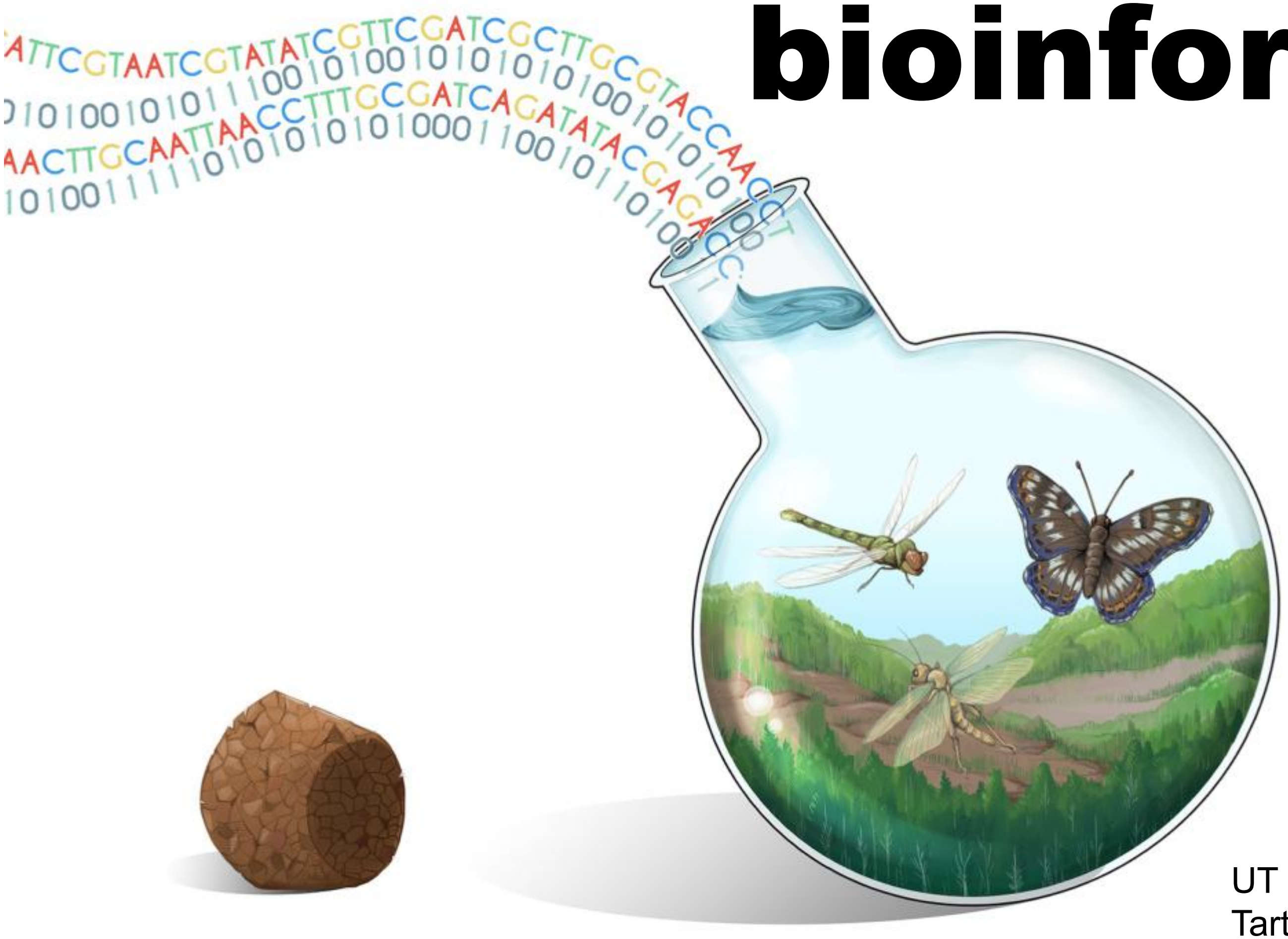
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

MYCOLOGY
& MICRO
BIOLOGY

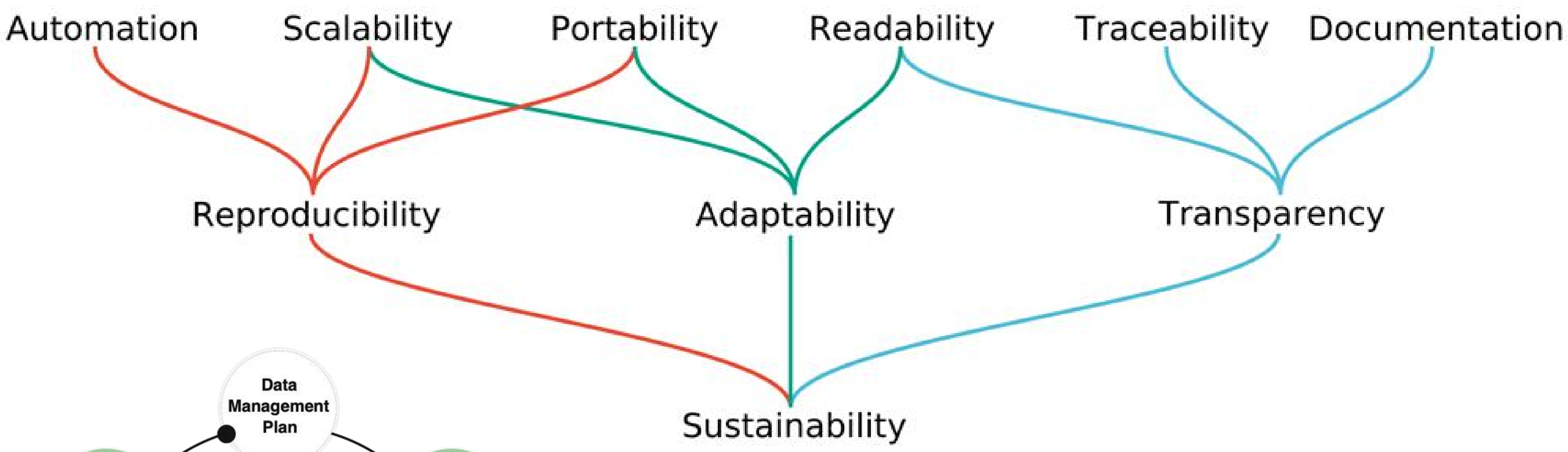
Modern technologies in bioinformatics

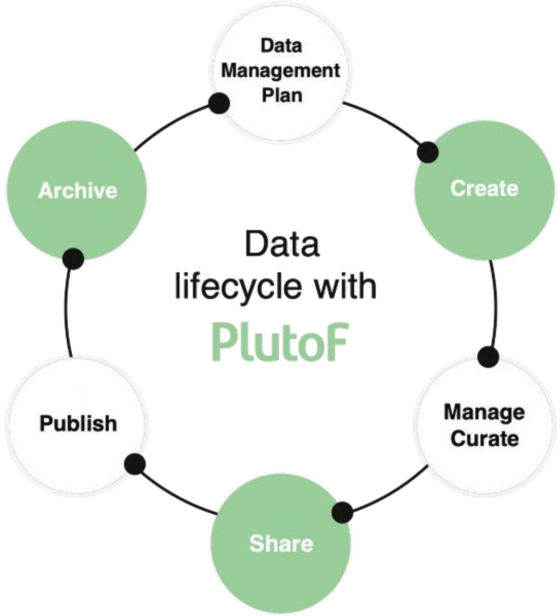


Vladimir Mikryukov

UT International Summer University,
Tartu, August 01-05 2022

Aspects of sustainable data analysis

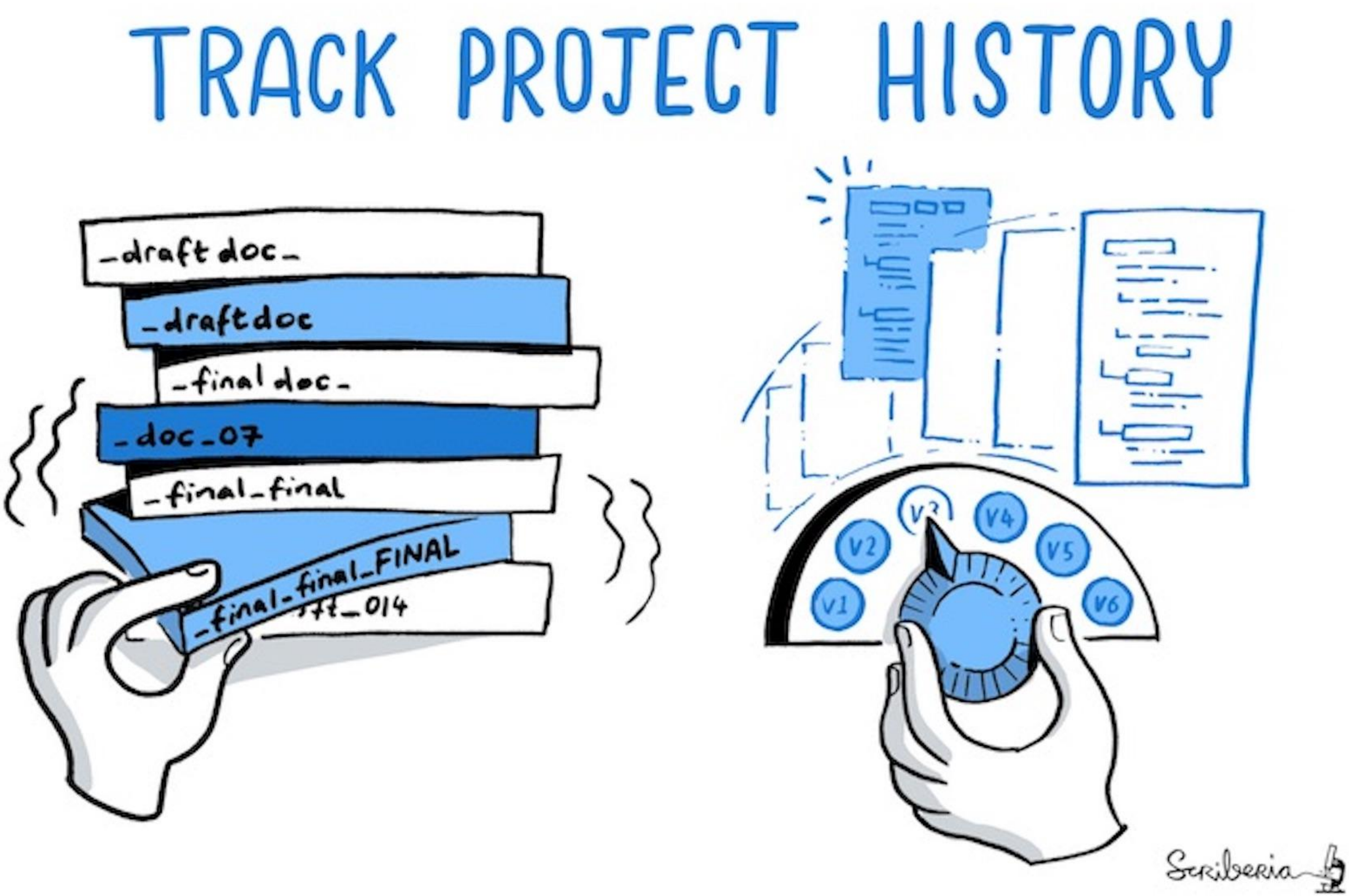




Abarenkov et al. (2010)
DOI:10.4137/EBO.S6271

Mölder et al. (2021)
DOI:10.12688/f1000research.29032.2

Version control

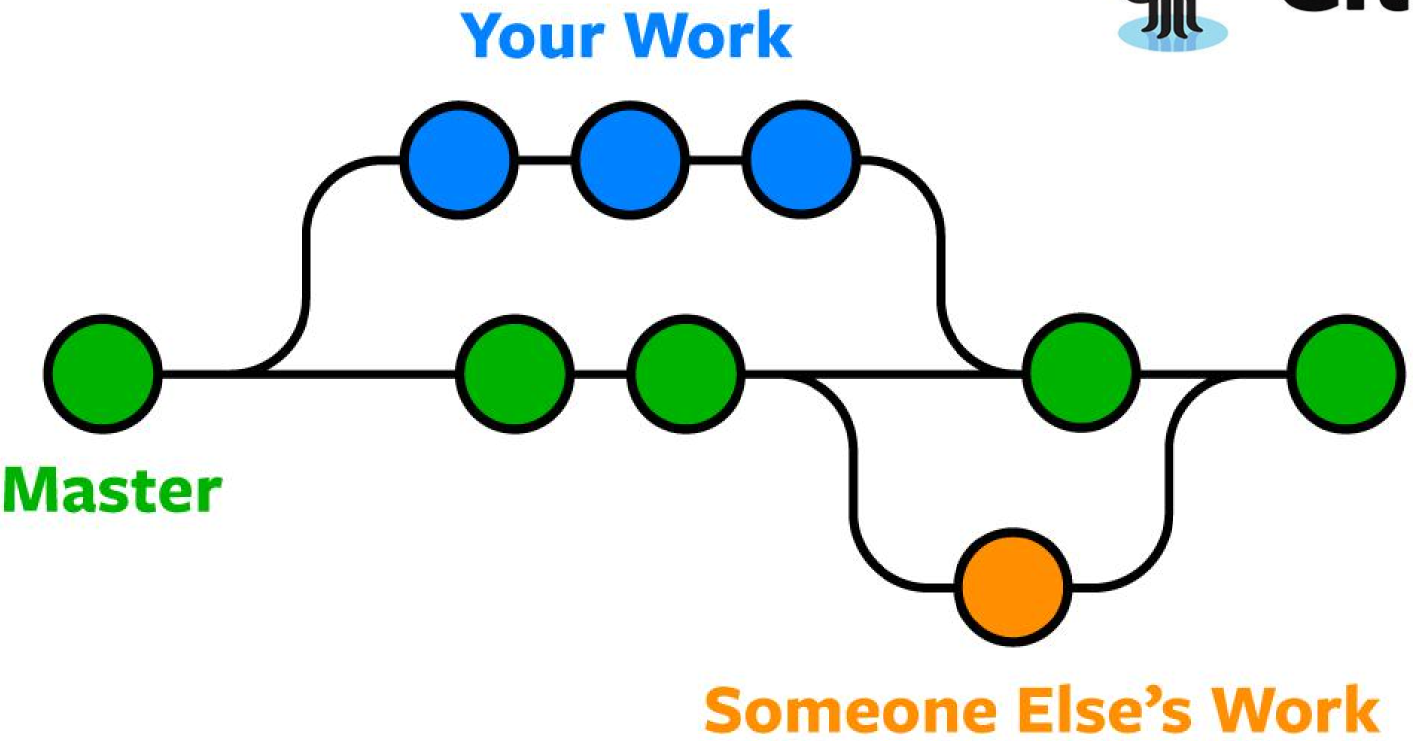


The Turing Way // Scriberia. DOI: 10.5281/zenodo.3332807

Version control, collaborative working

git

GitHub



https://www.nobledesktop.com/learn/git/what-is-git

In case of fire

1. git commit
2. git push
3. leave building

https://github.com/hendrixroa/in-case-of-fire

Software installation

Manual installation

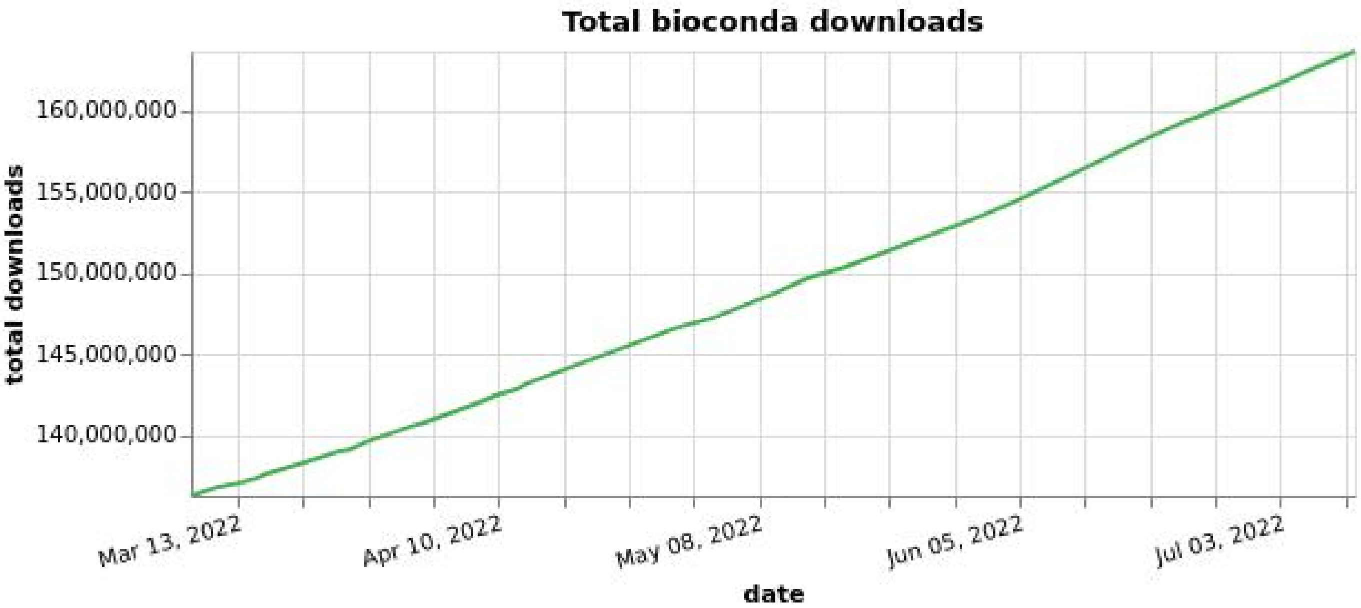
```
sudo apt-get install build-essential autoconf automake libtool
git clone https://github.com/xflouris/PEAR.git
cd PEAR
./autogen.sh
./configure
make
sudo make install
```

CONDA

```
conda install -c bioconda pear
```

Software installation

CONDA
BIOCONDA®



Grüning et al. (2018)
DOI:10.1038/s41592-018-0046-7

https://bioconda.github.io/

Software environments

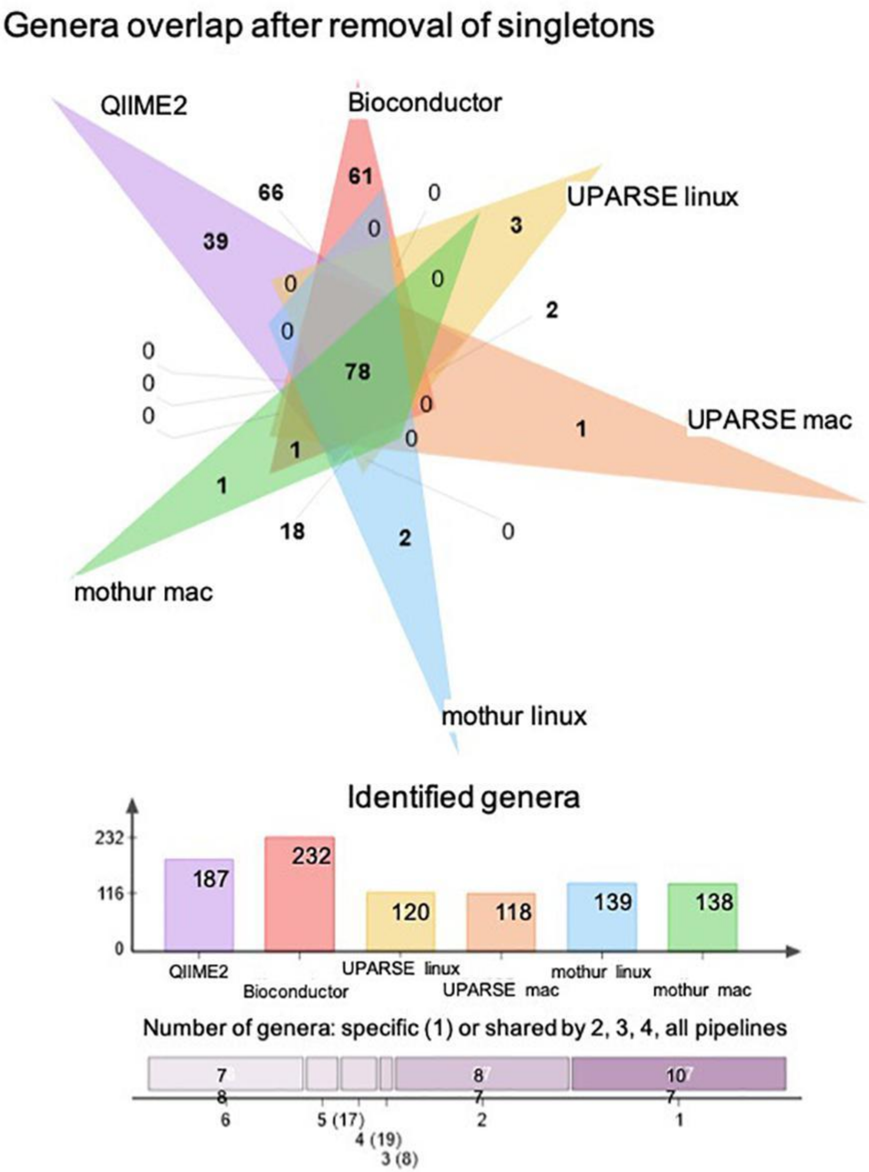
- Package, dependency, and environment management
- Large ecosystem of pre-packaged software
- Specific versions

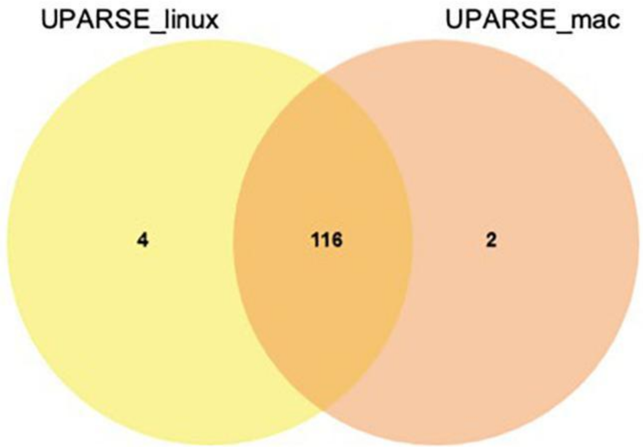
```
conda install -c bioconda blast=2.13.0
```

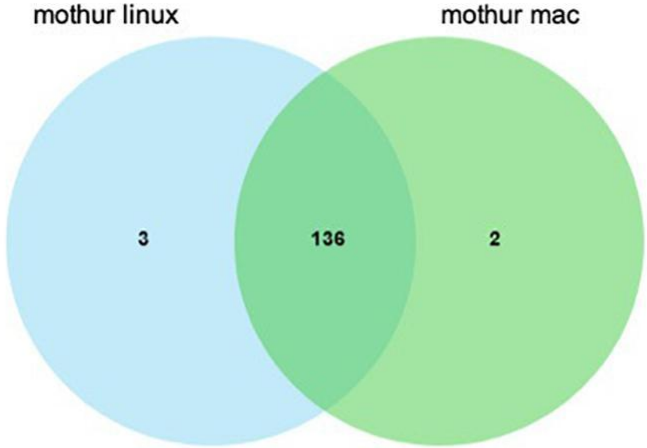
- Multiple environments

```
conda --name OLDBLAST -c bioconda blast=2.13.0
conda --name NEWBLAST -c bioconda blast=2.5.0

conda activate OLDBLAST
blastn --version

conda activate NEWBLAST
blastn --version
```

Reproducibility vs OS (Linux and Mac)

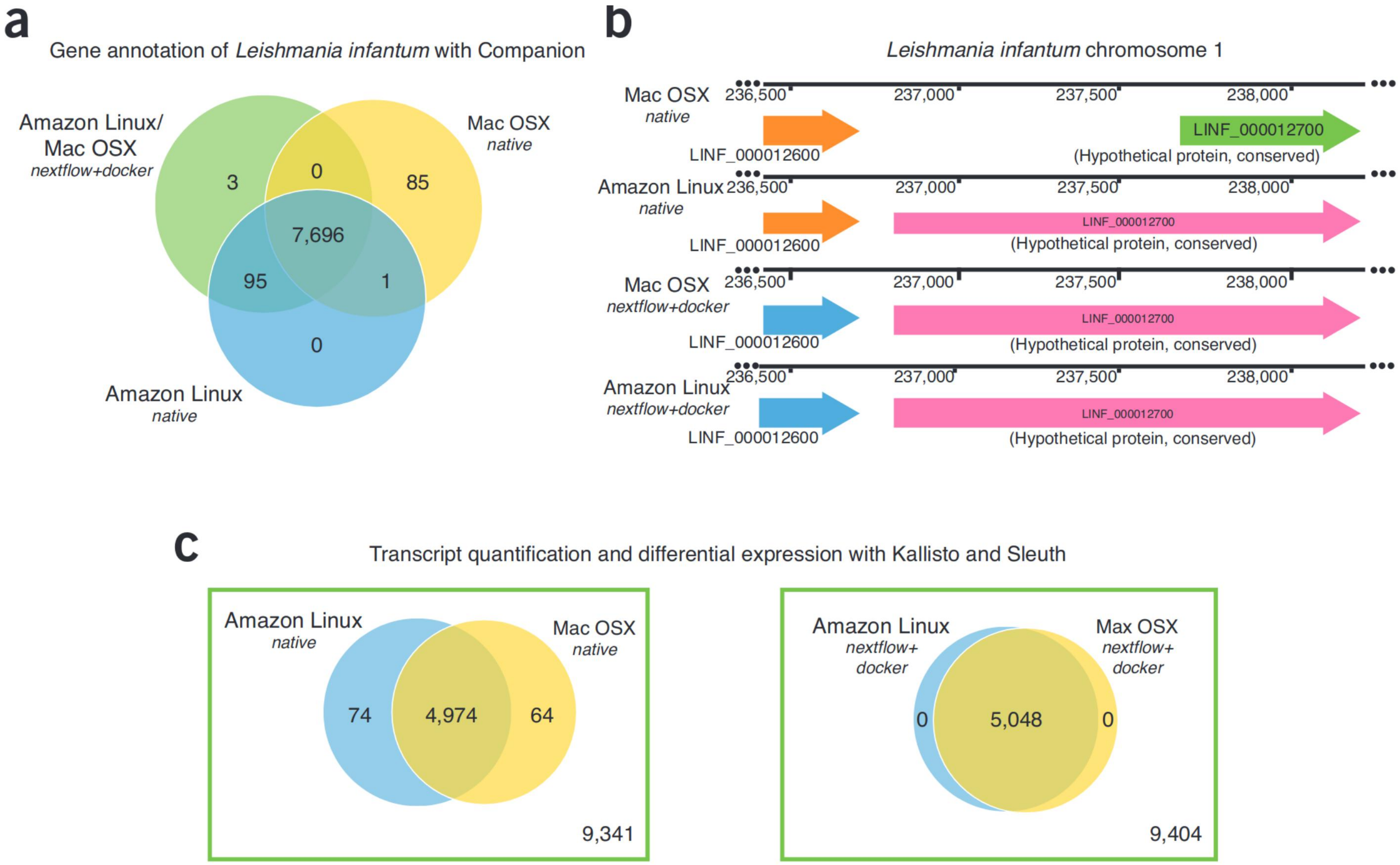

Genera overlap after removal of singletons





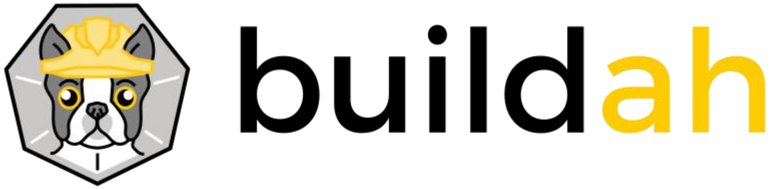
Marizzoni et al. (2020)
DOI:10.3389/fmicb.2020.01262

Stable analyses on different platforms in Dockerized environment

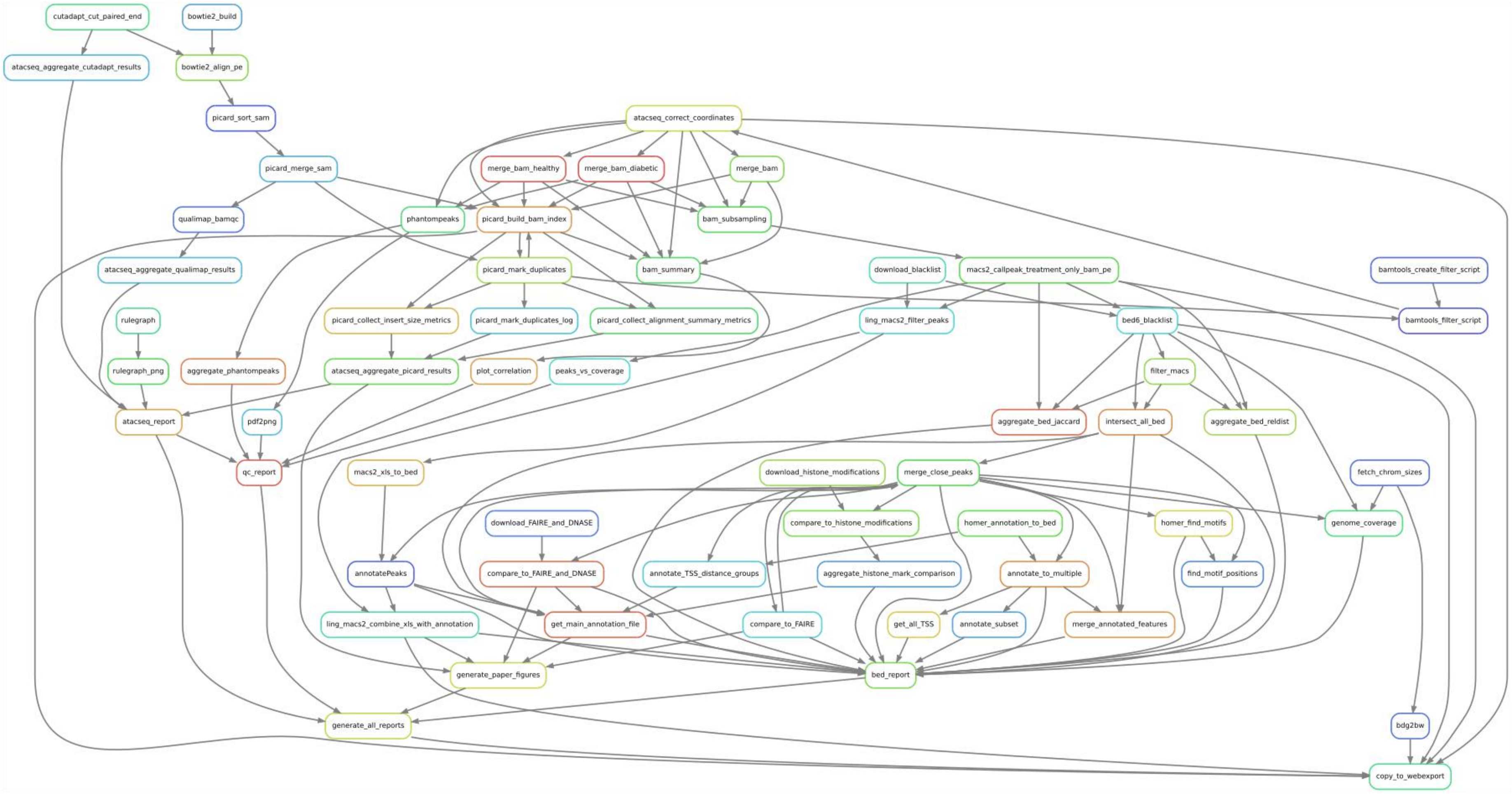


Di Tommaso et al. (2017) DOI:10.1038/nbt.3820

Containers



Complex pipelines

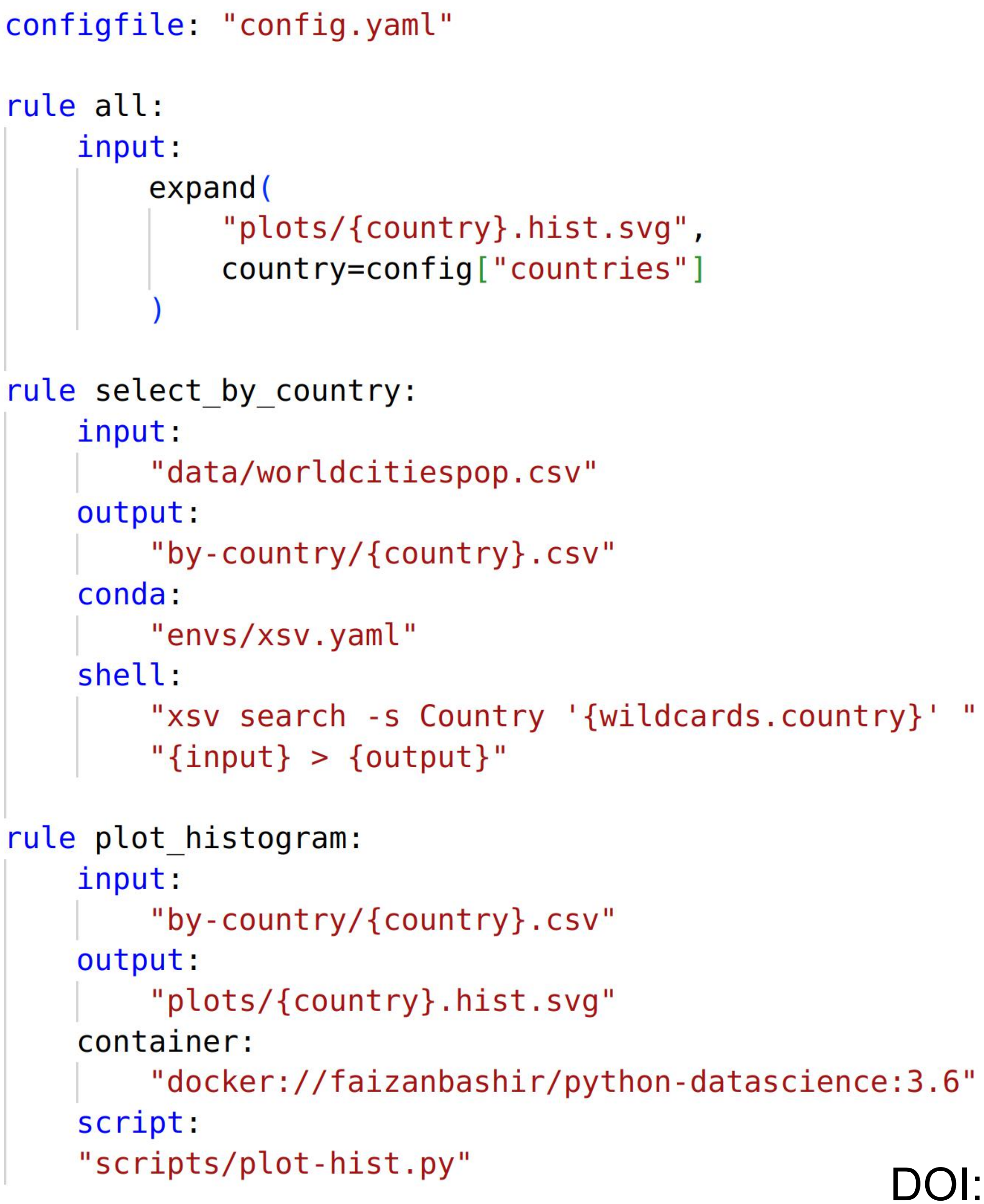


Workflow
management systems

Snakemake
https://snakemake.github.io/
Mölder et al. (2021)

Nexflow
https://nextflow.io/
Di Tommaso et al. (2017)
DOI:10.1038/nbt.3820

Targets
https://docs.ropensci.org/targets/
Landau (2021)
DOI:10.21105/joss.02959

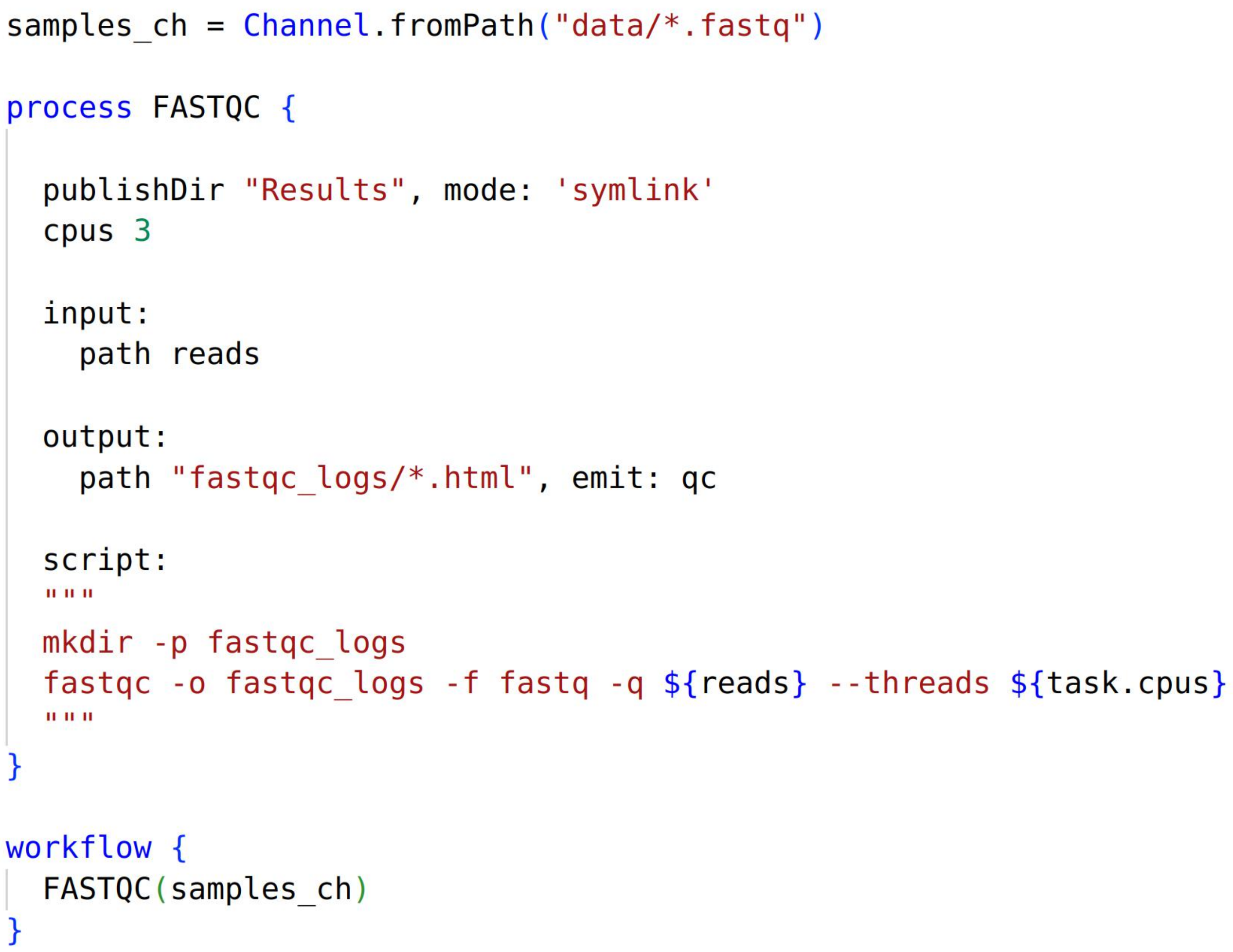


```
configfile: "config.yaml"

rule all:
    input:
        expand(
            "plots/{country}.hist.svg",
            country=config["countries"]
        )

rule select_by_country:
    input:
        "data/worldcitiespop.csv"
    output:
        "by-country/{country}.csv"
    conda:
        "envs/xsv.yaml"
    shell:
        "xsv search -s Country '{wildcards.country}' "
        "{input} > {output}"

rule plot_histogram:
    input:
        "by-country/{country}.csv"
    output:
        "plots/{country}.hist.svg"
    container:
        "docker://faizanbashir/python-datascience:3.6"
    script:
    "scripts/plot-hist.py"
```

snakemake
A framework for reproducible data analysis

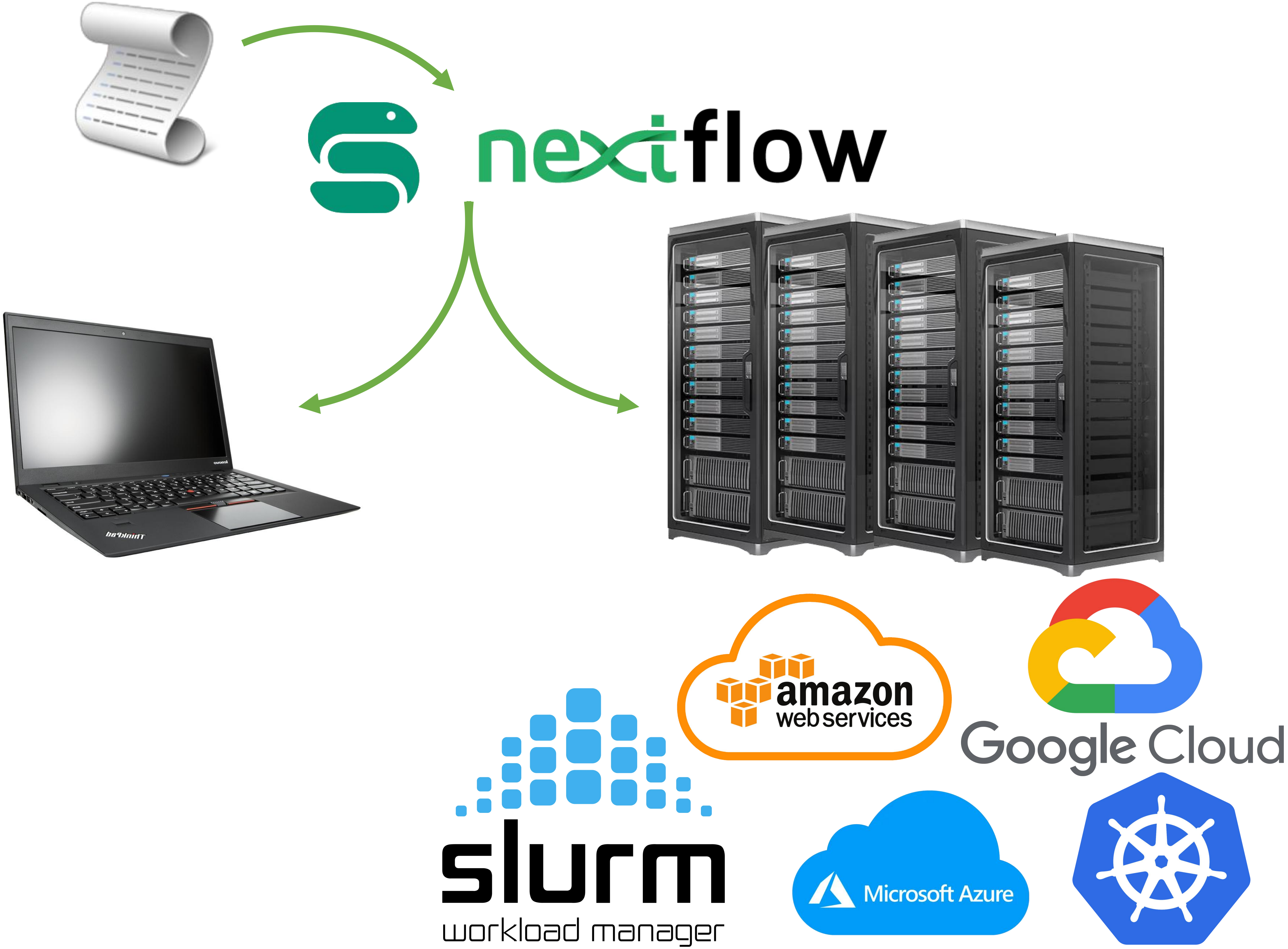
Mölder et al. (2021)
DOI:10.12688/f1000research.29032.2

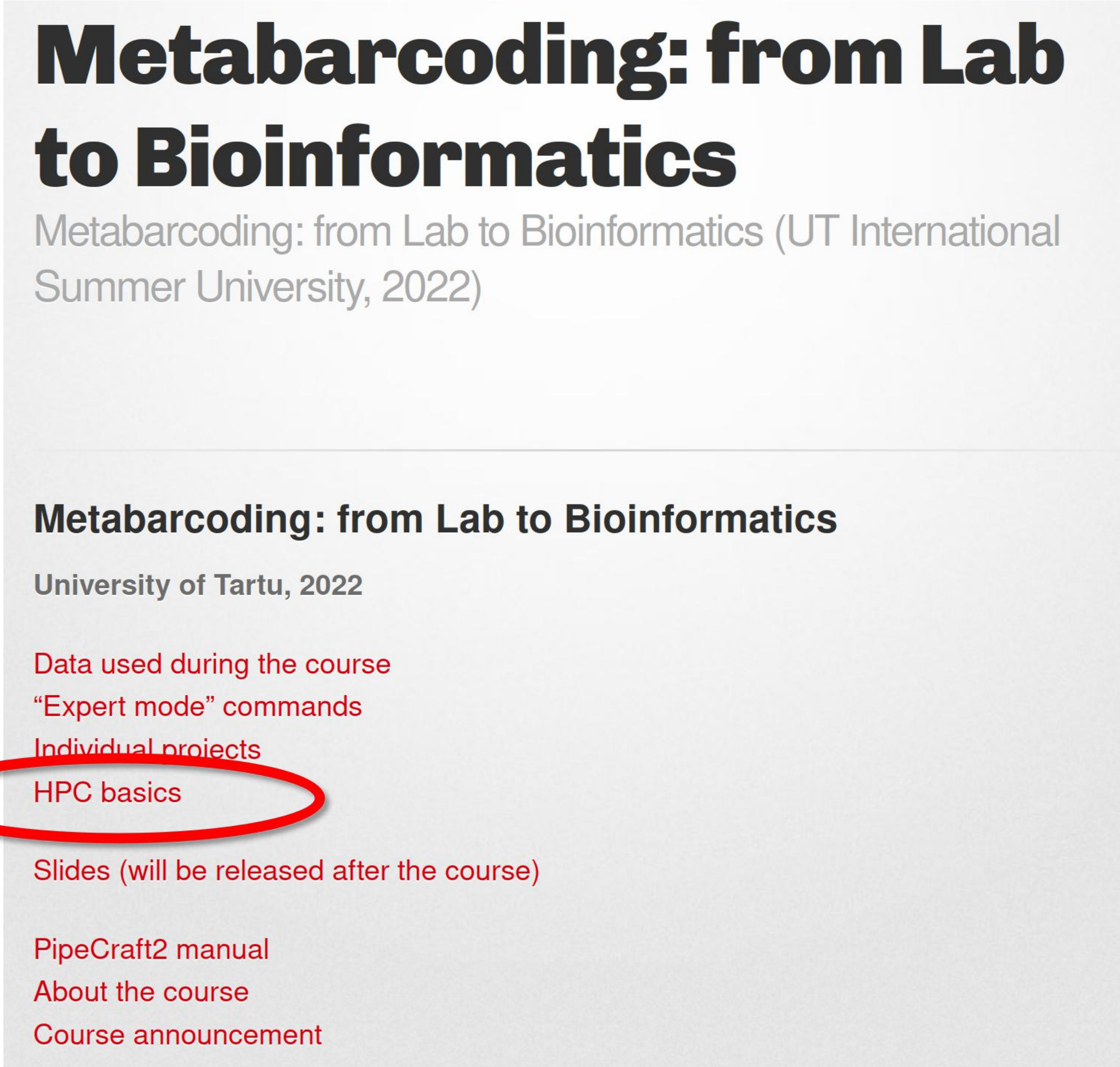
https://snakemake.github.io/



```
samples_ch = Channel.fromPath("data/*.fastq")

process FASTQC {

  publishDir "Results", mode: 'symlink'
  cpus 3

  input:
    path reads

  output:
    path "fastqc_logs/*.html", emit: qc

  script:
  """
  mkdir -p fastqc_logs
  fastqc -o fastqc_logs -f fastq -q ${reads} --threads ${task.cpus}
  """
}

workflow {
  FASTQC(samples_ch)
}
```

Di Tommaso et al. (2017)
DOI:10.1038/nbt.3820
https://nextflow.io/


nextflow
amazon web services
Google Cloud
slurm workload manager
Microsoft Azure



Metabarcoding: from Lab to Bioinformatics

Metabarcoding: from Lab to Bioinformatics (UT International Summer University, 2022)

Metabarcoding: from Lab to Bioinformatics

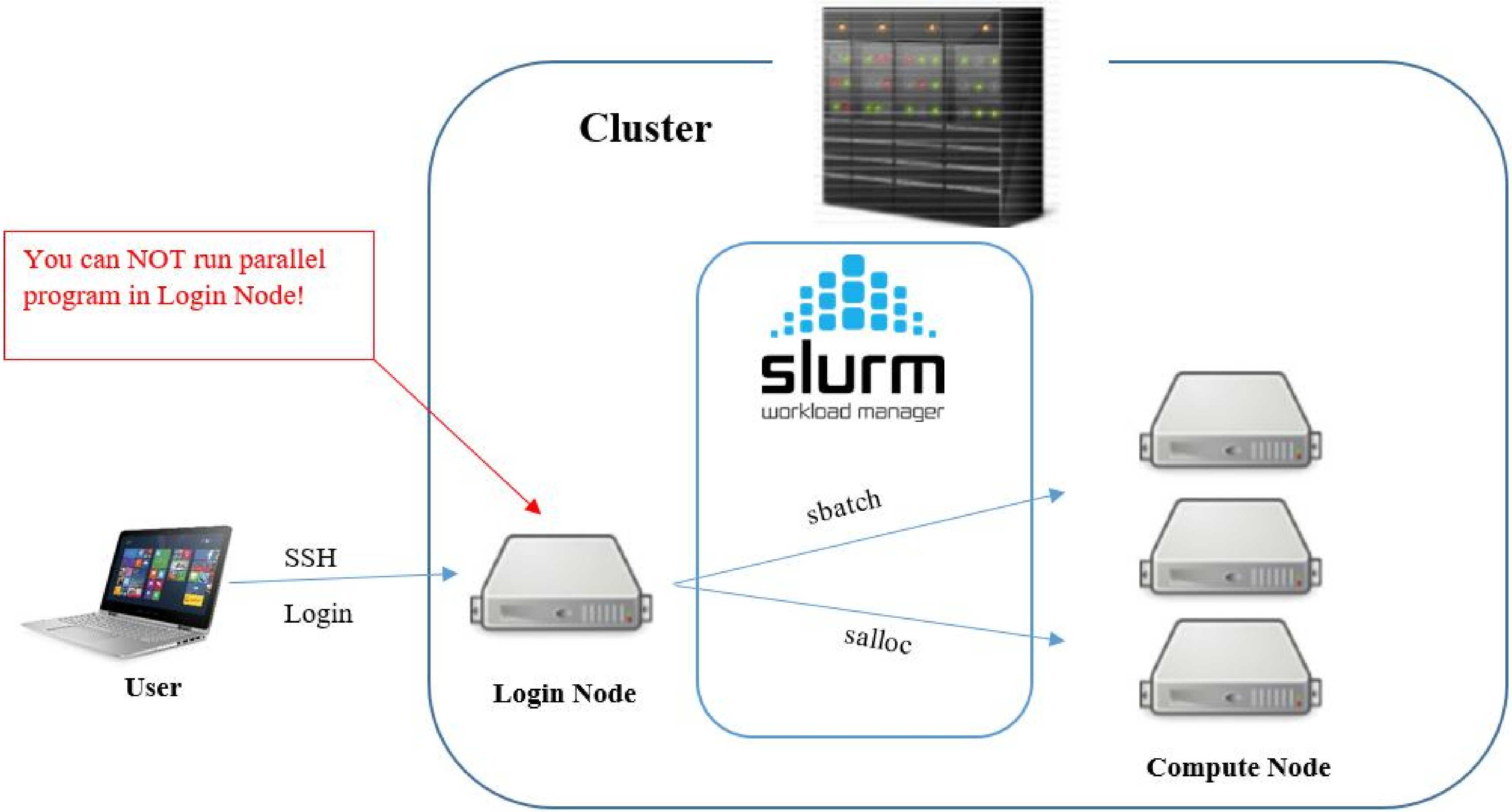
University of Tartu, 2022

Data used during the course
“Expert mode” commands
Individual projects
HPC basics

Slides (will be released after the course)

PipeCraft2 manual
About the course
Course announcement

High performance computing (HPC)



Cluster

You can NOT run parallel program in Login Node!

SSH

Login

User

Login Node

sbatch

salloc

Compute Node

slurm
workload manager

SLURM = Simple Linux Utility for Resource Management

https://pdc-support.github.io/hpc-intro/09-scheduling/

Working environment on HPC cluster

- Software installed by system administrator

```
module load blast-plus/2.12.0
```

- User-installed software

```
conda install -c bioconda blast=2.13.0
```

- Containerized software

```
singularity pull docker://ncbi/blast
singularity exec blast_latest.sif blastn
```

Scheduling a task on a cluster

```
#!/bin/bash
#SBATCH --job-name=my_job
#SBATCH --cpus-per-task=4
#SBATCH --nodes=1
#SBATCH --mem=10G
#SBATCH --partition amd
#SBATCH --time=48:00:00

my_program \
  -i input.data \
  -o output_1.data \
  --threads 4
```

Scheduling a task on a cluster

```
sbatch my_job.sh

sbatch
  --job-name=my_job
  --ntasks-per-node=4
  --nodes=1
  --mem=10G
  -p amd
  --time=48:00:00
  some_script.sh input.data
```

Job management

```
squeue -u $USER

scancel <JOBID>
scancel --name my_job
scancel -u $USER
```