Genome assembly and annotation

MMB-114

Schedule

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

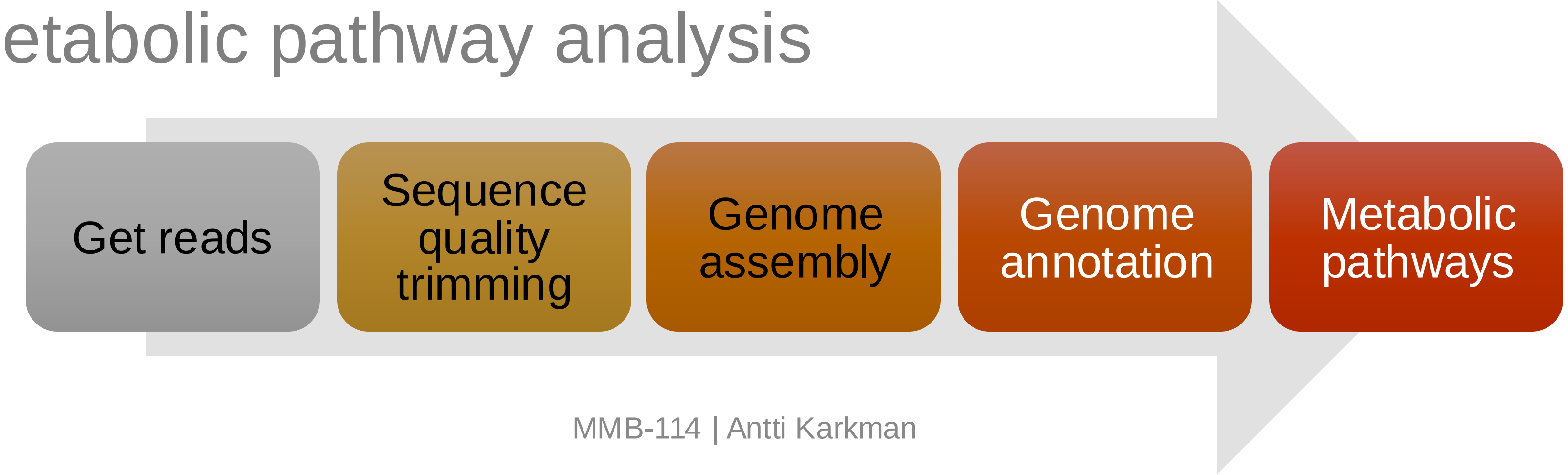
Day 2: Handling of Nanopore/Illumina data

Day 3: Check-up

Day 4: Genome assembly

Day 5: Genome annotation

Day 6: Metabolic pathway analysis



Get reads

Sequence quality trimming

Genome assembly

Genome annotation

Metabolic pathways

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

Practical things

- **All exercises and presentations can be found from:**

  https://github.com/karkman/MMB-114_Genomics

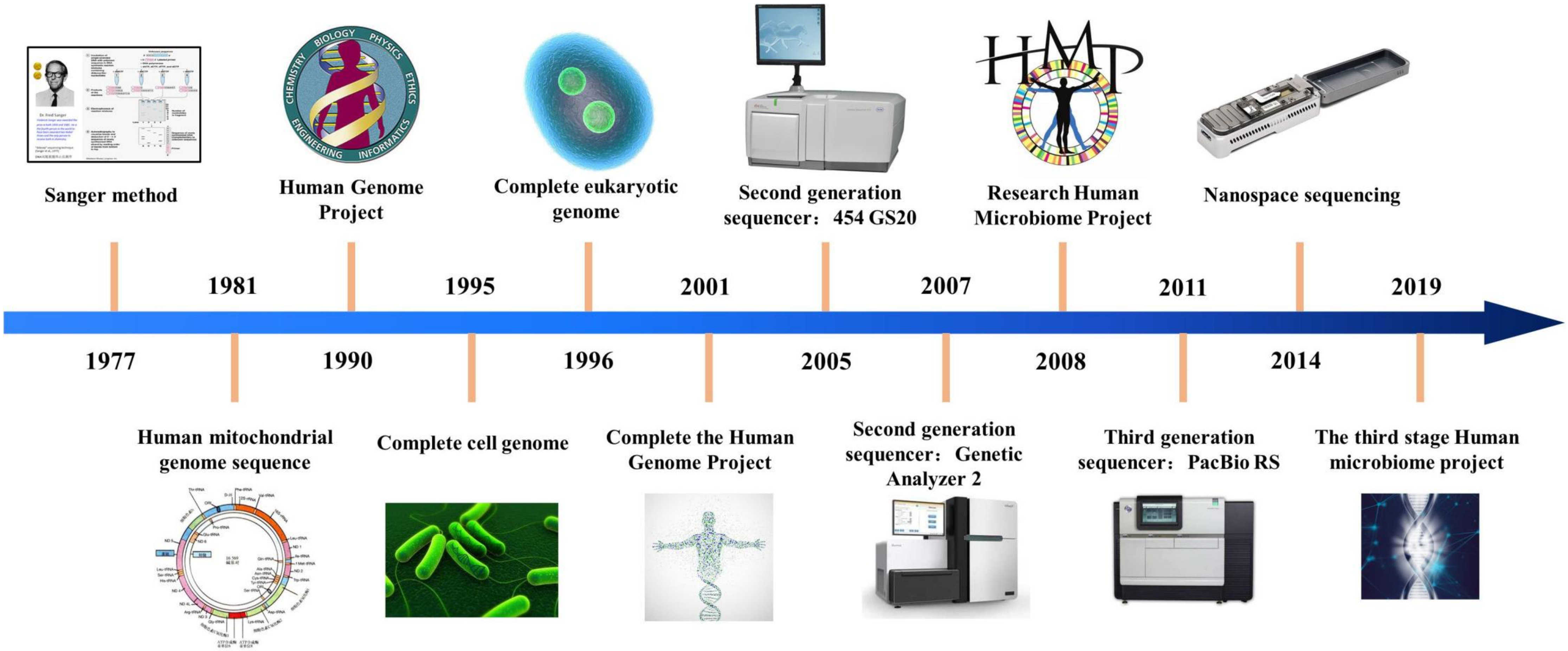
- We will start every day at 10.00 with an introductory lecture and then some exercises

- REP7: Materials and methods for bioinformatics part

  DL Friday 8.11.

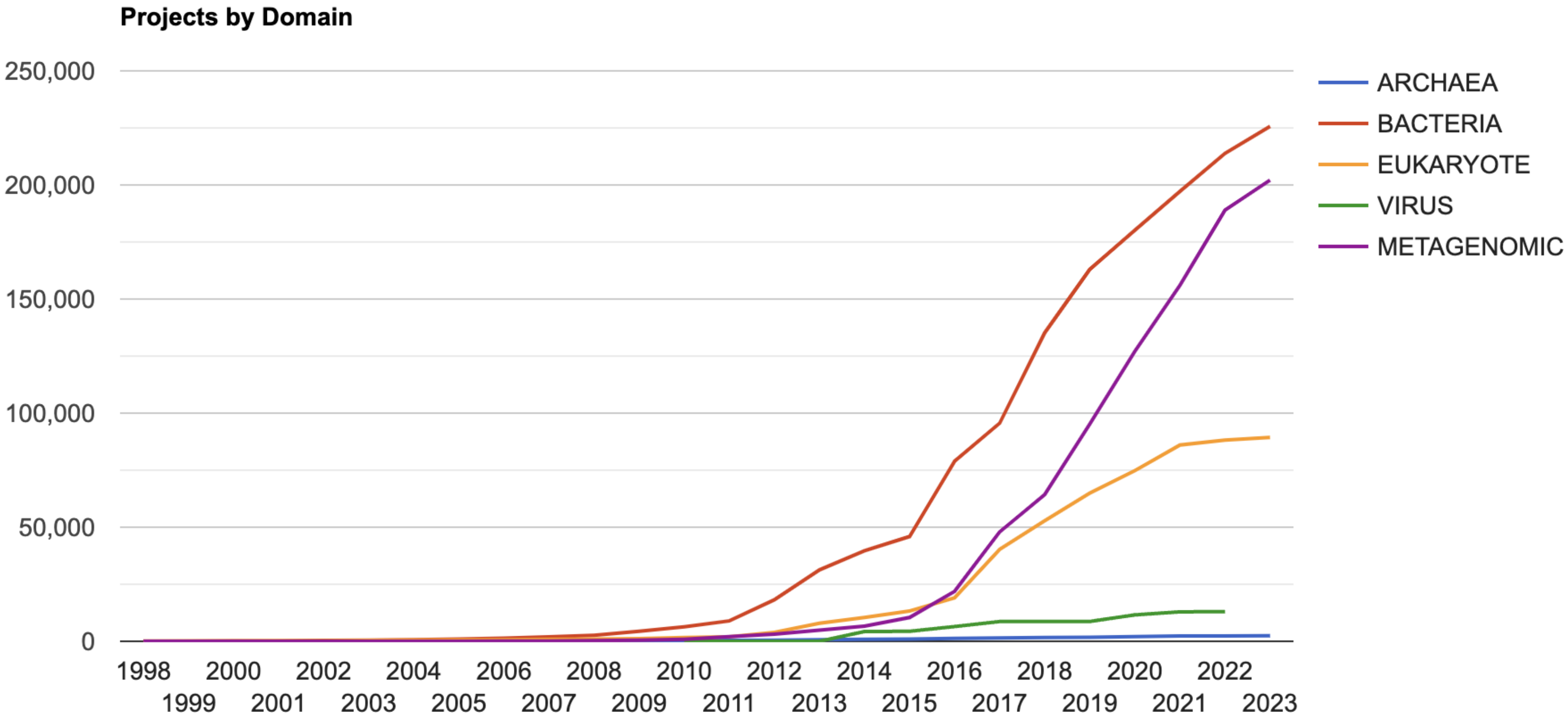
Whole genome sequencing (WGS)

The evolution of sequencing


1977
Sanger method
1981
Human mitochondrial genome sequence
1990
Human Genome Project
1995
Complete cell genome
1996
Complete eukaryotic genome
2001
Complete the Human Genome Project
2005
Second generation sequencer: 454 GS20
2007
Second generation sequencer: Genetic Analyzer 2
2008
Research Human Microbiome Project
2011
Third generation sequencer: PacBio RS
2014
Nanospace sequencing
2019
The third stage Human microbiome project
HMP
BIOLOGY PHYSICS ETHICS INFORMATICS ENGINEERING CHEMISTRY

https://doi.org/10.3389/fbioe.2020.01032

Genomes OnLine (GOLD) database

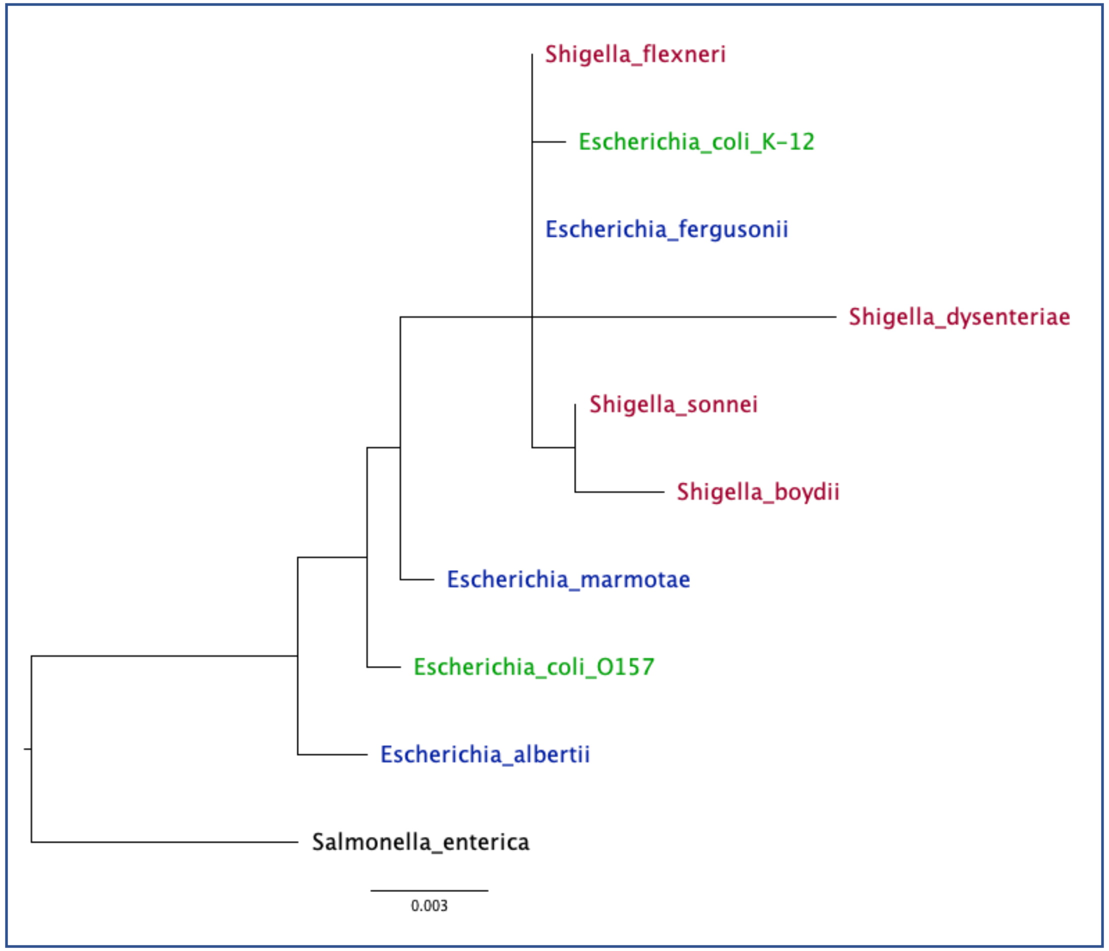


https://gold.jgi.doe.gov/statistics

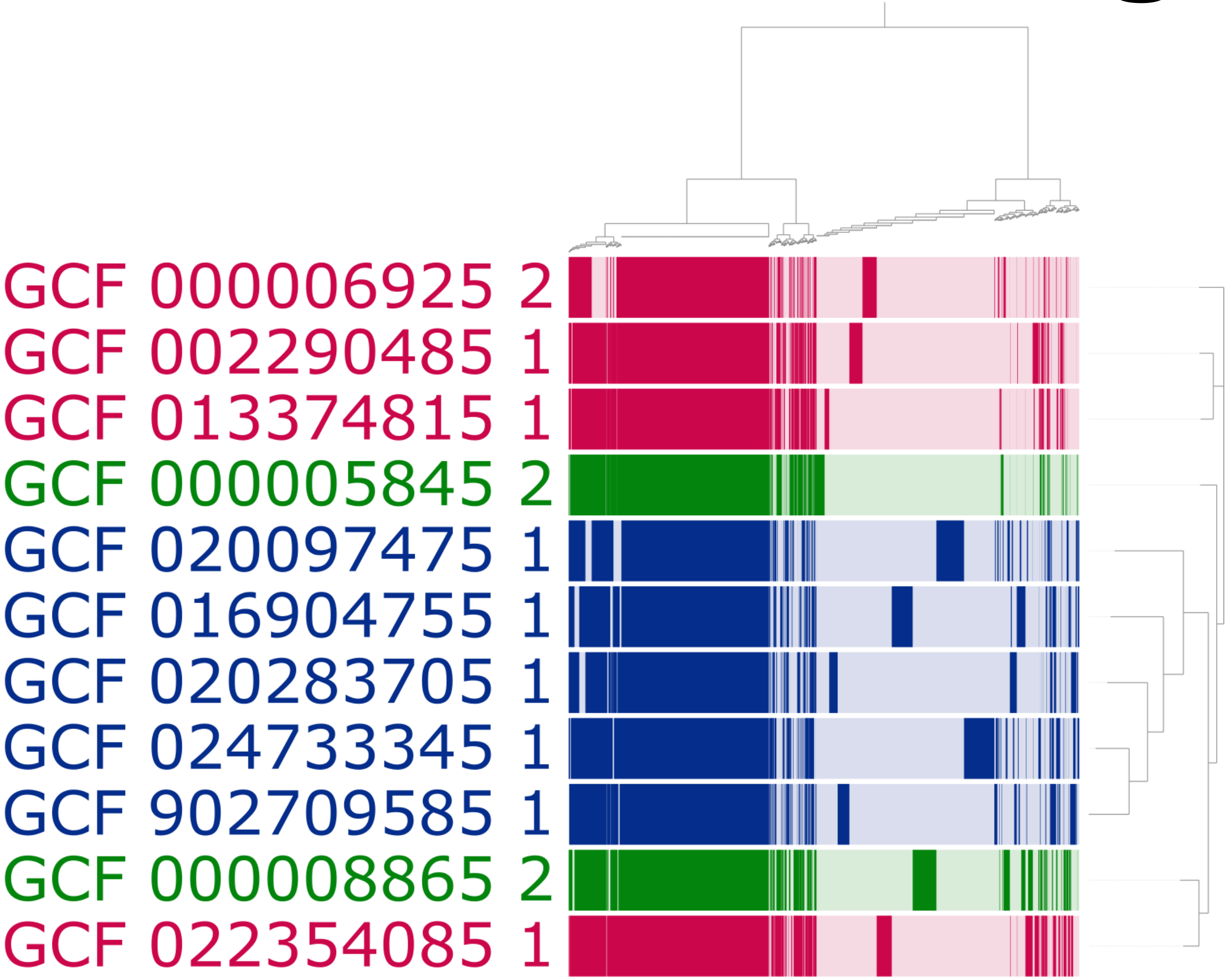
Why whole genome sequencing?

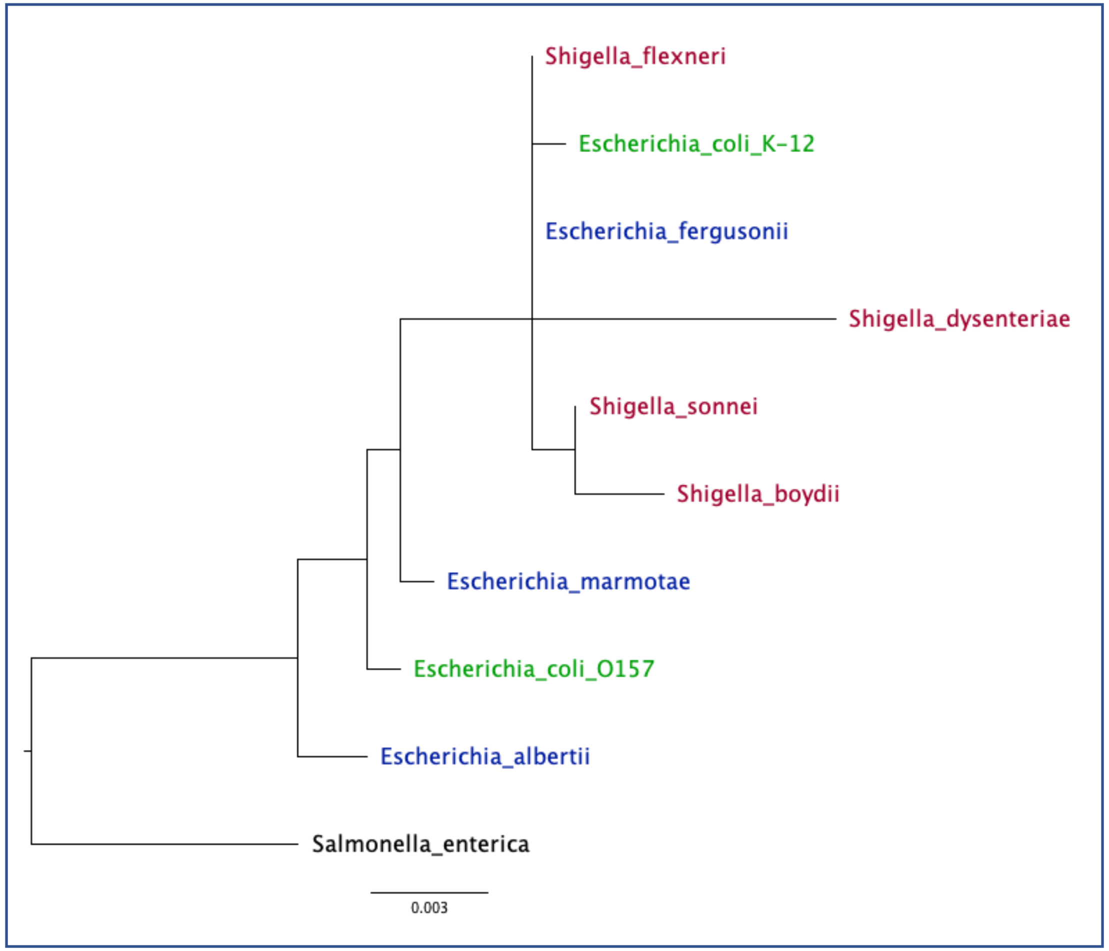
- *Escherichia* and *Shigella* as an example
- 16S rRNA gene cannot resolve the two genera
  (or these belong to the same genus)

- **What can you tell about your bacteria based on the 16S rRNA gene?**



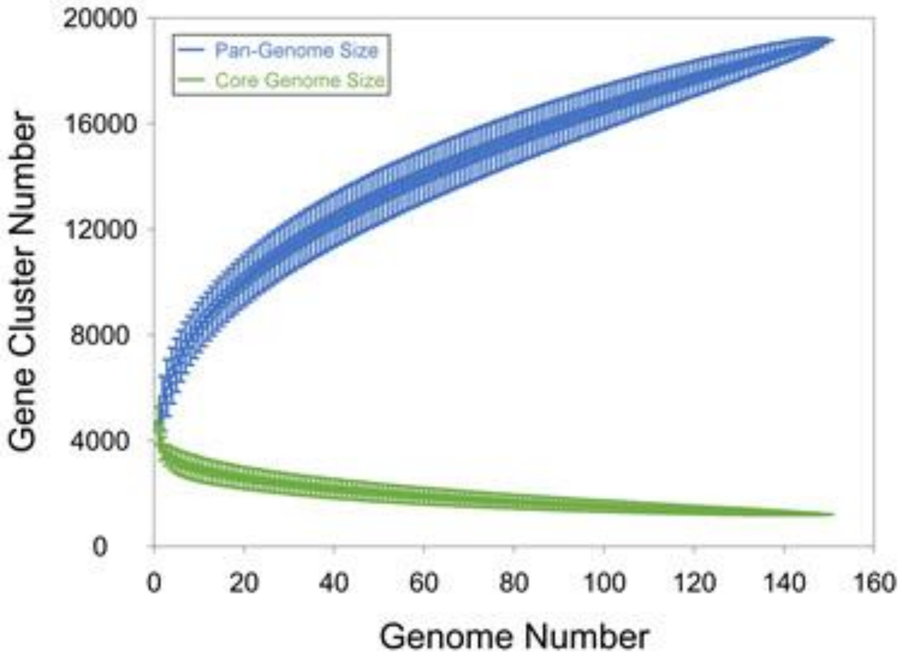
Escherichia–Shigella WGS vs. 16S



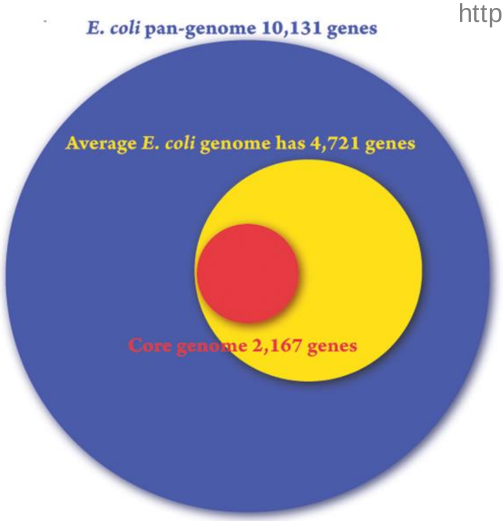


Pangenomes

- Bacterial species can have very diverse genomes
- Bacterial lifestyle affects pangenome size
- **Pangenome:** All genes from the studies strains
- **Core-genome:** genes shared by all strains
- **Accessory genome:** genes shared by only one or few strains



https://pangp.zhaopage.com/

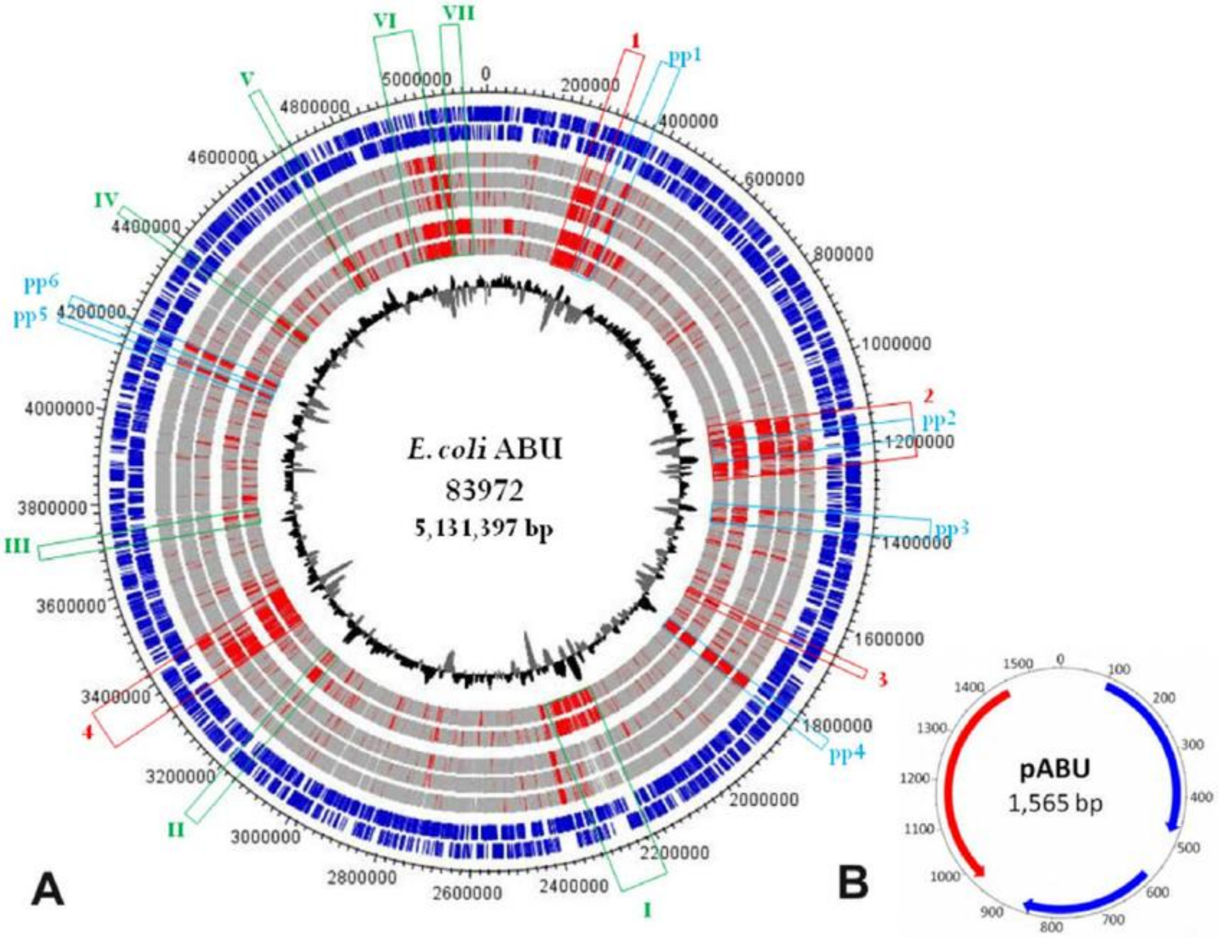


E. coli pan-genome 10,131 genes
Average *E. coli* genome has 4,721 genes
Core genome 2,167 genes

https://doi.org/10.1371/journal.pgen.1000335

Genomics can be used to study

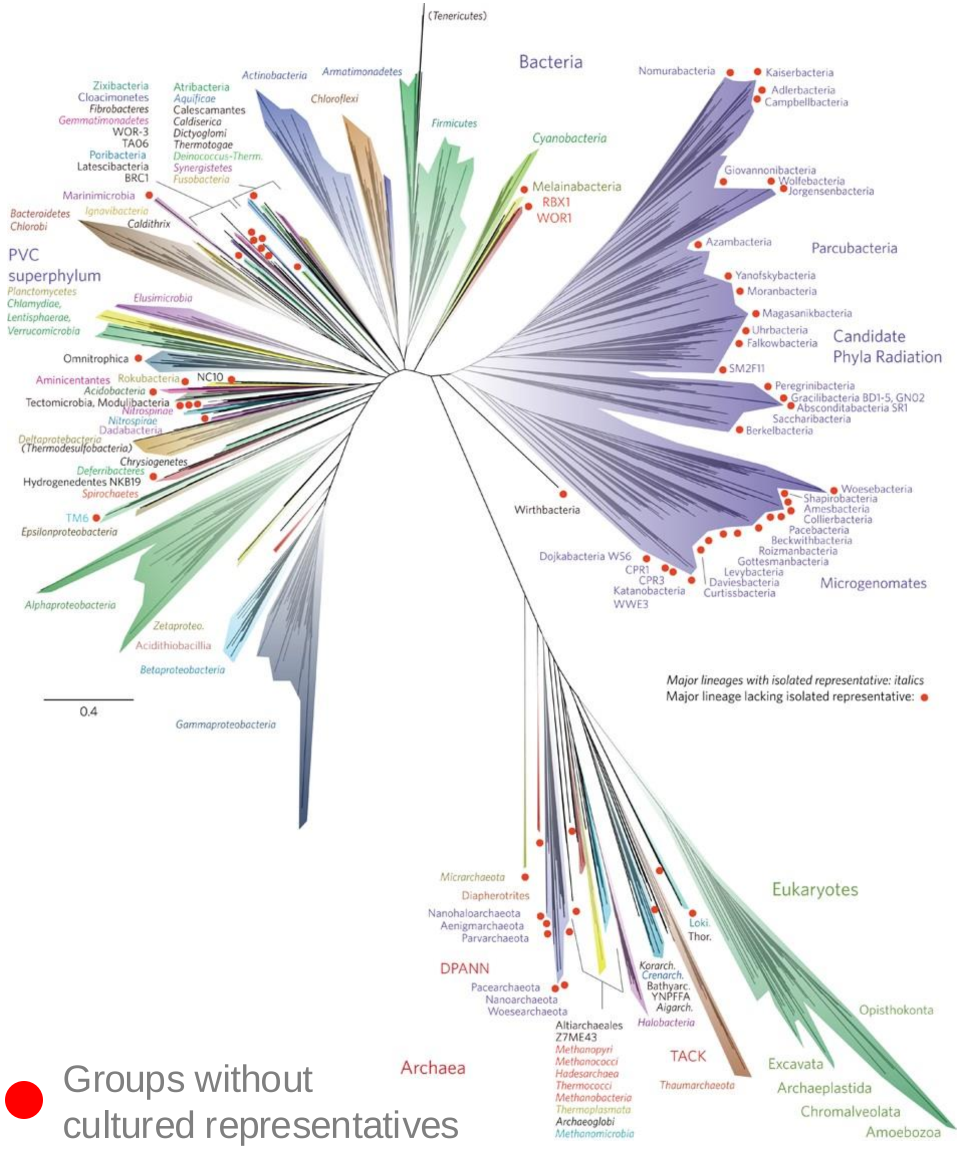
- Physiology
- Evolution
- Pathogenesis
- Novel industrial processes
- Novel diagnostic/ epidemiologic tests
- Extra-chromosomal elements



https://doi.org/10.1371/journal.ppat.1001078

The new 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

https://doi.org/10.1038/nmicrobiol.2016.48

Your strains

What was done by you in the lab?

- Isolation of the strain
- DNA extraction
- DNA measurements
  - Purity, integrity, contamination

What was done at the sequencing lab?

- Quality control (size, amount)
- Library preparation with the Nextera kit
- Sequencing with Illumina MiSeq

Illumina library preparation

Size selection

Correct amount of input DNA to avoid under- and overtagmentation

Adapter ligation

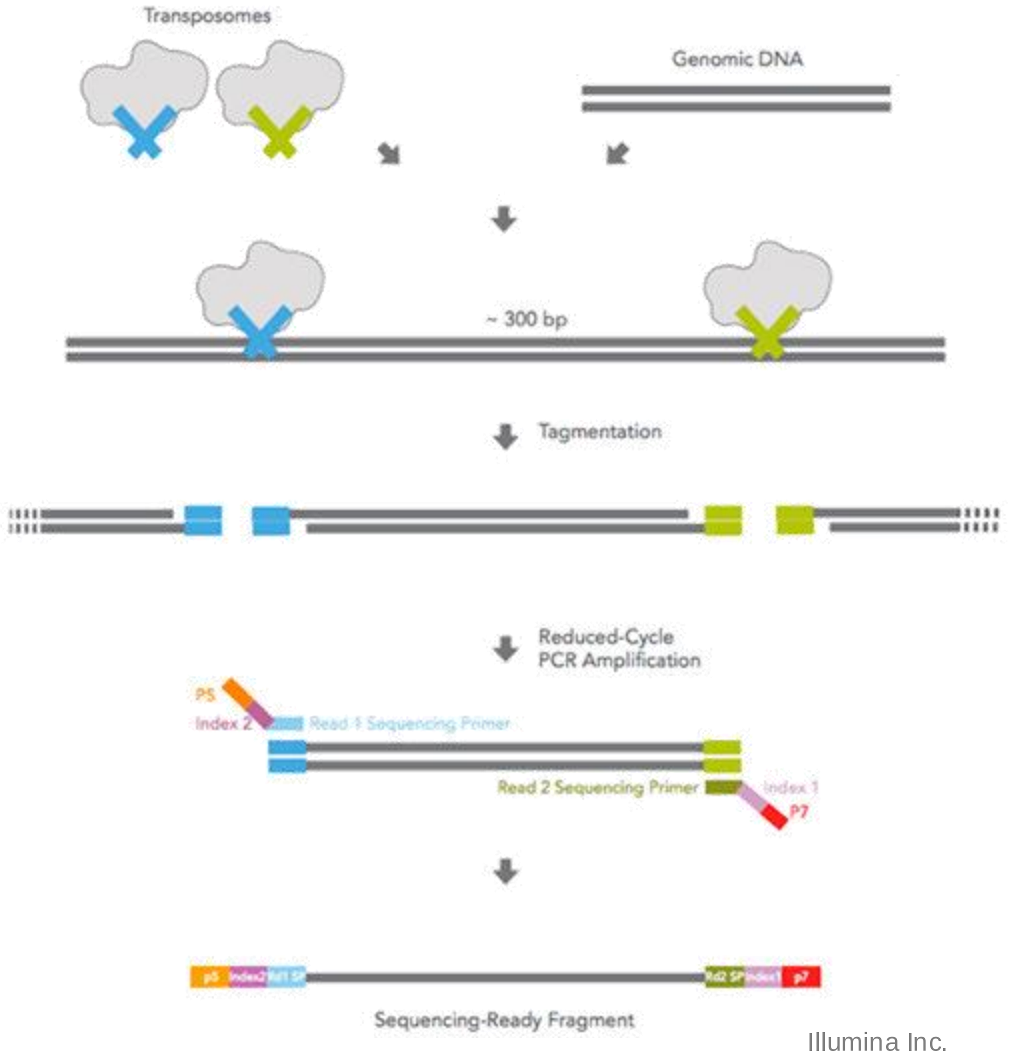
Flow cell adapters

Sequencing primers

Sequencing indexes

Optional indexes for multiplexing

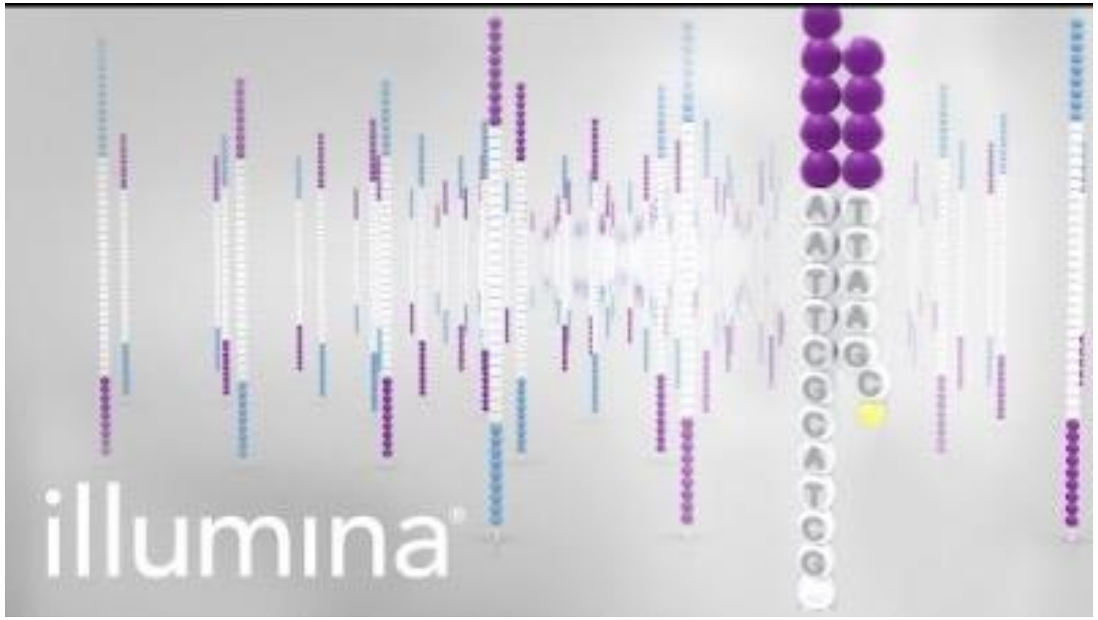
Biases in the first bases have been observed



Illumina Inc.

Illumina sequencing

- 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



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

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 and CSC

UNIX

- Family of computer operating systems (OS)
  - Linux, macOS, Solaris, OpenBSD

- Key characteristics
  - Multitasking
  - Multiuser
  - Multiprocessing
  - Portable

The UNIX philosophy

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

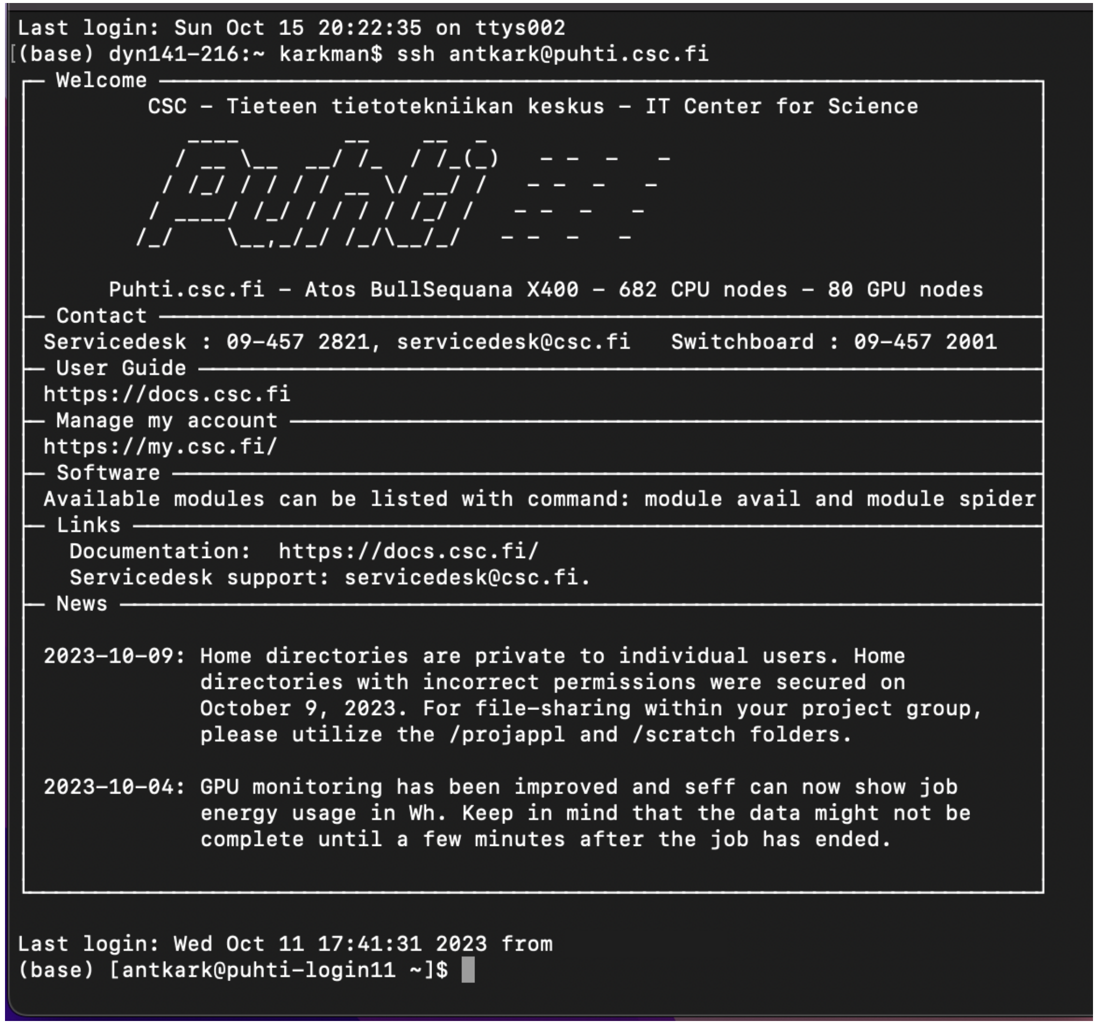
- Use of plain text for storing data
- Use of a hierarchical filesystem
- Use of a large number of simple programs performing a limited, well-defined function
- Use of a command-line interpreter (“shell”) to combine these programs to perform complex tasks

The UNIX 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



```
Last login: Sun Oct 15 20:22:35 on ttys002
[(base) dyn141-216:~ karkman$ ssh antkark@puhti.csc.fi
┌─ Welcome ────────────────────────────────────────────────────────────────────┐
│         CSC - Tieteen tietotekniikan keskus - IT Center for Science          │
│            ____        __    __  _                                           │
│           / __ \__  __/ /_  / /_(_)   - -  -   -                             │
│          / /_/ / / / / __ \/ __/ /   - -  -   -                              │
│         / ____/ /_/ / / / / /_/ /   - -  -   -                               │
│        /_/    \__,_/_/ /_/\__/_/   - -  -   -                                │
│                                                                              │
│      Puhti.csc.fi - Atos BullSequana X400 - 682 CPU nodes - 80 GPU nodes     │
├─ Contact ────────────────────────────────────────────────────────────────────┤
│ Servicedesk : 09-457 2821, servicedesk@csc.fi   Switchboard : 09-457 2001    │
├─ User Guide ─────────────────────────────────────────────────────────────────┤
│ https://docs.csc.fi                                                          │
├─ Manage my account ──────────────────────────────────────────────────────────┤
│ https://my.csc.fi/                                                           │
├─ Software ───────────────────────────────────────────────────────────────────┤
│ Available modules can be listed with command: module avail and module spider │
├─ Links ──────────────────────────────────────────────────────────────────────┤
│   Documentation:  https://docs.csc.fi/                                       │
│   Servicedesk support: servicedesk@csc.fi.                                   │
├─ News ───────────────────────────────────────────────────────────────────────┤
│                                                                              │
│ 2023-10-09: Home directories are private to individual users. Home           │
│             directories with incorrect permissions were secured on           │
│             October 9, 2023. For file-sharing within your project group,     │
│             please utilize the /projappl and /scratch folders.               │
│                                                                              │
│ 2023-10-04: GPU monitoring has been improved and seff can now show job       │
│             energy usage in Wh. Keep in mind that the data might not be      │
│             complete until a few minutes after the job has ended.            │
│                                                                              │
└──────────────────────────────────────────────────────────────────────────────┘

Last login: Wed Oct 11 17:41:31 2023 from
(base) [antkark@puhti-login11 ~]$
```

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 ≠ 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 (the key)

History (up arrow)

How to learn UNIX?

- Bu using it!
  - Trial and error
  - Don’t copy/paste, type yourself
- Ask the internet
  - http://stackoverflow.com/
  - http://stackexchange.com/
  - http://askubuntu.com/
  - Search engines
  - **chatGPT**

- Online courses/tutorials
  - http://codecademy.com
- Cheat sheets
  - https://www.stationx.net/unix-commands-cheat-sheet/
- Manual pages
  - `man <program>`

CSC - IT CENTER FOR SCIENCE

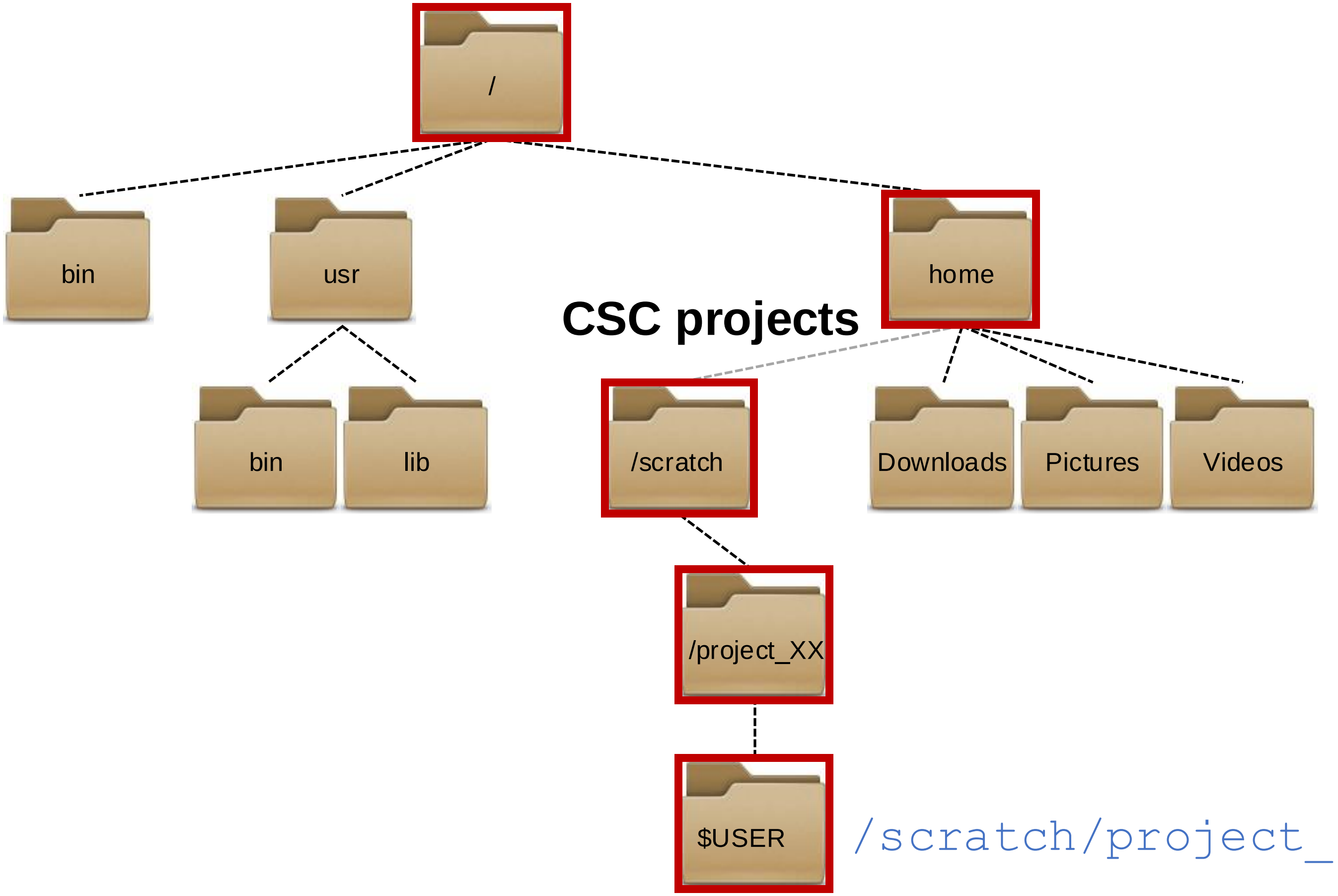
- ”Non-profit state enterprise with special tasks”
- Provides computing services for academics, research institutes and companies
- Free for academic research
- Owned by Finnish state (70 %) and higher education institutions (30 %)
- Funding from the Ministry of education and culture

Puhti

- One of the supercomputers at CSC
  - Atos BullSequana X400 cluster based on Intel CPU
- The best suited for bioinformatics
- Has CPU and GPU nodes
- Large collection of pre-installed software (modules)
- Interactive use and batch job scheduling system (SLURM)

- Read more: https://docs.csc.fi/support/tutorials/puhti_quick/

The filesystem in Puhti

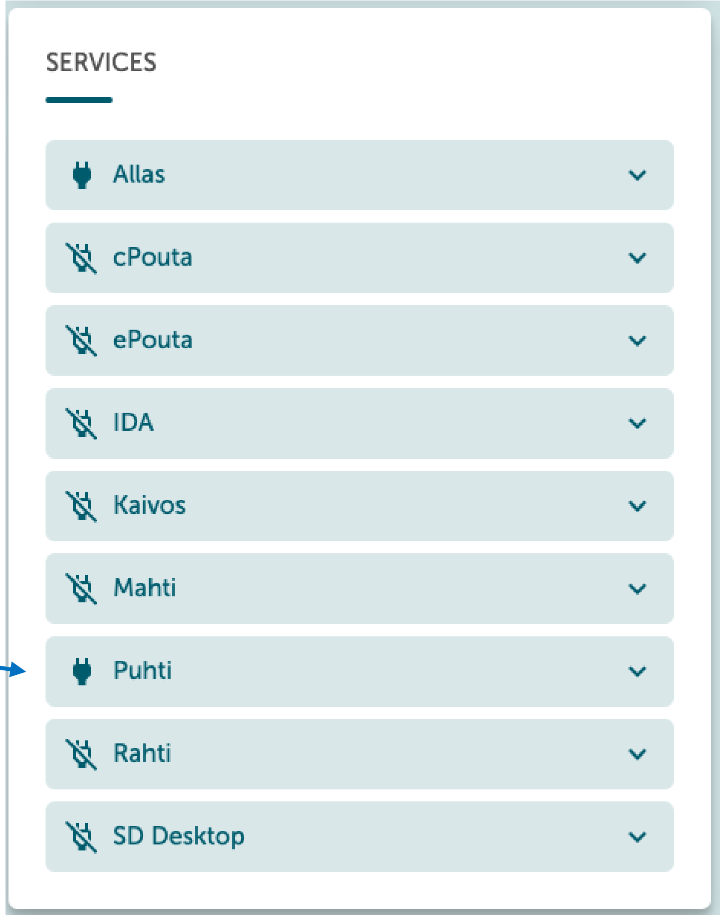


/scratch/project_2006616/$USER

Before we start, check if everything is OK in CSC

- Login to https://my.csc.fi/
- Go to **Projects**
- Click on **MMB-114_Genomics**
- Scroll down to **Services** on the right side and see if access to Puhti is enabled

- **Make sure you know your CSC username and password**


SERVICES

- Allas
- cPouta
- ePouta
- IDA
- Kaivos
- Mahti
- Puhti
- Rahti
- SD Desktop

Connecting to Puhti

- Launch Visual Studio Code
- Down left corner you will have a (green) button with "><”, **click it**
    - **If not:** Open Extensions (one of the icons on the left) and install Remote - SSH
- Choose "Connect Current Window to Host..."
- Type **YOUR_USER_NAME**@puhti.csc.fi and hit "Enter"
- Type your password and hit "Enter"
- In the following dialogue, type yes and hit "Enter"
- When the down left corner says SSH:puhti.csc.fi, you're connected.

UNIX exercises

https://github.com/karkman/MMB-114_Genomics

Day 1: UNIX and CSC