

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

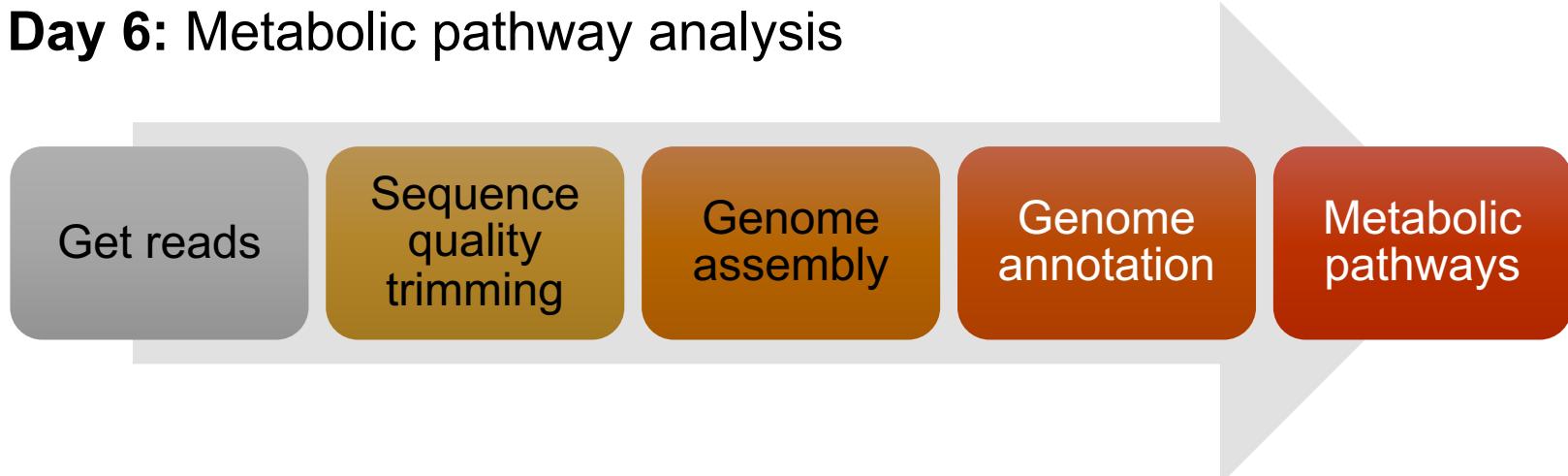
Day 2: Handling of Illumina data

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:

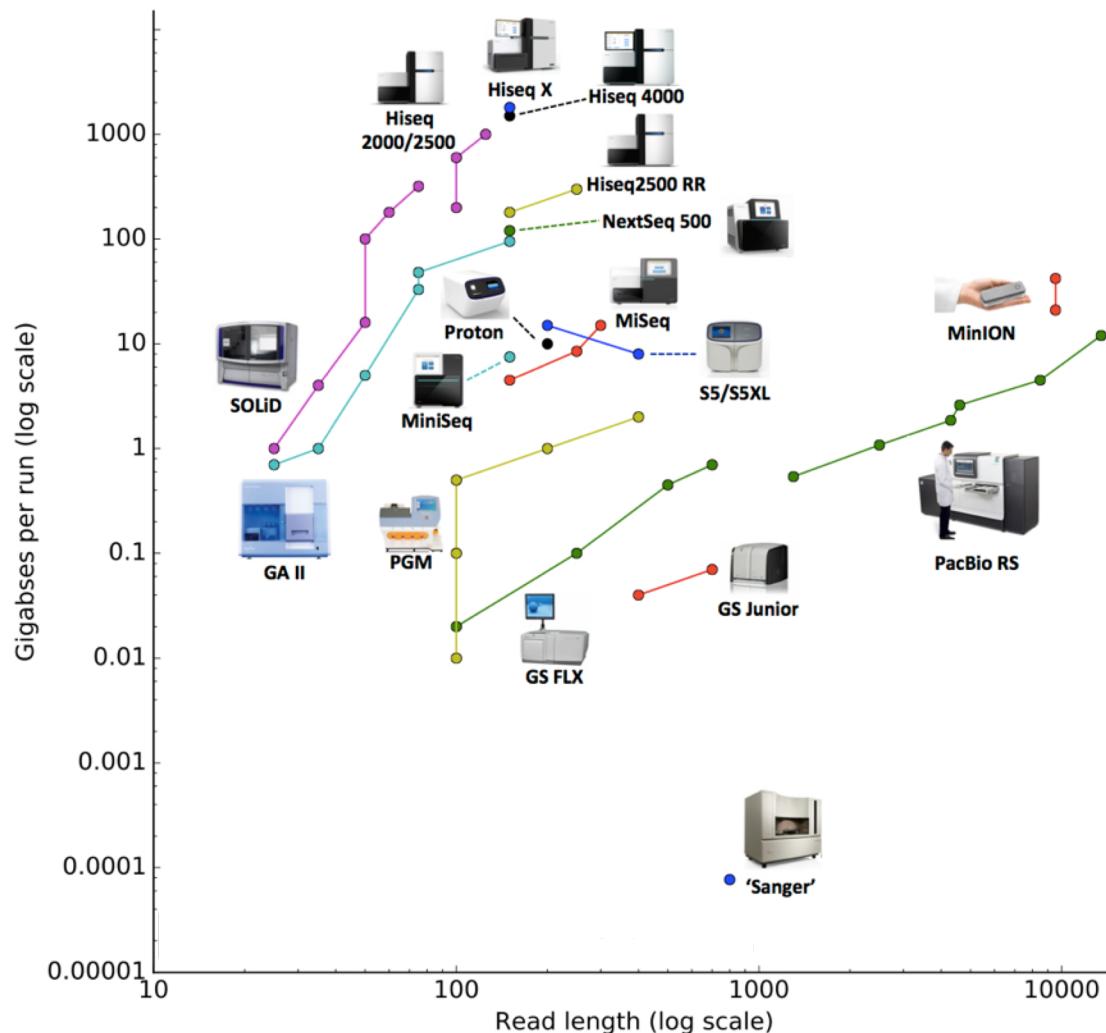
<https://github.com/igorspp/MMB-114>

We will start everyday at 10.00 with an introductory lecture

You can (and should) interrupt me any time

Those joining by Zoom, you can leave your camera on or off,
as you wish

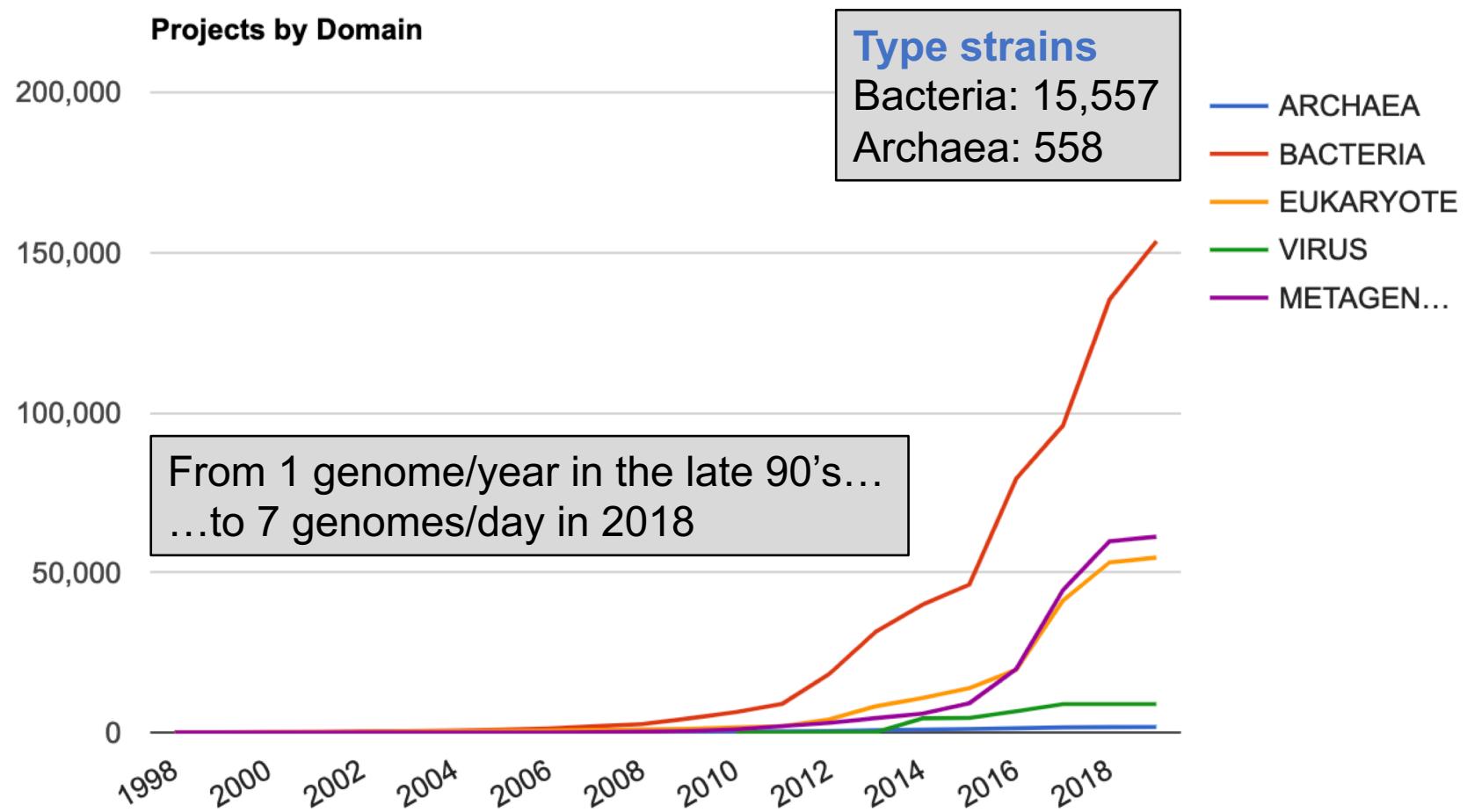
The evolution of sequencing



Features

- Read size
- Read depth (output)
- Error rates
- Price

Genomes OnLine (GOLD) database



Genomics can be used to study

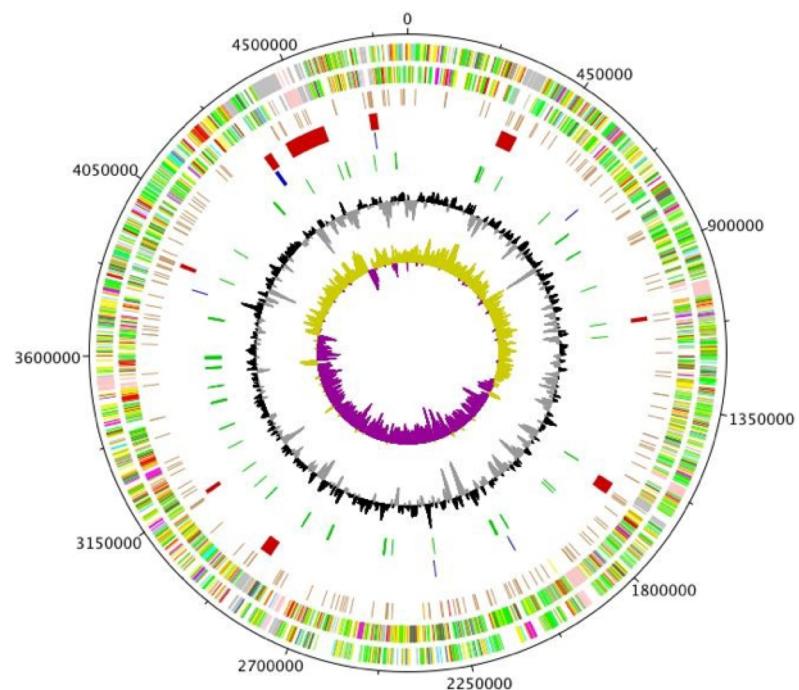
Physiology

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

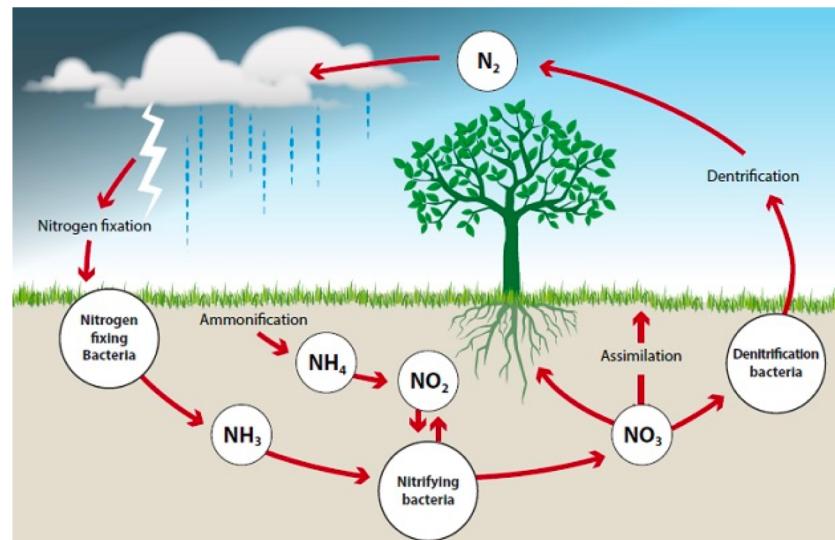
Recent genomics studies have revealed the importance of ammonia-oxidizing archaea

Thaumarchaeota

Novel gene inventory

Higher substrate affinity

Different pathways



Genome evolution

Mutations

Synonymous and non-synonymous

Small insertions and deletions

Can change the function of a gene

Gene recombination, deletion, duplication

Rearrange genes, can change expression

Increase in gene copy numbers

Deletion of genes

Horizontal gene transfer

Acquiring material from non-parental organism

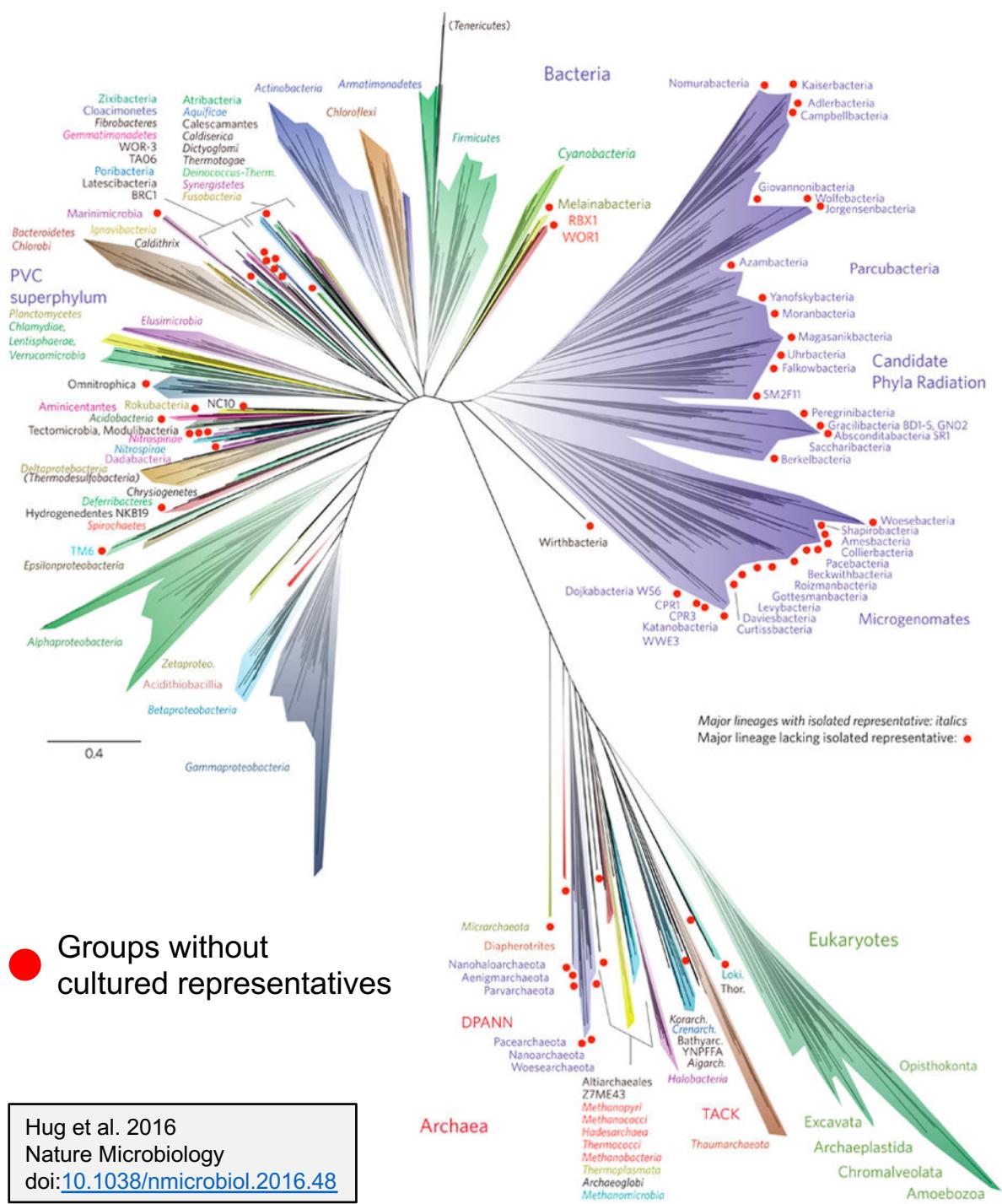
Antibiotic and xenobiotic resistance, toxin production

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)

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

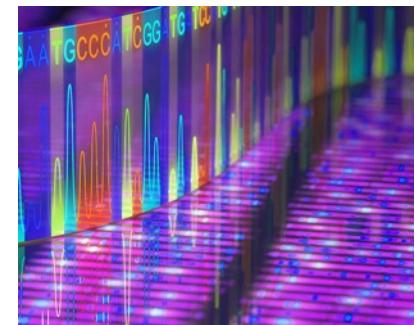


What was done in the sequencing facility?

Quality control (size, amount)

Library preparation with the Nextera kit

Sequencing with Illumina MiSeq



Illumina library preparation and sequencing

Illumina MiSeq and HiSeq

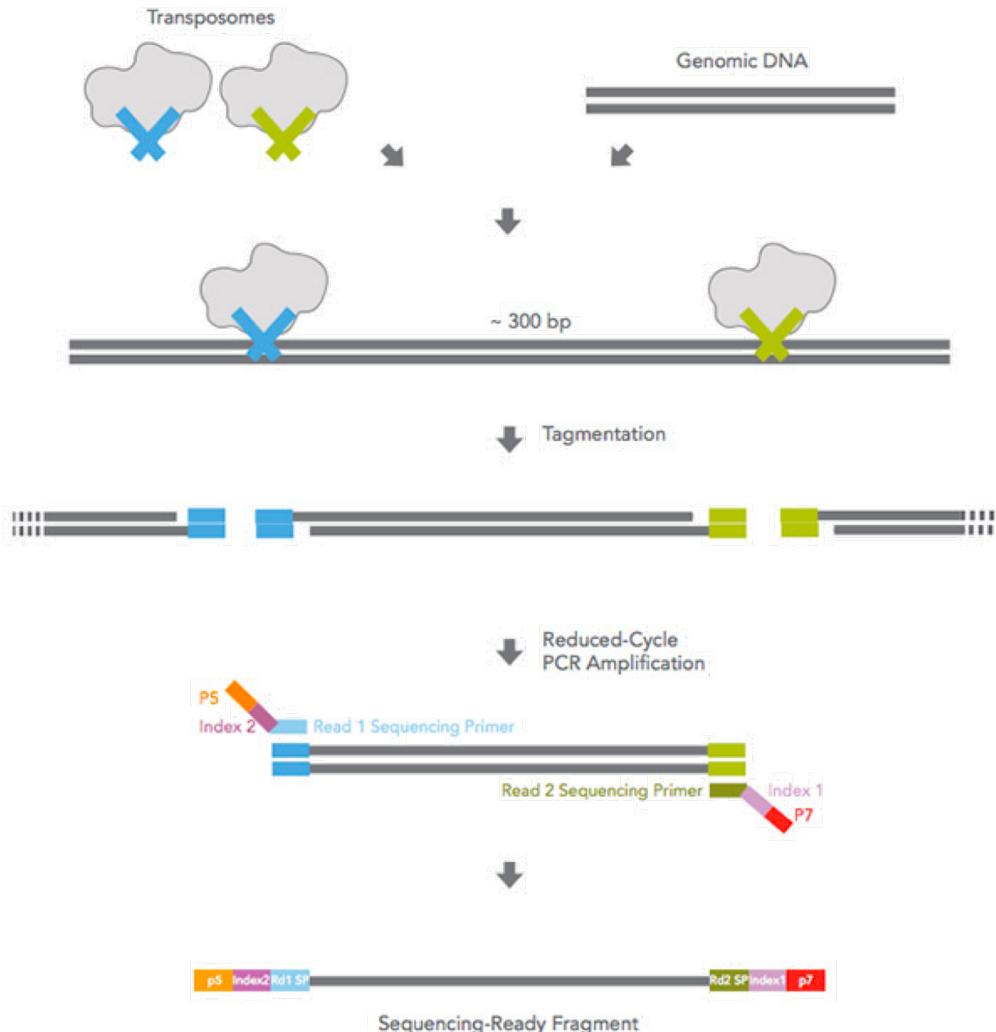
Good compromise between size, amount and error rate of reads

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

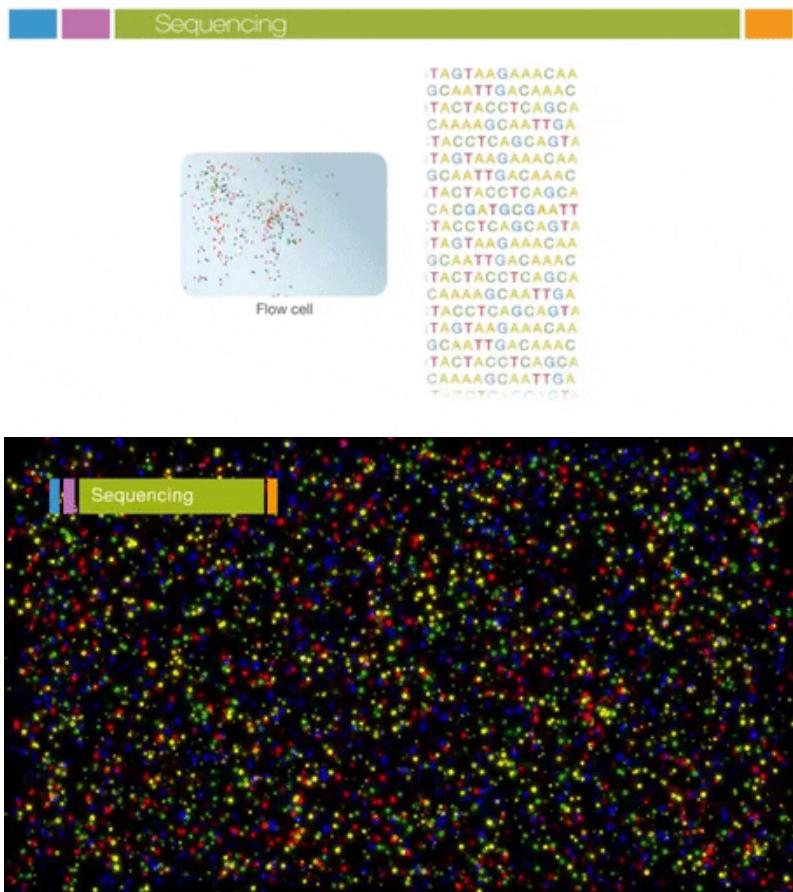
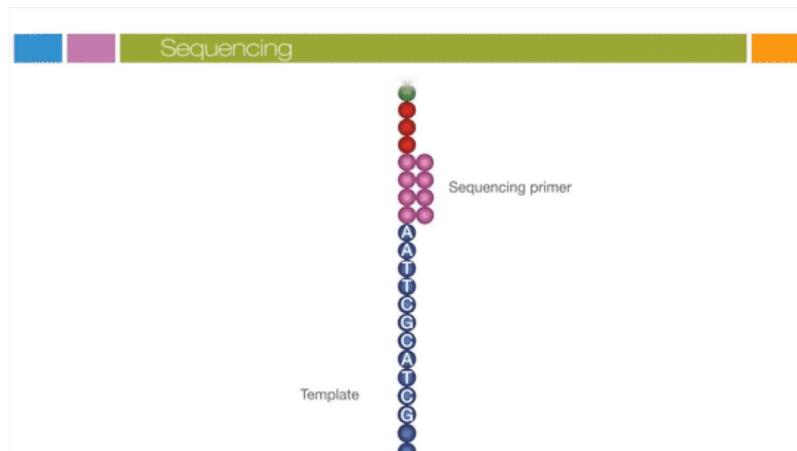
Correct amount of input DNA to avoid under- and overtagmentation

Adapter ligation

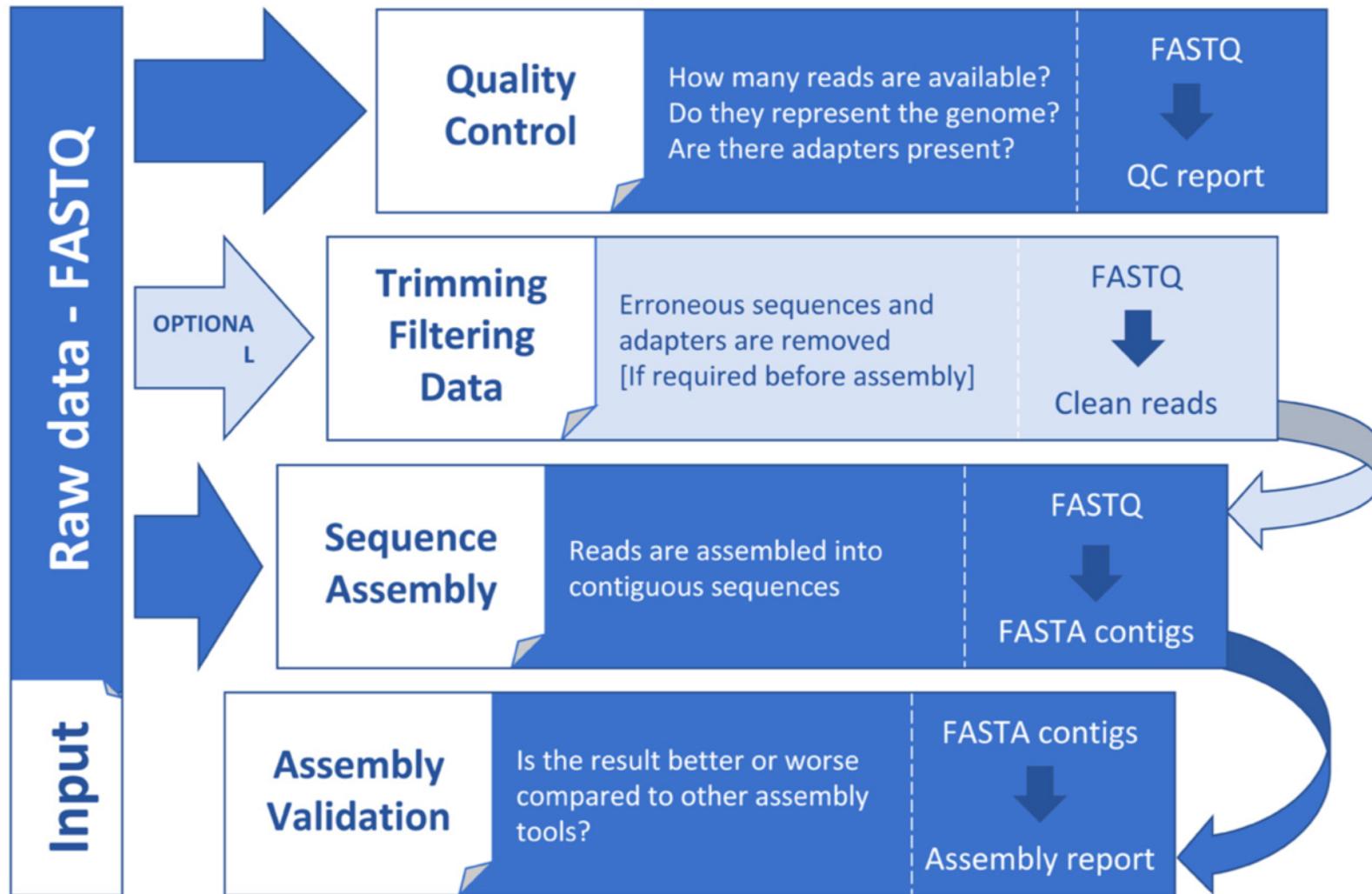
Flow cell adapters
Sequencing primers
Sequencing indexes
Optional indexes for multiplexing

Biases in the first bases have been observed

Sequencing



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

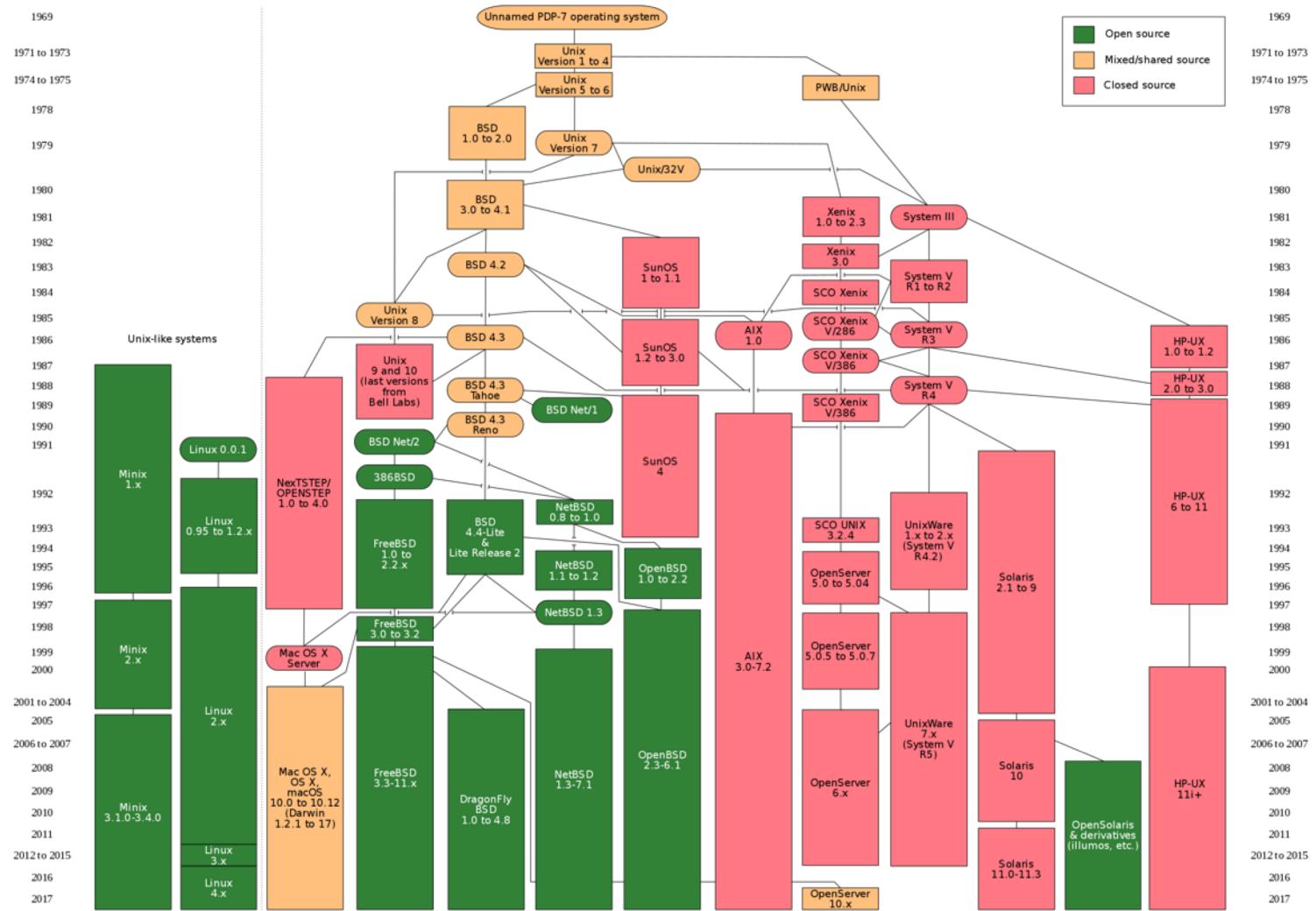
Multiprocessing

- Capable of supporting and utilizing more than one computer processor

Portable

- Can be used in various hardware architectures

History of UNIX



Linux



A UNIX-based OS
Leading OS on servers, supercomputers
Free and open source

A composite screenshot of a Linux desktop environment. On the left, a terminal window displays command-line output related to the Linux kernel version and release. In the center, a file manager window shows a 'Bioinformatics Applications' folder containing icons for various bioinformatics tools like Artemis, Cytoscape, and BLAST. On the right, a large green background image features a close-up of a green caterpillar on a leaf, with the text 'Bio-Linux' overlaid.

The UNIX philosophy

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

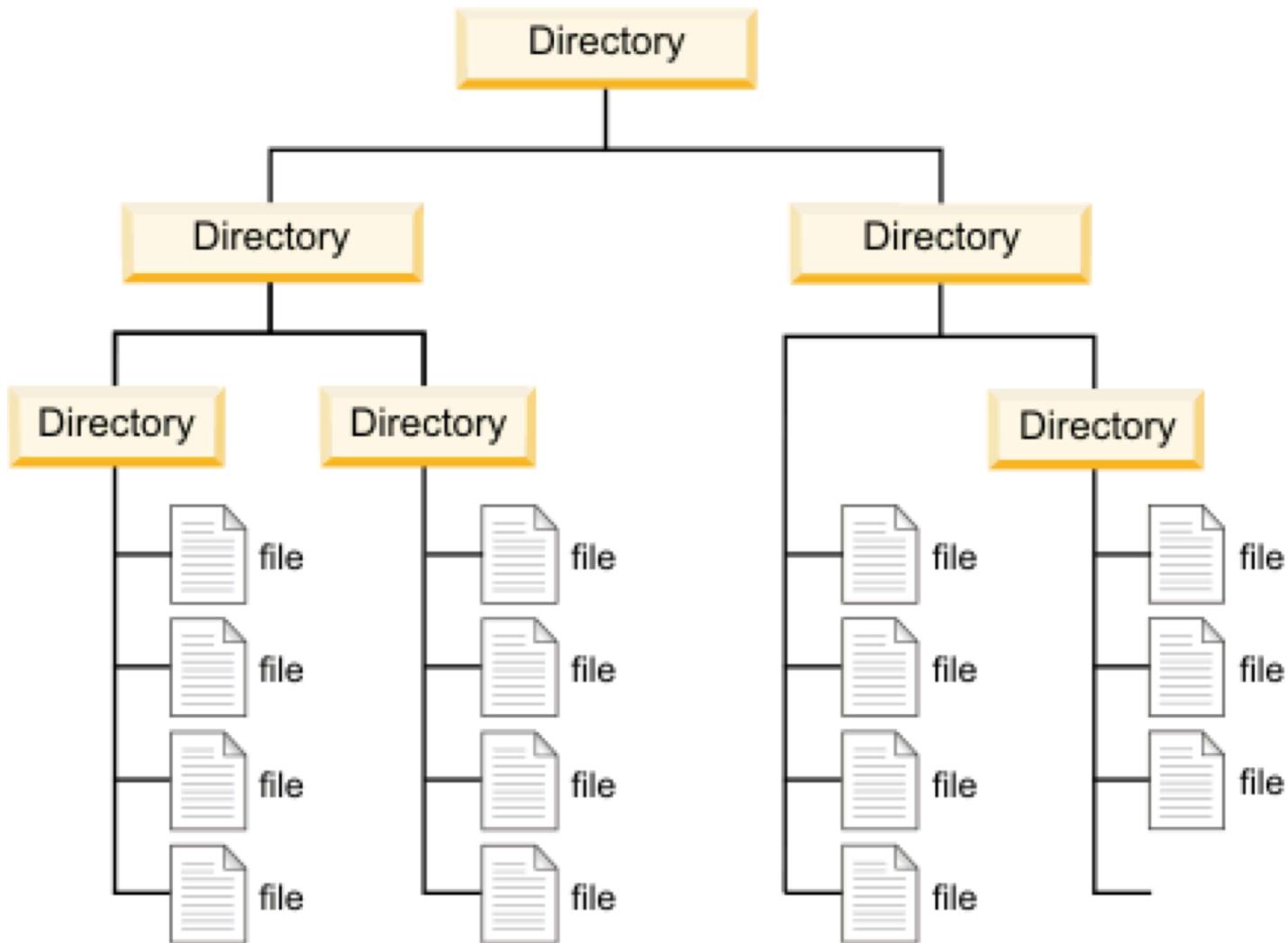
Use of plain text for storing data

Use of a hierarchical filesystem

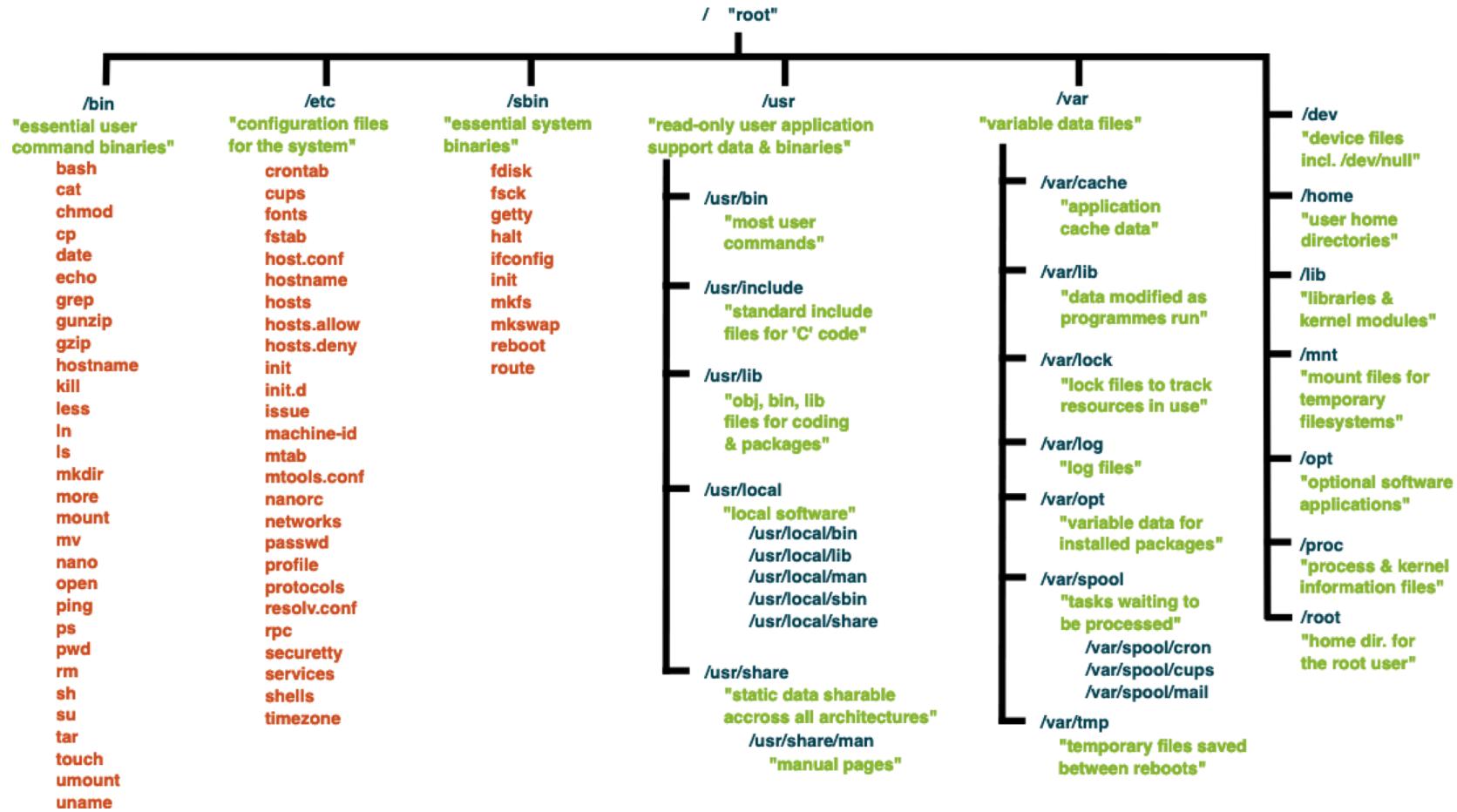
Use of a large number of simple programs performing a limited, well-defined function

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

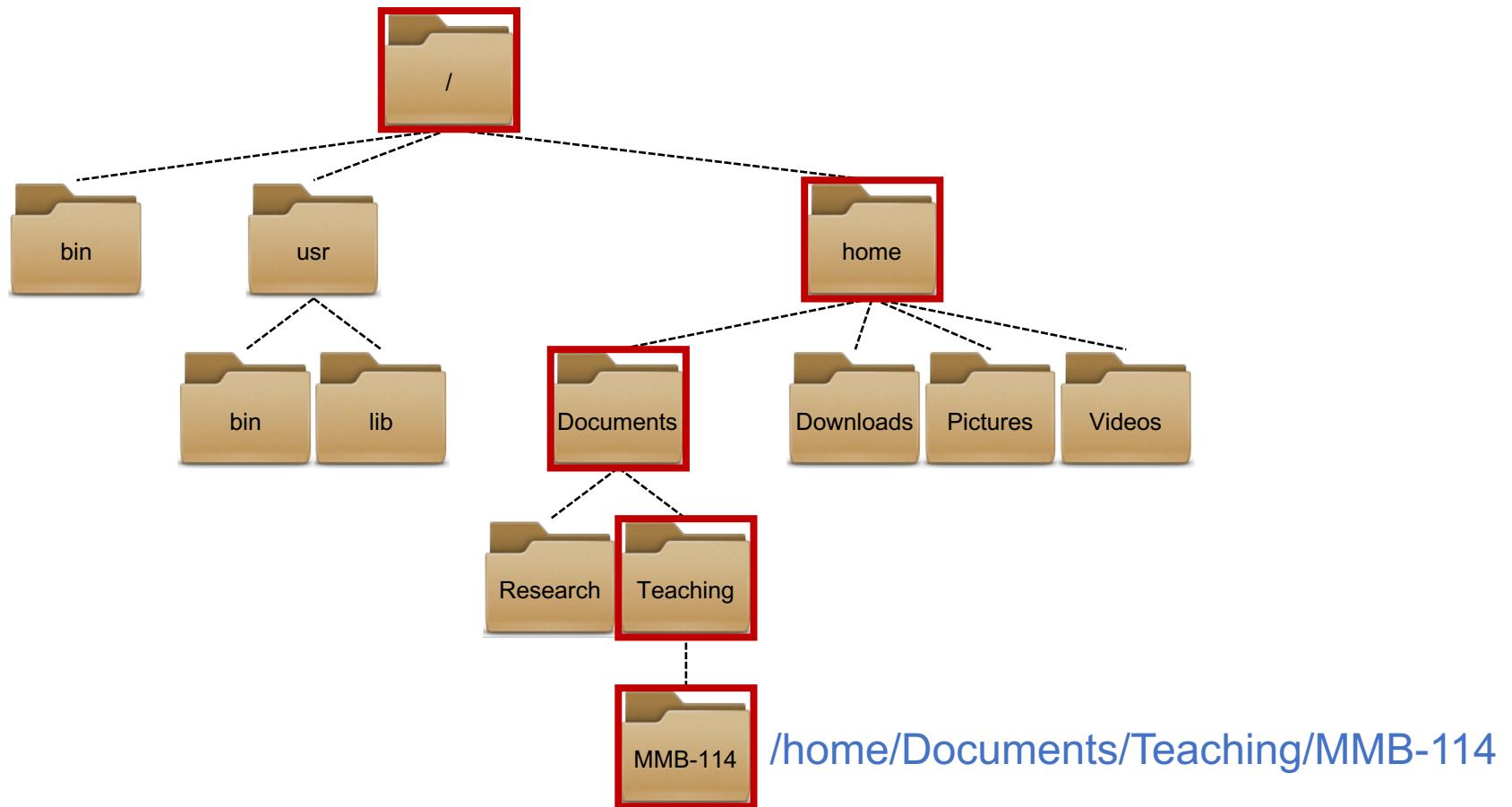
The UNIX hierarchical filesystem



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

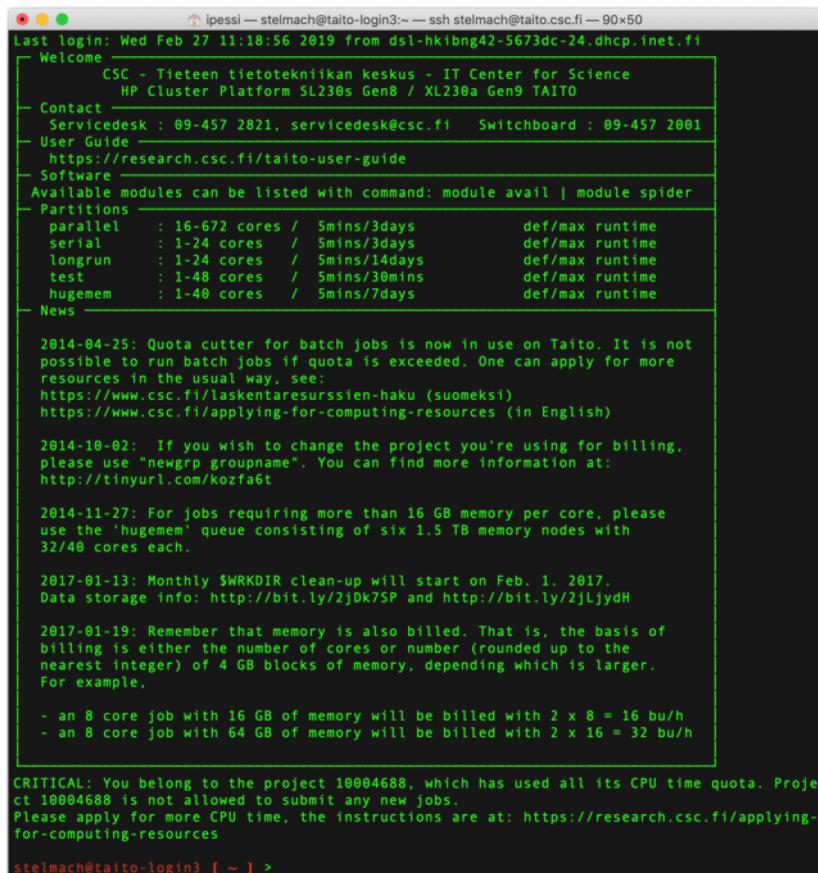
Primary interface before graphical user interfaces (GUIs) appeared

Still widely used today

Efficient

Low memory footprint

Advanced scripting



A screenshot of a terminal window titled 'stelmach@taito-login3:~'. The window displays a manpage for the 'module' command. The title bar shows the session details: 'ipessi — stelmach@taito-login3:~ — ssh stelmach@taito.csc.fi — 90x50'. The manpage content includes sections for 'Welcome', 'Contact', 'Software', 'Partitions', and 'News'. The 'News' section contains several announcements about resource usage and billing rules, such as 'Quota cutter for batch jobs is now in use on Taito', 'Change project for billing', and 'Monthly SWRKDIR clean-up'. At the bottom of the window, there is a message about a CPU quota being reached and a link to a resource management guide.

```
ipessi — stelmach@taito-login3:~ — ssh stelmach@taito.csc.fi — 90x50
Last login: Wed Feb 27 11:10:56 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 SWRKDIR clean-up will start on Feb. 1. 2017.
  Data storage info: http://bit.ly/2jDk7SP 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://mally.stanford.edu/~sr/computing/basic-unix.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, change the billing project in CSC

1. Login to <https://my.csc.fi/>
2. Go to My Profile
3. Under “Default billing project” select “2001379 MMB-114 Exploratory microbial research - lab course 2019”
4. Click “Save” to confirm

The screenshot shows the 'My Profile' page of the CSC portal. The left sidebar has 'My Profile' selected. The main area is titled 'Your personal details'. It includes fields for CSC Username (stelmech), Field of Science (Natural Sciences), Choose sub science area (Biological sciences), Choose optional second level sub science area, Default billing project (which is highlighted with a red box and contains the value '2001379 MMB-114 Exploratory microbial res ...'), Email (igor.pessi@helsinki.fi), First Name (Igor), Last Name (Stelmach Pessi), Citizenship (Brazil), Preferred language (English), and Job Title (Postdoctoral Researcher). To the right, there is an 'Address' section with fields for Street (P.O. Box 56), Postal Code (00014), City (Helsinki), and Mobile phone (+358413645678). At the bottom right are 'Save' and 'Change password' buttons.

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)