

Working with R in the RE

Emily Perry and Eleni Christodoulou

Research Engagement Manager

11th March 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



Eleni Christodoulou
Solutions Specialist -
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



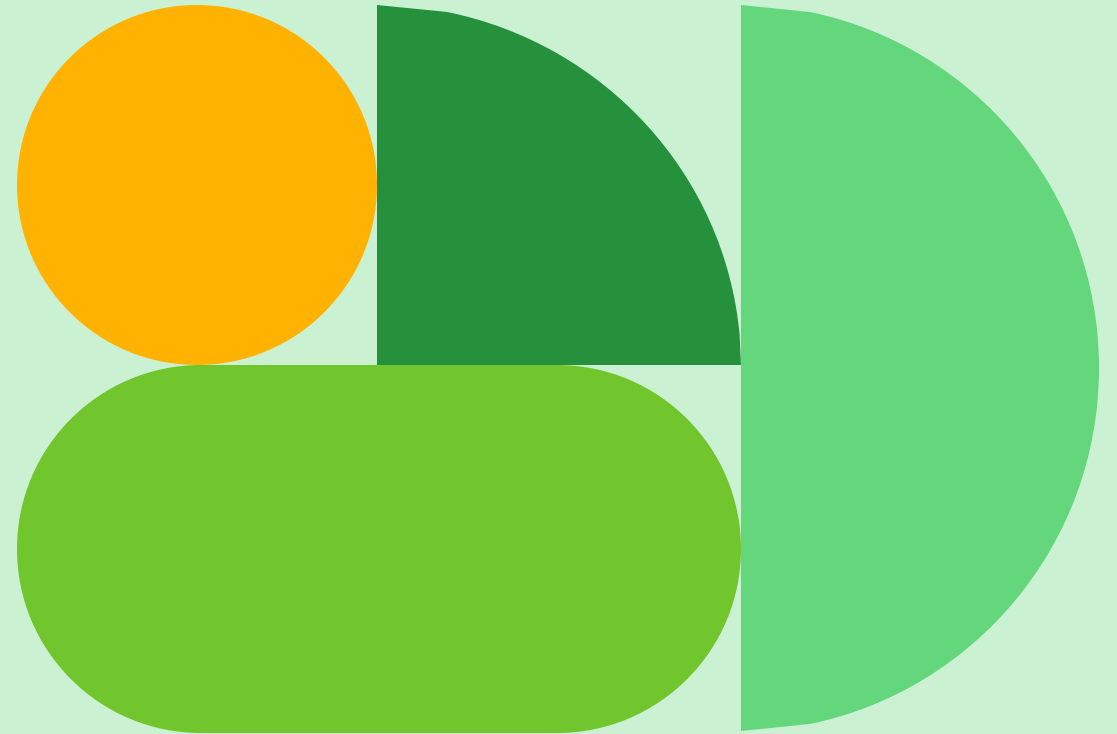
**Magdalena
Drożdż**
Bioinformatician -
Research Services



**Hamzah
Syed**
Solutions
Manager -
Lifebit

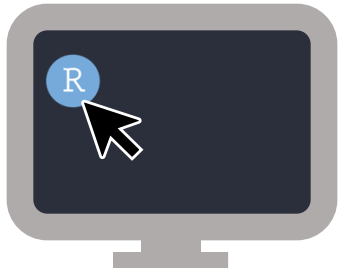
Agenda

- 1 Introduction and admin
- 2 Working with RStudio on the HPC
- 3 Plotting in R
- 4 Working with R libraries
- 5 Working with RStudio in CloudOS interactive sessions
- 6 Query clinical data with R
- 7 Help and questions



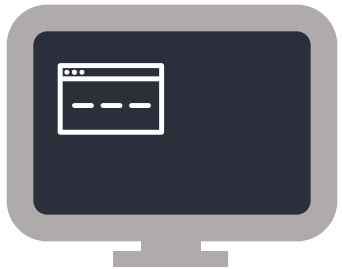
2. Working with RStudio on the HPC

RStudio in the RE



Opens a default version of R

Can launch an HPC job



Choose your preferred R version



A job on the inter queue

```
bsub -q inter -P <your_project_code> -R  
rusage[mem=1000] -M 1000 -n 1 -Is bin/bash
```

Do some work

```
bjobs (to find job number)  
bkill <job number>
```

Rstudio demo

The desktop environment features a grid of application icons on the left side, including:

- 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

On the right side, the **Genomics England** logo is displayed, featuring a stylized DNA sequence graphic. A mouse cursor is visible near the logo. A large grey circle and a blue horizontal bar are also present on the desktop.

3. Plotting in R



No GUI on the HPC

```
png("my_image.png")
```



```
Unable to start device PNG  
or  
Unable to open connection  
to X11 display
```

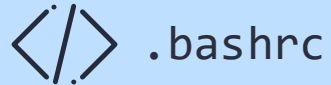
Use an X Virtual Frame Buffer



```
> module load  
R/<version>  
> xvfb-run -a R
```



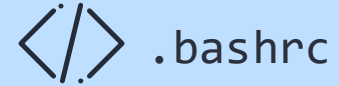
```
png("my_image.png")  
plot(1)  
dev.off()
```



```
.bashrc  
  
alias R='xvfb-run -a  
R'
```



```
png("my_image.png")  
plot(1)  
dev.off()
```



```
.bashrc  
  
function r_headless(){  
  module purge  
  module load lang/R/<version>  
  xvfb-run -a Rscript $1  
}
```

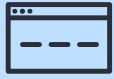


```
png("my_image.png")  
plot(1)  
dev.off()
```



```
r_headless script.R
```

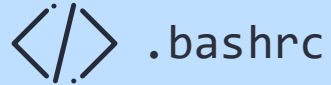
Use an X Virtual Frame Buffer



```
> module load  
R/<version>  
> xvfb-run -a R
```



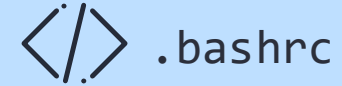
```
library(ggplot2)  
fig_1 <- ggplot(*some_data*)  
ggsave(fig_1, type = "cairo",  
file="fig_1.<extension>")
```



```
.bashrc  
  
alias R='xvfb-run -a  
R'
```



```
library(ggplot2)  
fig_1 <- ggplot(*some_data*)  
ggsave(fig_1, type = "cairo",  
file="fig_1.<extension>")
```



```
.bashrc  
  
function r_headless(){  
  module purge  
  module load lang/R/<version>  
  xvfb-run -a Rscript $1  
}
```



```
library(ggplot2)  
fig_1 <- ggplot(*some_data*)  
ggsave(fig_1, type = "cairo",  
file="fig_1.<extension>")
```



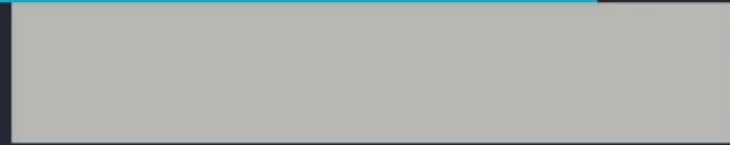
r_headless script.R

Plotting demo

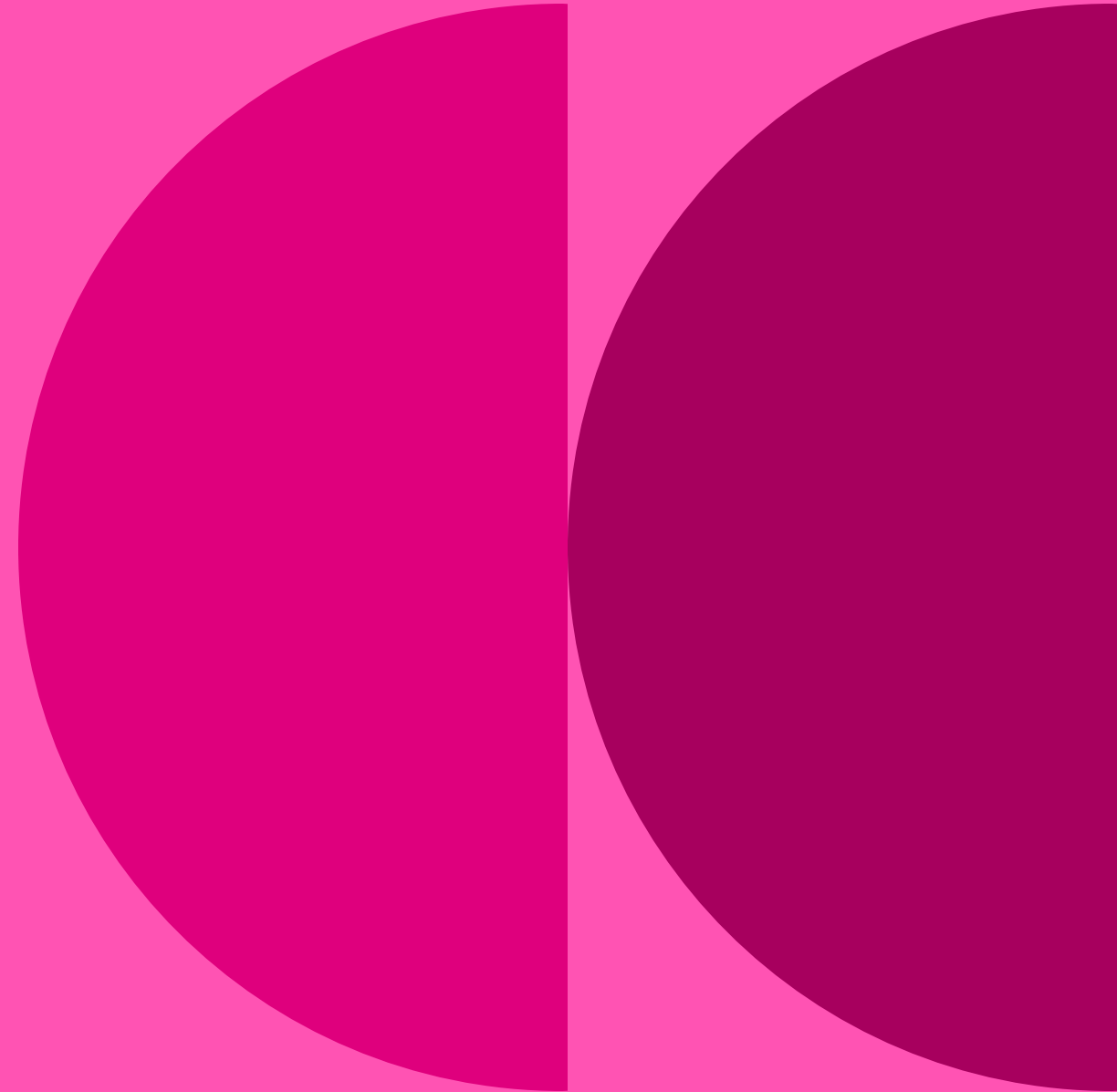
The desktop environment features a grid of application icons on the left side, including:

- 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

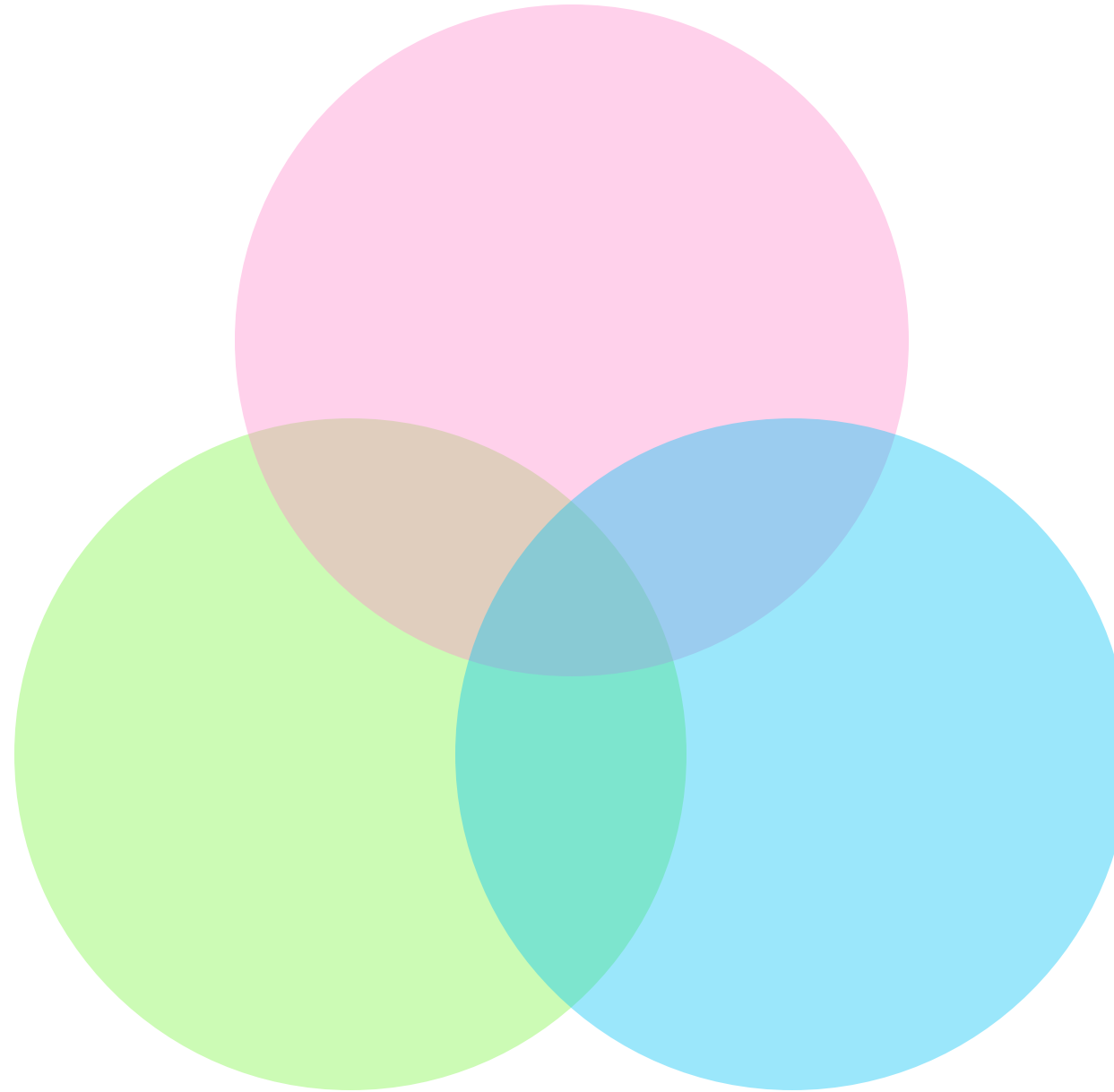
The system tray at the bottom left contains icons for Applications, Places, System, and the Dash. The top right corner displays the system status bar with the date and time: Fri 17 Jan 13:28.



4. Working with R libraries



R packages



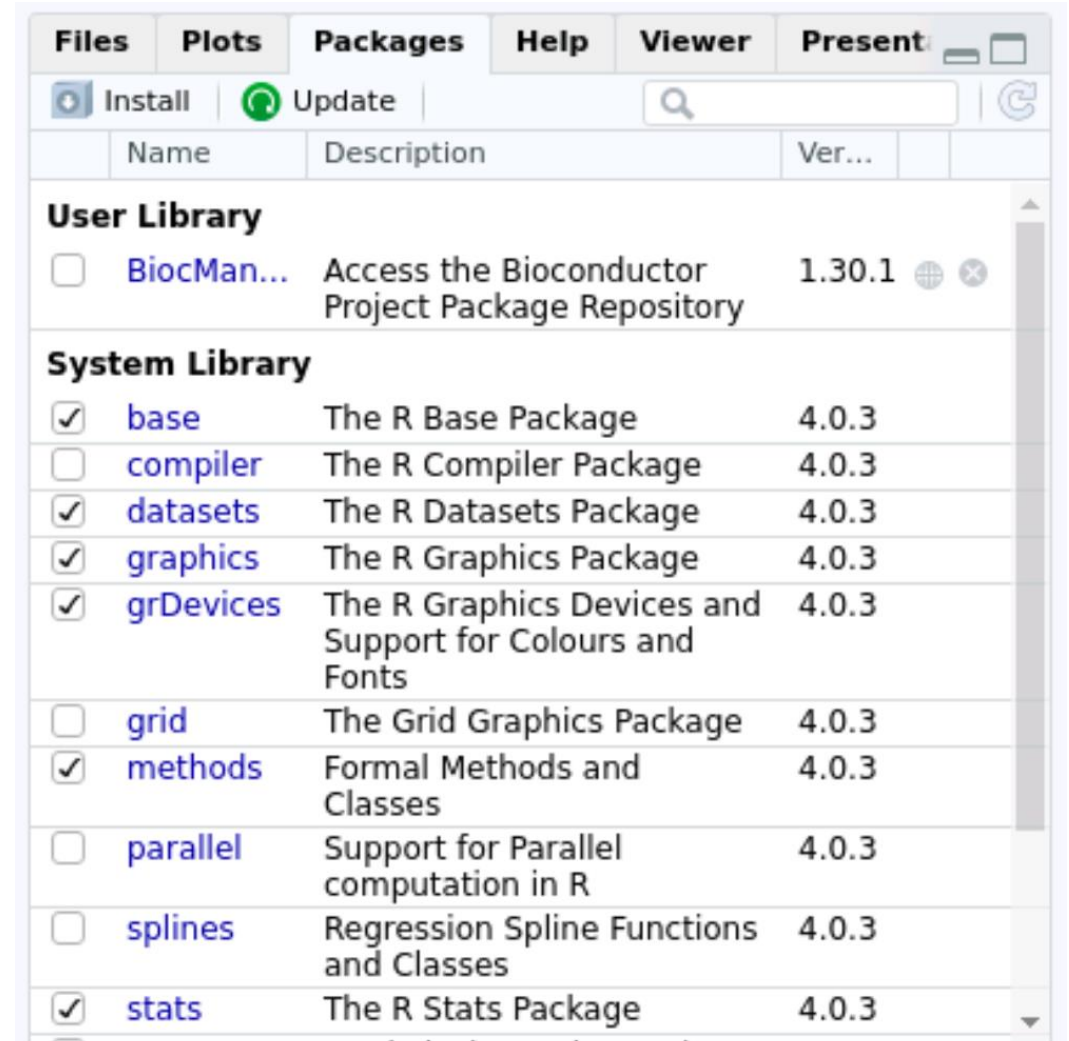
R/3.6.3

R/4.2.1

R/4.3.3

Loading libraries

```
library(library_name)
```



The screenshot shows the 'Packages' window in R. It has tabs for 'Files', 'Plots', 'Packages', 'Help', 'Viewer', and 'Present'. Below the tabs are buttons for 'Install' and 'Update', and a search bar. The main area is a table with columns for 'Name', 'Description', and 'Ver...'. The table is divided into two sections: 'User Library' and 'System Library'. In the 'User Library' section, 'BiocMan...' is listed with a description 'Access the Bioconductor Project Package Repository' and version '1.30.1'. In the 'System Library' section, several packages are listed with their descriptions and version '4.0.3'. Checkmarks in the first column indicate which packages are installed.

	Name	Description	Ver...
User Library			
<input type="checkbox"/>	BiocMan...	Access the Bioconductor Project Package Repository	1.30.1
System Library			
<input checked="" type="checkbox"/>	base	The R Base Package	4.0.3
<input type="checkbox"/>	compiler	The R Compiler Package	4.0.3
<input checked="" type="checkbox"/>	datasets	The R Datasets Package	4.0.3
<input checked="" type="checkbox"/>	graphics	The R Graphics Package	4.0.3
<input checked="" type="checkbox"/>	grDevices	The R Graphics Devices and Support for Colours and Fonts	4.0.3
<input type="checkbox"/>	grid	The Grid Graphics Package	4.0.3
<input checked="" type="checkbox"/>	methods	Formal Methods and Classes	4.0.3
<input type="checkbox"/>	parallel	Support for Parallel computation in R	4.0.3
<input type="checkbox"/>	splines	Regression Spline Functions and Classes	4.0.3
<input checked="" type="checkbox"/>	stats	The R Stats Package	4.0.3

Bioconductor

R 4.33

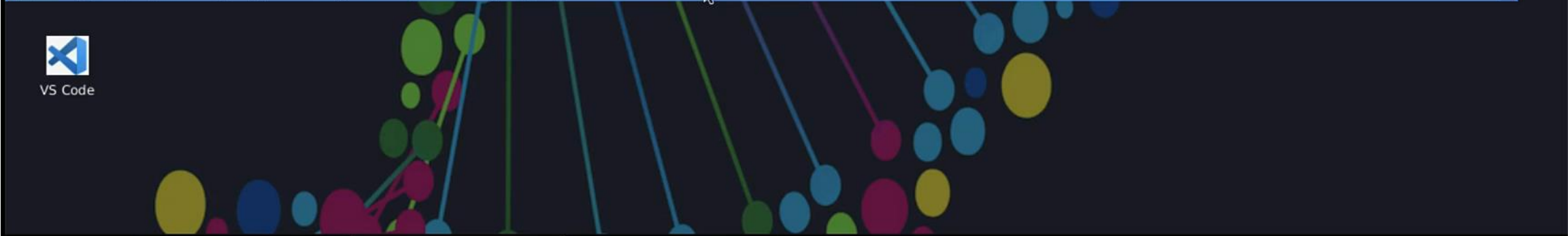
```
options(  
  BIOCONDUCTOR_CONFIG_FILE =  
  "https://artifactory.aws.gel.ac:443/artifactory/bioconductor.org-  
  cache/config.yaml"  
)  
  
library("BiocManager")  
BiocManager::install("<package_name>")  
library(<package_name>)
```

R 3.6.0
to 4.3.2

```
.libPaths(c( .libPaths(), "/tools/aws-workspace-apps/ce/R/4.0.2" ))  
library("BiocManager")  
BiocManager::install("<package_name>")  
library(<package_name>)
```

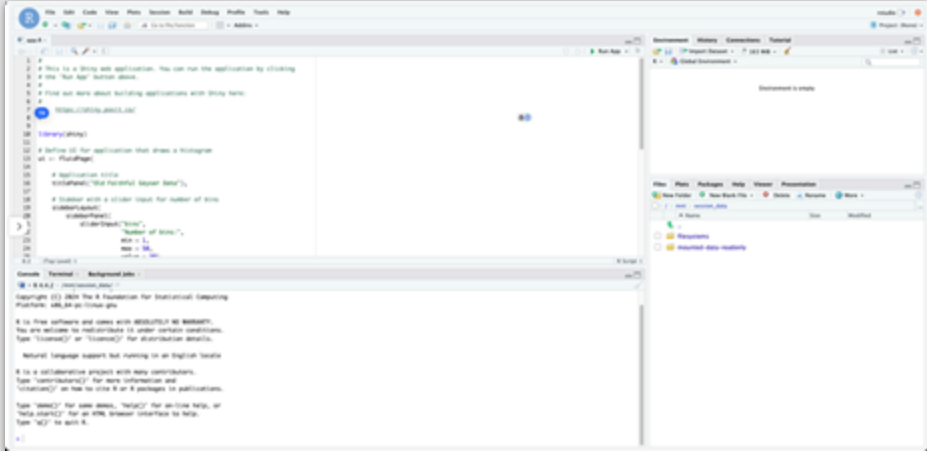
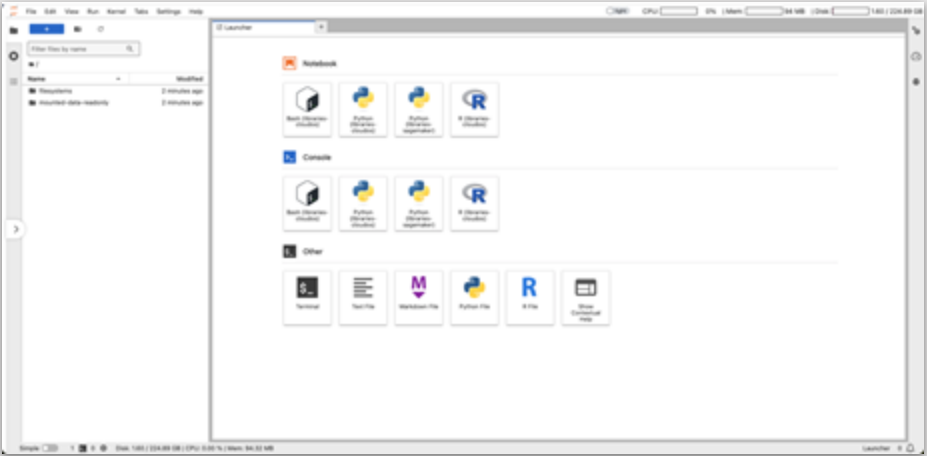
CRAN/Bioconductor demo

```
Terminal - eperry@corp.gel.ac@phpgridzlogn003:/gel_data_resources/software_catalogues/R_catalogue
File Edit View Terminal Tabs Help
.2
| 21 | testpy2_7_12nopyi | /resources/conda/miniconda3/envs/testpy2_7_12nopyi | pybedtools | 0.8.
1
| 22 | testpy2_7_12nopyi | /resources/conda/miniconda3/envs/testpy2_7_12nopyi | bedtools | 2.29
.2
| 23 | testpy2_7_12pyi | /resources/conda/miniconda3/envs/testpy2_7_12pyi | pybedtools | 0.8.
1
| 24 | testpy2_7_12pyi | /resources/conda/miniconda3/envs/testpy2_7_12pyi | 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
|----|:-----|:-----|:-----|
| 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]$
```



5. Working with RStudio in CloudOS interactive sessions

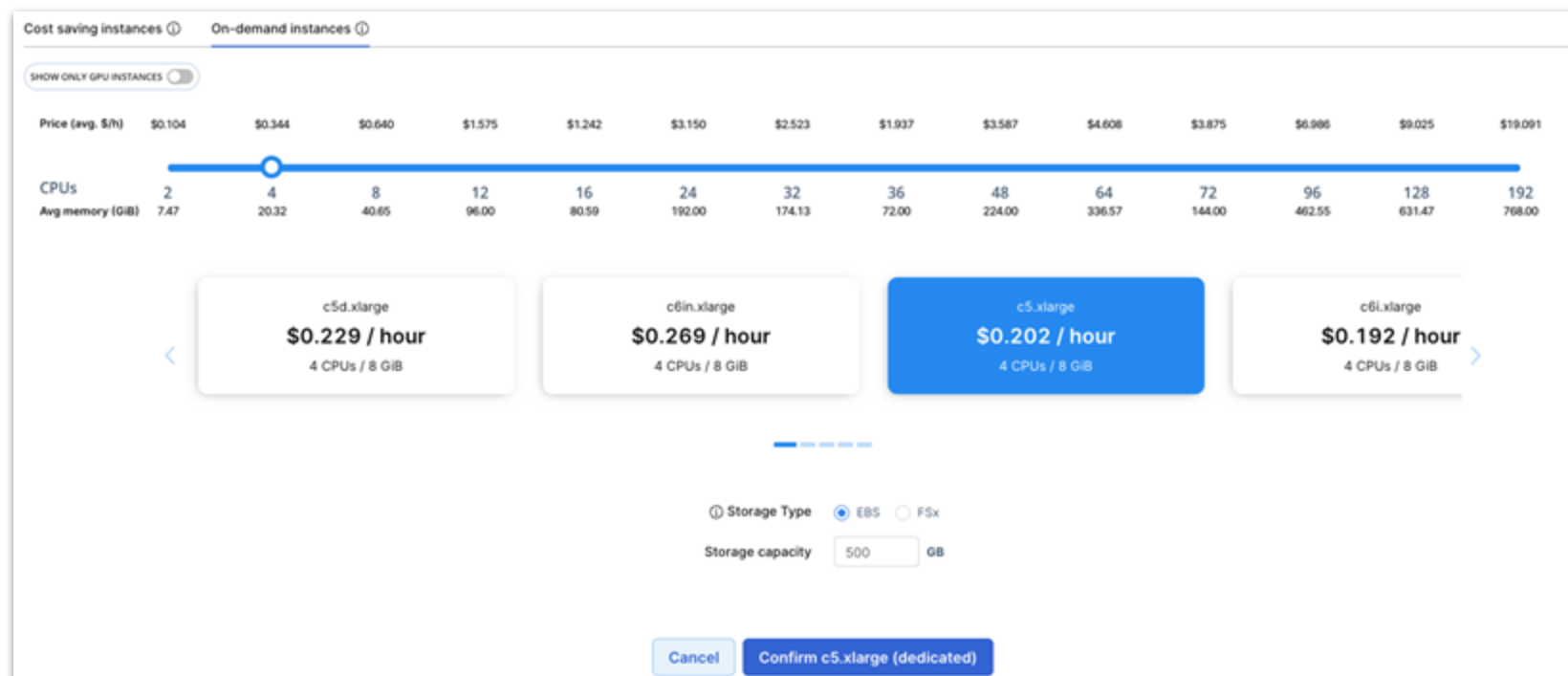
Command line and GUI



Compute options

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

GPUs available on demand



The screenshot shows the AWS On-demand instances selection interface. It features a price slider at the top, a table of instance configurations, and a carousel of instance options. The selected instance is c5.xlarge.

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
CPU	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

Instance Type	Price / hour	CPU	Memory
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

Buttons: Cancel, Confirm c5.xlarge (dedicated)

Installing packages

- Full Flexibility with installing packages from CRAN, Bioconductor 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

The image displays the lifebit interface, which is used for developing, running, and sharing code. The top part shows a sidebar with navigation options: Data, Manhattan & QQ Plot, LocusZoom Plots, and Forrester Plot. The main area is titled "Customize your Forrester plot" and includes a "Design Options" section with fields for "Change title", "x-axis label", "Plot File Name", and "Odds Ratio". A "Forrester Plot" button is visible. Below this is a preview of a Forrester plot showing several horizontal bars with error bars, representing different genetic variants.

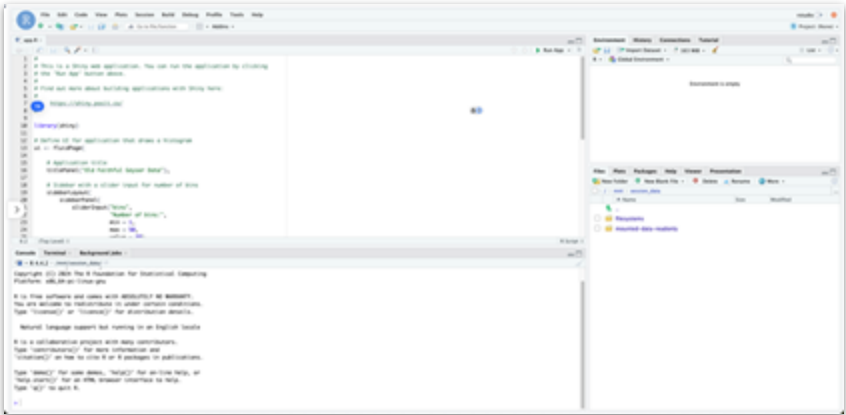
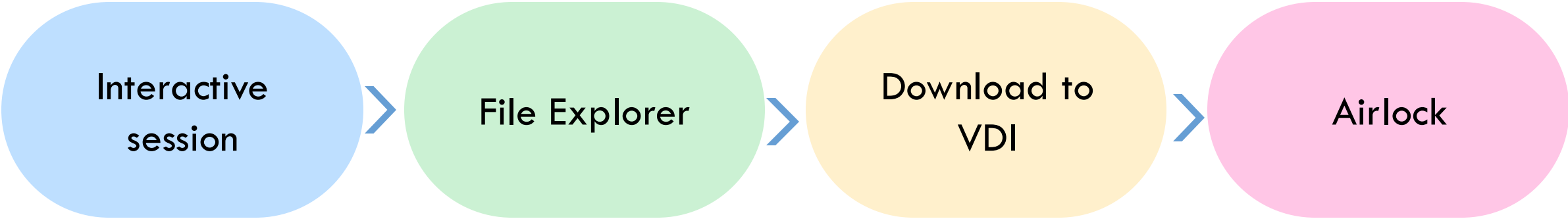
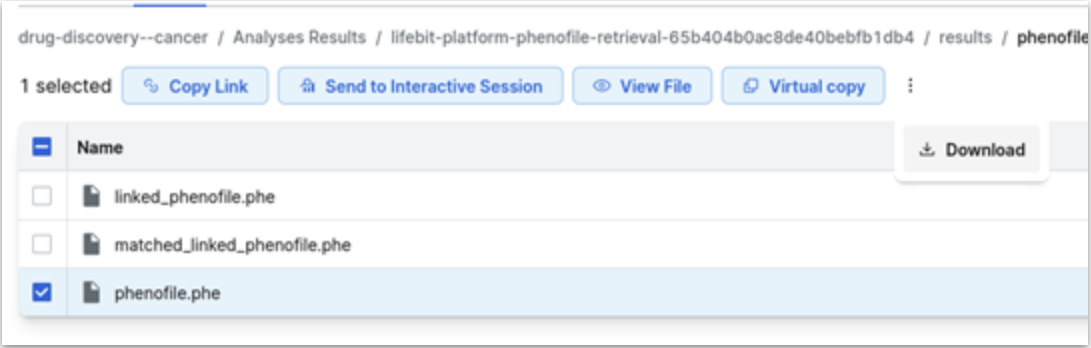
The bottom part of the image shows a code editor with the following R code:

```
lrf  
# 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$vitalsstatus), main = "Distribution by Vital Status")  
barplot(table(data$cancer_disease_type), main = "Distribution by Cancer Disease Type")  
barplot(table(data$smoking_status), main = "Distribution by Smoking Status")  
...
```

Below the code editor is a preview of a bar chart titled "Distribution by Smoking Status". The chart shows the distribution of smoking status across four categories: Current smoker, Ex smoker, Never smoked, and Unknown. The y-axis represents the count, ranging from 0 to 100.

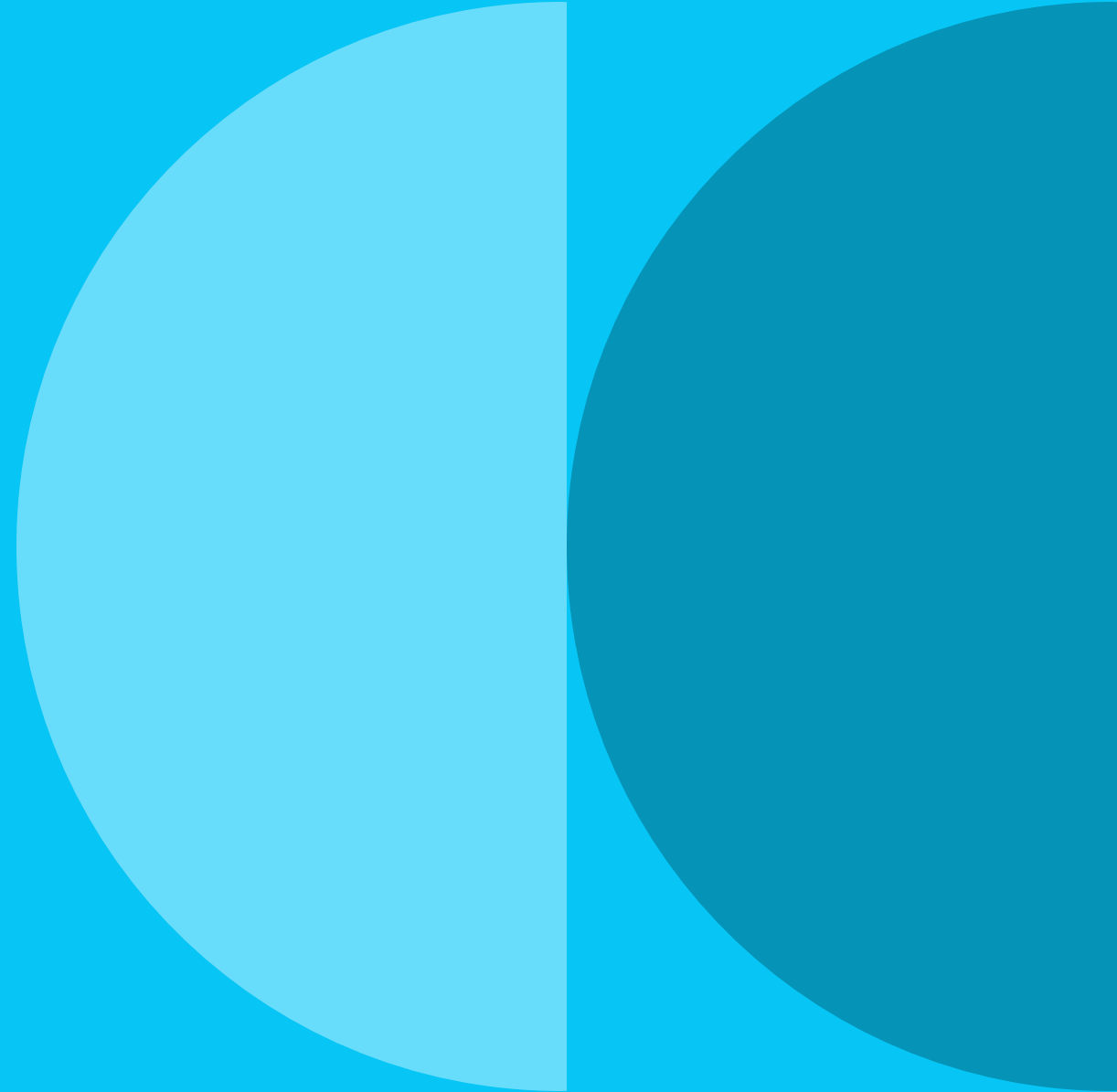
Smoking Status	Count
Current smoker	20
Ex smoker	100
Never smoked	55
Unknown	10

Access your results



CloudOS demo

6. Query clinical data with R

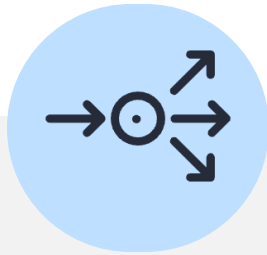


LabKey

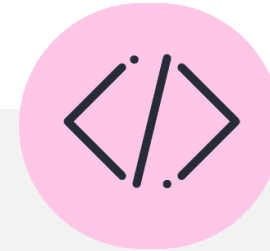
- Central database of:
 - Clinical data
 - Results of bioinformatics analysis
 - Locations of genomic files
- Point and click interface
- API



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

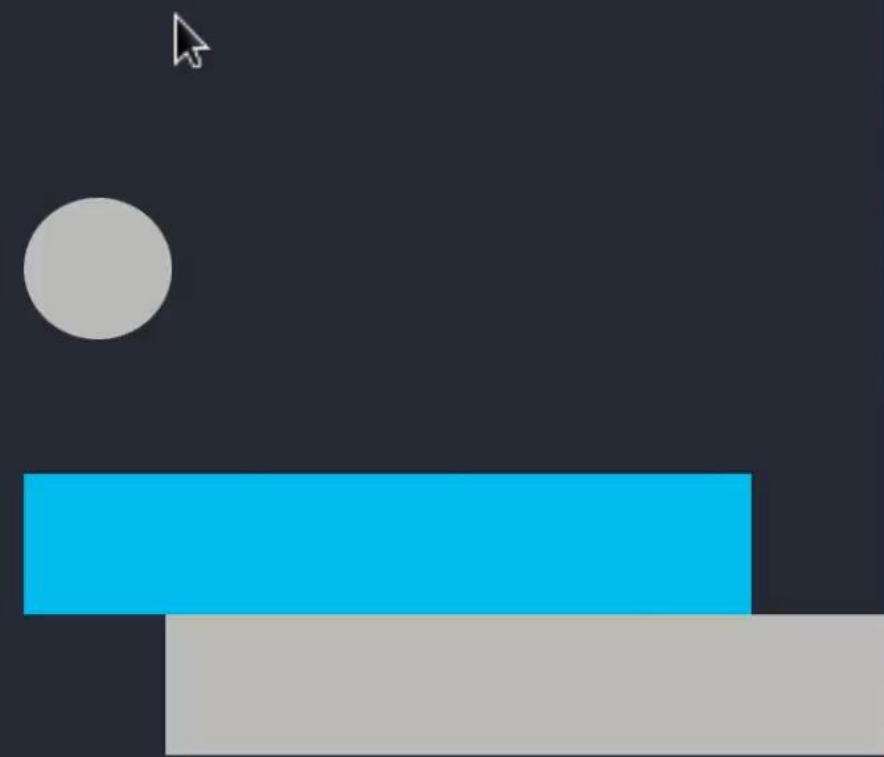
LabKey .netrc

- 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

LabKey API demo



- 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



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

18/3

15/4

20/5

22/7

19/8

16/9



Materials from
past training
all online

Training sessions

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

14/10

What tools and workflows should I use to fulfil an overall goal?



Materials from
past training
all online

Feedback



Thank you

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