio de archivos temporales.
	/usr	Aquí se encuentran la mayoría de los archivos del sistema, aplicaciones, librerías, manuales, juegos... Es un espacio compartido por todos los usuarios.
	/var	Contiene archivos administrativos y datos que cambian con frecuencia: registro de errores, bases de datos, colas de impresión, etc.
	/root	Directorio de trabajo del administrador del sistema (usuario root).
	/proc	Aquí se almacenan datos del kernel e información sobre procesos.

ESTRUCTURA DE FICHEROS

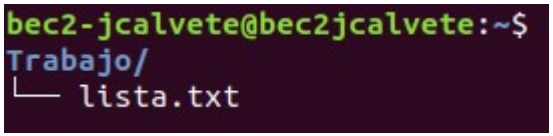
/home/luisjose	/Documentos			
	/Escritorio			
	/Imágenes			
	/Música			
	/matematicas	**/curso_01**	**/algebra**	**/examenes_antiguos** **/apuntes**
			/analisis	
			/fisica	**/libros_de_ejercicios** **/videos**
			/informatica	**/compiladores_pascal**
	/Video			

RUTA ABSOLUTAS

Localización del archivo **en el árbol de directorios**.

p.ej para referirnos al fichero lista.txt en el directorio *Trabajo* de mi directorio personal, la ruta absoluta es:

"/home/jdurban/Trabajo/lista.txt"



RUTAS RELATIVAS

Localización del archivo **desde el directorio actual**.

p.ej para referirnos al fichero lista.txt en el directorio Trabajo desde mi directorio personal, la ruta relativa es:

"/Trabajo/lista.txt"

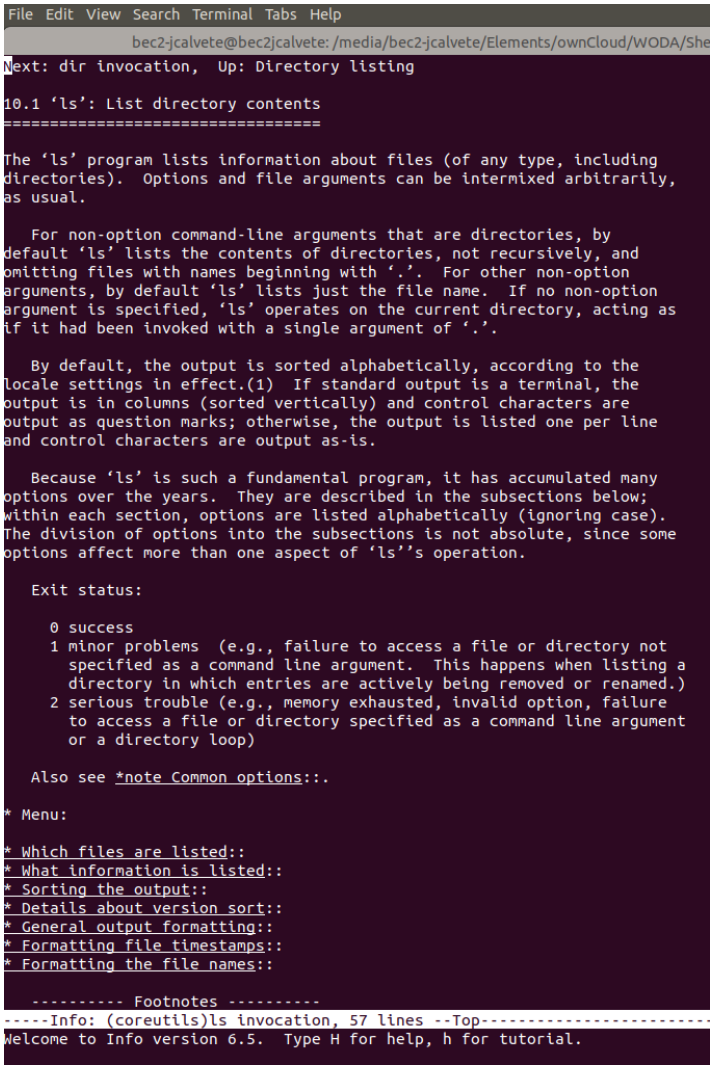
COMANDOS MÁS HABITUALES: VISUALIZACIÓN

Comando	*Acción*	*Ejemplo*
pwd	muestra el directorio actual	pwd
ls	lista ficheros y directorios	ls -l
cd	cambia de directorio	cd mp3/wim_mertens
mkdir	crea uno o varios directorios	mkdir cartas facturas
cat	visualiza un fichero	cat /var/log/dmesg
more	visualiza un fichero pantalla a pantalla	more /var/log/dmesg
less	visualiza un fichero pantalla a pantalla y permite retroceder	less /var/log/dmesg
head	visualiza las primeras filas de un fichero	head -n5 /var/log/dmesg
tail	visualiza las últimas filas de un fichero	tail /var/log/dmesg
touch	crea un fichero vacío	touch listado.txt
ee	editor de textos muy simple	ee listado.txt
mcedit	editor de textos que forma parte de Midnight Commander	mcedit listado.txt
vi	editor de textos muy potente	vi listado.txt
apt-get	instala y desinstala programas	apt-get install mc
man	muestra ayuda sobre un determinado comando	man ls

COMANDOS MÁS HABITUALES: MANIPULACIÓN

Comando	*Acción*	*Ejemplo*
cp	copia archivos o directorios	cp *.txt correspondencia/
mv	mueve o renombra archivos o directorios	mv palabras.txt texto.txt
rm	borra archivos o directorios	rm -R cosas/basurilla
rmdir	borra directorios	rmdir viejo

AYUDA : *man* e *info*



AYUDA : *man* e *info*

google is

google is **skynet**
google is **going to take over the world**
google is **watching you**
google is **your friend**
google is **making us stupid**
google is **hiring at home workers**
google is **down**
google is **hiring**
google is **a number**
google is **the devil**

Google Search I'm Feeling Lucky

https://ja.cat/e8jfd