Genomics England

Working with Python in the RE

Emily Perry

Research Engagement Manager

8th April 2025

Data security

- This training session will include data from the GEL Research Environment
- As part of your IG training you have agreed to not distribute these data in any way
- If you are joining virtually, you are not allowed to:
    - Invite colleagues to watch this training with you
    - Take any screenshots or videos of the training
    - Share your webinar link (we will remove anyone who is here twice)

Presenters

Emily Perry
Research Engagement Manager

Sangram Keshari Sahu
LifeBit

Questions

All your microphones are muted

Use the Zoom Q&A to ask questions

Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Helpers

Elena Bernabeu
Genomics Data Scientist - Research Services

Christian Bouwens
Bioinformatician - Research Services

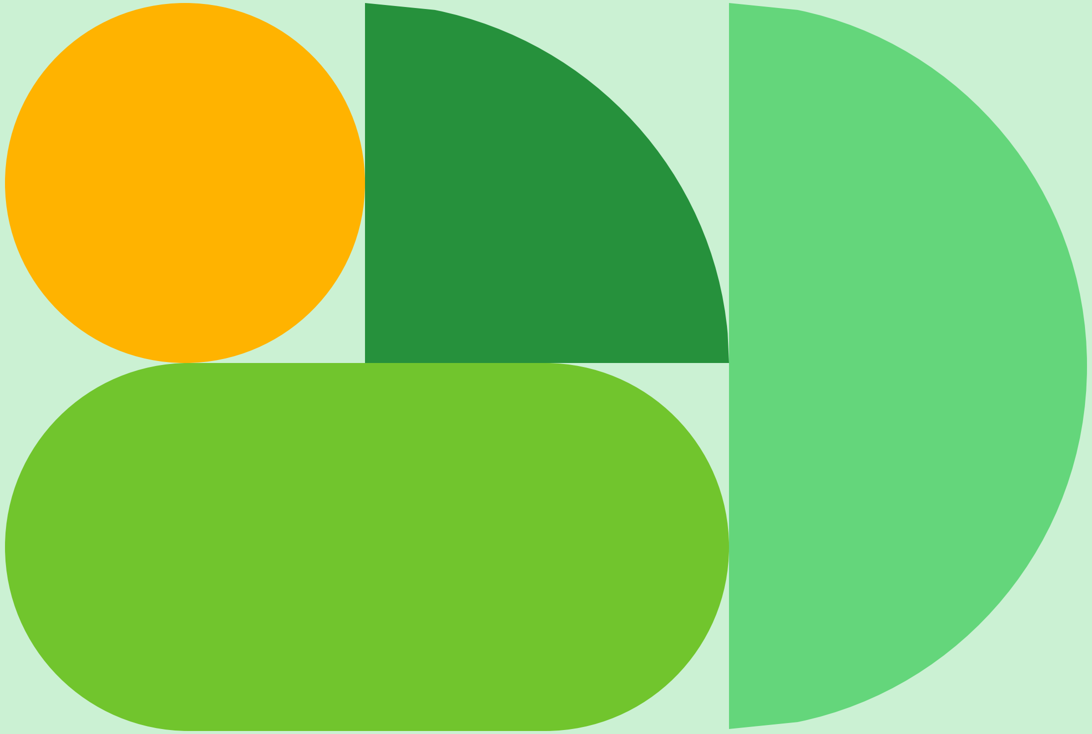
Matthieu Vizuete-Forster
Bioinformatician - Research Services

Eleni Kyriakou
LifeBit

Hamzah Syed
Lifebit

Agenda

1. Introduction and admin
2. Pre-built conda environments in the Research Environment
3. Create conda environments in the Research Environment
4. Working with Jupyter on the HPC
5. Working with Jupyter notebooks in CloudOS interactive sessions
6. Query clinical data with Python
7. Help and questions



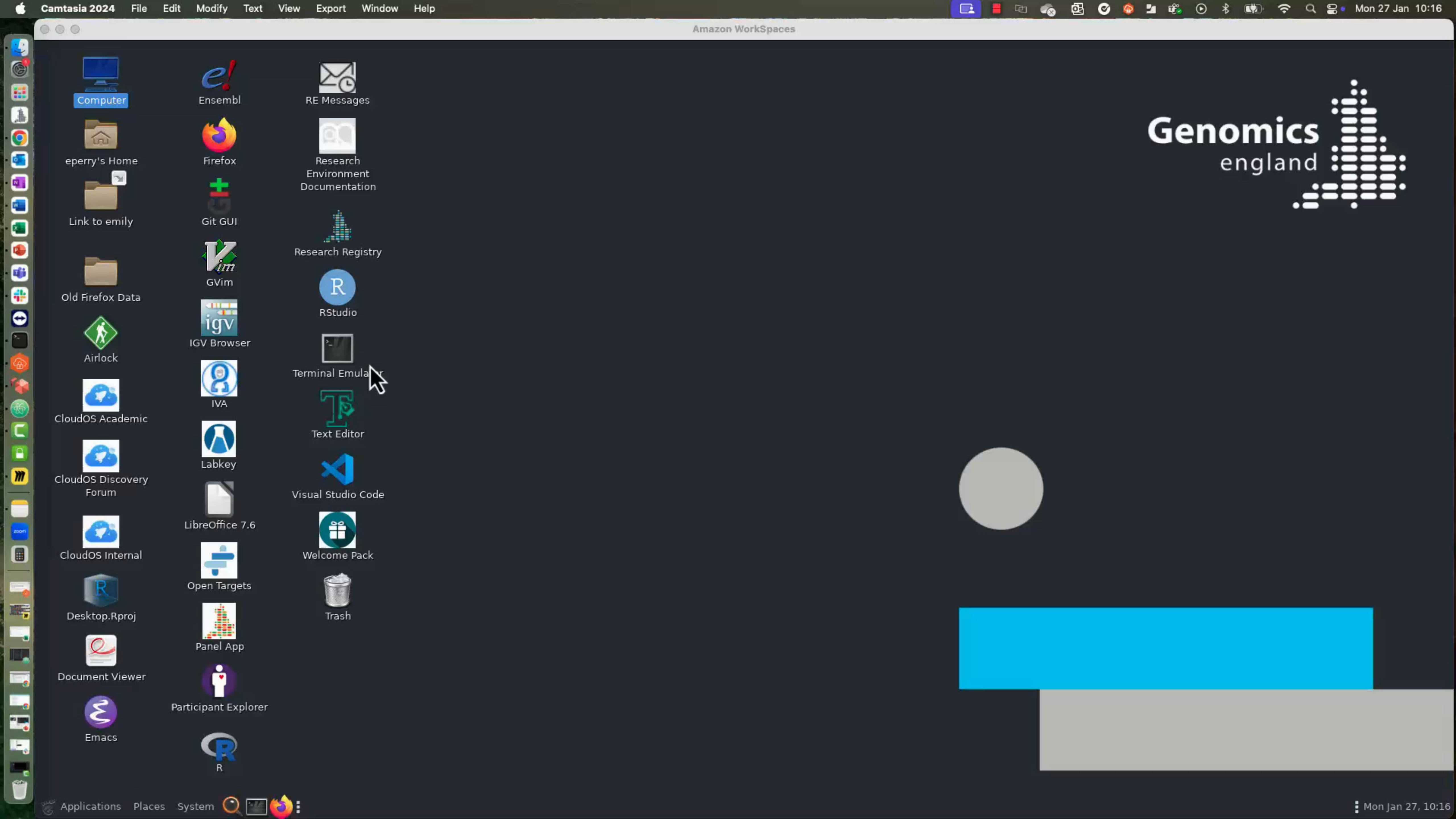
2. Pre-built conda environments in the Research Environment

Python packages in conda environments

idppy3
idppy2
ipy3nopypirev1
ipy3pypirev1
ipy3nopypi
ipy3pypi
ipy3tf2
ipy3keras
ipy3mavis
py3nopypirev1
py3nopypi
py3pypi
py2_7_12nopypi
py2_7_12pypi
idpcorepy3_6_5rev1
idpcorepy2_7_12rev1
idpcorepy3tf2rev1

https://re-docs.genomicsengland.co.uk/hpc_conda

Conda demo


Camtasia 2024
File
Edit
Modify
Text
View
Export
Window
Help
Mon 27 Jan 10:16
Amazon WorkSpaces
Computer
eperry's Home
Link to emily
Old Firefox Data
Airlock
CloudOS Academic
CloudOS Discovery Forum
CloudOS Internal
Desktop.Rproj
Document Viewer
Emacs
Ensembl
Firefox
Git GUI
GVim
IGV Browser
IVA
Labkey
LibreOffice 7.6
Open Targets
Panel App
Participant Explorer
R
RE Messages
Research Environment Documentation
Research Registry
RStudio
Terminal Emulator
Text Editor
Visual Studio Code
Welcome Pack
Trash
Genomics england
Applications
Places
System
Mon Jan 27, 10:16

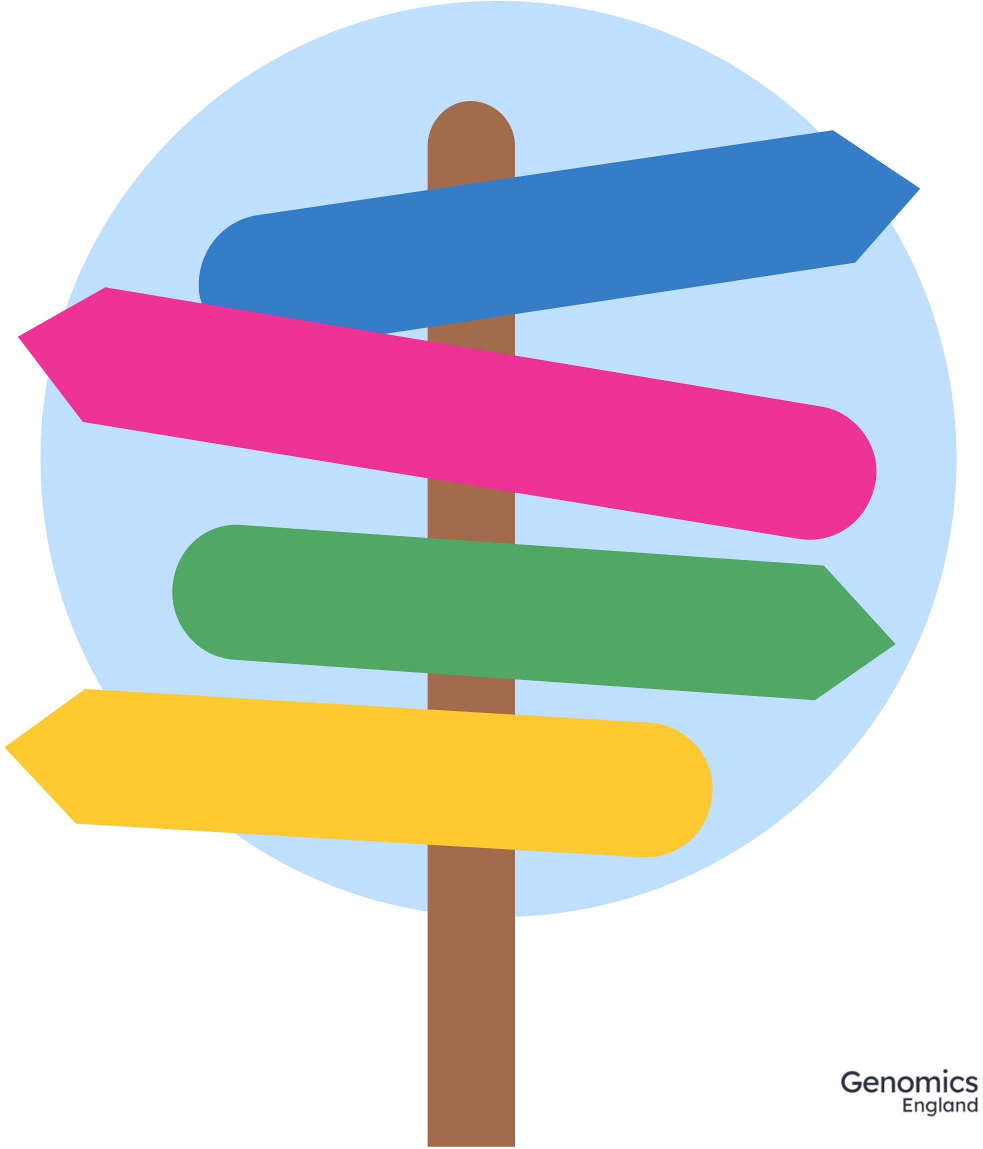
3. Create conda environments in the Research Environment

Authorised channels

anaconda

bioconda

conda-forge



https://re-docs.genomicsengland.co.uk/hpc_conda/

Create an environment

Copy config file to your folder:

```
cp /gel_data_resources/example_config_files/Helix/.condarc ~/.
```

Create your environment

```
conda create python==<version_number> --prefix
/path/to/env/location
```

Use the env path to activate:

```
source /resources/conda/miniconda3/bin/activate
conda activate /path/to/env/location
```

https://re-docs.genomicsengland.co.uk/hpc_conda/

Add packages

with conda (old style)

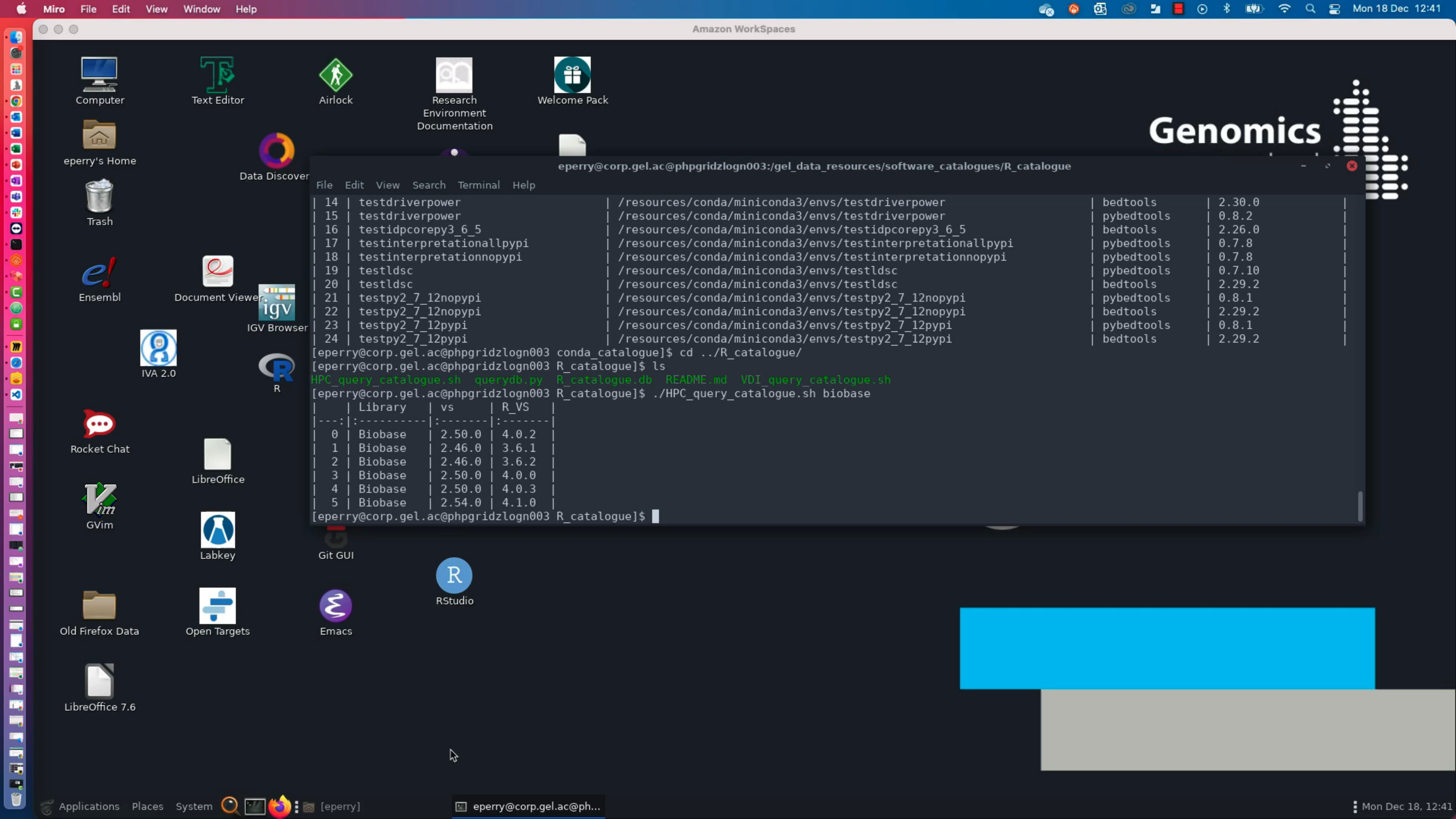
```
conda install -c conda-anaconda <package_1> -c conda-bioconda
<package_2> -conda-conda-forge <package_3> -c conda-r <package_4>
```

with pip (via artifactory)

```
pip install <package_name> --index-url
https://artifactory.aws.gel.ac/artifactory/api/pypi/pypi/simple
```

https://re-docs.genomicsengland.co.uk/hpc_conda/

Conda demo



```
| 14 | testdriverpower                     | /resources/conda/miniconda3/envs/testdriverpower                      | bedtools       | 2.30.0            |
| 15 | testdriverpower                     | /resources/conda/miniconda3/envs/testdriverpower                      | pybedtools     | 0.8.2             |
| 16 | testidpcorepy3_6_5                  | /resources/conda/miniconda3/envs/testidpcorepy3_6_5                   | bedtools       | 2.26.0            |
| 17 | testinterpretationallpypi           | /resources/conda/miniconda3/envs/testinterpretationallpypi            | pybedtools     | 0.7.8             |
| 18 | testinterpretationnopypi            | /resources/conda/miniconda3/envs/testinterpretationnopypi             | pybedtools     | 0.7.8             |
| 19 | testldsc                            | /resources/conda/miniconda3/envs/testldsc                             | pybedtools     | 0.7.10            |
| 20 | testldsc                            | /resources/conda/miniconda3/envs/testldsc                             | bedtools       | 2.29.2            |
| 21 | testpy2_7_12nopypi                  | /resources/conda/miniconda3/envs/testpy2_7_12nopypi                   | pybedtools     | 0.8.1             |
| 22 | testpy2_7_12nopypi                  | /resources/conda/miniconda3/envs/testpy2_7_12nopypi                   | bedtools       | 2.29.2            |
| 23 | testpy2_7_12pypi                    | /resources/conda/miniconda3/envs/testpy2_7_12pypi                     | pybedtools     | 0.8.1             |
| 24 | testpy2_7_12pypi                    | /resources/conda/miniconda3/envs/testpy2_7_12pypi                     | bedtools       | 2.29.2            |
[eperry@corp.gel.ac@phpgridzlogn003 conda_catalogue]$ cd ../R_catalogue/
[eperry@corp.gel.ac@phpgridzlogn003 R_catalogue]$ ls
HPC_query_catalogue.sh  querydb.py  R_catalogue.db  README.md  VDI_query_catalogue.sh
[eperry@corp.gel.ac@phpgridzlogn003 R_catalogue]$ ./HPC_query_catalogue.sh biobase
|    | Library   | vs      | R_VS    |
|---:|:----------|:--------|:--------|
|  0 | Biobase   | 2.50.0  | 4.0.2   |
|  1 | Biobase   | 2.46.0  | 3.6.1   |
|  2 | Biobase   | 2.46.0  | 3.6.2   |
|  3 | Biobase   | 2.50.0  | 4.0.0   |
|  4 | Biobase   | 2.50.0  | 4.0.3   |
|  5 | Biobase   | 2.54.0  | 4.1.0   |
[eperry@corp.gel.ac@phpgridzlogn003 R_catalogue]$ 
```

4. Working with Jupyter on the HPC

Jupyter in the RE

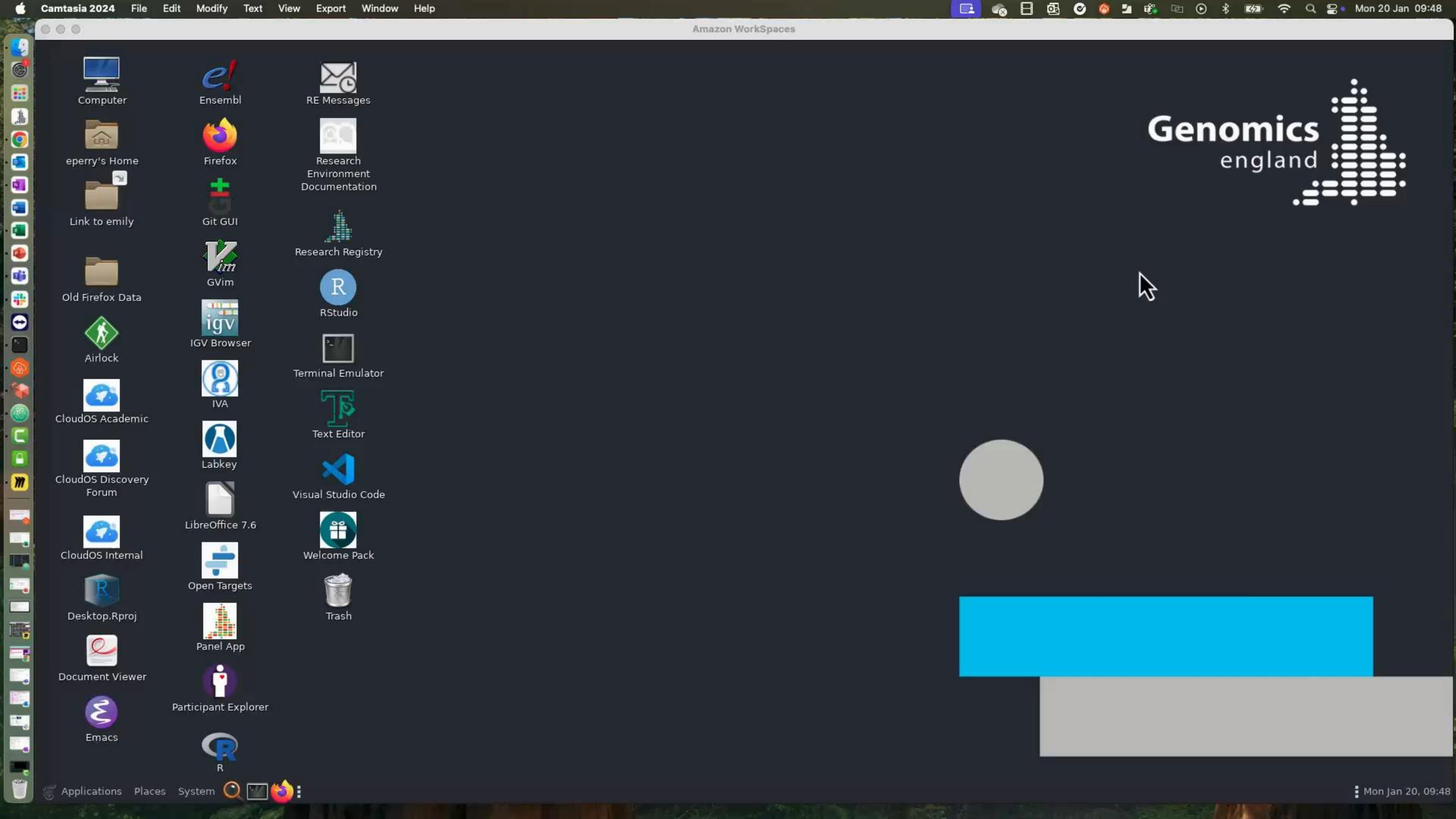
Launch from command line in VDI

Launch interactive job on HPC

Launch Jupyter on HPC

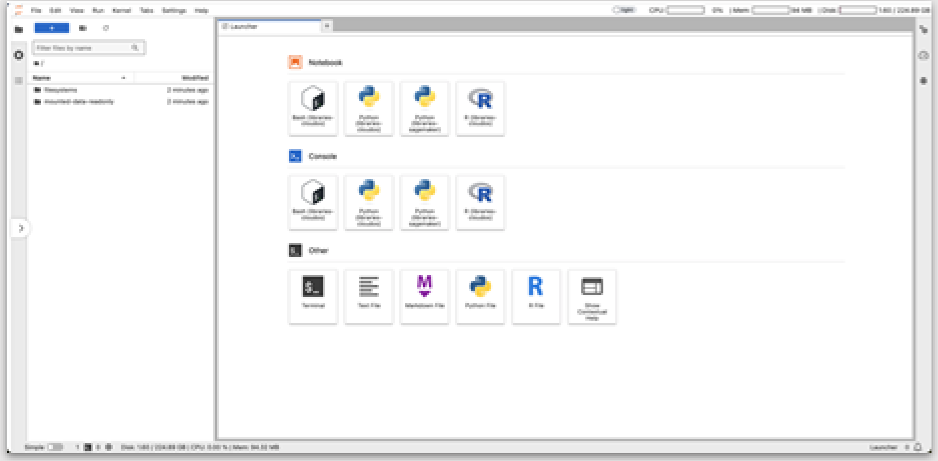
Tunnel in to open notebook

Jupyter demo


Camtasia 2024
File
Edit
Modify
Text
View
Export
Window
Help
Mon 20 Jan 09:48
Amazon WorkSpaces
Computer
eperry's Home
Link to emily
Old Firefox Data
Airlock
CloudOS Academic
CloudOS Discovery Forum
CloudOS Internal
Desktop.Rproj
Document Viewer
Emacs
Ensembl
Firefox
Git GUI
GVim
IGV Browser
IVA
Labkey
LibreOffice 7.6
Open Targets
Panel App
Participant Explorer
R
RE Messages
Research Environment Documentation
Research Registry
RStudio
Terminal Emulator
Text Editor
Visual Studio Code
Welcome Pack
Trash
Genomics
england
Applications
Places
System
Mon Jan 20, 09:48

5. Working with Jupyter notebooks in CloudOS interactive sessions

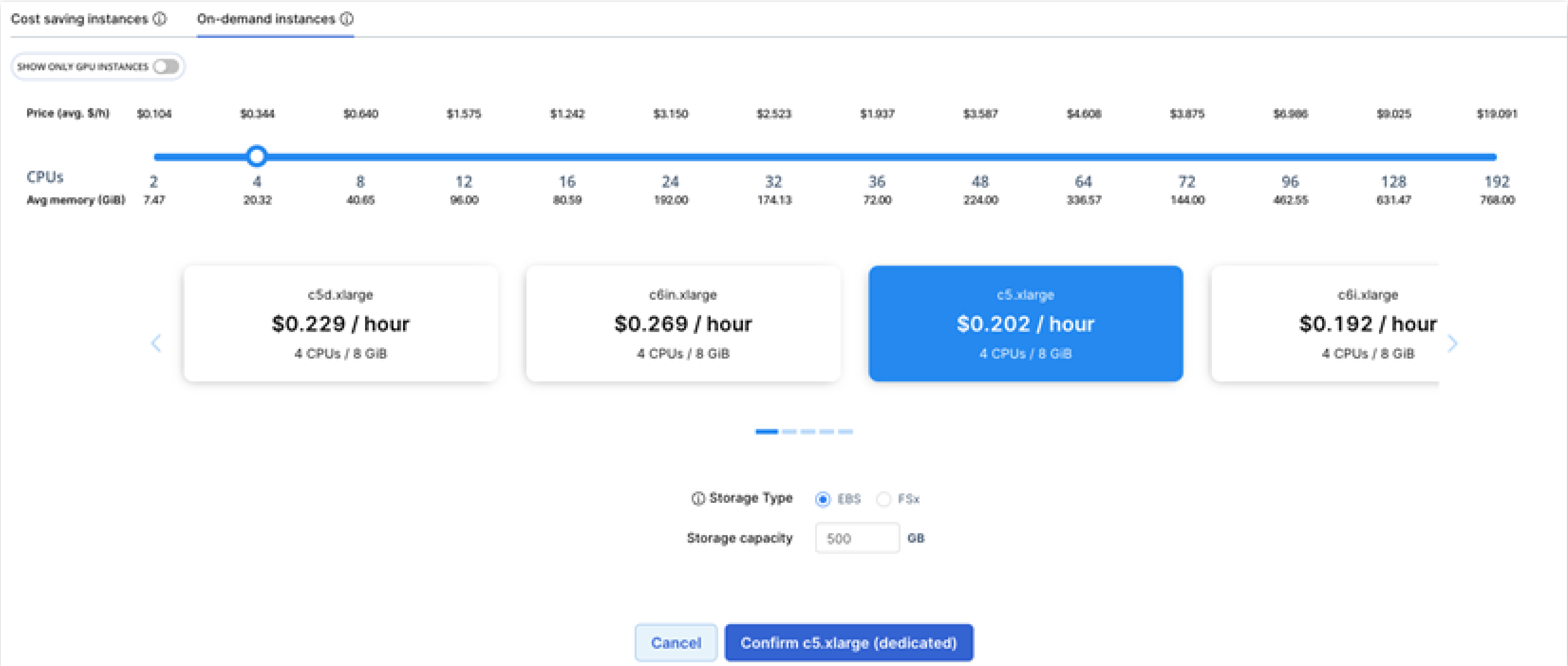
Command line and GUI



Compute options

Scale your compute resources to match your data size and analysis needs

GPUs available on demand



Price (avg. $/h)	$0.104	$0.344	$0.640	$1.575	$1.242	$3.150	$2.523	$1.937	$3.587	$4.608	$3.875	$6.986	$9.025	$19.091
CPUs	2	4	8	12	16	24	32	36	48	64	72	96	128	192
Avg memory (GiB)	7.47	20.32	40.65	96.00	80.59	192.00	174.13	72.00	224.00	336.57	144.00	462.55	631.47	768.00

c5d.xlarge
$0.229 / hour
4 CPUs / 8 GiB

c6in.xlarge
$0.269 / hour
4 CPUs / 8 GiB

c5.xlarge
$0.202 / hour
4 CPUs / 8 GiB

c6i.xlarge
$0.192 / hour
4 CPUs / 8 GiB

ⓘ Storage Type ● EBS ○ FSx

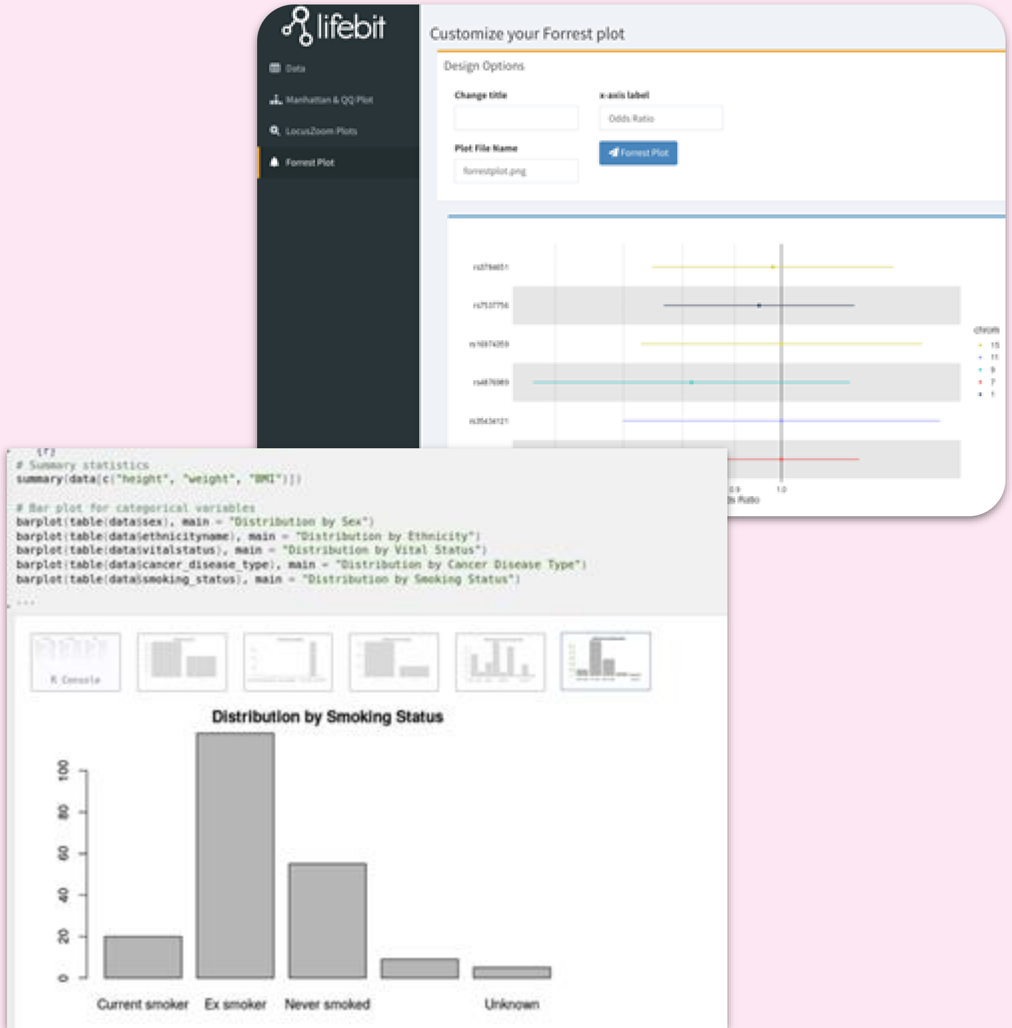
Storage capacity: 500 GB

Installing packages

- Full flexibility with installing packages from pip, biopython and Conda
- Save snapshots of environments

Develop, run and share code

- Write scripts, notebooks, apps and more…
- Collaborate in real time with multiple users
- Share code with others in your workspace

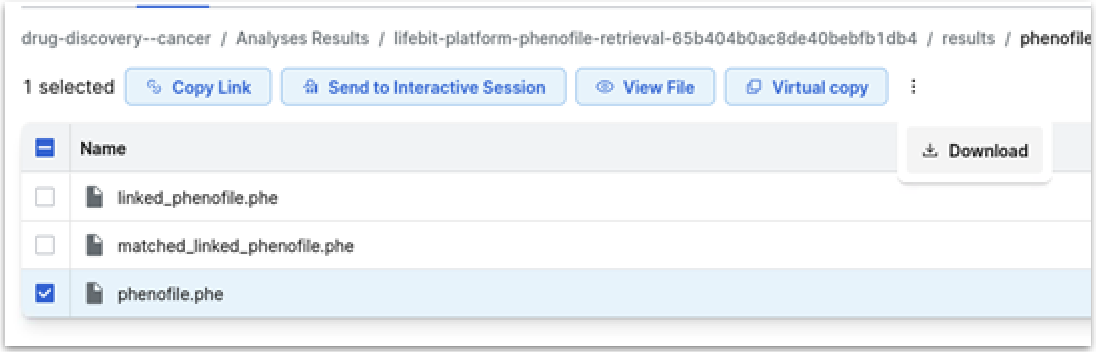


~~~
{r}
# Summary statistics
summary(data[c("height", "weight", "BMI")])

# Bar plot for categorical variables
barplot(table(data$sex), main = "Distribution by Sex")
barplot(table(data$ethnicityname), main = "Distribution by Ethnicity")
barplot(table(data$vitalstatus), main = "Distribution by Vital Status")
barplot(table(data$cancer_disease_type), main = "Distribution by Cancer Disease Type")
barplot(table(data$smoking_status), main = "Distribution by Smoking Status")

```
~~~

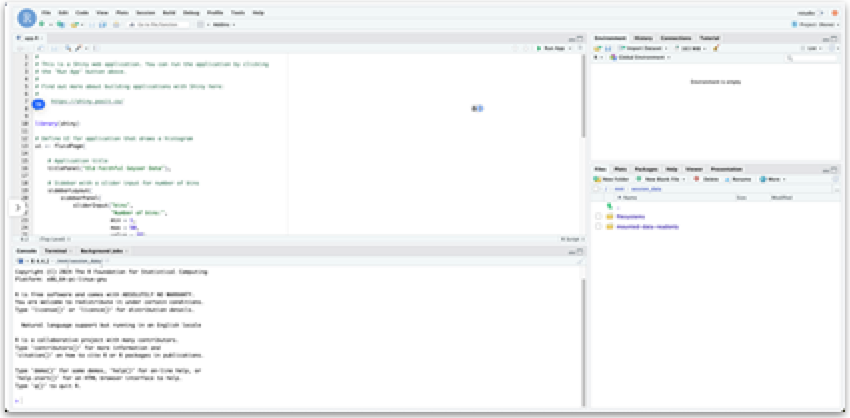
Access your results


drug-discovery--cancer / Analyses Results / lifebit-platform-phenofile-retrieval-65b404b0ac8de40bebfb1db4 / results / phenofile

1 selected

☐	Name
☐	linked_phenofile.phe
☐	matched_linked_phenofile.phe
☑	phenofile.phe

Interactive session › File Explorer › Download to VDI › Airlock



CloudOS demo

6. Query clinical data with Python

LabKey

- Central database of:
    - Clinical data
    - Results of bioinformatics analysis
    - Locations of genomic files
- Point and click interface
- API

https://re-docs.genomicsengland.co.uk/labkey/

LabKey API

Combine queries between tables

Work in a variety of programming languages (support for Python and R) using SQL queries

Replicate queries between releases and analyses

Work locally and on the HPC

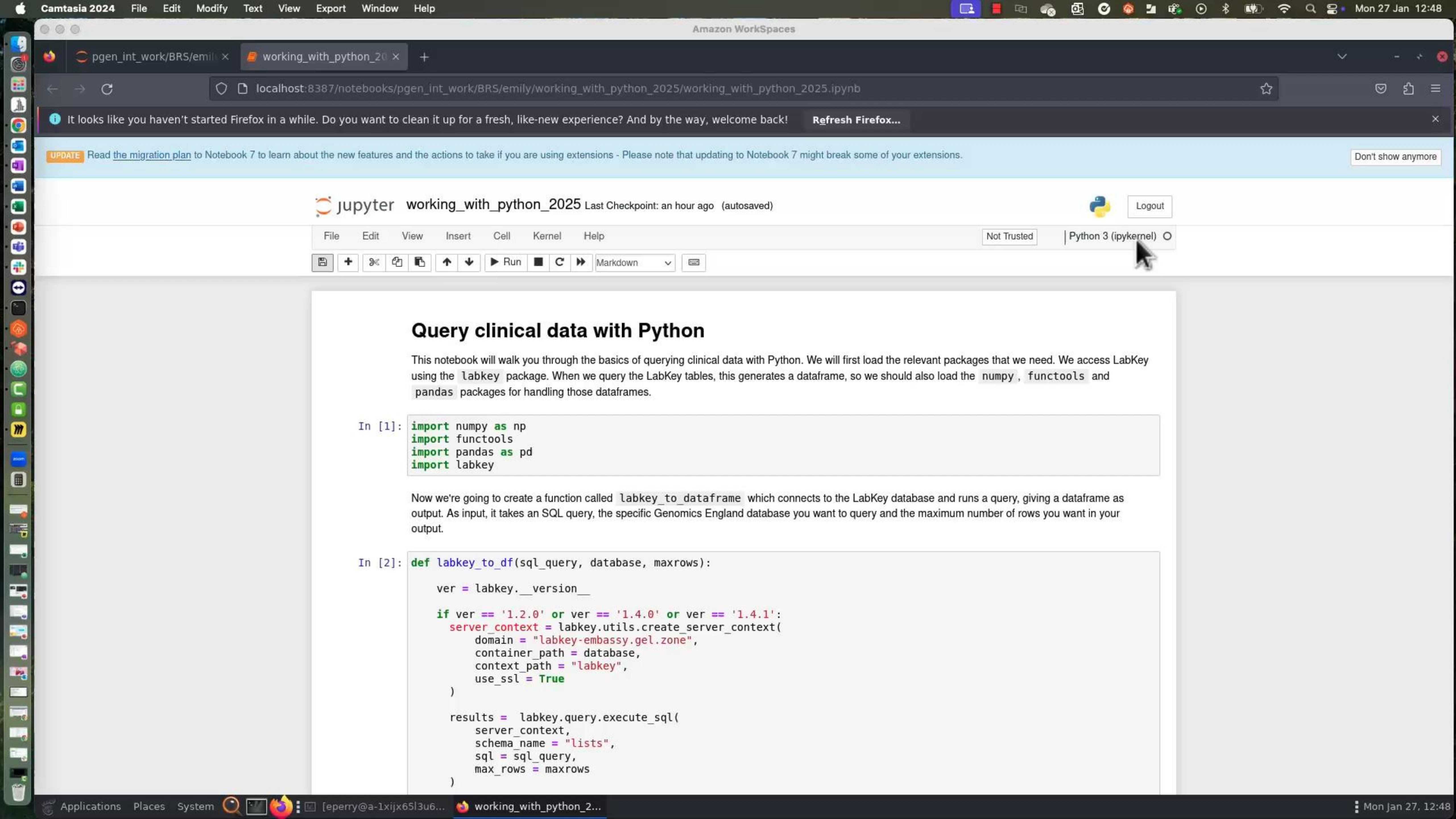
https://re-docs.genomicsengland.co.uk/labkey_api/

LabKey .netrc

- You can access the same data via the LabKey API as you can through other means
- You will need to configure access to the LabKey API with your username and password
    - In your home directory
    - On the HPC
- You do this by editing a file called .netrc

https://re-docs.genomicsengland.co.uk/labkey_api/

LabKey API demo



It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!

Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

working_with_python_2025 Last Checkpoint: an hour ago (autosaved)

Query clinical data with Python

This notebook will walk you through the basics of querying clinical data with Python. We will first load the relevant packages that we need. We access LabKey using the `labkey` package. When we query the LabKey tables, this generates a dataframe, so we should also load the `numpy`, `functools` and `pandas` packages for handling those dataframes.

In [1]:

```
import numpy as np
import functools
import pandas as pd
import labkey
```

Now we're going to create a function called `labkey_to_dataframe` which connects to the LabKey database and runs a query, giving a dataframe as output. As input, it takes an SQL query, the specific Genomics England database you want to query and the maximum number of rows you want in your output.

In [2]:

```
def labkey_to_df(sql_query, database, maxrows):

    ver = labkey.__version__

    if ver == '1.2.0' or ver == '1.4.0' or ver == '1.4.1':
      server_context = labkey.utils.create_server_context(
          domain = "labkey-embassy.gel.zone",
          container_path = database,
          context_path = "labkey",
          use_ssl = True
      )

      results =  labkey.query.execute_sql(
          server_context,
          schema_name = "lists",
          sql = sql_query,
          max_rows = maxrows
      )
```

7. Getting help and questions

Getting help

Check our documentation:
https://re-docs.genomicsengland.co.uk/
Click on the documentation icon in the environment

Contact our Service Desk:
https://jiraservicedesk.extge.co.uk/plugins/servlet/desk

Training sessions

3rd Tuesday every month

Introduction to the RE

15/4
20/5
22/7
19/8
16/9
21/10

https://re-docs.genomicsengland.co.uk/upcoming/

Materials from past training all online

Training sessions

13/5	Building cancer cohorts and survival analysis
10/6	Building rare disease cohorts with matching controls
8/7	Finding participants based on genotypes
9/9	Getting medical records for participants
14/10	What tools and workflows should I use to fulfil an overall goal?
11/11	Using GEL data for publications and reports

https://re-docs.genomicsengland.co.uk/upcoming/

Materials from past training all online

Genomics England

Research Summit 2025

Tuesday 17 June 2025

Feedback

Thank you

Visit: https://re-docs.genomicsengland.co.uk/