Getting started with command-line bioinformatics

Following "Ten simple rules for getting started with command-line bioinformatics" by Brandies and Hogg, 2021
https://doi.org/10.1371/journal.pcbi.1008645

Sophia Derdak

Core Facilities MedUni Wien

sophia.derdak@meduniwien.ac.at

1. Get familiar with basic computing terminology

We are going to use the interactive Q&A tool Mentimeter, please connect here (from your phone or computer):

https://www.menti.com

[CODE:]

[GABIO_20240417_10rules]

2. Get familiar with basic bash commands in the terminal

```
ls          list directory contents

more        a filter for paging through text one screenful at a time
less        reads and displays only the first screenful of a file first

cat         concatenate files and print on the standard output

head        output the first part of files
tail        output the last part of files
```

2. Get familiar with basic bash commands in the terminal

`grep` **"global regular expression print", print lines that match patterns**

Example:
`grep "chr5" myfile`   Prints all lines that contain the string "chr5" in the file "myfile"

`awk` **pattern scanning and processing language, for column-based files**

Example:
`awk '$1==5 {print $2, $3}' myfile`   For rows in the file "myfile" where the value in column 1 is equal to 5, print columns 2 and 3

`sed` **stream editor for filtering and transforming text**

Example:
`sed 's/sample1/ID4711/g' myfile`   Replace all occurences of "sample1" with "ID4711" in the file "myfile" and print the result

3. Find keys for special characters on your computer keyboard

```
{ }

|

[ ]

`   `   "

@

#

/    \

~   –   _

^

>   <

$

&
```

PS.: … and start using a plain text editor for everything command-line related!
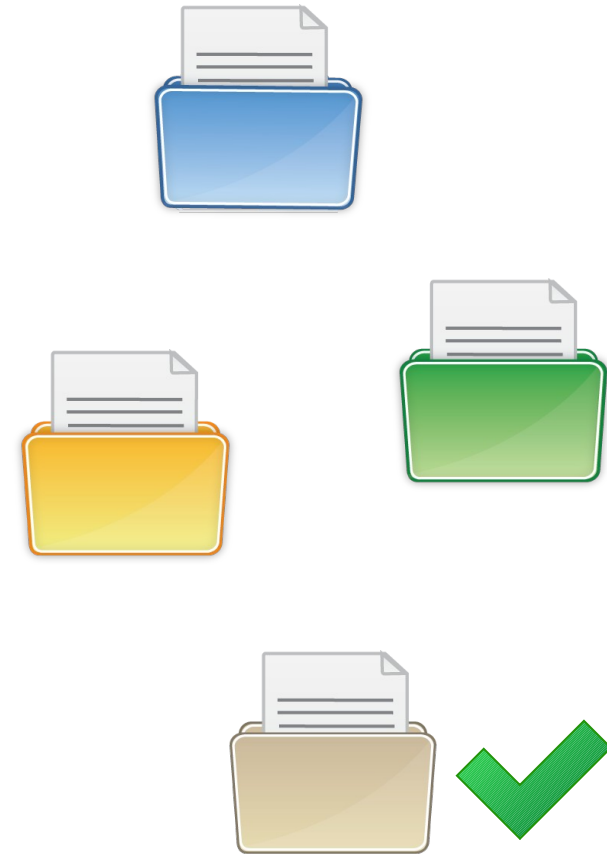
4. Know your data and select analysis tools
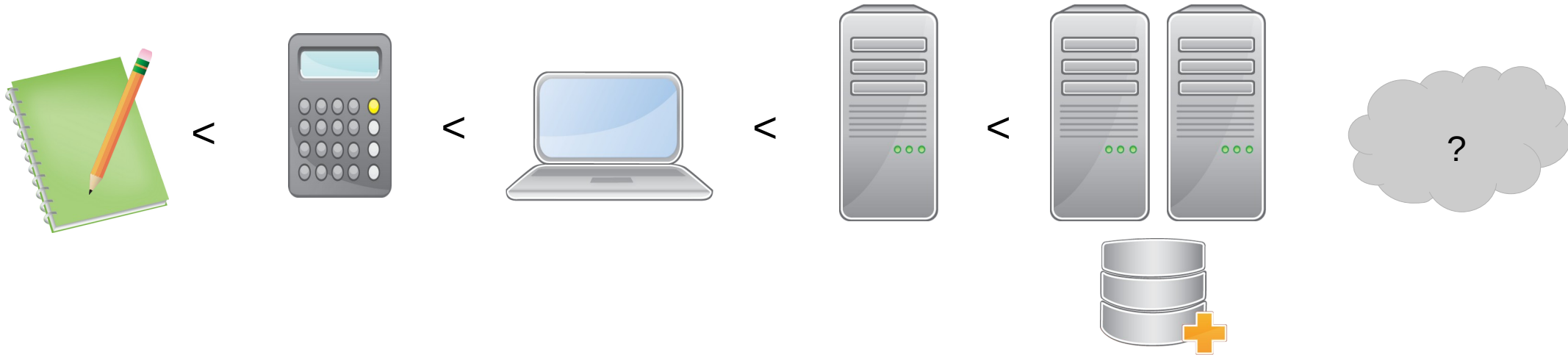

Your research question

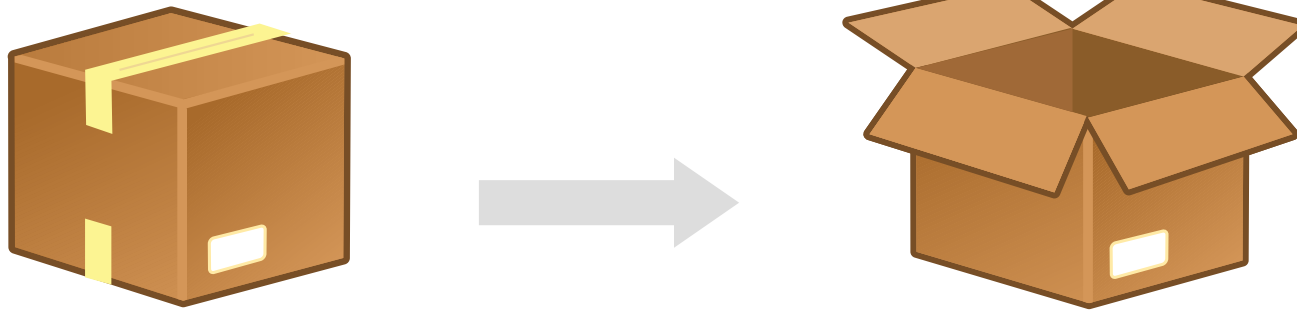Data type and quality

Available computing resources

Bioinformatic software selection

5. Estimate your computing requirements and explore computing options

- Storage for input and output data
- RAM
- CPU/cores
- GPU?
- Local vs "cloud"
- Get to know the file system

6. Understand the basics of software installation

- Do I have enough permissions on the system?
- Download executables (and make them executable)
- Use of package managers (apt, yum, conda, pip...)
- Build and install from source code (make)
- Containers (docker, singularity...)

7. Carefully compose and test commands and scripts

1. Did the installation work?

```
mysoftware -h
mysoftware --help
```

2. Read the documentation (online, command-line, README...)

3. Configure an example command offline in a text editor, then paste it to the terminal and run it

4. Configure and run your "real" command and check the output

5. Combine several commands to a "pipeline", run and check the output, "debug" if necessary

6. Monitor and optimize resource usage when working with "real" data

8. Take notes and document your commands

- Use a plain text editor for taking notes and drafting commands

- Keep log files where you save:
  - software installation procedures
  - data downloads
  - directory structure in your file system
  - all drafts and commands you configured throughout the process of setting up the analysis pipeline
  - errors and workarounds
  - runtimes

- Comment copiously on the use of certain parameters, variants of the command etc.

- Later: use version control (git ...) for your scripts

9. Join the bioinformatics community

- Find allies for discussion (colleagues, other learners, or those who know already)

- Learn how to write and do post questions in forums, on github etc., when you are stuck

- Post answers in forums on topics you know about

- Explain your work to others

- Improve your interdisciplinary language (biology-medicine-(bio)informatics)
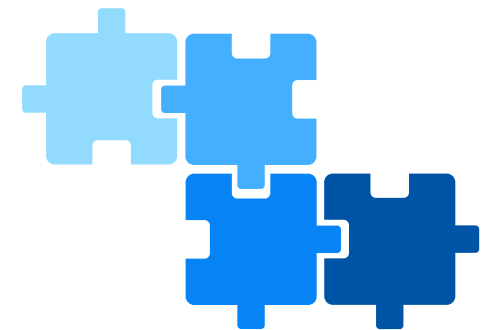
10. Endurance and day-to-day work

Double-check all data, scripts etc. passed on to you before you start working with them

Beware of changing versions of tools and databases you use

Document your own work fully

Write READMEs for files and folders

Record a log of everything you run on the computer

Backup!

Thank you!

Sophia Derdak

Core Facilities MedUni Vienna

sophia.derdak@meduniwien.ac.at