Curso de Iniciación a la Secuenciación Masiva

BU-ISCIII

Práctica 1 día 2: Manejo y gestión de ficheros

17-21 Junio 2019, 7a Edición, Programa Formación Continua, ISCIII

Descripción

Uno de los puntos fuertes de los sistemas Linux estriba en la facilidad con la que se analizan los ficheros de texto. Estos sistemas incluyen una serie de herramientas que permiten realizar una gran cantidad de manipulaciones en estos ficheros sin necesidad de instalar ninguna herramienta especializada.

Ficheros de texto y binarios

Antes de comenzar a analizar este tipo de ficheros hay que aclarar qué es y qué no es un fichero de texto. Un fichero de texto es un fichero dividido en líneas y cuyo contenido es texto. A pesar de lo que pudiese parecer a priori, un documento de Microsoft Office o de LibreOffice no es un fichero de texto. La información contenida en estos documentos es binaria y sólo los programas especialmente creados para abrir estos ficheros pueden acceder a ella de un modo inteligible. En un documento como en un fichero Word además de texto se guarda la información sobre el formato, imágenes, tablas, etc... Por el contrario en un fichero de texto sólo hay caracteres alfanuméricos (letras y números), retornos de carro y tabuladores.

Los ficheros de texto pueden ser abiertos e inspeccionados sin necesidad de hacer uso de un software especial diseñado para trabajar con ellos. Un documento Word no puede ser leído sino tenemos el Office o LibreOffice instalado pero un fichero de texto se puede ver y editar con las herramientas que vienen instaladas por defecto en el sistema operativo.

Estas herramientas de manejo de ficheros de texto permiten realizar complejas manipulaciones de un modo muy sencillo y son uno de los principales atractivos de los sistemas Linux para el manejo de grandes cantidades de información.

Para esta práctica se va a usar el fichero microarray_adenoma_hk69.csv (/home/alumno/curso_NGS/dia2). En este fichero están almacenados los resultados de un experimento de expresión diferencial en el que se han analizado distintos adenomas.

Este es un fichero tabular en el que la información se representa dividiendo los campos mediante tabuladores (formato tabla). En este caso cada fila del fichero corresponde a una sonda de microarray y cada columna a una propiedad sobre la sonda o sobre el resultado de la hibridación sobre ella.

Lo primero que podemos hacer con un fichero de texto es abrirlo para ver sus contenidos. Existen editores de texto que funcionan en ventanas, como gedit, y editores que funcionan en la terminal, como nano. Por desgracia a veces los ficheros con los que vamos a trabajar son tan grandes que incluso los buenos editores de texto pueden tener problemas para abrirlos.

Otra forma de acceder a los contenidos del fichero es visualizarlo en la terminal utilizando el comando cat:

```
cat microarray_adenoma_hk69.csv
```

Pero si intentáis hacerlo la terminal quedará bloqueada durante bastante tiempo puesto que el fichero es muy grande.

Nota: Si habéis ejecutado el comando anterior y ahora queréis terminar (o matar) el programa que está ejecutándose en el terminal (en este caso cat) podéis utilizar la combinación de teclas ctrl + c. Esto suele hacer que los programas terminen lo que estén haciendo inmediatamente, se apaguen y vuelva a mostrarse el prompt.

Para hacerse una idea del contenido del fichero sin bloquear la terminal se puede mostrar en pantalla tan solo una parte utilizando los comandos head o tail. Tanto a head como a tail se le puede añadir el parámetro –n para ver el número de líneas que desees.

```
head microarray_adenoma_hk69.csv
tail - n 2 microarray_adenoma_hk69.csv
head - n 4 microarray_adenoma_hk69.csv
```

Para abrir ficheros de texto inmensos sin problemas se usan los comandos more o less. No se puede editar el fichero, pero sí navegar por su contenido. Son programas interactivos por lo que cuando se ejecute se abrirá ocupando el terminal y haciendo desaparecer el prompt. En cualquier momento se puede salir pulsando la tecla "q".

```
less microarray_adenoma_hk69.csv
more microarray_adenoma_hk69.csv
```

Algunas de las tareas más habituales en el tratamiento de ficheros de texto van a ser:

- Seleccionar diversas líneas (comando grep)

En archivos de texto a veces necesitamos localizar rápidamente las líneas que contienen cierto tipo de identificador, o localizar una entrada en particular. Esto lo podemos hacer fácilmente con el comando grep, que busca en el archivo una expresión regular o cadena de caracteres determinada y devuelve las líneas que la contienen. Una expresión regular es una secuencia de caracteres que forman un patrón de búsqueda, como pueden ser nuestro ya conocidos '*' para expresar cualquier cadena de caracteres.

Grep también posee cantidad de opciones que modifican su funcionamiento de diversas maneras. Caben destacar las opciones –v, que devuelve las líneas que no contienen el patrón de búsqueda; -r se usa en lugar de indicar un archivo en el que buscar y sirve para buscar en todos los archivos del directorio de trabajo; -i ignora las mayúsculas o minúsculas; -w busca solo palabras enteras; y –n que devuelve el número de la línea donde encuentra el resultado de la búsqueda.

```
grep -w Experiment microarray_adenoma_hk69.csv
```

- Contar líneas, palabras y caracteres (comando wc)

Saber cuántas líneas tiene un archivo es indispensable para saber las dimensiones d datos con las que trabajas. wc puede hacer esto con su opción -l, además de contar caracteres (-m), palabras (-w) o bytes (-c).

```
wc -l microarray_adenoma_hk69.csv
```

- Si el fichero está dividido en campos (como en el caso de la tabla usada en la práctica), seleccionar campos de la tabla (comando cut)

Las tablas con las que se suele trabajar en bioinformática son demasiado grandes para que Excel pueda abrirlas, por lo que hay que utilizar otras herramientas para trabajar con ellas. Una de las más comunes para su exploración es cut, que nos permite escoger determinadas columnas. En combinación con grep (u otras herramientas que seleccionan filas rápidamente) podemos extraer información útil de grandes tablas de datos. El comando cut requiere siempre de las opciones -f, que es el número de columna a extraer (pueden ser rangos o listas separadas por comas) y -d para especificar el separador de campos en el archivo de texto (el tabulador es el separado por defecto y no hace falta especificarlo).

```
grep -w Experiment microarray_adenoma_hk69.csv | cut -f 2 -d'='
```

- Ordenar el contenido del fichero (comando sort)

Con sort podemos finalmente ordenar los una tabla de texto por una de sus columnas, ya sea alfabéticamente en caso de campos de caracteres o en orden ascendente o descendente en caso de numéricos. Para ello, se vale de las opciones -k (número de columna por la que ordenar), -n (especifica que la columna contiene números y no caracteres), -r (ordenar en orden inverso) y -u (eliminar duplicados y mostrar solo elementos únicos).

```
head microarray_adenoma_hk69.csv | sort -k1 -r
```

Ejercicios

1) Saber cuál es la expresión de los genes relacionados con la leucemia en el fichero del microarray. Para ello buscaremos en el archivo las líneas que contienen leukemia ignorando diferencias entre mayúsculas y minúsculas. Si las líneas que coinciden con el patrón son muchas, usar una tubería (|) y pasar la salida del comando grep como entrada al comando less o more.
2) Buscar la palabra leukemia en todos los ficheros presentes el directorio home (/home/usuario).
3) ¿Qué posiciones del fichero están las líneas que cumplen con el patrón leukemia?
4) Buscar la palabra leukemia en las primeras cien líneas del fichero.
5) Buscar la palabra leukemia en las primeras cien líneas del fichero y guardar el resultado en un fichero.
6) Como el contenido del fichero está dividido en campos, seleccionar de la búsqueda anterior solo el nombre y la descripción del gen.
7) Ordenar alfabéticamente los genes relacionados con la leucemia.
8) Contar cuantos genes relacionados con la leucemia hay en el fichero.

Nota: Recordad usar los comandos pwd, cd y ls para conocer vuestra localización, moveros entre directorios y listar el contenido de los directorios respectivamente. Necesitaréis ir al directorio donde se encuentra el archivo 'microarray_adenoma_hk69.csv' para poder trabajar con él más fácilmente (/home/alumno/curso_NGS/dia2).

Soluciones

```
#1
grep -i leukemia microarray_adenoma_hk69.csv | less
-----------------------------------------------------
#2
grep -r leukemia ~ #o grep -r leukemia
-----------------------------------------------------
#3
grep -n leukemia microarray_adenoma_hk69.csv | cut -f 1 -d':'
-----------------------------------------------------
#4
head -n 100 microarray_adenoma_hk69.csv | grep leukemia
-----------------------------------------------------
#5
head -n 100 microarray_adenoma_hk69.csv | grep leukemia > busqueda_leukemia_100.txt
-----------------------------------------------------
#6
grep leukemia microarray_adenoma_hk69.csv | cut -f 3,4
-----------------------------------------------------
#7
grep leukemia microarray_adenoma_hk69.csv | cut -f 3,4 | sort
-----------------------------------------------------
#8
grep leukemia microarray_adenoma_hk69.csv | cut -f 3,4 | sort -u | wc -l
```

Nota final

- Podéis practicar estos ejercicios en cualquier ordenador, no solo en la máquian virtual del curso.
- Estos comandos son universales y funcionan en toda máquina que corra linux y similares, incluyendo macs y WSL.

- La manera más sencilla de practicarlos sin instalar nada es vía http://www.webminal.org/. En esta web puedes crearte un usuario de forma gratuita y abrir una terminal en una máquina remota, todo a través de vuestro explorador web. También contiene tutoriales complementarios que os pueden servir para afianzar lo aprendido hoy o repasar los comandos cuando tengáis necesidad de usarlos.

```
Visita Webminal para practicar en casa: http://www.webminal.org/
```