Finding participants based on genotypes

Emily Perry

8th July 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

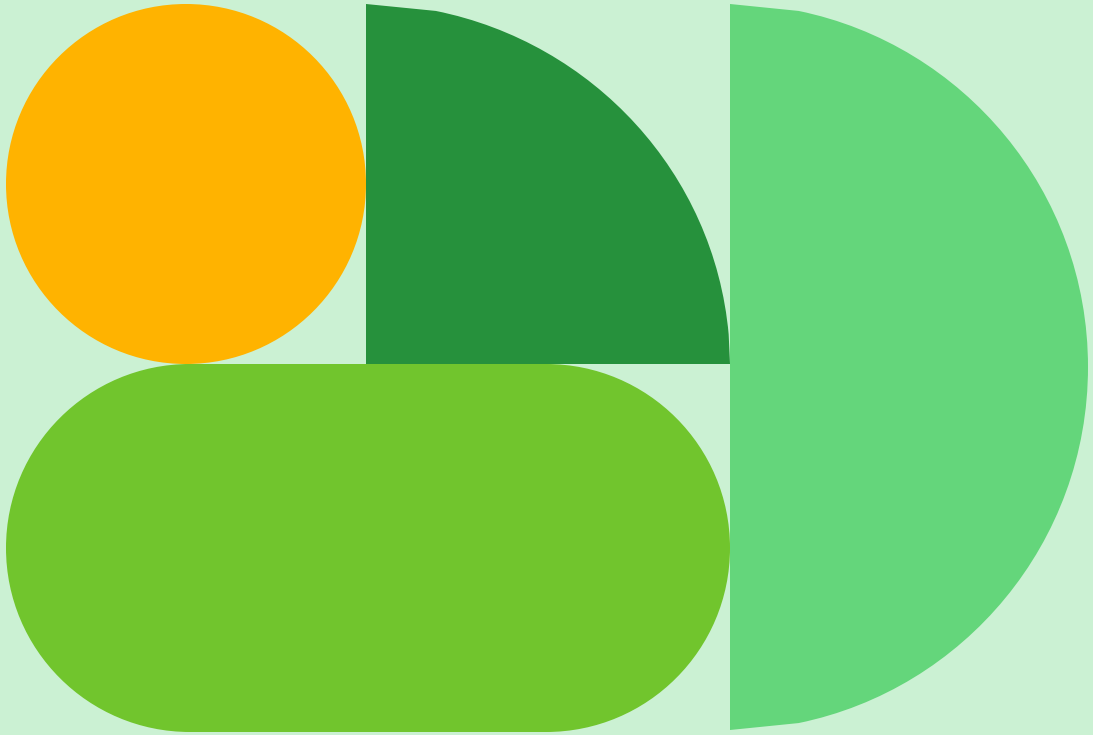
Upvote your favourite questions: if we are short on time we will prioritise those with the most votes

Questions

Magdalena Drożdż
Bioinformatician - Research Services

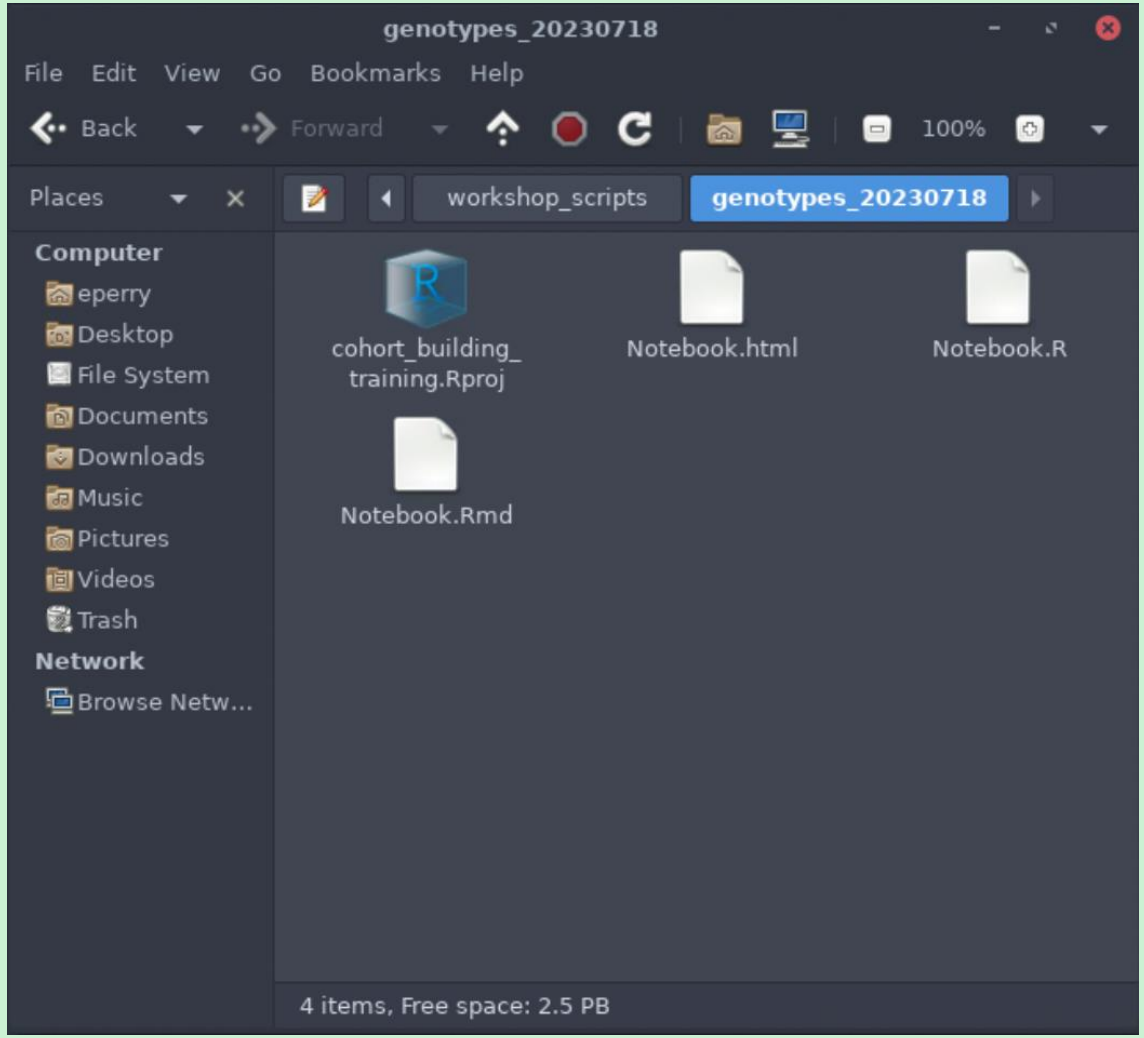
Agenda

1. Introduction and admin
2. Genome assembly
3. LabKey tables of variant genotypes
4. Finding genotypes with IVA and Cohort Browser
5. The Small Variant and Structural Variant workflows
6. Aggregated variant files
7. When/why you would use each method
8. 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/genotypes_2025`



2. Genome Assembly

100,000 Genomes Project

Rare disease

GRCh37 (hg19)
GRCh38 (hg38)

Cancer

Germline GRCh37 (hg19)
Somatic GRCh37 (hg19)
Germline GRCh38 (hg38)
Somatic GRCh38 (hg38)

100,000 Genomes Project

Rare disease

GRCh37 (hg19)

GRCh38 (hg38)

Cancer

~~Germline GRCh37 (hg19)~~

~~Somatic GRCh37 (hg19)~~

Germline GRCh38 (hg38)

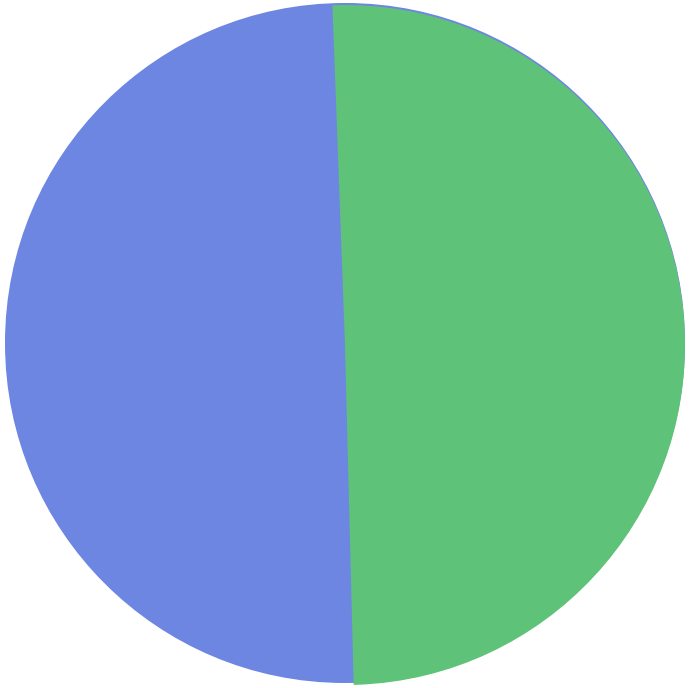
Somatic GRCh38 (hg38)

NHS GMS

Rare disease

GRCh38 (hg38)

Cancer



Germline GRCh38 (hg38)

Somatic GRCh38 (hg38)

Genome assembly coordinates

	GRCh37 (hg19)	GRCh38 (hg38)
ZAR1L	13:32,877,837-32,889,481	chr13:32,303,699- 32,315,363
ENST00000345108.6:c.931T>C	13:32,878,051	chr13:32,303,914

Converting between assemblies

Inside the RE

- Liftover tool on HPC
- Chain files in `public_data_resources`

Outside the RE

- Ensembl Assembly converter
- UCSC Liftover

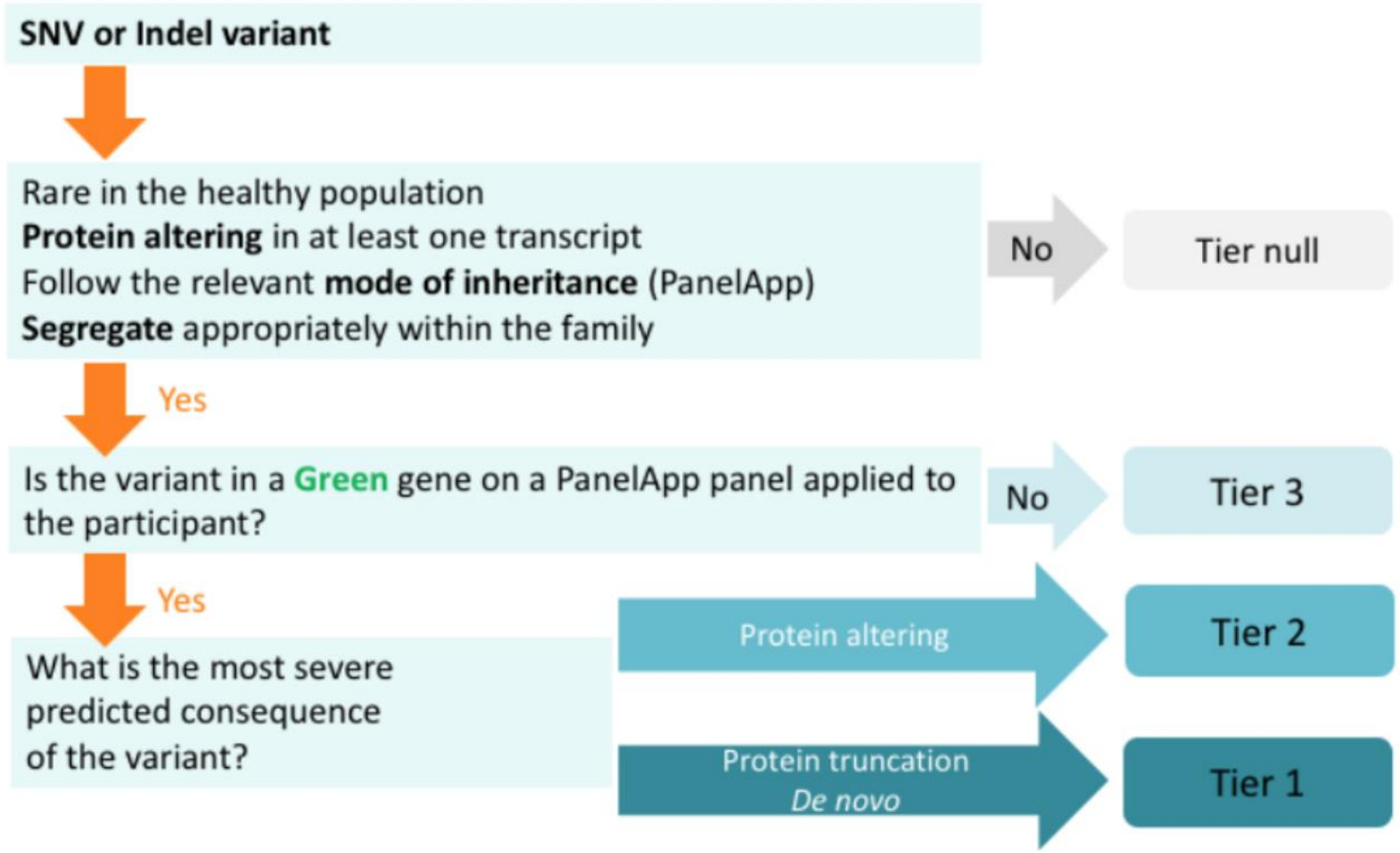
3. LabKey tables of variant genotypes

LabKey

- Participant details and family relationships
- Sample details
- Genomic file locations
- Clinical data
- Bioinformatics analysis results
  - Rare disease tiering
  - Cancer tiering
  - Exomiser

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

Rare disease tiering



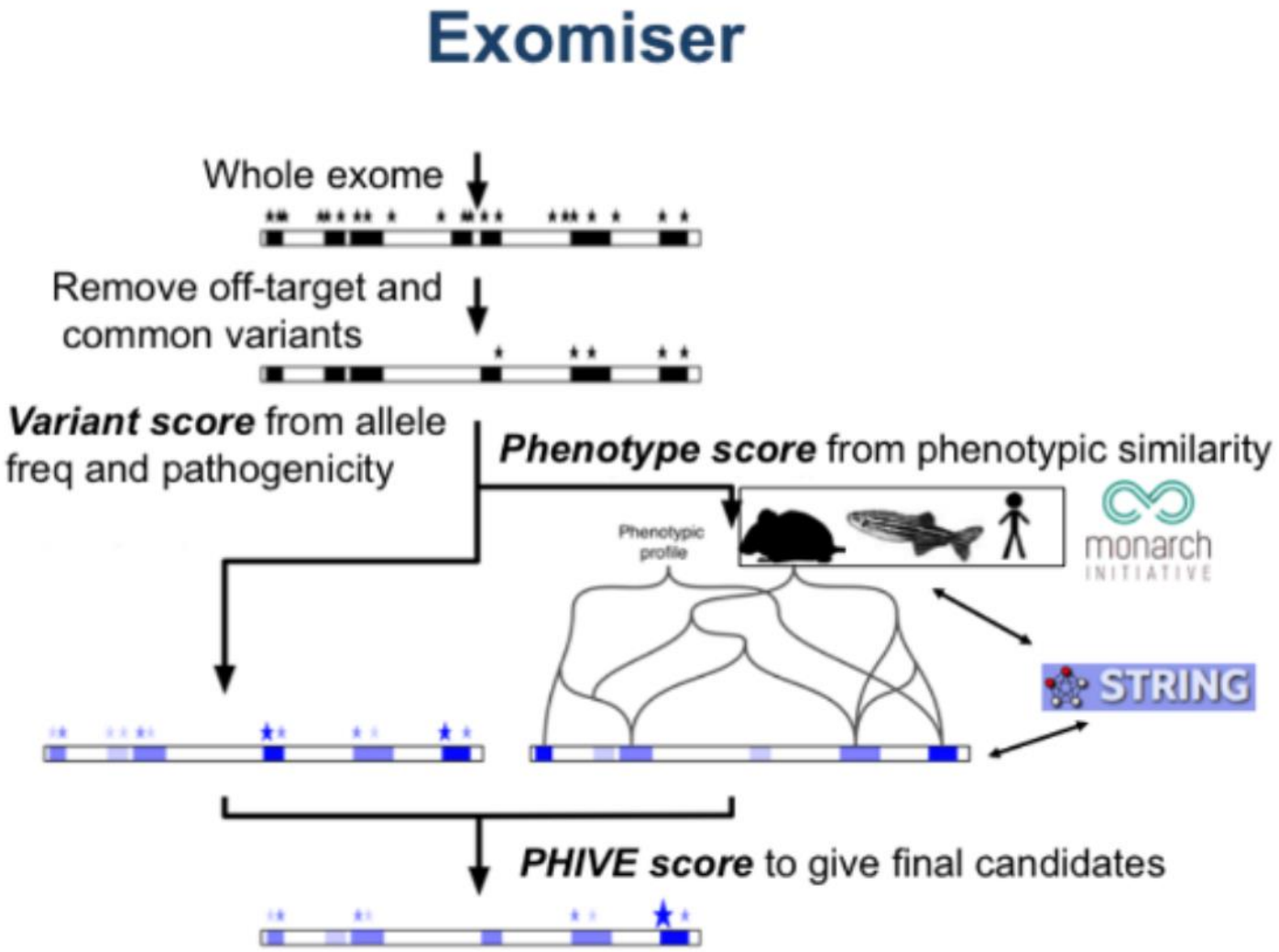
https://re-docs.genomicsengland.co.uk/tiering/

Rare disease tiering based on PanelApp genes

List ↑	Entity	Reviews	Mode of inheritance	Details
Filter Entities				8 Entities
Green	ATP1A3	1 review 1 green	MONOALLELIC, autosomal or pseudoautosomal, NOT imprinted	**Sources** • Expert Review • Expert Review Green **Phenotypes** • 601338 • 614820 **Tags**
Green	DFNB59	2 reviews 1 green	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review • Expert Review Green **Phenotypes** • 610219 **Tags** new-gene-name
Green	OPA1	2 reviews 1 green	MONOALLELIC, autosomal or pseudoautosomal, NOT imprinted	**Sources** • Eligibility statement prior genetic testing • Expert Review Green **Phenotypes** • Optic atrophy 1, OMIM:165500 • Optic atrophy plus syndrome, OMIM:125250 **Tags**
Green	OTOF	1 review 1 green	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Green • Radboud University Medical Center, Nijmegen **Phenotypes** • 601071 **Tags**
Amber	DIAPH3	3 reviews 1 red	BOTH monoallelic and biallelic, autosomal or pseudoautosomal	**Sources** • Expert Review Amber • Radboud University Medical Center, Nijmegen **Phenotypes** • Auditory neuropathy, autosomal dominant, 1, 609129 **Tags**

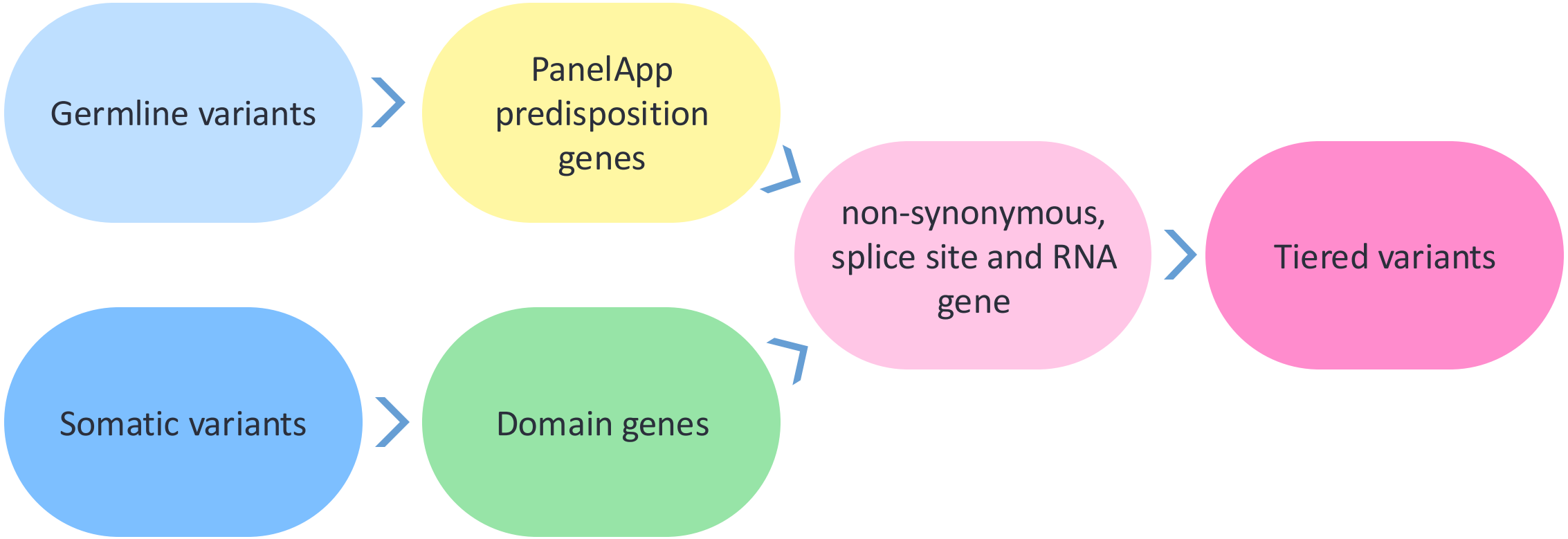
https://re-docs.genomicsengland.co.uk/panelapp/

Rare disease Exomiser



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

Cancer tiering



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

100k tiering

Rare disease

GRCh37 (hg19)

GRCh38 (hg38)

Cancer

Germline GRCh38 (hg38)

Somatic GRCh38 (hg38)

Tiering tables genome assembly – 100k

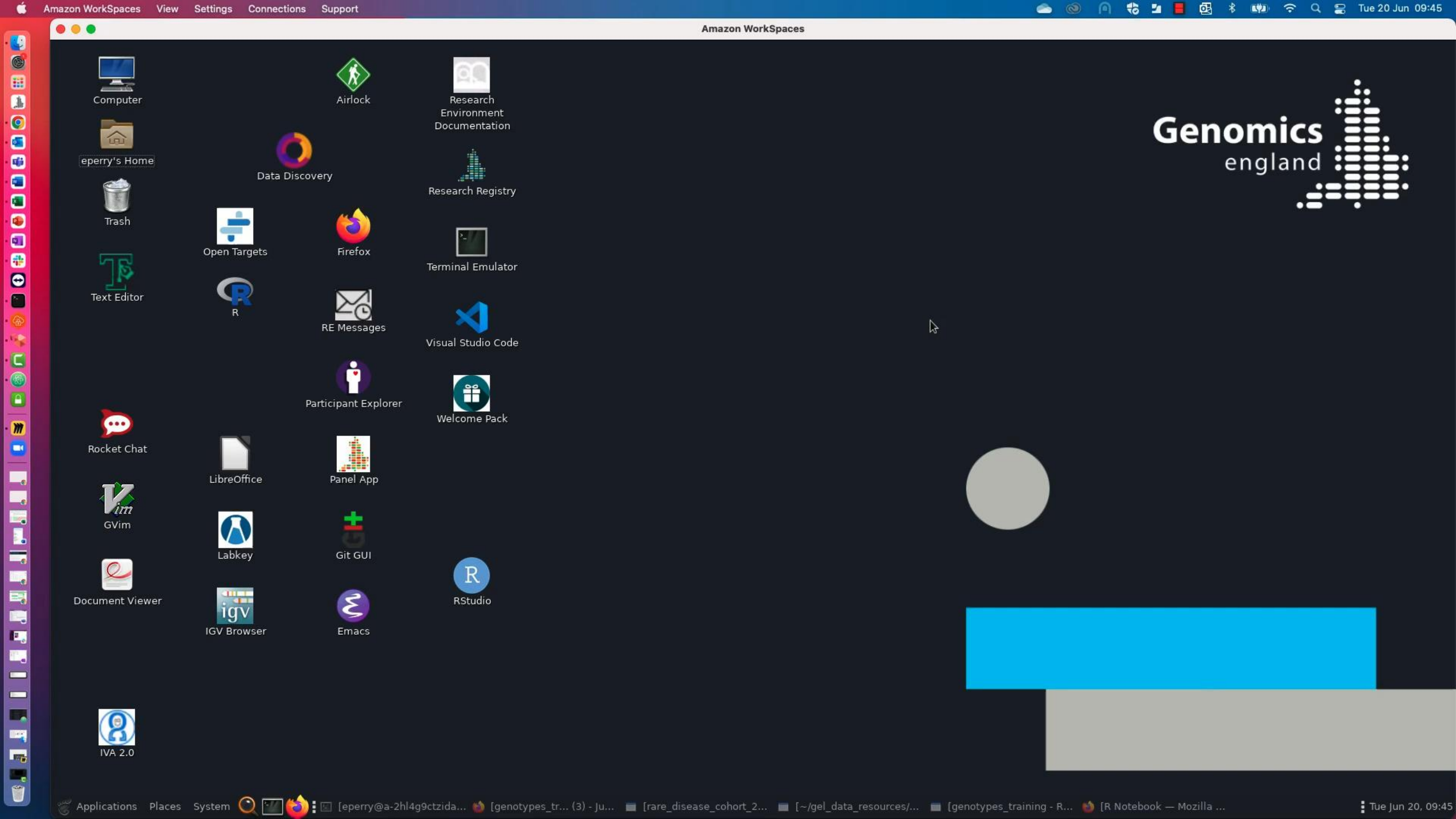
Search by	
gene	Should find all filter-passing variants in the gene on either assembly
coordinate	You must also specify the genome assembly
HGVS (exomiser only)	Should find all filter-passing variants that match your string on either assembly

NHS GMS tiering

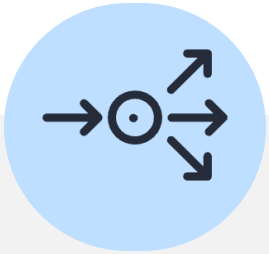
Rare disease

GRCh38 (hg38)

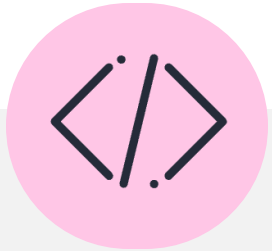
Variant data in LabKey demo


Amazon WorkSpaces View Settings Connections Support
Tue 20 Jun 09:45
Amazon WorkSpaces
Computer
Airlock
Research Environment Documentation
eperry's Home
Data Discovery
Research Registry
Trash
Open Targets
Firefox
Terminal Emulator
Text Editor
R
RE Messages
Visual Studio Code
Participant Explorer
Welcome Pack
Rocket Chat
LibreOffice
Panel App
GVim
Labkey
Git GUI
Document Viewer
RStudio
IGV Browser
Emacs
IVA 2.0
Genomics england
Applications Places System
[eperry@a-2hl4g9ctzida...
[genotypes_tr... (3) - Ju...
[rare_disease_cohort_2...
[~/gel_data_resources/...
[genotypes_training - R...
[R Notebook — Mozilla ...
Tue Jun 20, 09:45

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/

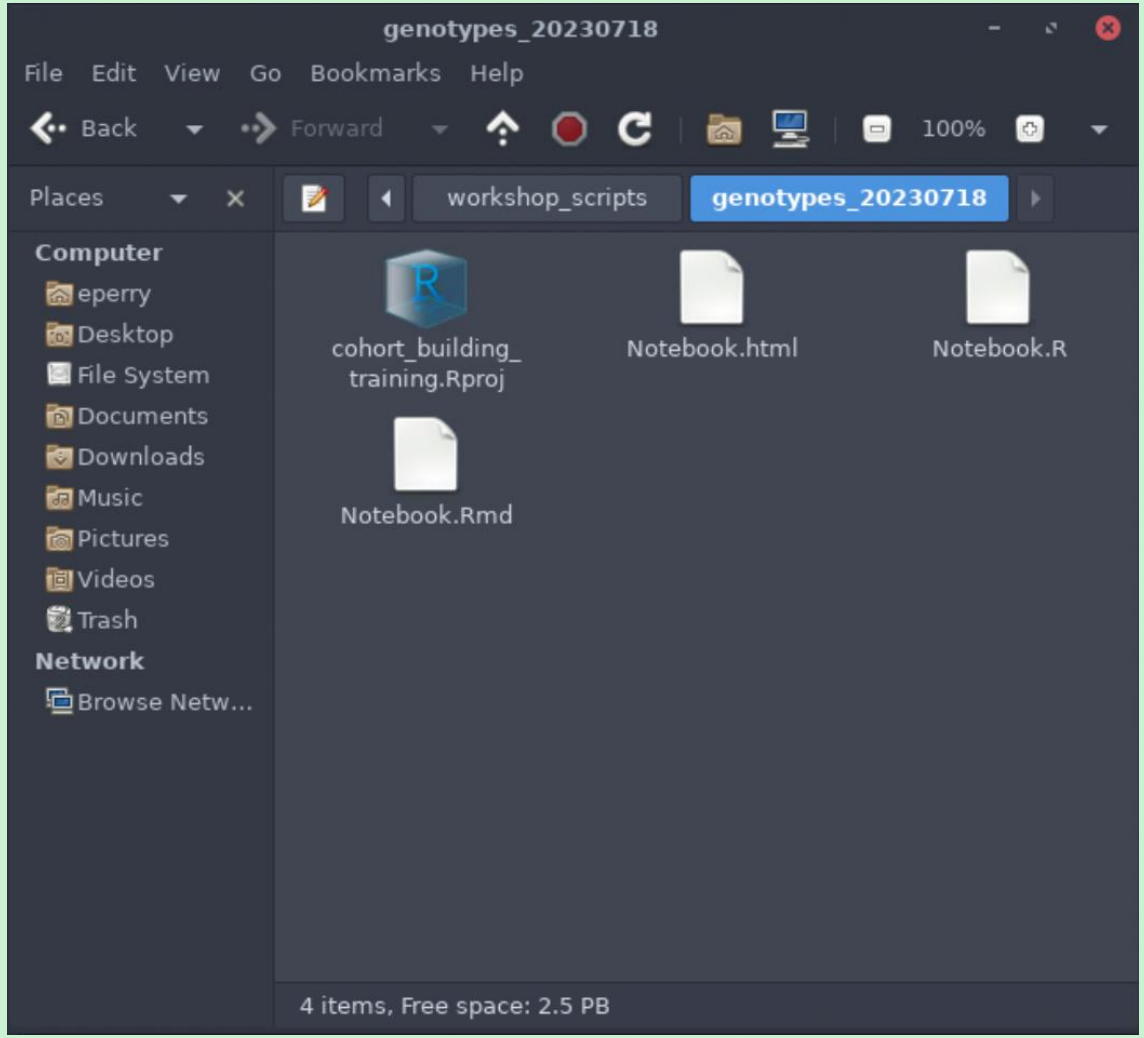
LabKey .netrc

- 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/

Materials

- Slides and video will be sent out to you after the session
- Scripts available in `/gel_data_resources/example_scripts/workshop_scripts/genotypes_2025`



Variant data in LabKey API demo

genotypes_tr... (3) - JupyterLab — Mozilla Firefox

Jupyter Lab on Helix - Ge | R Notebook | genotypes_tr... (3) - Jupy

127.0.0.1:9537/lab/tree/genotypes_training.ipynb

File Edit View Run Kernel Tabs Settings Help

GENOTYPES_TRAINING.IPYNB

rd_cohort_building_trainir | genotypes_training.ipynb | cohort_building_training.ip

Markdown | Python 3 (ipykernel)

Finding participants by genotypes in Python

This notebook will walk you through finding participants by genotypes. You are welcome to copy/paste any code from this notebook for your own scripts/notebooks.

Contents:

Use the LabKey API

Import Python modules you need

[32]:
```
import numpy as np
import functools
import labkey
import pandas as pd
```

Helper function to access the LabKey API with

Simple 0 3 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 genotypes_training.ipynb

R Notebook — Mozilla Firefox

R Notebook

file:///nas/weka.gel.zone/pgen_int_work/BRS/emily/genotypes_2023/genotyp

R Notebook

Finding participants by genotypes in R

This notebook will walk you through finding participants by genotypes. You are welcome to copy/paste any code from this notebook for your own scripts/notebooks.

Contents:

Use the LabKey API

Import R libraries you need

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(Rlabkey)
```

4. Finding genotypes with IVA and Cohort Browser

Interactive variant analysis (IVA)

- Point-and-click interface to explore variants
- Filter by loci, consequences, population frequencies and inheritance
- Find participant genotypes

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

100k in IVA

Rare disease

GRCh37 (hg19)

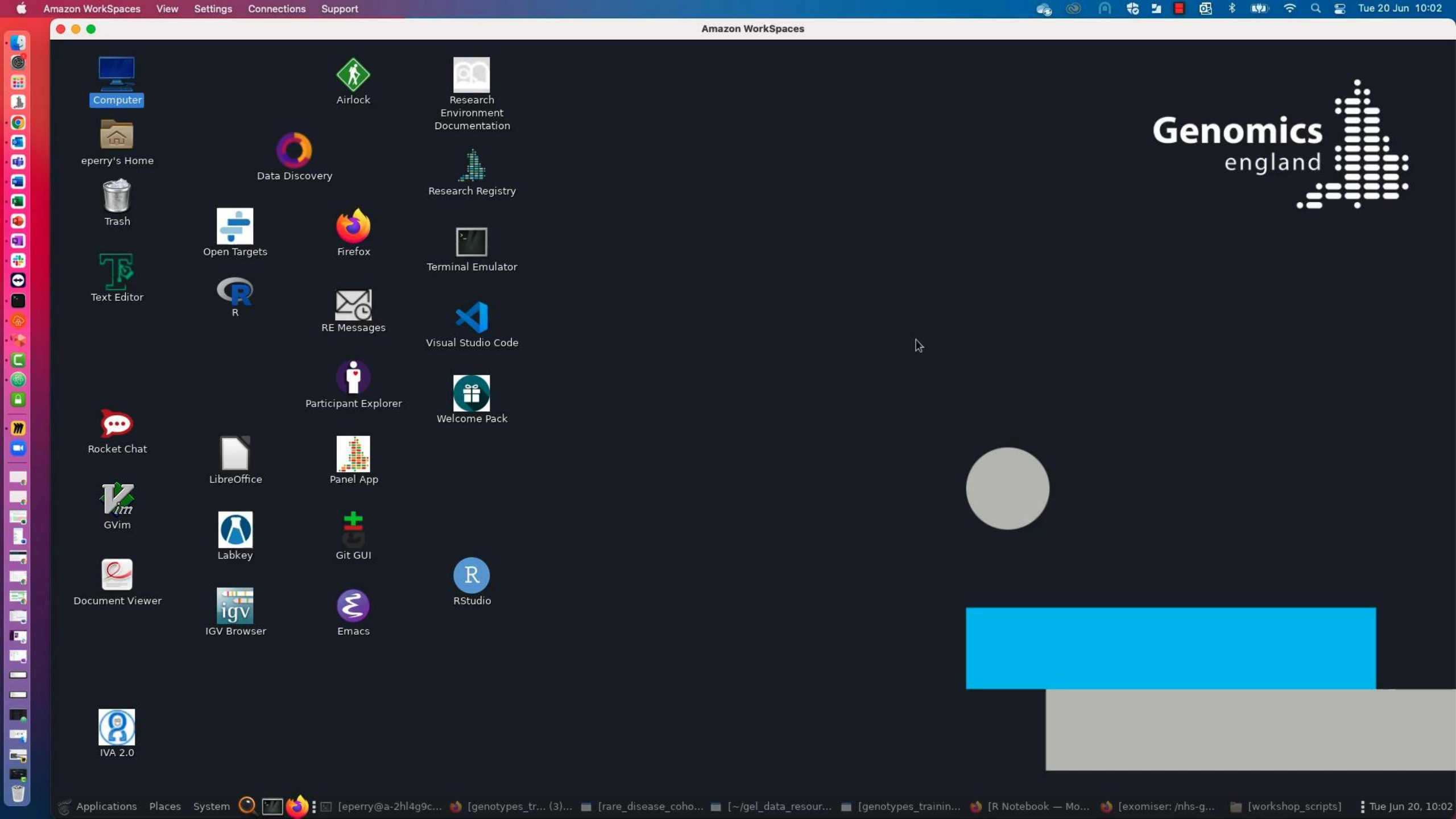
GRCh38 (hg38)

Cancer

Germline GRCh38 (hg38)

Somatic GRCh38 (hg38)

IVA demo


Amazon WorkSpaces View Settings Connections Support
Tue 20 Jun 10:02
Amazon WorkSpaces
Computer
Airlock
Research Environment Documentation
eperry's Home
Data Discovery
Research Registry
Trash
Open Targets
Firefox
Terminal Emulator
Text Editor
R
RE Messages
Visual Studio Code
Participant Explorer
Welcome Pack
Rocket Chat
LibreOffice
Panel App
GVim
Labkey
Git GUI
Document Viewer
RStudio
IGV Browser
Emacs
IVA 2.0
Genomics england
Applications Places System
[eperry@a-2hl4g9c...
[genotypes_tr... (3)...
[rare_disease_coho...
[~/gel_data_resour...
[genotypes_trainin...
[R Notebook — Mo...
[exomiser: /nhs-g...
[workshop_scripts]
Tue Jun 20, 10:02

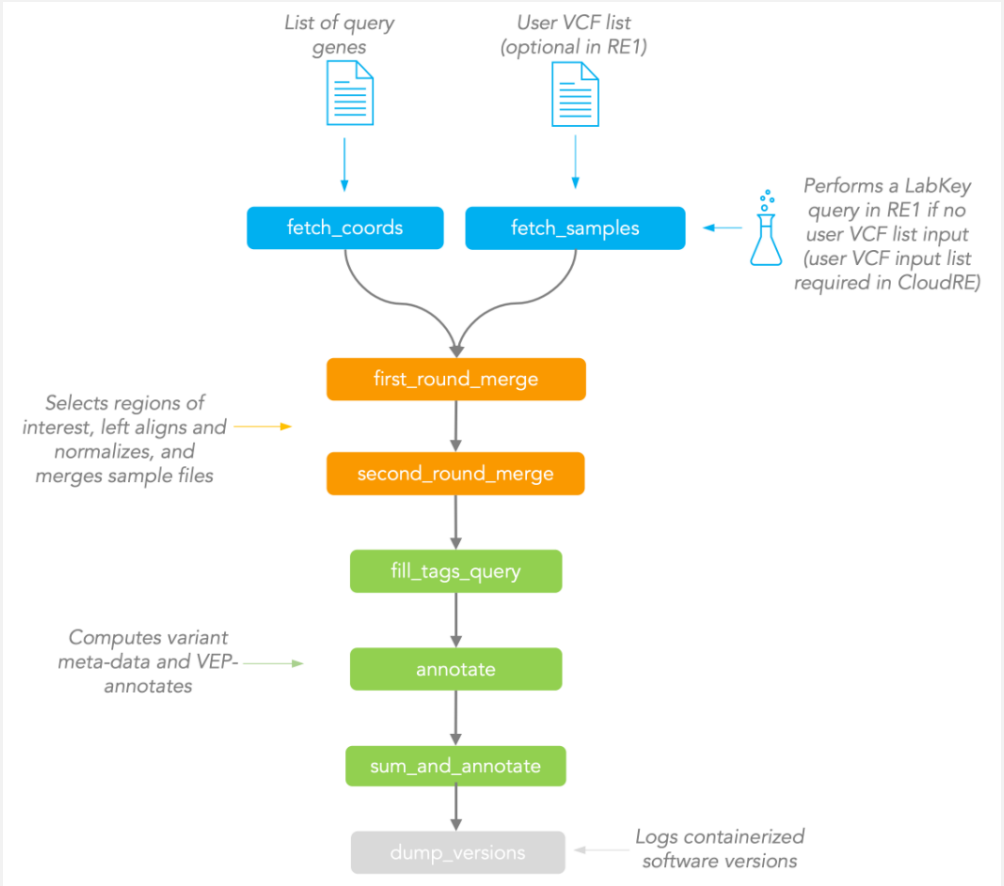
5. The Small Variant and Structural Variant workflows

Small variant workflow

Submit a list of genes

Find all short variants in these genes

Get 100k participants with these variants



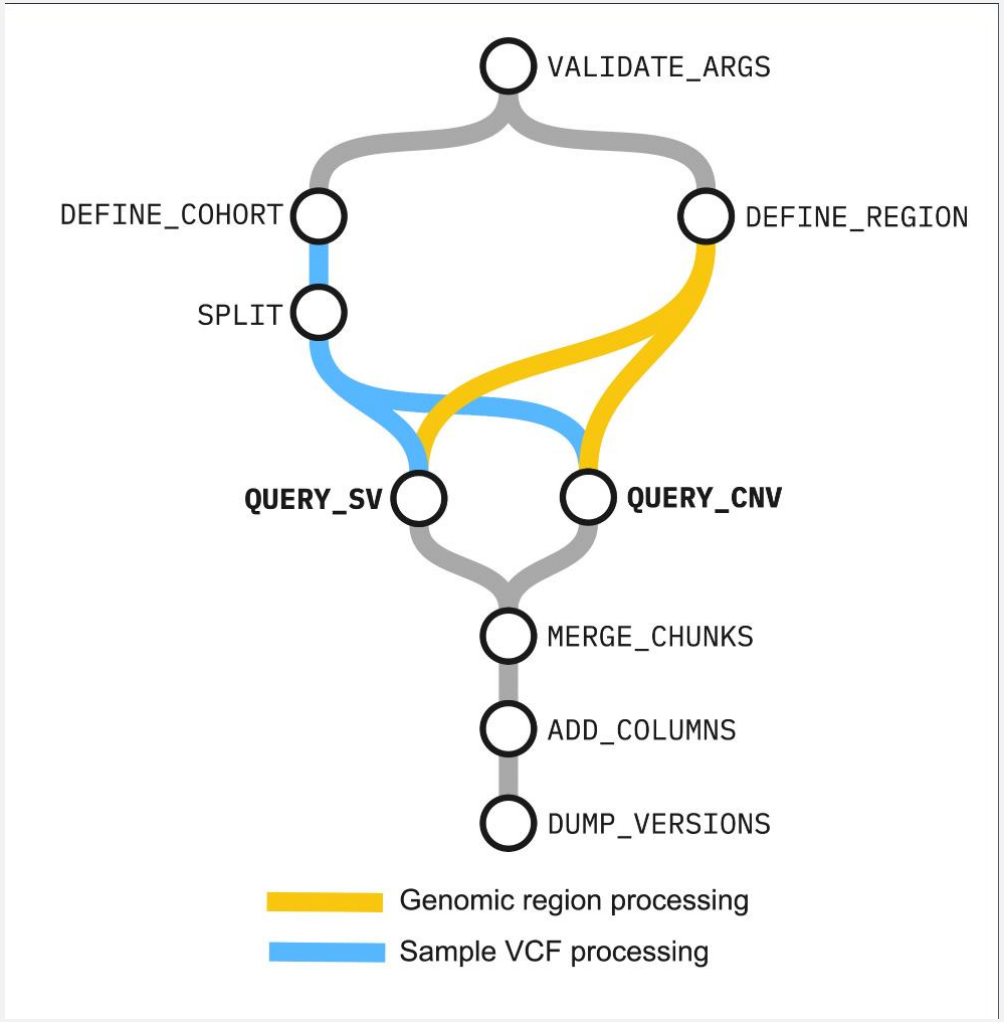
https://re-docs.genomicsengland.co.uk/small_variant/

Structural variant workflow

Submit a list of genes or regions

Find all structural variants overlapping these genes

Get 100k participants with these variants



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

Workflows genome assembly – 100k

Search by	
gene	Should find all variants in the gene(s) on either assembly
coordinates (structural only)	You must also specify the genome assembly

Running workflows on the HPC demo

6. Aggregate variant files

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/

AggV2 – germline samples

Rare disease

GRCh37 (hg19)

GRCh38 (hg38)

Cancer

Germline GRCh38 (hg38)

Somatic GRCh38 (hg38)

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

somAgg

Rare disease

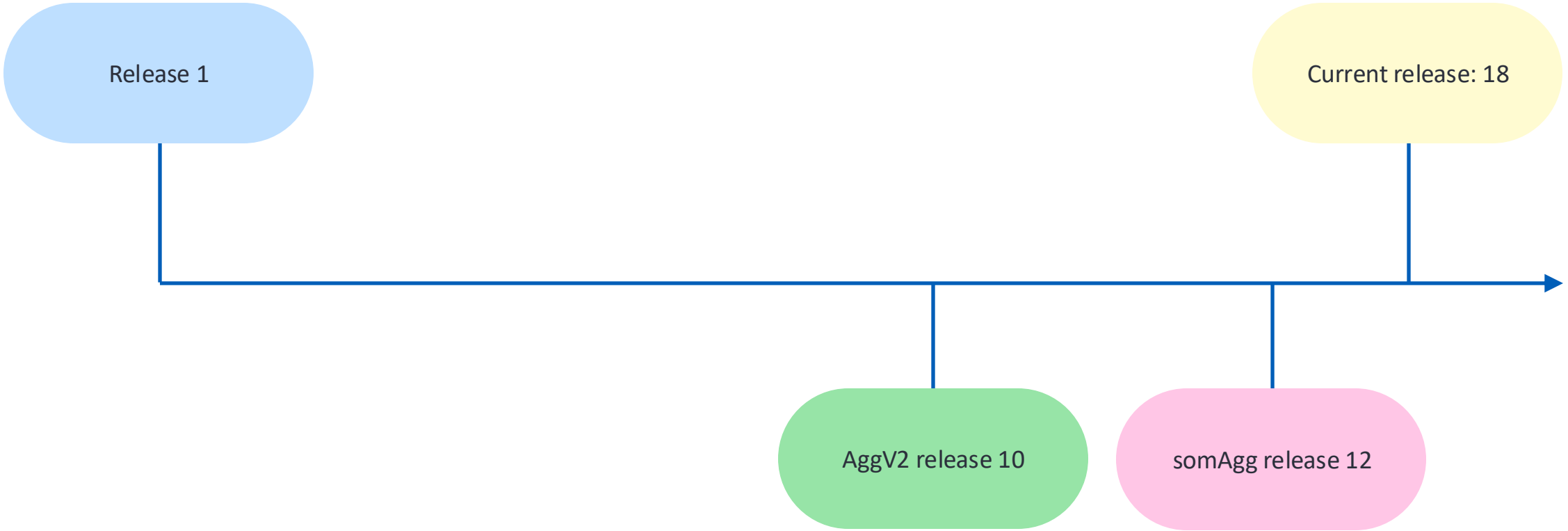
Cancer

GRCh37 (hg19)
GRCh38 (hg38)

Germline GRCh38 (hg38)
Somatic GRCh38 (hg38)

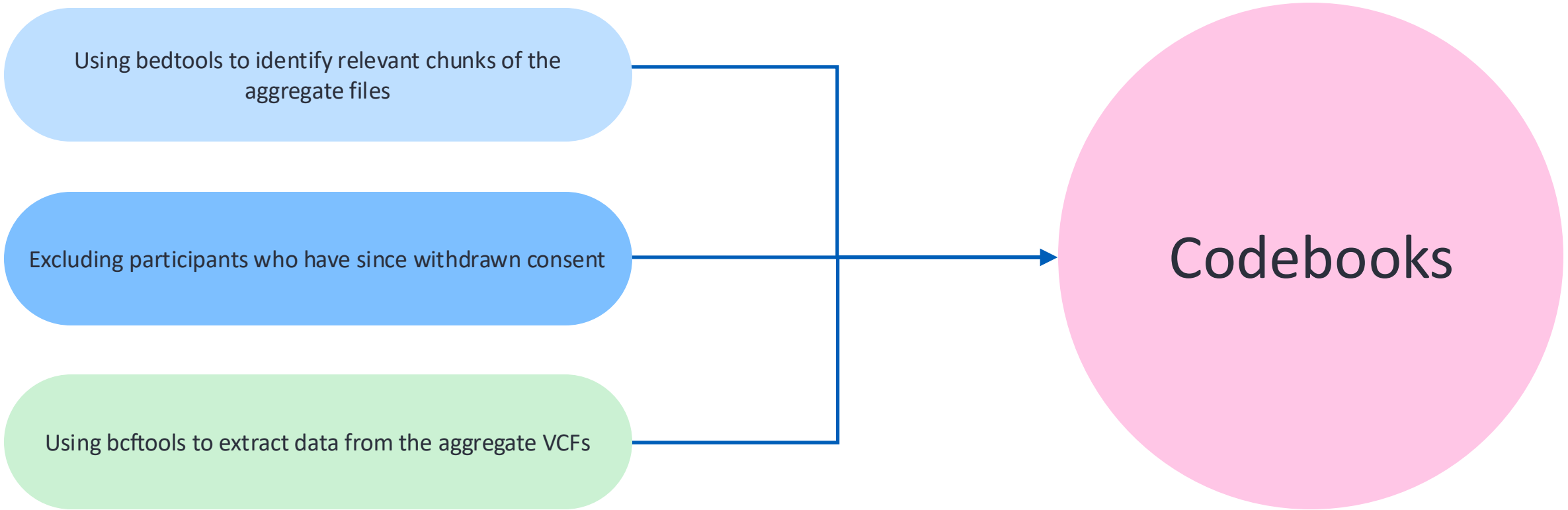
https://re-docs.genomicsengland.co.uk/somAgg/

Aggregate VCFs



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

Aggregate VCFs



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

Aggregate VCF chunks

- Locus-based queries must query the correct chunk file
- BED file of chunks available
- Create a sorted BED file of your own regions
- Intersect with BEDtools
- Code books with more information
- Also available in Plink2 format

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

Using bcftools on the HPC demo

7. When/why you would use each method

Platform

Tiering and exomiser tables

LabKey®

IVA

INTERACTIVE VARIANT ANALYSIS

Small/ Structural variant workflows

Querying the aggregates

Search by

Tiering and exomiser tables

HGVS HUMAN GENOME VARIATION SOCIETY

IVA

dbSNP Short Genetic Variations

Small/ Structural variant workflows

Querying the aggregates

Variants included

Tiering and exomiser tables

Variants that have passed filters

IVA

All

Small/ Structural variant workflows

All

Querying the aggregates

Variants from GRCh38-aligned genomes from release 8 or 12

Datasets

Tiering and exomiser tables

100k and NHS GMS

IVA and Cohort browser omics

100k

Small/ Structural variant workflows

100k

Querying the aggregates

100k

Genome assembly

Tiering and exomiser tables

GRCh37 and GRCh38
Assembly as a separate column

IVA

GRCh37 and GRCh38 in separate databases

Small/ Structural variant workflows

GRCh37 and GRCh38 queries simultaneously

Querying the aggregates

GRCh38 only

Underlying VCFs

Tiering and exomiser tables

Rare disease: Platypus
Cancer: Strelka

IVA

Rare disease: Platypus
Cancer: Strelka

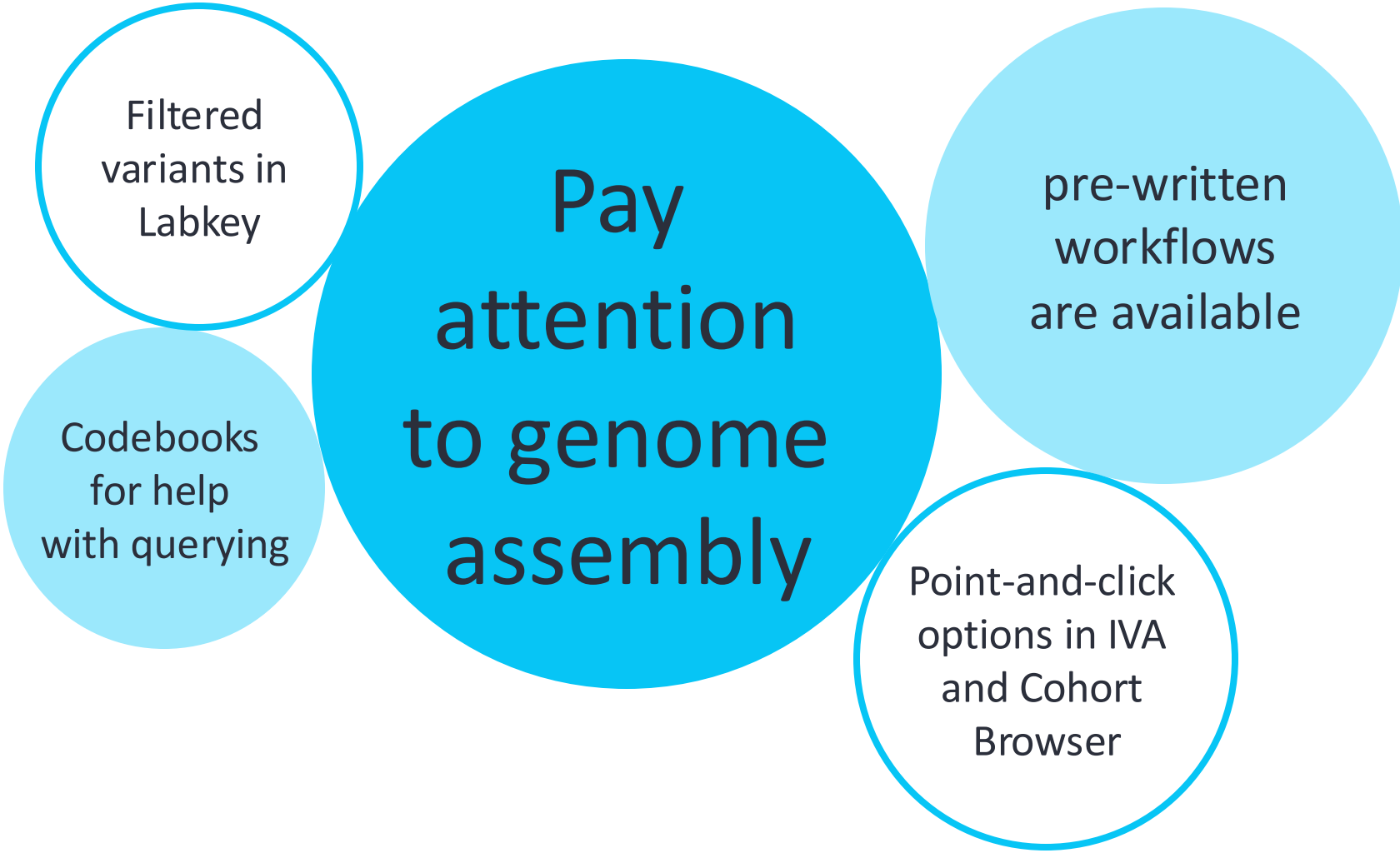
Small/ Structural variant workflows

Strelka

Querying the aggregates

Strelka

Key takeways



8. 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

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

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/