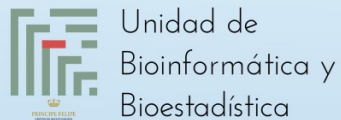


GNU/Linux on Bioinformatics

Jordi Durban

Institut de Biomedicina de València (CSIC)

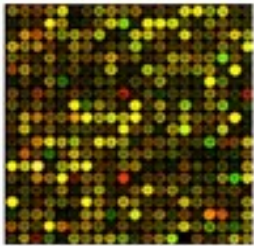
17 Junio 2019



WODA

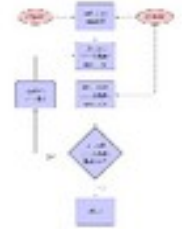
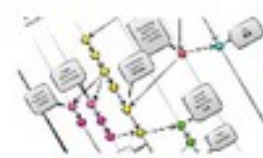
WEB-BASED OMICS DATA ANALYSIS

GNU/Linux on Bioinformatics



```
jose@leamtop:~$ ls  
almacen  devel  
analysis @inera.com.zip  
articulos @ocencia  
bin      eclipse
```

```
def split_mates(seq_fhands, l):  
    "it splits the input seqs"  
  
    if linkers is None:  
        linkers = LINKERS  
  
    for seq_fhand in seq_fhands:  
        matcher = _blastMatcher  
        for seqrec in read_seq
```



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

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

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)

bec2-jcalvete@bec2jcalvete: /media/bec2-jcalvete/Elements/ownCloud/WODA/Shell

"Spot"	"Clone ID"	"Gene Symbol"	"Gene Name"	"Cluster ID"	"Accession"	"Preferred name"	"Locuslink ID"	"Name"	"Sequence Type"	"X Grid Coordinate (within sector)"	"Y Grid Coordinate (within sector)"	"Sector"	"Failed"	"Plate Number"	"Plate Row"	"Plate Column"	"Clone Source"	"Is Verified"	"Is Contaminated"	"Luid"	"% CH1 PIXELS > BG + 15D"	"Box Bottom"	"Box Left"	"Box Right"	"Box Top"	"Ch1 Background (Median)"	"Ch1 Intensity (Mean)"	"Ch1 Net (Mean)"	"Ch2 Background (Median)"	"Ch2 Intensity (Mean)"	"Ch2 Net (Mean)"	"Ch2 Normalized Background (Median)"	"Ch2 Normalized Intensity (Mean)"	"Ch2 Normalized Net (Mean)"	"Channel 1 Background (Mean)"	"Channel 1 Mean Intensity / Median Background Intensity"	"Channel 2 Background (Mean)"	"Channel 2 Normalized (Mean Intensity / Median Background Intensity)"	"G/R (Mean)"	"G/R Normalized (Mean)"	"Log(base2) of R/G Normalized Ratio (Mean)"	"Number of Background Pixels"	"Number of Spot Pixels"	"Overall Intensity A (sqrt(RG)) (Means)"	"R/G (Mean)"	"R/G Median (per pixel)"	"R/G"																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																				
1	"IMAGE:753234"	"ZFX"	"Zinc finger protein, X-linked"	"Hs.336681"	"AA406372"		7543	13762	"CDNA"	1	1	1	0	14918	"A"	1	565	503	59	9246	85	9113	".707"	".934"	"GF200:96(1A1):384(1A1)"	224	32	"486.22"	1415	1447	01/07/10	".96"	1359	100619	0	7692	13766	"CDNA"	2	1	1	0	14918	"A"	5	"GF200:96(1A3):384(1A5)"	1932	".95"	304	52	468388	".684"	".684"	".518"	".95"	".631"	115139	0	59	7375	86	4097	1411	1859	".895"	220	32	261809	".709"	".718"	".538"	".93"	".641"	113006	0	413	357	83	336	253	62	254	192	"IMAGE:302190"	"MLL"	"Myeloid/Lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)"	"Hs.258855"	"N77807"		4297	13770	"CDNA"	3	1	1	0	14918	"A"	9	"GF200:96(1A5):384(1A9)"	59	7375	86	4097	1411	1859	".895"	220	32	261809	".709"	".718"	".538"	".93"	".641"	113006	0	413	357	83	336	253	62	254	192	"IMAGE:51408"	"DSCR11"	"Down syndrome critical region gene 1-like 1"	"Hs.440168"	"H19439"		10231	13774	"CDNA"	4	1	1	0	14918	"A"	13	"GF200:96(1A7):384(1A13)"	4437	".329"	".434"	1203	98	12	160823	3038	3042	2302	".91"	2612	105279	0	161	106	94	416	322	71	315	244	59	2927	105	"IMAGE:324901"	"WNT5A"	"Wingless-type MMTV integration site family, member 5A"	"Hs.643085"	"W49672"		7474	13778	"CDNA"	5	1	1	0	14918	"A"	17	"GF200:96(1A9):384(1A17)"	3169	3796	5007	-2324	307	52	315509	".263"	".245"	".2"	".88"	".243"	119812	0	761	706	86	272	186	65	206	141	97	13836	174	"IMAGE:234856"	"VHL"	"Von Hippel-Lindau tumor suppressor"	"Hs.517792"	"H73053"		7428	13782	"CDNA"	6	1	1	0	14918	"A"	21	"GF200:96(1A11):384(1A21)"	2249	-1169	295	52	559417	".586"	".595"	".445"	".96"	".558"	117660	0	893	839	79	571	492	59	432	373	54	16537	82	7322	1705	"IMAGE:366893"	"TCF3"	"Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)"	"Hs.371282"	"AA026102"		6929	13953	"CDNA"	7	1	1	0	14918	"I"	1	"GF200:96(1E1):384(1I1)"	56	07/08/10	78	5017	1214	1605	".683"	176	32	295198	".824"	".787"	".623"	".89"	".769"	101678	0	429	374	77	385	308	58	291	233	"IMAGE:823864"	"TCN2"	"Transcobalamin II; macrocytic anemia"	"Hs.417948"	"AA490459"		6948	13957	"CDNA"	8	1	1	0	14918	"I"	5	"GF200:96(1E3):384(1I5)"	1	0	14918	"I"	13	"GF200:96(1E7):384(1I13)"	2947	2890	55	12778	80	51702	".167"	".22"	2184	279	52	1355743	5998	5989	4544	".98"	5914	106334	0	690	636	76	3891	3815	57	"IMAGE:205185"	"THBD"	"Thrombosmodulin"	"Hs.2030"	"H59861"		7056	13969	"CDNA"	11	1	1	0	14918	"I"	17	"GF200:96(1E9):384(1I17)"	104	12	236749	2053	2378	1553	".93"	".954"	98591	0	82	472	390	62	357	295	101	4393	117	5758	".487"	".644"	".635	"IMAGE:241530"	"EPHA2"	"EPH receptor A2"	"Hs.171596"	"H84480"		1969	13973	"CDNA"	12	1	1	0	14918	"I"	21	"GF200:96(1E11):384(1I21)"	301	52	575356	".811"	".815"	".614"	".97"	".784"	118742	0	675	595	60	511	451	68	14593	93	8517	1234	1627	".703	"IMAGE:345559"	"RBMS"	"RNA binding motif protein 5"	"Hs.439480"	"W73892"		10181	14402	"CDNA"	13	1	1	0	14919	"A"	1	"GF200:96(5A1):384(2A1)"	2661	-1412	227	32	592164	".496"	".489"	".376"	".66"	".546"	116589	0	66	429	363	57	18889	92	06/05/10	2017	"IMAGE:47681"	"SFRS10"	"Splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila)"	"Hs.533122"	"H11720"		6434	14405	"CDNA"	14	1	1	0	14919	"A"	1	"GF200:96(5A3):384(2A5)"	0	14919	"A"	5	"GF200:96(5A3):384(2A5)"	1694	56	122981	100	24528	2891	3816	-1932	370	52	3309337	".346"	".339"	".262"	".98"	".35"	116348	0	6518	6465	96	2332	2236	72	1766	"IMAGE:141562"	"SLC16A4"	"Solute carrier family 16, member 4 (monocarboxylic acid transporter 5)"	"Hs.351306"	"R73003"		9122	14409	"CDNA"	15	1	1	0	14919	"A"	1	"GF200:96(5A5):384(2A9)"	0	14919	"A"	9	"GF200:96(5A5):384(2A9)"	0	14919	"A"	9	"GF200:96(5A5):384(2A9)"

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? Ctrl + C , Ctrl + 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?

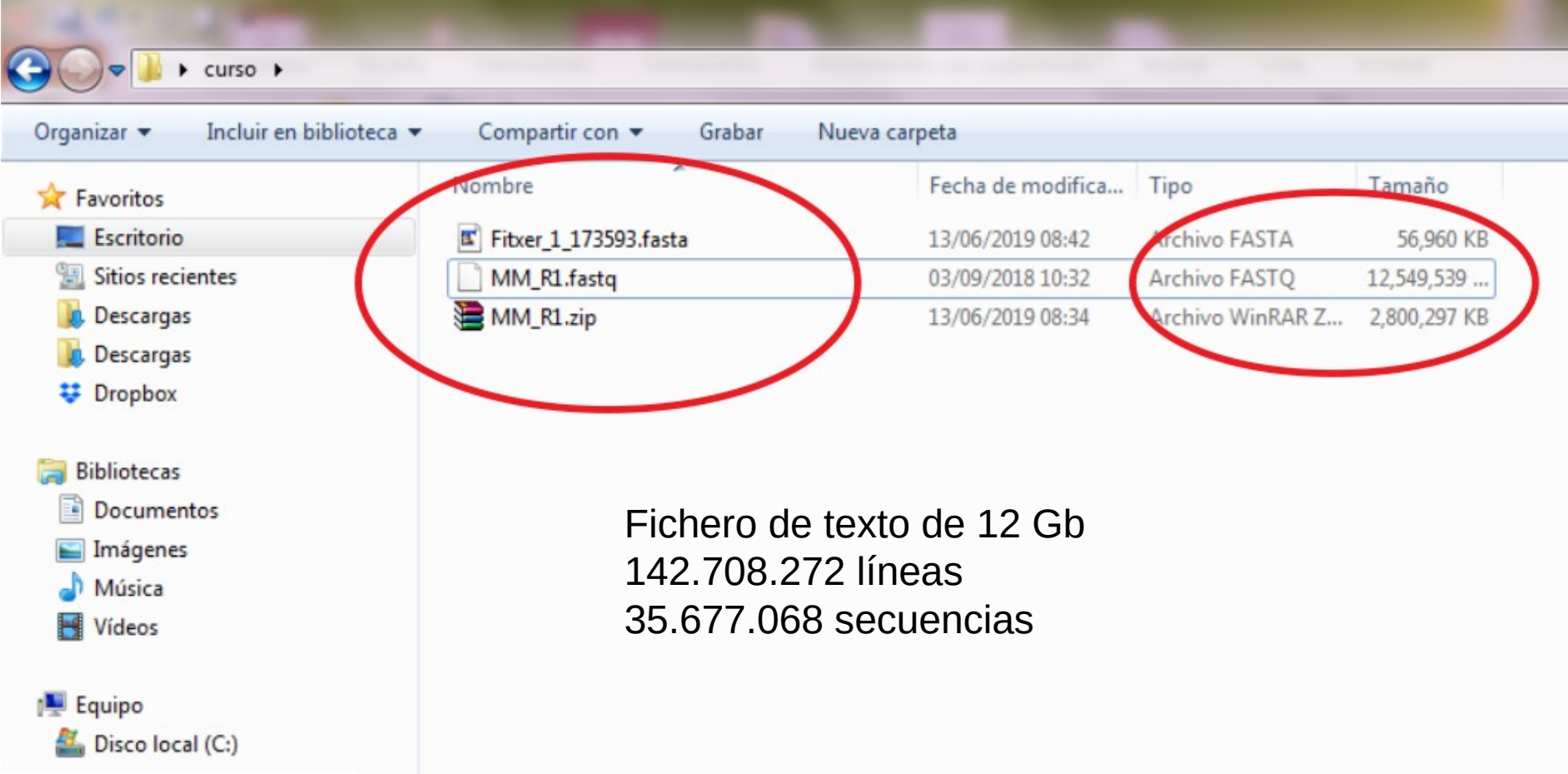
```
awk -F "\t" '{if ($4 ~ /ZXD family zinc finger C/) print}' microarray_adenoma_hk69.csv
```

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

```
grep -i leukemia microarray_adenoma_hk69.csv
```

GNU/Linux on Bioinformatics

EJEMPLO: FASTQ



Organizar ▾ Incluir en biblioteca ▾ Compartir con ▾ Grabar Nueva carpeta

Nombre	Fecha de modifica...	Tipo	Tamaño
Fitxer_1_173593.fasta	13/06/2019 08:42	Archivo FASTA	56,960 KB
MM_R1.fastq	03/09/2018 10:32	Archivo FASTQ	12,549,539 ...
MM_R1.zip	13/06/2019 08:34	Archivo WinRAR Z...	2,800,297 KB

Fichero de texto de 12 Gb
142.708.272 líneas
35.677.068 secuencias

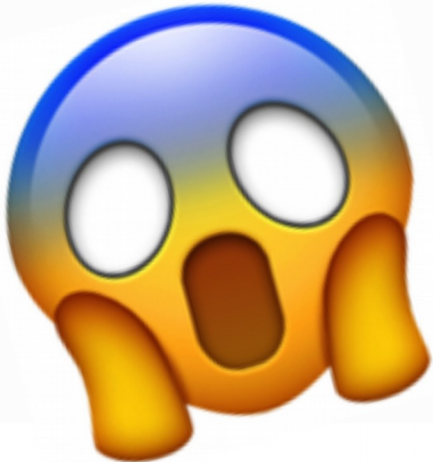
GNU/Linux on Bioinformatics

EJEMPLO: FASTQ

Organizar ▾ Incluir en biblioteca ▾ Compartir con ▾ Grabar Nueva carpeta

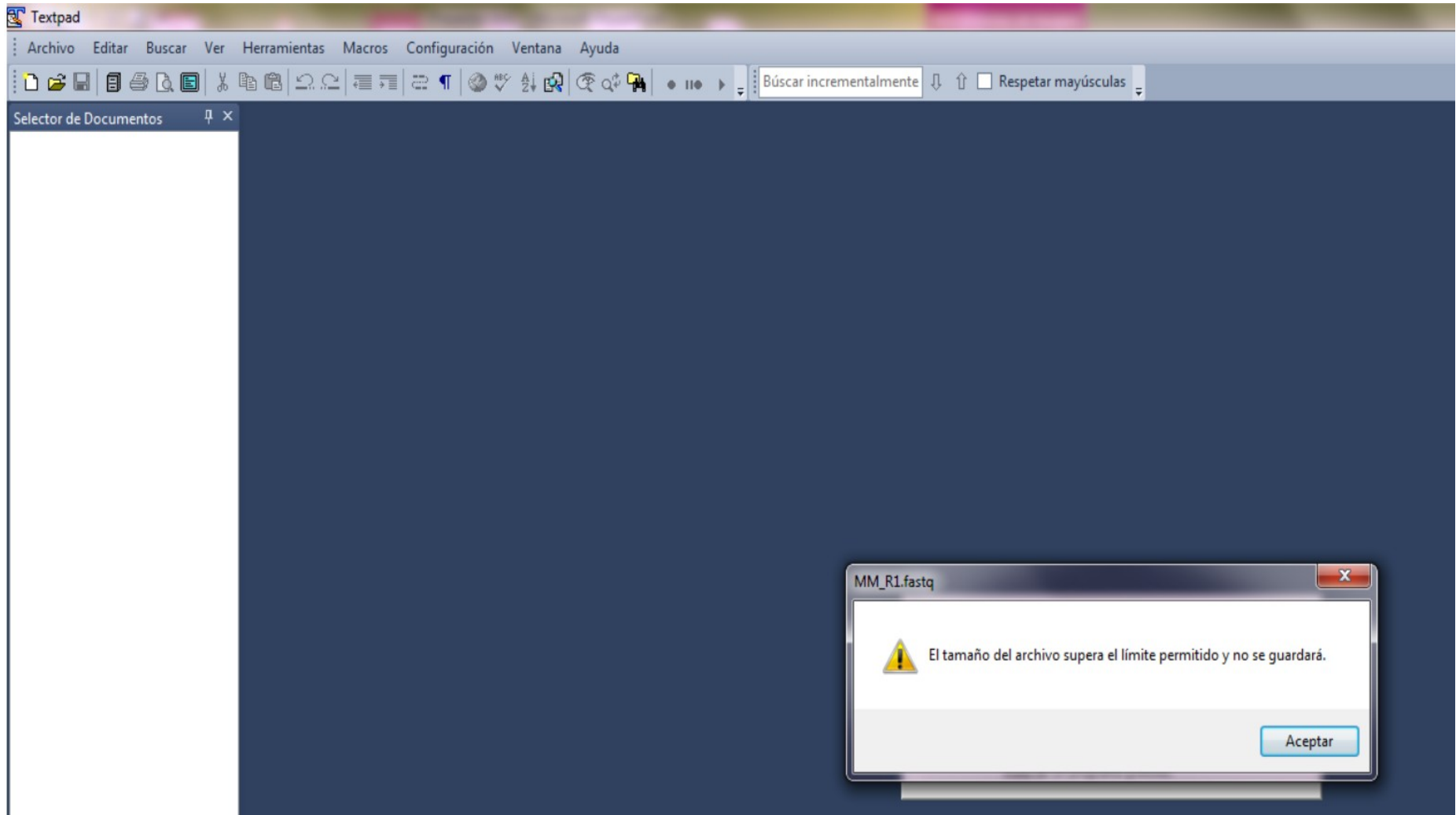
Nombre	Fecha de modifica...	Tipo	Tamaño
Fitxer_1_173593.fasta	13/06/2019 08:42	Archivo FASTA	56,960 KB
MM_R1.fastq	03/09/2018 10:32	Archivo FASTQ	12,549,539 ...
MM_R1.zip	13/06/2019 08:34	Archivo WinRAR Z...	2,800,297 KB

Fichero de texto de 12 Gb
142.708.272 líneas
35.677.068 secuencias



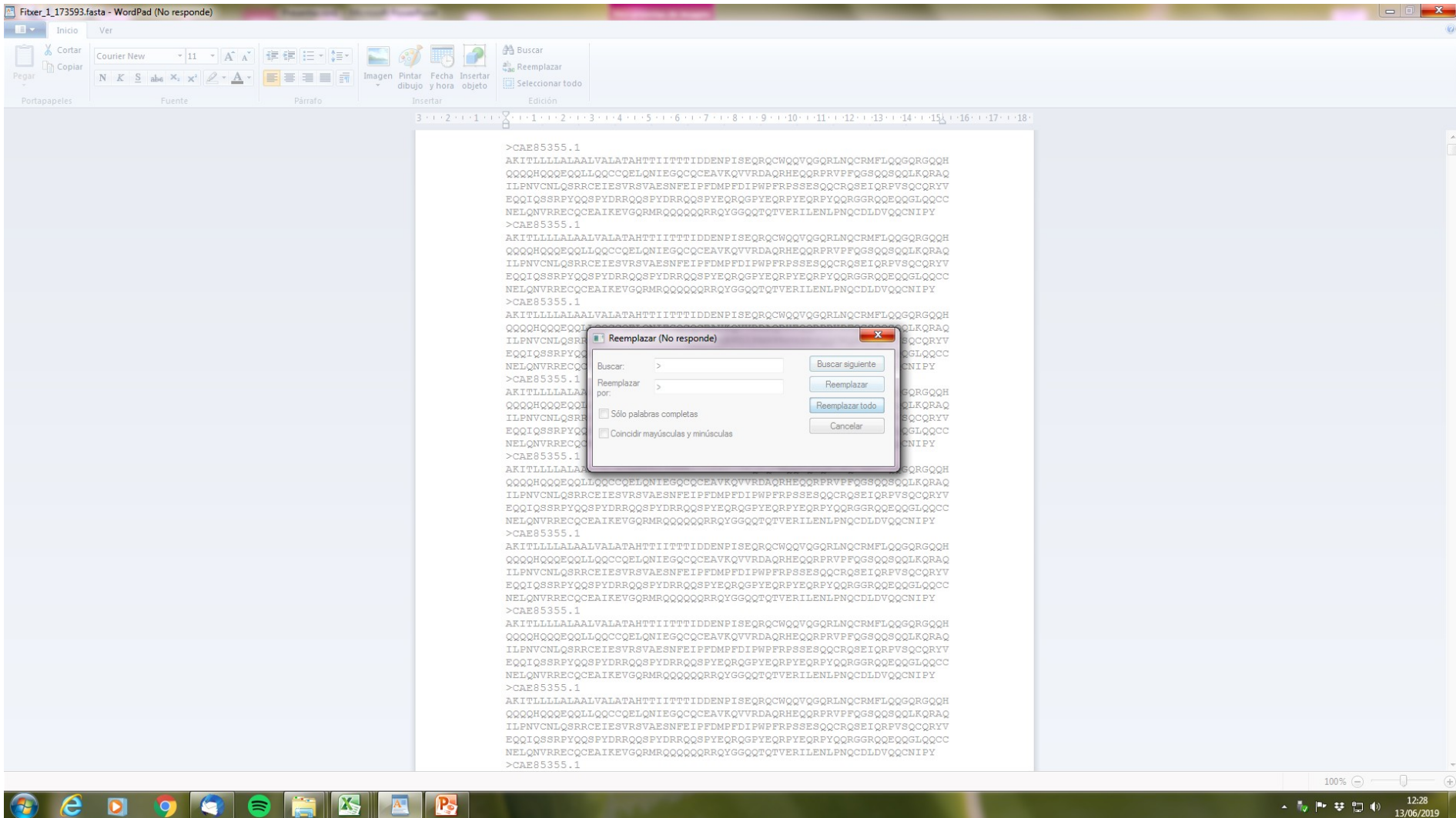
GNU/Linux on Bioinformatics

EJEMPLO: FASTQ



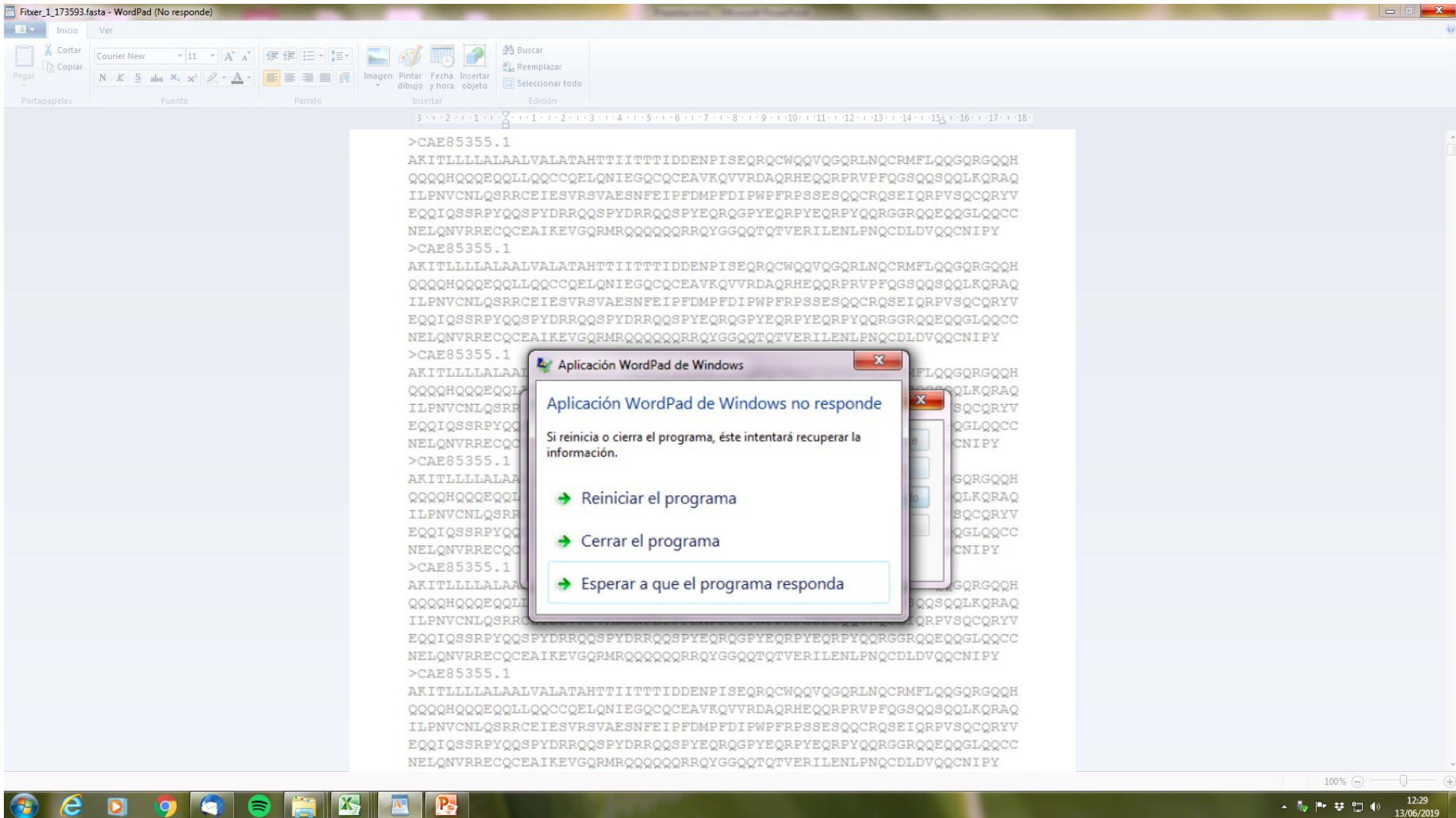
GNU/Linux on Bioinformatics

EJEMPLO: FASTA de 173593 secuencias



GNU/Linux on Bioinformatics

EJEMPLO: FASTA de 173593 secuencias

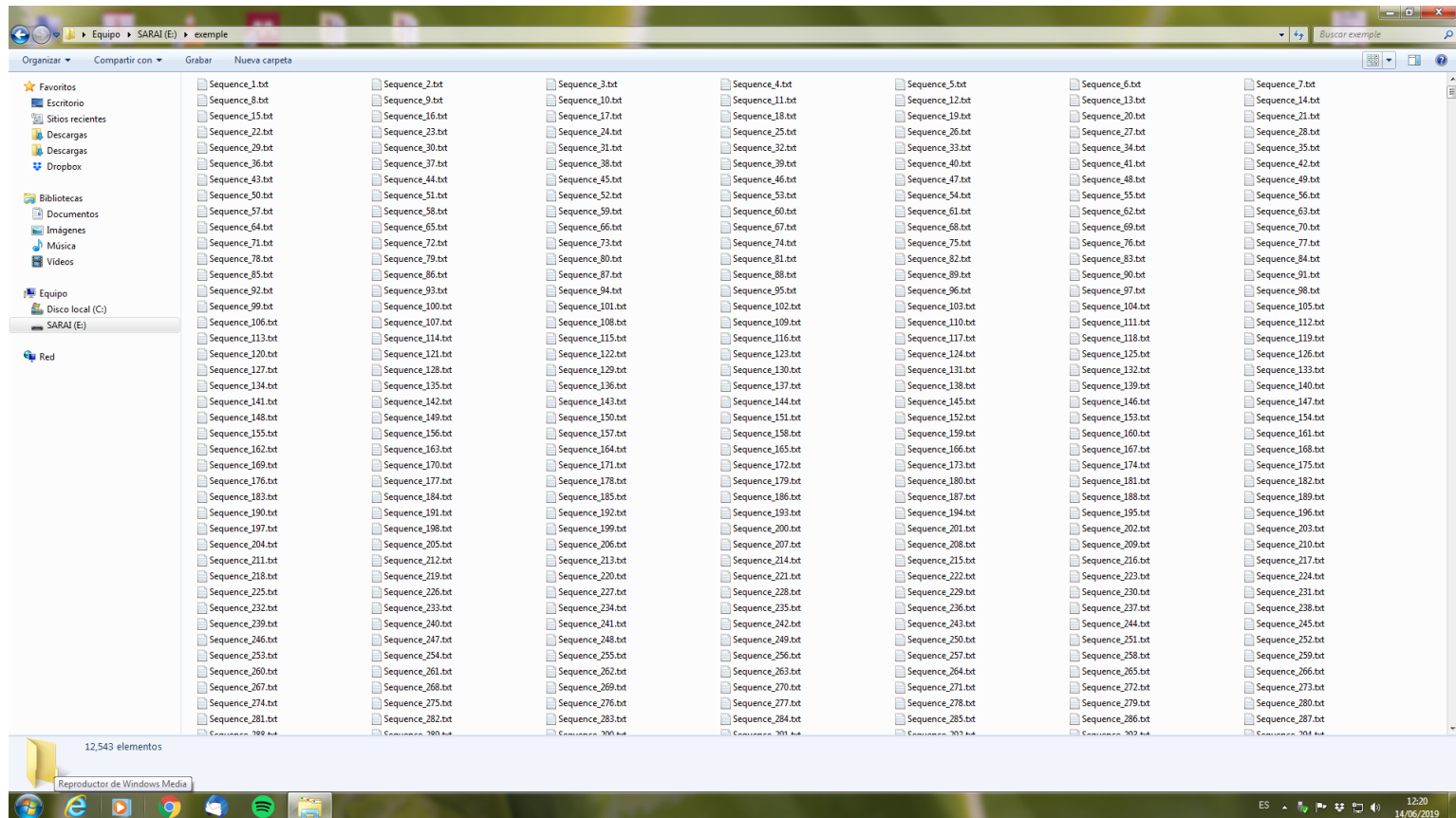


GNU/Linux on Bioinformatics

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



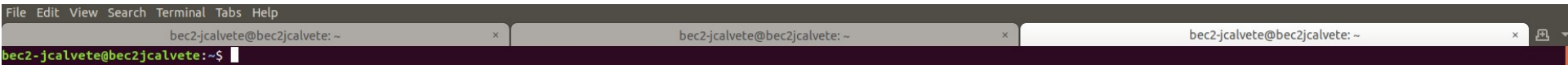
EJEMPLO: RENOMBRAR

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.

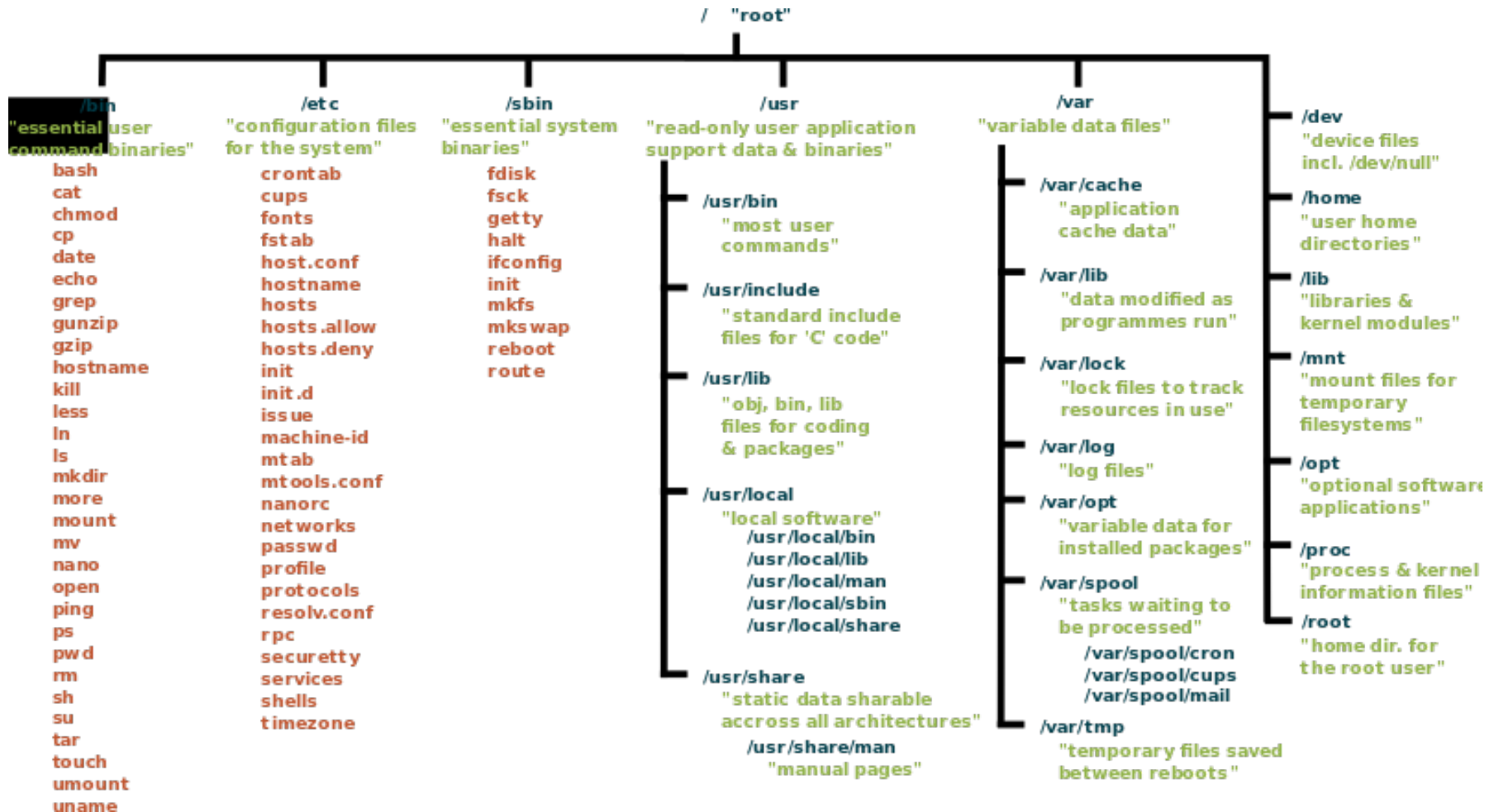
Reproducibilidad. Documentar y repetir el proceso seguido para realizar un análisis con un programa gráfico es muy costoso puesto que es difícil describir la secuencia de clicks y doble clicks que hemos realizado. Por el contrario, los procesos realizados mediante la línea de comandos son muy fáciles 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 <code>/home/</code> .
	/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	/exámenes_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"

```
bec2-jcalvete@bec2jcalvete:~$  
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

<i>Comando</i>	<i>Acción</i>	<i>Ejemplo</i>
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

<i>Comando</i>	<i>Acción</i>	<i>Ejemplo</i>
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

GNU/Linux on Bioinformatics

AYUDA : *man e info*

```
File Edit View Search Terminal Tabs Help
bec2-jcalvete@bec2jcalvete: /media/bec2-jcalvete/Elements/ownCloud/WODA/She
Next: dir invocation, Up: Directory listing

10.1 'ls': List directory contents
=====

The 'ls' program lists information about files (of any type, including
directories).  Options and file arguments can be intermixed arbitrarily,
as usual.

For non-option command-line arguments that are directories, by
default 'ls' lists the contents of directories, not recursively, and
omitting files with names beginning with '.'.  For other non-option
arguments, by default 'ls' lists just the file name.  If no non-option
argument is specified, 'ls' operates on the current directory, acting as
if it had been invoked with a single argument of '.'.

By default, the output is sorted alphabetically, according to the
locale settings in effect.(1)  If standard output is a terminal, the
output is in columns (sorted vertically) and control characters are
output as question marks; otherwise, the output is listed one per line
and control characters are output as-is.

Because 'ls' is such a fundamental program, it has accumulated many
options over the years.  They are described in the subsections below;
within each section, options are listed alphabetically (ignoring case).
The division of options into the subsections is not absolute, since some
options affect more than one aspect of 'ls''s operation.

Exit status:

0 success
1 minor problems (e.g., failure to access a file or directory not
specified as a command line argument.  This happens when listing a
directory in which entries are actively being removed or renamed.)
2 serious trouble (e.g., memory exhausted, invalid option, failure
to access a file or directory specified as a command line argument
or a directory loop)

Also see *note Common options::.

* Menu:
* Which files are listed::
* What information is listed::
* Sorting the output::
* Details about version sort::
* General output formatting::
* Formatting file timestamps::
* Formatting the file names::

----- Footnotes -----
-----Info: (coreutils)ls invocation, 57 lines --Top-----
Welcome to Info version 6.5.  Type H for help, h for tutorial.
```

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>