Genomics England

Building a cohort based on phenotypes

Emily Perry

Research Engagement Manager

24th May 2022

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

https://research-help.genomicsengland.co.uk/display/GERE/9.+Data+Security+and+You

Questions

Your microphones are all 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

Matthieu Vizuete-Forster
Bioinformatician - Research Services

Christian Bouwens
Bioinformatician - Research Services

Ronnie Rodrigues Pereira
Bioinformatician - Research Services

Ana Lisa Taylor Tavares
Senior Clinical Fellow in Rare Disease Genomics - Research Engagement

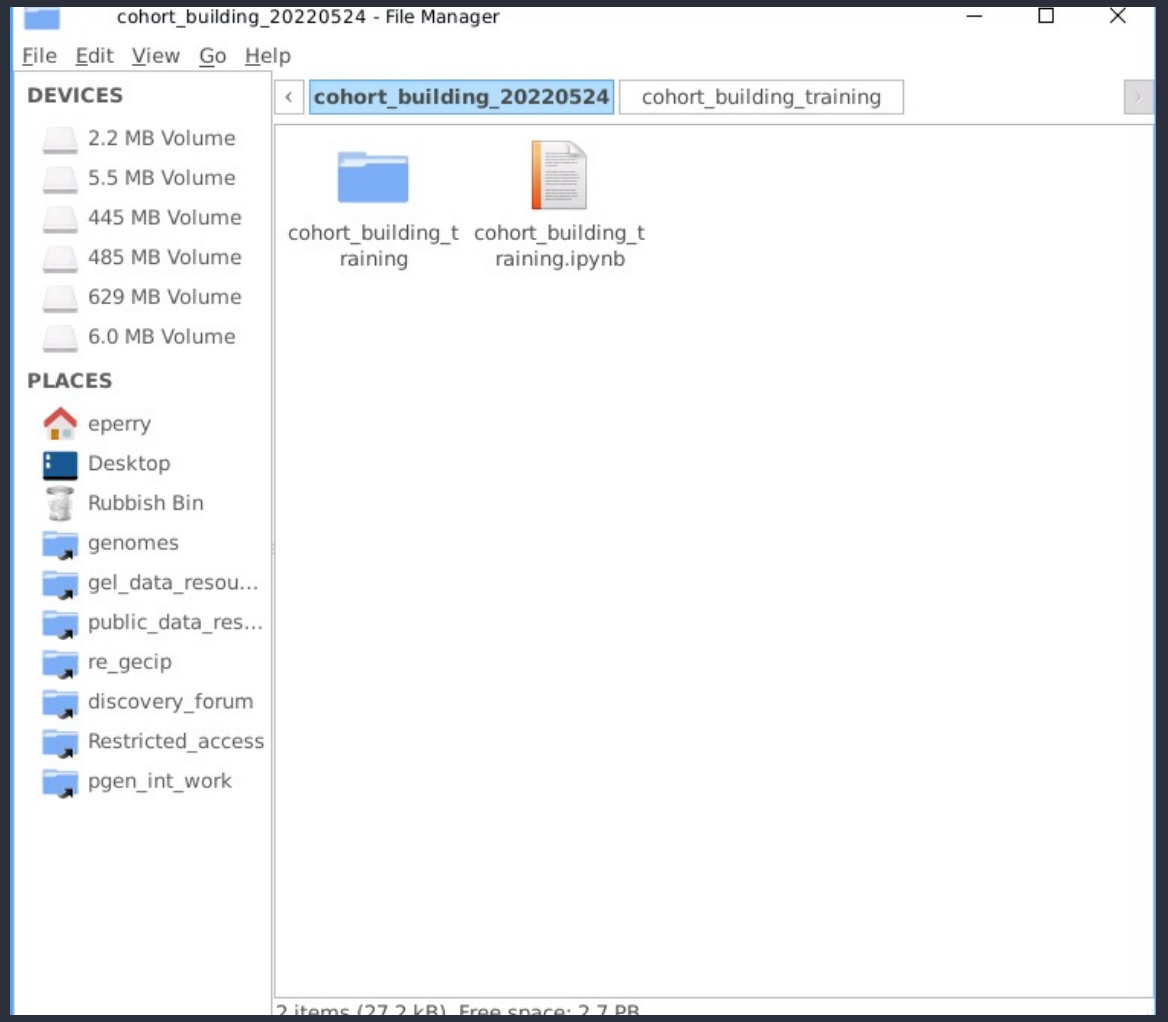
Roel Bevers
Senior Bioinformatician - Research Services

Agenda

1. Introduction and admin
2. Parameters and considerations for building a cohort
3. Point-and-click cohort building with Participant Explorer
4. Labkey tables for cohort building
5. Covariates in cohort building
6. Using the Labkey API in Python and R to build cohorts
7. Fetching data for downstream analysis
8. Getting help and questions

Materials

- Slides and video will be sent out to you after the session
- Scripts available in /gel_data_resources/example_scripts/workshop_scripts/cohort_building_20220524



2. Parameters and considerations for building a cohort

Phenotype data to filter by

Cancer

- Disease type
  - Recruited disease
  - Diagnosis codes in health records
- Staging
- Treatment

Rare disease

- Recruited disease
- HPO terms
- Solved? Alive?

Common disease

- Rare disease relatives and cancer participants
- Diagnosis codes in health records

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

Participant Explorer

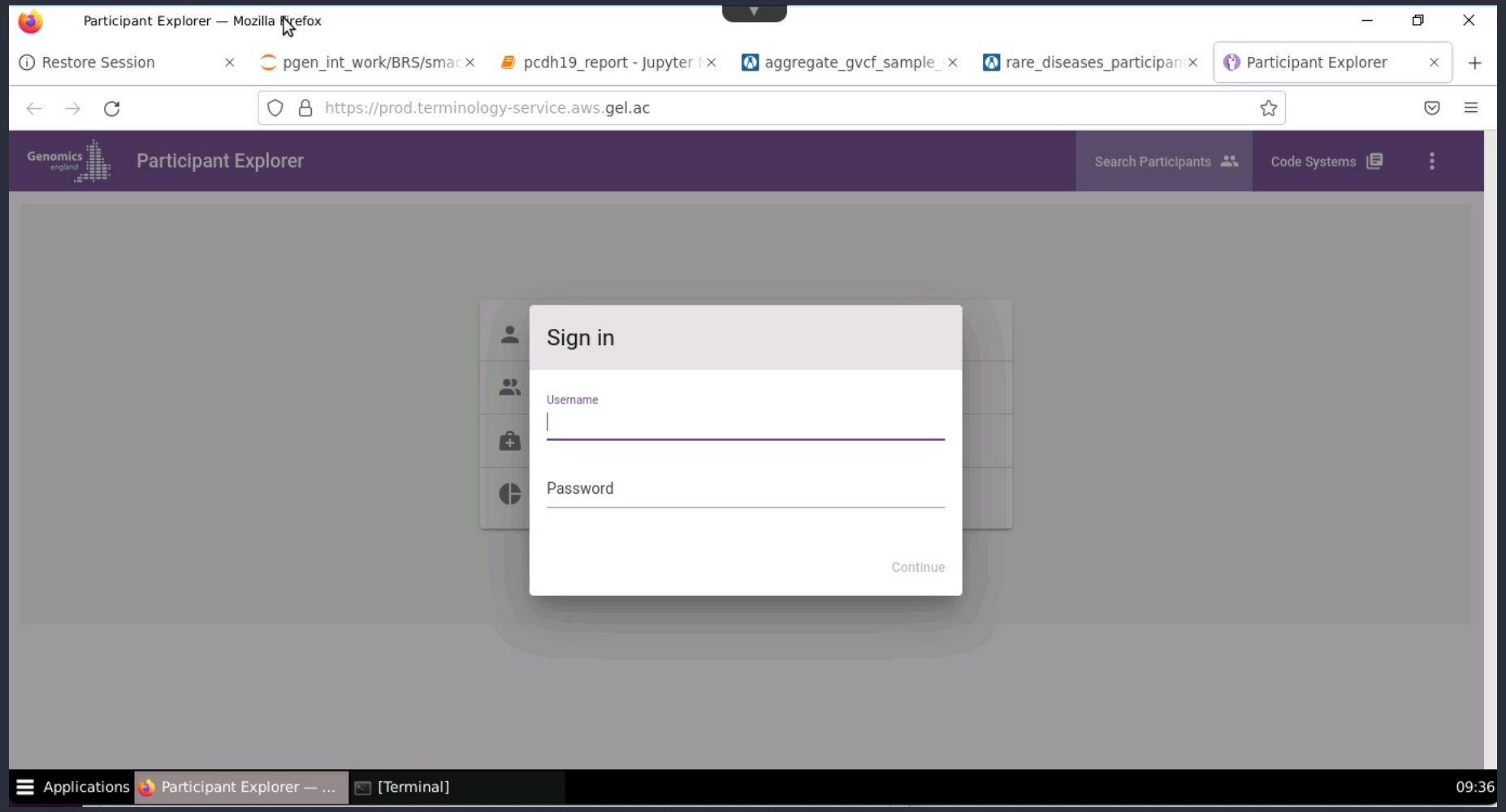
- Point-and-click tool
- Ontology-aware
- Filter by phenotypes
- Combine multiple filters

https://research-help.genomicsengland.co.uk/display/GERE/Participant+Explorer

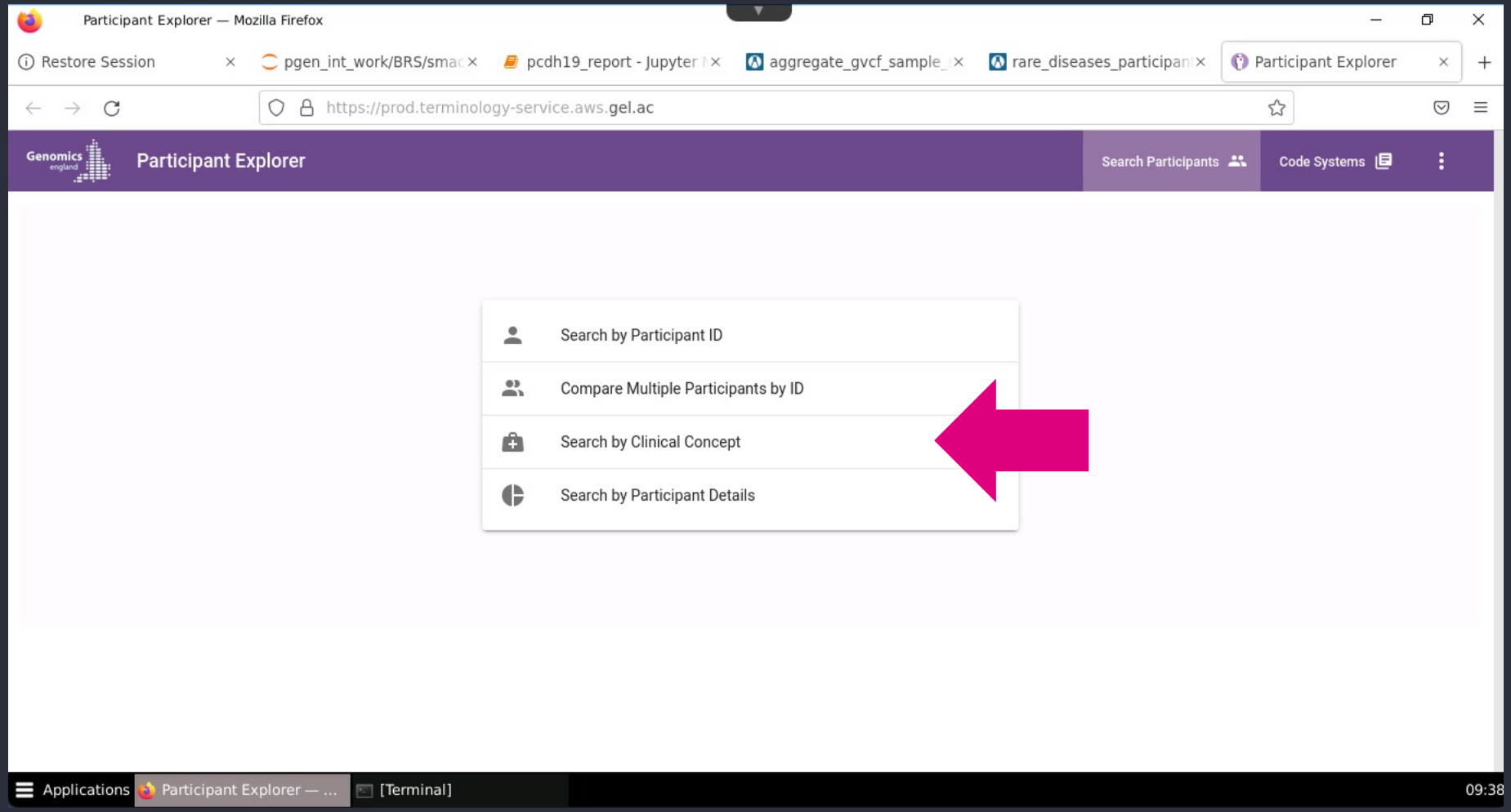
Demo: Participant Explorer

https://research-help.genomicsengland.co.uk/display/GERE/Participant+Explorer

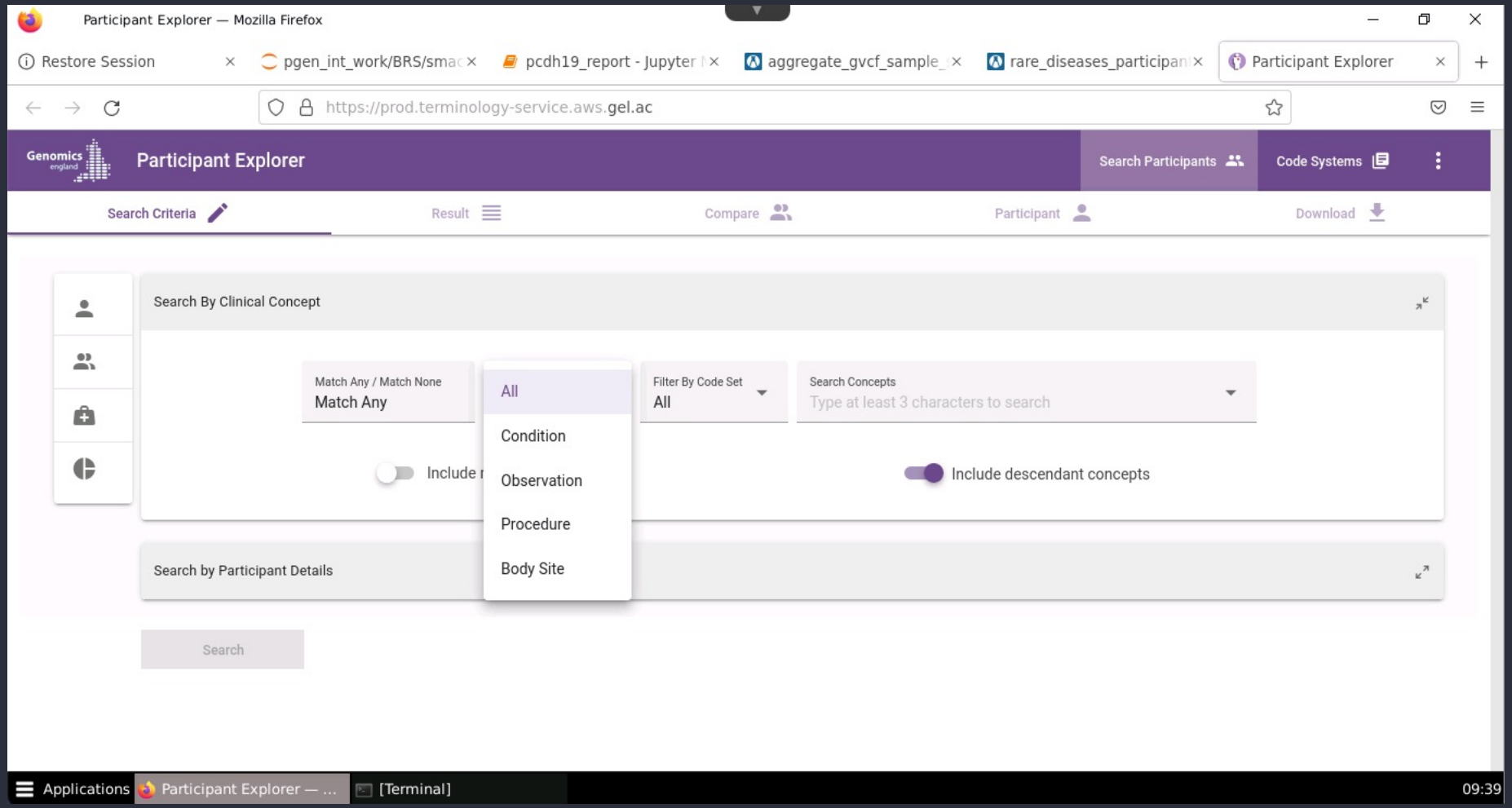
Launch Participant Explorer


Participant Explorer — Mozilla Firefox
Restore Session
pgen_int_work/BRS/sma
pcdh19_report - Jupyter
aggregate_gvcf_sample_
rare_diseases_participan
Participant Explorer
https://prod.terminology-service.aws.gel.ac
Genomics england
Participant Explorer
Search Participants
Code Systems
Sign in
Username
Password
Continue
Applications
Participant Explorer — ...
[Terminal]
09:36

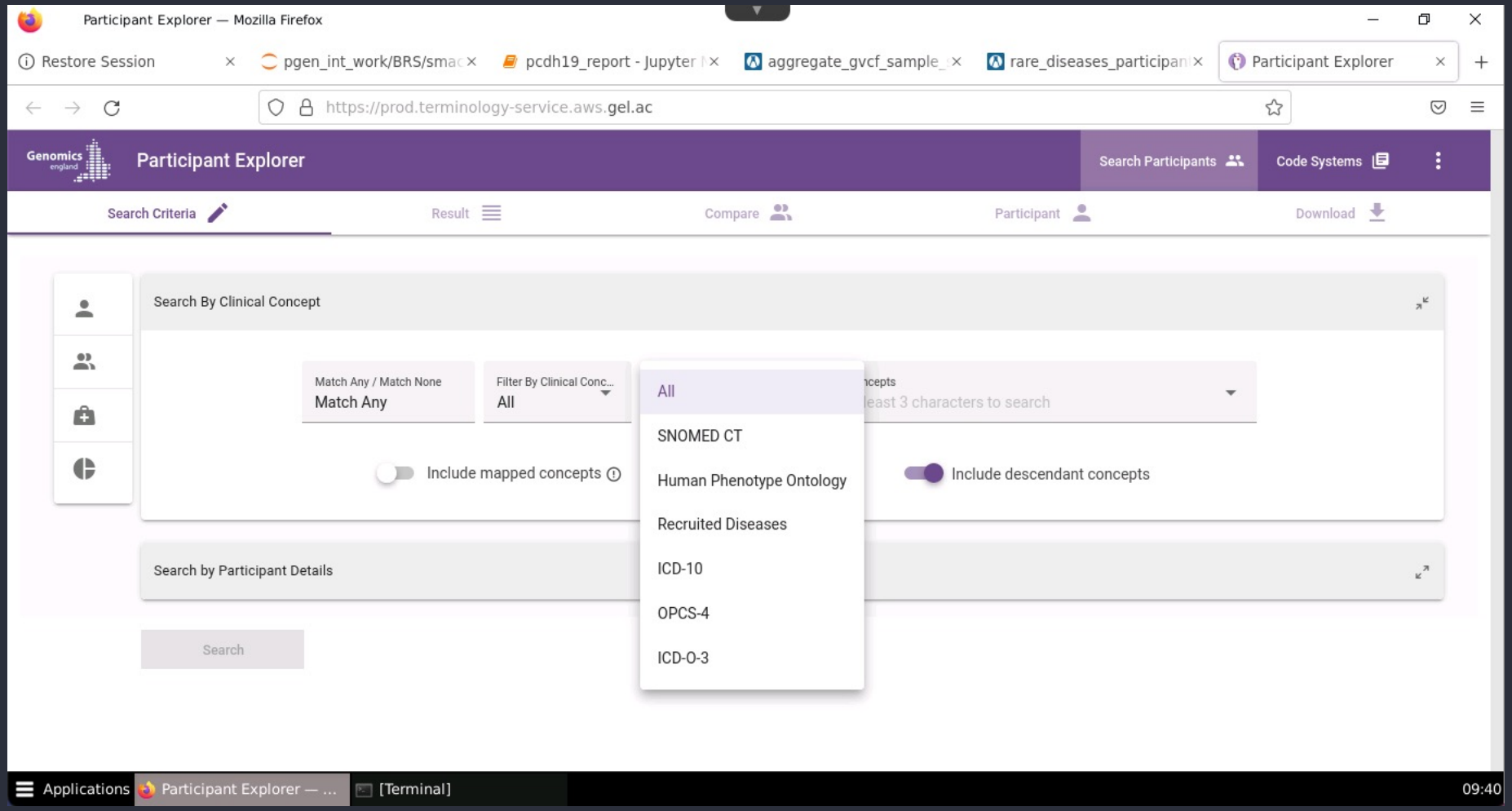
Sign in



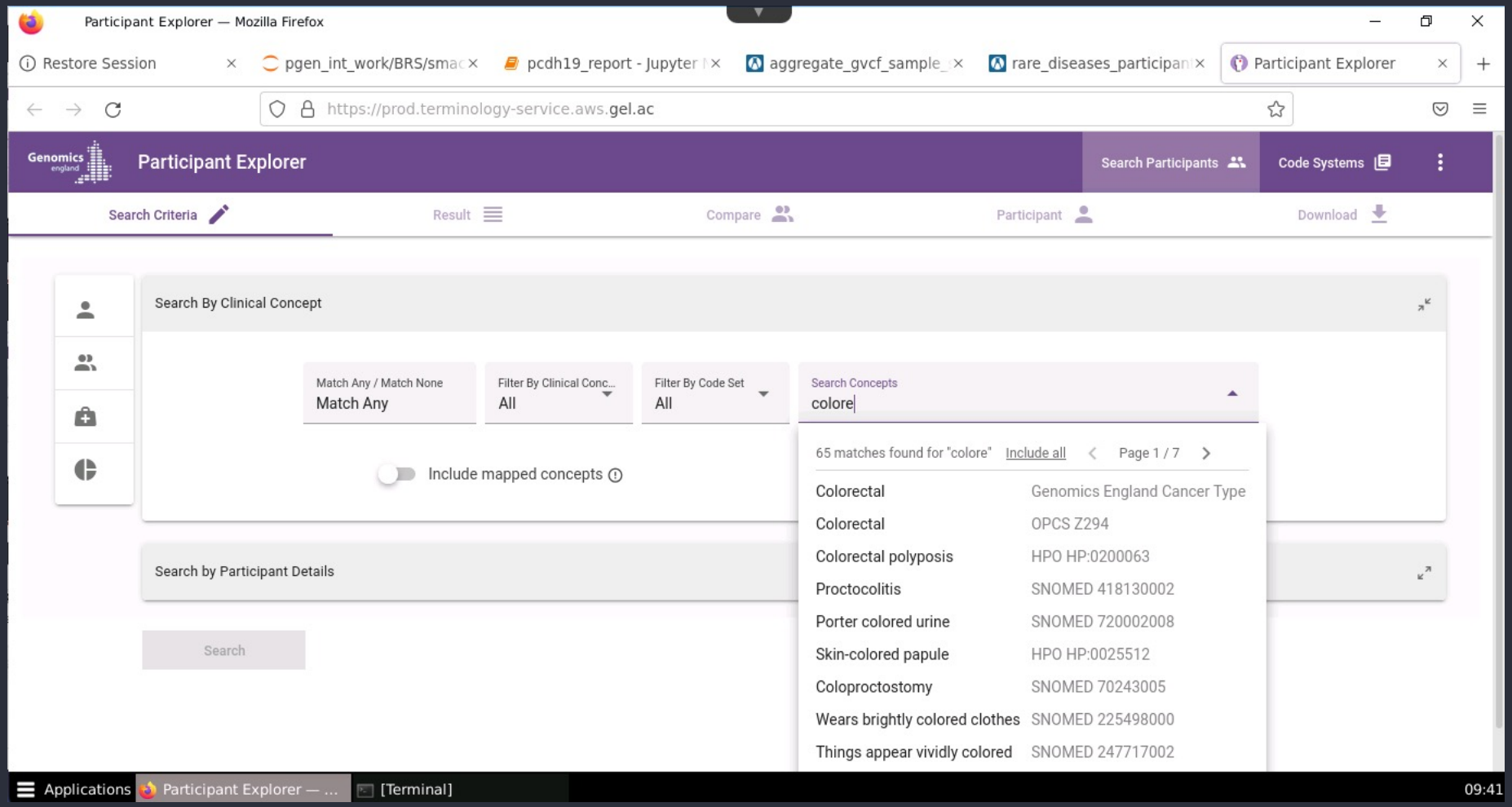
Search by Clinical Concept



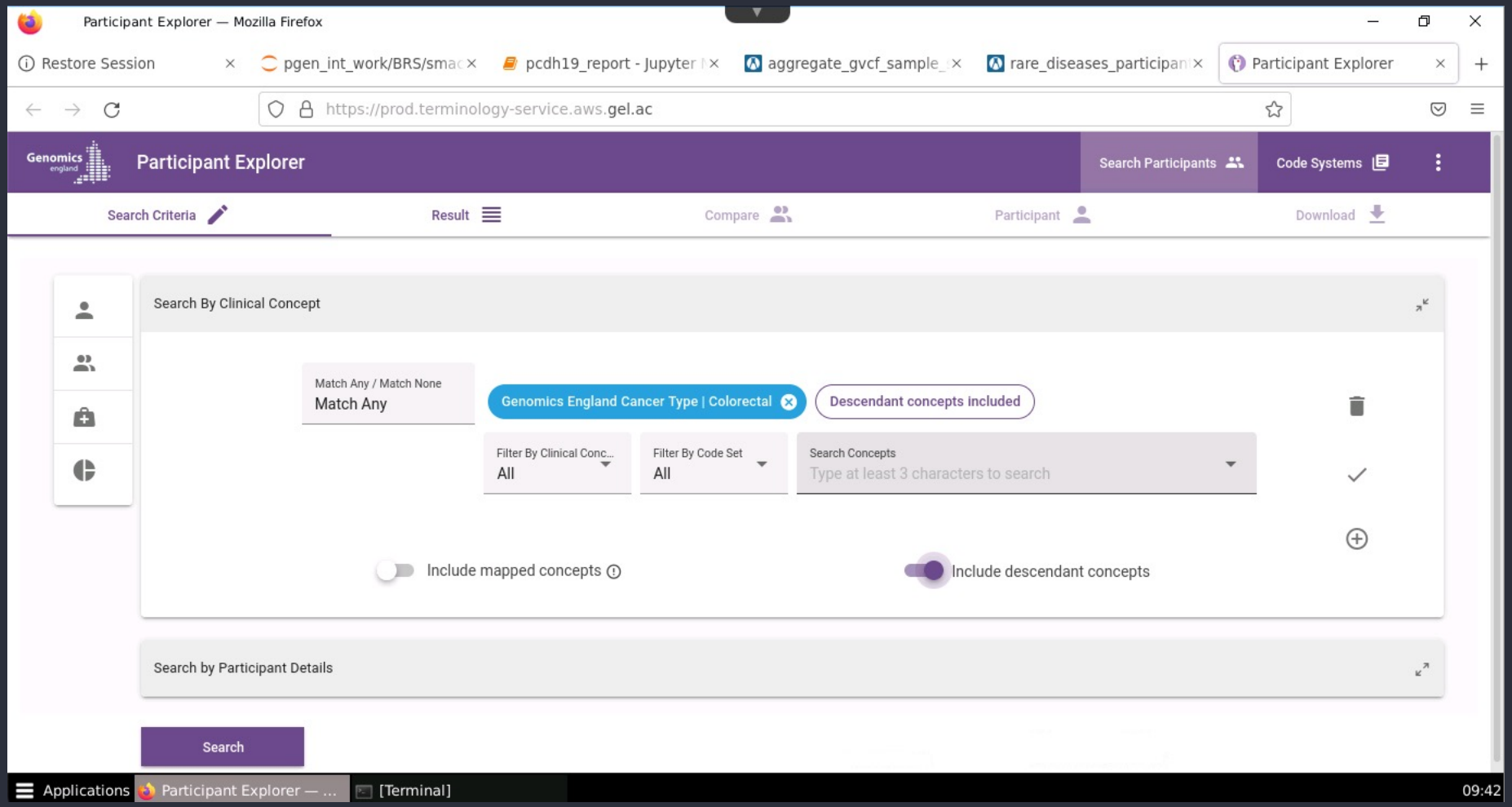
Choose what type of clinical concept



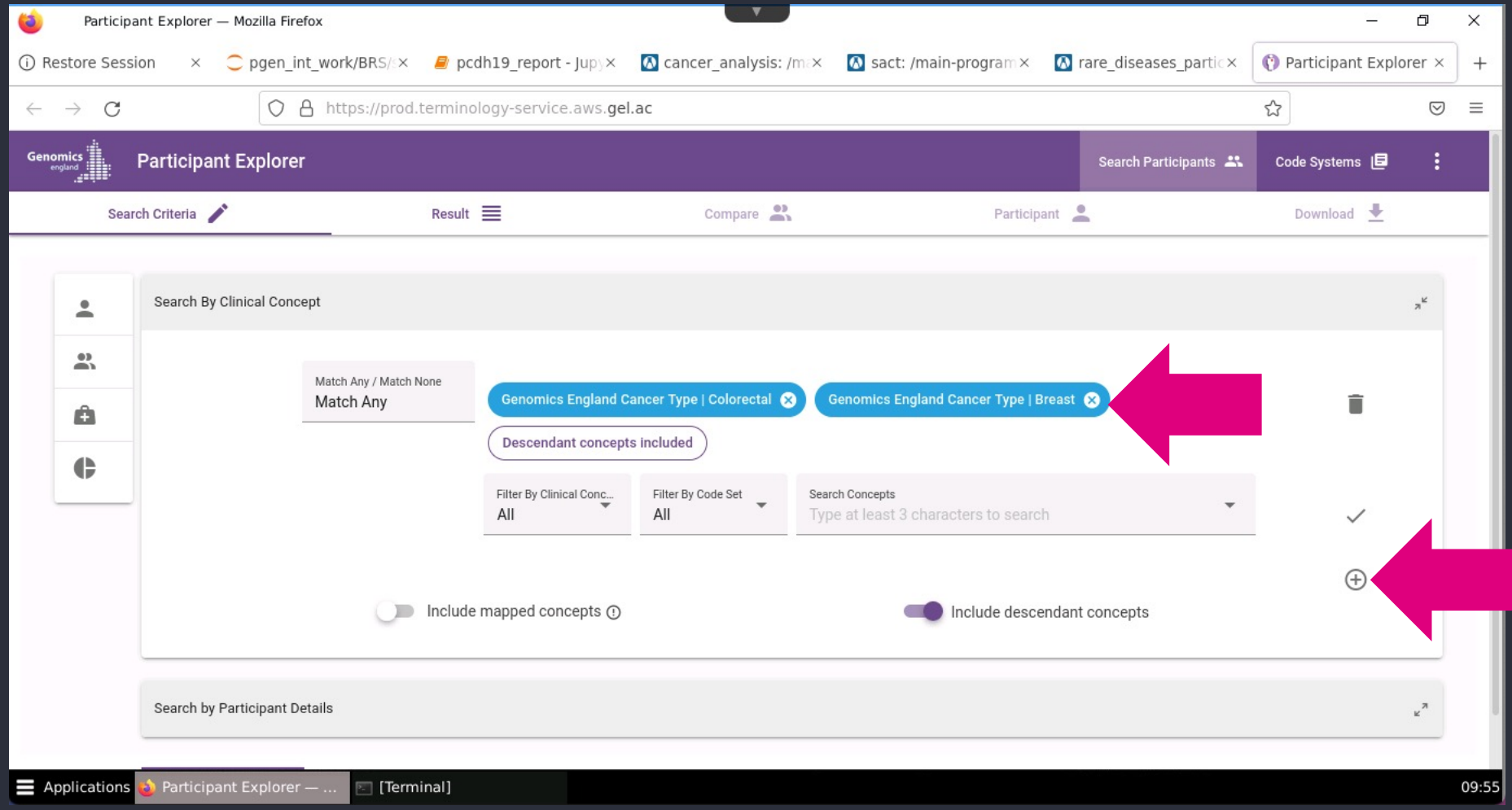
Choose which vocabulary to search with



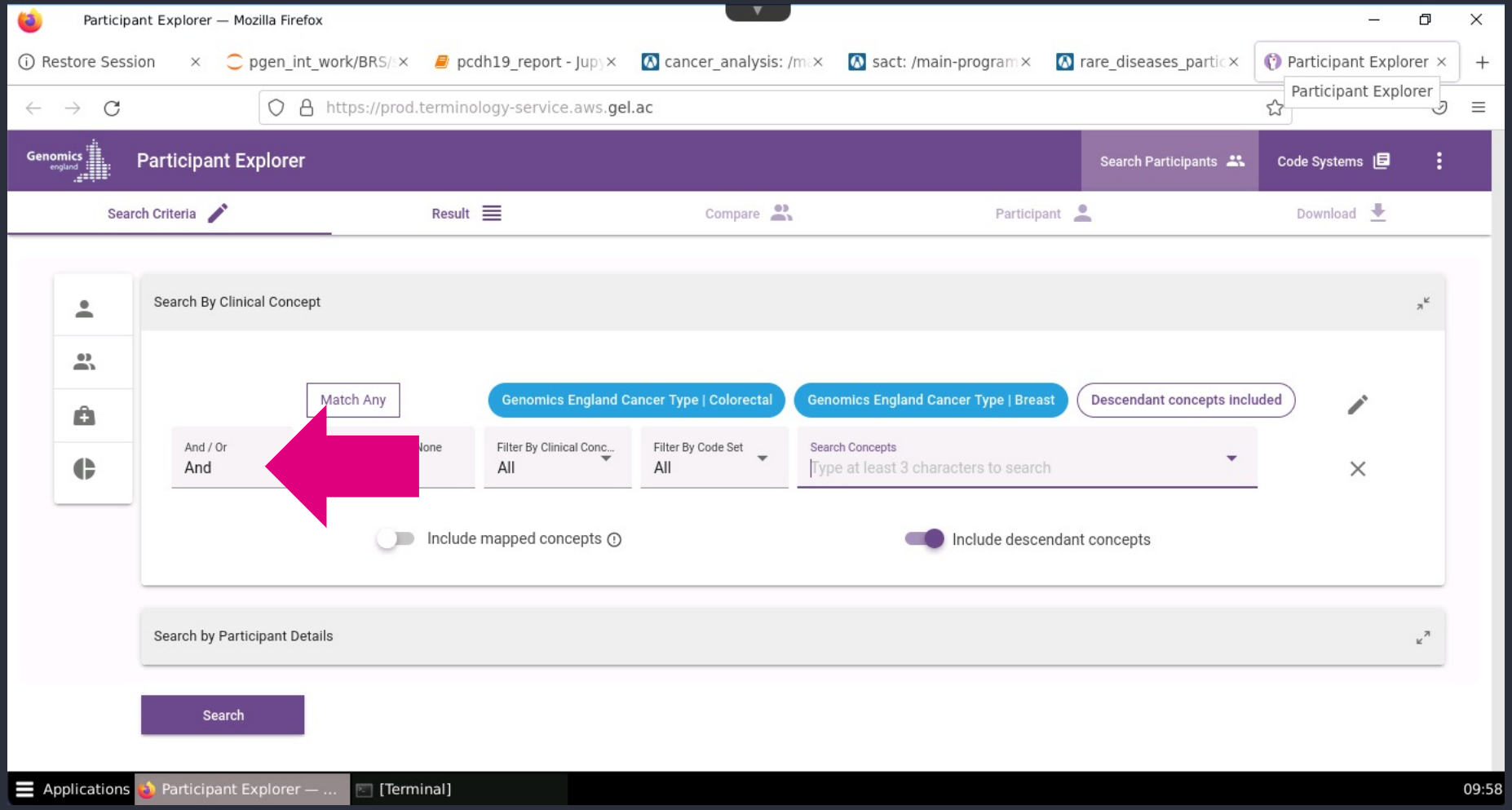
Type in to search. Choose colorectal cancer.



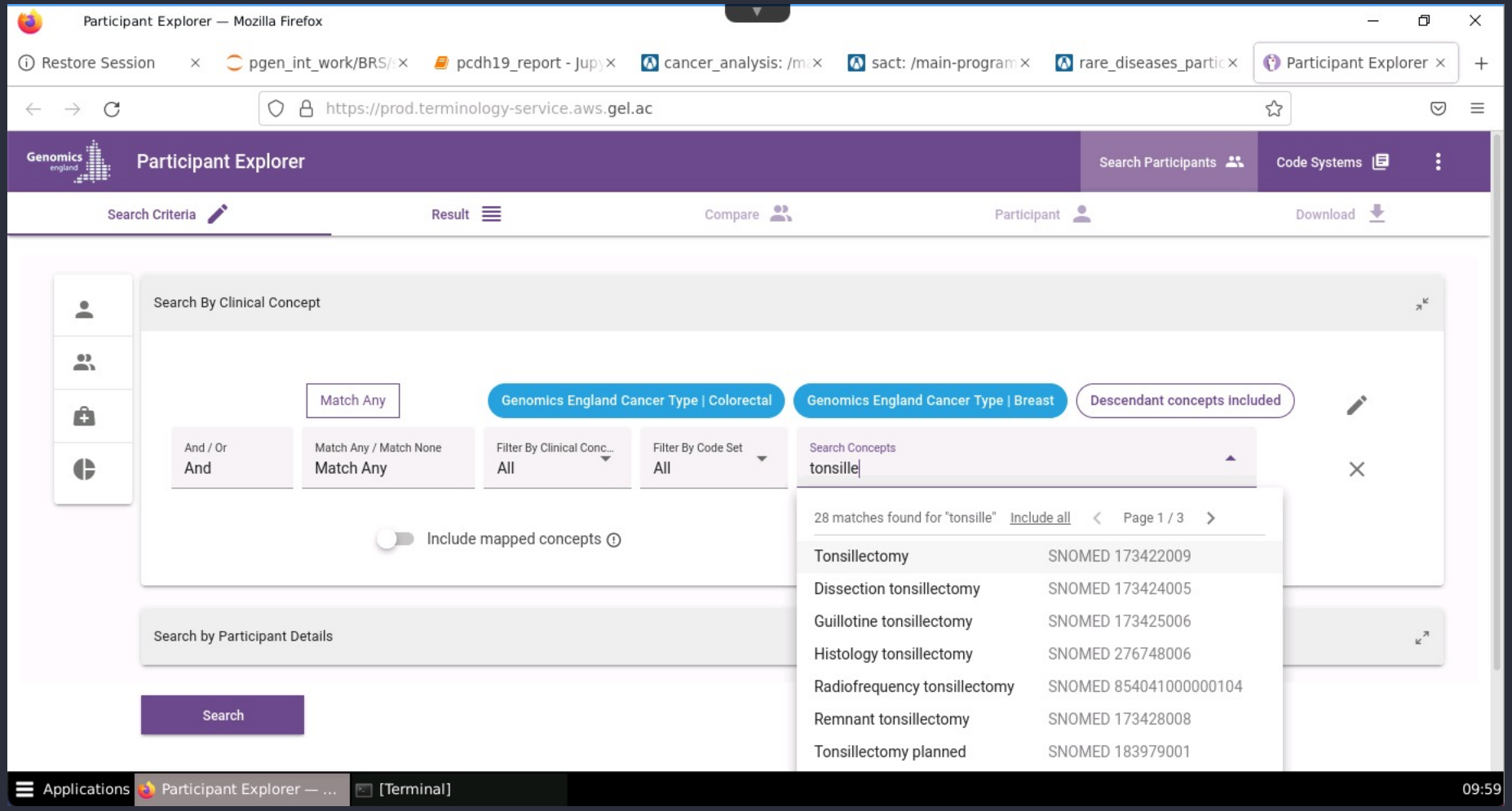
Search for more clinical concepts



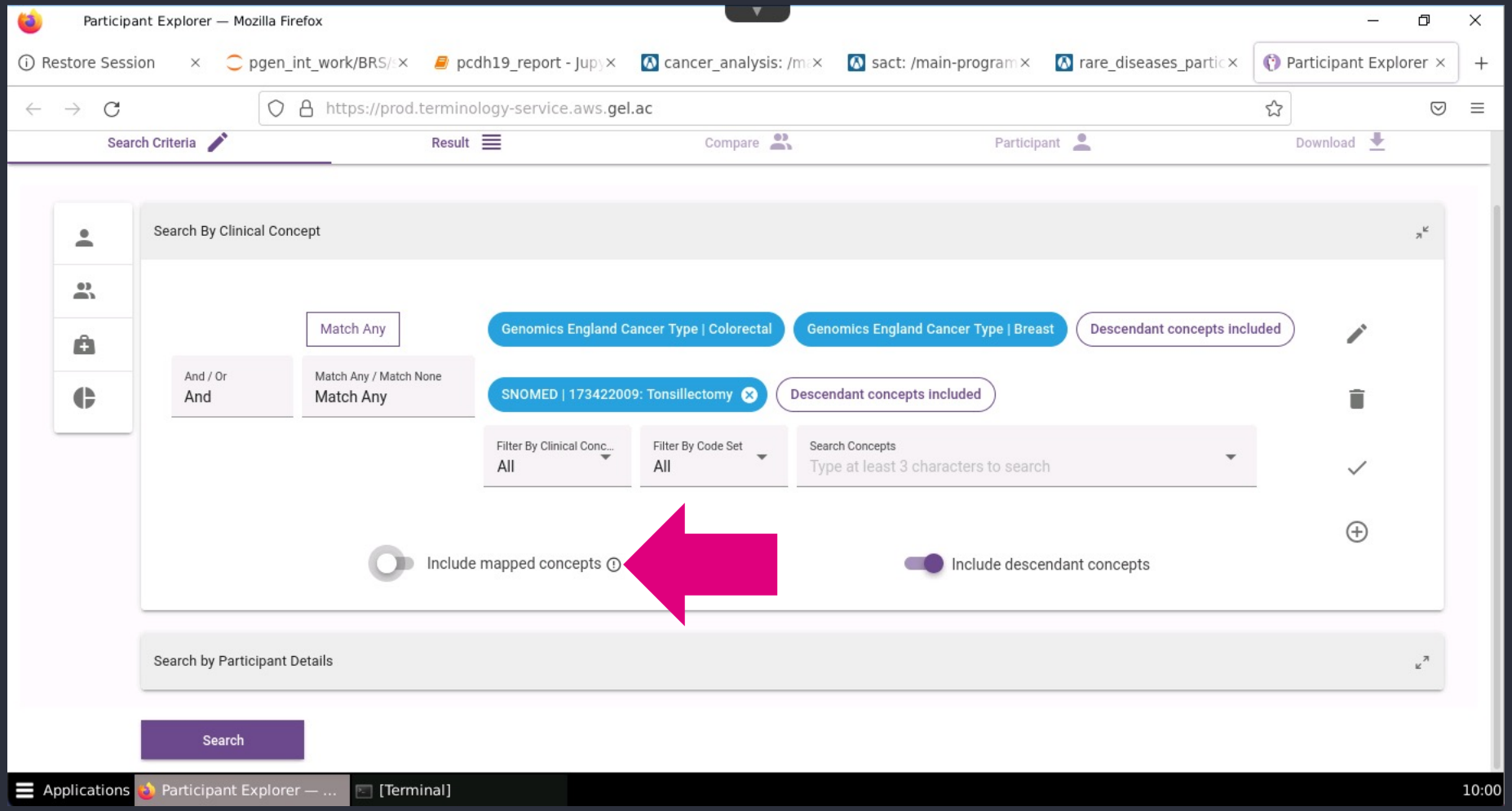
Terms appear together with Match Any/None logic. For AND logic, click on +



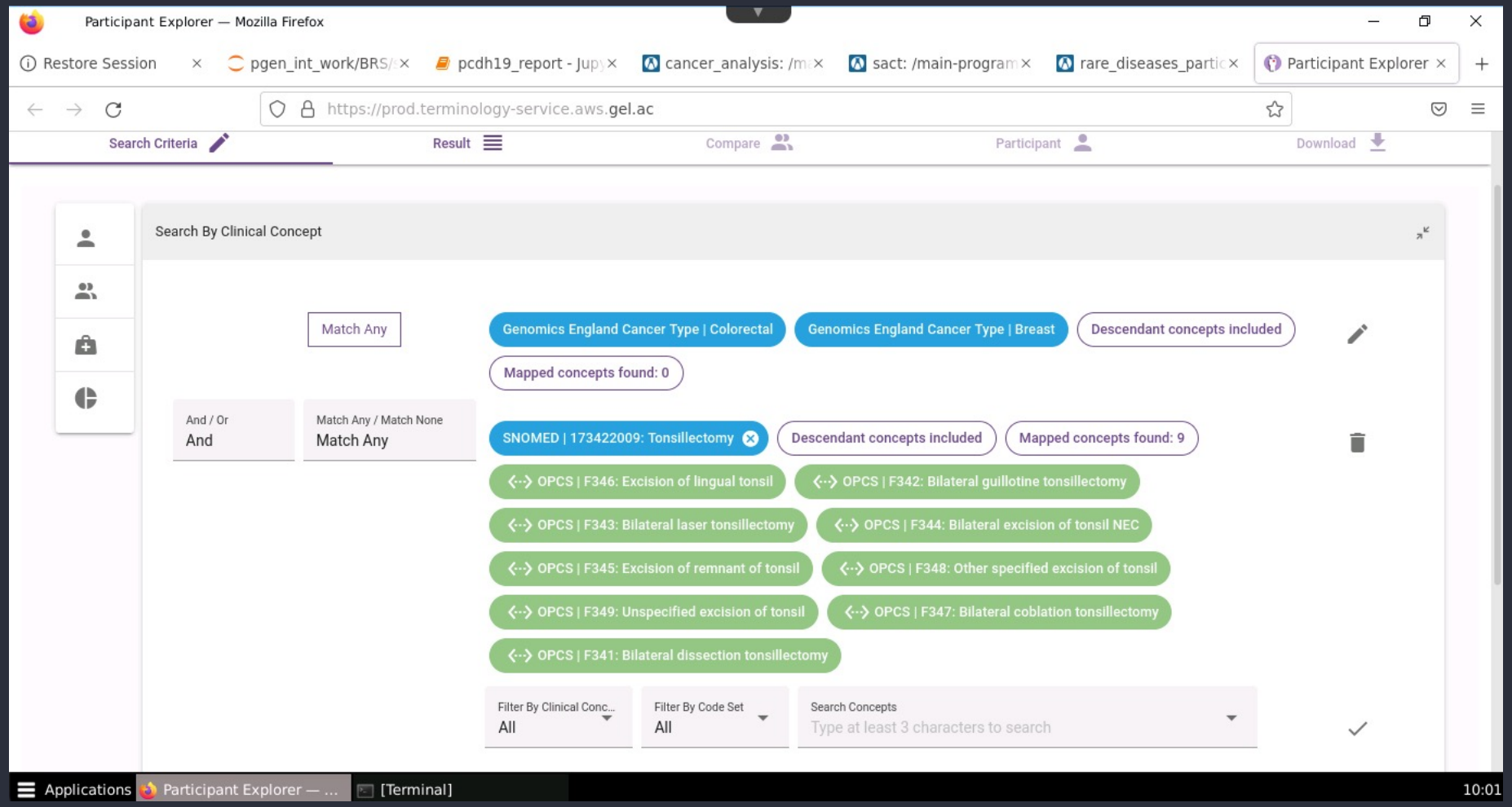
Search the same but now includes and And/Or option



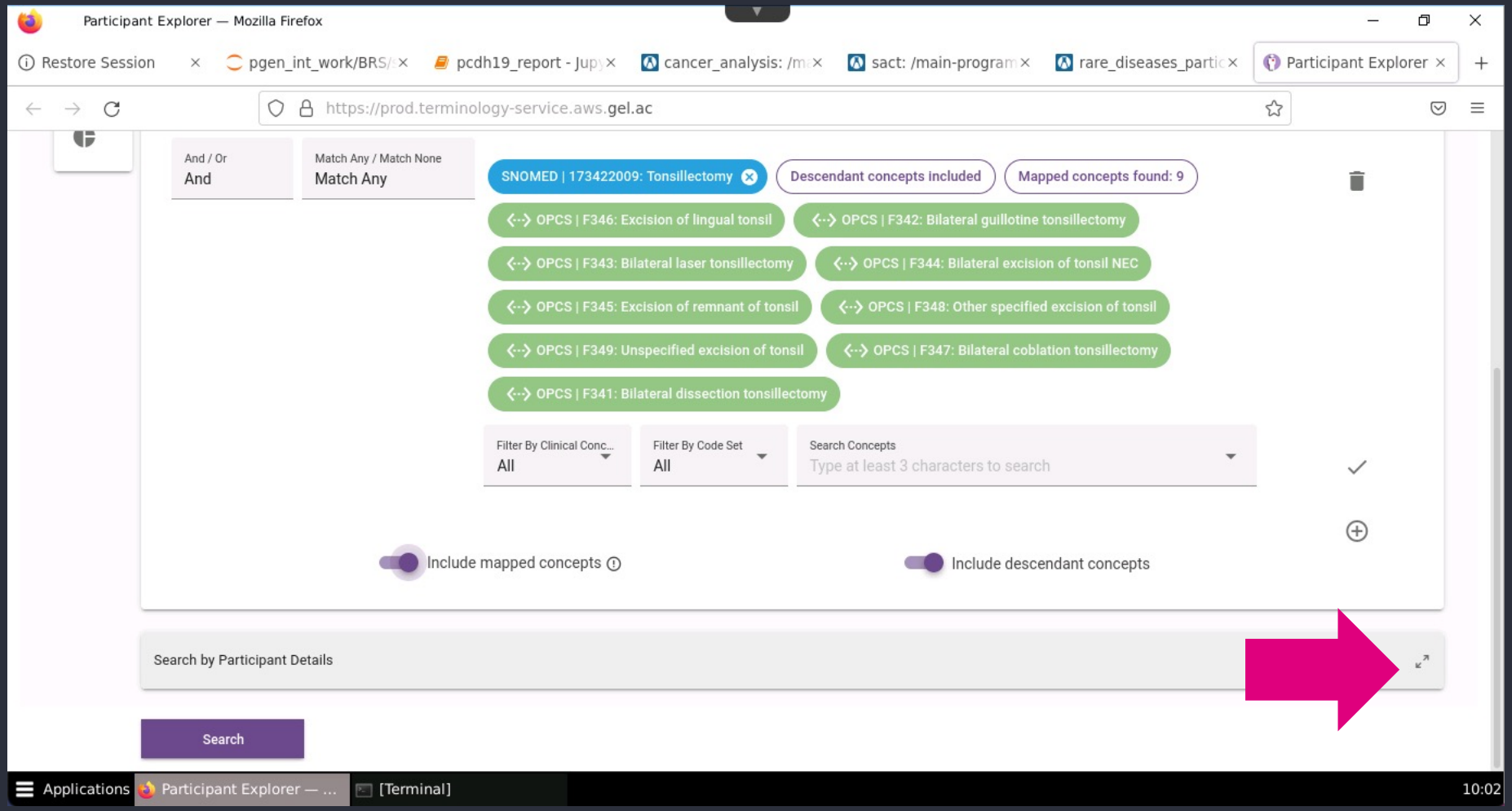
Search for tonsillectomy



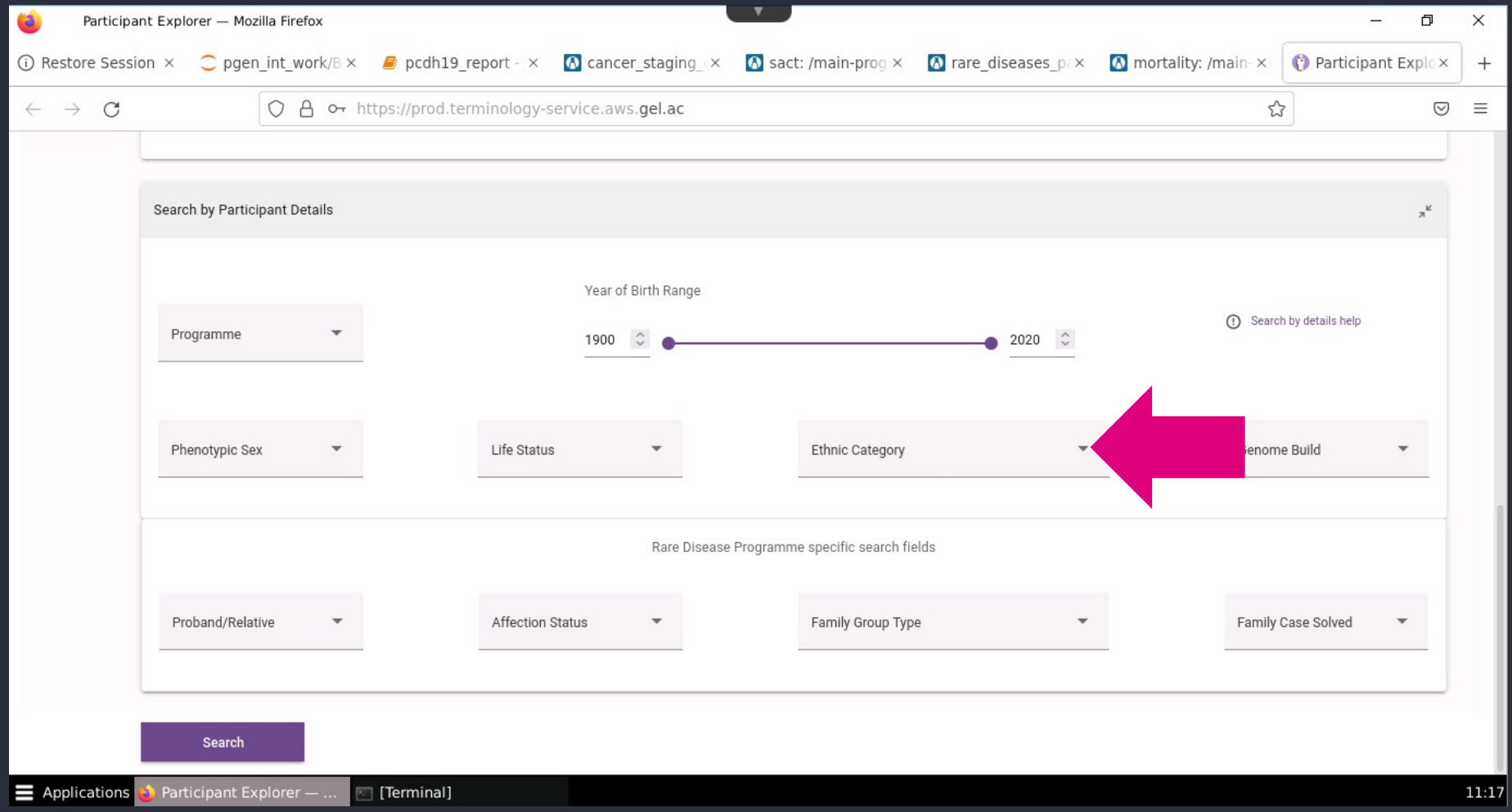
Select Include mapped concepts



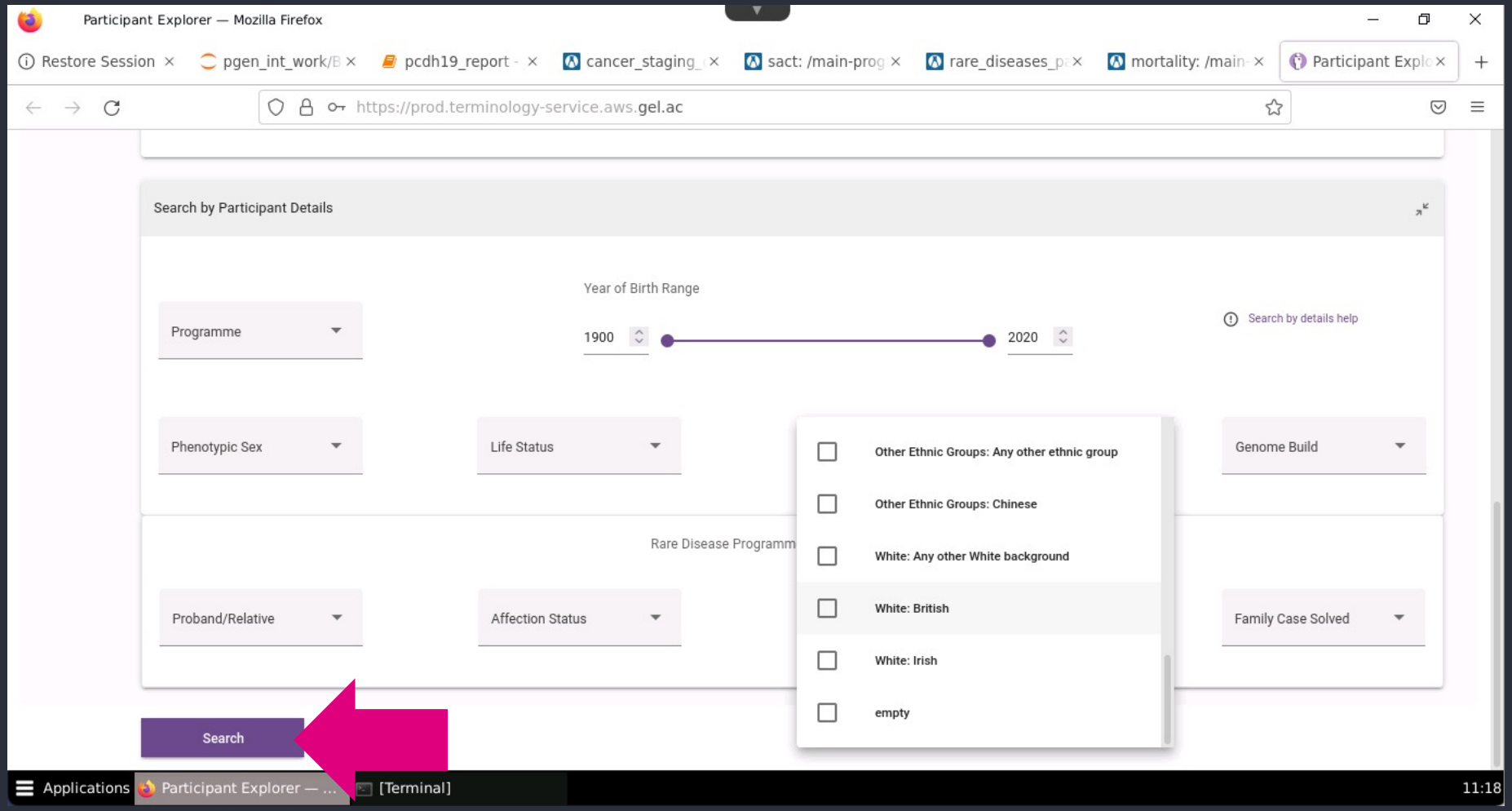
Terms from other coding systems will now also be searched



Expand Search by Participant Details



Expand any of the categories and select



Select White British then search

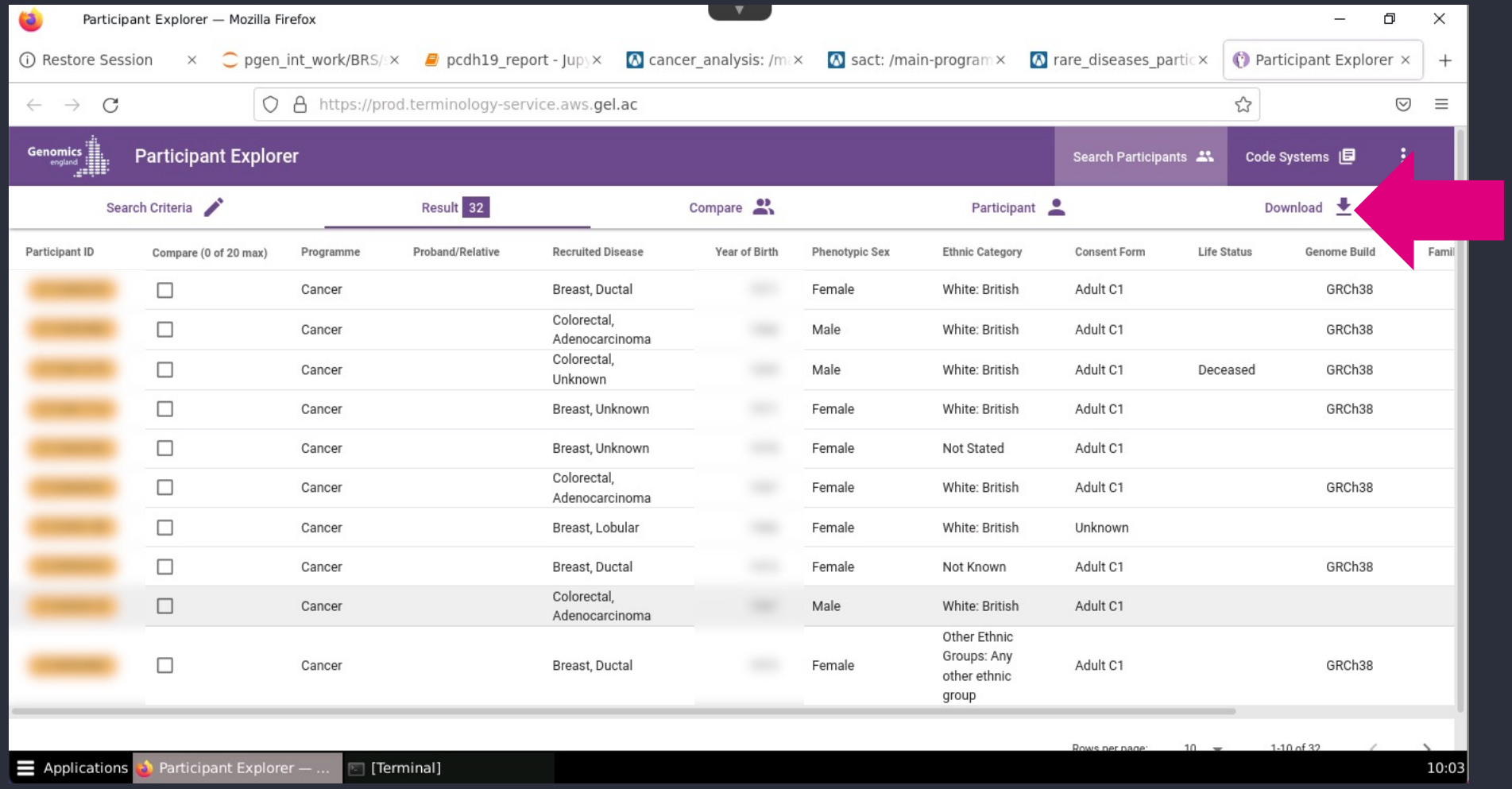
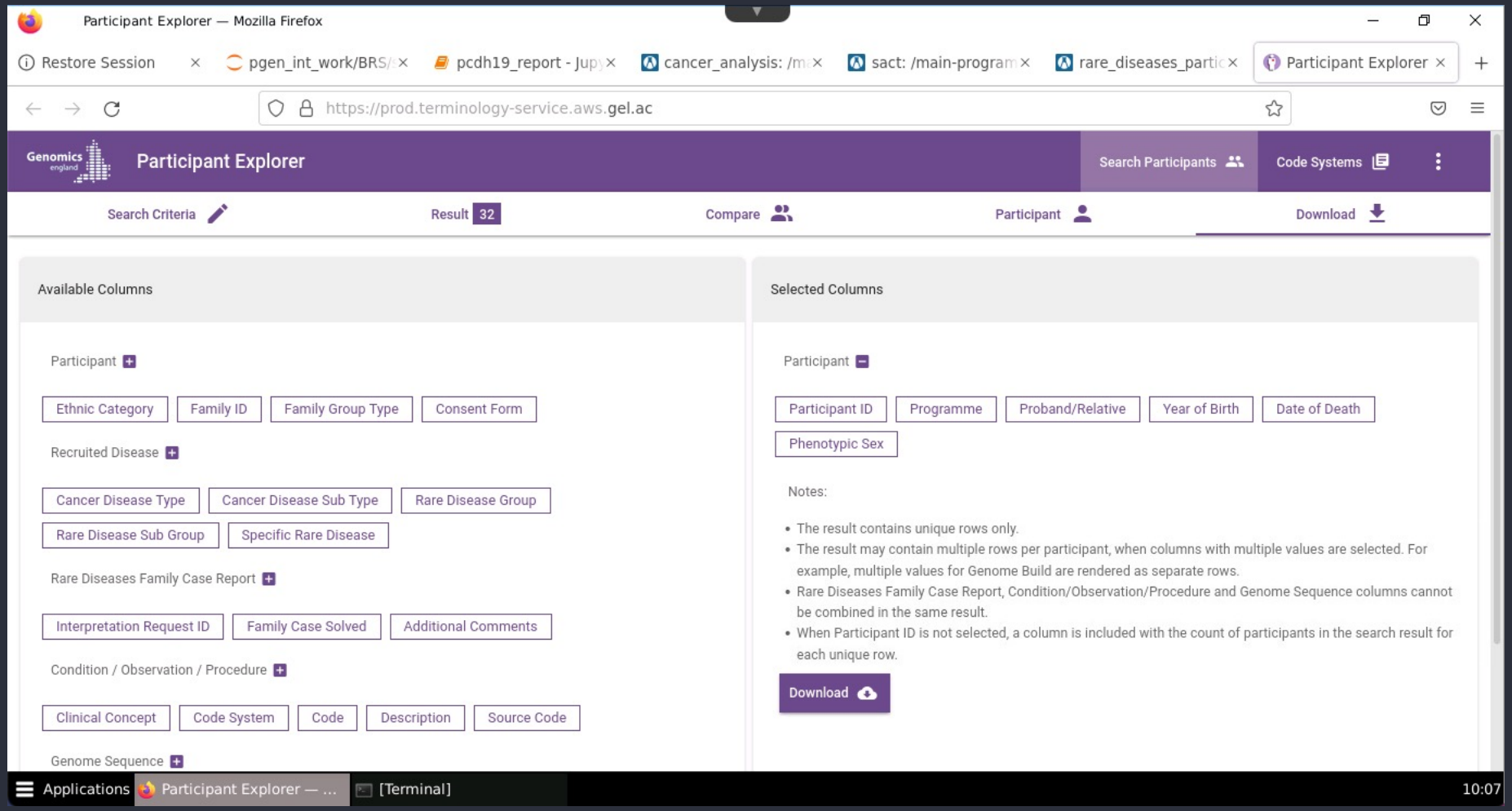
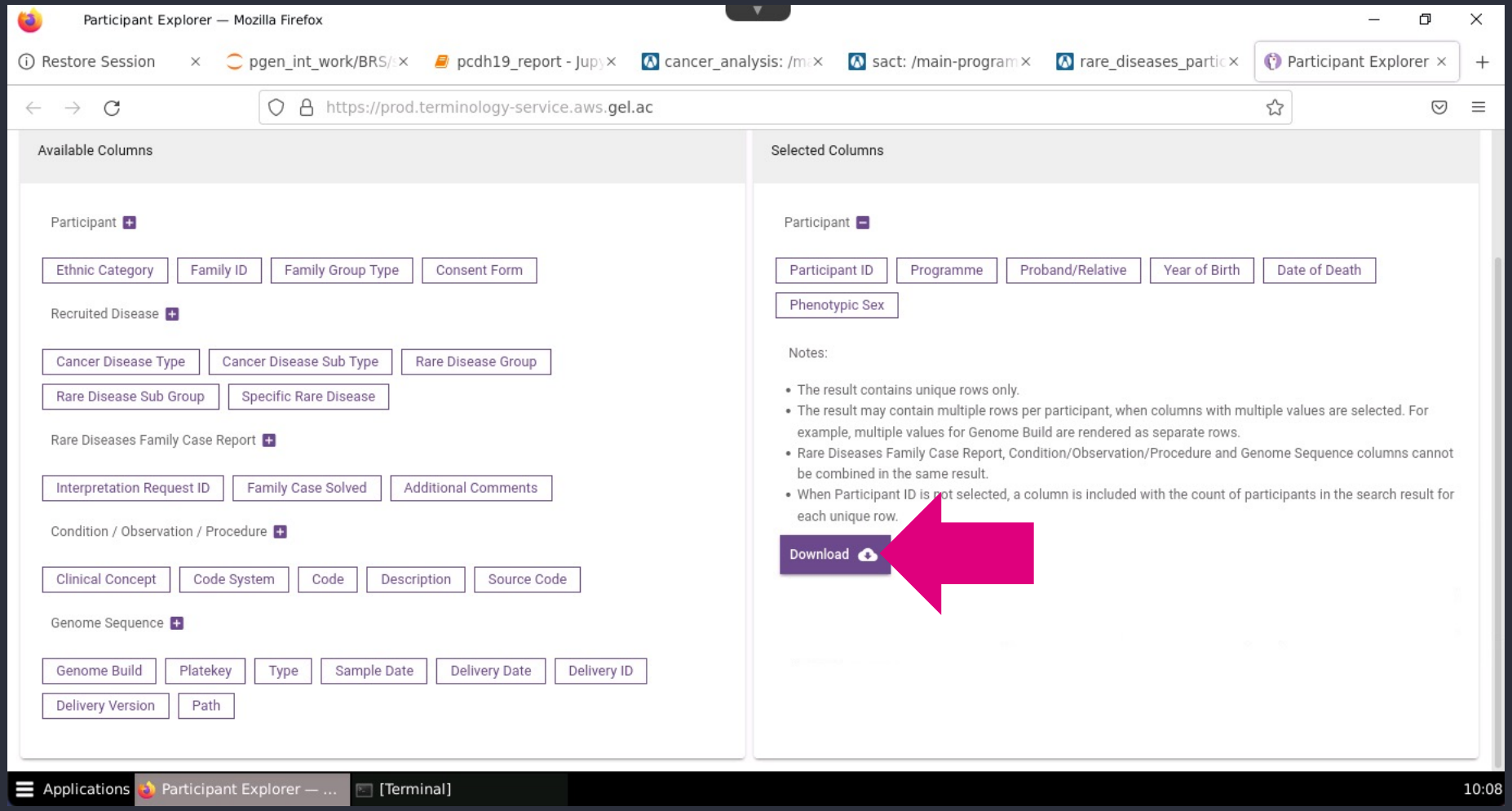


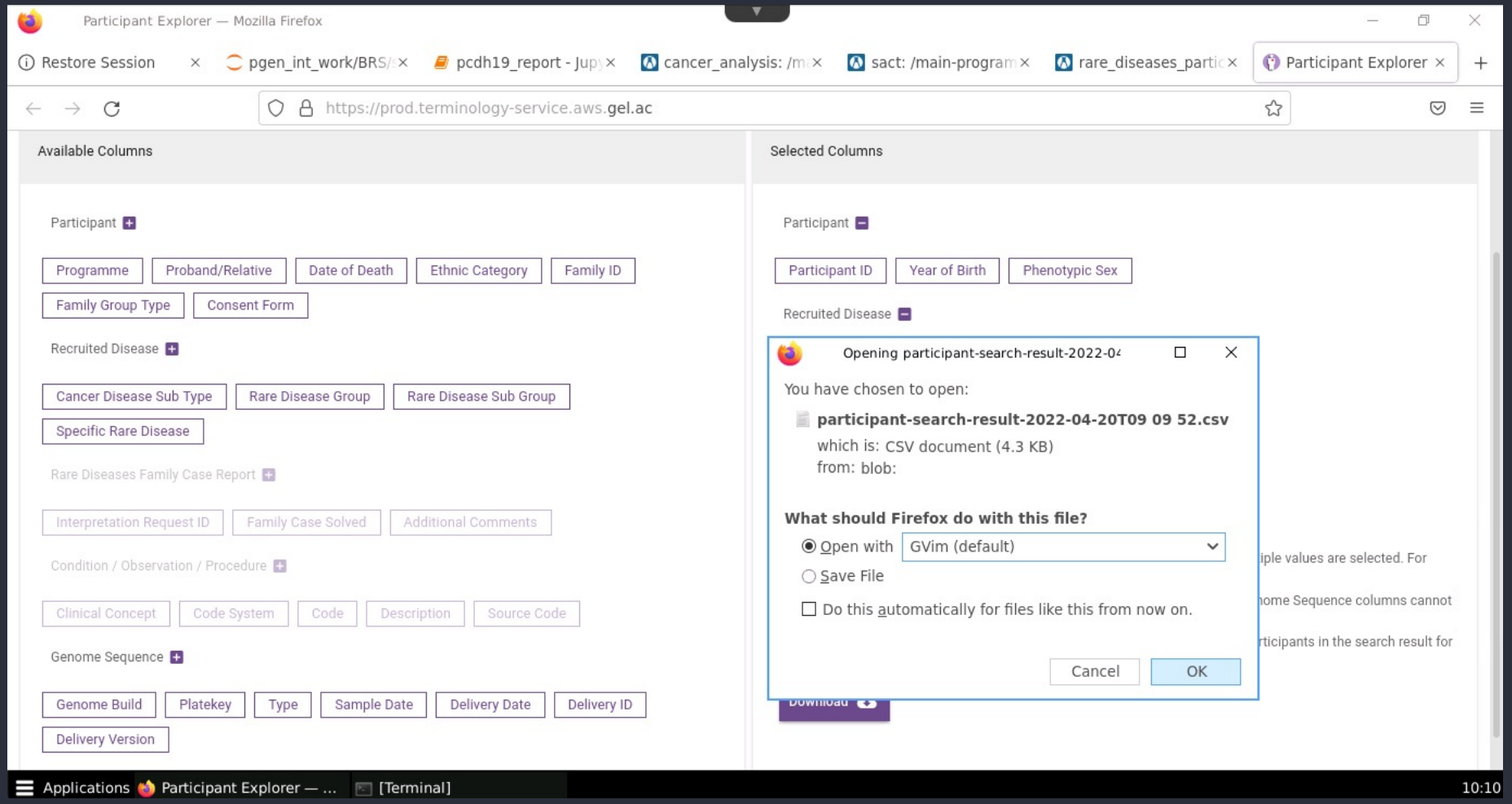
Table of matching participants. Click on download



Select which columns you want to include in your downloaded table



If you want to access the genome files, select their location. Select download when ready


Participant Explorer — Mozilla Firefox
Restore Session
pgen_int_work/BRS/
pcdh19_report - Jup
cancer_analysis: /m
sact: /main-program
rare_diseases_partic
Participant Explorer
https://prod.terminology-service.aws.gel.ac
Available Columns
Selected Columns
Participant
Programme
Proband/Relative
Date of Death
Ethnic Category
Family ID
Family Group Type
Consent Form
Recruited Disease
Cancer Disease Sub Type
Rare Disease Group
Rare Disease Sub Group
Specific Rare Disease
Rare Diseases Family Case Report
Interpretation Request ID
Family Case Solved
Additional Comments
Condition / Observation / Procedure
Clinical Concept
Code System
Code
Description
Source Code
Genome Sequence
Genome Build
Platekey
Type
Sample Date
Delivery Date
Delivery ID
Delivery Version
Participant ID
Year of Birth
Phenotypic Sex
Opening participant-search-result-2022-04
You have chosen to open:
participant-search-result-2022-04-20T09 09 52.csv
which is: CSV document (4.3 KB)
from: blob:
What should Firefox do with this file?
Open with
GVim (default)
Save File
Do this automatically for files like this from now on.
Cancel
OK
iple values are selected. For
nome Sequence columns cannot
rticipants in the search result for
Applications
Participant Explorer — ...
[Terminal]
10:10

Participant Explorer pros/cons

✅ Easy to use point and click interface

✅ Natural language search, don't need to know the codes

✅ Can combine clinical concept filters with participant details such as sex, life status and ethnicity

✅ Many concepts that are disparate in Labkey are consolidated in PXA (eg death, diagnoses)

❌ Not all of the data in the LabKey tables available to be searched

❌ Not possible to save searches and return to them for new releases or small tweaks

❌ Underlying data not exposed in results, making it difficult to verify in bulk

4. Labkey tables for cohort building

Labkey

- Participant details and family relationships
- Sample details
- Genomic file locations
- Clinical data
  - Primary data: GEL
  - Secondary data: NHS, PHE and ONS
- Bioinformatics analysis results

Cancer disease

- Recruited disease
  - cancer_analysis.disease_type
- Diagnosis codes
  - Hospital episode statistics: hes_###.diag##
    - Accident and emergency: hes_ae
    - Admitted patient care: hes_apc
    - Critical care : hes_cc
    - Outpatient: hes_op

Common to use cancer_analysis.disease_type to find participants, then verify with diagnosis codes from secondary data

Cancer staging

- cancer_staging_consolidated
  - TNM (Tumour, Node, Metastasis)
  - AJCC (American Joint Committee on Cancer)
  - Dukes (bowel)
  - Gleason (prostate)
  - FIGO (uterine)
  - HER, ER and PR status (breast)
- cancer_participant_tumour, sact and av_tumour
  - cancer_staging_consolidated comprises data from these tables
  - cancer_staging_consolidated does not have all participants and you may find more participants that fit your criteria by expanding your search to these tables

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38047040

Cancer treatment

- Systemic anti-cancer therapy: sact
  - analysis_group – all the drugs the participant was treated with
  - drug_group – the drug being referred to in this line of the table
  - Details of how/when the drug was administered

Rare disease phenotype

- rare_disease_analysis.normalised_specific_disease
- rare_diseases_participant_disease
  - Name and categorisation of the disease
- rare_diseases_participant_phenotype
  - HPO term and definition
  - HPO term present?
  - Onset, progression and severity

Rare disease cases

If you're looking for rare disease cases that have not been solved yet, or you want to learn more about solved cases (e.g., for eligibility to participate in a clinical trial), you can filter by:

- gmc_exit_questionnaire.case_solved_family
- death_details

Common disease

- Diagnosis codes
  - Hospital episode statistics: hes_###.diag##
    - Accident and emergency: hes_ae
    - Admitted patient care: hes_apc
    - Critical care : hes_cc
    - Outpatient: hes_op

5. Covariates in cohort building

Covariates to consider

- Age
- Sex
- Ethnicity
- Alive?

Age

- We don’t have Age stored
  - You need to calculate it from year_of_birth (participant table)
  - Age will always be an approximation, since we only have year and not full date: this is particularly important for anything in small children
- What age do you want to know?
  - Age now (current year – participant.year_of_birth)
  - Age at sampling (clinic_sample.clinic_sample_datetime – participant.year_of_birth)
  - Age at diagnosis (hes_###.###date – participant.year_of_birth)
  - Age at death (death_details.death_date – participant.year_of_birth

Sex

- participant.participant_phenotypic_sex: male/female/indeterminate
- participant.participant_karyotypic_sex: XX/XY + aneuploidies
- participant.participant_stated_gender: male/female/not stated

Ethnicity

- participant.participant_ethnic_category: what they have ticked on a form
- aggregate_gvcf_sample_stats.pred_ethnicity_ancestry: 0-1
  - Ethnicities are 1000 Genomes super-populations: African, South Asian, East Asian, European, American
  - If score ≥ 0.8, participant is this population
  - If all scores <0.8, participant is admixed

Alive?

- death_details
- mortality
- cen (cohort event notification)
- av_patient
- ons (Office of National Statistics)

We can be certain that a participant is dead, but not that they are alive. They could have died since the last data freeze.

6. Using the Labkey API in Python and R to build cohorts

LabKey API

Labkey API allows you to:

- Combine data and filters from multiple tables
- Work in a variety of programming languages, but most support for Python and R
- Work both locally and on the HPC

https://research-help.genomicsengland.co.uk/display/GERE/Using+the+LabKey+API

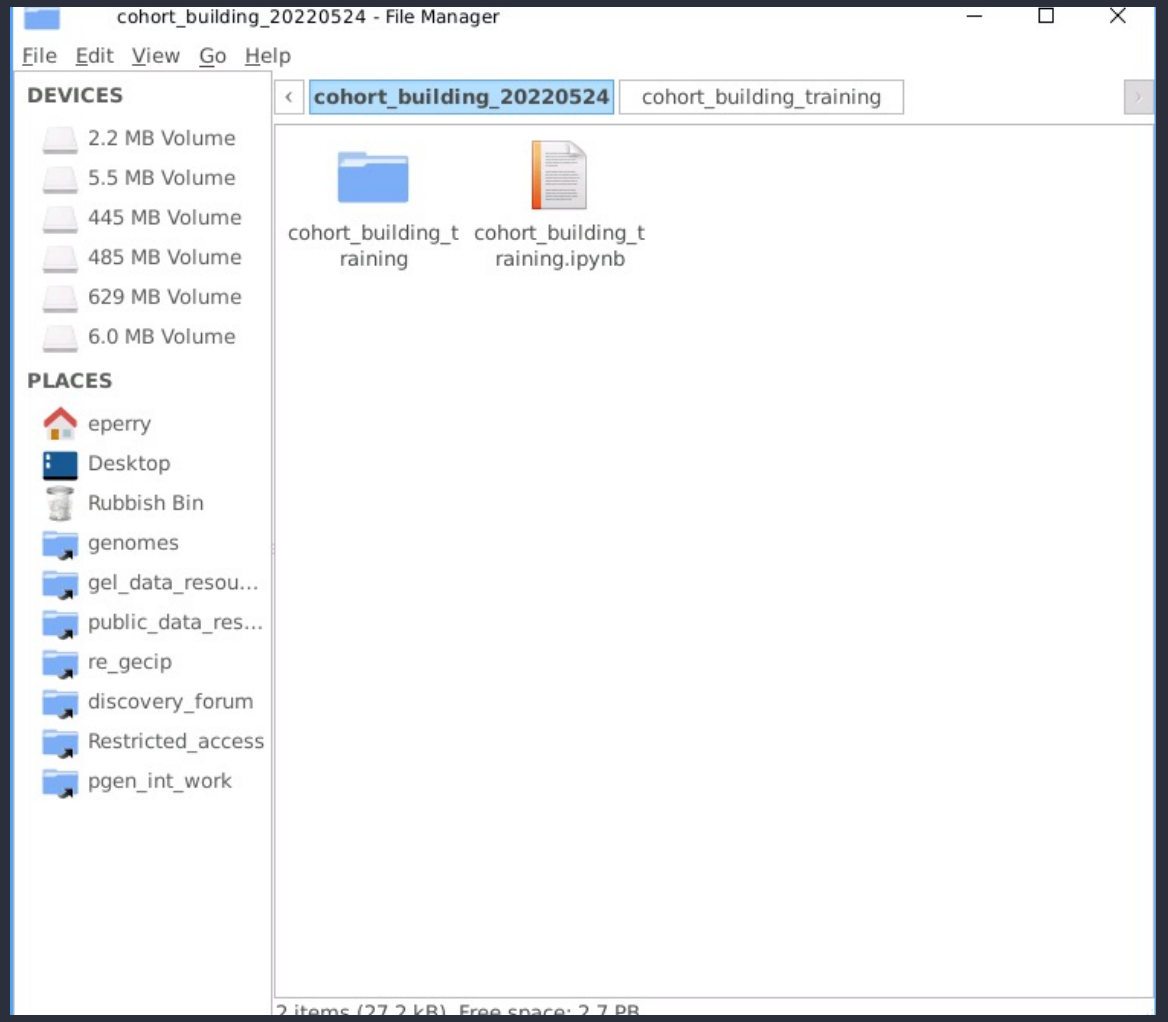
Set up .netrc

- 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://research-help.genomicsengland.co.uk/display/GERE/Using+the+LabKey+API#UsingtheLabKeyAPI-Configuringaccess

Materials

- Slides and video will be sent out to you after the session
- Scripts available in /gel_data_resources/example_scripts/workshop_scripts/cohort_building_20220524



Accessing the notebooks

Python

```
module load python/3.8.1
jupyter notebook --port
<four digit port number>
```

R

```
module load R/4.0.2
rstudio
```

Open link in browser

Demo: Labkey API

https://research-help.genomicsengland.co.uk/display/GERE/Participant+Explorer

7. Fetching data for downstream analysis

Data to include

- genome_file_paths_and_types
  - gvcfs
  - SV/CNV files
- For filtering out data from gene-centric pipelines (Gene-Variant workflow, gene-centric SNV report)
  - participant_id
- Phenotype files (for AVT and GWAS downstream analysis):
  - Platekey
  - Phenotype: 0, 1 or score
  - Covariates: age, sex etc

8. Getting help and questions

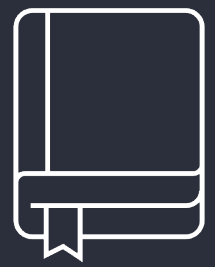
Key takeaways

Use Participant Explorer for point-and-click

Labkey tables contain loads of data for creating cohorts

Copy-and-paste whatever code snippets you need from the notebooks!

Getting help



Check our documentation:

- https://research-help.genomicsengland.co.uk/
- Click on the documentation icon in the environment

Contact our Service Desk:

- ge-servicedesk@genomicsengland.co.uk

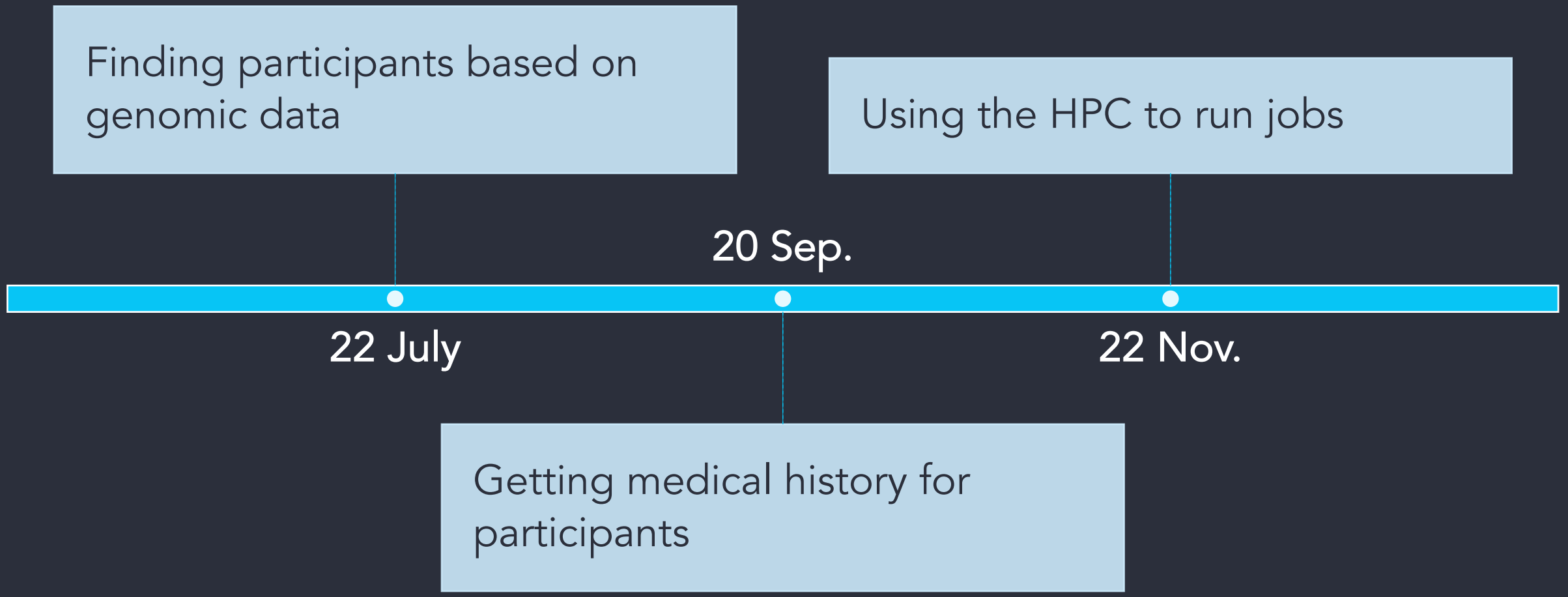
Questions

Your microphones are all 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

Future sessions



Feedback

Genomics England

Thank you