UNIVERSITÀ DEGLI STUDI DELLA TUSCIA

BIOINFORMATICA 1

LINUX:
FILE SYSTEM – SHELL E COMANDI PRINCIPALI;
APPLICAZIONE AI DATI GENOMICI

Tiziana Castrignanò

LINUX

Comandi: prelievo documenti dalla rete (WGET)

- Il comando **wget** (World Wide Web get) recupera documenti digitali da internet.

In ambiente Linux il comando `wget` è uno strumento utilizzato per scaricare contenuti da server web. Viene eseguito dalla riga di comando e può scaricare file, directory e pagine web intere. Esegue download in background e supporta il download ripetuto automaticamente in caso di interruzione.

Può essere utilizzato con una semplice sintassi come `wget [URL]`, dove `[URL]` è l'indirizzo del file o della pagina che desideri scaricare.

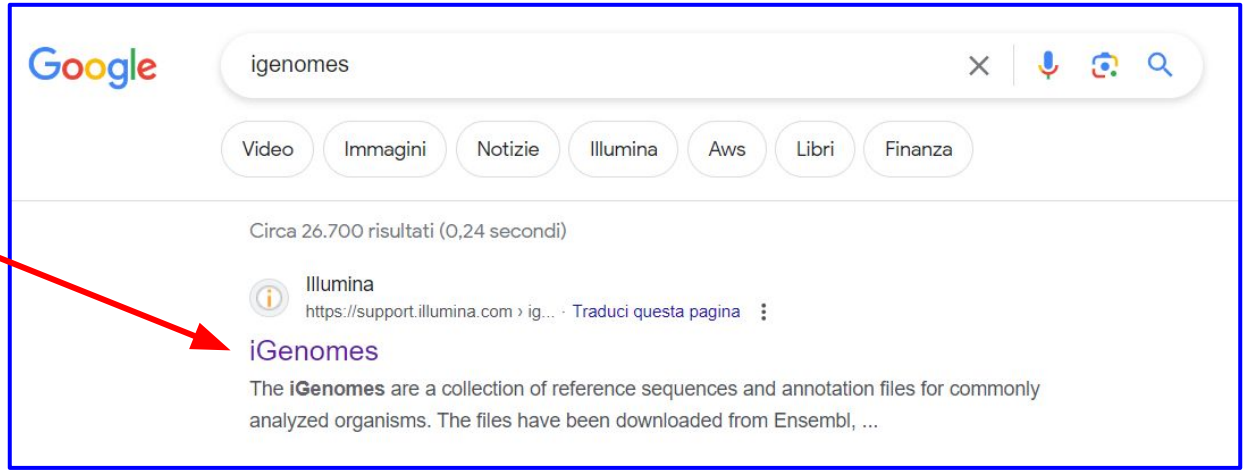
Ad esempio, puoi usare `wget -r [URL]` per scaricare ricorsivamente tutti i file accessibili da una specifica pagina web o directory.

LINUX

Comandi: prelievo documenti dalla rete (WGET)

Vediamo come utilizzare wget nella pratica.

- Per scaricare un file di una release di genoma per prima cosa andiamo alla pagina **igenomes** (Ready-To-Use Reference Sequences and Annotations) di Illumina



LINUX

Comandi: uso di WGET

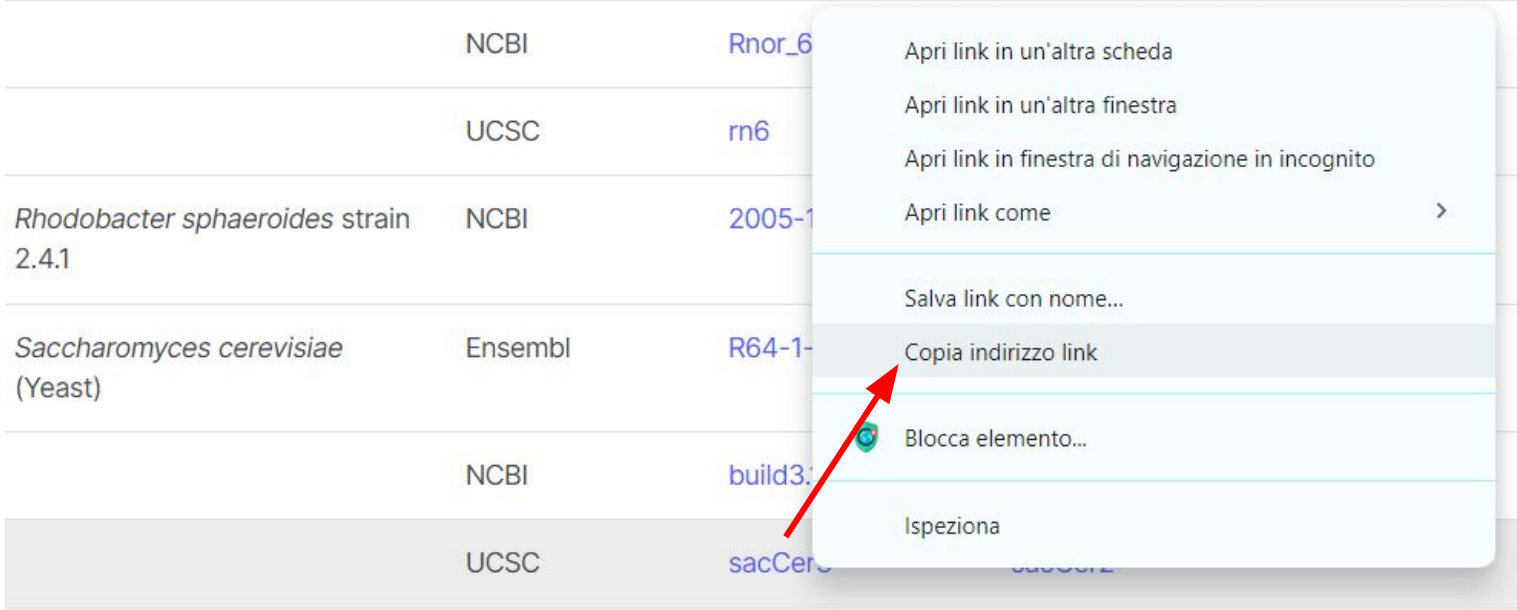
- Ci focalizziamo sulla distribuzione di UCSC di una release del genoma di Saccharomyces cerevisiae (sacCer3).

Rhodobacter sphaeroides strain 2.4.1	NCBI	2005-10-07		
Saccharomyces cerevisiae (Yeast)	Ensembl	R64-1-1	EF4	EF3
	NCBI	build3.1	build2.1	
	UCSC	sacCer3	sacCer2	
Schizosaccharomyces pombe	Ensembl	EF2	EF1	

LINUX

Comandi: uso di WGET

- Posizionare il puntatore del mouse su sacCer3 e spingere sul tasto destro per aprire la tendina e copiare l'indirizzo link



LINUX

Comandi: uso di WGET

- Scaricare in locale sul server il file della distribuzione sacCer3 con il comando wget e tasto destro del mouse. Nella shell quindi scrivete il comando wget digitate uno spazio e cliccate sul tasto destro del mouse per incollare il link copiato precedentemente sul browser.

wget

http://igenomes.illumina.com.s3-website-us-east-1.amazonaws.com/Saccharomyces_cerevisiae/UCSC/sacCer3/Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz

LINUX

Comandi: uso di WGET

```
castri@raganella:/data/TEMP$ wget http://igenomes.illumina.com.s3-website-us-east-1.amazonaws.com/Saccharomyces_cerevisiae/UCSC/sacCer3/Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz
--2024-03-11 22:56:32--  http://igenomes.illumina.com.s3-website-us-east-1.amazonaws.com/Saccharomyces_cerevisiae/UCSC/sacCer3/Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz
Risoluzione di igenomes.illumina.com.s3-website-us-east-1.amazonaws.com (igenomes.illumina.com.s3-website-us-east-1.amazonaws.com)... 52.217.162.61, 54.231.168.157, 52.216.37.53, ...
Connessione a igenomes.illumina.com.s3-website-us-east-1.amazonaws.com (igenomes.illumina.com.s3-website-us-east-1.amazonaws.com)|52.217.162.61|:80... connesso.
Richiesta HTTP inviata, in attesa di risposta... 200 OK
Lunghezza: 71237136 (68M) [application/x-tar]
Salvataggio in: "Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz"

Saccharomyces_cerevisiae_UCSC_sacCer3. 100%[=========================================================================>]  67,94M  21,0MB/s    in 3,2s

2024-03-11 22:56:41 (21,0 MB/s) - "Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz" salvato [71237136/71237136]
```

Con il comando ***ls*** controllate di aver scaricato la distribuzione del genoma ***sacCer3***

```
castri@raganella:/data/TEMP$ ls *gz
Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz
castri@raganella:/data/TEMP$
```

LINUX

Comandi: uso di TAR

```
castri@raganella:/data/TEMP$ ls *gz
Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz
castri@raganella:/data/TEMP$
```

Un file con estensione `.tar.gz` è un archivio compresso in Linux e UNIX. Il formato `.tar` (Tape Archive) raggruppa più file e directory in un unico file (archivio), mentre l'estensione `.gz` indica che l'archivio è stato compresso usando la compressione gzip. Quindi, un file `.tar.gz` (talvolta chiamato "tarball") è un archivio che è stato sia aggregato sia compresso, rendendolo più piccolo e più facile da trasferire o archiviare.

LINUX

Comandi: uso di TAR

Il comando `***tar zxvf file.tar.gz***` è utilizzato per estrarre il contenuto di un archivio .tar.gz. Qui, z indica la decompressione con gzip, x sta per "estrai", v per "verbose" (fornisce un output dettagliato dell'operazione), e f specifica il nome del file archivio.

tar zxvf Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz

LINUX

Comandi: uso di TAR

```
castri@raganella:/data/TEMP$ tar zxvf Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz
Saccharomyces_cerevisiae/UCSC/sacCer3/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Genes
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/README.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/SmallRNA
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/Genes/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/Genes/ChromInfo.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/Genes/refGene.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/README.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/SmallRNA/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2012-03-09-08-20-35/Variation/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-current
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/Genes/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/Genes/ChromInfo.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/Genes/refGene.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/README.txt
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/SmallRNA/
Saccharomyces_cerevisiae/UCSC/sacCer3/Annotation/Archives/archive-2013-03-06-20-12-39/Variation/
Saccharomyces_cerevisiae/UCSC/sacCer3/Sequence/
```

...

LINUX

Comandi: uso di TAR

Vediamo con ***ls -l*** che si è creata una cartella Saccharomyces_cerevisiae che contiene una sottocartella UCSC

```
castri@raganella:/data/TEMP$ ls Sac*
Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz

Saccharomyces_cerevisiae:
UCSC
castri@raganella:/data/TEMP$ ls -l Sac*
-rw-rw-r-- 1 castri castri 71237136 mar  4  2017 Saccharomyces_cerevisiae_UCSC_sacCer3.tar.gz

Saccharomyces_cerevisiae:
totale 4
drwxrwxr-x 3 castri castri 4096 mar 11 23:20 UCSC
castri@raganella:/data/TEMP$
```

...

LINUX

Comandi: TREE

Il comando `***tree***` in Linux visualizza in modo grafico la struttura delle directory e dei file, mostrandoli come un albero ramificato. Questo comando è utile per ottenere una panoramica immediata della gerarchia dei file e delle cartelle in una directory.

Se `***tree***` non è già installato nel tuo sistema, potresti doverlo installare con il gestore di pacchetti della tua distribuzione.

LINUX

Comandi: uso di TREE

Con il comando ***tree Saccharomyces_cerevisiae*** otteniamo la struttura delle directory e dei file, mostrandoli come un albero:

```
castri@raganella:/data/TEMP$ tree Saccharomyces_cerevisiae
Saccharomyces_cerevisiae
└── UCSC
    └── sacCer3
        ├── Annotation
        │   ├── Archives
        │   │   ├── archive-2012-03-09-08-20-35
        │   │   │   ├── Genes
        │   │   │   │   ├── ChromInfo.txt
        │   │   │   │   └── refGene.txt
        │   │   │   ├── README.txt
        │   │   │   ├── SmallRNA
        │   │   │   └── Variation
        │   │   ├── archive-2013-03-06-20-12-39
        │   │   │   ├── Genes
        │   │   │   │   ├── ChromInfo.txt
        │   │   │   │   └── refGene.txt
        │   │   │   ├── README.txt
        │   │   │   ├── SmallRNA
        │   │   │   └── Variation
        │   │   └── archive-current -> archive-2013-03-06-20-12-39
        │   ├── Genes -> Archives/archive-current/Genes
        │   ├── README.txt -> Archives/archive-current/README.txt
        │   └── SmallRNA -> Archives/archive-current/SmallRNA
        └── Sequence
            ├── AbundantSequences
            │   ├── adapter_contam1.fa
            │   ├── chrM.fa -> ../Chromosomes/chrM.fa
```

LINUX

Comandi: GREP

Il comando `***grep***` in Linux è uno strumento da linea di comando usato per cercare stringhe di testo all'interno di file.

Puoi usarlo per cercare una specifica parola o pattern in uno o più file.

La sintassi base è `***grep 'stringa' file***`.

`***grep***` può anche supportare espressioni regolari, che lo rendono molto potente per ricerche testuali complesse.

LINUX

Comandi: set environmental variable

Cerchiamo nell'albero delle dir la directory WholeGenomeFasta.

Una volta identificata definire la variabile WGF ed usarla per entrare nella directory WholeGenomeFasta:

WGF="/data/TEMP/Saccharomyces_cerevisiae/UCSC/sacCer3/Sequence/WholeGenomeFasta/"

cd $WGF

ls *fa produce in standard output

genome.fa

LINUX

Comandi: GREP

Usare il ***grep*** per determinare il numero di sequenze cromosomiche nel file genome.fa

```
castri@raganella:/data/TEMP/Saccharomyces_cerevisiae/UCSC/sacCer3/Sequence/WholeGenomeFasta$ grep ">" genome.fa
>chrI
>chrII
>chrIII
>chrIV
>chrIX
>chrM
>chrV
>chrVI
>chrVII
>chrVIII
>chrX
>chrXI
>chrXII
>chrXIII
>chrXIV
>chrXV
>chrXVI
castri@raganella:/data/TEMP/Saccharomyces_cerevisiae/UCSC/sacCer3/Sequence/WholeGenomeFasta$
```

LINUX

Comandi: `|` (pipe)

Il simbolo `|` in Linux è chiamato "pipe". Viene usato per passare l'output di un comando come input al successivo. Ad esempio, `comando1 | comando2` prende l'output di `comando1` e lo utilizza come input per `comando2`. È un potente strumento per combinare comandi e filtrare i dati.

LINUX

Comandi: `|` (pipe)

Come esempio di utilizzo utilizziamo il comando grep precedente e concateniamolo con il comando wc -l che conta le righe di un file di input.

```
castri@raganella:/data/TEMP/Saccharomyces_cerevisiae/UCSC/sacCer3/Sequence/WholeGenomeFasta$ grep ">" genome.fa|wc -l
17
castri@raganella:/data/TEMP/Saccharomyces_cerevisiae/UCSC/sacCer3/Sequence/WholeGenomeFasta$
```

In output otteniamo il numero 17, pari al numero di cromosomi di Saccharomyces cerevisiae (16 cromosomi nucleari+1 mitocondriale)


Fine