Building rare disease cohorts with matching controls

Emily Perry

Research Engagement Manager

10th June 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 are recording and will distribute the censored video later

Questions

All your microphones are muted

Use the Zoom Q&A to ask questions

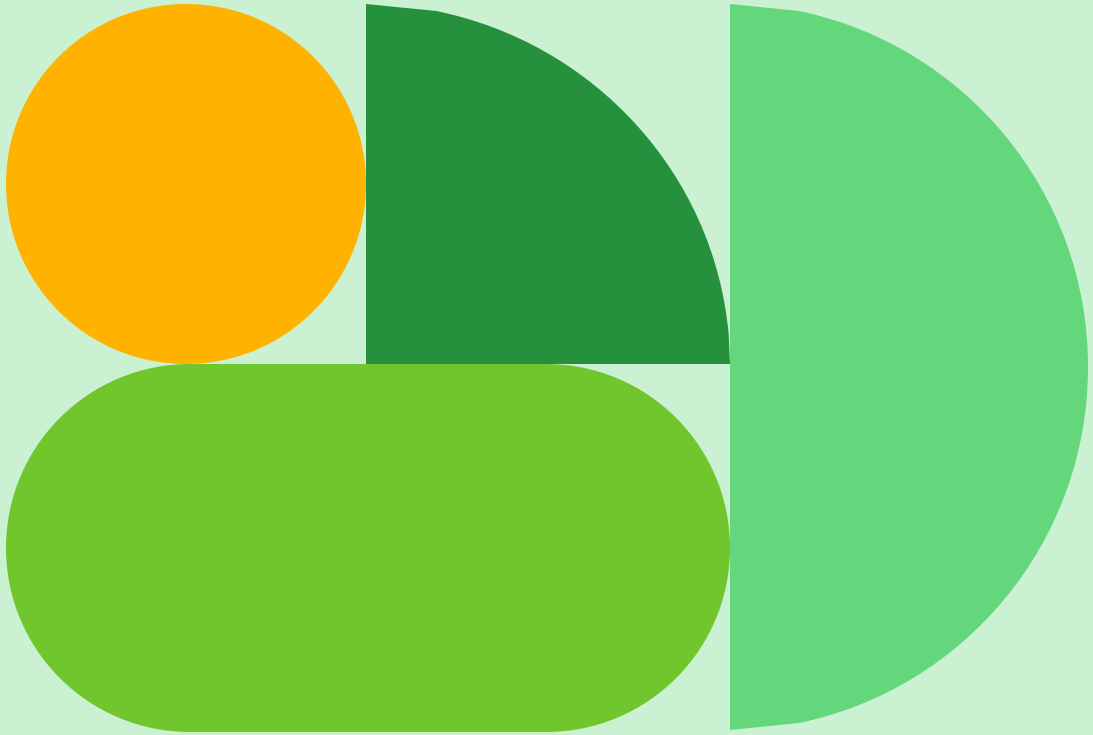
Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Helpers

Matthieu Vizuete-Forster
Bioinformatician - Research Services

Agenda

1. Introduction and admin
2. Parameters and considerations for building a cohort
3. Point-and-click cohort building with Participant Explorer
4. Tables for cohort building in rare disease
5. Programmatic cohort building in Python and R
6. Creating a matched control cohort
7. Getting genomic filepaths for your cohort
8. Using your cohort with aggregate VCFs
9. Help and questions



2. Parameters and considerations for building a cohort

Recruited disease

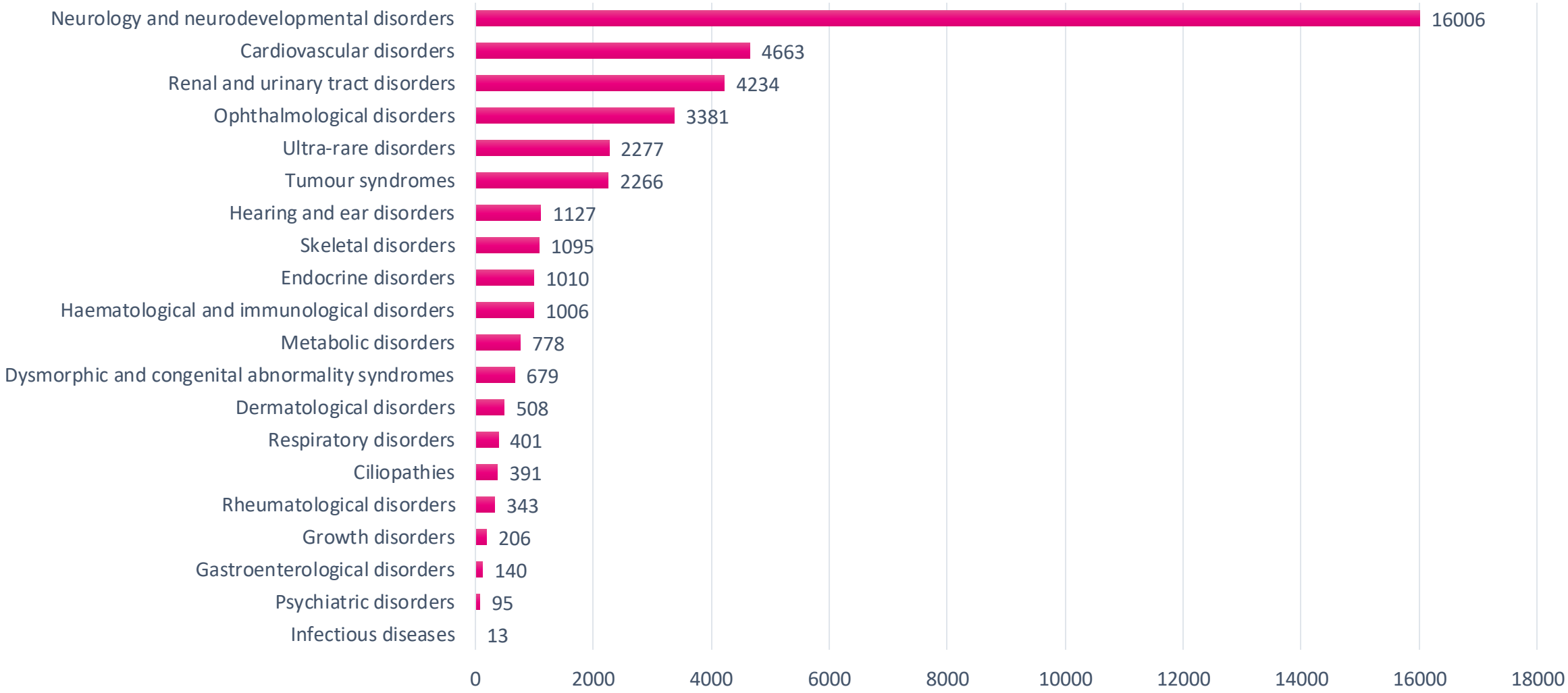
100kGP

Disease group	Disease sub-group	Specific disease
Normalised	Normalised	Normalised

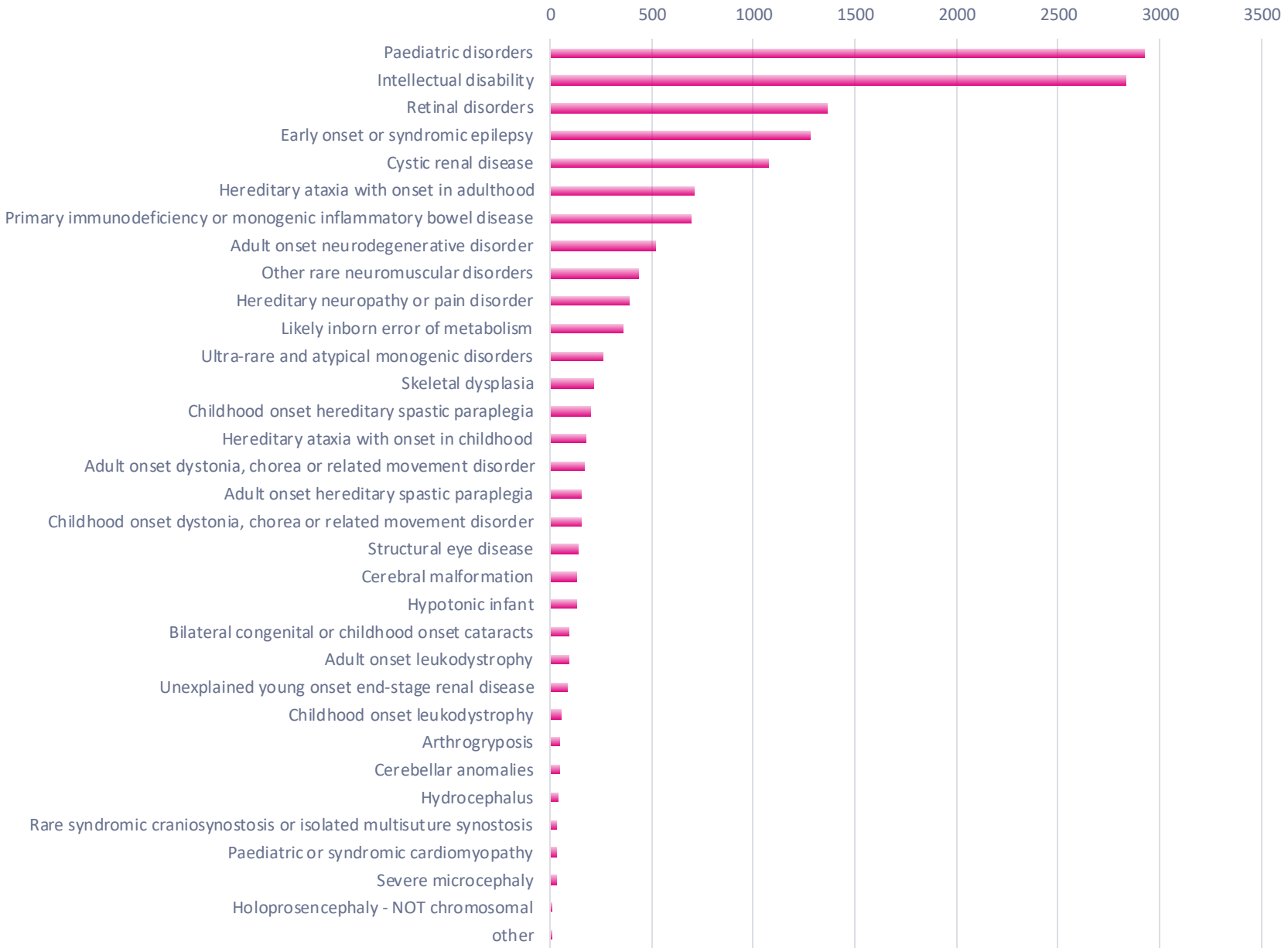
NHS GMS

Clinical indication

100,000 Genomes rare disease



NHS GMS rare disease



Phenotypes

HPO terms assessed on recruitment

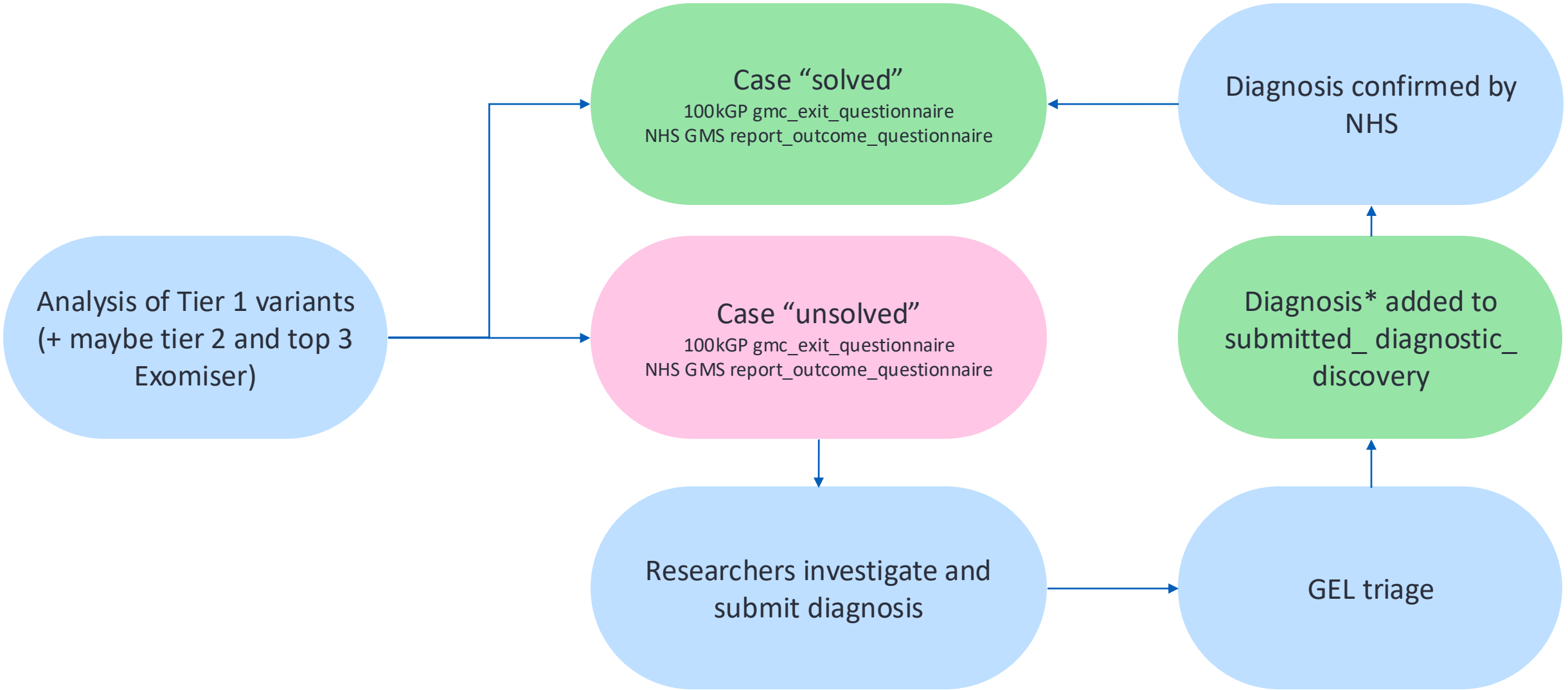
ICD-10 codes in medical records

Rare disease phenotyping – 100kGP

- 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/

Solved cases



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

Solved cases

Use unsolved cases for diagnostic discovery

Use solved cases for clinical trials

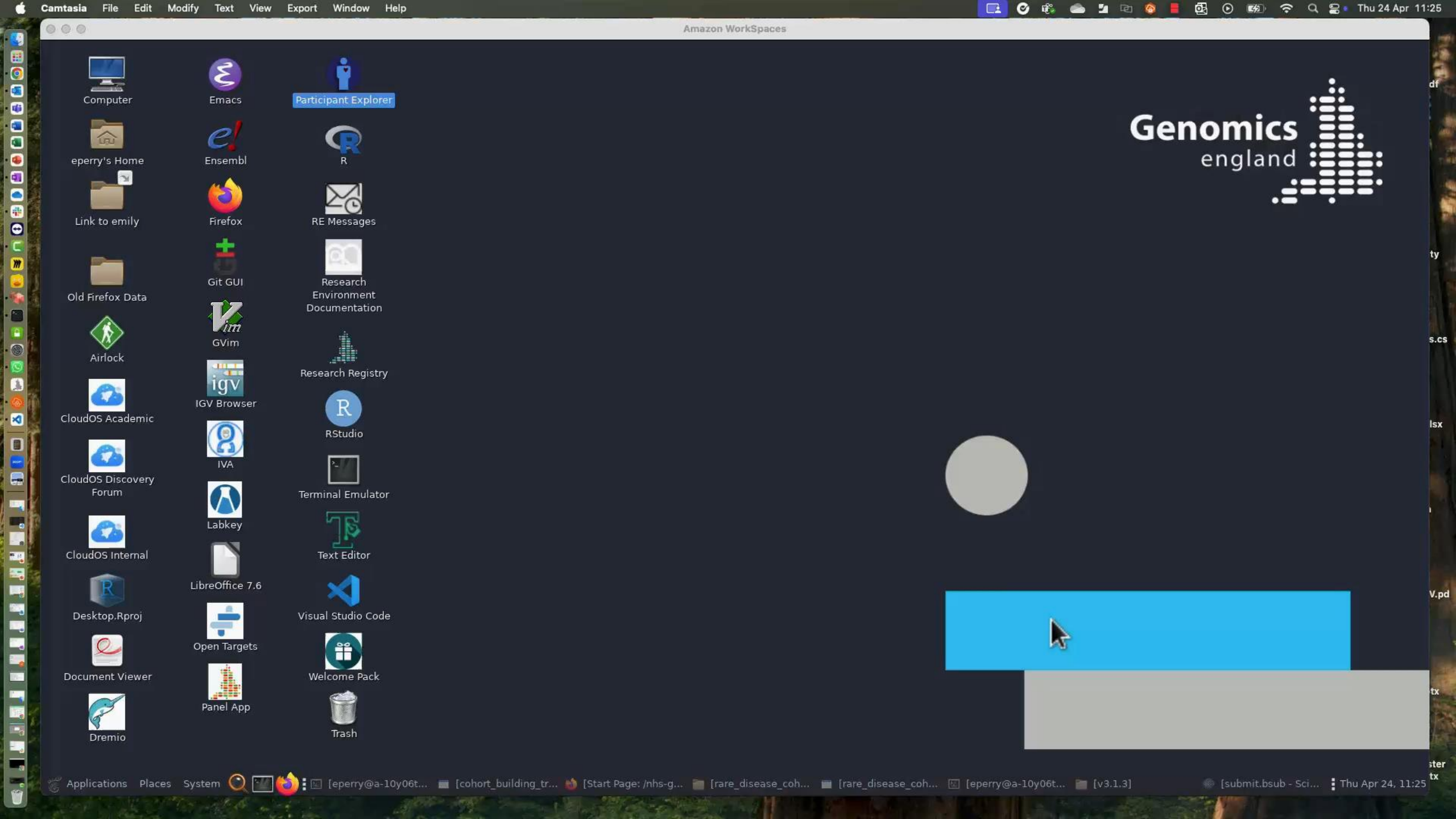
3. Point-and-click cohort building with Participant Explorer

Participant Explorer

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

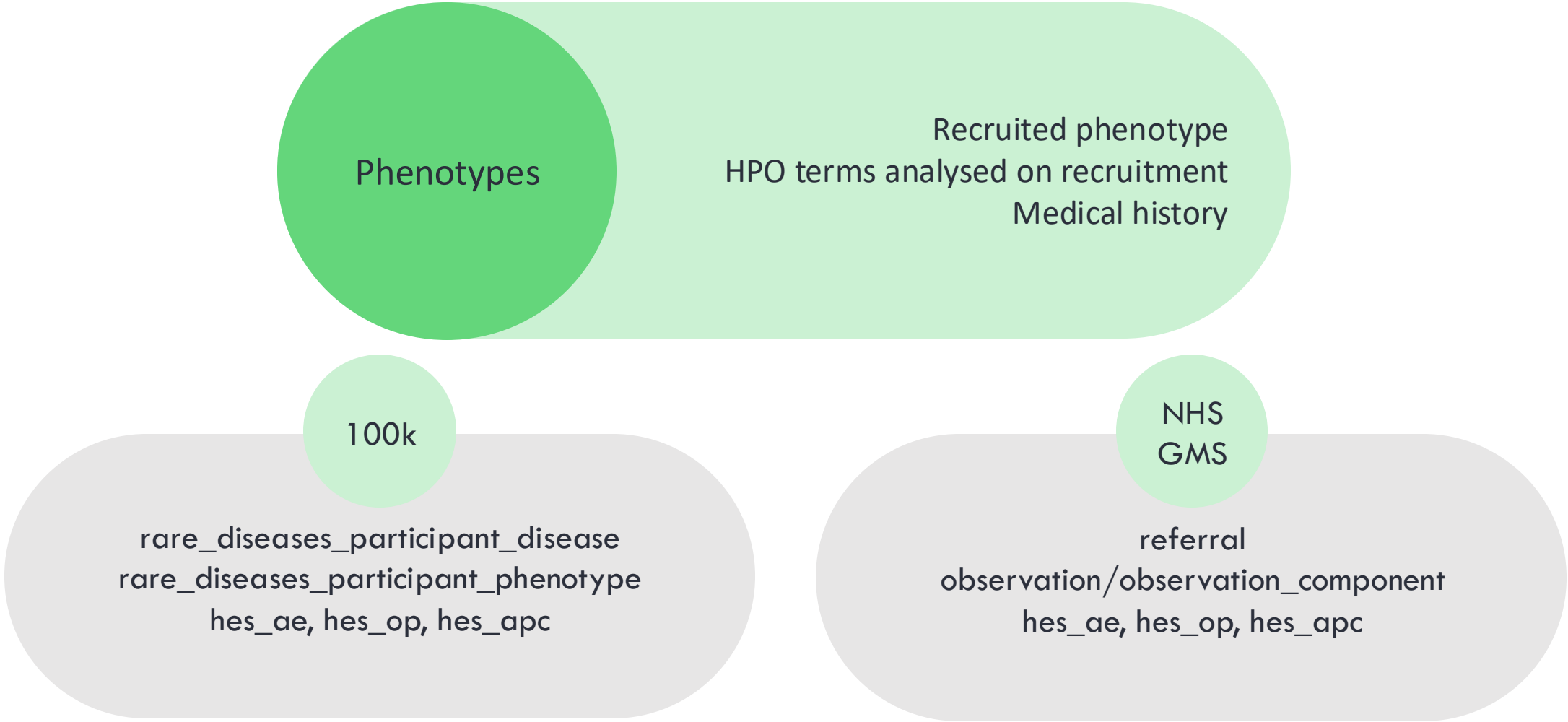
https://re-docs.genomicsengland.co.uk/pxa/

Participant Explorer demo

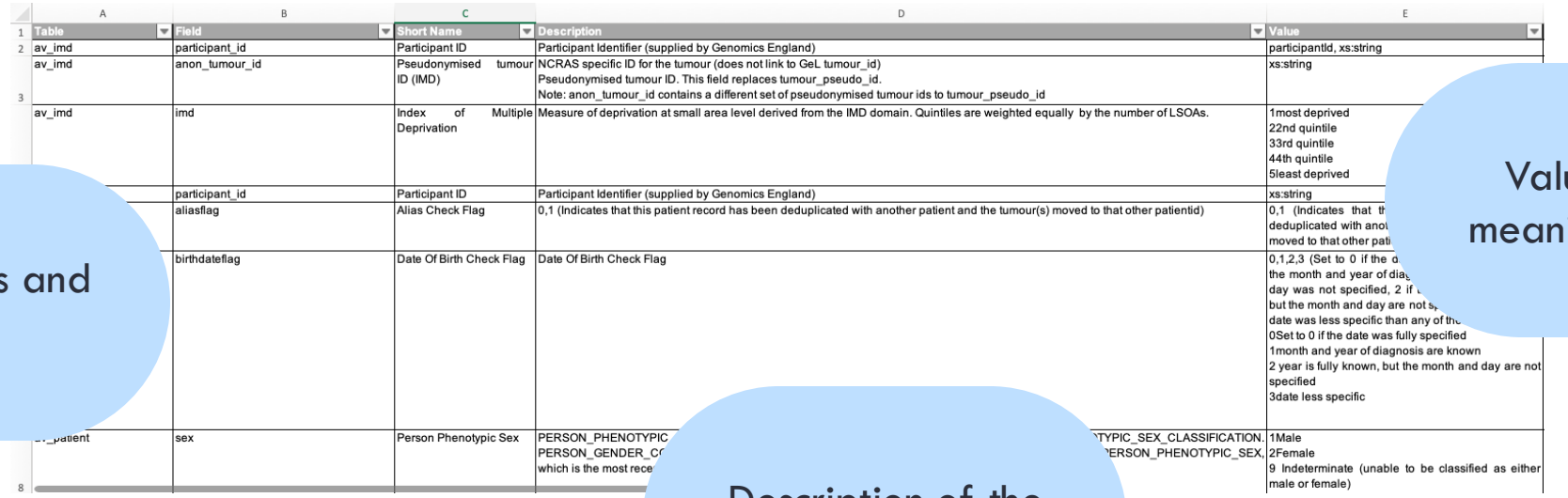

Camtasia File Edit Modify Text View Export Window Help
Thu 24 Apr 11:25
Amazon WorkSpaces
Computer
Emacs
Participant Explorer
eperry's Home
Ensembl
R
Link to emily
Firefox
RE Messages
Git GUI
Old Firefox Data
Research Environment Documentation
GVim
Airlock
Research Registry
IGV Browser
CloudOS Academic
RStudio
IVA
CloudOS Discovery Forum
Terminal Emulator
Labkey
CloudOS Internal
Text Editor
LibreOffice 7.6
Desktop.Rproj
Visual Studio Code
Open Targets
Document Viewer
Welcome Pack
Panel App
Dremio
Trash
Genomics england
Applications Places System
[eperry@a-10y06t...
[cohort_building_tr...
[Start Page: /nhs-g...
[rare_disease_coh...
[rare_disease_coh...
[eperry@a-10y06t...
[v3.1.3]
[submit.bsub - Sci...
Thu Apr 24, 11:25

4. Tables for cohort building in rare disease

Rare disease cohort parameters



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 anot 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 diag day was not specified, 2 if but the month and day are not s date was less specific than any of the 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)

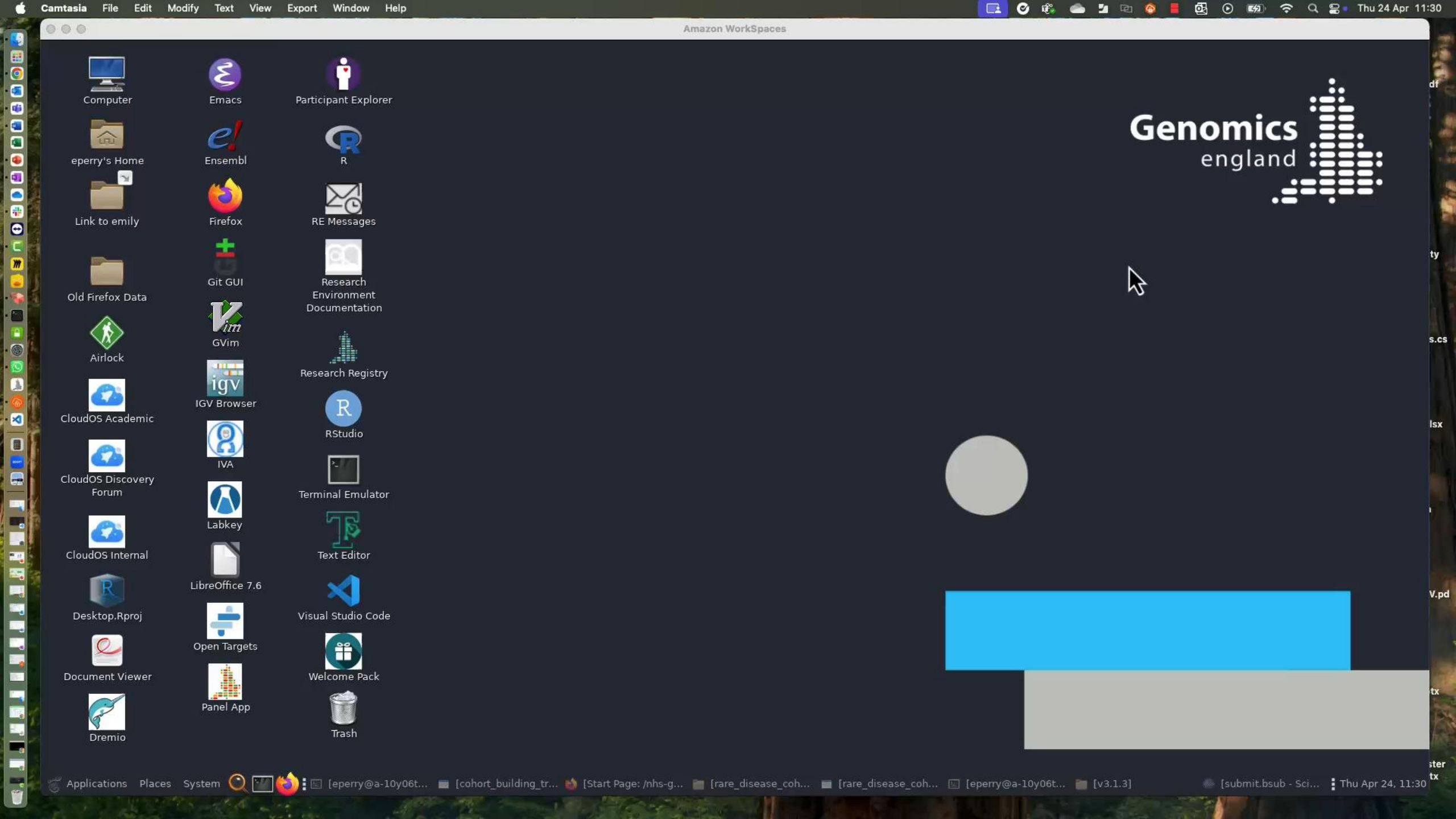
Lists of tables and columns

Value type or meaning of codes

Description of the data

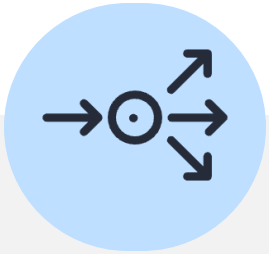
https://re-docs.genomicsengland.co.uk/100k_data_dictionary https://re-docs.genomicsengland.co.uk/GMS_data_dictionary

Tables demo

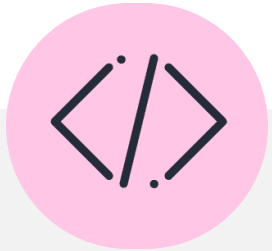

Camtasia File Edit Modify Text View Export Window Help
Thu 24 Apr 11:30
Amazon WorkSpaces
Computer
Emacs
Participant Explorer
eperry's Home
Ensembl
R
Link to emily
Firefox
RE Messages
Git GUI
Old Firefox Data
Research Environment Documentation
GVim
Airlock
Research Registry
IGV Browser
CloudOS Academic
RStudio
IVA
CloudOS Discovery Forum
Terminal Emulator
Labkey
CloudOS Internal
Text Editor
LibreOffice 7.6
Desktop.Rproj
Visual Studio Code
Open Targets
Document Viewer
Welcome Pack
Panel App
Dremio
Trash
Genomics england
Applications Places System
Thu Apr 24, 11:30

5. Programmatic cohort building in Python and R

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/

Demo notebooks

/gel_data_resources/example_scripts/workshop_scripts/rare_disease_cohorts_2025

Programming demo

Building rare disease cohorts with matching controls in Python

This notebook will walk you through building rare disease cohorts using the LabKey API in Python. You are welcome to copy/paste any code from this notebook for your own scripts/notebooks.

Contents:

Import Python modules you need

In [138]:

```
1 import numpy as np
2 import functools
3 import labkey
4 import pandas as pd
```


Code

Building rare disease cohorts with matching controls in R

Building rare disease cohorts with matching controls in R

This notebook will walk you through building rare disease cohorts using the LabKey API in R. You are welcome to copy/paste any code from this notebook for your own scripts/notebooks.

Import R libraries you need

Hide

```
library(tidyverse)
```

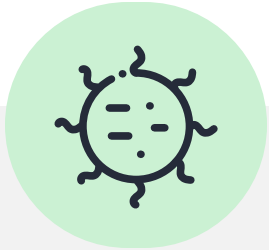
6. Creating a matched control cohort

NOT phenotype

NOT recruited disease or related disease

NOT HPO terms or related HPO terms

NOT ICD10 codes or related ICD10 codes



NOT cancer in related tissues

General inclusion criteria

karyotypic = phenotypic sex
No aneuploidy

(for 100k)
aligned to GRCh38
in AggV2

assess kinship

blood
EDTA extraction
PCR-free

include MZ twins?

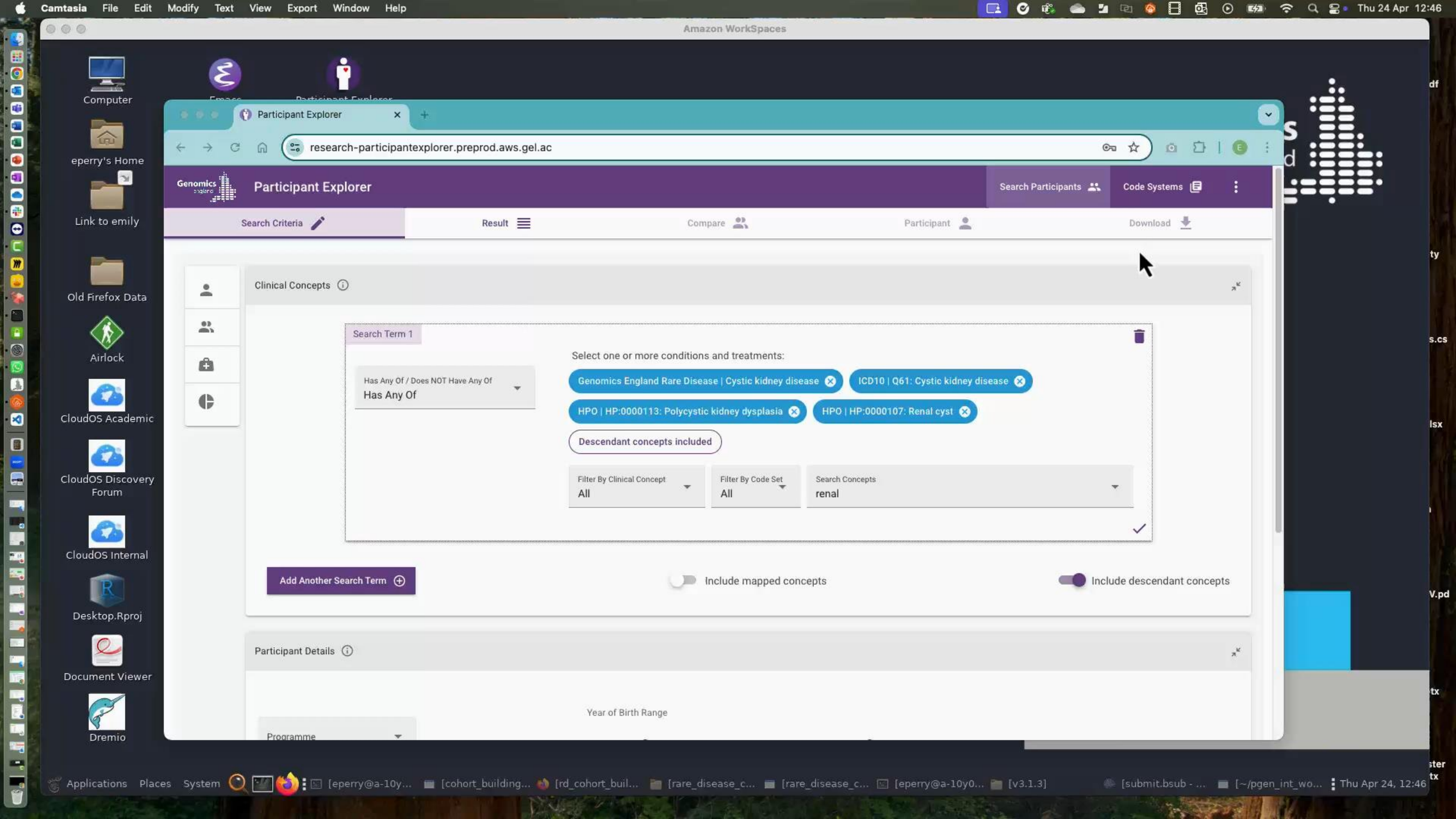
Additional QC

Match case/control

work with a single ethnicity

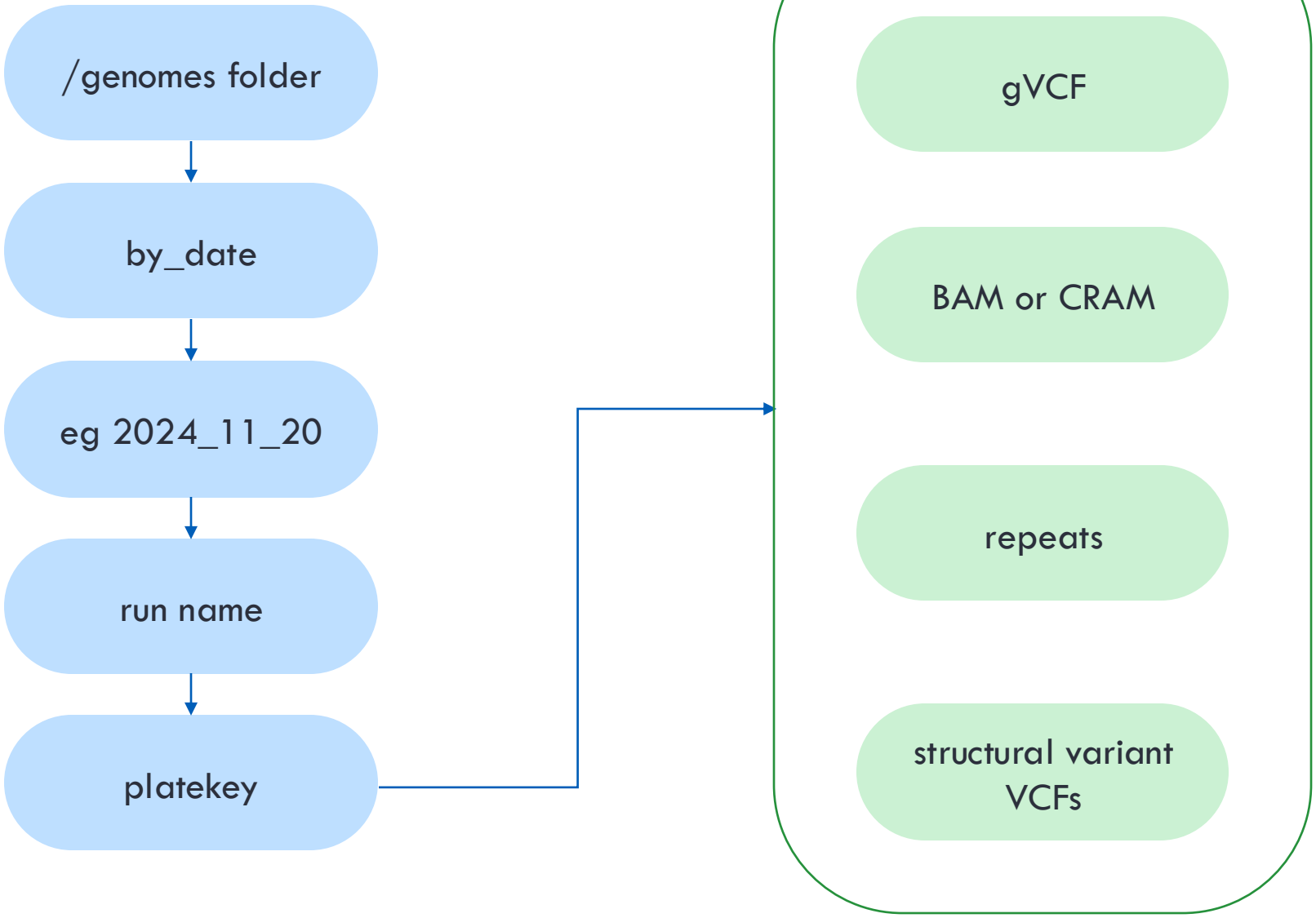
match sex ratios and age distribution

Control demo


Camtasia File Edit Modify Text View Export Window Help
Thu 24 Apr 12:46
Amazon WorkSpaces
Computer
eperry's Home
Link to emily
Old Firefox Data
Airlock
CloudOS Academic
CloudOS Discovery Forum
CloudOS Internal
Desktop.Rproj
Document Viewer
Dremio
Participant Explorer
research-participantexplorer.preprod.aws.gel.ac
Genomics England
Participant Explorer
Search Participants
Code Systems
Search Criteria
Result
Compare
Participant
Download
Clinical Concepts
Search Term 1
Select one or more conditions and treatments:
Has Any Of / Does NOT Have Any Of
Has Any Of
Genomics England Rare Disease | Cystic kidney disease
ICD10 | Q61: Cystic kidney disease
HPO | HP:0000113: Polycystic kidney dysplasia
HPO | HP:0000107: Renal cyst
Descendant concepts included
Filter By Clinical Concept
All
Filter By Code Set
All
Search Concepts
renal
Add Another Search Term
Include mapped concepts
Include descendant concepts
Participant Details
Year of Birth Range
Programme
Applications Places System
[eperry@a-10y...
[cohort_building...
[rd_cohort_buil...
[rare_disease_c...
[rare_disease_c...
[eperry@a-10y0...
[v3.1.3]
[submit.bsub - ...
[~/pgen_int_wo...
Thu Apr 24, 12:46

6. Getting genomic filepaths for your cohort

Genomes location

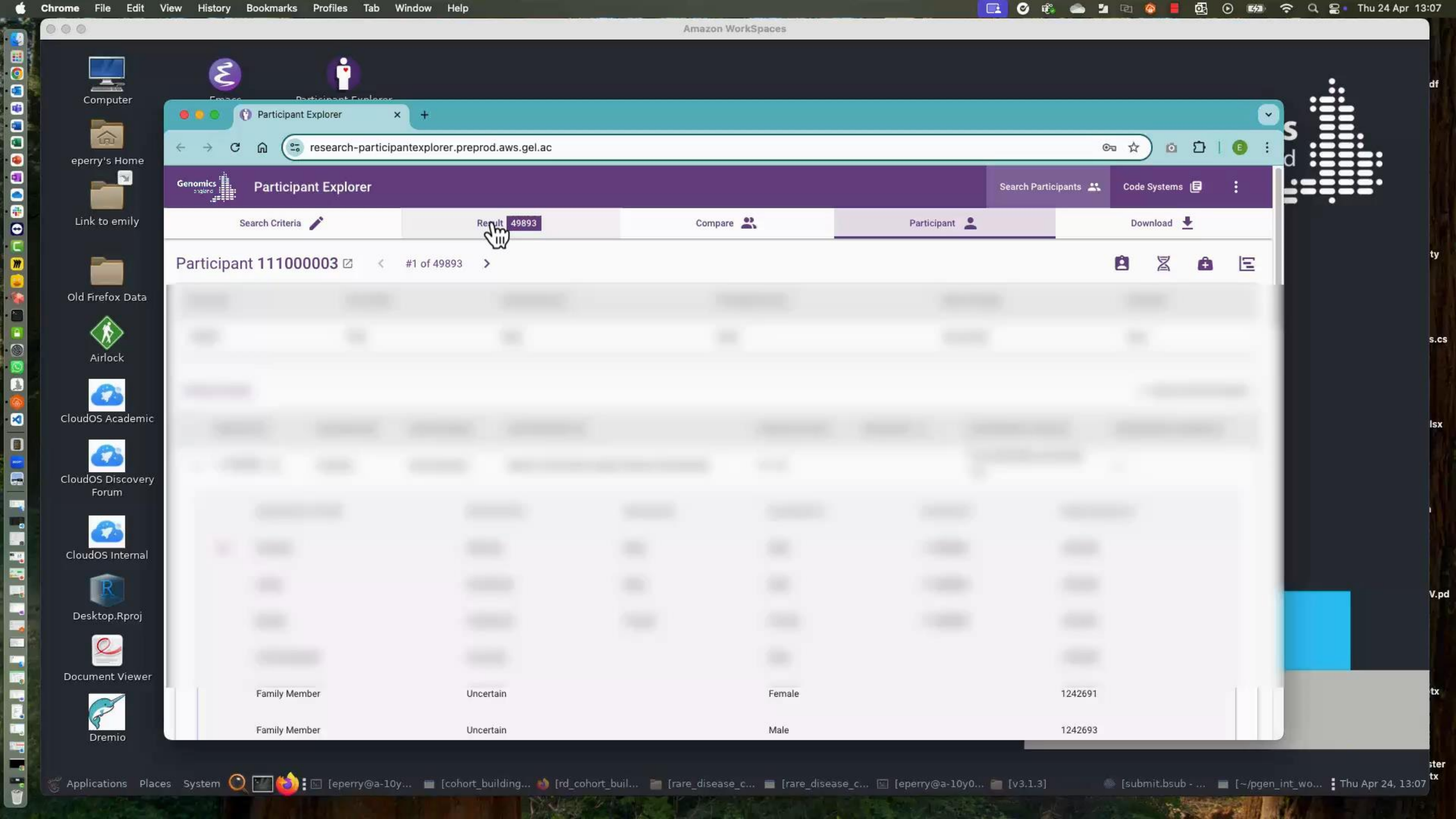


https://re-docs.genomicsengland.co.uk/genomic_data/

File locations

Participant ID	Platekey	Filepath	Filename	File sub-type
12345678	LP12345-DNA_A01	/genomes/by_date/2025-04-08/000111222333/LP12345-DNA_A01/Variations/LP12345-DNA_A01.vcf.gz	LP12345-DNA_A01.vcf.gz	Genomic VCF
12345678	LP12345-DNA_A01	/genomes/by_date/2025-04-08/000111222333/LP12345-DNA_A01/Assembly/LP12345-DNA_A01.cram	LP12345-DNA_A01.cram	CRAM
12345678	LP12345-DNA_A01	/genomes/by_date/2025-04-08/000111222333/LP12345-DNA_A01/Variations/LP12345-DNA_A01.SV.vcf.gz	LP12345-DNA_A01.SV.vcf.gz	Structural VCF

Filepaths demo


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Thu 24 Apr 13:07
Amazon WorkSpaces
Computer
eperry's Home
Link to emily
Old Firefox Data
Airlock
CloudOS Academic
CloudOS Discovery Forum
CloudOS Internal
Desktop.Rproj
Document Viewer
Dremio
Participant Explorer
research-participantexplorer.preprod.aws.gel.ac
Genomics England
Participant Explorer
Search Participants
Code Systems
Search Criteria
Result 49893
Compare
Participant
Download
Participant 111000003
#1 of 49893
Family Member Uncertain Female 1242691
Family Member Uncertain Male 1242693
Applications Places System
Thu Apr 24, 13:07

7. Using your cohort with aggregate VCFs

Aggregate VCFs

Combined gVCFs

Quality metrics

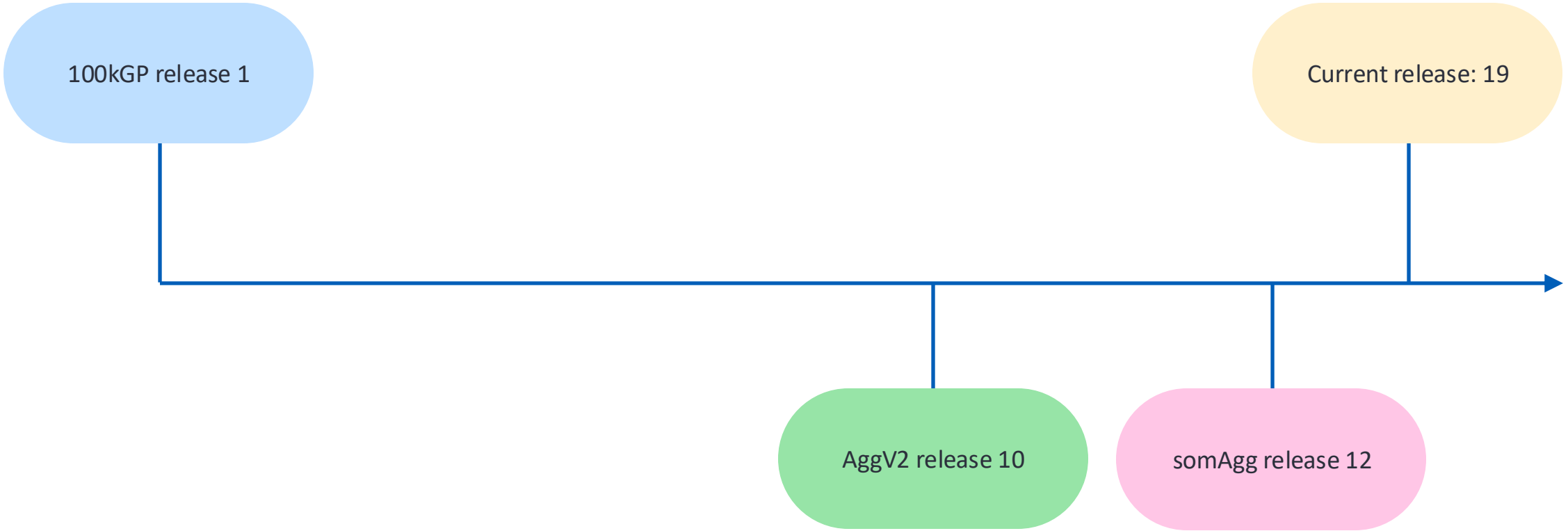
Functional annotation files

Genotype files – platekeys

1371 chunks

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

Aggregate VCFs



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

Aggregate VCFs

Using bedtools to identify relevant chunks of the aggregate files

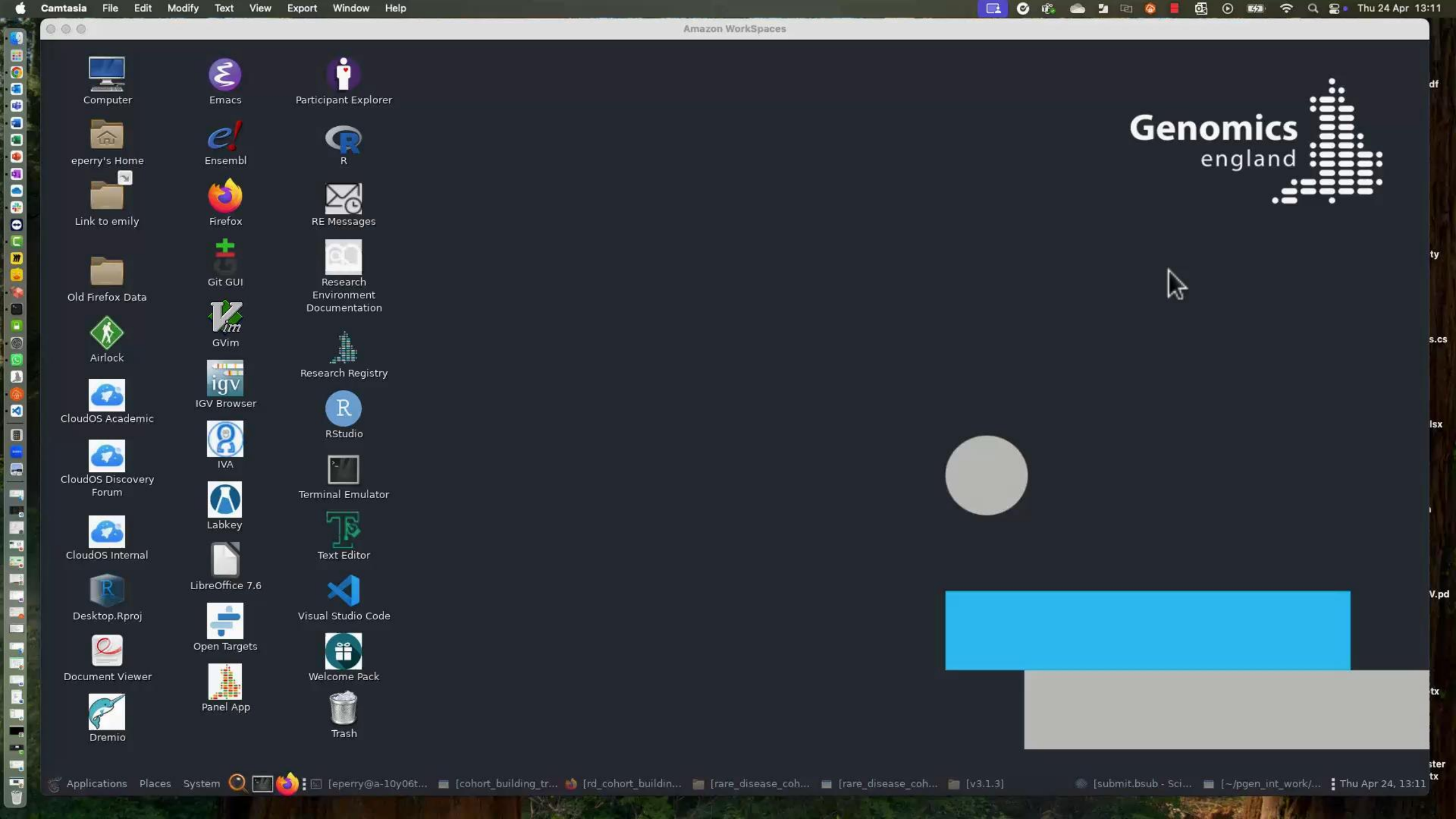
Excluding participants who have since withdrawn consent

Using bcftools to extract data from the aggregate VCFs

Codebooks

https://re-docs.genomicsengland.co.uk/aggv2_code_book/ https://re-docs.genomicsengland.co.uk/somAgg_code_book/

Aggregate demo


Camtasia File Edit Modify Text View Export Window Help
Thu 24 Apr 13:11
Amazon WorkSpaces
Computer
eperry's Home
Link to emily
Old Firefox Data
Airlock
CloudOS Academic
CloudOS Discovery Forum
CloudOS Internal
Desktop.Rproj
Document Viewer
Dremio
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
[eperry@a-10y06t...
[cohort_building_tr...
[rd_cohort_buildin...
[rare_disease_coh...
[rare_disease_coh...
[v3.1.3]
[submit.bsub - Sci...
[~/pgen_int_work/...
Thu Apr 24, 13:11

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

In-person training

30/6 For non-coders

1/7 For coders

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

Training sessions

3rd Tuesday every month

Introduction to the RE

22/7 19/8 16/9

21/10 18/11 16/12

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

Materials from past training all online

Training sessions

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
9/12	Running workflows on the HPC and Cloud

Materials from past training all online

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

Feedback

Thank you

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