

INTRODUCTION TO BIOINFORMATICS

PRESENTED BY: PRAISSY ZEFI J (DTU)

CONTACT: pzeje@dtu.dk

LEARNING OBJECTIVES

Understanding sequencing output and files generated

File storing, naming, sharing and protection guidelines

Introduction to Operating systems

Introduction to Bioinformatics

Evading conflict issues using docker containers, virtual machines and environments

SEQUENCING OUTPUT

Data generated after sequencing DNA → in the forms of sequence with quality scores

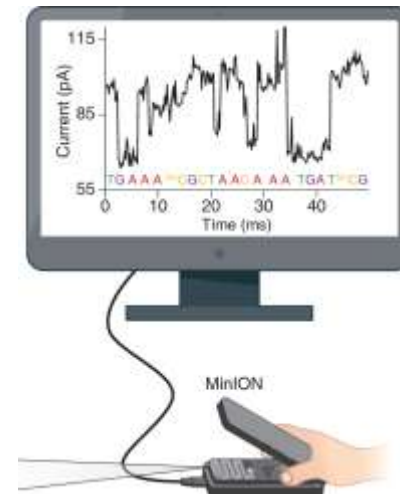
Can be in the form of chromatograms, text files or tables → files will be raw reads, mapped reads, aligned reads and corresponding quality scores

Analyzed using bioinformatic tools to interpret it biologically

ILLUMINA → FASTQ format (file123.fq)

ONT → FAST5 , POD5 format, FASTQ format

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
@ERR000589.42 EAS139_45:5:1:2:1293/1
AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
+
IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```



ILLUMINA SEQUENCING – OUTPUT FILE

FASTQ file: text-based sequencing data file format → stores raw sequence data and quality scores viewed using text-based editors or command line systems

SEQUENCE IDENTIFIER

SEQUENCE

SEPARATOR

QUALITY SCORES

```

1 @ERR000589.41 EAS139_45:5:1:2:111/1
2 CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
3 +
4 3IIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
5 @ERR000589.42 EAS139_45:5:1:2:1293/1
6 AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTAAAAGAAAT
7 +
8 IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I

```

**FASTQ ORA Sequence File Format → binary compressed file format of the text-based FASTQ file format
Up to 5x smaller than corresponding fastq.gz files without compromising data integrity**

Quality score, $Q = -10 \log_{10}(e)$

$e \rightarrow$ estimated probability that base call is wrong

(Q20) \rightarrow error rate of 1 in 100 (meaning every 100 bp sequencing read may contain an error) i.e. accuracy = 99%.

QUALITY SCORE	ERROR PROBABILITY
Q40	0.0001 (1 in 10000)
Q30	0.001 (1 in 1000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

SEQUENCE IDENTIFIER	1	@ERR000589.41 EAS139_45:5:1:2:111/1
SEQUENCE	2	CTTTCCTCCCTGCTTTCCTGGCCCCACCATTCGACCGAAGCTCTCTGAT
SEPARATOR	3	+
QUALITY SCORES	4	3IIIIIIIIIIII>1IIIF9BG08E00I9
	5	@ERR000589.42 EAS139_45:5:1:2:1
	6	AGTTGTTAAAATCCAAGCCAATTAAGATAGT
	7	+
	8	IIIIIGII.AIIII=?I9G-/II=+I=4?76

Q SCORE BINS	Example of empirically mapped Q scores
N (no call)	N (no call)
2-9	6
10-19	15
20-24	22
25-29	27

Each basecall \rightarrow quality predictor values

Quality table / Quality list \rightarrow lists combination of quality predictor scores

Quality score binning

ONT SEQUENCING – OUTPUT FILES

Basecalling → process of converting the electrical signals generated by a DNA strand passing through the nanopore → corresponding base sequence of the strand

Raw data → changes in ionic current → measured by MiniKNOW software

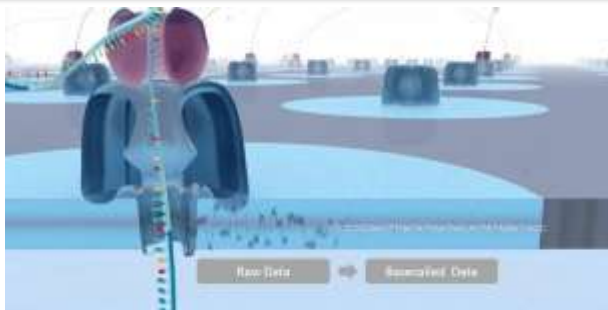
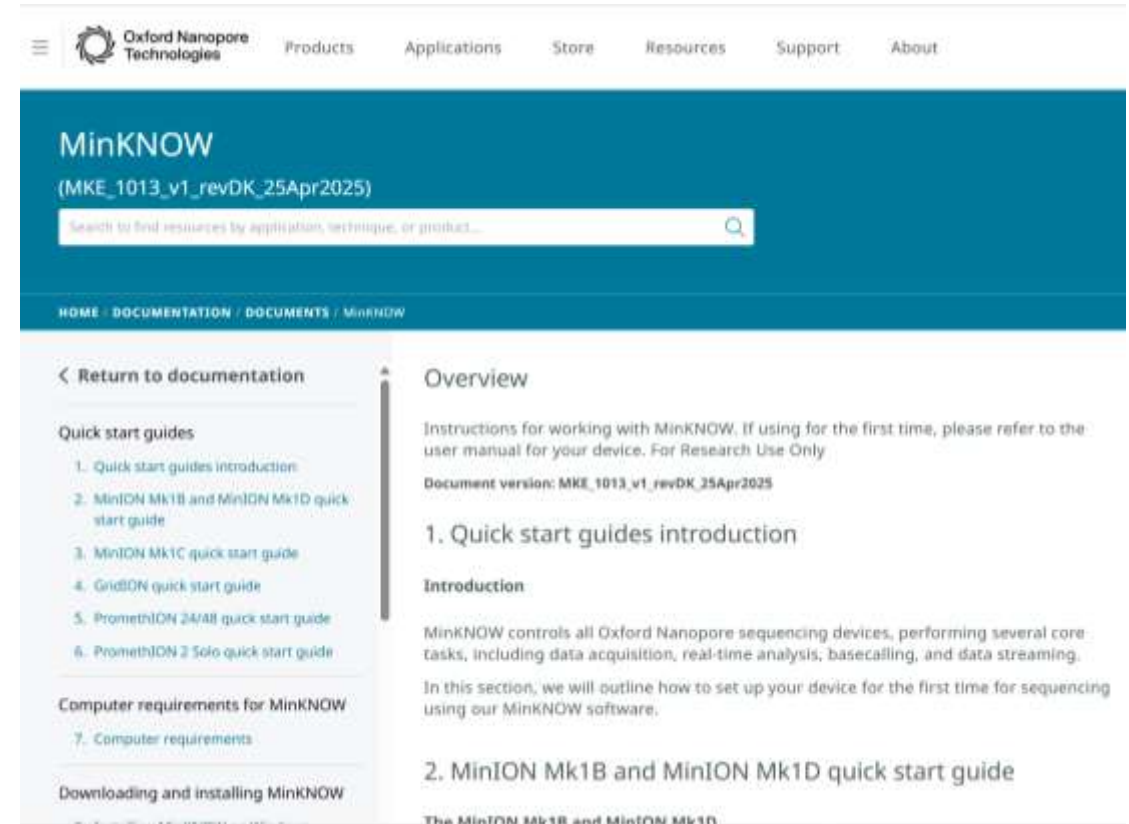
“Signals” → “reads” → POD5 output

FASTQ files also produced

Basecalling algorithm → 4000 reads / file

ONT Basecallers → MiniKNOW, Dorado, Research algorithms

Output files → POD5 , FAST5 , FASTQ

Oxford Nanopore Technologies

Products Applications Store Resources Support About

MinKNOW

(MKE_1013_v1_revDK_25Apr2025)

Search to find resources by application, technique, or product...

HOME · DOCUMENTATION · DOCUMENTS / MinKNOW

< Return to documentation

Quick start guides

1. Quick start guides introduction
2. MinION Mk1B and MinION Mk1D quick start guide
3. MinION Mk1C quick start guide
4. GridION quick start guide
5. PromethION 2A/4B quick start guide
6. PromethION 2 Solo quick start guide

Computer requirements for MinKNOW

7. Computer requirements

Downloading and installing MinKNOW

Overview

Instructions for working with MinKNOW. If using for the first time, please refer to the user manual for your device. For Research Use Only

Document version: MKE_1013_v1_revDK_25Apr2025

1. Quick start guides introduction

Introduction

MinKNOW controls all Oxford Nanopore sequencing devices, performing several core tasks, including data acquisition, real-time analysis, basecalling, and data streaming.

In this section, we will outline how to set up your device for the first time for sequencing using our MinKNOW software.

2. MinION Mk1B and MinION Mk1D quick start guide

The MinION Mk1B and MinION Mk1D

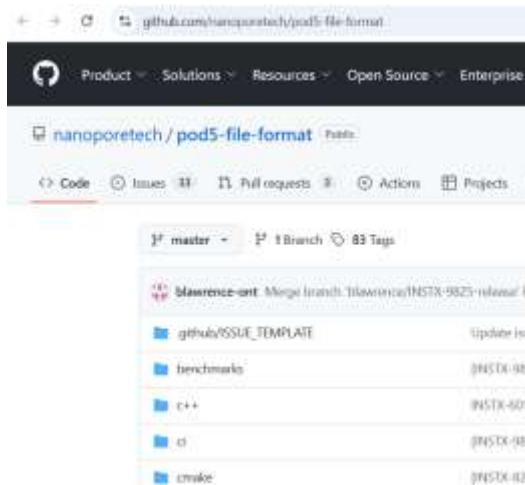
<https://nanoporetech.com/document/experiment-companion-minknow>

ONT SEQUENCING – OUTPUT FILES

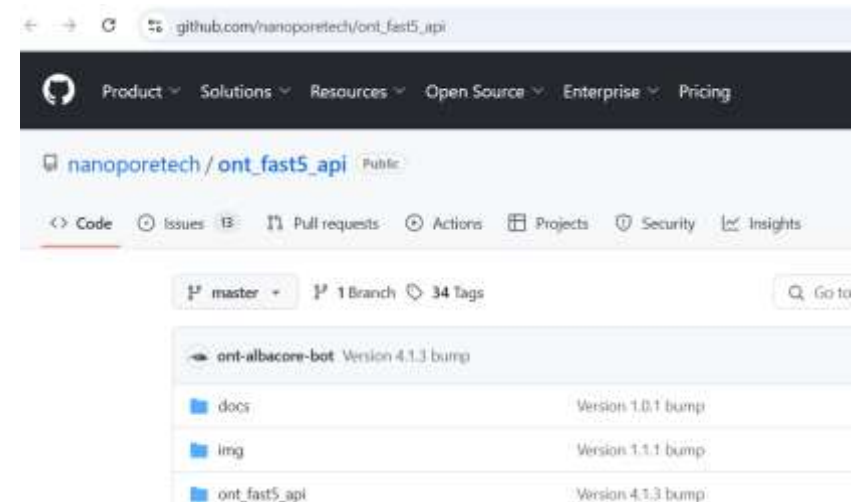
POD5 files

- More accessible
- Reads and writes faster
- Occupies less space
- Data stored using *Apache Arrow*

- Fast5 - implementation of HDF5 file format, with specific data schemas for ONT sequencing data
- HDF5 file format - a portable file format for storing and managing data
- FASTQ files



<https://github.com/nanoporetech/pod5-file-format>



https://github.com/nanoporetech/ont_fast5_api

FILE NAMING GUIDELINES

IDENTIFY YOUR FILE TYPE

USE VERSIONING

SORTING AND RETRIEVING FILES

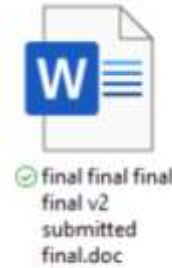
FILE NAMES

CONFIDENTIALITY AND SECURITY

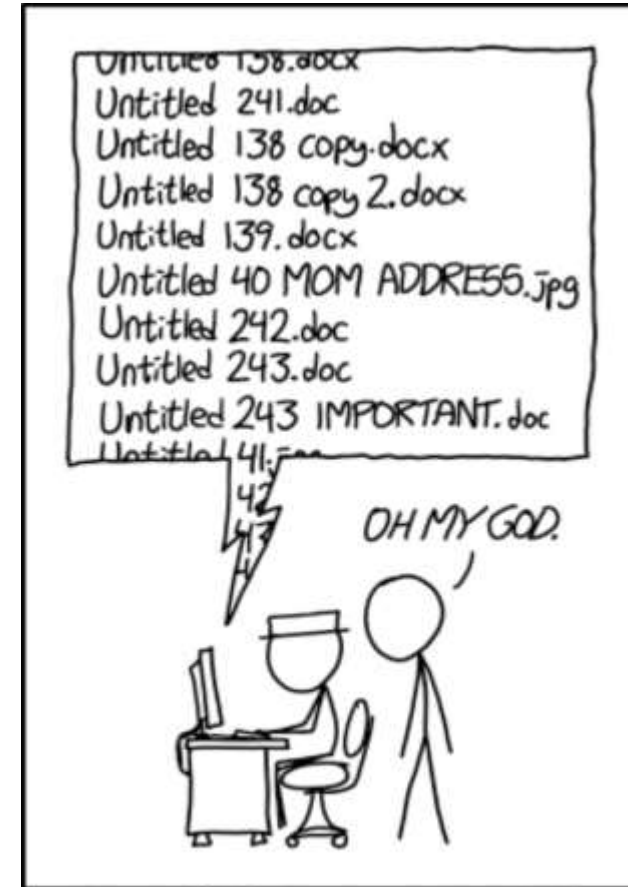
How it started



How it's going



Name	Date Modified
finished	
<ul style="list-style-type: none"> FINALS <ul style="list-style-type: none"> Final.pdf THIS ONE_FINAL_LATEST <ul style="list-style-type: none"> Screen Shot 20...9 at 1.59.20 PM Final.psd FINAL_FOR_REAL <ul style="list-style-type: none"> Final_print.pdf Final_print.ai 	<ul style="list-style-type: none"> Today, 6:14 PM Today, 12:12 PM Today, 6:15 PM Yesterday, 1:59 PM Today, 12:12 PM Today, 6:14 PM Today, 12:13 PM Today, 12:20 PM



PRO TIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

FILE NAMING GUIDELINES



Identify multiple versions of the same file

EXAMPLE:

Bacteria_203.fasta

salmonella_22.fastq

samplefromasia.fasta

EXAMPLE:

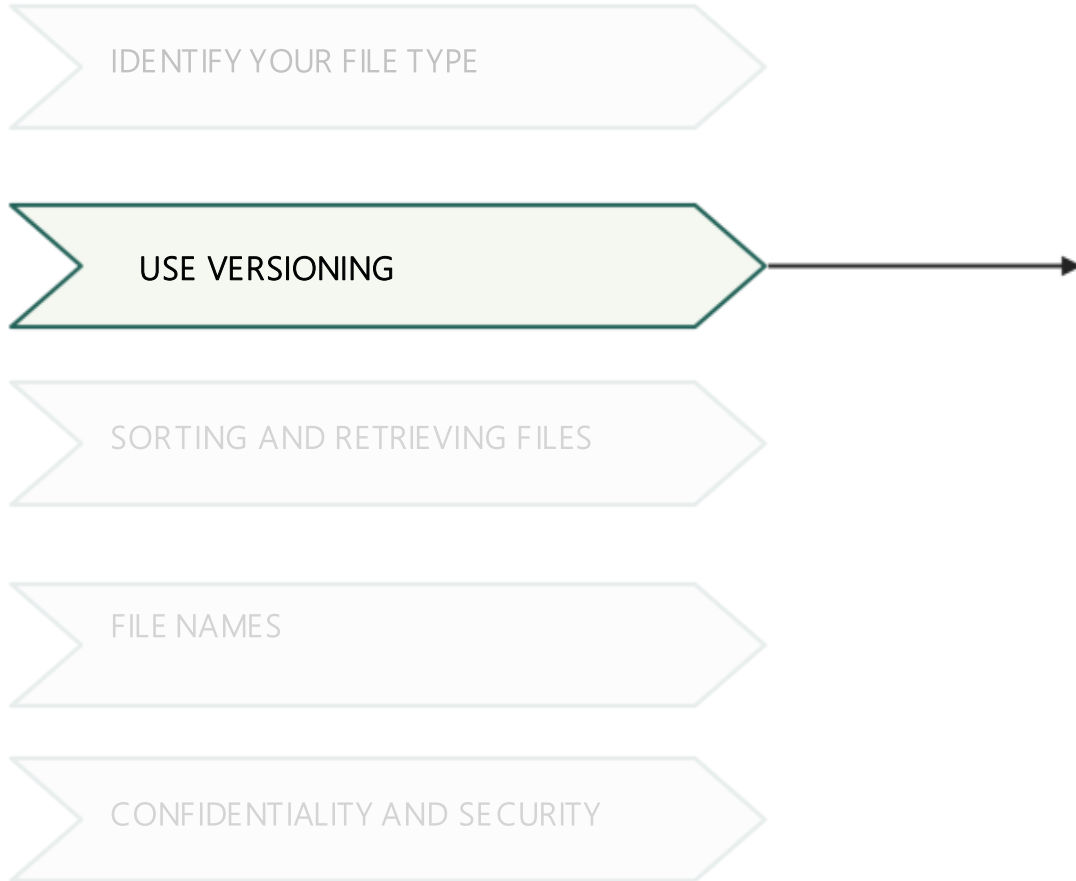
salmonella_bhutan_20190911_024.fastq

salmonella_bhutan_20190911_025.fastq

ecoli_asia_0022.fasta

ecoli_asia_0228.fasta

FILE NAMING GUIDELINES



- Identifying sample type
- Identifying sample origin
- Identifying sample ID
- identifying duplicates

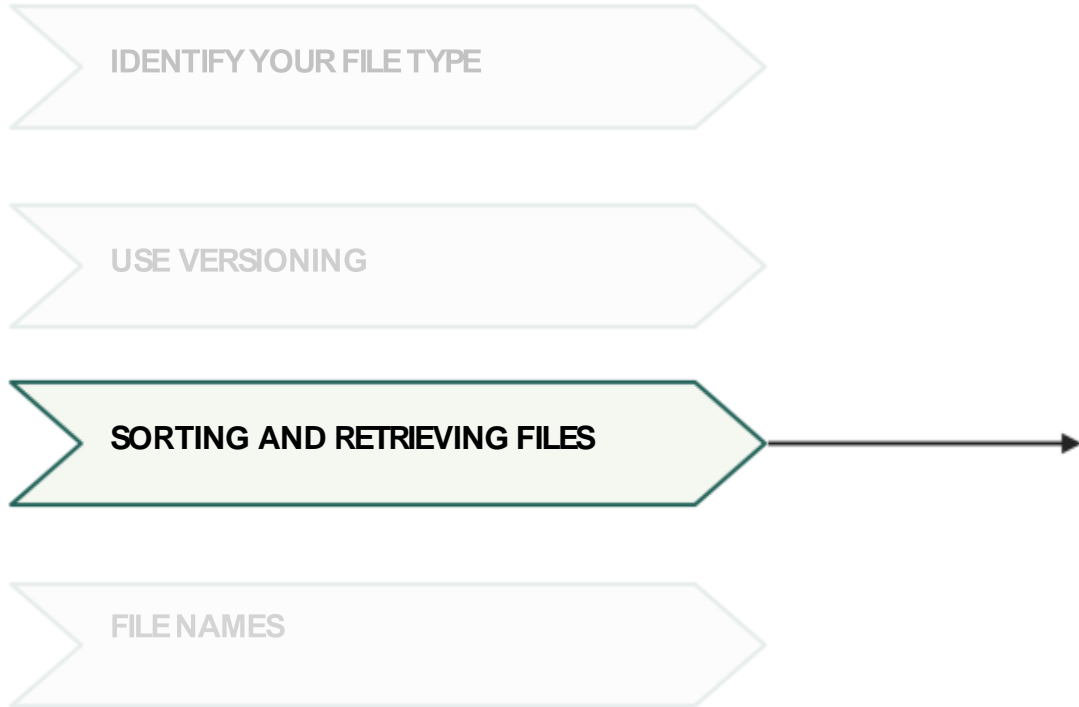
EXAMPLE:

Salmonella_illumina_batch_final.sh
 Salmonella_illumina_batch_final_FINAL.sh
 Salmonella_ente_illumina_batch_final.sh

EXAMPLE:

salmonella_illumina_batch_v01.sh
 salmonella_illumina_batch_v02.sh
 salmonella_illumina_batch_v01_01.sh

FILE NAMING GUIDELINES



- date , isolate ID , sample name , sequencing method
- sequencing method, sample name , isolate ID , date

EXAMPLE:

salmonella_2017.fasta
 salmonella_20171.fasta
 illumina_salmonella_2017.fasta

 *Not sorted correctly*

file1.csv
file10.csv
file2.csv

 *Sorted correctly*

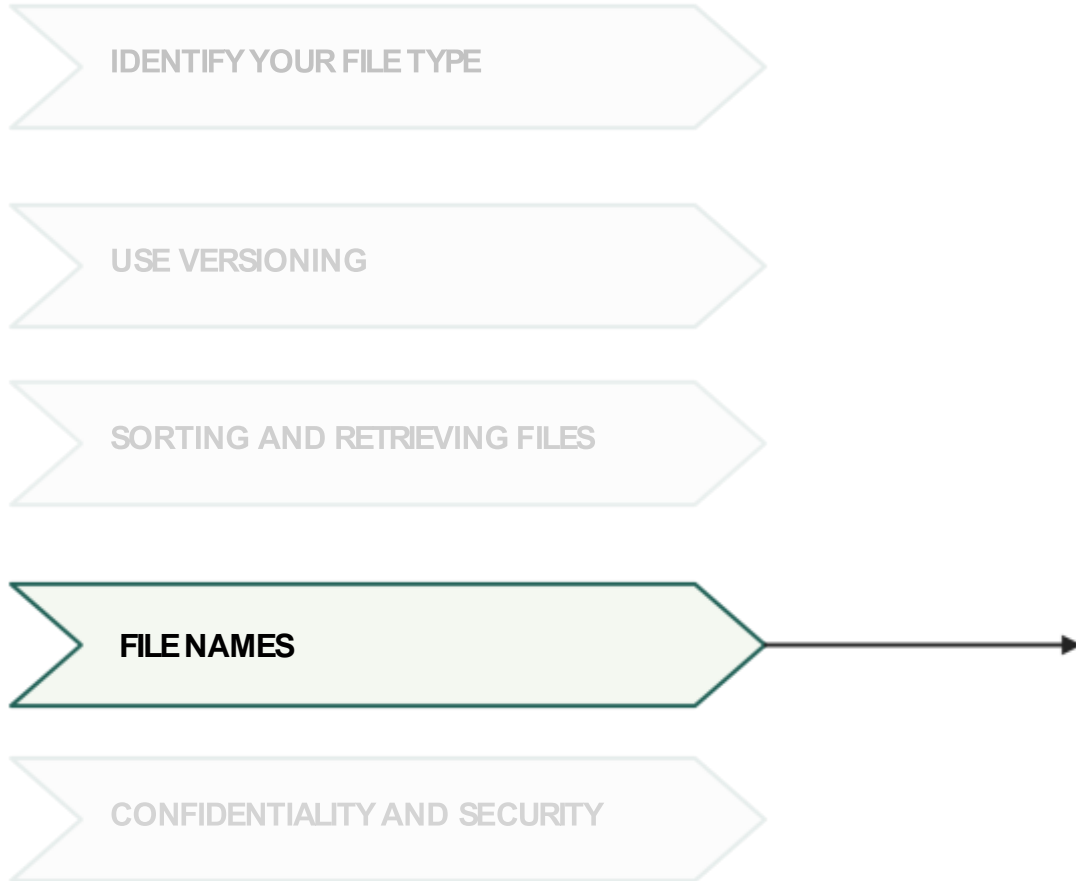
file01.csv
file02.csv
file10.csv

EXAMPLE:

salmonella_illumina_20171203_0022.fasta
 salmonella_illumina_20171203_0023.fasta

20171203_salmonella_illumina_001.fasta
 20171203_salmonella_illumina_002.fasta
 20171203_salmonella_illumina_003.fasta
 20171203_ecoli_illumina_001.fasta
 20171203_ecoli_illumina_002.fasta

FILE NAMING GUIDELINES



- Avoid vague words
- Avoid full stops
- Avoid special characters
- Avoid long names
- Maybe abbreviations
- Follow consistent date format (YYYYMMDD)

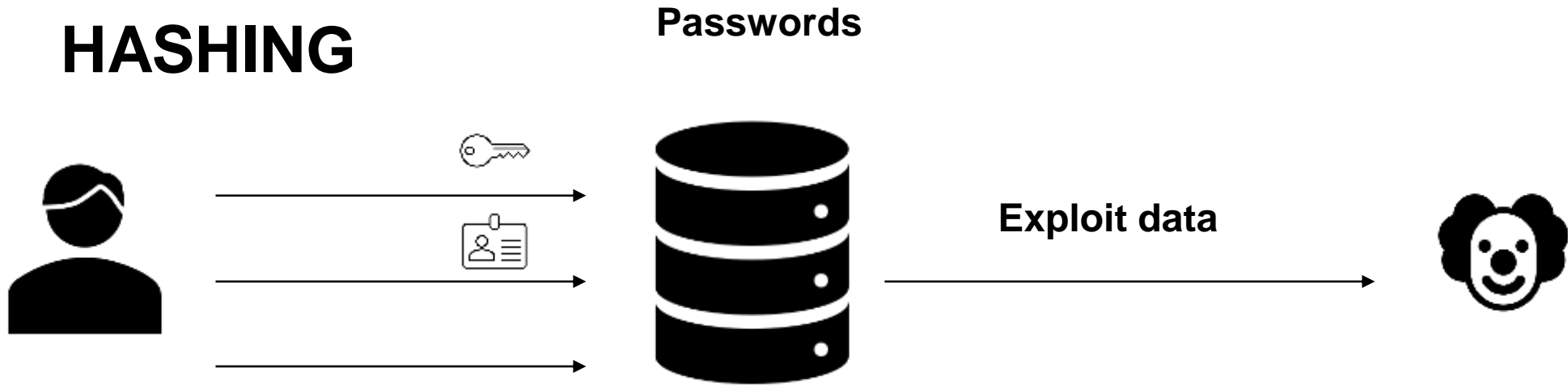
EXAMPLE:

```
salmonella_sample_first.fasta
salmonella_sample_second_last.fasta
salmonella_sample_last.fasta
salmonella_sample_last_LAST.fasta
```

EXAMPLE:

```
salmonella_india_20231212_001.fasta
salmonella_india_20231212_098.fasta
salmonella_india_20231212_099.fasta
salmonella_india_20231212_100.fasta
```

HASHING



`2cf24dba5fb0a30e26e83b2ac5b`

MD5 HASH

MD5 Hash Generator

Use this generator to create an MD5 hash of a string:

hello

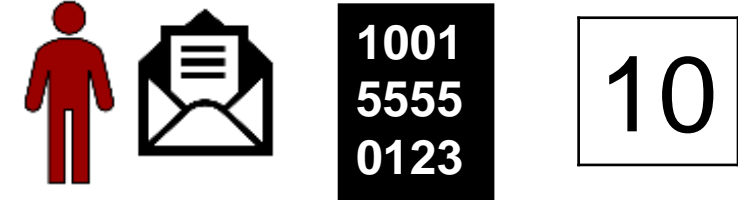
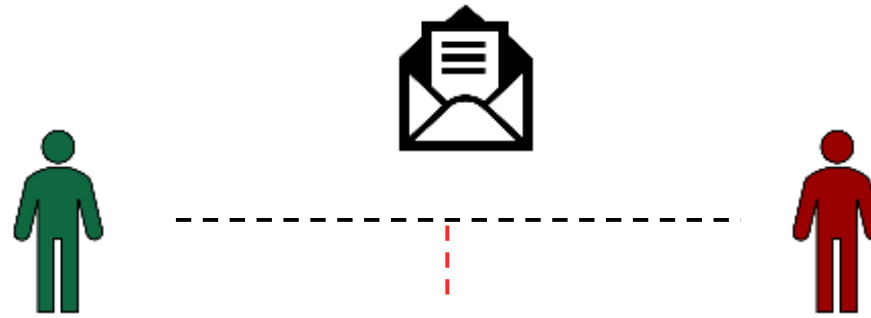
Generate →

Your String	hello
MD5 Hash	5d411402abc4b2a76b9719d911017c592 Copy

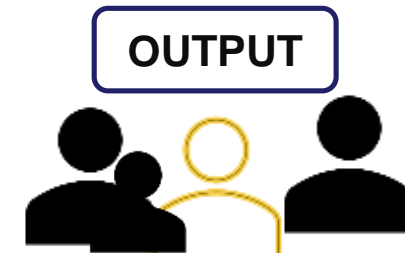
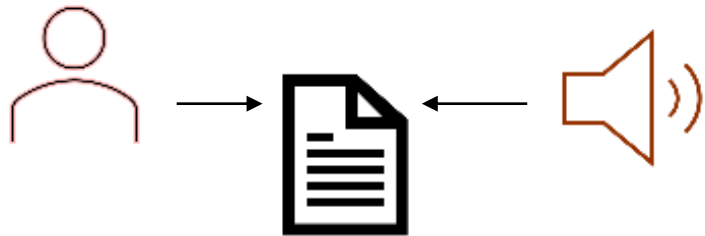
DATA + HASHING FUNCTION = HASH
Cannot get original data from a HASH
Same DATA = Same HASH ; Different data = Different Hash
verified by using HASH
Has the HASH been altered?

<https://www.md5hashgenerator.com/>

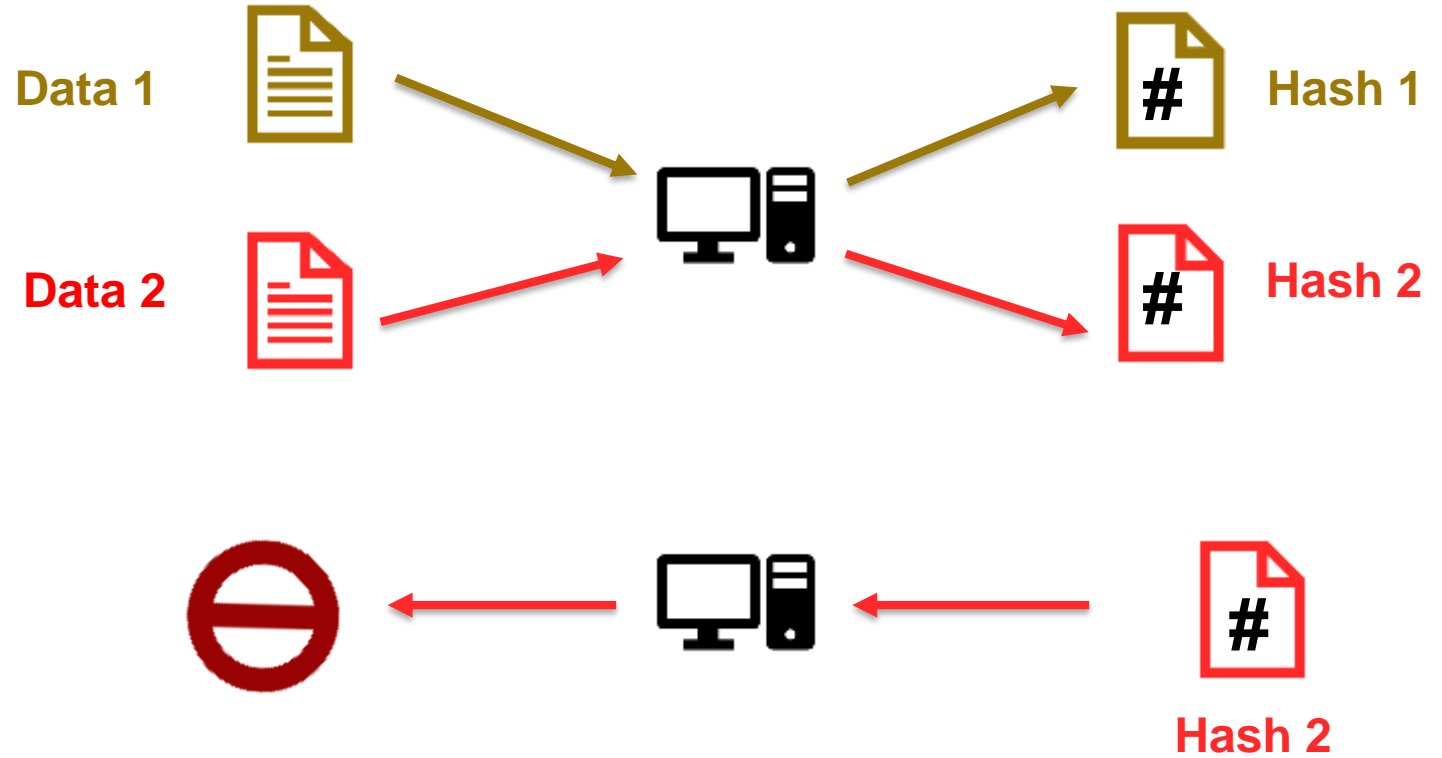
CHECKSUMS



CHECKSUM → complex cryptographic hash functions



- **SHA-256** → Secure Hash Algorithm 256-bit
- Outputs a fixed 256-bit (64 hexadecimal characters) hash from any input.
- Input data → passed through a cryptographic hash function → outputs a unique hash
- Even a 1-bit change in input = a completely different hash
- same input = same hash
- 4 bits = 1 character

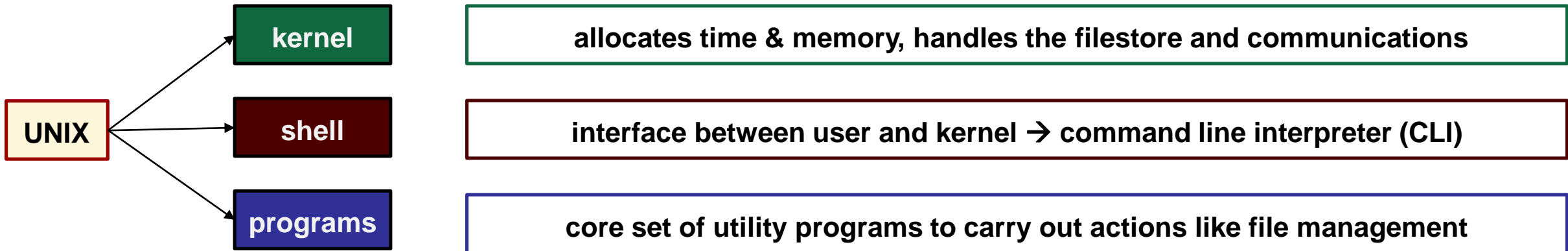


SHA family	Bit size	Hash characters
SHA - 1	160 bits	40 characters
SHA - 224	224 bits	56 characters
SHA - 256	256 bits	64 characters
SHA - 384	384 bits	96 characters
SHA - 512	512 bits	128 characters

LINUX / UNIX operating systems



- the suite of programs which make the computer work
- It is stable, multi-user, multi-tasking system for servers, desktops and laptops



LINUX vs UNIX → share the same idea of the shell and kernel

UNIX → used for high end operations

LINUX → easily downloadable and operable

```

mark@linux-desktop: ~
File Edit View Search Terminal Help
mark@linux-desktop:~$ pwd
/home/mark
mark@linux-desktop:~$
  
```

UNIX COMMANDS

Allows users to customize shell environment and write own shell scripts

Usage of pipes (|) → links workflows unidirectionally

Eg: `ls | grep file.txt`

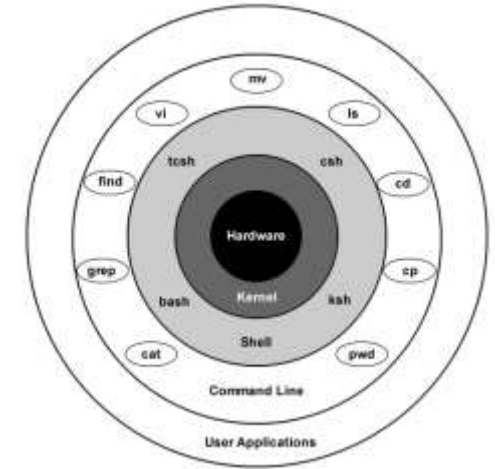
```

root@gfgubun1:/home/jayesh# ls
Desktop      file.txt      Pictures      Templates
Documents    'hello jayesh this side'  Public        Videos
Downloads    Music         snap
root@gfgubun1:/home/jayesh# ls | grep file.txt
file.txt
root@gfgubun1:/home/jayesh# █

```

More commands

<https://www.w3schools.com/bash/>



Linux / Unix TUTORIAL

[JupyterHub](#)

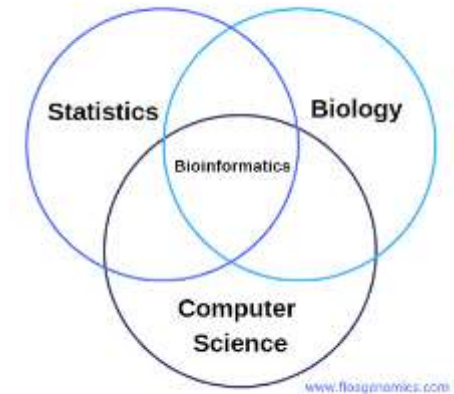


BIOINFORMATICS

Using computer technology to collect, store, analyze and disseminate biological data

GOALS:

- Development of efficient algorithms
- Extension of experimental data by predictions
- Increase understanding of biological processes
- Interpretation of biological data to correlate results
- Solve practical problems in data storage, management and sharing



DATA IN BIOINFORMATICS

- Classic data: DNA sequences of genes, full genomes , amino acid sequence of proteins
- “omics” data: transcriptomics, proteomics, interactomics, metabolomics
- Metagenomics and Metaproteomics



COMPUTATIONAL COMPONENTS



HARDWARE

High-performance computing (HPC) clusters
 Cloud-based servers → AWS, Google Cloud, Azure
 Local machines → workstations or laptops

OPERATING SYSTEMS

Windows, Mac, Linux, UNIX

SOFTWARE TOOLS

Sequence alignment tools → BLAST, BWA
 Data visualization tools → IGV, UCSC Genome Browser
 Statistical tools → R, Bioconductor

PROGRAMMING LANGUAGES

Python, Perl, R, Bash

PACKAGES

Conda, Bioconda,

DATABASES

NCBI, UniProt, EMBL,

WORKFLOW MANAGEMENT SYSTEMS

Nextflow, Snakemake, Galaxy



BIOCONDA





HARDWARE

High-performance
Cloud-based servers
Local machines →



! COMMON ISSUES !

- i. Dependency conflicts
- ii. Installation issues
- iii. Reproducibility concerns across systems



PACKAGES

... mac, Linux, UNIX
... tools → BLAST, BWA
... → IGV, UCSC Genome Browser
... tools → R, Bioconductor
Python, Perl, R, Bash

BIOCONDA



SOLUTION
Docker containers, Virtual Machines
Or using environments !

WORKFLOWS



Nextflow, Snakemake, Galaxy



Developing Oracle weblogic software



Testing out the software



Deploying the software

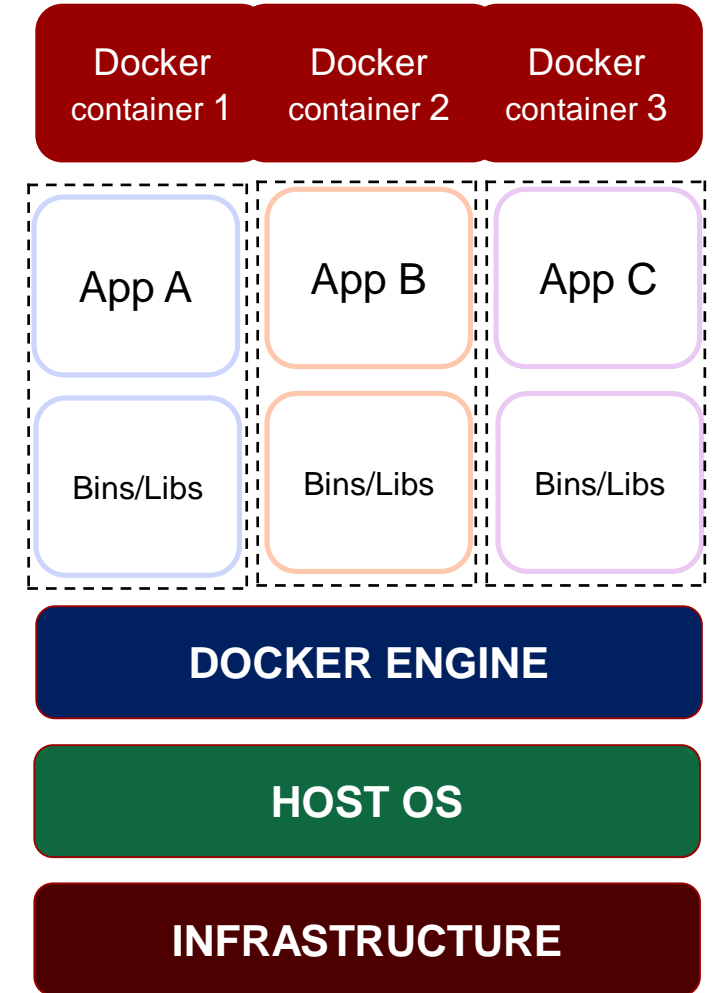
Installation – 3x !!!!

DOCKER CONTAINERS

- Open-source containerization software platform
- Create deploy and manage applications in containers
- Applications + environment → provided in parallel

Steps:

- i. Develop application and supporting components using containers
- ii. Container → unit for distributing and testing out application
- iii. Deploy application into production environment (local data center / a cloud provider / hybrid)

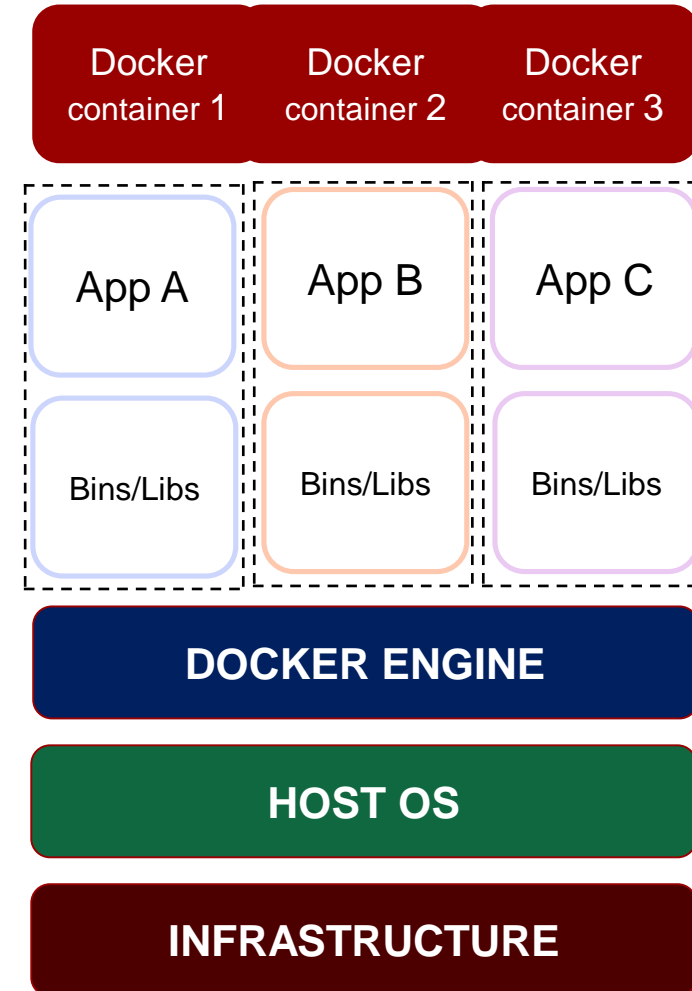


CONTAINERS

- Containers help in creating an isolated environment → “sandbox”
- Virtualize the host OS and isolate an applications dependencies from other containers running on the same machine
- Uses → portability, lightweight, less disk space, secure, speed
- Container engine + container image → package of application + its dependencies

Advantage

- ✓ No need to run separate operating system / application
- ✓ Lower costs
- ✓ Higher resource utilization

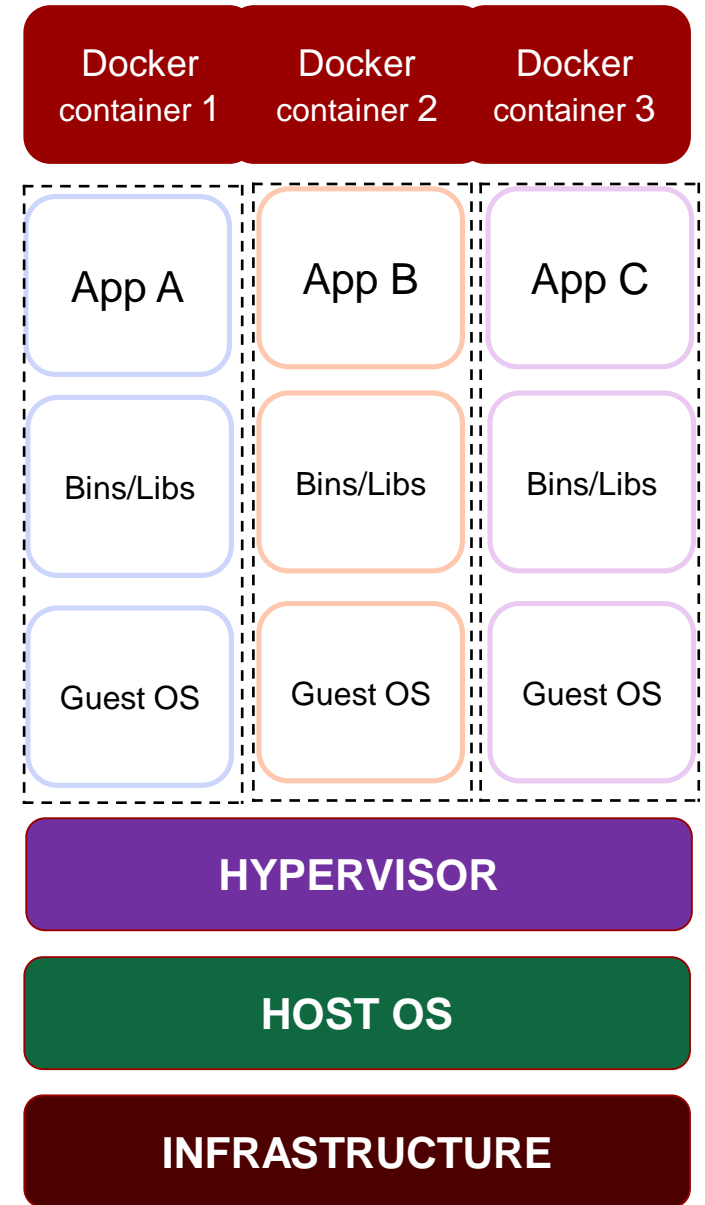


VIRTUAL MACHINES

- VM → individual computer inside a host machine (computer)
- Multiple VMs can live inside a single host machine
- Uses host machine's hardware
- Hypervisor → piece of software that virtualizes the host's hardware and manages feeding resources to the VMs

Disadvantages:

- ✓ computational overhead
- ✓ Own OS installed → requires several Gb of storage
- ✓ Takes time to install → has to be done each time



COMPUTING ENVIRONMENTS IN BIOINFORMATICS

environments make it easy to install a wide variety of command-line tools → prevents them from interfering with one another

The logo for Conda, featuring the word "CONDA" in green capital letters with a registered trademark symbol. The letter "C" is stylized with a circular pattern of dots.

Main framework



Speeds up conda installations

The logo for Bioconda, featuring the word "BIOCONDA" in green capital letters with a registered trademark symbol.

Additional packages in Bioinformatics

COMPUTING ENVIRONMENTS IN BIOINFORMATICS

environments make it easy to install a wide variety of command-line tools → prevents them from interfering with one another



Main framework

- Create environments
- Listing environments
- Installing packages
- Specifying channels



Speeds up conda installations

```
conda create -n <env-name>
```

```
conda info --envs
```

```
# via environment activation
conda activate myenvironment
conda install matplotlib
```

```
conda install conda-forge::numpy
```



Additional packages in Bioinformatics

```
conda create -n myenvironment python numpy pandas
```

```
conda environments:
```

```
base          /home/username/Anaconda3
myenvironment * /home/username/Anaconda3/envs/myenvironment
```

```
# via command line option
conda install --name myenvironment matplotlib
```

<https://docs.conda.io/projects/conda/en/stable/user-guide/getting-started.html>

COMPUTING ENVIRONMENTS IN BIOINFORMATICS

environments make it easy to install a wide variety of command-line tools → prevents them from interfering with one another



Main framework



Speeds up conda installations



Additional packages in Bioinformatics

- Create environments
- Listing environments and installing packages

```
mamba create -n nameofmyenv <list of packages>
```

```
mamba install
```

```
mamba create -n myjlabenv jupyterlab -c conda-forge
mamba activate myjlabenv # activate our environment
jupyter lab # this will start up jupyter lab and open a browser
```

https://mamba.readthedocs.io/en/latest/user_guide/mamba.html

COMPUTING ENVIRONMENTS IN BIOINFORMATICS

environments make it easy to install a wide variety of command-line tools → prevents them from interfering with one another



Main framework

Speeds up conda installations

Additional packages in Bioinformatics

ANACONDA.ORG

Search Anaconda.org

About Anaconda Help Download Anaconda Sign In

bioconda / packages

Filters
 Type: all Access: public Label: all

Package Name	Access	Summary	Updated
orthanc	public	Uncertainty aware HLA typing and general haplotype quantification.	2025-05-28
heasoft	public	NASA High Energy Astrophysics Software (HEASoft)	2025-05-28
alignor	public	A tool for creating alignment plots from bam files.	2025-05-28
metagraph	public	Ultra Scalable Framework for DNA Search, Alignment, Assembly.	2025-05-28
gpsw	public	GPSW is a tool for analysing Global Protein Stability Profiling data.	2025-05-28
openstructure	public	Open-Source Computational Structural Biology Framework	2025-05-28
salmon	public	Highly-accurate & biased fast transcript-level quantification from RNA-seq reads using selective alignment	2025-05-28
smalgenomutilities	public	A collection of scripts that are useful for dealing with viral RNA NGS data.	2025-05-28
magmax	public	MAGmax is a robust tool for dereplicating MAGs through bin merging and reassembly.	2025-05-28

- Install conda first
- always cite it

```
conda config --add channels bioconda
conda config --add channels conda-forge
conda config --set channel_priority strict
```

<https://bioconda.github.io/>

THANK YOU FOR LISTENING, ANY QUESTIONS ?

Kindly ensure to drop your email address on the chat to receive the certificates