HeaDS

HPC & pipes

Center for Health Data Science

Health Data Science Sandbox

Overview

1 Intro to HPCs

2 UCloud platform & **setup**

3 **Pipelines, workflows** & nf-core

HPC

What is high performance computing (HPC)?

Using a supercomputer or a cluster of computers to perform jobs too computationally intensive for a personal computer (laptop or desktop)

Compute-heavy tasks

RAM-heavy tasks

GPU-based tasks

Safely shared tasks

When would an HPC platform help you in biotech / health data science?

HPC

Tasks that benefit from HPC

Compute-heavy: many easily broken down tasks to be run – sequential or parallel

RAM-heavy: large datasets need to be processed in close proximity – i.e. genome alignment

GPU-based: large ML models (neural networks) and multi-dim models (biomechanics, weather)

Safely shared: you're working with sensitive data and/or many users need to access the same private datasets in parallel

Compute-heavy tasks

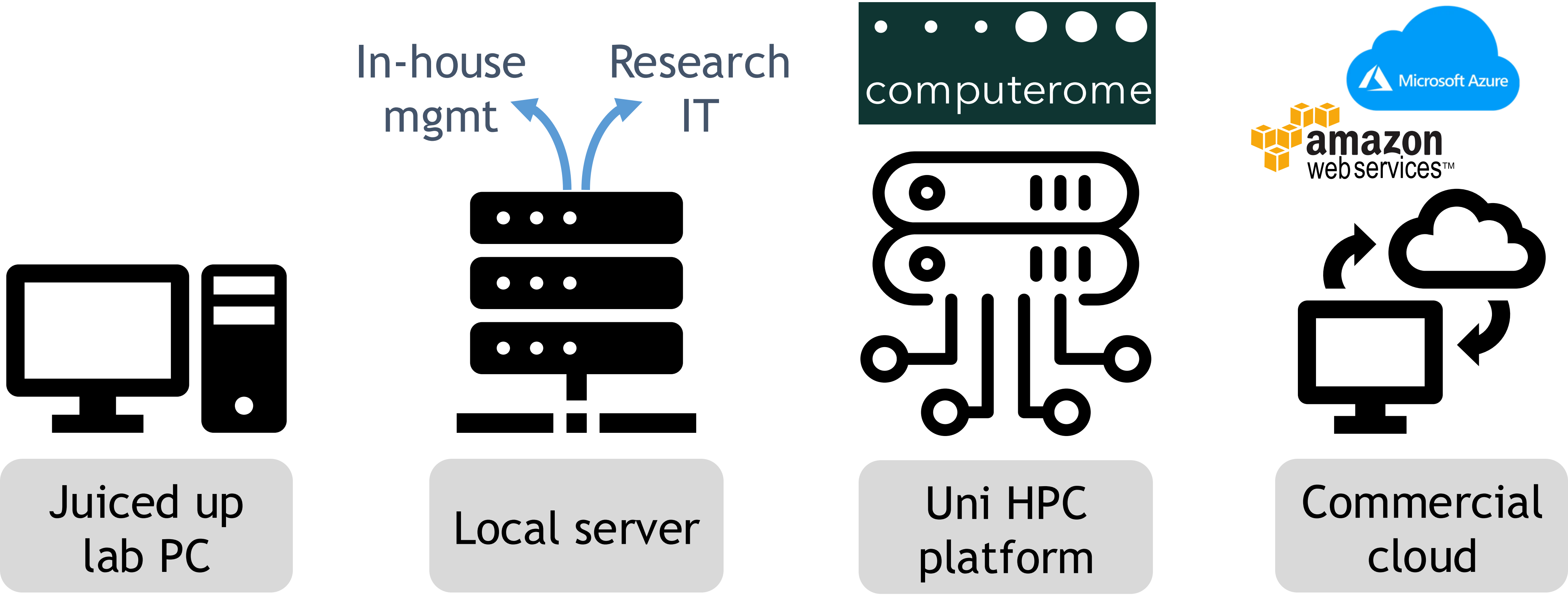
RAM-heavy tasks

GPU-based tasks

Safely shared tasks

HPC

An HPC poll – computing experience?



In-house mgmt

Research IT

Juiced up lab PC

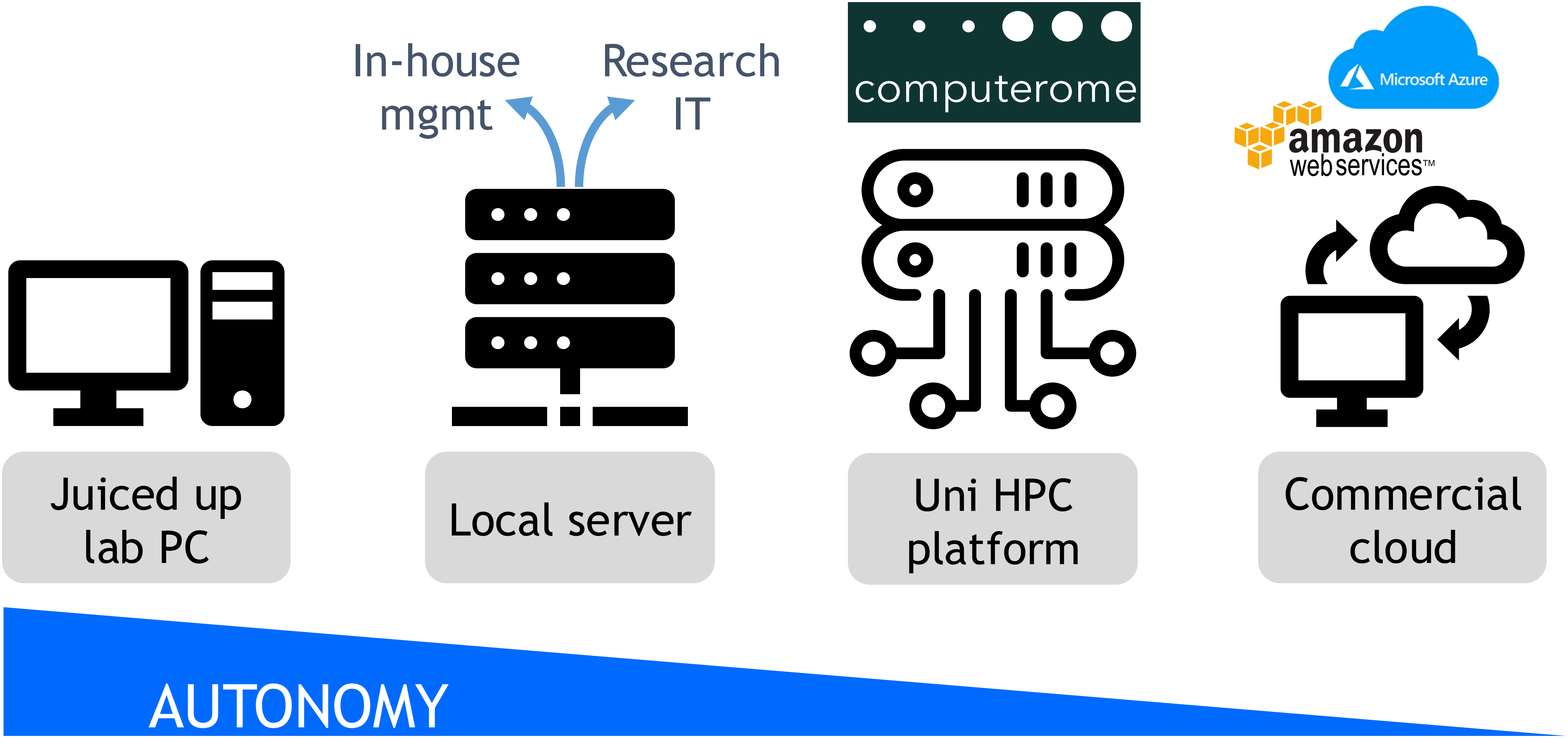
Local server

Uni HPC platform

Commercial cloud

HPC

Important considerations



In-house mgmt

Research IT

computerome

Microsoft Azure

amazon web services™

Juiced up lab PC

Local server

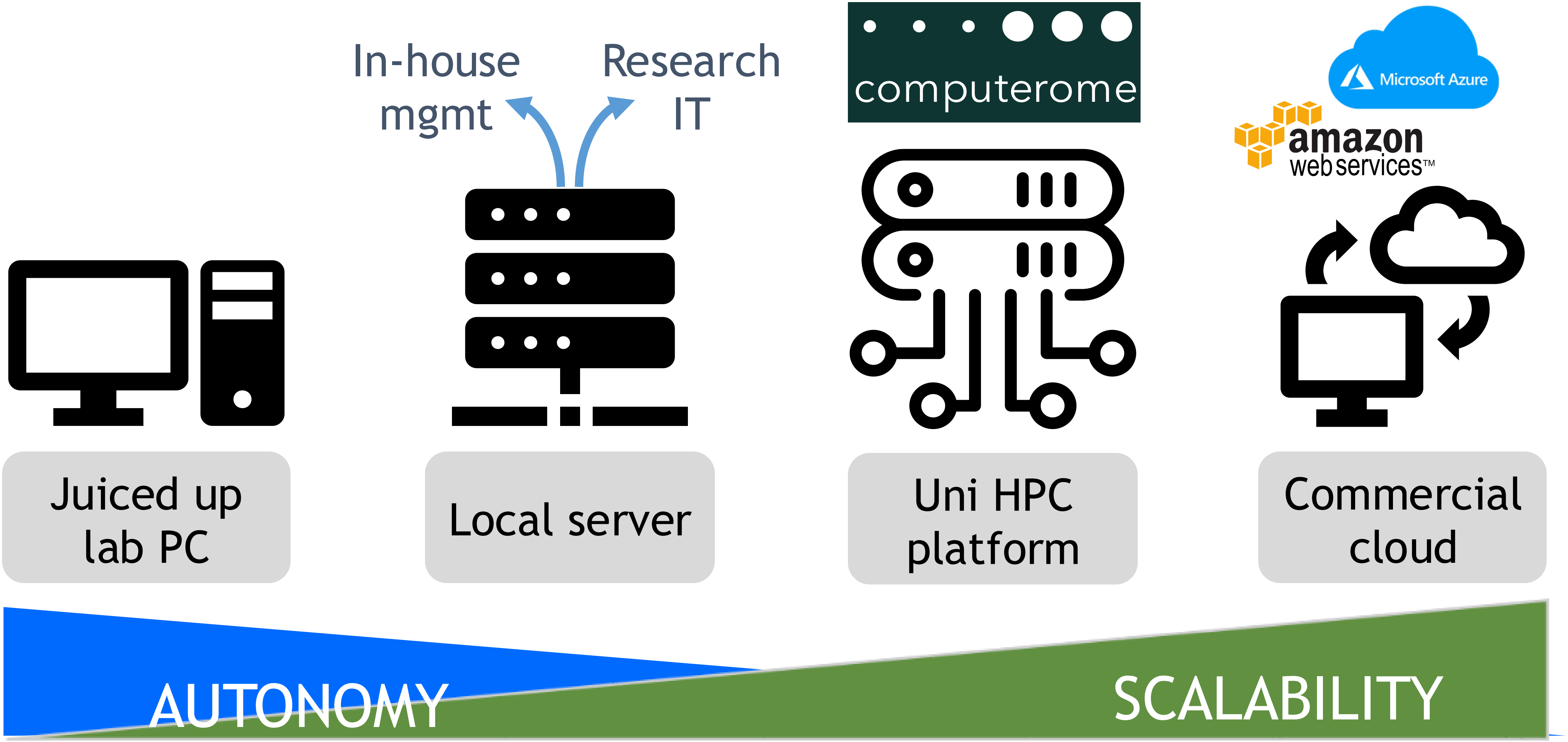
Uni HPC platform

Commercial cloud

AUTONOMY

HPC

Important considerations



In-house mgmt

Research IT

Juiced up lab PC

Local server

Uni HPC platform

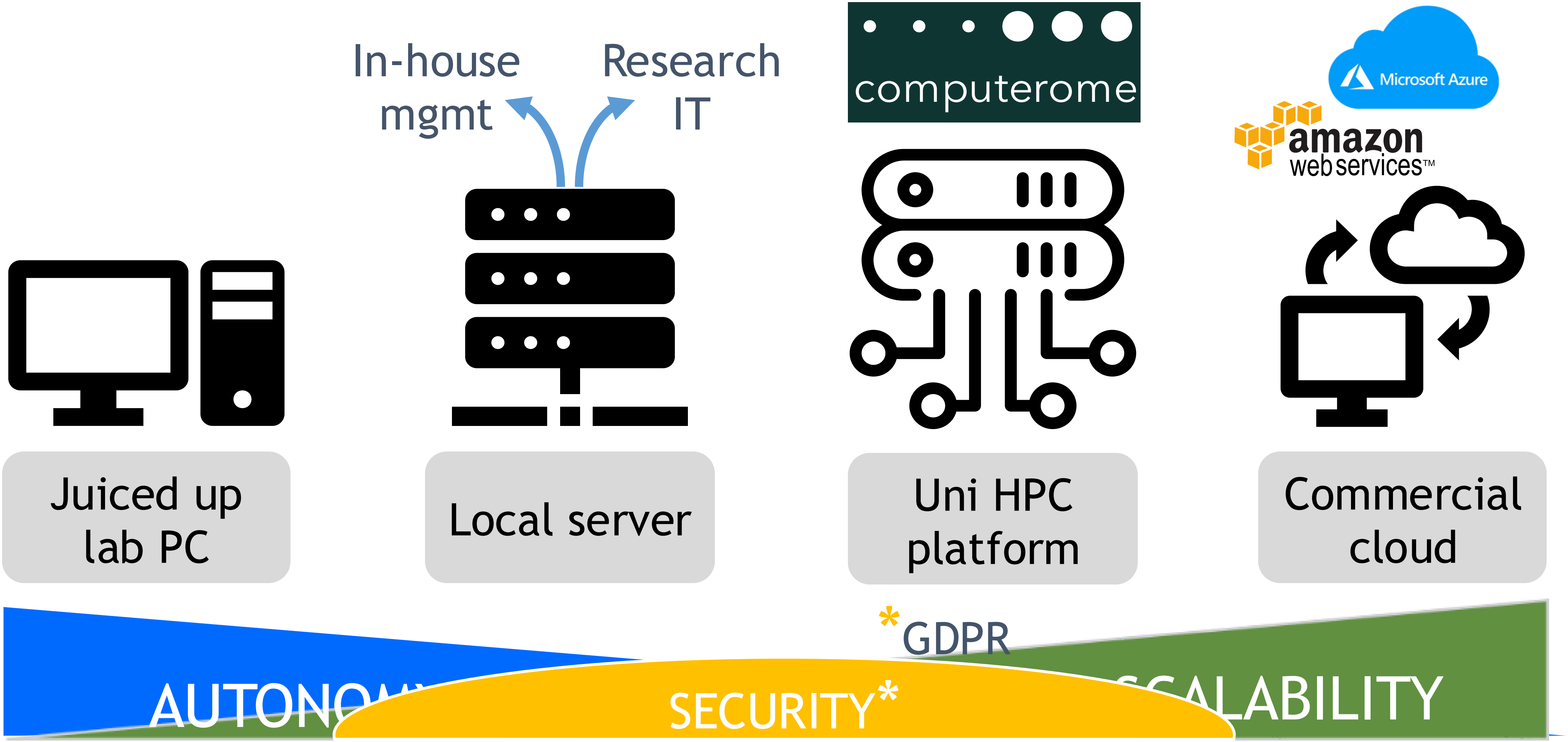
Commercial cloud

AUTONOMY

SCALABILITY

HPC

Important considerations



In-house mgmt

Research IT

computerome

Microsoft Azure

amazon web services™

Juiced up lab PC

Local server

Uni HPC platform

Commercial cloud

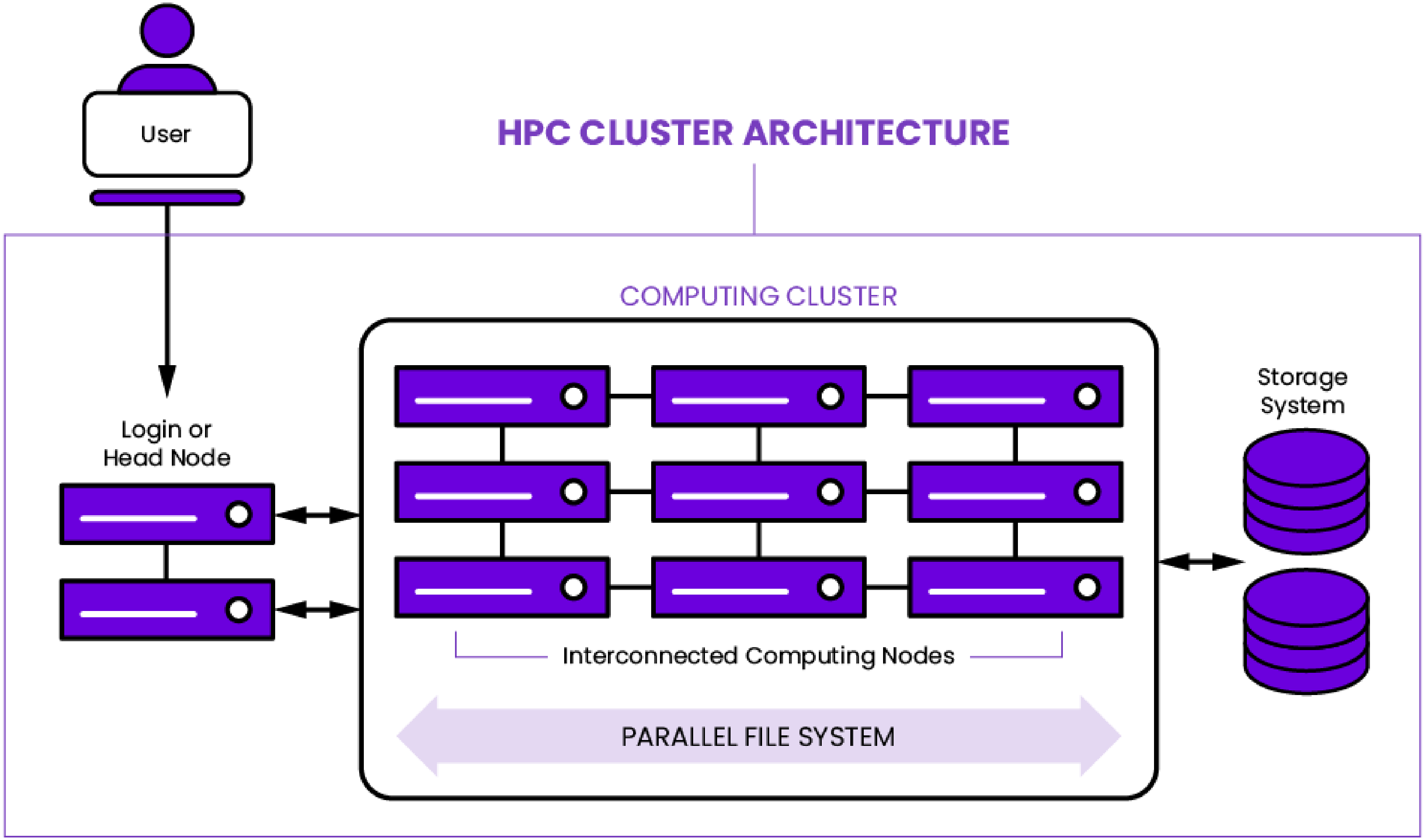
*GDPR

AUTONOMY

SECURITY*

SCALABILITY

Components of an HPC

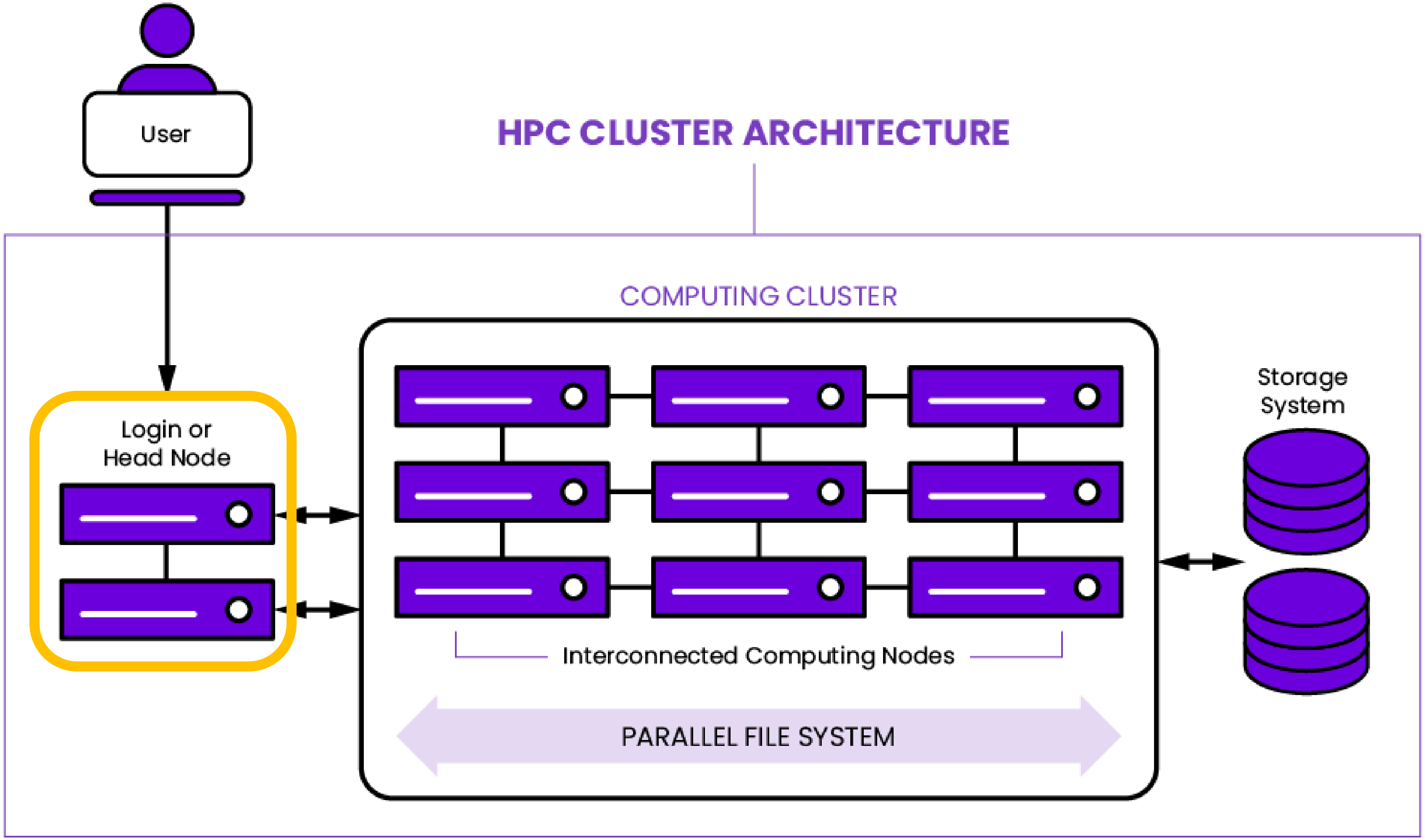


https://www.weka.io/learn/hpc/what-are-hpc-and-hpc-clusters/

Nodes

- ➢ Any physical device that can send, receive, or pass information
- ➢ In HPC, term usually references a compute node or a login node
- ➢ Nodes are interconnected to facilitate communication and data transfer between them

Components of an HPC

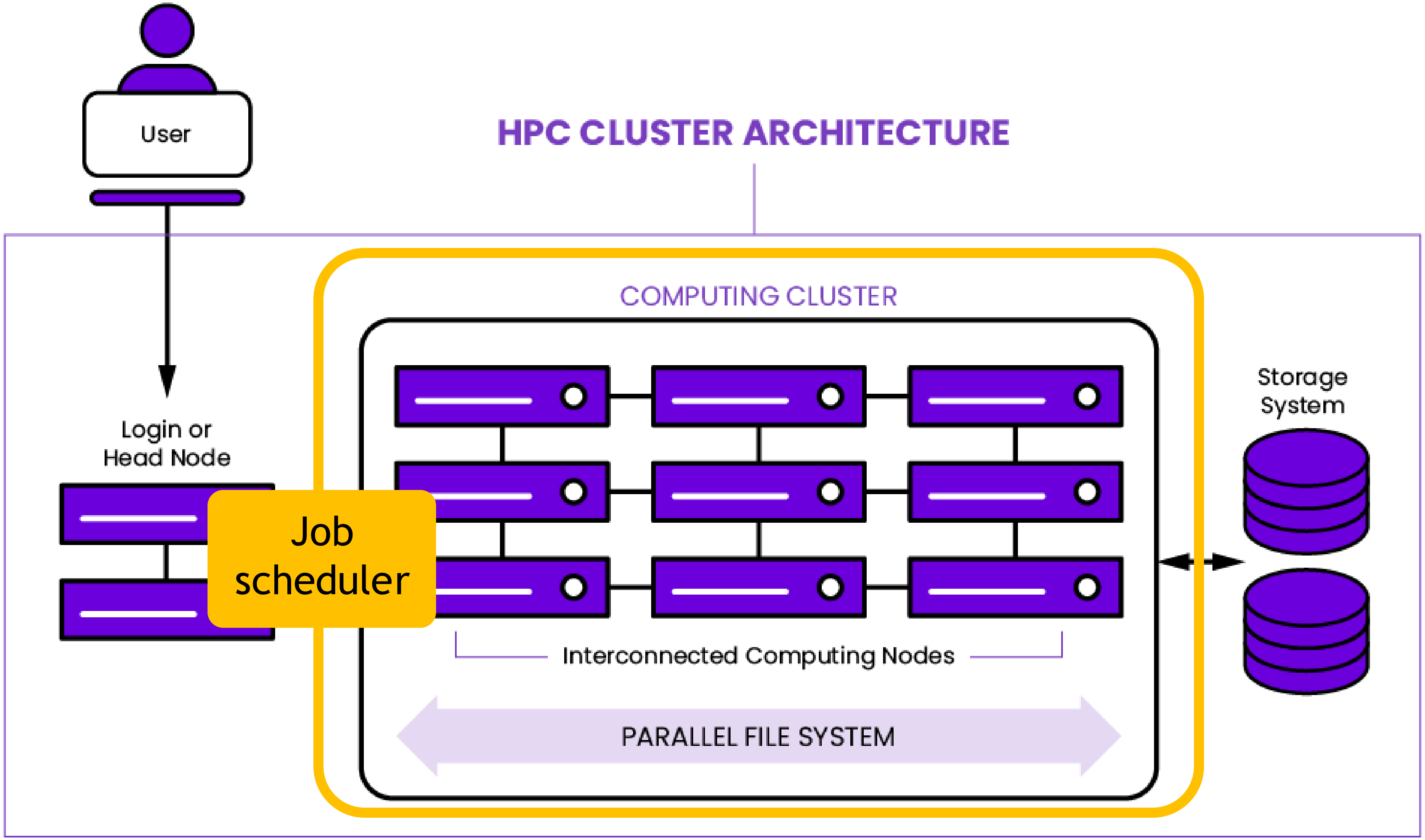


https://www.weka.io/learn/hpc/what-are-hpc-and-hpc-clusters/

Login Node

- A computer that acts as the front end to the HPC system, where users access (request) cluster resources and submit tasks for the computing nodes to perform

Components of an HPC

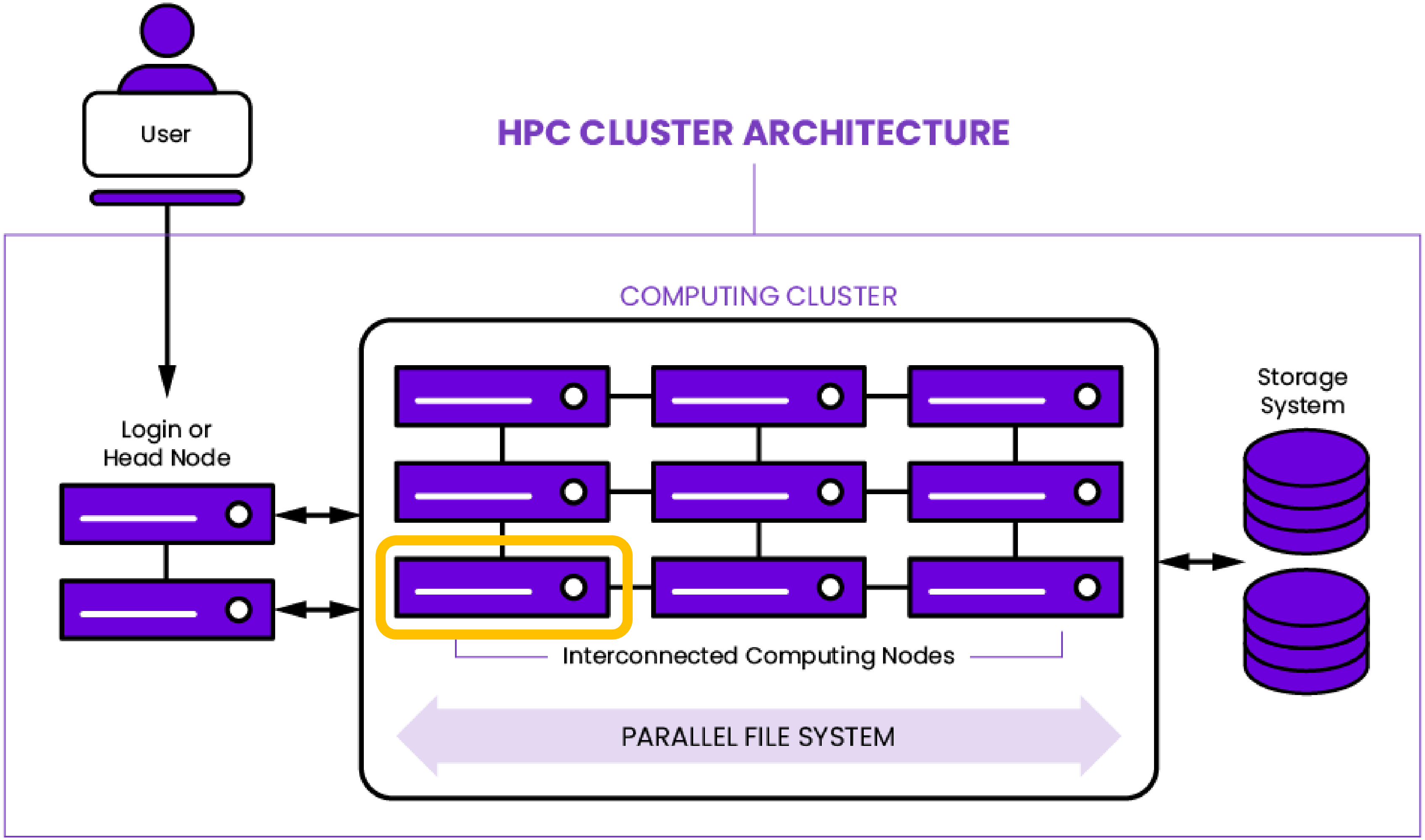


https://www.weka.io/learn/hpc/what-are-hpc-and-hpc-clusters/

Computing Cluster

- A (large) group of closely interconnected computers that work together as a single system to complete jobs
- Job scheduler manages and schedules jobs (tasks) across compute nodes allocating resources to complete the job

Components of an HPC

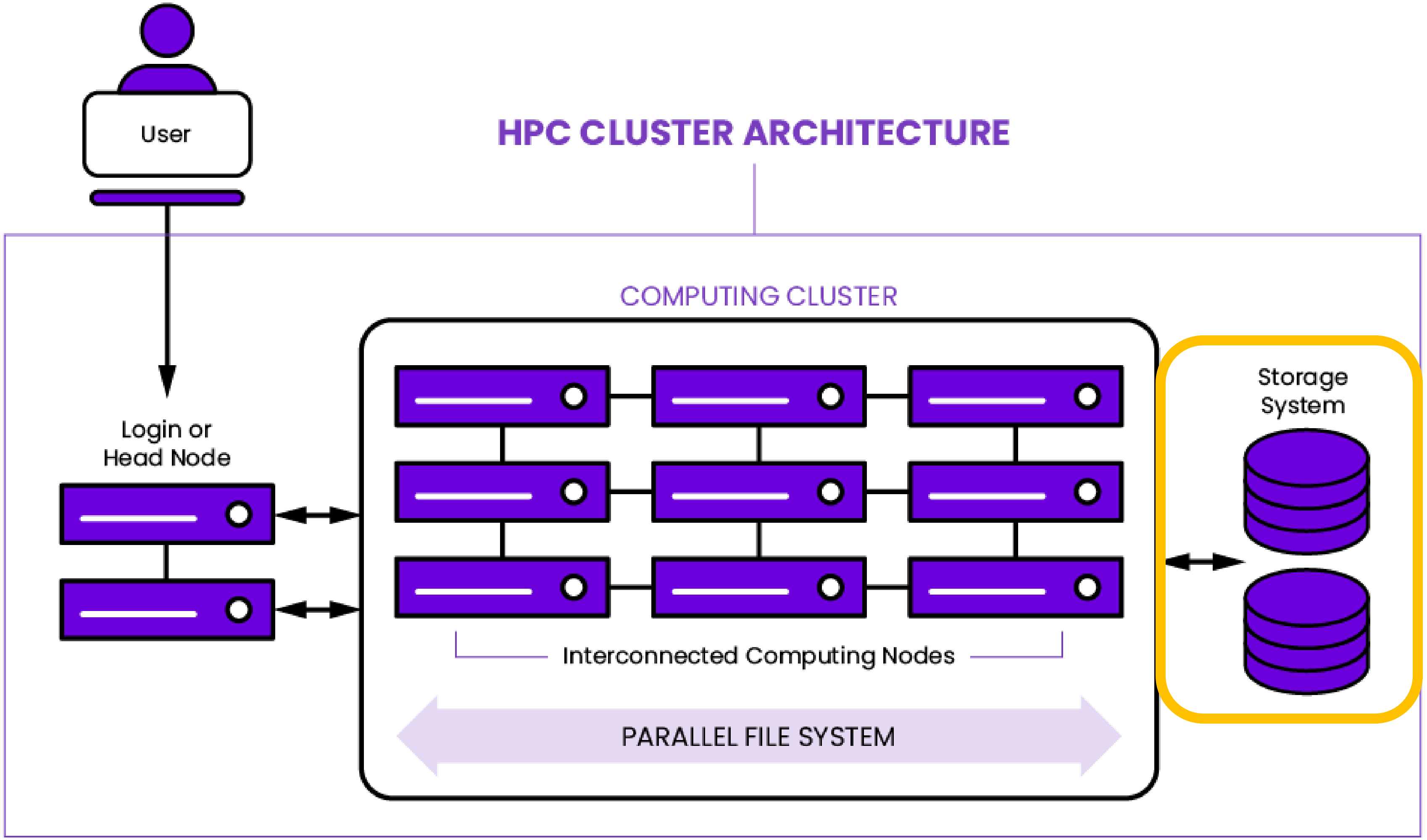


Computing Node

- an individual computer within the compute cluster made up of a set of processors and their local memory
- 'Size' of the node traditionally varies by number of processors and amount of memory

HeaDS

Components of an HPC

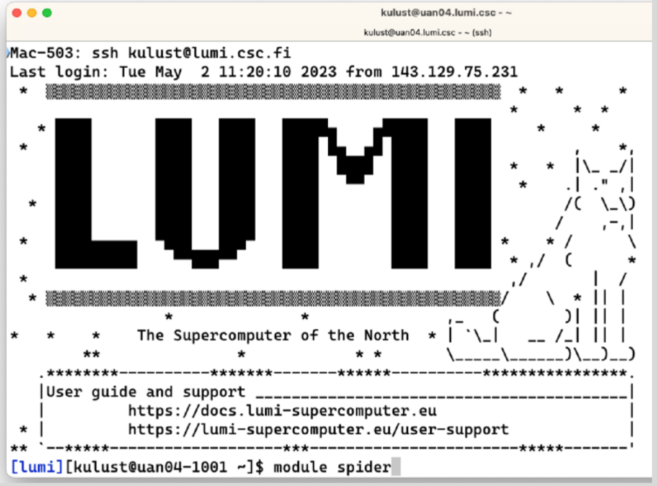


https://www.weka.io/learn/hpc/what-are-hpc-and-hpc-clusters/

Storage system

➢ Provides persistent storage for data and programs used by the HPC system

HPC interfaces you might encounter



```
Mac-503: ssh kulust@lumi.csc.fi
Last login: Tue May  2 11:20:10 2023 from 143.129.75.231

 *   *   *     The Supercomputer of the North  *
      **              *              * *
   .********----------*******-------******----------****************.
   |User guide and support _________________________________________|
   |          https://docs.lumi-supercomputer.eu                    |
 * |          https://lumi-supercomputer.eu/user-support            |
** `--*****--------------------***---------------------------*****-------'
[lumi][kulust@uan04-1001 ~]$ module spider
```

Terminal / command line

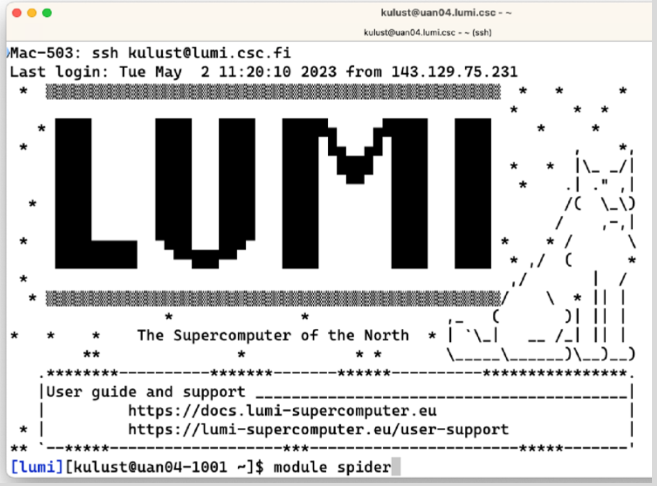
VMware Horizon Client

Virtual Machine (usually linux OS)

UCloud

Custom Graphical User Interface

HPC interfaces you might encounter



```
Mac-503: ssh kulust@lumi.csc.fi
Last login: Tue May  2 11:20:10 2023 from 143.129.75.231

                     The Supercomputer of the North

   .********----------*******-------******----------****************.
   |User guide and support ________________________________________|
   |         https://docs.lumi-supercomputer.eu                     |
 * |         https://lumi-supercomputer.eu/user-support             |
** `--*****------------------***-------------------------*****-------'
[lumi][kulust@uan04-1001 ~]$ module spider
```

Terminal / command line

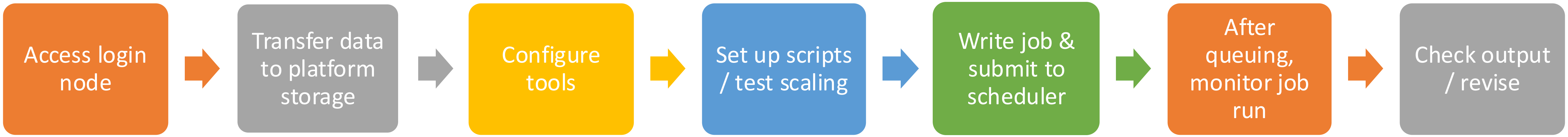
VMware Horizon Client

Virtual Machine (usually linux OS)

UCloud

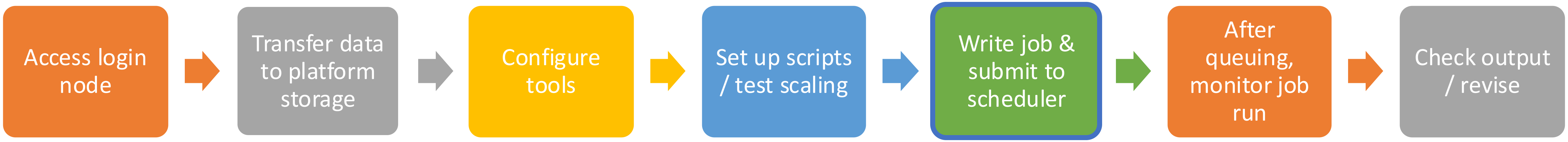
Custom Graphical User Interface

A standard HPC workflow



Access login node → Transfer data to platform storage → Configure tools → Set up scripts / test scaling → Write job & submit to scheduler → After queuing, monitor job run → Check output / revise

A standard HPC workflow



Access login node

Transfer data to platform storage

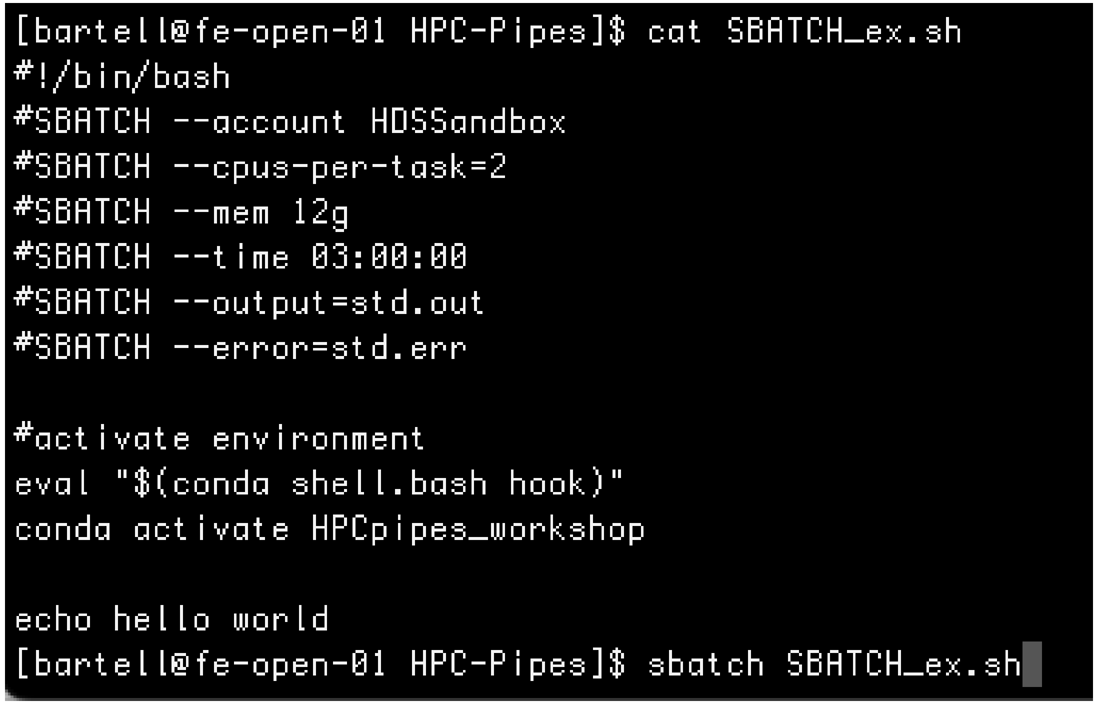
Configure tools

Set up scripts / test scaling

Write job & submit to scheduler

After queuing, monitor job run

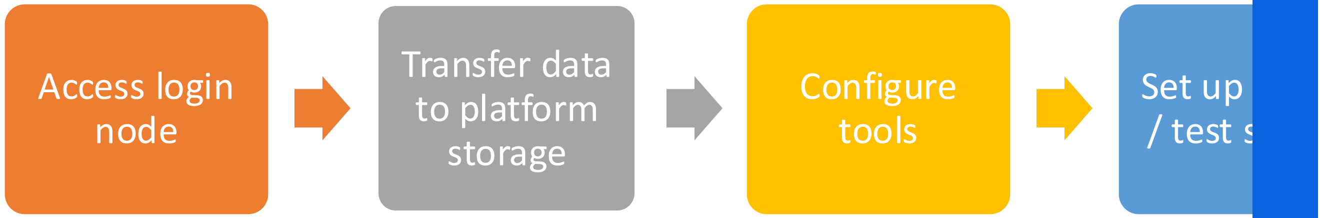
Check output / revise



```
[bartell@fe-open-01 HPC-Pipes]$ cat SBATCH_ex.sh
#!/bin/bash
#SBATCH --account HDSSandbox
#SBATCH --cpus-per-task=2
#SBATCH --mem 12g
#SBATCH --time 03:00:00
#SBATCH --output=std.out
#SBATCH --error=std.err

#activate environment
eval "$(conda shell.bash hook)"
conda activate HPCpipes_workshop

echo hello world
[bartell@fe-open-01 HPC-Pipes]$ sbatch SBATCH_ex.sh
```

HeaDS

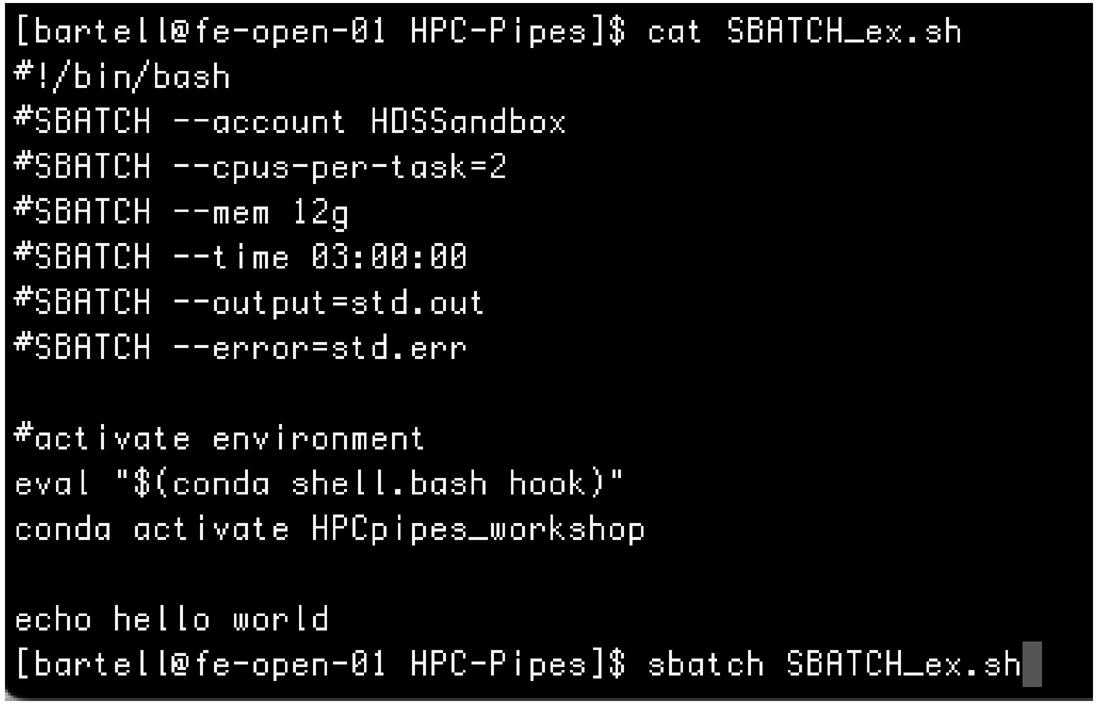
A standard HPC workfl


Access login node

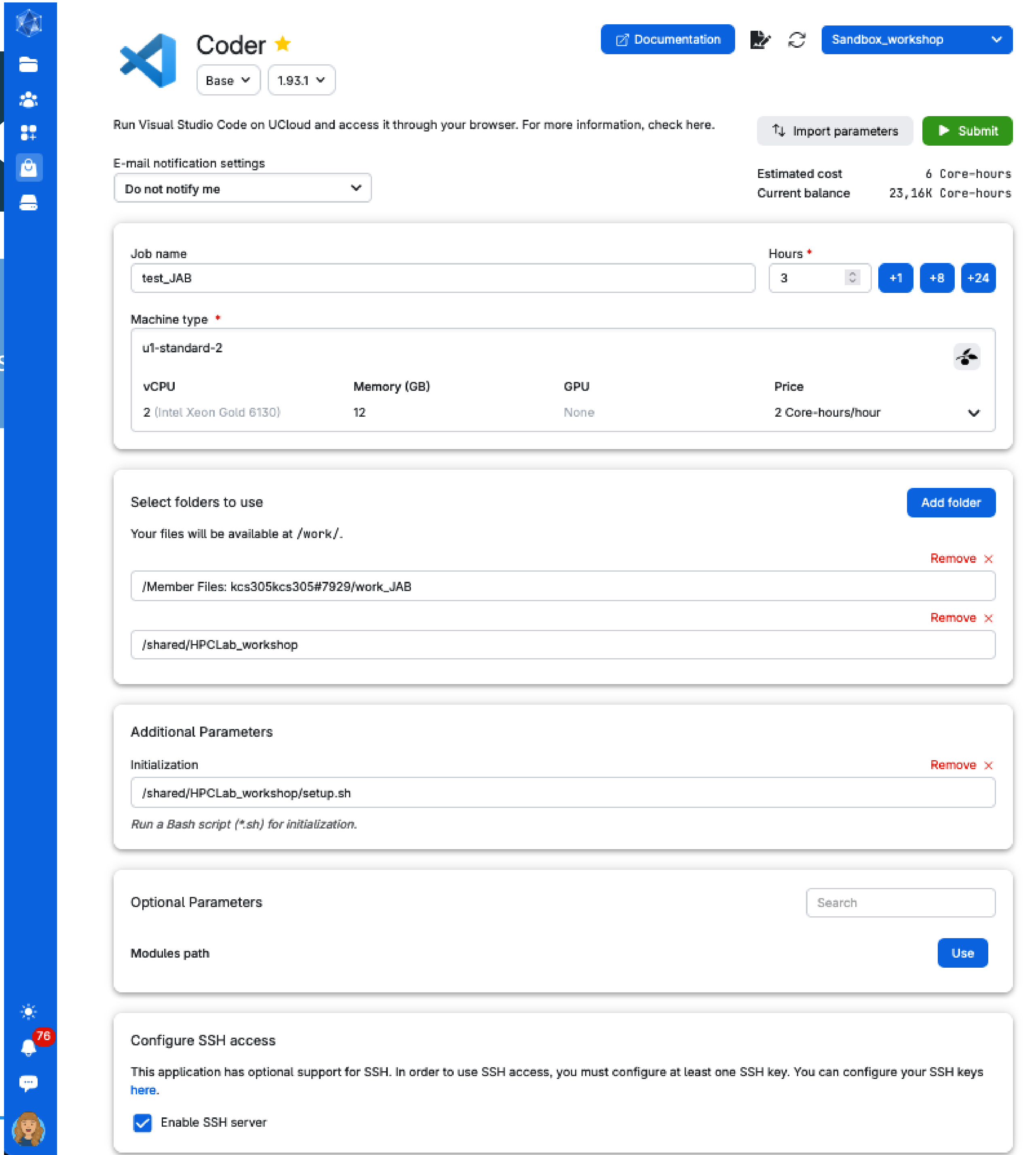
Transfer data to platform storage

Configure tools

Set up / test s



```
[bartell@fe-open-01 HPC-Pipes]$ cat SBATCH_ex.sh
#!/bin/bash
#SBATCH --account HDSSandbox
#SBATCH --cpus-per-task=2
#SBATCH --mem 12g
#SBATCH --time 03:00:00
#SBATCH --output=std.out
#SBATCH --error=std.err

#activate environment
eval "$(conda shell.bash hook)"
conda activate HPCpipes_workshop

echo hello world
[bartell@fe-open-01 HPC-Pipes]$ sbatch SBATCH_ex.sh
```



Coder ★

Base 1.93.1

Run Visual Studio Code on UCloud and access it through your browser. For more information, check here.

E-mail notification settings: Do not notify me

Estimated cost 6 Core-hours
Current balance 23,16K Core-hours

Job name: test_JAB

Hours *: 3

Machine type *: u1-standard-2

vCPU: 2 (Intel Xeon Gold 6130) | Memory (GB): 12 | GPU: None | Price: 2 Core-hours/hour

Select folders to use

Your files will be available at /work/.

/Member Files: kcs305kcs305#7929/work_JAB — Remove ×

/shared/HPCLab_workshop — Remove ×

Additional Parameters

Initialization: /shared/HPCLab_workshop/setup.sh — Remove ×

Run a Bash script (.sh) for initialization.*

Optional Parameters

Search

Modules path

Configure SSH access

This application has optional support for SSH. In order to use SSH access, you must configure at least one SSH key. You can configure your SSH keys here.

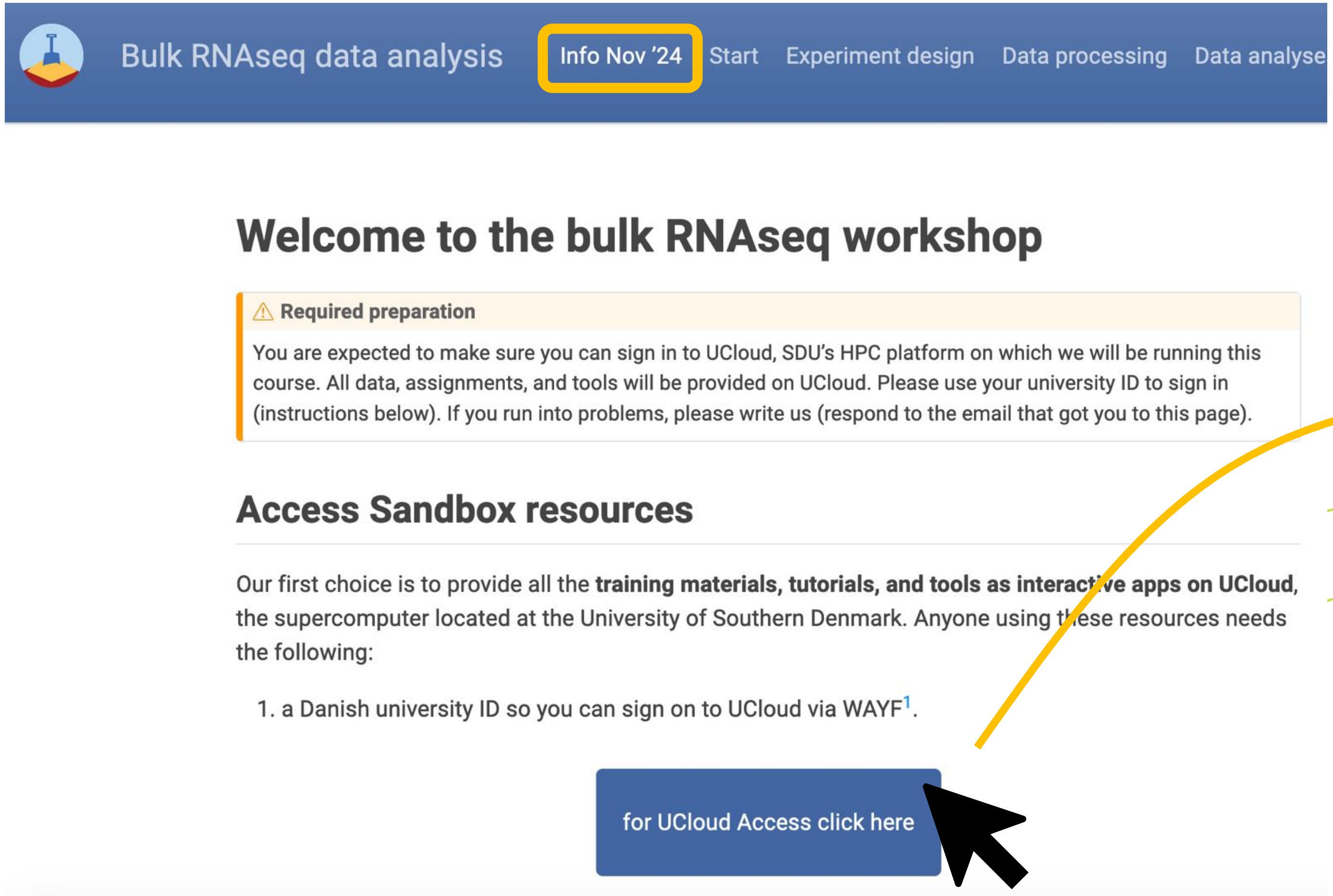
☑ Enable SSH server

UCloud

- UCloud is a danish High Performance Computing environment
  - Lots of storage, lots of cpus and RAM (computing power)
- Danish institutions have access to it
  - You personally have 1000dkk in computing resources
- UCloud works in apps, giving you access to different programs
  - All apps have documentation on how to use them!
- This means everyone is using the same versions of software
  - Makes teaching much much easier as results are reproducible

UCloud access

https://hds-sandbox.github.io/bulk_RNAseq_course/develop/



Welcome to the bulk RNAseq workshop

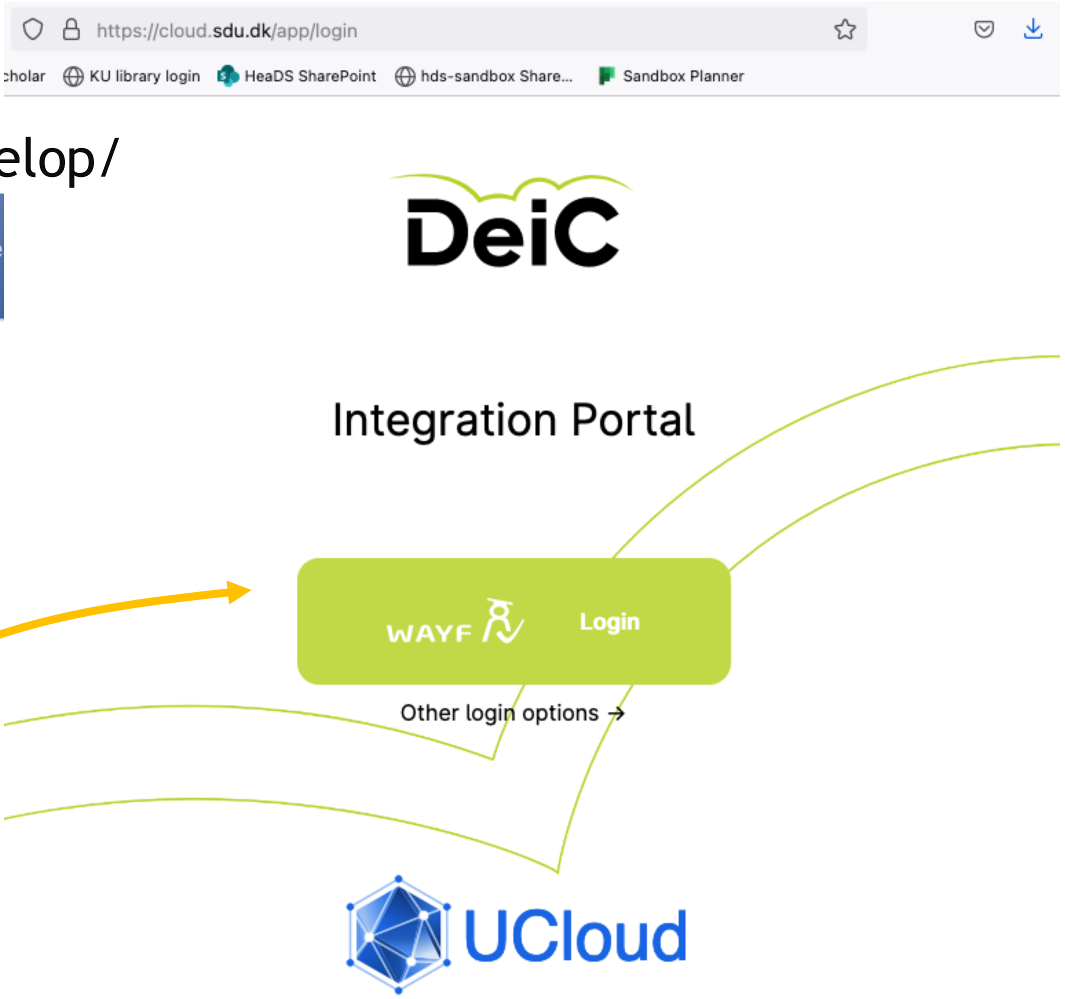
⚠ **Required preparation**

You are expected to make sure you can sign in to UCloud, SDU's HPC platform on which we will be running this course. All data, assignments, and tools will be provided on UCloud. Please use your university ID to sign in (instructions below). If you run into problems, please write us (respond to the email that got you to this page).

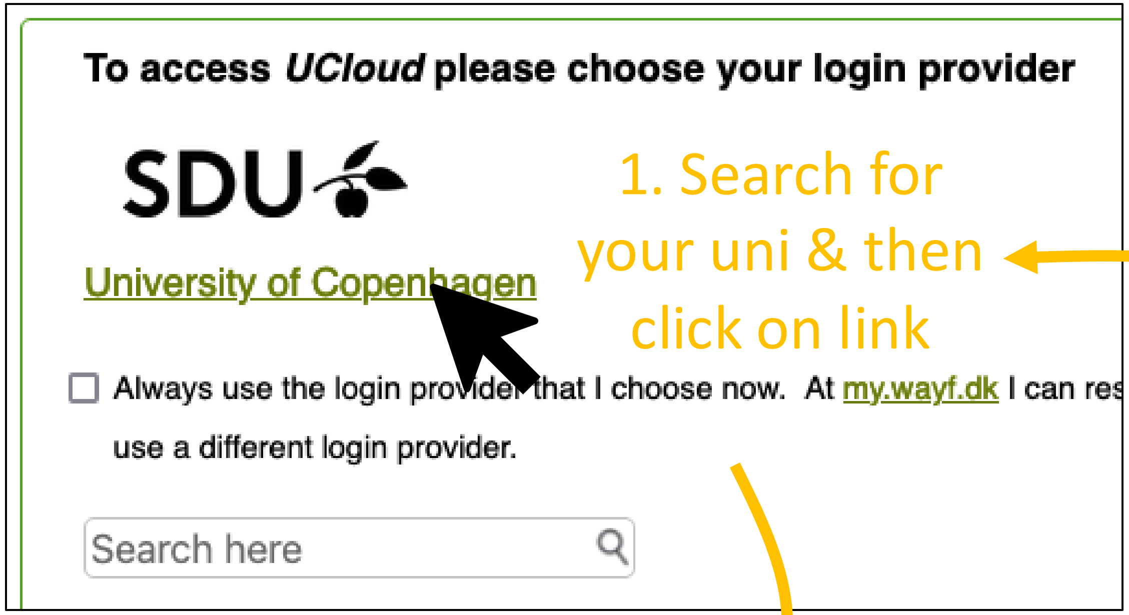
Access Sandbox resources

Our first choice is to provide all the **training materials, tutorials, and tools as interactive apps on UCloud**, the supercomputer located at the University of Southern Denmark. Anyone using these resources needs the following:

1. a Danish university ID so you can sign on to UCloud via WAYF[1].



UCloud log-in



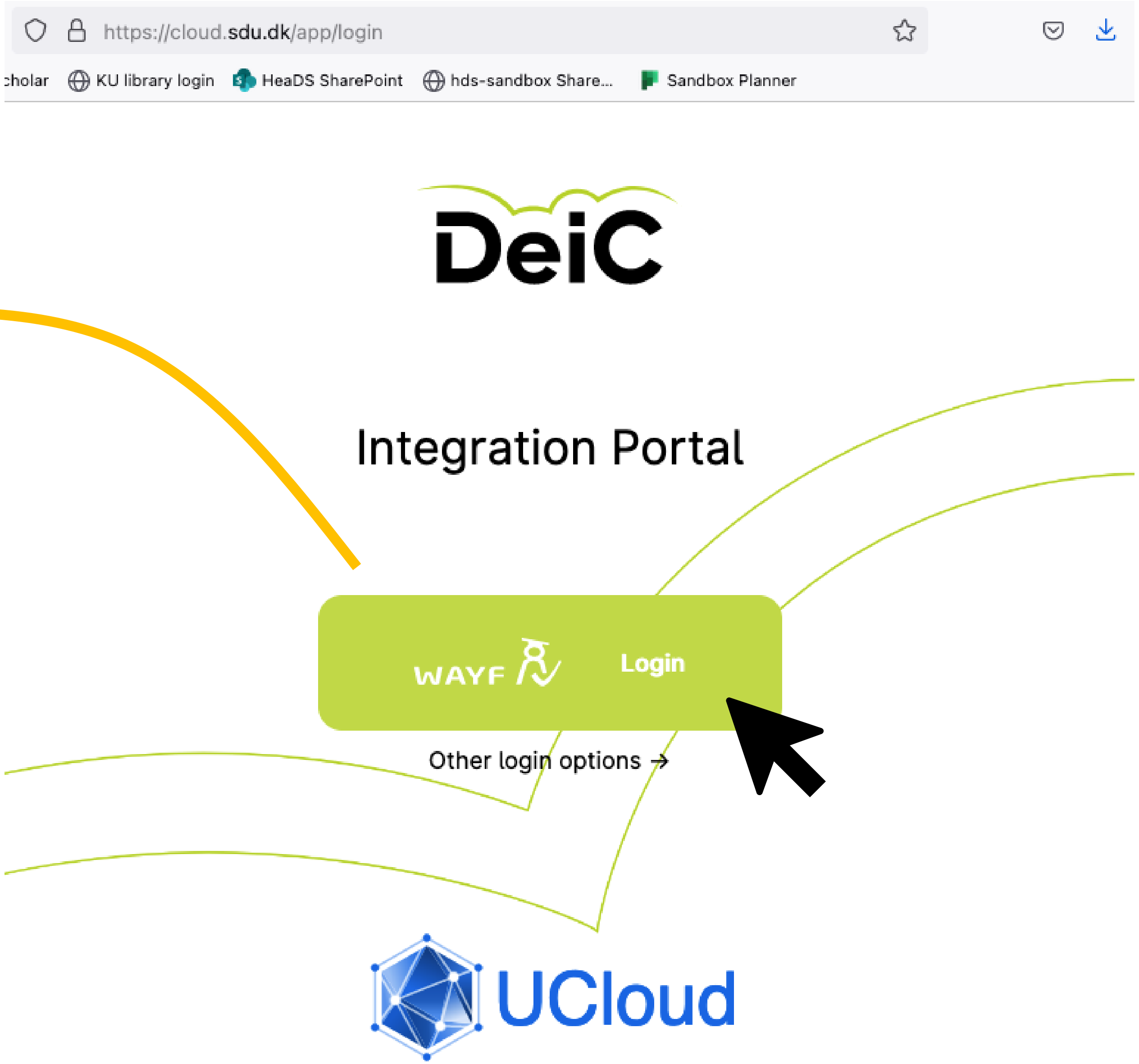
To access *UCloud* please choose your login provider

University of Copenhagen

□ Always use the login provider that I choose now. At my.wayf.dk I can res
use a different login provider.

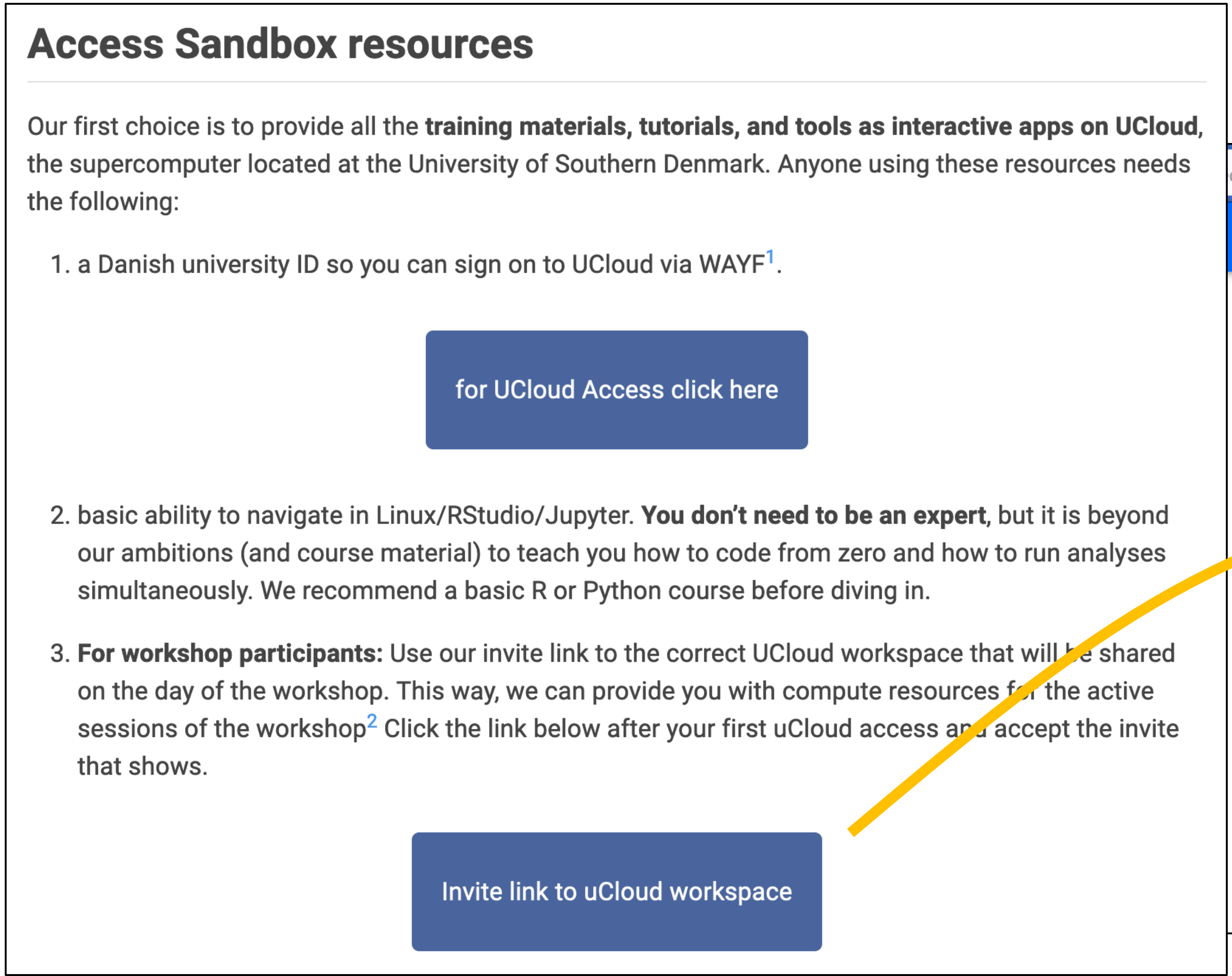
1. Search for your uni & then click on link

2. Sign-in via your uni portal



UCloud log-in

Back to the Info Nov '24 page…



Access Sandbox resources

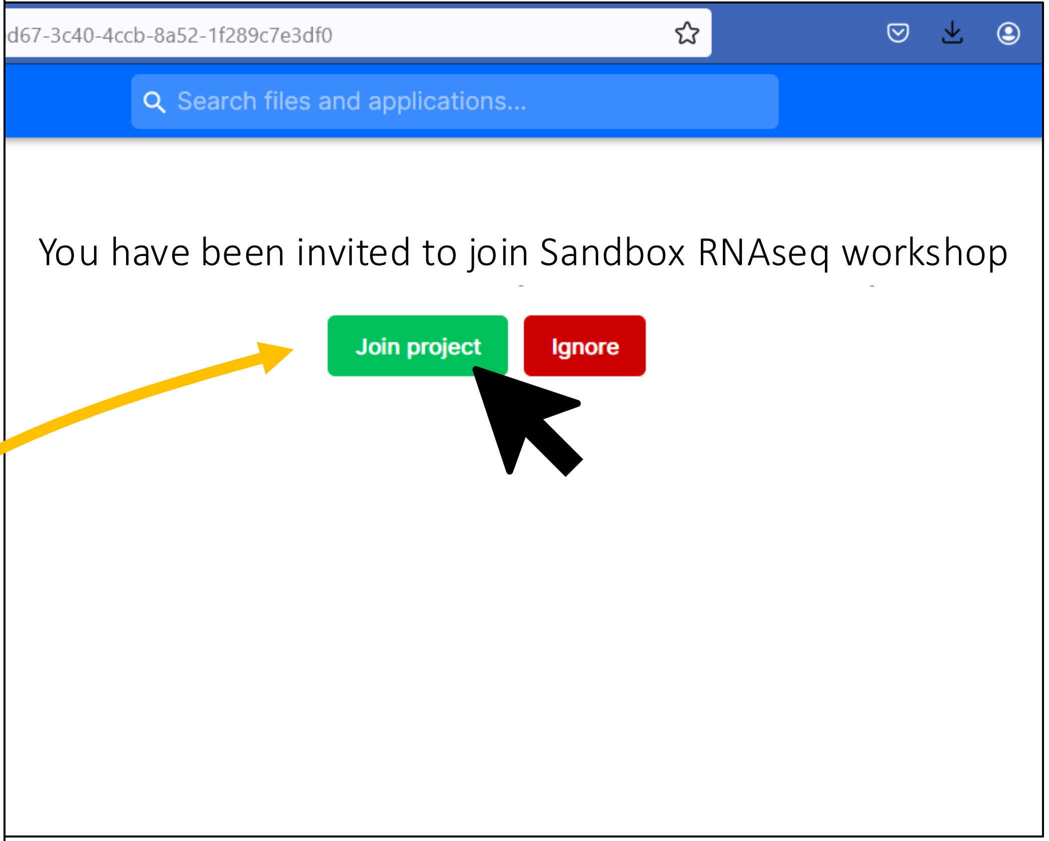
Our first choice is to provide all the **training materials, tutorials, and tools as interactive apps on UCloud**, the supercomputer located at the University of Southern Denmark. Anyone using these resources needs the following:

1. a Danish university ID so you can sign on to UCloud via WAYF[1].

for UCloud Access click here

2. basic ability to navigate in Linux/RStudio/Jupyter. **You don't need to be an expert**, but it is beyond our ambitions (and course material) to teach you how to code from zero and how to run analyses simultaneously. We recommend a basic R or Python course before diving in.
3. **For workshop participants:** Use our invite link to the correct UCloud workspace that will be shared on the day of the workshop. This way, we can provide you with compute resources for the active sessions of the workshop[2] Click the link below after your first uCloud access and accept the invite that shows.

Invite link to uCloud workspace

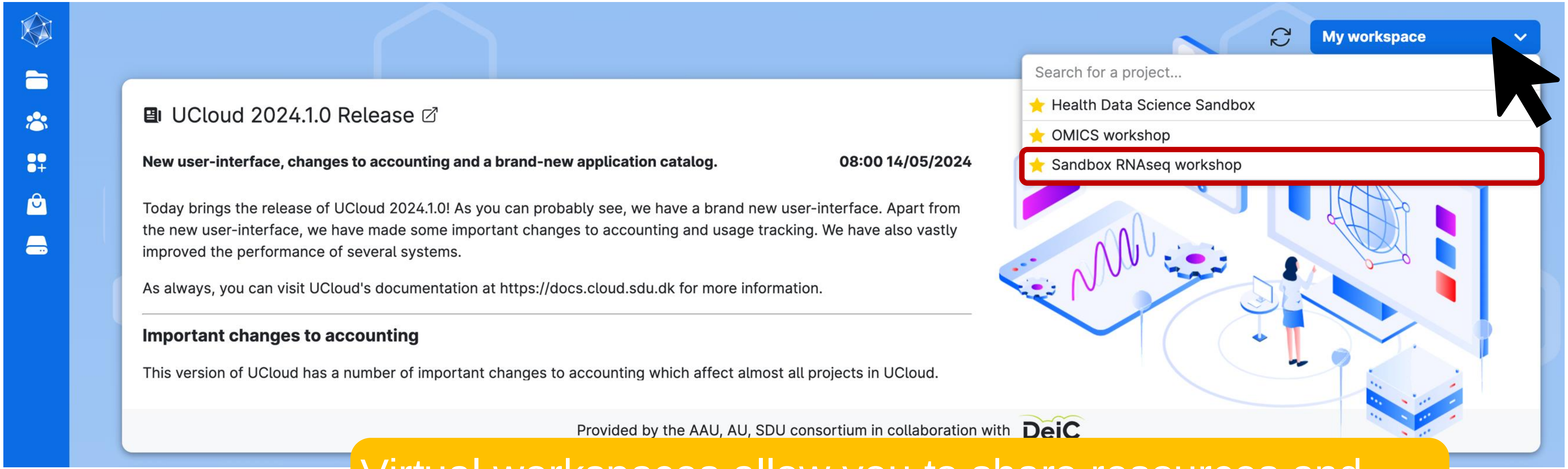


d67-3c40-4ccb-8a52-1f289c7e3df0

You have been invited to join Sandbox RNAseq workshop

10 min break!

Workspaces



UCloud 2024.1.0 Release

New user-interface, changes to accounting and a brand-new application catalog. **08:00 14/05/2024**

Today brings the release of UCloud 2024.1.0! As you can probably see, we have a brand new user-interface. Apart from the new user-interface, we have made some important changes to accounting and usage tracking. We have also vastly improved the performance of several systems.

As always, you can visit UCloud's documentation at https://docs.cloud.sdu.dk for more information.

Important changes to accounting

This version of UCloud has a number of important changes to accounting which affect almost all projects in UCloud.

Provided by the AAU, AU, SDU consortium in collaboration with DeiC

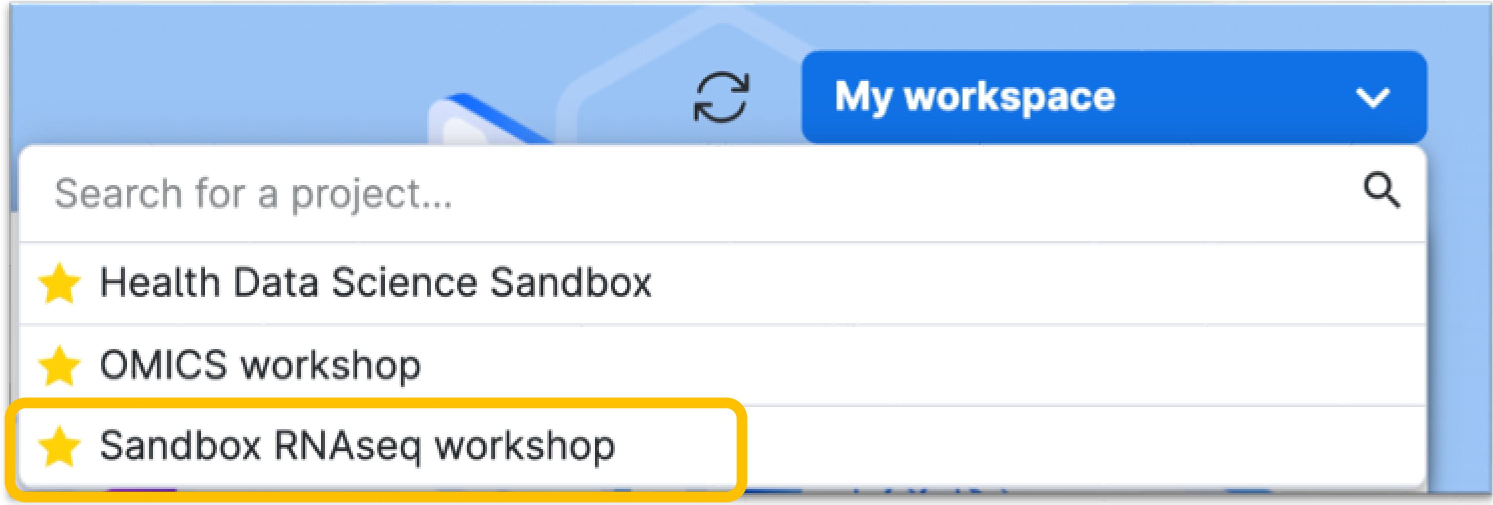
Virtual workspaces allow you to share resources and work together with project collaborators

HeaDS

UCloud usage

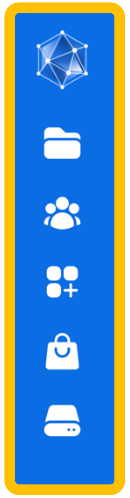
During this course, utilize **Sandbox RNAseq workshop workspace** (resources have been requested for this purpose)

Following the workshop, switch to “My workspace



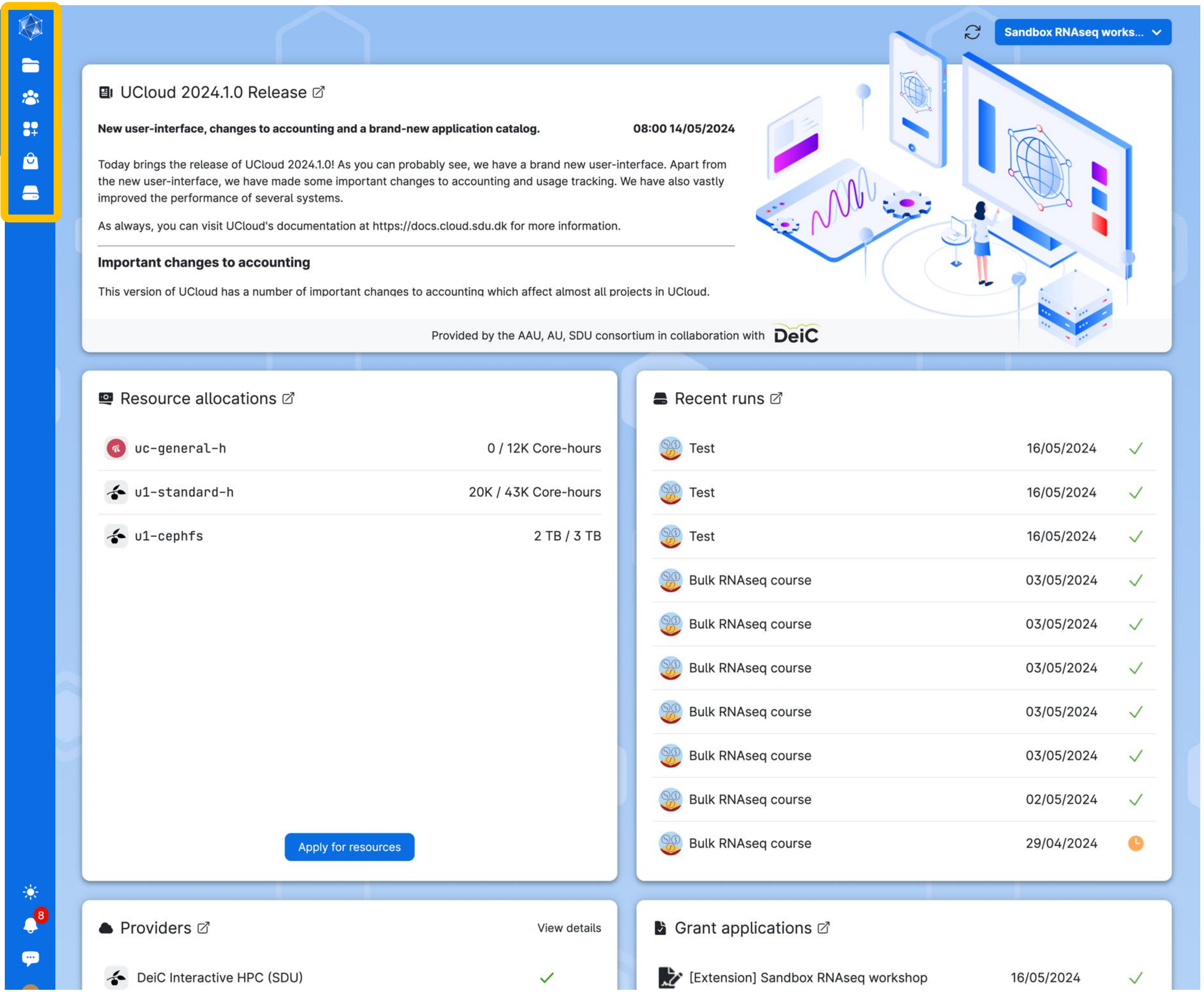
Dashboard

Interactive HPC with UCloud



Left-side menu:

- Drives/Files
- Projects
- Resources
- Applications
- Runs



Sandbox RNAseq works...

UCloud 2024.1.0 Release

New user-interface, changes to accounting and a brand-new application catalog. **08:00 14/05/2024**

Today brings the release of UCloud 2024.1.0! As you can probably see, we have a brand new user-interface. Apart from the new user-interface, we have made some important changes to accounting and usage tracking. We have also vastly improved the performance of several systems.

As always, you can visit UCloud's documentation at https://docs.cloud.sdu.dk for more information.

Important changes to accounting

This version of UCloud has a number of important changes to accounting which affect almost all projects in UCloud.

Provided by the AAU, AU, SDU consortium in collaboration with DeiC

Resource allocations

uc-general-h	0 / 12K Core-hours
u1-standard-h	20K / 43K Core-hours
u1-cephfs	2 TB / 3 TB

Apply for resources

Recent runs

Test	16/05/2024	✓
Test	16/05/2024	✓
Test	16/05/2024	✓
Bulk RNAseq course	03/05/2024	✓
Bulk RNAseq course	03/05/2024	✓
Bulk RNAseq course	03/05/2024	✓
Bulk RNAseq course	03/05/2024	✓
Bulk RNAseq course	03/05/2024	✓
Bulk RNAseq course	02/05/2024	✓
Bulk RNAseq course	29/04/2024	●

Providers View details

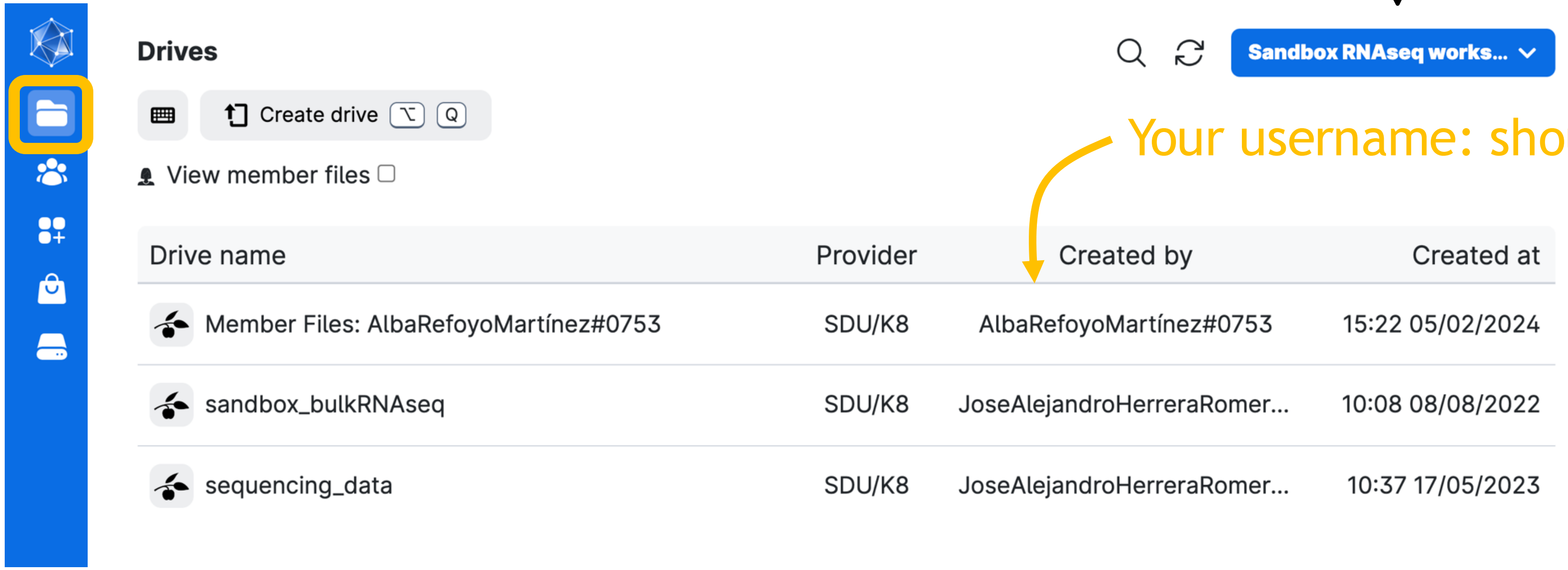
DeiC Interactive HPC (SDU)	✓

Grant applications

[Extension] Sandbox RNAseq workshop	16/05/2024	✓

Drives

Project folders, files, etc. that only belong to the **active workspace** will be accessible from the menu at the left



Drives

View member files ☐

Drive name	Provider	Created by	Created at
Member Files: AlbaRefoyoMartínez#0753	SDU/K8	AlbaRefoyoMartínez#0753	15:22 05/02/2024
sandbox_bulkRNAseq	SDU/K8	JoseAlejandroHerreraRomer...	10:08 08/08/2022
sequencing_data	SDU/K8	JoseAlejandroHerreraRomer...	10:37 17/05/2023

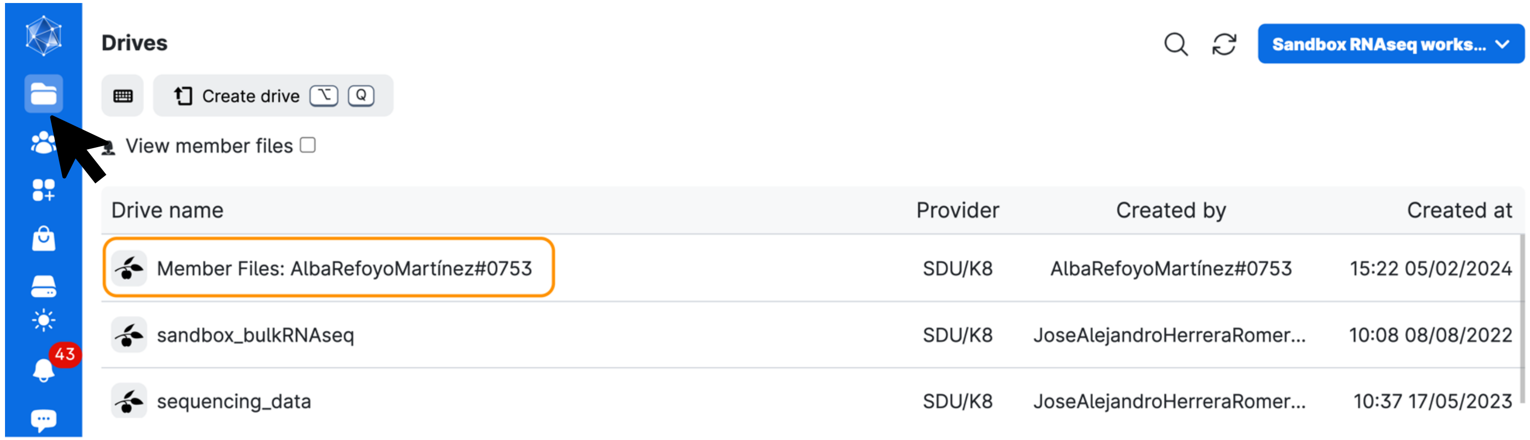
Your username: should be FirstLast#0000...

You have a personal drive

You have shared drives

Drives

Access file structure (shared and generated in previous jobs)

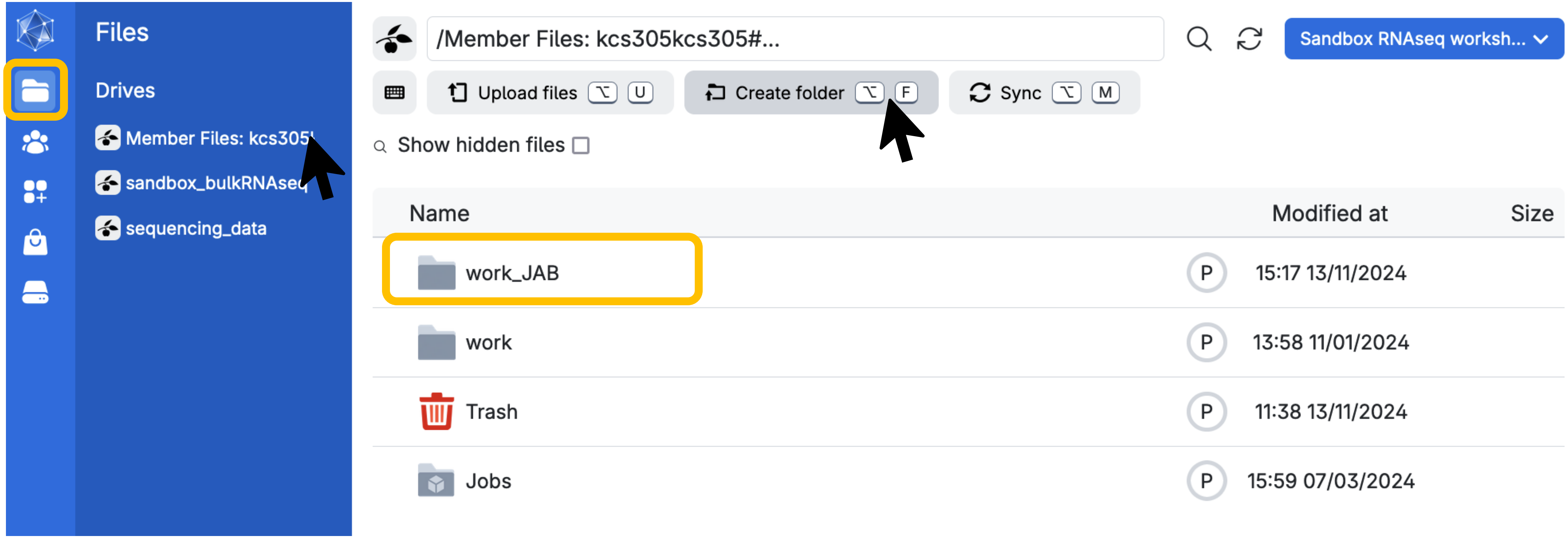


Drive name	Provider	Created by	Created at
Member Files: AlbaRefoyoMartínez#0753	SDU/K8	AlbaRefoyoMartínez#0753	15:22 05/02/2024
sandbox_bulkRNAseq	SDU/K8	JoseAlejandroHerreraRomer...	10:08 08/08/2022
sequencing_data	SDU/K8	JoseAlejandroHerreraRomer...	10:37 17/05/2023

- Personal workspace folder "Member Files:username”: results will go here
  - Jobs folder
    - Subfolders with Apps names
    - App name: All runs (the job’s name) results
- "sandbox_bulkRNAseq" : contains some course material for teachers
- "sequencing_data" : contains fastq files for preprocessing (nf-core RNAseq)

Drives

Make your own custom working directory… we'll use this later!

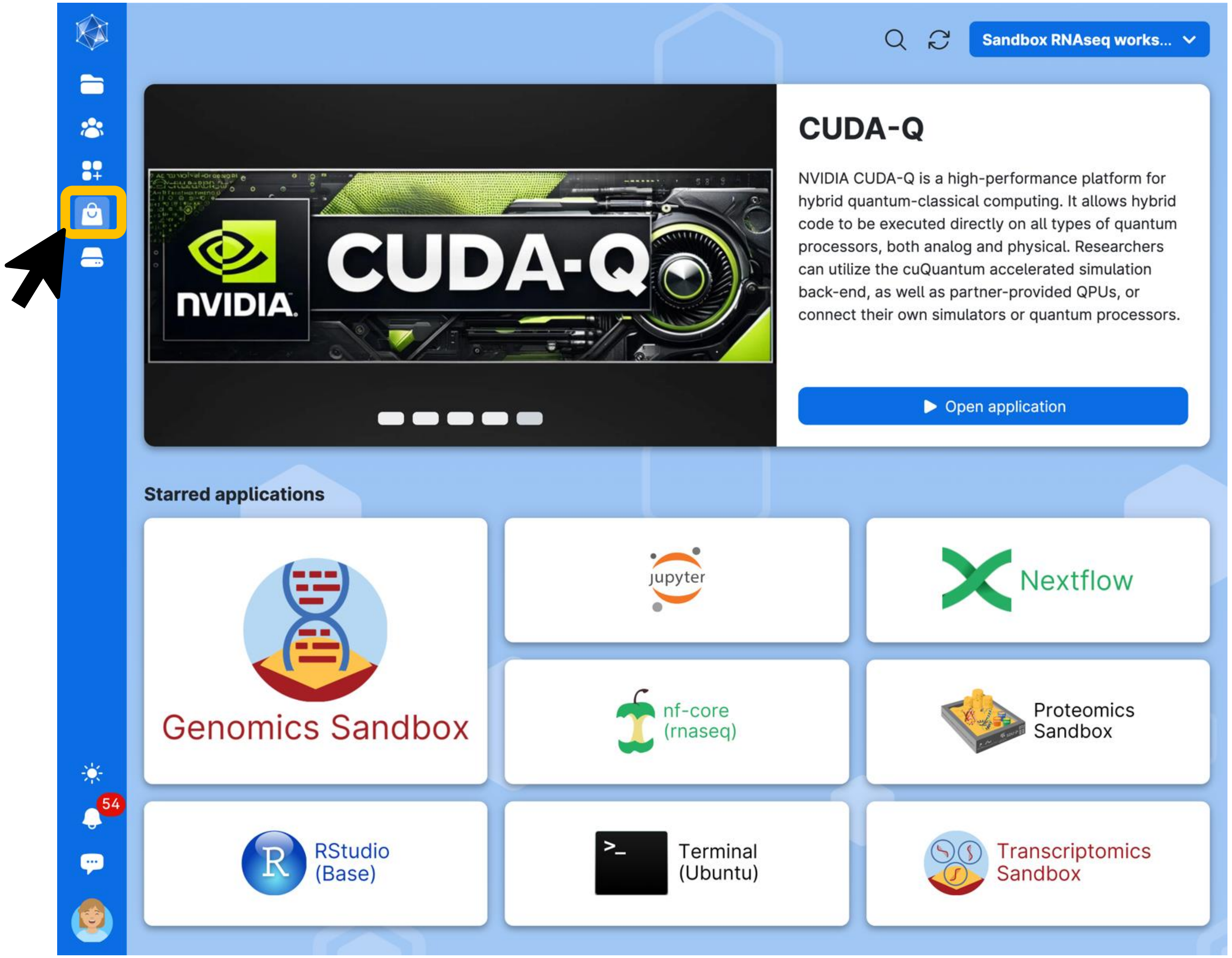


Name		Modified at	Size
work_JAB	P	15:17 13/11/2024	
work	P	13:58 11/01/2024	
Trash	P	11:38 13/11/2024	
Jobs	P	15:59 07/03/2024	

Applications

There is a wide variety of applications.

Here are some of my favorites!

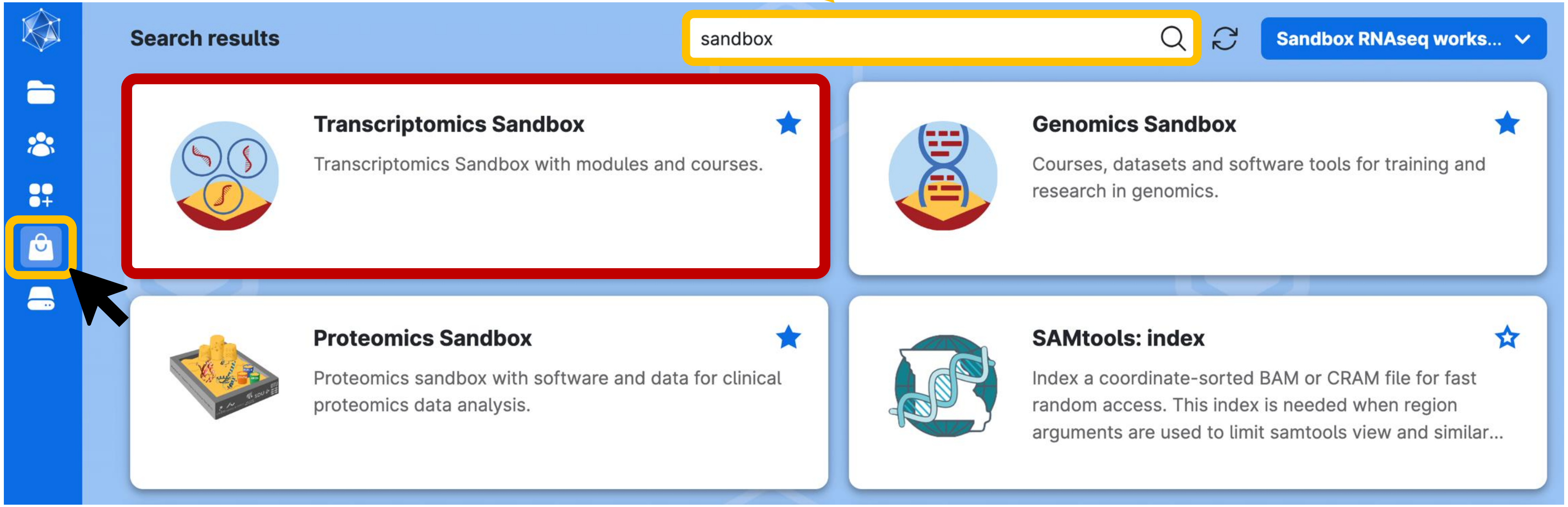


CUDA-Q

NVIDIA CUDA-Q is a high-performance platform for hybrid quantum-classical computing. It allows hybrid code to be executed directly on all types of quantum processors, both analog and physical. Researchers can utilize the cuQuantum accelerated simulation back-end, as well as partner-provided QPUs, or connect their own simulators or quantum processors.

Apps

Search for Sandbox apps



Search results

sandbox

Transcriptomics Sandbox

Transcriptomics Sandbox with modules and courses.

Genomics Sandbox

Courses, datasets and software tools for training and research in genomics.

Proteomics Sandbox

Proteomics sandbox with software and data for clinical proteomics data analysis.

SAMtools: index

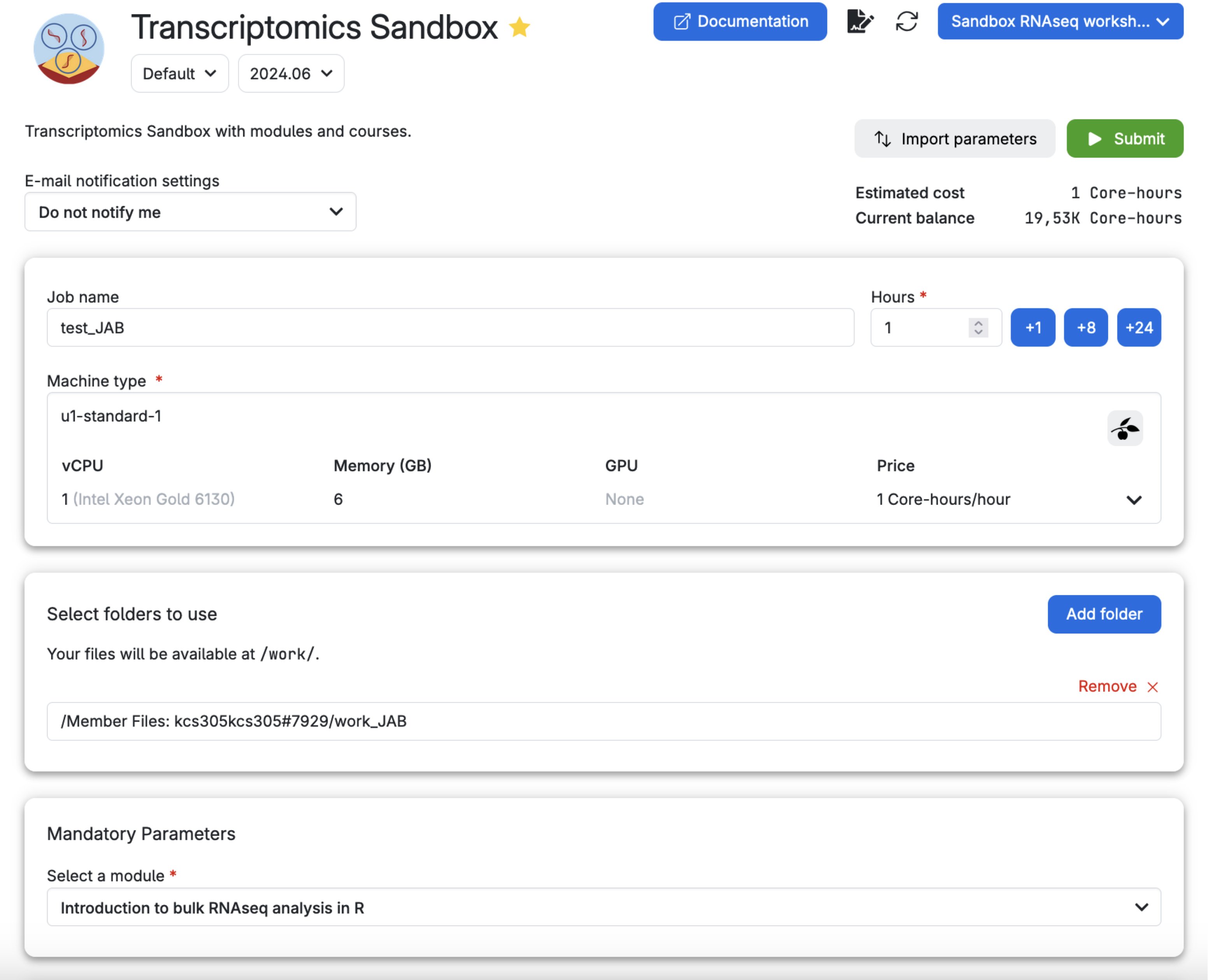
Index a coordinate-sorted BAM or CRAM file for fast random access. This index is needed when region arguments are used to limit samtools view and similar...

HeaDS

Jobs

How to submit a Sandbox app job?

Let's set up the app together!



Transcriptomics Sandbox ★

Default | 2024.06

Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost	1 Core-hours
Current balance	19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Your files will be available at /work/.

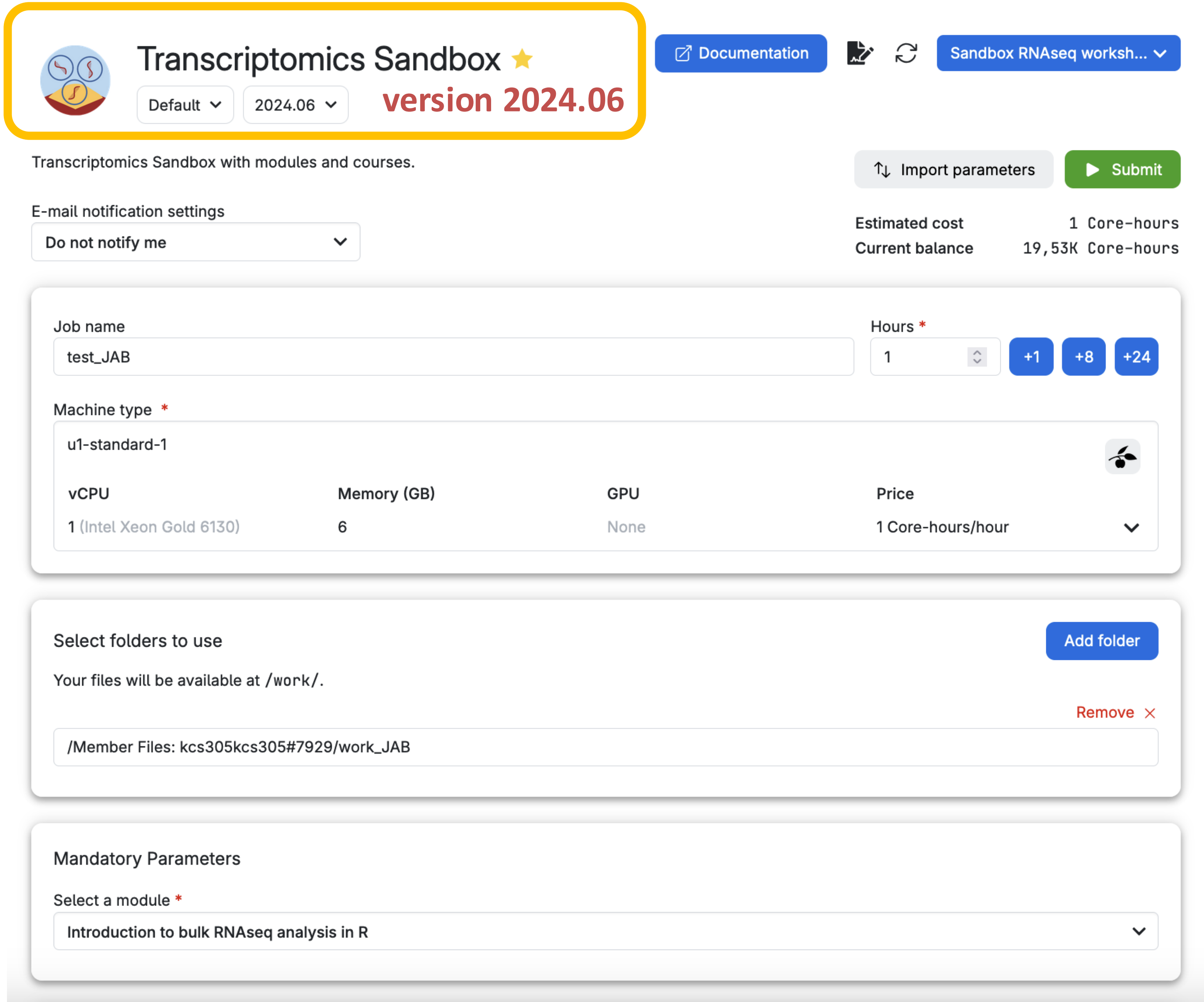
/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost	1 Core-hours
Current balance	19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Your files will be available at `/work/`.

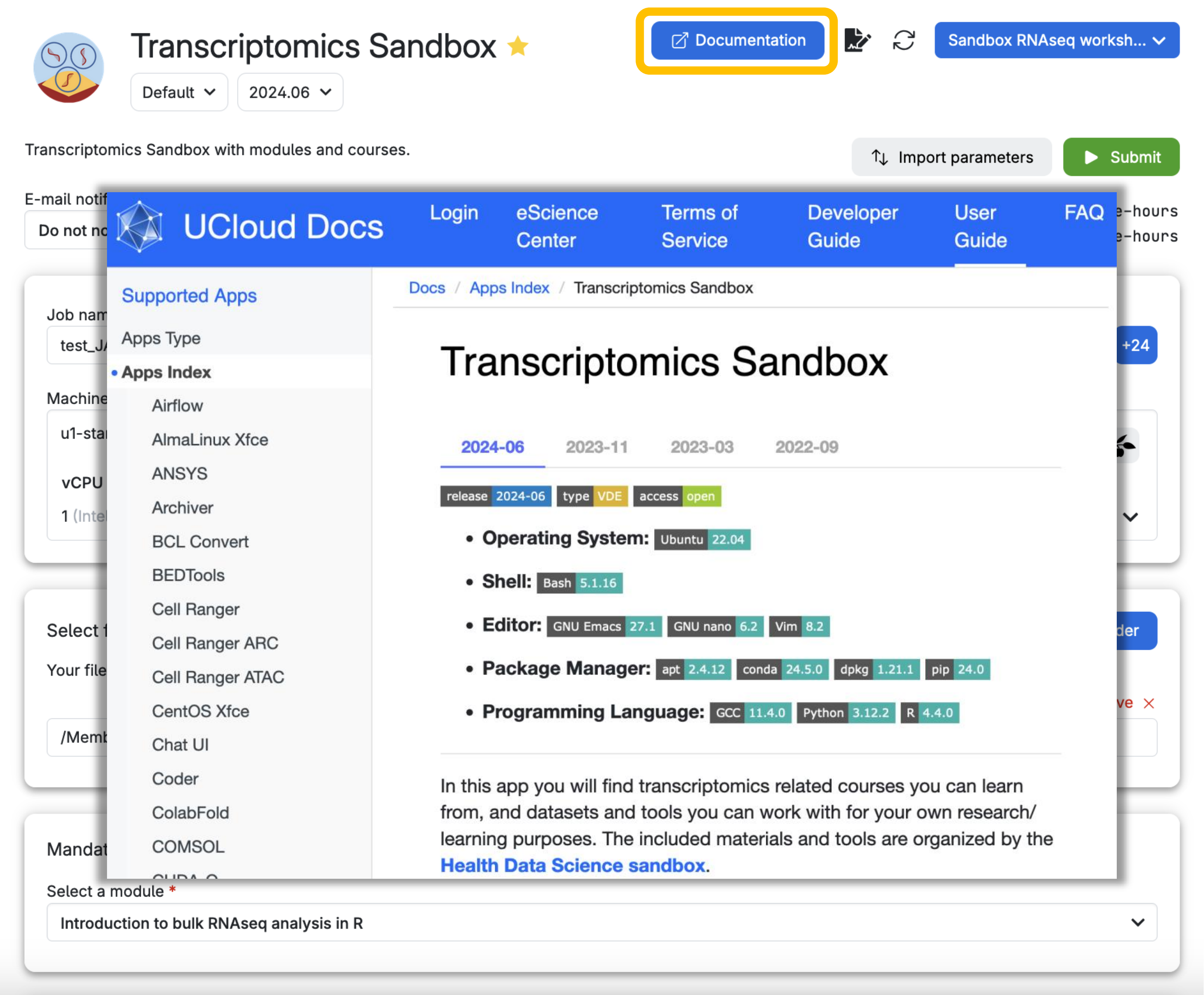
/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox ★

Default ⌄ 2024.06 ⌄

Documentation

Sandbox RNAseq worksh... ⌄

Transcriptomics Sandbox with modules and courses.

Import parameters

▶ Submit

E-mail notif

Do not no

e-hours

e-hours

Job nam

test_J

Machine

u1-sta

vCPU

1 (Inte

+24

Select f

Your file

/Memb

der

ve ✕

Mandat

Select a module *

Introduction to bulk RNAseq analysis in R ⌄

UCloud Docs

Login | eScience Center | Terms of Service | Developer Guide | User Guide | FAQ

Supported Apps

Apps Type

• Apps Index

Airflow

AlmaLinux Xfce

ANSYS

Archiver

BCL Convert

BEDTools

Cell Ranger

Cell Ranger ARC

Cell Ranger ATAC

CentOS Xfce

Chat UI

Coder

ColabFold

COMSOL

Docs / Apps Index / Transcriptomics Sandbox

Transcriptomics Sandbox

2024-06 | 2023-11 | 2023-03 | 2022-09

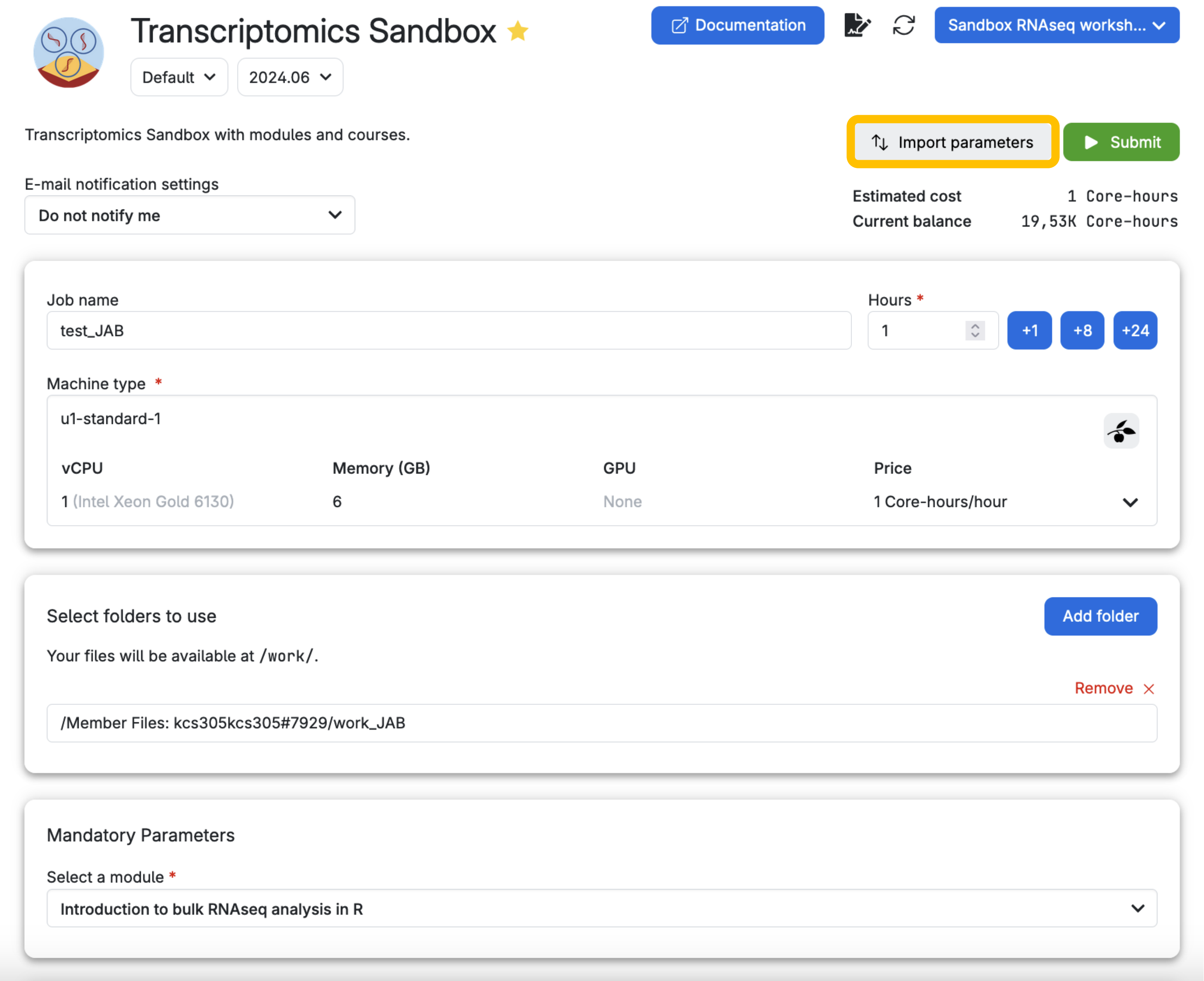
release 2024-06 | type VDE | access open

- **Operating System:** Ubuntu 22.04
- **Shell:** Bash 5.1.16
- **Editor:** GNU Emacs 27.1 | GNU nano 6.2 | Vim 8.2
- **Package Manager:** apt 2.4.12 | conda 24.5.0 | dpkg 1.21.1 | pip 24.0
- **Programming Language:** GCC 11.4.0 | Python 3.12.2 | R 4.4.0

In this app you will find transcriptomics related courses you can learn from, and datasets and tools you can work with for your own research/learning purposes. The included materials and tools are organized by the **Health Data Science sandbox**.

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox ★

Default | 2024.06

Transcriptomics Sandbox with modules and courses.

Estimated cost	1 Core-hours
Current balance	19,53K Core-hours

E-mail notification settings: Do not notify me

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Your files will be available at `/work/`.

Remove ✕

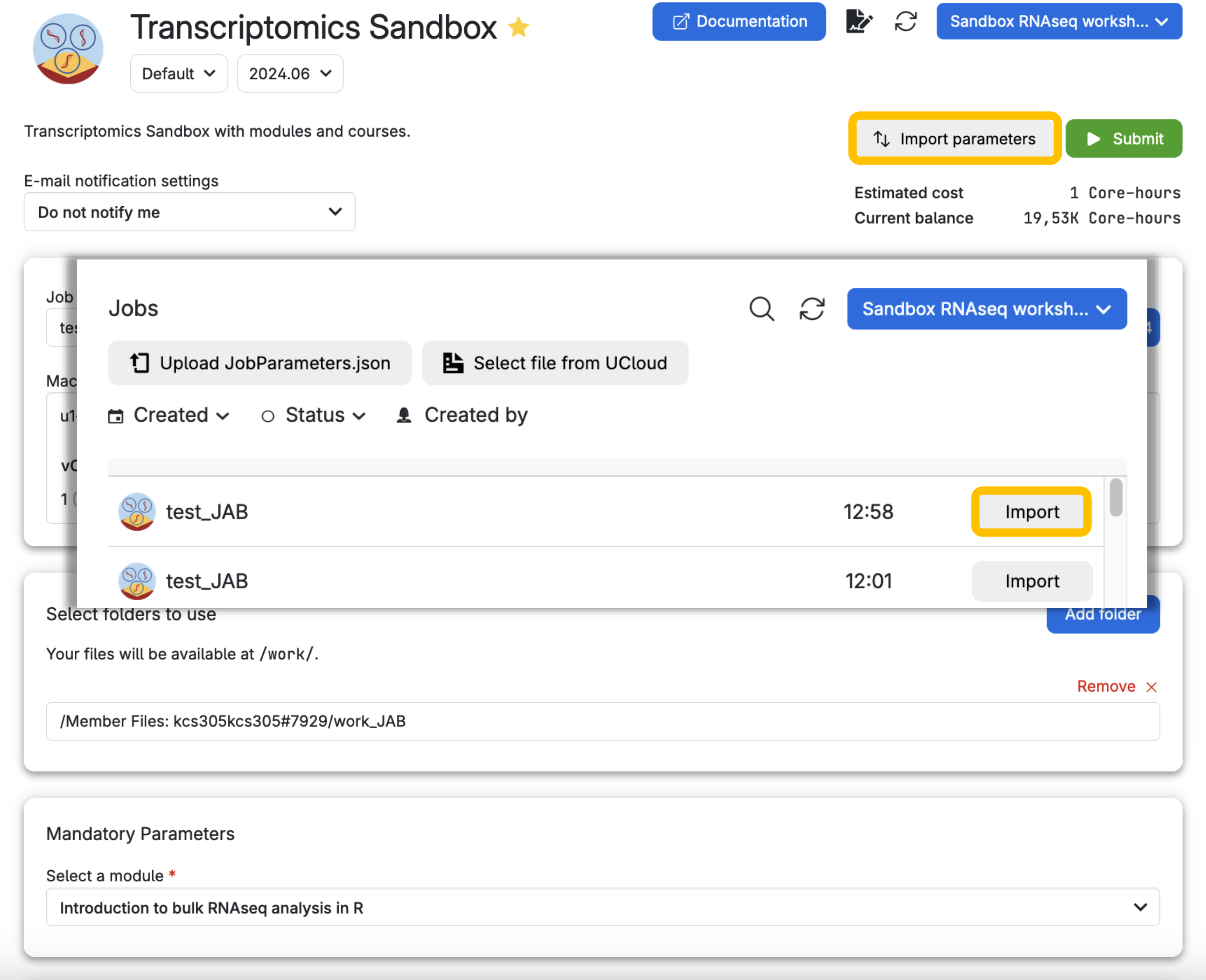
/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost	1 Core-hours
Current balance	19,53K Core-hours

test_JAB	12:58	Import
test_JAB	12:01	Import

Select folders to use

Your files will be available at /work/.

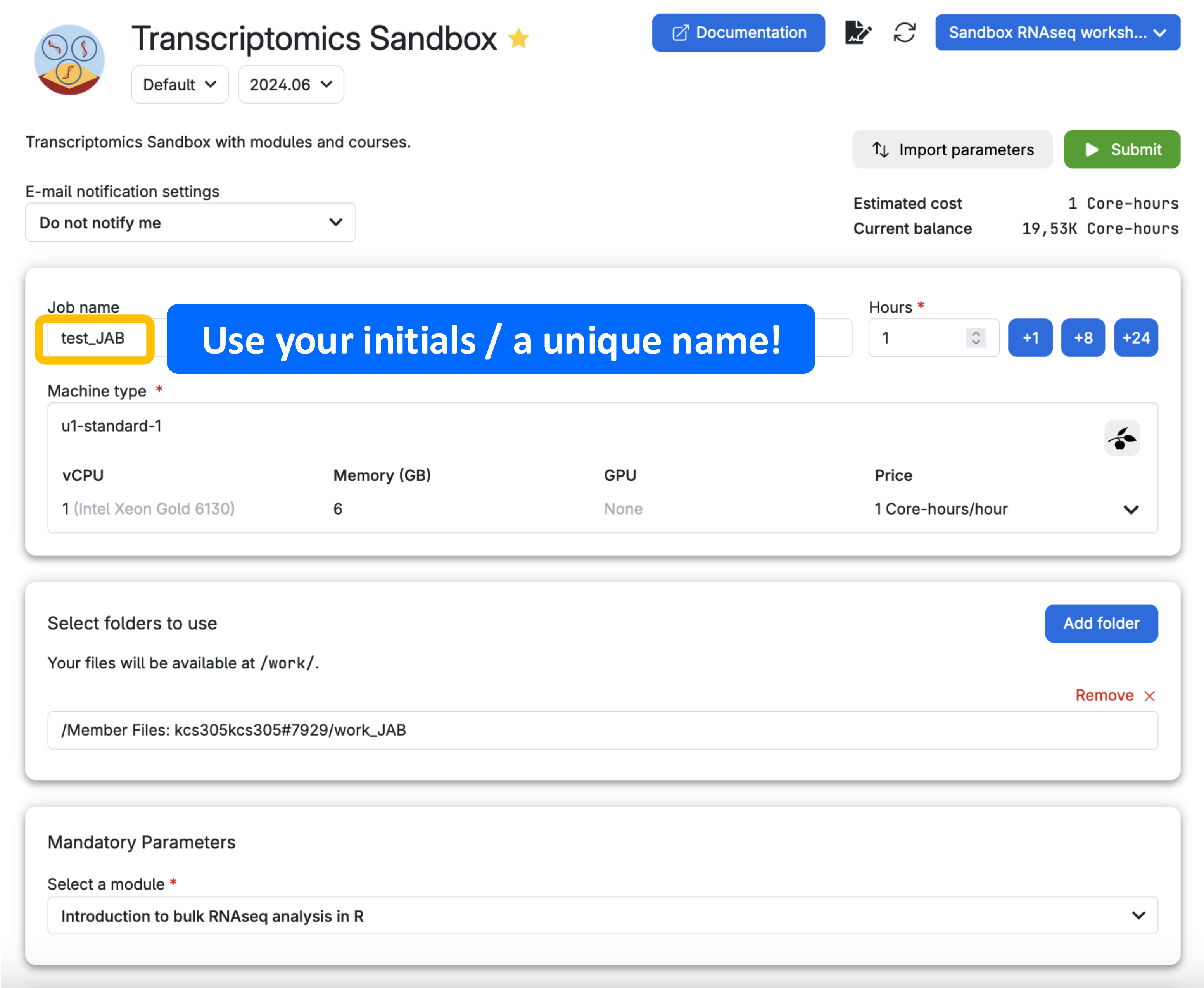
/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. **Job name, hours, and machine type (resources set-up)**
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox ★

Default ⌄ | 2024.06 ⌄

Documentation | Sandbox RNAseq worksh... ⌄

Transcriptomics Sandbox with modules and courses.

Import parameters | ▶ Submit

E-mail notification settings: Do not notify me

Estimated cost: 1 Core-hours

Current balance: 19,53K Core-hours

Job name: test_JAB

Use your initials / a unique name!

Hours *: 1 | +1 | +8 | +24

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use | Add folder

Your files will be available at `/work/`.

Remove ✕

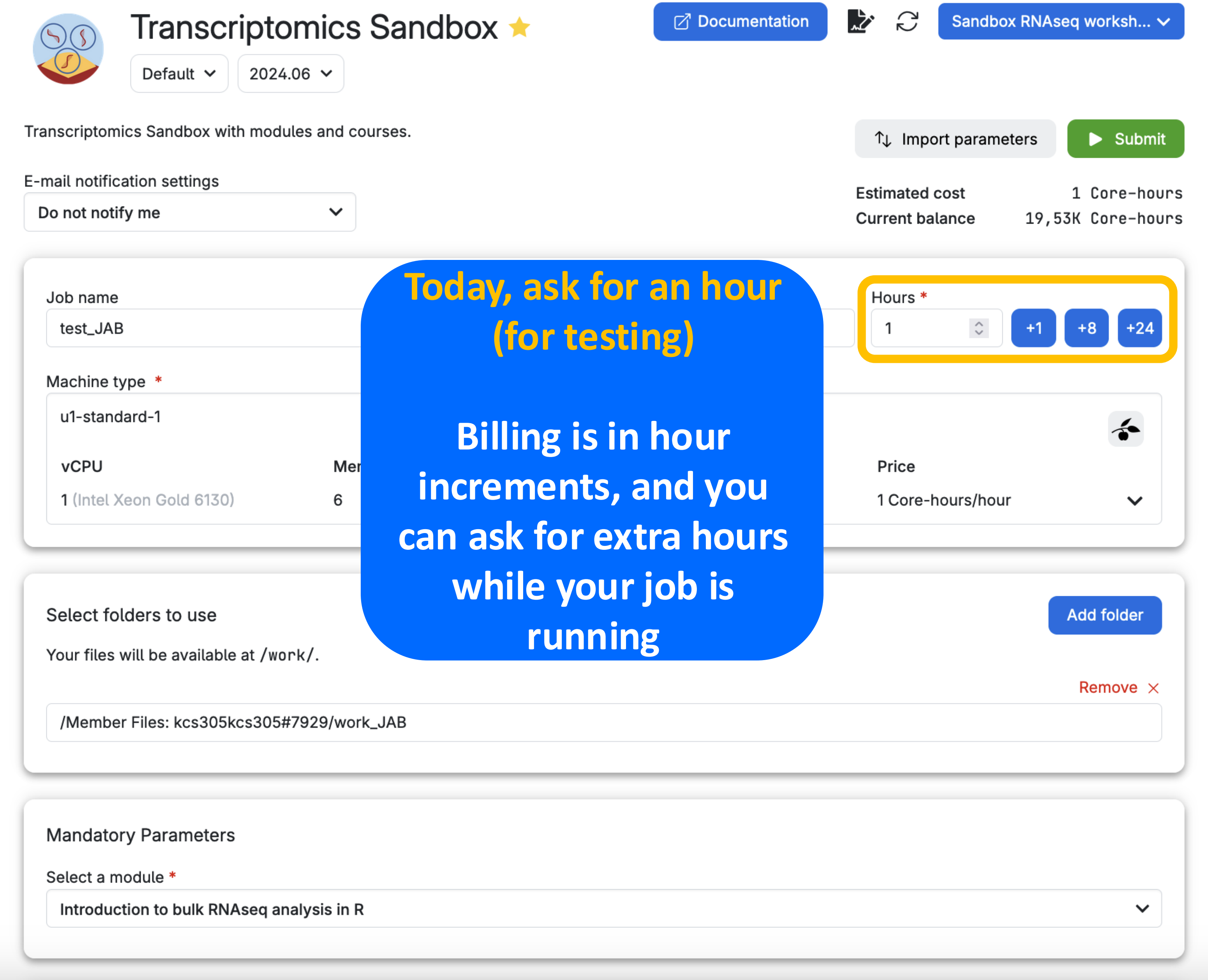
/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost 1 Core-hours
Current balance 19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU: 1 (Intel Xeon Gold 6130)

Mem: 6

Price: 1 Core-hours/hour

Today, ask for an hour (for testing)

Billing is in hour increments, and you can ask for extra hours while your job is running

Select folders to use

Your files will be available at `/work/`.

/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1 App & version (dropdown menu to change it)

2 Read documentation before using it

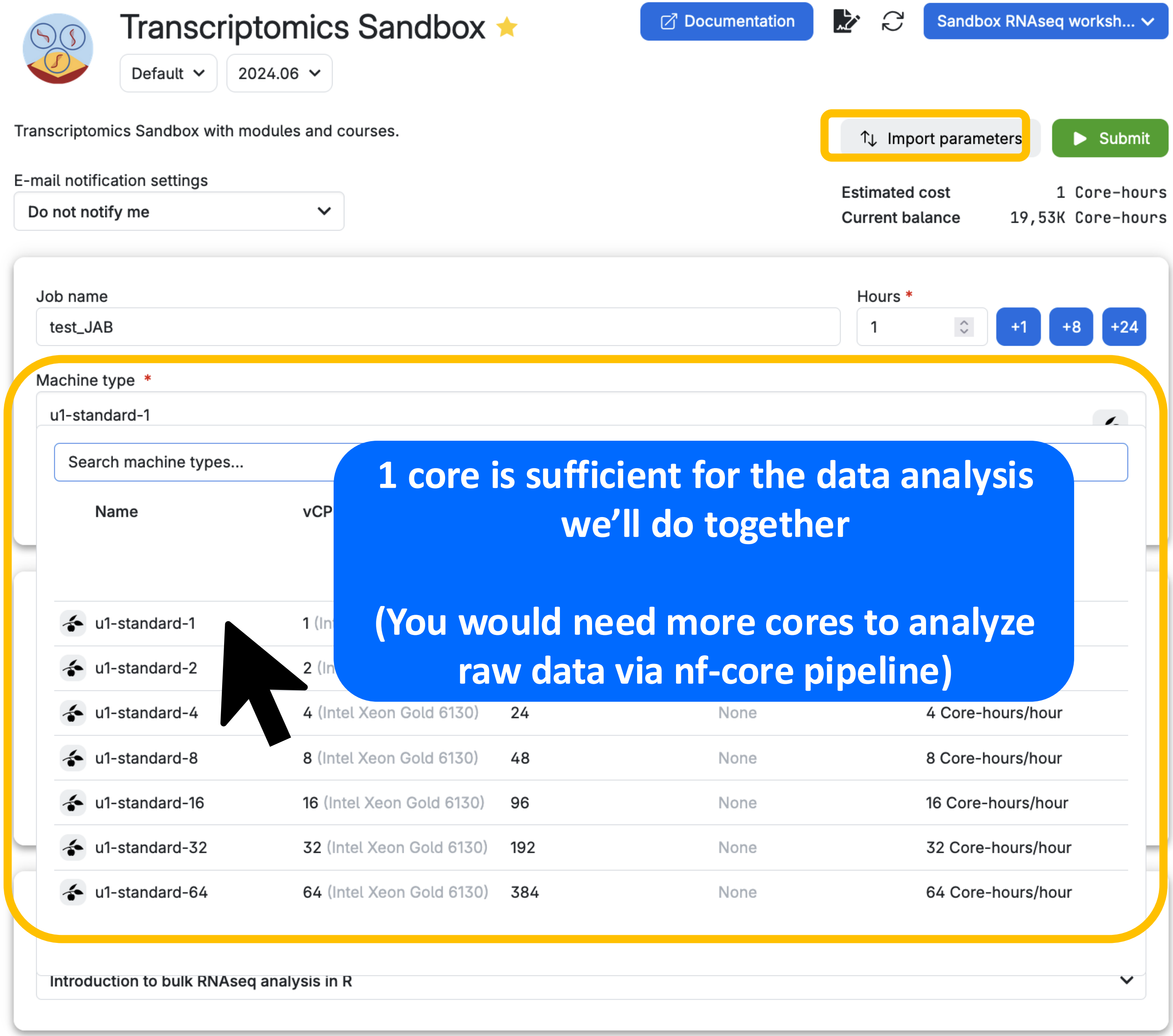
3 Import parameters (if wanted from a previous job)

4 Job name, hours, and machine type (resources set-up)

5 Folders to access while running this particular job

6 Module to use (which includes Notebooks & Data)

7 ▶ Submit



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost: 1 Core-hours

Current balance: 19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

Search machine types...

Name	vCP			
u1-standard-1	1 (In			
u1-standard-2	2 (In			
u1-standard-4	4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour
u1-standard-8	8 (Intel Xeon Gold 6130)	48	None	8 Core-hours/hour
u1-standard-16	16 (Intel Xeon Gold 6130)	96	None	16 Core-hours/hour
u1-standard-32	32 (Intel Xeon Gold 6130)	192	None	32 Core-hours/hour
u1-standard-64	64 (Intel Xeon Gold 6130)	384	None	64 Core-hours/hour

Introduction to bulk RNAseq analysis in R

1 core is sufficient for the data analysis we'll do together

(You would need more cores to analyze raw data via nf-core pipeline)

Submitting a job with a Sandbox app

1 App & version (dropdown menu to change it)

2 Read documentation before using it

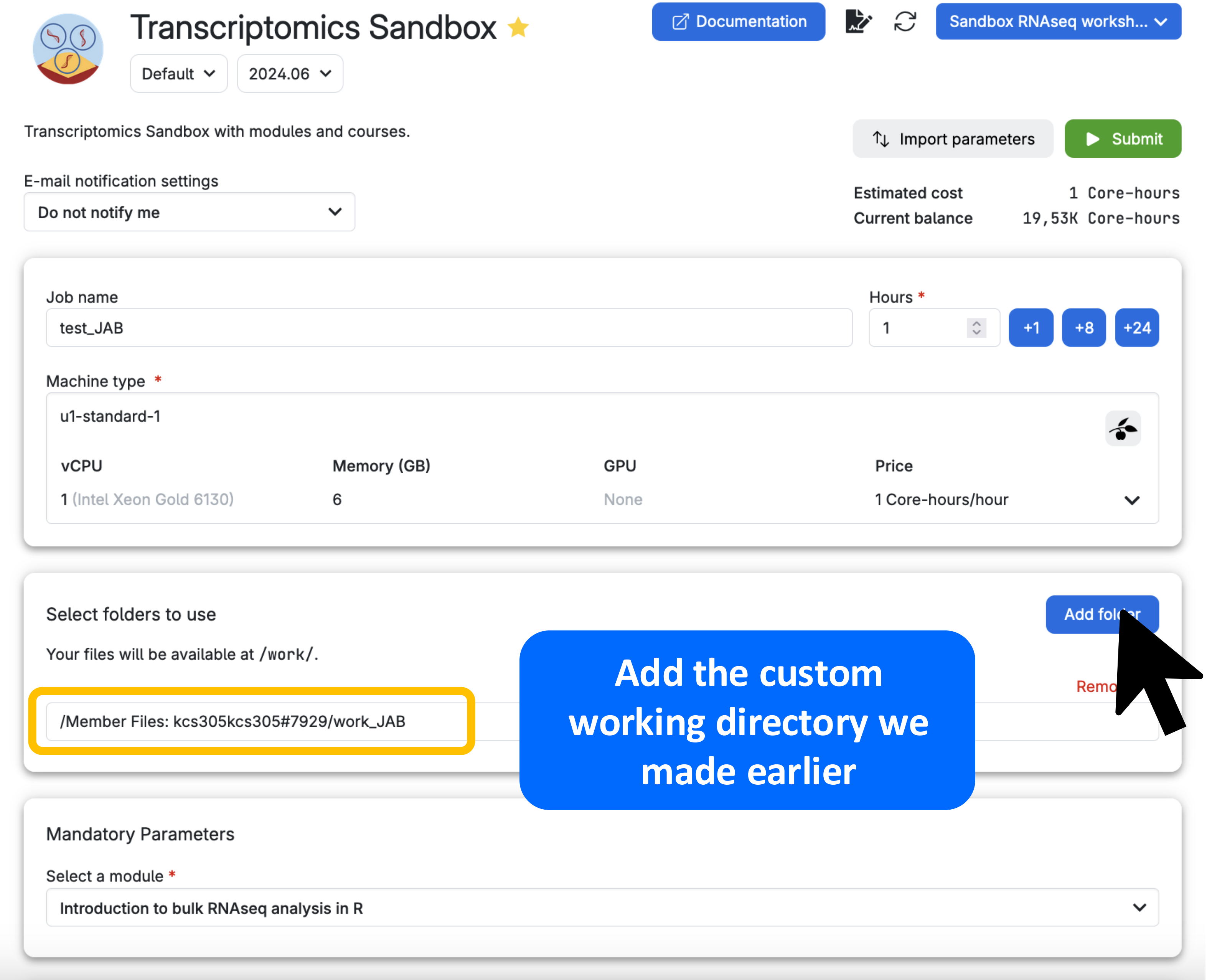
3 Import parameters (if wanted from a previous job)

4 Job name, hours, and machine type (resources set-up)

5 Folders to access while running this particular job

6 Module to use (which includes Notebooks & Data)

7 ▶ Submit



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost: 1 Core-hours
Current balance: 19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Your files will be available at /work/.

/Member Files: kcs305kcs305#7929/work_JAB

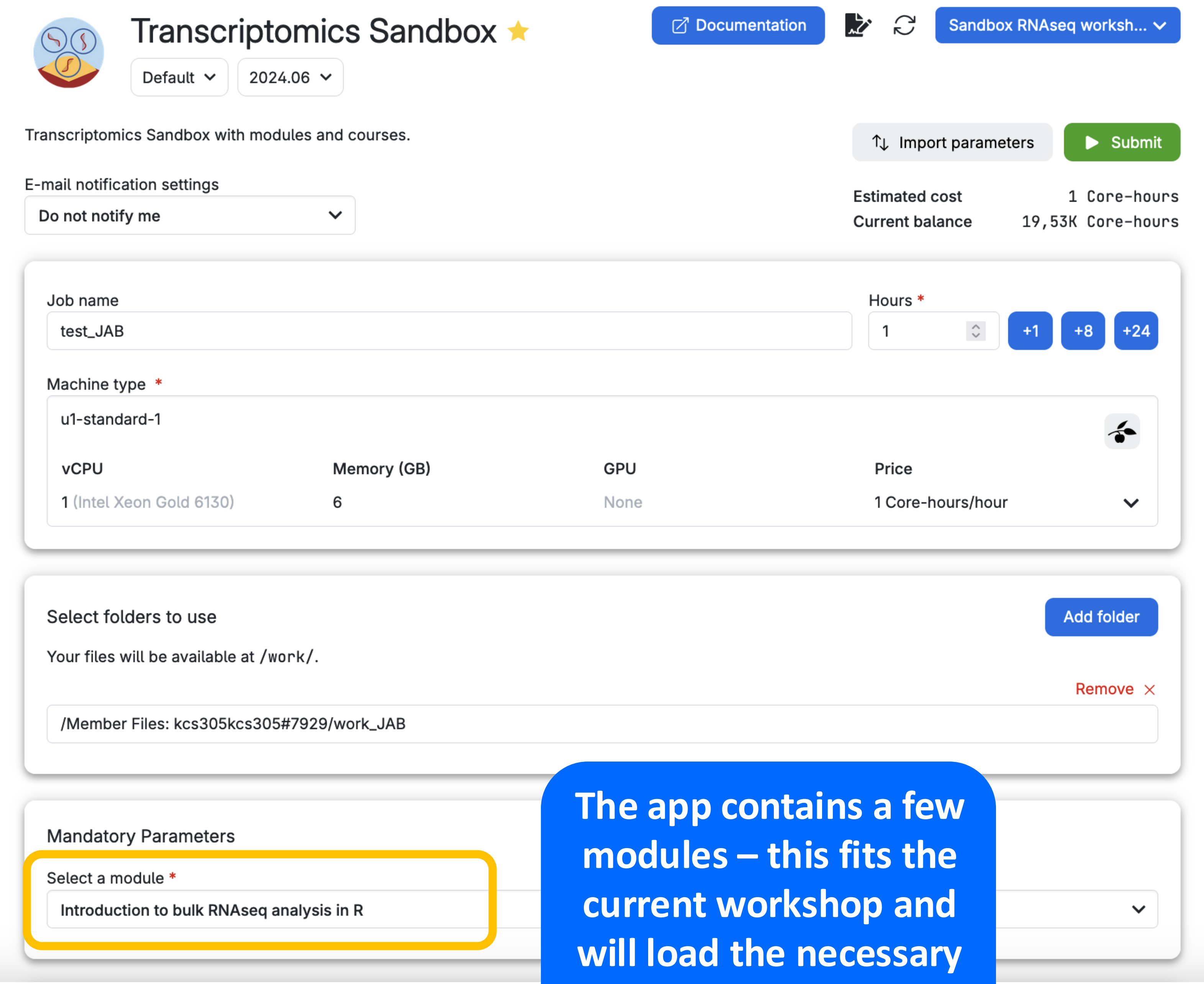
Add the custom working directory we made earlier

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. ▶ Submit



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost: 1 Core-hours
Current balance: 19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Your files will be available at `/work/`.

/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

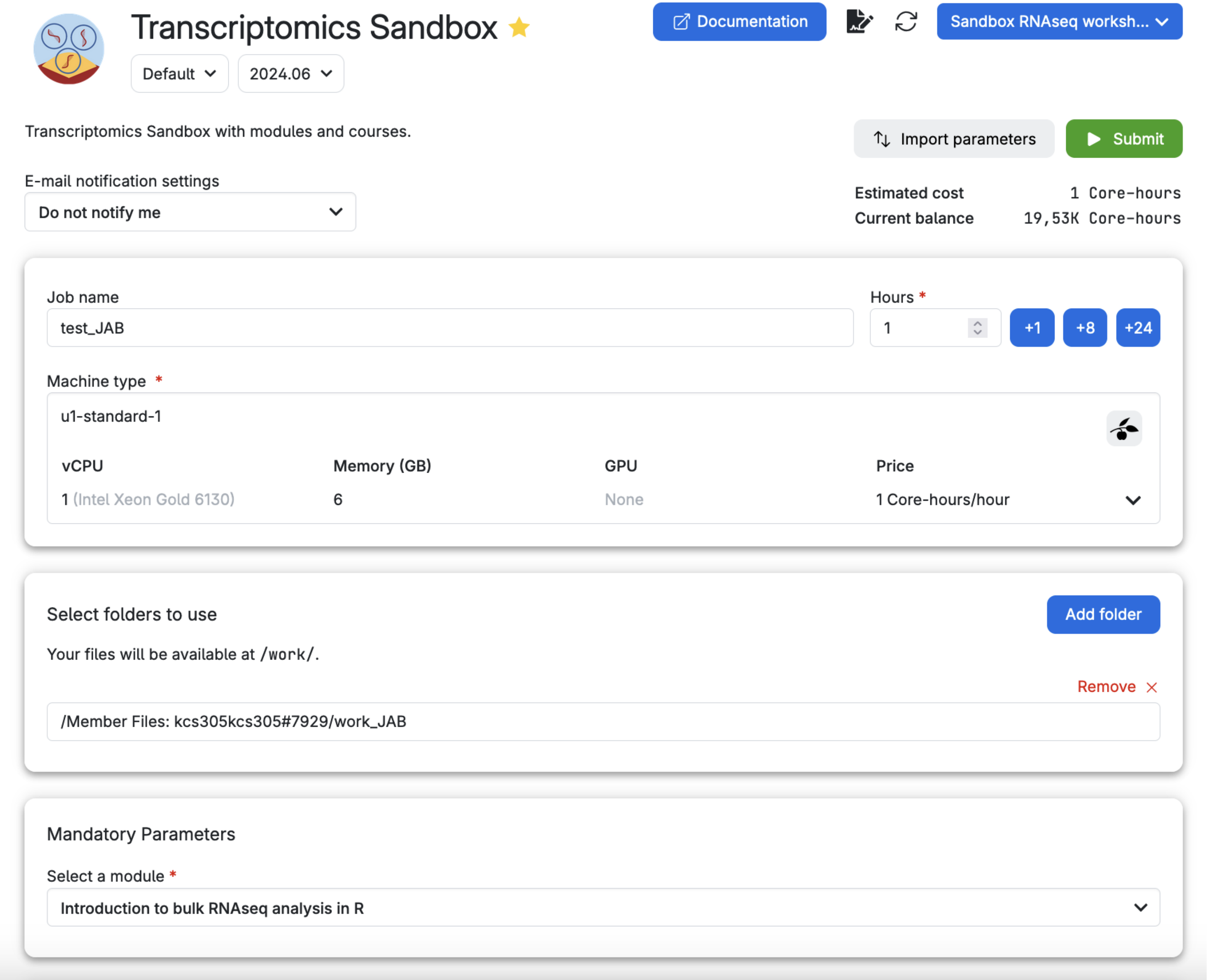
The app contains a few modules – this fits the current workshop and will load the necessary tools and notebooks

Submitting a job with a Sandbox app

For those with some HPC experience…

Do these steps look familiar?

Perhaps similar to a job's bash script that you submit using a workload manager like SLURM or PBS?



Transcriptomics Sandbox ★

Default 2024.06

Documentation

Sandbox RNAseq worksh...

Transcriptomics Sandbox with modules and courses.

Import parameters

Submit

E-mail notification settings: Do not notify me

Estimated cost: 1 Core-hours

Current balance: 19,53K Core-hours

Job name: test_JAB

Hours *: 1

+1 +8 +24

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Add folder

Your files will be available at /work/.

Remove ✕

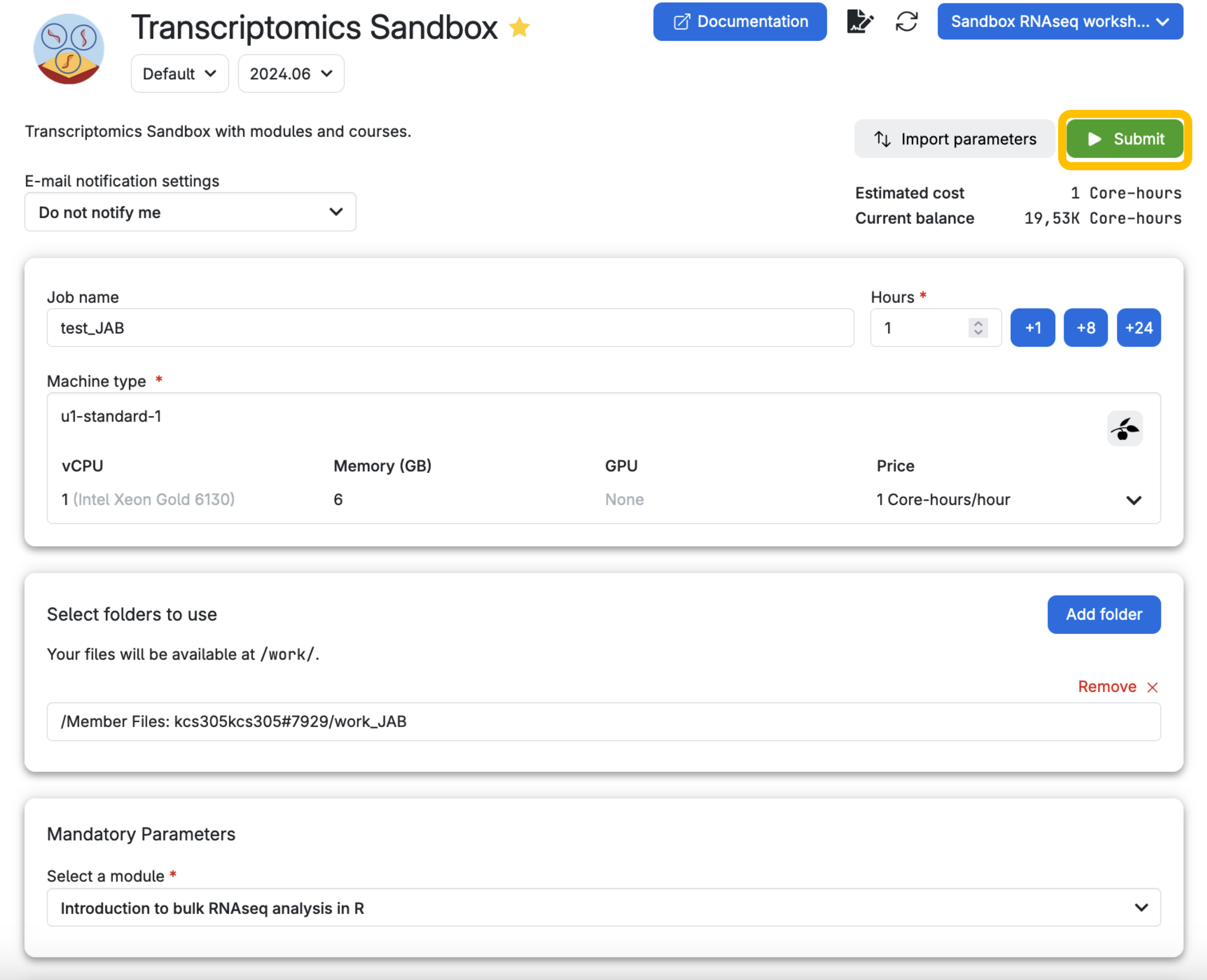
/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. **Review & Submit**



Transcriptomics Sandbox ★

Default | 2024.06

Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost: 1 Core-hours
Current balance: 19,53K Core-hours

Job name: test_JAB

Hours *: 1

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

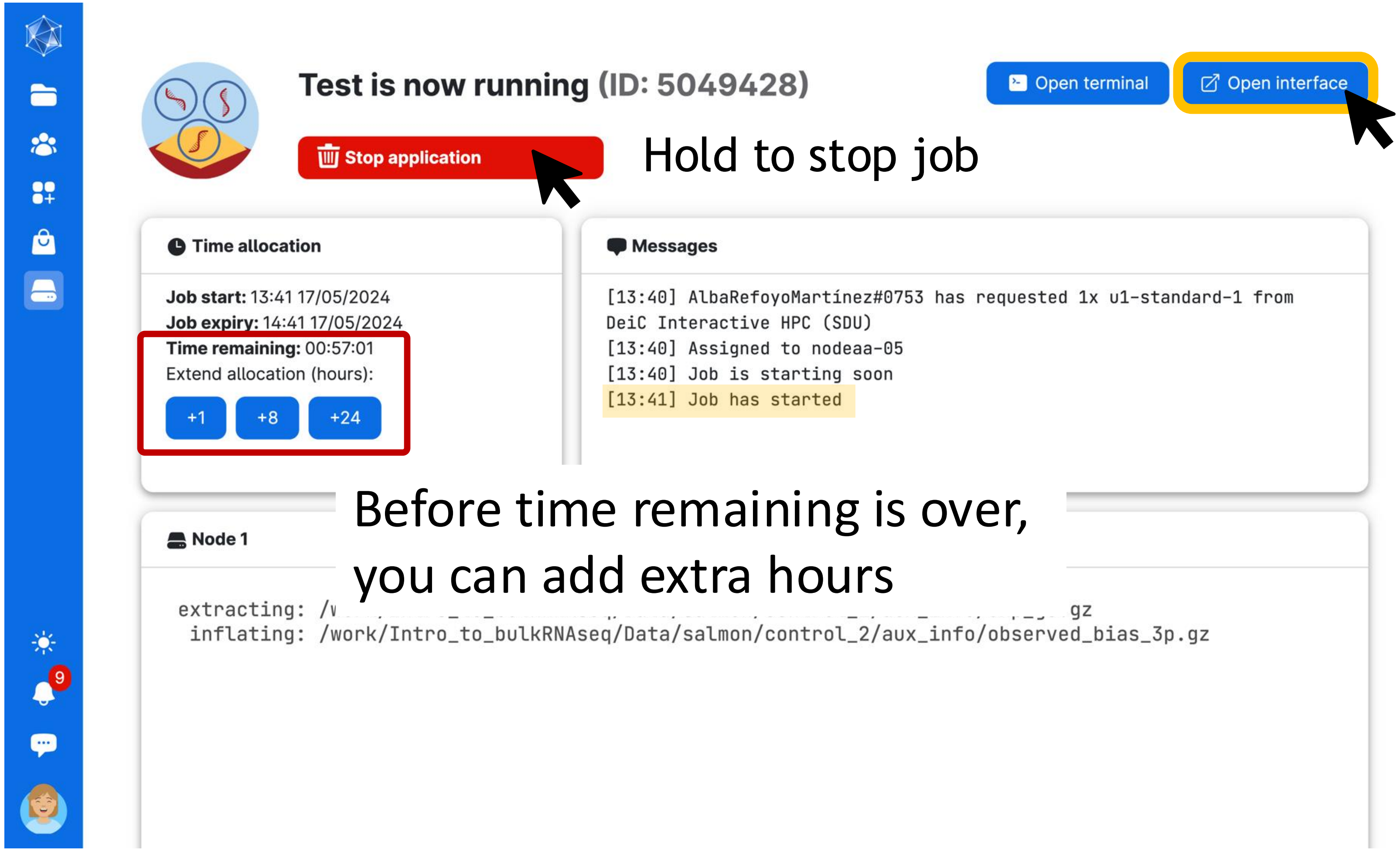
Your files will be available at `/work/`.

/Member Files: kcs305kcs305#7929/work_JAB

Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Jobs



Test is now running (ID: 5049428)

Hold to stop job

Time allocation

Job start: 13:41 17/05/2024
Job expiry: 14:41 17/05/2024
Time remaining: 00:57:01
Extend allocation (hours):

+1 +8 +24

Messages

```
[13:40] AlbaRefoyoMartínez#0753 has requested 1x u1-standard-1 from
DeiC Interactive HPC (SDU)
[13:40] Assigned to nodeaa-05
[13:40] Job is starting soon
[13:41] Job has started
```

Node 1

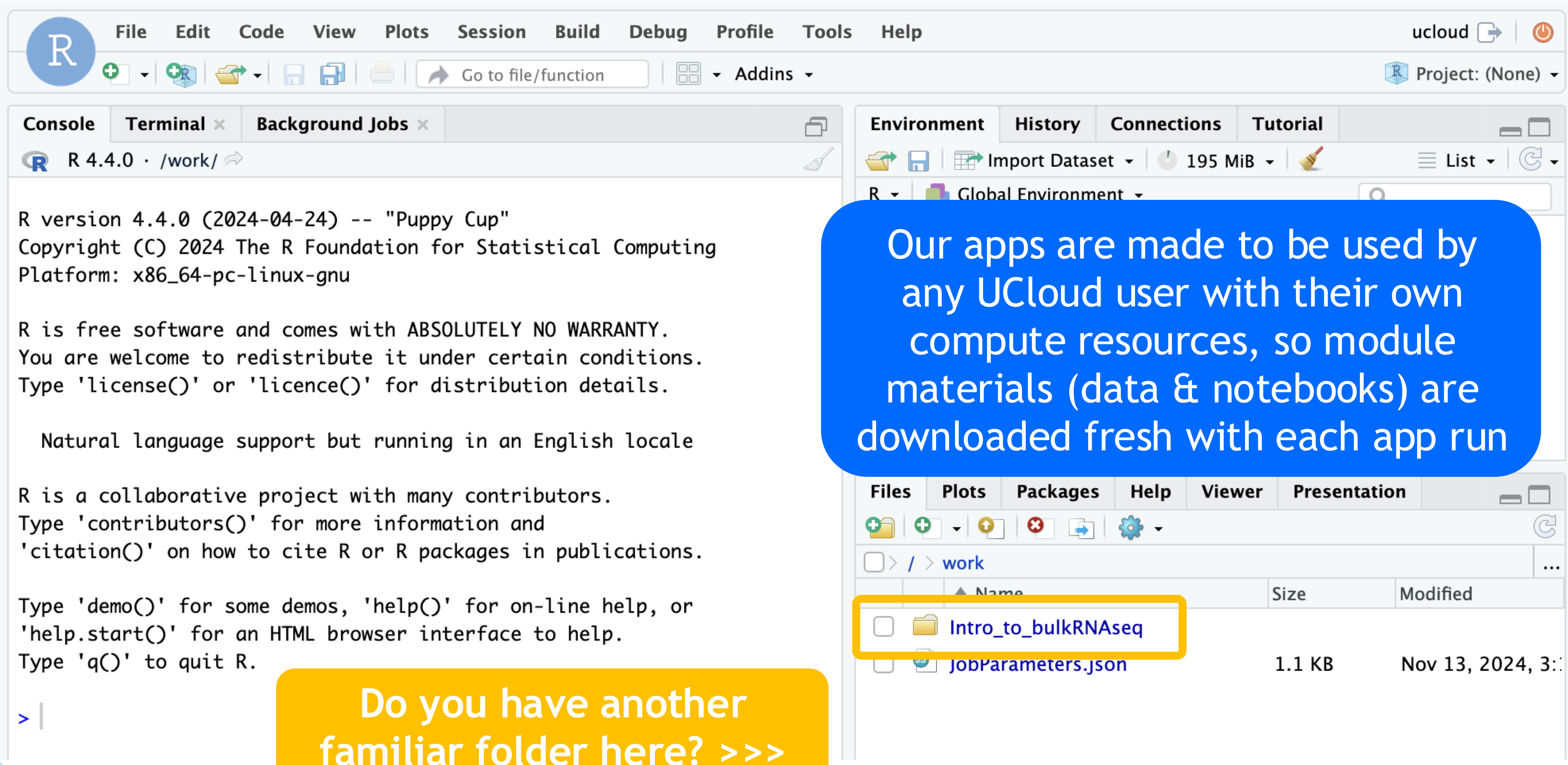
```
extracting: /w                                                          gz
 inflating: /work/Intro_to_bulkRNAseq/Data/salmon/control_2/aux_info/observed_bias_3p.gz
```

Before time remaining is over, you can add extra hours

If your ‘Open interface button’ doesn’t go dark blue after you get a message that your ‘Job has started’, then hit refresh (browser)!

Do the same on the new tab that pops up if it just spins, too

Jobs



```
R version 4.4.0 (2024-04-24) -- "Puppy Cup"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

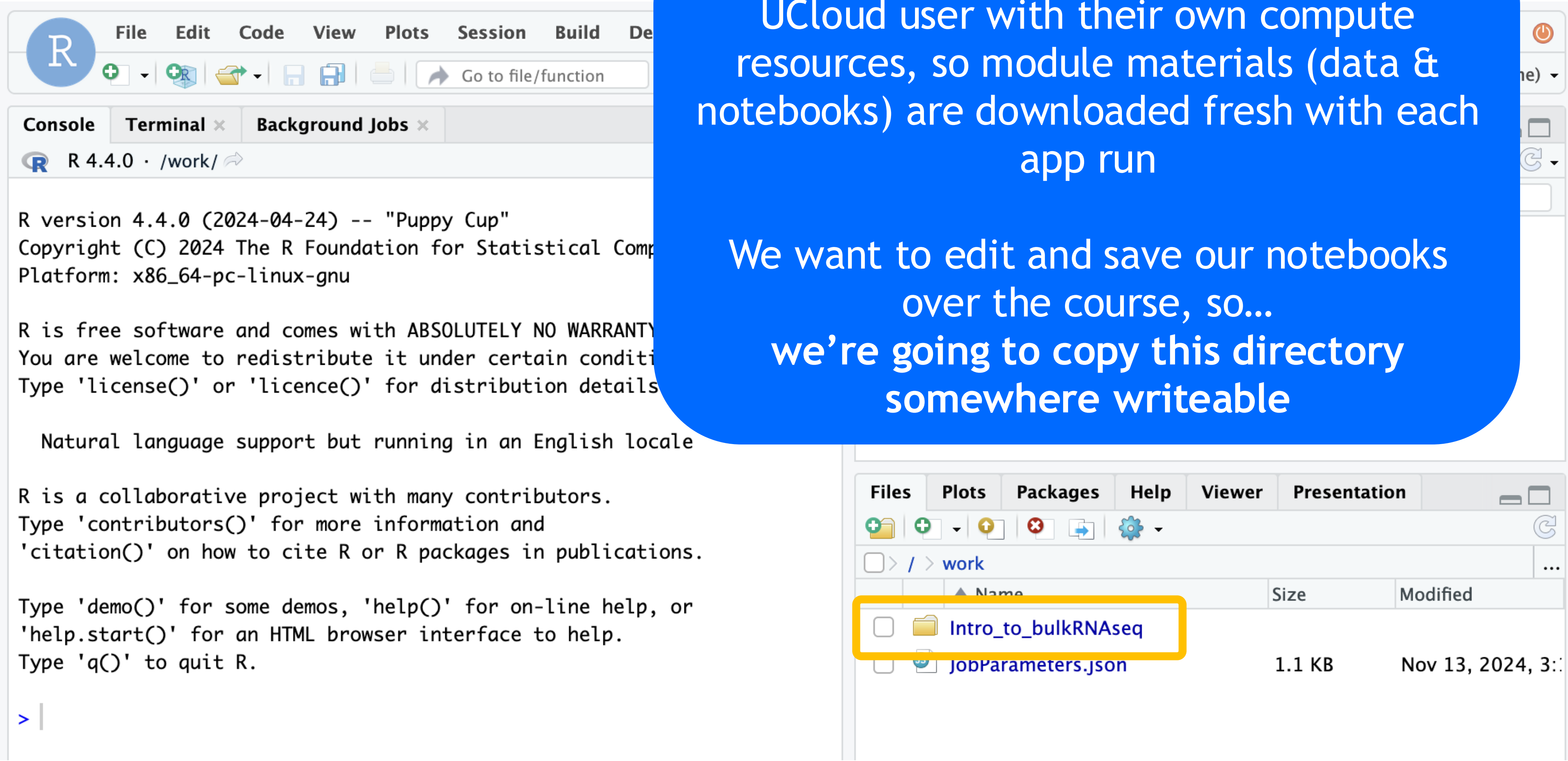
Name	Size	Modified
Intro_to_bulkRNAseq		
JobParameters.json	1.1 KB	Nov 13, 2024, 3::

Our apps are made to be used by any UCloud user with their own compute resources, so module materials (data & notebooks) are downloaded fresh with each app run

Do you have another familiar folder here? >>>

HeaDS

Jobs



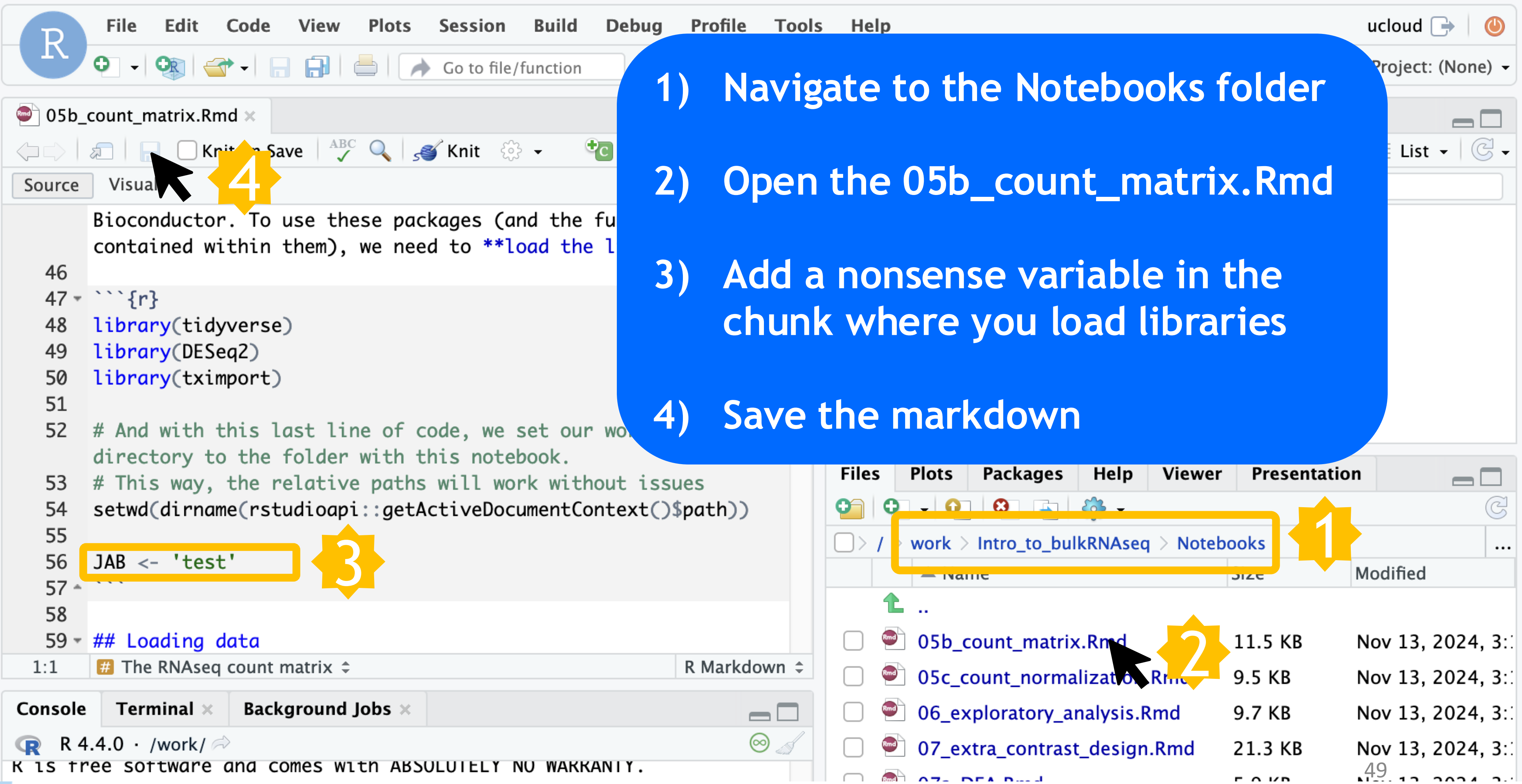
```
R version 4.4.0 (2024-04-24) -- "Puppy Cup"
Copyright (C) 2024 The R Foundation for Statistical Comp
Platform: x86_64-pc-linux-gnu

R is free software and comes with ABSOLUTELY NO WARRANTY
You are welcome to redistribute it under certain conditi
Type 'license()' or 'licence()' for distribution details

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Name	Size	Modified
Intro_to_bulkRNAseq		
JobParameters.json	1.1 KB	Nov 13, 2024, 3:

Our apps are made to be used by any UCloud user with their own compute resources, so module materials (data & notebooks) are downloaded fresh with each app run

We want to edit and save our notebooks over the course, so…
we’re going to copy this directory somewhere writeable

Jobs



~~~
Bioconductor. To use these packages (and the fu
contained within them), we need to **load the l

```{r}
library(tidyverse)
library(DESeq2)
library(tximport)

# And with this last line of code, we set our wo
directory to the folder with this notebook.
# This way, the relative paths will work without issues
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

JAB <- 'test'
```

## Loading data
~~~

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
```

/ › work › Intro_to_bulkRNAseq › Notebooks

Name	Size	Modified
..		
05b_count_matrix.Rmd	11.5 KB	Nov 13, 2024, 3:
05c_count_normalization.Rmd	9.5 KB	Nov 13, 2024, 3:
06_exploratory_analysis.Rmd	9.7 KB	Nov 13, 2024, 3:
07_extra_contrast_design.Rmd	21.3 KB	Nov 13, 2024, 3:

1) Navigate to the Notebooks folder

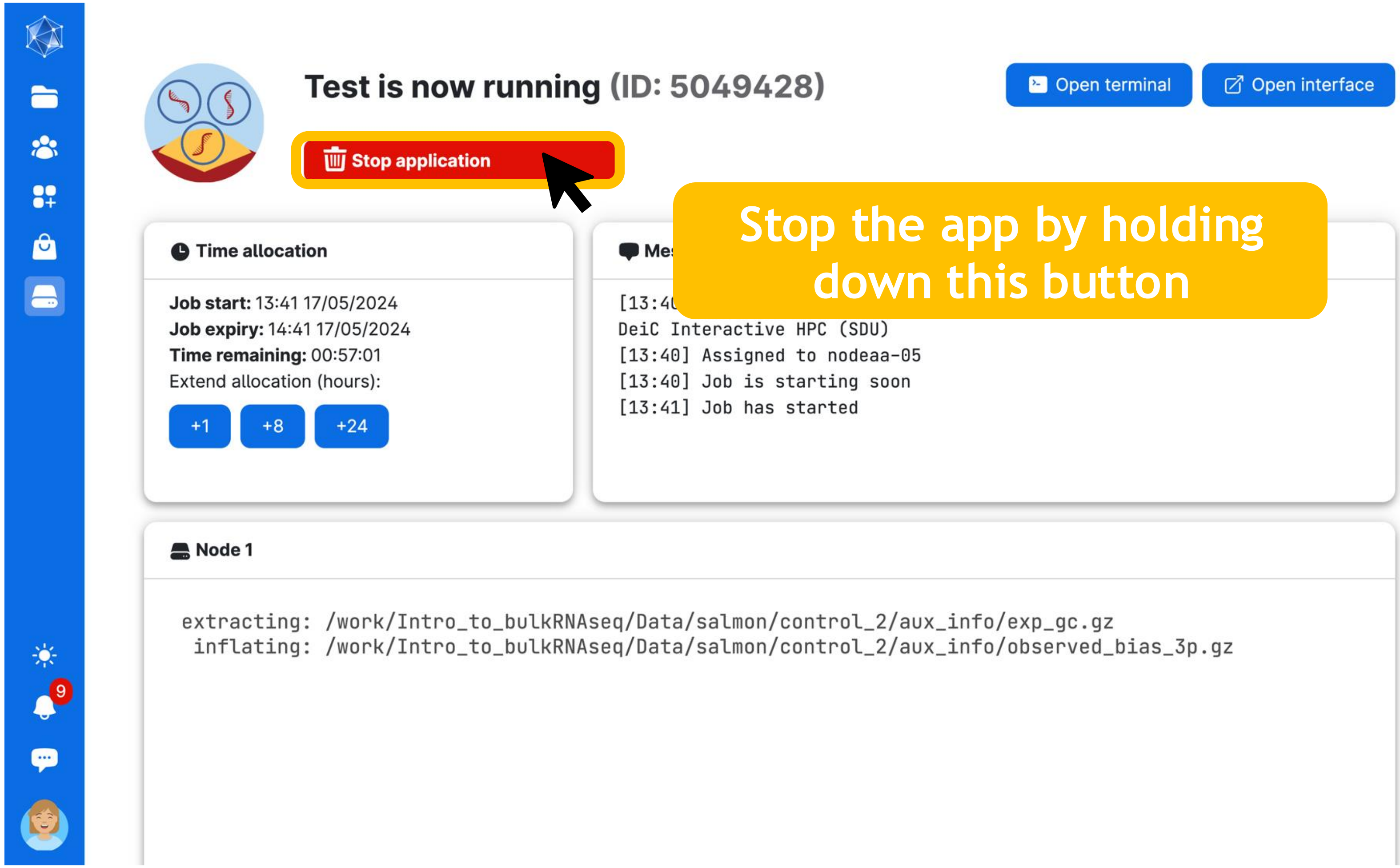
2) Open the 05b_count_matrix.Rmd

3) Add a nonsense variable in the chunk where you load libraries

4) Save the markdown

HeaDS

Jobs



Test is now running (ID: 5049428)

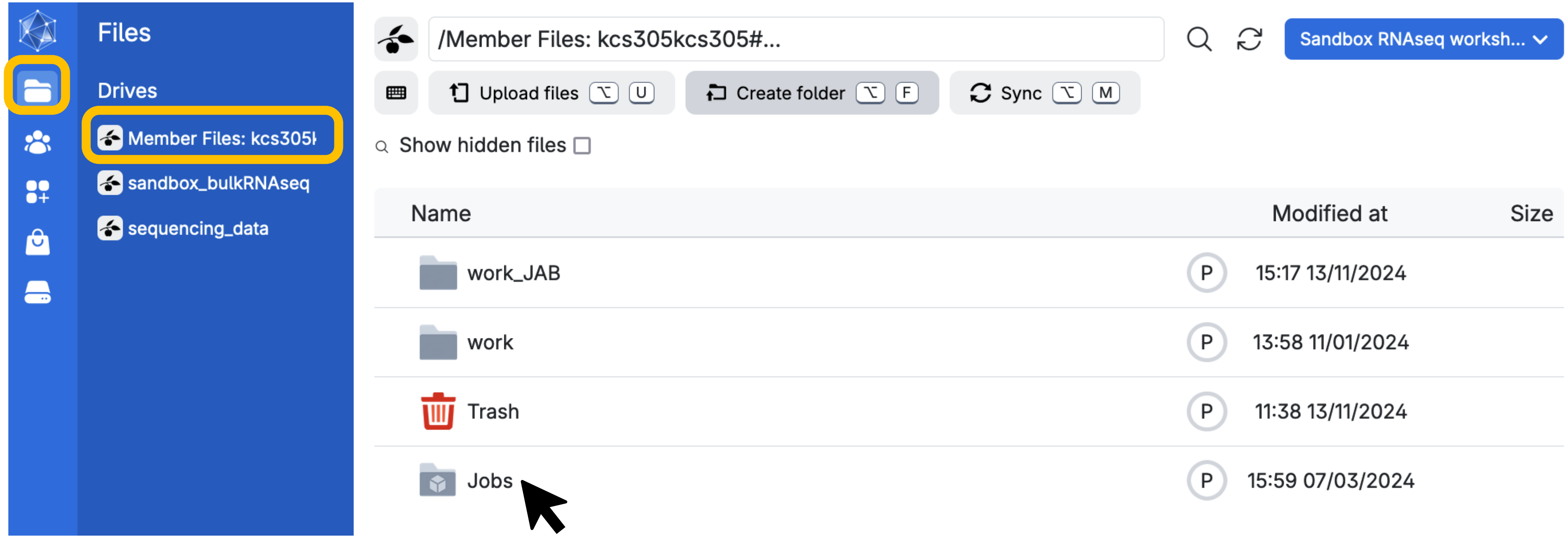
Job start: 13:41 17/05/2024
Job expiry: 14:41 17/05/2024
Time remaining: 00:57:01
Extend allocation (hours):

```
[13:40
DeiC Interactive HPC (SDU)
[13:40] Assigned to nodeaa-05
[13:40] Job is starting soon
[13:41] Job has started
```

```
extracting: /work/Intro_to_bulkRNAseq/Data/salmon/control_2/aux_info/exp_gc.gz
 inflating: /work/Intro_to_bulkRNAseq/Data/salmon/control_2/aux_info/observed_bias_3p.gz
```

Stop the app by holding down this button

Jobs



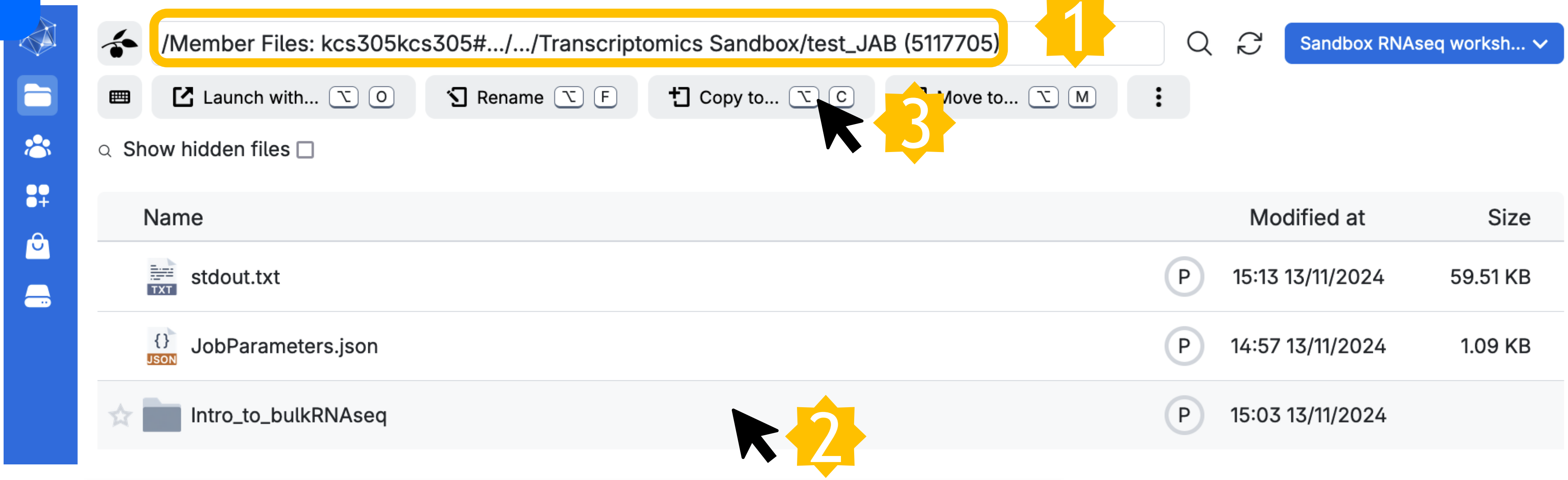
Name		Modified at	Size
work_JAB	P	15:17 13/11/2024	
work	P	13:58 11/01/2024	
Trash	P	11:38 13/11/2024	
Jobs	P	15:59 07/03/2024	

Go to your Jobs folder in your personal drive

Jobs

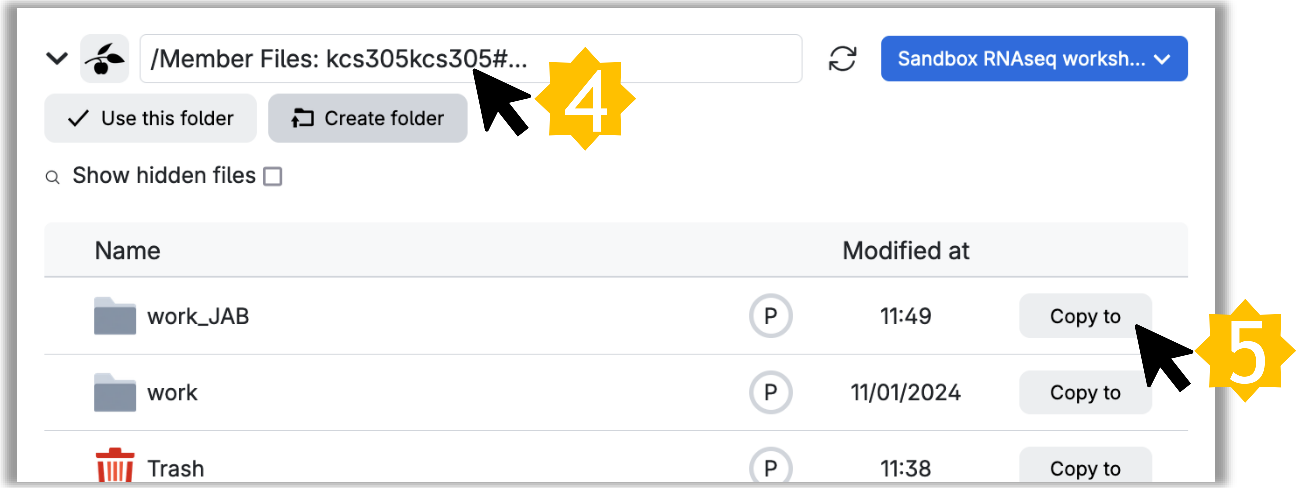
1. Go into Transcriptomics Sandbox and pick the job we just ran
2. Click ONCE to select Intro_to_bulkRNAseq
3. Click ‘Copy to…’
4. Click on your root member files directory
5. Copy to the custom working directory you made

Now you will always have your own, writeable directory when using the app



/Member Files: kcs305kcs305#.../.../Transcriptomics Sandbox/test_JAB (5117705)

Name		Modified at	Size
stdout.txt	P	15:13 13/11/2024	59.51 KB
JobParameters.json	P	14:57 13/11/2024	1.09 KB
Intro_to_bulkRNAseq	P	15:03 13/11/2024	



/Member Files: kcs305kcs305#...

Name		Modified at	
work_JAB	P	11:49	Copy to
work	P	11/01/2024	Copy to
Trash	P	11:38	Copy to

Jobs

1. Go into Transcriptomics Sandbox and pick the job we just ran

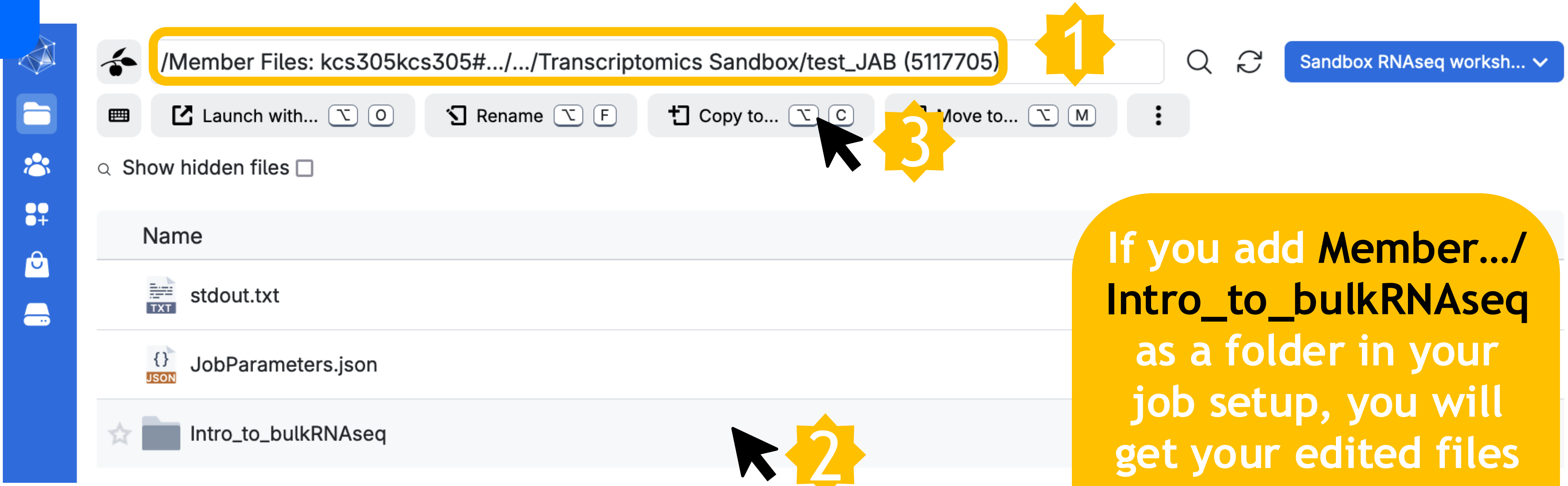
2. Click ONCE to select Intro_to_bulkRNAseq

3. Click 'Copy to…'

4. Click on your root member files directory

5. Copy to the custom working directory you made

Now you will always have your own, writeable directory when using the app

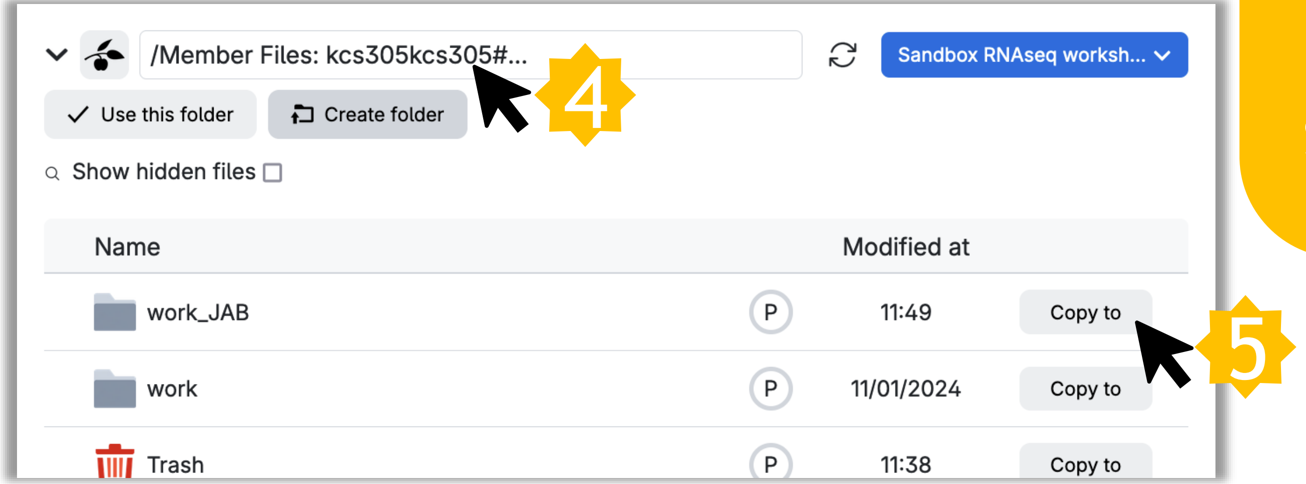


/Member Files: kcs305kcs305#…/…/Transcriptomics Sandbox/test_JAB (5117705)

Name
stdout.txt
JobParameters.json
Intro_to_bulkRNAseq

If you add **Member…/ Intro_to_bulkRNAseq** as a folder in your job setup, you will get your edited files

If you don't, you'll get clean files from the app

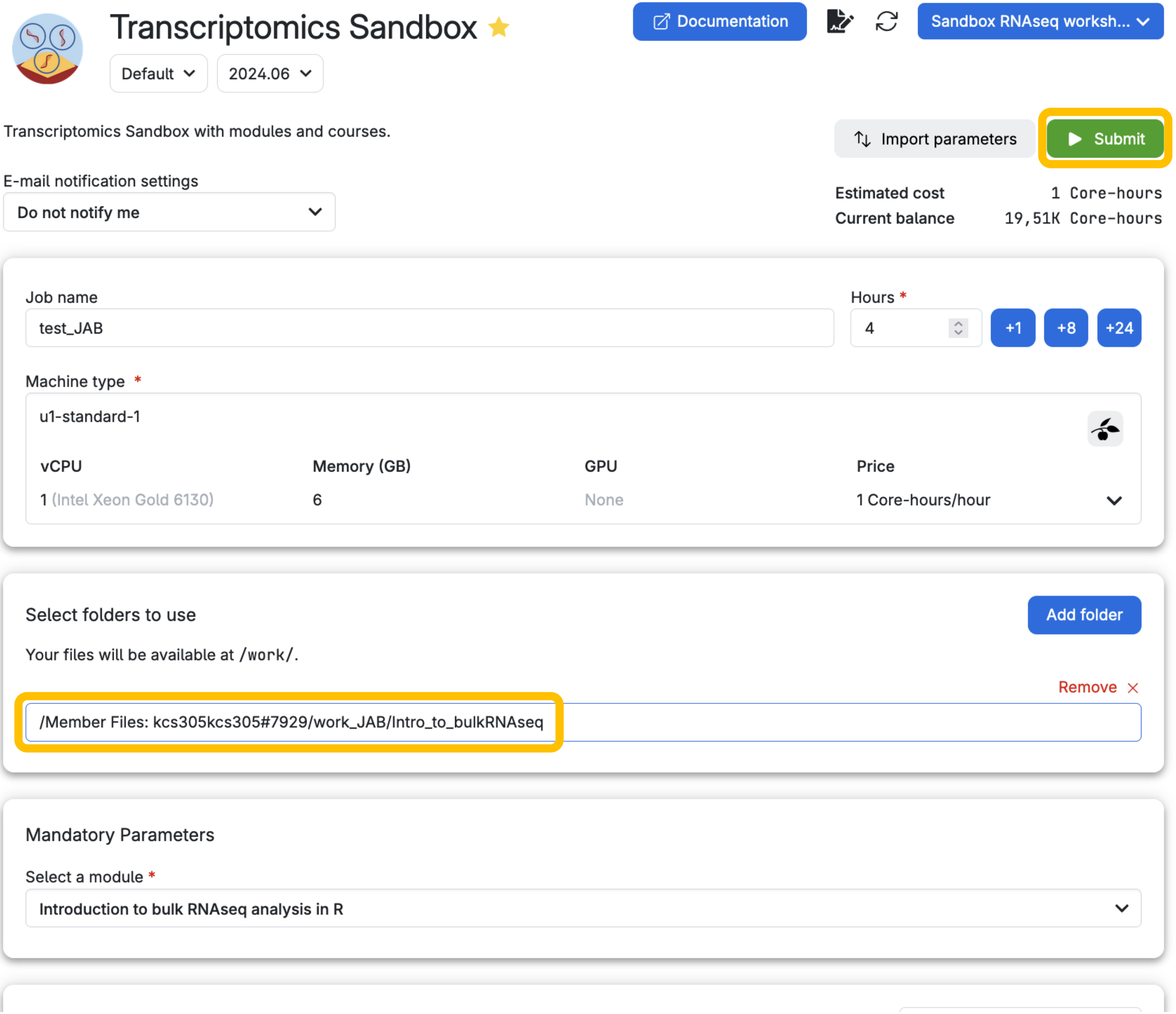


/Member Files: kcs305kcs305#…

Name		Modified at	
work_JAB	P	11:49	Copy to
work	P	11/01/2024	Copy to
Trash	P	11:38	Copy to

Submitting a job with a Sandbox app

1. App & version (dropdown menu to change it)
2. Read documentation before using it
3. Import parameters (if wanted from a previous job)
4. Job name, hours, and machine type (resources set-up)
5. Folders to access while running this particular job
6. Module to use (which includes Notebooks & Data)
7. **Review & Submit**



Transcriptomics Sandbox with modules and courses.

E-mail notification settings: Do not notify me

Estimated cost 1 Core-hours
Current balance 19,51K Core-hours

Job name: test_JAB

Hours *: 4 (+1, +8, +24)

Machine type *: u1-standard-1

vCPU	Memory (GB)	GPU	Price
1 (Intel Xeon Gold 6130)	6	None	1 Core-hours/hour

Select folders to use

Your files will be available at `/work/`.

/Member Files: kcs305kcs305#7929/work_JAB/Intro_to_bulkRNAseq

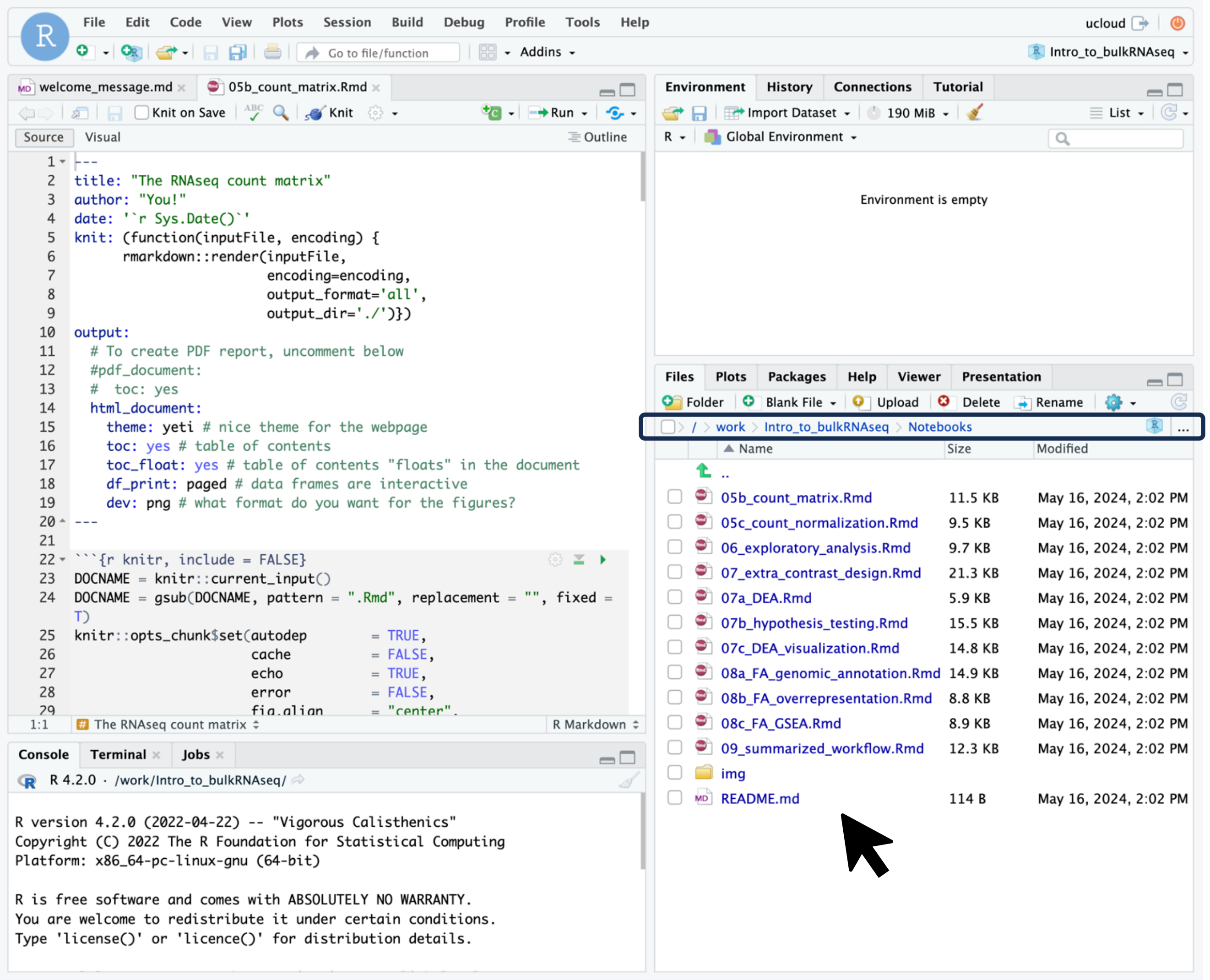
Mandatory Parameters

Select a module *: Introduction to bulk RNAseq analysis in R

Sandbox RStudio-based app

In the transcriptomics app, you will get RStudio in your web browser.

Click on the *Intro_to_bulkRNAseq* folder and navigate through the R Markdown notebooks to run the analyses.



~~~
---
title: "The RNAseq count matrix"
author: "You!"
date: '`r Sys.Date()`'
knit: (function(inputFile, encoding) {
      rmarkdown::render(inputFile,
                        encoding=encoding,
                        output_format='all',
                        output_dir='./')})
output:
  # To create PDF report, uncomment below
  #pdf_document:
  #  toc: yes
  html_document:
    theme: yeti # nice theme for the webpage
    toc: yes # table of contents
    toc_float: yes # table of contents "floats" in the document
    df_print: paged # data frames are interactive
    dev: png # what format do you want for the figures?
---

```{r knitr, include = FALSE}
DOCNAME = knitr::current_input()
DOCNAME = gsub(DOCNAME, pattern = ".Rmd", replacement = "", fixed =
T)
knitr::opts_chunk$set(autodep        = TRUE,
                      cache          = FALSE,
                      echo           = TRUE,
                      error          = FALSE,
                      fig.align      = "center",
~~~

```
R 4.2.0 · /work/Intro_to_bulkRNAseq/

R version 4.2.0 (2022-04-22) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
```

/ › work › Intro_to_bulkRNAseq › Notebooks

Name	Size	Modified
..		
05b_count_matrix.Rmd	11.5 KB	May 16, 2024, 2:02 PM
05c_count_normalization.Rmd	9.5 KB	May 16, 2024, 2:02 PM
06_exploratory_analysis.Rmd	9.7 KB	May 16, 2024, 2:02 PM
07_extra_contrast_design.Rmd	21.3 KB	May 16, 2024, 2:02 PM
07a_DEA.Rmd	5.9 KB	May 16, 2024, 2:02 PM
07b_hypothesis_testing.Rmd	15.5 KB	May 16, 2024, 2:02 PM
07c_DEA_visualization.Rmd	14.8 KB	May 16, 2024, 2:02 PM
08a_FA_genomic_annotation.Rmd	14.9 KB	May 16, 2024, 2:02 PM
08b_FA_overrepresentation.Rmd	8.8 KB	May 16, 2024, 2:02 PM
08c_FA_GSEA.Rmd	8.9 KB	May 16, 2024, 2:02 PM
09_summarized_workflow.Rmd	12.3 KB	May 16, 2024, 2:02 PM
img		
README.md	114 B	May 16, 2024, 2:02 PM

HeaDS

UCloud jobs

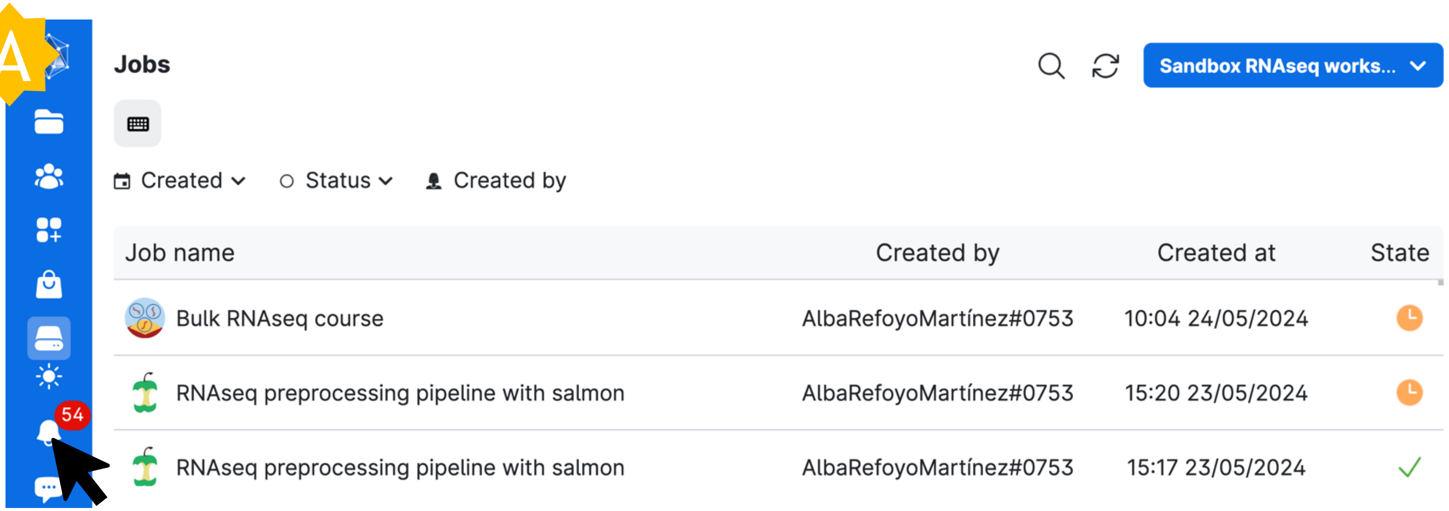
Access past and current running jobs

A Click on a job to stop it or rerun (ensuring the use of the same parameters)

or

B Access old jobs from 'Import parameters' button on job setup page

A



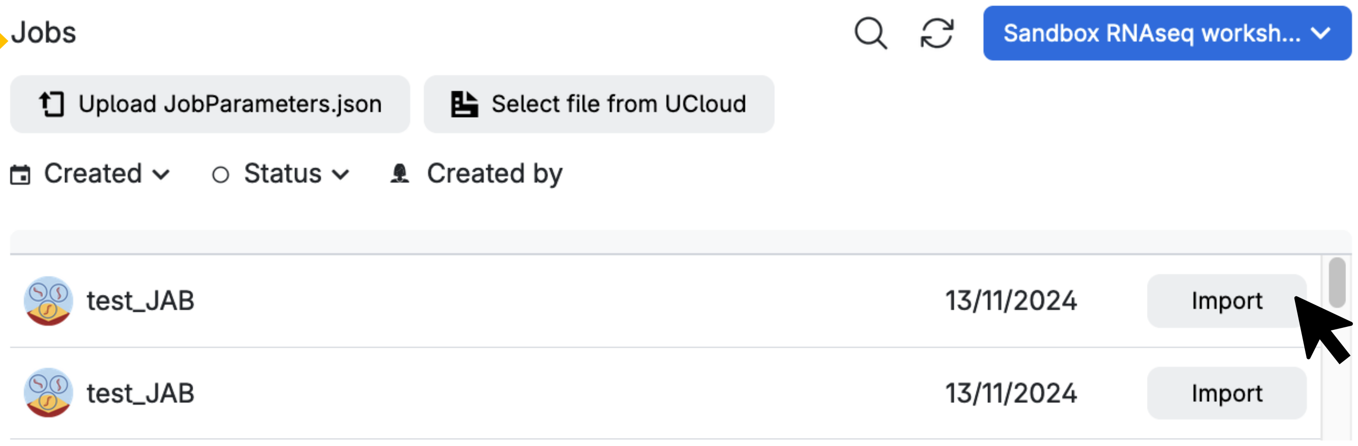
Jobs

Sandbox RNAseq works...

Created ⌄ ○ Status ⌄ Created by

Job name	Created by	Created at	State
Bulk RNAseq course	AlbaRefoyoMartínez#0753	10:04 24/05/2024	◷
RNAseq preprocessing pipeline with salmon	AlbaRefoyoMartínez#0753	15:20 23/05/2024	◷
RNAseq preprocessing pipeline with salmon	AlbaRefoyoMartínez#0753	15:17 23/05/2024	✓

B



Jobs

Sandbox RNAseq worksh...

Upload JobParameters.json

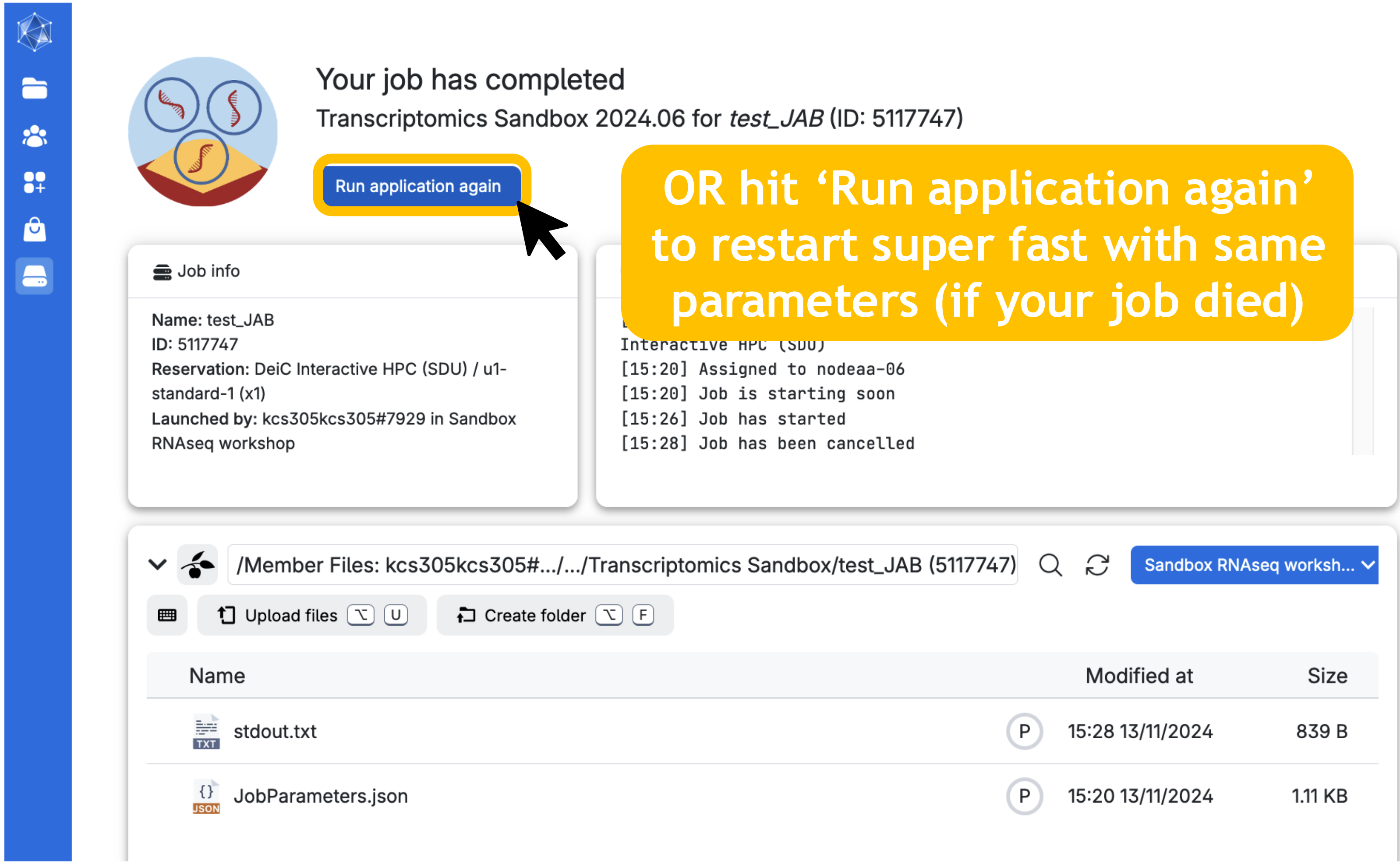
Select file from UCloud

Created ⌄ ○ Status ⌄ Created by

test_JAB	13/11/2024	Import
test_JAB	13/11/2024	Import

HeaDS

UCloud jobs



Your job has completed

Transcriptomics Sandbox 2024.06 for *test_JAB* (ID: 5117747)

Job info

Name: test_JAB
ID: 5117747
Reservation: DeiC Interactive HPC (SDU) / u1-standard-1 (x1)
Launched by: kcs305kcs305#7929 in Sandbox RNAseq workshop

```
Interactive HPC (SDU)
[15:20] Assigned to nodeaa-06
[15:20] Job is starting soon
[15:26] Job has started
[15:28] Job has been cancelled
```

/Member Files: kcs305kcs305#.../.../Transcriptomics Sandbox/test_JAB (5117747)

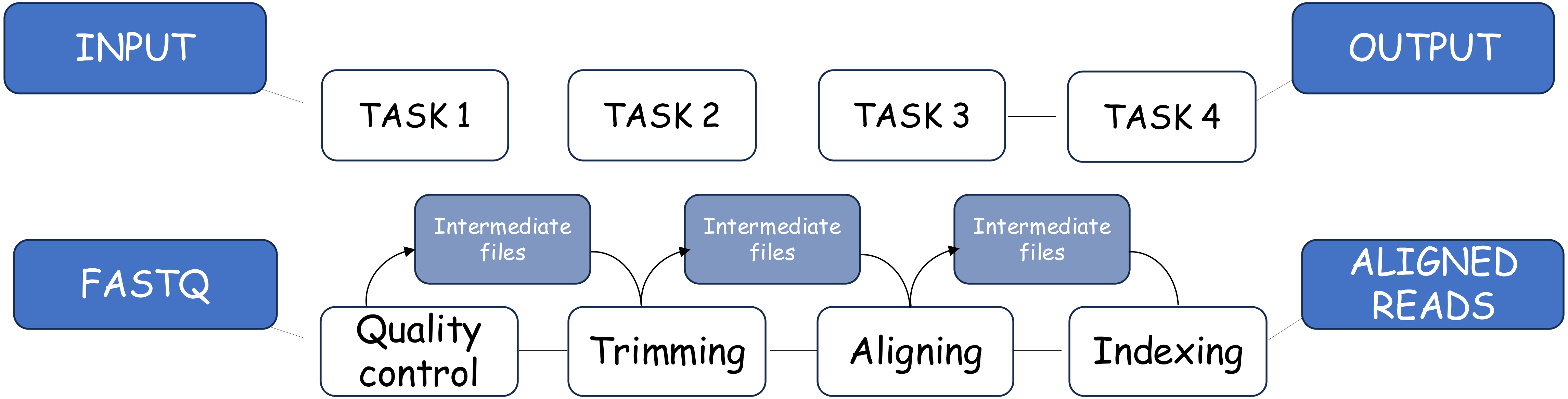
Name		Modified at	Size
stdout.txt	P	15:28 13/11/2024	839 B
JobParameters.json	P	15:20 13/11/2024	1.11 KB

OR hit ‘Run application again’ to restart super fast with same parameters (if your job died)

15 min break!

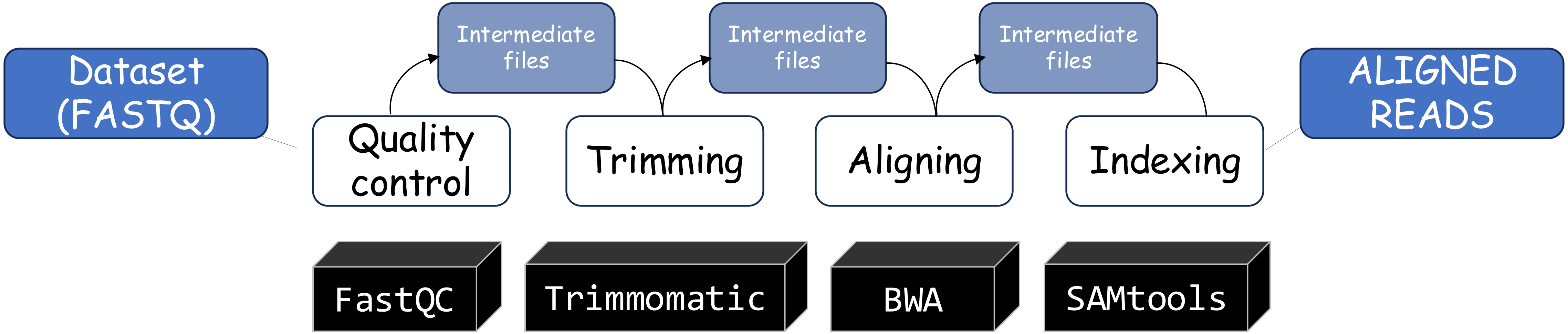
Why use pipelines/ workflows

A pipeline (workflow) is a series of programmatic steps to transform raw data into processed results, figures, and insights.



Why use pipelines/ workflows

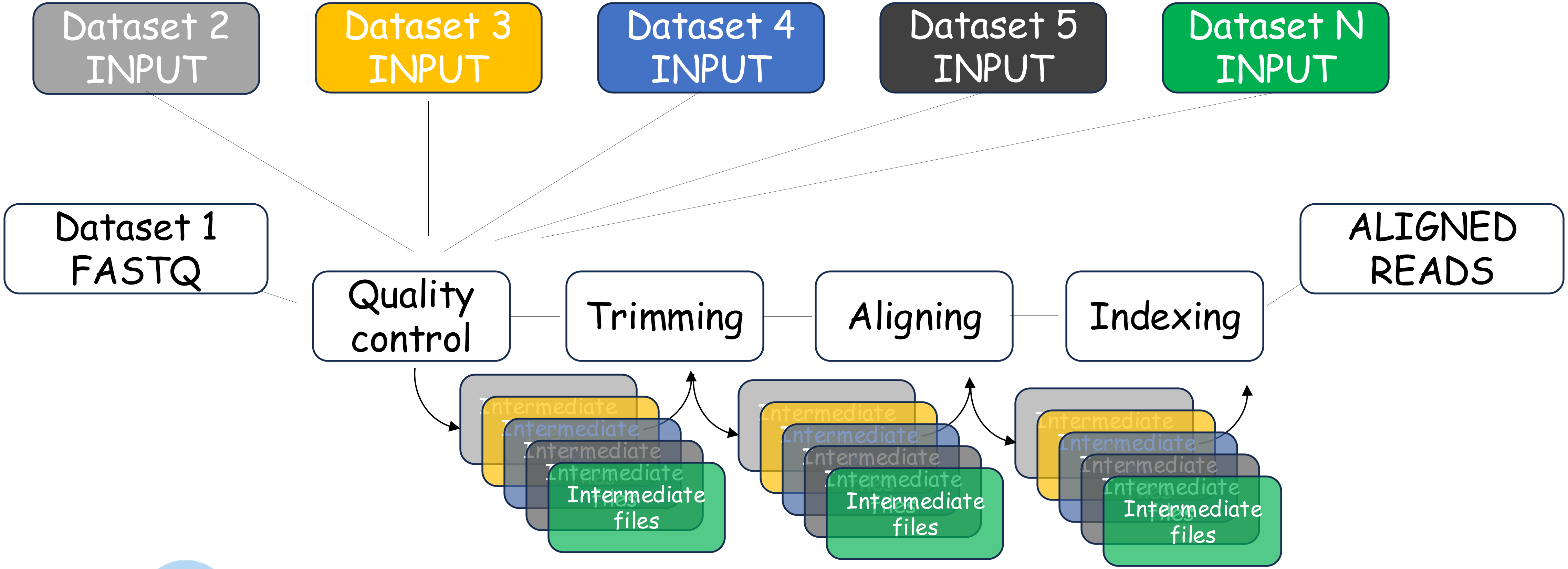
Each step involving **different tools, parameters, reference databases, and specific requirements.**



Let me do this by hand via single tool calls in the terminal…

Why use pipelines/ workflows

Now apply the same analysis to new data... 😟



Dataset 2 INPUT

Dataset 3 INPUT

Dataset 4 INPUT

Dataset 5 INPUT

Dataset N INPUT

Dataset 1 FASTQ

Quality control

Trimming

Aligning

Indexing

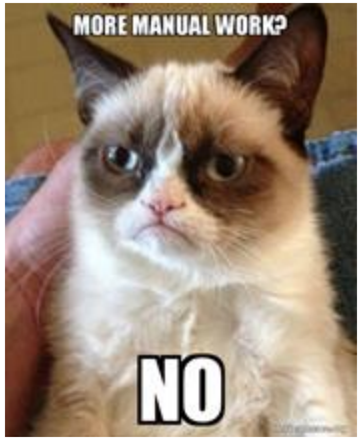
ALIGNED READS

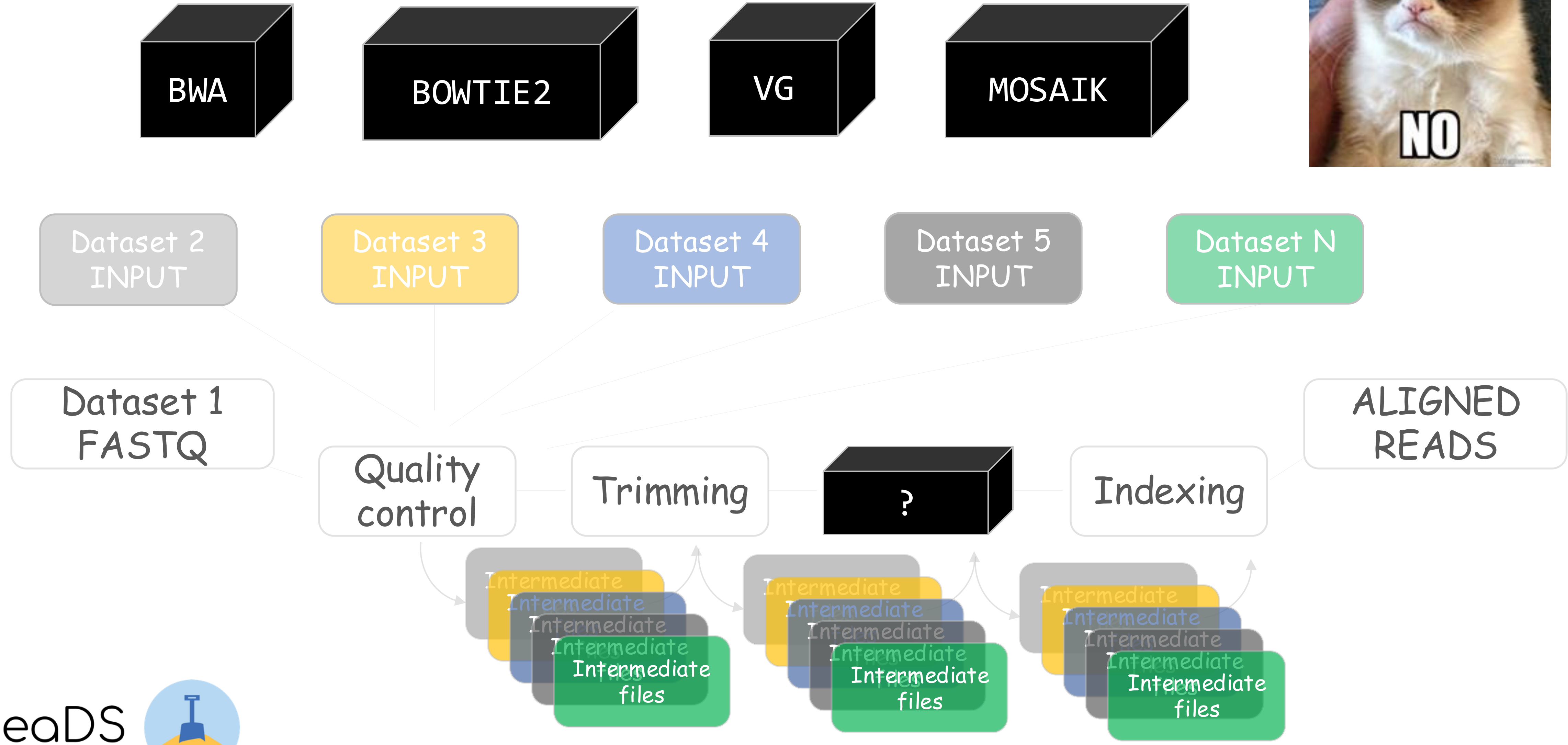
Intermediate files

Intermediate files

Intermediate files

Wait, should I use a different mapping tool? Which one?





Why use pipelines/ workflows

SOFTWARE

- **Multiple softwares** required, and sometimes even more **parameters** to tweak
- Small changes in the parameters software can cause a large difference in the results
- Differences in **program resource** needs at each step (compute power, data inputs, software dependencies, etc.)

DATA

- Many files are being generated (also intermediate files) and the size of the data files can be large
- Differences in **data type, shape and scale**

Bioinformatics workflows are complex... and reproducibility can be very challenging

Workflow managers (languages)

nextflow

Based on Groovy (Java dialect)

Pipeline order controlled by channel flow (allows diff outputs, FIFO)

Maintained by venture-funded co.

Native cloud support ($) +++ many plug-ins

Standardized portable modules

X Excessive temp dirs / symlinks

snakemake

Galaxy PROJECT

WfM languages

nextflow

Based on Groovy (Java dialect)

Pipeline order controlled by channel flow (allows diff outputs, FIFO)

Maintained by venture-funded co.

Native cloud support ($) +++ many plug-ins

Standardized portable modules

X Excessive temp dirs / symlinks

```
params.samples = 'data/samples/*.fastq'
params.outdir = 'output'
params.genome = 'data/genome.fa'

process FASTP {

    publishDir "${params.outdir}/QC", pattern: "*.html"

    input:
    tuple val(id), path(reads)

    output:
    path '*.json', emit: json
    path '*.html', emit: html

    """
    fastp -i ${reads} \ --json ${id}.fastp.json \   --html ${id}.fastp.html
    """
}


process MINIMAP2 {

    cpus 2

    input:
    tuple val(id), path(reads)
    path genome

    output:
    tuple val(id), path("*.sam")

    """
    minimap2 -t ${task.cpus} \
```

HeaDS

WfM languages

nextflow

Based on Groovy (Java dialect)

Pipeline order controlled by channel flow (allows diff outputs, FIFO)

Maintained by venture-funded co.

Native cloud support ($) +++ many plug-ins

Standardized portable modules

X Excessive temp dirs / symlinks

```
params.samples = 'data/samp
params.outdir = 'output'
params.genome = 'data/geno

process FASTP {

    publishDir "${params.outd

    input:
    tuple val(id), path(reads)

    output:
    path '*.json', emit: json
    path '*.html', emit: html

    """
    fastp -i ${reads} \ --json ${
    """
}


process MINIMAP2 {

    cpus 2

    input:
    tuple val(id), path(reads)
    path genome

    output:
    tuple val(id), path("*.sam"

    """
    minimap2 -t ${task.cpus} \
```

```
workflow {

    // Get our samples into a channel
    ch_samples = channel.fromPath(
params.samples )
        | map { [ it.simpleName, it ] }

    // Invoke fastp and put output into a
channel
    ch_fastp = ch_samples | FASTP

    ch_genome = channel.value(
file(params.genome, checkIfExists: true) )

    ch_flagstat = MINIMAP2( ch_samples,
ch_genome )
        | SAMTOOLS_VIEW
        | SAMTOOLS_FLAGSTAT


    ch_files = ch_flagstat
        | map { it[1] }
        | mix( ch_fastp.json )
        | collect


    MULTIQC ( ch_files,
file("${workflow.projectDir}/assets/multiq
c_config.yml", checkIfExists: true) )

}
```

Workflow managers

- WfMS are software tools designed to **automate** and **optimize complex data analysis workflows.**
- WfMS allow users to **define, run, and track tasks and dependencies** in a modular, scalable way, simplifying the **management and reproducibility** of analyses.

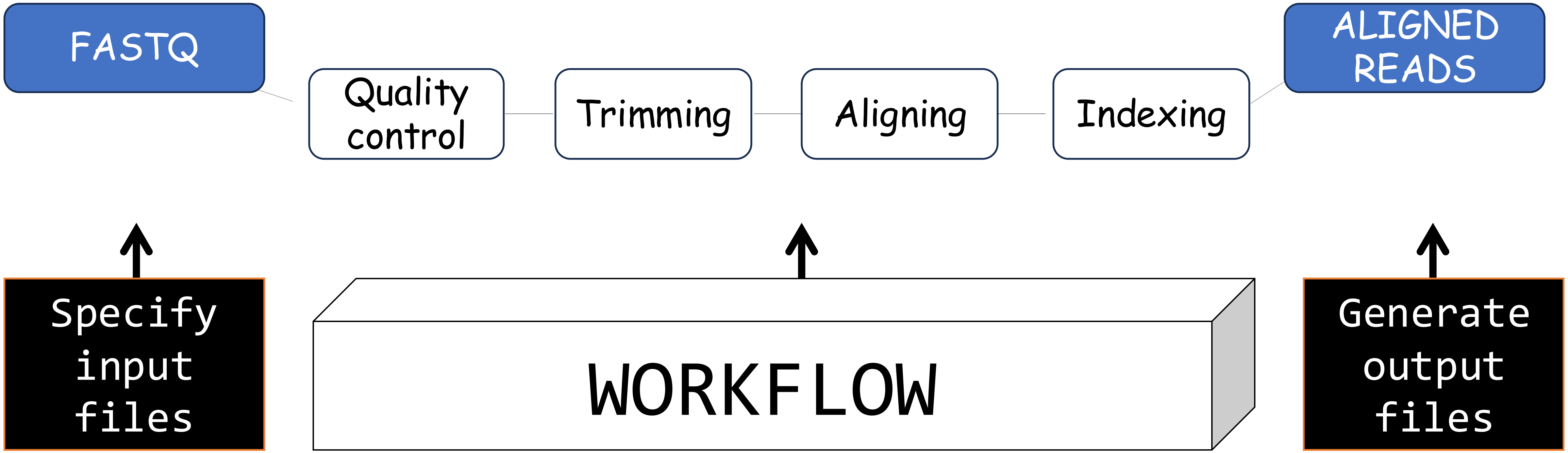
•Why do we want to use them?

- Using WfMS helps **reduce errors and inconsistencies** in analyses, while enhancing the efficiency and reproducibility of research.

Workflow managers

Execution without manual intervention

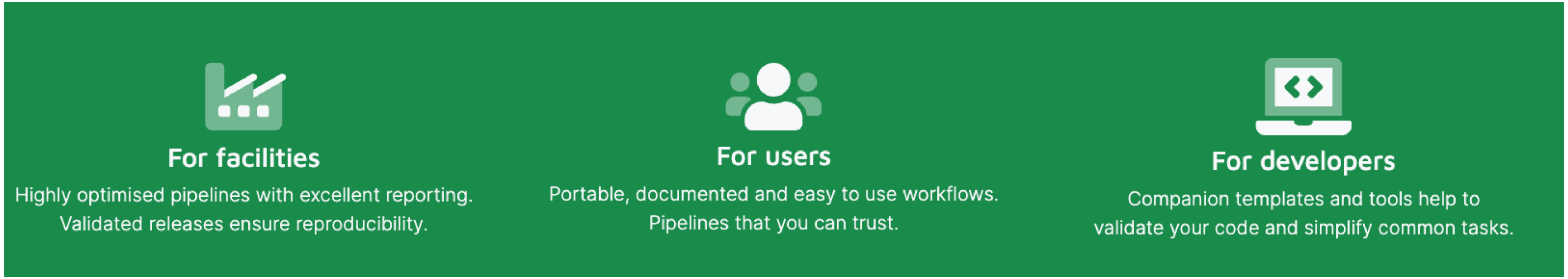
Choose the pipeline that does the work for you! (or define them yourself)



nf-core

https://nf-co.re/pipelines

A community effort to collect a curated set of analysis pipelines built using Nextflow.



For facilities

Highly optimised pipelines with excellent reporting.
Validated releases ensure reproducibility.

For users

Portable, documented and easy to use workflows.
Pipelines that you can trust.

For developers

Companion templates and tools help to
validate your code and simplify common tasks.

⇧ We will have a look at the results folder in a sec!

⇧ Big community, routinely maintained/updated.
Slack channel for questions & discussions (get helped and advised!)

⇧ Use their template!

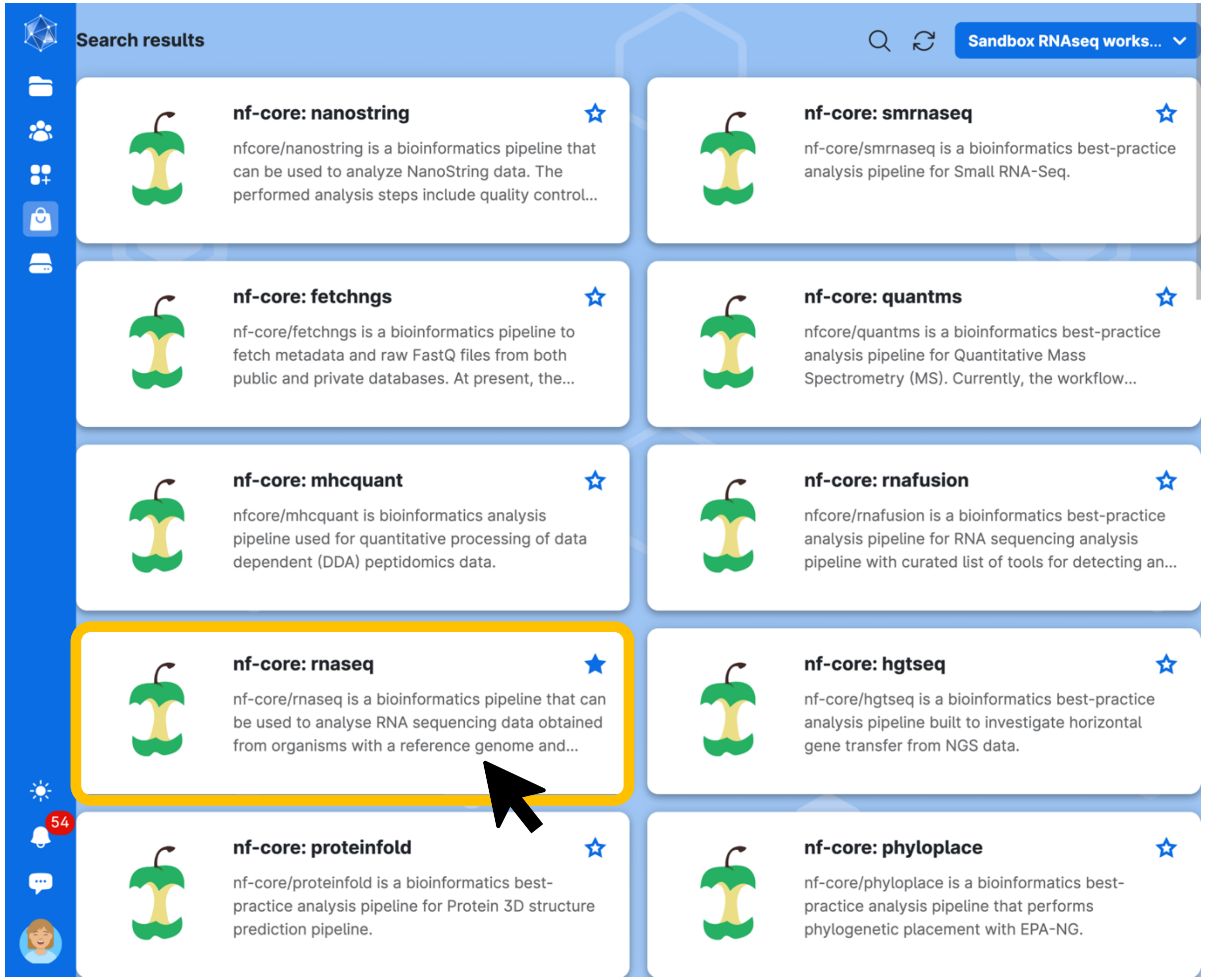
nf-core

https://nf-co.re/pipelines

- Bioinformatics community for curated pipelines written in nextflow
  - **Bulk RNAseq**
  - **Single Cell RNAseq**
  - **ATACseq**
  - **ChIPseq**
  - **HICseq**
- Completely reproducible, following gold standards (best practices) and open source
- Easy to implement (packaged software). They work on any computational infrastructures
- Very well documented
- More and more pipelines are being introduced and updated

Pipelines on UCloud

Search for nf-core apps



nf-core: nanostring
nfcore/nanostring is a bioinformatics pipeline that can be used to analyze NanoString data. The performed analysis steps include quality control...

nf-core: smrnaseq
nf-core/smrnaseq is a bioinformatics best-practice analysis pipeline for Small RNA-Seq.

nf-core: fetchngs
nf-core/fetchngs is a bioinformatics pipeline to fetch metadata and raw FastQ files from both public and private databases. At present, the...

nf-core: quantms
nfcore/quantms is a bioinformatics best-practice analysis pipeline for Quantitative Mass Spectrometry (MS). Currently, the workflow...

nf-core: mhcquant
nfcore/mhcquant is bioinformatics analysis pipeline used for quantitative processing of data dependent (DDA) peptidomics data.

nf-core: rnafusion
nfcore/rnafusion is a bioinformatics best-practice analysis pipeline for RNA sequencing analysis pipeline with curated list of tools for detecting an...

nf-core: rnaseq
nf-core/rnaseq is a bioinformatics pipeline that can be used to analyse RNA sequencing data obtained from organisms with a reference genome and...

nf-core: hgtseq
nf-core/hgtseq is a bioinformatics best-practice analysis pipeline built to investigate horizontal gene transfer from NGS data.

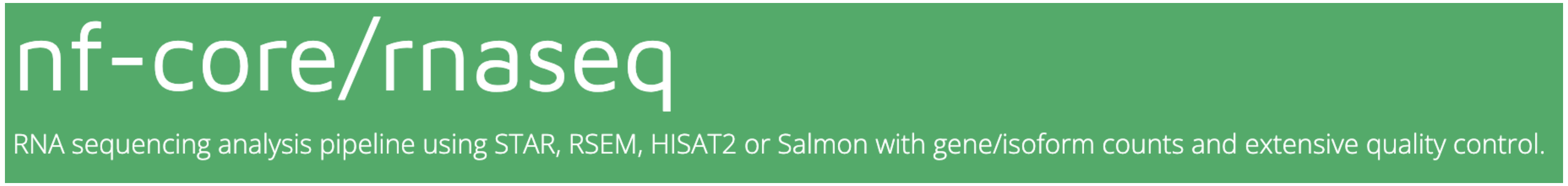
nf-core: proteinfold
nf-core/proteinfold is a bioinformatics best-practice analysis pipeline for Protein 3D structure prediction pipeline.

nf-core: phyloplace
nf-core/phyloplace is a bioinformatics best-practice analysis pipeline that performs phylogenetic placement with EPA-NG.

Your problem:

To perform quality control and quantify the expression of genes in a genome (bulk RNA sequencing)

The solution:



nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

It performs quality control (QC), trimming and (pseudo-)alignment, and produces a gene expression matrix and extensive QC report

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

All you need is a samplesheet and FASTQ files!

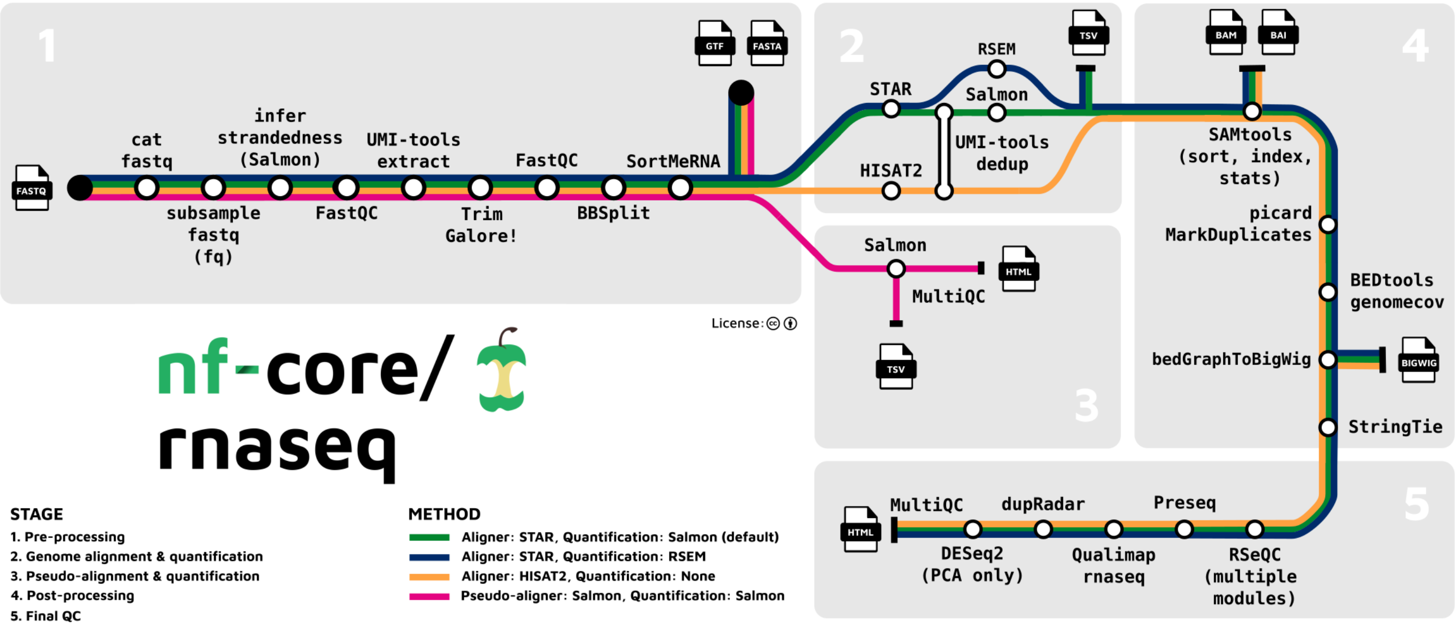
```
> cat samplesheet.csv

sample,fastq_1,fastq_2,strandedness
CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,auto
CONTROL_REP1,AEG588A1_S1_L003_R1_001.fastq.gz,AEG588A1_S1_L003_R2_001.fastq.gz,auto
CONTROL_REP1,AEG588A1_S1_L004_R1_001.fastq.gz,AEG588A1_S1_L004_R2_001.fastq.gz,auto
```

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Check their website!

All you need is a samplesheet and FASTQ files!



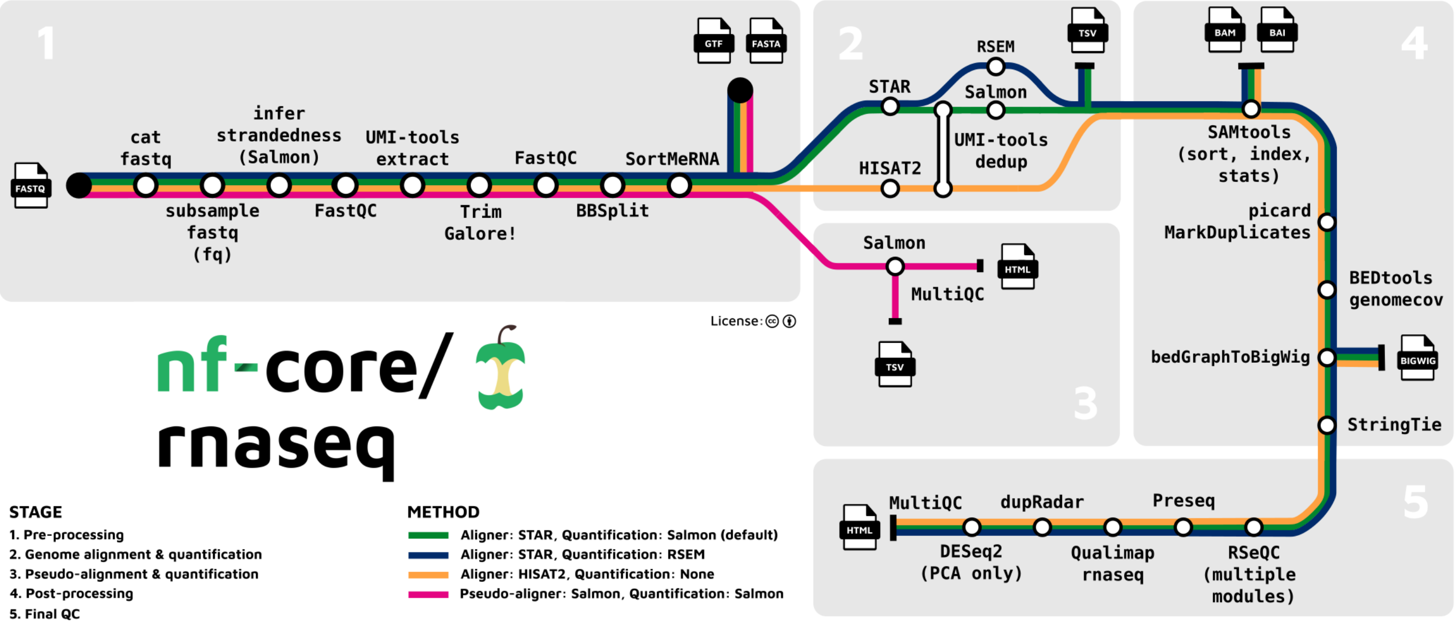
1. Merge re-sequenced FastQ files (cat)
2. Sub-sample FastQ files and auto-infer strandedness (fq, Salmon)
3. Read QC (FastQC)
4. UMI extraction (UMI-tools)
5. Adapter and quality trimming (Trim Galore!)
6. Removal of genome contaminants (BBSplit)
7. Removal of ribosomal RNA (SortMeRNA)
8. Choice of multiple alignment and quantification routes:
    1. STAR -> Salmon
    2. STAR -> RSEM
    3. HiSAT2 -> **NO QUANTIFICATION**
9. Sort and index alignments (SAMtools)
10. UMI-based deduplication (UMI-tools)
11. Duplicate read marking (picard MarkDuplicates)
12. Transcript assembly and quantification (StringTie)
13. Create bigWig coverage files (BEDTools, bedGraphToBigWig)
14. Extensive quality control:
    1. RSeQC
    2. Qualimap
    3. dupRadar
    4. Preseq
    5. DESeq2
15. Pseudoalignment and quantification (Salmon or 'Kallisto'; *optional*)
16. Present QC for raw read, alignment, gene biotype, sample similarity, and strand-specificity checks (MultiQC, R)

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

The nf-core/rnaseq pipeline will perform:

1. **Pre-processing**: trim and clean your reads
2. **Alignment and quantification of expression levels**
   - Alignment **post-processing**: mark duplicates
3. **Pseudoalignment and quantification**: align your reads and create count matrix
4. Postprocessing
   - Create BigWigs (coverage files)
5. **QC (for all the steps)**



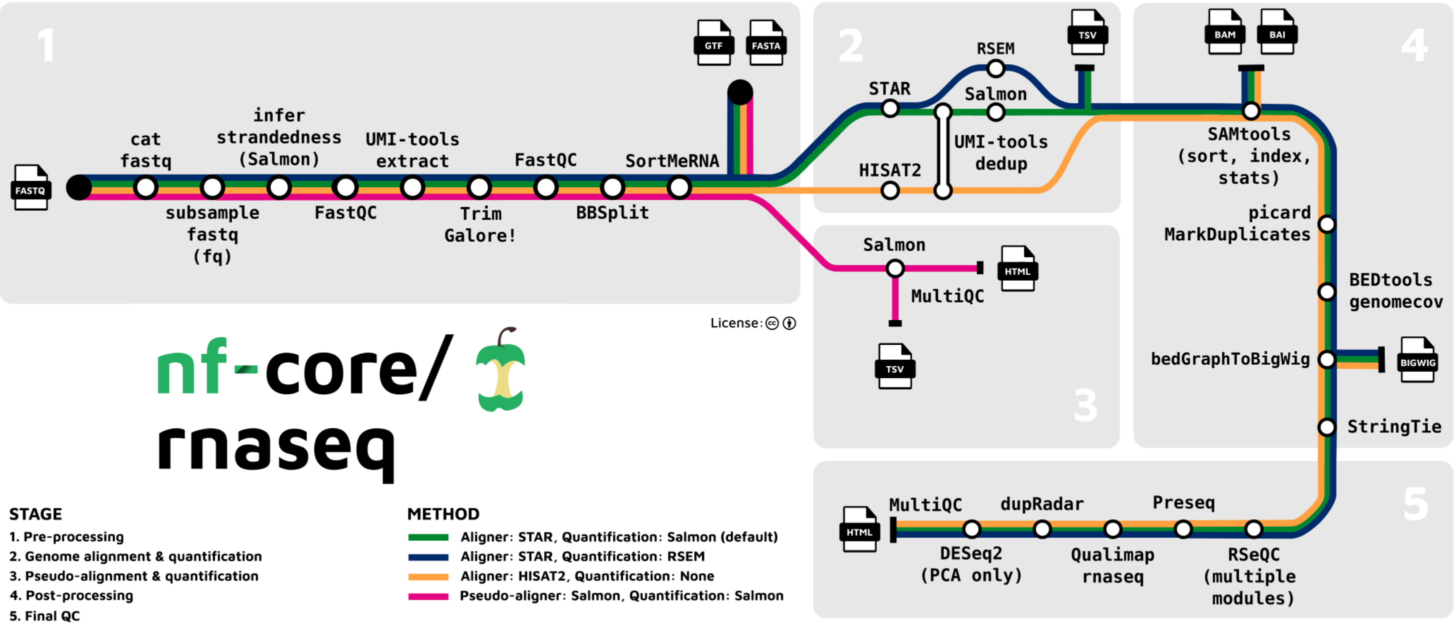
nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

This takes some time, so we have run the pipeline for you!!

You have looked at some of the output files already

1. **Pre-processing**: trim and clean your reads
2. **Alignment and quantification of expression levels**
   - Alignment **post-processing**: mark duplicates
3. **Pseudoalignment and quantification**: align your reads and create count matrix
4. Postprocessing
   - Create BigWigs (coverage files)
5. **QC (for all the steps)**



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟪 Pseudo-aligner: Salmon, Quantification: Salmon

Pipelines and workflows

The pipeline generates:

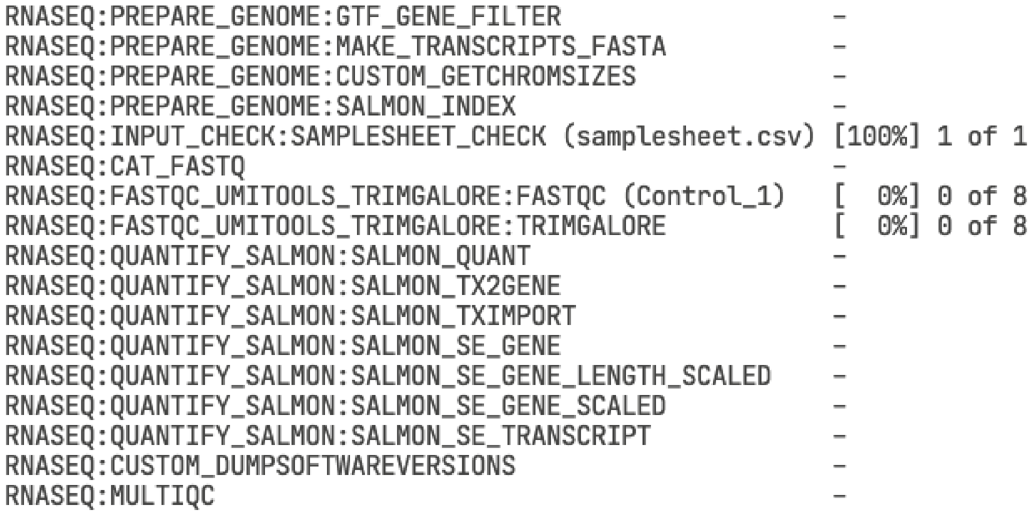
- Results (e.g. count matrix, QC plots…)
- Other output files and reports:
  - Checkpoints, logs and progress

⇩

Why workflow reporting files?

- troubleshoot errors with the running
- For every launch commands -> run times and resource usage.

nextflow



```
RNASEQ:PREPARE_GENOME:GTF_GENE_FILTER                          -
RNASEQ:PREPARE_GENOME:MAKE_TRANSCRIPTS_FASTA                   -
RNASEQ:PREPARE_GENOME:CUSTOM_GETCHROMSIZES                     -
RNASEQ:PREPARE_GENOME:SALMON_INDEX                             -
RNASEQ:INPUT_CHECK:SAMPLESHEET_CHECK (samplesheet.csv) [100%] 1 of 1
RNASEQ:CAT_FASTQ                                               -
RNASEQ:FASTQC_UMITOOLS_TRIMGALORE:FASTQC (Control_1)   [  0%] 0 of 8
RNASEQ:FASTQC_UMITOOLS_TRIMGALORE:TRIMGALORE           [  0%] 0 of 8
RNASEQ:QUANTIFY_SALMON:SALMON_QUANT                            -
RNASEQ:QUANTIFY_SALMON:SALMON_TX2GENE                          -
RNASEQ:QUANTIFY_SALMON:SALMON_TXIMPORT                         -
RNASEQ:QUANTIFY_SALMON:SALMON_SE_GENE                          -
RNASEQ:QUANTIFY_SALMON:SALMON_SE_GENE_LENGTH_SCALED            -
RNASEQ:QUANTIFY_SALMON:SALMON_SE_GENE_SCALED                   -
RNASEQ:QUANTIFY_SALMON:SALMON_SE_TRANSCRIPT                    -
RNASEQ:CUSTOM_DUMPSOFTWAREVERSIONS                             -
RNASEQ:MULTIQC                                                 -
```

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

All pipelines are run like this (test example):

```
nextflow run nf-core/rnaseq -r 3.17.0 -profile test --outdir <OUTDIR>
```

Arguments from **Nextflow** start with `-`
-r: pipeline version
-profile: docker, conda, etc
-resume: restart failed job

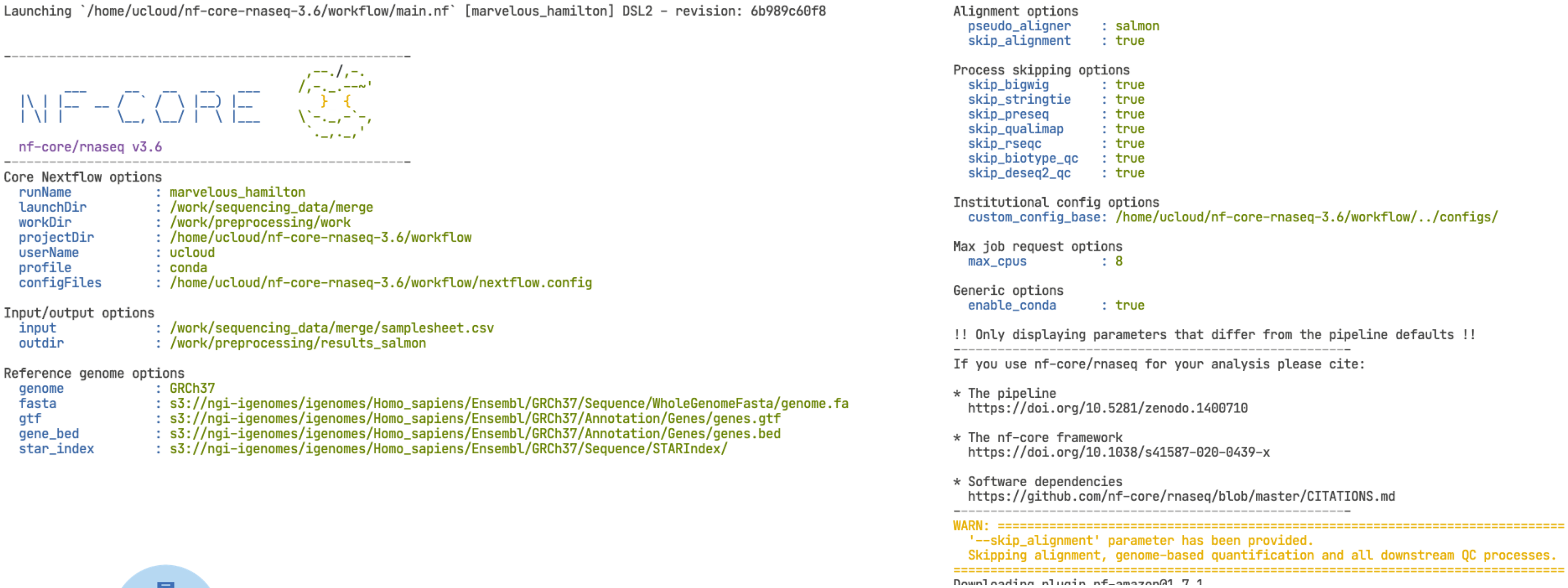
Arguments from **nf-core** start with `--`
--input: csv file with sample and read metadata
--outdir: results folder
--genome: reference genome to use
--aligner: select aligner
--skip_<X>: skip X process from pipeline

https://nf-co.re/rnaseq/3.17.0/parameters

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

After starting the workflow, you will see this:



```
Launching `/home/ucloud/nf-core-rnaseq-3.6/workflow/main.nf` [marvelous_hamilton] DSL2 - revision: 6b989c60f8


------------------------------------------------------
                                        ,--./,-.
        ___     __   __   __   ___     /,-._.--~'
  |\ | |__  __ /  ` /  \ |__) |__         }  {
  | \| |       \__, \__/ |  \ |___     \`-._,-`-,
                                        `._,._,'
  nf-core/rnaseq v3.6
------------------------------------------------------
Core Nextflow options
  runName                   : marvelous_hamilton
  launchDir                 : /work/sequencing_data/merge
  workDir                   : /work/preprocessing/work
  projectDir                : /home/ucloud/nf-core-rnaseq-3.6/workflow
  userName                  : ucloud
  profile                   : conda
  configFiles               : /home/ucloud/nf-core-rnaseq-3.6/workflow/nextflow.config

Input/output options
  input                     : /work/sequencing_data/merge/samplesheet.csv
  outdir                    : /work/preprocessing/results_salmon

Reference genome options
  genome                    : GRCh37
  fasta                     : s3://ngi-igenomes/igenomes/Homo_sapiens/Ensembl/GRCh37/Sequence/WholeGenomeFasta/genome.fa
  gtf                       : s3://ngi-igenomes/igenomes/Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.gtf
  gene_bed                  : s3://ngi-igenomes/igenomes/Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.bed
  star_index                : s3://ngi-igenomes/igenomes/Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex/
```

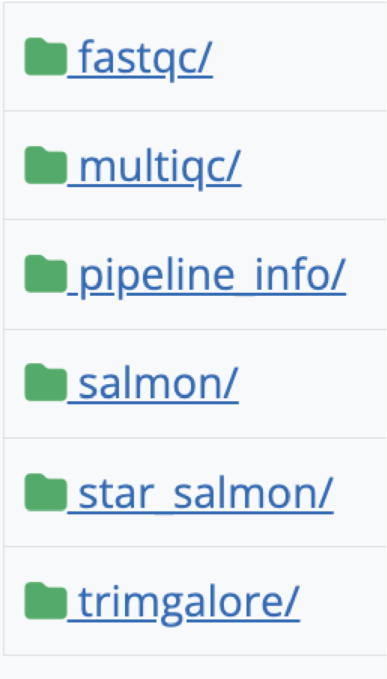
```
Alignment options
  pseudo_aligner     : salmon
  skip_alignment     : true

Process skipping options
  skip_bigwig        : true
  skip_stringtie     : true
  skip_preseq        : true
  skip_qualimap      : true
  skip_rseqc         : true
  skip_biotype_qc    : true
  skip_deseq2_qc     : true

Institutional config options
  custom_config_base: /home/ucloud/nf-core-rnaseq-3.6/workflow/../configs/

Max job request options
  max_cpus           : 8

Generic options
  enable_conda       : true

!! Only displaying parameters that differ from the pipeline defaults !!
------------------------------------------------------
If you use nf-core/rnaseq for your analysis please cite:

* The pipeline
  https://doi.org/10.5281/zenodo.1400710

* The nf-core framework
  https://doi.org/10.1038/s41587-020-0439-x

* Software dependencies
  https://github.com/nf-core/rnaseq/blob/master/CITATIONS.md
------------------------------------------------------
WARN: ===================================================================================
  '--skip_alignment' parameter has been provided.
  Skipping alignment, genome-based quantification and all downstream QC processes.
=====================================================================================
Downloading plugin nf-amazon@1.7.1
```

nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

https://nf-co.re/rnaseq/3.17.0/docs/output/

Let’s check out the results folder:

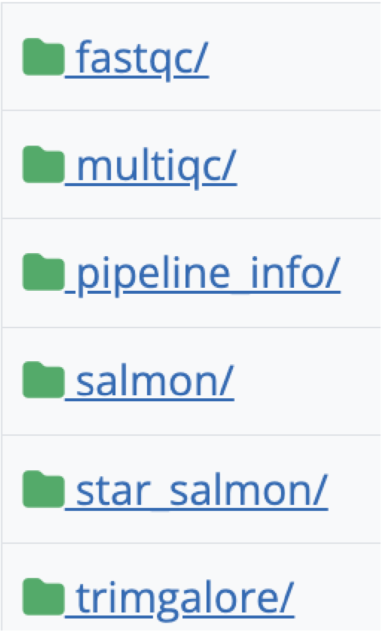


fastqc/	Raw reads quality control
multiqc/	Full quality control report
pipeline_info/	Pipeline information
salmon/	Results from salmon pseudoaligner
star_salmon/	Results from STAR aligner and quantification with salmon
trimgalore/	Trimming and cleaning of reads + fastqc

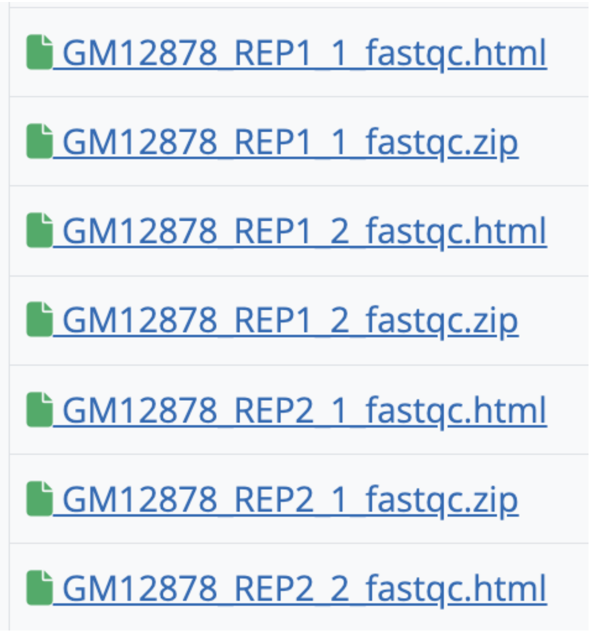
nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Let’s check out the results folder: **fastqc**



- fastqc/
- multiqc/
- pipeline_info/
- salmon/
- star_salmon/
- trimgalore/



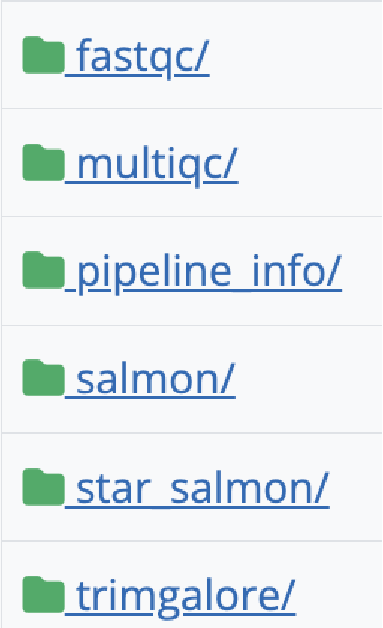
- GM12878_REP1_1_fastqc.html
- GM12878_REP1_1_fastqc.zip
- GM12878_REP1_2_fastqc.html
- GM12878_REP1_2_fastqc.zip
- GM12878_REP2_1_fastqc.html
- GM12878_REP2_1_fastqc.zip
- GM12878_REP2_2_fastqc.html

- Individual fastqc reports for raw reads
- Also in zip form

nf-core/rnaseq

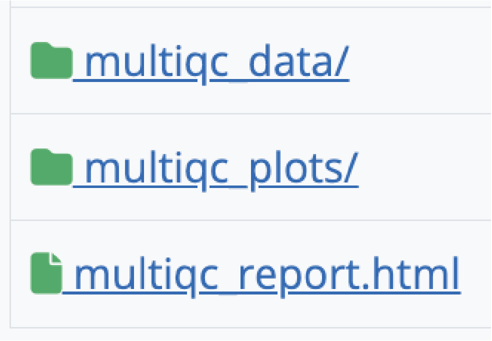
RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Let’s check out the results folder: **multiqc**



- fastqc/
- multiqc/
- pipeline_info/
- salmon/
- star_salmon/
- trimgalore/

→



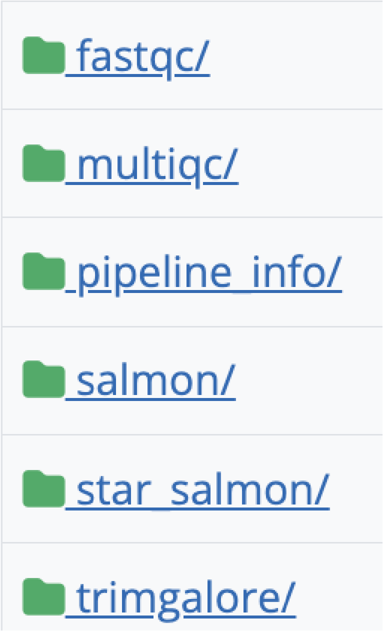
multiqc_data/	Multiqc report data
multiqc_plots/	Multiqc individual plots
multiqc_report.html	Full MultiQC report

HeaDS

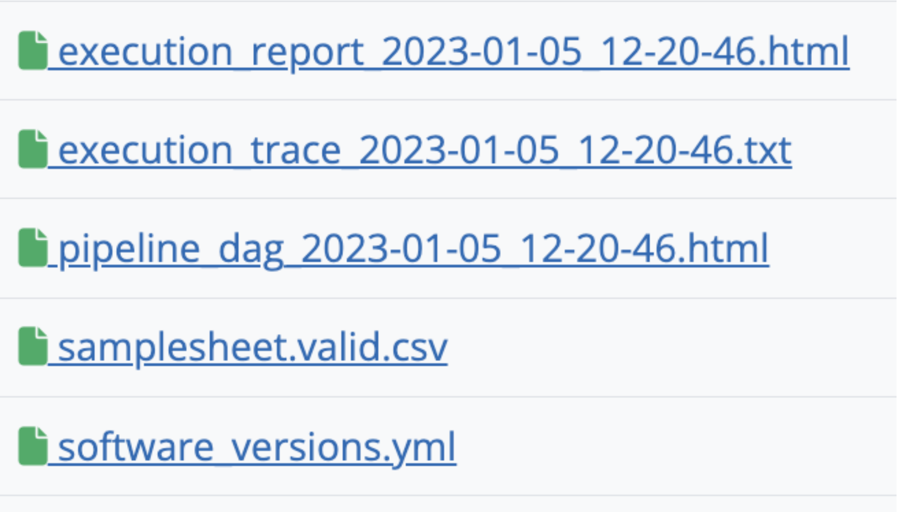
nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Let’s check out the results folder: **pipeline_info**



- fastqc/
- multiqc/
- pipeline_info/
- salmon/
- star_salmon/
- trimgalore/



- execution_report_2023-01-05_12-20-46.html
- execution_trace_2023-01-05_12-20-46.txt
- pipeline_dag_2023-01-05_12-20-46.html
- samplesheet.valid.csv
- software_versions.yml

Execution timeline (for each task!)

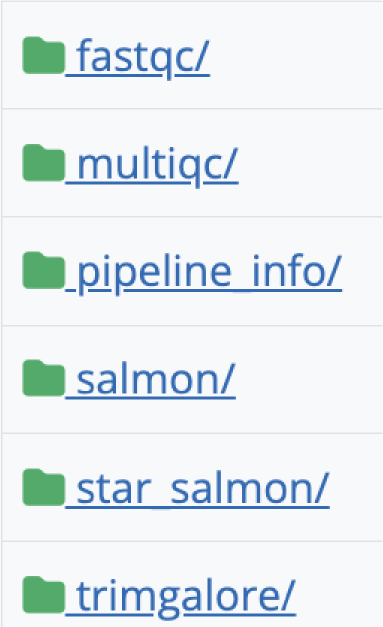
Pipeline diagram

Saved for reproducibility

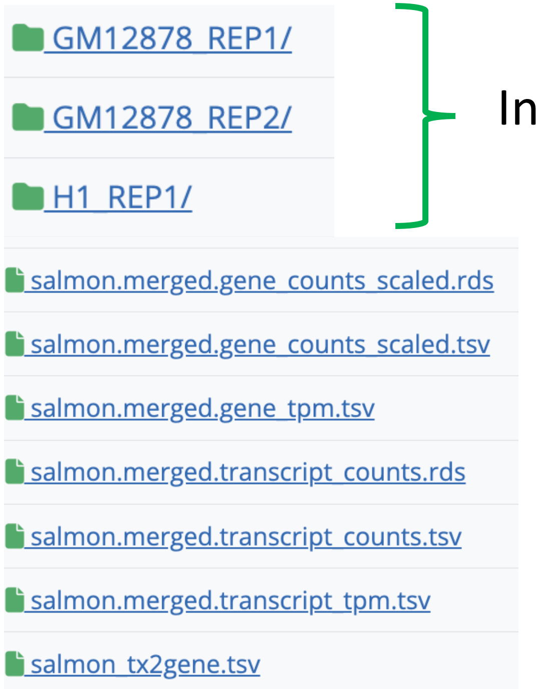
nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Let’s check out the results folder: **salmon**



- fastqc/
- multiqc/
- pipeline_info/
- salmon/
- star_salmon/
- trimgalore/



- GM12878_REP1/
- GM12878_REP2/
- H1_REP1/
- salmon.merged.gene_counts_scaled.rds
- salmon.merged.gene_counts_scaled.tsv
- salmon.merged.gene_tpm.tsv
- salmon.merged.transcript_counts.rds
- salmon.merged.transcript_counts.tsv
- salmon.merged.transcript_tpm.tsv
- salmon_tx2gene.tsv

Individual salmon results

All samples together. Different formats. Genes' quantification

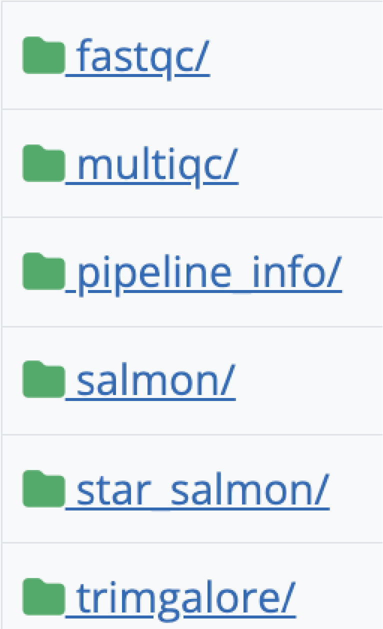
All samples together. Different formats. Transcripts' quantification

Transcript to gene table

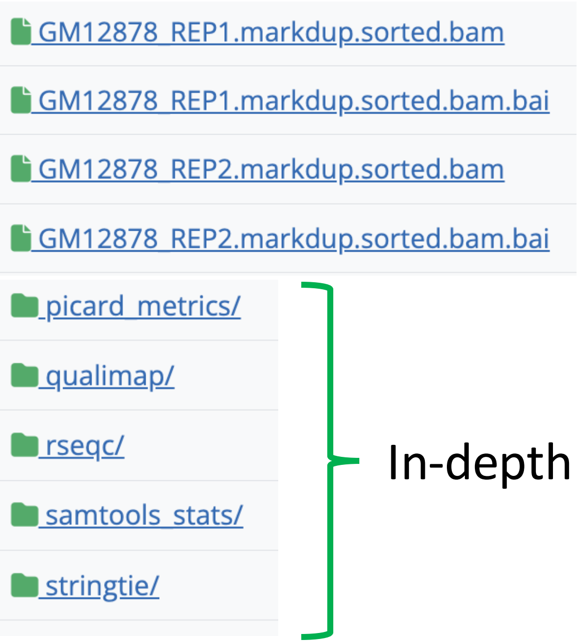
nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Let’s check out the results folder: **star_salmon**



- fastqc/
- multiqc/
- pipeline_info/
- salmon/
- star_salmon/
- trimgalore/



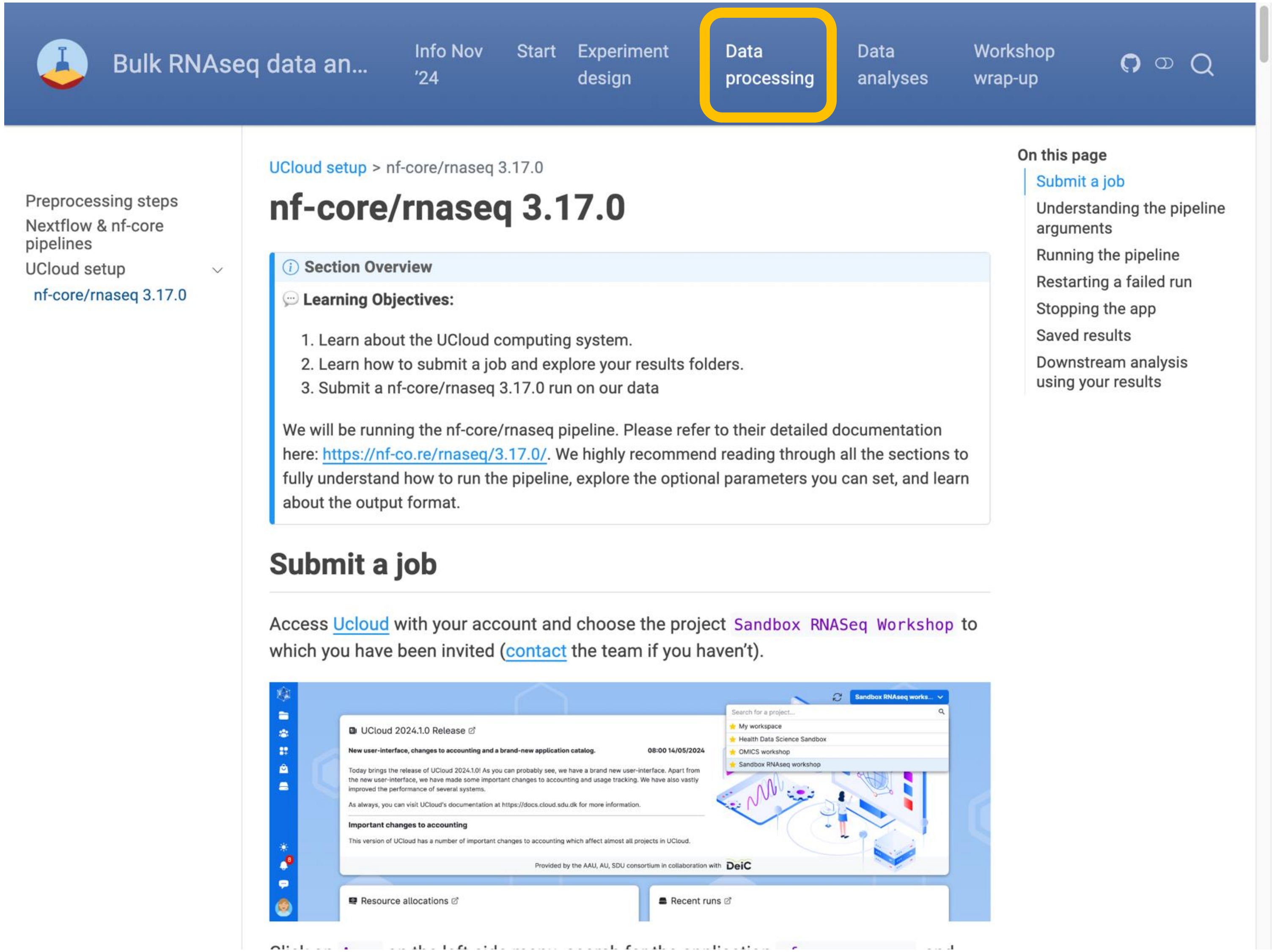
- GM12878_REP1.markdup.sorted.bam
- GM12878_REP1.markdup.sorted.bam.bai
- GM12878_REP2.markdup.sorted.bam
- GM12878_REP2.markdup.sorted.bam.bai
- picard_metrics/
- qualimap/
- rseqc/
- samtools_stats/
- stringtie/

Individual mapped reads

In-depth QC tools

star_salmon also contains same results as salmon

Run nf-core on UCloud



UCloud setup > nf-core/rnaseq 3.17.0

nf-core/rnaseq 3.17.0

Section Overview

Learning Objectives:

1. Learn about the UCloud computing system.
2. Learn how to submit a job and explore your results folders.
3. Submit a nf-core/rnaseq 3.17.0 run on our data

We will be running the nf-core/rnaseq pipeline. Please refer to their detailed documentation here: https://nf-co.re/rnaseq/3.17.0/. We highly recommend reading through all the sections to fully understand how to run the pipeline, explore the optional parameters you can set, and learn about the output format.

On this page

- Submit a job
- Understanding the pipeline arguments
- Running the pipeline
- Restarting a failed run
- Stopping the app
- Saved results
- Downstream analysis using your results

Submit a job

Access Ucloud with your account and choose the project `Sandbox RNASeq Workshop` to which you have been invited (contact the team if you haven't).

UCloud 2024.1.0 Release

New user-interface, changes to accounting and a brand-new application catalog. 08:00 14/05/2024

Today brings the release of UCloud 2024.1.0! As you can probably see, we have a brand new user-interface. Apart from the new user-interface, we have made some important changes to accounting and usage tracking. We have also vastly improved the performance of several systems.

As always, you can visit UCloud's documentation at https://docs.cloud.sdu.dk for more information.

Important changes to accounting

This version of UCloud has a number of important changes to accounting which affect almost all projects in UCloud.

Provided by the AAU, AU, SDU consortium in collaboration with DeiC

UCloud Files

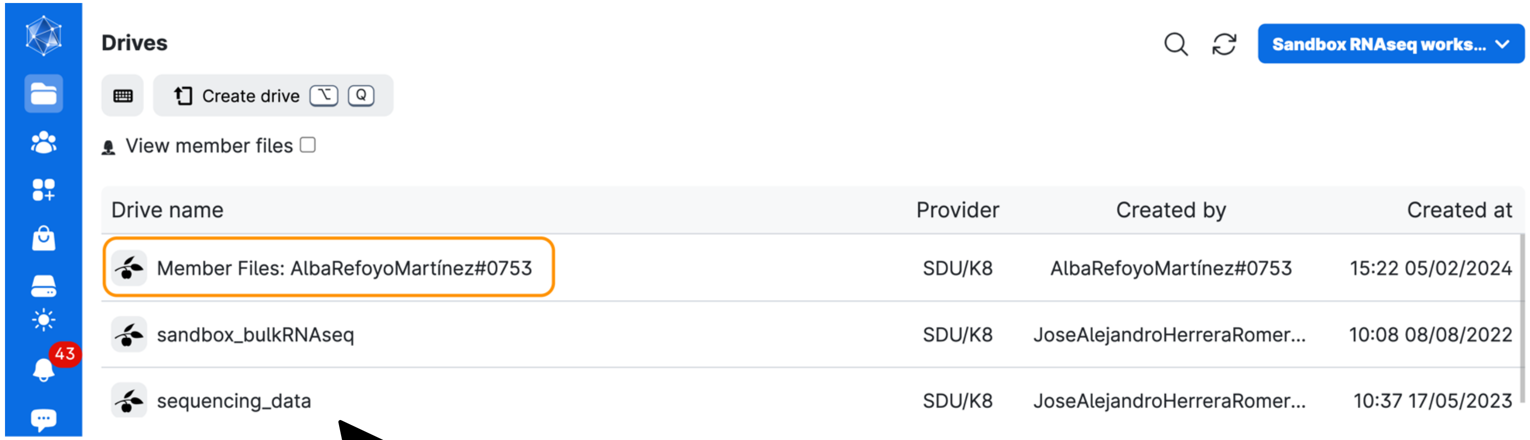
Salmon multiqc report

In **Files:**

- *Member Files:username:* your personal space
  - Work results will be here

  `Path: Member Files:username/nf-core rnaseq/<runName>/results_salmon`
- *sandbox_bulkRNAseq*: contains some course material for teachers
- *sequencing_data*: contains fastq files for preprocessing **(Don’t try to modify… write-protected!)**

  `Path: sequencing_data/preprocessing_results_salmon/results_salmon`



Drives

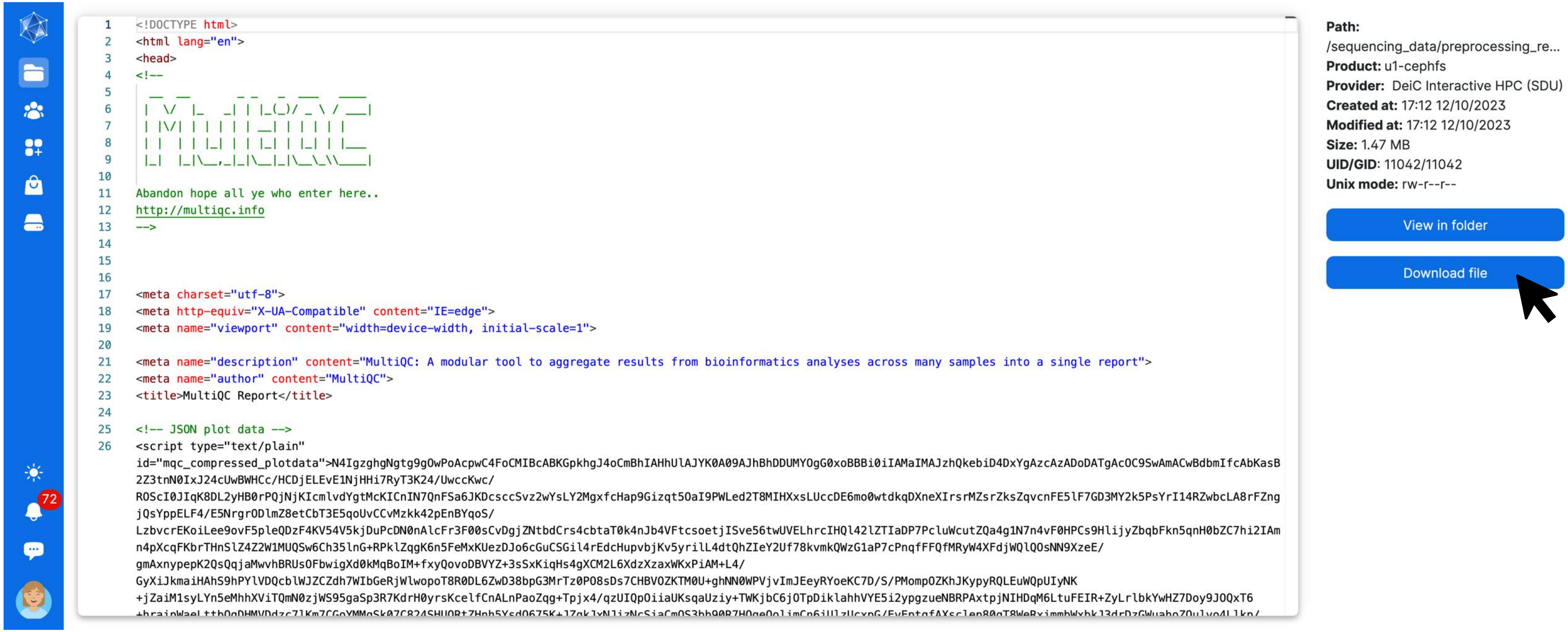
Sandbox RNAseq works...

Create drive ⌥ Q

View member files ☐

Drive name	Provider	Created by	Created at
Member Files: AlbaRefoyoMartínez#0753	SDU/K8	AlbaRefoyoMartínez#0753	15:22 05/02/2024
sandbox_bulkRNAseq	SDU/K8	JoseAlejandroHerreraRomer...	10:08 08/08/2022
sequencing_data	SDU/K8	JoseAlejandroHerreraRomer...	10:37 17/05/2023

sequencing_data (778339) > preprocessing_results_salmon > results_salmon/multiqc > multiqc_report.html

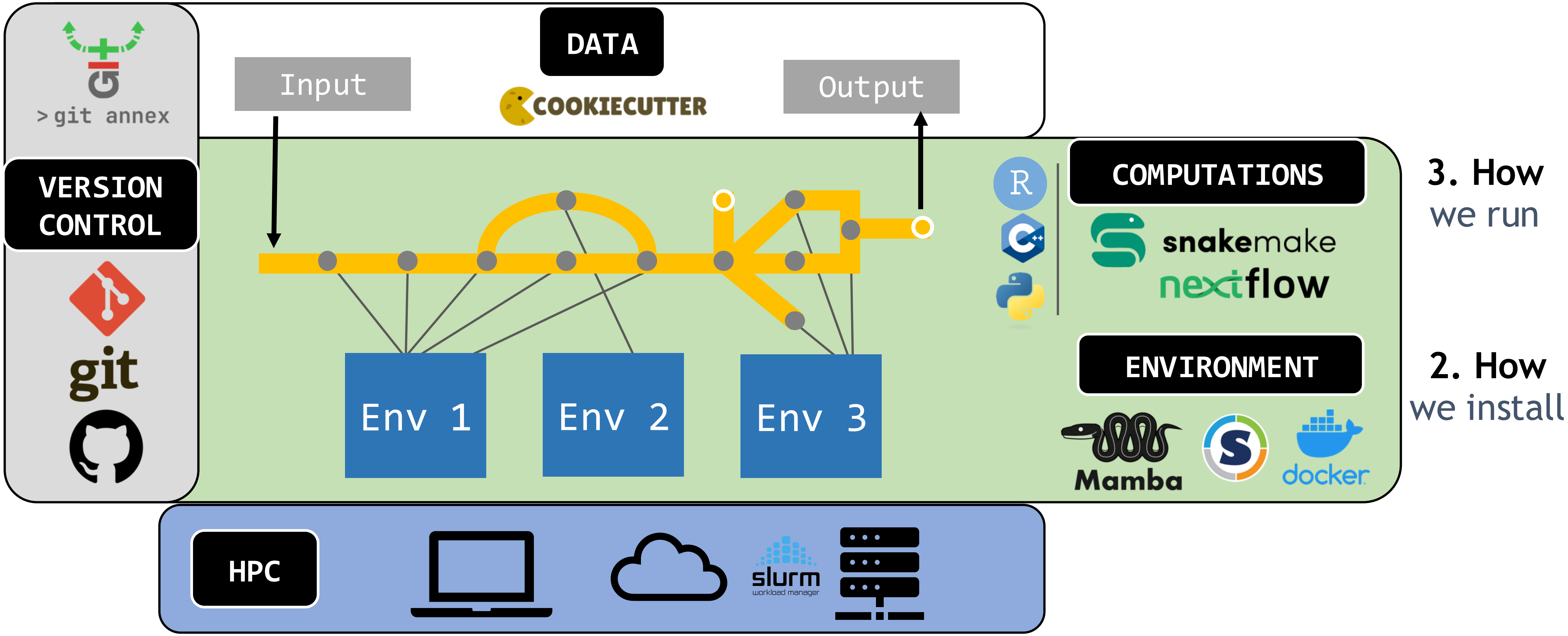


```
<!DOCTYPE html>
<html lang="en">
<head>
<!--
 __  __       _ _   _  ___   ____
|  \/  |_   _| | |_(_)/ _ \ / ___|
| |\/| | | | | | __| | | | | |
| |  | | |_| | | |_| | |_| | |___
|_|  |_|\__,_|_|\__|_|\__\_\\____|

Abandon hope all ye who enter here..
http://multiqc.info
-->



<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="viewport" content="width=device-width, initial-scale=1">

<meta name="description" content="MultiQC: A modular tool to aggregate results from bioinformatics analyses across many samples into a single report">
<meta name="author" content="MultiQC">
<title>MultiQC Report</title>

<!-- JSON plot data -->
<script type="text/plain"
id="mqc_compressed_plotdata">
```

N4IgzghgNgtg9gOwPoAcpwC4FoCMIBcABKGpkhgJ4oCmBhIAHhUlAJYK0A09AJhBhDDUMYOgG0xoBBBi0iIAMaIMAJzhQkebiD4DxYgAzcAzADoDATgAcOC9SwAmACwBdbmIfcAbKasB
2Z3tnN0IxJ24cUwBWHCc/HCDjELEvE1NjHHi7RyT3K24/UwccKwc/
ROScI0JIqK8DL2yHB0rPQjNjKIcmlvdYgtMcKICnIN7QnFSa6JKDcsccSvz2wYsLY2MgxfcHap9Gizqt5OaI9PWLed2T8MIHXxsLUccDE6mo0wtdkqDXneXIrsrMZsrZksZqvcnFE5lF7GD3MY2k5PsYrI14RZwbcLA8rFZng
jQsYppELF4/E5NrgrODlmZ8etCbT3E5qoUvCCvMzkk42pEnBYqoS/
LzbvcrEKoiLee9ovF5pleQDzF4KV54V5kjDuPcDN0nAlcFr3F00sCvDgjZNtbdCrs4cbtaT0k4nJb4VFtcsoetjISve56twUVELhrcIHQl42lZTIaDP7PcluWcutZQa4g1N7n4vF0HPCs9HlijyZbqbFkn5qnH0bZC7hi2IAm
n4pXcqFKbrTHnSlZ4Z2W1MUQSw6Ch35lnG+RPklZqgK6n5FeMxKUezDJo6cGuCSGil4rEdcHupvbjKv5yrilL4dtQhZIeY2Uf78kvmkQWzG1aP7cPnqfFFQfMRyW4XFdjWQlQOsNN9XzeE/
gmAxnypepK2QsQqjaMwvhBRUsOFbwigXd0kMqBoIM+fxyQovoDBVYZ+3sSxKiqHs4gXCM2L6XdzXzaxWKxPiAM+L4/
GyXiJkmaiHAhS9hPYlVDQcblWJZCZdh7WIbGeRjWlwopoT8R0DL6ZwD38bpG3MrTz0PO8sDs7CHBVOZKTM0U+ghNN0WPVjvImJEeyRYoeKC7D/S/PMompOZKhJKypyRQLEuWQpUIyNK
+jZaiM1syLYn5eMhhXViTQmN0zjWS95gaSp3R7KdrH0yrsKcelfCnALnPaoZqg+Tpjx4/qzUIQpOiiaUKsqaUziy+TWKjbC6jOTpDiklahhVYE5i2ypgzueNBRPAxtpjNIHDqM6LtuFEIR+ZyLrlbkYwHZ7Doy9JOQxT6

Path: /sequencing_data/preprocessing_re...
Product: u1-cephfs
Provider: DeiC Interactive HPC (SDU)
Created at: 17:12 12/10/2023
Modified at: 17:12 12/10/2023
Size: 1.47 MB
UID/GID: 11042/11042
Unix mode: rw-r--r--

Components of a bioinformatics pipeline



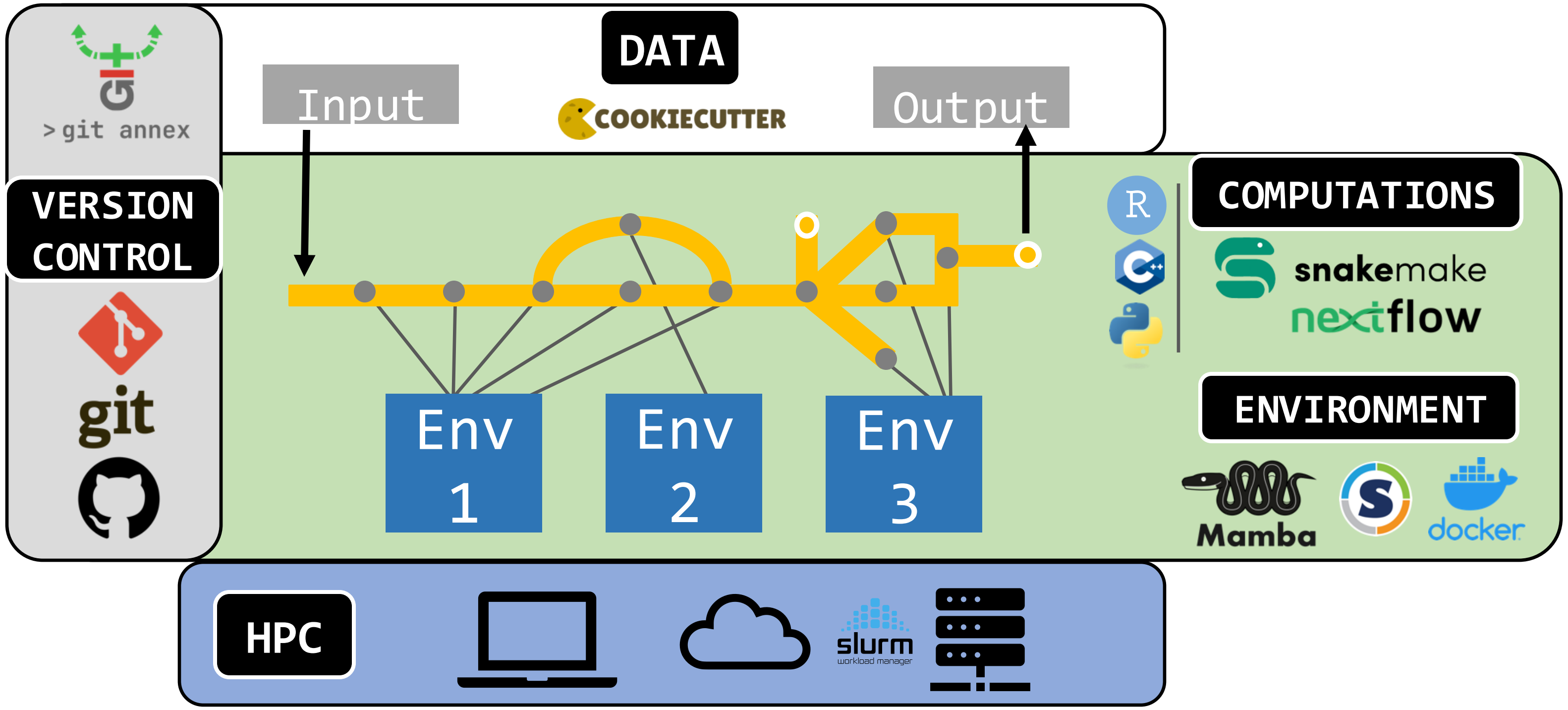
3. **How** we run

2. **How** we install

1. **Where** we set up

More pipeline help

https://heads.ku.dk/course/



Git & Github
- Code management
- Version control

Bash & Unix
- Operating from the terminal

HPC-Launch
- Omics data management
- Using DK HPCs

HPC-Pipes
- Software envs
- Pipeline management