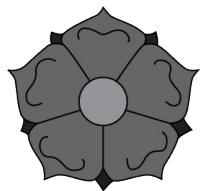


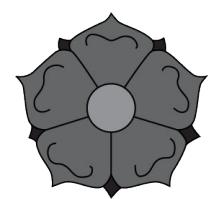
Workshop on Genomics 2026

unix



Mercè Montoliu Nerín

January 13th, 2026



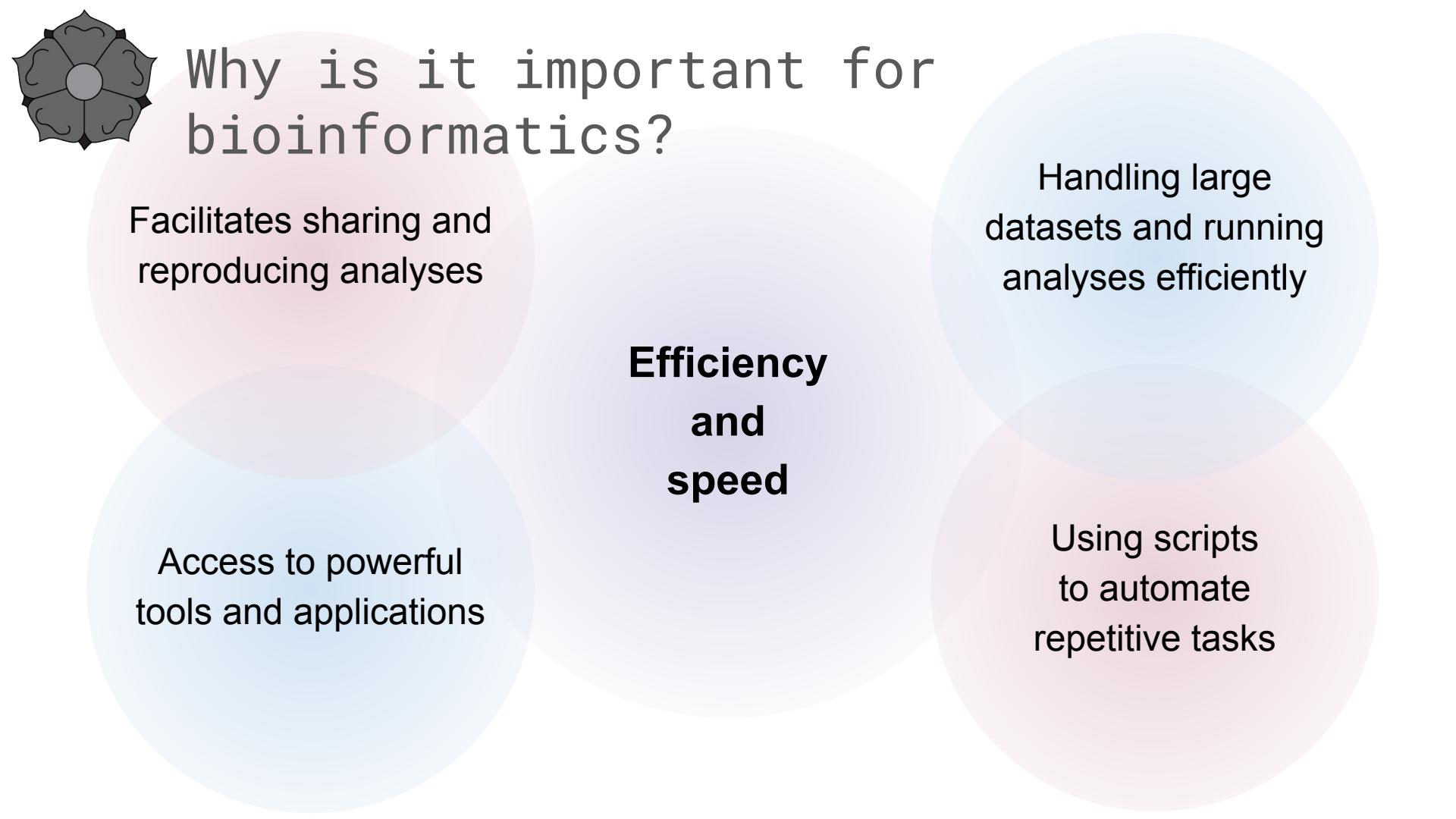
What is UNIX?

Operating system

powerful

multi-user

multitasking



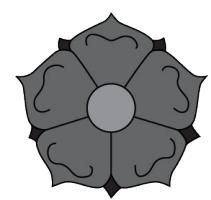
Why is it important for bioinformatics?

Facilitates sharing and reproducing analyses

Handling large datasets and running analyses efficiently

Access to powerful tools and applications

Using scripts to automate repetitive tasks

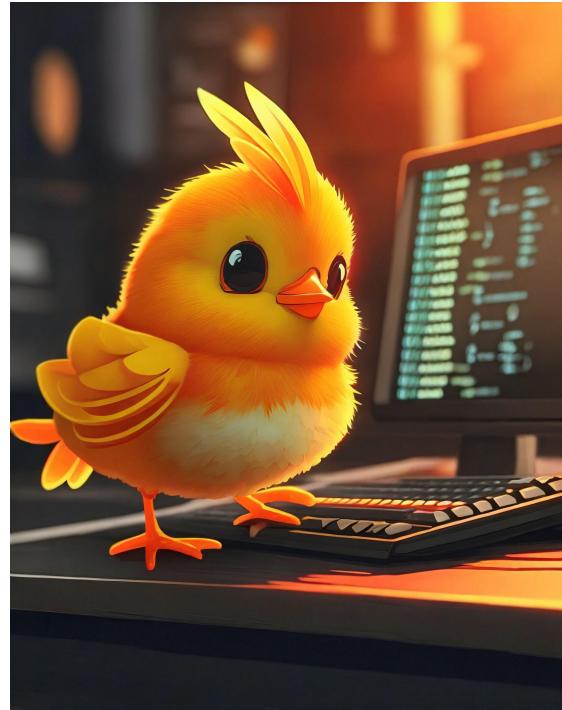


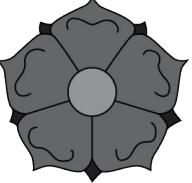
The terminal



Make it comfortable to work in

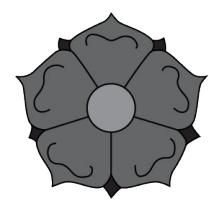
- Resize the window
- Change the font size
- Open multiple terminal windows (or tabs)
- Make sure you have the right combination of colours that work for **you**.



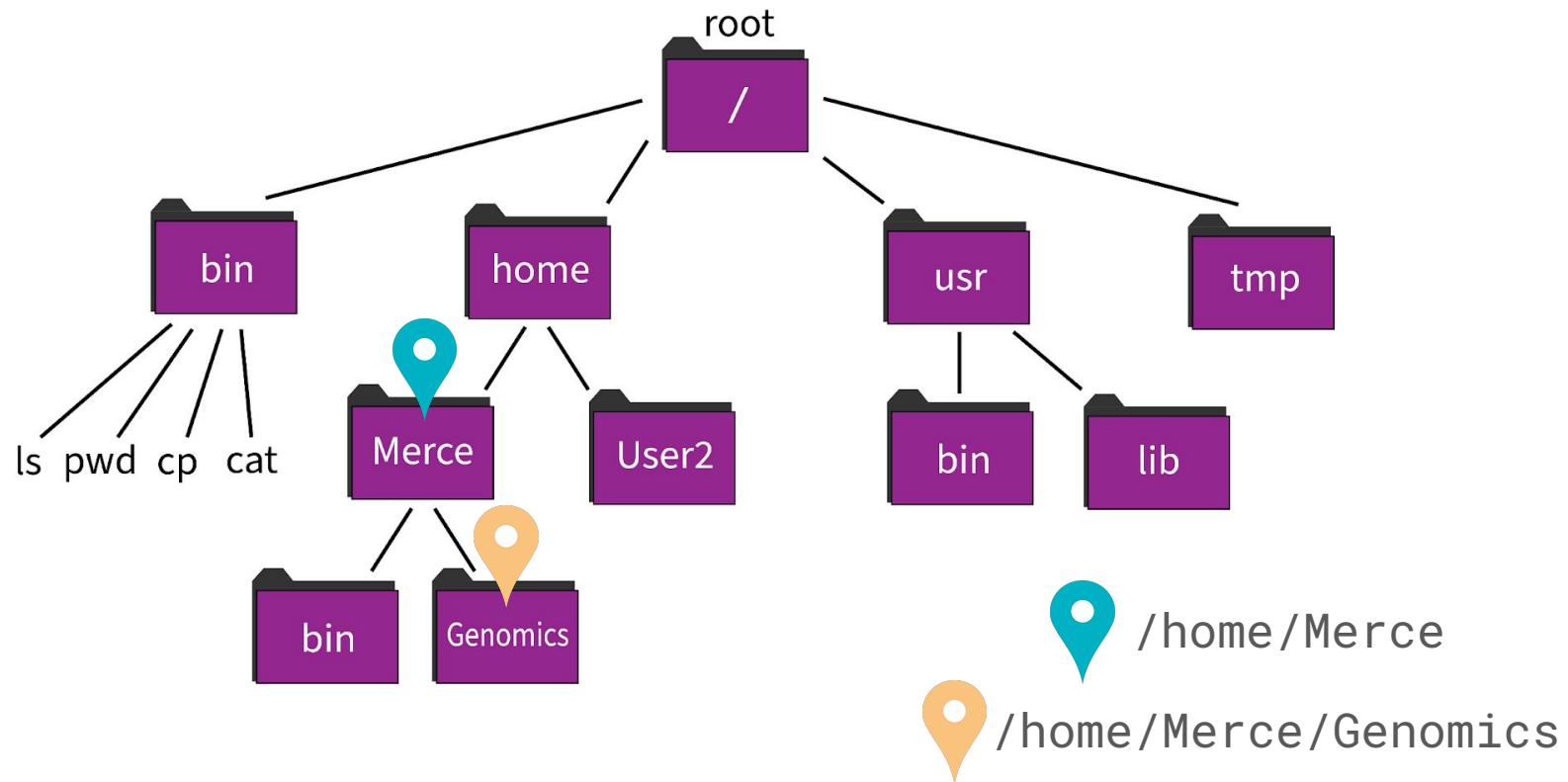


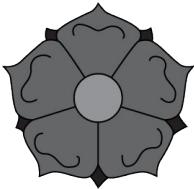
The terminal

```
... - ~ - genomics@161.116.37.121 - ~ - sah_ferrer@161.116.37.121 - ~ - Google Drive - bash +  
dwxr-xr-x@ 342 apache staff 13K 21 Aug 2022 Boda Joan i Josie 22 de desembre de 2020  
dwxr-xr-x@ 13 apache staff 4169 21 Aug 2022 potential_cover_images  
dwxr-xr-x@ 8 apache staff 2568 21 Aug 2022 Anillament_divulgacio  
dwxr-xr-x@ 11 apache staff 3528 21 Aug 2022 Leis  
dwxr-xr-x@ 4 apache staff 1288 21 Aug 2022 Hercules  
dwxr-xr-x@ 3 apache staff 946 21 Aug 2022 Castells  
dwxr-xr-x@ 43 apache staff 1,3K 21 Aug 2022 Concurs_TGN  
dwxr-xr-x@ 36 apache staff 1,1K 21 Aug 2022 Images_baldrigues  
dwxr-xr-x@ 5 apache staff 1468 21 Aug 2022 Sacolla_2019  
dwxr-xr-x@ 38 apache staff 1,2K 21 Aug 2022 TFM  
dwxr-xr-x@ 3 apache staff 968 21 Aug 2022 Public  
dwxr-xr-x@ 11 apache staff 3528 21 Aug 2022 Photos  
dwxr-xr-x@ 19 apache staff 6988 24 Aug 2022 Veus_procel  
dwxr-xr-x@ 9 apache staff 2888 24 Aug 2022 Magnetoreception in Birds  
dwxr-xr-x@ 6 apache staff 1928 1 Oct 2022 rmuu_pongan  
dwxr-xr-x@ 3 apache staff 968 12 Nov 2022 Herencia  
dwxr-xr-x@ 18 apache staff 3298 28 Nov 15:54 xavi TFM  
dwxr-xr-x@ 9 apache staff 2888 14 Jan 09:54 Comandes-pressupostos  
dwxr-xr-x@ 55 apache staff 1,7K 20 Jan 17:07 Tesi  
dwxr-xr-x@ 3 apache staff 968 20 Jan 17:29 Pares  
dwxr-xr-x@ 20 apache staff 6468 25 Jan 20:00 CV  
dwxr-xr-x@ 29 apache staff 9288 25 Jan 22:08 calo.pillars_hercules  
dwxr-xr-x@ 4 apache staff 1288 26 Jan 14:18 Gliquidoro_PhD  
dwxr-xr-x@ 3 apache staff 968 30 Jan 20:13 seafan  
dwxr-xr-x@ 38 apache staff 1,2K 3 Mar 12:05 Fotos_ocells_text_Borja  
dwxr-xr-x@ 11 apache staff 3528 3 Mar 13:07 Lobsters_Josie  
dwxr-xr-x@ 2891 apache staff 98K 6 Mar 15:21 Camera Uploads  
dwxr-xr-x@ 3 apache staff 968 19 Mar 18:21 World_bird_list_JFerrer  
dwxr-xr-x@ 13 apache staff 4168 22 Mar 20:59 Seguencies-2  
dwxr-xr-x@ 3 apache staff 968 3 Apr 09:56 Prizes  
dwxr-xr-x@ 22 apache staff 7948 5 Apr 10:53 Protocols  
dwxr-xr-x@ 21 apache staff 6728 5 Apr 12:09 Postdoc_Milan  
dwxr-xr-x@ 11 apache staff 3528 7 Apr 09:28 papers_reviewed  
-rw-r--r--@ 2 apache staff 0B 4 May 11:24 Icon?  
dwxr-xr-x@ 33 apache staff 1,0K 7 May 17:54 Joan_Josie_life_paperwork  
dwxr-xr-x@ 43K apache staff 3528 15 May 16:42 Screenshots  
base@genomics-MacBook-Pro.local:~[appleDropbox]$
```

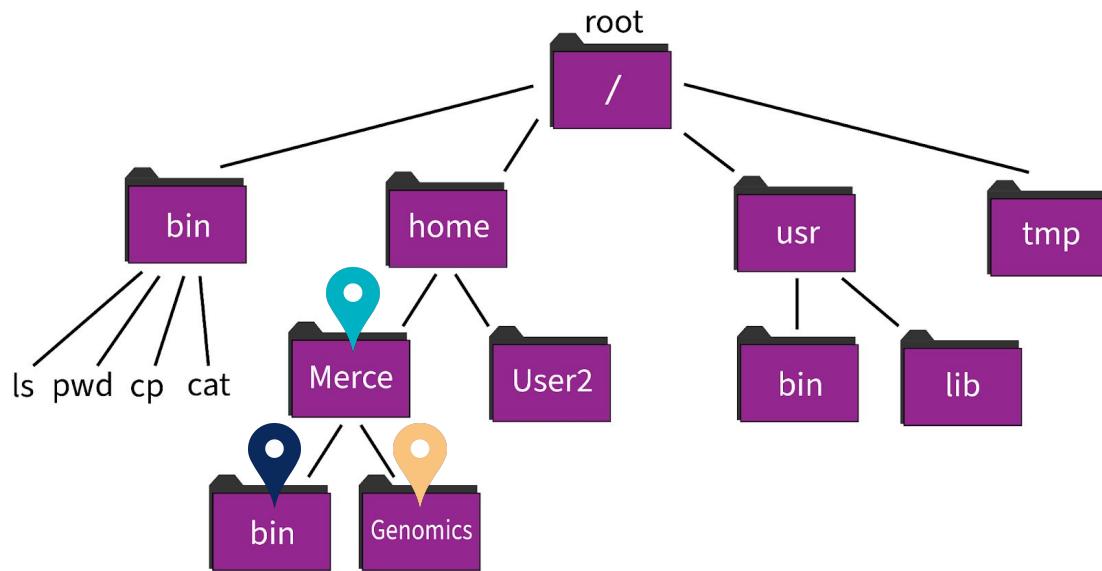


File system organization





Paths - Absolute vs Relative

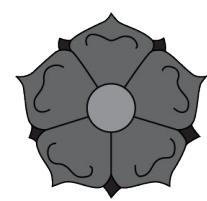


Absolute paths

-  /home/Merce
-  /home/Merce/Genomics
-  /home/Merce/bin

. refers to our current location

... refers to the location above us



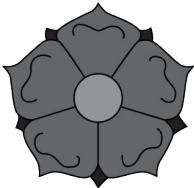
File system navigation

pwd - where am I?



cd - change directory





File system navigation

pwd - where am I?

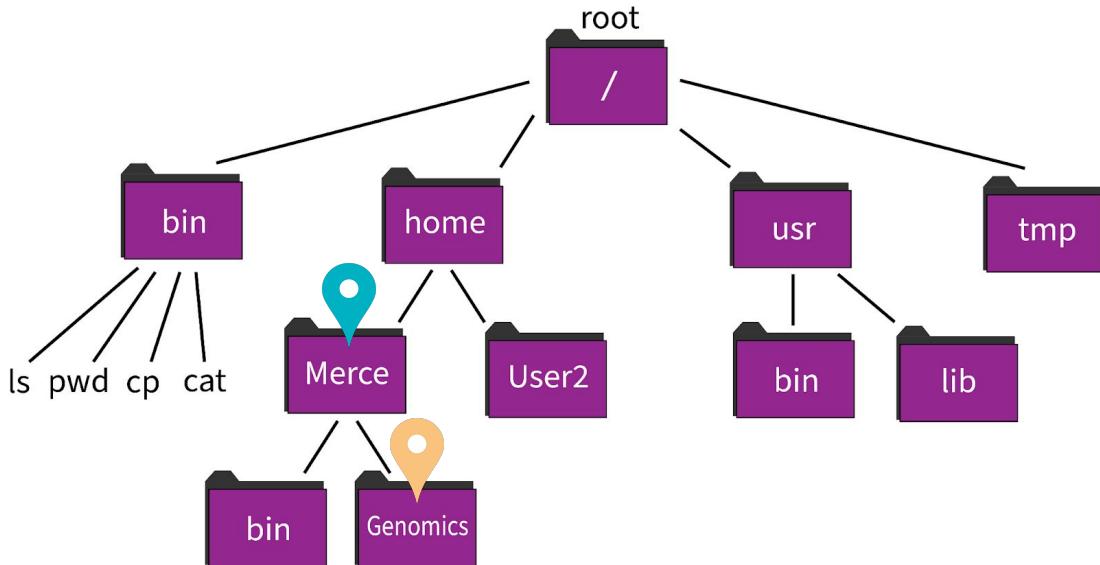
📍 /home/Merce

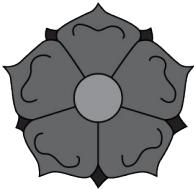


cd - change directory



```
> cd /home/Merce/Genomics  
> cd ./Genomics
```





File system navigation

pwd - where am I?



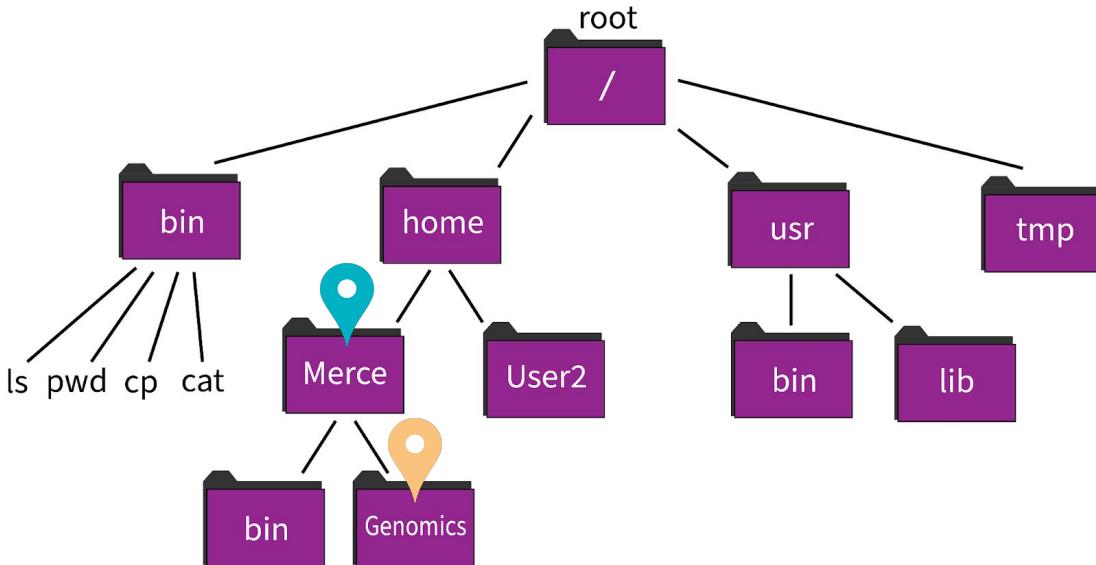
/home/Merce/Genomics

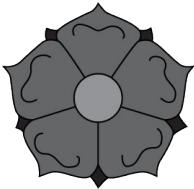


cd - change directory



```
> cd /home/Merce  
> cd ../
```

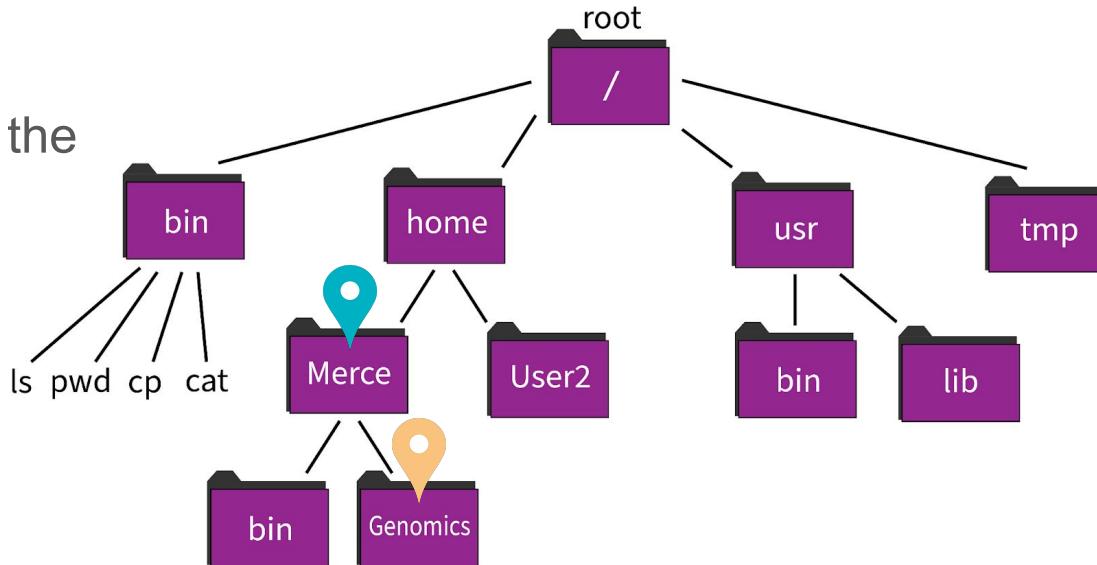




File system visualization

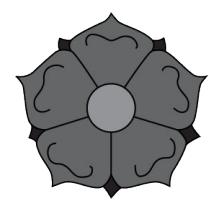
ls - shows you the contents the directory you are in

```
> ls  
> ls .  
> ls ./
```



```
> ls ../
```





Create, copy, move, and remove files and folders

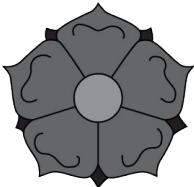
mkdir - create new directory

cp - copy file

mv - move file or directory

rm - remove file

"Unix was not designed to stop its users from doing stupid things, as that would also stop them from doing clever things." - Doug Gwyn



Symbolic links

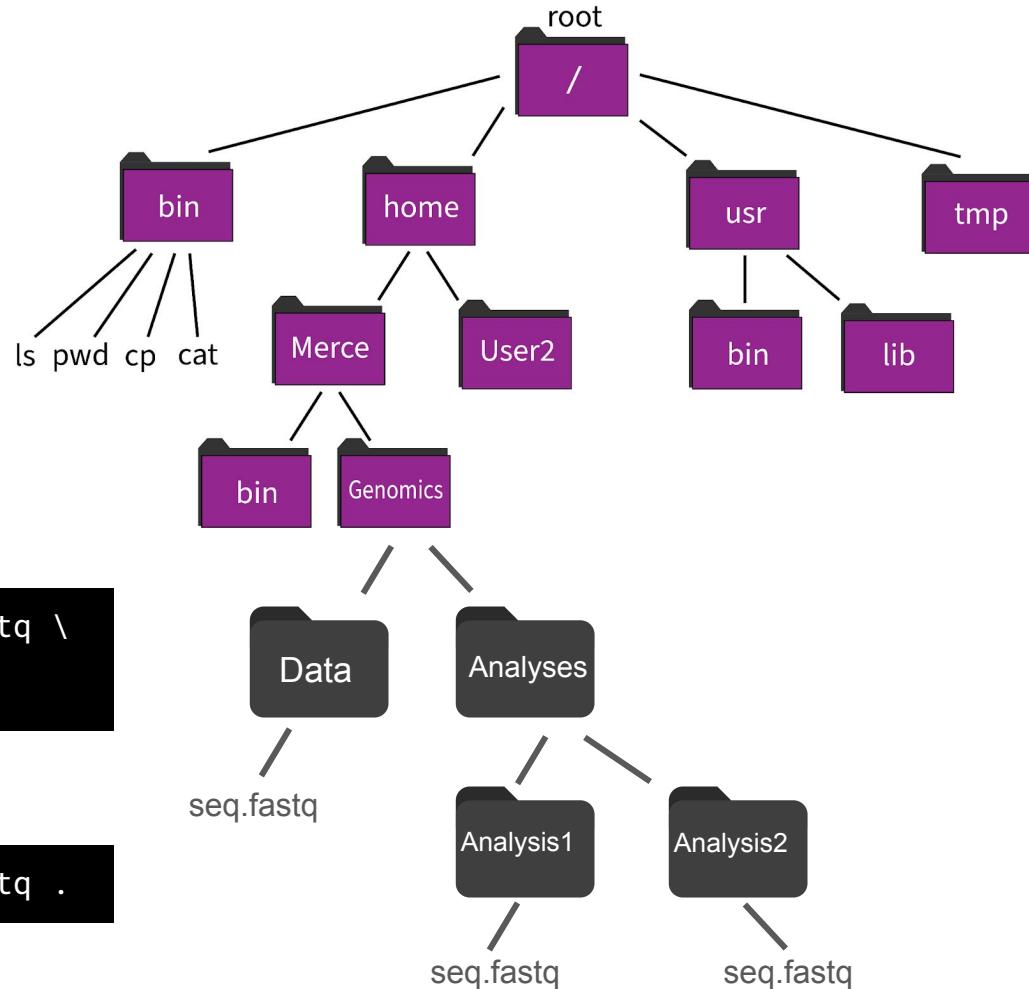
In **-s /path/to/file** link

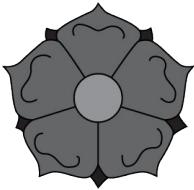
create a symlink of file

```
> ln -s /home/Merce/Genomics/Data/seq.fastq \
  /home/Merce/Genomics/Analyses/Analysis1/
```

If we are already inside the folder Analysis1:

```
> ln -s /home/Merce/Genomics/Data/seq.fastq .
```





Manual

man *command* - manual of the command

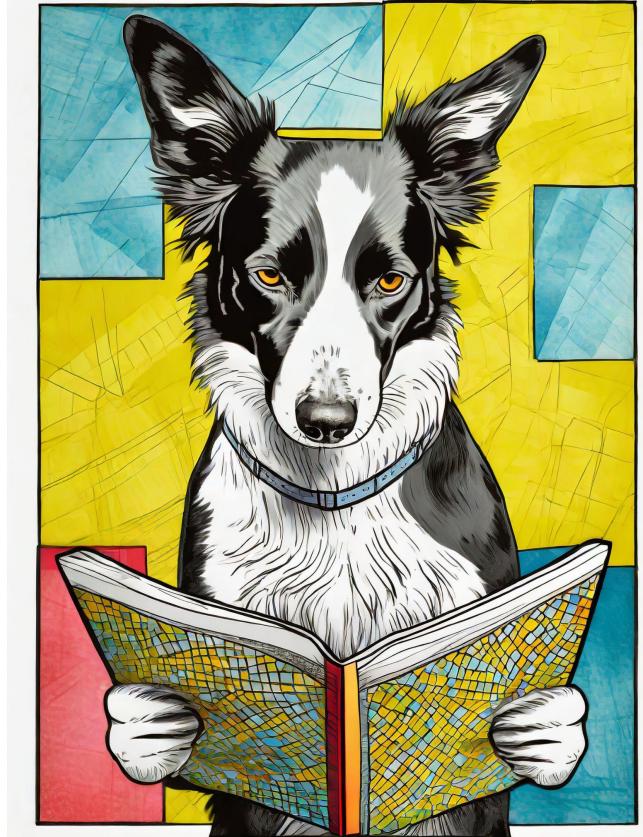
```
> man ls
```

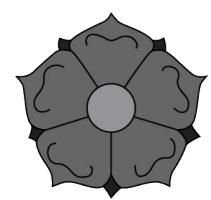


ls -l formatted list

ls -h “human” formatted list

ls -lh combination of flags





Explore file content

wc - word count (-l lines, -c characters, -w words)

less - visualize file contents in your terminal screen (press q to exit)

cat - prints contents of your file as *standard output* in your terminal

head - visualize the first 10 lines of a file

tail - visualize the last 10 lines of a file

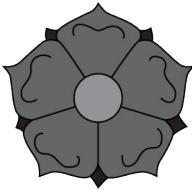
The character | (pipe) is used to concatenate commands, so that we can run one command after the other, avoiding the creation of intermediate files.

```
command1 input | command2 > output
```

Instead of :

```
command1 input > output1  
command2 output1 > output2
```

Using pipe, the output of running *command1* on a given input gets directly piped into *command2*, and we obtain an output of these two consecutive commands, generating only one output.



A bit more advanced file-handling commands

cat - prints contents of your file as *standard output* in your terminal

redirect to a command

`cat fileA | command > output.txt`



`cat | grep`



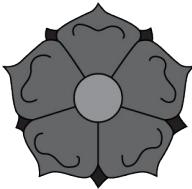
`grep`

concatenate files

`cat fileA fileB >> fileC`

`cat fileA > fileC`

`cat fileB >> fileC`



A bit more advanced file-handling commands

sort - puts in certain order a series of lines in our file

sort -r fileA sorts in reverse order

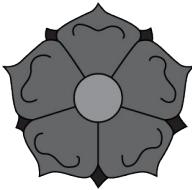
sort -n fileA sorts lines in fileA numerically

sort -k 2 fileA sort fileA by column 2

sort -k 2nr fileA sort fileA by column 2, numerically and in reverse order

sort -V fileA sort lines in fileA numerically natural.

sort -u fileA sort lines and removes duplicates -> sort fileA | uniq



A bit more advanced file-handling commands

Are these two files different?

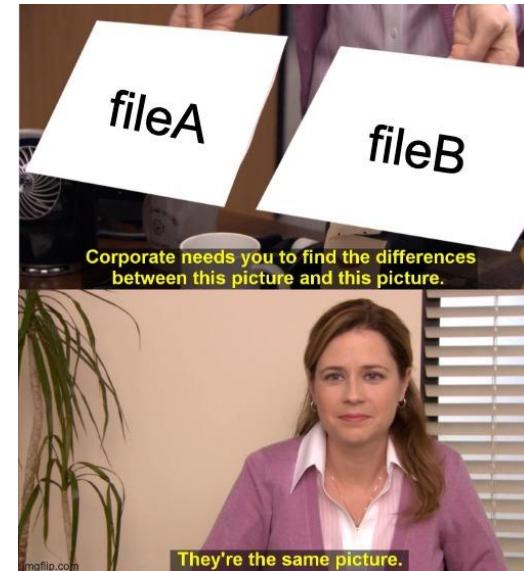
diff - can tell us if there are differences between two files

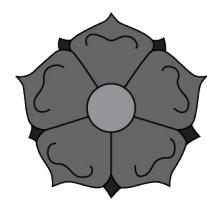
`diff -q fileA fileB`

“Files fileA and fileC differ”

`diff fileA fileB`

prints differences





A bit more advanced file-handling commands

Splitting a file

split - split a given file into multiple files (default 1000)

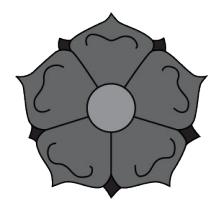
```
split -l 20 fileA
```

produce x number of files from fileA, each containing 20 lines.

cut - extract specific parts of a file

```
cut -c 2 fileA
```

extract specific columns from a file



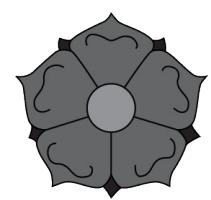
Text editors

Nano - The simpler option of text editor. All commands within the nano text editor are given by pressing the Control-key, usually represented as ^

`^S` save current file

`^O` save to (a different file)

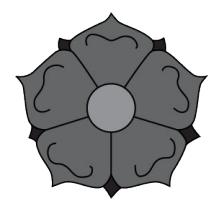
`^X` exit from nano



Text editors

Vim - a highly configurable text editor built to make creating and changing any kind of text very efficient

- i start insert mode (you can start typing after where your cursor is)
- ESC exits insert mode (also Ctrl + C)
- :w save file without exiting
- :q exit file (if there are unsaved changes, it fails)
- :wq save and exit
- :q! exit without saving changes

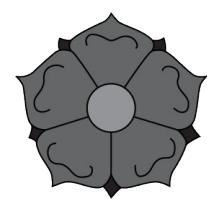


Text editors

emacs - a text editor characterized by its extensibility and configurability. Some essential commands get activated by typing Control + X, then the command (while holding the control key), but there is a wide range of key combinations to be used to move and edit the text

Ctrl + x + s save file

Ctrl + x + c exit editor (if not saved, it ask if you want to save, then type "yes")



Bash scripts

Shell scripts often have the suffix .sh

Shell scripts must be executable `chmod 755` or `chmod +x`

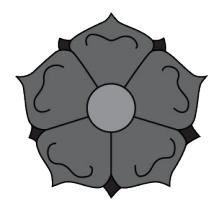
Comments can be written in scripts with a #

Variables can be used to shorten long paths

Shell loops can be used to process lots of files

\ can be used to wrap long commands across multiple lines

`#!/bin/bash` must be the first line, it specifies interpreter



Bash scripts - Variables

We can save variables under almost any name.

Variables can be string type:

```
evomics="Workhop_on_genomics_2026"  
data="genome_assembly_file.fasta"  
path="/home/genomics/workshop_materials/unix"
```

Integer type:

```
num=5
```

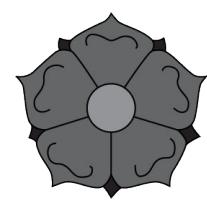
or float type:

```
pi=3.14
```

We can refer to the variables using a dollar sign:

`$evomics`

`${evomics}`



Bash scripts - *for* loops

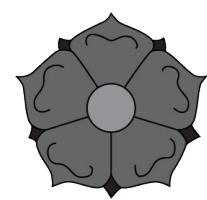
Loop over files inside a directory:

```
for file in ./unix/working_directory/*fastq
do
    commands $file
done
```

Loop over files that we stored inside a variable:

```
files="file1
file2
file3
file4"

for file in $files
do
    commands $file
done
```

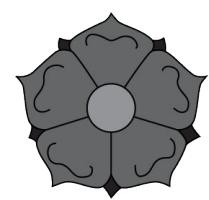


Bash scripts - *while* loops

while loops

```
while read line
do
    command $line
done
```

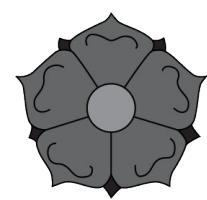
We can pipe the command `ls -l` to this script to run the command on each of the files listed.



Bash scripts - Stay informed!

If we want to print messages to the standard output while the script is running we can do that using the echo command. This is specially useful when running a long pipeline of multiple commands, so that we can keep track of the stage that is currently running.

```
for file in ./unix/working_directory/*fastq
do
    echo "Command 1 is running on $file"
    command1 $file
    echo "Command 2 is running on $file"
    command2 $file
done
```



Bash scripts - Some tricks!

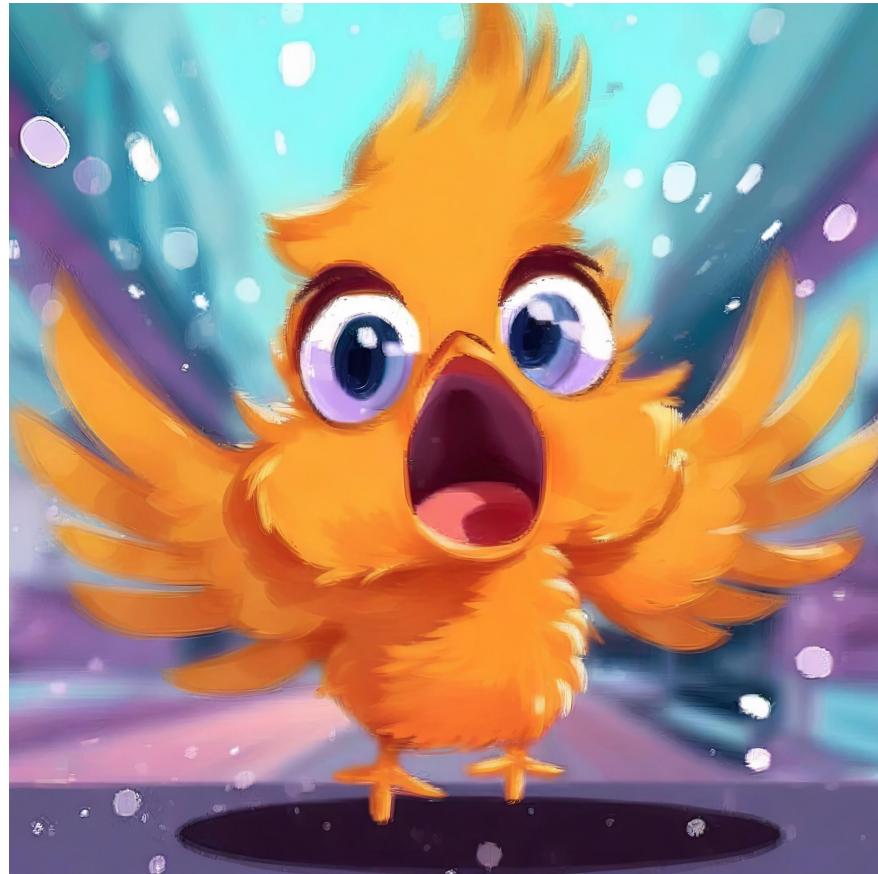
```
for file in ./unix/working_directory/*fastq
do
    file_name=$(basename $file ".fastq")

    command1 $file -out ${file_name}_command1.fastq
done
```

It is the same as:

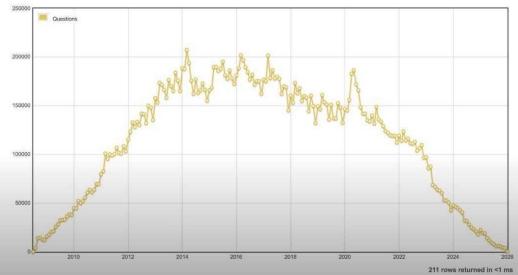
```
for file in ./unix/working_directory/*fastq
do
    file_name=$(basename $file ".fastq")

    command1 $file -out "$file_name"_command1.fastq
done
```





What my family and friends
think I do



What I actually do



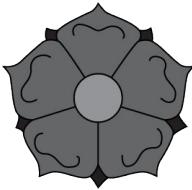
What my supervisor thinks I do



ChatGPT



stack overflow



Cheat-sheet

Navigation

pwd show current path / directory
ls list directory
cd *dir* change directory to *dir*
cd ~ go home
cd - change to previous directory
. current directory
.. parent directory

file & dir

mkdir *dir* create directory *dir*
cp *file1* *dir/* copy *file1* to directory *dir*
mv *file1* *dir/* move *file1* to directory *dir*
mv *file1* *file2* rename *file1* to *file2*

rm *file1* delete *file1*
rm -r *dir* delete directory *dir*

file exploration

wc count
(-l lines, -w words, -c characters)
tail *file* output last 10 lines of file
head *file* output first 10 lines of file
less *file* visualise contents of file
cat *file* output file to standard output
sort sort rows
uniq keep unique rows
diff *fileA* *fileB* differences?
cut -c 2 cut column 2

archive & permissions

tar -xzf archive.tar.gz uncompress tar.gz
tar -czf archive.tar.gz archive compress
archive to tar.gz
gzip compress .gz
gunzip uncompress .gz

chmod +x makes file executable
 (-r readable, -w writable)

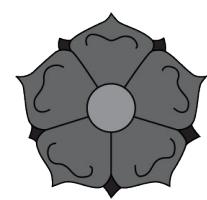
remote

wget web-address-to-file download file
curl -O web-address-to-file into current dir

ssh user@server connect to cluster

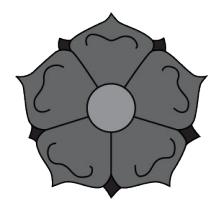
shortcuts & useful commands

Ctrl + C halts current command
Ctrl + Shift + C copy (linux)
Cmd + C (mac) copy (mac)
Ctrl + Shift + V paste (linux)
Cmd + V (mac) paste (mac)
Ctrl + W erases one word
Ctrl + U erases whole line
Ctrl + A go to beginning of line
Ctrl + E go to end of line
exit exit
history log out of current session
man command prints your past commands
manual for command

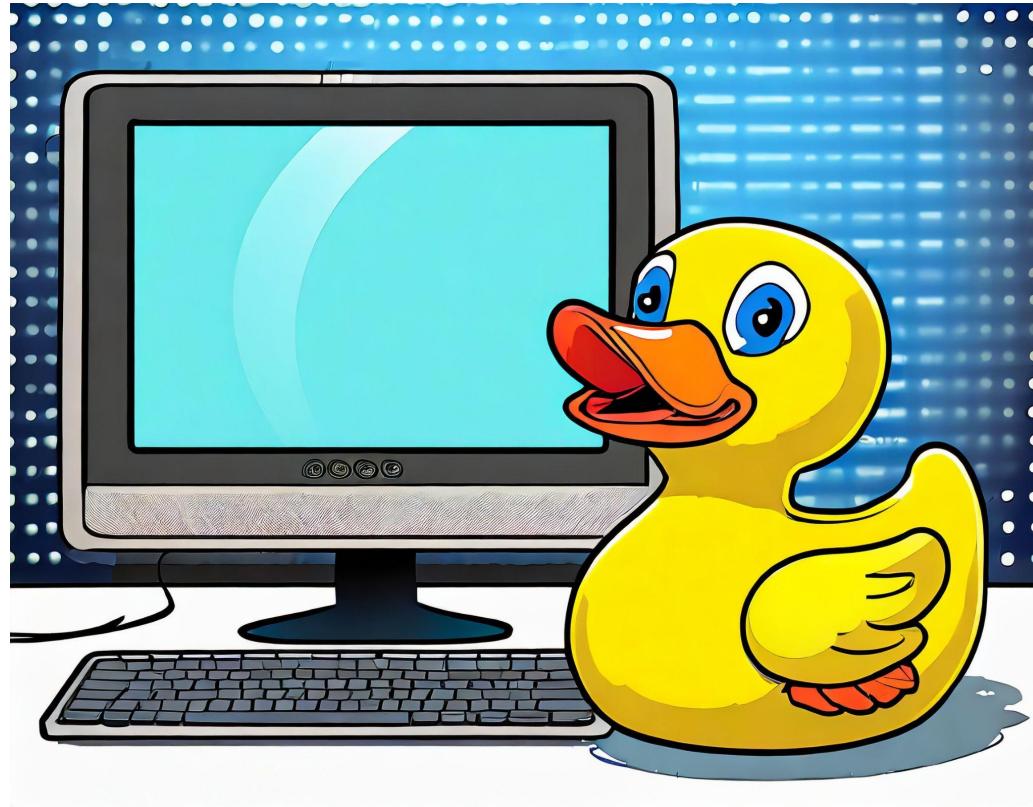


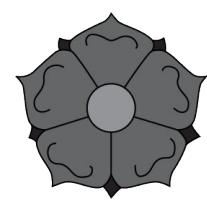
Cheat-sheet



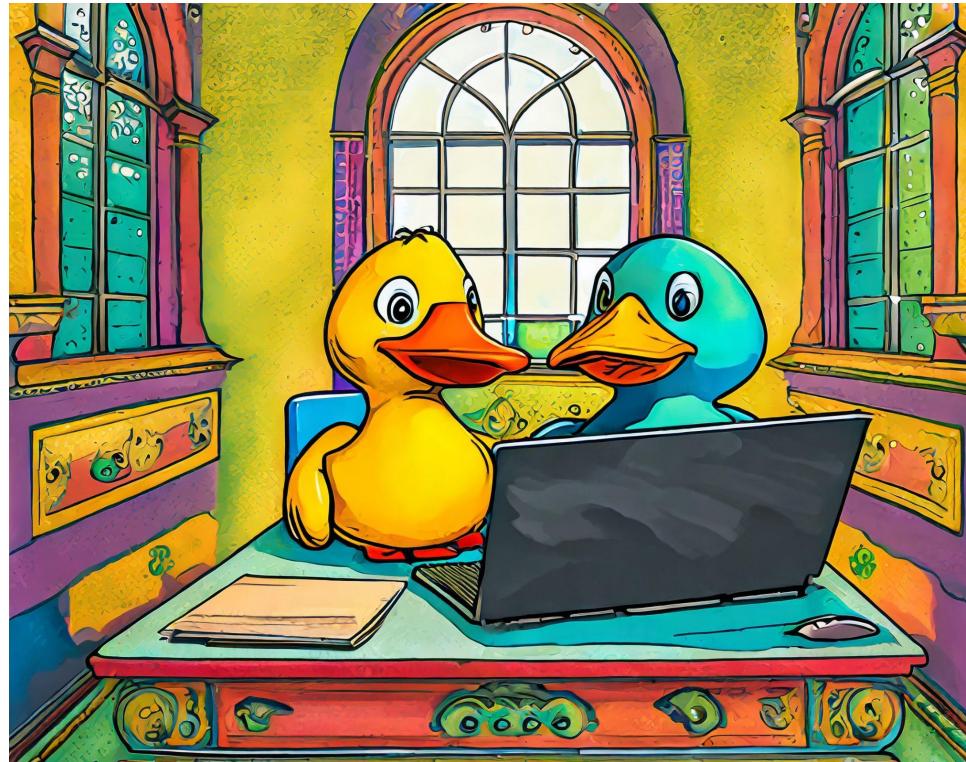


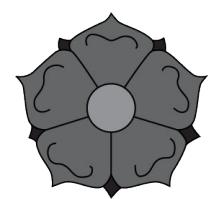
Rubber duck





Now it is your turn to practice!
And become each other's ducks!





Bioinformatic Files

Parsing and Editing

FastQ

FastQ format is a text file in which each entry is composed of a set of four lines:

Line 1 begins with @ and is followed by the sequence identifier.

Line 2 is the raw sequence nucleotide letters

Line 3 is a + character alone (in some cases, it may be followed again by the sequence identifier)

Line 4 encodes the quality values for each letter in the sequence (Line 2 and 4 must be the same length)

For example:

FastA

FastA format is a text file in which each entry is composed of a set of two lines:

Line 1 begins with > and is followed by the sequence identifier.

Line 2 is the sequence nucleotide letters.

For example:

```
>NC_054070.1 Falco naumanni isolate bFalNau1 chromosome 17, bFalNau1.pat, whole genome shotgun
sequence
aaccctaagagcctgagtctaacccctaaccctcaccctaagagcctgagtctaacccctaagagcctgagtctaacccctaaccgaa
gagcctga
```

BED

BED format is a text file in which each entry is composed of one line with genomic coordinates and any associated annotations.

A BED file contains a minimum of three columns, with the first three columns being the chromosome, start, and stop coordinates of the regions considered.

The start coordinate is zero-based, and the stop coordinate is non-inclusive.

For example:

NC_054069.1	12	4897	-8209981	+	(CTAACCC)n	Simple_repeat	1	4516
0		119312						

In this file, the first three columns refer to the genomic region, and the remaining columns provide additional annotations specific to this region.

GTF/GFF

[GTF](#) or [GFF](#) formats are text files in which each entry is composed of one line, used for describing genes and other features of DNA, RNA and protein sequences.

Both formats contain 9 columns, which are tab-separated.

Column 1 contains the seqID, which gives the name of the sequence.

Column 2 contains the source, which is the procedure that generated the feature.

Column 3 contains the type of feature, such as 'gene' or 'exon.'

Column 4 contains the start position of the feature, which is 1-based (different from BED format).

Column 5 contains the stop position of the feature.

Column 6 contains the score, which gives the confidence from the source in the annotated feature.

Column 7 contains the strand, which can be '+' '-' or '?' for an unknown orientation.

Column 8 contains the phase, which will be 0, 1 or 2 for CDS features or '.' for anything else.

Column 9 contains attributes, which are semicolon separated and provide additional information about the feature. The format and information contained in this column differs between GFF3 and GTF formats.

For example, a GFF file may look like this:

```
NC_054069.1      RefSeq  region  1      8214878 .      +      .
ID=NC_054069.1:1..8214878;Dbxref=taxon:148594;Name=16;chromosome=16;collected-by=Diego Rubolini;collection-date=2016-06-25;country=Italy: Matera;dev-stage=juvenile;gbkey=Src;genome=chromosome;isolate=bFalNau1;lat-lon=40.67 N 16.60 E;mol_type=genomic DNA;sex=female;specimen-voucher=P51P2 - H187058 (nest 2016-P51);tissue-type=blood
```

VCF

VCF format is a text file in which each entry is composed of one line and is used for storing sequence variation.

A VCF file generally starts with a header that provides metadata describing the content of the file, which is denoted by starting with `#` and special keywords denoted by `##`.

After the header follows the body of the VCF file, which contains a mandatory 8 columns, and unlimited optional columns.

Column 1 contains the name of the sequence where the variant is located, usually the chromosome.

Column 2 contains the position of the variant, which is 1-based.

Column 3 contains the identifier of the variant, which will be `"` if unknown.

Column 4 contains the reference sequence at the variant position.

Column 5 contains a list of alternative alleles at the variant position.

Column 6 contains the quality score associated with the inference of the given alleles.

Column 7 contains a flag giving information on filters that the variant has failed to pass, or `PASS` if the variant has passed filters.

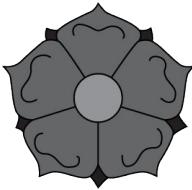
Column 8 contains a list of fields with information describing the variant. The fields may vary depending upon the method of variant detection, with fields separated by semicolons.

Column 9 is optional, but included if there are sample columns included in the VCF. This column provides a list of fields describing the information contained in the samples.

After column 9 is an unlimited number of columns describing the samples described in the file.

For example, a VCF file including one sample:

NC_054069.1	86611	521906:71:-	A	G	.	PASS	NS=93;AF=0.038;SF=0,1	GT:DP:AD:GQ:GL
0/0:49:49,0:40:-0,-15.74,-125.17								



Regular Expressions

Regular expressions (or RegEx) are like special codes used to search for patterns in text.

Special characters in RegEx

We can match any character using regular expressions, except those that have a special meaning in RegEx.

The below listed characters are special characters in RegEx:

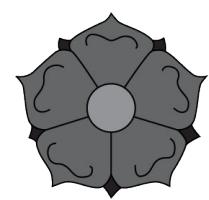
.	+	*	?	{}	^
\$	()	[]		\	

HOW TO REGEX

STEP 1: OPEN YOUR FAVORITE EDITOR
@ GARABATOKID

STEP 2: LET YOUR CAT PLAY ON YOUR KEYBOARD

/^([A-Z0-9_\.-])+/

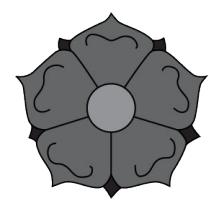


Regular Expressions

The dot (.) represents a single character, any single character. It can be a digit, a letter, a symbol, and even a space.

If we want to match our sequence "ATG", but this time we want to include the next nucleotide, we can do that with "ATG .", in this case, the character after 'ATG' was a 'T', so we find 'ATGT':

ATAGCATCAAATGTAGCATTACGTAGTAGCTATAGCTATTACGTAGGGCTACTTATAGCATCAAATCTAGCATCTACG
TAGCATCAAATCTAGCACGTACGTAGTAGCTCATGCTATTACGTAGCGAACAGCTAACCTCAGGCTACTTATAGCAT
CAAATCTAGCATTA



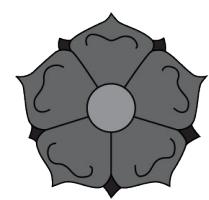
Regular Expressions

The star (*) means "zero or more occurrences of the previous character."

We can combine it with the dot in the following way ".*". So we match zero or more occurrences of any character.

If you wanted to find all sequences that start with "ATG" and end with "TAA", you could use a regular expression like "ATG .*TAA", which means "find 'ATG', followed by zero or more of any character, followed by 'TAA':

ATAGCATCAAATGTAGCATTACGTAGTAGCTATAGCTATTACGTAGGGCTACTTATAGCATCAAATCTAGCATCTACGTAGCATC
AAATCTAGCACGTACGTAGTAGCTCATGCTATTACGTAGCGCAACAGCTAACCTCAGGCTACTTATATCTAGCATTAACGTAGTA



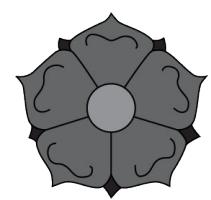
Regular Expressions

The plus (+) means "one or more occurrences of the previous character".

We can combine it with the dot in the following way ".+". So we match one or more occurrences of any character.

If you wanted to find all sequences that start with "ATG" and end with "TAA", you could use a regular expression like "ATG .+TAA", which means "find 'ATG', followed by one or more of any character, followed by 'TAA':

ATAGCATCAAATGTAGCATTACGTAGTAGCTATAGCTATTACGTAGGGCTACTTATAGCATCAAATCTAGCATCTACGTAGCATC
AAATCTAGCACGTACGTAGTAGCTCATGCTATTACGTAGCGCAACAGCTAACCTCAGGCTACTTATATCTAGCATTAACGTAGTA



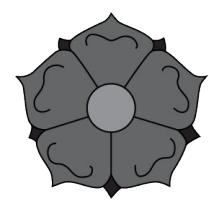
Regular Expressions

The plus (+) means "one or more occurrences of the previous character".

But, what would happen if we have this sequence? Can we use ATG.+TAA to match it?

ATAGCATCAAATGTAA CATTACGTAGTAGCTATAGCTATTACGTAGGGCTACTTATAGCATCAAATCTAGCATCTACGTAGCATC
AAATCTAGCACGTACGTAGCTCATGCTATTACGTAGCGAACAGCTAACCTCAGGCTACTTATAGCATCAAATCTAGCATT

"ATG.*TAA" would be able to match it, but not "ATG.+TAA", as it requires that there is at least one character in between 'ATG' and 'TAA'



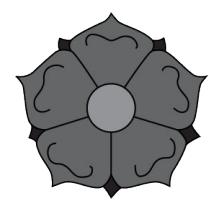
Regular Expressions

The question mark (?) matches zero or one time the previous character

If we want to match our sequence starting with "ATG" and ending with TAA, and we know sometimes there is a T after ATG, but sometimes not, we can do that with "ATGT?TAA", in this case, the character after 'ATG' can be a T, or can be nothing, and both the following sequences would be matched:

ATAGCATCAAATGTAA CATTACGTAGTAGCTATAGCTATTACGTAGGGCTACTTATAGCATCAAATCTAGCATCTACGTAGCATC
AAATCTAGCACGTACGTAGTAGCTCATGCTATTACGTAGCGAACAGCTAACCTCAGGCTACTTATAGCATCAAATCTAGCATT

ATAGCATCAAATGTTAA CATTACGTAGTAGCTATAGCTATTACGTAGGGCTACTTATAGCATCAAATCTAGCATCTACGTAGCAT
CAAATCTAGCACGTACGTAGTAGCTCATGCTATTACGTAGCGAACAGCTAACCTCAGGCTACTTATAGCATCAAATCTAGCATT



Regular Expressions

The curly brackets ({}) can reference the amount of times we expect the previous character to occur. It has three main configurations:

{m} - previous character exactly m number of times

{m,n} - previous character m to n number of times

{m,} - previous character m or more number of times

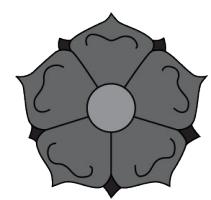
If we want to find 3 'A' in a row, we could use "A{3 }"

ATAGCATCATAATGTAGCATTACGTAGTAGCTATAGCTATTACGTAGGGCTAAAAATAGCATCATATCTAG

We can also specify that we want to find 'A' a minimum of 3 times and a maximum of 5 with "A{3 , 5 }", then we would match:

ATAGCATCATAATGTAGCATTACGTAGTAGCTATAGCTATTACGTAGGGCTAAAAATAGCATCATATCTAG

(We could get the same result in this specific sequence, by using "A{3 , }")



Regular Expressions

The caret (^) is used to match the beginning of the line.

So that if we want to match sequences that start with "ATG", we can use ^ATG.

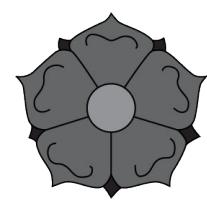
For example, in the following sequences:

ATGATAGCTAACATTACGTAGTAGCTATAGCTATT

GTCATGAGCTATTAGCATCACATCTAGCACGTTCATG

ATGCTATGAAGTCTACTTTATAGCATCAAATCTAGTA

The regular expression ^ATG matches ATG only in the first and third lines because they begin with "ATG"



Regular Expressions

The dollar sign (\$) is used to match the end of a line or string.

If we want to match sequences that end with "TAA", we can use "TAA\$".

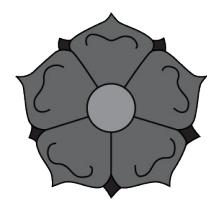
For example, in the following sequences:

TATAGCTAAAGTCTACTTATAATCAATGATAGCTTAA

ATGAGCTATTAGCATCACATCTAGCAGTCATGAGCTAT

GTAGCATTACGTAGCTAGCTATGCTAAGAAGT

The regular expression TAA\$ matches TAA only in the first line because it ends with "TAA"



Regular Expressions

Square brackets ([]) are used to define a set of characters to match.

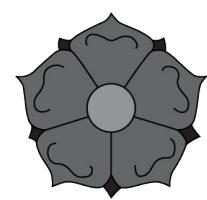
For example, [ACGT] matches any single character that is A, C, G, or T.

If we want to find sequences where "ATG" is followed by a C or G, we can use ATG[CG].

In the sequence:

ATAGCATCAAATGCTAACATTTACGTAGTAGCTATAGCTATTACGTATGGCTACTTTATAGCATCAAATCT

The regular expression ATG[CG] matches both ATGC and ATGG.



Regular Expressions

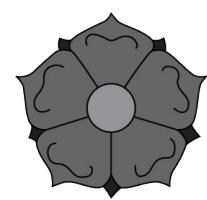
The pipe (|) means "or" and allows you to specify alternative patterns.

For instance, ATG|TAA matches either "ATG" or "TAA"

In the sequences:

ATGACGACGTAGCGAACAGCTAACCTCAGGCTACTTATAGCATCAAATCTAGCATTTAAATAG
TCTAGATGACGACGTAGCGAACAGCTAACCTCAATAGCTATTACGTAGTGCAATGTACTATTA
ACCTCAGGCTACTTATATAGCTATTACGTAGAGCATCAAATCTAGCATTTAAATAGCCGTATCC

The regular expression ATG|TAA matches ATG or TAA when found.



Regular Expressions

Parentheses () are used for grouping and capturing.

If we want to capture sequences that follow the pattern "ATG" followed by any two characters and "TAA," we can use "(ATG .. TAA)" to match and capture the sequence "ATGCTTAA":

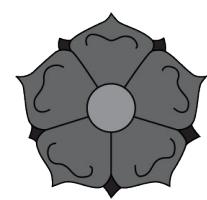
ATGCTTAAATGCCAGTAA

Captured groups can be referenced later in the same regular expression or used in programming languages. To find repeated sequences like "ATGTACTAA", you can use:

(ATGTACTAA).*\1

It would match in the following sequence:

GTAAATGTACTAACAGTAACGTAGCGATGTACTAAACCTCAATAG



Regular Expressions

The backslash (\) is used as an escape character to treat special characters literally.

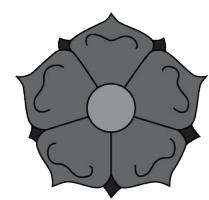
For example, if you want to match a literal dot, use "\ ." instead of "."

If we want to find "A . T" as it appears (with the dot), we can use "A\ . T"

A.TGACTTAAG.A.T

The regular expression "A\ . T" matches A.T twice in the sequence.

The backslash is also used to represent special characters that are not written literally. For example, \n represents a new line, and \t represents a tab, \s matches any whitespace, \d any digit, and \w any word.



grep

The command grep will print the lines matching a given pattern.

`grep PATTERN file`

`grep -e PATTERN file` (Pattern uses regex)



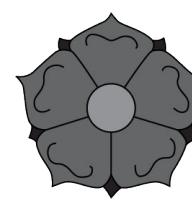
Understanding grep with a simple fasta file:

`>contig1`

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT

`>contig2`

ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG



grep

Understanding grep with a simple fasta file:

>contig1

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT

>contig2

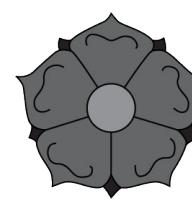
ATTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG

Find a specific sequence within our sequences:

```
grep "AGGGG" file.fasta
```

-> will print only the first sequence:

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT



grep

Find sequence headers only:

```
grep ">" file.fasta
```

-> will print all fasta headers:

```
>contig1
```

```
>contig2
```

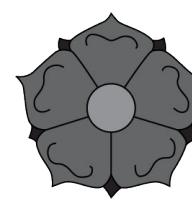
Understanding grep with a simple fasta file:

```
>contig1
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTTAT
```

```
>contig2
```

```
ATTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```



grep

Understanding grep with a simple fasta file:

>contig1

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTTAT

>contig2

ATTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG

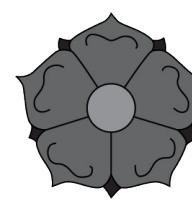
Count number of sequences:

```
grep ">" file.fasta | wc -l
```

-> will count how many lines contain ">", which will match with the number of sequences: 2

we can also use the flag -c in grep to do the same:

```
grep -c ">" file.fasta
```



grep

Understanding grep with a simple fasta file:

>contig1

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTTAT

>contig2

ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG

Print the DNA sequences with no headers:

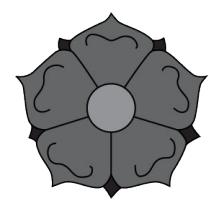
```
grep -v ">" file.fasta
```

-> will print all lines that do not contain ">":

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTTAT

ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG

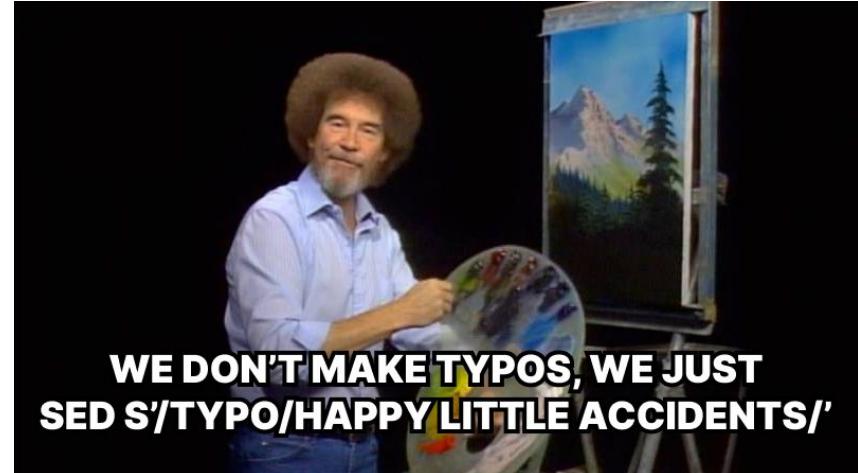


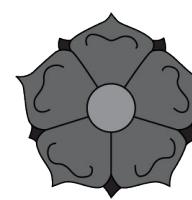


sed

sed ("stream editor") is a tool that can parse a file line by line, and transform text, using a compact programming language that can fit in one line. Sed is a powerful tool with a big array of possible commands, but the most common one is the substitution, in which we find a pattern and substitute it for another string.

```
sed 's/patternA/patternB/' file.txt
```





sed

Understanding sed with a simple fasta file:

>contig1 assembled a

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGCTACTTAT

>contig2 assembled b

ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG

Modify the fasta header to contain "sequence" instead of "contig":

```
sed 's/contig/sequence/' file.fasta
```

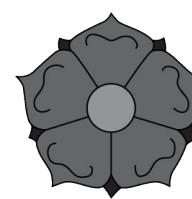
We will obtain the entire file with the replacement:

>sequence1 assembled 2025 a

AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGCTACTTAT

>sequence2 assembled 2025 b

ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG



sed

Understanding sed with a simple fasta file:

```
>contig1 assembled a
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGCTACTTAT
```

```
>contig2 assembled b
```

```
ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```

Substitute spaces for underscores, in order to avoid problems with other programs:

In this case we add the flag "g" at the end, to make sure it replaces each occurrence even if there is multiple within the same line:

```
sed 's/ /_/g' file.fasta
```

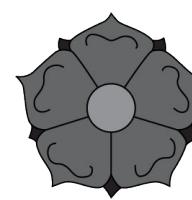
We will obtain the entire file with the replacements:

```
>sequence1_assembled_2025_a
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGCTACTTAT
```

```
>sequence2_assembled_2025_b
```

```
ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```



sed

Understanding sed with a simple fasta file:

```
>contig1 assembled a
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT
```

```
>contig2 assembled b
```

```
ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```

Simplify a fasta header:

```
sed 's/ .*/' file.fasta
```

We use regular expressions to match the first space we find in a line, followed by any character (.), any number of times (*)

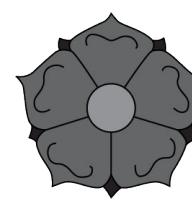
We will obtain the entire file with the replacement:

```
>contig1
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT
```

```
>contig2
```

```
ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```



sed

Understanding sed with a simple fasta file:

```
>contig1 assembled a
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT
```

```
>contig2 assembled b
```

```
ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```

Simplify a fasta header:

```
sed 's/ .*/' file.fasta
```

We use regular expressions to match the first space we find in a line, followed by any character (.), any number of times (*)

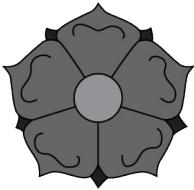
We will obtain the entire file with the replacement:

```
>contig1
```

```
AATCTAGCATTACGTAGTAGCTAAAGCTAACCTCAGGGGCTACTTAT
```

```
>contig2
```

```
ATTTACGTAGCATCAAATCTAGCATTACGTAGTAGCTAAAGCTATTACG
```



AWK

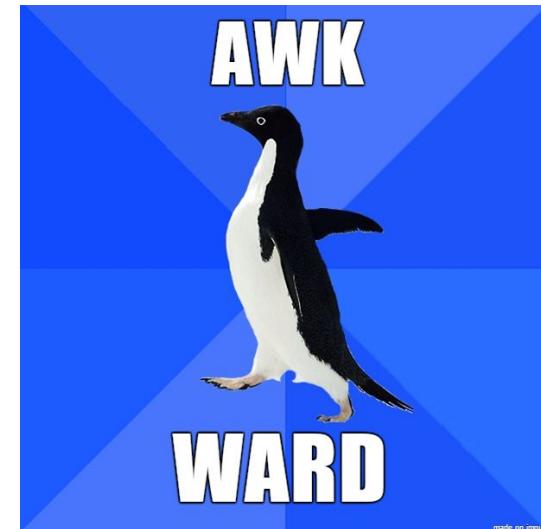
AWK is a language designed for text processing, like sed and grep. AWK is a standard feature of most Unix-like operating systems. AWK reads one line at a time, searching for a specific pattern to execute the desired action. It requires a condition, and an action:

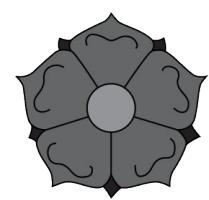
```
awk condition {action} file.txt
```

AWK is a language fieldaware (column aware):

\$0 refers to the whole line

$\$1, \$2, \$3 \dots$ refers to columns 1, 2, 3 ...





AWK

Print only the lines containing genes in contig1:

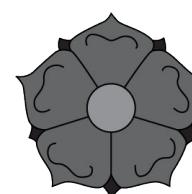
```
awk '$1="contig1" {print}' file.bed
```

We would get the following printed out:

```
contig1 20 1305 gene1 . +
contig1 4674 8563 gene4 . -
```

Understanding awk with a simple BED file:

contig1	20	1305	gene1	.	+
contig1	4674	8563	gene4	.	-
contig2	1239	5387	gene6	.	-
contig3	546	3524	gene9	.	+



AWK

Understanding awk with a simple BED file:

contig1	20	1305	gene1	.	+
contig1	4674	8563	gene4	.	-
contig2	1239	5387	gene6	.	-
contig3	546	3524	gene9	.	+

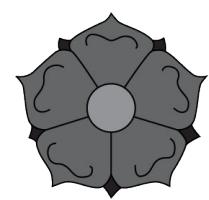
Count how many genes we have in our file:

```
awk '$1="contig"' file.bed | wc -l
```

or we can create a counter after each condition is met, in this case finding the word contig in column 1. And we use the function END to mark that an extra action is done when all lines are finished being parsed:

```
awk '$1="contig" {count++} END {print count}' file.bed
```

both these commands will print: 4



AWK

Understanding awk with a simple BED file:

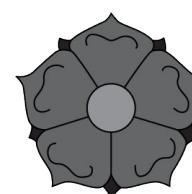
```
contig1 20 1305 gene1 . +
contig1 4674 8563 gene4 . -
contig2 1239 5387 gene6 . -
contig3 546 3524 gene9 . +
```

We can also use the function BEGIN to add an action before we start parsing the lines in our file:

```
awk 'BEGIN {print "We have these many genes:"} $1="contig" {count++} END {print count}' file.bed
```

This command will print:

We have these many genes:



AWK

Understanding awk with a simple BED file:

contig1	20	1305	gene1	.	+
contig1	4674	8563	gene4	.	-
contig2	1239	5387	gene6	.	-
contig3	546	3524	gene9	.	+

Finally, we can combine information in multiple columns to create our conditions.

Print out the gene names of all genes that are larger than 2000 bp:

We need can use the information in column 2 and 3, which marks the start and end of each gene, and we will print the information in column 4 (gene name) if column 3 - column 2 is larger than 2000:

```
awk '$3 - $2 > 2000 {print $4}' file.bed
```

It will print:

gene4

gene6



Exercises

Time to put into practice everything you learned!

Inside the folder `/home/genomics/workshop_materials/unix_tutorial`, you will find a folder called `data`. It contains a dataset that includes:

- `genome_assembly.fasta`
- `sequencing_reads.fastq`
- `variant_analysis.vcf`
- `annotation.gff`

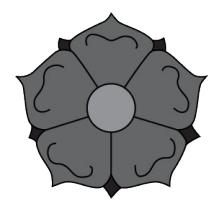
Extra slides

Inputs and outputs

`stdin` It stands for standard input, and is used for taking text as an input.

`stdout` It stands for standard output, and is used to text output of any command you type in the terminal, and then that output is stored in the `stdout` stream.

`stderr` It stands for standard error. It is invoked whenever a command faces an error, then that error message gets stored in this data stream.



stdin, stdout, stderr

command *stdin*

if it works: prints in our terminal the *stdout*
if it fails: prints in our terminal the *stderr*

command *stdin > stdout*

if it works: *stdout* is redirected to a file
if it fails: prints in our terminal the *stderr*

command2 *stdin2 > stdout*

stdout is redirected to a file and rewrites its contents

command2 *stdin2 >> stdout*

stdout is redirected to a file and appended after its contents



stdin, stdout, stderr

command *stdin*

if it works: prints in our terminal the *stdout*
if it fails: prints in our terminal the *stderr*

command file1 > output.txt
stdin *stdout*

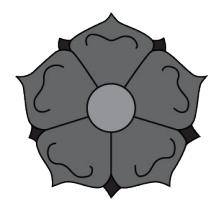
if it works: *stdout* is redirected to a file
if it fails: prints in our terminal the *stderr*

command2 file2 > output.txt
stdin *stdout*

stdout is redirected to a file and rewrites its contents

command2 file2 >> output.txt
stdin *stdout*

stdout is redirected to a file and appended after its contents



stdin, stdout, stderr

command file1 2> errors.txt
stdin stderr

if it works: prints in our terminal the *stdout*
if it fails: *stderr* is redirected to a file

command file1 &> output.txt
stdin stdout&stderr

redirects both *stdout* and *stderr* to a file

command file1 > output.txt 2> errors.txt
stdin stdout stderr

redirects both *stdout* and *stderr* to a separate file each.

Key shortcuts

Ctrl + C halts current command

Ctrl + Shift + C copy (linux) - Cmd + C (mac)

Ctrl + Shift + V paste (linux) - Cmd + V (mac)

Ctrl + W erases one word in current line

Ctrl + U erases whole line

Ctrl + A go to begining of line

Ctrl + E go to end of line

Type exit log out of current session