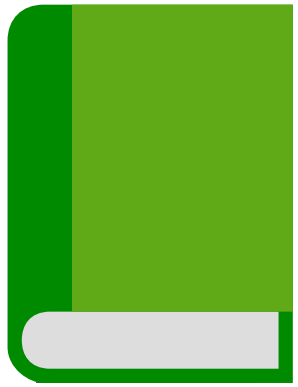
Getting started with command-line bioinformatics: An introductive note for a beginner

@ange_omics

25th November, 2022

Ten rules to get started!

1.  Terminology
2.  Tool selection
3.  Resources needed
4.  Platform selection
5.  Software installation
6.  Scripting
7.  Monitoring
8.  File handling
9.  Take note again and again!
10.  No 빨리빨리... Be Patient



1

Terminology

1

Terminology

CONTAINER

WALLTIME

RAM

HPC

THREAD

PIPELINE

EXECUTABLE

CPU

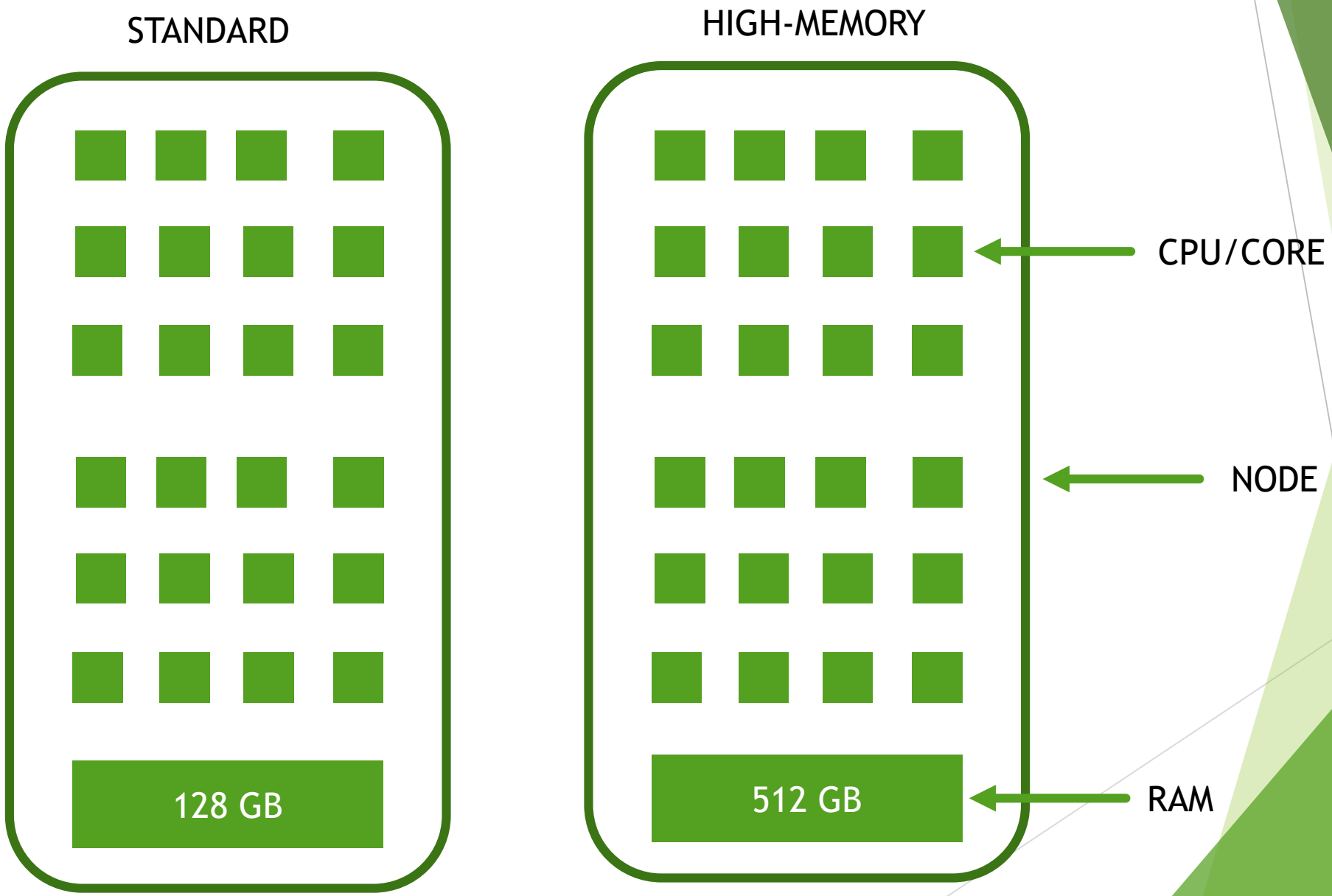
PATH

DEPENDENCY

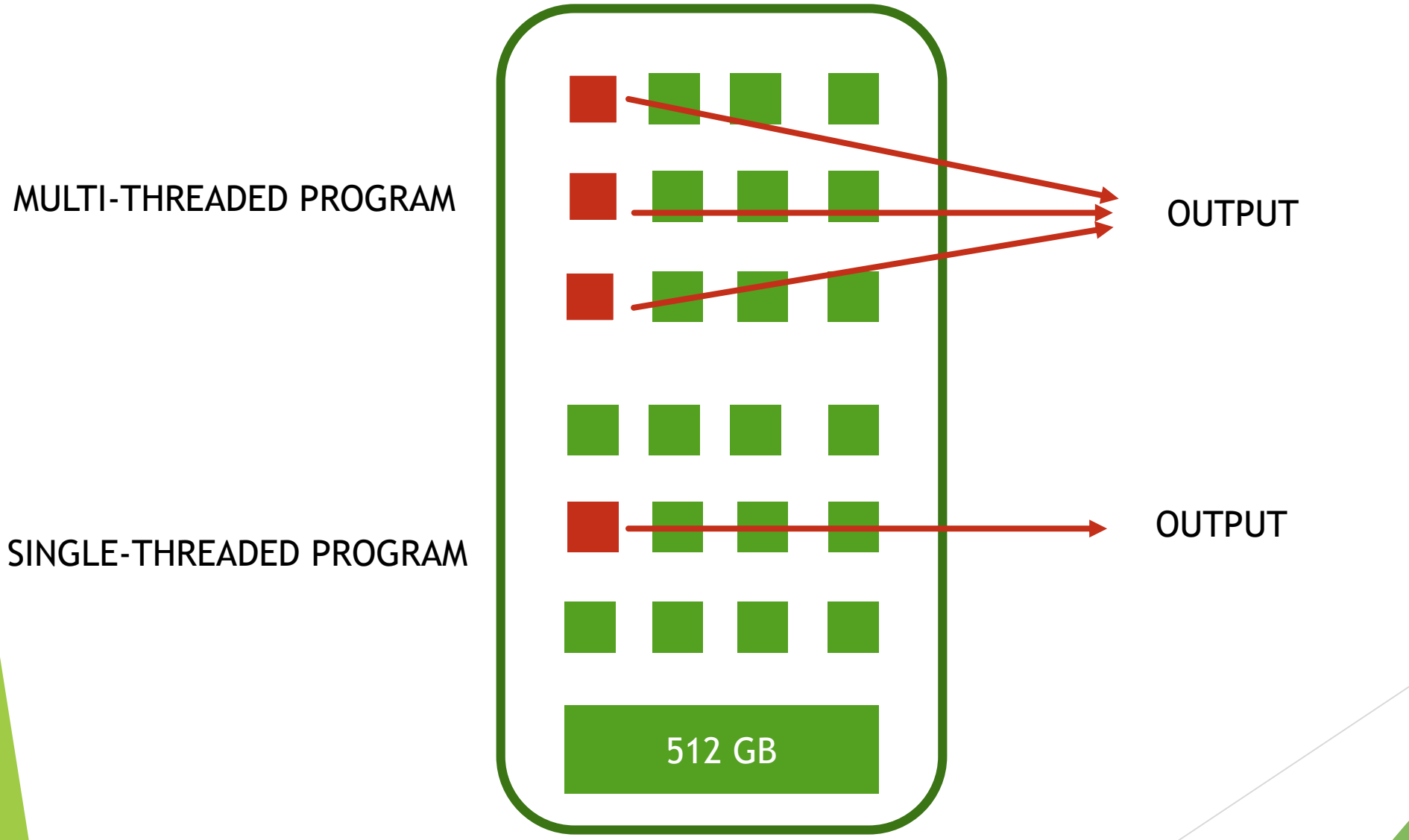
DEPENDENCY

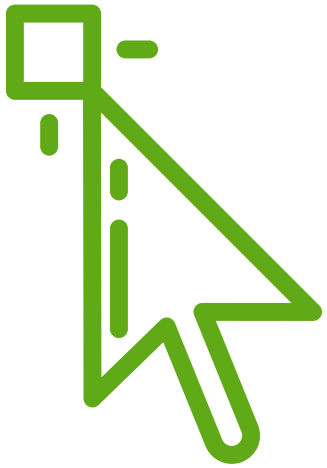
1

Terminology



Terminology





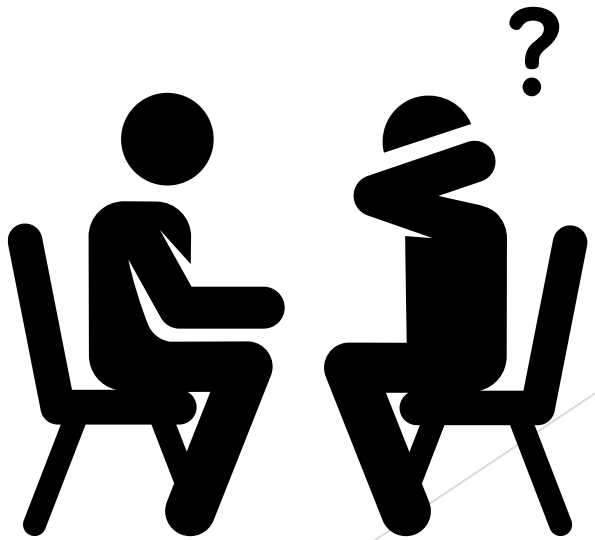
2

Tool selection

Tool selection

Know your data and assess your need

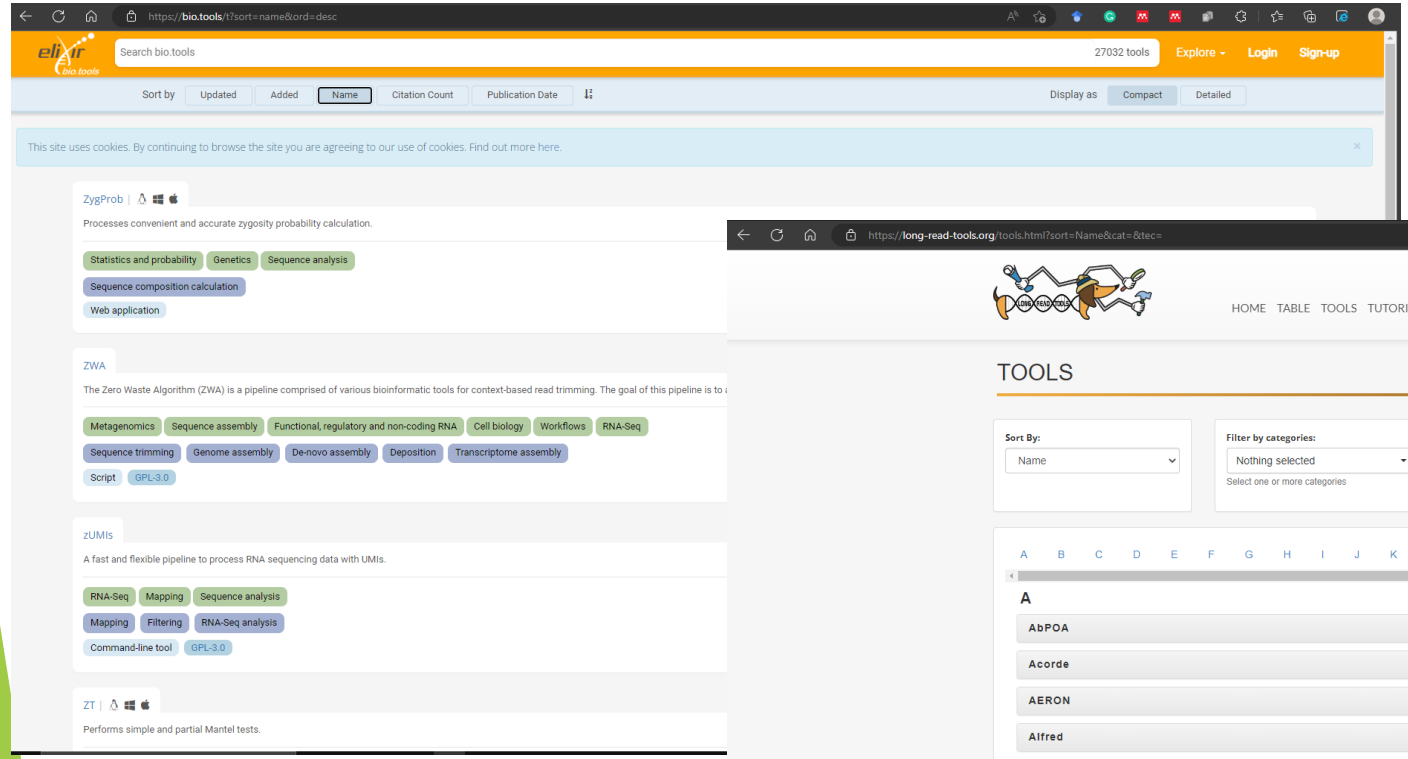
- Data type
- Target species
- Available computing power
- Already tested tools?

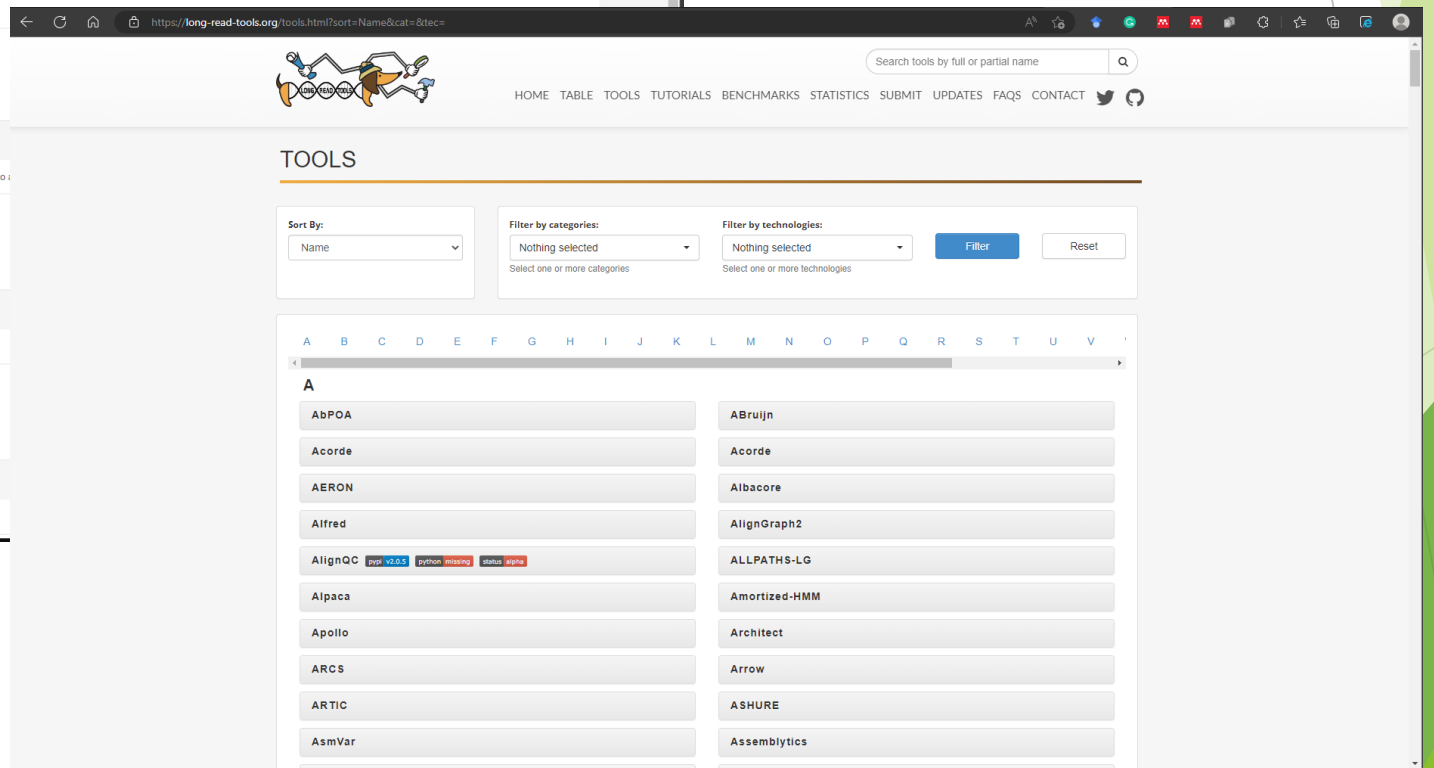


2

Tool selection

Know your data and assess your need - Tools available?





3

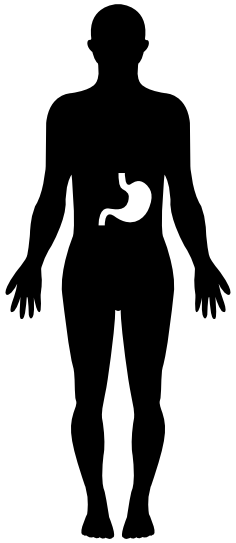
Resources needed

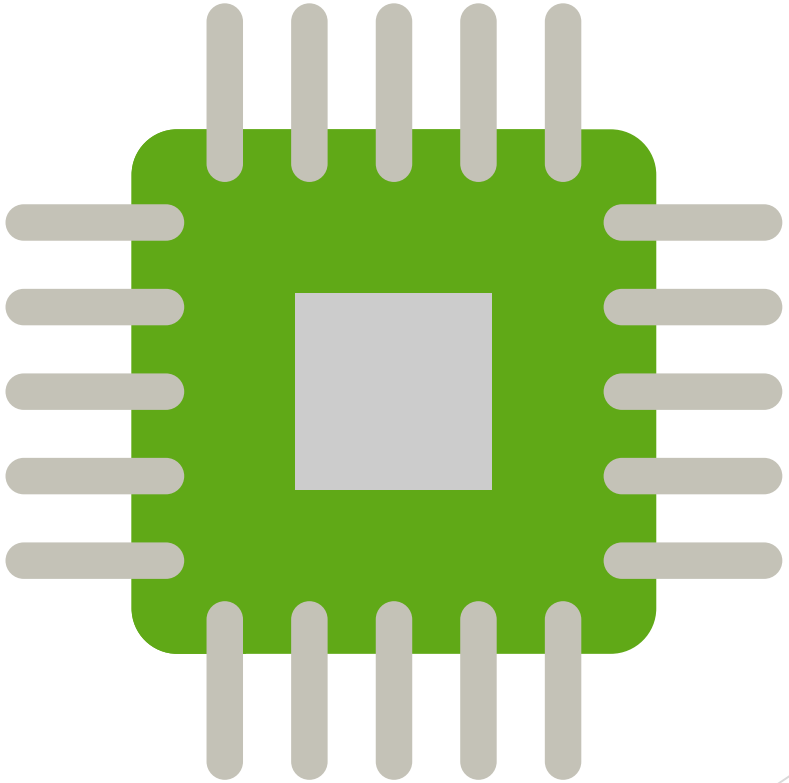
Resources needed

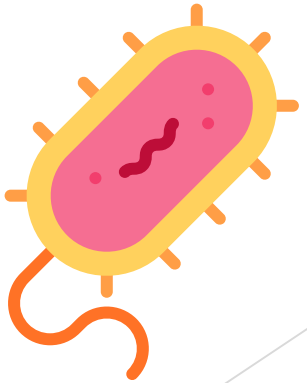
Estimating your computing requirements

- Tool documentation + datasets
- Tools benchmarking papers
- Ask colleagues

32 CPUs
128 RAM
1TB Storage







4

Platform selection

Platform selection

Cloud

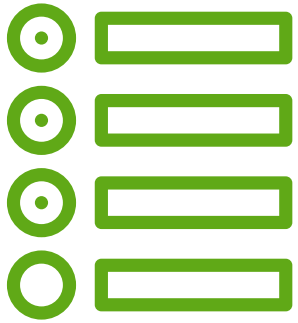
Customised computing resources
Unshared resource (live programming)
Unshared resource (live programming)
Both free and paid commercial options

https://ronin.cloud/
https://aws.amazon.com/
https://cloud.google.com/

Personal Computer/work station

Fixed computing resources
Resource not shared(live programming)
Compute environment self controlled
Cost for acquisition but flexibility

Shared HPC



Fixed computing resources
Shared resource (scheduled jobs)
Compute environment largely controlled by IT
Usually freely available based on institution/merit

5

Software installation

Software installation

Manual

- Download
- Unpack
- Configure
- Build
- Install

Package managers

- APT –Debian/Ubuntu
- Yum – RedHat/Centos
- Conda/Bioconda/Mamba

Containers

- Docker
- Singularity

Workflow managers

- Nextflow
- Snakemake
- Galaxy

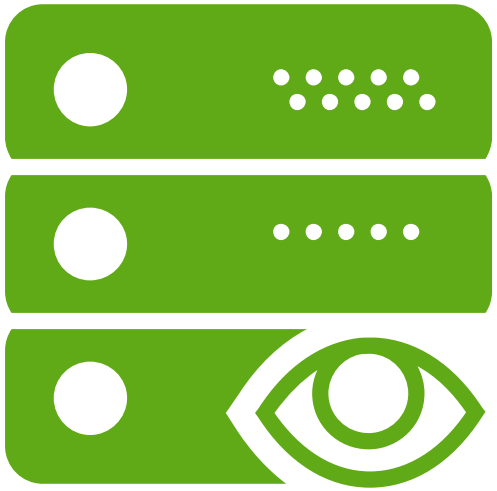
6 Scripting

Scripting

- Write your script
- Check syntax and minor scripting error
- Software well installed
- Dependencies well installed
- Resource sufficient?

https://github.com/

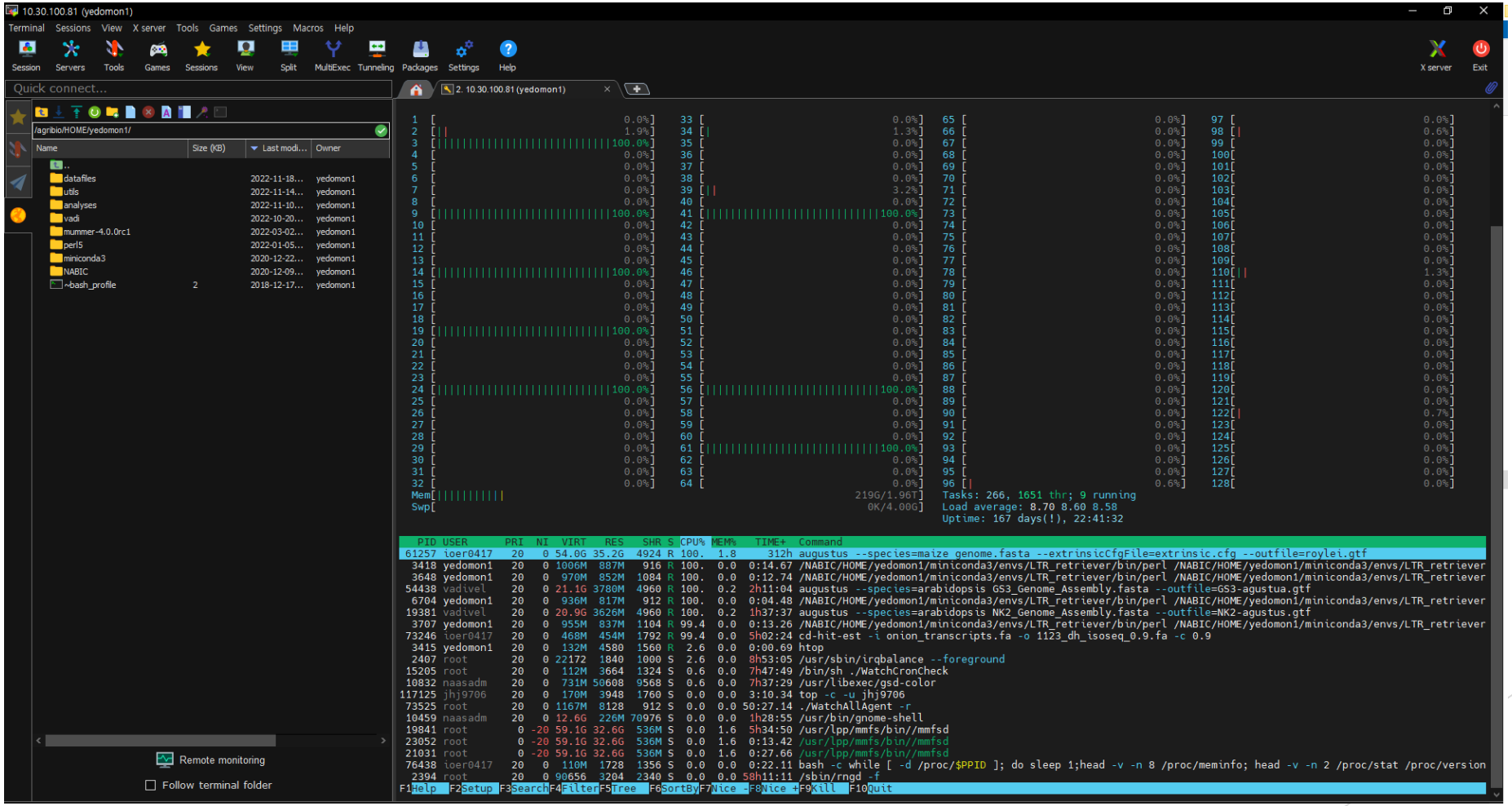
https://snyk.io/code-checker/



7 Monitoring

Monitoring

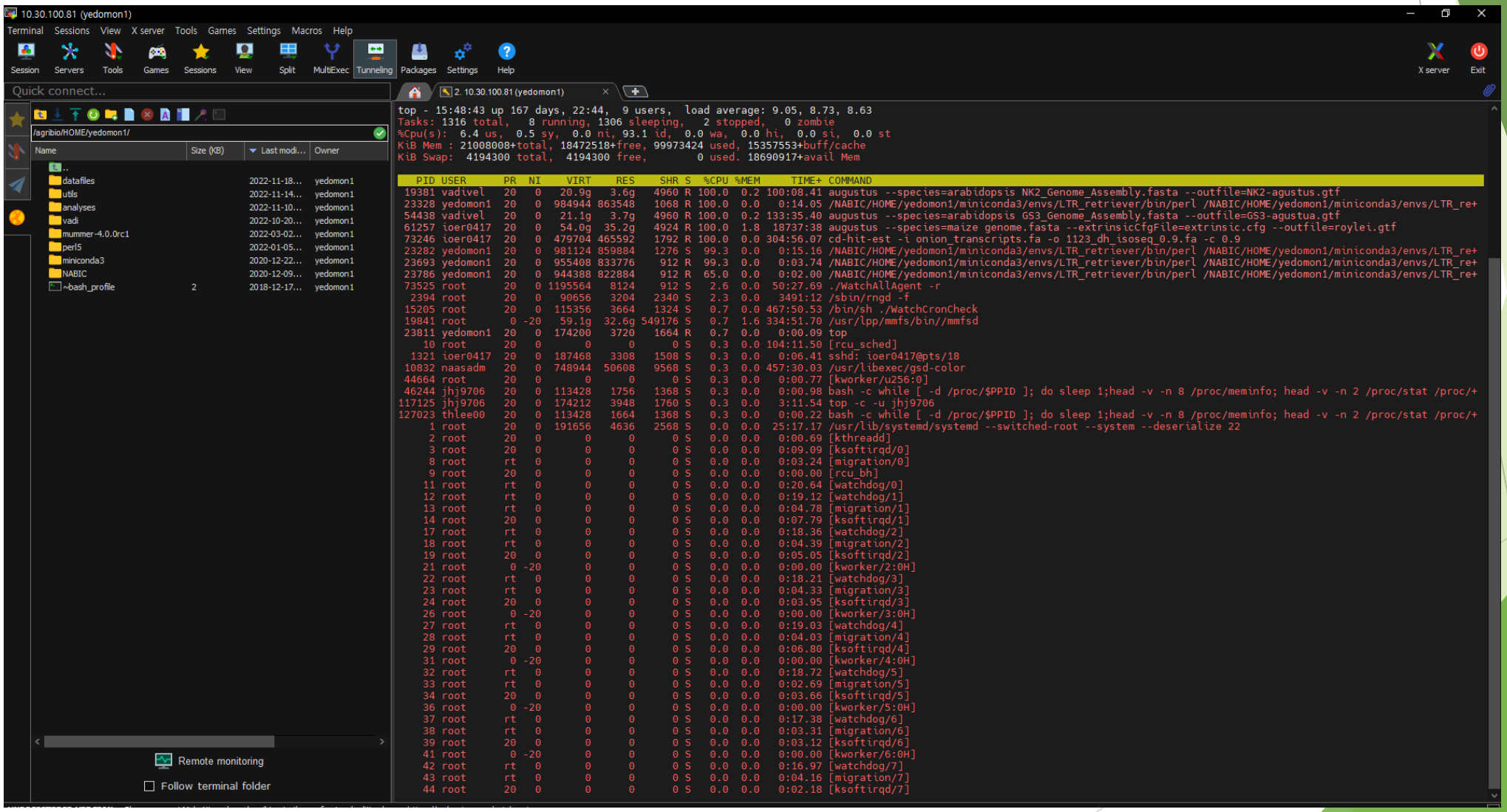
HTOP



7

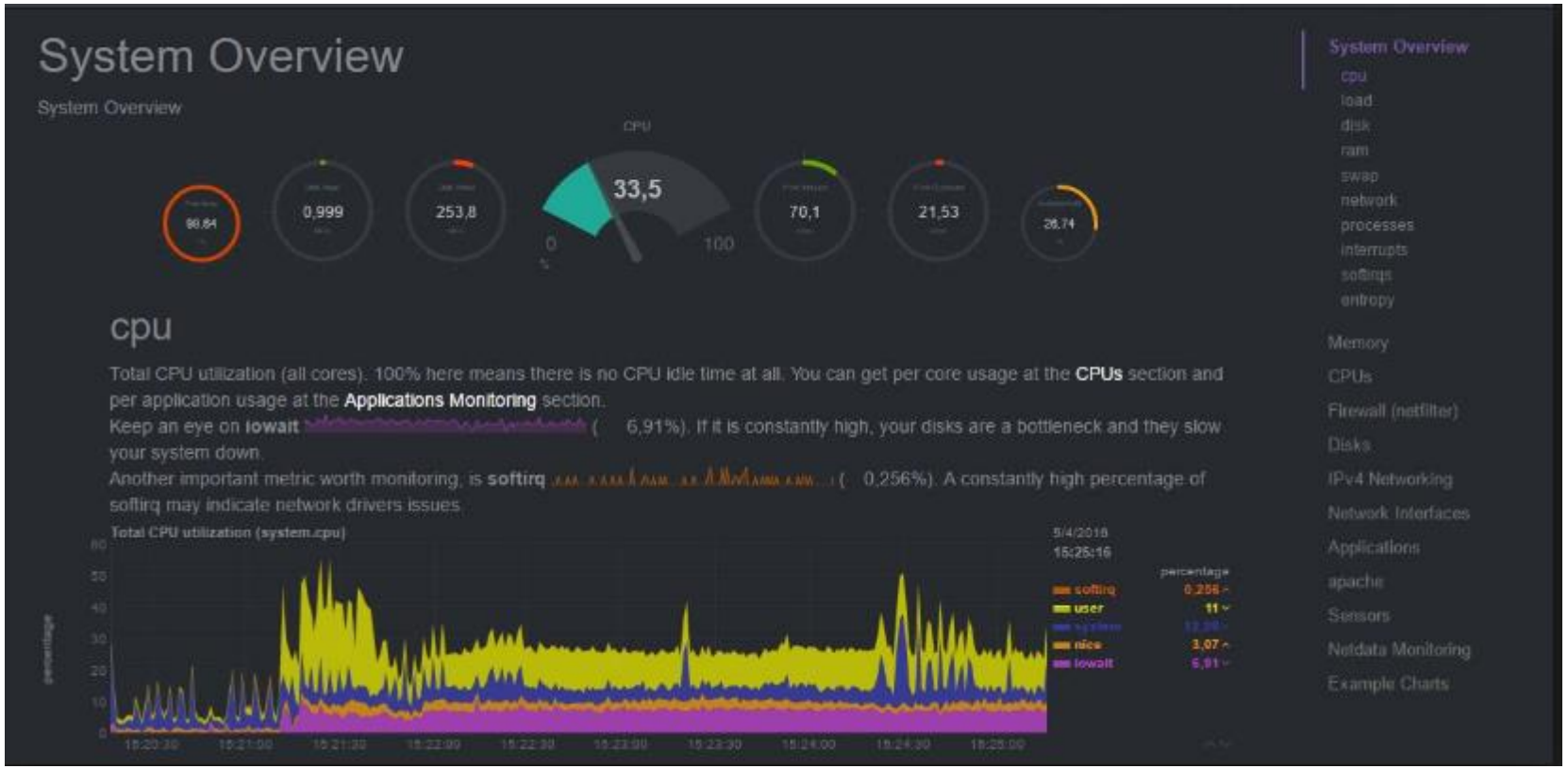
Monitoring

TOP



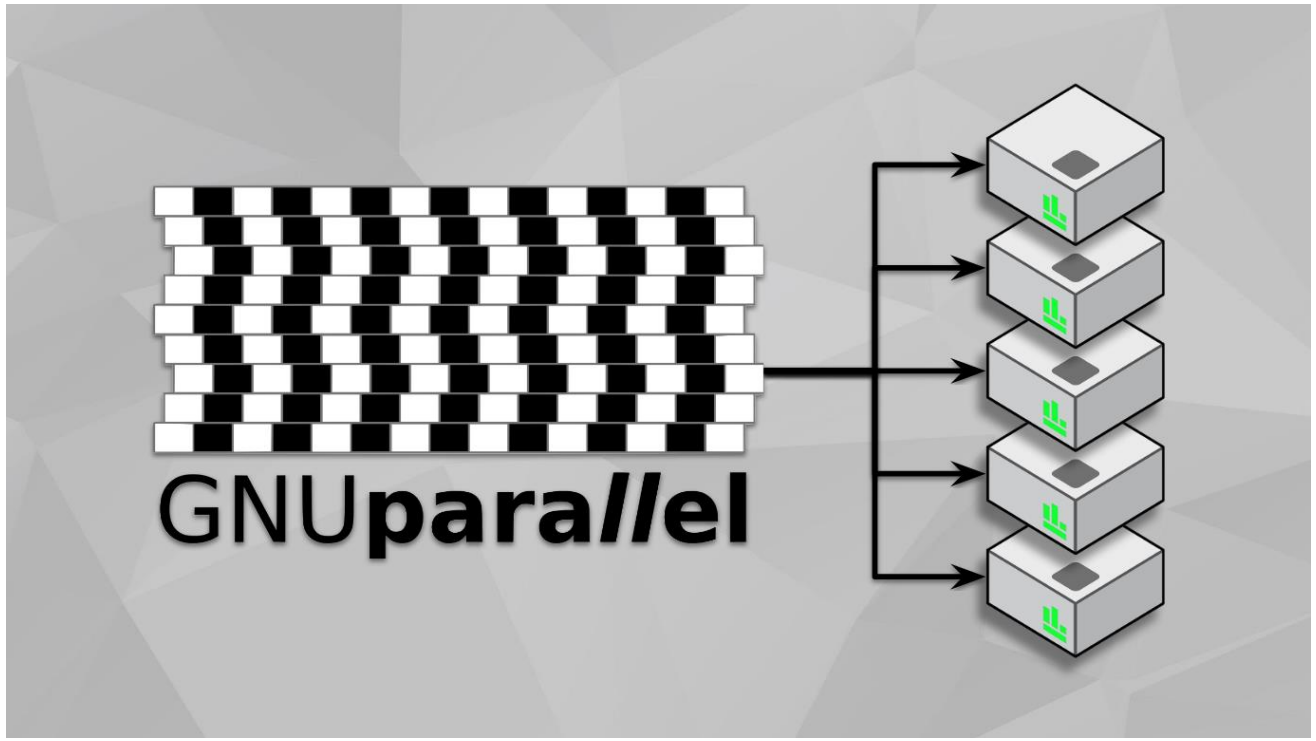
Monitoring

NETDATA



Monitoring

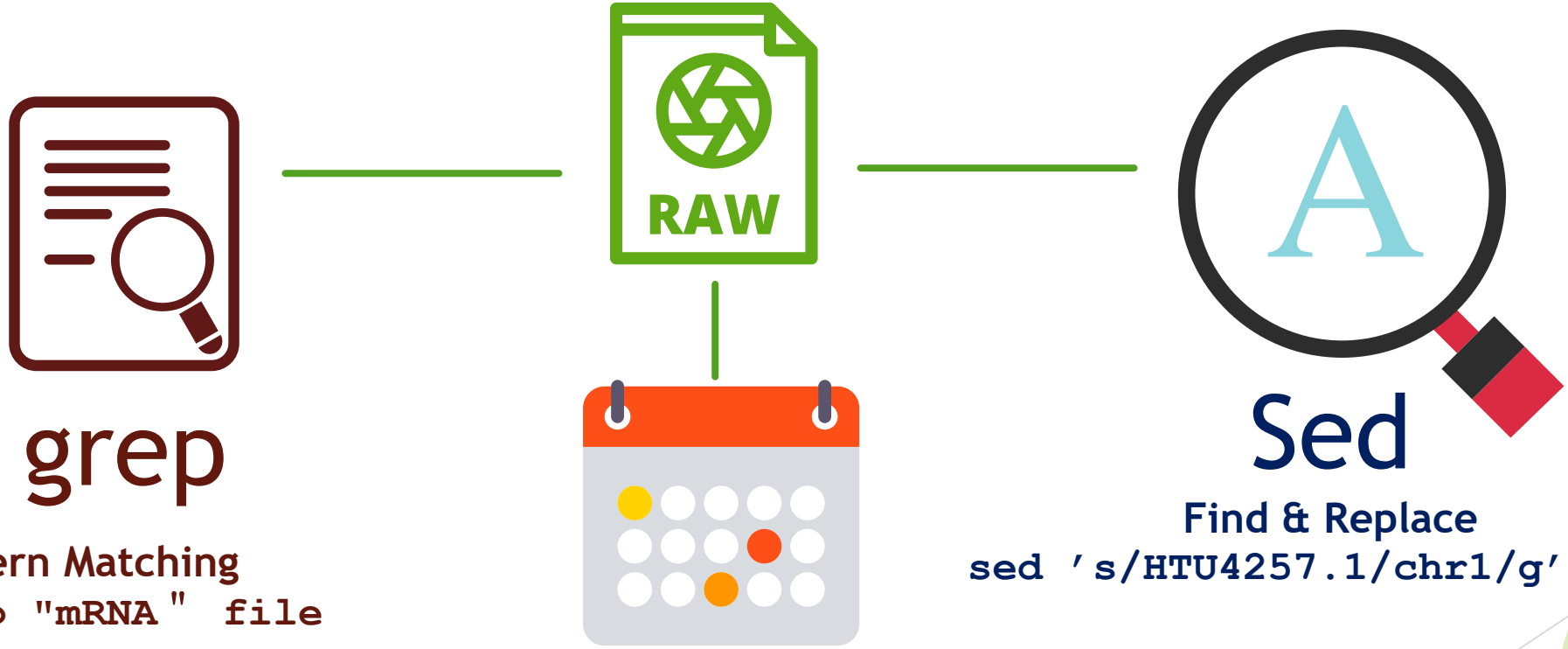
You say optimization? Get more thing dome with GNU parallel



8 File handling

File handling

Frequently used programs to parse files in Linux – Get familiar with them



grep

Pattern Matching

```
grep "mRNA" file
```

Awk

Data processing

```
awk '$1 == 100 {print $2, $3}' file
```

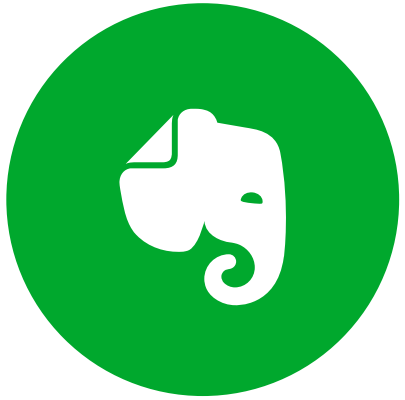
Sed

Find & Replace

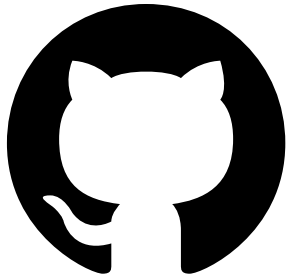
```
sed 's/HTU4257.1/chr1/g' file
```

9 Note taking

Note taking

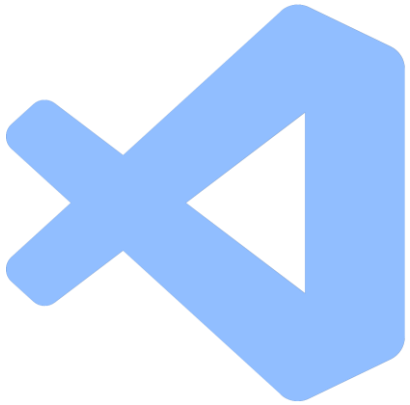


Evernote

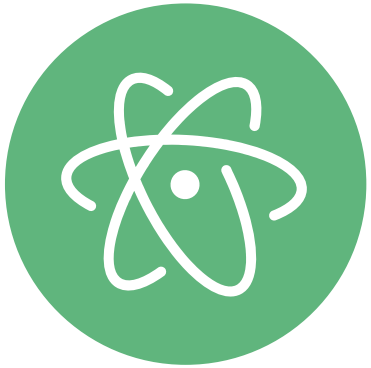


Github

Jupyter Notebook



Visual Studio Code



atom

10

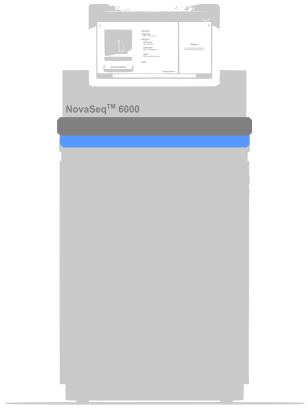
Be patient

Be patient

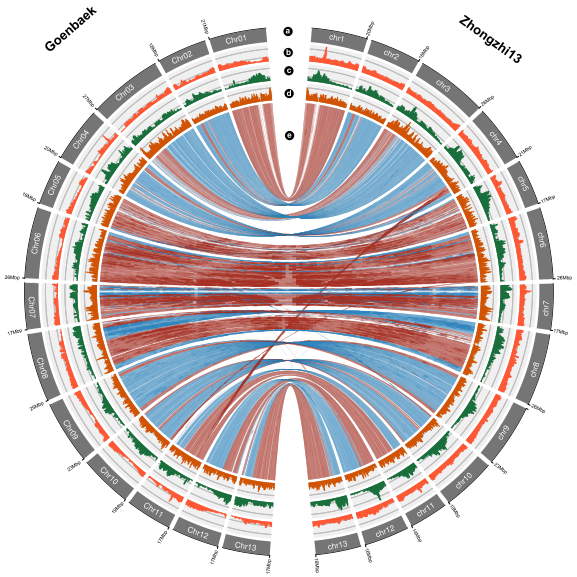
No 빨리빨리...
Be Patient

Bonus | Make a reference genome

RNA | DNA



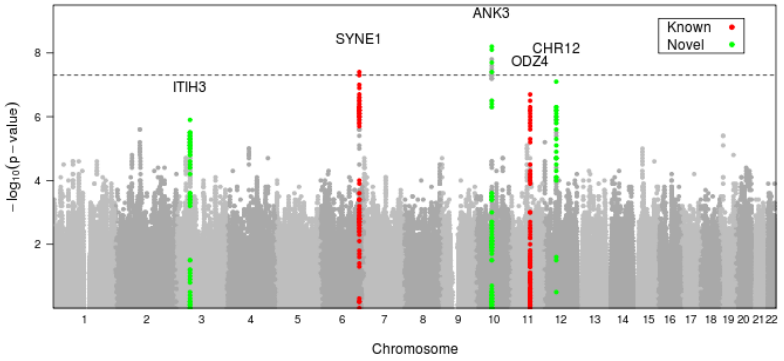
Genome assembly → Genome annotation



QTL MAPPING

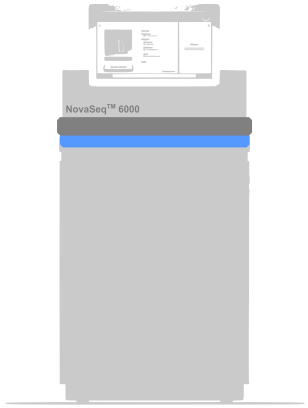
GWAS

POPULATION GENOMICS



Bonus | Application

DNA



E.g.: https://github.com/josieparis/gatk-snp-calling/

Acknowledgments

Brandies PA, Hogg CJ (2021) Ten simple rules for getting started with command-line bioinformatics. PLoS Comput Biol 17(2): e1008645.

https://doi.org/10.1371/journal.pcbi.1008645

감사합니다☺