

Building rare disease cohorts with matching controls

Emily Perry

Research Engagement Manager

12th May 2026



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

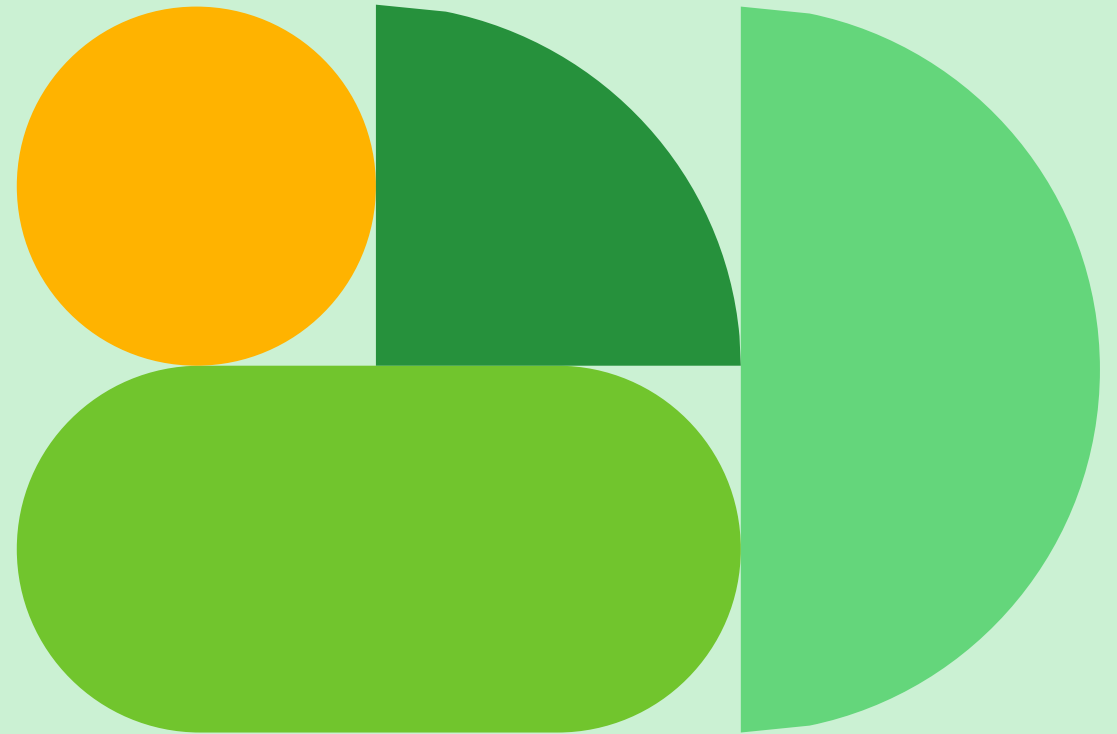
Helpers



**Matthieu
Vizquete-Forster**
Learning designer

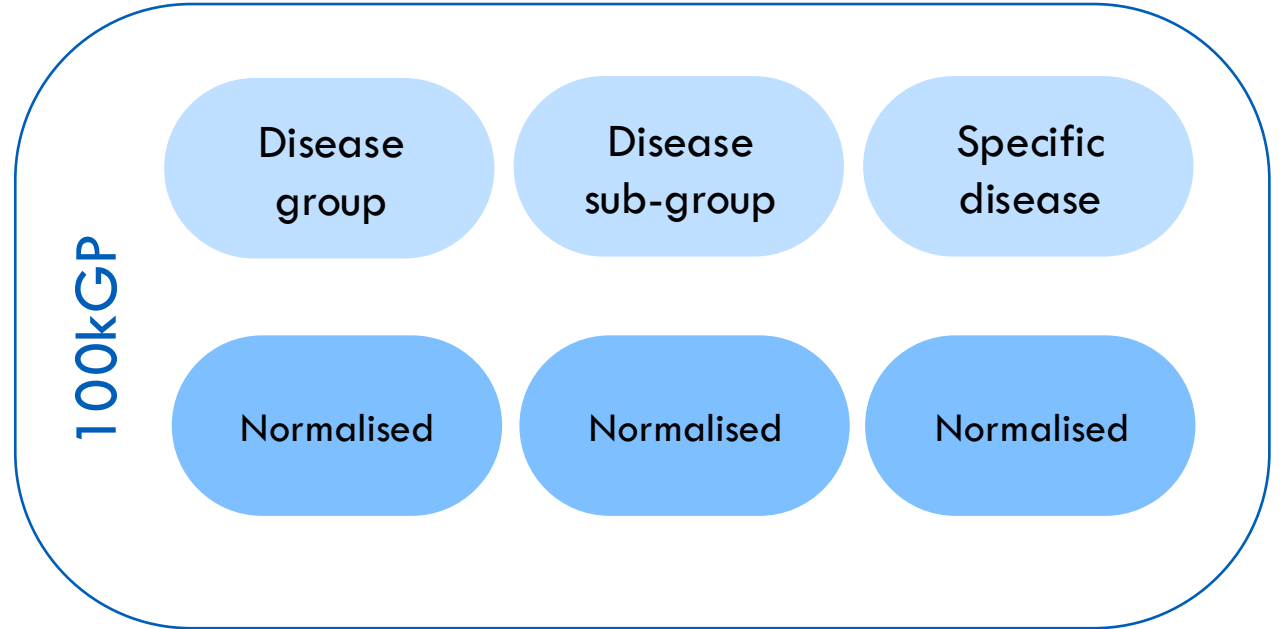
Agenda

- 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

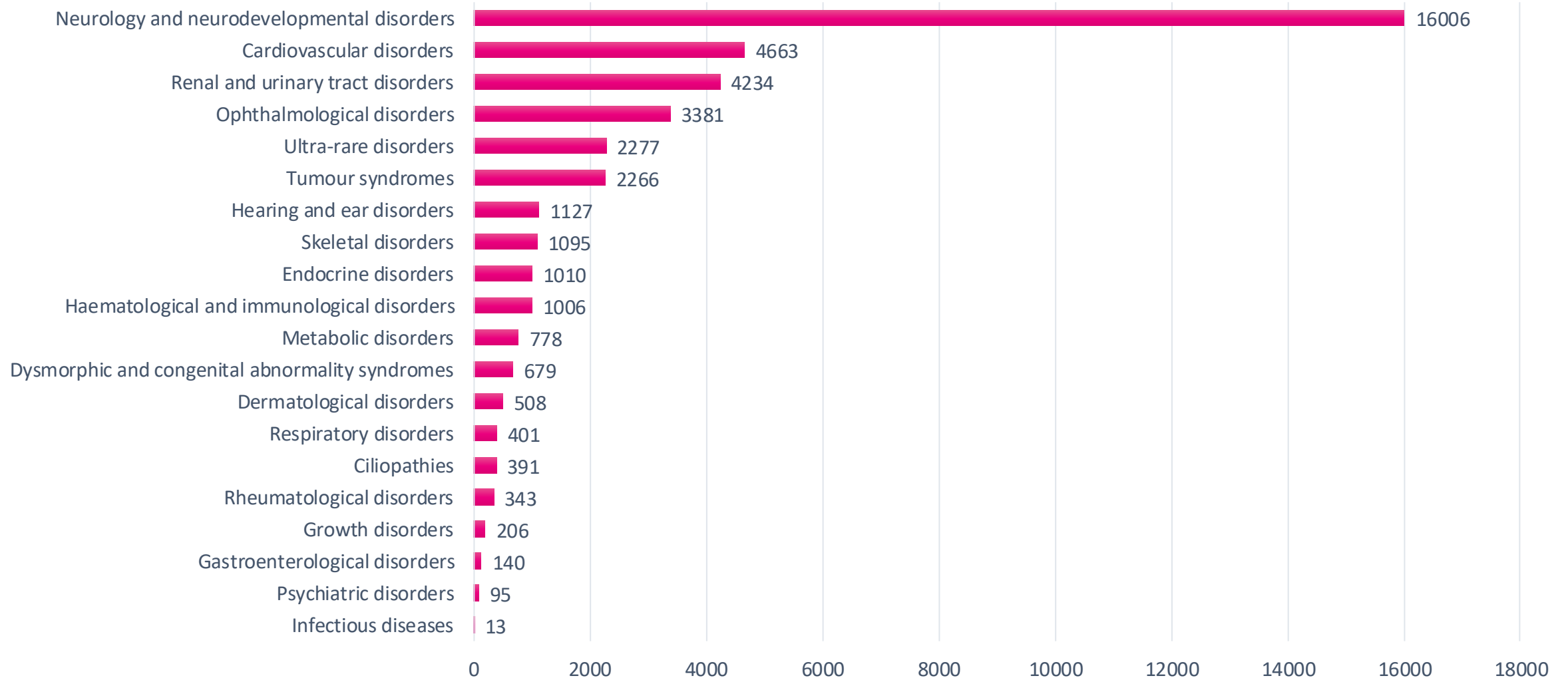


2. Parameters and considerations for building a cohort

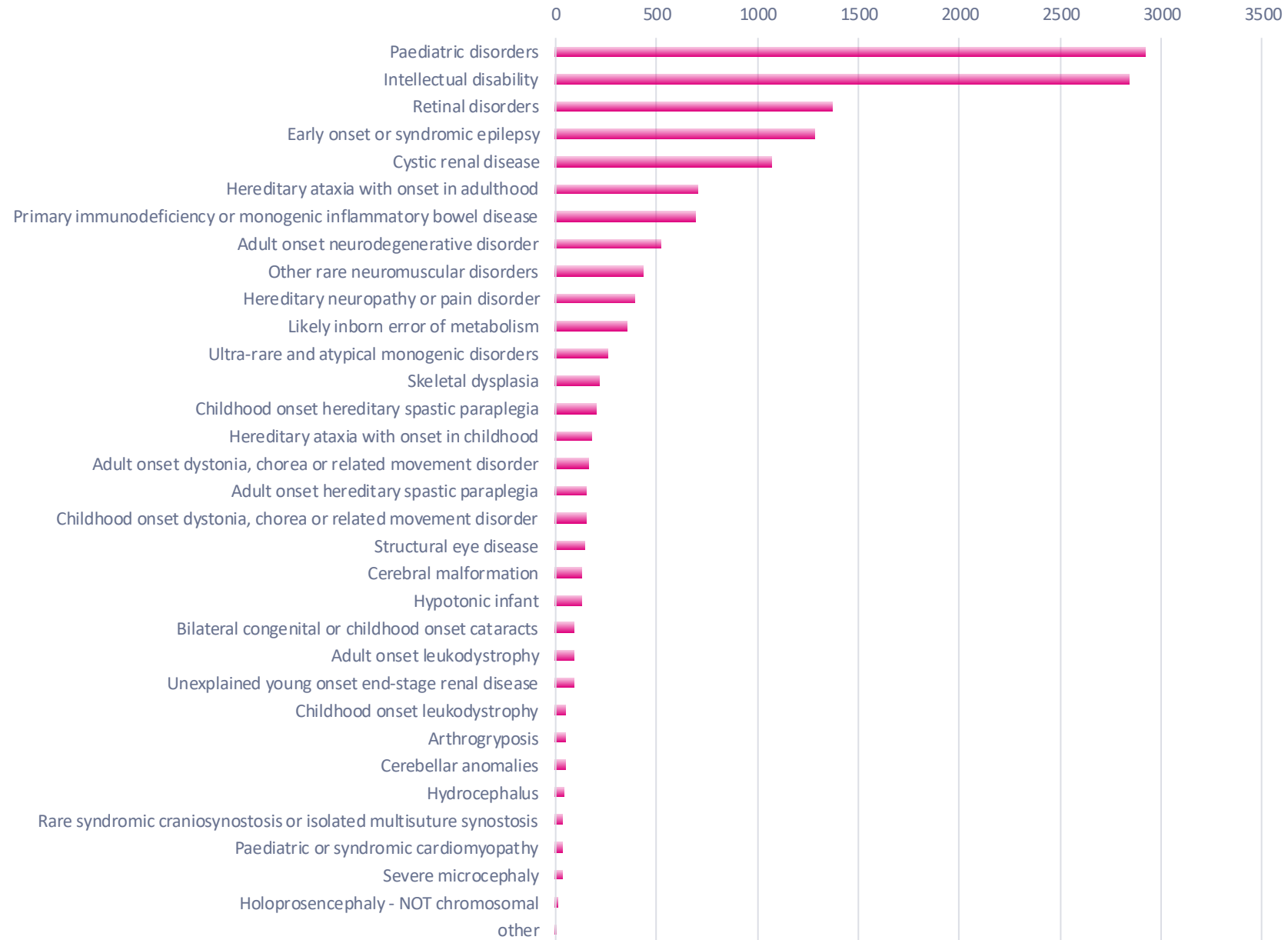
Recruited disease



100,000 Genomes rare disease



NHS GMS rare disease



Phenotypes



HPO terms
assessed on
recruitment



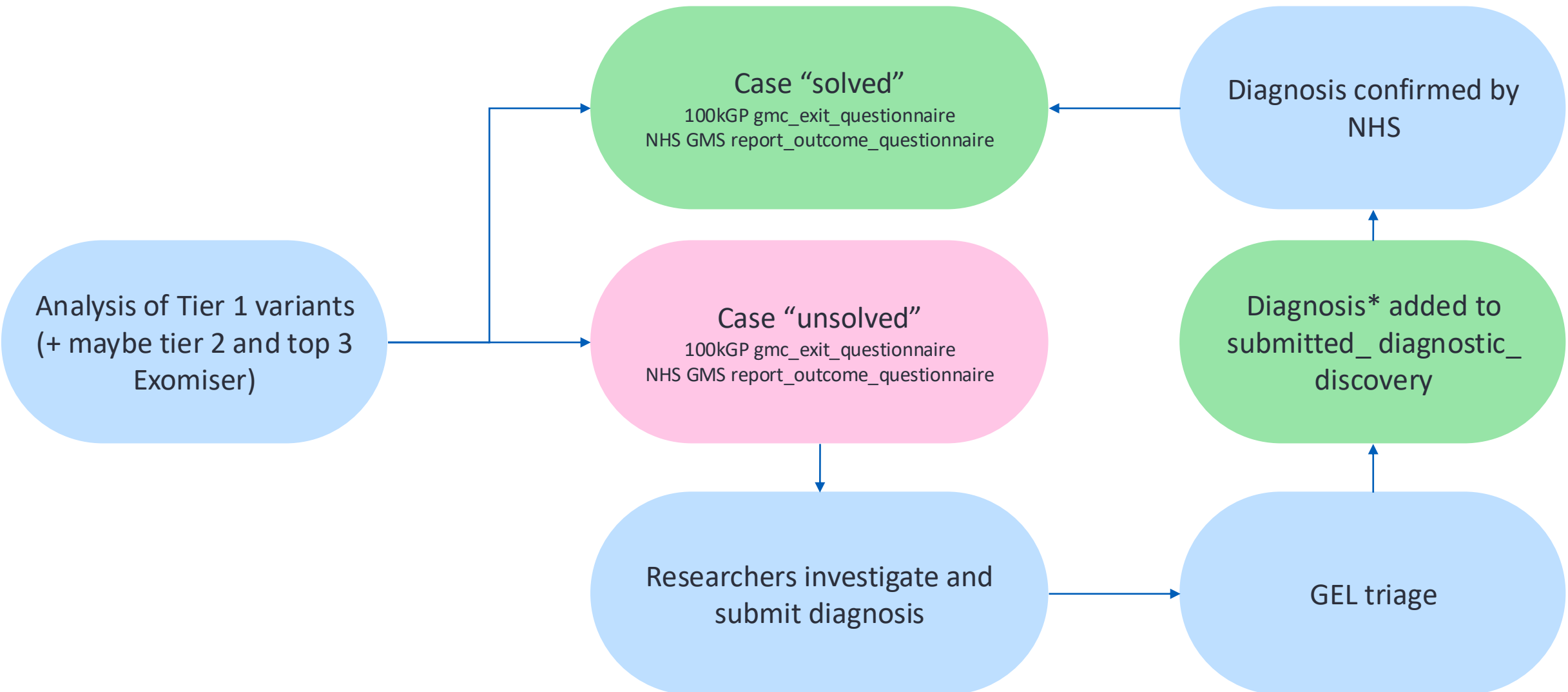
ICD-10 codes
in medical
records

Rare disease phenotyping – 100kGP

- Measurements and observations (not universal)
 - general measurements
 - early childhood observations
 - details of imaging (but not results)
 - genetic tests
 - lab tests



Solved cases



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

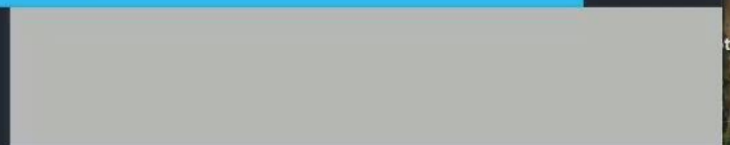
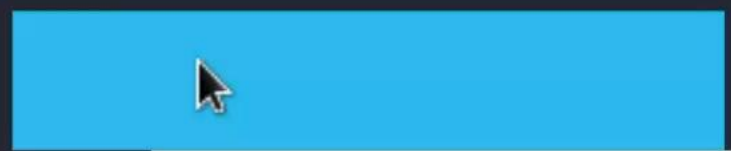
- Search for participants by:
 - IDs
 - Clinical concepts
 - diagnoses
 - treatments
 - ontology-aware
 - Personal details
- View/compare medical histories



Participant Explorer demo

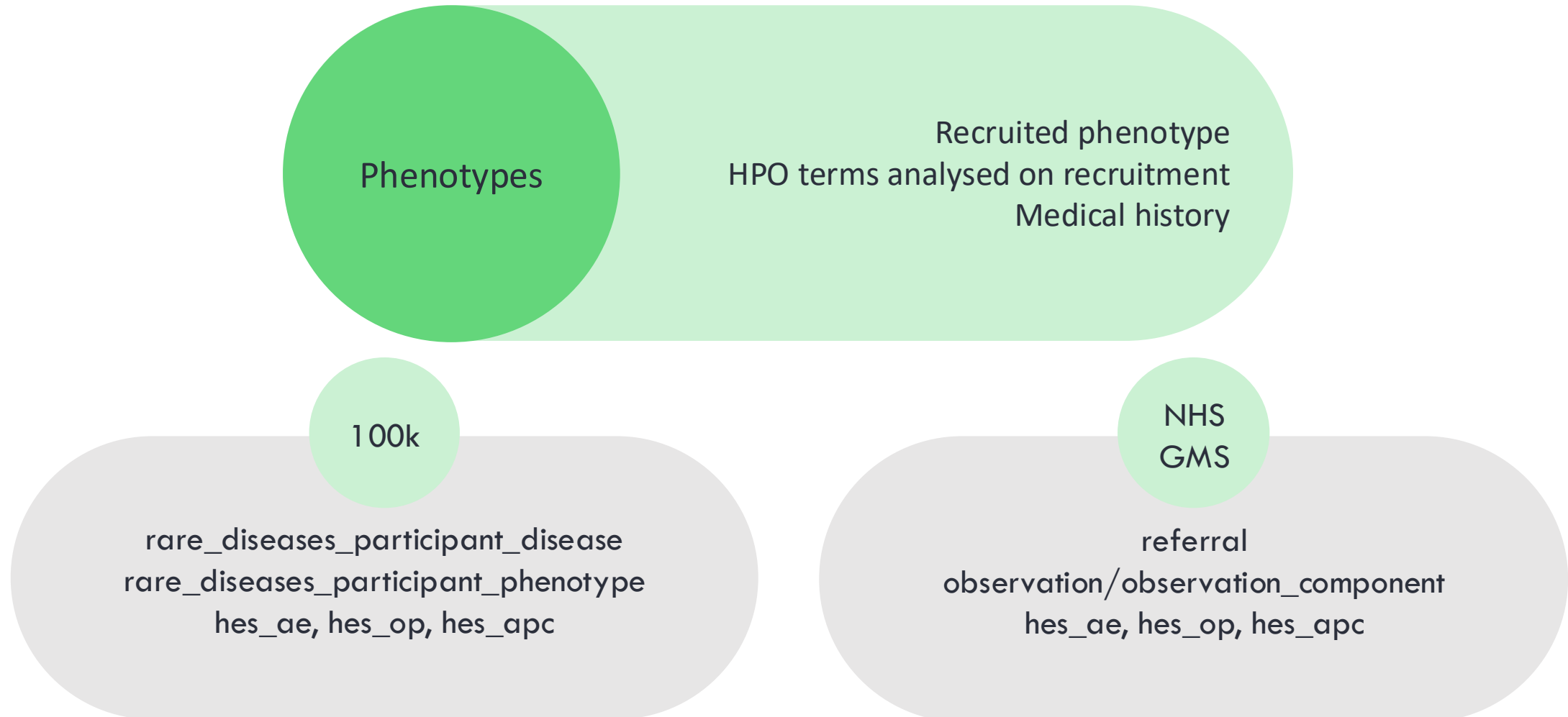


- Computer
- Emacs
- Participant Explorer
- eperry's Home
- Ensembl
- R
- Link to emily
- Firefox
- RE Messages
- Old Firefox Data
- Git GUI
- Research Environment Documentation
- Airlock
- GVim
- Research Registry
- CloudOS Academic
- IGV Browser
- RStudio
- CloudOS Discovery Forum
- IVA
- Terminal Emulator
- CloudOS Internal
- Labkey
- Text Editor
- Desktop.Rproj
- LibreOffice 7.6
- Visual Studio Code
- Document Viewer
- Open Targets
- Welcome Pack
- Dremio
- Panel App
- Trash



4. Tables for cohort building in rare disease

Rare disease cohort parameters



Data dictionary

Table	Field	Short Name	Description	Value
av_imd	participant_id	Participant ID	Participant Identifier (supplied by Genomics England)	participantid, xs:string
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 the record has been deduplicated with another patient record and the tumour(s) moved to that other patientid)
	birthdateflag	Date Of Birth Check Flag	Date Of Birth Check Flag	0,1,2,3 (Set to 0 if the date of diagnosis is not specified, 1 if the month and year of diagnosis are known but the month and day are not specified, 2 if the month and year of diagnosis are known but the date was less specific than any of the above, 3 if the date was fully specified) 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
av_patient	sex	Person Phenotypic Sex	PERSON_PHENOTYPIC_SEX_CLASSIFICATION PERSON_GENDER_CODE which is the most recent	1Male 2Female 9 Indeterminate (unable to be classified as either male or female)

Lists of tables and columns

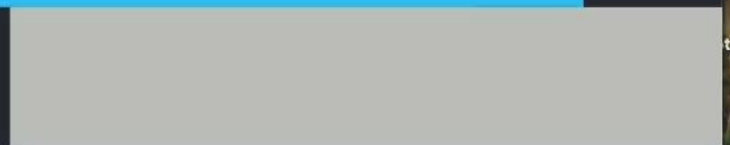
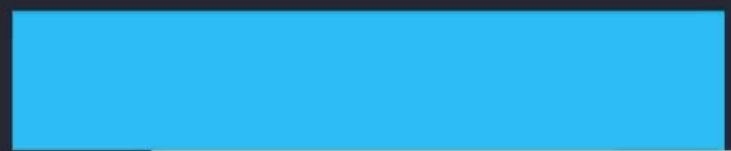
Value type or meaning of codes

Description of the data

Tables demo

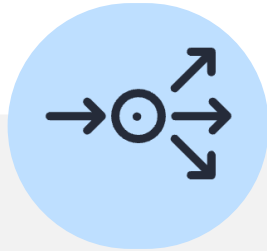


- Computer
- Emacs
- Participant Explorer
- eperry's Home
- Ensembl
- R
- Link to emily
- Firefox
- RE Messages
- Old Firefox Data
- Git GUI
- Research Environment Documentation
- Airlock
- GVim
- Research Registry
- CloudOS Academic
- IGV Browser
- RStudio
- CloudOS Discovery Forum
- IVA
- Terminal Emulator
- CloudOS Internal
- Labkey
- Text Editor
- Desktop.Rproj
- LibreOffice 7.6
- Visual Studio Code
- Document Viewer
- Open Targets
- Welcome Pack
- Dremio
- Panel App
- Trash

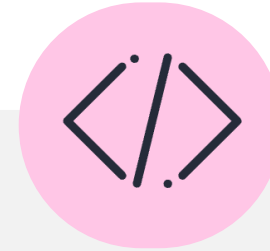


5. Programmatic cohort building in Python and R

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

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

Programming tools in the RE



Demo notebooks



`/gel_data_resources/example_scripts/
workshop_scripts/rare_disease_cohorts_2026`

Programming demo

localhost:8387/notebooks/pgen_int_work/BRS/emily/rare_dise...

jupyter rd_cohort_building_training (unsaved changes) Logout

File Edit View Insert Cell Kernel Help Trusted Python 3 (ipykernel)

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
- Helper function to access the LabKey API with Python
- 100kGP or Main programme
 - Case cohort
 - Recruited disease
 - HPO terms
 - ICD10 codes
 - Unsolved cases
 - Control cohort
 - NOT phenotype
 - Match demographics
 - General inclusion criteria
 - Filepaths
- NHS GMS

Import Python modules you need

```
In [138]: 1 import numpy as np
          2 import functools
          3 import labkey
          4 import pandas as pd
```

~/pgen_int_work/BRS/emily/rare_disease_cohorts_2024/rare_disease_cohorts_2025.nb.html

rare_disease_cohorts_2025.nb.html Open in Browser Find Publish

Building rare disease cohorts with matching controls in R

- Building rare disease cohorts with matching controls in R
 - Import R libraries you need
 - Helper function to access the LabKey API with R
 - Case cohort
 - Recruited disease
 - HPO terms
 - ICD10 codes
 - Unsolved cases
 - Control cohort
 - NOT phenotype
 - Combine and exclude
 - Match demographics
 - General inclusion criteria
 - Filepaths
 - Working with aggregate VCFs

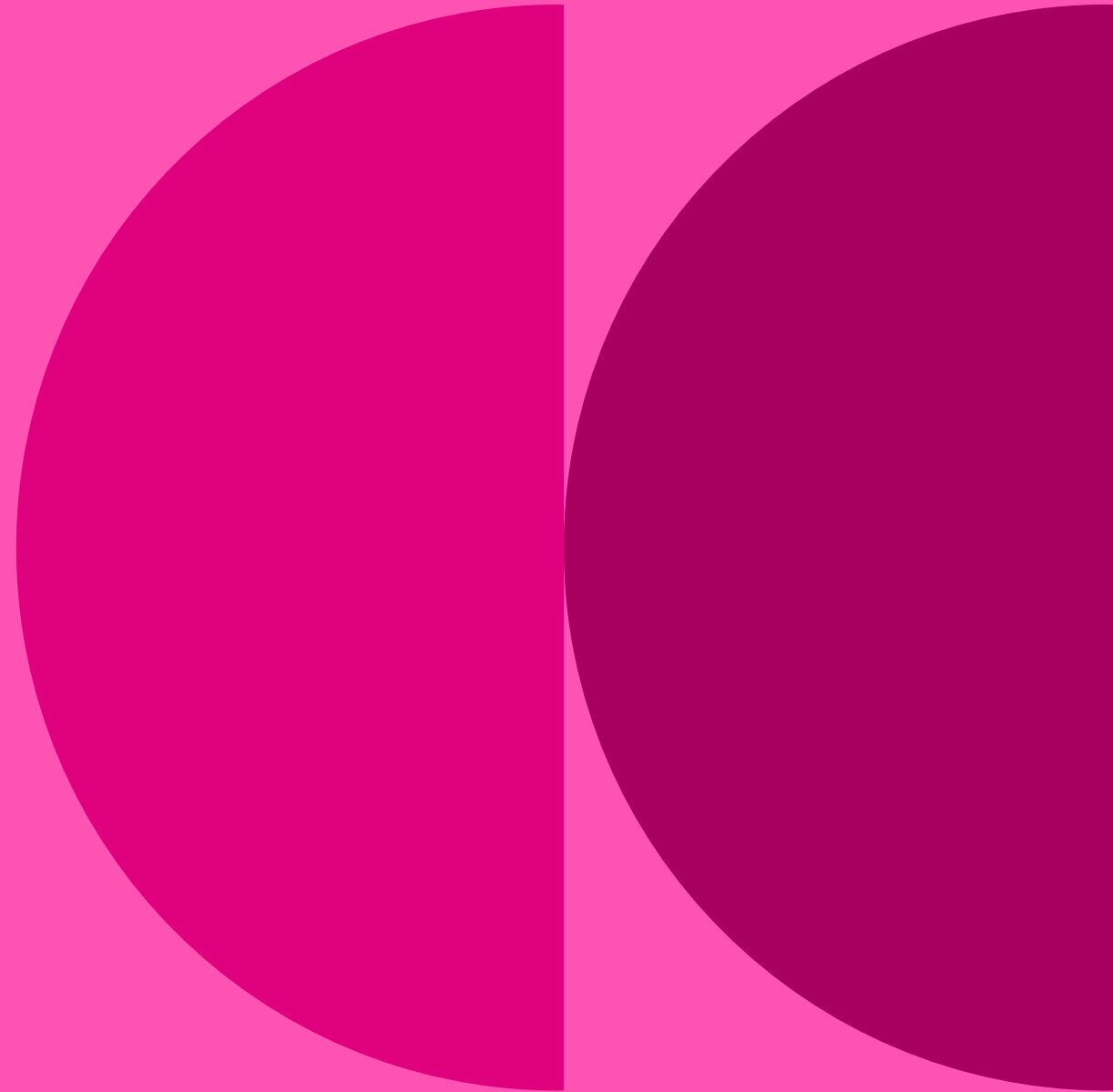
Building rare disease cohorts with matching controls in R

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

```
library(tidyverse)
```

6. Creating a matched control cohort



NOT phenotype



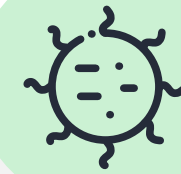
NOT recruited disease or related disease



NOT HPO terms or related HPO terms

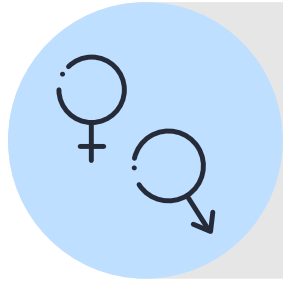


NOT ICD10 codes or related ICD10 codes

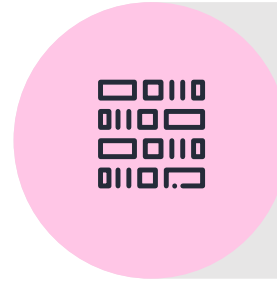


NOT cancer in related tissues

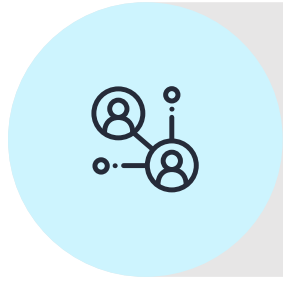
General inclusion criteria



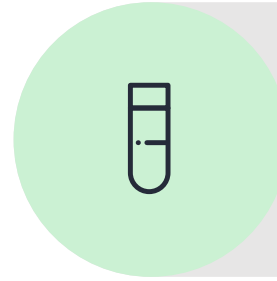
karyotypic = phenotypic sex
No aneuploidy



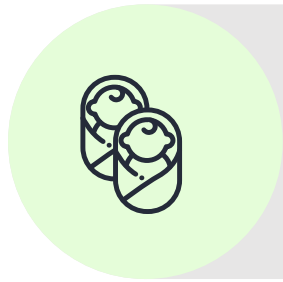
(for 100k)
aligned to GRCh38
in AggV2



assess kinship



blood
EDTA extraction
PCR-free



include MZ twins?



Additional QC

Match case/control



work with a single ethnicity

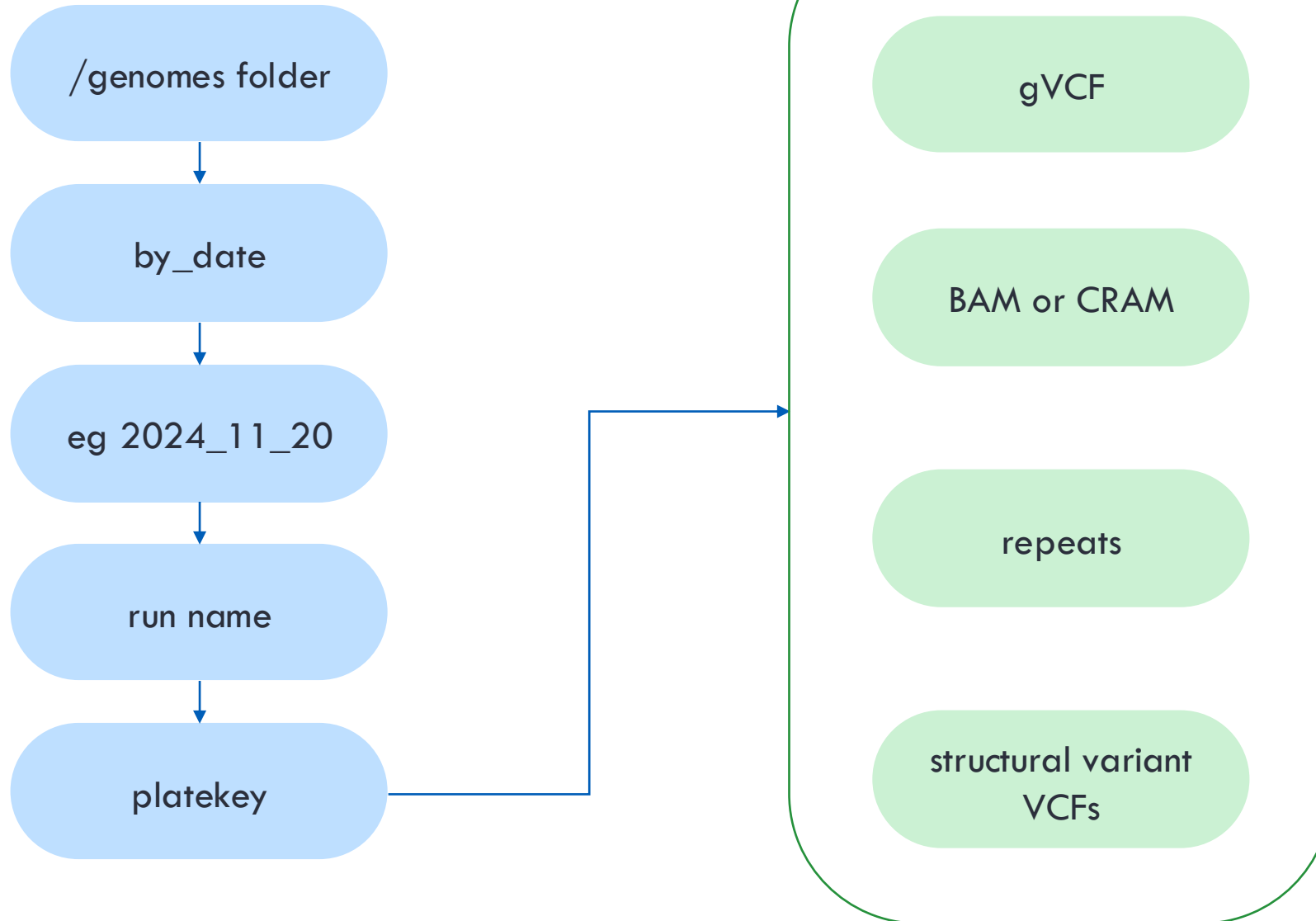


match sex ratios and age distribution

Control demo

6. Getting genomic filepaths for your cohort

Genomes location



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

Participant Explorer

research-participantexplorer.preprod.aws.gel.ac

Genomics Participant Explorer

Search Participants Code Systems

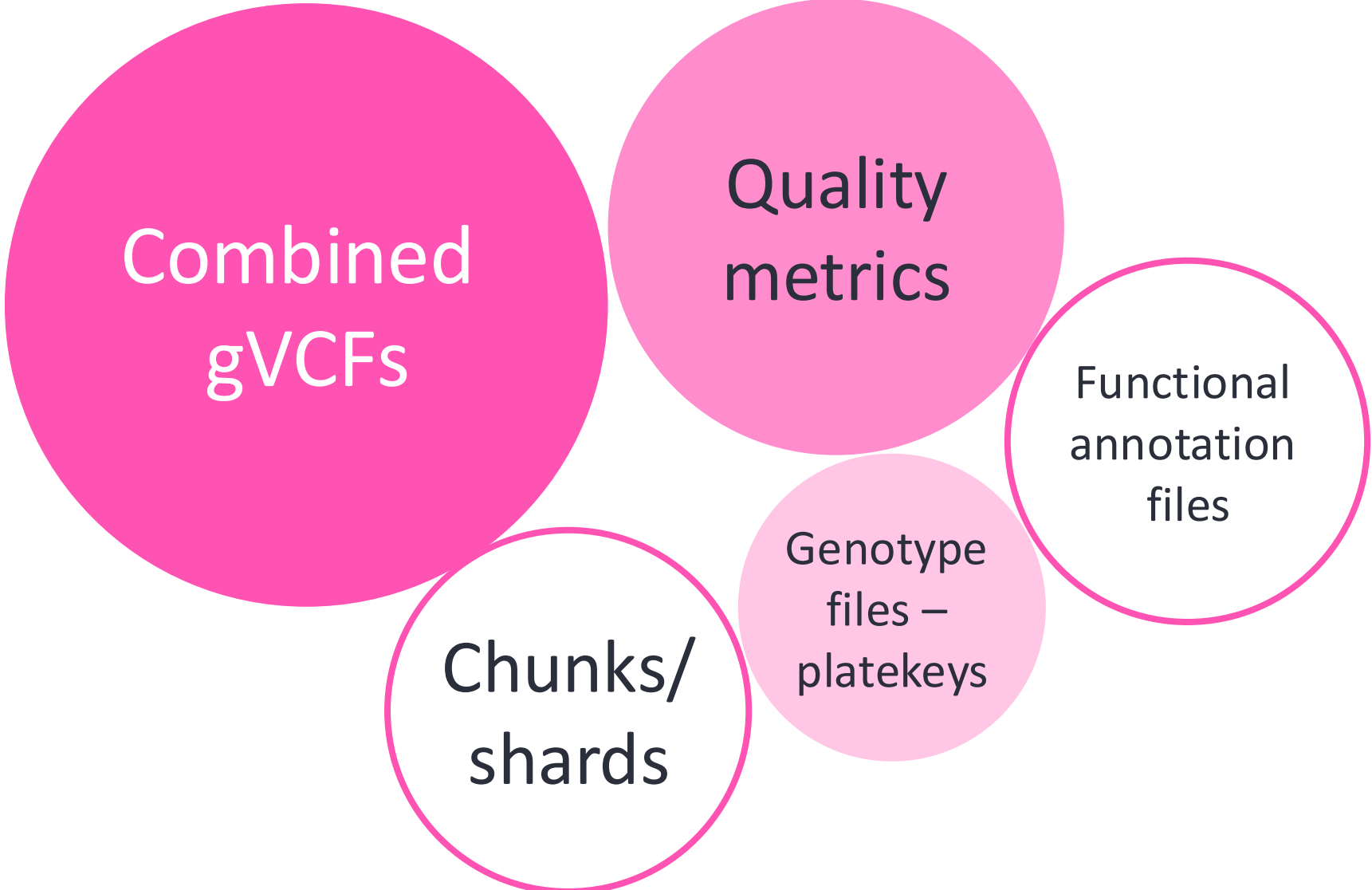
Search Criteria Result 49893 Compare Participant Download

Participant 111000003 #1 of 49893

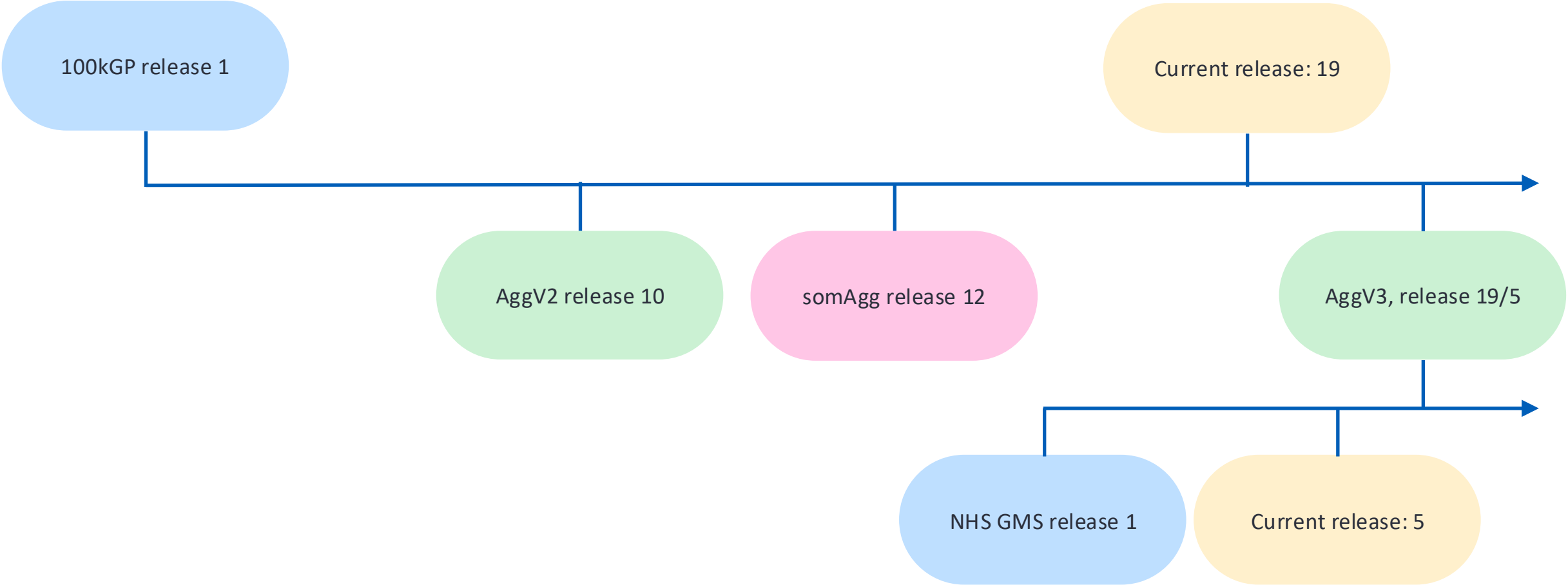
Family Member	Uncertain	Female	1242691
Family Member	Uncertain	Male	1242693

7. Using your cohort with aggregate VCFs

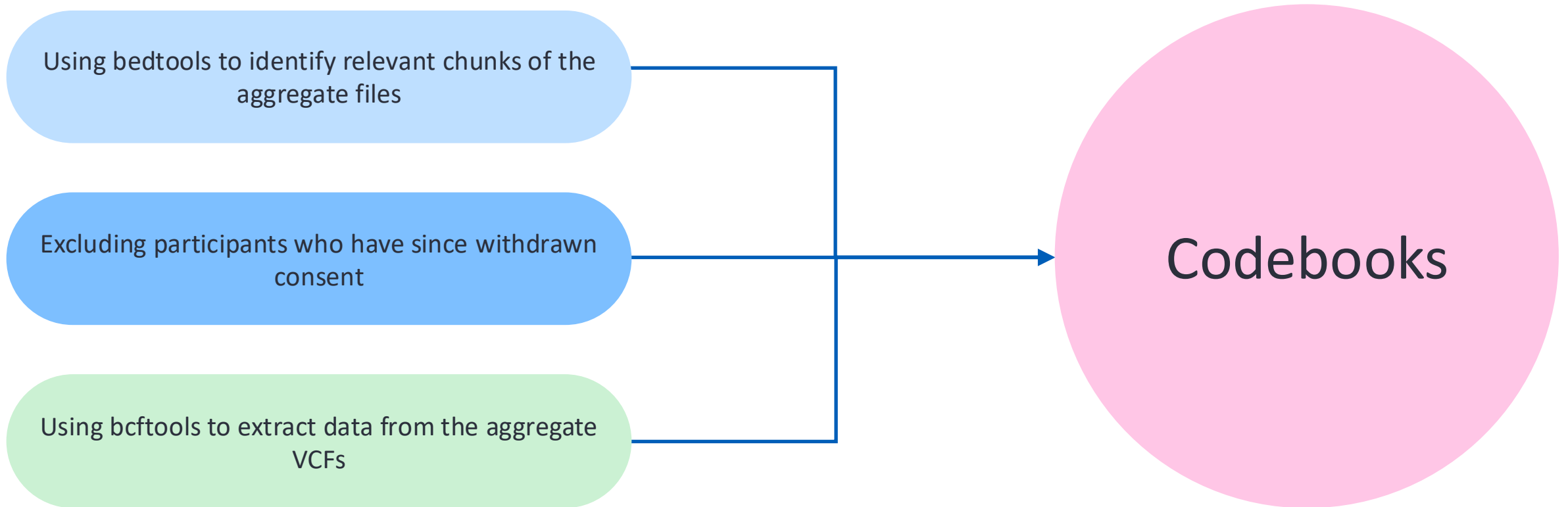
Aggregate VCFs



Aggregate VCFs



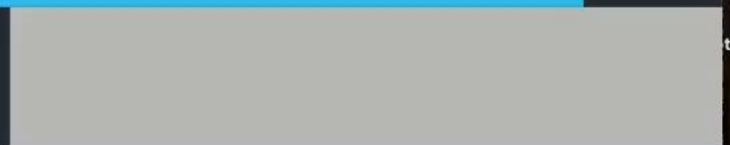
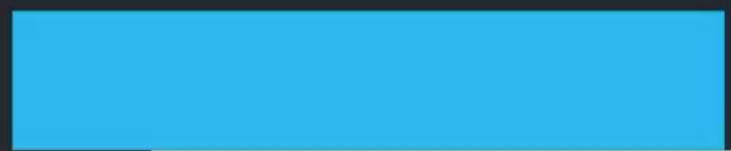
Aggregate VCFs



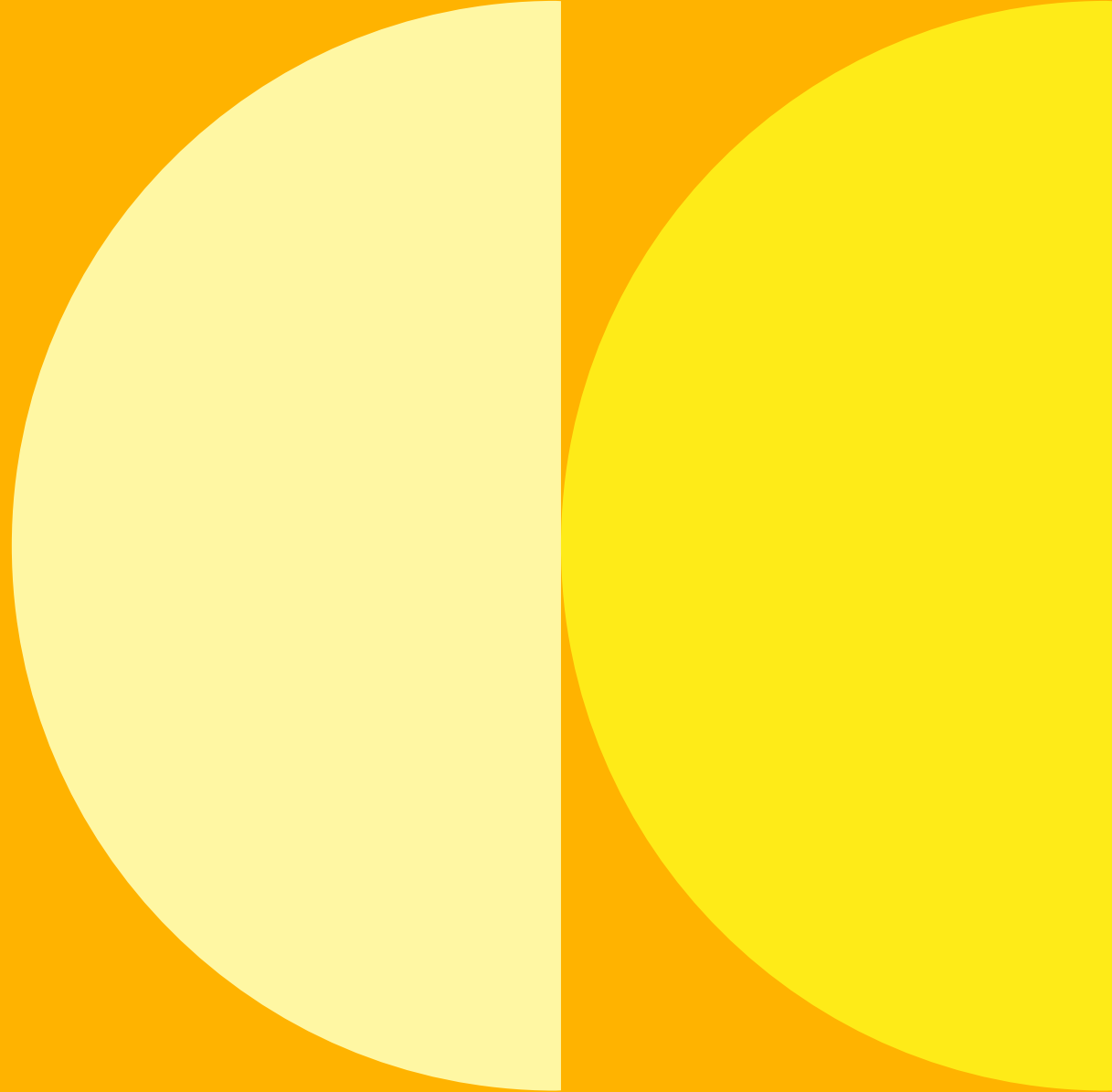
Aggregate demo



- Computer
- Emacs
- Participant Explorer
- eperry's Home
- Ensembl
- R
- Link to emily
- Firefox
- RE Messages
- Old Firefox Data
- Git GUI
- Research Environment Documentation
- Airlock
- GVim
- Research Registry
- CloudOS Academic
- IGV Browser
- RStudio
- CloudOS Discovery Forum
- IVA
- Terminal Emulator
- CloudOS Internal
- Labkey
- Text Editor
- Desktop.Rproj
- LibreOffice 7.6
- Visual Studio Code
- Document Viewer
- Open Targets
- Welcome Pack
- Dremio
- Panel App
- Trash



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>

Training sessions

3rd Tuesday every month

Introduction to the RE

19/5



Materials from
past training
all online

Training sessions

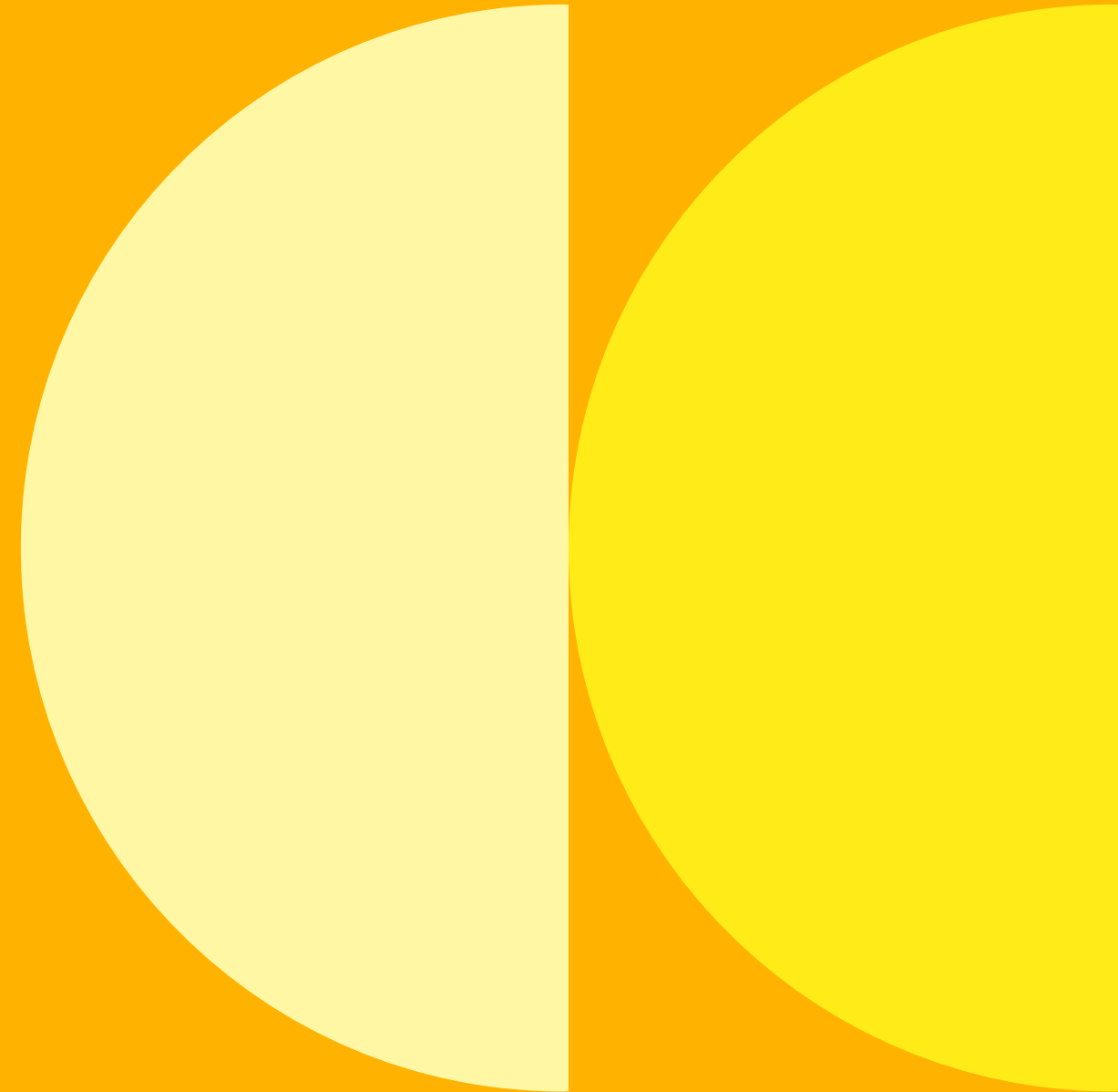
9/6

Finding participants by genotype



Materials from
past training
all online

Feedback



Thank you

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