Introduction of Bioinformatics

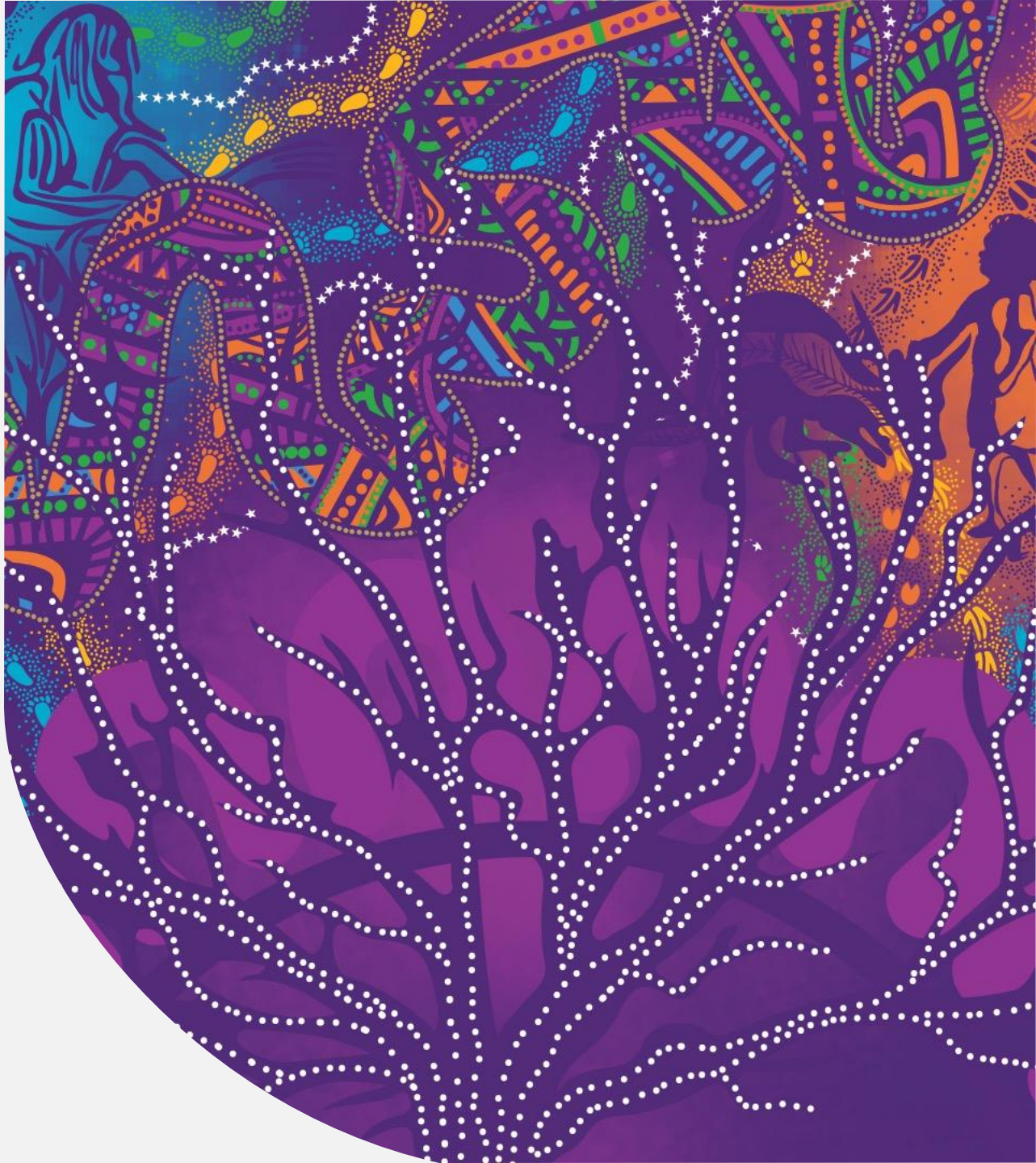
QAAFI Student Association

Acknowledgment of **Country**

The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.



QAAFI
STUDENT ASSOCIATION

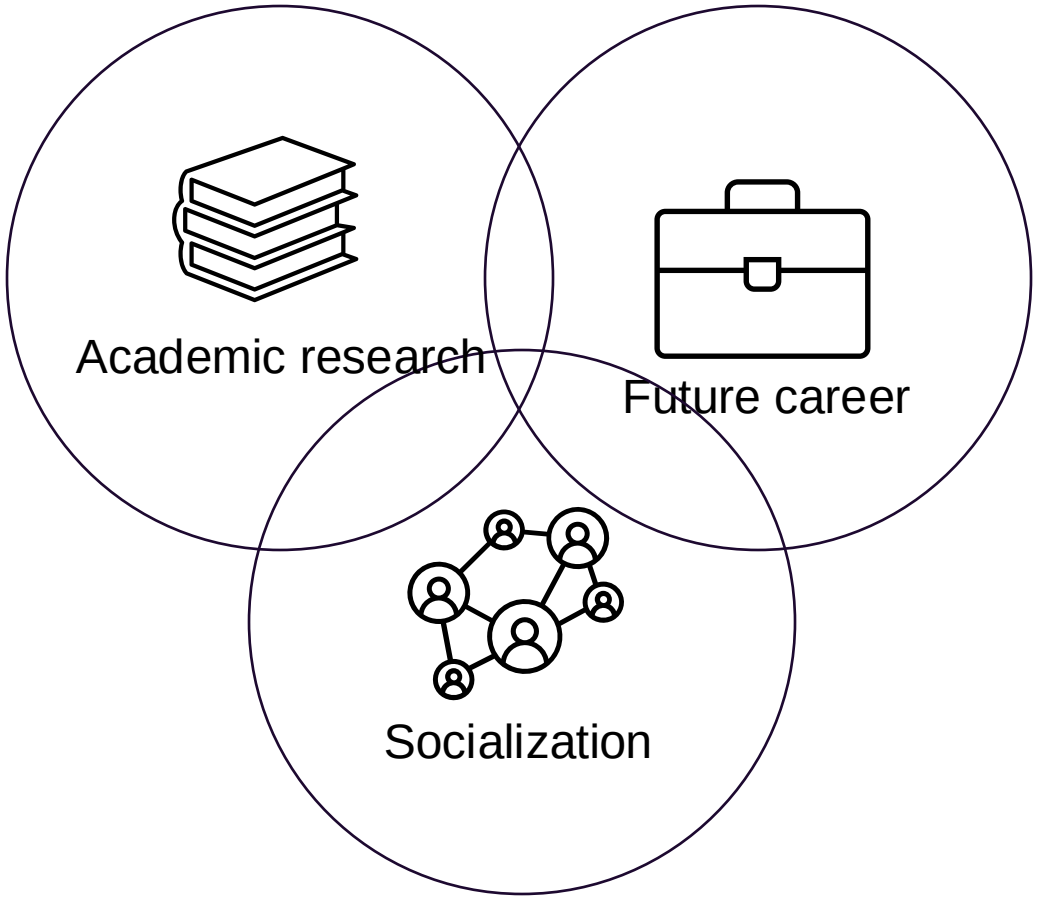
Enriching the professional and social experience of QAAFI students

As HDRs, we face challenges around...

Research – Life balance

Mental wellbeing

Employment Security



This year is a big year for us, we have achieved...

Seminar: How to ACE your PhD?

Market day at St Lucia & Gatton

Launch of QSA branded Merch

Get in touch and stay up to date!

QAAFI STUDENT ASSOCIATION

qsa@uq.edu.au

@qsaqaafi

LinkedIn

Introduction of Bioinformatics

QAAFI Student Association

Ziming Chen

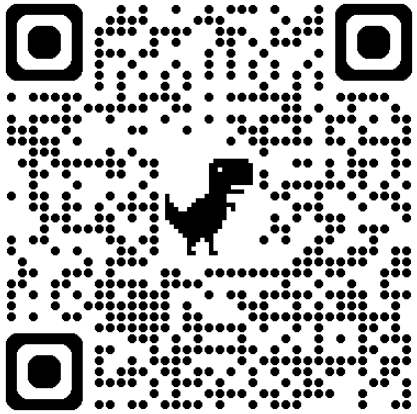
Third-year PhD Student

Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation,

The University of Queensland, St Lucia, QLD, Australia

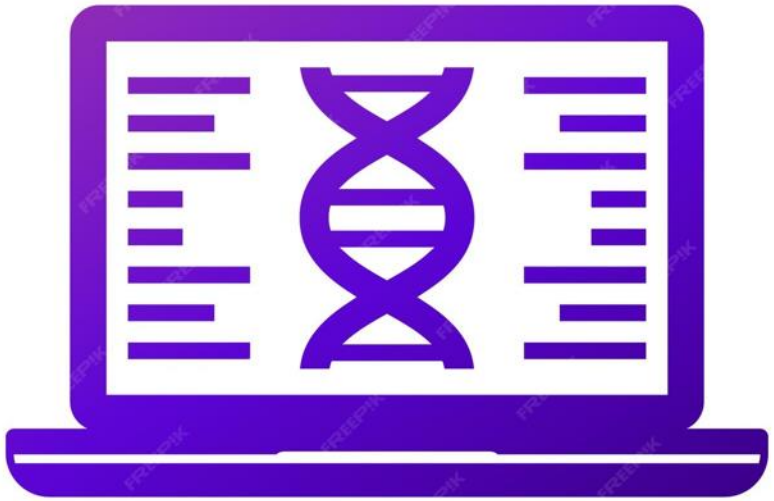
Workshop Timeline

Time	Topic
9:10 - 9:35	Concepts in Bioinformatics
9:35 - 10:00	Introduction of High-Performing Computing
10:00 - 10:30	Linux Command Lines
10:30 - 10:45	Break
10:45 - 11:15	Linux Command Lines
11:15 – 11:45	Analysis Workflows
11:45 - 12:00	Wrap-up
12:00 - 13:00	Lunch Time



Overview

- What is **bioinformatics**?
- **What** can you do with bioinformatics?
- What is **High-Performing Computing**?
- Why do we use **High-Performing Computing**?
- What is the **Command Line interface**?
- What do you need to **consider** during **analysis**?

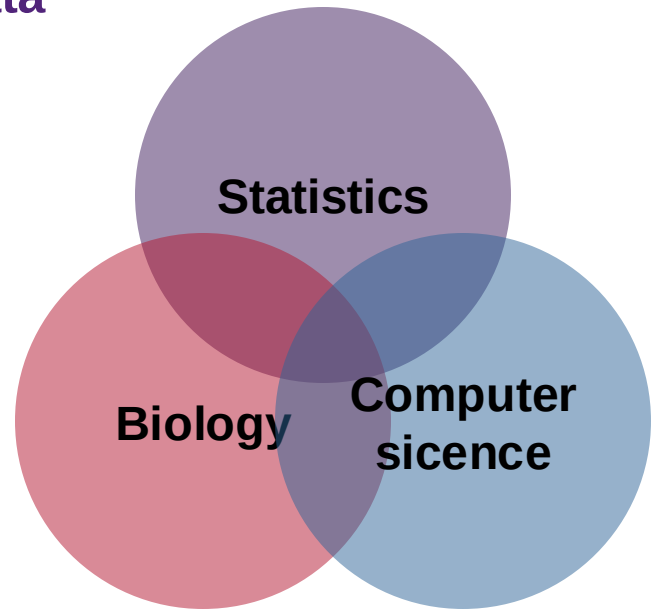


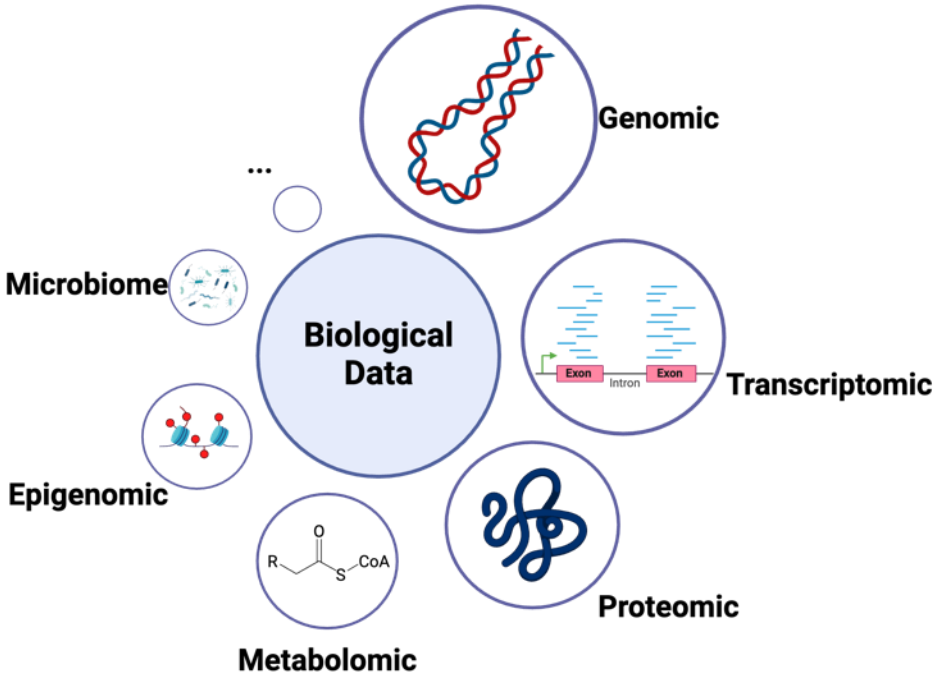
Bioinformatics

Bioinformatics

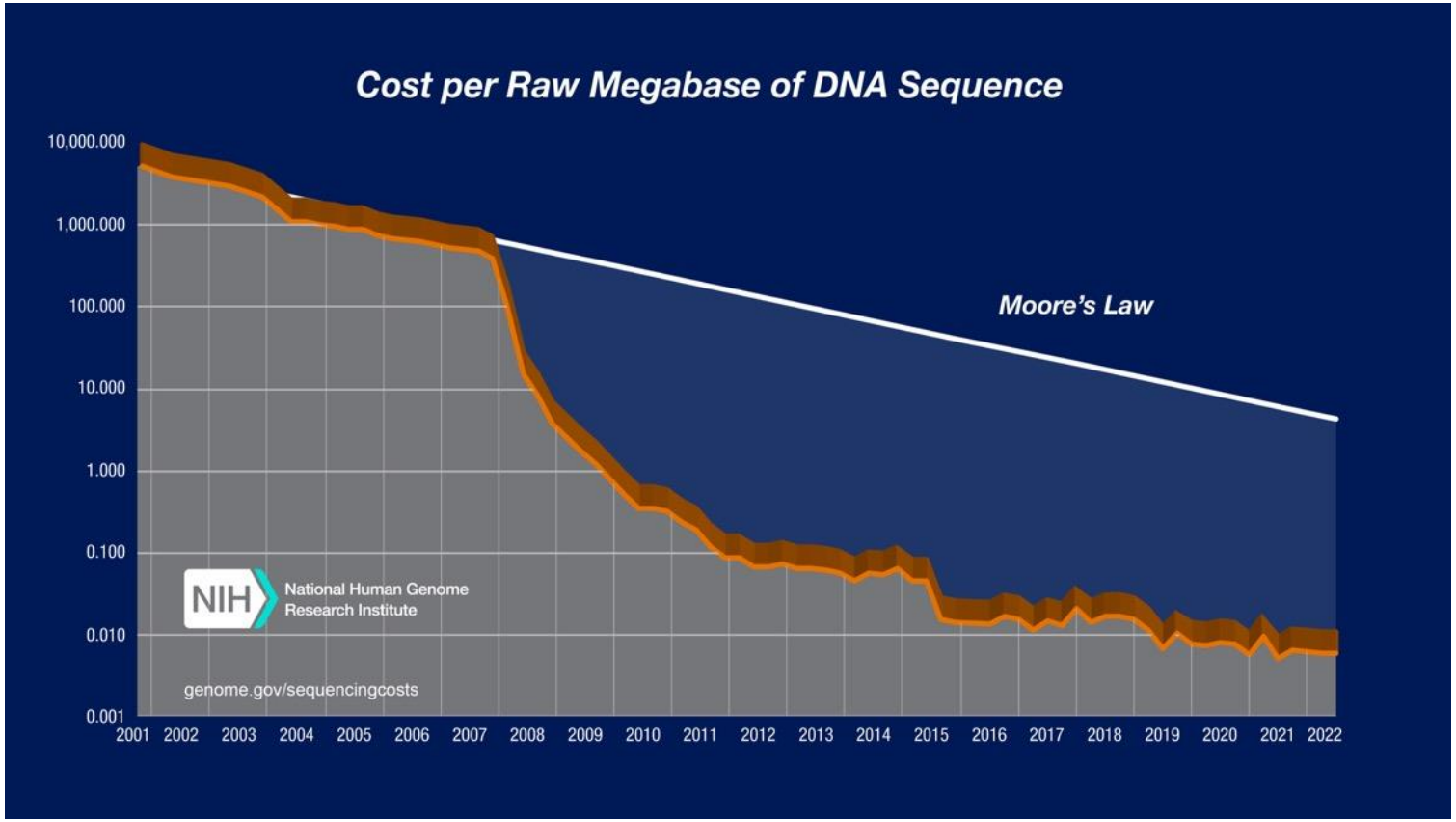
An **interdisciplinary** field that **analyses** and **interprets biological data**.

- Biology background
- Computing and statistics
- **Biological data**





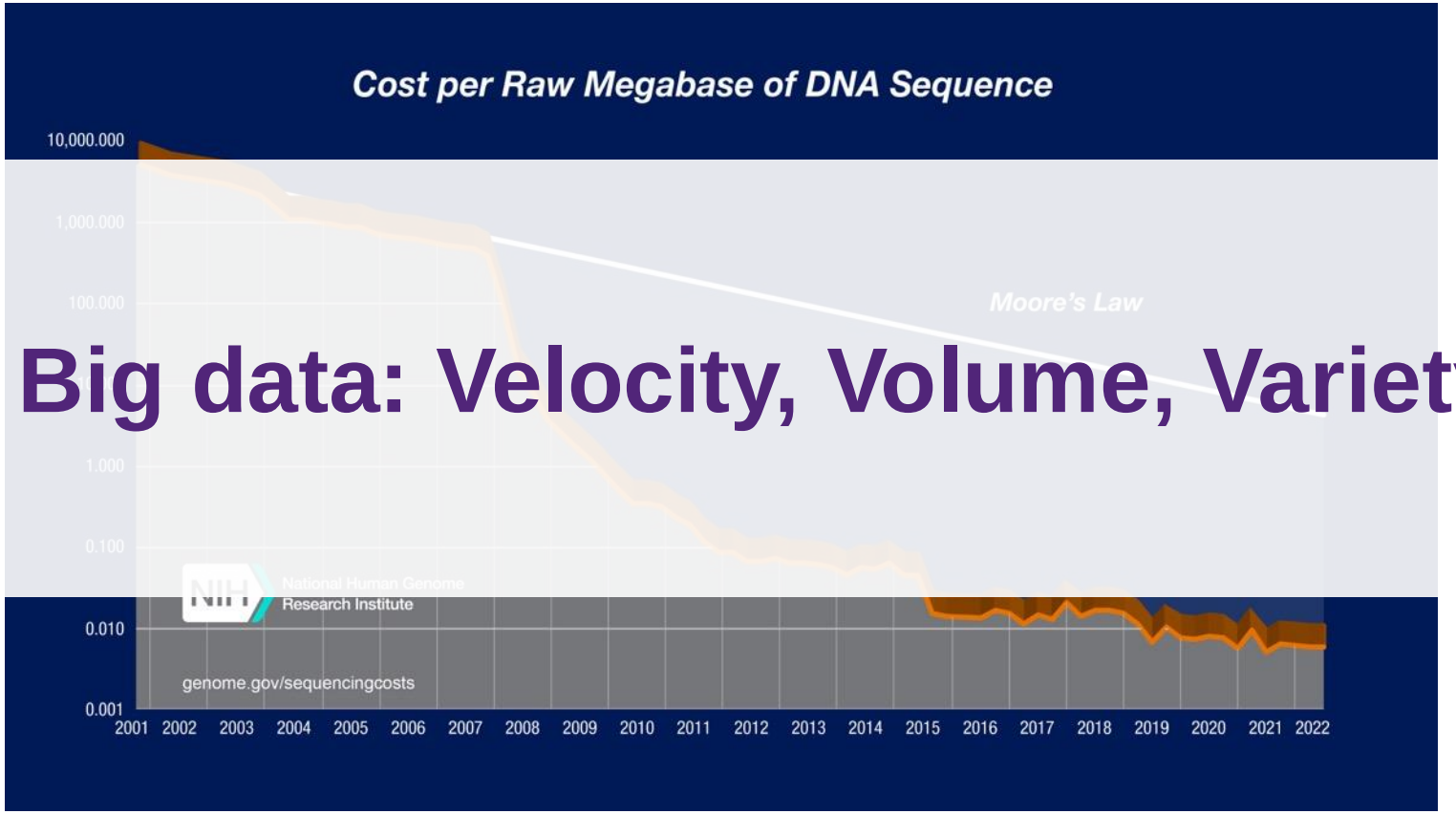
Lower Cost



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

DOI:https://doi.org/10.4137/BII.S31559

Lower Cost



Big data: Velocity, Volume, Variety

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

DOI:https://doi.org/10.4137/BII.S31559

Velocity - Increasing Speed

Oxford Nanopore DNA Sequencing by MinION device

- **Portable**
- **Real-time data**
- Long read

Velocity - Increasing Speed

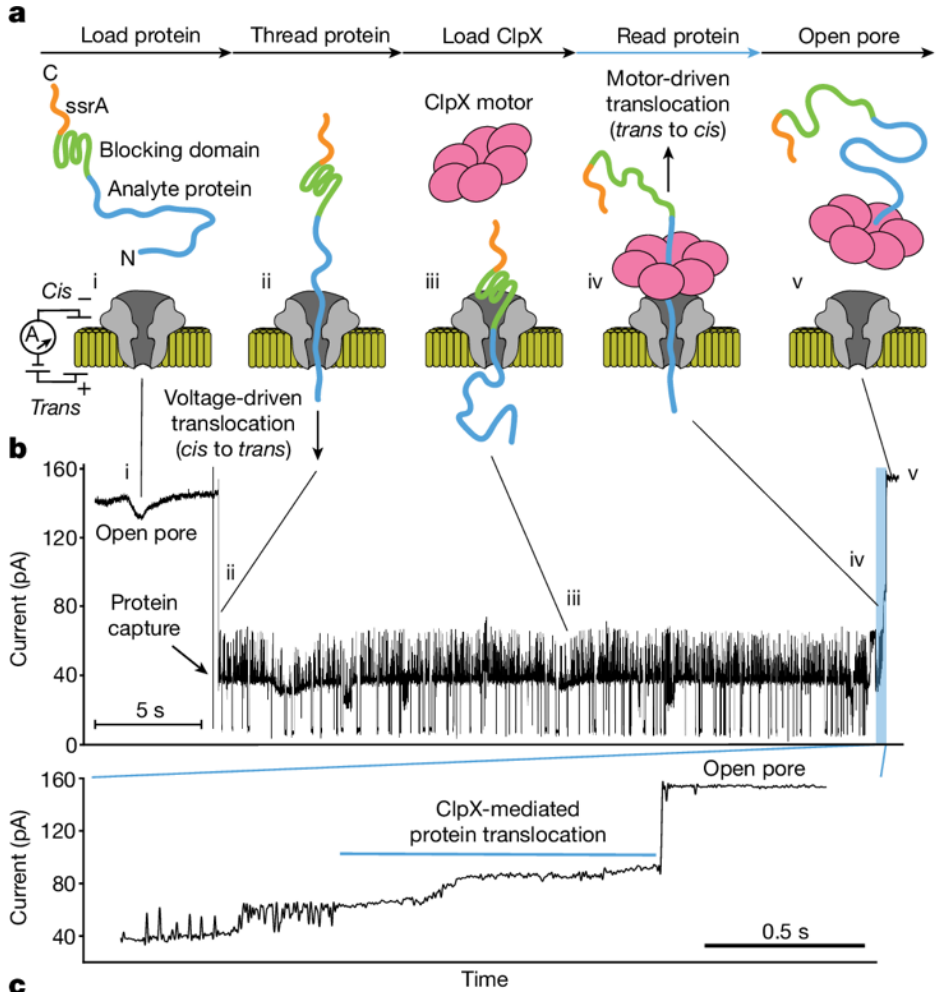
Article | Open access | Published: 11 September 2024

Multi-pass, single-molecule nanopore reading of long protein strands

Keisuke Motone, Daphne Kontogiorgos-Heintz, Jasmine Wee, Kyoko Kurihara, Sangbeom Yang, Gwendolin Roote, Oren E. Fox, Yishu Fang, Melissa Queen, Mattias Tolhurst, Nicolas Cardozo, Miten Jain & Jeff Nivala

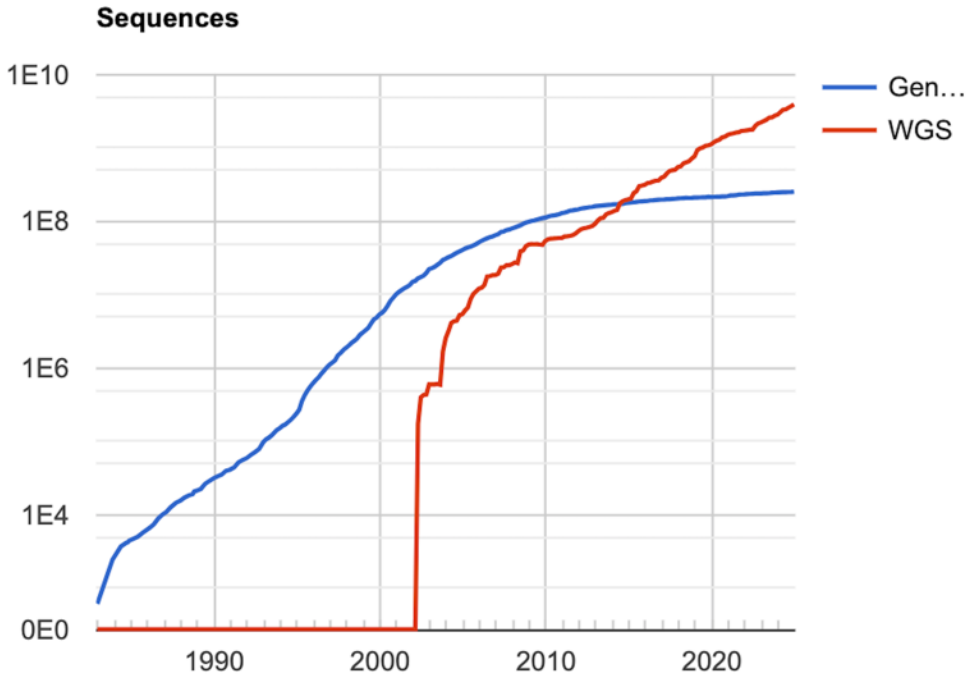
Nature **633**, 662–669 (2024) | Cite this article

57k Accesses | **12** Citations | **348** Altmetric | Metrics

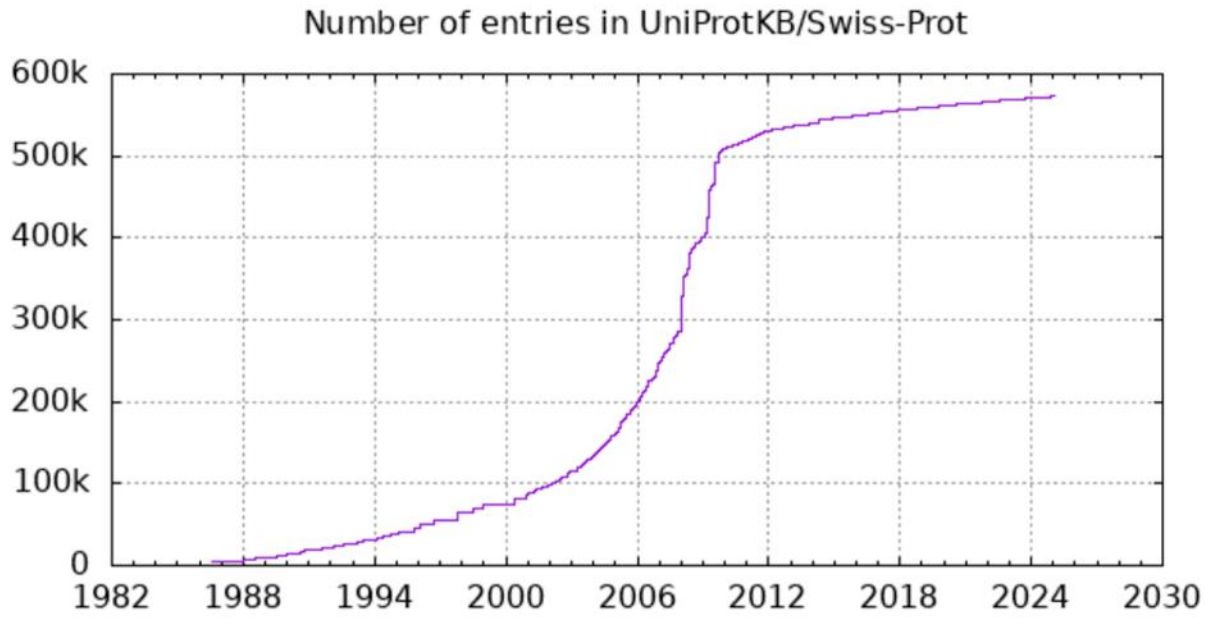


DOI:https://doi.org/10.1038/s44222-024-00260-8

Volume - Increasing Size

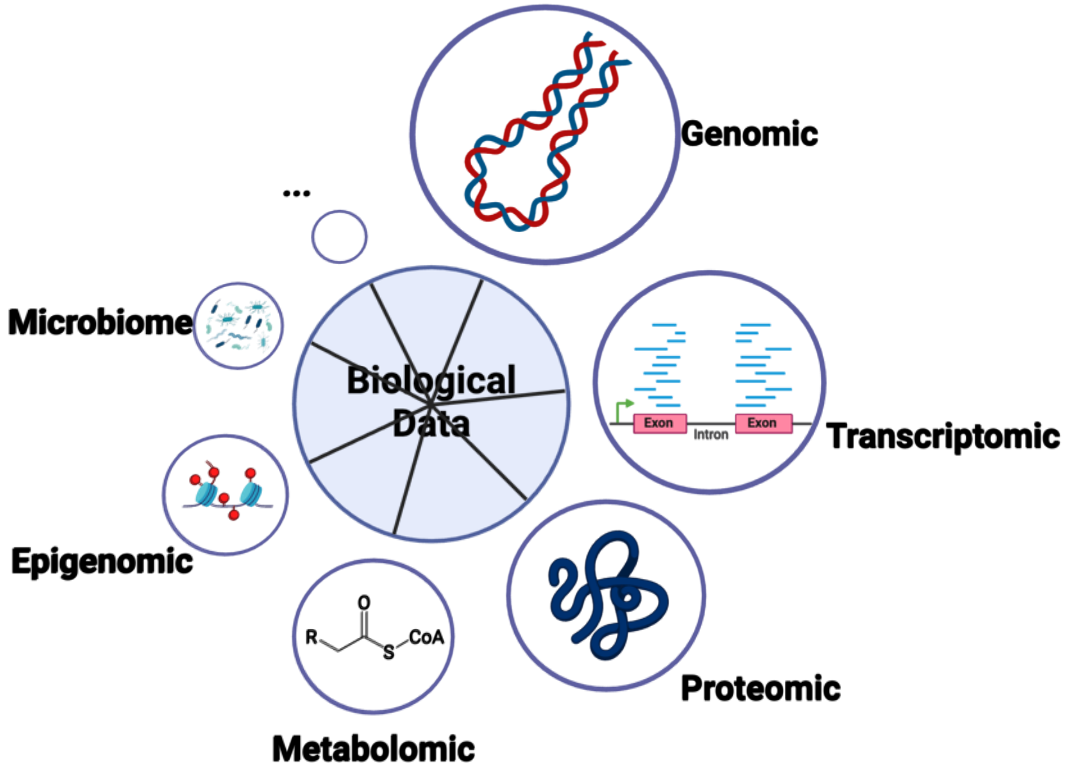


GenBank Sequence Database

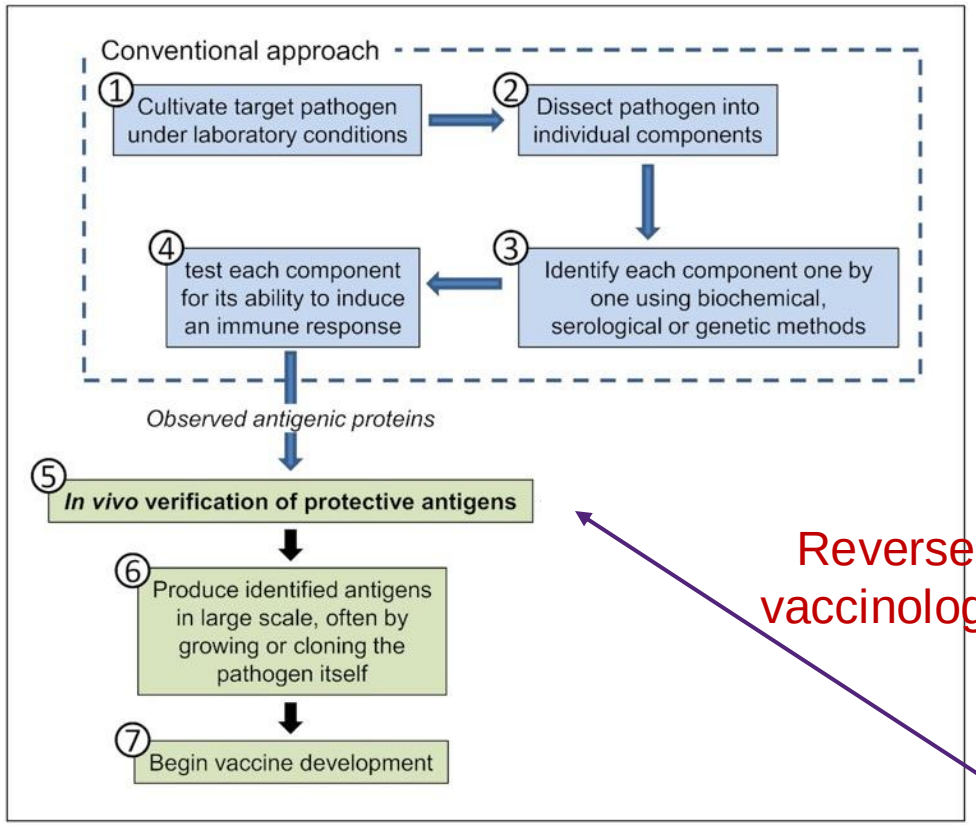


UniProtKB/Swiss-Prot Protein Database

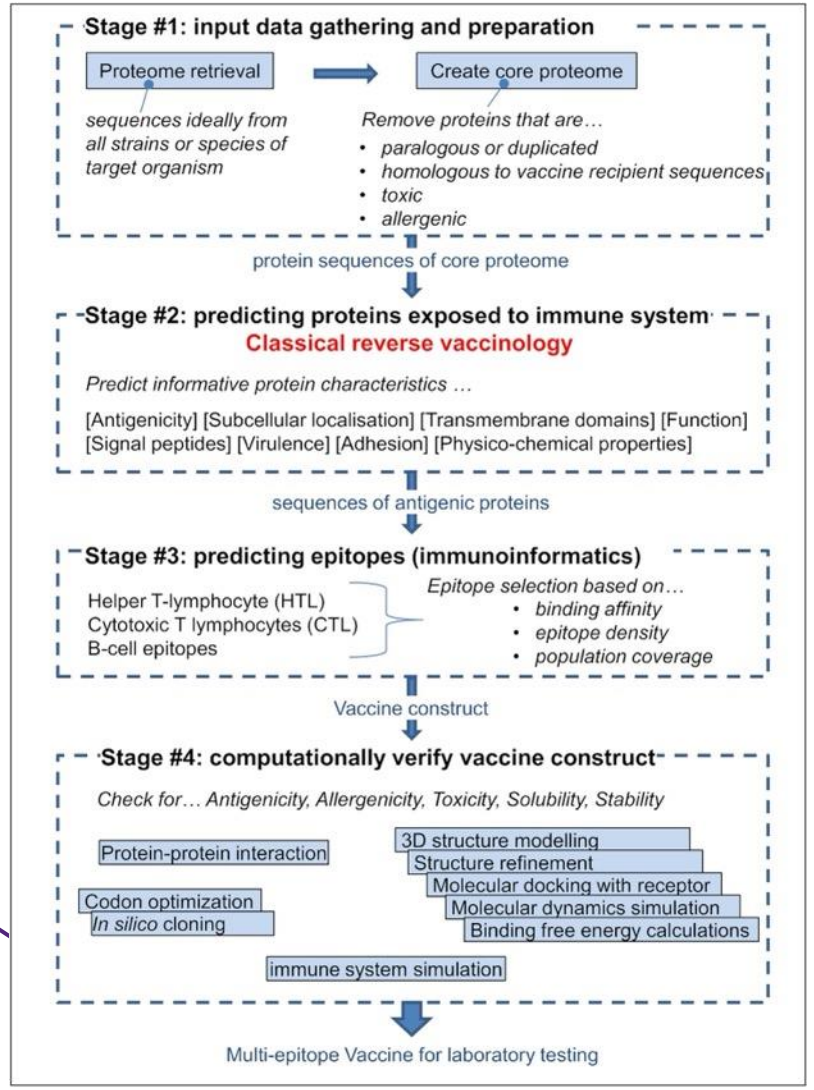
Variety - Increasing Complexity



Can Proteomic Data Accelerate the Vaccine Development?

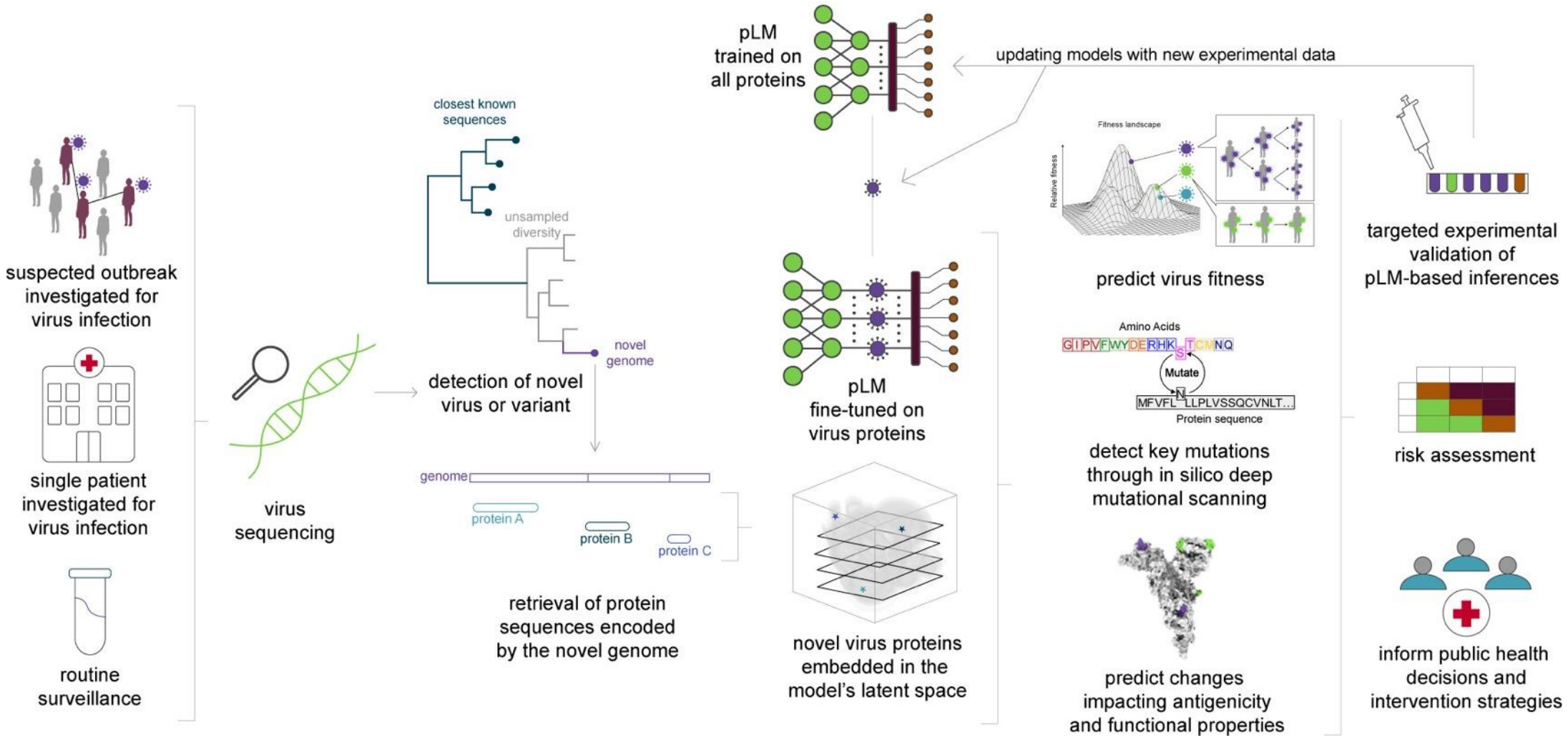


Reverse vaccinology



DOI: https://doi.org/10.1093/femsre/fuad004

Lessons from COVID-19: Interventions Before a Global Pandemic?



DOI: https://doi.org/10.1128/jvi.01601-24

Research Questions

- What **biological questions** do you want to answer?
- What **biological data** can you use?
- **How** do you analyze the **biological data** (**big data**)?
- What are the expected **biological insights**, and how can they be **applied**?

https://www.geeksforgeeks.org/bioinformatics-application/

High-Performing Computing

The Queensland Alliance for Agriculture and Food Innovation (QAAFI) is a research institute of The University of Queensland (UQ), supported by the Queensland Government.

CRICOS code 00025B

Speed and Efficiency

10 hours, 5 people

10 hours, 30 people

1.5 hours, 170 people

High-performing Computing (HPC)

The use of **supercomputers** and **computer clusters** to solve problems requiring massive computation.

UQ acquires new supercomputer

6 Jun 2022

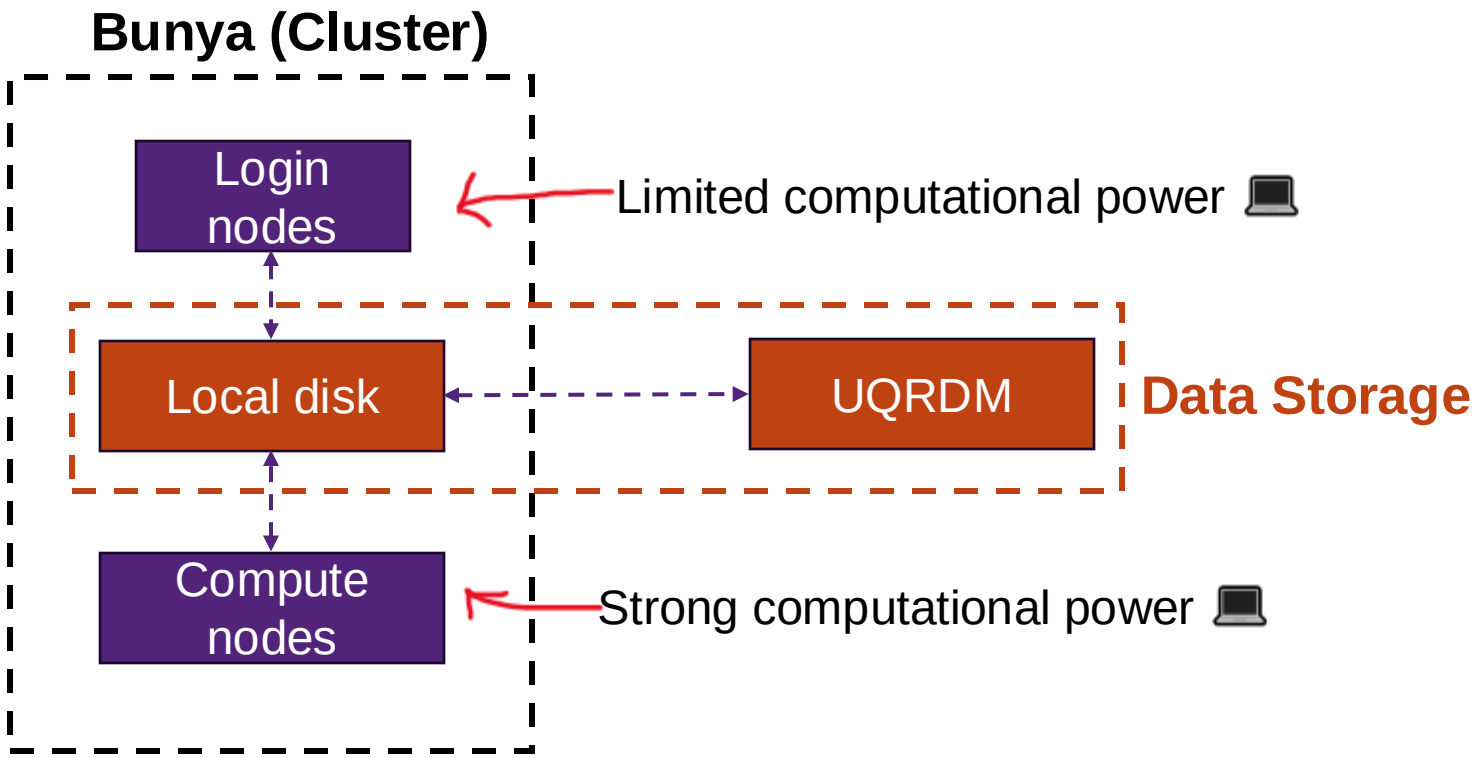
- Introduction to Bunya webinar
- Bunya's technical specifications

The University of Queensland has purchased a new supercomputer that is faster, multifaceted and more efficient than its current high-performance computers (HPCs).

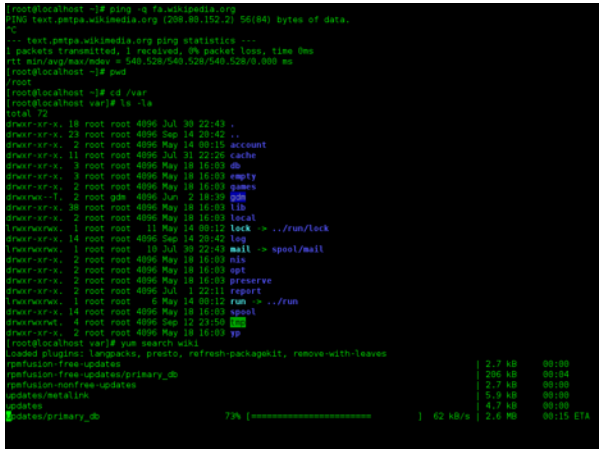
The HPC, named 'Bunya', after the native South-East Queensland tree, was bought from Dell Technologies and is estimated to become operational in July this year.

The Polaris Data Centre in Springfield, Queensland.

General Structure of Bunya

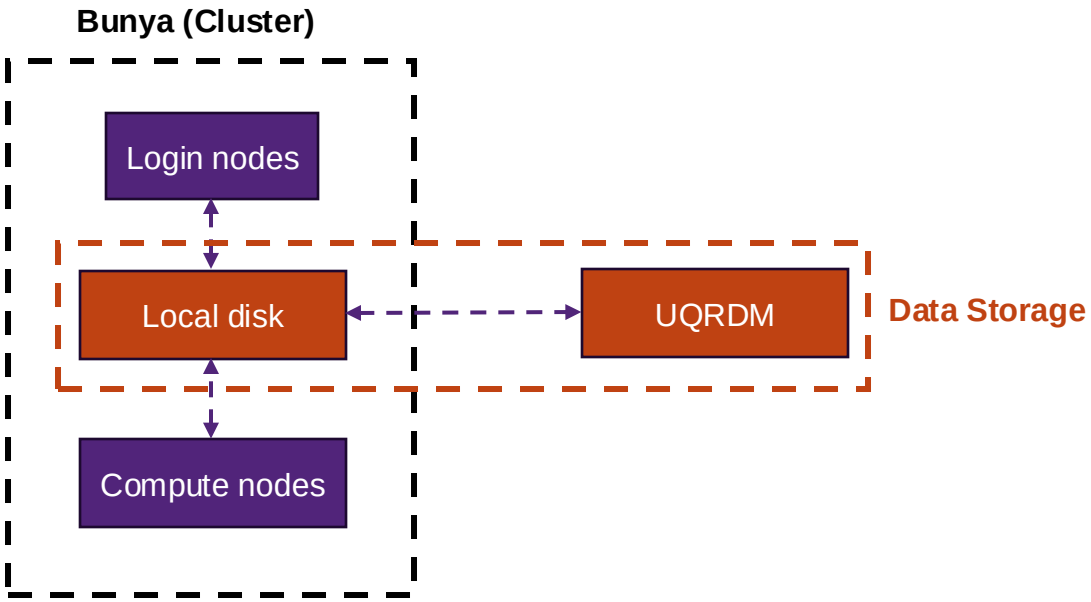


Terminal Interface to Bunya



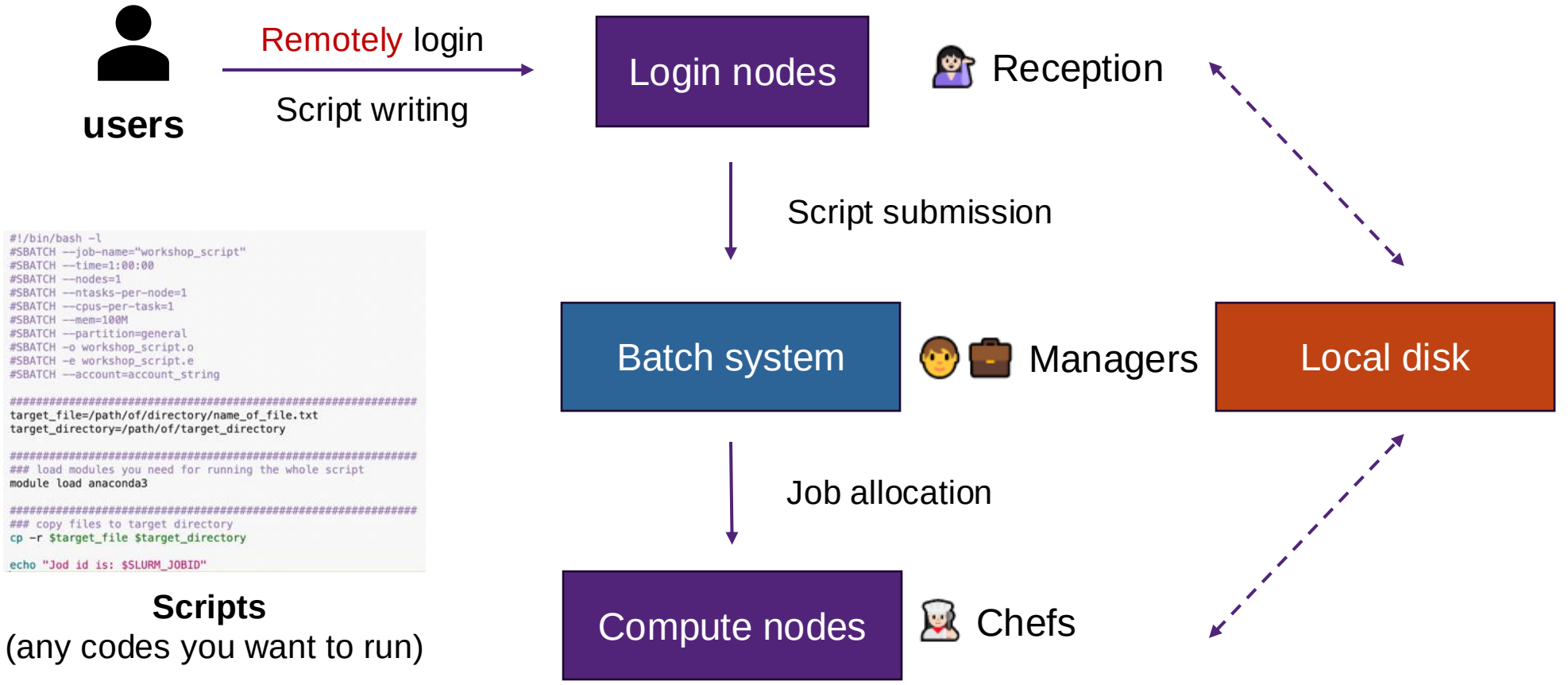
Terminal Interface/emulator

Securely
Remote Access



Once **remotely connected**, you are navigating the HPC until you **disconnect**.

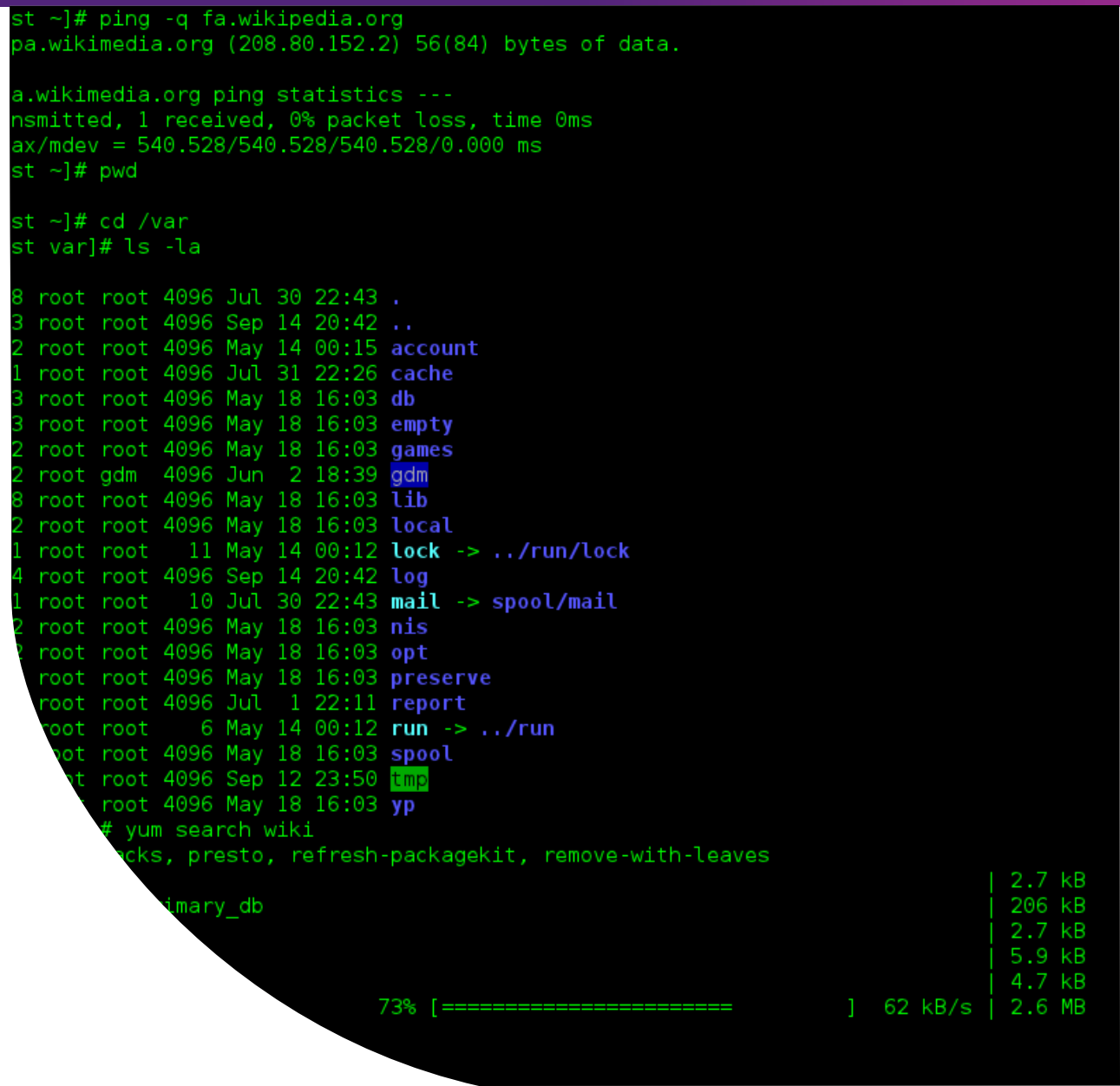
General Workflow of Bunya




```
#!/bin/bash -l
#SBATCH --job-name="workshop_script"
#SBATCH --time=1:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --mem=100M
#SBATCH --partition=general
#SBATCH -o workshop_script.o
#SBATCH -e workshop_script.e
#SBATCH --account=account_string

############################################################
target_file=/path/of/directory/name_of_file.txt
target_directory=/path/of/target_directory

############################################################
### load modules you need for running the whole script
module load anaconda3

############################################################
### copy files to target directory
cp -r $target_file $target_directory

echo "Jod id is: $SLURM_JOBID"
```

Scripts
(any codes you want to run)

Linux Command Lines Interface

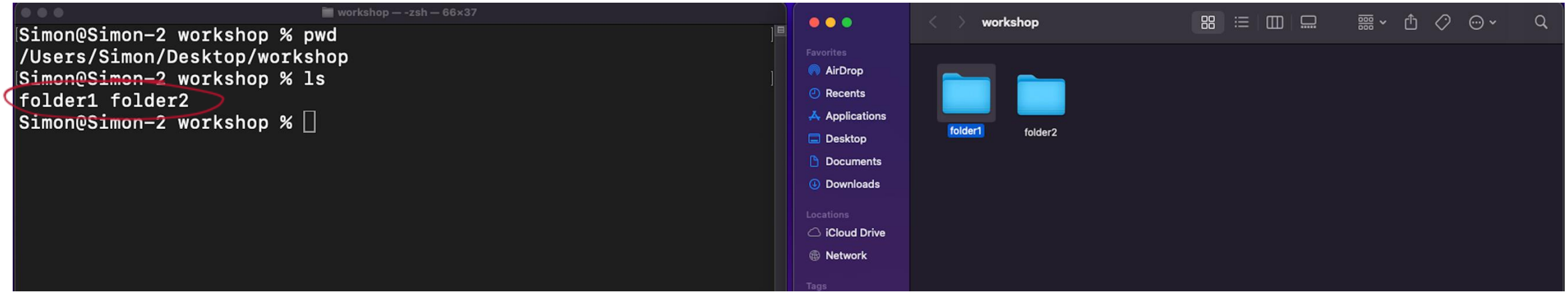


The Queensland Alliance for Agriculture and Food Innovation (QAAFI) is a research institute of The University of Queensland (UQ), supported by the Queensland Government.

CRICOS code 00025B

Linux Command Line Interface

Interacting with a **Unix-like computer system** by entering **text-based commands**.

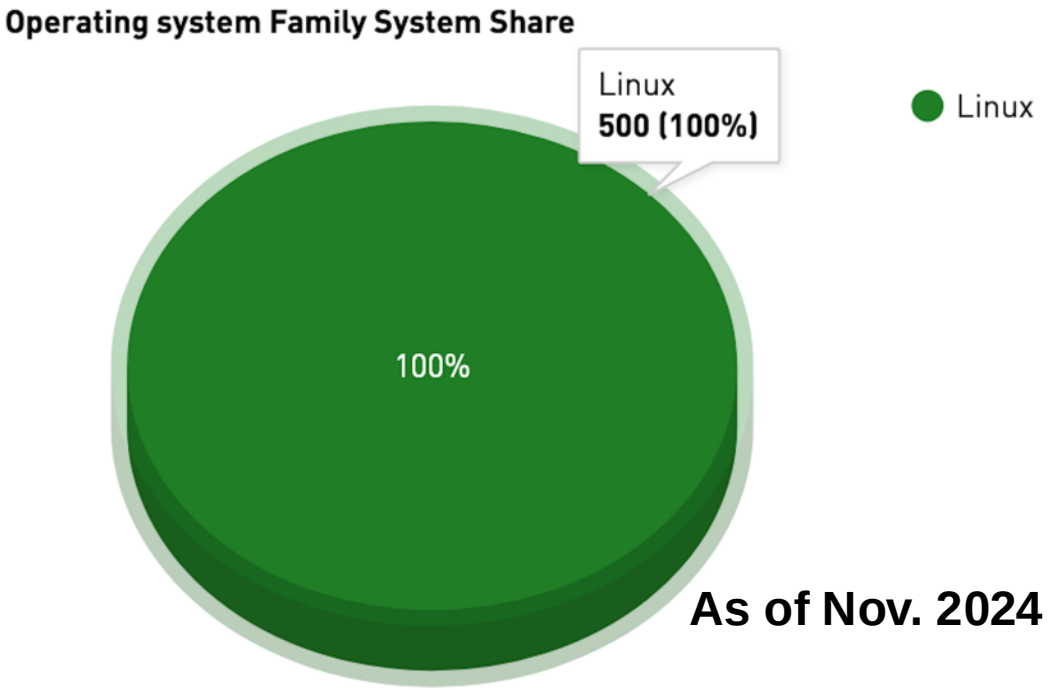


Command Line Interface

Graphic User Interface

Why Linux Command Lines

- **Resource efficiency -** A lighter way system interaction
- **Automation -** repetitive tasks using scripts
- **Efficient Data Handling** - efficiently process, transfer, and manage data
- **Remote Access** to HPC



As of Nov. 2024

https://www.top500.org/statistics/list/

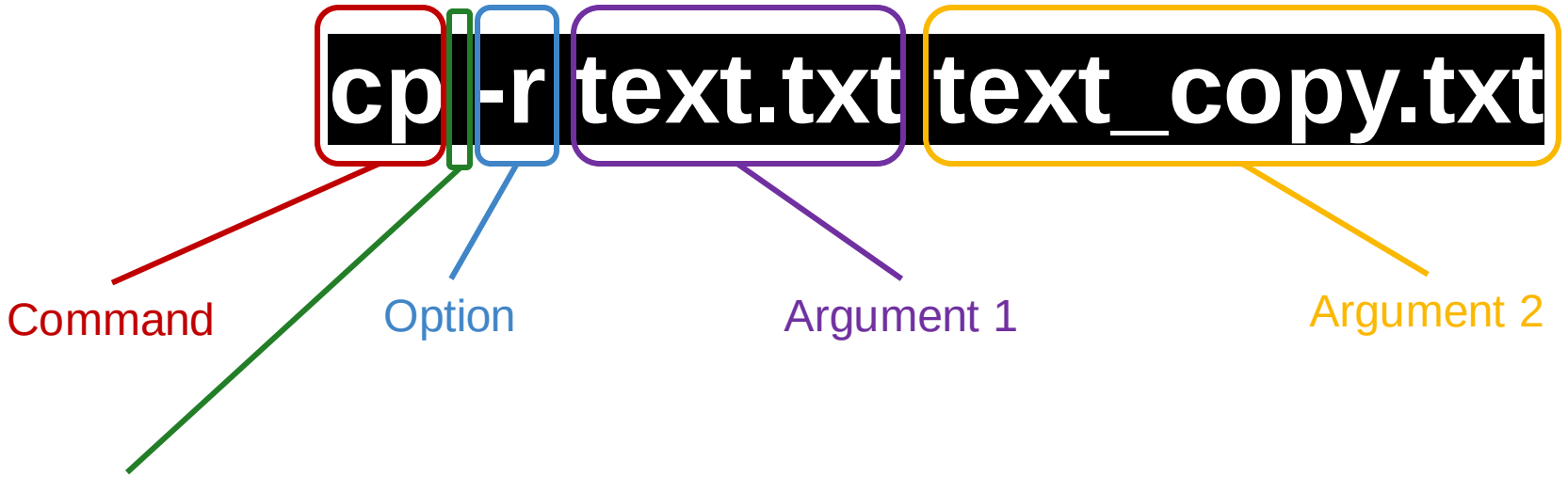
Linux Command Lines Practice

```
st ~]# ping -q fa.wikipedia.org
pa.wikimedia.org (208.80.152.2) 56(84) bytes of data.

a.wikimedia.org ping statistics ---
nsmitted, 1 received, 0% packet loss, time 0ms
ax/mdev = 540.528/540.528/540.528/0.000 ms
st ~]# pwd

st ~]# cd /var
st var]# ls -la

8 root root 4096 Jul 30 22:43 .
3 root root 4096 Sep 14 20:42 ..
2 root root 4096 May 14 00:15 account
1 root root 4096 Jul 31 22:26 cache
3 root root 4096 May 18 16:03 db
3 root root 4096 May 18 16:03 empty
2 root root 4096 May 18 16:03 games
2 root gdm  4096 Jun  2 18:39 gdm
8 root root 4096 May 18 16:03 lib
2 root root 4096 May 18 16:03 local
1 root root   11 May 14 00:12 lock -> ../run/lock
4 root root 4096 Sep 14 20:42 log
1 root root   10 Jul 30 22:43 mail -> spool/mail
2 root root 4096 May 18 16:03 nis
2 root root 4096 May 18 16:03 opt
 root root 4096 May 18 16:03 preserve
 root root 4096 Jul  1 22:11 report
oot root    6 May 14 00:12 run -> ../run
ot root 4096 May 18 16:03 spool
t root 4096 Sep 12 23:50 tmp
 root 4096 May 18 16:03 yp
# yum search wiki
cks, presto, refresh-packagekit, remove-with-leaves
                                                                | 2.7 kB
imary_db                                                        | 206 kB
                                                                | 2.7 kB
                                                                | 5.9 kB
                                                                | 4.7 kB
                                  73% [=======================        ]  62 kB/s | 2.6 MB
```

The Queensland Alliance for Agriculture and Food Innovation (QAAFI) is a research institute of The University of Queensland (UQ), supported by the Queensland Government.

CRICOS code 00025B

Linux Command Line Structure



Space is important for separating commands, options or arguments

Notes:

- **Avoid** using space in your file name. Use "_" to replace space.
- Type both **file name** and **file format** if the argument is a file.

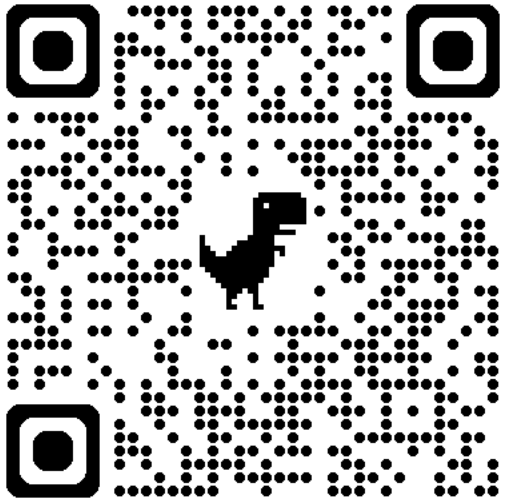
Linux Command Line Interface Programs

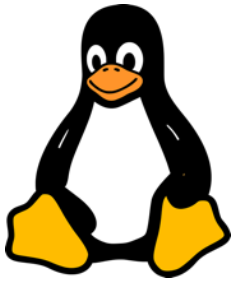
Windows user:

- Scan the QR code below to download **Git Bash**
- Download the **"Windows Portable"** one for installation.

macOS/Linux user:

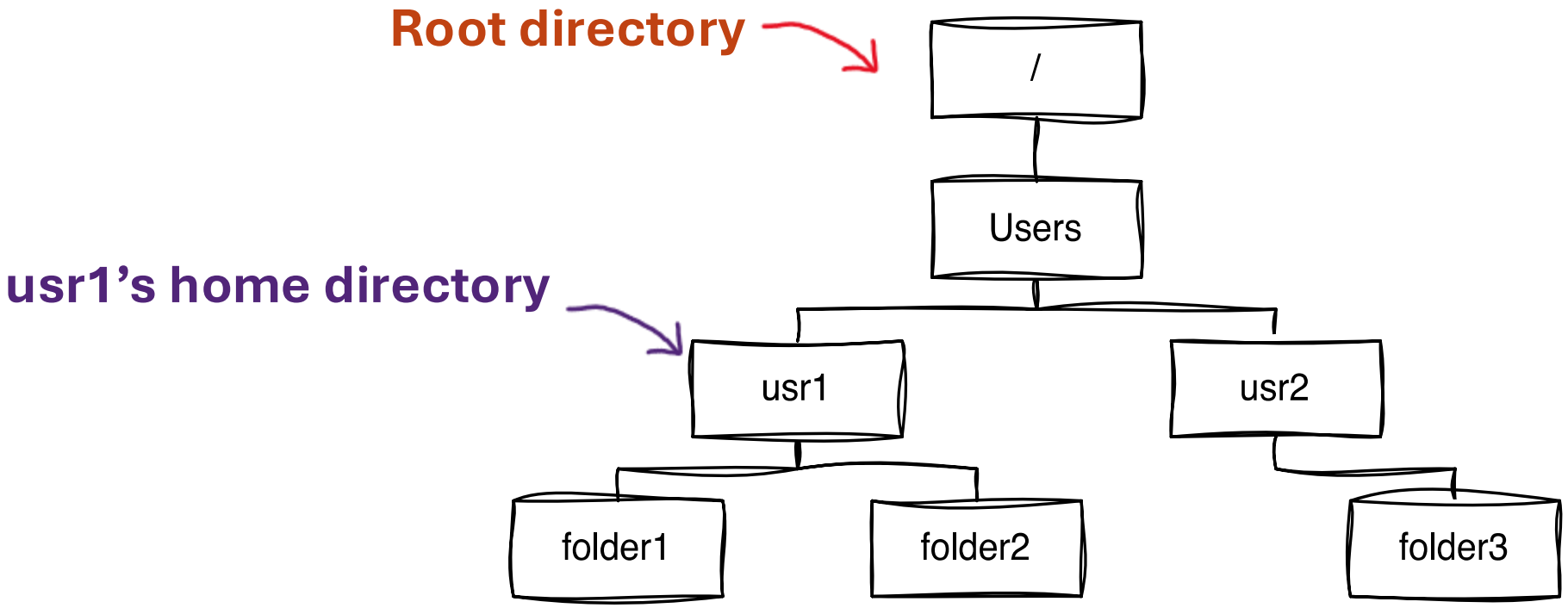
- Open the built-in **Terminal** app in your laptop





Navigating

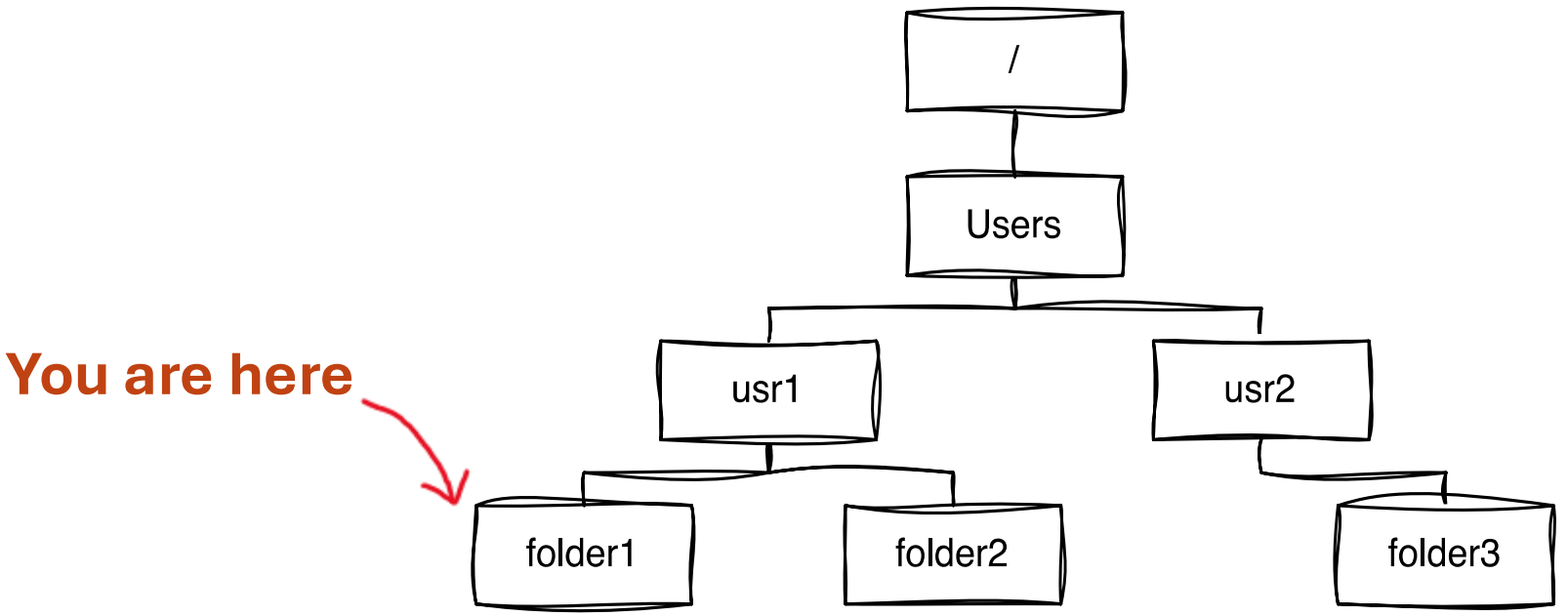
The **file system** of a computer system follows a **hierarchical tree structure**, with directories branching into **subdirectories** and **files**.



Navigating

- **pwd**

Print the **full/absolute path** of the **current working directory**



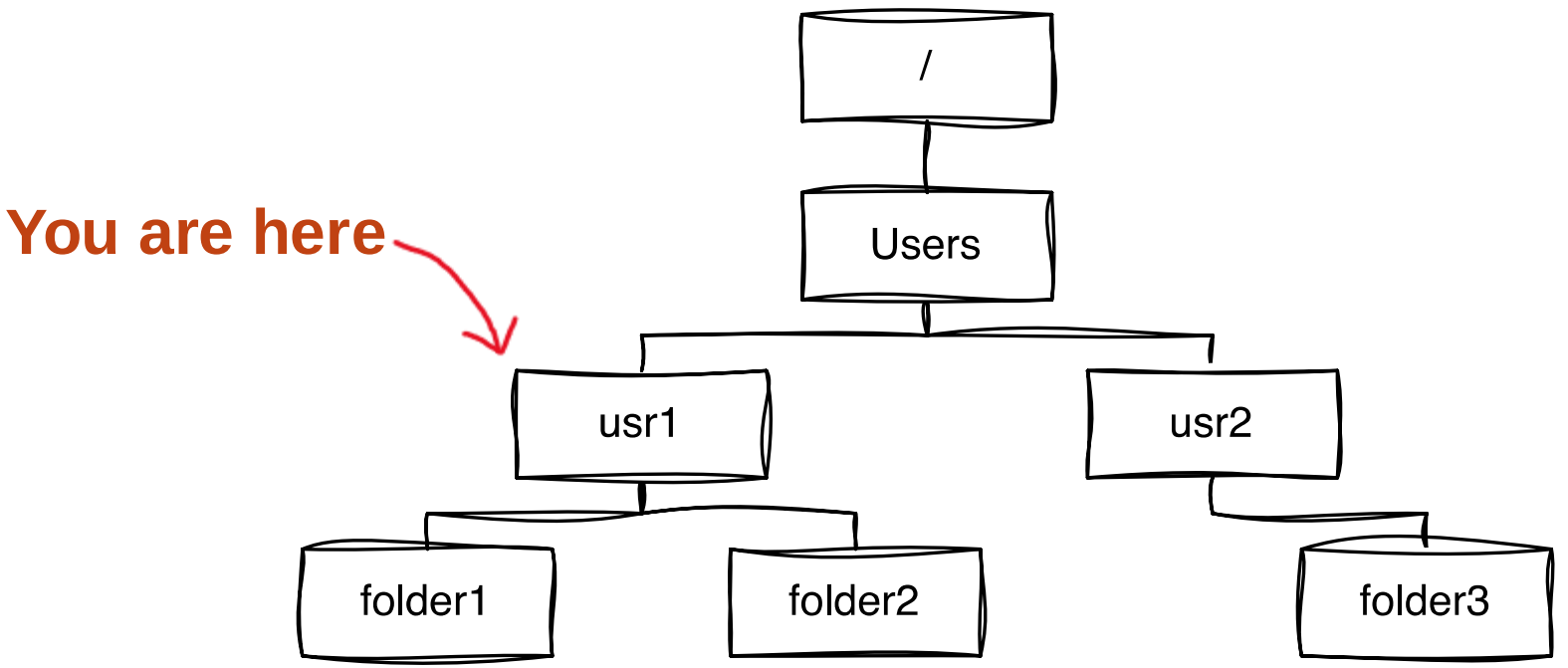
/Users/usr1/folder1

"/" is a directory separator

Navigating

- **ls**

List the contents of a directory

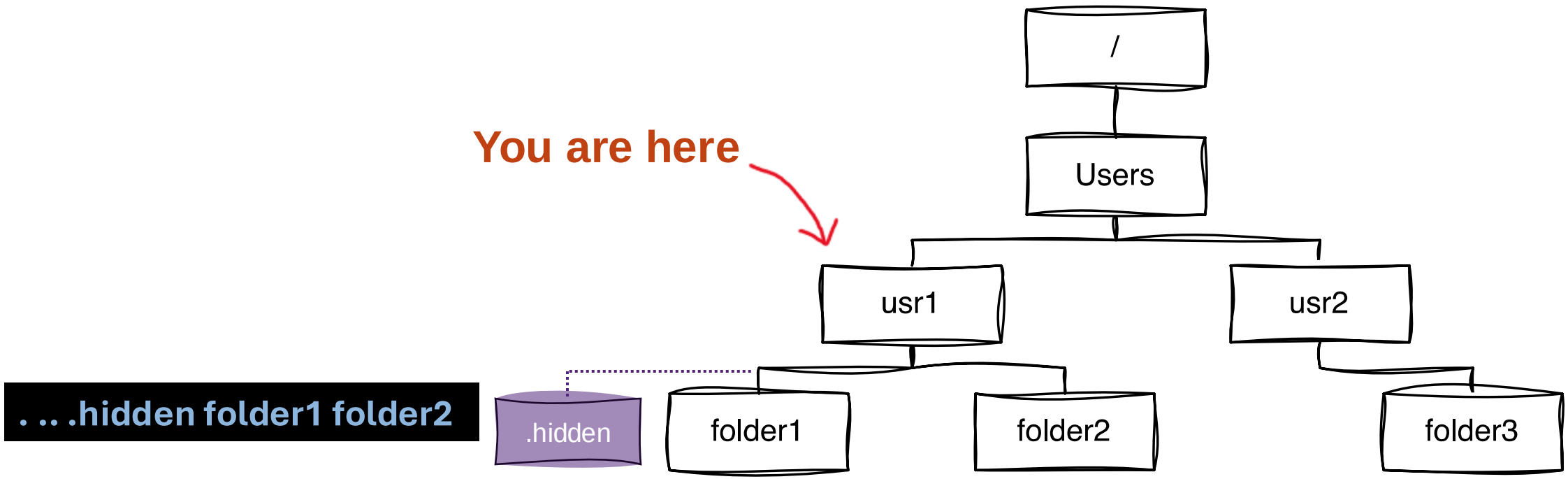



```
folder1 folder2
```

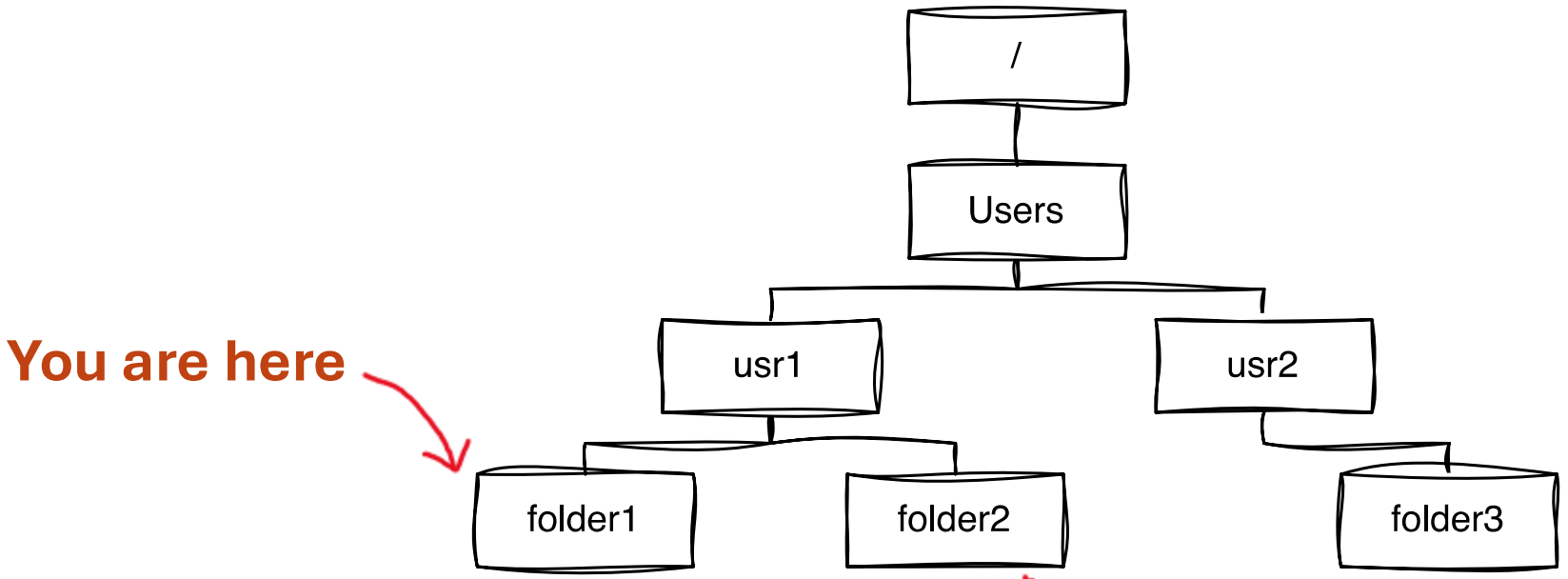
Navigating

- **ls -a**

List the **all** contents of a directory



Navigating



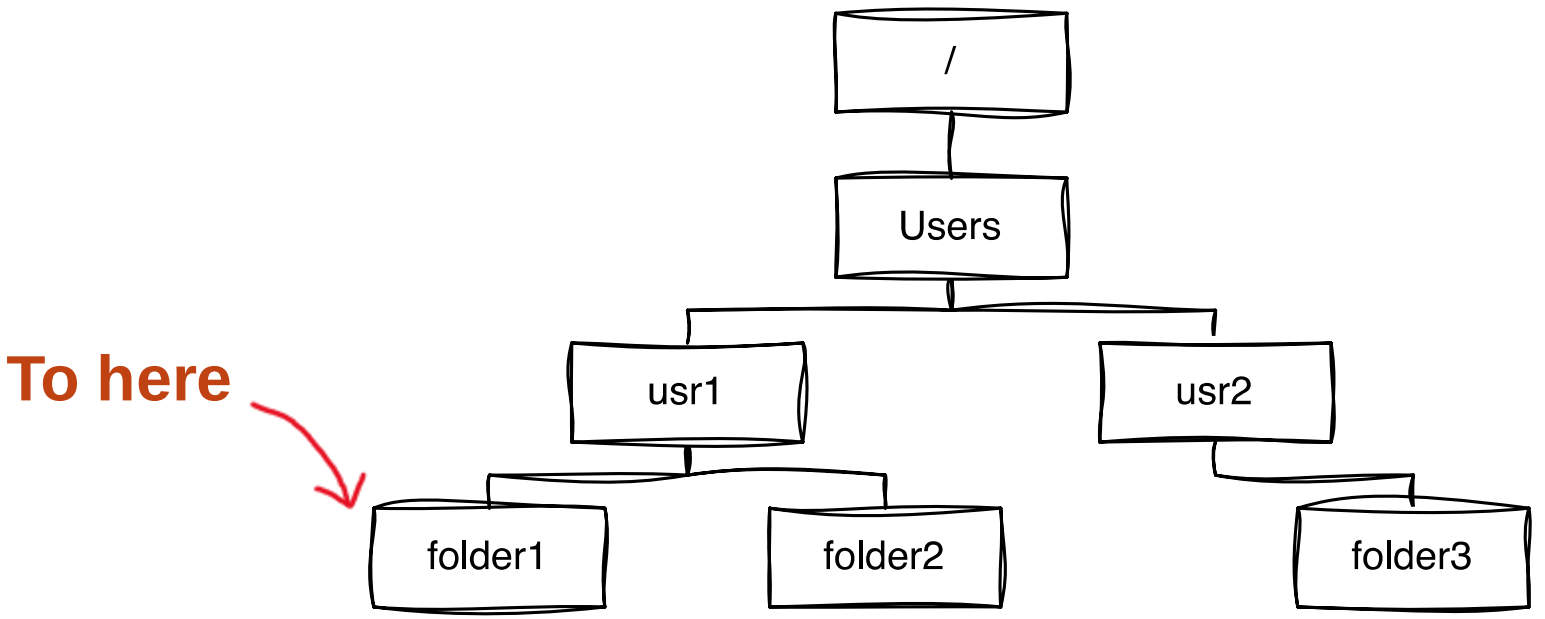
The "**/**" here tells the computer that it is from a root directory

- **Absolute path of folder2:** /Users/usr1/folder2
- **Relative path of folder2 based on folder1:** ../folder2
- "**..**" means the parent directory of the current directory

Navigating

- **cd** /path/of/directory

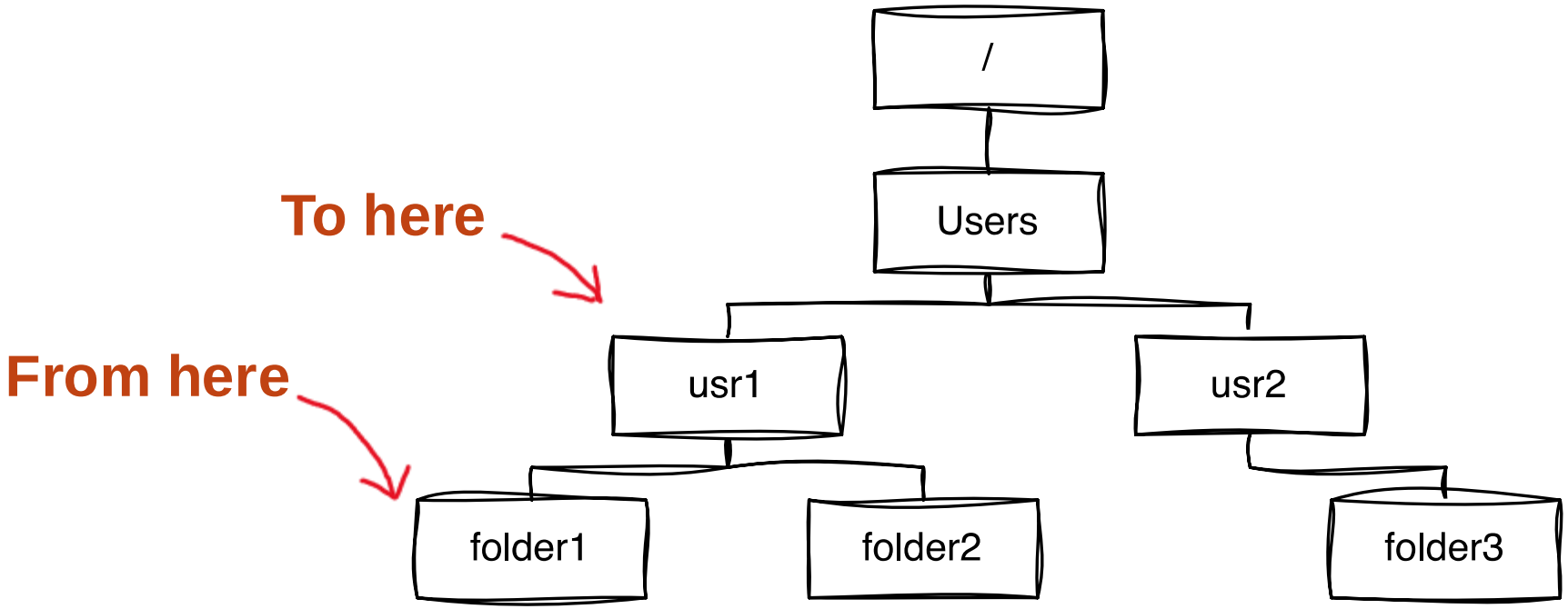
  Change the working directory to the target directory



cd /Users/usr1/folder1

Navigating

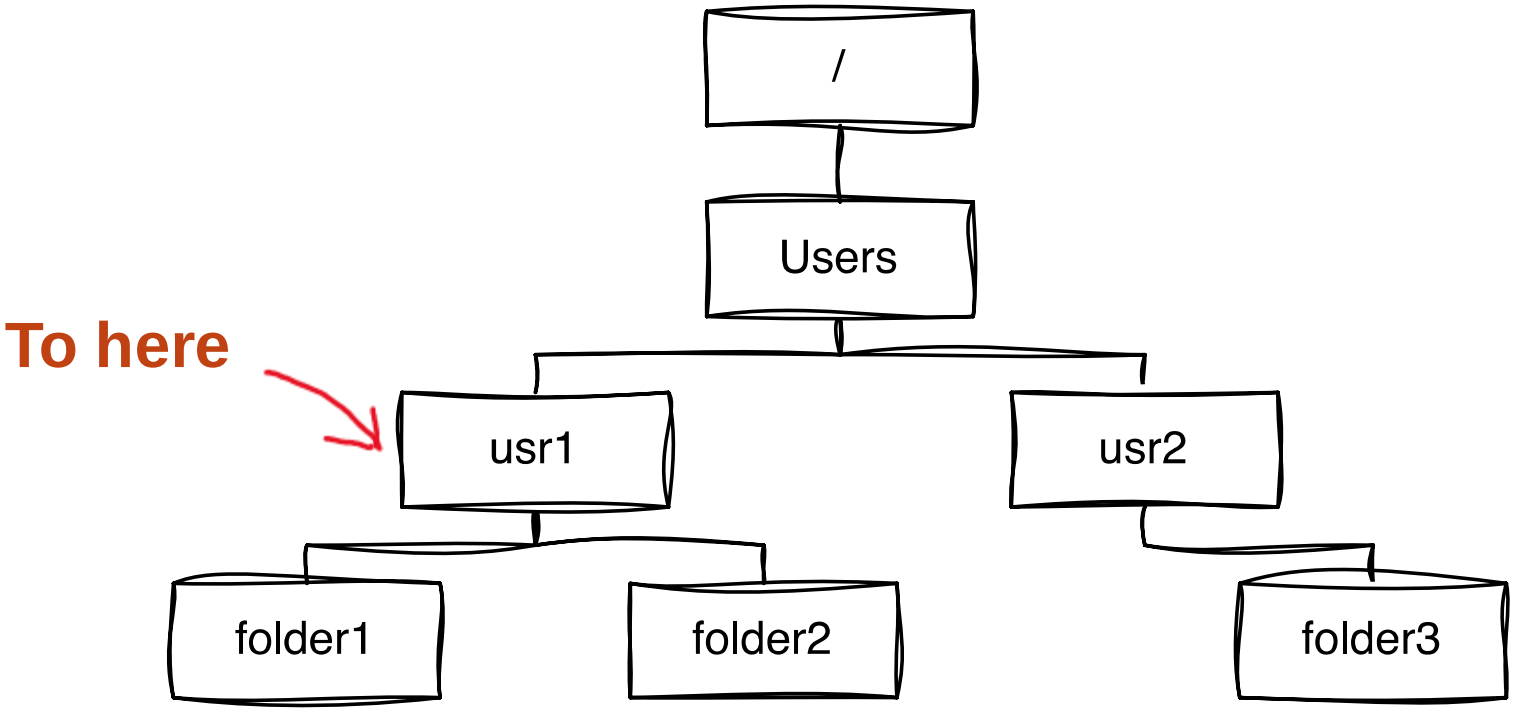
- **cd ..**
- **cd /Users/usr1**



Navigating

- `cd ~`

  "~" means **user's home** directory



15-minute break

Changing the File System

- **mkdir** name_of_folder

  Make(create) a directory with a specific name at the current working directory

- **rmdir** name_of_folder

  Remove an **empty** directory at the current working directory

- **nano** target_file

  Text editor to create text file (1. Ctr+O write out; 2. Enter; 3. Ctr+X exit)

Changing the File System

- **cp** target_file /path/of/directory

  Copy files to a target directory

- **cp** filename_1 filename_2

  Copy filename_1 to filename_2

Changing the File System

- **mv** target_file /path/of/directory

  Move files to a target directory

- **mv** filename_1 filename_2

  Rename filename_1 to filename_2

Changing the File System

- **rm** target_file

  Remove target files (**once delete, no restore**)

- **rm -r** name_of_folder

  Remove target folder that is **not empty** (**once delete, no restore**)

  ”-r” means remove directories and their contents recursively

Viewing

- **less** target_file

  View the contents of a file **on one page** at a time

  Press "Q" to exit

- **cat** target_file

  Concatenate and display the content of a file

- **grep** 'pattern' target_file

  "**Global regular expression print**", searching matched patterns in files

- **head/tail** target_file

  Print the **first/last 10** lines of each file

Other usages

- **sort –n** target_file

  Sort or merge records (lines) of text files with numeric data

- **sort -u** target_file

  Sort or merge records and print the **unique** keys in a file

- **>**

  Output redirection symbol, save the output of a command into a file.

  uniq filename_1 > filename_2

Manual and help

- **<command> --help**

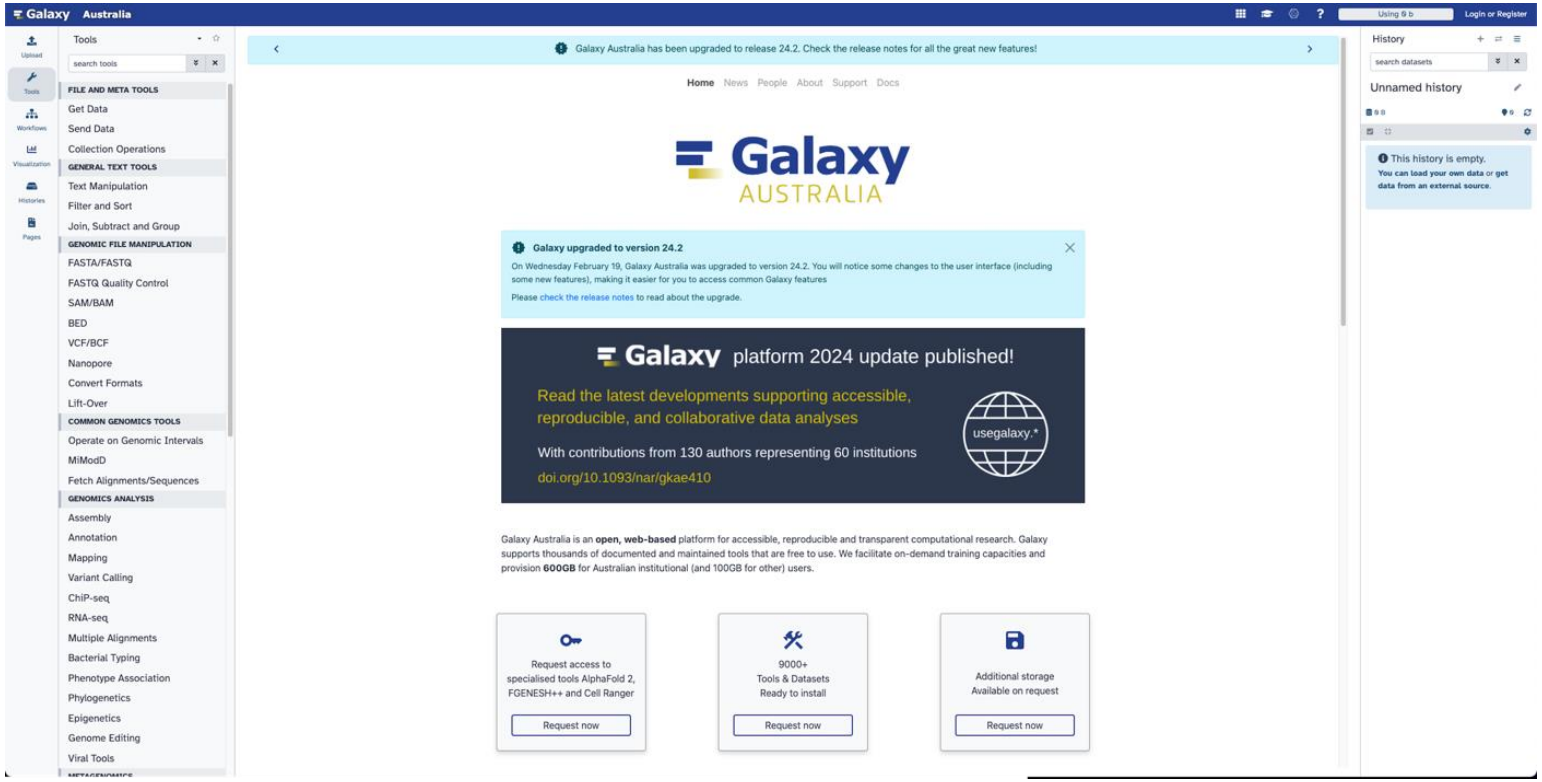
  Display **help information** for the usage of a command

- **man <command>**

  Open a **manual** page for a command

  Press “Q” to exit

No Like Linux Command Lines?



An **open** and **web-based** platform for computational research

https://usegalaxy.org.au/

Limitations of Galaxy Australia

- Sharing **computational resources** with researchers across Australia.
- **Limited data storage** quota during analysis.
- **Not ideal** for handling **numerous** file.
- **Not for long-term** data storage.

Analysis and Workflow

The Queensland Alliance for Agriculture and Food Innovation (QAAFI) is a research institute of The University of Queensland (UQ), supported by the Queensland Government.

CRICOS code 00025B

Analysis Workflow

- Github, such as: https://github.com/epi2me-labs/wf-transcriptomes
- Galaxy Australia: https://training.galaxyproject.org/training-material/
- Specific workshops, such as https://genomicsaotearoa.github.io/metagenomics_summer_school/
- Research papers that match your field.

Things you need to consider

1. What type of data do you have?

- DNA? RNA? …
- Paired-end or Single-end sequencing data?
- Long-read or short-read sequencing data? …

2. Are the tools suitable for your data?

- Can you run direct-RNA data mapping using a DNA alignment tool?
- Can you run your short-read data on a tool for long-read data? ...

Things you need to consider

3. Is the statistics suitable for your data?

- Student t-tests or Wilcoxon test?
- Paired or non-paired t-tests?
- What statistic models should be used for your data?
- Do you need to consider random effects in your model? …

4. Can the pipeline you choose help answer your biological questions?

- What are your hypotheses?
- Do you need a customised workflow?
- Do you need a customised script to handle the data?...

5. …

Training and Supports

The Queensland Alliance for Agriculture and Food Innovation (QAAFI) is a research institute of The University of Queensland (UQ), supported by the Queensland Government.

CRICOS code 00025B

Training and Supports

- QCIF

https://www.qcif.edu.au/training-resources

- LinkedIn Learning

https://www.linkedin.com/learning/

- Hacky Hour UQ

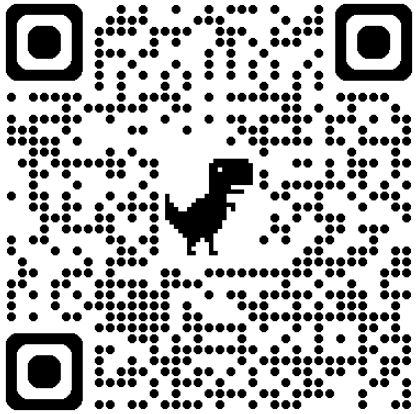
https://rcc.uq.edu.au/training-support/meetups#Hacky%20Hour%20UQ%C2%A0%C2%A0

- UQ Bunya

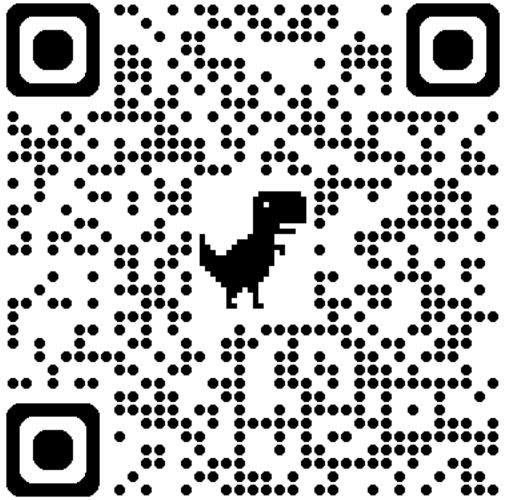
https://rcc.uq.edu.au/systems/high-performance-computing/bunya

Take Home Messages

- Select the appropriate **biological data** that can answer your **biological questions**.
- Select the **workflows, statistics,** or **models** suitable for your data analysis.
- Utilise HPC to accelerate your analysis.



Thank you



- Your **feedback** is important for our improvement!
- Want more **activities** or **bioinformatic workshops** from QSA? Let us know what you want!