

Day2_Linux_Exercice1_CORRECTED

October 3, 2023

1 First steps on Linux - On Your Own

The program is : * using linux commands seen previously and new linux commands * Counting number of reads, of files * Plot some basics stats * Recover sequences of interest from a GFF (capture seq) * Compare two files (BED) * Bonus: create a mini workflow with pipe

2 where are you? - pwd

Check where you are in the linux filetree

```
[1]: pwd
```

```
/home/jovyan/NOTEBOOK/DAY1-2
```

3 List the content of the directory SV_DATA - ls

Detailed list of files

```
[3]: ls -l ../../SV_DATA
```

```
total 16
drwxr-sr-x 2 jovyan users 4096 Jun 17  2022 ASSEMBLIES
drwxr-sr-x 4 jovyan users 4096 Oct  1 18:59 LONG_READS
drwxr-sr-x 3 jovyan users 4096 Oct  2 14:23 REF
drwxr-sr-x 2 jovyan users 4096 Oct  1 19:54 SHORT_READS
```

3.0.1 With -R, -recursive list subdirectories recursively

```
[17]: ls -R ../../SV_DATA
```

```
../../SV_DATA:
ASSEMBLIES  LONG_READS  REF  SHORT_READS

../../SV_DATA/ASSEMBLIES:
5417_assembly.fasta  A8_assembly.fasta  assembly-stats.txt

../../SV_DATA/LONG_READS:
5417  A8
```

```

../SV_DATA/LONG_READS/5417:
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_0_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_10_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_1_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_11_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_12_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_13_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_14_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_2_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_3_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_4_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_5_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_6_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_7_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_8_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_9_0.fastq.gz

```

```

../SV_DATA/LONG_READS/A8:
fastq_runid_023874091309670a42f7a7c51d363b94fa86dd50_0_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_10_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_1_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_11_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_2_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_3_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_4_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_5_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_6_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_7_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_8_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_9_0.fastq.gz

```

```

../SV_DATA/REF:
GCA_002220235.1_ASM222023v1_genomic.fna
GCA_002220235.1_ASM222023v1_genomic.fna.0123
GCA_002220235.1_ASM222023v1_genomic.fna.amb
GCA_002220235.1_ASM222023v1_genomic.fna.ann
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
GCA_002220235.1_ASM222023v1_genomic.fna.fai
GCA_002220235.1_ASM222023v1_genomic.fna.pac
GCA_002220235.1_ASM222023v1_genomic.gff
qtl.bed

```

```

../SV_DATA/SHORT_READS:
1613_R1.fastq.gz 685_R2.fastq.gz B218_R1.fastq.gz D119_R2.fastq.gz
1613_R2.fastq.gz 716_R1.fastq.gz B218_R2.fastq.gz E2_R1.fastq.gz
1868_R1.fastq.gz 716_R2.fastq.gz B8_R1.fastq.gz E2_R2.fastq.gz
1868_R2.fastq.gz A1_R1.fastq.gz B8_R2.fastq.gz E318_R1.fastq.gz

```

```

4752_R1.fastq.gz  A1_R2.fastq.gz  C218_R1.fastq.gz  E318_R2.fastq.gz
4752_R2.fastq.gz  A8_R1.fastq.gz  C218_R2.fastq.gz  G11_R1.fastq.gz
5417_R1.fastq.gz  A8_R2.fastq.gz  C2_R1.fastq.gz    G11_R2.fastq.gz
5417_R2.fastq.gz  B1_R1.fastq.gz  C2_R2.fastq.gz    H44_R1.fastq.gz
685_R1.fastq.gz   B1_R2.fastq.gz  D119_R1.fastq.gz  H44_R2.fastq.gz

```

4 List the content of the directory SHORT_READS - ls -l

```
[13]: ls -l ../../SV_DATA/SHORT_READS
```

```

total 7067920
-rwxr-x--- 1 jovyan users 115855630 Oct  1 18:24 1613_R1.fastq.gz
-rwxr-x--- 1 jovyan users 129668496 Oct  1 18:24 1613_R2.fastq.gz
-rwxr-x--- 1 jovyan users 108366349 Oct  1 18:24 1868_R1.fastq.gz
-rwxr-x--- 1 jovyan users 120549974 Oct  1 18:24 1868_R2.fastq.gz
-rwxr-x--- 1 jovyan users 106516571 Oct  1 18:24 4752_R1.fastq.gz
-rwxr-x--- 1 jovyan users 118981710 Oct  1 18:24 4752_R2.fastq.gz
-rwxr-x--- 1 jovyan users 123634824 Oct  1 18:24 5417_R1.fastq.gz
-rwxr-x--- 1 jovyan users 149650318 Oct  1 18:24 5417_R2.fastq.gz
-rwxr-x--- 1 jovyan users 113973867 Oct  1 18:24 685_R1.fastq.gz
-rwxr-x--- 1 jovyan users 127925521 Oct  1 18:24 685_R2.fastq.gz
-rwxr-x--- 1 jovyan users 108607336 Oct  1 18:24 716_R1.fastq.gz
-rwxr-x--- 1 jovyan users 121545822 Oct  1 18:24 716_R2.fastq.gz
-rwxr-x--- 1 jovyan users 399056884 Oct  1 18:24 A1_R1.fastq.gz
-rwxr-x--- 1 jovyan users 423279636 Oct  1 18:24 A1_R2.fastq.gz
-rwxr-x--- 1 jovyan users 126125733 Oct  1 18:24 A8_R1.fastq.gz
-rwxr-x--- 1 jovyan users 152151244 Oct  1 18:24 A8_R2.fastq.gz
-rwxr-x--- 1 jovyan users 126130397 Oct  1 18:24 B1_R1.fastq.gz
-rwxr-x--- 1 jovyan users 151606359 Oct  1 18:24 B1_R2.fastq.gz
-rwxr-x--- 1 jovyan users 313591987 Oct  1 18:24 B218_R1.fastq.gz
-rwxr-x--- 1 jovyan users 326574109 Oct  1 18:24 B218_R2.fastq.gz
-rwxr-x--- 1 jovyan users 131613912 Oct  1 18:24 B8_R1.fastq.gz
-rwxr-x--- 1 jovyan users 157774699 Oct  1 18:24 B8_R2.fastq.gz
-rwxr-x--- 1 jovyan users 391619599 Oct  1 18:24 C218_R1.fastq.gz
-rwxr-x--- 1 jovyan users 408700056 Oct  1 18:24 C218_R2.fastq.gz
-rwxr-x--- 1 jovyan users 123634531 Oct  1 18:25 C2_R1.fastq.gz
-rwxr-x--- 1 jovyan users 149384144 Oct  1 18:25 C2_R2.fastq.gz
-rwxr-x--- 1 jovyan users 261162790 Oct  1 18:25 D119_R1.fastq.gz
-rwxr-x--- 1 jovyan users 274437605 Oct  1 18:25 D119_R2.fastq.gz
-rwxr-x--- 1 jovyan users 124037165 Oct  1 18:25 E2_R1.fastq.gz
-rwxr-x--- 1 jovyan users 149177136 Oct  1 18:25 E2_R2.fastq.gz
-rwxr-x--- 1 jovyan users 338337601 Oct  1 18:25 E318_R1.fastq.gz
-rwxr-x--- 1 jovyan users 355161556 Oct  1 18:25 E318_R2.fastq.gz
-rwxr-x--- 1 jovyan users 122978817 Oct  1 18:25 G11_R1.fastq.gz
-rwxr-x--- 1 jovyan users 147336917 Oct  1 18:25 G11_R2.fastq.gz
-rwxr-x--- 1 jovyan users 314797219 Oct  1 18:25 H44_R1.fastq.gz
-rwxr-x--- 1 jovyan users 323490585 Oct  1 18:25 H44_R2.fastq.gz

```

```
[12]: # with -h, --human-readable and only print the 10th lines of the list files
ls -lh ../../SV_DATA/SHORT_READS | head
```

```
total 6.8G
-rwxr-x--- 1 jovyan users 111M Oct  1 18:24 1613_R1.fastq.gz
-rwxr-x--- 1 jovyan users 124M Oct  1 18:24 1613_R2.fastq.gz
-rwxr-x--- 1 jovyan users 104M Oct  1 18:24 1868_R1.fastq.gz
-rwxr-x--- 1 jovyan users 115M Oct  1 18:24 1868_R2.fastq.gz
-rwxr-x--- 1 jovyan users 102M Oct  1 18:24 4752_R1.fastq.gz
-rwxr-x--- 1 jovyan users 114M Oct  1 18:24 4752_R2.fastq.gz
-rwxr-x--- 1 jovyan users 118M Oct  1 18:24 5417_R1.fastq.gz
-rwxr-x--- 1 jovyan users 143M Oct  1 18:24 5417_R2.fastq.gz
-rwxr-x--- 1 jovyan users 109M Oct  1 18:24 685_R1.fastq.gz
```

EXERCICES:

- 1 - List the content of the directory REF by using its absolute path in first then its relative path - `ls` command
- 2 - List the content of the directory LONG_READS with the `ls` command and the option `-R`
- 3 - List the content of the directory ifbdata/public/swissprot with the `ls` command and the option `-al` or `-a -l`

1 - List the content of the directory REF by using its absolute path in first then its relative path - `ls` command

`-t` sort by time, newest first;

`-r`, `--reverse` reverse order while sorting

```
[19]: #1
ls -lrt /home/jovyan/SV_DATA/REF
```

```
total 101648
-rw-r--r-- 1 jovyan users      65 Oct  2 14:17 qtl.bed
-rw-r--r-- 1 jovyan users    681 Sep 25 19:37
GCA_002220235.1_ASM222023v1_genomic.fna.fai
-rw-r--r-- 1 jovyan users 48991633 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
-rw-r--r-- 1 jovyan users 30148640 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.0123
-rw-r--r-- 1 jovyan users     308 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.amb
-rw-r--r-- 1 jovyan users    1587 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.ann
-rw-r--r-- 1 jovyan users 3768582 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.pac
```

```
-rw----- 1 jovyan users 5886829 Sep 23 15:25
GCA_002220235.1_ASM222023v1_genomic.gff
-rw----- 1 jovyan users 15263945 Sep 23 01:29
GCA_002220235.1_ASM222023v1_genomic.fna
```

2 - List the content of the directory LONG_READS with the ls command and the option -R

```
[21]: ls -R ../../SV_DATA/LONG_READS
```

```
../../SV_DATA/LONG_READS:
5417  A8
```

```
../../SV_DATA/LONG_READS/5417:
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_0_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_10_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_1_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_11_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_12_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_13_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_14_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_2_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_3_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_4_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_5_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_6_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_7_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_8_0.fastq.gz
fastq_runid_d9b45d7d98e41184ec96f40f59021521f703067d_9_0.fastq.gz
```

```
../../SV_DATA/LONG_READS/A8:
fastq_runid_023874091309670a42f7a7c51d363b94fa86dd50_0_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_10_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_1_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_11_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_2_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_3_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_4_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_5_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_6_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_7_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_8_0.fastq.gz
fastq_runid_243e56cb3c6676a3d235987c3ae536c7646d460a_9_0.fastq.gz
```

3 - List the content of the directory ifbdata/public/swissprot with the ls command and the option -al or -a -l

```
[ ]: ls ~/ifbdata/public/
```

```
[20]: ls --help
```

Usage: ls [OPTION]... [FILE]...

List information about the FILES (the current directory by default).

Sort entries alphabetically if none of -cftuvSUX nor --sort is specified.

Mandatory arguments to long options are mandatory for short options too.

-a, --all	do not ignore entries starting with .
-A, --almost-all	do not list implied . and ..
--author	with -l, print the author of each file
-b, --escape	print C-style escapes for nongraphic characters
--block-size=SIZE	with -l, scale sizes by SIZE when printing them; e.g., '--block-size=M'; see SIZE format below
-B, --ignore-backups	do not list implied entries ending with ~
-c	with -lt: sort by, and show, ctime (time of last modification of file status information); with -l: show ctime and sort by name; otherwise: sort by ctime, newest first
-C	list entries by columns
--color[=WHEN]	colorize the output; WHEN can be 'always' (default if omitted), 'auto', or 'never'; more info below
-d, --directory	list directories themselves, not their contents
-D, --dired	generate output designed for Emacs' dired mode
-f	do not sort, enable -aU, disable -ls --color
-F, --classify	append indicator (one of */=>@) to entries
--file-type	likewise, except do not append '*'
--format=WORD	across -x, commas -m, horizontal -x, long -l, single-column -1, verbose -l, vertical -C
--full-time	like -l --time-style=full-iso
-g	like -l, but do not list owner
--group-directories-first	group directories before files; can be augmented with a --sort option, but any use of --sort=none (-U) disables grouping
-G, --no-group	in a long listing, don't print group names
-h, --human-readable	with -l and -s, print sizes like 1K 234M 2G etc.
--si	likewise, but use powers of 1000 not 1024
-H, --dereference-command-line	follow symbolic links listed on the command line
--dereference-command-line-symlink-to-dir	follow each command line symbolic link that points to a directory
--hide=PATTERN	do not list implied entries matching shell PATTERN (overridden by -a or -A)
--hyperlink[=WHEN]	hyperlink file names; WHEN can be 'always'

(default if omitted), 'auto', or 'never'

--indicator-style=WORD append indicator with style WORD to entry names:
 none (default), slash (-p),
 file-type (--file-type), classify (-F)

-i, --inode print the index number of each file

-I, --ignore=PATTERN do not list implied entries matching shell PATTERN

-k, --kibibytes default to 1024-byte blocks for disk usage;
 used only with -s and per directory totals

-l use a long listing format

-L, --dereference when showing file information for a symbolic
 link, show information for the file the link
 references rather than for the link itself

-m fill width with a comma separated list of entries

-n, --numeric-uid-gid like -l, but list numeric user and group IDs

-N, --literal print entry names without quoting

-o like -l, but do not list group information

-p, --indicator-style=slash append / indicator to directories

-q, --hide-control-chars print ? instead of nongraphic characters
 --show-control-chars show nongraphic characters as-is (the default,
 unless program is 'ls' and output is a terminal)

-Q, --quote-name enclose entry names in double quotes
 --quoting-style=WORD use quoting style WORD for entry names:
 literal, locale, shell, shell-always,
 shell-escape, shell-escape-always, c, escape
 (overrides QUOTING_STYLE environment variable)

-r, --reverse reverse order while sorting

-R, --recursive list subdirectories recursively

-s, --size print the allocated size of each file, in blocks

-S sort by file size, largest first
 --sort=WORD sort by WORD instead of name: none (-U), size (-S),
 time (-t), version (-v), extension (-X)

--time=WORD change the default of using modification times;
 access time (-u): atime, access, use;
 change time (-c): ctime, status;
 birth time: birth, creation;
 with -l, WORD determines which time to show;
 with --sort=time, sort by WORD (newest first)

--time-style=TIME_STYLE time/date format with -l; see TIME_STYLE below

-t sort by time, newest first; see --time

-T, --tabsize=COLS assume tab stops at each COLS instead of 8

-u with -lt: sort by, and show, access time;
 with -l: show access time and sort by name;
 otherwise: sort by access time, newest first

-U do not sort; list entries in directory order

-v natural sort of (version) numbers within text

-w, --width=COLS set output width to COLS. 0 means no limit

-x list entries by lines instead of by columns

```

-X                      sort alphabetically by entry extension
-Z, --context           print any security context of each file
-1                      list one file per line.  Avoid '\n' with -q or -b
--help                 display this help and exit
--version              output version information and exit

```

The SIZE argument is an integer and optional unit (example: 10K is 10*1024). Units are K,M,G,T,P,E,Z,Y (powers of 1024) or KB,MB,... (powers of 1000). Binary prefixes can be used, too: KiB=K, MiB=M, and so on.

The TIME_STYLE argument can be full-iso, long-iso, iso, locale, or +FORMAT. FORMAT is interpreted like in date(1). If FORMAT is FORMAT1<newline>FORMAT2, then FORMAT1 applies to non-recent files and FORMAT2 to recent files. TIME_STYLE prefixed with 'posix-' takes effect only outside the POSIX locale. Also the TIME_STYLE environment variable sets the default style to use.

Using color to distinguish file types is disabled both by default and with --color=never. With --color=auto, ls emits color codes only when standard output is connected to a terminal. The LS_COLORS environment variable can change the settings. Use the dircolors command to set it.

Exit status:

- 0 if OK,
- 1 if minor problems (e.g., cannot access subdirectory),
- 2 if serious trouble (e.g., cannot access command-line argument).

GNU coreutils online help: <<https://www.gnu.org/software/coreutils/>>
 Full documentation <<https://www.gnu.org/software/coreutils/ls>>
 or available locally via: info '(coreutils) ls invocation'

5 Working with the wildcard *

What is a wildcard ?

- It is a character that can be used as a substitute for any of a class of characters in a search
- Can be used with numerous linux commands

How to use the wildcard *

- * can represent zero characters, all single characters or any string

Example

ls *fastq will display **only the files ended by fastq** in the directory where I am working

Try the following commands using (or not) the wildcard *

```
[1]: ls /home/jovyan/SV_DATA/REF
```

```
GCA_002220235.1_ASM222023v1_genomic.fna
GCA_002220235.1_ASM222023v1_genomic.fna.0123
GCA_002220235.1_ASM222023v1_genomic.fna.amb
GCA_002220235.1_ASM222023v1_genomic.fna.ann
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
GCA_002220235.1_ASM222023v1_genomic.fna.fai
GCA_002220235.1_ASM222023v1_genomic.fna.pac
GCA_002220235.1_ASM222023v1_genomic.gff
qtl.bed
```

```
[2]: ls /home/jovyan/SV_DATA/REF/*fna
```

```
/home/jovyan/SV_DATA/REF/GCA_002220235.1_ASM222023v1_genomic.fna
```

EXERCICES

- 4 - List the content of the directory `SHORT_READS`. - `ls` command
- 5 - List only the files starting by B (in the directory `SHORT_READS`) - `ls` command & *
- 6 - List only the files with the word R1 (in the directory `SHORT_READS`) - `ls` command & *

4 - List the content of the directory `SHORT_READS`. - `ls` command

```
[5]: ls ~/SV_DATA/SHORT_READS
```

```
1613_R1.fastq.gz 685_R2.fastq.gz B218_R1.fastq.gz D119_R2.fastq.gz
1613_R2.fastq.gz 716_R1.fastq.gz B218_R2.fastq.gz E2_R1.fastq.gz
1868_R1.fastq.gz 716_R2.fastq.gz B8_R1.fastq.gz E2_R2.fastq.gz
1868_R2.fastq.gz A1_R1.fastq.gz B8_R2.fastq.gz E318_R1.fastq.gz
4752_R1.fastq.gz A1_R2.fastq.gz C218_R1.fastq.gz E318_R2.fastq.gz
4752_R2.fastq.gz A8_R1.fastq.gz C218_R2.fastq.gz G11_R1.fastq.gz
5417_R1.fastq.gz A8_R2.fastq.gz C2_R1.fastq.gz G11_R2.fastq.gz
5417_R2.fastq.gz B1_R1.fastq.gz C2_R2.fastq.gz H44_R1.fastq.gz
685_R1.fastq.gz B1_R2.fastq.gz D119_R1.fastq.gz H44_R2.fastq.gz
```

5 - List only the files starting by B (in the directory `SHORT_READS`) - `ls` command & *

```
[7]: ls ~/SV_DATA/SHORT_READS/B*
```

```
/home/jovyan/SV_DATA/SHORT_READS/B1_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/B1_R2.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/B218_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/B218_R2.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/B8_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/B8_R2.fastq.gz
```

starting by B or C

```
[8]: ls ~/SV_DATA/SHORT_READS/[BC]*
```

```
/home/jovyan/SV_DATA/SHORT_READS/B1_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B1_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B218_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B218_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B8_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B8_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C218_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C218_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C2_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C2_R2.fastq.gz
```

starting by a letter between B and E

```
[9]: ls ~/SV_DATA/SHORT_READS/[B-E]*
```

```
/home/jovyan/SV_DATA/SHORT_READS/B1_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B1_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B218_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B218_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B8_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B8_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C218_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C218_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C2_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/C2_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/D119_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/D119_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/E2_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/E2_R2.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/E318_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/E318_R2.fastq.gz
```

6 -List only the files with the word R1 (in the directory SHORT_READS) - ls command & *

```
[12]: ls ~/SV_DATA/SHORT_READS/*R1*
```

```
/home/jovyan/SV_DATA/SHORT_READS/1613_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/1868_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/4752_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/5417_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/685_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/716_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/A1_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/A8_R1.fastq.gz  
/home/jovyan/SV_DATA/SHORT_READS/B1_R1.fastq.gz
```

```

/home/jovyan/SV_DATA/SHORT_READS/B218_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/B8_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/C218_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/C2_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/D119_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/E2_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/E318_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/G11_R1.fastq.gz
/home/jovyan/SV_DATA/SHORT_READS/H44_R1.fastq.gz

```

Check that we have the same number of fastq files “Reverse” (R1) and “Forward” (R2)

```
[13]: ls ~/SV_DATA/SHORT_READS/*R1* | wc -l
```

```
18
```

```
[14]: ls ~/SV_DATA/SHORT_READS/*R2* | wc -l
```

```
18
```

Display the begiing of the fastq file

```
[17]: zcat ~/SV_DATA/SHORT_READS/C2_R1.fastq.gz | head -n 5
```

```

@M02713:279:GW200319000:1:2106:14798:1364 1:N:0:GGATCTGA+GGTAAGCT
AACAGTCGCAAAGATTCTCTCAGAGAATGAGAAGACTTCCTTTTCTCGTCGTCATCGTTCTTCTTCACGCCATTTCCACC
TCCACCACCACCACCTTGTGCCATGCTGCTGCTGCTGCTGATGAAACGATGATCGTCTCTCTGACTCGCTCTGACTC
TCTTACTGTGAGGAAACAAACGACAAGTTTTTTTTCCTTCTTTTCTCATCAGACGACGAGAGAGTTTCTCGCAAGTCAA
AATCACCTT
+
AAAAAFFADDADGBG1F3F3BGGGGHHBFFFBGFFGFFFFHHH1GHHEGA/EEHFBG/FEFHGFHB2GHHGGGHFHFHFGF
HFGFHHH/FFAEAGFFHFFHGFHHHHGHFH111FB>GCHGFHGHFFHHG/CCFHHEGHEFFHHHHHFFFCFHHFFGHF
HFHHHHHH<D1FHHHGGEHFHGCGG-CDHHHGGGG/CGHHGGGGGGGGGGGGFG-
AA@-@B@FFFBFFFEF@F?FFFFBBFFFFF/
@M02713:279:GW200319000:1:2106:16854:1377 1:N:0:GGATCTGA+GGTAAGCT

```

```
gzip: stdout: Broken pipe
```

6 Moving into file system using cd and ls command

EXERCICES

- 7 - Go to the directory *REF* and check in the prompt you have changed correctly your working directory `pwd`
- 8 - List the dir content with `ls`

```
[24]: #?
echo "Where am I?"
```

```
pwd

cd /home/jovyan/SV_DATA/REF

echo -e "\nWhere am I?"
pwd
```

Where am I?
/home/jovyan/SV_DATA/REF

Where am I?
/home/jovyan/SV_DATA/REF

[26]: `ls -lrt`

```
total 101648
-rw----- 1 jovyan users 15263945 Sep 23 01:29
GCA_002220235.1_ASM222023v1_genomic.fna
-rw----- 1 jovyan users  5886829 Sep 23 15:25
GCA_002220235.1_ASM222023v1_genomic.gff
-rw-r--r-- 1 jovyan users  3768582 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.pac
-rw-r--r-- 1 jovyan users    1587 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.ann
-rw-r--r-- 1 jovyan users    308 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.amb
-rw-r--r-- 1 jovyan users 30148640 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.0123
-rw-r--r-- 1 jovyan users 48991633 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
-rw-r--r-- 1 jovyan users    681 Sep 25 19:37
GCA_002220235.1_ASM222023v1_genomic.fna.fai
-rw-r--r-- 1 jovyan users    65 Oct  2 14:17 qtl.bed
```

Let's use the command `cd ..`

- `cd ..` : go one folder up
- `cd ../../` : go two folders up

EXERCICES

- 9 - Go to the *LONG_READ/A8* directory using `../`
 - 10 - Go to the *SHORT_READS*. From this directory, and without any change in your working dir, list what's in *REF* directory
 - 11 - List the *ifbdata/public/* directory using `-R` option. What is there in this dir ?
 - 12 - Come back to your home directory using `cd`
-

```
[27]: #9
      cd ../LONG_READS/A8/
      pwd
```

/home/jovyan/SV_DATA/LONG_READS/A8

```
[28]: #10
      cd ../../SHORT_READS
      pwd
```

/home/jovyan/SV_DATA/SHORT_READS

```
[29]: ls ../REF
```

```
GCA_002220235.1_ASM222023v1_genomic.fna
GCA_002220235.1_ASM222023v1_genomic.fna.0123
GCA_002220235.1_ASM222023v1_genomic.fna.amb
GCA_002220235.1_ASM222023v1_genomic.fna.ann
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
GCA_002220235.1_ASM222023v1_genomic.fna.fai
GCA_002220235.1_ASM222023v1_genomic.fna.pac
GCA_002220235.1_ASM222023v1_genomic.gff
qtl.bed
```

```
[30]: #12
      #7
      echo "Where am I?"
      pwd

      cd

      echo -e "\nWhere am I?"
      pwd
```

Where am I?

/home/jovyan/SV_DATA/SHORT_READS

Where am I?

/home/jovyan

7 Some Commands

pwd	Display the full path of the current directory
ls	List all files/directories
cd DIR_NAME	Change the working directory
rm nom_fichier	Remove a directory
cp file1 file2	Make a copy of file1 and calls it file2

cp FILE_NAME DIR_NAME	Copy the file FILE_NAME in the directory DIR_NAME
cp FILE DIR/NEW_FILE	Mix of 2

8 Manipulating Files and Folders

Exercices {Note} Be careful with the ``rm`` command, in linux, there is no bin !!!

- 13 - Remove the file SV-DATA.tar.gz within your home directory - `rm`
- 14 - Copy the files Day1_Get_Data.ipynb and myFirstJupyterbook.ipynb into the directory NOTEBOOK
- 15 - Remove the files Day1_Get_Data.ipynb and myFirstJupyterbook.ipynb in your home directory - `rm command`
- 16 - Create a directory called SCRIPT in your home directory - `mkdir`

```
[32]: ls /home/jovyan -l
```

```
total 7403360
-rw-r--r-- 1 jovyan users      9447 Oct  2 13:35 Day1_Get_Data.ipynb
drwxr-xr-x 4 root  root      4096 Sep 29 06:35 ifbdata
drwxr-sr-x 3 jovyan users     4096 Oct  2 13:20 NOTEBOOK
drwxr-sr-x 6 jovyan users     4096 Oct  1 19:00 SV_DATA
-rw-r--r-- 1 jovyan users 7581002472 Oct  1 21:10 SV_DATA.tar.gz
-rw-r--r-- 1 jovyan users     1425 Oct  2 13:22 Untitled.ipynb
drwxr-xr-x 2 jovyan 1000      4096 Sep 29 06:35 work
```

```
[34]: rm /home/jovyan/SV_DATA.tar.gz
ls /home/jovyan/ -l
```

```
rm: cannot remove '/home/jovyan/SV_DATA.tar.gz': No such file or directory
total 32
-rw-r--r-- 1 jovyan users 9447 Oct  2 13:35 Day1_Get_Data.ipynb
drwxr-xr-x 4 root  root  4096 Sep 29 06:35 ifbdata
drwxr-sr-x 3 jovyan users 4096 Oct  2 13:20 NOTEBOOK
drwxr-sr-x 6 jovyan users 4096 Oct  1 19:00 SV_DATA
-rw-r--r-- 1 jovyan users 1425 Oct  2 13:22 Untitled.ipynb
drwxr-xr-x 2 jovyan 1000 4096 Sep 29 06:35 work
```

9 Let's practice some bioinformatics tools

TIPS

For saving the output of a command to a file, we will use the `>` character

9.1 1 - Manipulation of gff and bed files

- Recover sequences of interest from a GFF (capture seq)
- Compare two files (BED)

Download the file that contains the rice genome annotation (gff format) and the sequence of the 12 chromosomes (fasta format) - wget

- the url of the fasta file (all.chr.con): http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/all.chrs.con
- the url of the gff file (all.gff3) : http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/all.gff3

EXERCICE :

- 1 - Go to the bank directory and type the correct wget command
- 2 - Download the gff file (all.gff3) and the fasta file (all.chrs.con) in the bank directory
- 3 - List the content of the bank directory
- 4 - Count the line number of the gff file
- 5 - Displays the firsts and lasts lines of the gff file

```
[17]: #1
      pwd
      cd ../../SV_DATA/REF
      ls -lrt
```

```
/home/jovyan/NOTEBOOK/DAY1-2
total 101644
-rw----- 1 jovyan users 15263945 Sep 23 01:29
GCA_002220235.1_ASM222023v1_genomic.fna
-rw----- 1 jovyan users  5886829 Sep 23 15:25
GCA_002220235.1_ASM222023v1_genomic.gff
-rw-r--r-- 1 jovyan users  3768582 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.pac
-rw-r--r-- 1 jovyan users    1587 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.ann
-rw-r--r-- 1 jovyan users    308 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.amb
-rw-r--r-- 1 jovyan users 30148640 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.0123
-rw-r--r-- 1 jovyan users 48991633 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
-rw-r--r-- 1 jovyan users    681 Sep 25 19:37
GCA_002220235.1_ASM222023v1_genomic.fna.fai
```

```
[18]: #2
      wget http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/
      ↪pseudomolecules/version_7.0/all.dir/all.chrs.con
      wget http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/
      ↪pseudomolecules/version_7.0/all.dir/all.gff3
```

```
--2023-10-03 12:59:54-- http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sati
va/annotation_dbs/pseudomolecules/version_7.0/all.dir/all.chrs.con
Resolving rice.uga.edu (rice.uga.edu)... 128.192.162.131
Connecting to rice.uga.edu (rice.uga.edu)|128.192.162.131|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 380710511 (363M)
Saving to: 'all.chrs.con'
```

```
all.chrs.con          100%[=====>] 363.07M  5.69MB/s    in 3m 35s
```

```
2023-10-03 13:03:29 (1.69 MB/s) - 'all.chrs.con' saved [380710511/380710511]
```

```
--2023-10-03 13:03:29-- http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sati
va/annotation_dbs/pseudomolecules/version_7.0/all.dir/all.gff3
Resolving rice.uga.edu (rice.uga.edu)... 128.192.162.131
Connecting to rice.uga.edu (rice.uga.edu)|128.192.162.131|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 81498659 (78M)
Saving to: 'all.gff3'
```

```
all.gff3              100%[=====>] 77.72M  2.78MB/s    in 25s
```

```
2023-10-03 13:03:55 (3.07 MB/s) - 'all.gff3' saved [81498659/81498659]
```

```
[19]: ls -lt
```

```
total 553024
-rw-r--r-- 1 jovyan users      681 Sep 25 19:37
GCA_002220235.1_ASM222023v1_genomic.fna.fai
-rw-r--r-- 1 jovyan users 48991633 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.bwt.2bit.64
-rw-r--r-- 1 jovyan users 30148640 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.0123
-rw-r--r-- 1 jovyan users      308 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.amb
-rw-r--r-- 1 jovyan users      1587 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.ann
-rw-r--r-- 1 jovyan users 3768582 Sep 24 10:45
GCA_002220235.1_ASM222023v1_genomic.fna.pac
-rw----- 1 jovyan users 5886829 Sep 23 15:25
GCA_002220235.1_ASM222023v1_genomic.gff
-rw----- 1 jovyan users 15263945 Sep 23 01:29
GCA_002220235.1_ASM222023v1_genomic.fna
-rw-r--r-- 1 jovyan users 81498659 Feb  7  2012 all.gff3
-rw-r--r-- 1 jovyan users 380710511 Oct 31  2011 all.chrs.con
```

More information about gff format

- Counts the number of gene (line with the word “gene”, column 3)

- ```
[20]: head -n 20 all.chrs.con
```

```
[21]: tail -n 20 all.chrs.con
```

17

```

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNN

```

```
[22]: wc -l all.chrs.con
```

```
7464929 all.chrs.con
```

Search and count for > motif in the rice genome (fasta)

```
[23]: grep ">" all.chrs.con
```

```

>Chr1
>Chr2
>Chr3
>Chr4
>Chr5
>Chr6
>Chr7
>Chr8
>Chr9
>Chr10
>Chr11
>Chr12

```

```
[24]: grep ">" all.chrs.con -c
```

```
12
```

Search for the keyword exon in the gff file

```
[25]: grep "exon" all.gff3 | head
```

```

Chr1 MSU_osa1r7 exon 2903 3268 . + .
ID=LOC_Os01g01010.1:exon_1;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 3354 3616 . + .
ID=LOC_Os01g01010.1:exon_2;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 4357 4455 . + .
ID=LOC_Os01g01010.1:exon_3;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 5457 5560 . + .
ID=LOC_Os01g01010.1:exon_4;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 7136 7944 . + .
ID=LOC_Os01g01010.1:exon_5;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 8028 8150 . + .
ID=LOC_Os01g01010.1:exon_6;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 8232 8320 . + .
ID=LOC_Os01g01010.1:exon_7;Parent=LOC_Os01g01010.1

```

```

Chr1 MSU_osa1r7 exon 8408 8608 . + .
ID=LOC_Os01g01010.1:exon_8;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 9210 9617 . + .
ID=LOC_Os01g01010.1:exon_9;Parent=LOC_Os01g01010.1
Chr1 MSU_osa1r7 exon 10104 10187 . + .
ID=LOC_Os01g01010.1:exon_10;Parent=LOC_Os01g01010.1
grep: write error: Broken pipe

```

```
[26]: grep "exon" all.gff3 -c
```

```
312567
```

```
[27]: grep -e "\sexon\s" all.gff3 -c
```

```
312495
```

## EXERCICE

- 1 - search/select only the lines with the gene keyword in the gff file - **grep**
- 2 - save in a new file (all\_gene.gff3) all the lines with the gene keyword in the gff file all.gff3

Compare two bed files

We have found 3 genomic regions of interest from a previous analysis in our lab. The file qtl.bed contains informations about these regions (chromosome, start, stop) / bed format

```
[21]: #1
grep -e "\sgene\s" all.gff3 | head
```

```

Chr1 MSU_osa1r7 gene 2903 10817 . + . ID=LOC_Os01g01010;Name=LOC_Os01g01010;Note=TBC%20domain%20containing%20protein%2C%20expressed
Chr1 MSU_osa1r7 gene 11218 12435 . + . ID=LOC_Os01g01019;Name=LOC_Os01g01019;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 12648 15915 . + . ID=LOC_Os01g01030;Name=LOC_Os01g01030;Note=monocopper%20oxidase%2C%20putative%2C%20expressed
Chr1 MSU_osa1r7 gene 16292 20323 . + . ID=LOC_Os01g01040;Name=LOC_Os01g01040;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 22841 26971 . + . ID=LOC_Os01g01050;Name=LOC_Os01g01050;Note=R3H%20domain%20containing%20protein%2C%20expressed
Chr1 MSU_osa1r7 gene 27136 28651 . + . ID=LOC_Os01g01060;Name=LOC_Os01g01060;Note=40S%20ribosomal%20protein%20S5%2C%20putative%2C%20expressed
Chr1 MSU_osa1r7 gene 29818 34493 . + . ID=LOC_Os01g01070;Name=LOC_Os01g01070;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 35581 41180 . + . ID=LOC_Os01g01080;Name=LOC_Os01g01080;Note=decarboxylase%2C%20putative%2C%20expressed
Chr1 MSU_osa1r7 gene 44595 47526 . + . ID=LOC_Os01g01090;Name=LOC_Os01g01090;Note=expressed%20protein

```

```
s01g01090;Name=LOC_0s01g01090;Note=retrotransposon%20protein%2C%20putative%2C%20
unclassified%2C%20expressed
Chr1 MSU_osa1r7 gene 47856 53412 . - . ID=LOC_0
s01g01100;Name=LOC_0s01g01100;Note=retrotransposon%20protein%2C%20putative%2C%20
Ty3-gypsy%20subclass%2C%20expressed
grep: write error: Broken pipe
```

```
[23]: #1
grep -e "\sgene\s" all.gff3 -c
grep -e "\sgene\s" all.gff3 > all_genes.gff3
```

55986

```
[24]: wc -l all_genes.gff3
```

55986 all\_genes.gff3

```
[]: wget https://itrop.ird.fr/trainingts/qt1.bed
```

```
[]: cat qt1.bed
```

Recover sequences of interest from a gff (capture seq)

We want to extract sequences for each genomic region using `bedtools getfasta` or `fastaFromBed`  
`fastFromBed -fi sequences.fasta -bed bed file`

Link : <https://bioweb.pasteur.fr/docs/modules/bedtools/2.17.0/content/fastafromBed.html>

```
[39]: fastaFromBed
```

Tool: `bedtools getfasta` (aka `fastaFromBed`)

Version: v2.30.0

Summary: Extract DNA sequences from a fasta file based on feature coordinates.

Usage: `bedtools getfasta [OPTIONS] -fi <fasta> -bed <bed/gff/vcf>`

Options:

|                     |                                                             |
|---------------------|-------------------------------------------------------------|
| <code>-fi</code>    | Input FASTA file                                            |
| <code>-fo</code>    | Output file (opt., default is STDOUT)                       |
| <code>-bed</code>   | BED/GFF/VCF file of ranges to extract from <code>-fi</code> |
| <code>-name</code>  | Use the name field and coordinates for the FASTA header     |
| <code>-name+</code> | (deprecated) Use the name field and coordinates for the     |

FASTA header

|                        |                                                                                             |
|------------------------|---------------------------------------------------------------------------------------------|
| <code>-nameOnly</code> | Use the name field for the FASTA header                                                     |
| <code>-split</code>    | Given BED12 fmt., extract and concatenate the sequences from the BED "blocks" (e.g., exons) |
| <code>-tab</code>      | Write output in TAB delimited format.                                                       |
| <code>-bedOut</code>   | Report extract sequences in a tab-delimited BED format                                      |

instead of in FASTA format.

```
-s - Default is FASTA format.
 Force strandedness. If the feature occupies the
antisense, strand, the sequence will be reverse complemented.
 - By default, strand information is ignored.
 -fullHeader Use full fasta header.
 - By default, only the word before the first space or
tab is used.
 -rna The FASTA is RNA not DNA. Reverse complementation handled
accordingly.
```

```
[6]: cd /home/jovyan/SV_DATA/REF
```

```
[20]: #index ref required by fastaFromBed
 samtools faidx all.chrs.con
```

```
[8]: fastaFromBed -fi all.chrs.con -bed qtl.bed -fo qtl.fasta
```

```
[10]: wc -l qtl.fasta
```

```
6 qtl.fasta
```

```
[16]: # Convert fasta generated by fastafrom bed (with one line that contained all
 ↪ the sequence extracted for each locus)
 # into a fasta with 60 pb per line
 seqtk seq -l 60 qtl.fasta > qtl.formatseqtk.fasta
```

```
[17]: wc -l qtl.formatseqtk.fasta
```

```
9919 qtl.formatseqtk.fasta
```

```
[19]: # print the line with > and the two following lines
 grep ">" qtl.formatseqtk.fasta -A 2
```

```
>Chr1:2615965-2700945
CAATTAAAATTTCAGGTTTAATAGGAGCAAGAAGGCAGGATGAACGAGGAGTGGCCCTAGCG
AATACACCCCTTACCACTCCCCCTTCTCCTTTTCTAGCCTCCACCCCTCATCAATGTTGG
--
>Chr1:9201049-9467749
TCACCCTAGGGCATAGCCCTGCGCCCCTATCCAAGCCGTTATCTCTATCCGTCGTCGTCG
CTCTTGTCGCGTCGCGTCGCCATCTTGTCGCCCCCGGTTGAGACCGCTGGGTAATGCGG
--
>Chr4:14902955-15146157
```

```
AGTACACCTTTCCCAACTTTTCAGAGGATAAAGGGATGCATTCTCAAGACCTCCCAGTCCA
AGGCTAGATAGGGCATTCTAGATAGGGCATTCTAATTGTTTAGTTCACAGTATAAAAC
```

Compare two files

We want to extract annotation information (all\_gene.gff3) for each genomic region (qtl.bed) using `bedtools intersect`

Link : <https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html>

```
[32]: # -wo Write the original A and B entries plus the number of base pairs
 ↪ of overlap between the two features.
 # Only A features with overlap are reported
 bedtools intersect -a all_genes.gff3 -b qtl.bed -wo | head
```

```
Chr1 MSU_osa1r7 gene 2613656 2623196 . - .
ID=LOC_Os01g05510;Name=LOC_Os01g05510;Note=expressed%20protein Chr1 2615965
2700945 7231
Chr1 MSU_osa1r7 gene 2629374 2630050 . + .
ID=LOC_Os01g05520;Name=LOC_Os01g05520;Note=expressed%20protein Chr1 2615965
2700945 677
Chr1 MSU_osa1r7 gene 2632554 2638081 . + .
ID=LOC_Os01g05530;Name=LOC_Os01g05530;Note=expressed%20protein Chr1 2615965
2700945 5528
Chr1 MSU_osa1r7 gene 2640937 2644665 . - .
ID=LOC_Os01g05540;Name=LOC_Os01g05540;Note=expressed%20protein Chr1 2615965
2700945 3729
Chr1 MSU_osa1r7 gene 2651134 2651589 . - .
ID=LOC_Os01g05550;Name=LOC_Os01g05550;Note=expressed%20protein Chr1 2615965
2700945 456
Chr1 MSU_osa1r7 gene 2652648 2655639 . + .
ID=LOC_Os01g05560;Name=LOC_Os01g05560;Note=expressed%20protein Chr1 2615965
2700945 2992
Chr1 MSU_osa1r7 gene 2660822 2661301 . + .
ID=LOC_Os01g05570;Name=LOC_Os01g05570;Note=expressed%20protein Chr1 2615965
2700945 480
Chr1 MSU_osa1r7 gene 2663539 2664072 . - . ID=LOC_0
s01g05580;Name=LOC_Os01g05580;Note=Exo70%20exocyst%20complex%20subunit%20family%
2C%20putative%2C%20expressed Chr1 2615965 2700945 534
Chr1 MSU_osa1r7 gene 2665412 2666458 . - . ID=LOC_0
s01g05585;Name=LOC_Os01g05585;Note=metallothionein%2C%20putative%2C%20expressed
Chr1 2615965 2700945 1047
Chr1 MSU_osa1r7 gene 2666504 2667785 . + . ID=LOC_0
s01g05590;Name=LOC_Os01g05590;Note=retrotransposon%20protein%2C%20putative%2C%20
unclassified%2C%20expressed Chr1 2615965 2700945 1282
```

```
[27]: # select in the result only some column : from 1st column to 5th plus 9th column
 bedtools intersect -a all_genes.gff3 -b qtl.bed -wo | cut -f1-5,9 | head
```

```

Chr1 MSU_osa1r7 gene 2613656 2623196
ID=LOC_Os01g05510;Name=LOC_Os01g05510;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2629374 2630050
ID=LOC_Os01g05520;Name=LOC_Os01g05520;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2632554 2638081
ID=LOC_Os01g05530;Name=LOC_Os01g05530;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2640937 2644665
ID=LOC_Os01g05540;Name=LOC_Os01g05540;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2651134 2651589
ID=LOC_Os01g05550;Name=LOC_Os01g05550;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2652648 2655639
ID=LOC_Os01g05560;Name=LOC_Os01g05560;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2660822 2661301
ID=LOC_Os01g05570;Name=LOC_Os01g05570;Note=expressed%20protein
Chr1 MSU_osa1r7 gene 2663539 2664072 ID=LOC_Os01g05580;Name=LOC_Os01g
05580;Note=Exo70%20exocyst%20complex%20subunit%20family%2C%20putative%2C%20expre
ssed
Chr1 MSU_osa1r7 gene 2665412 2666458 ID=LOC_Os01g05585;Name=LOC_Os01g
05585;Note=metallothionein%2C%20putative%2C%20expressed
Chr1 MSU_osa1r7 gene 2666504 2667785 ID=LOC_Os01g05590;Name=LOC_Os01g
05590;Note=retrotransposon%20protein%2C%20putative%2C%20unclassified%2C%20expres
sed
cut: write error: Broken pipe

```

```
[28]: bedtools intersect -a all_genes.gff3 -b qtl.bed -wo | cut -f1-5,9,11,12 | head
```

```

Chr1 MSU_osa1r7 gene 2613656 2623196
ID=LOC_Os01g05510;Name=LOC_Os01g05510;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2629374 2630050
ID=LOC_Os01g05520;Name=LOC_Os01g05520;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2632554 2638081
ID=LOC_Os01g05530;Name=LOC_Os01g05530;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2640937 2644665
ID=LOC_Os01g05540;Name=LOC_Os01g05540;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2651134 2651589
ID=LOC_Os01g05550;Name=LOC_Os01g05550;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2652648 2655639
ID=LOC_Os01g05560;Name=LOC_Os01g05560;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2660822 2661301
ID=LOC_Os01g05570;Name=LOC_Os01g05570;Note=expressed%20protein 2615965 2700945
Chr1 MSU_osa1r7 gene 2663539 2664072 ID=LOC_Os01g05580;Name=LOC_Os01g
05580;Note=Exo70%20exocyst%20complex%20subunit%20family%2C%20putative%2C%20expre
ssed 2615965 2700945
Chr1 MSU_osa1r7 gene 2665412 2666458 ID=LOC_Os01g05585;Name=LOC_Os01g
05585;Note=metallothionein%2C%20putative%2C%20expressed 2615965 2700945
Chr1 MSU_osa1r7 gene 2666504 2667785 ID=LOC_Os01g05590;Name=LOC_Os01g
05590;Note=retrotransposon%20protein%2C%20putative%2C%20unclassified%2C%20expres
sed 2615965 2700945

```

cut: write error: Broken pipe

```
[29]: bedtools intersect -a all_genes.gff3 -b qtl.bed -wo | cut -f1-5,9,11,12 > qtl.
 ↪inteseect.genes.bed
```

```
[33]: # count the number of genes in the three QTL
 wc -l qtl.inteseect.genes.bed
```

94 qtl.inteseect.genes.bed

```
[35]: # Go into the correct directory to download notebook - cd command
 mkdir -p /home/jovyan/NOTEBOOK/DAY3-4
 cd /home/jovyan/NOTEBOOK/DAY3-4
 wget https://itrop.ird.fr/trainings/notebook/DAY3-4/Day3_Mapping_Teaching_CORR.
 ↪ipynb --no-check-certificate
```

--2023-10-03 13:33:53--

https://itrop.ird.fr/trainings/notebook/DAY3-4/Day3\_Mapping\_Teaching\_CORR.ipynb

Resolving itrop.ird.fr (itrop.ird.fr)... 91.203.35.184

Connecting to itrop.ird.fr (itrop.ird.fr)|91.203.35.184|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 85197 (83K)

Saving to: 'Day3\_Mapping\_Teaching\_CORR.ipynb'

Day3\_Mapping\_Teachi 100%[=====>] 83.20K --.-KB/s in 0.03s

2023-10-03 13:33:53 (3.05 MB/s) - 'Day3\_Mapping\_Teaching\_CORR.ipynb' saved

[85197/85197]