Importing data and tools to use in the RE

Emily Perry

Research Engagement Manager

11th February 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

Hamzah Syed
Solutions Manager - 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

Questions

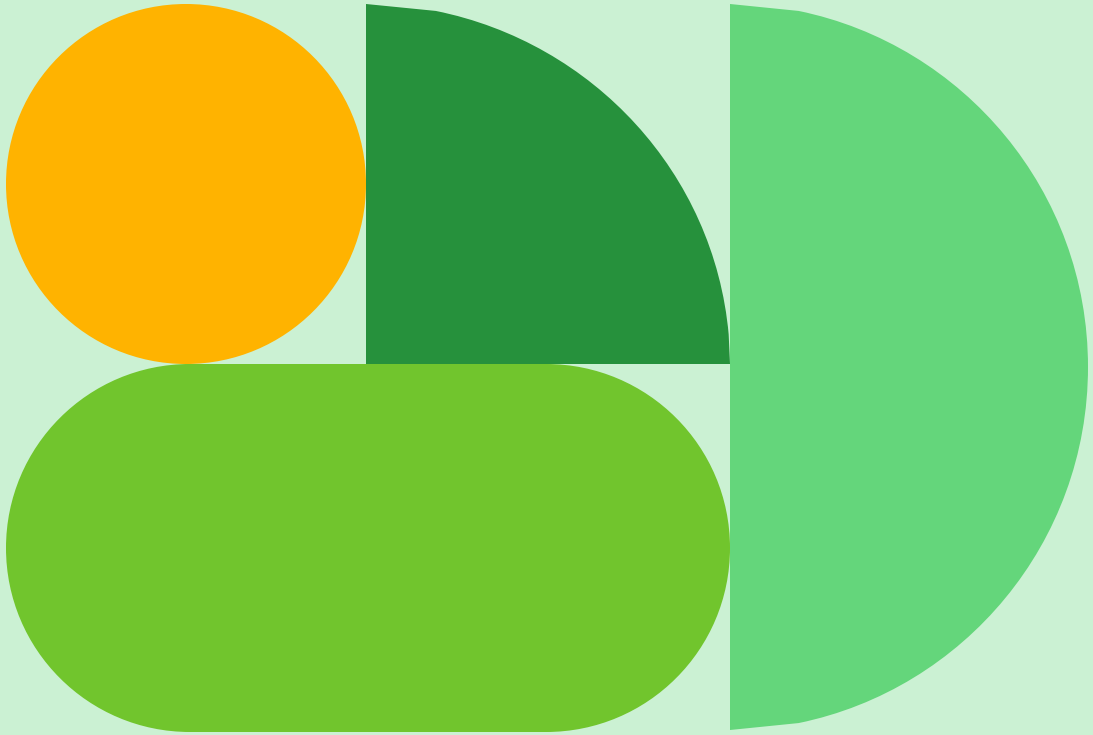
Miruna Carmen Barbu
Bioinformatician - Research Services

Eleni Kyriakou
Senior Client Success Manager - LifeBit

Eleni Christodoulou
Client Success Manager - LifeBit

Agenda

1. Introduction and admin
2. What is already available in the RE?
3. Personal conda environments
4. Importing R packages with CRAN and Bioconductor
5. Importing containers with Singularity
6. Using Airlock to bring data/software in
7. Making a software request
8. CloudOS – importing tools and pipelines on the Cloud
9. Software licensing and requirements
10. Help and questions



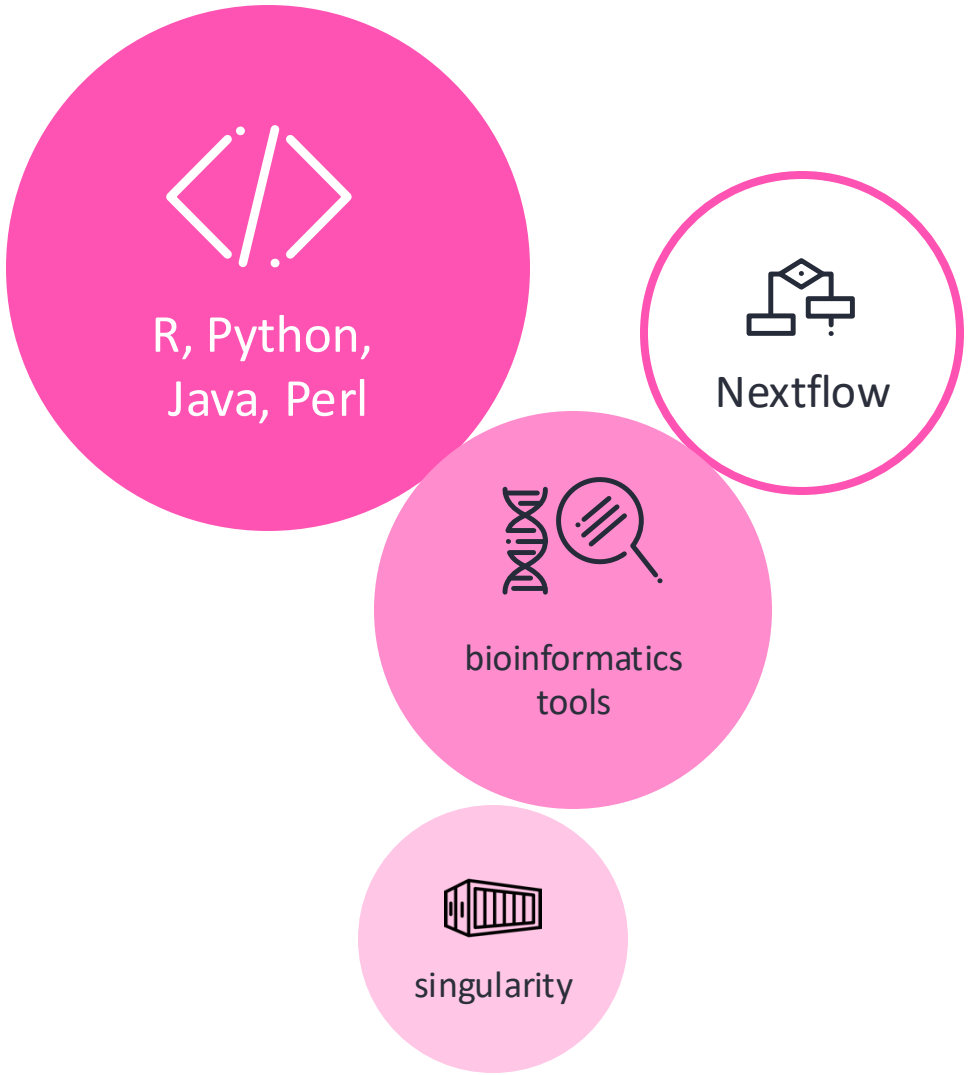
2. What is already in available the RE?

Public data

1000 Genomes
AmpliconArchitect
Annovar database
Battenberg
bigzips
BLAST
CADD
Centrifuge database
Centromeres and telomeres
Clair
ClinVar
COSMIC
dbNSFP
dbSNP
deepVariant
InterPro
Ensembl
eQTLGen
ExAC
Exomiser
ExpansionHunter
FannsDB
FAVORannotator
GATK
GENCODE
GERP
Genome in a Bottle
gnomAD
GTEx
HGDP
IGV
IMPC
Kraken
Liftover
mhcflurry
MutSigCV
NanoPlot
Nanopore WGS consortium
NCBI BLAST
Nextflow DB
GWAS Catalog
OncodriveFML
PathSeq
Phylop 100 way
Picard
Reference genomes
R packages
Rvtests
SAIGE
SeqSeek
SGDP
Slivar
SpliceAI
TCGA PCAWG
TOPMed
UCSC
VEP
wwPDB

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

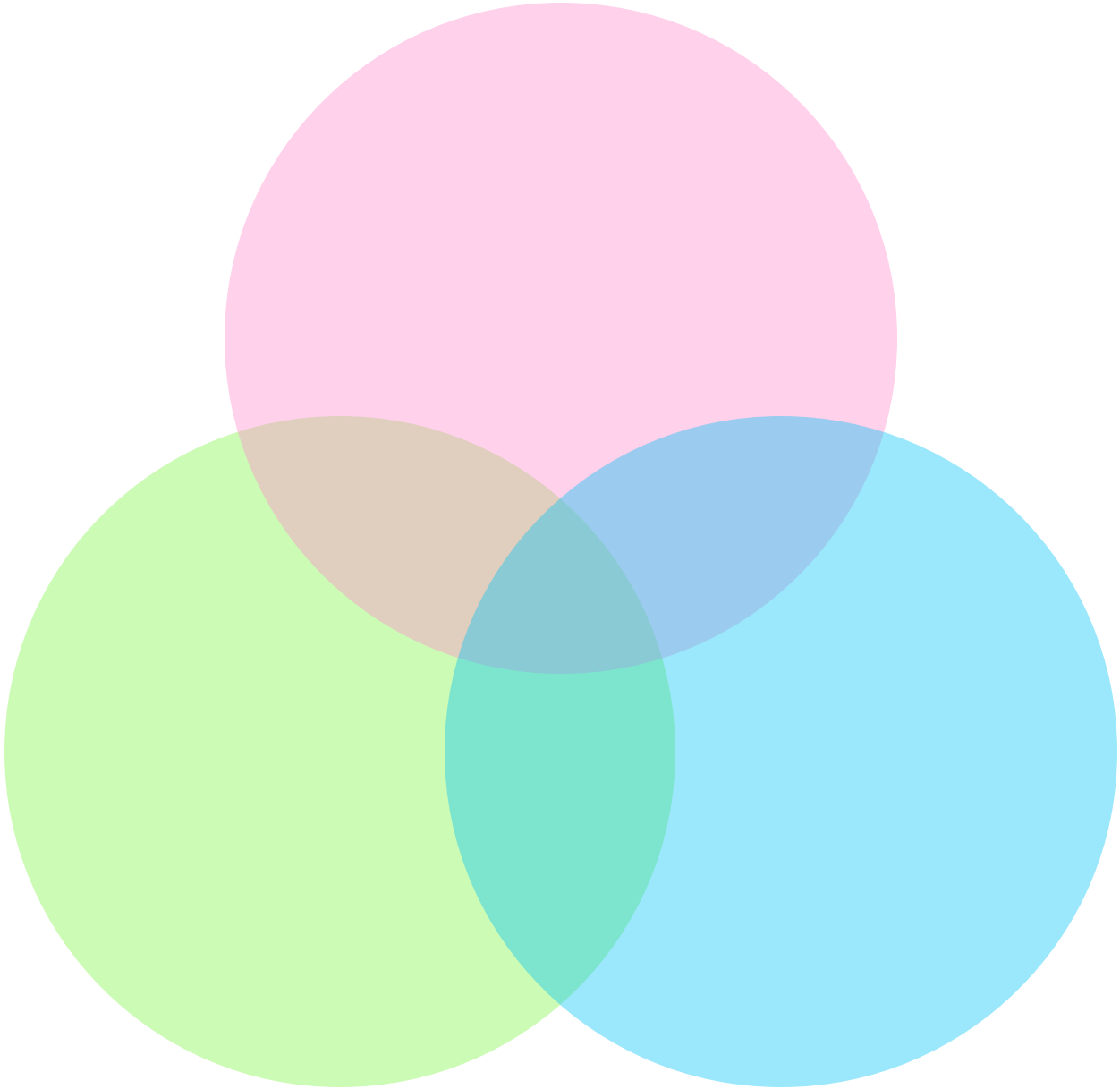
Software on the HPC



AdapterRemoval/2.3.3
aliview/1.28
ampliconArchitect/1.3.r7
ampliconClassifier/1.1.1
annotSV/3.3.7
annovar/2019Nov
annovar/2024-03-14
ant/1.9.16
apbs/3.4.1
asmc-asmc/2024-02-26
AutoDock_Vina/1.2.5
automake/1.15
aws-cli/2.15
bamtools/2.5.2
bcftools/1.16
beagle/5.4
bedops/2.4.41
bedtools/2.30.0
bedtools/2.31.0
BerkeleyDB/3.01
Bio-DB-HTS/3.01
blast+/2.15
blat/1.0
bolt-lmm/2.4.1
boost/1.83
bowtie2/2.5.2
BWA/0.7.17
CADD/1.6
canvas/1.40.0.1613
CaVEMan/1.15.3
circos/0.69-9
clang/16.0.6
cmake/3.24.3
CNView/1.0
CNVnator/0.4.1
cpan/1.7047
cromwell/v65
curl/7.81.0
cython/3.0.8
cytoscape/3.10.1
delly/1.2.6
denovoGear/1.1.1
discover/0.9.5
dotnet/2.0.0
dotnet/8.0.1
drop/1.2.4
eigen/3.3.9
exomiser/13.3.0
exonerate/2.2.0
ExpansionHunter/3.2.2
ExpansionHunter/4.0.2
ExpansionHunterDenovo/0.9.0
fastqc/0.12.1
fetk/1.9.3
ffmpeg/6.0
fribidi/1.0.12
gatk/4.5.0.0
gauchian/1.0.2
gcc/10.4.0
gcta/1.94
gdal/3.7.0
geos/3.12.1
gistic/2.0.23
gmp/6.2.1
gnu-parallel/20190222
gnu/4.4
gradle/8.5
GSL/2.7
guppy/3.4.5
gvcfgenotyper/2019.02.26
haplocheck/1.3.3
hipstr/0.7
hisat2/2.2.1
hla-la/1.0.3
hmftools/2024-02-06
homer/4.11
htslib/1.18
igv/2.17.1
imagemagick/7.1.0
java/1.8
java/11.0.2
java/17.0.2
java/19.0.2
jq/1.7.1
kallisto/0.50.1
king/2.3.2
kraken/1.1.1
kraken2/2.1.3
lapack/3.12.0
ldsc/1.0.1
ldstore/2.0
libdeflate/1.20
libgit2/1.6.2
libgit2/1.6.2
libtiff/3.4
libtiff/4.3.0
libtiff/4.5.0
libunwind/1.8.0
liftover/1.0
linasm/1.13
llvm/16.0.6
locuszoom/1.4
lollipop/0.3.0
lumpy/0.3.1
mafft/7.520
magma/1.10
manta/1.6.0
matlab/24.1
matlab/8.1
maven/3.9.6
MEDICC2/1.0.2
meme/5.5.5
metal/1.0
miniconda3/23.11.0
miniforge3/23.11.0-0
minimap2/2.26
mosaicHunter/2024-02-14
MPFR/4.2.0
mplayer/1.5
msisensor-pro/1.2.0
msisensor/0.6
multiqc/1.19
music2/0.2
mutserve/2.0.0-rc15
mutsig2cv/3.11
ncurses/6.4
new_fugue/2010-06-02
nextflow/22.10.5
nextflow/23.04
nextflow/23.10
nextflow/23.10-with-plugins
nextflow/24.04.2-with-plugins
nf-core/0.3.1
nf-test/0.7.3
nf-test/0.8.2
nf-test/0.9.0
nodejs/16.9.0
openrefine/3.7.4
openssl/1.1.1o
pandoc/3.3
perl/5.38.2
picard/3.1.1
pindel/0.2.5b8
platypus/0.8.1
plink_seq/0.10
plink/1.9
plink/2.0
plink/2.00a3.3LM
popdel/1.5.0
proj/8.2.1
prsice-2/2.3.5
pycircos/1.0.2
pysam/0.22.0
python/3.11
python/3.8
python/3.8.1
R/3.6.3
R/4.2.1
R/4.3.3
readline/8.0
regenie/3.4.1
repeatDetector/1.0
REViewer/0.2.7
rtg-tools/3.12.1
rvtests/2.1.0
saige/1.0.9
salmon/1.10.0
samtools/1.16.1
shapeit4/4.2.2
singularity/3.8.3
singularity/4.1.1
sniffles/1.0.11
somalier/0.2.19
sqlite3/3.40.0
squirls/2.0.1
stack/2.15.7
star/2.7.11a
star/2.7.2a
strelka/2.9.10
superSTR/1.0.1
svanna/1.0.4
tabix/1.18
trimmomatic/0.39
udunits/2.2.28
vcf2maf/1.6.21
vcfanno/0.3.4
vcflib/1.0.9
vcftools/0.1.16
verifyBamID/2.0.1
vt/0.57721
xz/5.4.7
zlib/1.3
zulu/21.0.1

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

R packages

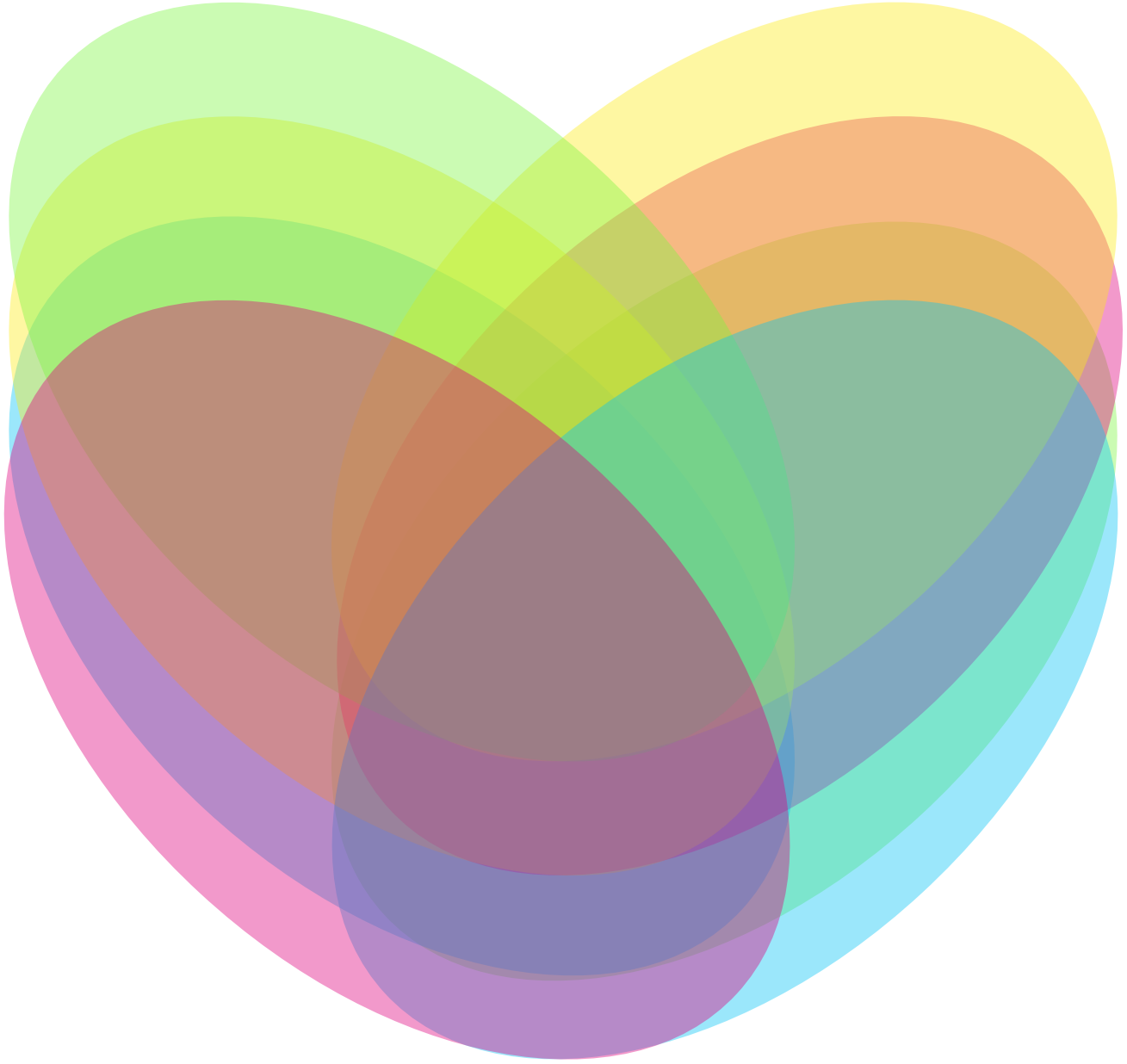


R/3.6.3
R/4.2.1
R/4.3.3

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

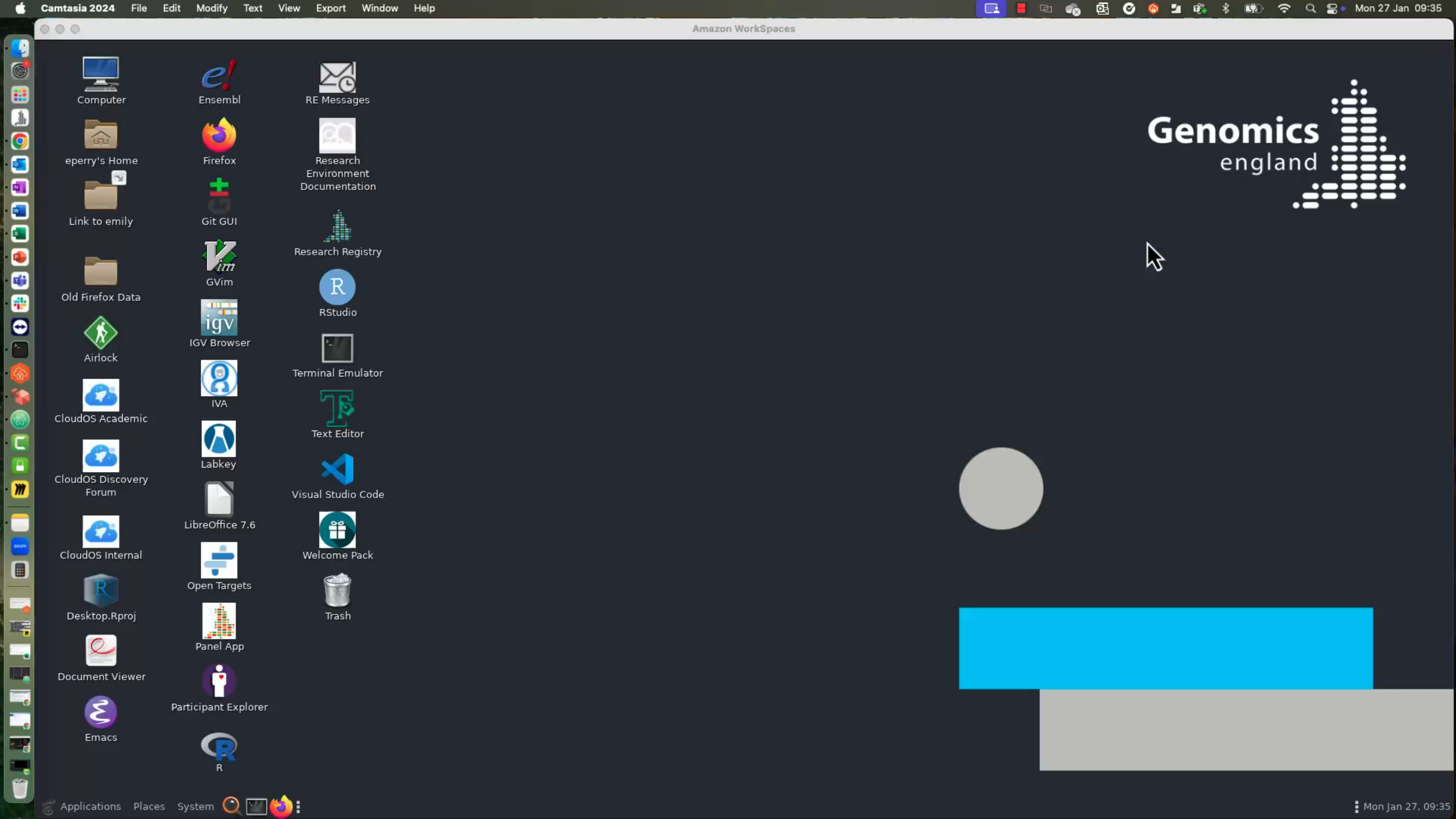
Python packages in conda environments

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

Existing data/tools demo


Camtasia 2024
File
Edit
Modify
Text
View
Export
Window
Help
Mon 27 Jan 09:35
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, 09:35

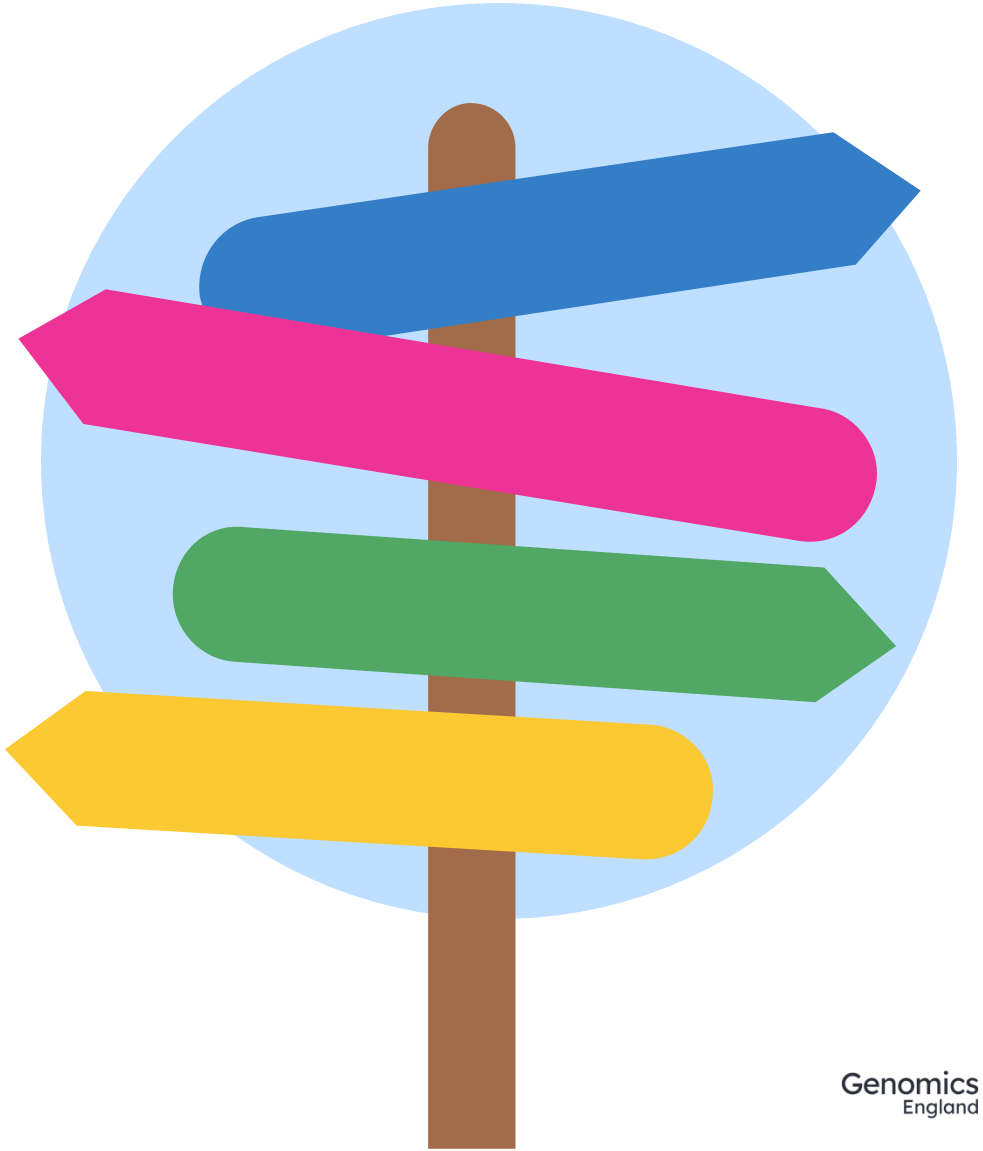
3. Personal conda environments

Authorised channels

anaconda

bioconda

conda-forge



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

Process

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:

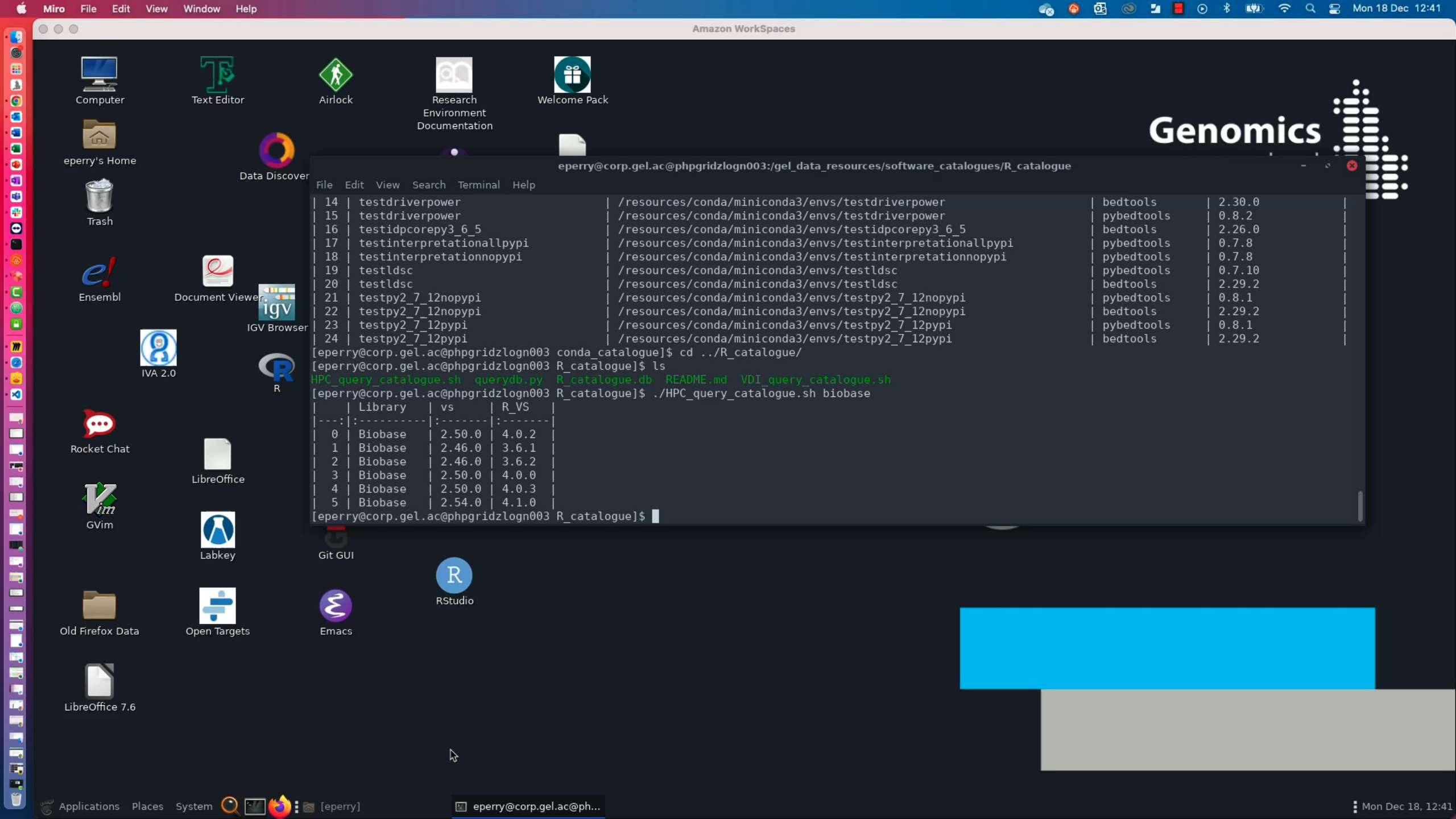
```
source /resources/conda/miniconda3/bin/activate conda activate
/path/to/env/location
```

Use pip to install further packages using a proxy path:

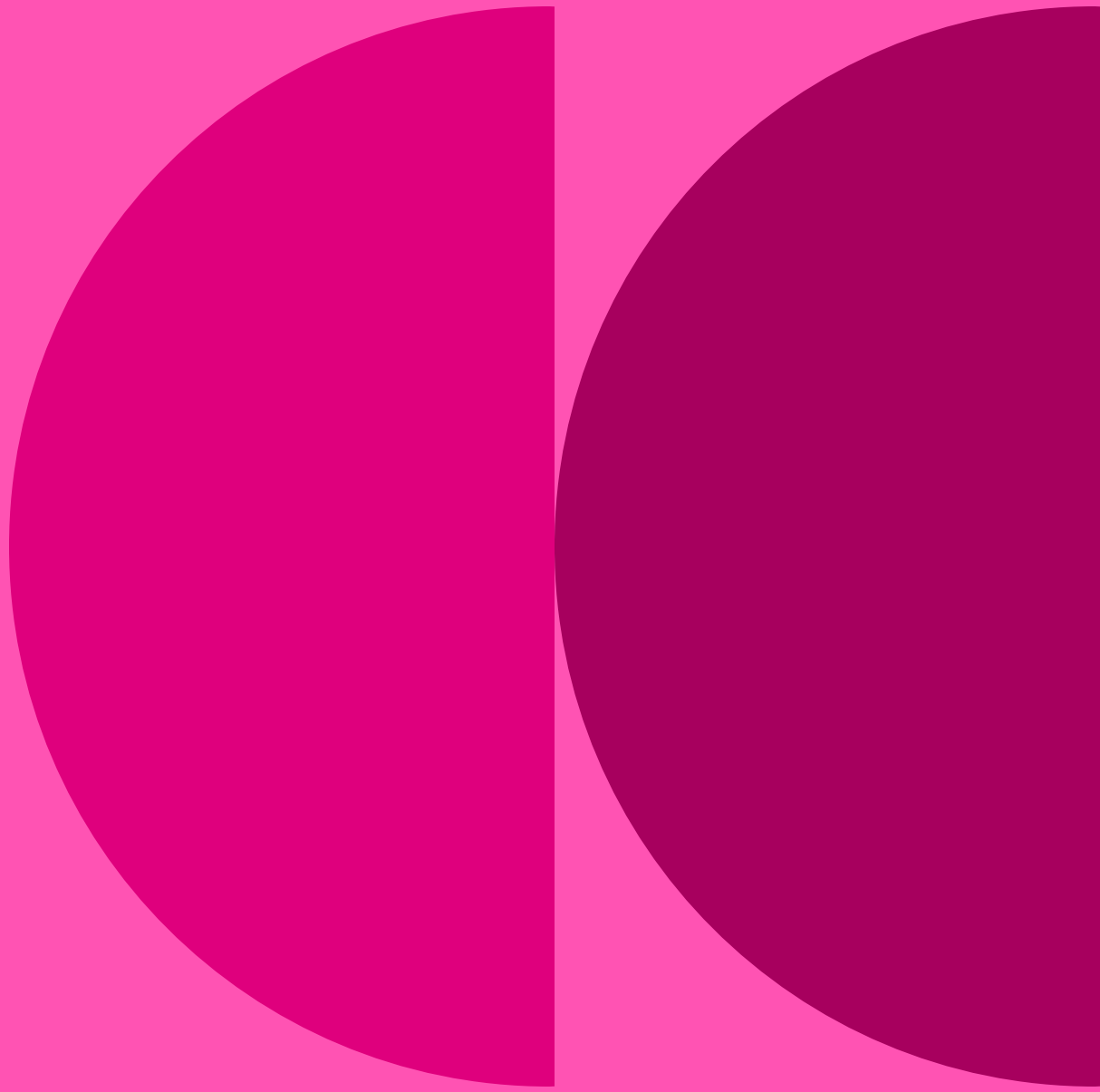
```
pip install <package_name> --index-url
https://artifactory.aws.gel.ac/artifactory/api/pypi/pypi/simple
```

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

Conda demo

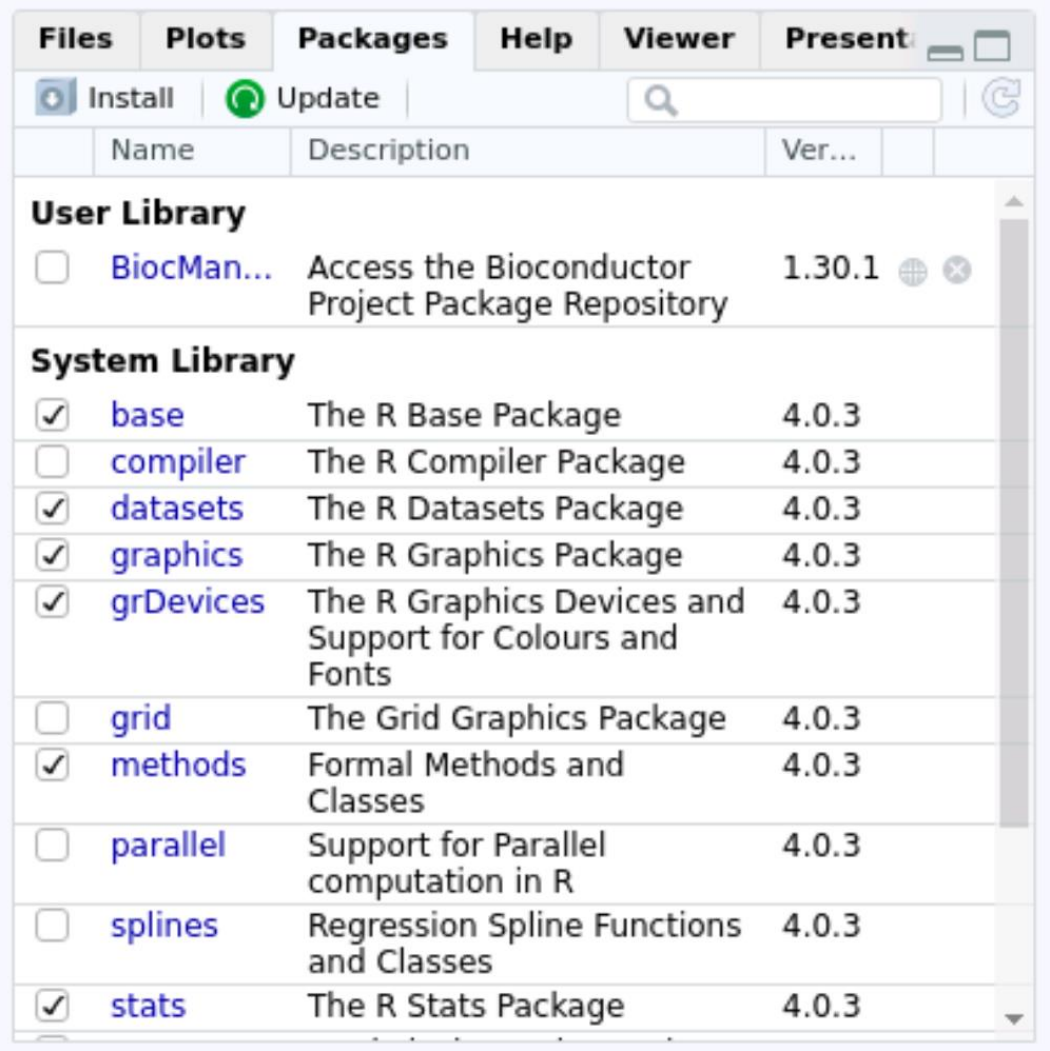

Miro File Edit View Window Help
Mon 18 Dec 12:41
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discover
Trash
Ensembl
Document Viewer
IGV Browser
IVA 2.0
R
Rocket Chat
LibreOffice
GVim
Labkey
Git GUI
RStudio
Old Firefox Data
Open Targets
Emacs
LibreOffice 7.6
Genomics
eperry@corp.gel.ac@phpgridzlogn003:/gel_data_resources/software_catalogues/R_catalogue
File Edit View Search Terminal Help
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]$
Applications Places System [eperry]
eperry@corp.gel.ac@ph...
Mon Dec 18, 12:41

4. Importing R packages with CRAN and Bioconductor



Loading libraries

```
library(library_name)
```



https://re-docs.genomicsengland.co.uk/r_packages/#installing-r-packages-from-cran

Bioconductor

R 4.33

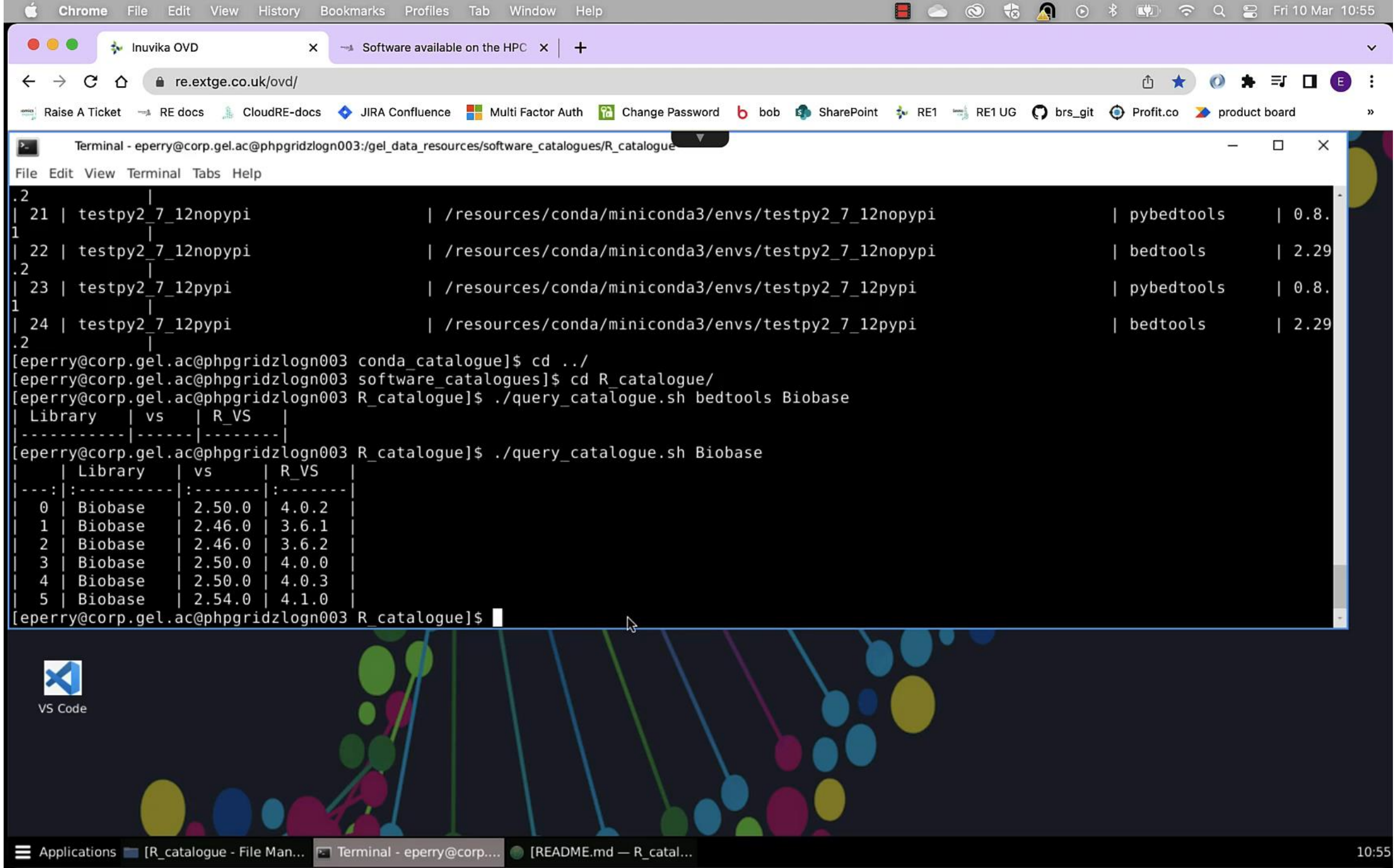
```
options(
BIOCONDUCTOR_CONFIG_FILE =
"https://artifactory.aws.gel.ac:443/artifactory/bioconductor.org-
cache/config.yaml"
)

library("BiocManager")
BiocManager::install("<package_name>")
library(<package_name>)
```

R 3.6.0
to 4.3.2

```
.libPaths(c( .libPaths(), "/tools/aws-workspace-apps/ce/R/<R_version>"))
library("BiocManager")
BiocManager::install("<package_name>")
library(<package_name>)
```

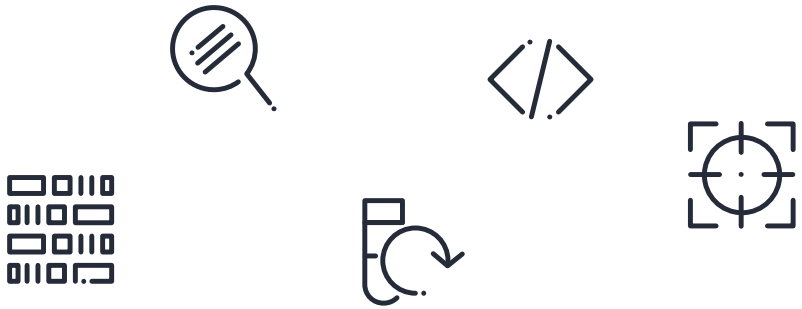
https://re-docs.genomicsengland.co.uk/r/#installing-and-configuring-packages-from-bioconductor

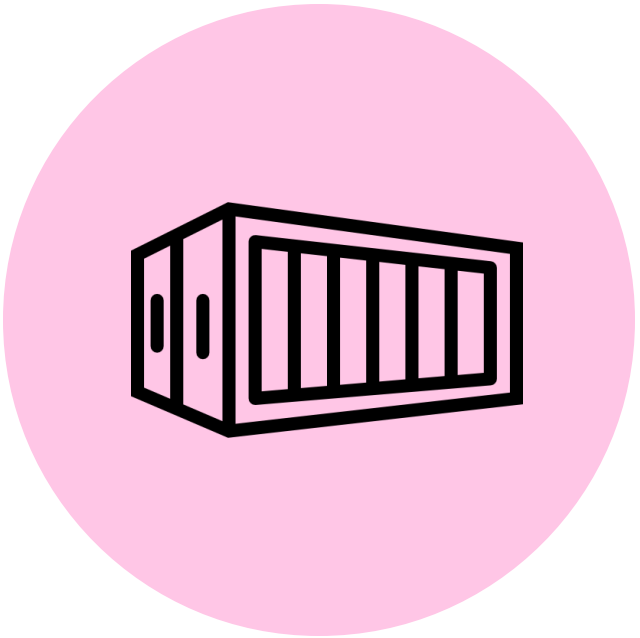
CRAN/Bioconductor demo

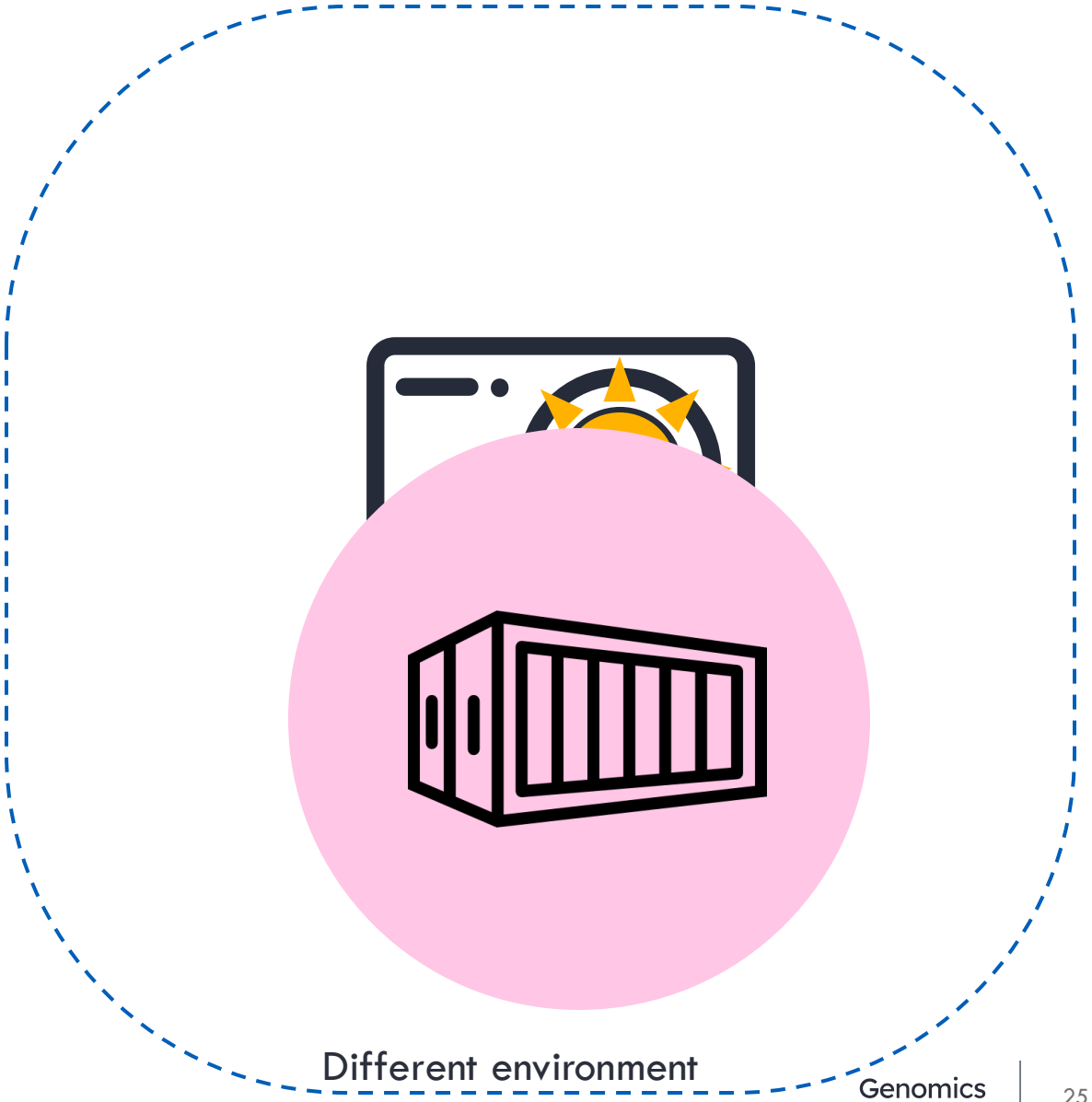

Chrome File Edit View History Bookmarks Profiles Tab Window Help
Fri 10 Mar 10:55
Inuvika OVD
Software available on the HPC
re.extge.co.uk/ovd/
Raise A Ticket
RE docs
CloudRE-docs
JIRA Confluence
Multi Factor Auth
Change Password
bob
SharePoint
RE1
RE1 UG
brs_git
Profit.co
product board
Terminal - eperry@corp.gel.ac@phpgridzlogn003:/gel_data_resources/software_catalogues/R_catalogue
File Edit View Terminal Tabs Help
.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 ../
[eperry@corp.gel.ac@phpgridzlogn003 software_catalogues]$ cd R_catalogue/
[eperry@corp.gel.ac@phpgridzlogn003 R_catalogue]$ ./query_catalogue.sh bedtools Biobase
Library	vs	R_VS
[eperry@corp.gel.ac@phpgridzlogn003 R_catalogue]$ ./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]$
VS Code
Applications
[R_catalogue - File Man...
Terminal - eperry@corp....
[README.md — R_catal...
10:55

5. Importing containers using Singularity

Containers

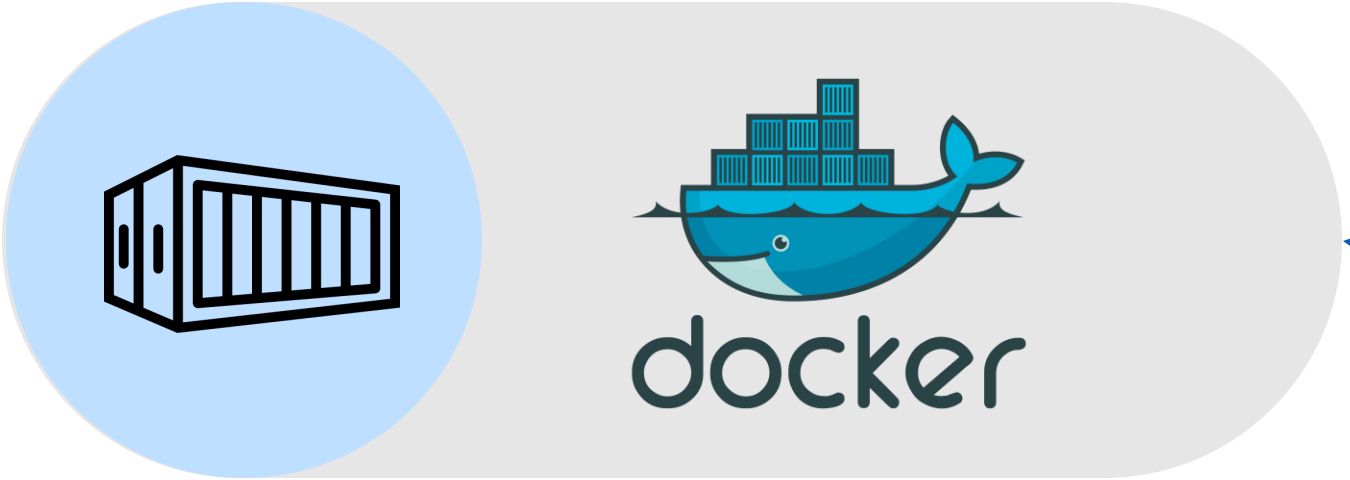






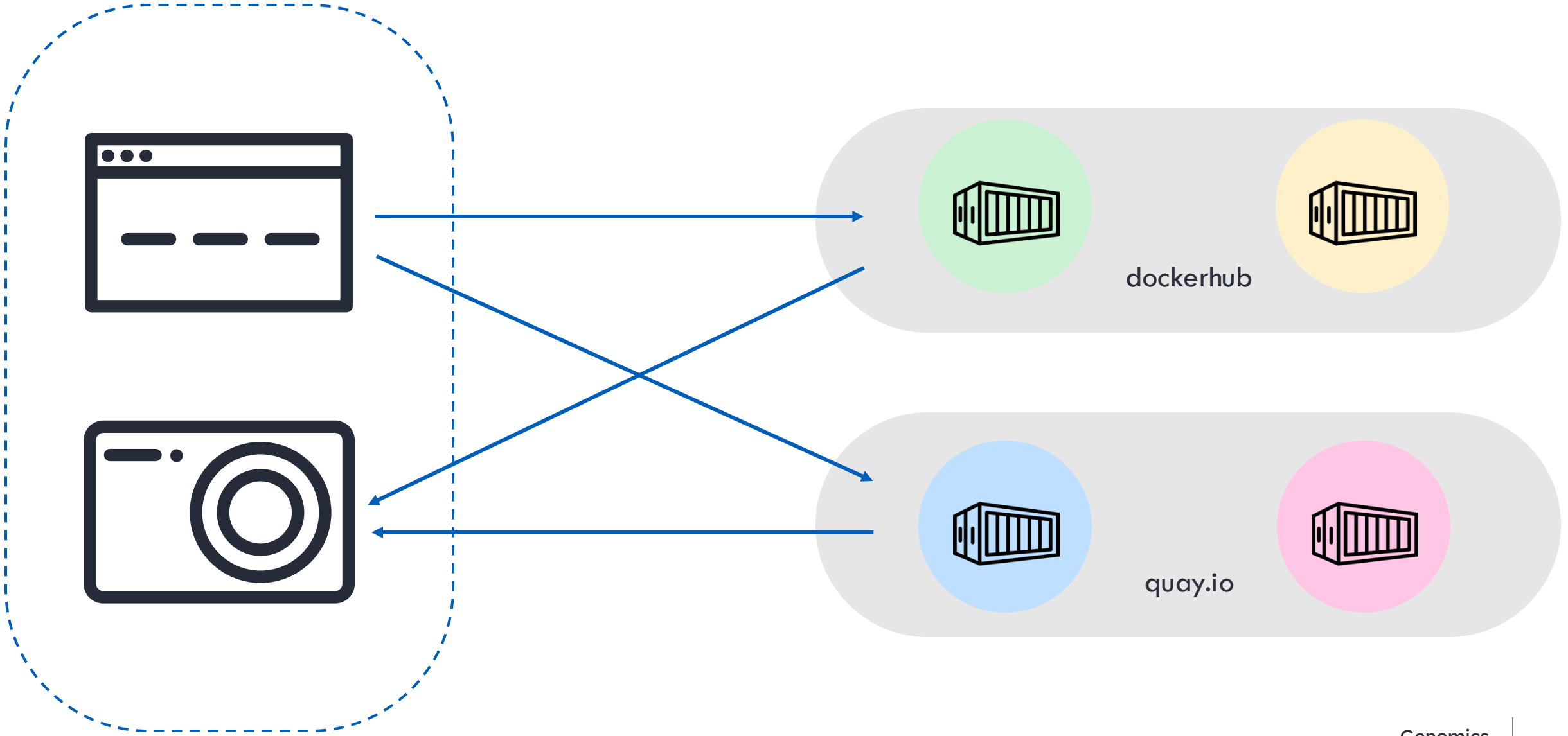
https://re-docs.genomicsengland.co.uk/hpc_containers/

Container types



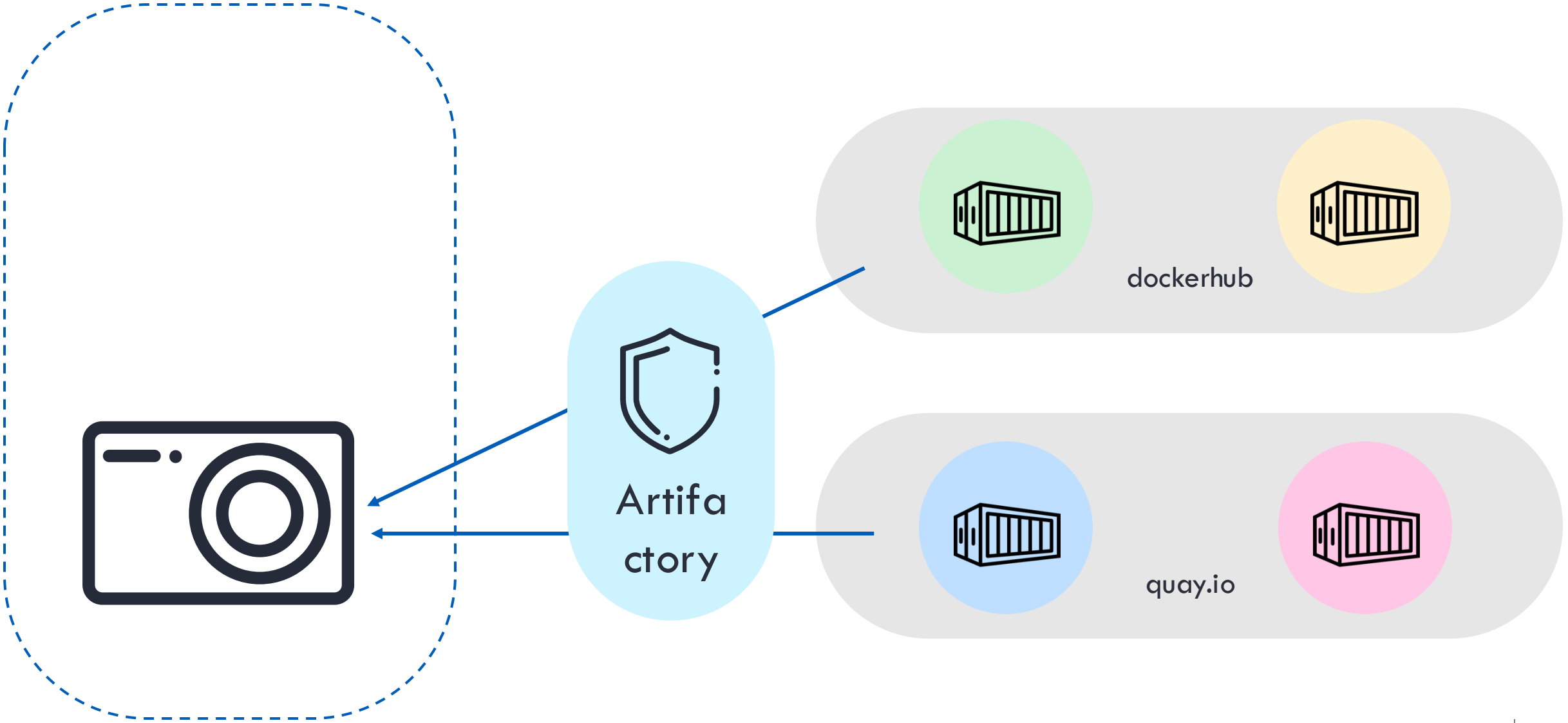
https://re-docs.genomicsengland.co.uk/hpc_containers/

Container repositories



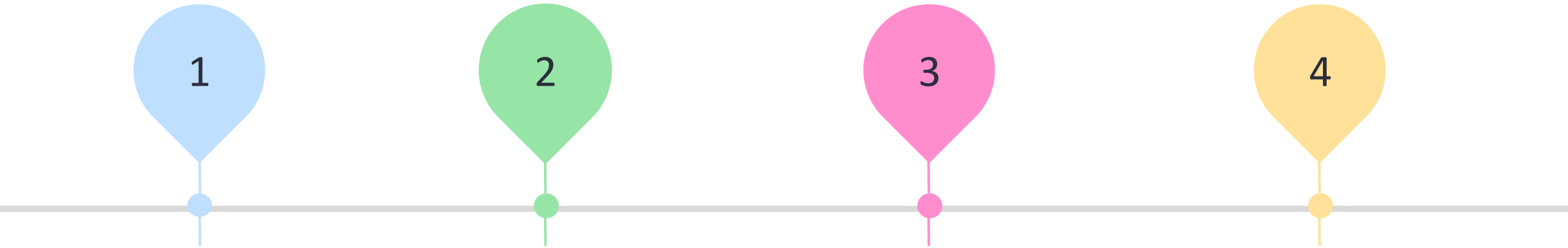
https://re-docs.genomicsengland.co.uk/hpc_containers/

Containers in the RE



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

Steps to import container



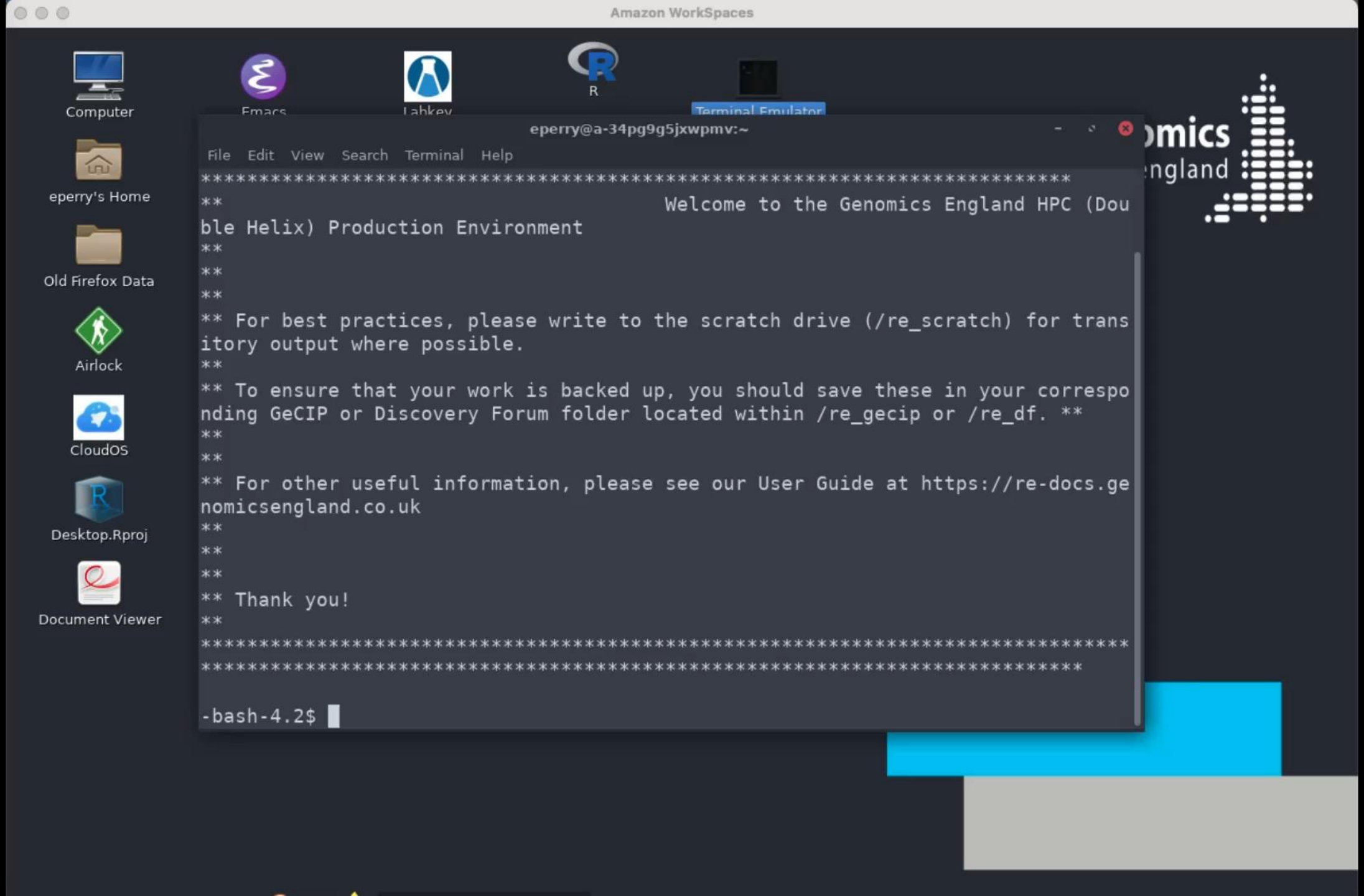
1. Find container in repository: identify docker or quay.io container location
2. modify import location to include artifactory redirect
3. pull container from an interactive job on the HPC
4. mount data and run analysis

Mounting data

```
--bind
/nas/weka.gel.zone/pgen_genomes:/nas/weka.gel.zone/p
gen_genomes --bind /genomes:/genomes
--bind
/nas/weka.gel.zone/pgen_int_data_resources:/nas/weka
.gel.zone/pgen_int_data_resources --bind
/gel_data_resources:/gel_data_resources
--bind
/nas/weka.gel.zone/pgen_public_data_resources:/nas/w
eka.gel.zone/pgen_public_data_resources --bind
/public_data_resources:/public_data_resources
/nas/weka.gel.zone/re_scratch:/nas/weka.gel.zone/re_
scratch --bind /re_scratch:/re_scratch
--bind
/nas/weka.gel.zone/re_gecip:/nas/weka.gel.zone/re_ge
cip --bind /re_gecip:/re_gecip
--bind
/nas/weka.gel.zone/discovery_forum:/nas/weka.gel.zon
e/discovery_forum --bind
/discovery_forum:/discovery_forum
```

You will need to bind the full path and the relative path

Singularity demo


Amazon WorkSpaces
Computer
Emacs
Labkey
R
Terminal Emulator
eperry's Home
Old Firefox Data
Airlock
CloudOS
Desktop.Rproj
Document Viewer
eperry@a-34pg9g5jxwpmv:~
File Edit View Search Terminal Help
** Welcome to the Genomics England HPC (Double Helix) Production Environment
** For best practices, please write to the scratch drive (/re_scratch) for transitory output where possible.
** To ensure that your work is backed up, you should save these in your corresponding GeCIP or Discovery Forum folder located within /re_gecip or /re_df. **
** For other useful information, please see our User Guide at https://re-docs.genomicsengland.co.uk
** Thank you!
-bash-4.2$

6. Using Airlock to bring data/tools in

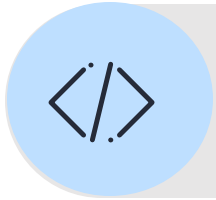
The Airlock

Data in the RE

Outside world

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

What you can import



Scripts and tools for analysing GEL data

Data you want to compare to GEL data

Data you have consent to use

Tools you are licensed to use

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

Airlock demo

Genomics England Research Environment

→ Getting started

How-to guides

Data in the Research Environment

Desktop applications in the Research Environment

High Performance Cluster (HPC)

Workflows, scripts and containers

Data security and Airlock

Training

December 12, 2024

ANNOUNCEMENTS

Current data release:

- 100kGP: /main-programme/main-programme_v19_2024-10-31
- NHS GMS: nhs-gms/nhs-gms-release_v4_2024-08-22

20th December - Secondary clinical data for NHS GMS

We now have medical record data for NHS GMS participants.

18th November - change to HPC address

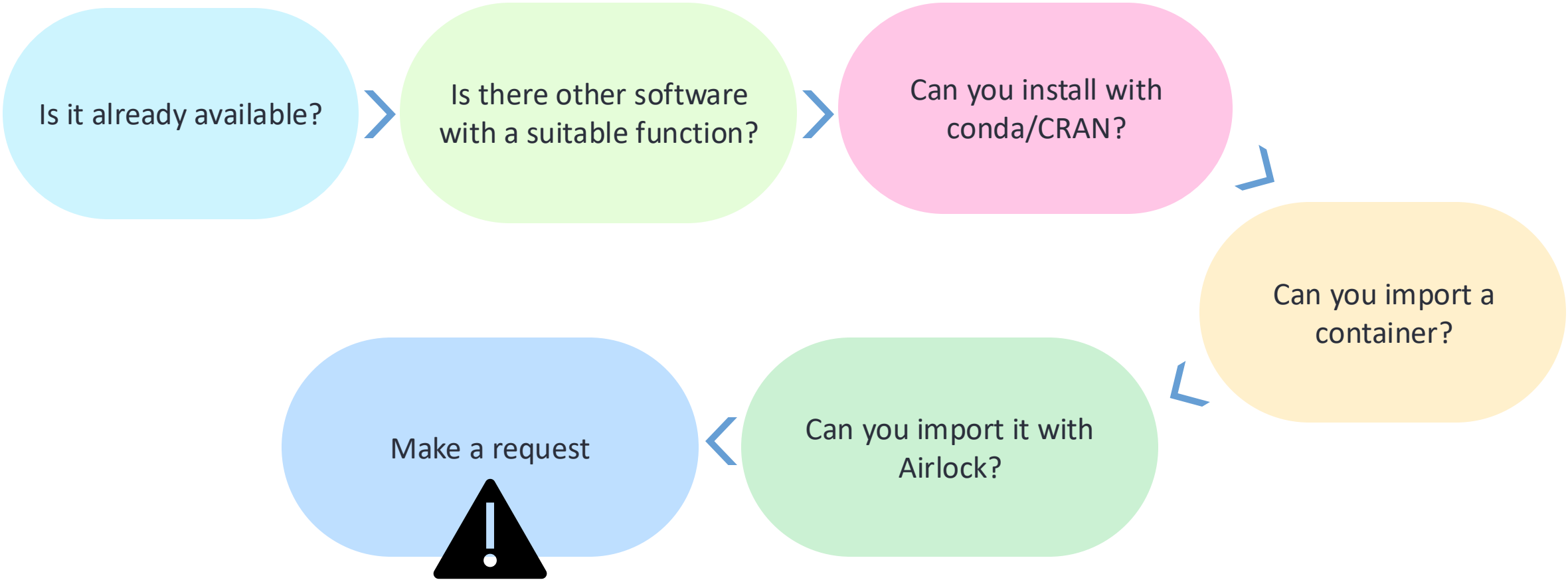
The HPC address has changed to `doublehelix.helix.prod.aws.gel.ac`.

--

6 November 2024 - Workflow update

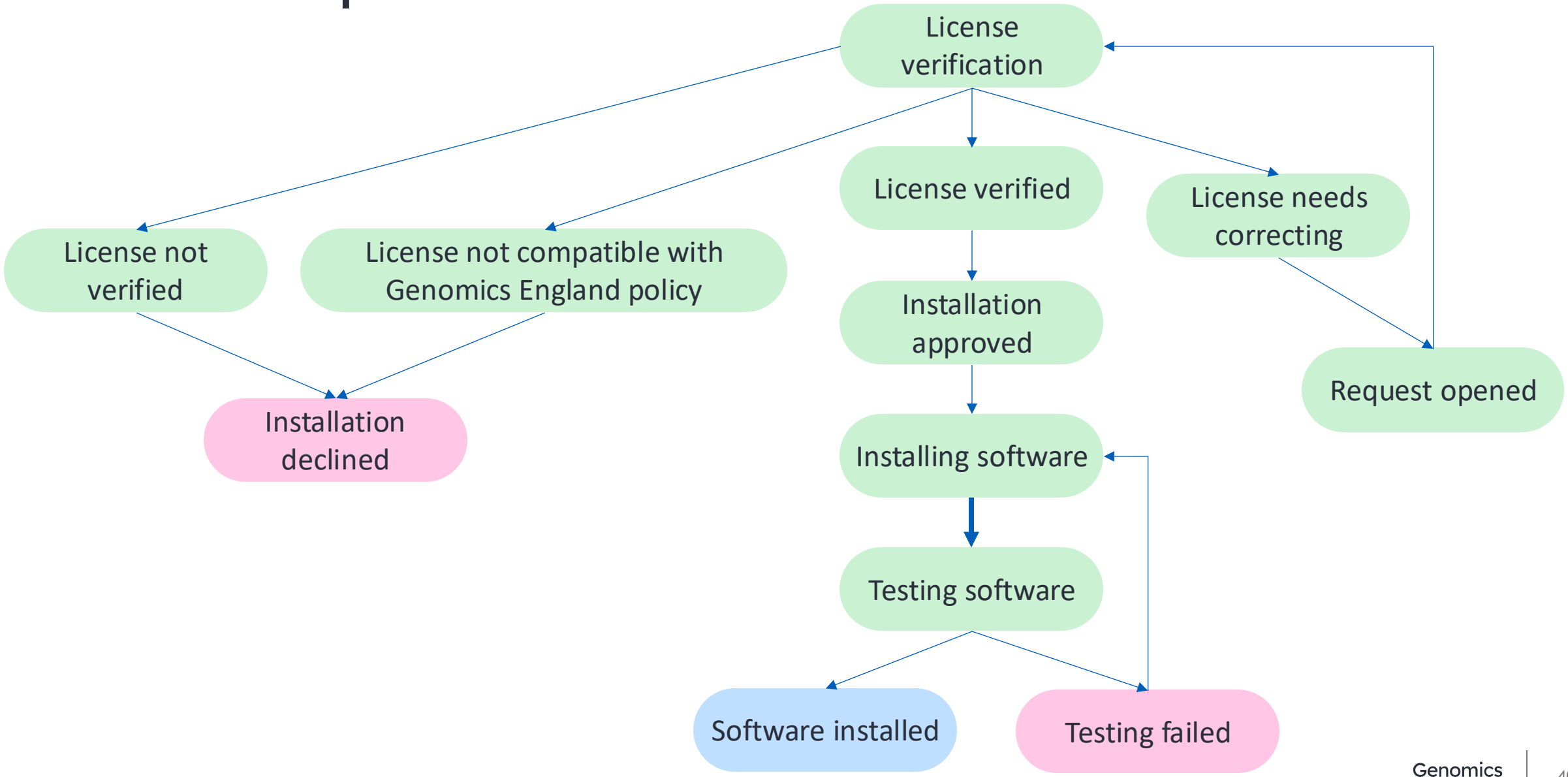
7. Make a software request

When to make a request



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

Installation process



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

Request demo

Genomics England Research Environment

- Getting started
- How-to guides
- Data in the Research Environment
- Desktop applications in the Research Environment
- High Performance Cluster (HPC)
- Workflows, scripts and containers
- Data security and Airlock
- Training

December 12, 2024

ANNOUNCEMENTS

Current data release:

- 100kGP: /main-programme/main-programme_v19_2024-10-31
- NHS GMS: nhs-gms/nhs-gms-release_v4_2024-08-22

20th December - Secondary clinical data for NHS GMS

We now have medical record data for NHS GMS participants.

18th November - change to HPC address

The HPC address has changed to `doublehelix.helix.prod.aws.gel.ac` .

--

6 November 2024 - Workflow update

8. CloudOS – importing tools and pipelines on the Cloud

Linking your account

Link private repositories

Access and use scripts and pipelines instantly

Share repos with collaborators in your workspace

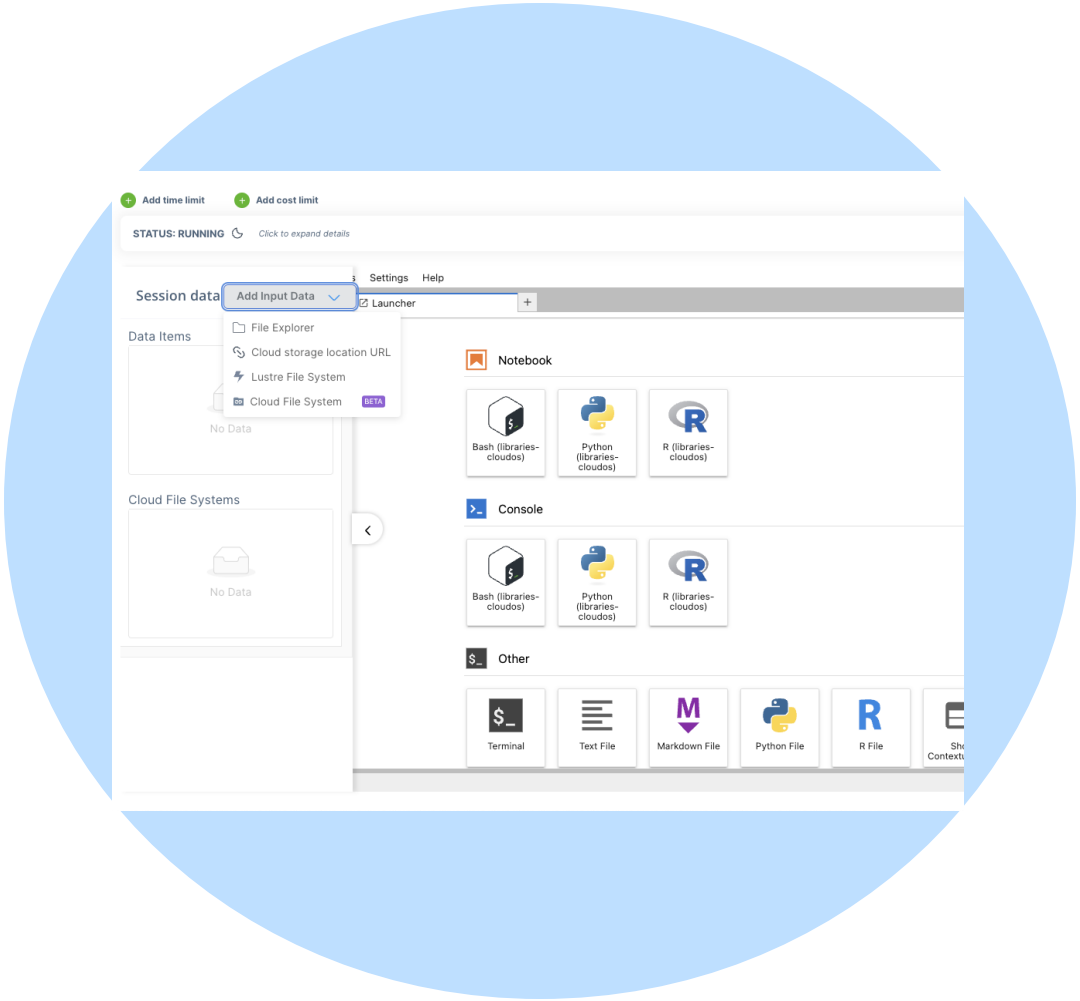
Import and run a pipeline

Easy to Setup and Configure (no command line needed)

Monitor, Clone & Resume analyses

Share & collaborate together

Run a pipeline interactively



Use `git clone` to pull your repo

Test and edit your Pipelines

Install and manage dependencies using Conda, Podman and Quay.io

Scale compute resources to your analysis

Add Data to your Workspace

Upload from the VDI

Import from S3

Link an S3 Bucket

HPC files

Web Browser files

1000 Genomes

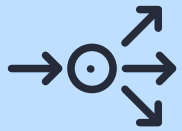
GNOMAD

Lifebit Datasets

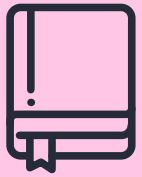
AWS Public Resources

Your own bucket

CloudOS Resources



Nextflow Template
https://github.com/lifebit-ai/template4users



Documentation
https://lifebit.atlassian.net/wiki/spaces/CD/overview?homepageId=168231575

Help Desk

CloudOS demo

9. Software licensing requirements

Responsibility

- Check for academic-only licenses (software and data)
- Purchase your own personal licenses
- We are not responsible if you use software or data you don't have a license for

10. 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

18/2 18/3 15/4

20/5 22/7 19/8

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

Training sessions

11/3	Working with R in the RE
8/4	Working with python in the RE
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

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

Materials from past training all online

Feedback

Thank you

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