Genome assembly and annotation

Day 1: UNIX and CSC

Igor Pessi

Department of Microbiology – UH

igor.pessi@helsinki.fi

Aims for this part of MMB-114

Day 1: Basics of UNIX and working with the command line

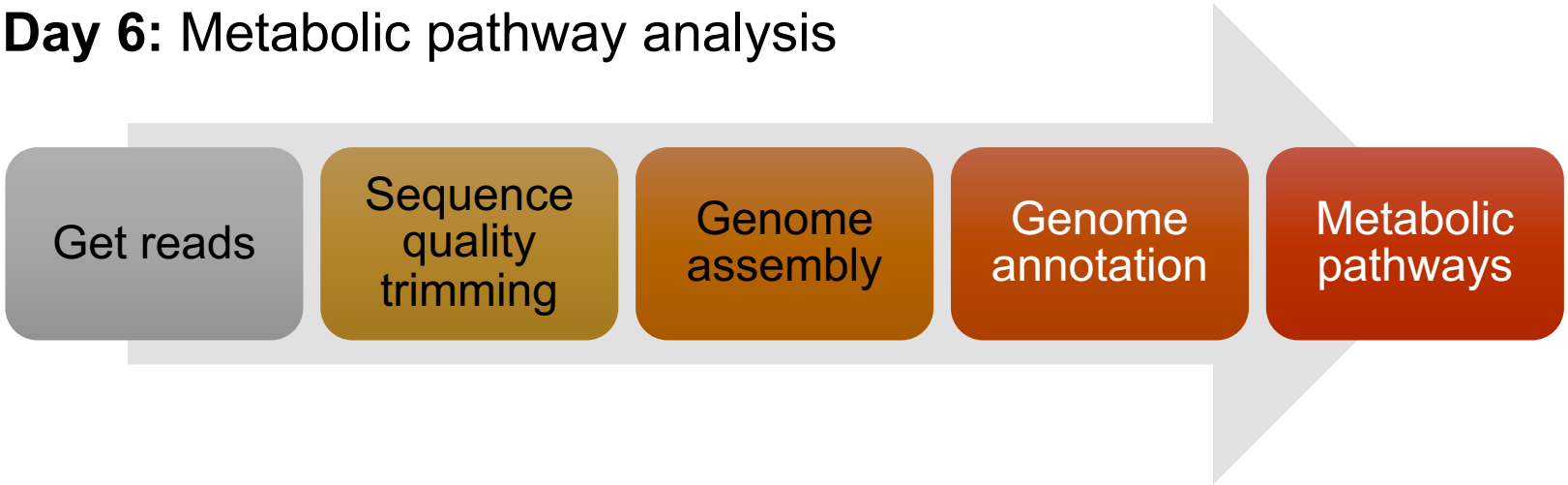
Day 2: Handling of Illumina data

Day 3: Genome assembly

Day 4: Check-up and report

Day 5: Genome annotation

Day 6: Metabolic pathway analysis



Learning outcomes

After completing this module, you will be able to:

Choose the most adequate platform for your genome sequencing experiment

Investigate and judge the quality of sequencing data

Make use of a variety of tools to:

- Process genome data
- Annotate genomes
- Predict metabolic pathways

Some practical things

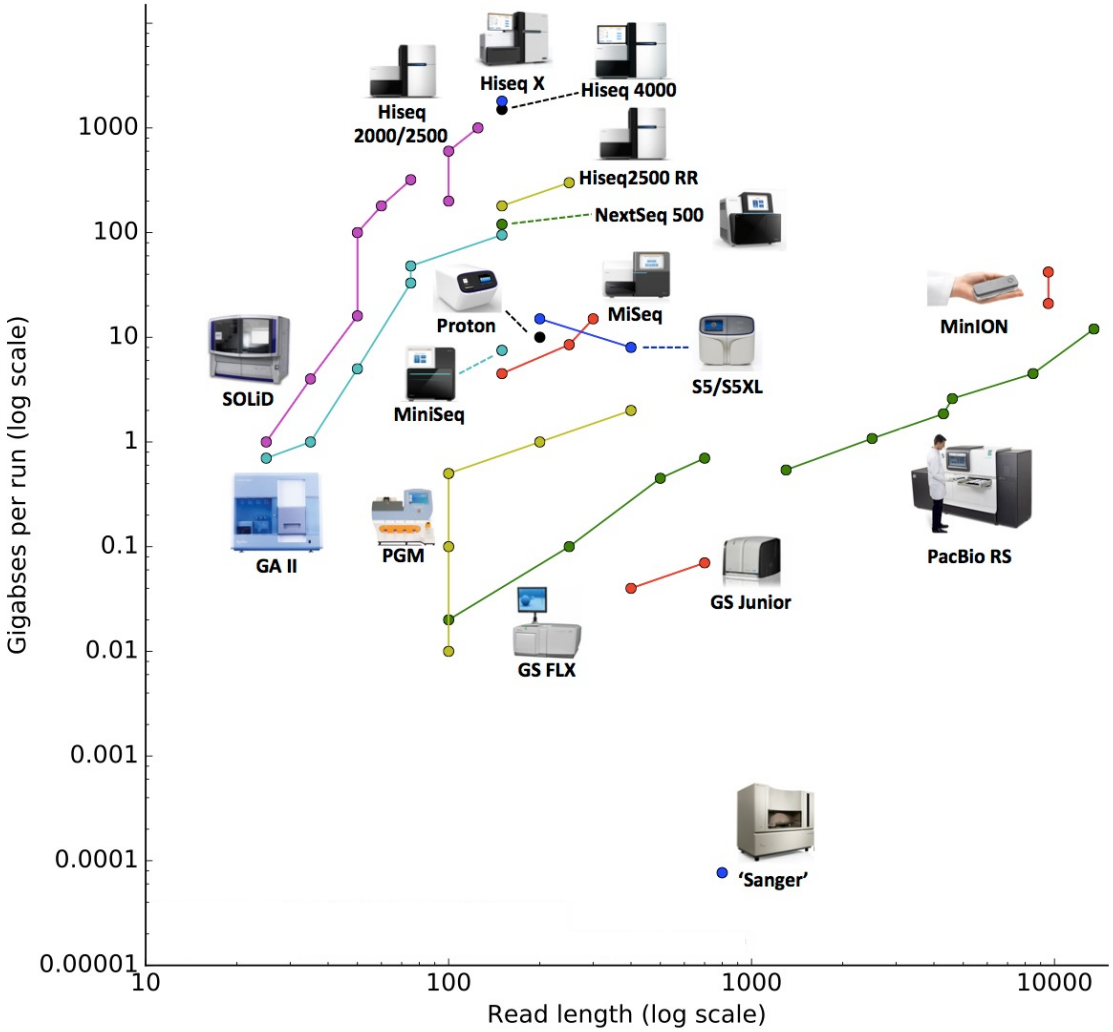
All exercises and presentations can be found in:

https://github.com/igorspp/MMB-114

We will start everyday at 10.00 with an introductory lecture

And then some exercises

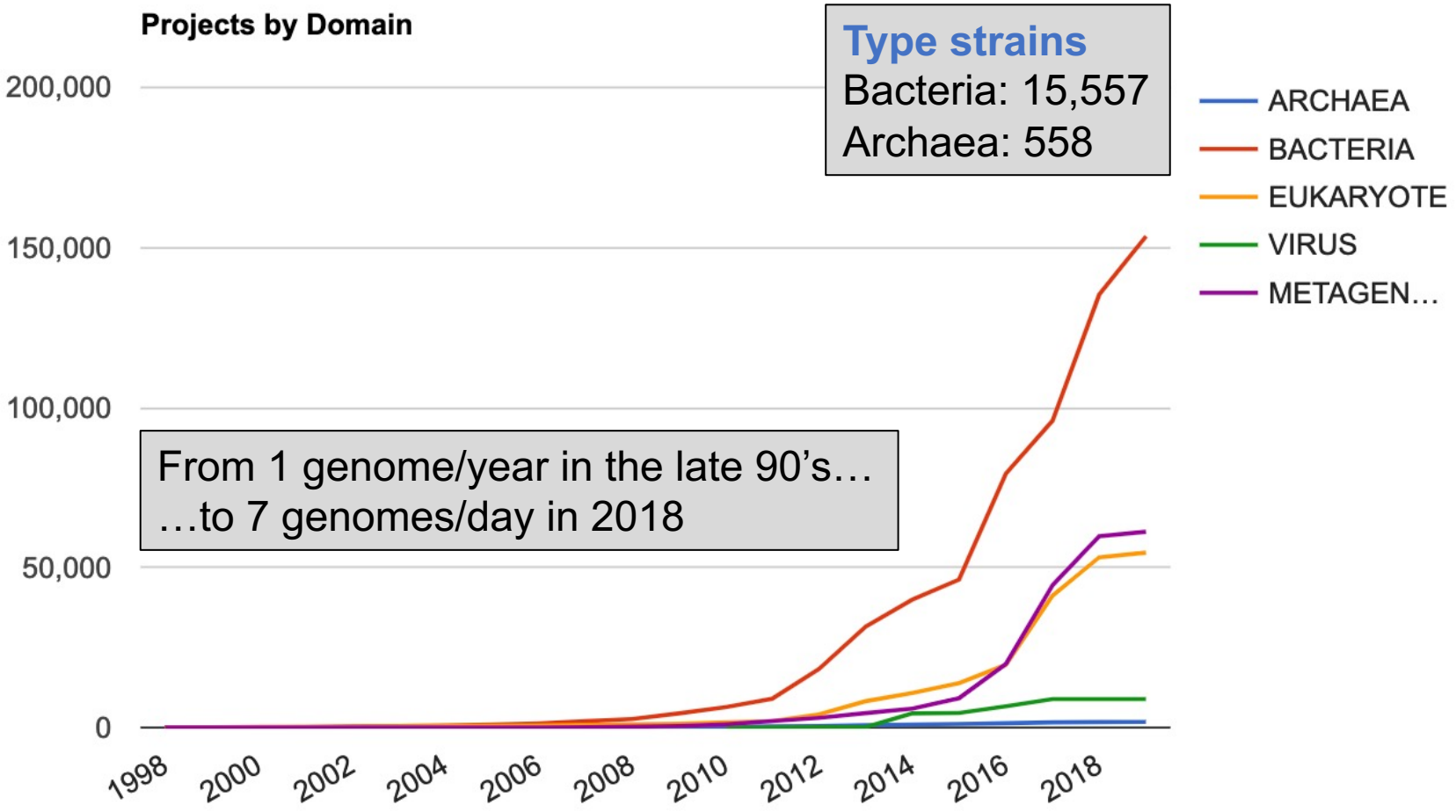
The evolution of sequencing



Features
Read size
Read depth (output)
Error rates
Price

http://dx.doi.org/10.6084/m9.figshare.100940

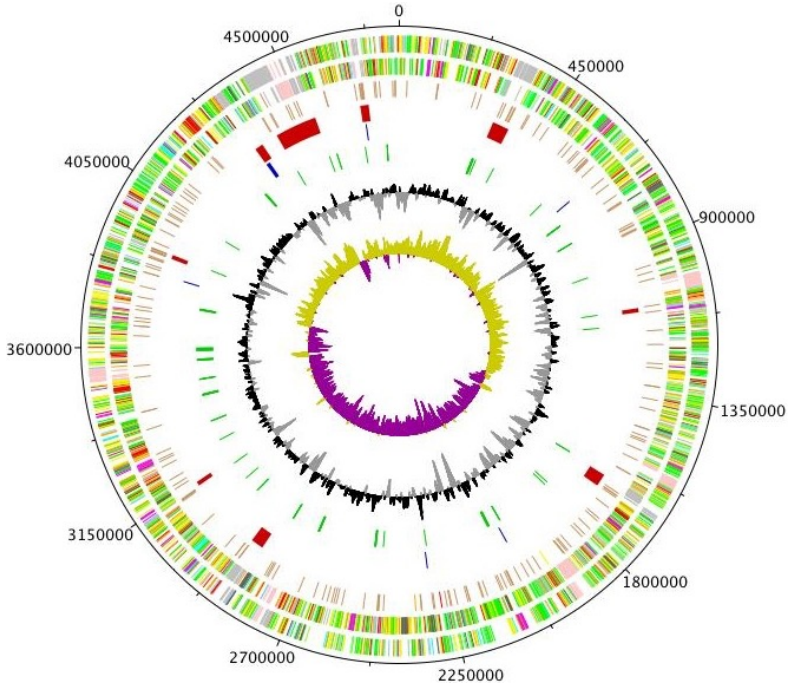
Genomes OnLine (GOLD) database



https://gold.jgi.doe.gov

Genomics can be used to study

- Physiology
- Evolution
- Pathogenesis
- Novel industrial processes
- Novel diagnostic/ epidemiologic tests



Ammonia-oxidizing archaea

Nitrifiers = obtain energy from the oxidation of ammonia

Initially thought thought to be performed mainly by bacteria

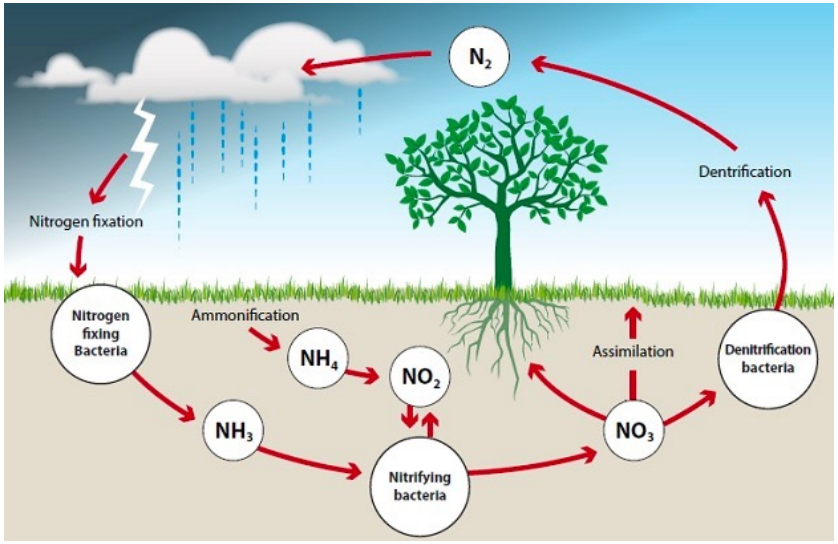
Recent genomics studies have revealed the importance of ammonia-oxidizing archaea

Thaumarchaeota

Novel gene inventory

Higher substrate affinity

Different pathways

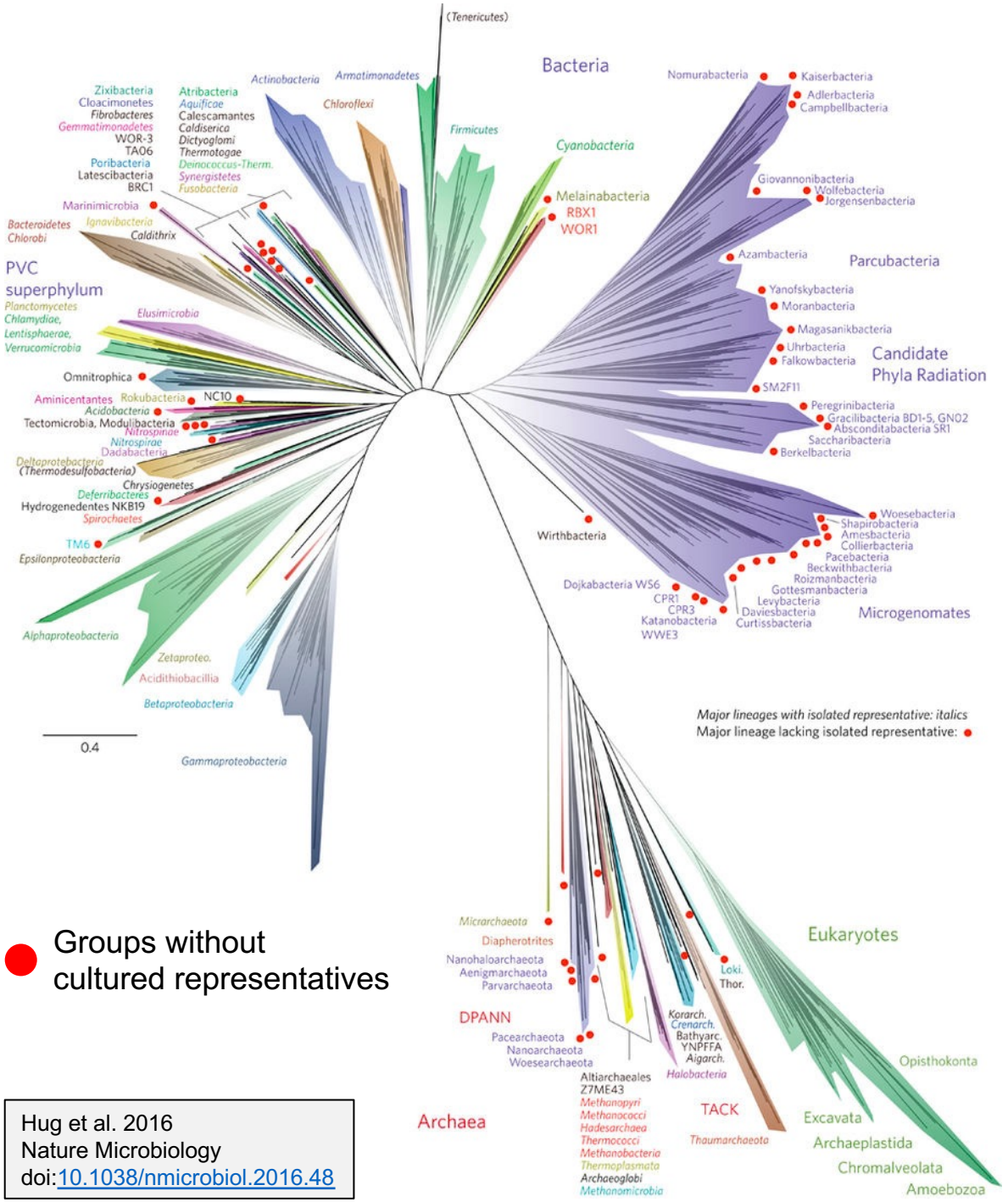


The (new) microbial tree of life

Isolation of most microorganisms from the environment is not trivial

Recent metagenomic assessments suggest the existence of up to 1 trillion microbial species in our planet

But < 10,000 bacterial species have been described so far



● Groups without cultured representatives

Hug et al. 2016
Nature Microbiology
doi:10.1038/nmicrobiol.2016.48

Whole genome sequencing in diagnostics and epidemiology

Pathogen identification and outbreak monitoring

Several phenotypic and molecular methods are currently used

Fail to distinguish closely related strains or detect virulence/resistance features

WGS soon will become the standard test in diagnostics and epidemiology

Increased speed and accuracy

Reduced cost and difficulty

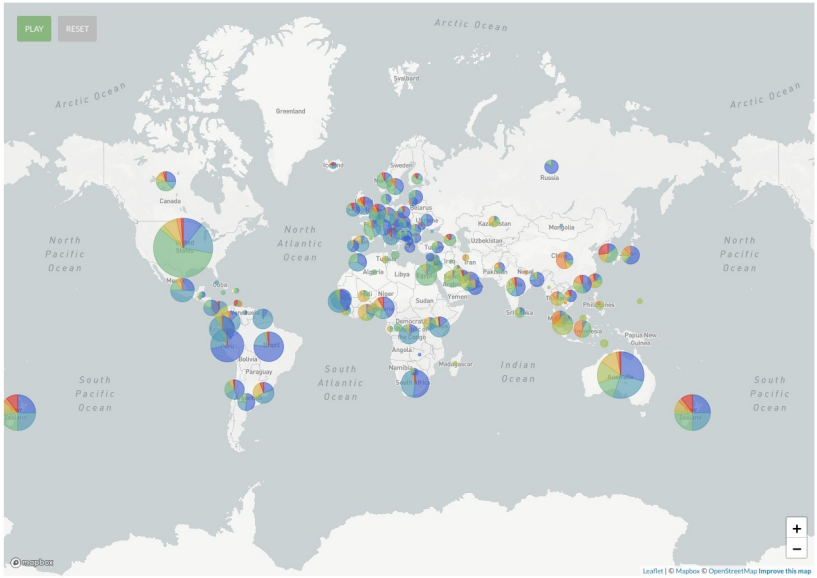
Trial phase of implementation in many countries

COVID-19

Tracking infections and outbreaks (also reinfections)

155,000 genomes sequenced

- https://www.gisaid.org



Your strains

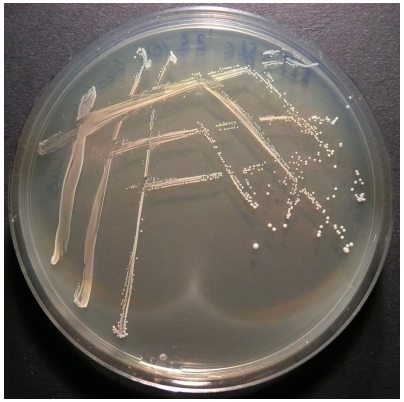
What was done by you in the lab?

Isolation/purification of the microorganism

DNA extraction

Important points?

- Purity, integrity, contamination DNA

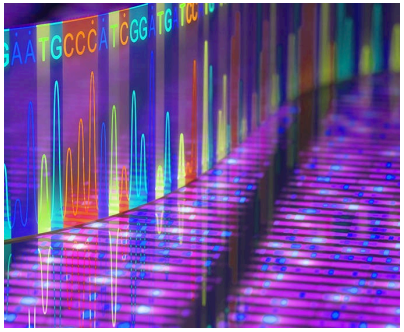


What was done in the sequencing facility?

Quality control (size, amount)

Library preparation with the Nextera kit

Sequencing with Illumina MiSeq



Illumina library preparation and sequencing

Illumina MiSeq and HiSeq

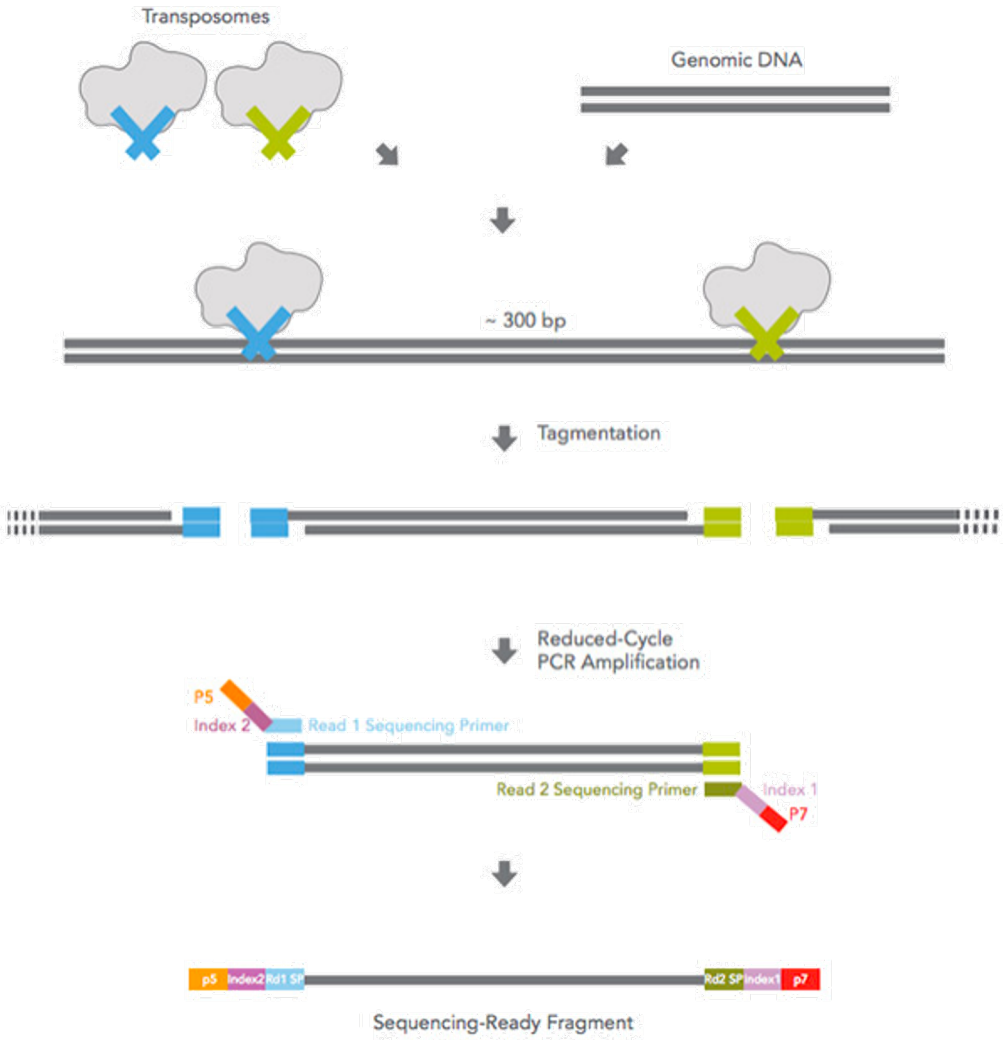
Good compromise between size, amount and error rate of reads

The longer the reads the better, but current long-read technologies produce too many errors

Let’s watch a video together and then we will discuss some features in more detail:

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Library preparation: Nextera tagmentation



Size selection

Correct amount of input DNA to avoid under- and overtagmentation

Adapter ligation

Flow cell adapters

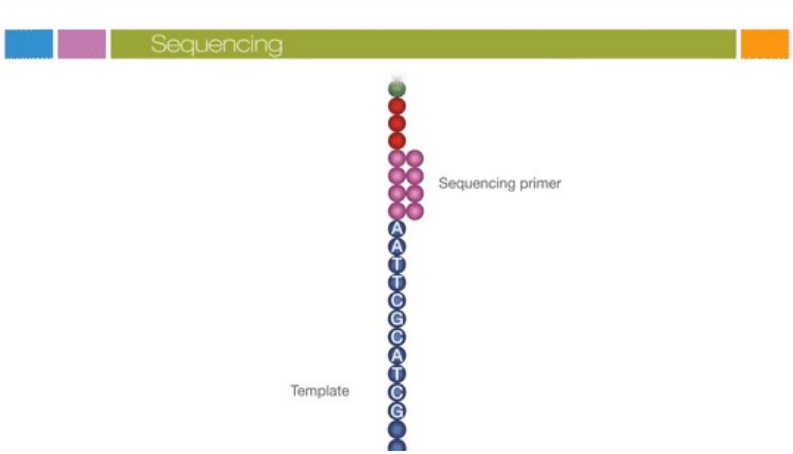
Sequencing primers

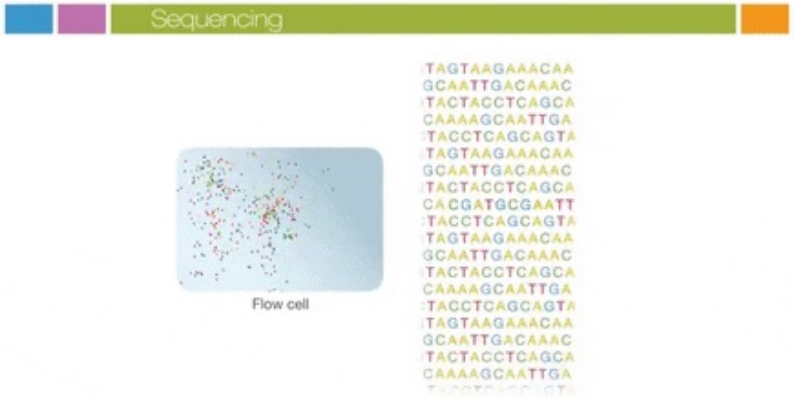
Sequencing indexes

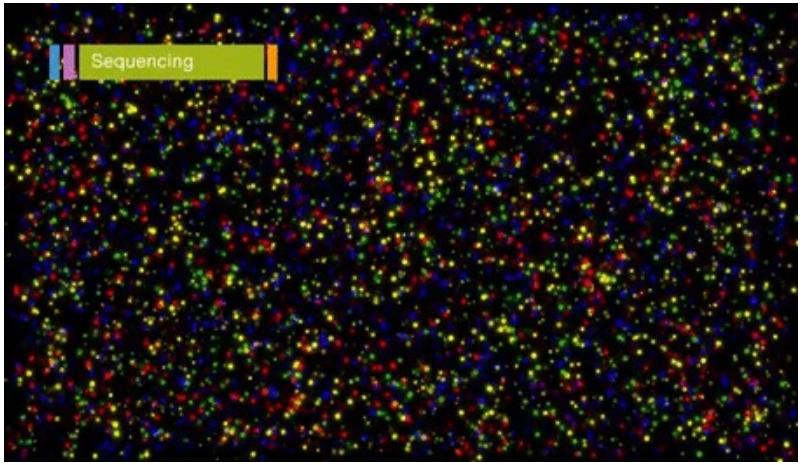
Optional indexes for multiplexing

Biases in the first bases have been observed

Sequencing

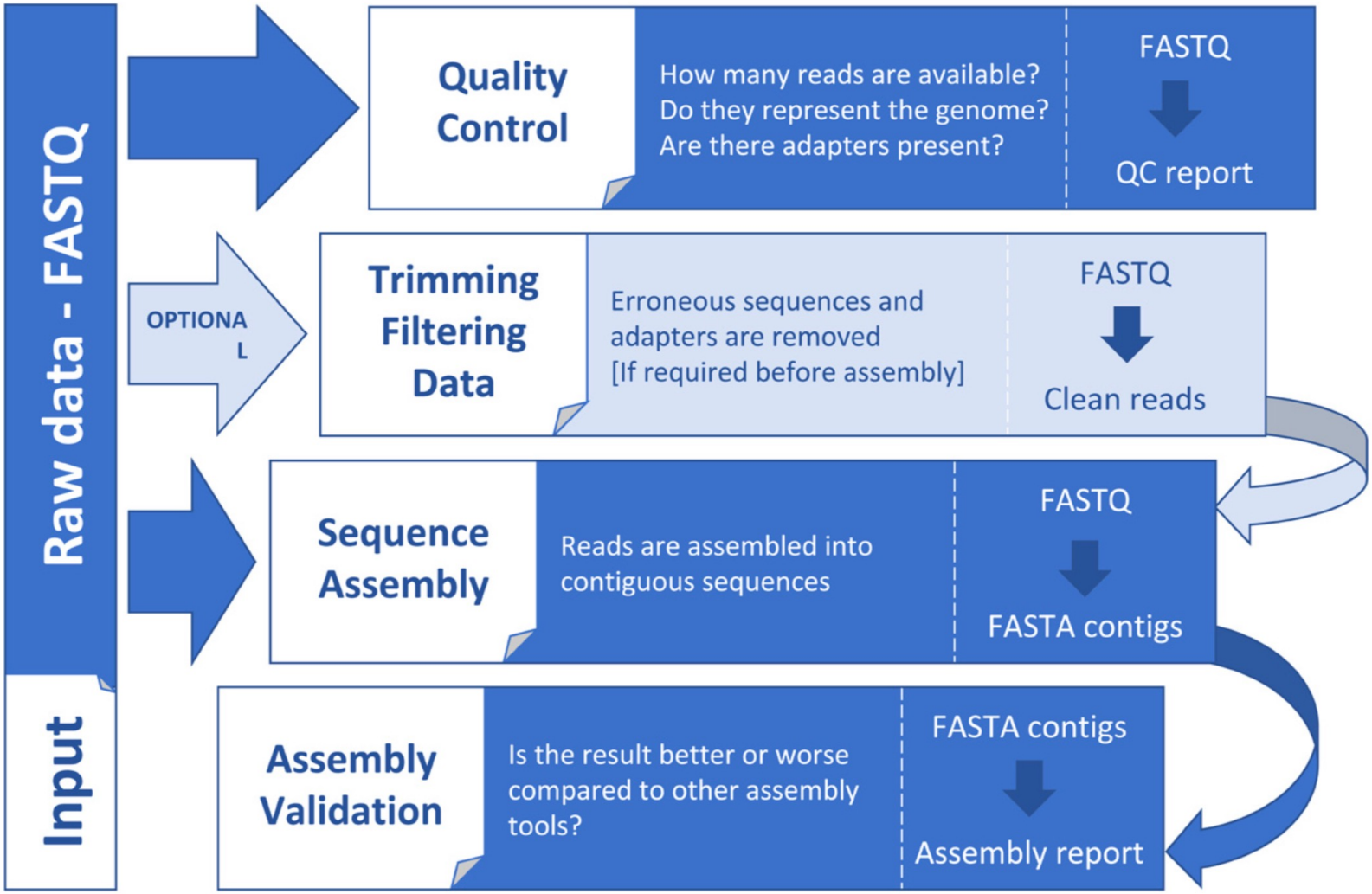

Sequencing
Sequencing primer
Template


Sequencing
Flow cell


Sequencing

Dominguez Del Angel et al. 2018
F1000Reseach
doi: 10.12688/f1000research.13598.1

What now?



Bioinformatics

Interdisciplinary field that develops and applies computational methods to analyse biological data such as DNA and protein sequences

Biology meets information science

A rapidly-evolving field

Software development

Amount and type of data generated

UNIX


A family of computer operating systems (OSs)

- Linux, MacOS, Solaris, OpenBSD

Key characteristics:

Multitasking

- Multiple software processes can run at the same time

Multiuser

- Several users can use the same computer at the same time

Multiprocessing

- Capable of supporting and utilizing more than one computer processor

Portable

- Can be used in various hardware architectures

The UNIX philosophy

“The idea that the power of a system comes more from the relationships among programs than from the programs themselves”

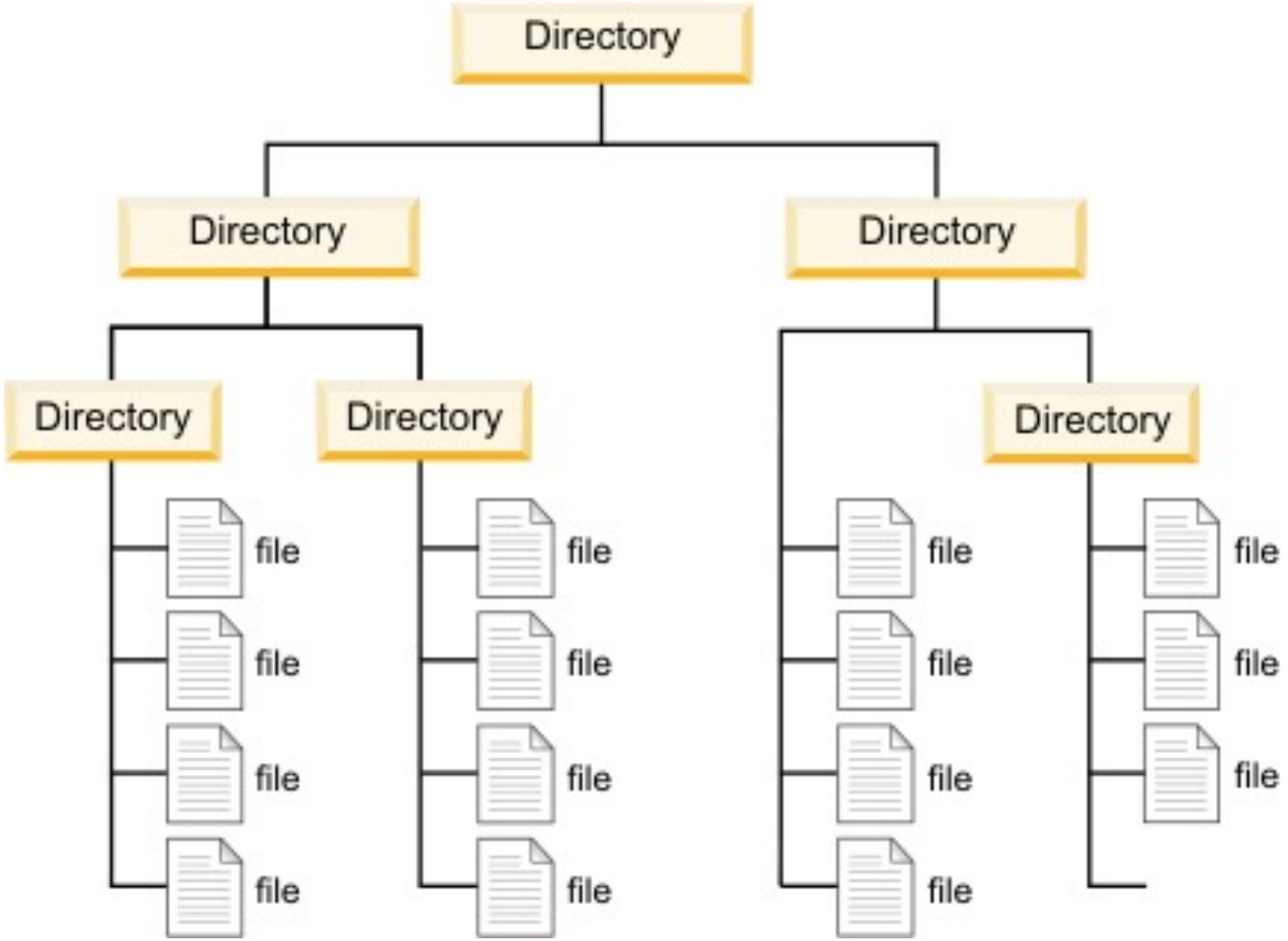
Use of plain text for storing data

Use of a hierarchical filesystem

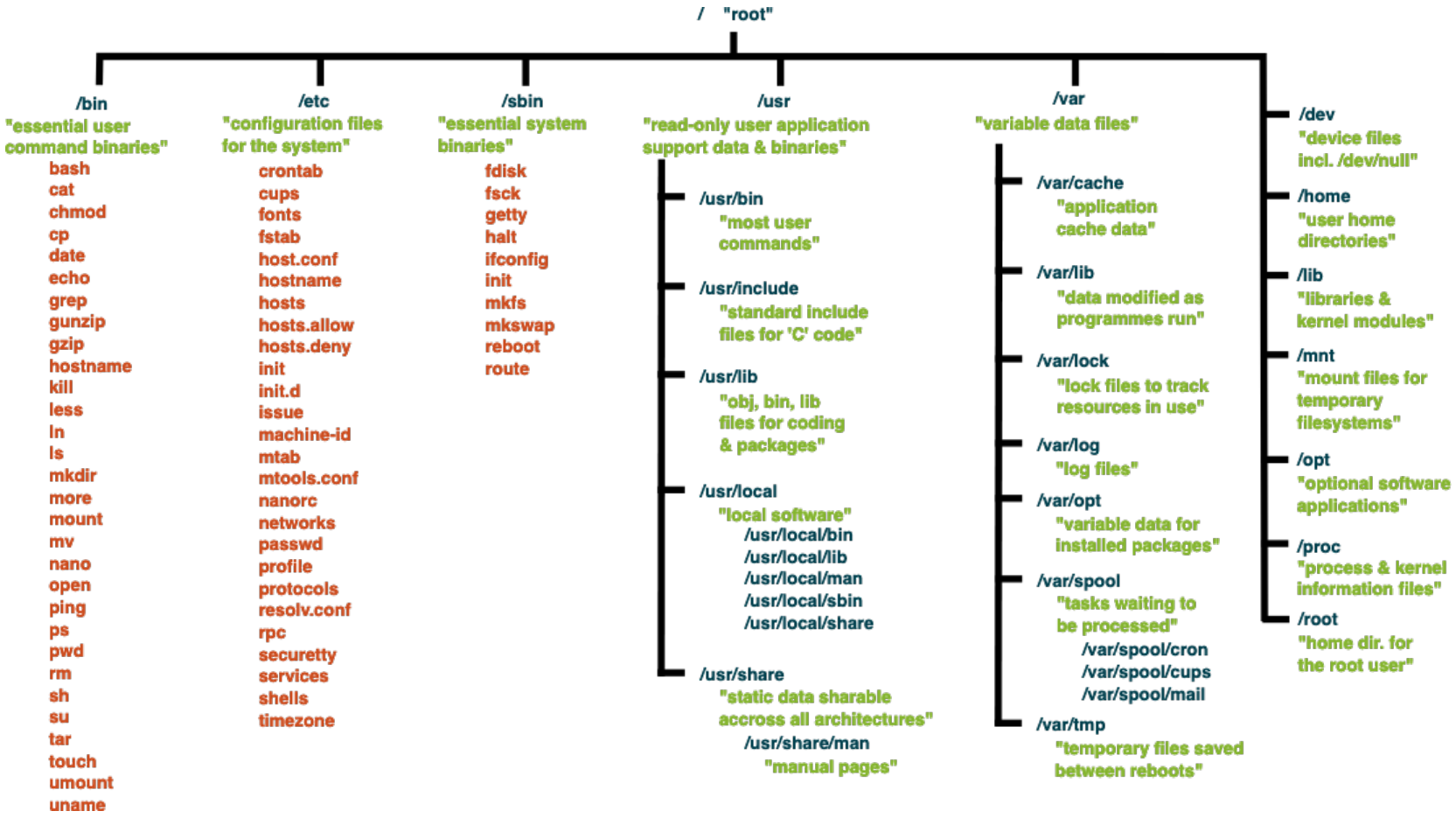
Use of a large number of simple programs performing a limited, well-defined function

Use of a command-line interpreter (“shell”) to combine these programs to perform complex tasks

The UNIX hierarchical filesystem

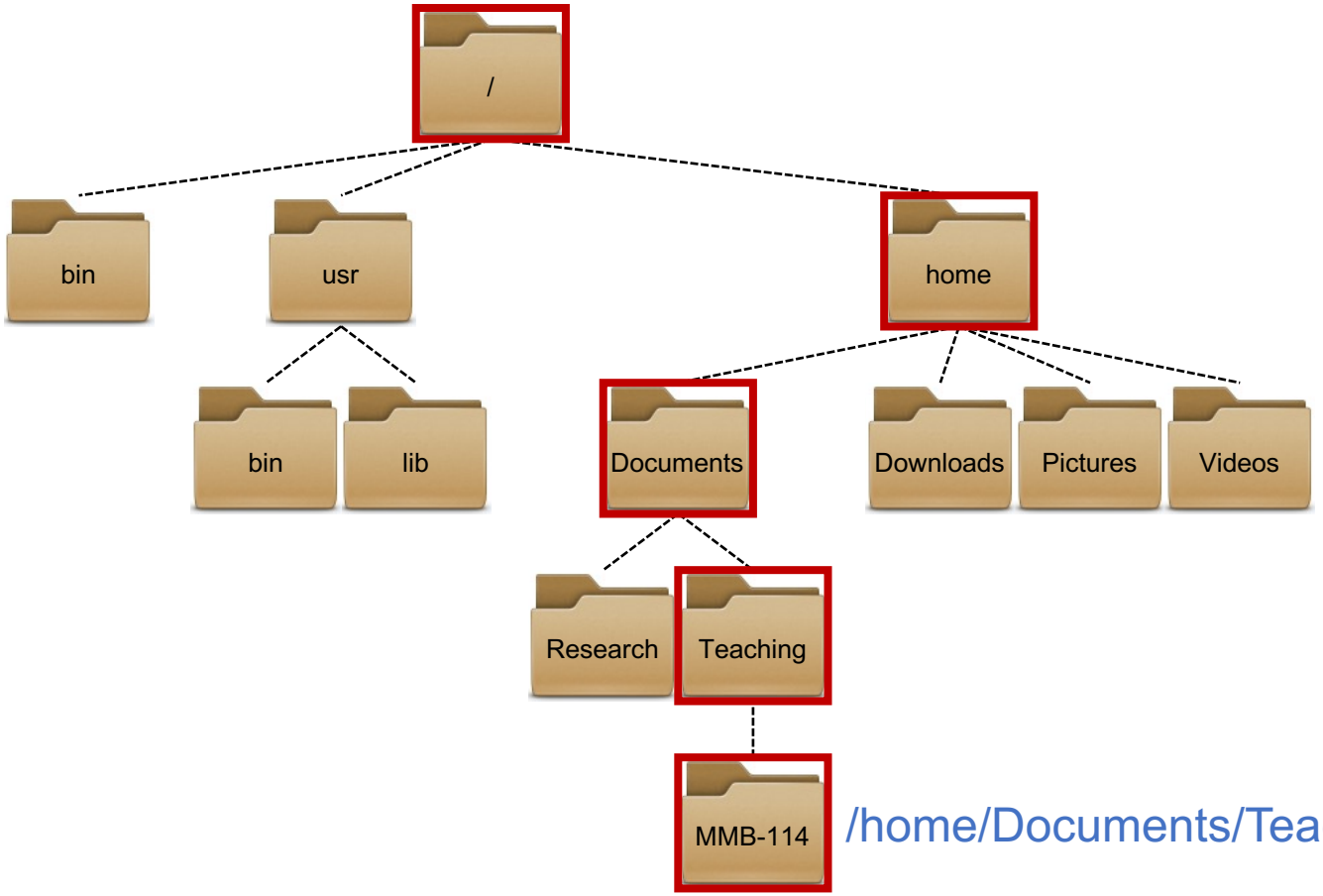


The UNIX hierarchical filesystem



https://www.linux.com/sites/lcom/files/styles/rendered_file/public/standard-unix-filesystem-hierarchy.png?itok=CVqmyk6P

The UNIX hierarchical filesystem



/home/Documents/Teaching/MMB-114

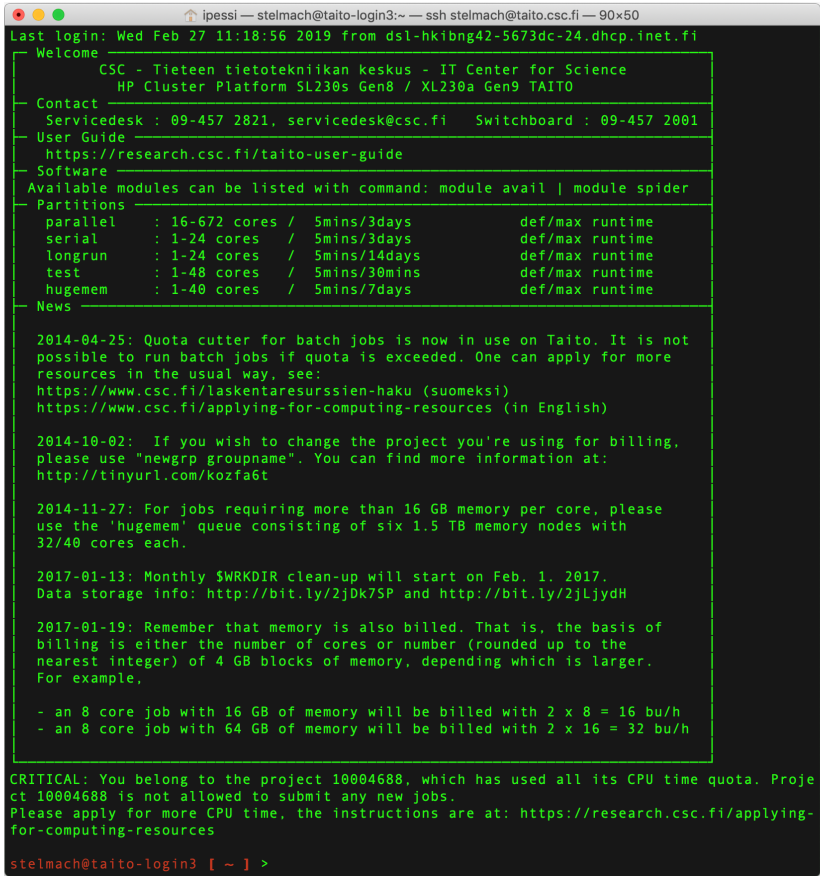
The UNIX shell

Command-line interpreter

Interprets sequences of text
Entered by a user
From a file
From a data stream

Primary interface before graphical user interfaces (GUIs) appeared

Still widely used today
Efficient
Low memory footprint
Advanced scripting



Some basic UNIX commands

`pwd:` **p**rint **w**orking **d**irectory (“where am I?”)

`ls:` **l**i**s**t (“show folder contents”)

`mkdir:` **m**ake **dir**ectory (a.k.a. folder)

`cd:` **c**hange **d**irectory (“go to folder”)

`cp:` **c**o**p**y

`mv:` **m**o**v**e

`rm:` re**m**ove

Some additional notes

Case-sensitive

`photo.jpg ≠ PHOTO.jpg`

Does not like spaces and special characters in file/folder names

`genome report.txt` ❌

`genome_report.txt` ✅

`väitöskirja.txt` ❌

`vaitoskirja.txt` ✅

Space after each “word” in the command

Commands have to be typed in a single line, one at a time

After each command, hit “Enter” to execute it

Lines starting with “#” are comments

A few tricks:

Tabulator

History

How to learn UNIX?

By using it!

Trial and error

Don't copy and paste it, type it

Ask the internet

http://stackoverflow.com/

http://stackexchange.com/

http://askubuntu.com/

Google!

Online courses/tutorials

http://codecademy.com

Cheat sheets

http://www.mathcs.emory.edu/~valerie/courses/fall10/155/resources/unix_cheatsheet.html

Manual ("man") pages

```
man cutadapt
```

CSC - IT CENTER FOR SCIENCE

Working with ‘omics data is computing-intensive

Use of high-performance computers (HPC)

- a.k.a. supercomputers

Before we start, check if everything is OK in CSC

1. Login to https://my.csc.fi/
2. Go to My Projects
3. Click on “MMB-114 Exploratory microbial research - lab course 2019”
4. Scroll down to “Services” and see if access to Puhti is enabled

Connecting to Puhti: Windows users

1. Launch PuTTY
2. In “Host Name (or IP address)”, type **puhti.csc.fi** and click “Open”
3. In the following dialogue window, choose “Yes”
4. Type your CSC username and hit “Enter”
5. Type your password and hit “Enter”
6. To logout just type **exit** and hit “Enter”

Connecting to Puhti: MacOS users

1. Launch Terminal
   (e.g. open the Launchpad and type **terminal**)
2. Type **ssh user@taito.csc.fi** and hit “Enter”
   (change “user” for your own CSC username)
3. In the following dialogue, type **yes** and hit “Enter”
4. Type your password and hit “Enter”
5. To logout first type **exit**, hit “Enter”,
   and close the Terminal window

Time to familiarize yourself with UNIX

https://github.com/igorspp/MMB-114

(**Day 1:** UNIX and CSC)