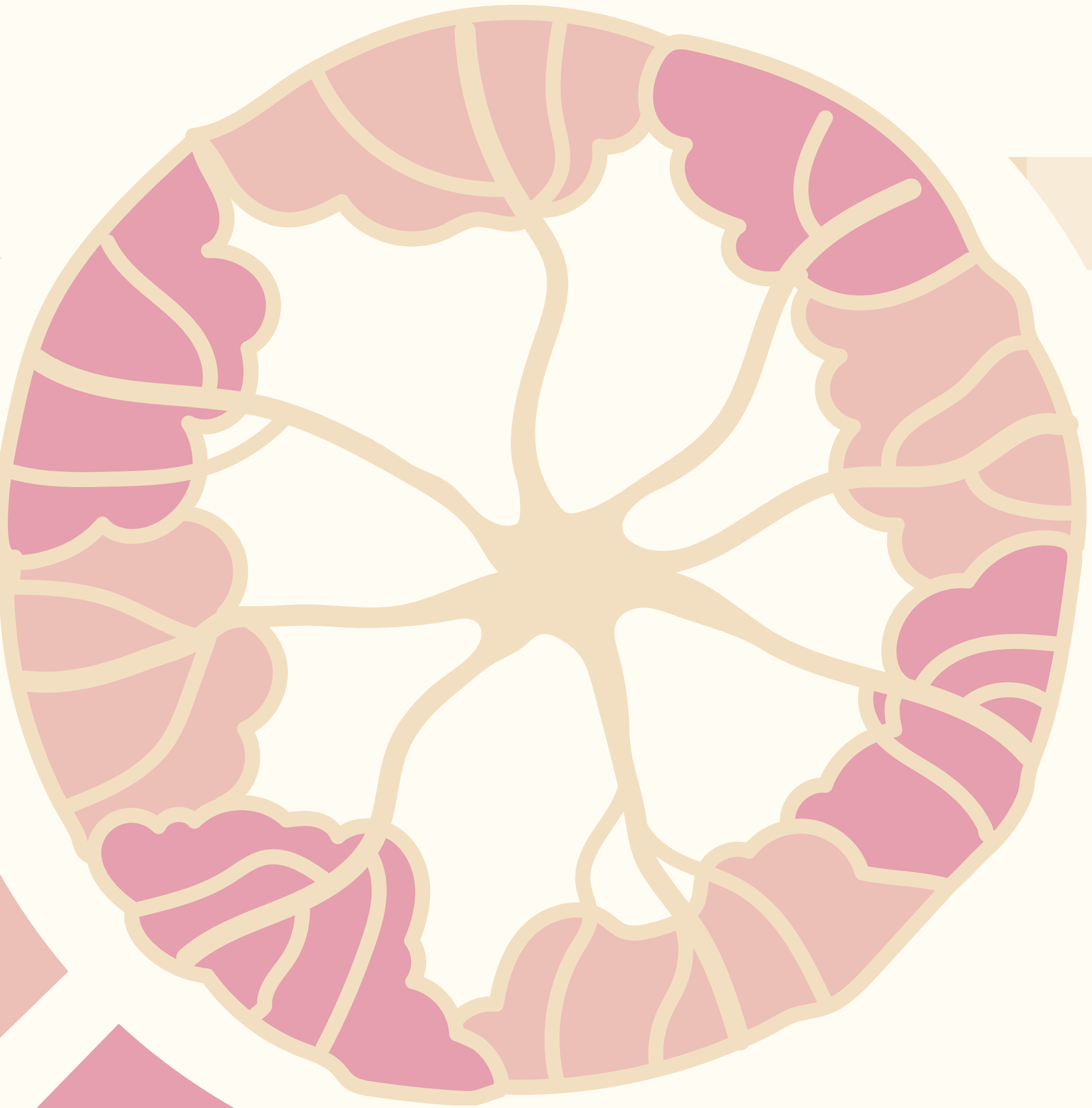
-Evolution for Everyone-

Digital Research Alliance of Canada



Bioinformatics for beginners

Land Acknowledgement

Our workshop is held on the UBC Point Grey (Vancouver) campus, which sits on the traditional, ancestral, unceded territory of the xʷməθkʷəy̓əm (Musqueam) First Nation.

Code of conduct

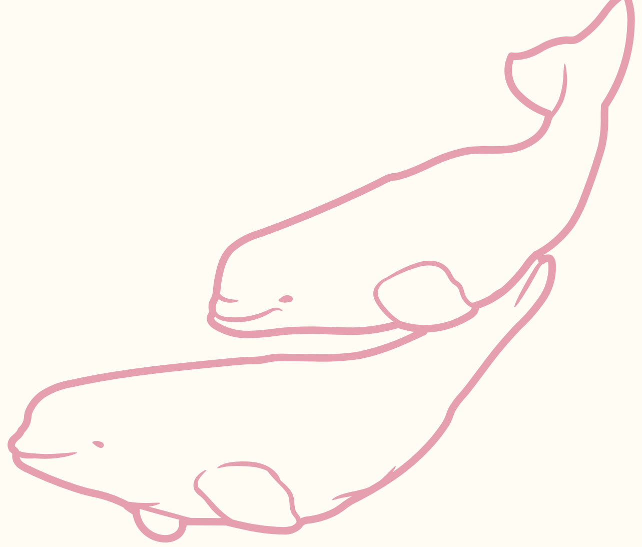
As a community we welcome everyone, and encourage a friendly and positive environment.

- Be friendly and patient
- Be welcoming
- Be considerate
- Be respectful
- Be careful in the words that we choose
- Try to understand why we disagree

(Based on the OLS Open Seeds code of conduct)

Overview

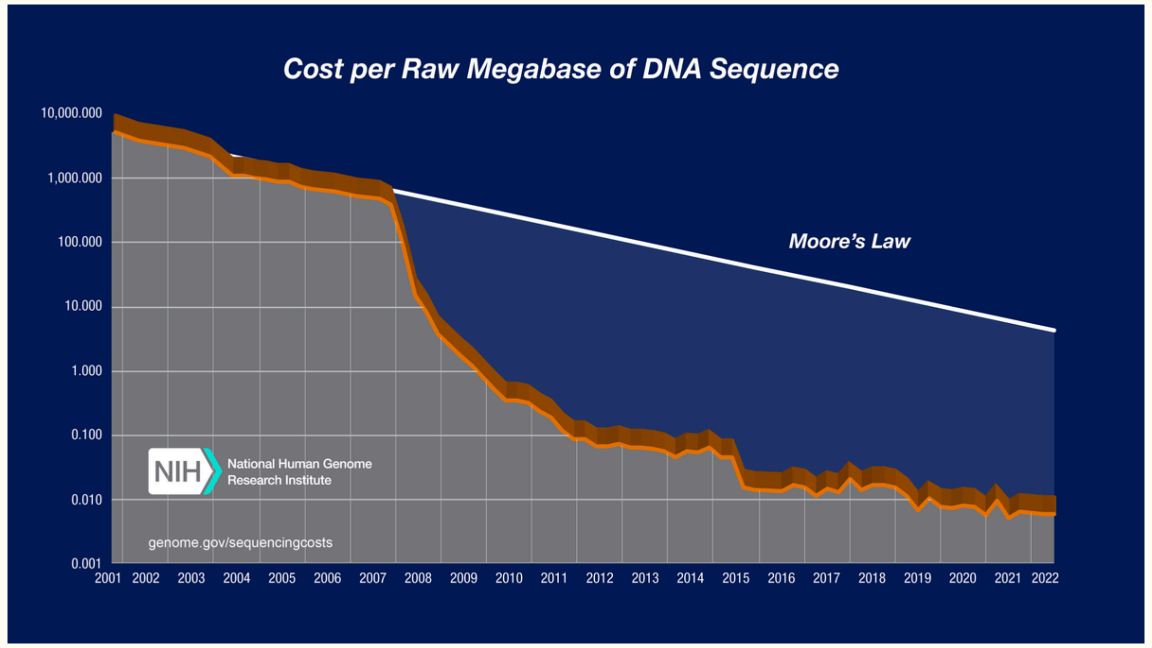
- Why Bioinformatics?
- The Linux/Unix system + why coding?
- Advanced research computing (working on a server)
- Digital Research Alliance of Canada
- Data management and reproducibility

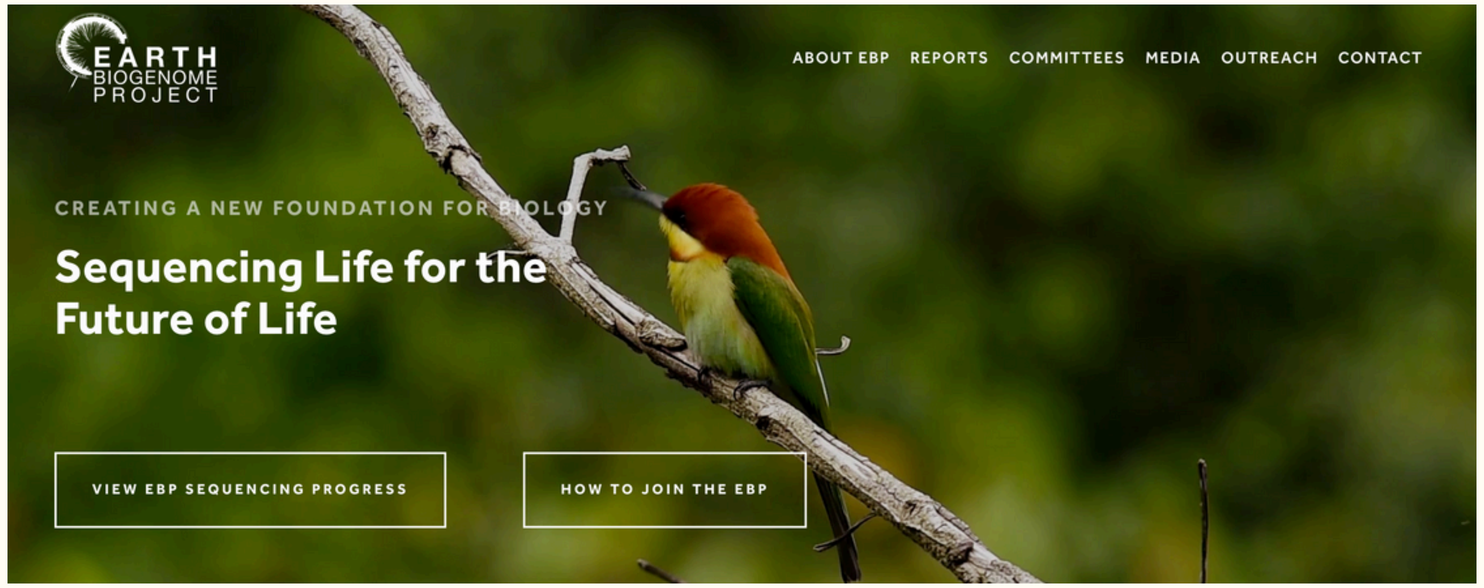


Why bioinformatics?

seriously, why?

As sequencing costs drop exponentially...





CREATING A NEW FOUNDATION FOR BIOLOGY

Sequencing Life for the Future of Life

we can sequence more genomes than ever.

There's SO MUCH DATA!!

Large-scale data requires powerful analysis tools!
That's where biology meets computer science ❤

Setup of Linux Torvalds creator of Linux
My Setup programming in python with ChatGTP

What is Linux?

Linux is an open source operating system (just like Windows and MacOS, but free!), created in 1991 by Linus Torvalds

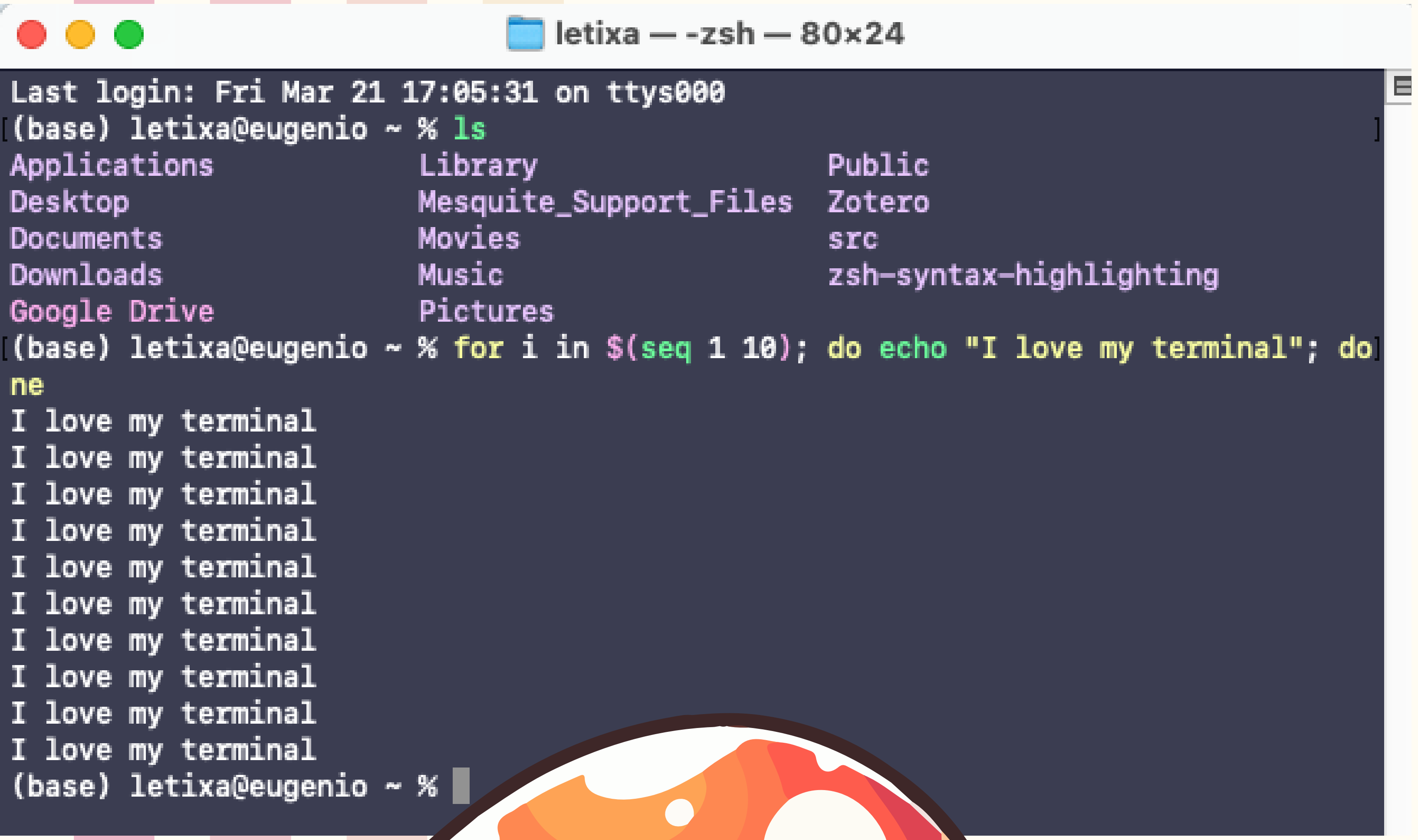
Different distributions and "flavours"

- Ubuntu and Mint are the most popular for first-time users

Many servers use Linux as their operating system, including the servers of the Digital Research Alliance of Canada (Beluga, Narval, Niagara, Cedar, Graham)

The terminal

You know what they say, once you go bash...



```
Last login: Fri Mar 21 17:05:31 on ttys000
(base) letixa@eugenio ~ % ls
Applications            Library                 Public
Desktop                 Mesquite_Support_Files  Zotero
Documents               Movies                  src
Downloads               Music                   zsh-syntax-highlighting
Google Drive            Pictures
(base) letixa@eugenio ~ % for i in $(seq 1 10); do echo "I love my terminal"; do
ne
I love my terminal
I love my terminal
I love my terminal
I love my terminal
I love my terminal
I love my terminal
I love my terminal
I love my terminal
I love my terminal
I love my terminal
(base) letixa@eugenio ~ %
```

🍄 What is the terminal?

The terminal is an app where you can communicate with your computer by typing commands

🍄 What can I do on the terminal?

Anything you do in your computer (and much MORE!)

🍄 What's so special about it?

The terminal allows you to run a lot of programs that don't have a graphic interface

The shell

You know what they say, once you go bash...

🍄 What is the shell?

The terminal runs a command-line interpreter, called shell, that passes your commands to the operating system

🍄 What's the difference between terminal and shell?

Terminal is the location, shell is the program that runs in the terminal

- *Think about a restaurant:* if the terminal is the table where you sit, the shell is the waiter who takes your order to the kitchen (operating system)

🍄 There are different types of shell!

Default shells of each operating system: bash (Linux), Zsh (macOS), Command Prompt (Windows)

Why coding?

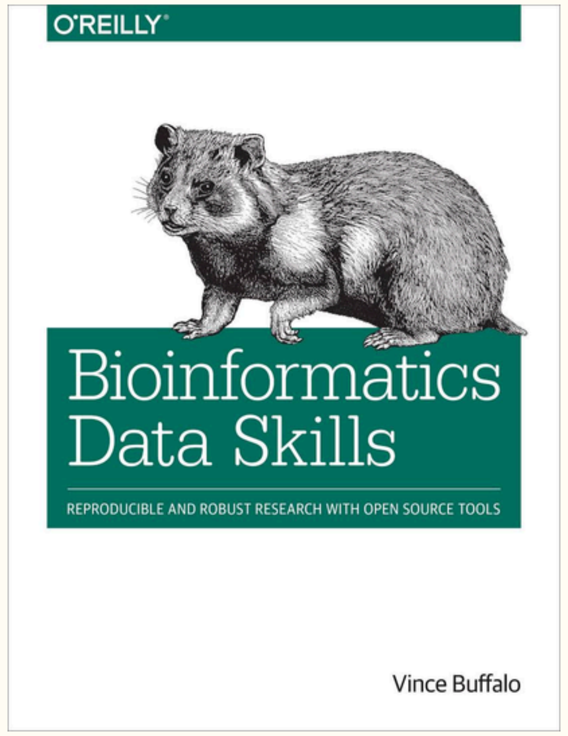
for power, of course

Popular programming languages in Bioinformatics:

- Bash
- Python
- R
- C++
- Matlab

How to get started?

a quick list of useful (and free!) resources

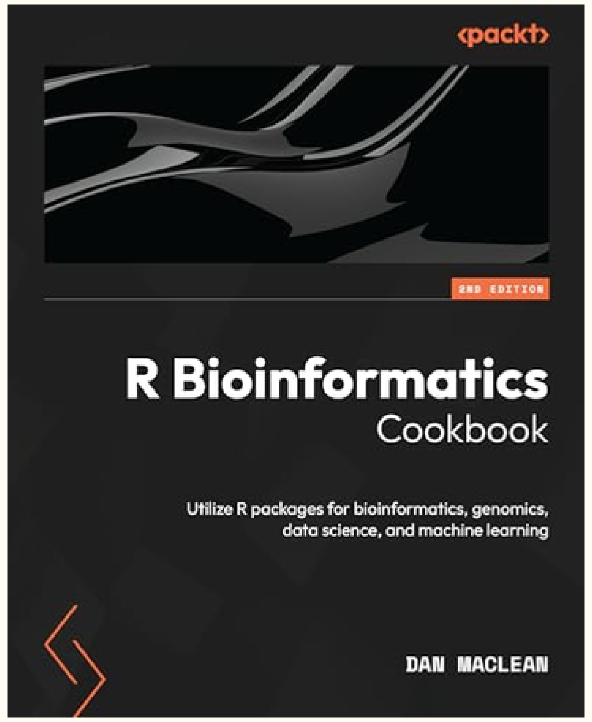


O'REILLY

Bioinformatics Data Skills

REPRODUCIBLE AND ROBUST RESEARCH WITH OPEN SOURCE TOOLS

Vince Buffalo



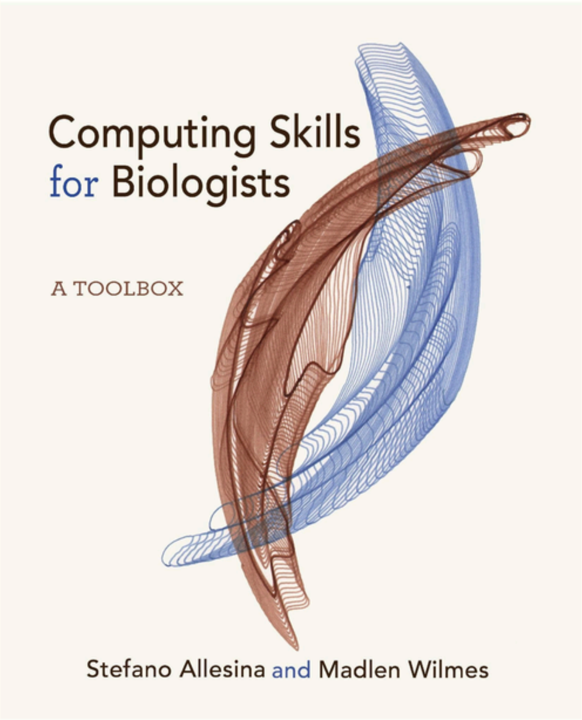
‹packt›

2ND EDITION

R Bioinformatics Cookbook

Utilize R packages for bioinformatics, genomics, data science, and machine learning

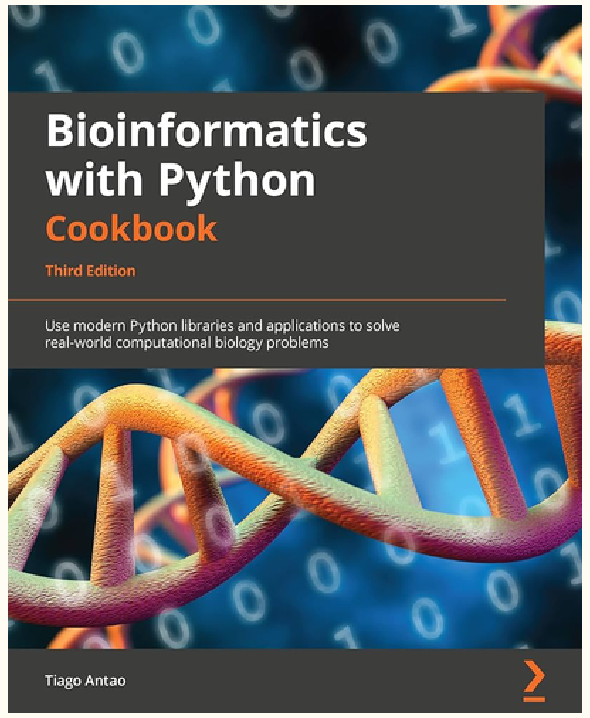
DAN MACLEAN



Computing Skills for Biologists

A TOOLBOX

Stefano Allesina and Madlen Wilmes



Bioinformatics with Python Cookbook

Third Edition

Use modern Python libraries and applications to solve real-world computational biology problems

Tiago Antao

Books!

(There's honestly so many, here are some personal recommendations)

How to get started?

a quick list of useful (and free!) resources

Online training (courses, workshops & tutorials):

- Digital Research Alliance of Canada
- Canadian Bioinformatics Hub
- ACENET
- Compute Ontario
- Calcul Québec
- The Carpentries
- Codeacademy
- Udemy
- Coursera
- Harvard free courses

How to get started?

a quick list of useful (and free!) resources

Online documentation + more tutorials:

- Introduction to R for biologists
- Introduction to Python for Bioinformatics
- Biopython
- PAML

Resources made by DRI-EDIA Champions:

- Applications of Digital Research Infrastructure (DRI) in Evolutionary Biology
- Evolution for Everyone
- Raise DRI

Advanced research computing

(in other words, what the **** is a server?)

We call it *Advanced Research Computing (ARC)* when your own computer is not enough (more computational power required) or you need expert advice (you don't know how to do the task)

🍄 **What is a server?**
A supercomputer sitting in room somewhere that you can access remotely, from your personal computer

🍄 **There are 3 main types of super-computers:**

- clusters (many individual computers linked together)
- multicore (one computer with many processing units)
- accelerators

Advanced research computing

What's cool about a server?

Power!!

servers are way larger and more powerful than our personal computers, so they can store more data, process things faster and deal with more complicated problems

Software

research servers have a vast software library, that is kept updated by the administrators. You don't have to install anything on your personal computer!

The Alliance

The **Digital Research Alliance of Canada** is an organization that supports advanced research in Canada providing infrastructure, tools, and expertise in high-performance computing (HPC), research data management (RDM), and research software.

They offer training, software and computing resources for any researcher affiliated with a Canadian institution.

By creating an account with the Alliance, you can access their servers:

- Beluga
- Cedar
- Graham
- Narval
- Niagara

Registering with the Alliance

🍄 Applying for a server account

To get access to the Alliance's servers, first you need to create an account by **registering** with the Alliance CCDB

- if you are a PI, you can apply directly for an account
- if you are a student/member of a research group, your supervisor needs have an account first, so they can sponsor your application at no cost.

Registering with the Alliance

🐞 Setting up your server account

Once your application has been approved (usually takes 1-2 business days), and before you can use the server, you need to set up:

1. Multi-factor Authentication
2. SSH-keys (recommended)
3. and for Niagara, you need to access the opt-in page and click on the join button. In the future, this step will be required for other servers as well.

Working on a server

🐞 Log in:

```
ssh [username]@[address].alliancecan.ca
```

Example: ssh leticiamagpali@narval.alliancecan.ca

🐞 Your folders:

- home (small storage, daily backup)
- scratch (huge storage, older files purged) → **submitting jobs!**
- project (large storage, daily backup) → **storing data!**

Working on a server

🐞 Transfer files:

1. Transferring a local file (from your computer) to the server

```
scp /path/to/local/file [username]@[address]:/path/to/remote/destination/folder
```

2. Transferring a remote file (from the server) to your computer

```
scp [username]@[server]:/path/to/remote/file /path/to/local/destination/folder
```

You need to run these commands from your local terminal (not the server!)

Working on a server

🐞 A server is a shared space!
That means it's built and used differently from your personal computer

It has specific rules to guarantee a fair sharing of resources by all users:

- Once you log in, you are directed to a login node (a shared space with a lot of other users, **not for running jobs**)
- You won't have permissions to do **everything** (for example, access to folders/users is restricted)
- To run an analysis/script/command on the compute nodes, you need to submit a job to a **scheduler** (a system that manages resource allocation to users).

📍 These are the rules of the Alliance servers. Always check your server's documentation to learn the rules and procedures!

Working on a server

🐞 How to submit a job:

To run a job, you need to write a *bash* script with your code and submit it to the scheduler (SLURM).

```
#!/bin/bash
#SBATCH--ntasks=1
#SBATCH--time=01:00:00
#SBATCH--mem=1000M

# Write your code below
```

The first 3 lines specify the resources you are asking for:

- ntasks = number of cores
- time = how long you'll be using these cores
- mem = how much memory you'd like to use

These lines are read by SLURM to schedule your job with the appropriate resources.

After writing the script, you can submit it using the command **sbatch <job_name.sh>,** and check the status with **squeue--me**

Data management & reproducibility

🐞 Good practices in data science:

- Always have backup(s)
  for example, GitHub, Drive/Cloud, local

- Document your work
  - Create REAME files
  - Annotate your code
  - Have a lab log / lab book

- Share your published data and code in open access databases! (GitHub, FigShare, Zenodo, Dryad)

Thank you!

Questions?

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

UBC

DALHOUSIE UNIVERSITY