UiO * Department of Biosciences

University of Oslo

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

2



UiO • Department of Biosciences
University of Oslo

Introduction to Unix and Linux

Anders K. Krabberød a.k.krabberod@ibv.uio.no



UiO Department of Biosciences University of Oslo

Why learn about UNIX and Linux?

- Bioinformatics tools and algorithms are very often developed for <u>UNIX/Linux</u> computers.
- Easy to set up **pipelines** and automate workflows.
- Powerful for handling text files, large files, and large number of files – very common in modern biology.
- Able to work on High Performance Computing platforms (Supercomputers) e.g. Saga

4

UiO Department of Biosciences University of Oslo

HPC (High Performance Computing)

- In Norway Sigma2 runs several HPC (High Performance Computing) clusters as well as other e-infrastructure facilities. https://www.sigma2.no/services-overview
- · Current clusters
 - Saga
 - Fram
 - Vilje
 - Stallo
 - Betzy

UiO • Department of Biosciences
University of Oslo

Saga

- Online as of fall 2019
- A computing cluster with 200 standard compute nodes, with 40 cores and 192 GiB memory each.
- · Total 9824 cores
- Access to resources is managed with a queue system: SLURM

· Runs linux: CentOS

Details	Saga
System	Hewlett Packard Enterprise - Apollo 2000/6500 Gen10
Number of Cores	9824
Number of nodes	244
Number of GPUs	32
CPU type	Intel Xeon-Gold 6138 2.0 GHz (normal) Intel Xeon-Gold 6130 2.1 GHz (bigmem) Intel Xeon-Gold 6126 2.6 GHz (accel)
GPU type	NVIDIA P100, 16 GiB RAM (accel)
Total max floating point performance, double	645 Teraflop/s (CPUs) + 150 Teraflop/s (GPUs)
Total memory	75 TiB
Total NVMe+SSD local disc	89 TiB + 60 TiB
Total parallel filesystem capacity	1 PB

6

UiO Department of Biosciences University of Oslo

The UNIX philosophy

- The success of the UNIX framework is because of the philosophy behind the development
 - Write programs that do **one** thing and **do it well**.
 - Write programs that will work together.
 - Write programs to handle **text** streams, because that is a universal interface.

UiO • Department of Biosciences

University of Oslo

- · Advantages with Linux
 - · Stability
 - Security
 - · Good control of users and processes
 - · Can handle large computing load
 - Open Source (you can see and modify the code)
 - · Several free options (Ubuntu, CentOS, Debian, etc.)
 - · Large community of developers
 - Large amount of free software (specially for bioinfo)
- Disadvantages
 - · Some hardware may not be supported
 - · Can be less user friendly than other options
 - (For CLI) It's not point-and-click

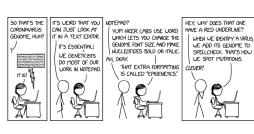
8

UiO * Department of Biosciences

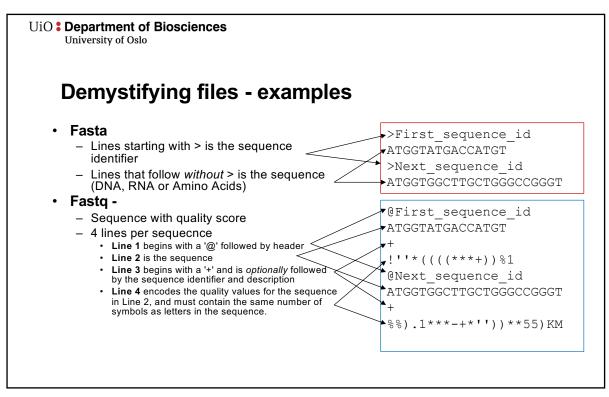
University of Oslo

Demystifying files

- Many files used in bioinformatics are just text files.
- Extension given to the file reflects the format, or how the "internal layout" of that file is.
- Often the format is formulated as set of rules that tells the user what the file should contain and how it should be organised



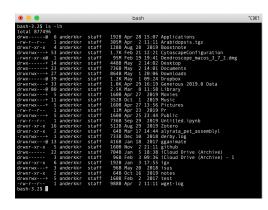
>Seq1 ATGGTATGACCATGT >Seq2 ATGGTGGCTTGCTG



10

The Shell / Terminal

UiO Department of Biosciences
University of Oslo



- The user writes commands on the command line usually in a terminal window
- · Line-by-line
- Step-by-step
- Logical work flow

Linux is case-sensitive!

UiO * Department of Biosciences

University of Oslo

The Shell

- · Getting a shell
- Windows: Windows Terminal
 - https://github.com/microsoft/terminal
- Mac and Linux: A shell is already installed (launch the Terminal application).

12

UiO: Department of Biosciences

University of Oslo

- If you've never used the terminal, you can start with simple commands and a good amount of patience ©
- There's plenty of resources and documentation online
 - Goolge is you friend!
- A set of instructions and exercises made for the course 3010 will help you:
- https://github.com/krabberod/bios3010-unix-linux

UiO • Department of Biosciences University of Oslo

Login into Saga

- Open terminal
- Type ssh <username>@saga.sigma2.no
- Type password

14

UiO * Department of Biosciences

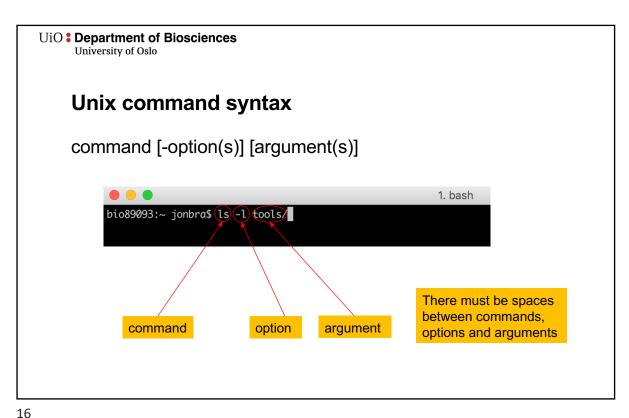
University of Oslo

General comments on commands

- The commands operates on text. Either files containing text or input from the keyboard (called standard input).
- · Commands in Linux usually takes the form

```
command [-options] inputfile
```

- The philosophy of Linux (and UNIX) programs is that they are supposed to be <u>simple</u> and <u>efficient</u> with a standardised way of taking inputs and giving outputs. Because of the standardisation, the output of one command can (often) be used as the input of another command. The advantage of this will be clear later.
- The full manual for a command can be accessed with man command



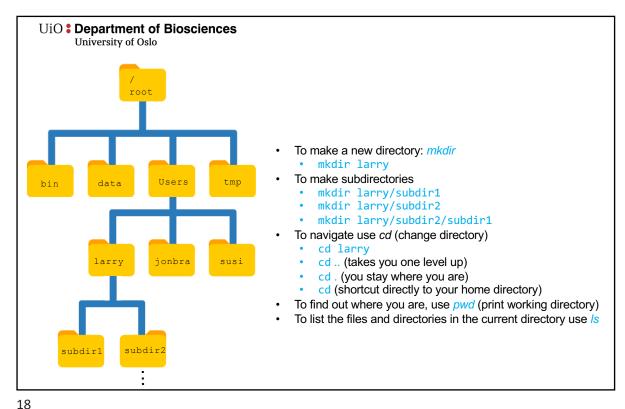
ΤO

UiO Department of Biosciences
University of Oslo

Navigating the filesystem

- cd changes directory
- 1s lists the files and directories in the current directory
- pwd prints the path of the current directory
- ~/ or \$HOME is your home directory (cd will take you there)
- is your current directory
- .. is the directory directly above the current

http://swcarpentry.github.io/shell-novice



10

UiO * Department of Biosciences

University of Oslo

Some of the most useful commands which will be covered in the exercises

- ssh Secure log on
- 1s List content
- pwd Print the working directory
- · cd change directory
- mkdir make directory
- cp Copy files and directories
- scp Secure copying of files and directories
- rsync Synchronise and transfer files and directories
- rm -Remove files and directories
- less Print content of file to screen
- more Print content of file to screen
- head Print the head of file(s)
- tail Print the end (tail) of file(s)

- cat Print and concatenate several files
- grep Search through text files
- sort Sort the content of files
- uniq Print unique (lines, or words)

Special operators:

- > (IO redirection)
- >> (IO append)
- (the vertical line is called pipe)
- Useful cheat sheet: https://cheatography.com/davechild/cheat-sheets/linux-command-line/

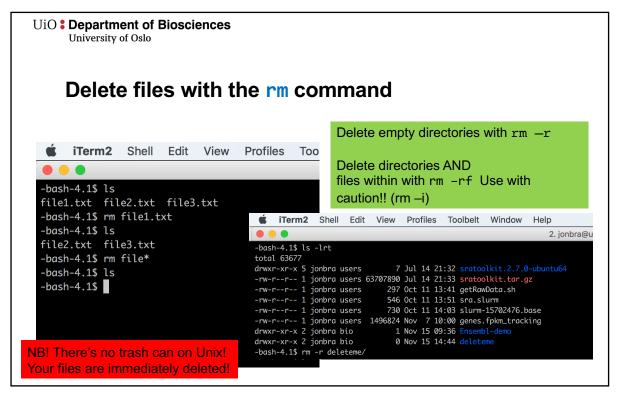
UiO Department of Biosciences University of Oslo

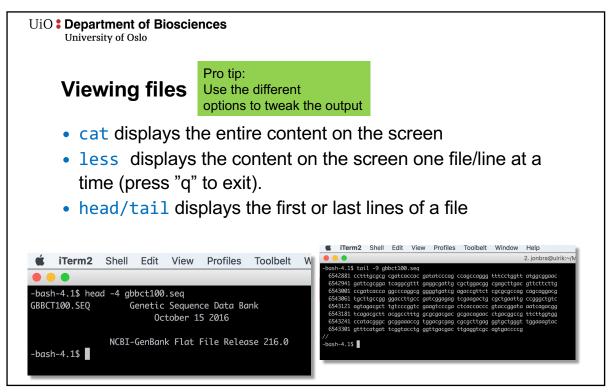
Remember

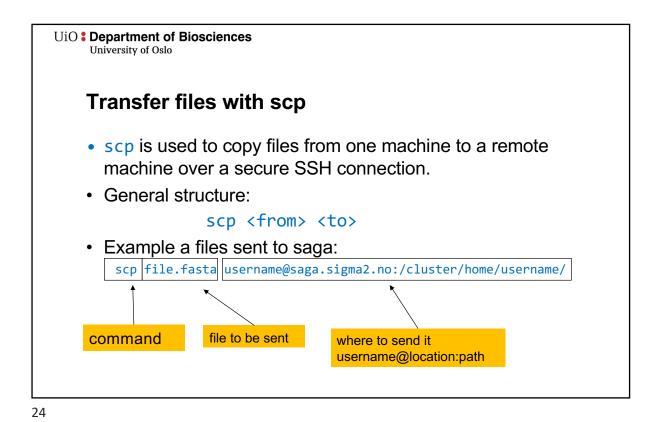
- Linux commands are case sensitive!
 - ssh
 - · is NOT the same as
 - Ssh
- Filenames and paths are also case sensitive!
 - huge file.fasta
 - is **NOT** the same as:
 - Huge_file.fasta
- Protip: Use the tab-key to autocomplete

20

```
UiO * Department of Biosciences
              University of Oslo
            Moving files with the mv command
                             iTerm2 Shell Edit View Profiles Toolbelt Window Help
                                                                              2. jonbra@ulrik:~/MBV-INFX410 (ssh)
                           -bash-4.1$ ls
total 204403
                         -rw-r--r-- 1 jonbra users 54707890 Jul 14 21:33 sratoolkit.2.7.0-uk
-rw-r--r-- 1 jonbra users 297 Oct 11 13:41 getRawData.sh
-rw-r--r-- 1 jonbra users 546 Oct 11 13:51 sra.slurm
-rw-r--r-- 1 jonbra users 730 Oct 11 14:03 slurm-15702476-base
-rw-r--r-- 1 jonbra bio 72051869 Oct 14 23:39 gbbct100_seq.gz.1
-rw-r--r-- 1 jonbra bio 72051869 Oct 14 23:39 gbbct100_seq.gz.1
-rw-r--r-- 1 jonbra users 1496824 Nov 7 10:00_genes.fpkm_tracking
drwxr-xr-x 2 jonbra bio 1 Nov 15 09:36 Ensembl-demo
drwxr-xr-x 2 jonbra bio 0 Nov_15 15:15 mytest
-bash-4.1$ my file1.txt mytest/
                                                                                 7 Jul 14 21:32 sratoolkit.2.7.0-ubuntu64
                                                                                                                                              Move file1.txt to directory mvtest
                            -bash-4.1$ cd mvtest/
                                                                                                                                      Pro tip:
                            -bash-4.1$ ls
                           file1.txt
-bash-4.1$ mv file1.txt file2.txt ◀
                                                                                                                                      mv command can also rename files!
                                                                                                                                      ("moves" file to a new filename)
                            -bash-4.1$ ls
                            file2.txt
                            -bash-4.1$
```







UiO • Department of Biosciences University of Oslo Downloading files from the internet Firefox File Edit View History Bookmarks Tools V The wget command make it File list - projects/clsi possible to download files from the () i A https://ftp.ncbi.nih.gov/genbank web Index of /genbank Last modified 14-Oct-2016 14:43
14-Nov-2016 01:33
05-Dec-2013 12:30
05-Dec-2013 12:30
14-Nov-2016 00:31
14-Nov-2016 00:31
14-Oct-2016 17:41
14-Oct-2016 17:41
14-Oct-2016 09:25
01-Jul-2013 11:20
01-Jul-2013 11:20
13-Nov-2016 05:51
13-Nov-2016 05:51 iTerm2 Shell Edit View Profiles Toolbelt Window Help catalog/ daily-nc/ docs/ 2. jonbra@ulrik:~/MBV-INFX410 (ssh) 2.jonbra@ulrk:-/MBV-NHX410(ssh)
-bash-4.1\$ wget https://ftp.ncbi.nih.gov/genbank/gbbct100.seq.gz
--2016-11-15 14:59:18-- https://ftp.ncbi.nih.gov/genbank/gbbct100.seq.gz
Resolving ftp.ncbi.nih.gov... 130.14.250.13, 2607:f220:41e:250::7
Connecting to ftp.ncbi.nih.govl130.14.250.131:443... connected.
HTTP request sent, owaiting response... 200 OK
Length: 72051869 (69M) [application/x-gzip]
Saving to: 'gbbct100.seq.gz' Release Number Right-click and copy the link!

| Sphot | 101 | Sept. GZ | 14-Oct - 2016 | 17:39 | 67M ==>] 72,051,869 19.8M/s in 5.8s 14-Oct-2016 17:39 14-Oct-2016 17:39 14-Oct-2016 17:39 gbbct101.seq.gz gbbct102.seq.gz 2016-11-15 14:59:25 (11.8 MB/s) - `gbbct100.seq.gz' saved [72051869/72051869]

UiO * Department of Biosciences

University of Oslo

Redirect and pipe

- · Since Linux (and UNIX) use streams of text it is easy to
- write the output of a command to a file.
 - Two versions:
 - The greater than symbol (>) is used to write a new files (will overwrite)

```
ls -lh > list_of_files.txt
```

• Two symbols (>>) will append the content to a file

```
ls -lh >> list_of_files.txt
```

26

UiO: Department of Biosciences

University of Oslo

Redirect and pipe

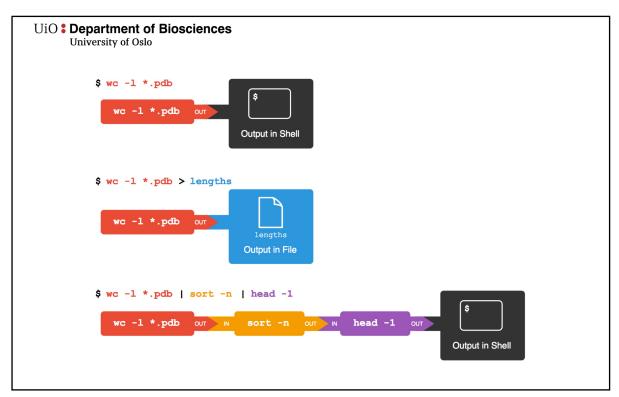
- Since Linux (and UNIX) use streams of text it is easy to
- Use the output of one command as input in another with the horizontal line called pipe |

```
wc -l *fasta | sort -n
```

First command

Pipe

Second command



28

UiO • Department of Biosciences University of Oslo Many of the commands takes extra arguments * = a wildcard! • ls -lh start_of_filena* cd (home) ; cd \overline{dir} (go to dir); cd .. (one level up, to the parent directory); cd ../../ (two levels up) mv old name new name mv /old/location/for/my/file /new/location/for/my/file • cp /path/to/file/to/copy/file /new/location/of/file cp -r /dir/to/copy/dir /location/for/dir/copy rm file $\label{local-contents} \mbox{rm -rf /remove/directory/and/all/contents (use carefully!)}$ cat file1 file2 > newcatfile paste file1 file2 > newpastefile chmod a+x executablefile chmod 755 file (owner:group:world permissions)

UiO Department of Biosciences University of Oslo A few other very useful commands: tail -n 300 file - (see last 300 lines file) • head -n 300 file - (see the first 300 lines of a file) • grep pattern file (search pattern in file) • grep -c pattern file (count pattern in file) • grep -f file_with_pattern file (look up patterns from one file in another) • ssh user@machine - (remote connection) Execute a program locally • ./executable

30

```
UiO • Department of Biosciences
     University of Oslo

    It is possible to combine commands with a "pipe" |

             • output | input
             • 1s | grep -c pattern (count patterns in filenames)
             • 1s | grep pattern > file

    Send output to file with ">"

            • grep pattern file > file2
         · Cut characters from a file
             • File1 (tab separated):
                · This is a test 1
                · This is a test 2
                • This is a test 3
             • Cut -f 1 file1
                • This
                • This
                • This
```

UiO • Department of Biosciences

University of Oslo

- Loops
 - Assuming a directory with:
 - file1, file2,..., filen
 - · We want to run a function over all files, then
 - for i in \$(ls file*); do cut -f 2 \$i > \$i.out; done
 - (cuts the field 2 of each file and sends it to an output)
- · Execute functions in bashscripts
 - Open a new file and edit it in vim, nano or emacs. Call it bashscript.sh

```
    #!/bin/bash
    for i in $(ls file*); do cut -f 2 $i > $i.out; done
    for i in $(ls *out); do wc $i > $i.wc: done
```

- · Make it executable:
 - · chmod a+x bashscript.sh
 - · ./bashscript.sh (execute)

32

UiO Department of Biosciences

University of Oslo

Piping commands together

Several commands in a row:

```
wc -1 * | sort -n | head -n 3 > results.txt
```

More on this in Exercise 3.

https://github.com/krabberod/bios3010-unix-linux

UiO Department of Biosciences University of Oslo

Some online resources

- Linux Cheat Sheet
 - https://www.cheatography.com/davechild/cheat-sheets/linux-command-line/
 - https://files.fosswire.com/2007/08/fwunixref.pdf
- · Vim Cheat Sheet
 - https://www.cheatography.com/nuggles/cheat-sheets/vim/
- Nano Cheat Sheet
 - https://www.cheatography.com/bipinthite/cheat-sheets/nano-editor/

34

UiO • Department of Biosciences

University of Oslo

Preinstalled programs on Saga

- Saga has already many preinstalled programs.
- Check available modules with
 - module avail
- These can be loaded by the command
 - module load <program>

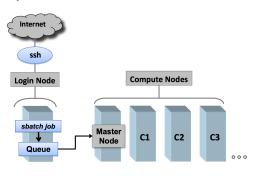
```
[anderkkr@login-1] - $ module avail

ABy55/2_0_2-gnopi-20199
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-1-foss-20180
AdapterRemovi/2_1-foss-20180-AdapterRemovi/2_1-foss-20180
Adapter/18-foss-20180-AdapterRemovi/2_1-foss-20180
Adapter/18-foss-20180-AdapterRemovi/2_1-foss-20180
Adapter/18-foss-20180-foss-20180
Adapter/18-foss-20180-foss-20180
Adapter/18-foss-20180-foss-20180
Adapter/18-foss-20180-foss-20180-foss-20180
Adapter/18-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-foss-20180-fo
```

UiO • Department of Biosciences University of Oslo

Queue system - Slurm

- Since Saga is used by hundreds of people it needs a queueing system to divide the resources
- The queue system is called Slurm.

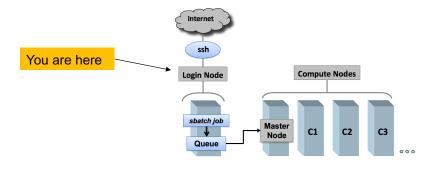


36

UiO : Department of Biosciences University of Oslo

Queue system - SLURM

- Since Saga is used by hundreds of people it needs a queueing system to divide the resources
- The queue system is called Slurm.



UiO • Department of Biosciences

University of Oslo

Typical SLURM script

```
#!/bin/sh
```

#SBATCH --job-name=job

#SBATCH --account=nn9338k

#SBATCH --output=slurm-%j.base

#SBATCH --cpus-per-task=16

#SBATCH --mem-per-cpu=6G

#SBATCH --time=48:00:00

module purge

module load SeqKit/0.11.0

<start the program>

- A slurm-script must start with the line #!/bin/sh
- You need to specify the account with computing hours nn9338k for OMG-peeps
- · You need to specify the amount of resources you need:
 - CPUs per task is the number of cores
 - Mem-per-CPU is the amount of memory per CPU in the previous line (i.e. 16*6Gb = 96 Gb)
- You also need to indicate the time the job will take. Your job will be killed when the time runs out!
- The total amount of CPU hours used is the number of cpus * the actuall running time.
- Then do the commands:
 - Example: load a module. Start a program.

38

UiO * Department of Biosciences

University of Oslo

SLURM script

- Finally send the script to the queue with:
 - sbatch your.script.slurm
- Check the gueue
 - squeue -u <username>
- Check a specific job:
 - scontrol show jobid <jobid nr>

UiO * Department of Biosciences

University of Oslo

Interactive terminal

- On the login node you can do small tasks (zipping, coping moving files). Larger tasks will be killed.
- If you need an interactive termial (i.e. resources without sending a script) you can ask for it with salloc togheter with the wanted amount of resources:

salloc --account=nn9338k --mem-per-cpu=10G --time=10:00:00 --cpus-per-task=8

40

UiO • Department of Biosciences

University of Oslo

Text editors in Linux

- Unix has many built-in ways editing text files.
- For instance nano (or vim)
- CTRL-x Exits the editor. If you are in the middle of editing a file the exit process will ask you if you want to save your work.
- CTRL-o Save file name and continue working.
- CTRL-w Search your text.
- CTRL-g Get help with Nano.



UiO • Department of Biosciences University of Oslo

- Overwhelmed?
- Keep calm and practice the simple stuff.
 - Test on real cases
- Do the exercises:
- https://github.com/krabberod/bios3010-unix-linux