Introduction to the Genomics England Research Environment

Emily Perry

Research Engagement Manager

21st January 2025

Questions

All your microphones are muted

Use the Zoom Q&A to ask questions

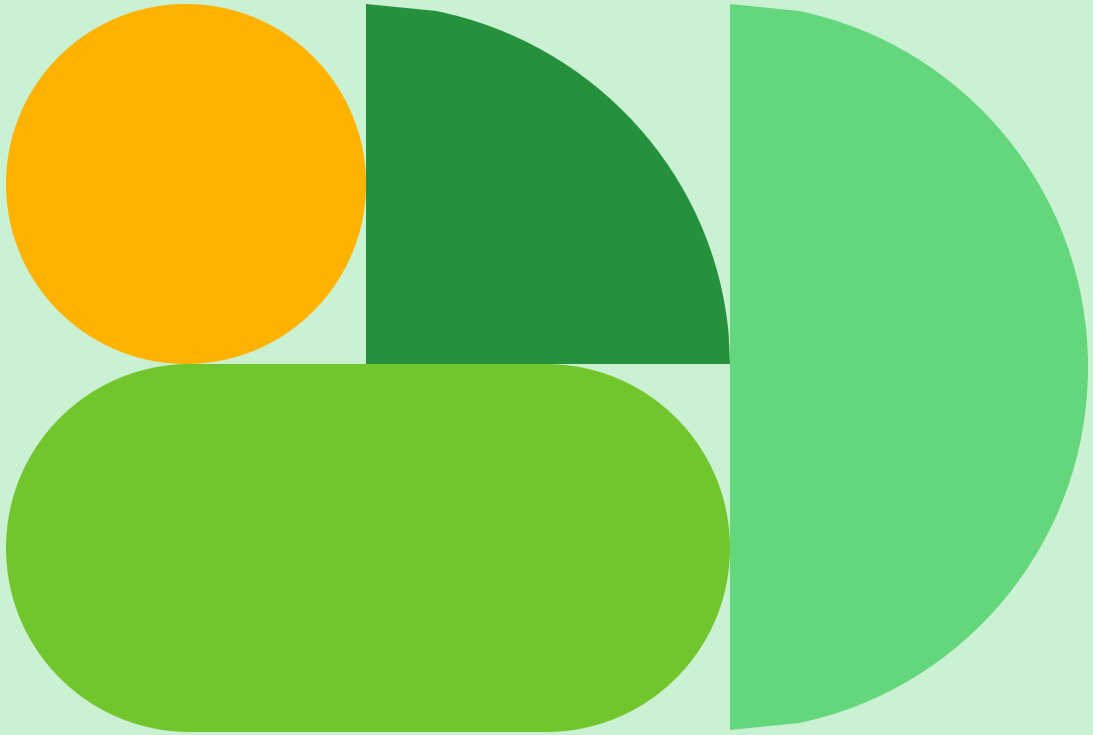
Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Helpers

Matthieu Vizuete-Forster
Bioinformatician - Research Services

Agenda

1. Introduction and admin
2. What is the RE?
3. Sources and types of data
4. Tools in the Research Environment
5. Programmatic access to NGRL data
6. Running command line tools and pipelines on the HPC
7. Import and export of data and tools
8. Help and questions



2. What is the RE?

A virtual machine

Genomic data

Phenotype data

Point and click tools

HPC and command line tools

A Trusted Research Environment

You can…

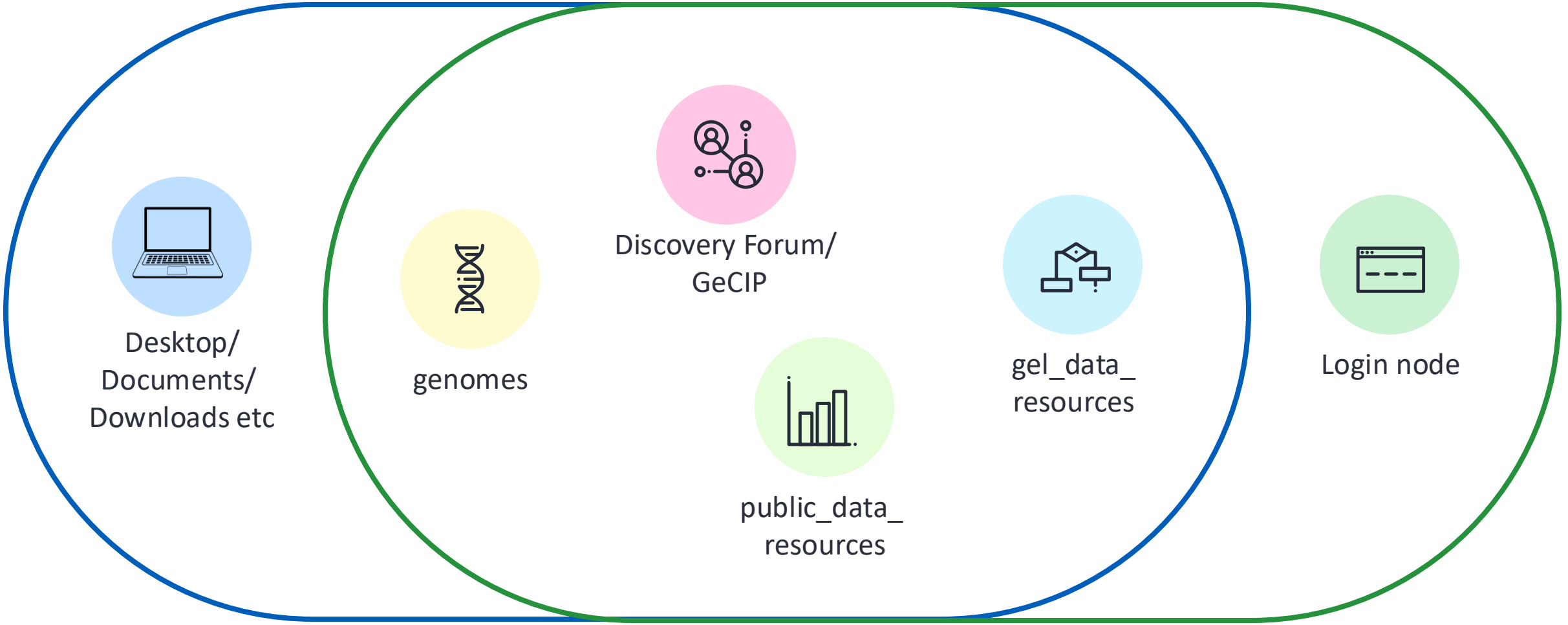
- Work with the data in the RE
- Copy/paste in
- Bring in Containers
- Access whitelisted websites
- Request to export the results of your analysis

You cannot…

- Share folders between your computer and the VM
- Copy/paste out
- Export files
- Access most of the internet

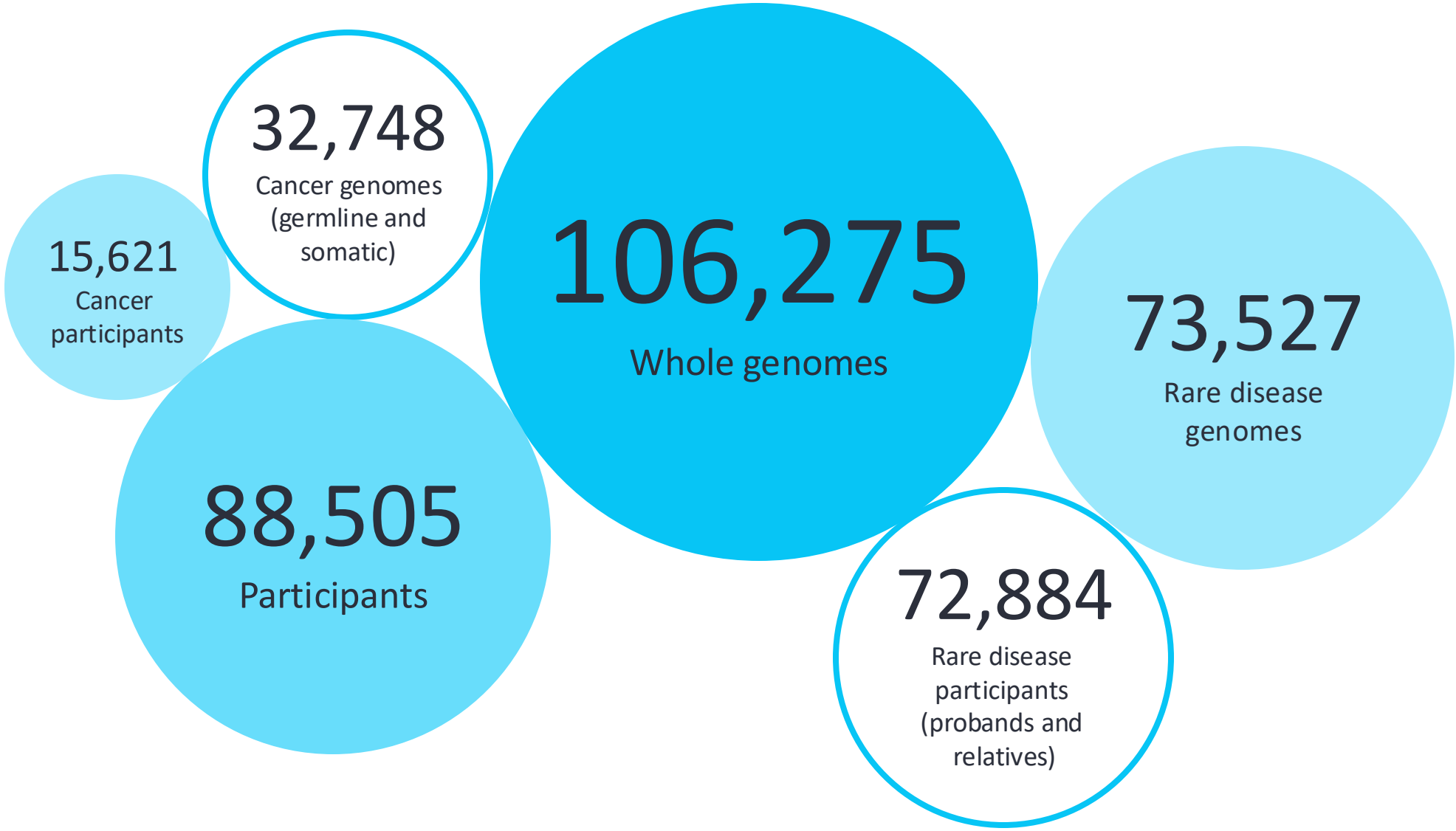
https://re-docs.genomicsengland.co.uk/data_security/

Files in the RE



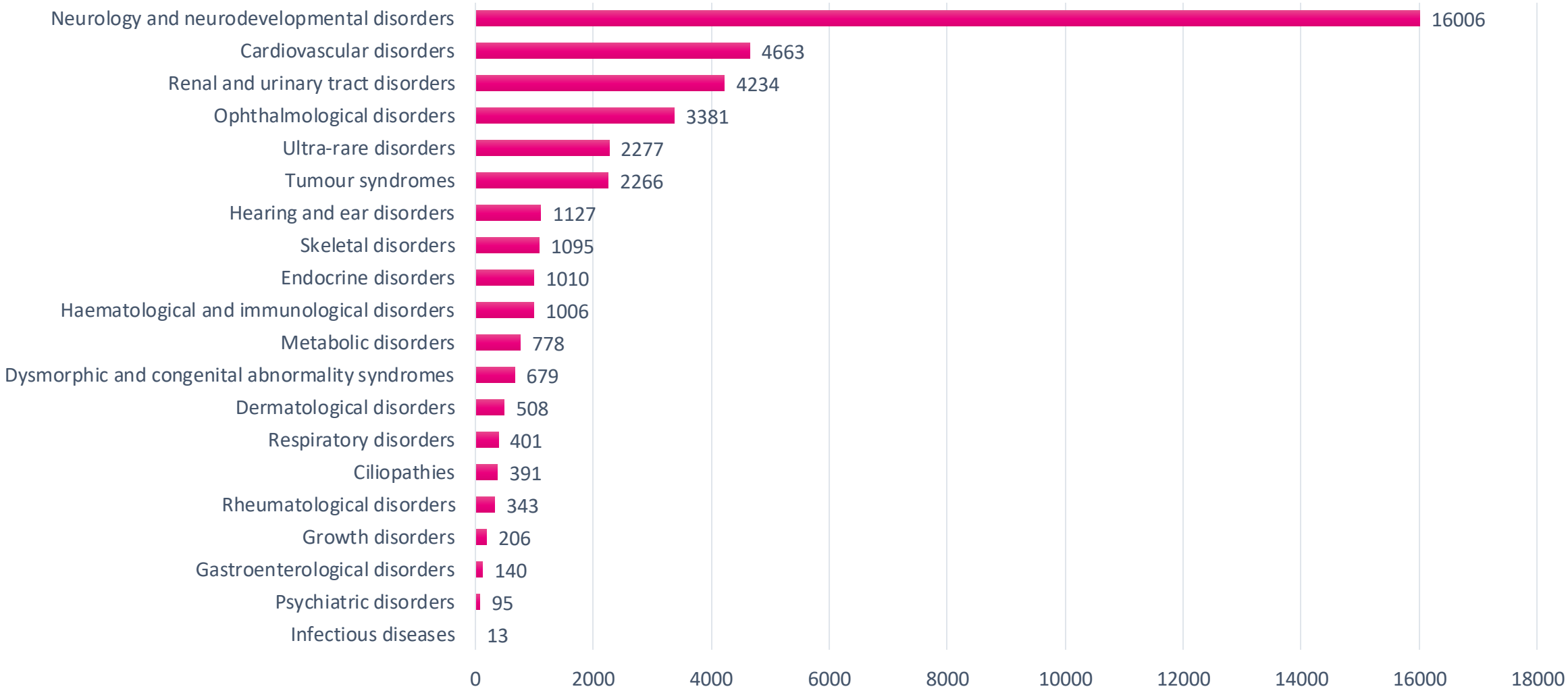
3. Sources and types of data

100,000 Genomes Project

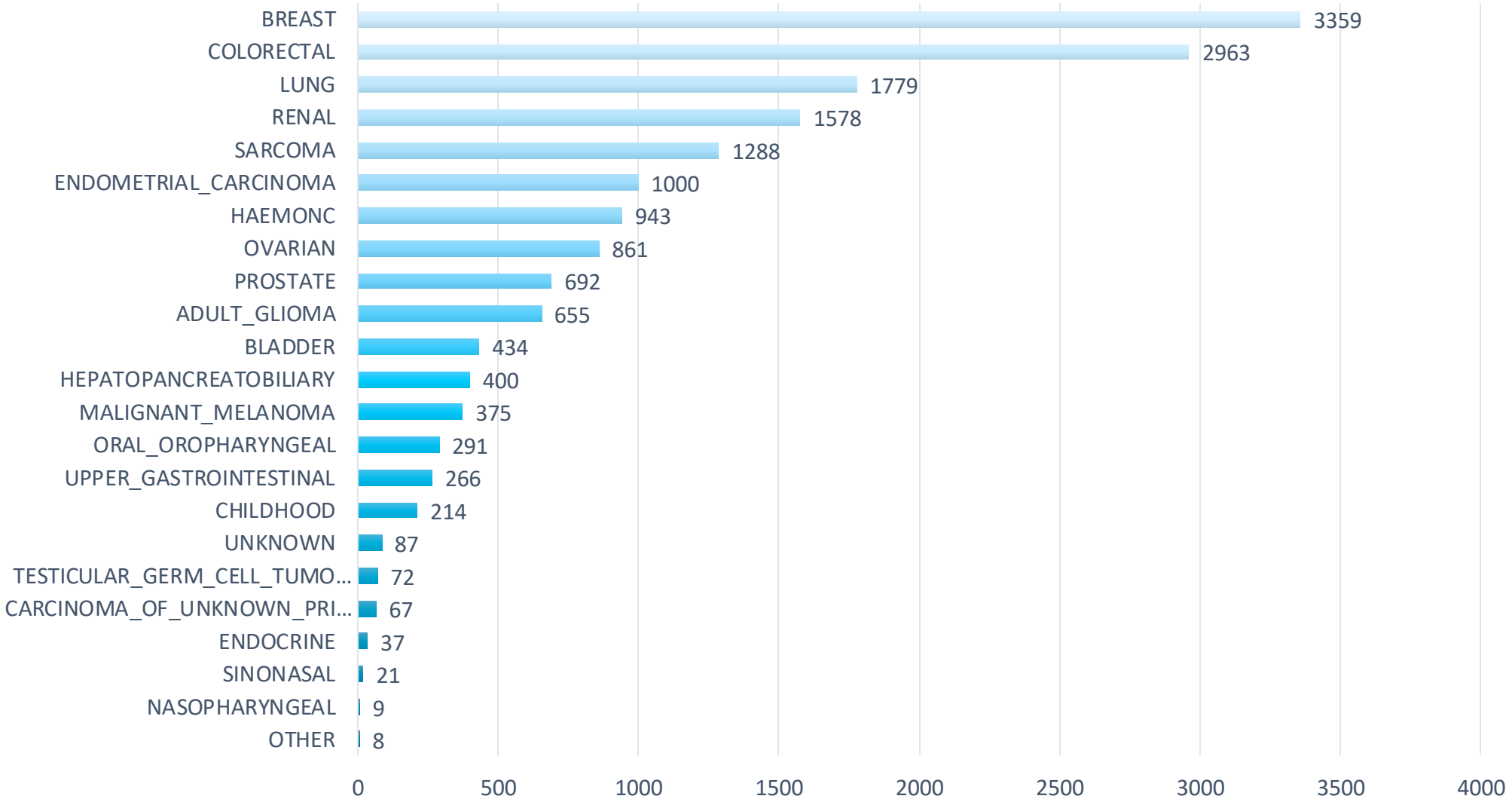


https://re-docs.genomicsengland.co.uk/release19/

100,000 Genomes rare disease



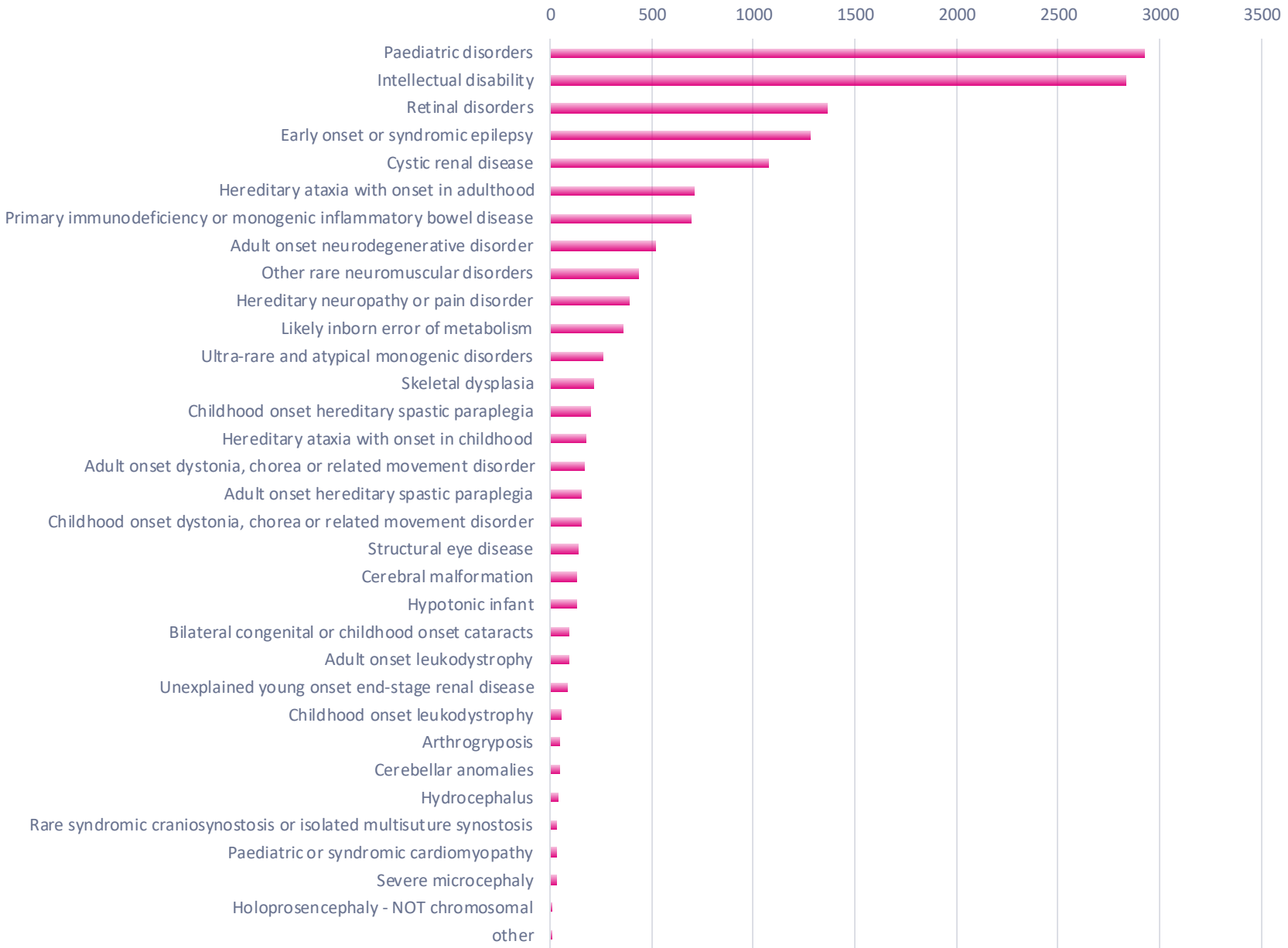
100,000 Genomes cancer



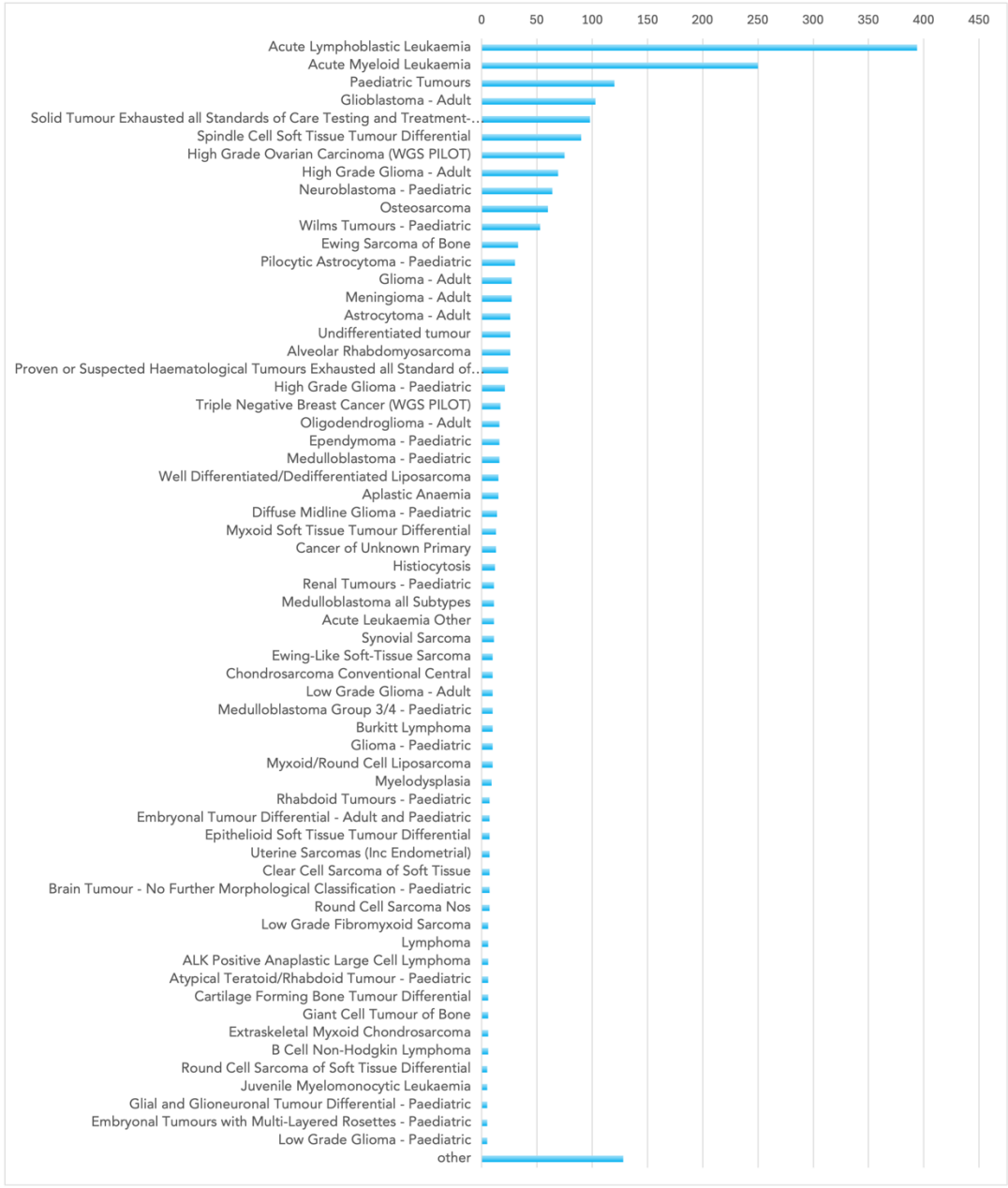
NHS GMS

https://re-docs.genomicsengland.co.uk/gms_release4/

NHS GMS rare disease



NHS GMS cancer



For EVERY genome

Alignment

as BAM or CRAM files

Variant calls

as VCFs, including gVCF, repeats VCFs, structural variant VCFs

Analysis

Variant tiering and tumour mutational signatures

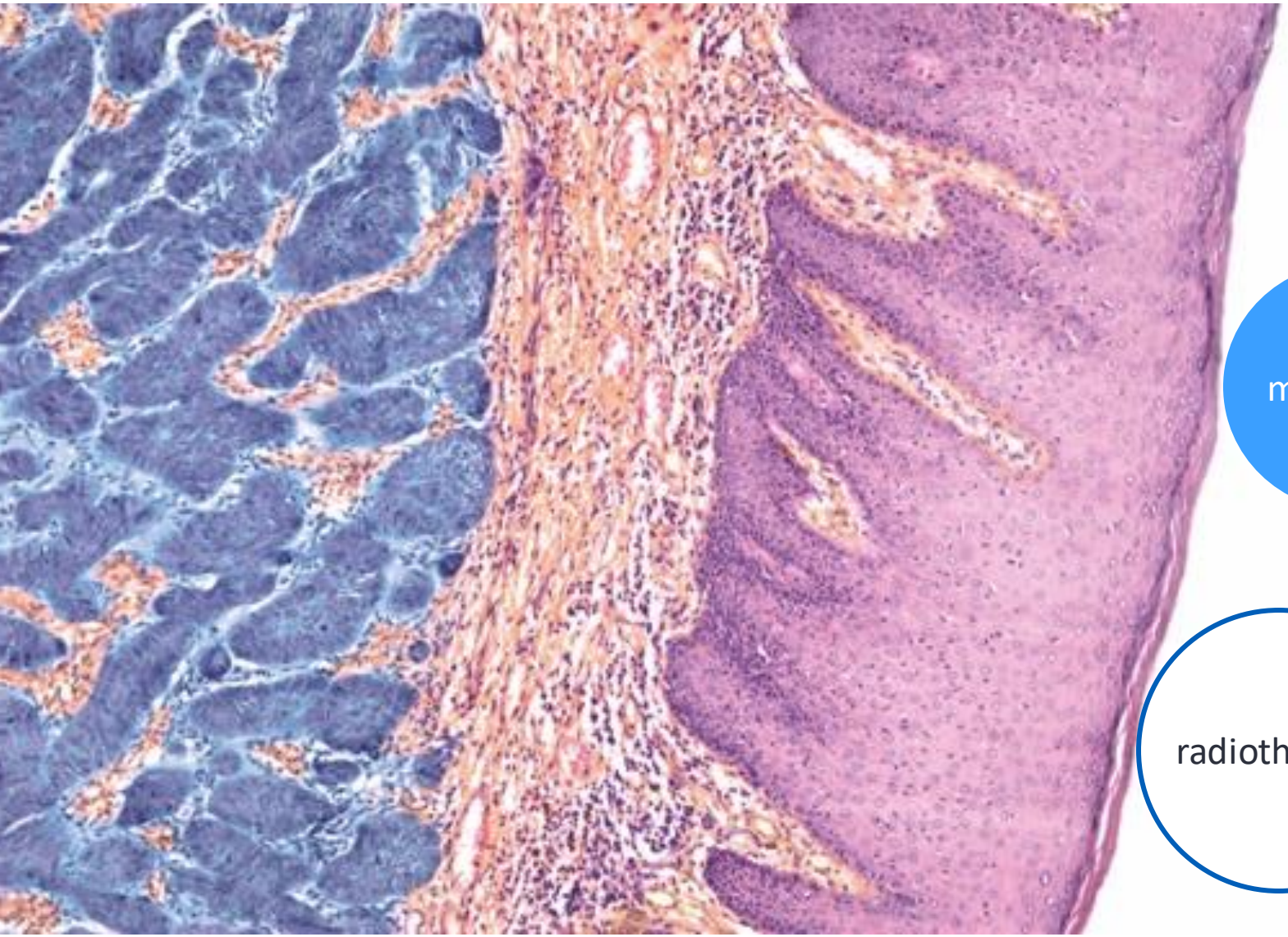
https://re-docs.genomicsengland.co.uk/genomic_data/

Rare disease phenotyping

- Disease classification
- HPO terms present/absent
- Measurements and observations (not universal)
    - general measurements
    - early childhood observations
    - details of imaging (but not results)
    - genetic tests
    - lab tests

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

Cancer tumour characterisation



risk factors

cancer type

surgeries

staging and grading

metastasis

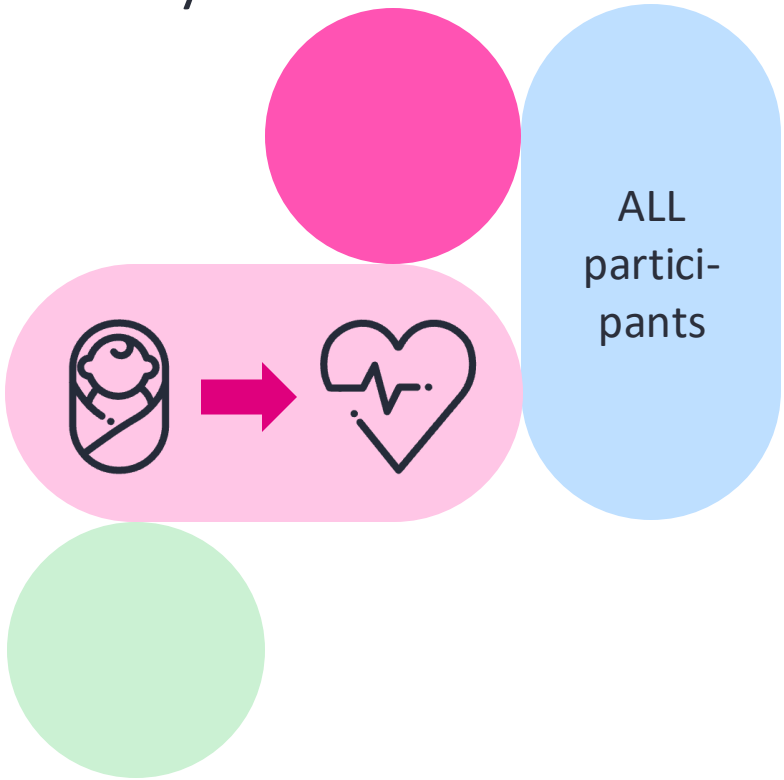
chemotherapy

radiotherapy

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

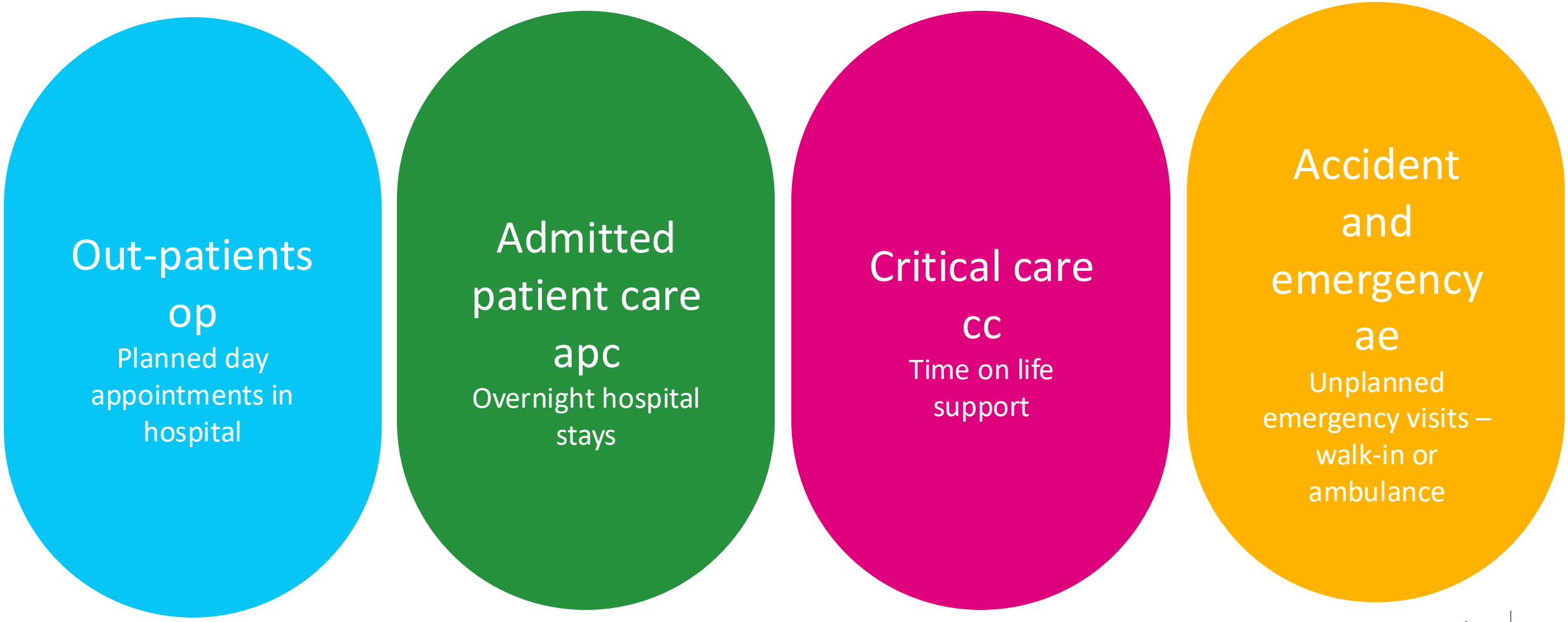
Medical history

- NHSE hospital episode statistics
- Mental health data
- Mortality



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

Hospital episode statistics

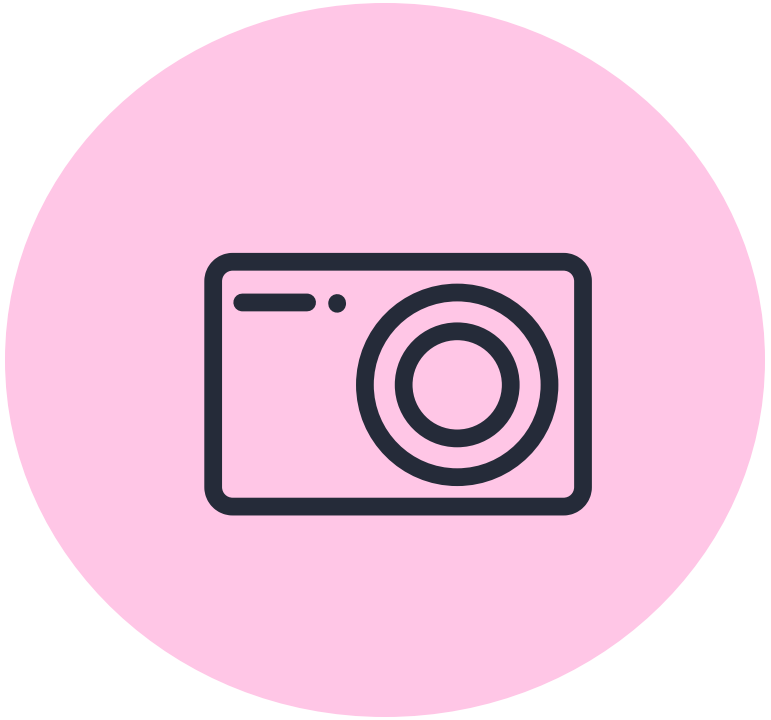


https://re-docs.genomicsengland.co.uk/general_clinical/#hospital-episodes-statistics-from-nhse

We don’t have…

Free text

GP data



diagnostic imaging

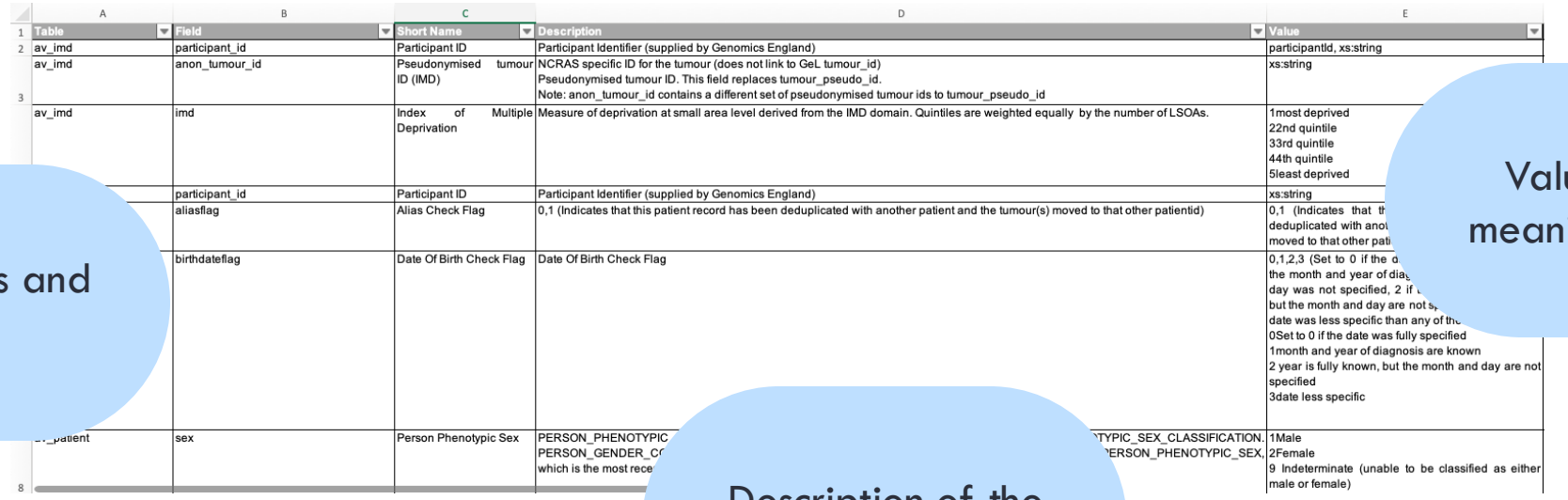
4. Tools in the Research Environment

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/

Data dictionary



	A	B	C	D	E
1	Table	Field	Short Name	Description	Value
2	av_imd	participant_id	Participant ID	Participant Identifier (supplied by Genomics England)	participantId, xs:string
3	av_imd	anon_tumour_id	Pseudonymised tumour ID (IMD)	NCRAS specific ID for the tumour (does not link to GeL tumour_id) Pseudonymised tumour ID. This field replaces tumour_pseudo_id. Note: anon_tumour_id contains a different set of pseudonymised tumour ids to tumour_pseudo_id	xs:string
	av_imd	imd	Index of Multiple Deprivation	Measure of deprivation at small area level derived from the IMD domain. Quintiles are weighted equally by the number of LSOAs.	1most deprived 22nd quintile 33rd quintile 44th quintile 5least deprived
		participant_id	Participant ID	Participant Identifier (supplied by Genomics England)	xs:string
		aliasflag	Alias Check Flag	0,1 (Indicates that this patient record has been deduplicated with another patient and the tumour(s) moved to that other patientid)	0,1 (Indicates that th deduplicated with ano moved to that other pati
		birthdateflag	Date Of Birth Check Flag	Date Of Birth Check Flag	0,1,2,3 (Set to 0 if the d the month and year of dia day was not specified, 2 if but the month and day are not s date was less specific than any of th 0Set to 0 if the date was fully specified 1month and year of diagnosis are known 2 year is fully known, but the month and day are not specified 3date less specific
8	_patient	sex	Person Phenotypic Sex	PERSON_PHENOTYPIC PERSON_GENDER_C which is the most rece	TYPIC_SEX_CLASSIFICATION. ERSON_PHENOTYPIC_SEX, 1Male 2Female 9 Indeterminate (unable to be classified as either male or female)

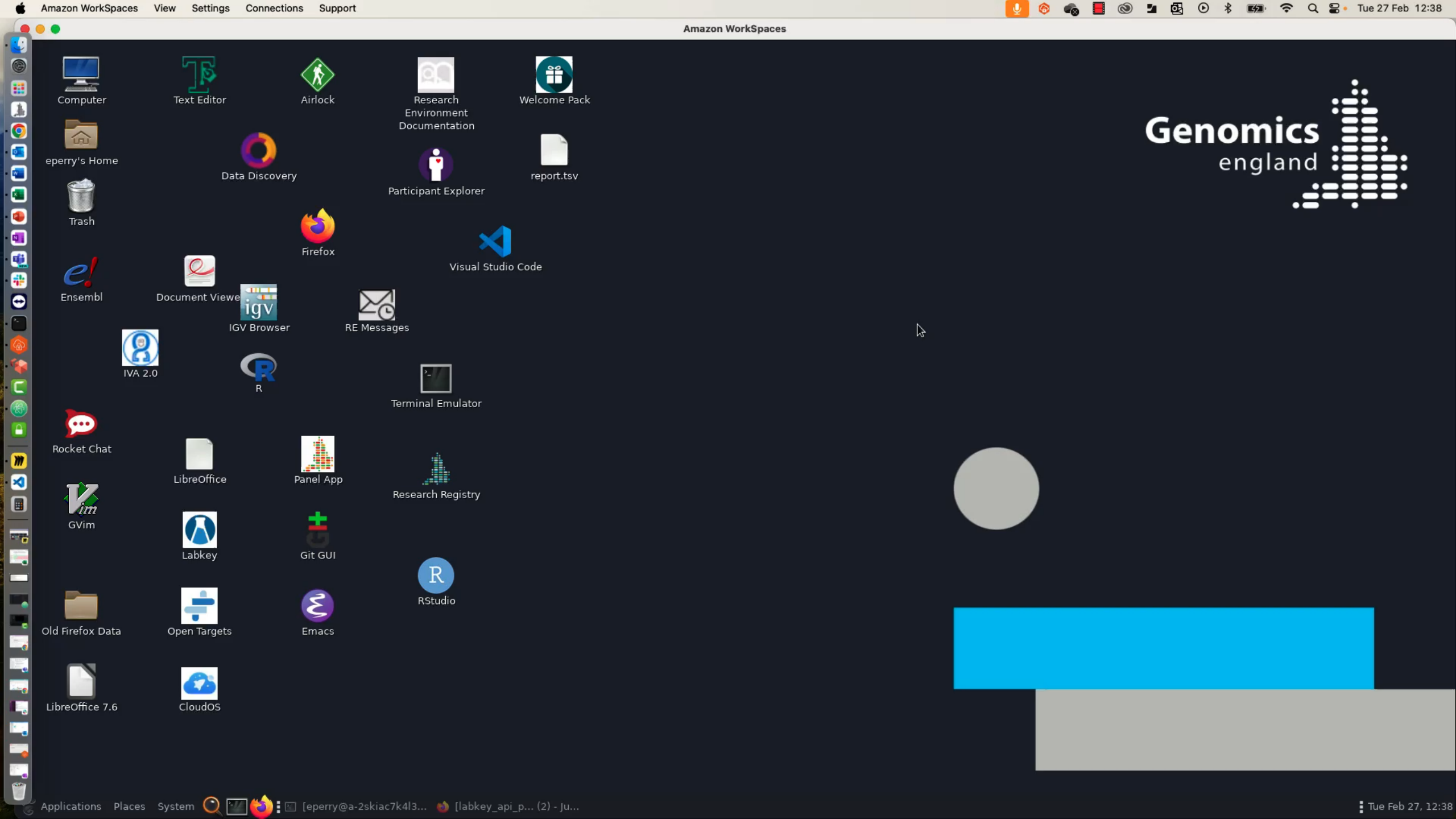
Lists of tables and columns

Value type or meaning of codes

Description of the data

https://re-docs.genomicsengland.co.uk/100k_data_dictionary https://re-docs.genomicsengland.co.uk/GMS_data_dictionary

LabKey demo

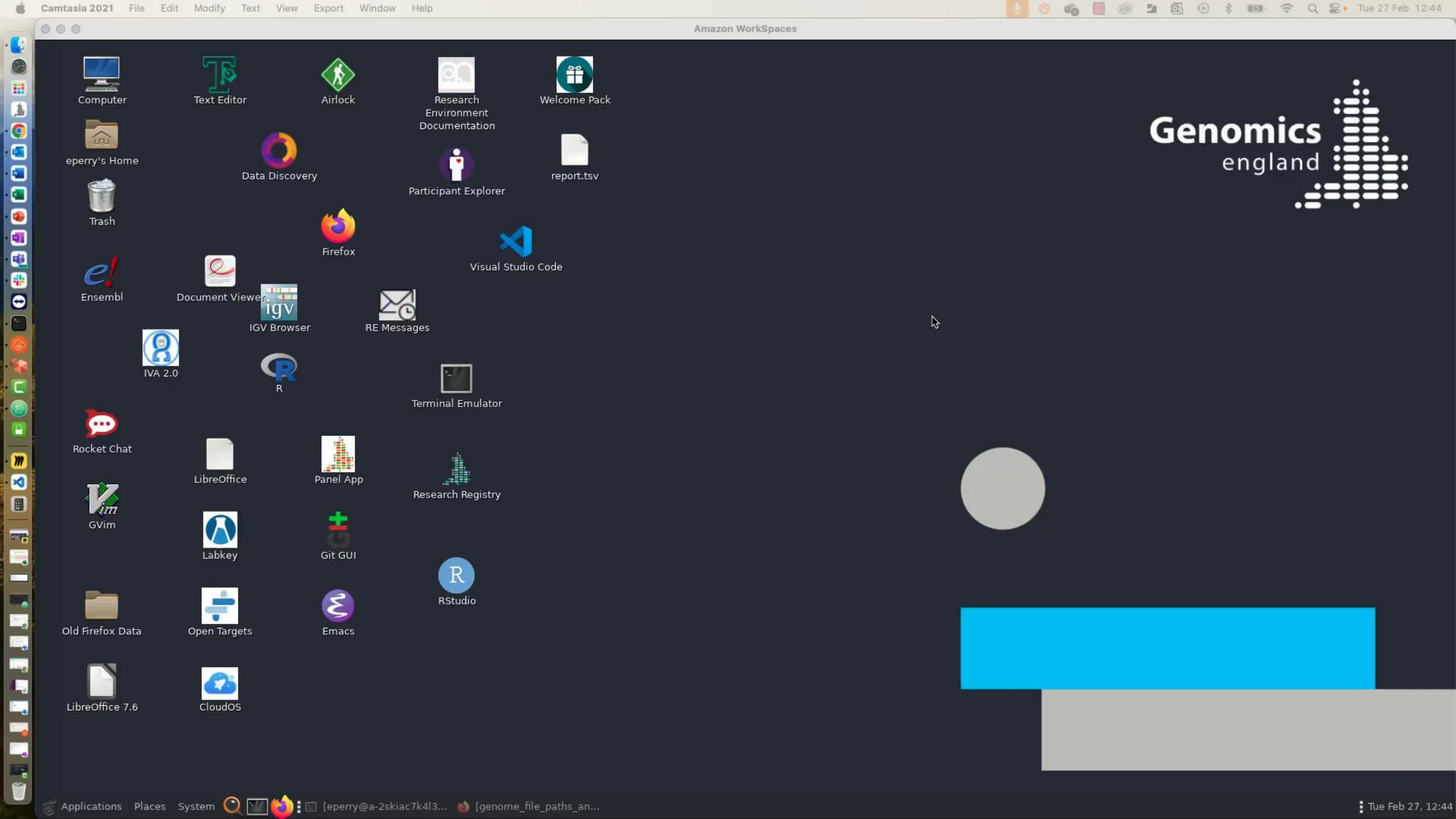

Amazon WorkSpaces View Settings Connections Support
Tue 27 Feb 12:38
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Ensembl
Document Viewer
IGV Browser
RE Messages
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
CloudOS
Genomics england
Applications Places System
[eperry@a-2skiac7k4l3...
[labkey_api_p... (2) - Ju...
Tue Feb 27, 12:38

Participant Explorer

- Search for participants by:
  - IDs
  - Clinical concepts
  - Personal details
- View/compare medical histories

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

Participant Explorer demo

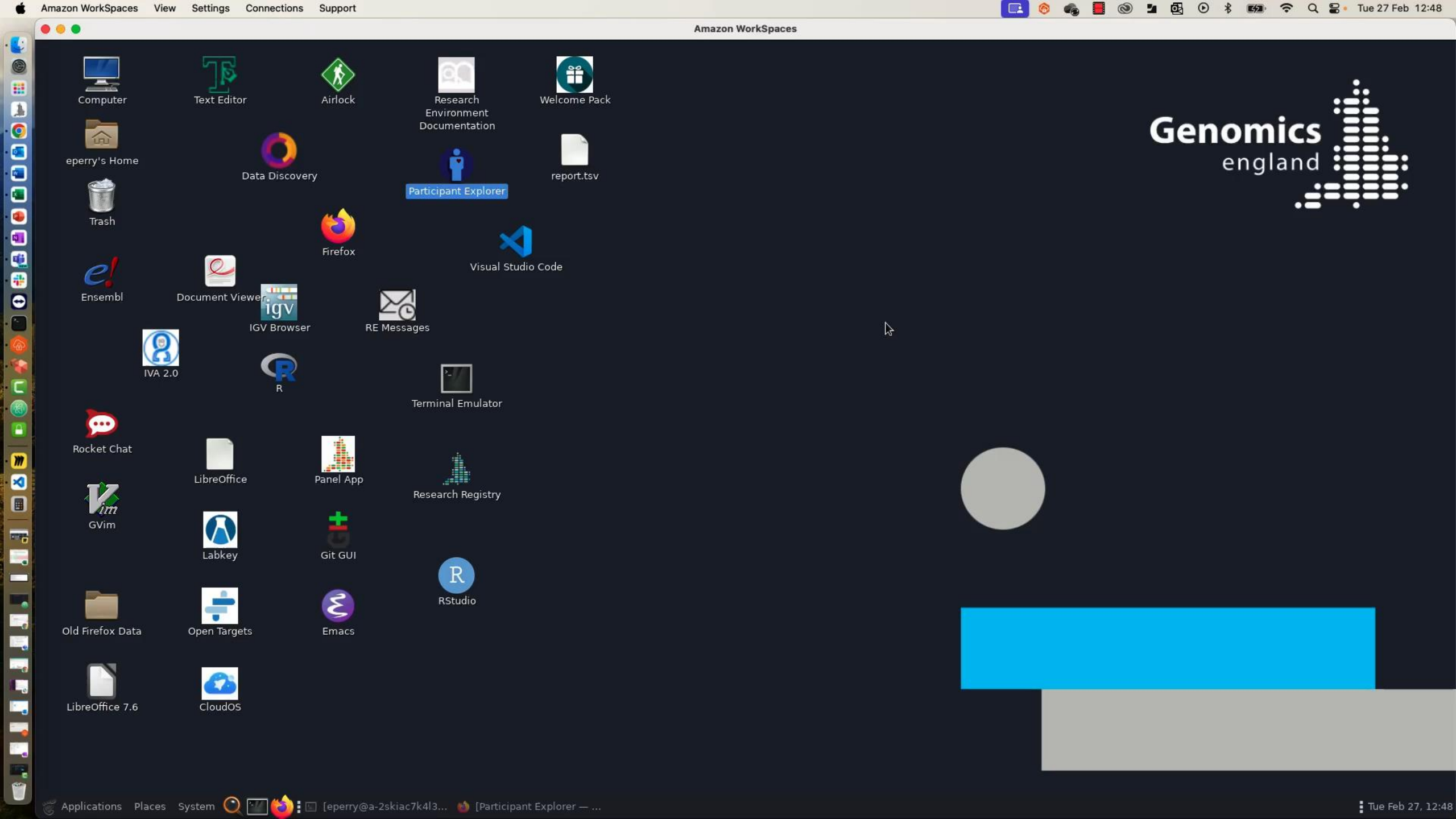

Camtasia 2021 File Edit Modify Text View Export Window Help
Tue 27 Feb 12:44
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Ensembl
Document Viewer
IGV Browser
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
CloudOS
Genomics england
Applications Places System
[eperry@a-2skiac7k4l3...
[genome_file_paths_an...
Tue Feb 27, 12:44

Interactive variant analysis (IVA)

- Find participants with genetic variants
- Filter variants in a participant by family genotypes
- Filter on genome features

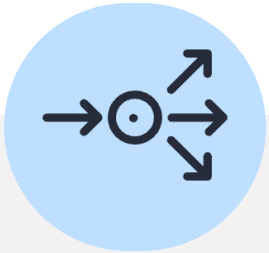
https://re-docs.genomicsengland.co.uk/iva/

IVA demo

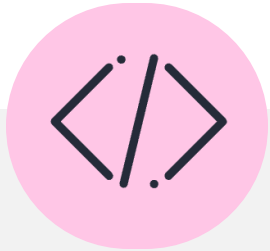

Amazon WorkSpaces View Settings Connections Support
Tue 27 Feb 12:48
Amazon WorkSpaces
Computer
Text Editor
Airlock
Research Environment Documentation
Welcome Pack
eperry's Home
Data Discovery
Participant Explorer
report.tsv
Trash
Firefox
Visual Studio Code
Ensembl
Document Viewer
IGV Browser
RE Messages
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
CloudOS
Genomics england
Applications Places System
[eperry@a-2skiac7k4l3...
[Participant Explorer — ...
Tue Feb 27, 12:48

5. Programmatic access to NGRL data

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

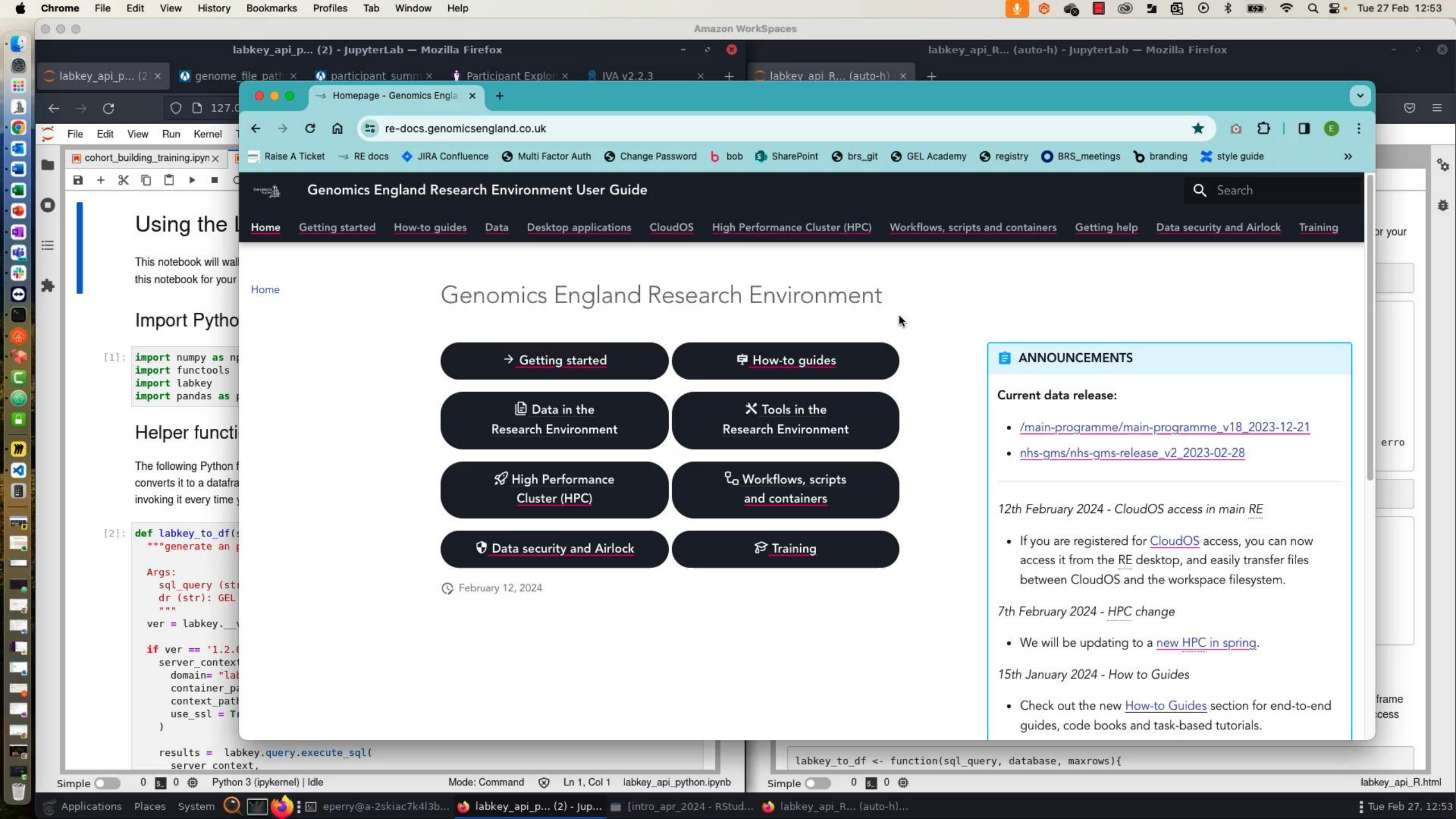
- You can access the same data via the LabKey API as you can through other means
- You will need to configure access to the LabKey API with your username and password
    - In your home directory
    - On the HPC
- You do this by editing a file called .netrc

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

Programming tools in the RE

https://re-docs.genomicsengland.co.uk/hpc_jupyter/ https://re-docs.genomicsengland.co.uk/enable_rstudio/

LabKey API demo


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Tue 27 Feb 12:53
Amazon WorkSpaces
labkey_api_p... (2) - JupyterLab — Mozilla Firefox
labkey_api_R... (auto-h) - JupyterLab — Mozilla Firefox
Homepage - Genomics Engla
re-docs.genomicsengland.co.uk
Raise A Ticket
RE docs
JIRA Confluence
Multi Factor Auth
Change Password
bob
SharePoint
brs_git
GEL Academy
registry
BRS_meetings
branding
style guide
Genomics England Research Environment User Guide
Search
Home
Getting started
How-to guides
Data
Desktop applications
CloudOS
High Performance Cluster (HPC)
Workflows, scripts and containers
Getting help
Data security and Airlock
Training
Genomics England Research Environment
Getting started
How-to guides
Data in the Research Environment
Tools in the Research Environment
High Performance Cluster (HPC)
Workflows, scripts and containers
Data security and Airlock
Training
February 12, 2024
ANNOUNCEMENTS
Current data release:
/main-programme/main-programme_v18_2023-12-21
nhs-gms/nhs-gms-release_v2_2023-02-28
12th February 2024 - CloudOS access in main RE
If you are registered for CloudOS access, you can now access it from the RE desktop, and easily transfer files between CloudOS and the workspace filesystem.
7th February 2024 - HPC change
We will be updating to a new HPC in spring.
15th January 2024 - How to Guides
Check out the new How-to Guides section for end-to-end guides, code books and task-based tutorials.
cohort_building_training.ipyn
Using the
Import Pytho
Helper functi
labkey_to_df <- function(sql_query, database, maxrows){
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
labkey_api_python.ipynb
labkey_api_R.html
Applications
Places
System
Tue Feb 27, 12:53

6. Running command line tools and pipelines on the HPC

What is an HPC?

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

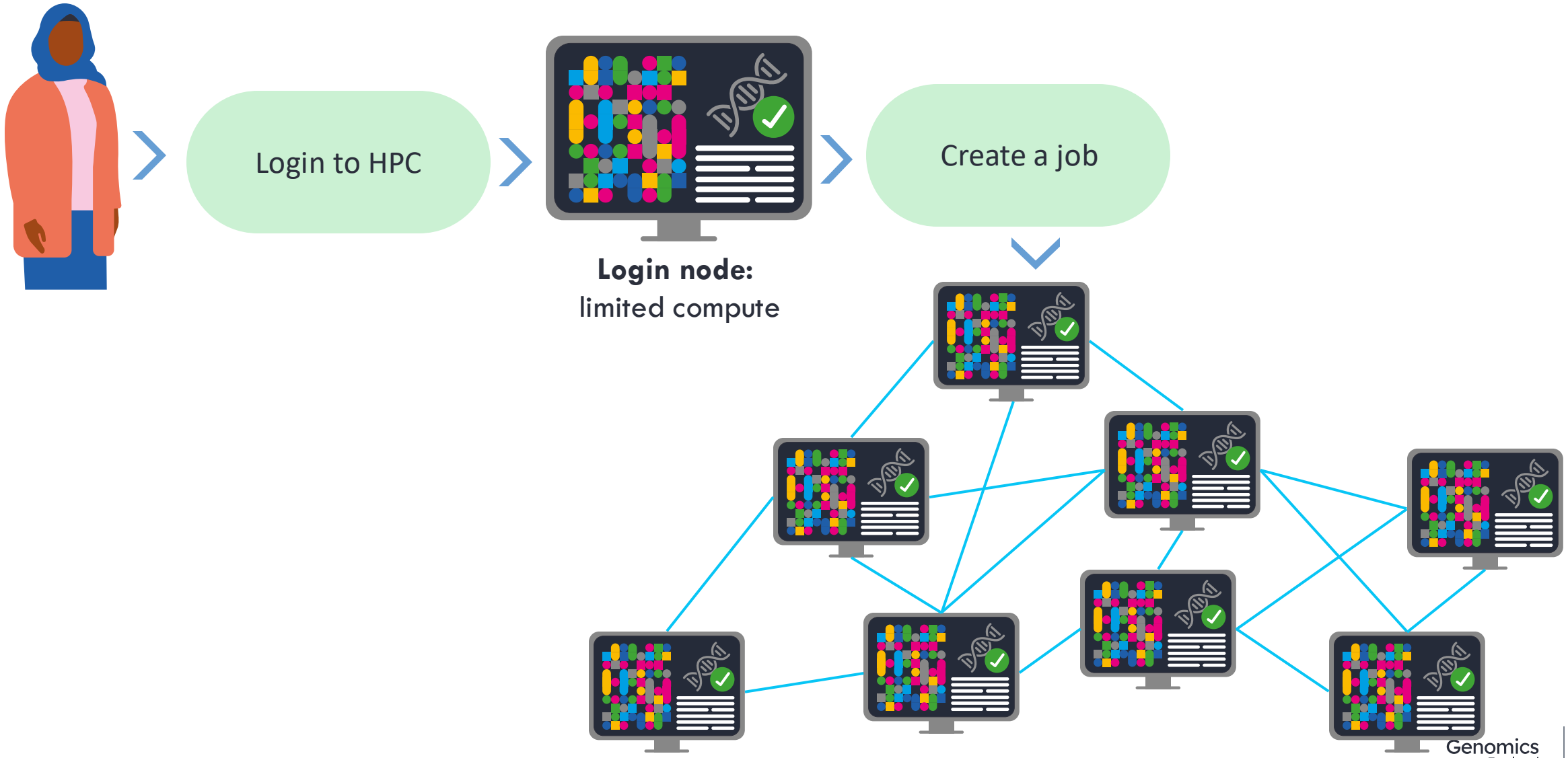
What is an HPC?

Lots of compute power

Shared with other researchers

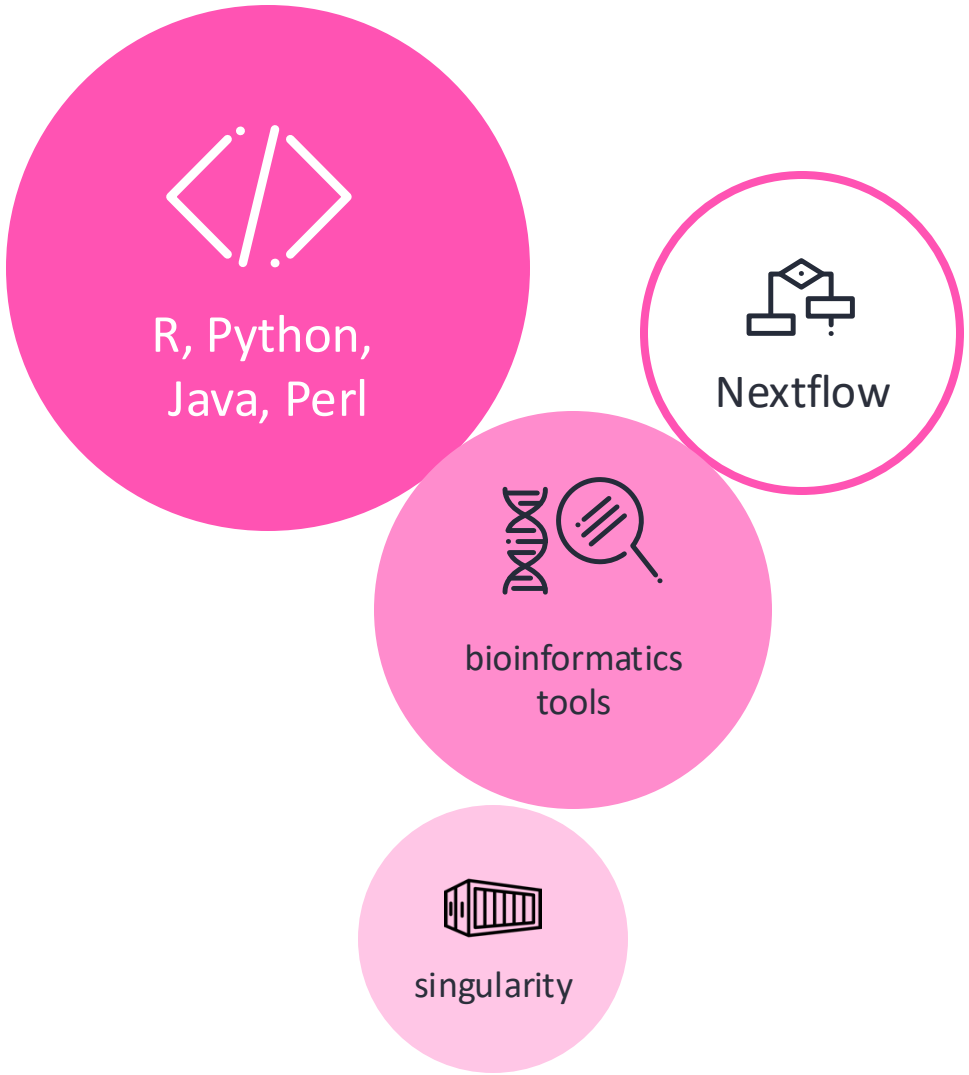
https://re-docs.genomicsengland.co.uk/hpc_about/

What is a ”job”?



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

Software on the HPC

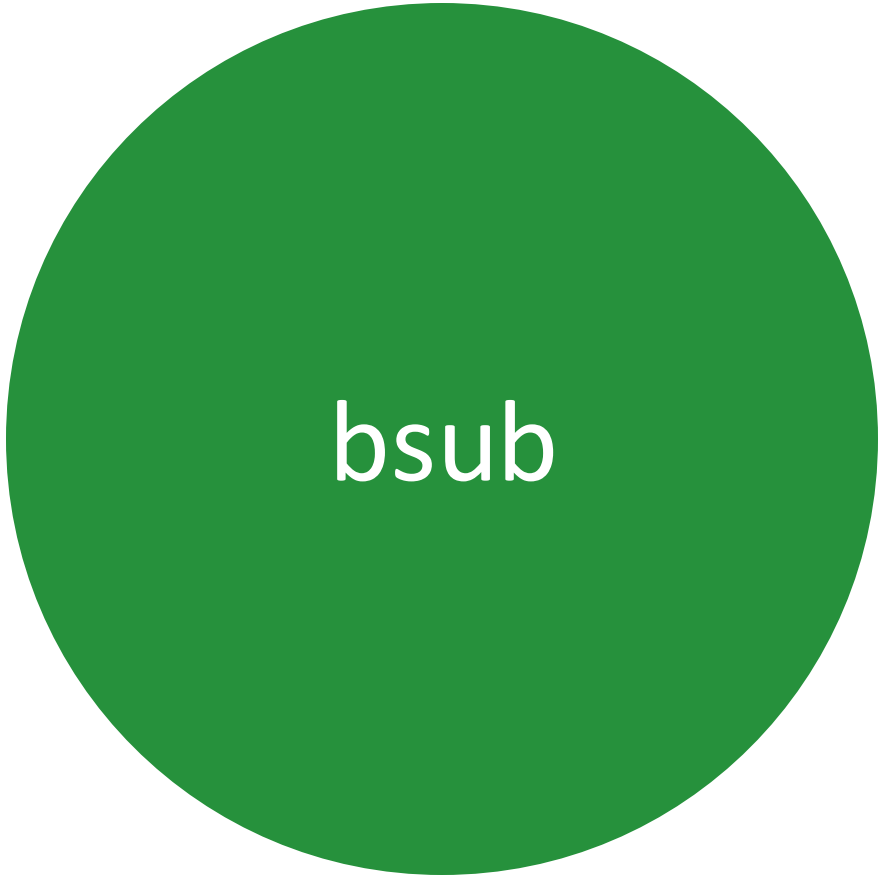


AdapterRemoval/2.3.3
aliview/1.28
ampliconArchitect/1.3.r7
ampliconClassifier/1.1.1
annotSV/3.3.7
annovar/2019Nov
annovar/2024-03-14
ant/1.9.16
apbs/3.4.1
asmc-asmc/2024-02-26
AutoDock_Vina/1.2.5
automake/1.15
aws-cli/2.15
bamtools/2.5.2
bcftools/1.16
beagle/5.4
bedops/2.4.41
bedtools/2.30.0
bedtools/2.31.0
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/

Creating a job - parameters

- -q — queue name: inter/short/medium/long
- -P — Your project ID
- -o — standard output location path_to/job.%J.out
- -e — standard error location path_to/job.%J.err
- -J — job name: useful for monitoring
- -R — Memory usage
- -M — Maximum memory
- -n — number of nodes
- -cwd — current working directory



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

Pre-built workflows/scripts to…

GWAS (common variants)
AVT (rare variants)

Small variant
Structural variant

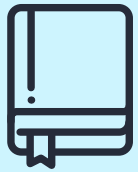
survival

VEP

Workflows/scripts provide

Code that runs with only minor tweaks to add your input

Optimised for use on our HPC and with our data



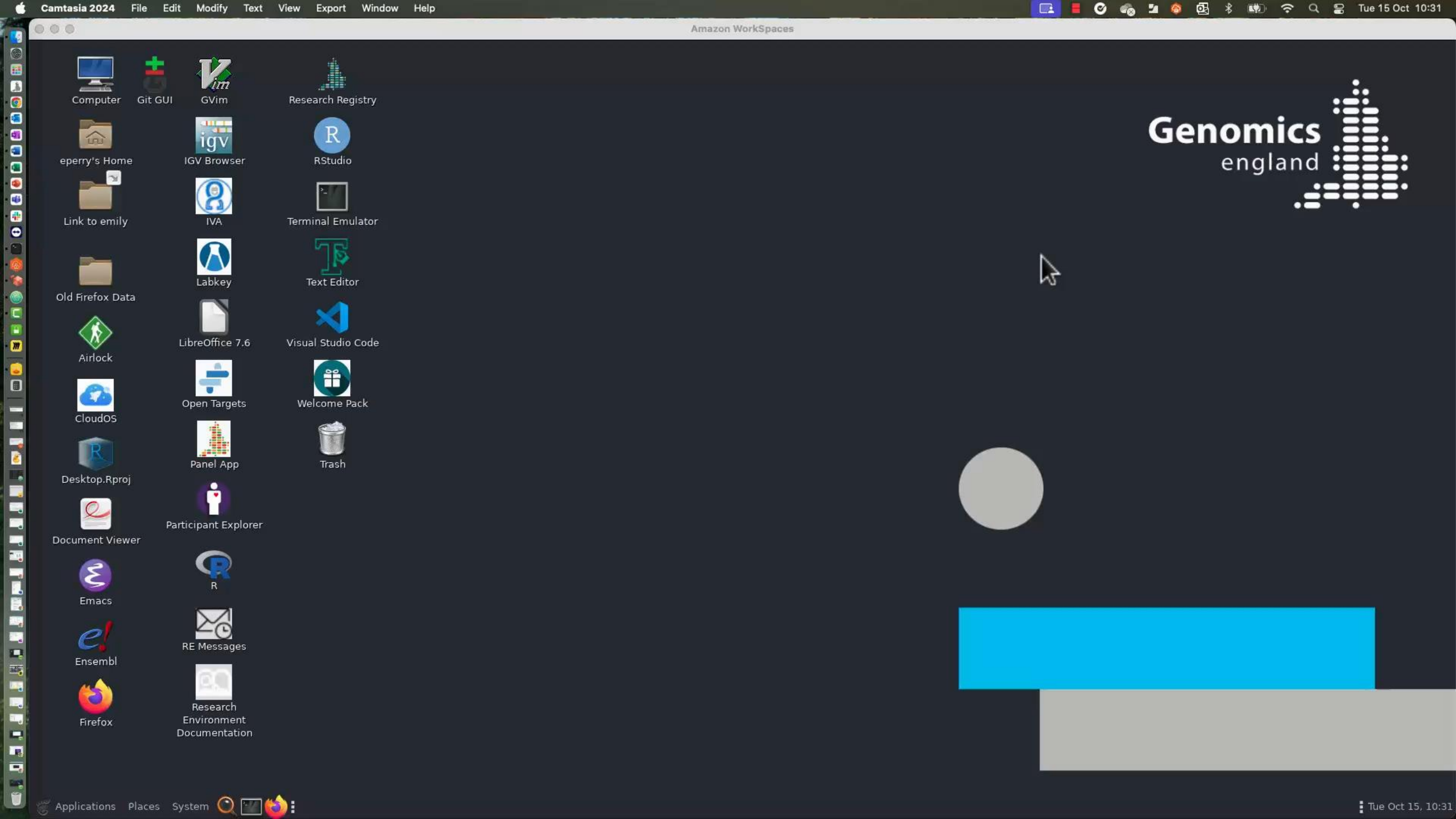
Step-by-step instructions for use

Output in standard or easy-to-interpret formats

Example input data and submission scripts

Customisable options

HPC demo


Camtasia 2024
File
Edit
Modify
Text
View
Export
Window
Help
Tue 15 Oct 10:31
Amazon WorkSpaces
Computer
Git GUI
GVim
Research Registry
eperry's Home
IGV Browser
RStudio
Link to emily
IVA
Terminal Emulator
Old Firefox Data
Labkey
Text Editor
Airlock
LibreOffice 7.6
Visual Studio Code
CloudOS
Open Targets
Welcome Pack
Desktop.Rproj
Panel App
Trash
Document Viewer
Participant Explorer
Emacs
R
Ensembl
RE Messages
Firefox
Research Environment Documentation
Genomics england
Applications
Places
System
Tue Oct 15, 10:31

7. Import and export of data and tools

The Airlock

Data in the RE

Outside world

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

Our contract with participants

"...although researchers can look at your data and ask questions about it, they can only take away the answers to their questions (their results). They can't copy or take away any of your individual data."

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

Forms in Airlock

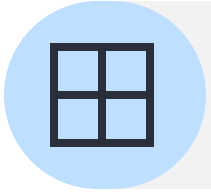
- Export findings
- Export analysis scripts and software
- Contact clinical team and/or report potential diagnosis

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

What should go through the Airlock?

Figures

Statistics or numbers for your text



Tables

Notes on the data

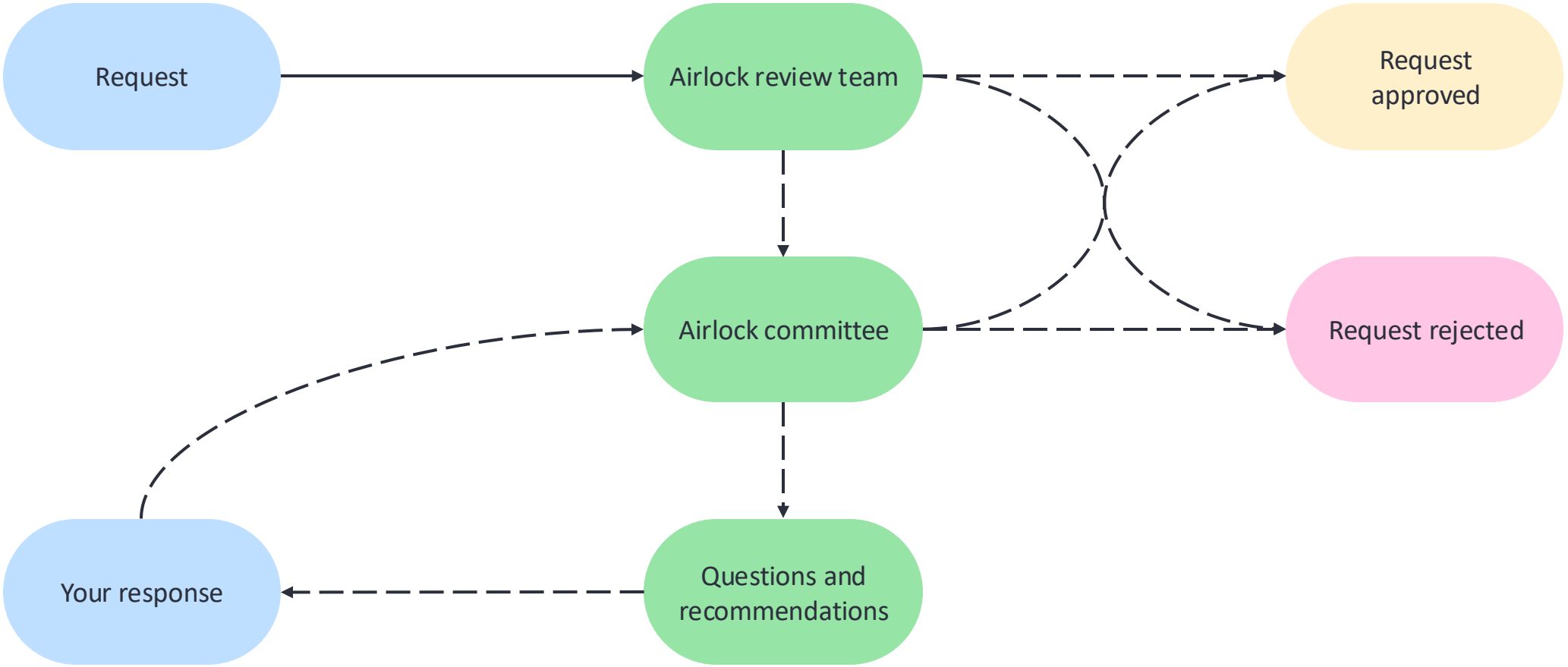
Airlock rules

Approved research project

Participants cannot be identified

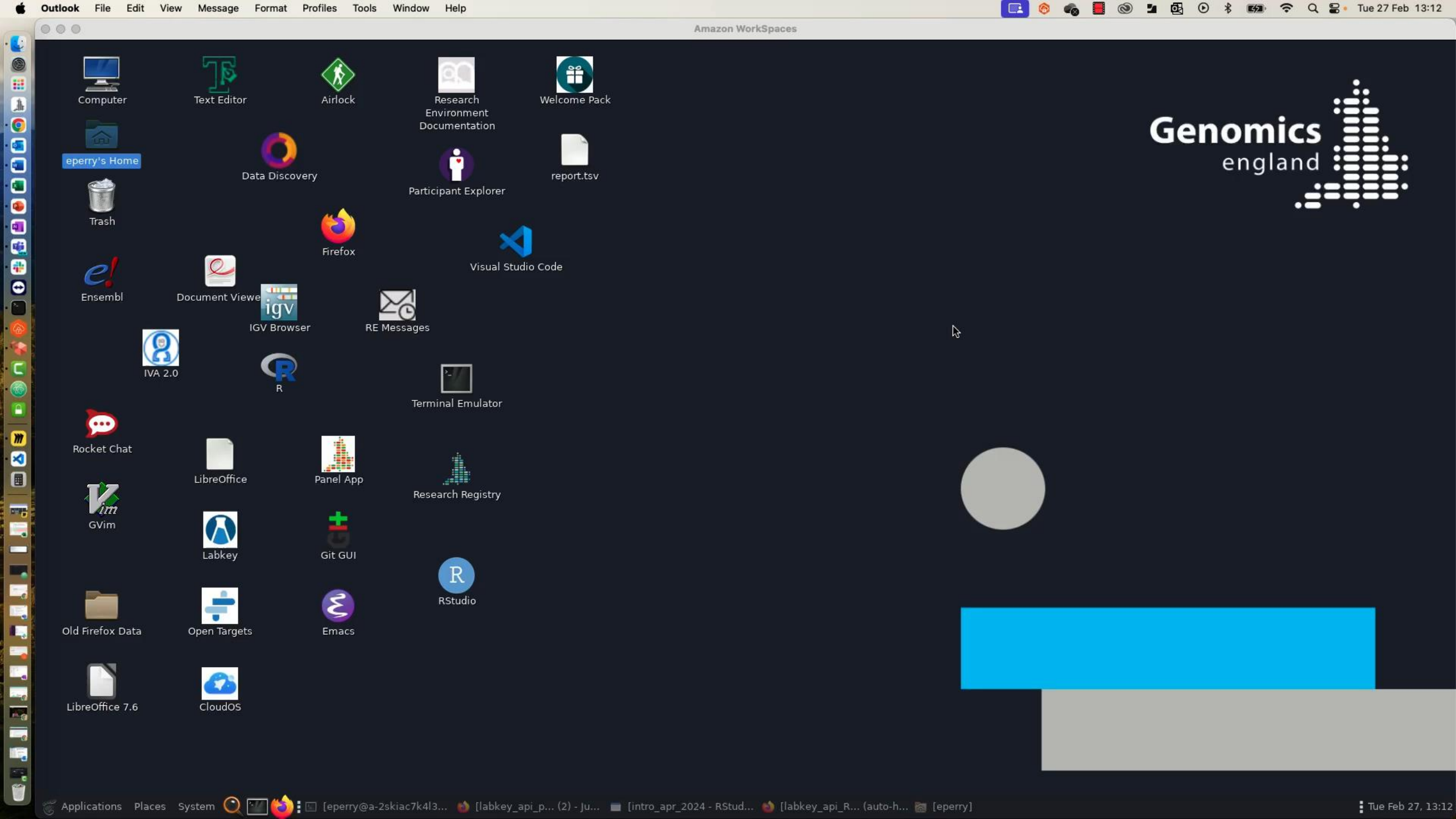
https://re-docs.genomicsengland.co.uk/airlock_rule/

What happens to my request?

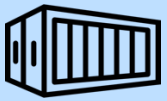


https://re-docs.genomicsengland.co.uk/airlock/

Airlock demo


Outlook File Edit View Message Format Profiles Tools Window Help
Tue 27 Feb 13:12
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
Ensembl
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
CloudOS
Genomics england
Applications Places System
[eperry@a-2skiac7k4l3...
[labkey_api_p... (2) - Ju...
[intro_apr_2024 - RStud...
[labkey_api_R... (auto-h...
[eperry]
Tue Feb 27, 13:12

Import



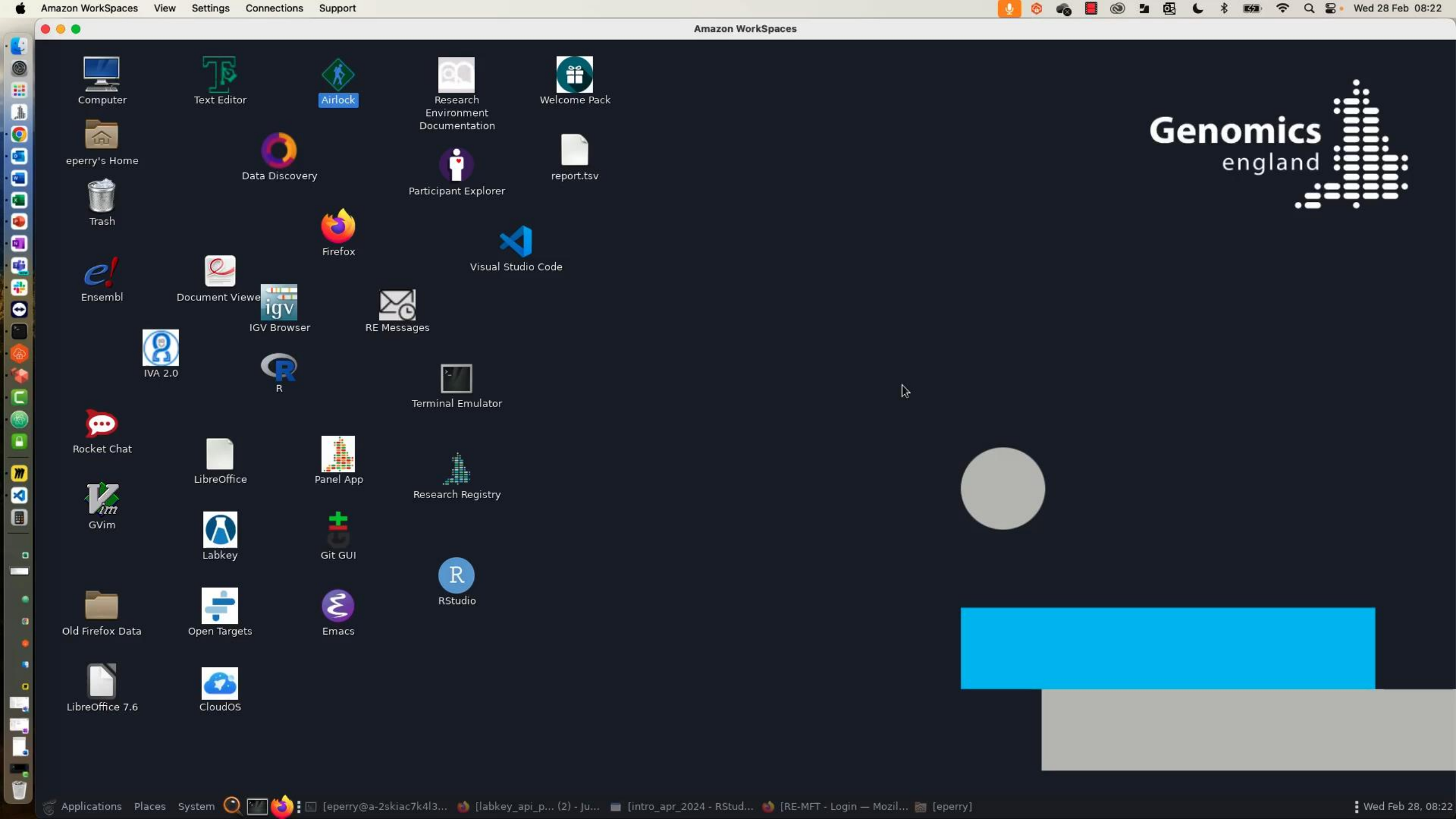
Containers

Copy/paste in

Airlock import

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

Import demo


Amazon WorkSpaces View Settings Connections Support
Wed 28 Feb 08:22
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
Ensembl
Document Viewe
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
CloudOS
Genomics england
Applications Places System
[eperry@a-2skiac7k4l3...
[labkey_api_p... (2) - Ju...
[intro_apr_2024 - RStud...
[RE-MFT - Login — Mozil...
[eperry]
Wed Feb 28, 08:22

8. 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

Thank you

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