Lesson 2

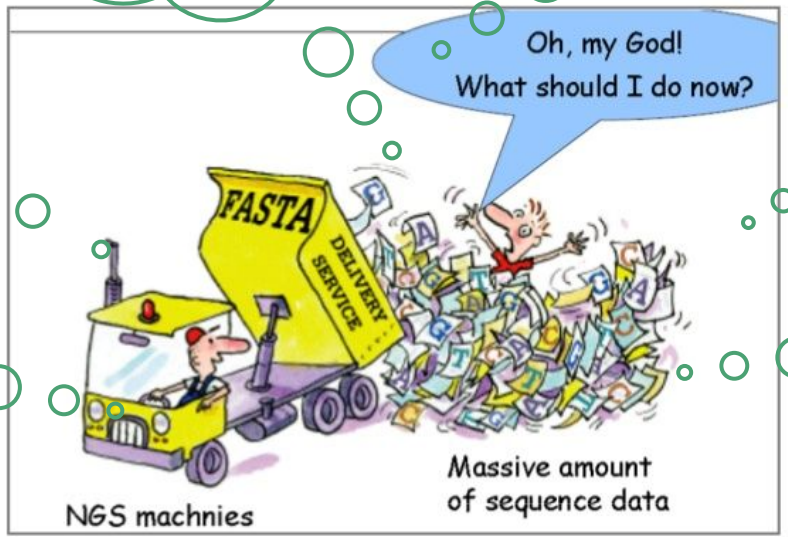
Linux

Uses of NGS

Detect genomic variation

Profile gene expression

Acquire new genomes

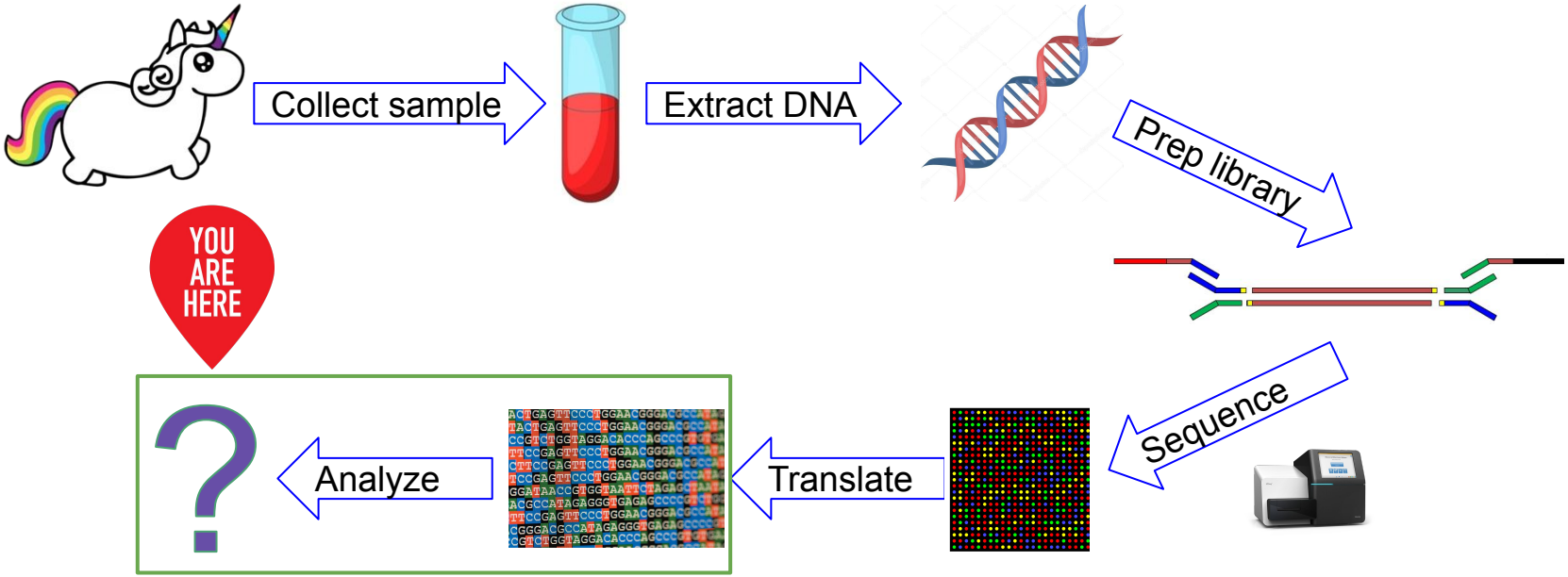


Genotype

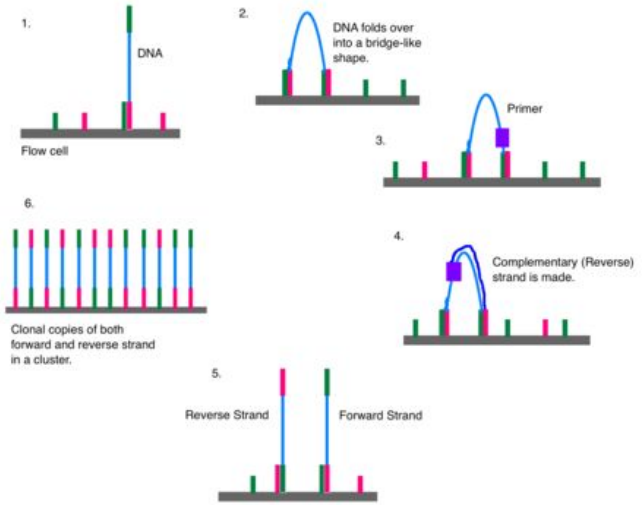
Characterize environments

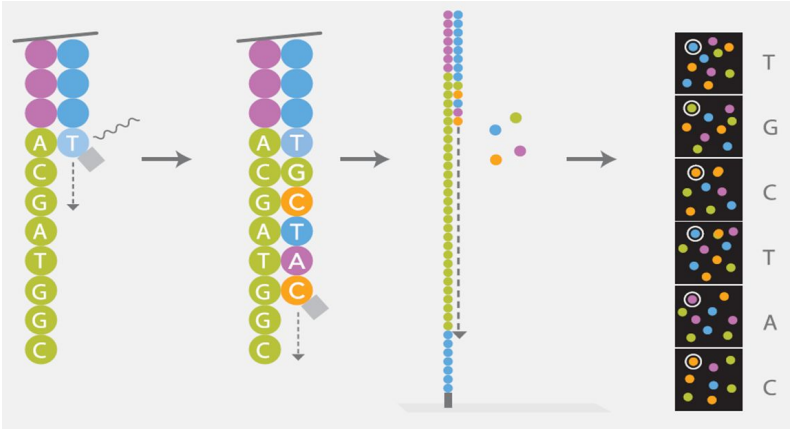
More exotic stuff...

NGS - from organism to sequence

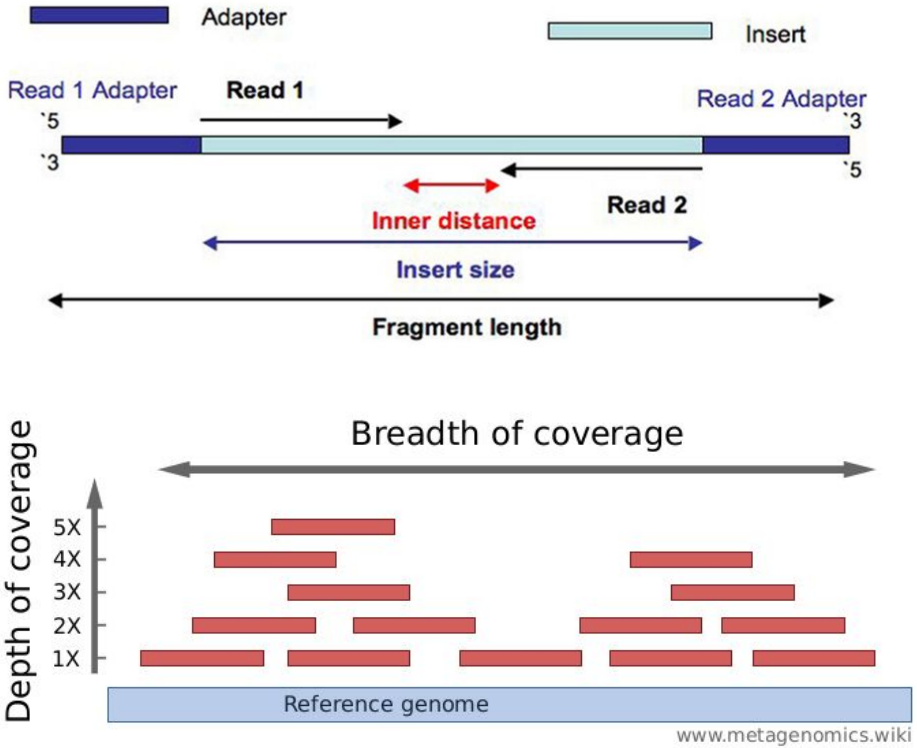


Illumina sequencing technology





Some basics

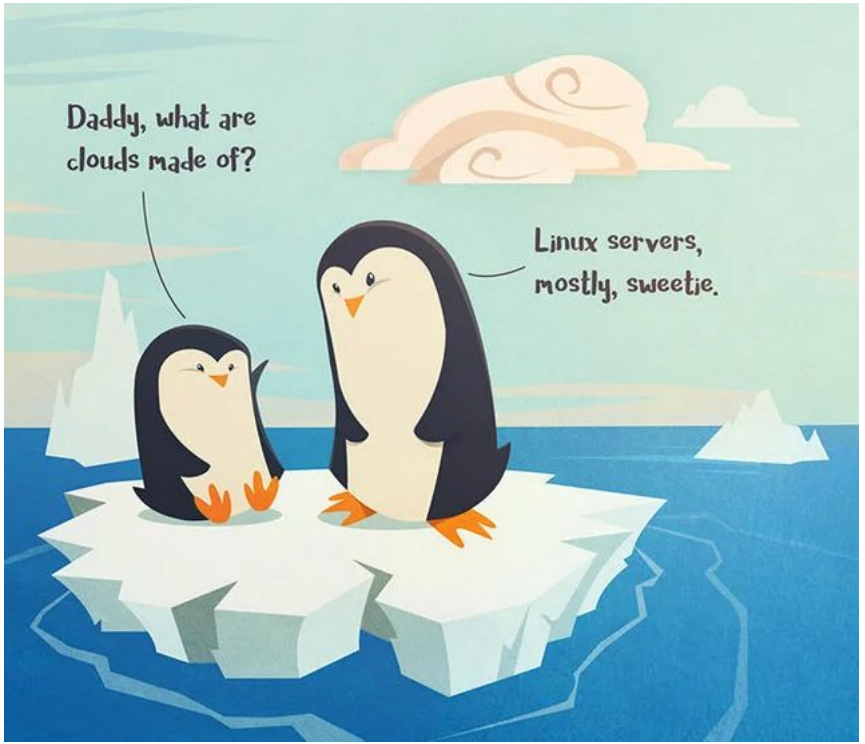

Adapter
Insert
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2
Inner distance
Insert size
Fragment length
Breadth of coverage
Depth of coverage
5X
4X
3X
2X
1X
Reference genome
www.metagenomics.wiki

By the end of this lesson you will...

- Initial familiarity with the Linux command line
- Understand the Linux file hierarchy and path convictions
- Be familiar with some basic Linux commands for:
  - File system navigation
  - File management
  - Text manipulation
- Know how to control processes (piping, redirection, running in background)
- Be able to use CLI for external software in Linux (e.g. Blast)

What is Linux?

- **An operating system** (OS) **kernel**
- Manages software to hardware interactions
- Unix based operating system
- Free and open source github.com/torvalds/linux
- GNU/Linux distribution - Windows, MacOS alternative (Ubuntu, Fedora, Manjaro, etc..)
- Very common - smartphones (Android), desktops and servers



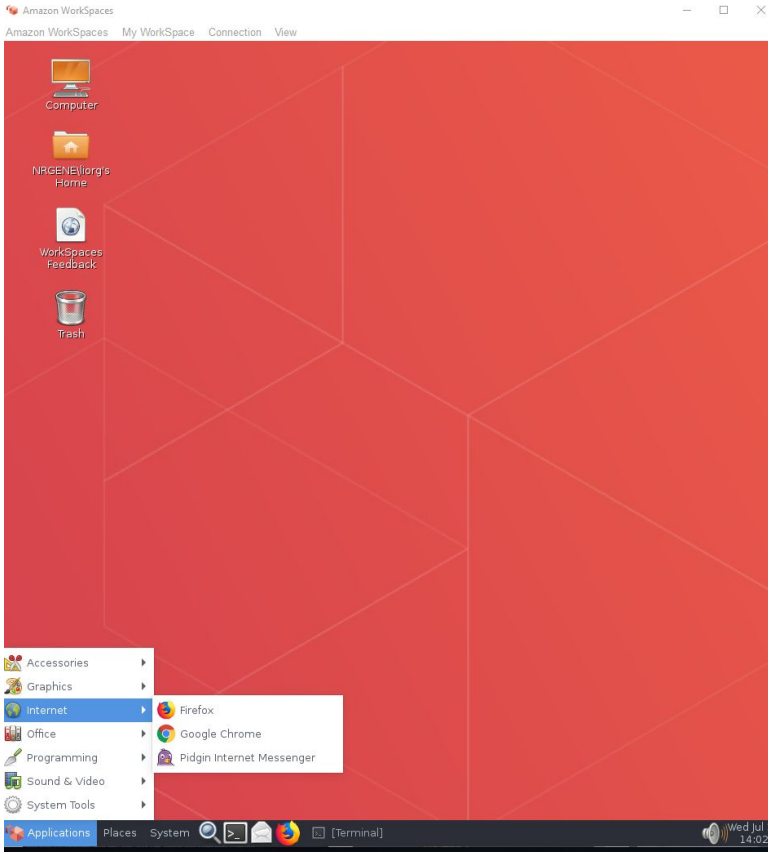
Why Linux?

- Great performance
- Natively runs many scientific software tools
- Provides powerful tools for data management
- Has a strong community of users and developers
- All free to use!

Linux - a (brief) history

- UNIX systems have been around since the 70s
- 1991 - Linus Torvalds releases the first Linux version
- Since, thousands of developers joined the effort and keep making releases
- Many distributions exist (flavors):

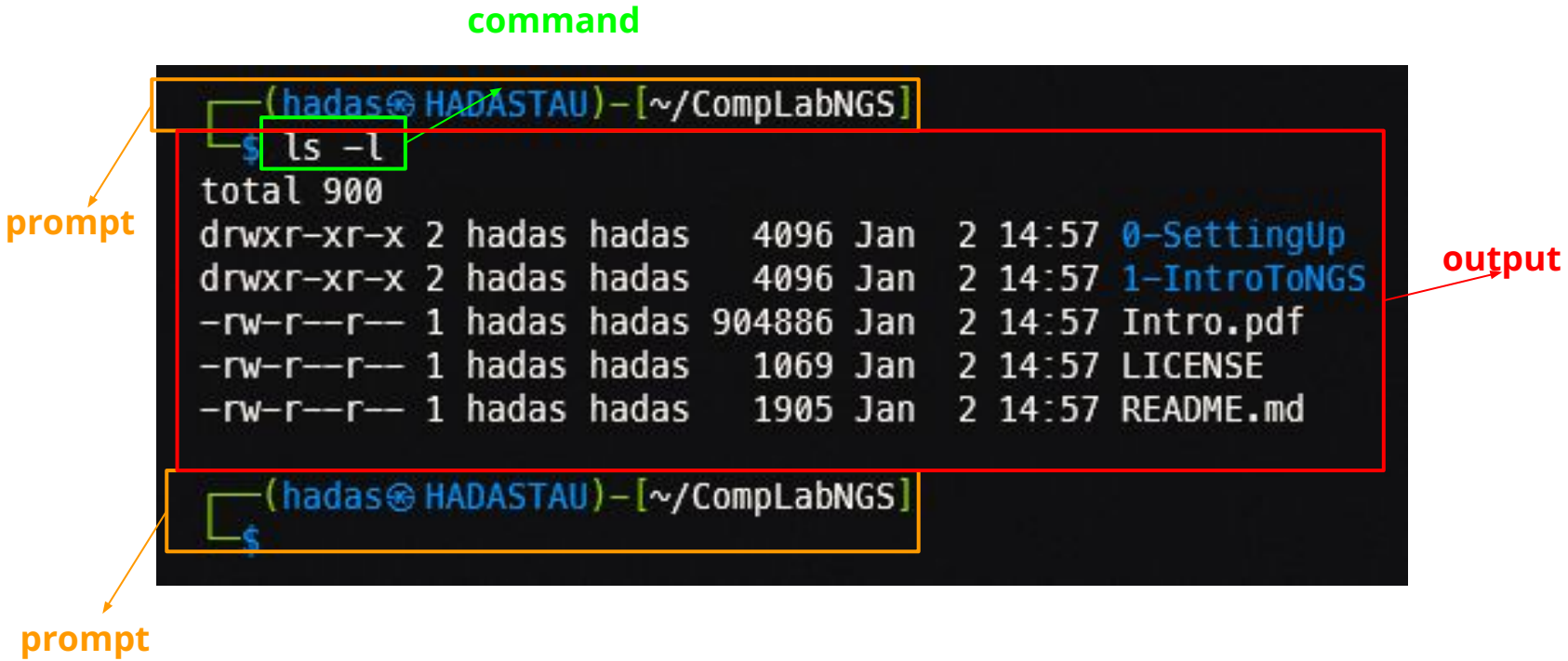
GUI



CLI

```
UDISKS_IGNORE=1
UDISKS_LVM2_PV_NUM_MDA=1
UDISKS_LVM2_PV_UUID=kUqsYj-SMdx-6m5I-FcHI-RYXb-vmvn-bHO3aG
UDISKS_LVM2_PV_VG_EXTENT_COUNT=476870
UDISKS_LVM2_PV_VG_EXTENT_SIZE=4194304
UDISKS_LVM2_PV_VG_FREE_SIZE=142606336
UDISKS_LVM2_PV_VG_LV_LIST=name=root;uuid=OhthbP-Zr4Y-Euey-kfPe-Zzqe-a88g-lZEsXT;size=1993565274112;;active=1
 name=swap_1;uuid=UcyETs-rUcd-p9Cx-QrVy-n5ed-SZNl-kt092Q;size=6429868032;;active=1
UDISKS_LVM2_PV_VG_NAME=ubuntu-vg
UDISKS_LVM2_PV_VG_PV_LIST=uuid=kUqsYj-SMdx-6m5I-FcHI-RYXb-vmvn-bHO3aG;size=2000137748480;allocated_size=1999
995142144
UDISKS_LVM2_PV_VG_SEQNUM=3
UDISKS_LVM2_PV_VG_SIZE=2000137748480
UDISKS_LVM2_PV_VG_UUID=sQg7X1-rTqJ-zIYu-B5q5-jenc-Q6X8-6k8xH3
UDISKS_PARTITION=1
UDISKS_PARTITION_ALIGNMENT_OFFSET=0
UDISKS_PARTITION_NUMBER=5
UDISKS_PARTITION_OFFSET=256901120
UDISKS_PARTITION_SCHEME=mbr
UDISKS_PARTITION_SIZE=2000141942784
UDISKS_PARTITION_SLAVE=/sys/devices/pci0000:00/0000:00:1f.2/ata1/host0/target0:0:0/0:0:0:0/block/sda
UDISKS_PARTITION_TYPE=0x8e
UDISKS_PRESENTATION_HIDE=1
UDISKS_PRESENTATION_NOPOLICY=0
USEC_INITIALIZED=68

bluepenguin@rampage-iii:~$ cat /var/log/udev
```

The command line / terminal



Your first Linux command - ls

List directory contents

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ ls
0-SettingUp  1-IntroToNGS  Intro.pdf  LICENSE  README.md
```

Default behavior is to simply list contents of current location

We can add command **arguments** to change behavior.

For example, specify another location:

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ ls 0-SettingUp/
HowToAsk.md  SettingUp.md
```

Your first Linux command - ls

Or we can use **options** to specify a different behavior.

Some options come without further information - these are called **flags**.

The -l flag provides a detailed contents list:

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ ls -l 0-SettingUp/
total 12
-rw-r--r-- 1 hadas hadas 2068 Jan  2 14:57 HowToAsk.md
-rw-r--r-- 1 hadas hadas 6056 Jan  2 14:57 SettingUp.md
```

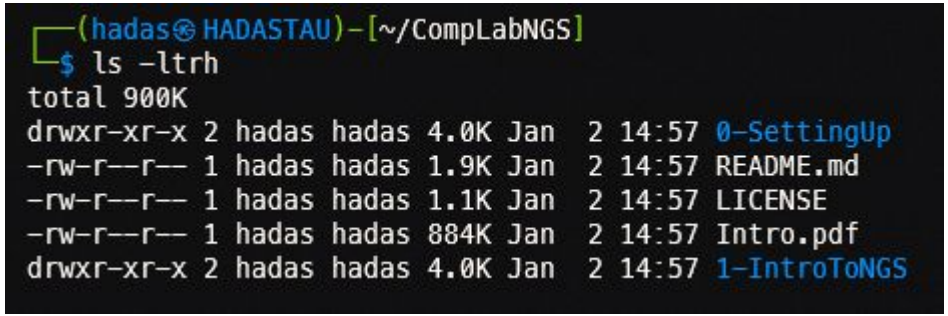
Other options take a value:

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ ls -l 0-SettingUp/ --time-style=full-iso
total 12
-rw-r--r-- 1 hadas hadas 2068 2024-01-02 14:57:46.993502467 +0200 HowToAsk.md
-rw-r--r-- 1 hadas hadas 6056 2024-01-02 14:57:46.993502467 +0200 SettingUp.md
```

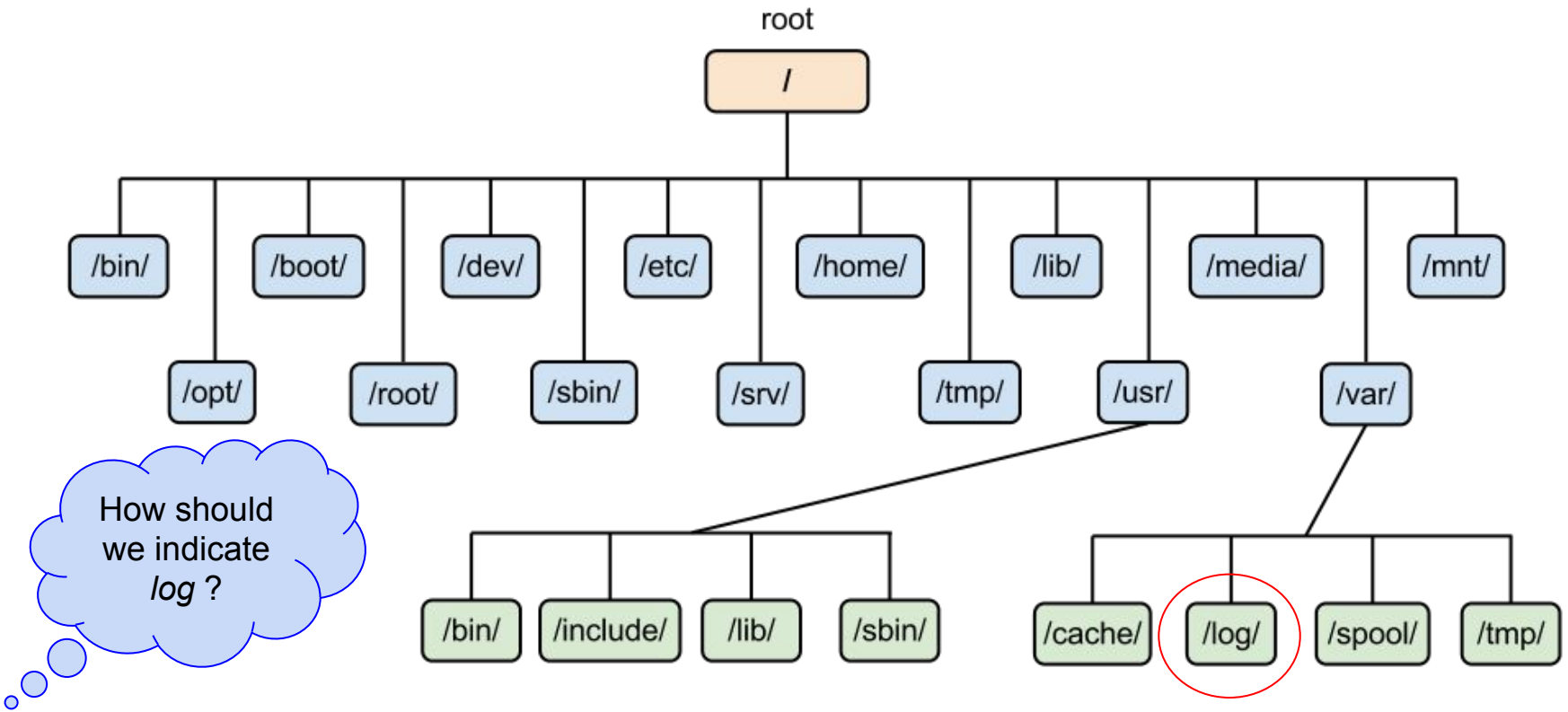
Your first Linux command - ls

You can combine multiple options together, e.g.:

-l - detailed list -t - sort by time -r - reverse order



The Linux file system - hierarchy



Navigating the files system

Where am I? - *pwd* - “present working directory”

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ pwd
/home/hadas/CompLabNGS
```

What’s in here? *ls* - “list”

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ ls
0-SettingUp  1-IntroToNGS  Intro.pdf  LICENSE  README.md
```

Move to directory - *cd* - ‘change directory“

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS]
└─$ cd 0-SettingUp/

┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ pwd
/home/hadas/CompLabNGS/0-SettingUp
```

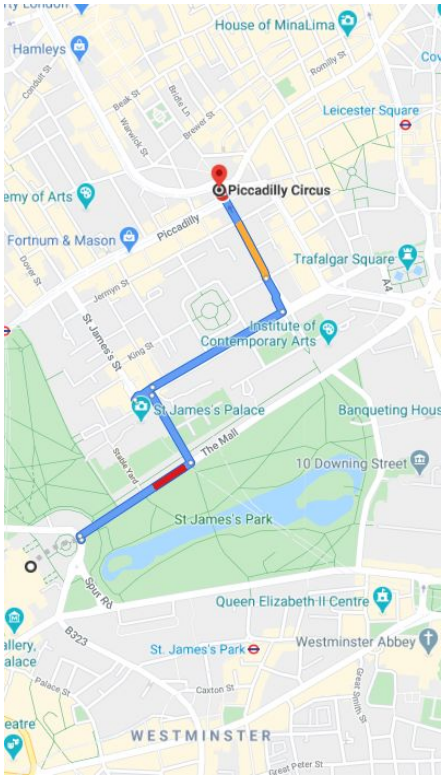
The path

- Each directory or file location is represented by a **path**.
- **Full path -** describes the full way from the root, e.g.:

  `/home/hadas/CompLabNGS/0-SettingUp`

- **Relative path** - relative to current location, e.g.:

  `CompLabNGS/0-SettingUp`



Some path tricks

/ - root

~/ - home directory

./ - current location

../ - parent directory

Get full path - *realpath*

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ pwd
/home/hadas/CompLabNGS/0-SettingUp

┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ cd ../..

┌──(hadas㉿HADASTAU)-[~]
└─$ pwd
/home/hadas
```

```
┌──(hadas㉿HADASTAU)-[~]
└─$ realpath CompLabNGS/0-SettingUp/
/home/hadas/CompLabNGS/0-SettingUp
```

Be lazy! Cut on typing (and typos)

- Use up/down arrows to browse command history

- Use tab to autocomplete paths and commands

More about files in Linux

- Text files vs. binary files
- Extensions (such as .txt) don’t matter! Are used to indicate file types
- Paths are case-sensitive
- Spaces in names - don’t do it - use _ instead

Working with files

Move or rename file
- *mv*

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ mv HowToAsk.md HowToAsk.mvd.md

┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ ls
HowToAsk.mvd.md  SettingUp.md
```

Copy file - *cp*

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ cp HowToAsk.md HowToAsk.copy.md

┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ ls
HowToAsk.copy.md  HowToAsk.md  SettingUp.md
```

Delete (remove) file
- *rm*

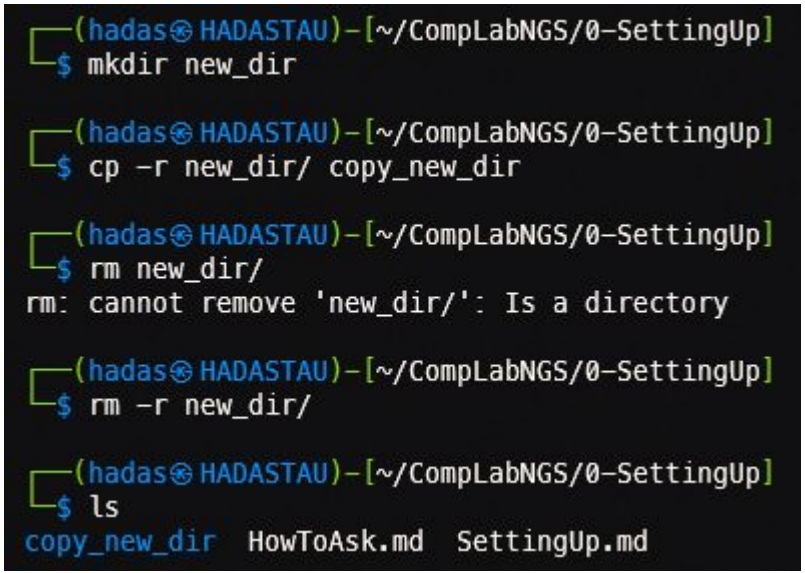
```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ rm HowToAsk.copy.md

┌──(hadas㉿HADASTAU)-[~/CompLabNGS/0-SettingUp]
└─$ ls
HowToAsk.md  SettingUp.md
```

Working with directories

Creating a new directory - *mkdir*

Copying a directory - *cp -r*

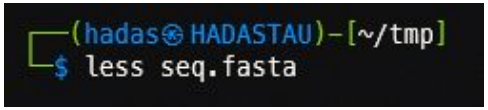
Deleting a directory - *rm -r*



Working with text files

Viewing content - *cat*

For longer files - better to use *less*



Exit with ‘q’

For editing (small files) we can use *vi*

Exit with ‘:q’

Manipulating text files

Use *head* to view the beginning of a file

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/2-Linux/data]
└─$ head UP000000625_83333.fasta
>sp|A0A385XJ53|INSA9_ECOLI Insertion element IS1 9 protein InsA OS=Escherichia coli (strain K12) OX=83333 GN=insA9 PE=3 SV=1
MASVSISCPSCSATDGVVRNGKSTAGHQRYLCSHCRKTWQLQFTYTASQPGTHQKIIDMA
MNGVGCRATARIMGVGLNTILRHLKNSGRSR
>sp|A0A385XJE6|INH21_ECOLI Transposase InsH for insertion sequence element IS5U OS=Escherichia coli (strain K12) OX=83333 GN=insH21 PE=3
MFVIWSHRTGFIMSHQLTFADSEFSSKRRQTRKEIFLSRMEQILPWQNMVEVIEPFYPKA
GNGRRPYPLETMLRIHCMQHWYNLSDGAMEDALYEIASMRLFARLSLDSALPDRTTIMNF
RHLLEQHQLARQLFKTINRWLAEAGVMMTQGTLVDATIIEAPSSTKNKEQQRDPEMHQTK
KGNQWHFGMKAHIGVDAKSGLTHSLVTTAANEHDLNQLGNLLHGEEQFVSADAGYQGAPQ
REELAEVDVDWLIAERPGKVRTLKQHPRKNKTAINIEYMKASIRARVEHPFRIIKRQFGF
VKARYKGLLKNDNQLAMLFTLANLFRADQMIRQWERSH
```

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/2-Linux/data]
└─$ head -n 3 UP000000625_83333.fasta
>sp|A0A385XJ53|INSA9_ECOLI Insertion element IS1 9 protein InsA OS=Escherichia coli (strain K12) OX=83333 GN=insA9 PE=3 SV=1
MASVSISCPSCSATDGVVRNGKSTAGHQRYLCSHCRKTWQLQFTYTASQPGTHQKIIDMA
MNGVGCRATARIMGVGLNTILRHLKNSGRSR
```

Manipulating text files

And *tail* to view the end

```
┌──(hadas㉿HADASTAU)-[~/CompLabNGS/2-Linux/data]
└─$ tail UP000000625_83333.fasta
MKIISKRRAMTIYRQHPESRIFRYCTGKYQWHGSVCHYTGRDVPDIAGVLAVYAERRQDR
NGPYTCLMSITLN
>sp|Q9Z3A0|YJGW_ECOLI Putative uncharacterized protein YjgW OS=Escherichia coli (strain K12) OX=83333 GN=yjgW PE=5 SV=1
MIRKNKWLRFQTVCRYIPLSLKNHNRLVIFVCQRIEWRYIFSTNTGHPLKNTCEMTGTDF
STDGLMRRYVTGGNNQWHTVASLHMTIMMNRIGTVADRQPKARKVKHGEMT
>sp|U3PVA8|IROK_ECOLI Protein IroK OS=Escherichia coli (strain K12) OX=83333 GN=iroK PE=4 SV=1
MKPALRDFIAIVQERLASVTA
>sp|V9HVX0|YPAA_ECOLI Uncharacterized protein YpaA OS=Escherichia coli (strain K12) OX=83333 GN=ypaA PE=4 SV=1
MTIAERLRQEGHQIGWQEGKLEGLHEQAIKIALRMLEQGFDRDQVLAATQLSEADLAANN
H

┌──(hadas㉿HADASTAU)-[~/CompLabNGS/2-Linux/data]
└─$ tail -n 2 UP000000625_83333.fasta
MTIAERLRQEGHQIGWQEGKLEGLHEQAIKIALRMLEQGFDRDQVLAATQLSEADLAANN
H
```

Manipulating text files

wc counts characters, words and lines of a file

If you only want line count, use *wc -l*

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ wc Homo_sapiens_assembly38.fasta
  32178545   32197477 3249912778 Homo_sapiens_assembly38.fasta

┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ wc -l Homo_sapiens_assembly38.fasta
32178545 Homo_sapiens_assembly38.fasta
```

We can also sort the lines of a file with *sort*

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ sort file1
Hello
NGS
TAU
```

Manipulating text files

Filter for lines containing some pattern with *grep*

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ grep TTTTTTTTTTTTTTTTTTTTTTTTCCAAAT Homo_sapiens_assembly38.fasta
AAGCTGTTACCTACATTGCAATTCAAAGTCCTCTTTACTTTTTCGTTGGCTTTTTTTTTTTTTTTTTTTTTTTTTCCAAATAGAAGCTGTCTTTCACCAT
ACCTAATCTCTTAAGTATGGAACCCATAACCAGGGTTCTGATTTCATTACCCATCTTTTTTTTTTTTTTTTTTTTTTTTTTCCAAATGGAGTTTCACCTT
AAATGTGATGTTTGACTTTGGGATGTGACTCAAGGTGACAATTCTTCCCTCCTTGCTGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCCAAATTTTCTTT
TATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCCAAATTGGAGTCAGATGTTTCTAAATATCTCATGGAATATTCCCAAAGATTTTACATGTGTAGTTTT
CAGCTCTTAGAAAAATGAATAGAGCTGGGTTTTTTTTTTTTTTTTTTTTTTTTTTCCAAATTAATGCCTATACCCATAGGTGAATAATGCTTCTTGCTAG
```

If you want all lines that **don't** contain the pattern, use *grep -v*

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ grep -v TTTTTTTTTTTTTTTTTTTTTTTTCCAAAT Homo_sapiens_assembly38.fasta
>chr1  AC:CM000663.2  gi:568336023  LN:248956422  rl:Chromosome  M5:6aef897c3d6ff0c78aff06ac189178dd
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Special characters

If the pattern you're searching for includes spaces - just use “”

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ grep "love NGS" file2
I love NGS a lot!
```

Also recommended if you need to include special characters:

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ grep ">" Homo_sapiens_assembly38.fasta
>chr1  AC:CM000663.2  gi:568336023  LN:248956422  rl:Chromosome  M5:6aef897c3d6ff0c78aff06ac189178dd  AS:GRCh38
>chr2  AC:CM000664.2  gi:568336022  LN:242193529  rl:Chromosome  M5:f98db672eb0993dcfdabafe2a882905c  AS:GRCh38
>chr3  AC:CM000665.2  gi:568336021  LN:198295559  rl:Chromosome  M5:76635a41ea913a405ded820447d067b0  AS:GRCh38
>chr4  AC:CM000666.2  gi:568336020  LN:190214555  rl:Chromosome  M5:3210fecf1eb92d5489da4346b3fddc6e  AS:GRCh38
>chr5  AC:CM000667.2  gi:568336019  LN:181538259  rl:Chromosome  M5:a811b3dc9fe66af729dc0dddf7fa4f13  AS:GRCh38
>chr6  AC:CM000668.2  gi:568336018  LN:170805979  rl:Chromosome  M5:5691468a67c7e7a7b5f2a3a683792c29  AS:GRCh38
>chr7  AC:CM000669.2  gi:568336017  LN:159345973  rl:Chromosome  M5:cc044cc2256a1141212660fb07b6171e  AS:GRCh38
>chr8  AC:CM000670.2  gi:568336016  LN:145138636  rl:Chromosome  M5:c67955b5f7815a9a1edfaa15893d3616  AS:GRCh38
>chr9  AC:CM000671.2  gi:568336015  LN:138394717  rl:Chromosome  M5:6c198acf68b5af7b9d676dfdd531b5de  AS:GRCh38
>chr10  AC:CM000672.2  gi:568336014  LN:133797422  rl:Chromosome  M5:c0eeee7acfdaf31b770a509bdaa6e51a  AS:GRCh38
```

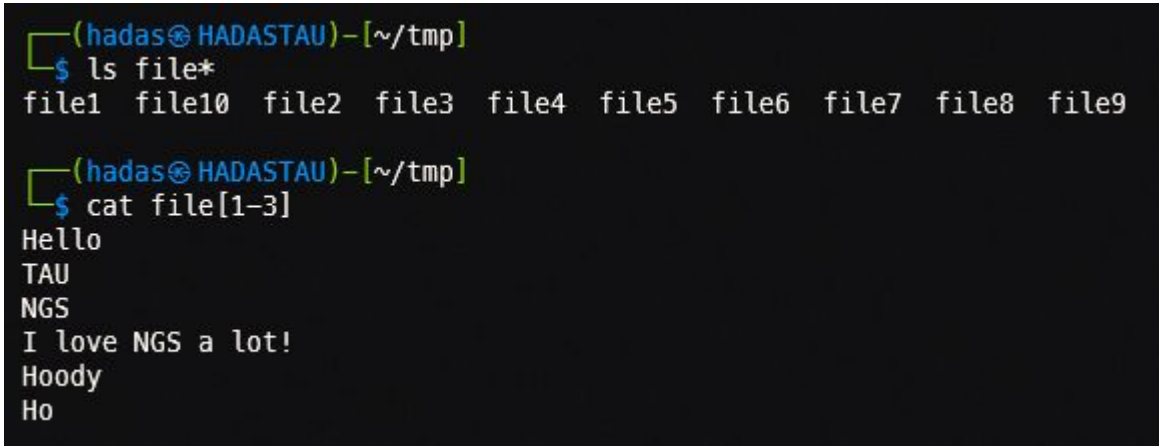
Using wildcards

Instead of providing full file/directory names, we can use wildcards:

* - any number of characters

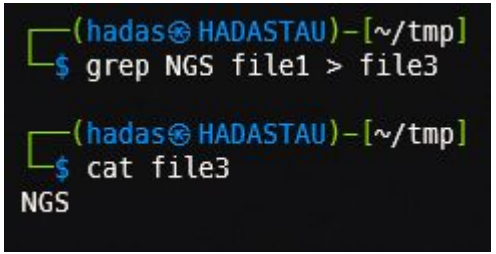
? - one character

[] - range



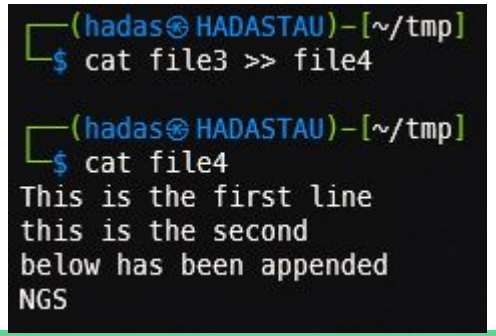
Redirecting command outputs

We can **redirect** output into new files, using the > helper



Careful! If the file already exists, you'll override it!

To **append** to an existing file, use >>



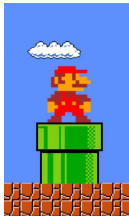
Piping command outputs

We can also **pipe** outputs - give them as inputs for another command.
We use |

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ ls -l | grep new
-rw-r--r-- 1 hadas hadas          0 Jan  3 10:27 I_am_a_new_file
-rw-r--r-- 1 hadas hadas          0 Jan  3 10:27 new_file_I_am
```

We can use multiple pipes

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat file* | grep "a" | sort -r
I love NGS a lot!
below has been appended
```



How can we save the first 100 lines of a file to a new file?

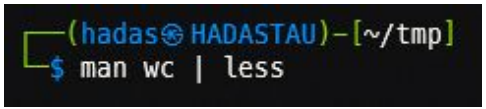
What if we just want to view the first 100 lines in *less* ?

What command will print the name of the file that was changed last to a new file, but only if its name contains ".fasta"?

CONFUSED?

Getting Linux help

- All Linux command have a manual (man page) - *man <command>*



- Some commands have a help flag such as *-h, -help, --help*

- Just ChatGPT/Google your question

- Stack Overflow/Exchange and other forums are a great place for validated answers

Running external software

- So far, we've only seen native Linux commands
- External software can be installed and run
- CLI similar to Linux commands
- Use *man <command>* or *<command> -h* to get help

Long-running commands

- *<command> [options] >out 2>err &*

  >out - redirect output to file ./out

  2>err - redirect error messages to ./err

  & - run process in background

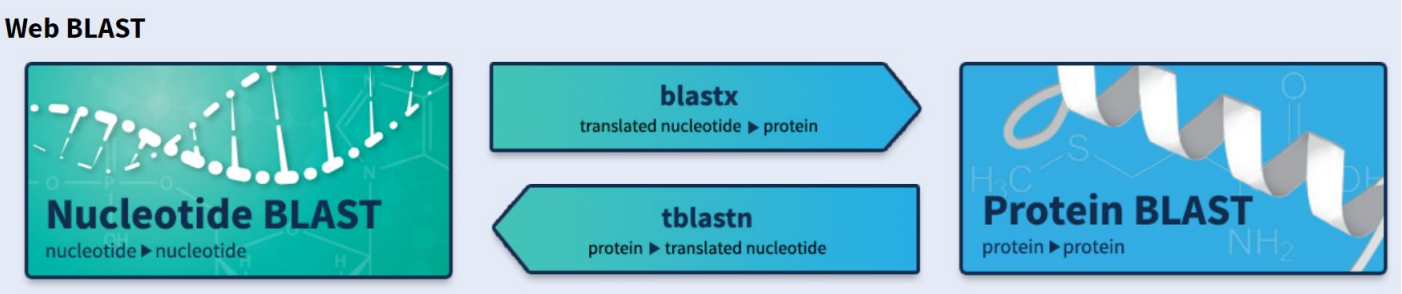
- `jobs` – what commands are running (in background)
- Ctl+C - terminate process
- Ctl+Z - suspend process

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat file2 >out 2>err &
[1] 1687

┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat out
I love NGS a lot!
[1]+  Done                    cat file2 > out 2> err

┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat err
```

```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat file11 >out 2>err &
[1] 1691

┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat out
[1]+  Exit 1                  cat file11 > out 2> err

┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ cat err
cat: file11: No such file or directory
```

Running external commands - example

- BLAST - Basic Search Alignment Tool
- Has a web interface
- But also has a standalone CLI
  - Better control
  - Better performance
  - Use your own DB
- Simply type e.g. *tblastn*



```
┌──(hadas㉿HADASTAU)-[~/tmp]
└─$ tblastn -h
USAGE
  tblastn [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-negative_seqidlist filename]
    [-taxids taxids] [-negative_taxids taxids] [-taxidlist filename]
    [-negative_taxidlist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-qcov_hsp_perc float_value] [-max_hsps int_value]
    [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-sum_stats bool_value] [-db_gencode int_value] [-ungapped]
    [-max_intron_length length] [-seg SEG_options]
    [-soft_masking soft_masking] [-matrix matrix_name]
    [-threshold float_value] [-culling_limit int_value]
    [-best_hit_overhang float_value] [-best_hit_score_edge float_value]
    [-subject_besthit] [-window_size int_value] [-lcase_masking]
    [-query_loc range] [-parse_deflines] [-outfmt format] [-show_gis]
    [-num_descriptions int_value] [-num_alignments int_value]
    [-line_length line_length] [-html] [-sorthits sort_hits]
    [-sorthsps sort_hsps] [-max_target_seqs num_sequences]
    [-num_threads int_value] [-mt_mode int_value] [-remote]
    [-comp_based_stats compo] [-use_sw_tback] [-in_pssm psi_chkpt_file]
    [-version]

DESCRIPTION
   Protein Query-Translated Subject BLAST 2.12.0+

Use '-help' to print detailed descriptions of command line arguments
```