

GWAS with the Genomics Sandbox

from the
Health Data Science
Sandbox



Alba Refoyo Martinez, PhD

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF
COPENHAGEN



Introductions



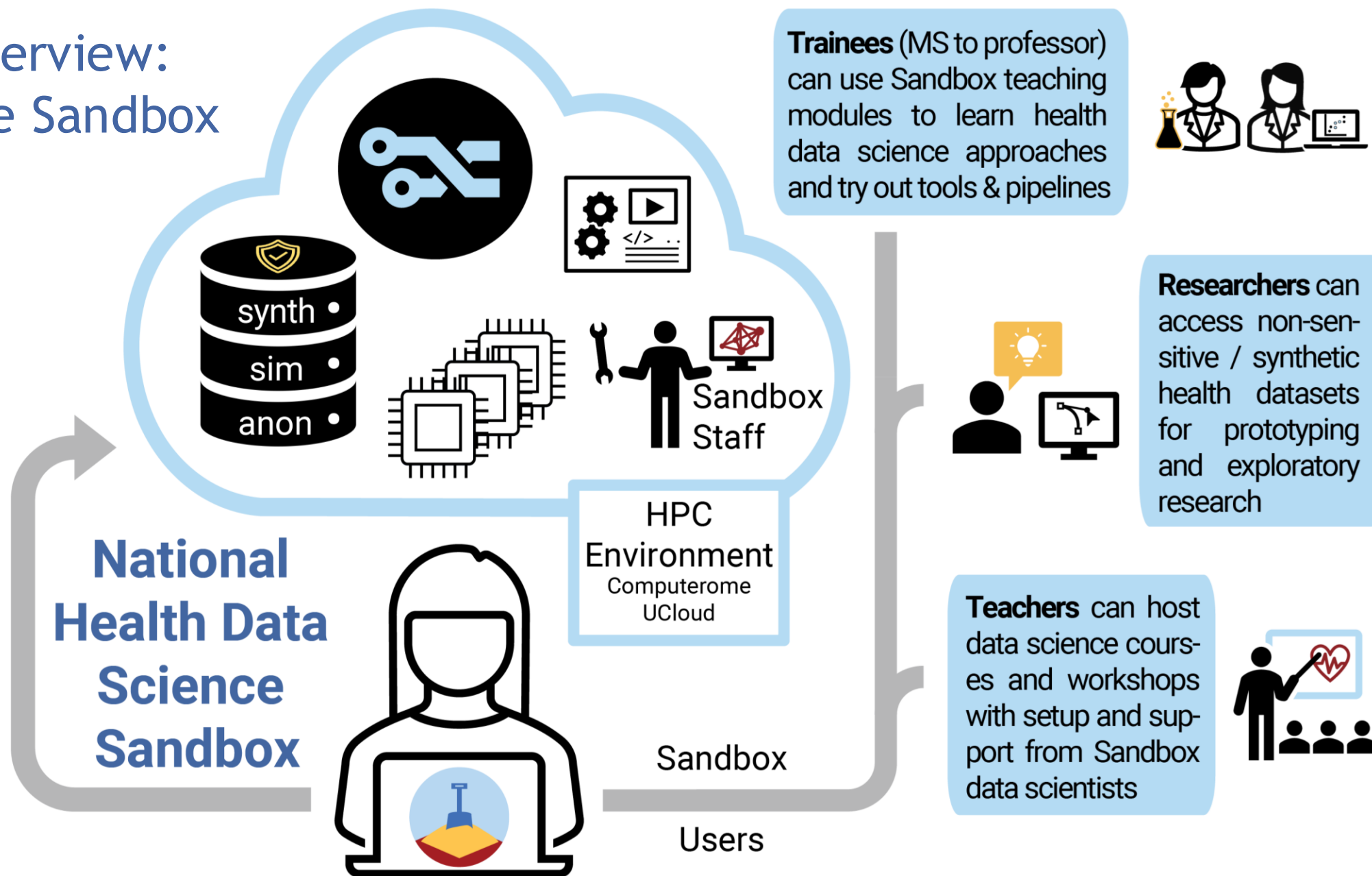
Samuele Soraggi, PhD
Sandbox data scientist



Alba Refoyo Martinez, PhD
Sandbox data scientist



Overview: the Sandbox

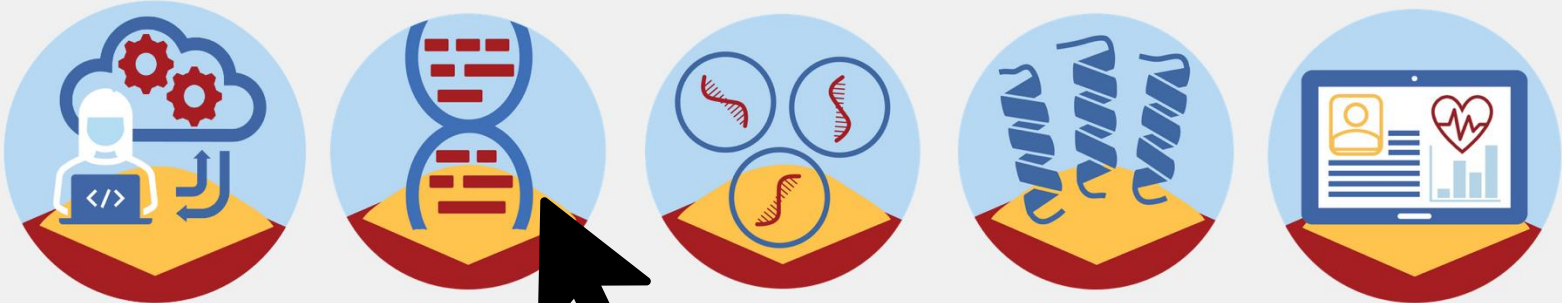


Resources on the Sandbox website

hds-sandbox.github.io

Welcome to the Health Data Science Sandbox






Access our training modules




HPC Lab HPC launch HPC pipes RDM	Genomics NGS data analysis Population Genomics GWAS	Transcriptomics Bulk RNAseq Single-cell RNAseq	Proteomics Clinical Proteomics ColabFold	Health records Synthetic data Personalized Medicine
--	---	---	---	--

Click 'Genomics' > GWAS
Click Workshop in top menu



**GWAS**[Home](#)[Access](#)[Tutorials](#)[Workshop](#)

Instructions


Introduction 

Genetic Theory Review

[What is a GWAS?](#)


Data collection

Data exploration

Quality Control 

Quality Control: initial steps

Quality Control: Relatedness & Population Stratification

Association Testing 

Association testing

Population stratification

Polygenic scores I

Polygenic scores II

Other tools & conclusions

[Introduction](#) > What is a GWAS?

What is a GWAS?

Information about this notebook

Over the past two decades, there has been an increasing interest in investigating the influence of genetic risk factors on various human traits. The technical and analytic tools required to conduct genetic studies have become increasingly accessible, which holds great promise. However, performing genetic association studies properly is complex and requires specific knowledge of genetics, statistics, and (bio)informatics.

This course will introduce key concepts and provide guidelines for building a reusable workflow for Genome-Wide association studies (GWAS) by progressively describing all necessary steps in a typical GWAS analysis. This notebook is read-only, while the others are hands-on practical exercises using R and the Linux command line.

Learning outcomes

- **Recognize and discuss** the basic terminology and GWAS principles
- **Identify** the main steps of a typical GWAS analysis

A Genome-Wide Association Study (GWAS) aims to link genotype and phenotype information by identifying genetic variants that have a statistical association with a trait. A **phenotype**, also called **trait**, can be any measured or observed property of an individual. Phenotypes can include quantitative traits, like standing height or body mass index (BMI), as well as binary traits, such as diagnoses of multiple sclerosis or schizophrenia.

On this page

[General concepts](#)

Key concepts

GWAS output

GWAS examples

Overview of GWAS Steps

Software overview for each GWAS Steps



Deployment on UCloud



cloud.sdu.dk



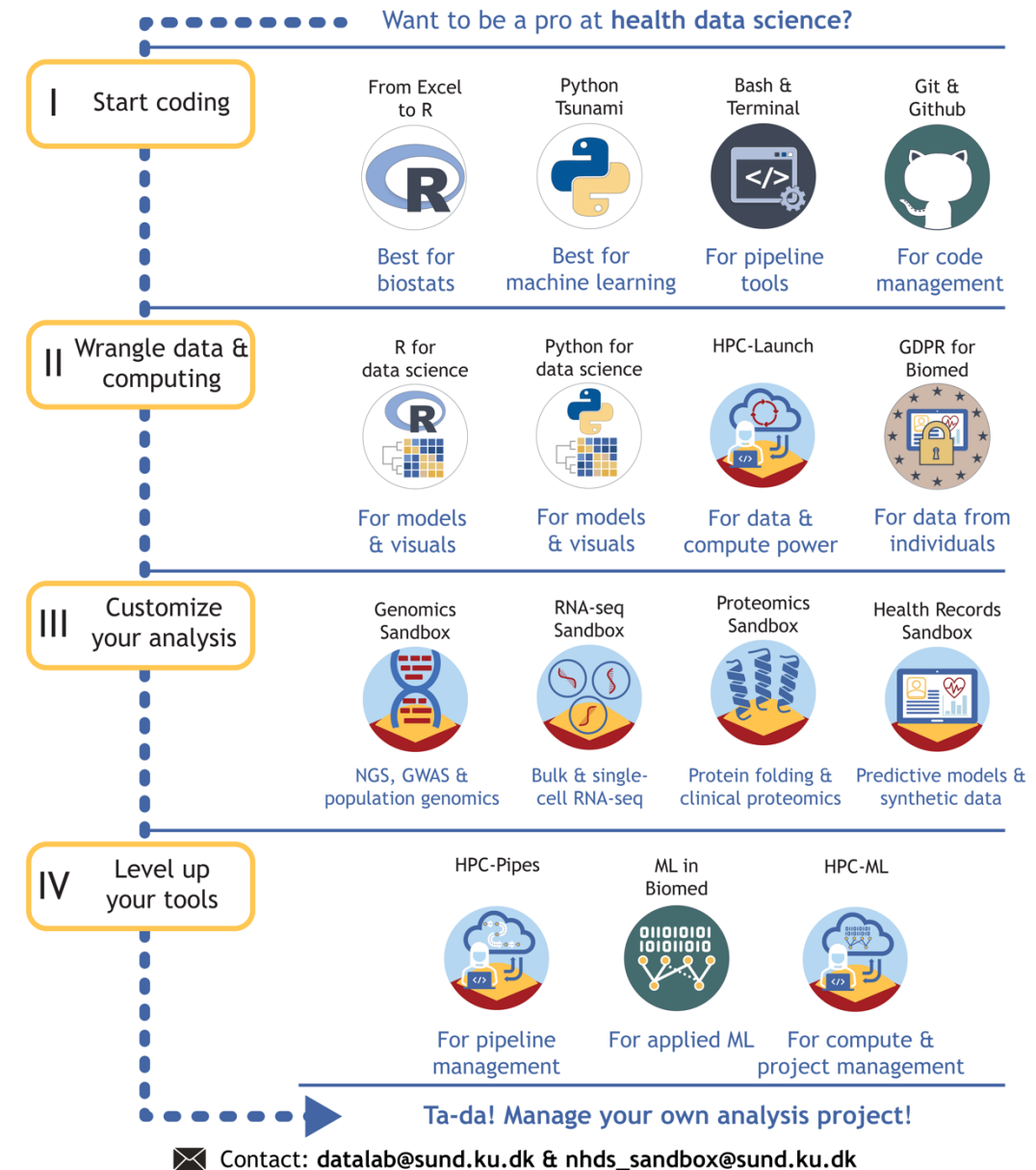
Sandbox + SUND Data Lab

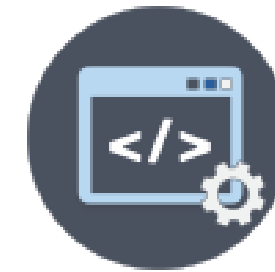
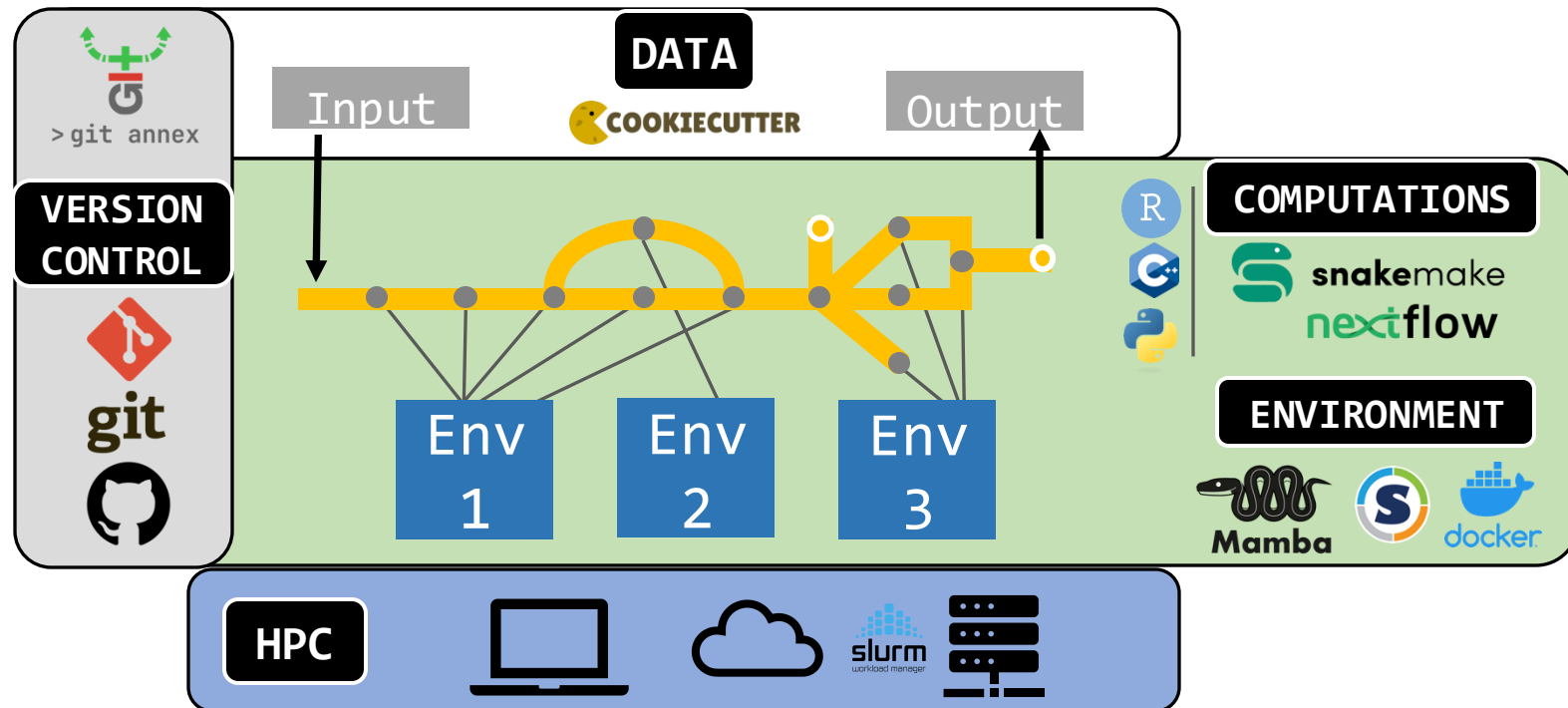
Sandbox: national team geared at more advanced users

- containerized training apps deployed on HPC platforms + web resources + workshops

SUND Data Lab: local KU team geared at beginners

- workshops + project mentorship + consulting





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



GWAS with the Genomics Sandbox



Course learning objectives

1. Explain key **population genetics concepts** and their relevance to genome-wide association studies.
2. Understand and apply **GWAS workflow**
 1. Data collection
 2. Quality control and preprocessing
 3. Imputation of missing genotypes
3. **Association testing:** perform association tests using linear regressions models. Understand linear mixed models and meta-analysis are particularly useful.
4. **Interpreting GWAS results:** analyzing GWAS findings and have a critical approach towards their limitations.
5. Practical **applications** of GWAS



GWAS with the Genomics Sandbox



Today's topics

1. Intro to UCloud
2. Intro to GWAS and data collection
3. GWAS - quality control & preprocessing
 - Part I
 - Part II
4. Practice on UCloud (SDU's teaching-focused HPC platform)





**Intro to
UCloud**

**from the
Health Data Science
Sandbox**



Alba Refoyo Martinez, PhD

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF
COPENHAGEN



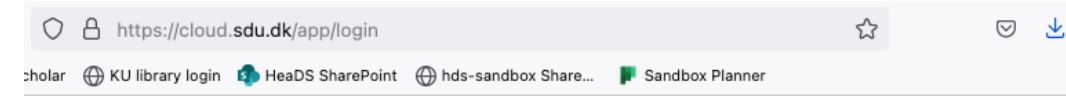
UCloud

- 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/GWAS_course/workshop



Welcome to GWAS with the Genomics Sandbox

⚠ 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).

Required preparation

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 WAYF¹.

for UCloud Access click here



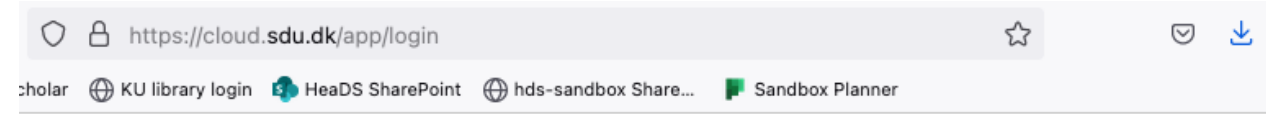
Integration Portal




Other login options →



UCloud log-in



To access *UCloud* please choose your login provider

SDU 
University of Copenhagen

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

Search here

1. Search for
your uni & then
click on link

2. Sign-in
via your uni
portal

DeiC

Integration Portal

WAYF  Login

Other login options →



UCloud log-in

Back to “Workshop” 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:

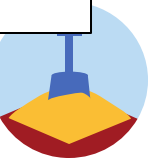
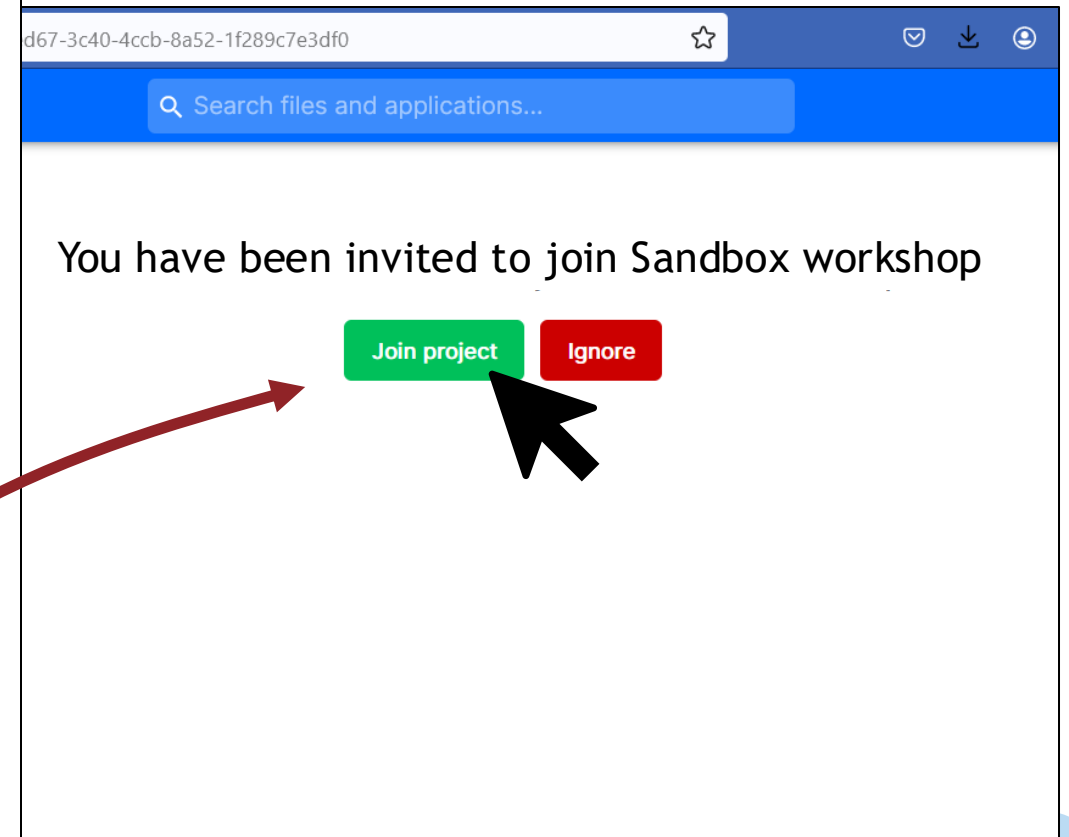
1. a Danish university ID so you can sign on to UCloud via WAYF¹.

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 workshop² Click the link below after your first uCloud access and accept the invite that shows.

Invite link to uCloud workspace





Workspace

The screenshot shows the UCloud workspace interface. On the left is a blue sidebar with icons for workspace, folder, users, add, shopping bag, and server. The main content area features a release announcement for 'UCloud 2025.2.0 Release' with details about improved Slurm integration and a new task system, dated 13:30 27/01/2025. On the right, a dropdown menu is open, showing a search bar and a list of projects: 'My workspace', 'Health Data Science Sandbox', 'OMICS workshop', 'Sandbox RNAseq workshop', and 'Sandbox_workshop'. The 'Sandbox_workshop' item is highlighted with a red border. A black mouse cursor is pointing at the dropdown arrow in the top right corner.

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

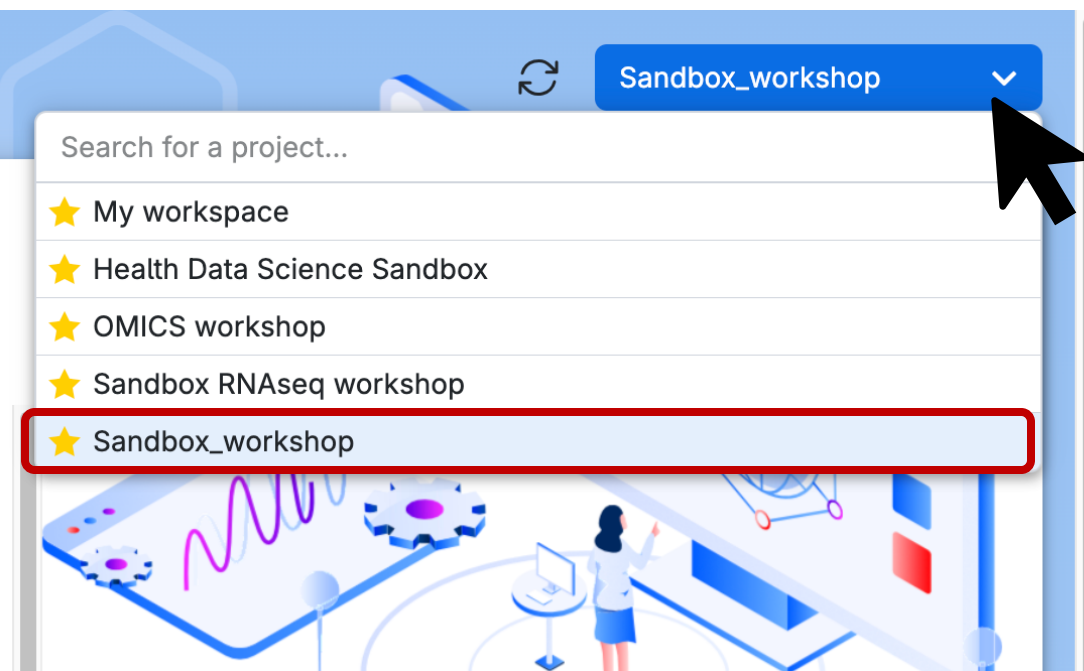




UCloud usage

During this course, utilize Sandbox 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

UCloud 2025.2.0 Release

Improved Slurm integration, application filters and new task system

13:30 27/01/2025

Today brings yet another release. Today we are focusing most of our energy on our Slurm integration. This release brings several new features while also fixing a vast number of bugs. A new system has been added allowing you to fully customize the scripts which are submitted to Slurm based systems. For Slurm service providers, it is now also possible to open a terminal in any folder directly from the file browser.

We have added several new applications for the Slurm based providers. To make them easier to find, a new filter for the application catalog has been added. This filter will allow you to find only the applications that are relevant for you. A new task system has been added, which allows you to track the progress of tasks running in the background.

Update 28/01/25: Bugs have been fixed related to job imports and grant applications.

Resource allocations

u1-standard-h

578,80 / 23,28K Core-hours (2%)

u1-cephfs

52 GB / 500 GB (10%)

Apply for resources

test_ARM

09:37

gwas qqman test

29/01/2025

smk day 2

13/11/2024

smk day 2

13/11/2024

smk day 2

13/11/2024

smk day 2

13/11/2024

test new version

11/11/2024

smk day 2

11/11/2024

snakemake exercise 3

05/11/2024

ESTH_day2

05/11/2024



Drives

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



Drives

Create drive

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



Sandbox RNAseq works... ▾

Your username: should be FirstLast#0000...

You have a personal drive

You have shared drives

In this workshop, you should only have access to your own personal drive



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



Drives

Inside your personal drive, make your own custom working directory... we'll use this later to save the results from the analyses!

Sandbox_workshop

Upload files

U

Create folder

F

Sync

M

Show hidden files ☒

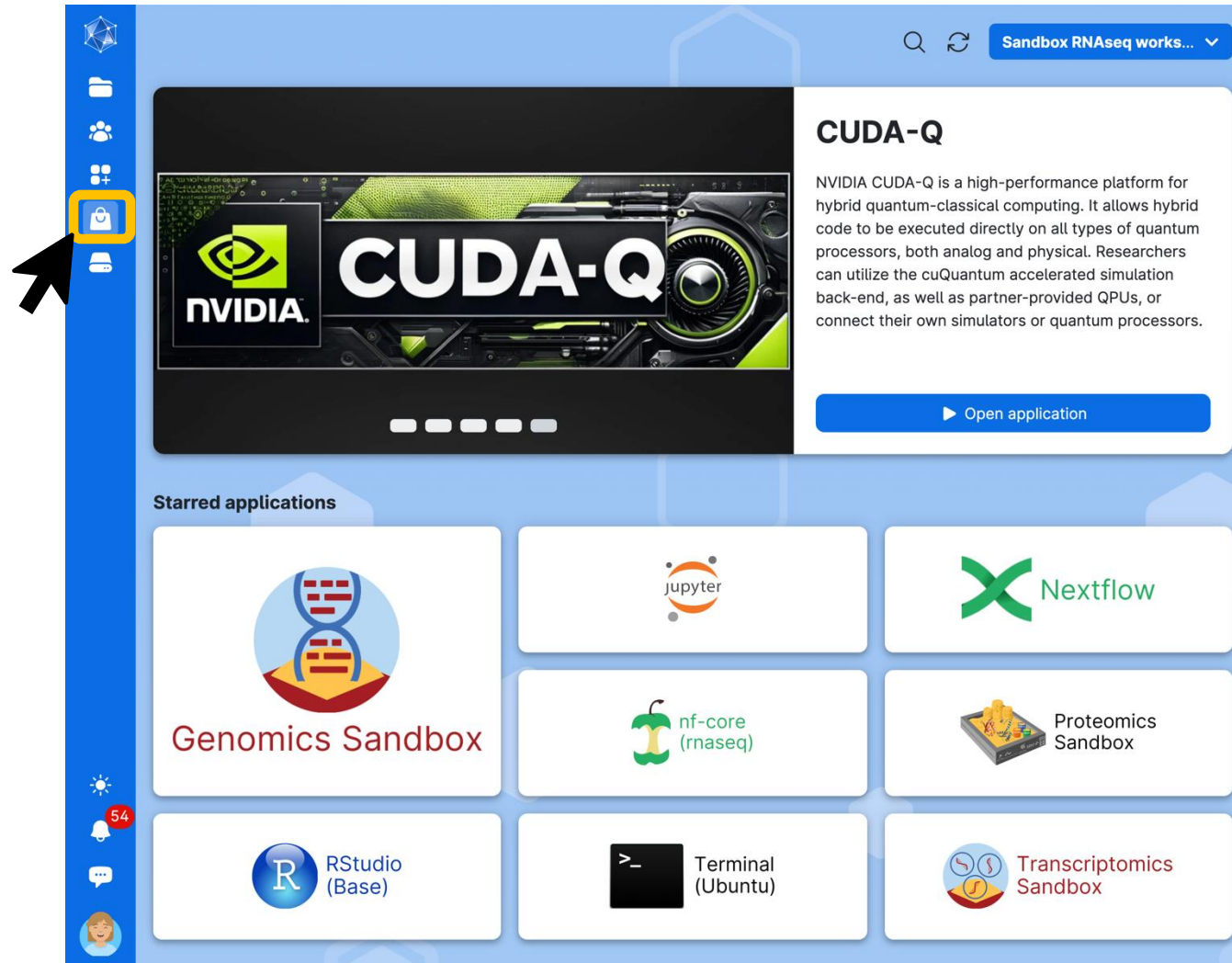
Name	↓ Modified at	Size
Jobs	09:37 25/02/2025	
work_ARM	09:35 25/02/2025	
Trash	02:01 19/11/2024	



Applications

There is a wide variety of applications.

Here are some of my favorites!





Apps

Search for Sandbox apps



The screenshot shows the HeaDS Apps interface. At the top, a search bar contains the text 'sandbox'. To the right of the search bar are icons for search and refresh, and a dropdown menu showing 'Sandbox_workshop'. Below the search bar, there are four app cards arranged in a 2x2 grid. Each card has an icon, a title, a description, and a blue star icon in the top right corner.

App Name	Description
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...





How to submit a Sandbox app job?

Let's set up the app together!



Health Data
Science
Sandbox

Documentation

Health Data Science San...

Genomics

2025.02

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

E-mail notification settings

Do not notify me

Import parameters

Submit

Estimated cost

-

Current balance

-

Job name

Example: Run with parameters XYZ

Hours *

1

+1

+8

+24

Machine type *

No machine type selected

Select folders to use


Add folder


If you need to use your files in this job then click "Add folder" to select the relevant files.

Mandatory Parameters

Select a course or module *

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 




Health Data Science

Sandbox

Genomics

2025.02

[Documentation](#) 

Sandbox_workshop

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

E-mail notification settings

Do not notify me

Import parameters

Submit

Estimated cost

Current balance

4 Core-hours

22, 70K Core-hours

Job name

test_ARM

Hours *

1

+1


+8

+24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour



Select folders to use

Add folder

If you need to use your [files](#) in this job then click "Add folder" to select the relevant files.

Mandatory Parameters

Select a course or module *


Introduction to GWAS

Configure custom links to your application

Add public link

If your job needs to be publicly accessible via a web-browser then click "Add public link" to select the correct link.

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 

Health Data Science
Sandbox

Genomics 2025.02

Documentation

Sandbox_workshop

Courses, d
E-mail notifi
Do not no

Job nam
test_A

Machine
u1-sta
vCPU
4 (Inte

Select
If you ne

Mandat

Select a
Introd

UCloud Docs

Login eScience Center Terms of Service Developer Guide User Guide FAQ

Docs / Apps Index / Genomics Sandbox

Supported Apps

Apps Type

Apps Index

Airflow
AlmaLinux Xfce
ANSYS
Archiver
BCL Convert
Cell Ranger
Chat UI
Coder
ColabFold
COMSOL
CUDA-Q
cuQuantum
CVAT
Dalton
Dash
Django
FastQC
FreeSurfer
Genomics
Gephi
GROMACS
Indico
JAMOV
JupyterLab

Genomics Sandbox

2025-02 2023-03-01 2022-06-01

Release 2025-02 Type JupyterLab Access Open

- Operating System: Ubuntu 24.04
- Terminal: tmux 3.19.0
- Shell: bash 5.2.30
- Editor: CNO Emacs 29.3 CNO nano 2.9 Vim 9.1
- Package Manager: apt 2.7.14 dpkg 1.20.6 rpm 4.16.0 pip 24.0
- Programming Language: GCC 13.2.0 Python 3.12.3

In this app you will find material for the Genomics sandbox of the Health Data Science sandbox. It contains course tutorials, datasets, and tools you can use for research or self-learning. Each course item of this sandbox is based on Jupyterlab. Jupyterlab is a web-based integrated development environment for Jupyter notebooks, code, and data.

Available material


Items are periodically added to this app and can be chosen from the menu. Each item can be a course, a setup to work with specific software, a research example, and comes with all necessary packages already installed, notebooks with code and explanations, and a dedicated webpage with additional material (notes, slides, recordings, ...).

Configure custom links to your application

Add public link

If your job needs to be publicly accessible via a web-browser then click "Add public link" to select the correct link.

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 



Health Data Science

Sandbox

Genomics

2025.02

Documentation

Sandbox_workshop

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

E-mail notification settings

Do not notify me

Import parameters

Submit

Estimated cost

Current balance

4 Core-hours

22,70K Core-hours

Job name

test_ARM

Hours *

1

+1

+8

+24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour

Select folders to use

Add folder

Your files will be available at /work/.

/Member Files: AlbaRefoyoMartínez#0753/work_ARM


Remove

Mandatory Parameters

Select a course or module *

Introduction to GWAS

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 

Health Data Science

Sandbox

Genomics

2025.02

Documentation

Sandbox_workshop

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

Import parameters

Submit

E-mail notification settings

Do not notify me

Estimated cost 4 Core-hours

Current balance 22.70K Core-hours

Jobs

Upload JobParameters.json

Select file from UCloud

Created

Status

Created by

test_ARM

09:37

Import

Your files will be available at /work/.

Remove


/Member Files: AlbaRefoyoMartínez#0753/work_ARM

Mandatory Parameters

Select a course or module *

Introduction to GWAS

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 



Health Data Science

Sandbox

Genomics

2025.02

Documentation

Sandbox_workshop

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

E-mail notification settings

Do not notify me

Estimated cost

Current balance

4 Core-hours

22,70K Core-hours

Import parameters

Submit

Job name

test_ARM

Hours *

1

+1

+8

+24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour

Use your initials / a unique name!

Select folders to use

Add folder

Your files will be available at /work/.

/Member Files: AlbaRefoyoMartínez#0753/work_ARM


Remove

Mandatory Parameters

Select a course or module *

Introduction to GWAS

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 



Health Data Science

Sandbox

Genomics

2025.02

Documentation

Sandbox_workshop

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

Import parameters

Submit

E-mail notification settings

Do not notify me

Estimated cost

Current balance

4 Core-hours

22,70K Core-hours

Job name

test_ARM

Machine type *

u1-standard-4

vCPU

4 (Intel Xeon Gold 6

Hours *

1

+1

+8

+24

Price

4 Core-hours/hour

Select folders to use

Your files will be available at /work/.

Add folder

/Member Files: AlbaRefoyoMartínez#0753/work_ARM

Remove x

Mandatory Parameters


Select a course or module *

Introduction to GWAS

Right now, ask for an hour (for testing)

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

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 



Health Data Science

Sandbox

Genomics

2025.02

Documentation

Sandbox_workshop

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

Import parameters

Submit

E-mail notification settings

Do not notify me

Estimated cost

Current balance

4 Core-hours

22,70K Core-hours

Job name

test_ARM

Hours *

1

+1

+8

+24

Machine type *

u1-standard-4

vCPU	Memory (GB)
4 (Intel Xeon Gold 6130)	24

We will ask for 4 cores to run the software to perform GWAS

Select folders to use

Add folder

Your files will be available at /work/.

/Member Files: AlbaRefoyoMartínez#0753/work_ARM


Remove

Mandatory Parameters

Select a course or module *

Introduction to GWAS

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 

Health Data Science

Sandbox

Genomics

2025.02

Documentation

Sandbox_workshop

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

E-mail notification settings

Do not notify me

Import parameters

Submit

Estimated cost

Current balance

4 Core-hours

22,70K Core-hours

Job name

test_ARM

Hours *

1

+1

+8

+24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24		hour

Select folders to use

Your files will be available at /work/.

/Member Files: AlbaRefoyoMartínez#0753/work_ARM

Remove x

Add folder

Mandatory Parameters


Select a course or module *

Introduction to GWAS

Add the custom working directory we made earlier



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 

Health Data Science
Sandbox

Genomics 2025.02

Documentation

Sandbox_workshop

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

E-mail notification settings
Do not notify me

Estimated cost
Current balance

4 Core-hours
22,70K Core-hours

Import parameters

Submit

Job name
test_ARM

Hours *
1

+1 +8 +24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour

Select folders to use
Your files will be available at /v

/Member Files: AlbaRefoyoM

Add folder

Remove x

Mandatory Parameters

Select a course or module *

Introduction to GWAS

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

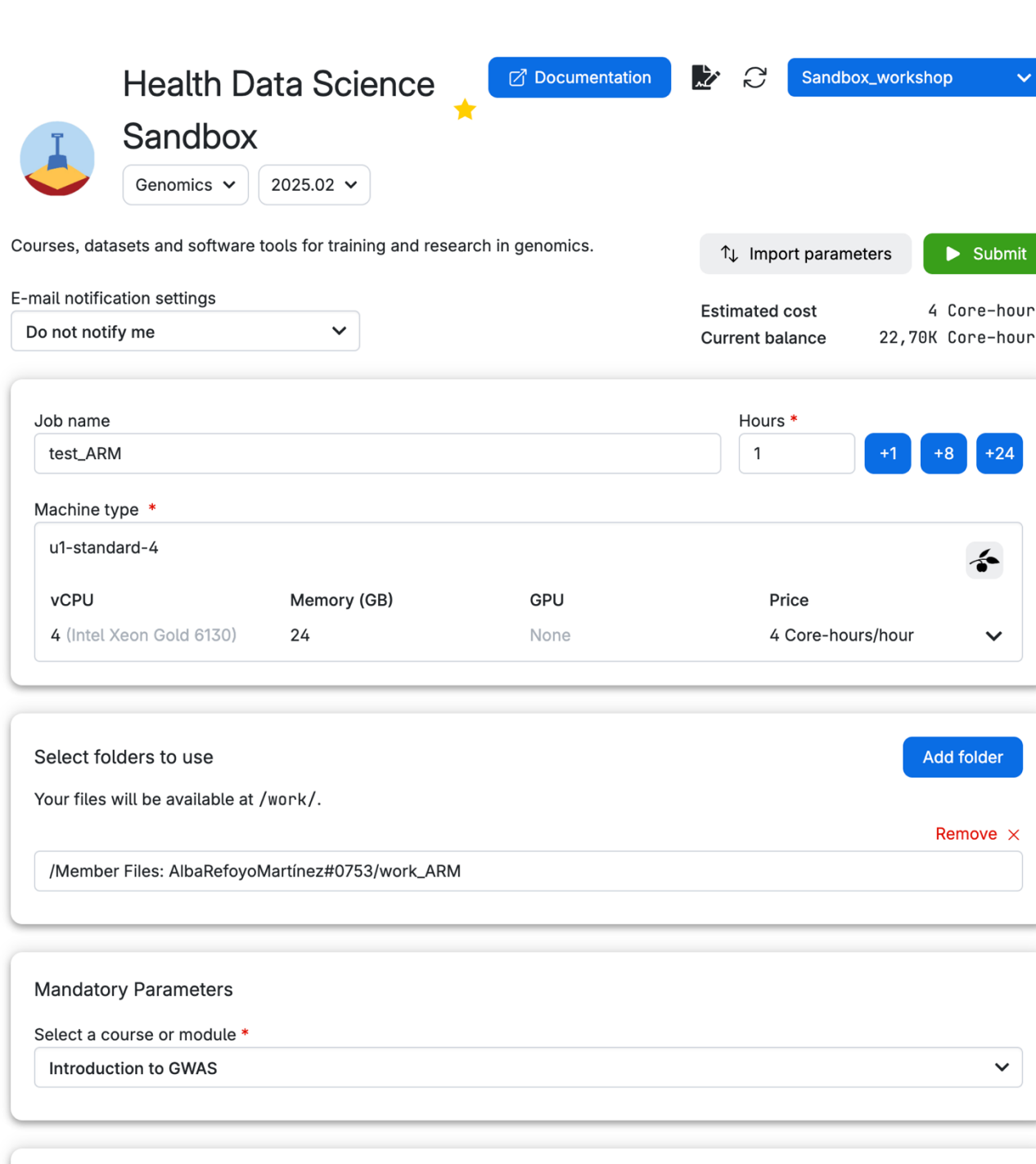


Submitting a job with a Sandbox app

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?



The screenshot displays the 'Health Data Science Sandbox' interface. At the top, there's a navigation bar with 'Documentation', a refresh icon, and a 'Sandbox_workshop' dropdown. The main header includes the 'Health Data Science Sandbox' title, a star icon, and dropdown menus for 'Genomics' and '2025.02'. Below this, a description states: 'Courses, datasets and software tools for training and research in genomics.' To the right are buttons for 'Import parameters' and 'Submit'. A section for 'E-mail notification settings' shows a dropdown set to 'Do not notify me'. On the right, 'Estimated cost' and 'Current balance' are listed, along with '4 Core-hours' and '22,70K Core-hours'. The main job configuration area includes a 'Job name' field with 'test_ARM', an 'Hours' field with '1', and buttons for '+1', '+8', and '+24'. Below this is the 'Machine type' section, showing 'u1-standard-4' with a table of specifications: vCPU (4 Intel Xeon Gold 6130), Memory (24 GB), GPU (None), and Price (4 Core-hours/hour). A 'Select folders to use' section shows a path '/Member Files: AlbaRefoyoMartínez#0753/work_ARM' with a 'Remove' button. The 'Mandatory Parameters' section has a dropdown set to 'Introduction to GWAS'.

Health Data Science
Sandbox

Genomics 2025.02

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

Import parameters Submit

E-mail notification settings
Do not notify me

Estimated cost
Current balance

4 Core-hours
22,70K Core-hours

Job name
test_ARM

Hours *
1 +1 +8 +24

Machine type *
u1-standard-4


vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour

Select folders to use
Your files will be available at /work/.
Remove x
/Member Files: AlbaRefoyoMartínez#0753/work_ARM

Mandatory Parameters
Select a course or module *
Introduction to GWAS

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**



Health Data Science
Sandbox

Genomics ▾2025.02 ▾

[Documentation](#)

Sandbox_workshop ▾

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

Import parameters

Submit

E-mail notification settings

Do not notify me ▾

Estimated cost

Current balance

4 Core-hours

22,70K Core-hours

Job name

test_ARM

Hours *

1

+1

+8

+24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price	
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour	▾

Select folders to use

Add folder

Your files will be available at /work/.

/Member Files: AlbaRefoyoMartínez#0753/work_ARM

Remove ×

Mandatory Parameters

Select a course or module *

Introduction to GWAS ▾



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: /work/]
inflating: /work/]
```

Wait a couple of min
before you click on
'Open interface'.

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

Before time remaining is over,
you can add extra hours





Jobs

Do you have another familiar folder here?

>>>

The screenshot shows the Databricks Launcher interface. At the top, there's a menu bar with File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. Below the menu is a sidebar with a file explorer showing a directory structure: /, Data (47s ago), Notebooks (45s ago), work_ARM (1h ago), and JobParameters.js... (9m ago). The main area is titled 'Launcher' and contains three sections: 'Notebook', 'Console', and 'Other'. Each section displays a grid of application tiles. The 'Notebook' section has tiles for Python 3 (ipykernel), Bash, R, and Xonsh. The 'Console' section also has tiles for Python 3 (ipykernel), Bash, R, and Xonsh. The 'Other' section has tiles for Terminal, Text File, Markdown File, Python File, and R File. At the bottom of the interface, there's a status bar showing 'Simple' mode, disk usage (854.50 / 1023.50 GB), CPU usage (0.00 %), and memory usage (117.47 MB). A small notification box in the bottom right corner asks 'Would you like to get notified about official Jupyter news?' with links to 'Open privacy policy', 'Yes', and 'No'.

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





Jobs

File Edit View Run Kernel Git Tabs Settings Help CPU: 0% | Mem: 117 MB | Disk: 854.50 / 1023.50 GB

Launcher

File Explorer

Name	Modified
Data	47s ago
Notebooks	45s ago
work_ARM	1h ago
JobParameters.js...	9m ago

Notebook

Python 3 (ipykernel) Bash R Xonsh

Console

Python 3 (ipykernel) Bash R X

Other

Terminal Text File

Simple 0 \$ 0 Disk: 854.50 / 1023.50 GB | CPU: 0.00 % | Mem: 117.47 MB Launcher 1

Would you like to get notified about official Jupyter news? [Open privacy policy](#) Yes No

As we have mentioned 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
(use the terminal or drag the data/nb into your own folder)





Jobs

1

2

3

1) Navigate to the Notebooks folder

2) Open the general-intro.ipynb

3) Scroll down, add a new cell, change the type of cell to code (from markdown) and type 'pwd'

1) Save the file

The screenshot shows the JupyterLab interface. On the left, the file explorer displays the directory structure: / work_ARM / Notebooks /. The file general-intro.ipynb is selected. In the center, the notebook content is visible, showing the title 'Getting Started with the GWAS Module' and the authors: Conor O'Hare, Samuele Soraggi, and Alba Refoyo Martinez. The notebook content includes a code cell with the following text:

```
format:
  html:
    number-sections: true
    toc: true
  ipynb:
    toc: true
    number-sections: false
```

Below the code cell, there is a paragraph of text: 'In this course, we will use JupyterLab, an open-source interactive development environment (IDE) for working with Jupyter Notebooks. If you're familiar with this interface, check out the [JupyterLab Documentation](#). It includes numerous screenshots and videos showcasing the cool of JupyterLab.'

At the bottom of the interface, there are two buttons: 'Choose the Bash kernel' and 'Choose the R kernel'. A small notification box in the bottom right corner asks: 'Would you like to get notified about official Jupyter news?' with links for 'Open privacy policy', 'Yes', and 'No'.





Kernels

We will shift between two kernels*, and along the notebook, every time we shift from one language to another.

1. Change kernel to Bash

*A Jupyter kernel provides a programming language runtime and the required dependencies for executing code.

The screenshot shows a Jupyter Notebook interface. The 'Kernel' menu is open, displaying options like 'Interrupt Kernel', 'Restart Kernel...', 'Restart Kernel and Clear Outputs of All Cells...', 'Restart Kernel and Run up to Selected Cell...', 'Restart Kernel and Run All Cells...', 'Restart Kernel and Debug...', 'Reconnect to Kernel', 'Shut Down Kernel', 'Shut Down All Kernels...', and 'Change Kernel...'. The 'Change Kernel...' option is highlighted with a yellow box. A dialog box titled 'Start R Kernel' is open, showing options: 'R', 'Use No Kernel', 'No Kernel', 'Start Kernel', 'Bash', 'Python 3 (ipykernel)', 'Xonsh', and 'Connect to Existing R Kernel'. The 'Bash' option is selected. Below the dialog box, there is a warning message: 'Warning: If a cell fails to run, you may need to restart the kernel. You can create new kernels from the menu at the top of the page, select Kernel -> Change Kernel, and then select the preferred one.' At the bottom of the notebook, there is a status bar showing 'Simple', '0', '1', 'R | Idle', 'Disk: 811.10 / 1023.50 GB | CPU: 0.00 % | Mem: 802.22 MB', 'Mode: Command', 'Ln 1, Col 1', 'general-intro.ipynb', and a notification bell icon.



Running a cell

Run the cell you created with the 'pwd' command

The screenshot shows a Jupyter Notebook interface. On the left is a file explorer showing a directory structure with folders like 'exercises', 'images', 'lib', and 'references', and files like 'general-intro.ipynb'. The main area displays a notebook titled 'general-intro.ipynb'. The notebook content includes a section titled 'General note on how to make the notebooks work' with several bullet points. Below this is a warning box. At the bottom, a code cell is shown with the command 'pwd' and its output '/work/work_ARM/Notebooks'. A yellow box highlights the command and its output. The top of the interface shows a menu bar with options like 'File', 'Edit', 'View', 'Run', 'Kernel', 'Git', 'Tabs', 'Settings', and 'Help'. The bottom status bar shows system information like CPU usage, memory, and disk space.

File Edit View Run Kernel Git Tabs Settings Help

CPU: 39% | Mem: 191 MB | Disk: 632.05 / 1023.50 GB


Launcher general-intro.ipynb


How to make the notebooks work

- How to make the notebooks work
- Managing kernels and terminals

General note on how to make the notebooks work

- We will primarily use R and bash command line programming languages, where R is used for statistical analysis of the output from various tools running in bash.
- Since we use multiple coding languages, you need to **choose a kernel every time we shift from one language to another**. A kernel contains a programming language and the necessary packages to run the course material. To choose a kernel, go to the menu at the top of the page, select **Kernel** --> **Change Kernel**, and then select the preferred one.
- We will shift between two kernels, and along the notebook, you will see a picture indicating when to change the kernel. The two pictures are shown below:

 Choose the Bash kernel

 Choose the R kernel

- You can run the code in each cell (grey background) by clicking the run cell sign in the toolbar, or simply by pressing **Shift + Enter**. When the code is done running, a small green check mark will appear on the left side.
- You need to **run the cells sequentially** to execute the analysis. Please do not run a cell until the one above is done running, and do not skip any cells.
- Textual descriptions accompany the code to help you understand what is happening. Please try not to focus on understanding the code itself in too much detail, but rather focus on the explanations and commands' output.
- You can create new code cells by pressing **+** in the Menu bar above or by pressing **B** after selecting a cell.

Warning

- If a cell fails to run, verify the kernel in use, shown in the top-right corner.
- You don't know the answer to the exercises? You can use Generative AI to help with the code. It can assist you when the tutorial alone isn't enough or if you want to explore additional concepts beyond the exercise.

```
[1]: pwd
/work/work_ARM/Notebooks
```

Would you like to get notified about official Jupyter news?

[Open privacy policy](#) Yes No

Simple 0 s 1 Bash | Idle Disk: 632.05 / 1023.50 GB | CPU: 0.00 % | Mem: 190.89 MB Mode: Command Ln 1, Col 1 general-intro.ipynb 1



Jobs



Test is now running (ID: 5049428)

Open terminal

Open interface

Stop application

Stop the app by holding down the 'stop' button

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] AlbaRefoyoMartinez#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





Jobs

Files

Drives

Member Files: AlbaRef

shared

Mer

/Member Files: AlbaRefoyoMar.../work_ARM

Sandbox_workshop

Upload files

Create folder

Sync

M

Show hidden files

☑

Name	↓ Modified at	Size
Notebooks	10:51 25/02/2025	
Data	10:40 25/02/2025	

Go to the new folder in your personal drive



Jobs

Access past and current running jobs



Click on 'Runs' and choose from the list of jobs, (*rerun application again* ensures the use of the same parameters)

or

Click in notifications (and choose the job you might want to stop it or rerun it)



Your job has completed

Genomics Sandbox 2025.02 for test_ARM (ID: 5179883)

Run application again



Job info

Name: test_ARM
ID: 5179883
Reservation: DeiC Interactive HPC (SDU/K8s) / u1-standard-4 (x1)
Launched by: AlbaRefoyoMartínez#0753 in Sandbox_workshop

Messages

[10:31] AlbaRefoyoMartínez#0753 has requested 1x u1-standard-4 from DeiC Interactive HPC (SDU/K8s)
[10:31] Assigned to nodeaa-20
[10:31] Job is starting soon
[10:41] Job has started
[11:07] Job has been cancelled



/Member Files: AlbaRefoyoMar.../.../Genomics Sandbox/test_ARM (5179883)



Sandbox_workshop







Upload files



Create folder



Name		↓ Modified at	Size
 stdout.txt		11:07 25/02/2025	6.10 KB
 JobParameters.json		10:31 25/02/2025	1.08 KB





Jobs



Your job has completed

Genomics Sandbox 2025.02 for test_ARM (ID: 5179883)


Run application again

Job info



Name: test_ARM
ID: 5179883
Reservation: DeIC Interactive HPC (SDU/K8s) / u1-standard-4 (x1)
Launched by: AlbaRefoyoMartínez#0753 in Sandbox_workshop

Messages

[10:31] AlbaRefoyoMartínez#0753 has requested 1x u1-standard-4 from DeIC Interactive HPC (SDU/K8s)
[10:31] Assigned to nodeaa-20
[10:31] Job is starting soon
[10:41] Job has started
[11:07] Job has been cancelled

▼  /Member Files: AlbaRefoyoMar.../.../Genomics Sandbox/test_ARM (5179883) 🔍 ↺ Sandbox_workshop ▼

📁 Upload files 📄 U 📁 Create folder 📄 F

Name	↓ Modified at	Size
 stdout.txt	(P) 11:07 25/02/2025	6.10 KB
 JobParameters.json	(P) 10:31 25/02/2025	1.08 KB

Hit 'Run application again' to restart super fast with same parameters





New job

Make sure you add your own Notebooks and Data folders (so you can modify and save them as you work on them)

📁

👤

⚙️

🛒

🔔

🌞

🕒

🔔

💬

👤

190

Health Data Science Sandbox ★

Genomics ▼ 2025.02 ▼

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

E-mail notification settings
Do not notify me ▼

[Documentation](#)

Sandbox_workshop ▼

Import parameters ↕

Submit ▶

Estimated cost
Current balance

4 Core-hours
22,76K Core-hours

Job name
test_ARM

Hours *
1 +1 +8 +24

Machine type *

u1-standard-4

vCPU	Memory (GB)	GPU	Price
4 (Intel Xeon Gold 6130)	24	None	4 Core-hours/hour

Select folders to use

Add folder

Your files will be available at /work/.

/Member Files: AlbaRefoyoMartinez#0753/work_ARM/Notebooks Remove ✕

/Member Files: AlbaRefoyoMartinez#0753/work_ARM/Data Remove ✕

Mandatory Parameters

Select a course or module *

Introduction to GWAS ▼

Configure custom links to your application

Add public link

If your job needs to be publicly accessible via a web-browser then click "Add public link" to select the correct link.



New job

Please submit the new job. Change the time for a few hours



Health Data Science Sandbox ★

Genomics ▾

2025.02 ▾

[Documentation](#)



Sandbox_workshop ▾

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

E-mail notification settings

Do not notify me ▾

↕ Import parameters

▶ Submit

Estimated cost
Current balance

4 Core-hours
22,76K Core-hours

Job name

test_ARM

Hours *

1

+1

+8

+24

Machine type *

u1-standard-4

vCPU

4 (Intel Xeon Gold 6130)

Memory (GB)

24

GPU

None

Price

4 Core-hours/hour

Select folders to use

Your files will be available at /work/.

Add folder

/Member Files: AlbaRefoyoMartinez#0753/work_ARM/Notebooks

Remove ✕

/Member Files: AlbaRefoyoMartinez#0753/work_ARM/Data

Remove ✕

Mandatory Parameters

Select a course or module *

Introduction to GWAS ▾

Configure custom links to your application

Add public link

If your job needs to be publicly accessible via a web-browser then click "Add public link" to select the correct link.