

# Genome assembly and annotation

**Day 1: UNIX and CSC**

**Antti Karkman**

Department of Microbiology – UH

[antti.karkman@helsinki.fi](mailto:antti.karkman@helsinki.fi)

# Aims for this part of MMB-114

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

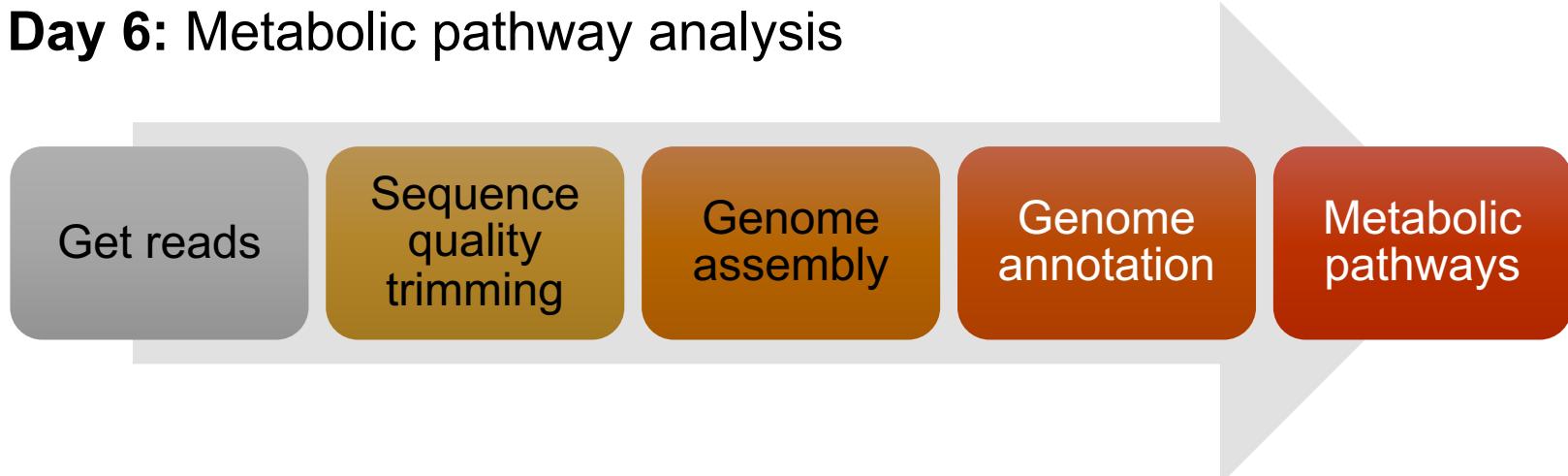
**Day 2:** Handling of Illumina data

**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 whole genome sequencing data
  - Assemble and annotate whole genome sequencing data
  - Predict metabolic pathways from assembled and annotated genomes

# **Some practical things**

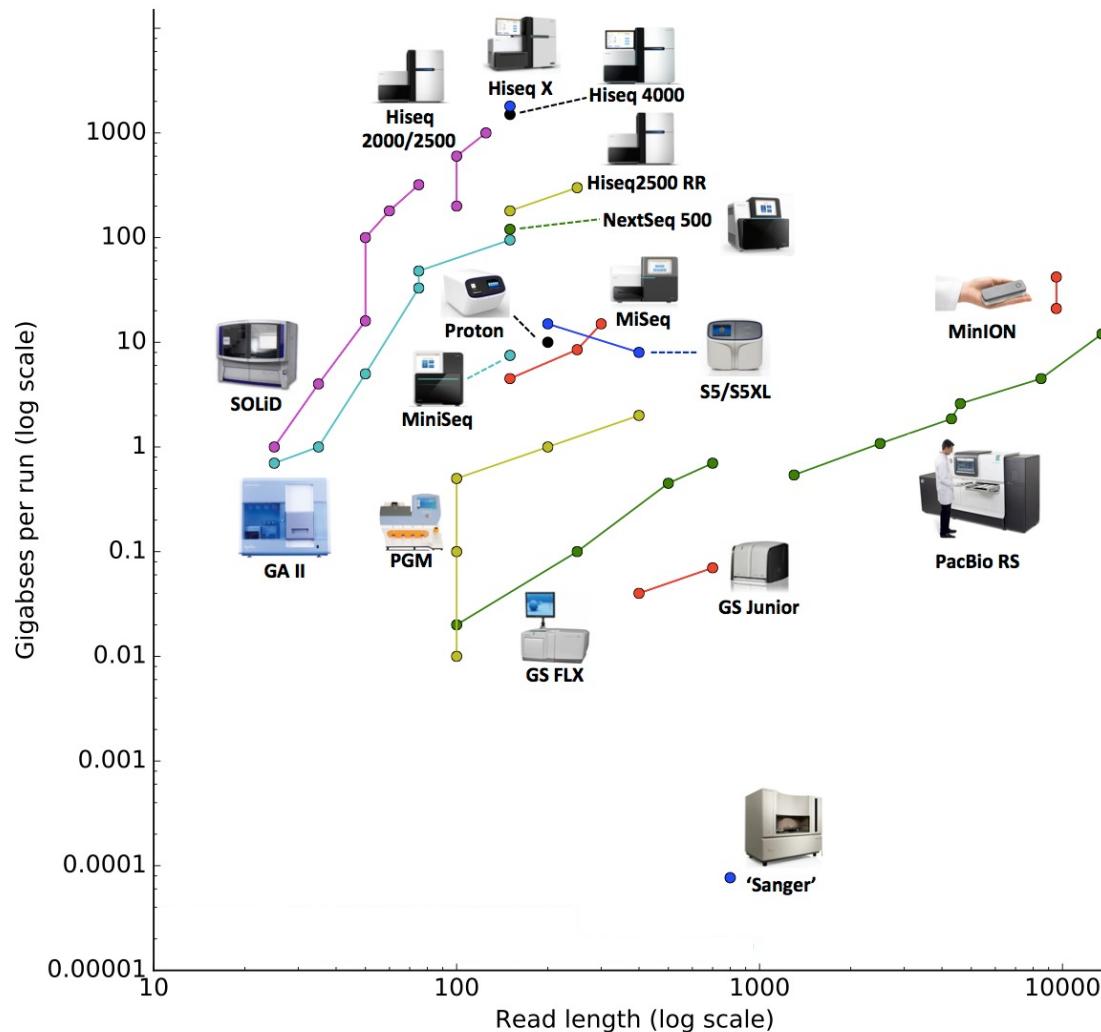
**All exercises and presentations can be found in:**

[https://github.com/karkman/MMB-114\\_Genomics](https://github.com/karkman/MMB-114_Genomics)

We will start everyday at 10.00 (Wed at 12.00) with an introductory lecture

And then some exercises

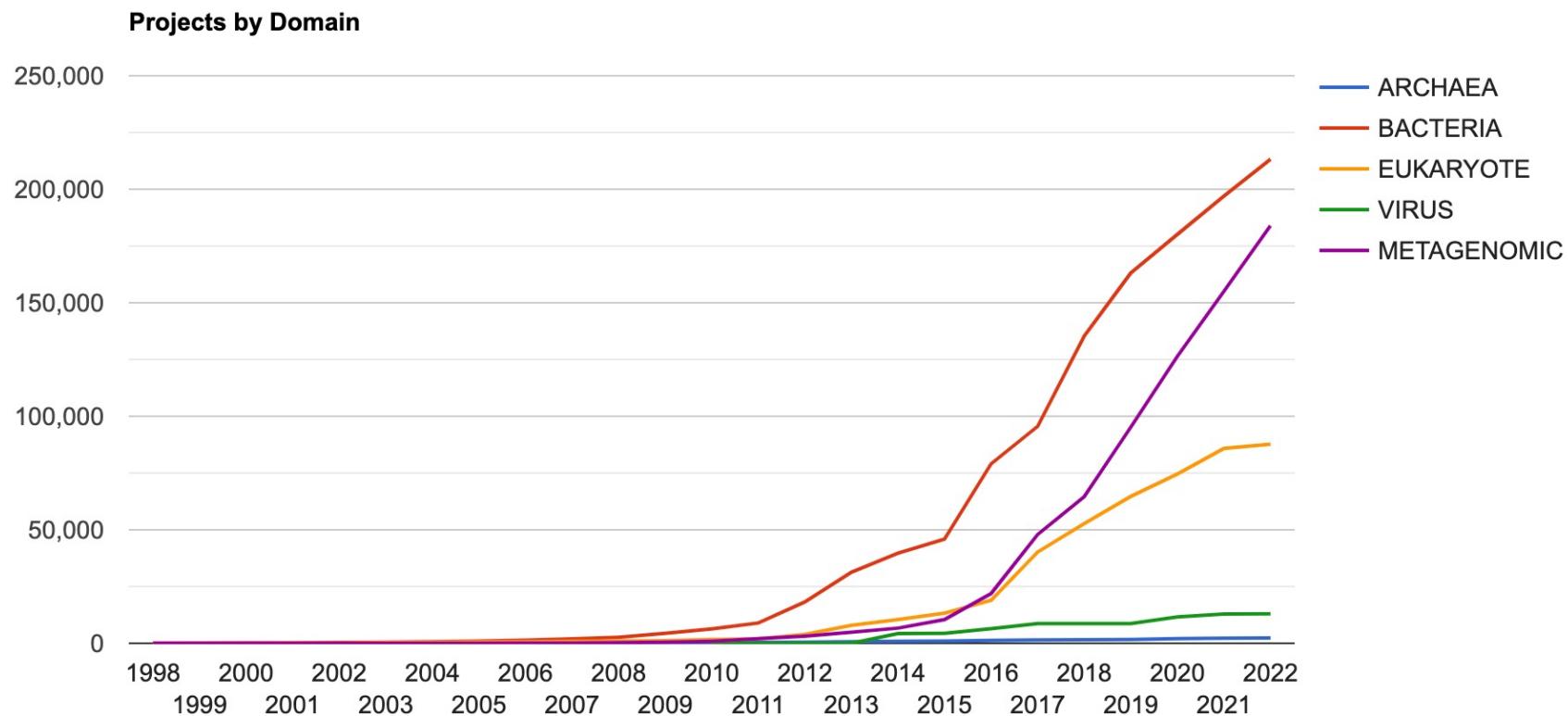
# The evolution of sequencing (2016)



## Features

- Read length
- Amount of data
- Error rate
- Price (per Gb)

# Genomes OnLine (GOLD) database



# **Genomics can be used to study**

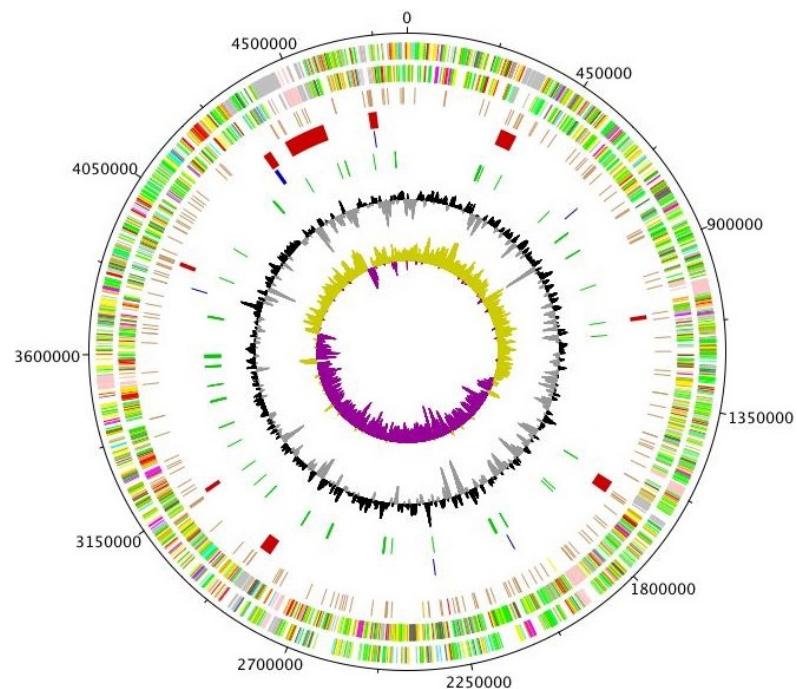
Physiology

Evolution

Pathogenesis

Novel industrial processes

Novel diagnostic/  
epidemiologic tests

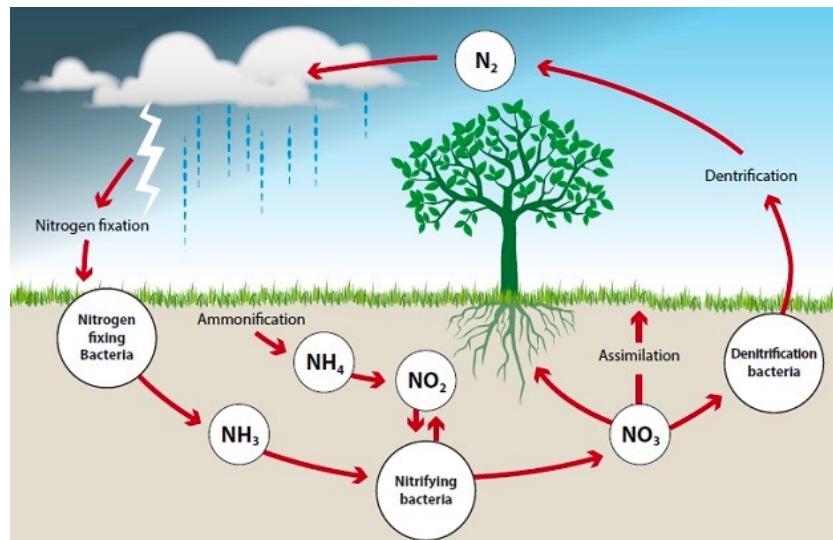


# Ammonia-oxidizing archaea

Nitrifiers = obtain energy from the oxidation of ammonia

Initially thought to be performed mainly by bacteria

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

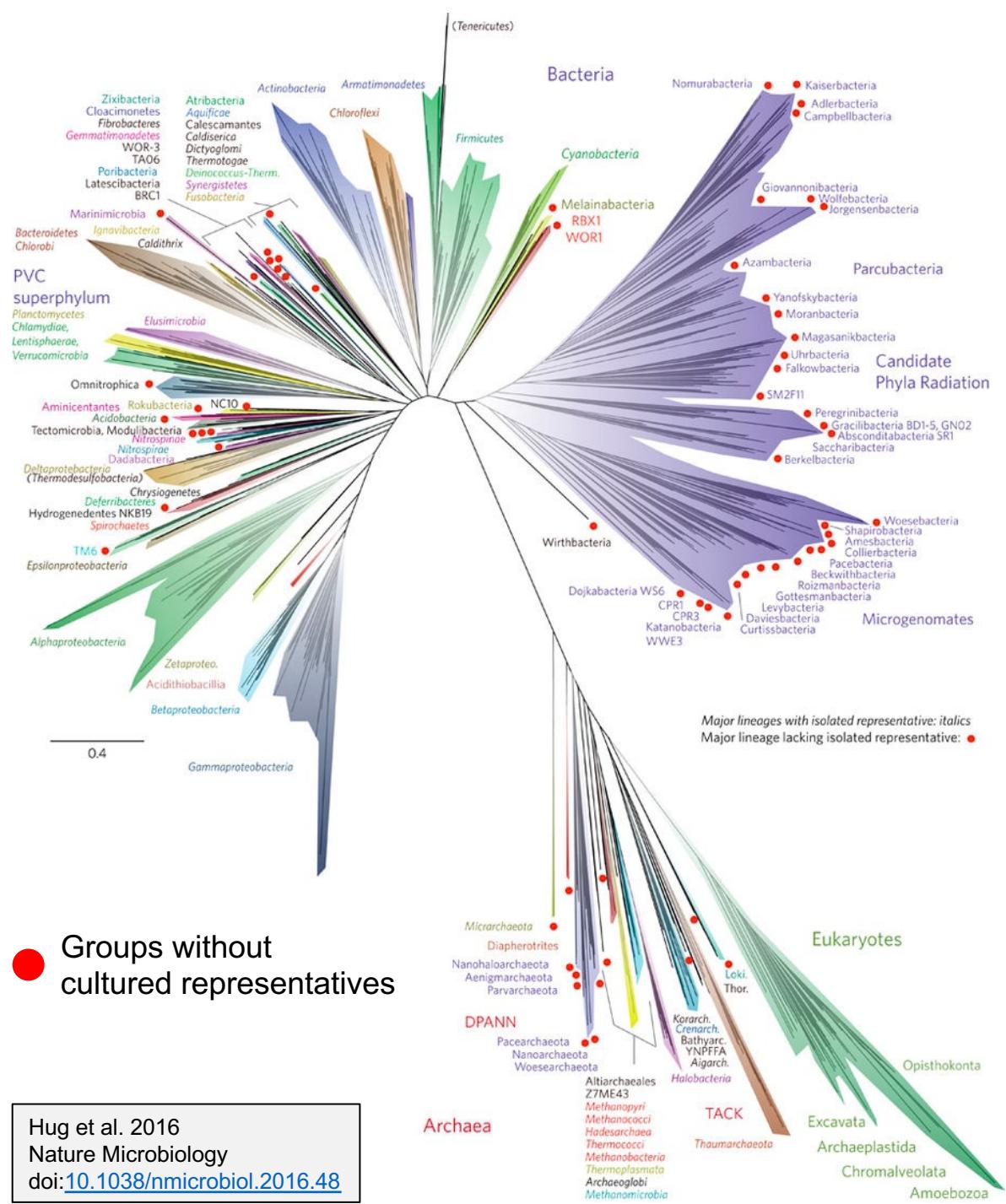


# The (new) microbial tree of life

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



# 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)  
**13 611 570** genomes sequenced (last year: 155 000)

- <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

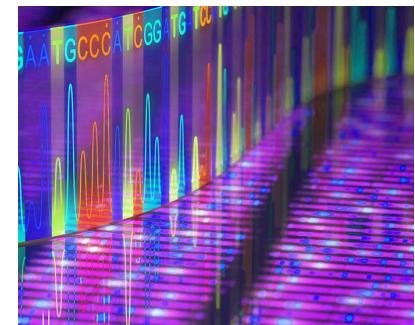


## What was done in the sequencing facility?

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

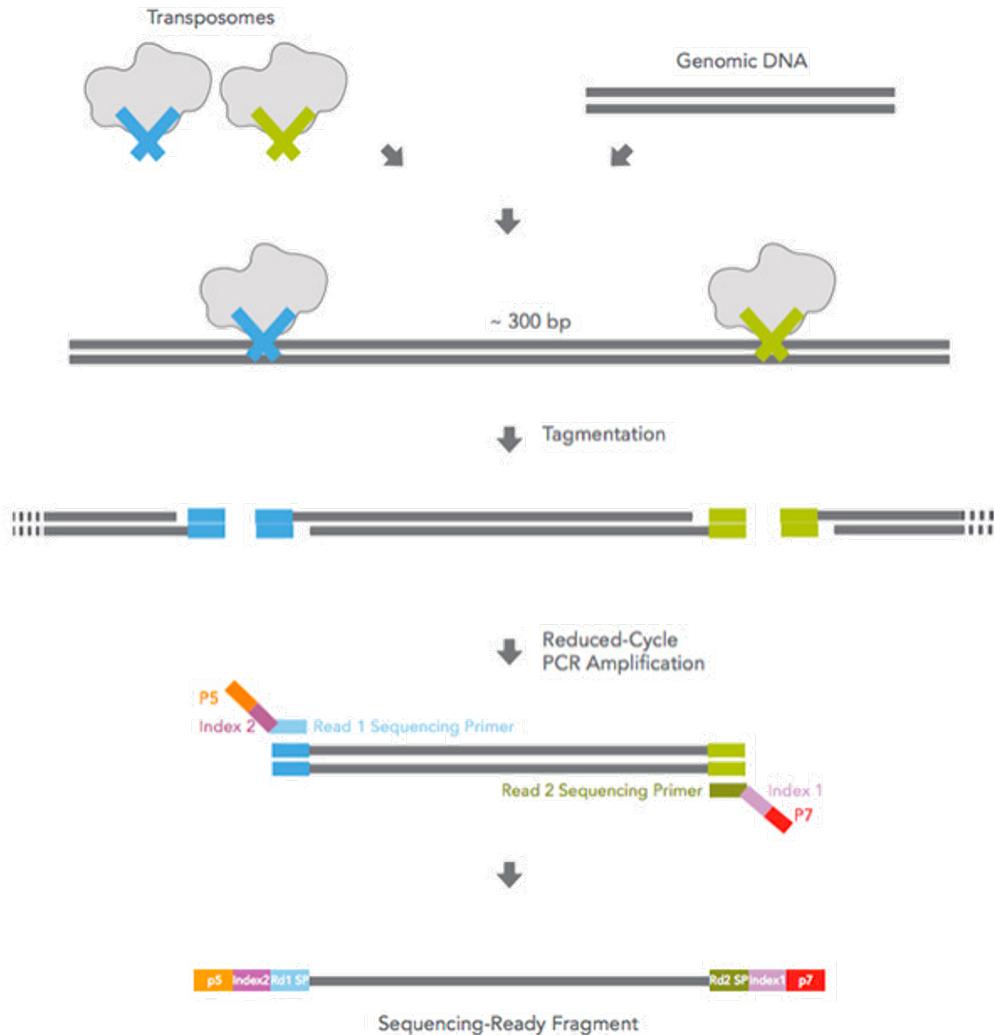
The longer the reads the better.

Current long-read technologies are becoming more relevant due to lower price and better sequence quality

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

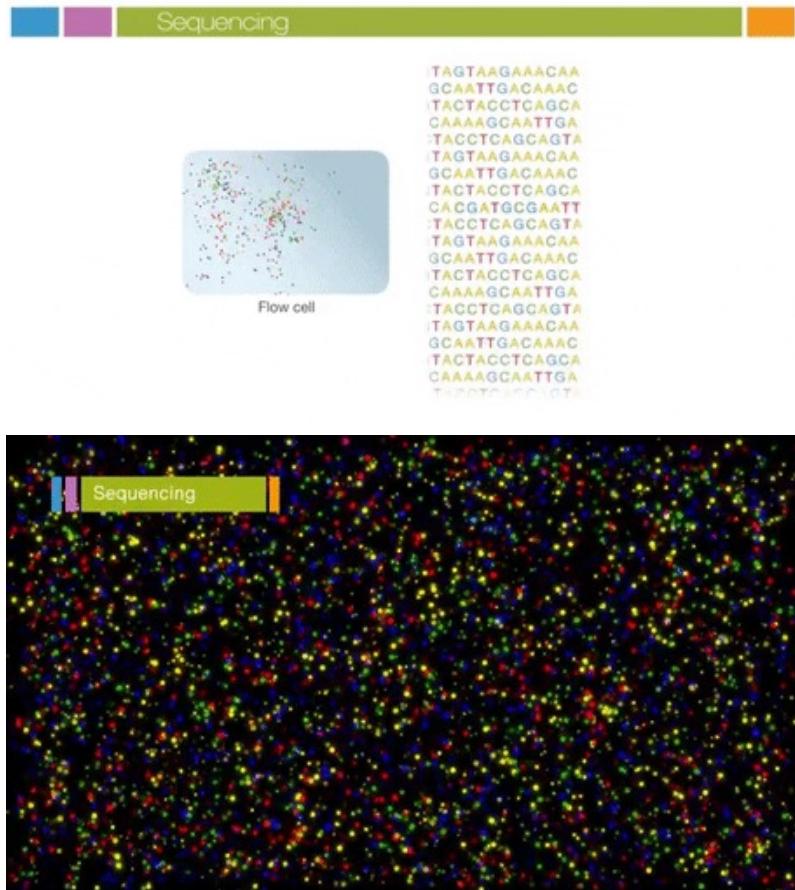
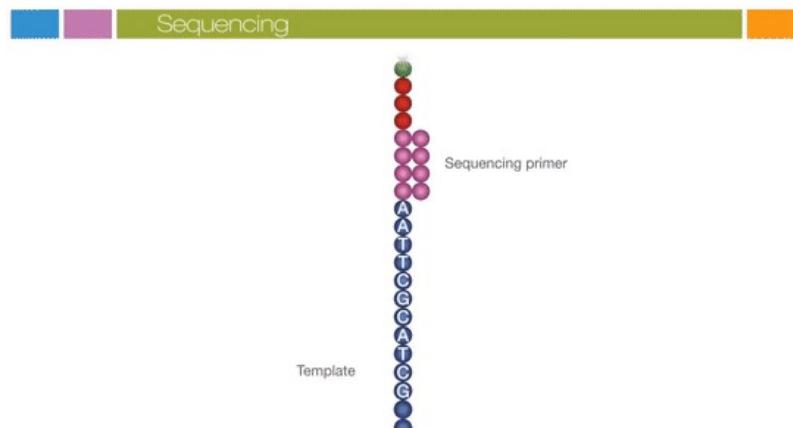
Correct amount of input DNA to avoid under- and overtagmentation

## Adapter ligation

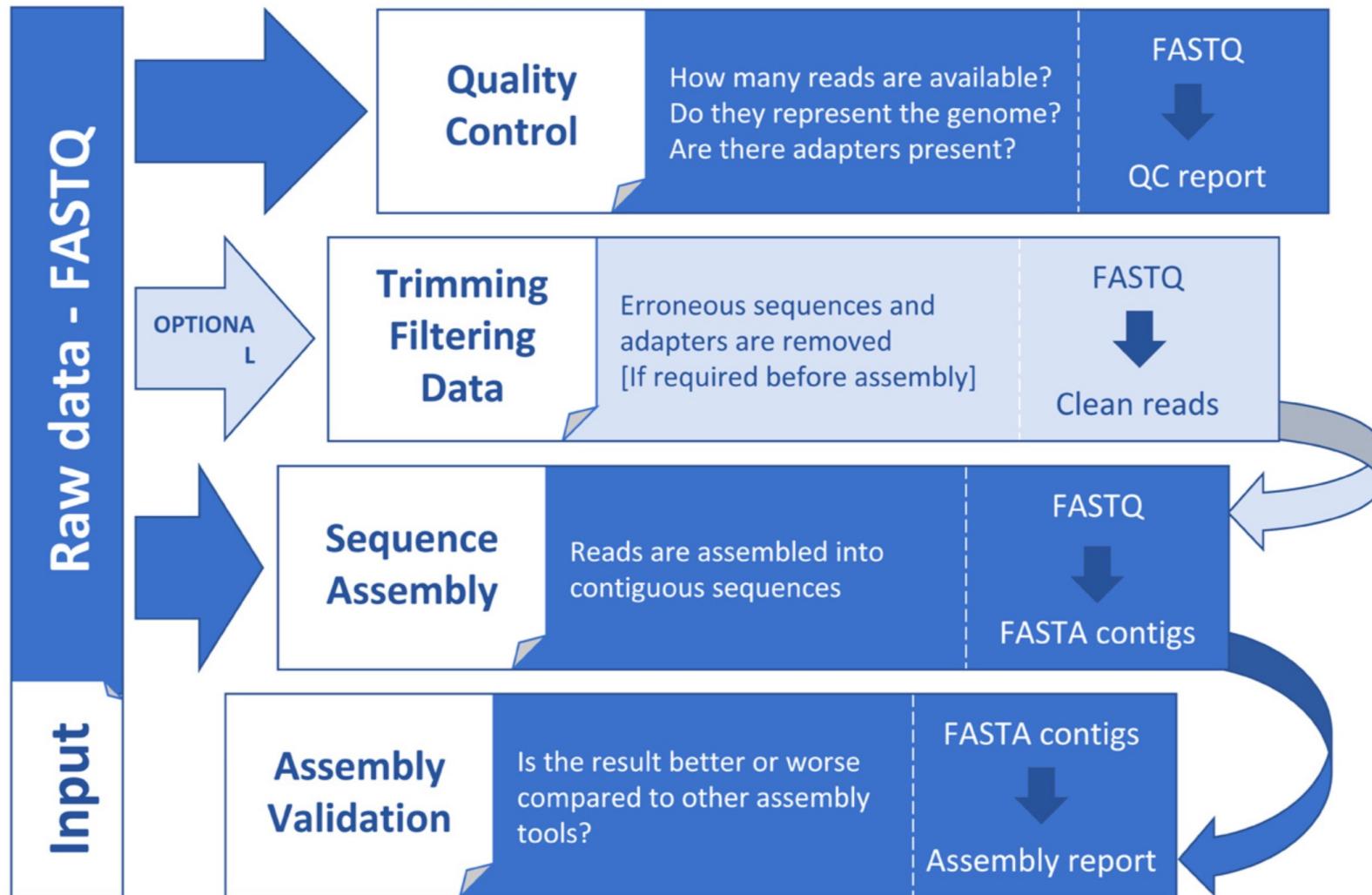
Flow cell adapters  
Sequencing primers  
Sequencing indexes  
Optional indexes for multiplexing

Biases in the first bases have been observed

# **Sequencing**



# 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”*

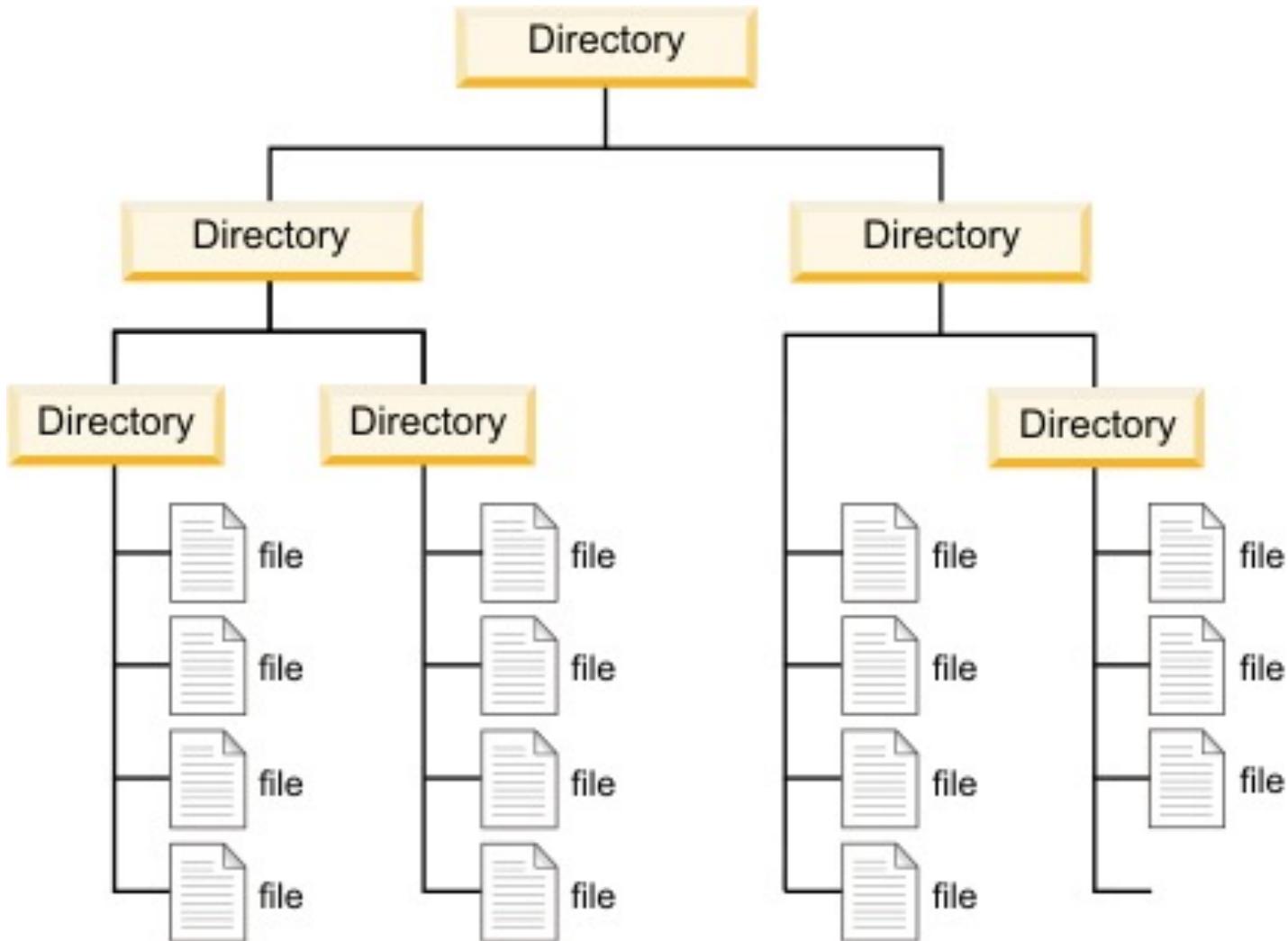
Use of plain text for storing data

Use of a hierarchical filesystem

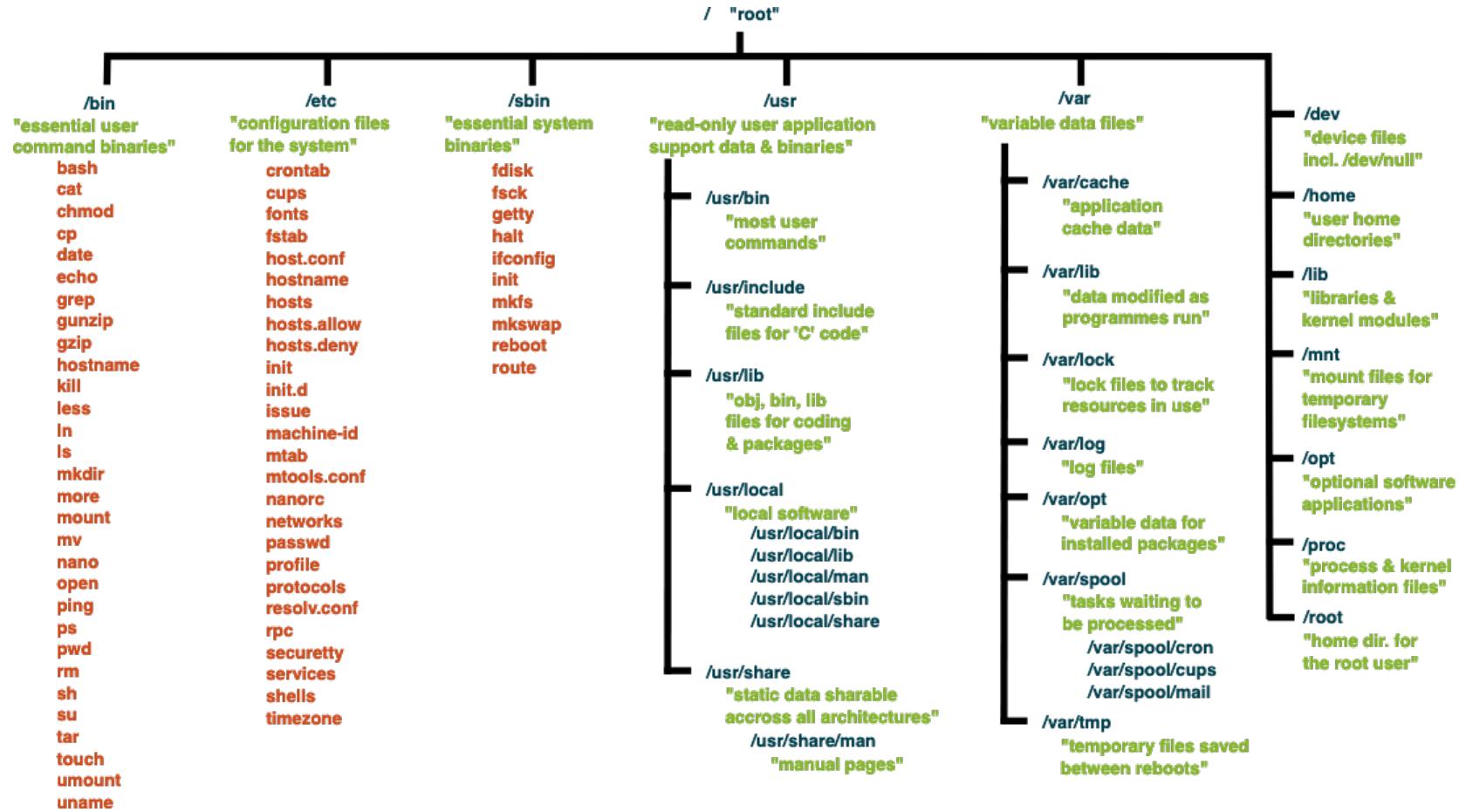
Use of a large number of simple programs performing a limited, well-defined function

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

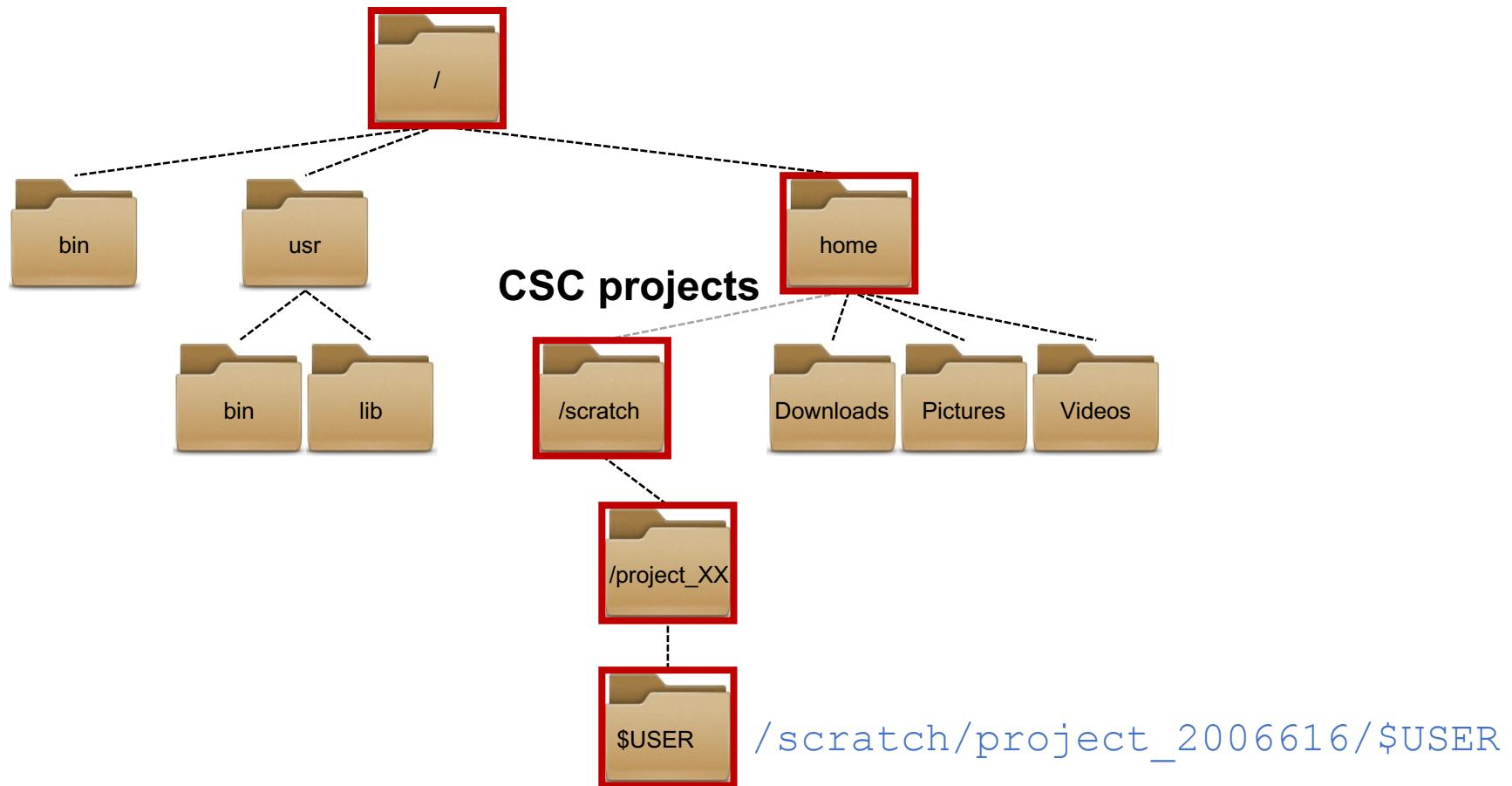
# The UNIX hierarchical filesystem



# The UNIX hierarchical filesystem



# The UNIX hierarchical filesystem



# The UNIX shell

Command-line interpreter

Interprets sequences of text

Entered by a user

From a file

From a data stream

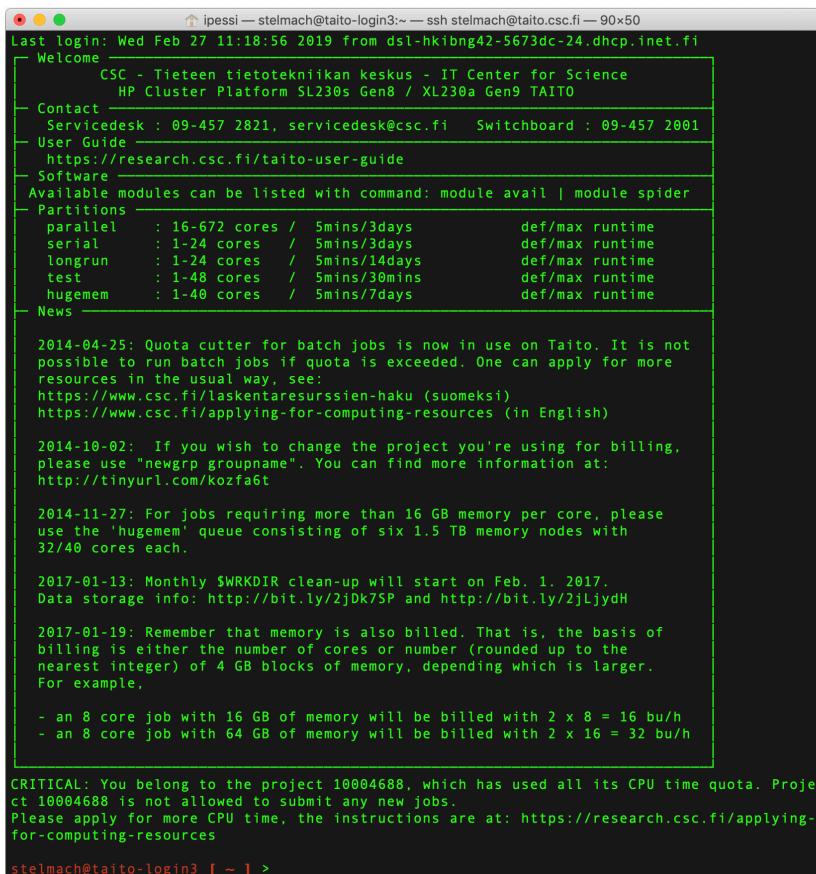
Primary interface before graphical user interfaces (GUIs) appeared

Still widely used today

Efficient

Low memory footprint

Advanced scripting



A screenshot of a terminal window titled 'stelmach@taito-login3:~'. The window shows a shell session with the following content:

```
ipessi — stelmach@taito-login3:~ — ssh stelmach@taito.csc.fi — 90x50
Last login: Wed Feb 27 11:18:56 2019 from dsl-hkibng42-5673dc-24.dhcp.inet.fi
Welcome
  CSC - Tieteen tietotekniikan keskus - IT Center for Science
  HP Cluster Platform SL230s Gen8 / XL230a Gen9 TAITO
Contact
  Servicedesk : 09-457 2821, servicedesk@csc.fi   Switchboard : 09-457 2001
User Guide
  https://research.csc.fi/taito-user-guide
Software
  Available modules can be listed with command: module avail | module spider
Partitions
  parallel    : 16-672 cores / 5mins/3days      def/max runtime
  serial      : 1-24 cores   / 5mins/3days      def/max runtime
  longrun     : 1-24 cores   / 5mins/14days     def/max runtime
  test        : 1-48 cores   / 5mins/30mins    def/max runtime
  hugemem     : 1-40 cores   / 5mins/7days     def/max runtime
News
  2014-04-25: Quota cutter for batch jobs is now in use on Taito. It is not possible to run batch jobs if quota is exceeded. One can apply for more resources in the usual way, see:
  https://www.csc.fi/laskentaresurssien-haku (suomeksi)
  https://www.csc.fi/applying-for-computing-resources (in English)

  2014-10-02: If you wish to change the project you're using for billing, please use "newgrp groupname". You can find more information at:
  http://tinyurl.com/kozfa6t

  2014-11-27: For jobs requiring more than 16 GB memory per core, please use the 'hugemem' queue consisting of six 1.5 TB memory nodes with 32/40 cores each.

  2017-01-13: Monthly $WRKDIR clean-up will start on Feb. 1. 2017.
  Data storage info: http://bit.ly/2jDk75P and http://bit.ly/2jLjydH

  2017-01-19: Remember that memory is also billed. That is, the basis of billing is either the number of cores or number (rounded up to the nearest integer) of 4 GB blocks of memory, depending which is larger.
  For example,
    - an 8 core job with 16 GB of memory will be billed with 2 x 8 = 16 bu/h
    - an 8 core job with 64 GB of memory will be billed with 2 x 16 = 32 bu/h

CRITICAL: You belong to the project 10004688, which has used all its CPU time quota. Project 10004688 is not allowed to submit any new jobs.
Please apply for more CPU time, the instructions are at: https://research.csc.fi/applying-for-computing-resources

stelmach@taito-login3 [ ~ ] >
```

# Some basic UNIX commands

`pwd`: print working directory (“where am I?”)

`ls`: list (“show folder contents”)

`mkdir`: make directory (a.k.a. folder)

`cd`: change directory (“go to folder”)

`cp`: copy

`mv`: move

`rm`: remove

# Some additional notes

## Case-sensitive

photo.jpg ≠ PHOT0.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](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

- 1) Launch Visual Studio Code
- 2) *Only on the first time:*  
*Open Extensions (one of the icons on the left) and install Remote - SSH*
- 3) Down left corner you will have a (green) button with "><" (hoover over it and it says "Open a Remote Window"), click it
- 4) Choose "Connect Current Window to Host..."
- 5) Type in the **USERNAME@puhti.csc.fi** and hit "Enter" (change "username" for your own CSC username)
- 6) Type your password and hit "Enter"
- 7) In the following dialogue, type **yes** and hit "Enter"
- 8) When the down left corner says **SSH:puhti.csc.fi**, you're connected.

# **Time to familiarize yourself with UNIX**

[https://github.com/karkman/MMB-114\\_Genomics](https://github.com/karkman/MMB-114_Genomics)

(Day 1: UNIX and CSC)