

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



25th November, 2022

Ten rules to get started!



Terminology



Tool selection



Resources needed



Platform selection



Software installation



Scripting



Monitoring



File handling

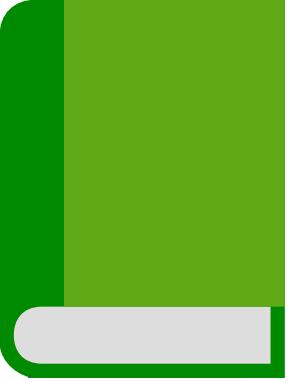


Take note again and again!



No 빨리빨리... Be Patient

1

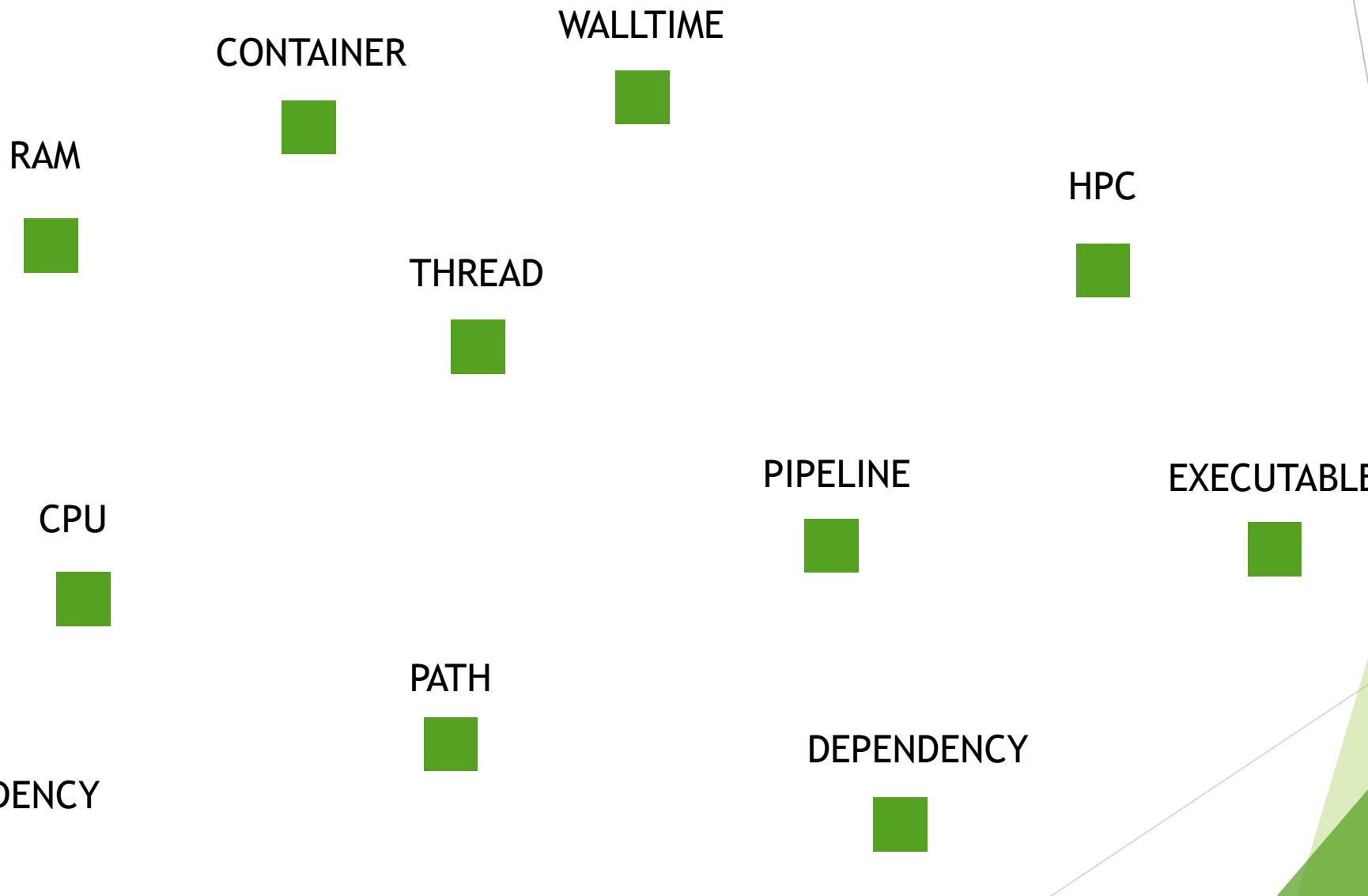


Terminology





Terminology

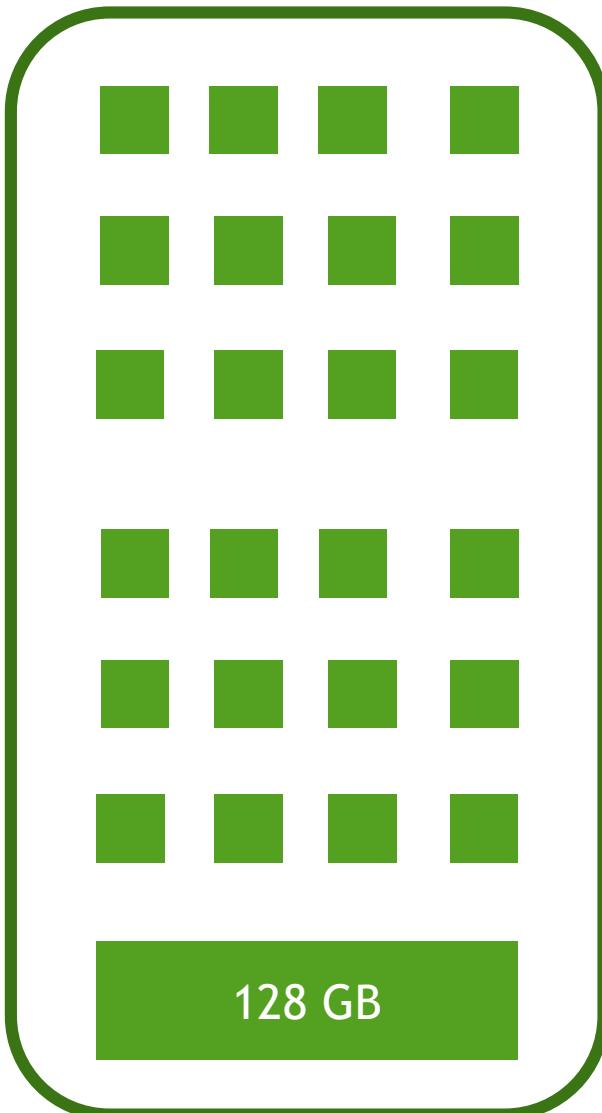




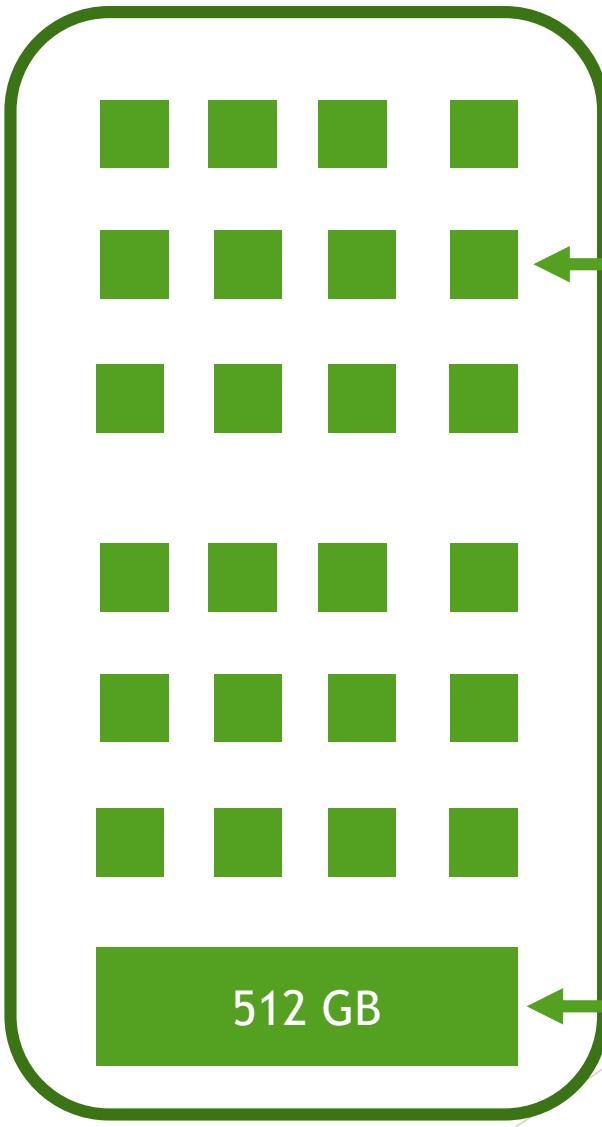
Terminology



STANDARD



HIGH-MEMORY

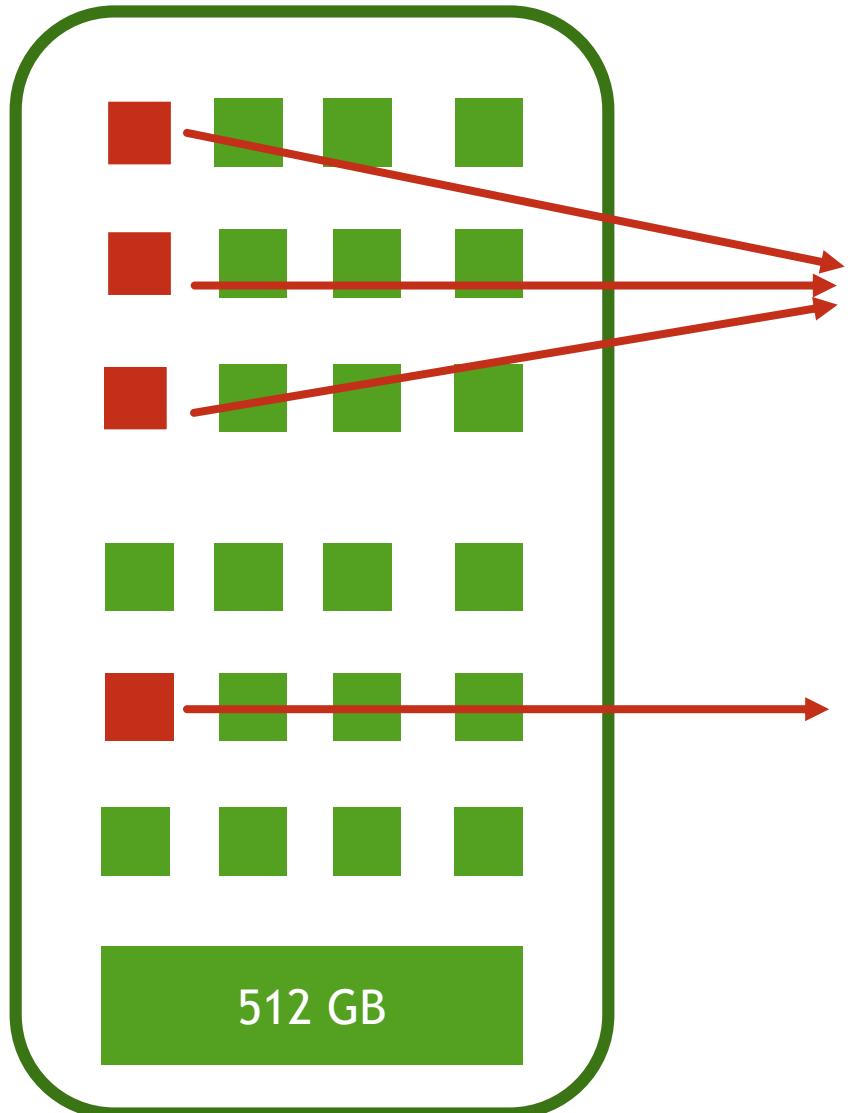




Terminology



MULTI-THREADED PROGRAM



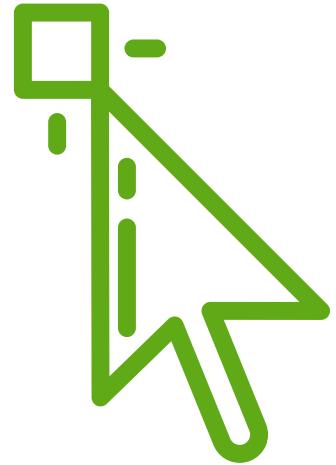
SINGLE-THREADED PROGRAM

OUTPUT

OUTPUT

512 GB

2



Tool selection





Tool selection

Know your data and assess your need

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



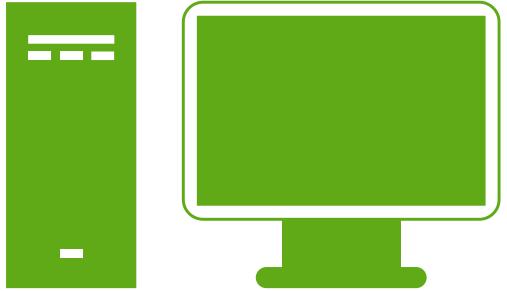


Tool selection

Know your data and assess your need - Tools available?

The screenshot displays two web pages side-by-side. The left page is the homepage of bio.tools (<https://bio.tools/>), featuring a search bar, a tool count of 27032, and a navigation menu with links like 'Explore', 'Login', and 'Sign-up'. Below the header, there's a message about cookie usage and a list of tools: ZygProb, ZWA, and ZT. Each tool has a brief description and a category bar below it. The right page is the homepage of long-read-tools.org (<https://long-read-tools.org/tools.html?sort=Name&cat=&tec=>), which includes a search bar, a cartoon character logo, and a large list of tools starting with 'A' (AbPOA, Acorde, AERON, Alfred, AlignQC, ALLPATHS-LG, Alpaca, Apollo, ARCS, ARTIC, AsmVar, Abrijn, Acorde, Albacore, AlignGraph2, Amortized-HMM, Architect, Arrow, ASHURE, Asemblytics). Both pages have a standard browser interface with tabs, back/forward buttons, and a top navigation bar.

3



Resources needed





Resources needed

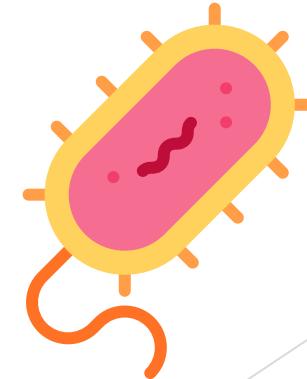
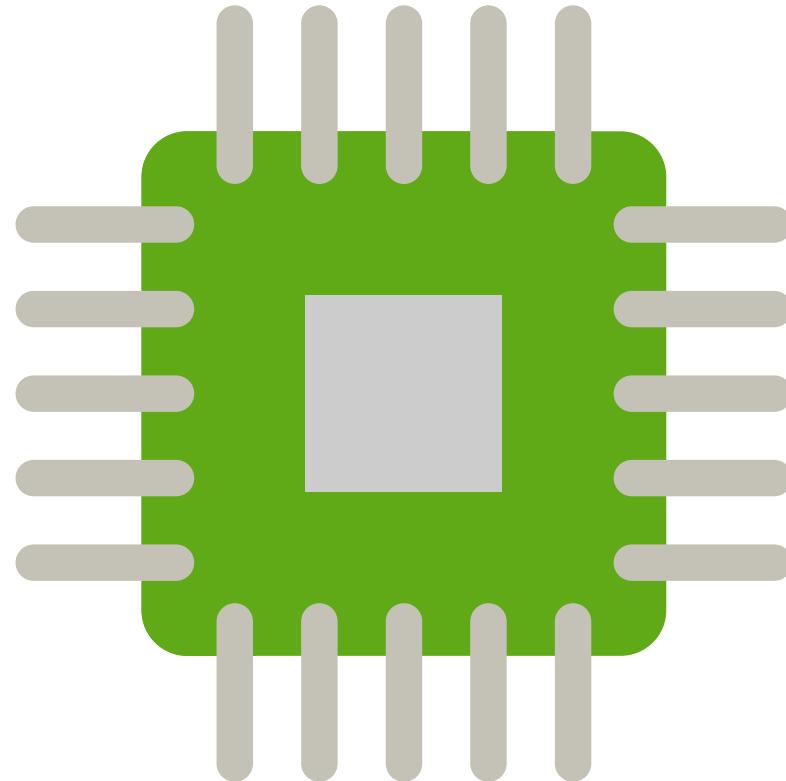
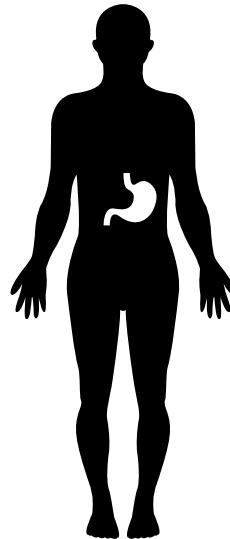


Estimating your computing requirements

- 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/>

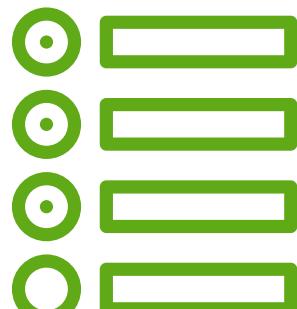
<https://aws.amazon.com/>

<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

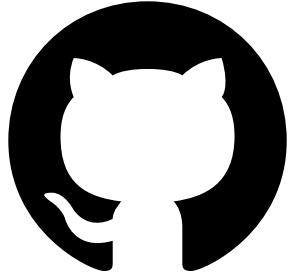




Scripting



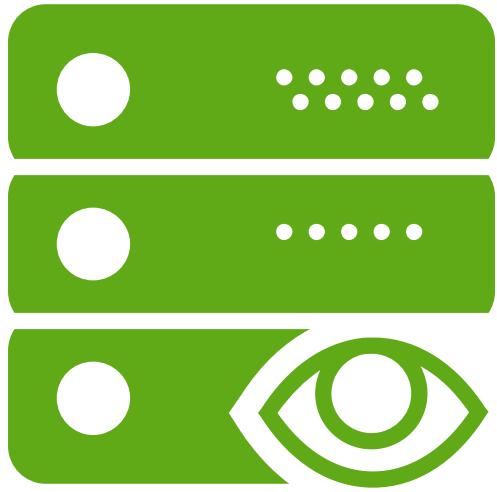
- Write your script
- Check syntax and minor scripting error
- Software well installed
- Dependencies well installed
- Resource sufficient?



<https://github.com/>



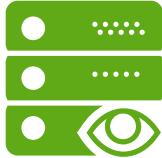
<https://snyk.io/code-checker/>



7

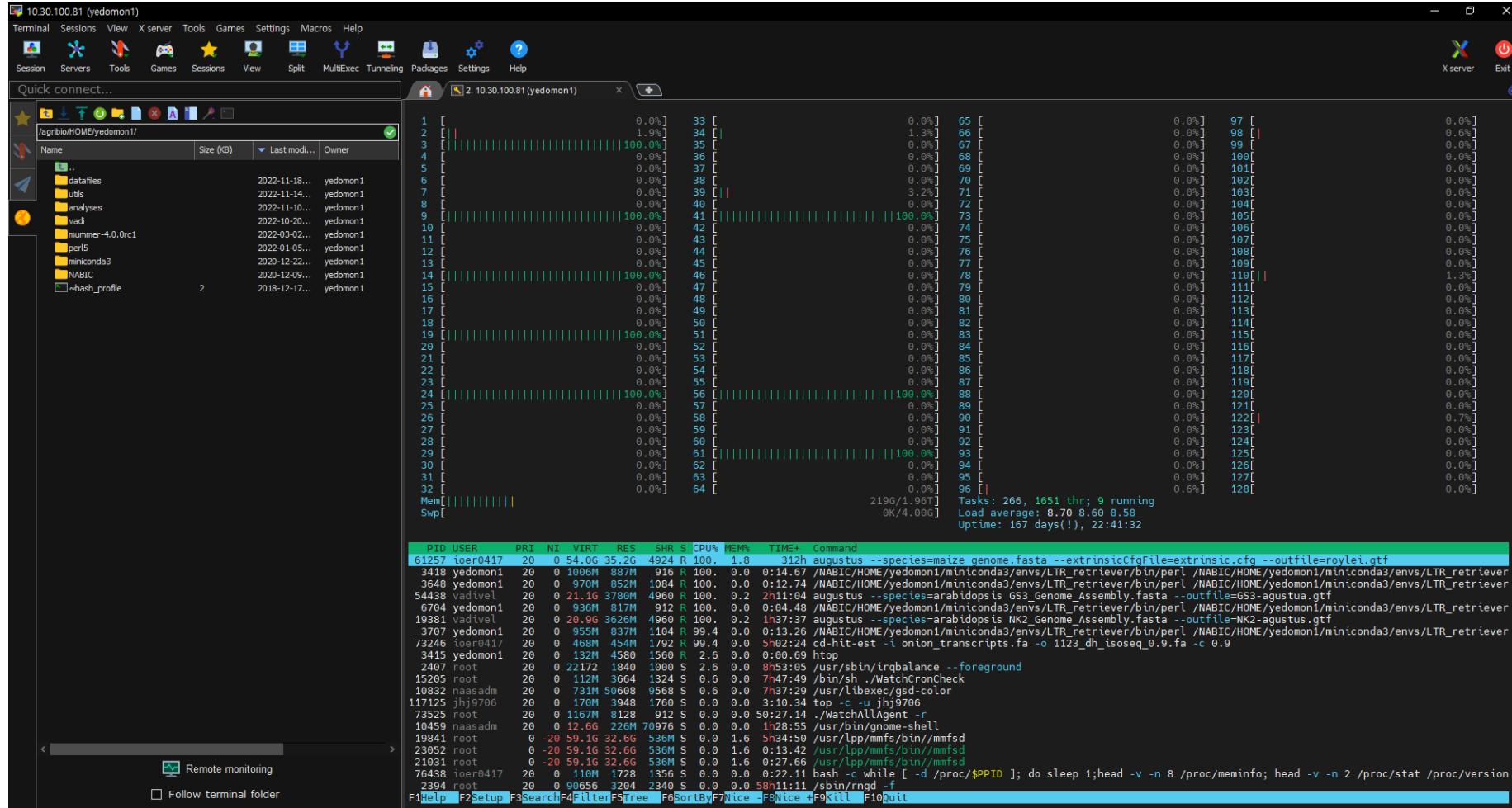
Monitoring

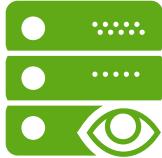




Monitoring

HTOP





Monitoring

TOP

10.30.100.81 (yedomon1)

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help

X server Exit

Quick connect... 2. 10.30.100.81 (yedomon1) +

/agribio/HOME/yedomon1/

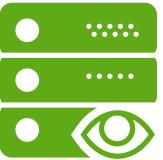
Name	Size (KB)	Last modi...	Owner
..		2022-11-18...	yedomon1
datafiles		2022-11-14...	yedomon1
utils		2022-11-10...	yedomon1
analyses		2022-10-20...	yedomon1
vadl		2022-03-02...	yedomon1
nummer-4.0rc1		2022-01-05...	yedomon1
perl5		2020-12-22...	yedomon1
miniconda3		2020-12-09...	yedomon1
NABIC		2018-12-17...	yedomon1
~bash_profile	2	2018-12-17...	yedomon1

```
top - 15:48:43 up 167 days, 22:44, 9 users, load average: 9.05, 8.73, 8.63
Tasks: 1316 total, 8 running, 1306 sleeping, 2 stopped, 0 zombie
%Cpu(s): 6.4 us, 0.5 sy, 0.0 ni, 93.1 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
KiB Mem : 2100800+total, 18472518+free, 99973424 used, 15357553+buff/cache
KiB Swap: 4194300 total, 4194300 free, 0 used. 18690917+avail Mem

PID USER PR NI VIRT RES SHR S %CPU %MEM TIME+ COMMAND
19381 vadivel 20 0 20.9g 3.6g 4960 R 100.0 0.2 100:08.41 augustus --species=arabidopsis NK2_Genome_Assembly.fasta --outfile=NK2-agustus.gtf
23328 yedomon1 20 0 984944 863548 1068 R 100.0 0.0 0:14.05 [/NABIC/HOME/yedomon1/miniconda3/envs/LTR_retriever/bin/perl /NABIC/HOME/yedomon1/miniconda3/envs/LTR_re+
54438 vadivel 20 0 21.1g 3.7g 4960 R 100.0 0.2 133:35.40 augustus --species=arabidopsis G3_Genome_Assembly.fasta --outfile=G3-agustus.gtf
61257 ioe0417 20 0 54.0g 35.2g 4924 R 100.0 1.8 1873:37.38 augustus --species=maize genome.fasta --extrinsiccfgFile=extrinsic.cfg --outfile=roylei.gtf
73246 ioe0417 20 0 479704 465592 1792 R 100.0 0.0 304:56.07 cd-hit-est -i onion_transcripts.fa -o 1123_dh_isoseq_0.9.fa -c 0.9
23282 yedomon1 20 0 981124 859884 1276 S 99.3 0.0 0:15.16 [/NABIC/HOME/yedomon1/miniconda3/envs/LTR_retriever/bin/perl /NABIC/HOME/yedomon1/miniconda3/envs/LTR_re+
23693 yedomon1 20 0 955408 833776 912 R 99.3 0.0 0:03.74 [/NABIC/HOME/yedomon1/miniconda3/envs/LTR_retriever/bin/perl /NABIC/HOME/yedomon1/miniconda3/envs/LTR_re+
23780 yedomon1 20 0 944388 822884 912 R 65.0 0.0 0:02.00 [/NABIC/HOME/yedomon1/miniconda3/envs/LTR_retriever/bin/perl /NABIC/HOME/yedomon1/miniconda3/envs/LTR_re+
73525 root 20 0 1195564 8124 912 S 2.6 0.0 50:27.69 ./WatchAllAgent -r
2394 root 20 0 90656 3204 2340 S 2.3 0.0 3491:12 /sbin/rngd -f
15205 root 20 0 115356 3664 1324 S 0.7 0.0 467:05.53 /bin/sh ./WatchCronCheck
19841 root 0 -20 59.1g 32.6g 549176 S 0.7 1.6 334:51.70 [/usr/lpp/mmfs/bin//mmfsd
23811 yedomon1 20 0 174200 3720 1664 R 0.7 0.0 0:00.09 top
10 root 20 0 0 0 0 0 0.3 0.0 104:11.50 [rcu_sched]
1321 ioe0417 20 0 187468 3308 1508 S 0.3 0.0 0:06.41 sshd: ioe0417@pts/18
10832 naasadm 20 0 748944 50608 9568 S 0.3 0.0 457:03.03 [/usr/libexec/gsd-color
44664 root 20 0 0 0 0 0 0.3 0.0 0:00.77 [kworker/u256:0]
46244 jhh9706 20 0 113428 1756 1368 S 0.3 0.0 0:00.98 bash -while [ -d /proc/$$PPID ]; do sleep 1;head -v -n 8 /proc/meminfo; head -v -n 2 /proc/stat /proc+
117125 jhh9706 20 0 174212 3948 1760 S 0.3 0.0 3:11.54 top -c -u jhh9706
127023 thlee00 20 0 113428 1664 1368 S 0.3 0.0 0:00.22 bash -while [ -d /proc/$$PPID ]; do sleep 1;head -v -n 8 /proc/meminfo; head -v -n 2 /proc/stat /proc+
1 root 20 0 191656 4636 2568 S 0.0 0.0 25:17.17 [/usr/lib/systemd/systemd --switched-root --system --deserialize 22
2 root 20 0 0 0 0 0 0.0 0.0 0:00.69 [kthreadd]
3 root 20 0 0 0 0 0 0.0 0.0 0:09.09 [ksoftirqd/0]
8 root rt 0 0 0 0 0 0.0 0.0 0:03.24 [migration/0]
9 root 20 0 0 0 0 0 0.0 0.0 0:00.00 [rcu_bh]
11 root rt 0 0 0 0 0 0.0 0.0 0:20.64 [watchdog/0]
12 root rt 0 0 0 0 0 0.0 0.0 0:19.12 [watchdog/1]
13 root rt 0 0 0 0 0 0.0 0.0 0:04.78 [migration/1]
14 root 20 0 0 0 0 0 0.0 0.0 0:07.79 [ksoftirqd/1]
17 root rt 0 0 0 0 0 0.0 0.0 0:18.36 [watchdog/2]
18 root rt 0 0 0 0 0 0.0 0.0 0:04.39 [migration/2]
19 root 20 0 0 0 0 0 0.0 0.0 0:05.05 [ksoftirqd/2]
21 root 0 -20 0 0 0 0 0.0 0.0 0:00.00 [kworker/2:0H]
22 root rt 0 0 0 0 0 0.0 0.0 0:18.21 [watchdog/3]
23 root rt 0 0 0 0 0 0.0 0.0 0:04.33 [migration/3]
24 root 20 0 0 0 0 0 0.0 0.0 0:03.95 [ksoftirqd/3]
26 root 0 -20 0 0 0 0 0.0 0.0 0:00.00 [kworker/3:0H]
27 root rt 0 0 0 0 0 0.0 0.0 0:19.03 [watchdog/4]
28 root rt 0 0 0 0 0 0.0 0.0 0:04.03 [migration/4]
29 root 20 0 0 0 0 0 0.0 0.0 0:06.80 [ksoftirqd/4]
31 root 0 -20 0 0 0 0 0.0 0.0 0:00.00 [kworker/4:0H]
32 root rt 0 0 0 0 0 0.0 0.0 0:18.72 [watchdog/5]
33 root rt 0 0 0 0 0 0.0 0.0 0:02.69 [migration/5]
34 root 20 0 0 0 0 0 0.0 0.0 0:03.66 [ksoftirqd/5]
36 root 0 -20 0 0 0 0 0.0 0.0 0:00.00 [kworker/5:0H]
37 root rt 0 0 0 0 0 0.0 0.0 0:17.38 [watchdog/6]
38 root rt 0 0 0 0 0 0.0 0.0 0:03.31 [migration/6]
39 root 20 0 0 0 0 0 0.0 0.0 0:03.12 [ksoftirqd/6]
41 root 0 -20 0 0 0 0 0.0 0.0 0:00.00 [kworker/6:0H]
42 root rt 0 0 0 0 0 0.0 0.0 0:16.97 [watchdog/7]
43 root rt 0 0 0 0 0 0.0 0.0 0:04.16 [migration/7]
44 root 20 0 0 0 0 0 0.0 0.0 0:02.18 [ksoftirqd/7]
```

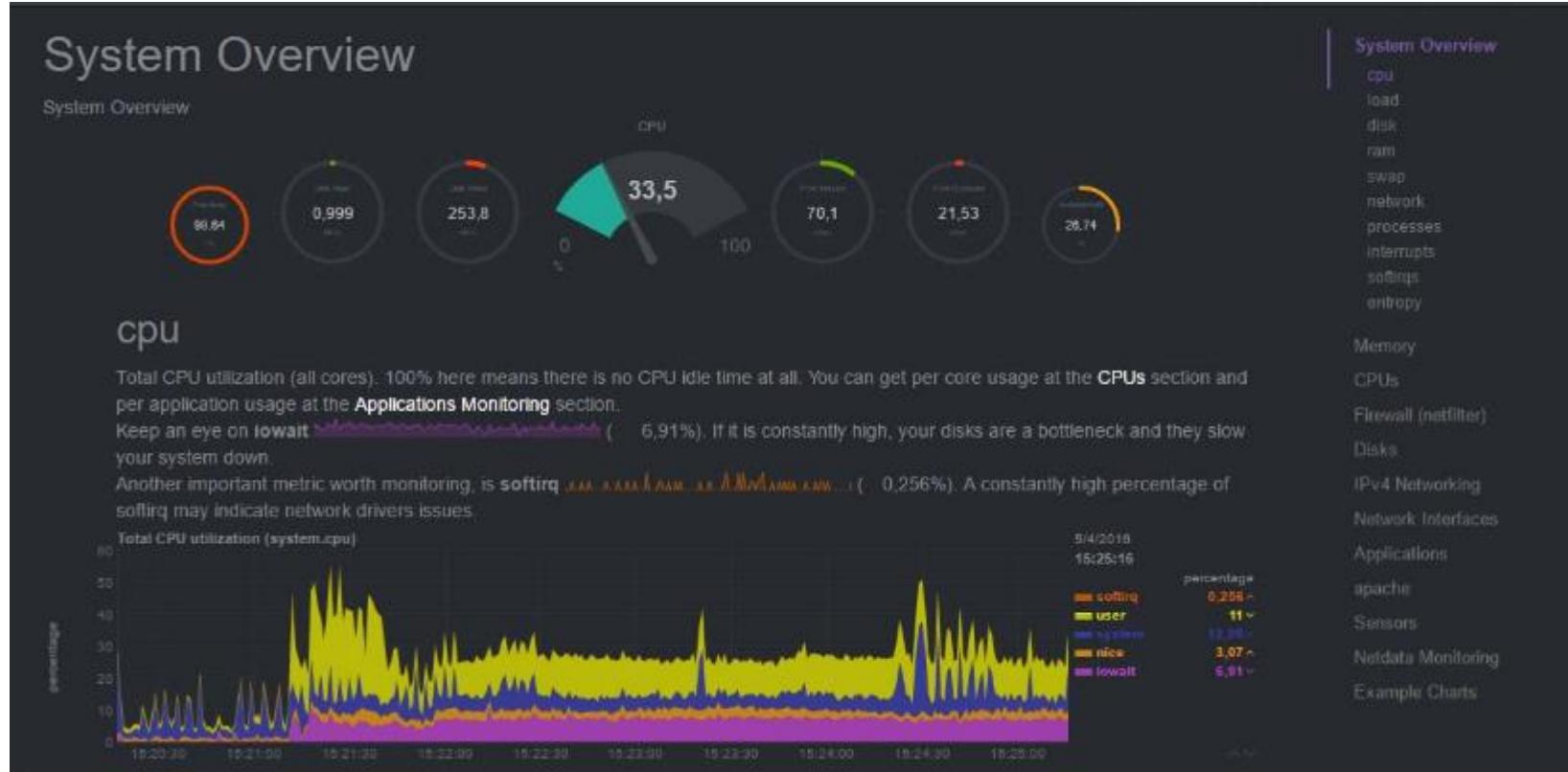
< >

Remote monitoring Follow terminal folder



Monitoring

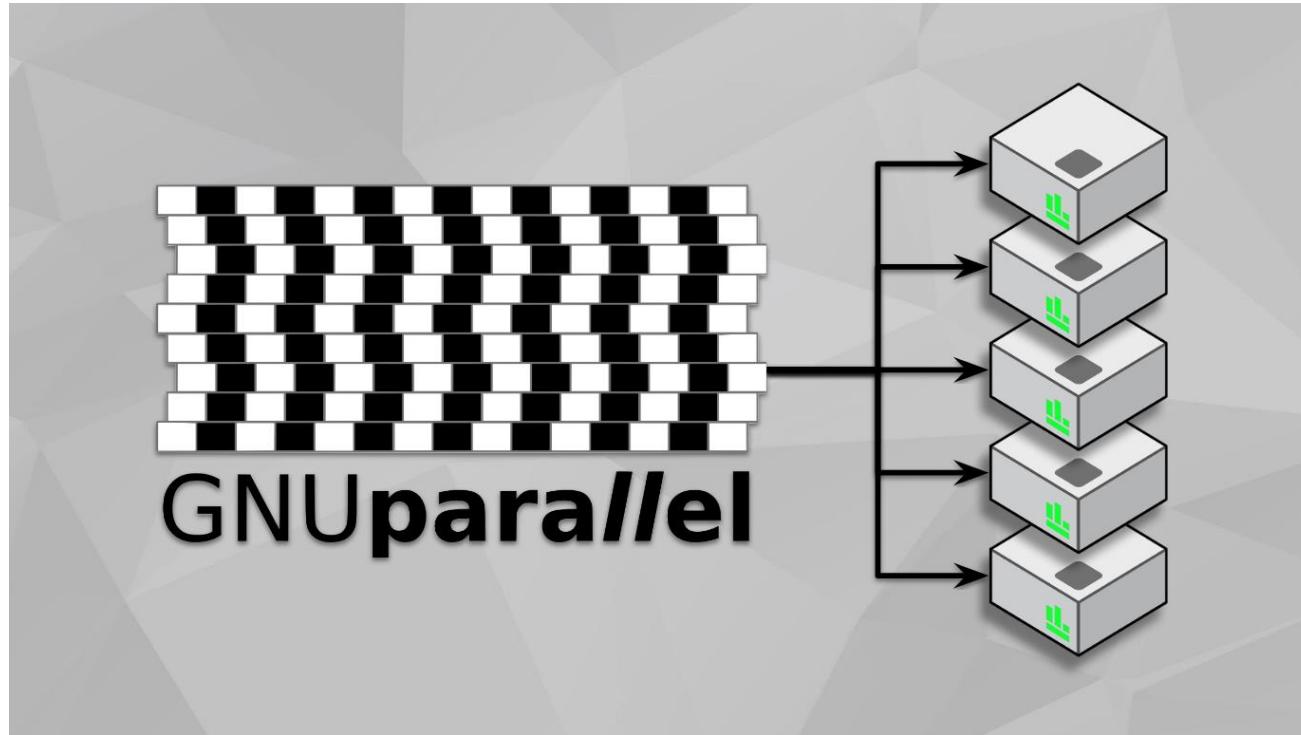
NETDATA





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`

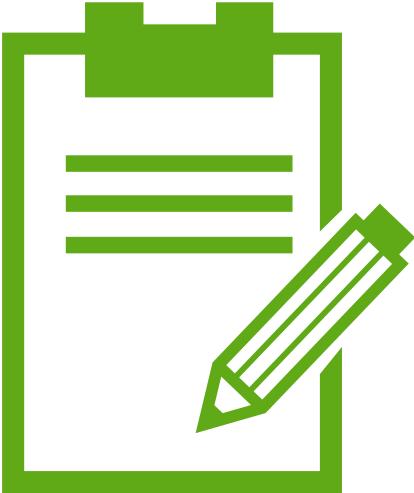


Sed

Find & Replace
`sed 's/HTU4257.1/chr1/g' file`

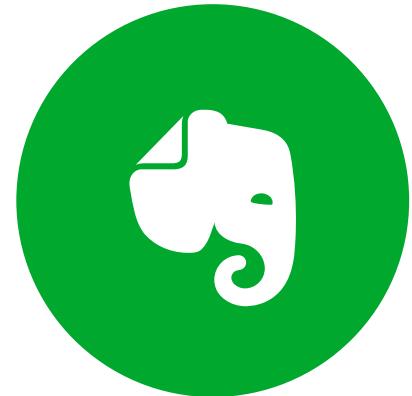
9

Note taking

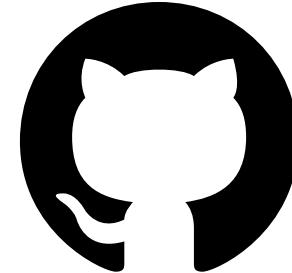




Note taking



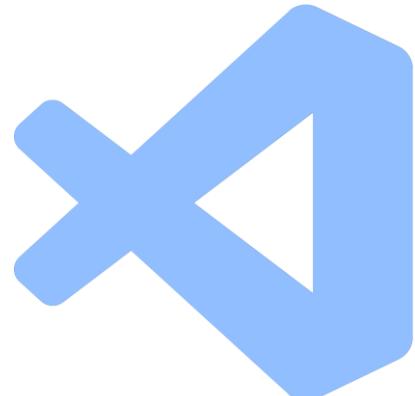
Evernote



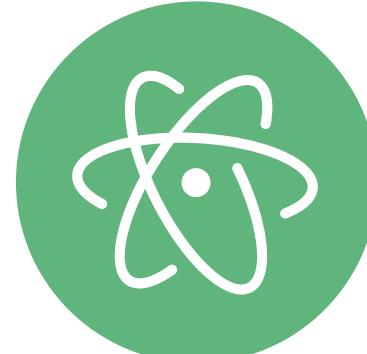
Github



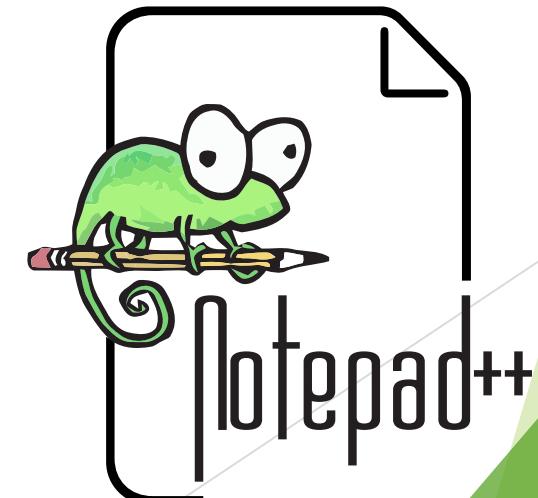
Jupyter Notebook



Visual Studio Code



atom



10



Be patient



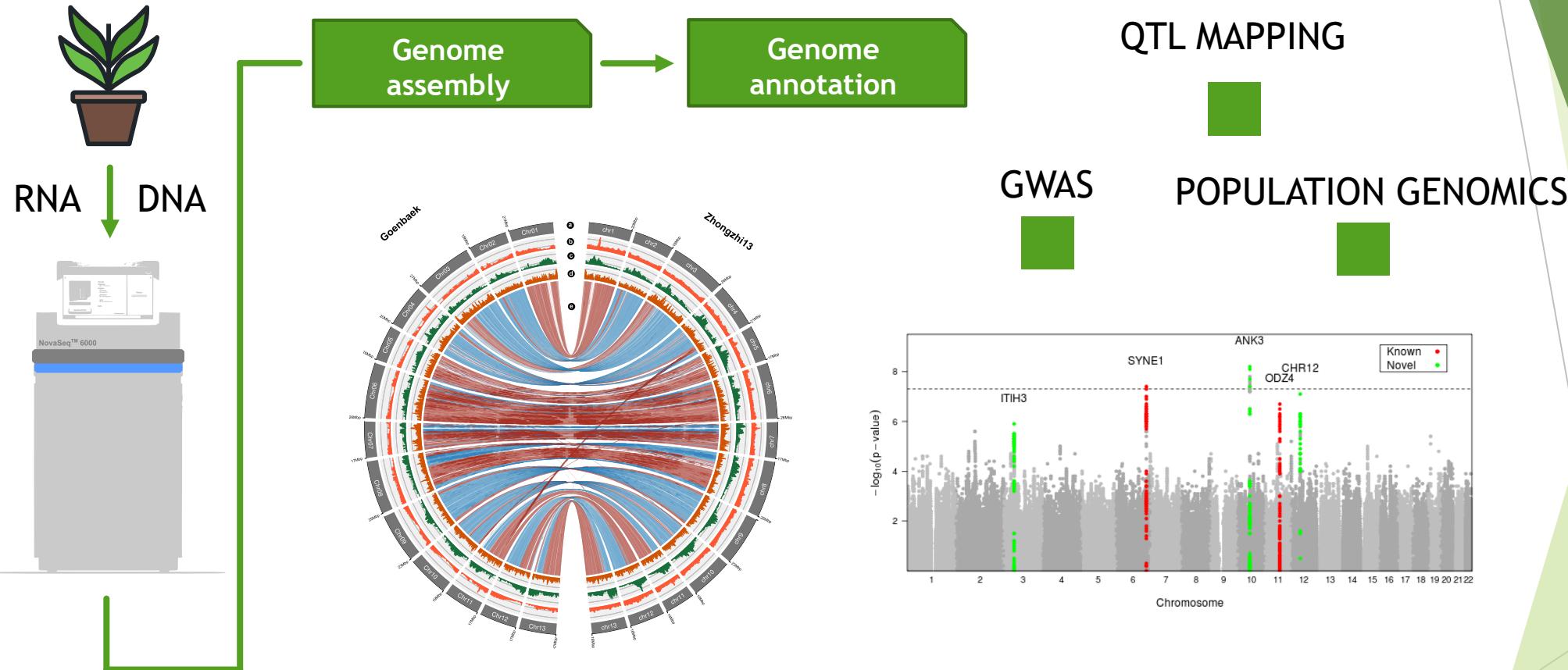


Be patient

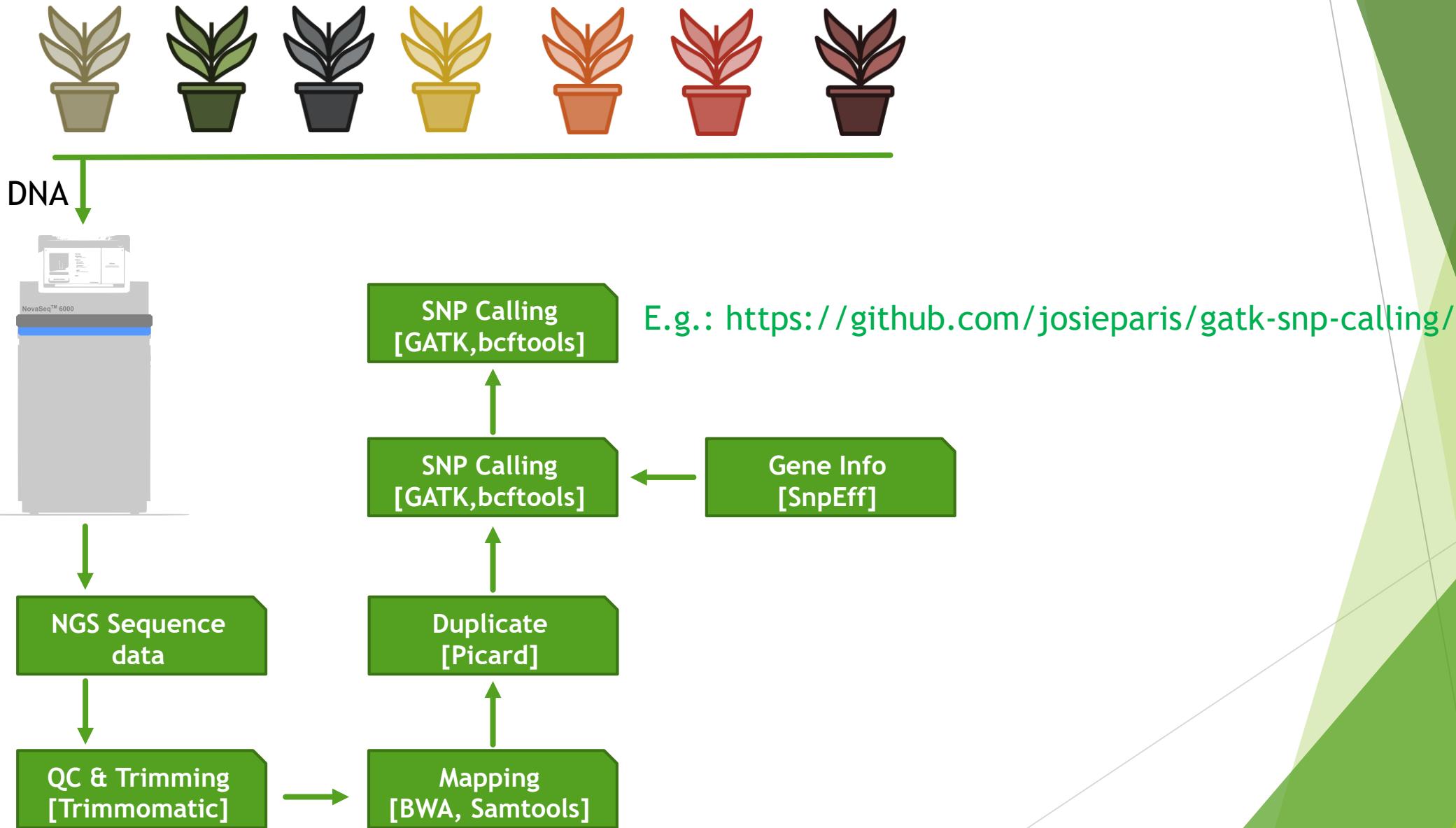


No 빨리빨리...
Be Patient

Bonus | Make a reference genome



Bonus | Application



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😊