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

Emily Perry

11th November 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
- 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)
- We will record this training and distribute the censored video afterwards

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

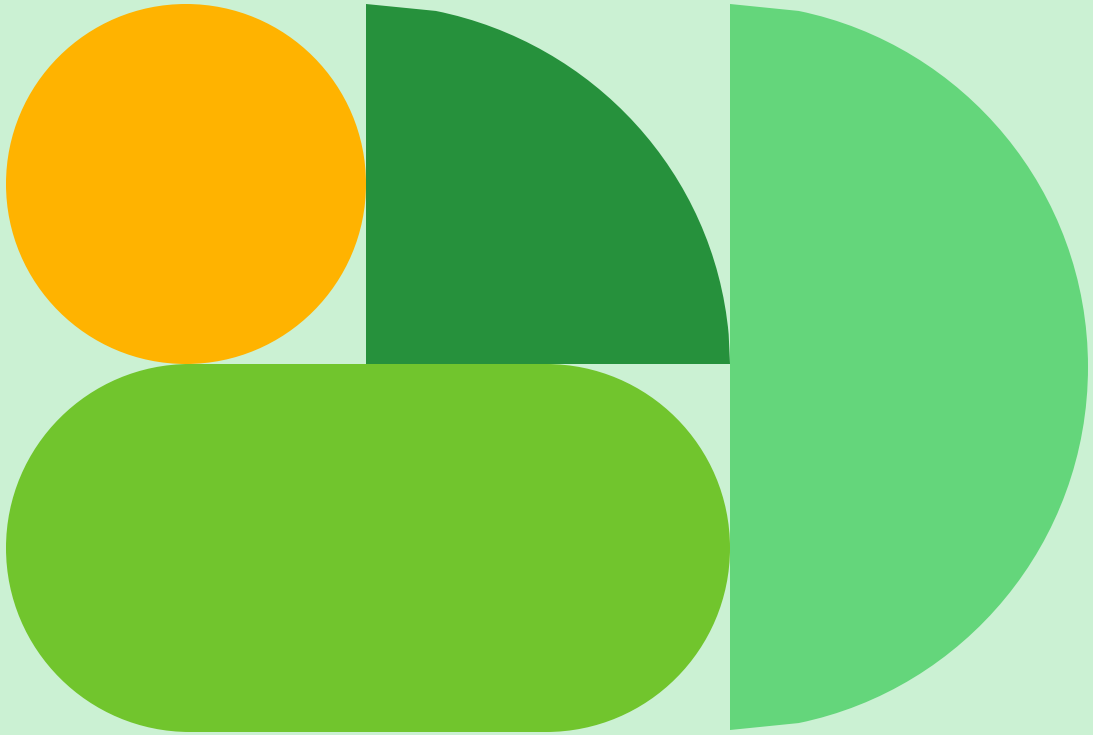
Questions

Miruna Carmen Barbu
Senior Genomics Data Scientist

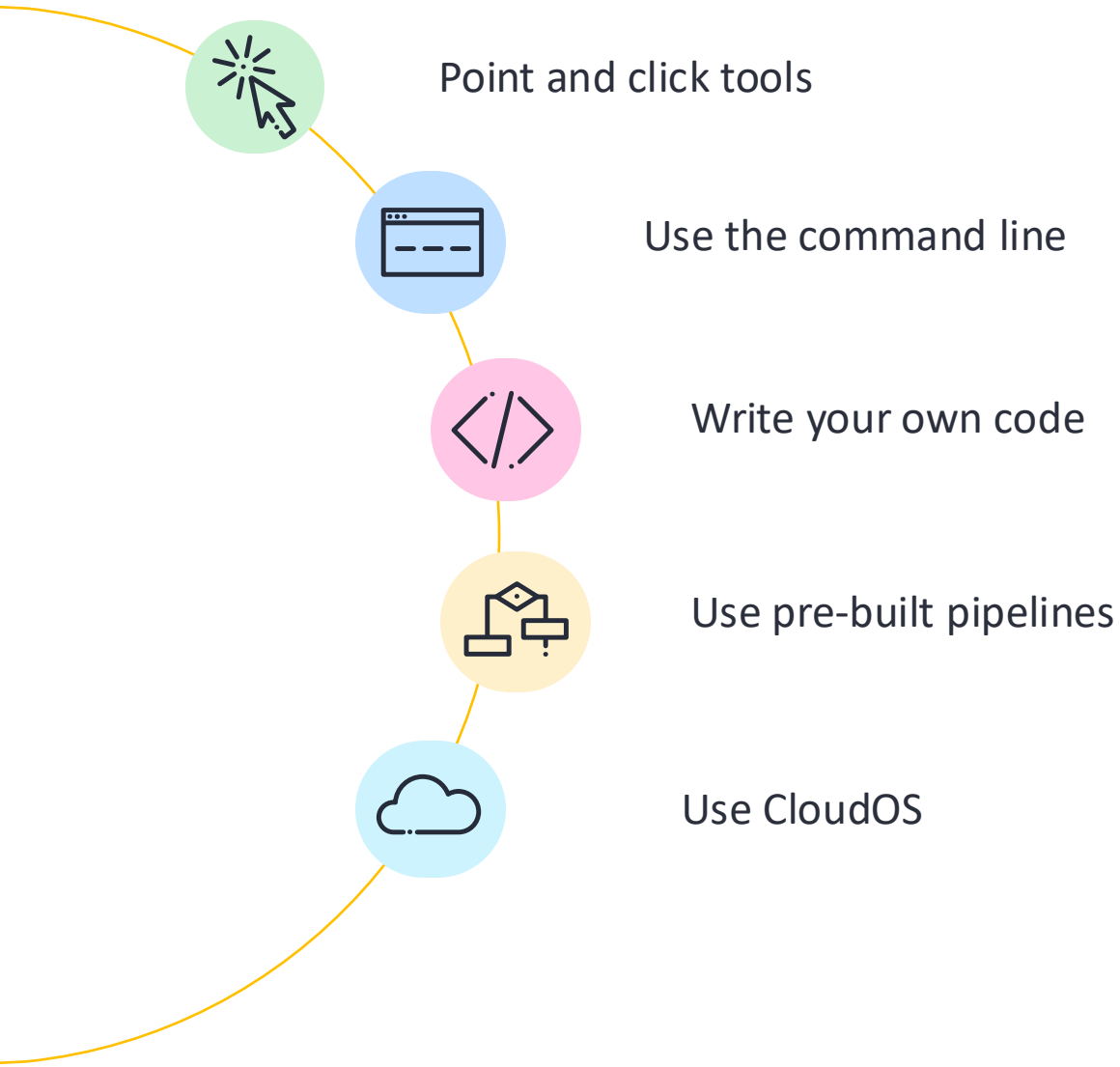
Christian Bouwens
Bioinformatician - Research Services

Agenda

1. Introduction and admin
2. Identifying variants associated with a phenotype
3. Identifying phenotypes associated with a gene
4. Finding diagnoses for patients who didn’t get one through primary clinical interpretation
5. Help and questions



This session



Point and click tools

Use the command line

Write your own code

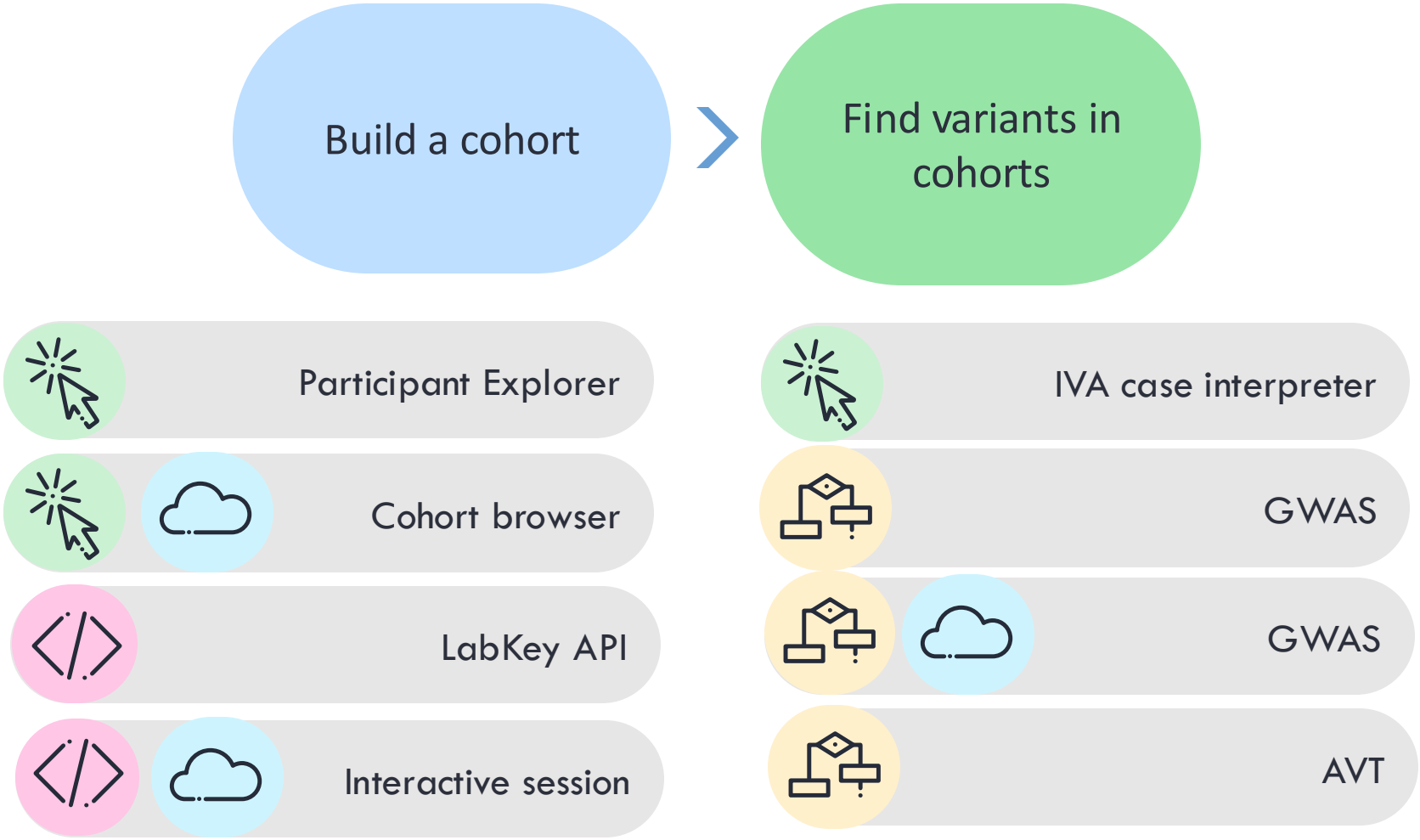
Use pre-built pipelines

Use CloudOS

Links to more detailed online tutorials – live sessions often also available

2. Identifying variants associated with a phenotype

Identifying variants associated with a phenotype



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

Build a cohort – Participant Explorer

- Filter participants by
  - Phenotypes, diagnoses
  - Personal details
- Reverse the search for a control cohort
- Download tables of search results

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

Build a cohort – Cohort Browser

Search for participants by

- Standardised OMOP tables
- Structured clinical data tables

Reverse the search for a control cohort

Create tables within CloudOS and download to VDI

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

Build a cohort – LabKey API

- Use SQL queries
- Tutorials in Python and R
- Cancer tutorial
- Rare disease tutorial

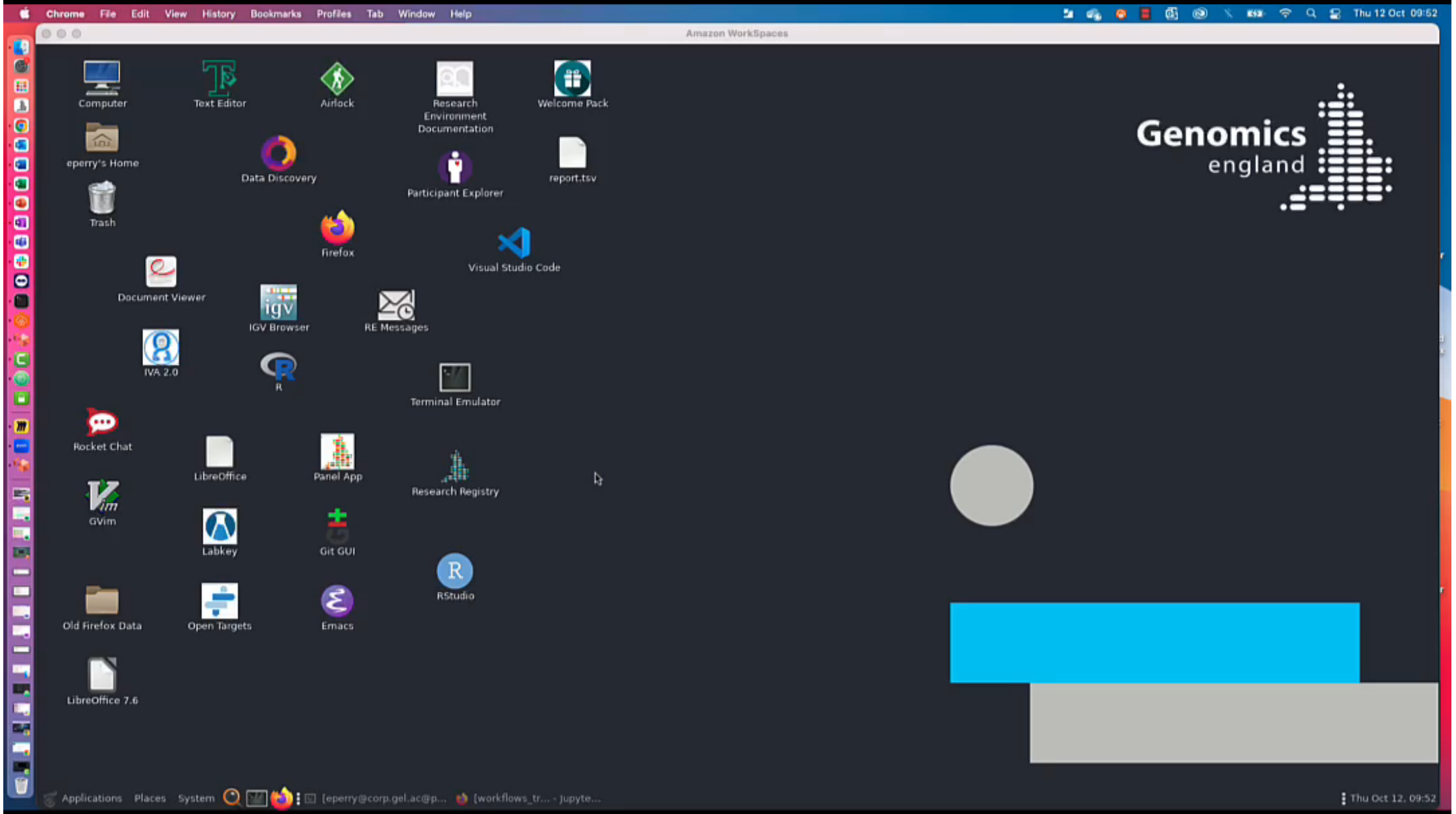
https://re-docs.genomicsengland.co.uk/cancer_cohorts/
https://re-docs.genomicsengland.co.uk/rd_cohorts/

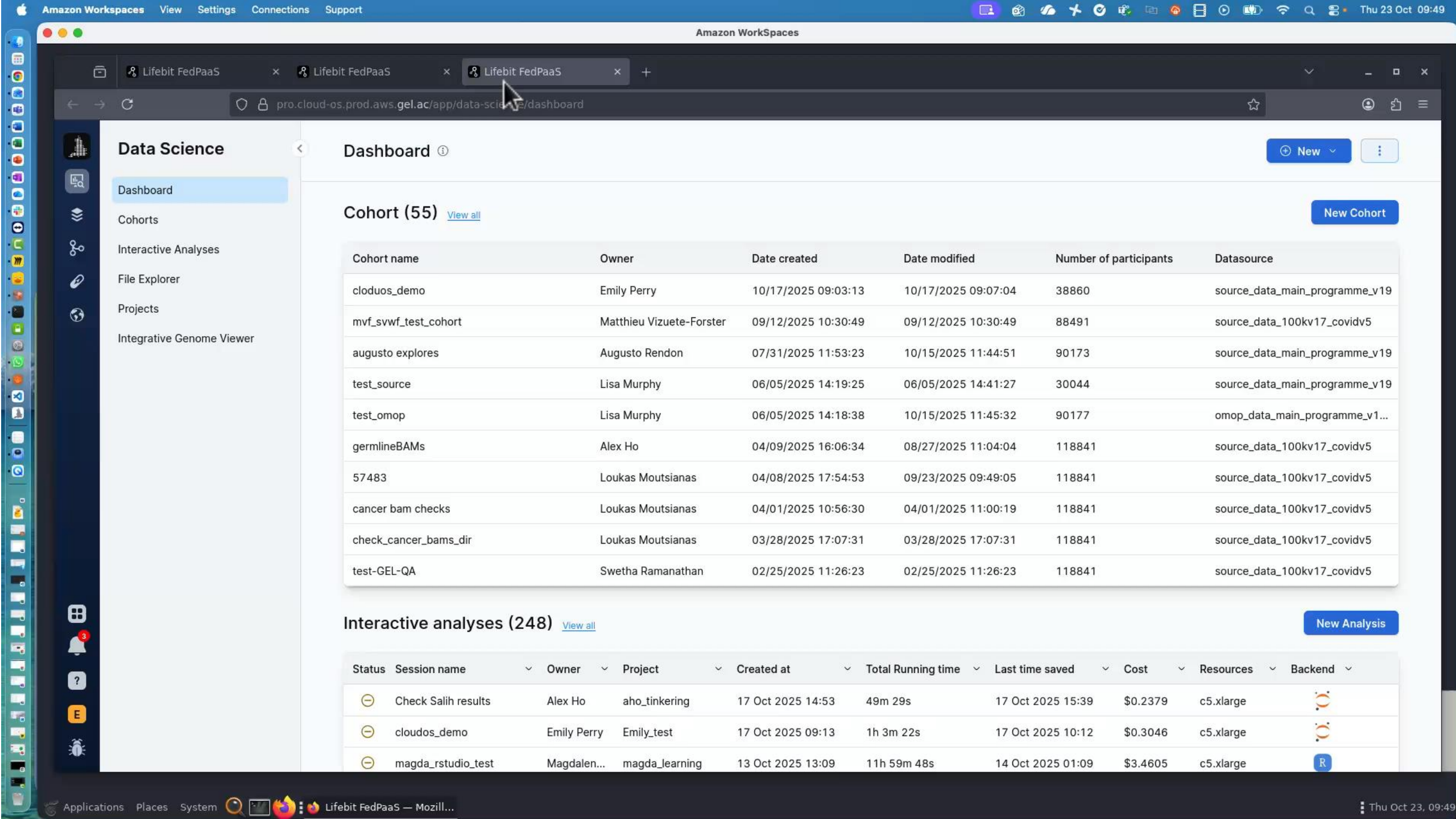
Build a cohort – Interactive session

- Use SQL queries
- Tutorials in Python and R
- Cancer tutorial
- Rare disease tutorial

https://re-docs.genomicsengland.co.uk/cancer_cohorts/
https://re-docs.genomicsengland.co.uk/rd_cohorts/

Build a cohort demo


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Thu 12 Oct 09:52
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Document Viewer
IGV Browser
RE Messages
IVA 2.0
R
Terminal Emulator
Rocket Chat
LibreOffice
Panel App
Research Registry
GVim
Labkey
Git GUI
RStudio
Old Firefox Data
Open Targets
Emacs
LibreOffice 7.6
Genomics england
Applications Places System
Thu Oct 12, 09:52


Amazon Workspaces View Settings Connections Support
Thu 23 Oct 09:49
Amazon WorkSpaces
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
pro.cloud-os.prod.aws.gel.ac/app/data-sci e/dashboard
Data Science
Dashboard
Cohorts
Interactive Analyses
File Explorer
Projects
Integrative Genome Viewer
Dashboard
New
Cohort (55) View all
New Cohort
Cohort name | Owner | Date created | Date modified | Number of participants | Datasource
cloduos_demo | Emily Perry | 10/17/2025 09:03:13 | 10/17/2025 09:07:04 | 38860 | source_data_main_programme_v19
mvf_svwf_test_cohort | Matthieu Vizuete-Forster | 09/12/2025 10:30:49 | 09/12/2025 10:30:49 | 88491 | source_data_100kv17_covidv5
augusto explores | Augusto Rendon | 07/31/2025 11:53:23 | 10/15/2025 11:44:51 | 90173 | source_data_main_programme_v19
test_source | Lisa Murphy | 06/05/2025 14:19:25 | 06/05/2025 14:41:27 | 30044 | source_data_main_programme_v19
test_omop | Lisa Murphy | 06/05/2025 14:18:38 | 10/15/2025 11:45:32 | 90177 | omop_data_main_programme_v1...
germlineBAMs | Alex Ho | 04/09/2025 16:06:34 | 08/27/2025 11:04:04 | 118841 | source_data_100kv17_covidv5
57483 | Loukas Moutsianas | 04/08/2025 17:54:53 | 09/23/2025 09:49:05 | 118841 | source_data_100kv17_covidv5
cancer bam checks | Loukas Moutsianas | 04/01/2025 10:56:30 | 04/01/2025 11:00:19 | 118841 | source_data_100kv17_covidv5
check_cancer_bams_dir | Loukas Moutsianas | 03/28/2025 17:07:31 | 03/28/2025 17:07:31 | 118841 | source_data_100kv17_covidv5
test-GEL-QA | Swetha Ramanathan | 02/25/2025 11:26:23 | 02/25/2025 11:26:23 | 118841 | source_data_100kv17_covidv5
Interactive analyses (248) View all
New Analysis
Status | Session name | Owner | Project | Created at | Total Running time | Last time saved | Cost | Resources | Backend
Check Salih results | Alex Ho | aho_tinkering | 17 Oct 2025 14:53 | 49m 29s | 17 Oct 2025 15:39 | $0.2379 | c5.xlarge
cloudos_demo | Emily Perry | Emily_test | 17 Oct 2025 09:13 | 1h 3m 22s | 17 Oct 2025 10:12 | $0.3046 | c5.xlarge
magda_rstudio_test | Magdalen... | magda_learning | 13 Oct 2025 13:09 | 11h 59m 48s | 14 Oct 2025 01:09 | $3.4605 | c5.xlarge
Applications Places System
Lifebit FedPaaS — Mozill...
Thu Oct 23, 09:49

Build a cohort – data available

Participant Explorer	100kGP		NHS GMS
Cohort browser	100kGP	Covid-19	
LabKey API	100kGP		NHS GMS
Interactive session	100kGP	Covid-19	

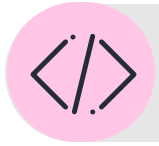
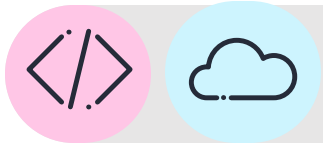
https://re-docs.genomicsengland.co.uk/phenotype_to_variants/

Build a cohort – search terms

Participant Explorer	Words	Ontology aware		
Cohort browser	Codes	Complex schema	OMOP	Can combine with omics
LabKey API	Codes	Complex schema		
Interactive session	Codes	Complex schema		

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

Build a cohort – output and saving

 Participant Explorer	Output to tables	Cannot save queries
 Cohort browser	Save queries	
 LabKey API	Flexible output	Reusable code
 Interactive session	Flexible output	Reusable code

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

Find variants in cohort – IVA case interpreter

See all variants in a participant

Filter variants by

Genes, consequences, variant frequency etc

Family genotypes

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

Find variants in cohort - GWAS

Run on the HPC

Run as a CloudOS batch job

GWAS summaries

Manhattan plots

https://re-docs.genomicsengland.co.uk/gwas/
https://lifebit.atlassian.net/wiki/spaces/CD/pages/813040563/GWAS+System+Tools

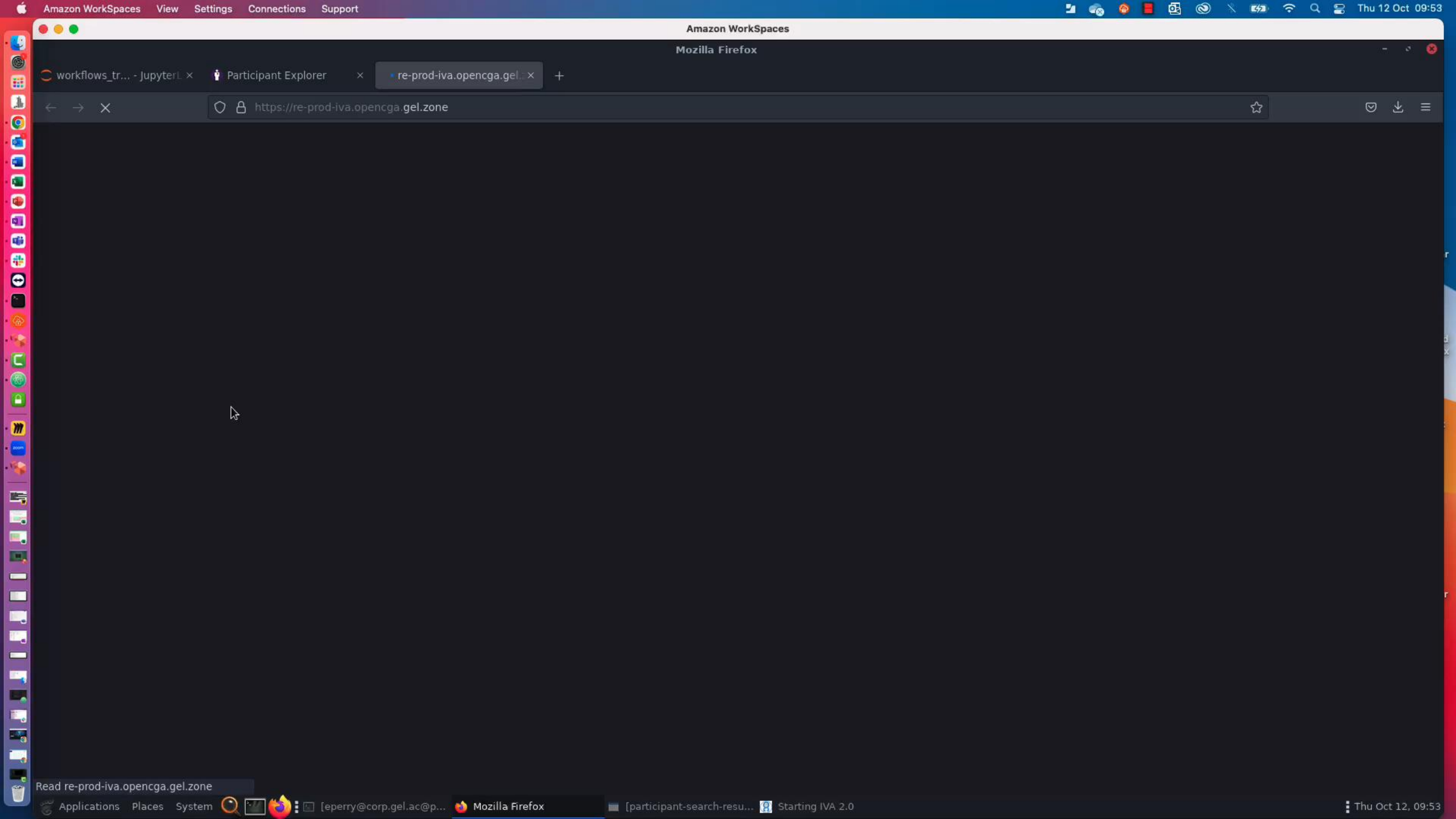
Find variants in cohort - AVT

Run on the HPC

Output depends on the tools used

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

Find variants in cohort demo


Amazon WorkSpaces View Settings Connections Support
Thu 12 Oct 09:53
Amazon WorkSpaces
Mozilla Firefox
workflows_tr... - JupyterL
Participant Explorer
re-prod-iva.opencga.gel.
https://re-prod-iva.opencga.gel.zone
Read re-prod-iva.opencga.gel.zone
Applications Places System
[eperry@corp.gel.ac@p...
Mozilla Firefox
[participant-search-resu...
Starting IVA 2.0
Thu Oct 12, 09:53

demo_cohort.csv | jovyan@2f6db97408dc: ~ | python_workflows_trainin | R_workflows_training.ipyn | gwas_phenofile_python.t | aggV2_chunk_names.bec

Markdown | R (libraries-query-tables)


```
gwas_input_table = distinct(gwas_input_table)
gwas_input_table = select(gwas_input_table, platekey, sex, pheno, age)
gwas_input_table["sex"][gwas_input_table["sex"] == "XX"] <- "1"
gwas_input_table["sex"][gwas_input_table["sex"] == "XY"] <- "0"
write_csv(gwas_input_table,"gwas_phenofile_r.csv")
```

Prepare a bgen file list

We also need a bgen file list to run the GWAS pipeline. This can be found in CloudOS at `gel-data-resources/aggregate_file_lists/aggV2_bgenlist_mergedbychr.txt`. Mount this file to your interactive session, then run.

[34]:
```
bgen_list <- read_tsv("aggV2_chunk_names.bed", col_names = FALSE)
bgen_list <- bgen_list[, c("X1", "X4")]
bgen_list <- filter(bgen_list, X1 != "chrM" & X1 != "chrY")
bgen_list$bgen <- paste('s3://512426816668-gel-data-resources/aggregations/gel_mainProgramme/aggV2/genomic/genomic_data/bgen_masked/bgen_bychunk/gel_mainProgramme_aggV2_',
bgen_list['index'] = paste('s3://512426816668-gel-data-resources/aggregations/gel_mainProgramme/aggV2/genomic/genomic_data/bgen_masked/bgen_bychunk/gel_mainProgramme_aggV2_'
bgen_list['sample'] = paste('s3://512426816668-gel-data-resources/aggregations/gel_mainProgramme/aggV2/genomic/genomic_data/bgen_masked/bgen_bychunk/gel_mainProgramme_aggV2_
names(bgen_list)[names(bgen_list) == "X1"] <- "chr"
bgen_list <- bgen_list[, c("chr", "bgen", "index", "sample")]
write_csv(bgen_list,"bgen_list_r.csv")
```

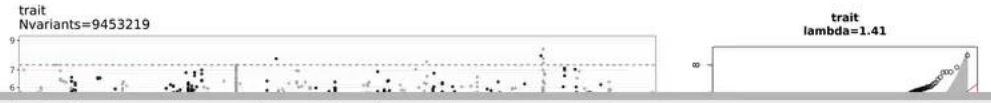
```
Rows: 1371 Columns: 7
── Column specification ──────────────────────────────────────────────────────
Delimiter: "\t"
chr (5): X1, X4, X5, X6, X7
dbl (2): X2, X3

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Run the GWAS pipeline

Now we can run the GWAS pipeline as a batch job. Use the BRS_tools_GWAS_nf pipeline.

This will generate a number of outputs, including a Manhattan plot and qq analysis:



Find variants in cohort – focus

IVA case interpreter	Single participant
GWAS	Common variants
GWAS	Common variants
AVT	Rare variants

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

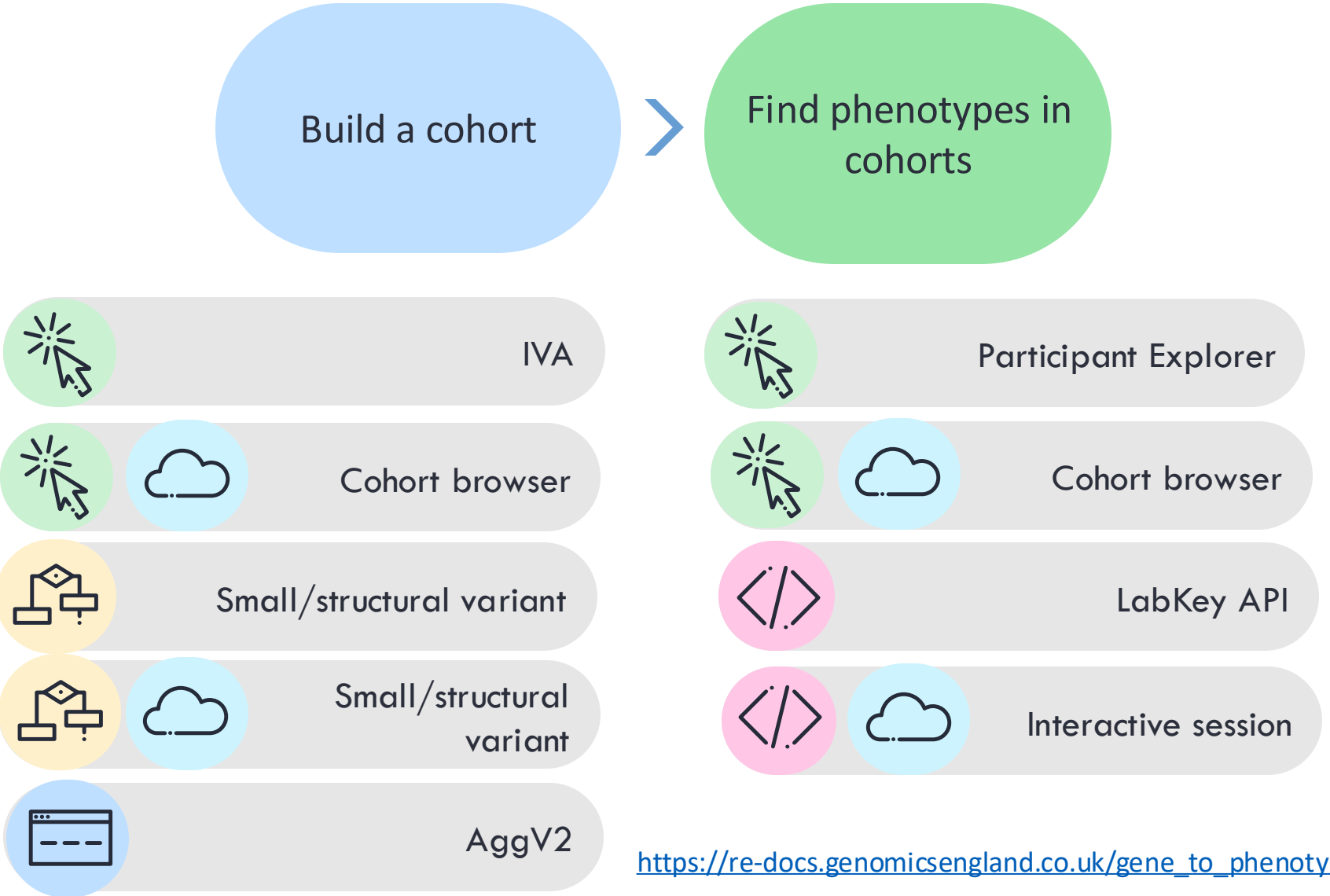
Find variants in cohort – input

IVA case interpreter	Participant ID
GWAS	Cohort file, Unrelated file, HQplinkfile, Genomic file list
GWAS	Cohort file, Unrelated file, HQplinkfile, Genomic file list
AVT	Cohort file

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

3. Identifying phenotypes associated with a gene

Identifying phenotypes associated with a gene



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

Build a cohort – IVA variant browser

- See all variants
- Filter variants by
  - Genes, consequences, variant frequency etc
- See all participants with variant

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

Build a cohort – Cohort browser

See all variants

Filter variants by

Genes, consequences, variant frequency etc

See all participants with variant

https://lifebit.atlassian.net/wiki/spaces/CD/pages/813107122/Filter+cohort+by+genomic+data

Build a cohort – Small/structural variant workflows

- Run on the HPC
- Run as a CloudOS batch job
- Input list of genes
- List of variants in gene(s)
- List of participants with variants

https://re-docs.genomicsengland.co.uk/structural_variant/
https://re-docs.genomicsengland.co.uk/small_variant/

Build a cohort – AggV2

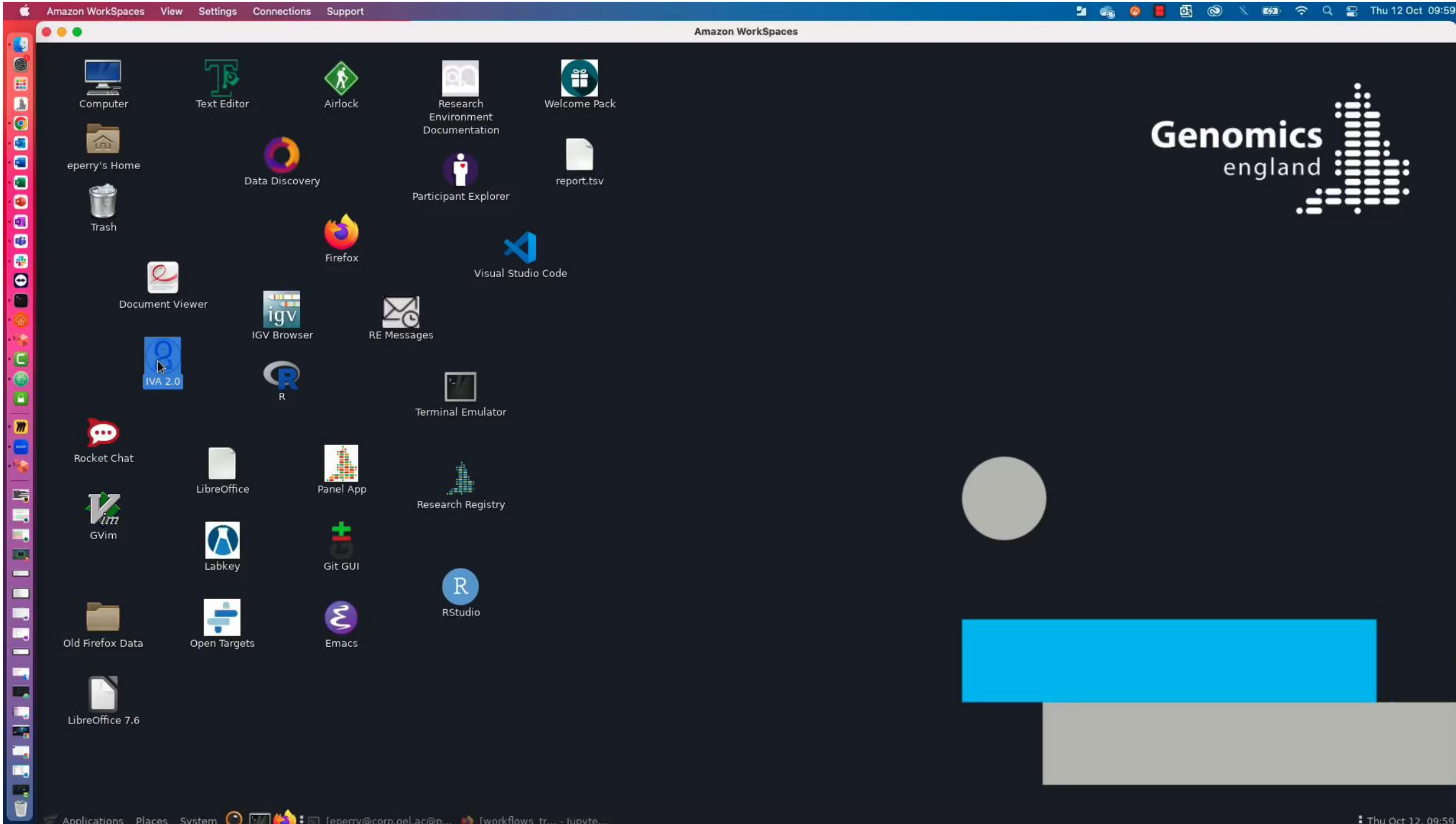
Aggregated VCFs of all variants

Participant genotypes

bcftools to pull out genotypes

https://re-docs.genomicsengland.co.uk/aggv2

Find participants with variants demo


Amazon WorkSpaces View Settings Connections Support
Thu 12 Oct 09:59
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Document Viewer
IGV Browser
RE Messages
IVA 2.0
R
Terminal Emulator
Rocket Chat
LibreOffice
Panel App
Research Registry
GVim
Labkey
Git GUI
RStudio
Old Firefox Data
Open Targets
Emacs
LibreOffice 7.6
Genomics england
Applications Places System
[eperry@corp.gel.ac@p...
[workflows_tr... - Jupyte...
Thu Oct 12, 09:59

demo_cohort.csv | jovyan@2f6db97408dc: | python_workflows_traini | R_workflows_training.ip | bgen_list_r.csv | gwas_phenofile_python. | aggV2_chunk_names.be

Markdown | Python (libraries-python-tables)

Filter files by name

/

Name	Modified
filesystems	14m ago
mounted-data-re...	14m ago
aggV2_chunk_na...	28d ago
bgen_list_python...	19d ago
bgen_list_r.csv	19d ago
demo_cohort.csv	6d ago
gel_mainProgra...	20d ago
gwas_phenofile_...	20d ago
gwas_phenofile_...	20d ago
python_workflow...	6m ago
R_workflows_trai...	6d ago
small_variant_vc...	6d ago
small_variant_vc...	6d ago
structual_variant...	20d ago
structual_variant...	20d ago
trait_manhattan....	20d ago
trait_qqplot.png	20d ago
Untitled.ipynb	6d ago

Run the Small Variant workflow

The first step is to find all participants with variants in a gene. We can do this using the small variant workflow.

You need to make a gene list file, which is just a list saved as `.txt`. In our example, I'm including only one gene:

```
SRY
```

You also need to create a sample list of all VCFs.

[5]:
```
vcf_list_sql = (f'''
    SELECT participant_id, platekey, genome_build, file_path, delivery_version, file_sub_type, type
    FROM genome_file_paths_and_types
    WHERE file_sub_type = 'Standard VCF'
    AND delivery_version = 'V4'
    AND platekey LIKE 'LP%'
    AND (type = 'rare disease germline' OR type = 'cancer germline')
    ''')

vcf_list_query = query_to_df(vcf_list_sql, version)
vcf_list_query = vcf_list_query.drop_duplicates(subset=['platekey'])
vcf_list_query.to_csv("small_variant_vcf_list_python.csv", sep=",", index=False)
```

Start a batch job and find the `gel-small-variant` workflow. Add the parameters:

`gene_input` - select using the file explorer menus to the right
`sample_input` - select using the file explorer menus to the right
`use_sample_input` - set this value to `true`

Run the batch job. Start an interactive job from the batch job results.

Find phenotypes associated with participants

Now we can open the output file and read the participant list. The code below also filters the output based on consequences. You can edit the list of consequences to only get those you're interested in. Finally, it pulls out all the individual platekeys.

[10]:
```
var_workflow = pd.concat([pd.read_table("mounted-data-readonly/GRCh38_SRY_ENSG00000184895_annotated_variants.tsv"), pd.read_table("mounted-data-readonly/GRCh37_SRY_ENSG00000
```

Find participant with variants - data

IVA	GRCh38	100kGP	
Cohort browser	GRCh37/GRCh 38	100kGP	Covid-19
Small/structural variant	GRCh37/GRCh 38	100kGP	
Small/structural variant			
AggV2	GRCh38	100kGP	NHS GMS

AggV3 coming soon

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

Find participant with variants - usage

IVA

Easy to use

Cohort browser

Combine with phenotype

Save queries

Small/structural variant

Submission script

Small/structural variant

Point-and-click

Extra input files

AggV2

bcftools

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

Find phenotypes in cohorts – Participant Explorer

Search for participant(s) by ID(s)

See medical history for a participant

mpare medical history between participants

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

Find phenotypes in cohorts – Cohort browser

- OMOP standardised format
- Source 100kGP clinical tables
- Plots of phenotype frequencies

https://lifebit.atlassian.net/wiki/spaces/CD/pages/813107011/Get+insights+on+your+cohort

Find phenotypes in cohorts – LabKey API

- Use SQL queries
- Tutorials in Python and R
- Detailed medical history tables

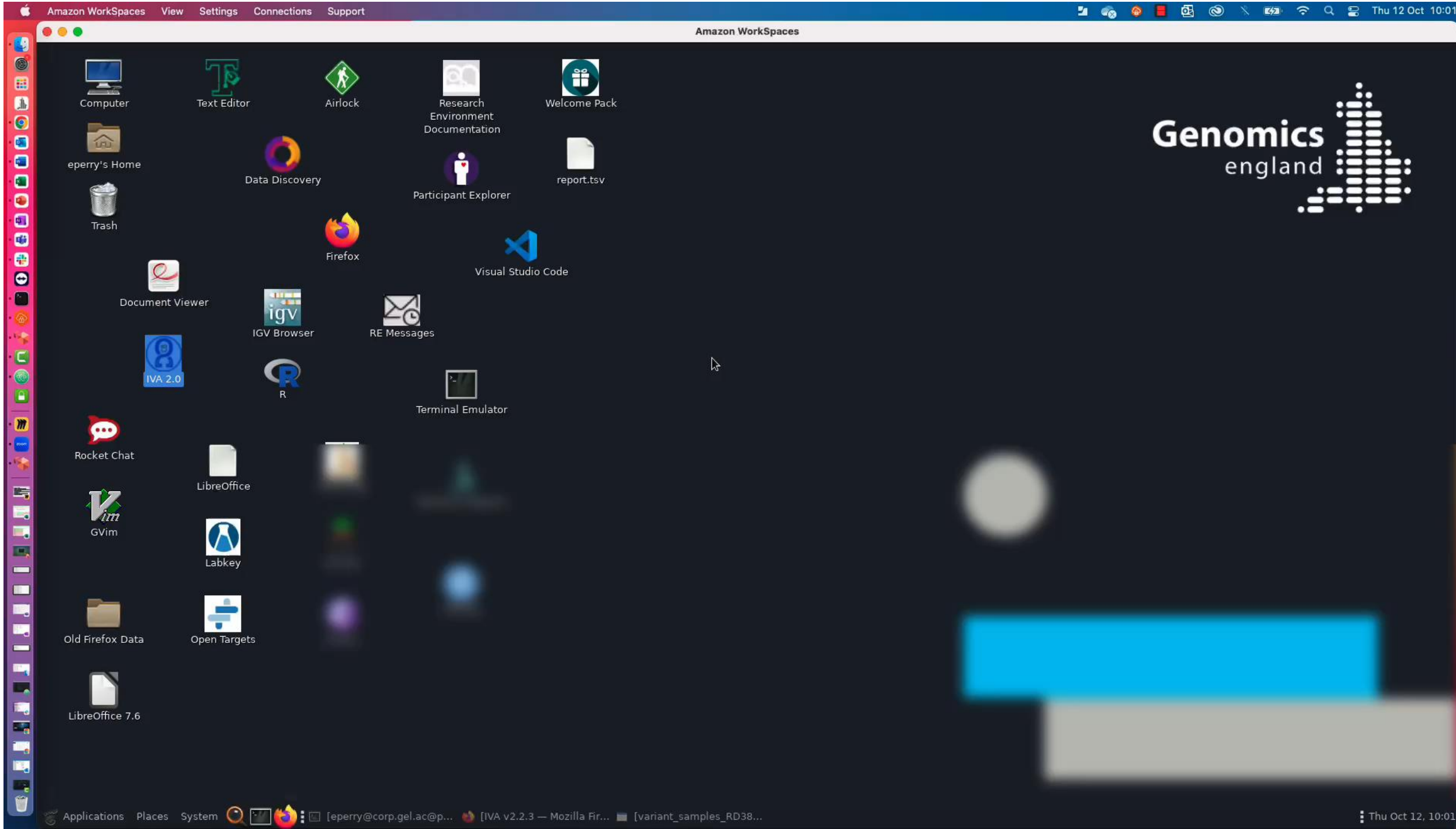
https://re-docs.genomicsengland.co.uk/medhist_api/

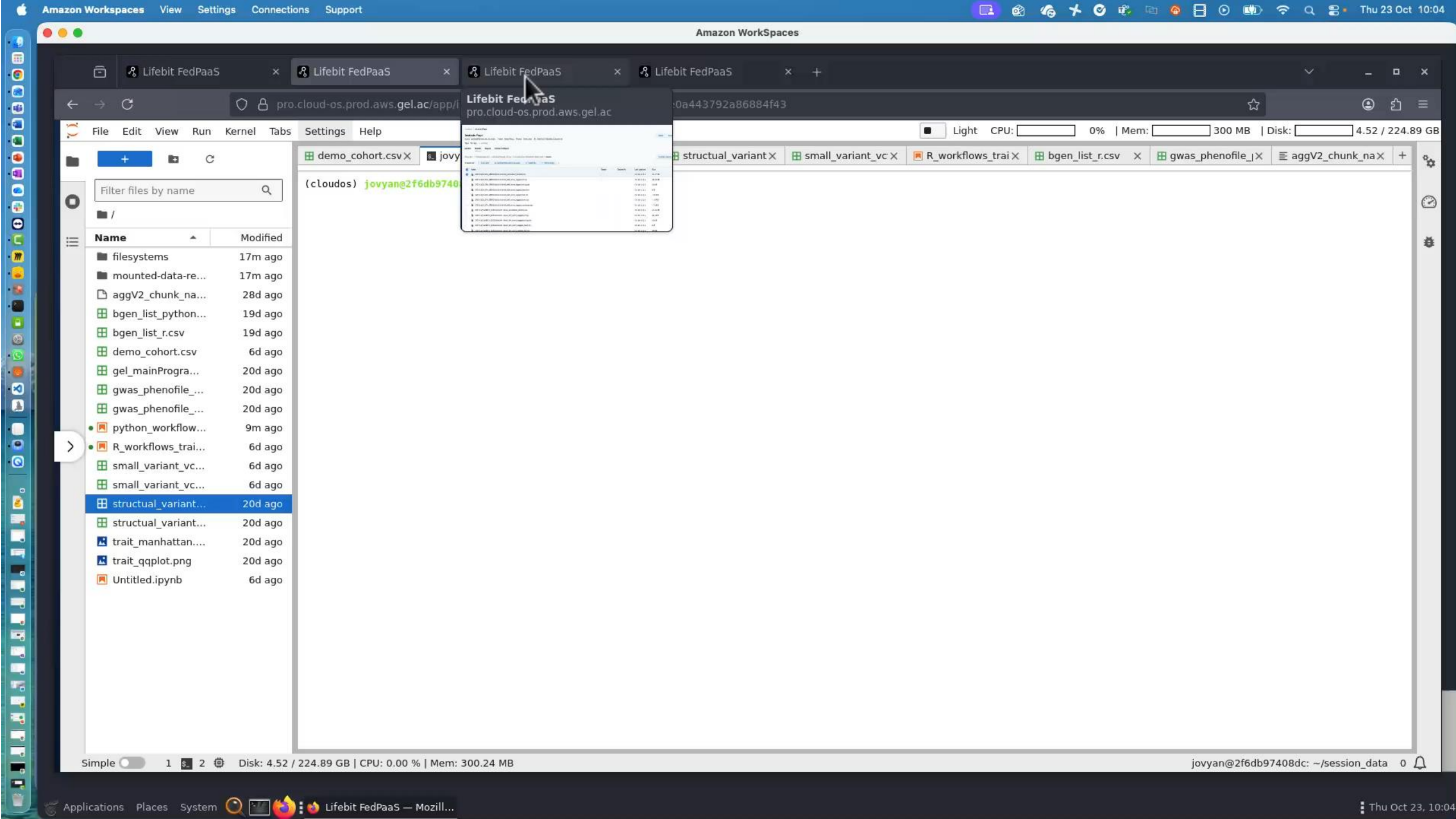
Find phenotypes in cohorts – Interactive session

- Use SQL queries
- Tutorials in python and R
- Detailed medical history tables

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

Find phenotypes in cohorts demo


Amazon WorkSpaces View Settings Connections Support
Thu 12 Oct 10:01
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Document Viewer
IGV Browser
RE Messages
IVA 2.0
R
Terminal Emulator
Rocket Chat
LibreOffice
GVim
Labkey
Old Firefox Data
Open Targets
LibreOffice 7.6
Genomics england
Applications Places System
[eperry@corp.gel.ac@p...
[IVA v2.2.3 — Mozilla Fir...
[variant_samples_RD38...
Thu Oct 12, 10:01

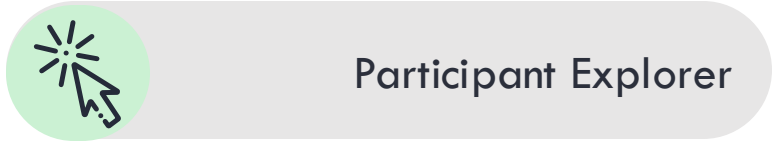

Amazon Workspaces View Settings Connections Support
Thu 23 Oct 10:04
Amazon WorkSpaces
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
Lifebit FedPaaS
pro.cloud-os.prod.aws.gel.ac
File Edit View Run Kernel Tabs Settings Help
Light CPU: 0% | Mem: 300 MB | Disk: 4.52 / 224.89 GB
demo_cohort.csv
structual_variant
small_variant_vc
R_workflows_trai
bgen_list_r.csv
gwas_phenofile_
aggV2_chunk_na
(cloudos) jovyan@2f6db9740
Filter files by name
/
Name
Modified
filesystems 17m ago
mounted-data-re... 17m ago
aggV2_chunk_na... 28d ago
bgen_list_python... 19d ago
bgen_list_r.csv 19d ago
demo_cohort.csv 6d ago
gel_mainProgra... 20d ago
gwas_phenofile_... 20d ago
gwas_phenofile_... 20d ago
python_workflow... 9m ago
R_workflows_trai... 6d ago
small_variant_vc... 6d ago
small_variant_vc... 6d ago
structual_variant... 20d ago
structual_variant... 20d ago
trait_manhattan.... 20d ago
trait_qqplot.png 20d ago
Untitled.ipynb 6d ago
Simple 1 2 Disk: 4.52 / 224.89 GB | CPU: 0.00 % | Mem: 300.24 MB
jovyan@2f6db97408dc: ~/session_data 0
Applications Places System
Lifebit FedPaaS — Mozill...
Thu Oct 23, 10:04

Find phenotypes in cohorts - phenotypes

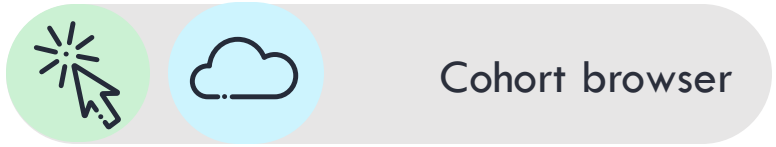
Participant Explorer	Words	Ontology aware
Cohort browser	Codes	Graphical output
LabKey API	Codes	
Interactive session	Codes	

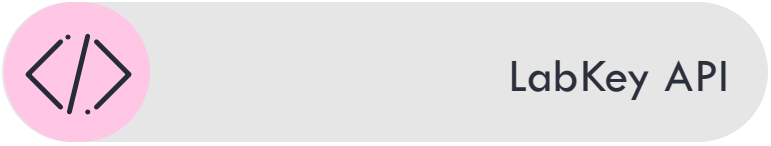
https://re-docs.genomicsengland.co.uk/phenotype_to_variants/

Find phenotypes in cohorts - scale



Limited

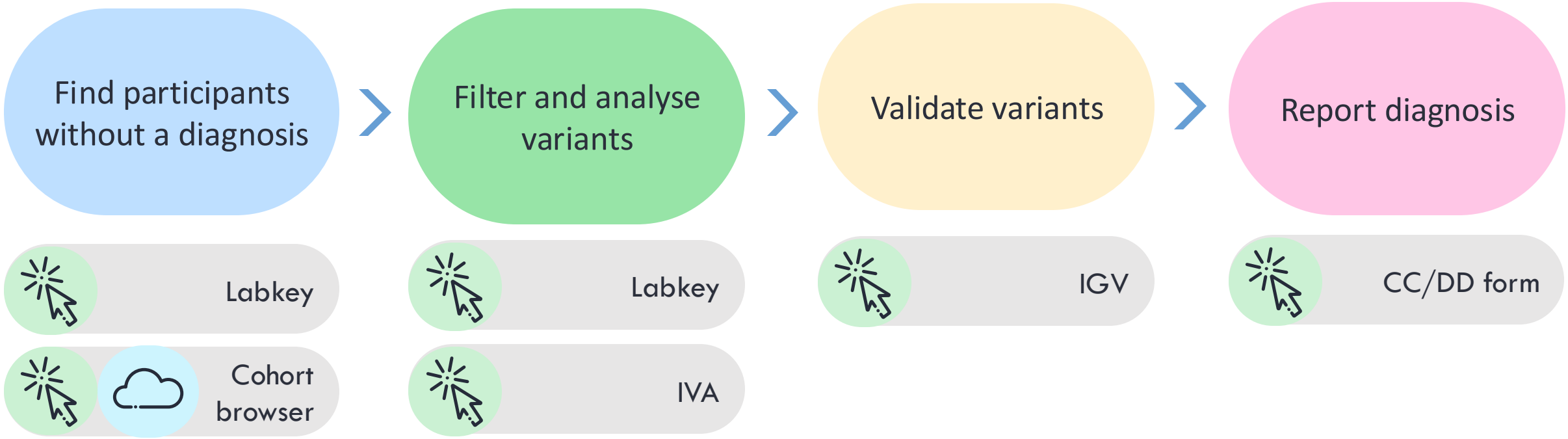




https://re-docs.genomicsengland.co.uk/phenotype_to_variants/

4. Finding diagnoses for patients who didn't get one through primary clinical interpretation

Finding diagnoses for patients who didn't get one through primary clinical interpretation



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

Find participants without a diagnosis - Labkey

Search for participants without solved cases

See unapproved diagnoses

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

Find participants without a diagnosis – Cohort Browser

Search for participants without solved cases

See unapproved diagnoses

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

Filter and analyse variants - Labkey

e variants prioritised by tiering and exomiser

Look at panels used

https://re-docs.genomicsengland.co.uk/tiering/
https://re-docs.genomicsengland.co.uk/exomiser/

Filter and analyse variants – IVA case interpreter

See all variants in a participant

Filter variants by

Genes, consequences, variant frequency etc

Family genotypes

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

Validate variants – IGV

- Identify family BAM file locations
- Visualise in IGV

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

Report diagnosis – CI/DD form

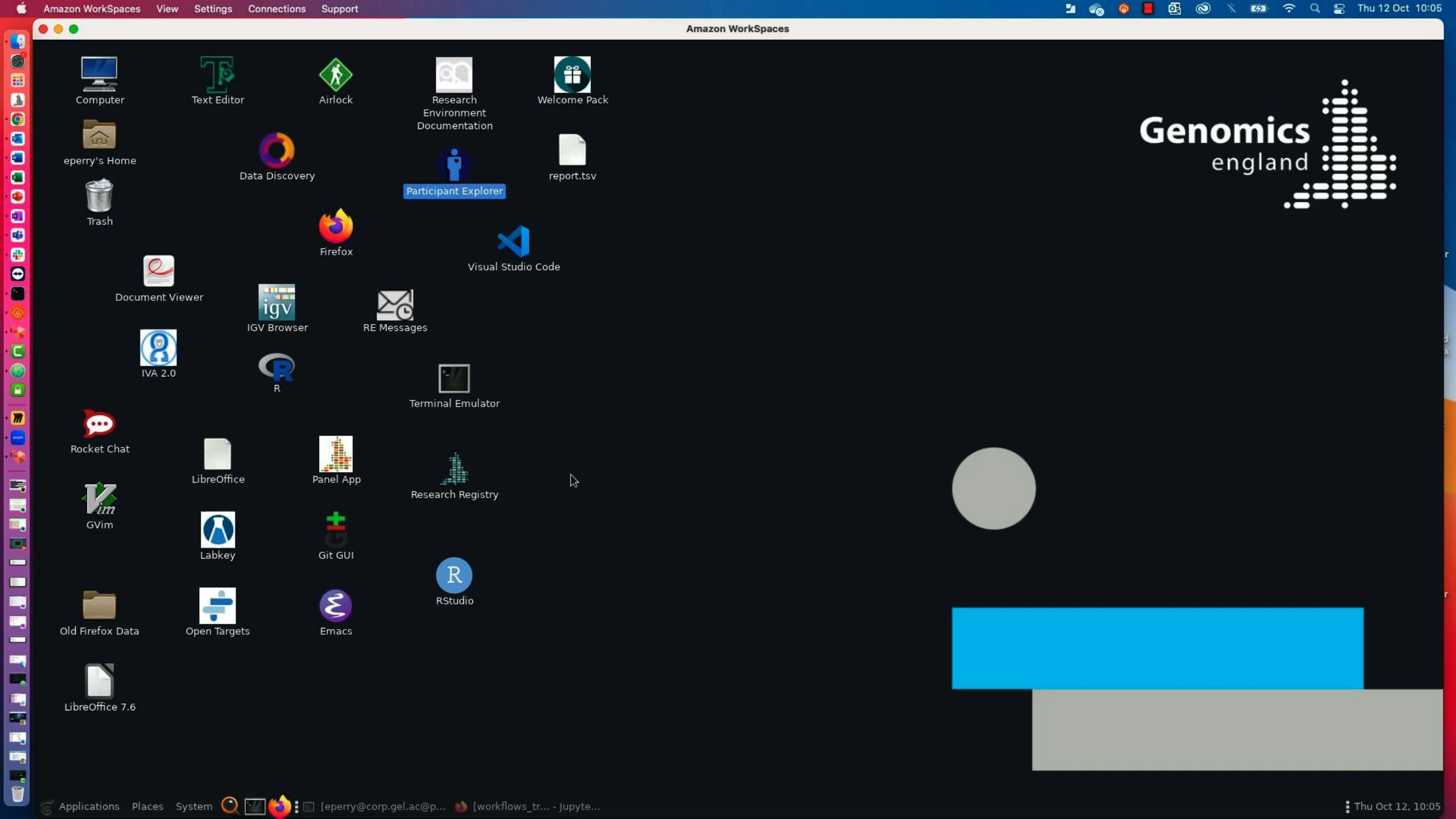
Request to contact clinician for:

- Consent to publish
- More samples or phenotype data

Report diagnosis to GEL

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

Diagnostic discovery demo


Amazon WorkSpaces View Settings Connections Support
Thu 12 Oct 10:05
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Document Viewer
IGV Browser
RE Messages
IVA 2.0
R
Terminal Emulator
Rocket Chat
LibreOffice
Panel App
Research Registry
GVim
Labkey
Git GUI
RStudio
Old Firefox Data
Open Targets
Emacs
LibreOffice 7.6
Genomics england
Applications Places System
[eperry@corp.gel.ac@p...
[workflows_tr... - Jupyte...
Thu Oct 12, 10:05

5. 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

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

Training sessions

3rd Tuesday every month

Introduction to the RE

18/11

16/12

Materials from past training all online

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

Training sessions

9/12	Running workflows on the HPC and Cloud
12/1	Using the Research Environment for clinical diagnostic discovery
9/2	Importing tools and data to use in the Research Environment

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

Help Shape the Future of the NGRL – Complete the User Survey

RN research-network@genomicsengland.co.uk
To Cerys Ballard

Unsubscribe

Dear Cerys,

Over 1,500 researchers access the National Genomic Research Library (NGRL) and we need your help to ensure we are best serving all your research needs.

We are launching an **annual user survey** and your input will help us:

- **Prioritise data and Research Environment enhancements**
- **Improve support** to maximise your research impact
- **Ensure responsible data use**

Complete the survey: https://surveys.genomicsengland.co.uk/s/Z43ZTG/?m=92073425yahlv

It will take less than 10 minutes and your feedback will directly inform strategic planning and resource allocation. Please note that your link is unique to your account, so please do not share it with others.

Survey closes 20 November 2025.

Thank you for helping us build a better research ecosystem.

Genomics England

Feedback

Thank you

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