Genomics England

Finding participants based on genotypes

Emily Perry

Research Engagement Manager

19th July 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

Ronnie Rodrigues Pereira
Bioinformatician -
Research Services

Alex Stuckey
Senior Bioinformatician -
Research Services

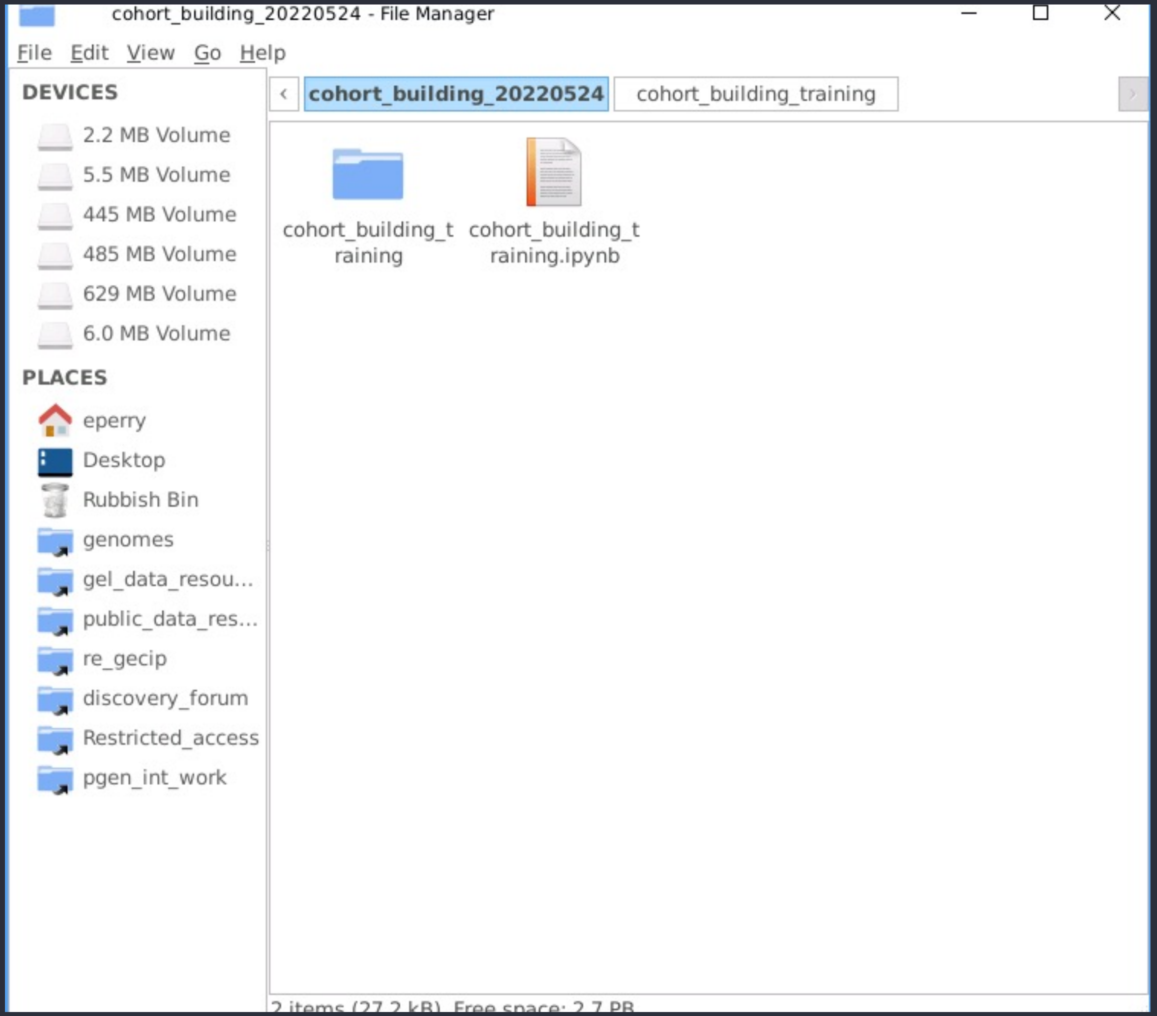
Christian Bouwens
Bioinformatician -
Research Services

Agenda

1. Introduction and admin
2. LabKey tables of variant genotypes
3. Finding genotypes with IVA
4. The Gene-Variant and SV/CNV workflows
5. Aggregated variant files
6. Using bcftools on the HPC
7. Getting help and questions

Materials

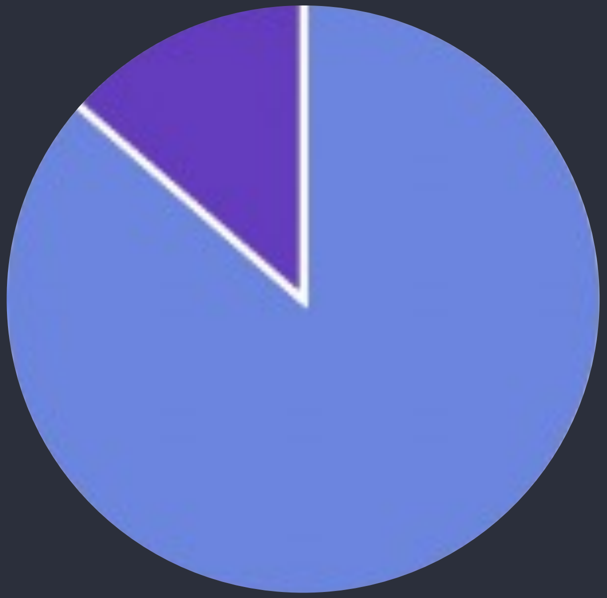
- Slides and video will be sent out to you after the session
- Scripts available in /gel_data_resources/example_scripts/workshop_scripts/cohort_building_20220524



2 items (27.2 kB), Free space: 2.7 PB

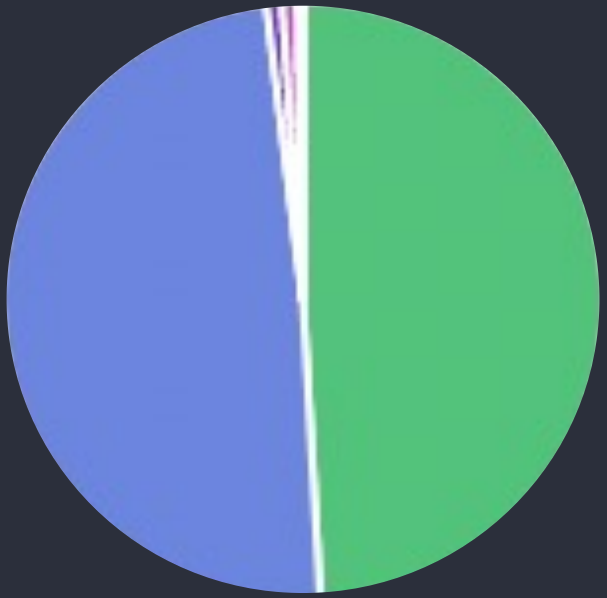
Genome assembly

Rare disease



GRCh38 (aka hg38)
GRCh37 (aka hg19)

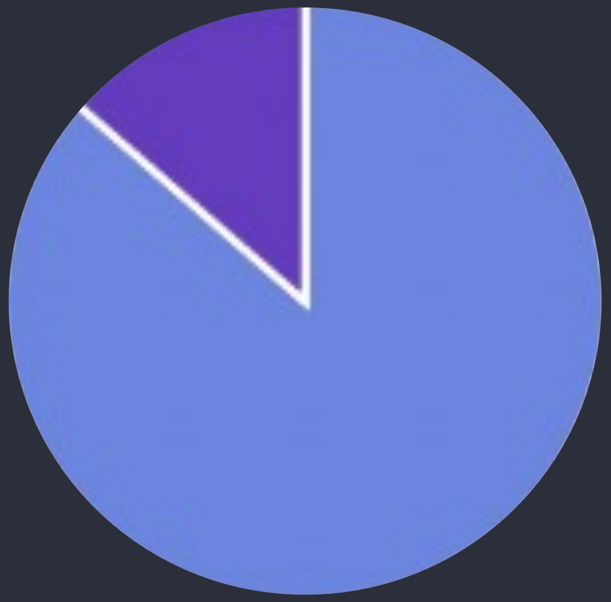
Cancer



Somatic GRCh38
Germline GRCh38
Germline GRCh37
Somatic GRCh37

Genome assembly

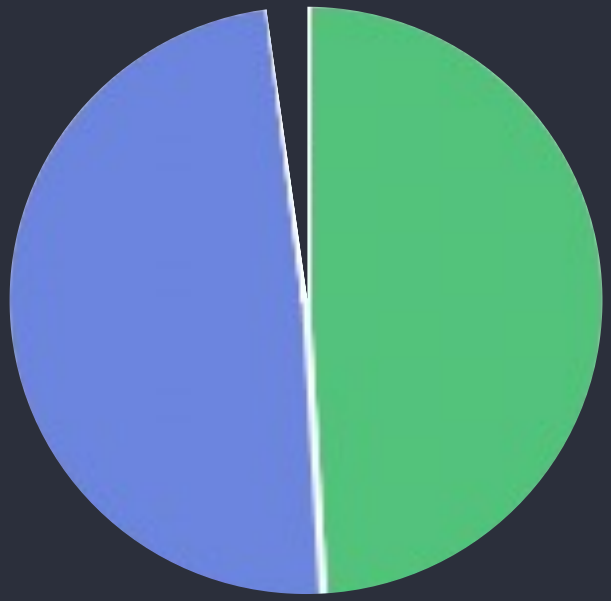
Rare disease



GRCh38 (aka hg38)

GRCh37 (aka hg19)

Cancer



Somatic GRCh38

Germline GRCh38

~~Germline GRCh37~~

~~Somatic GRCh37~~

Genome assembly

chr13 ZAR1L ENST00000345108.6:c.931T>C

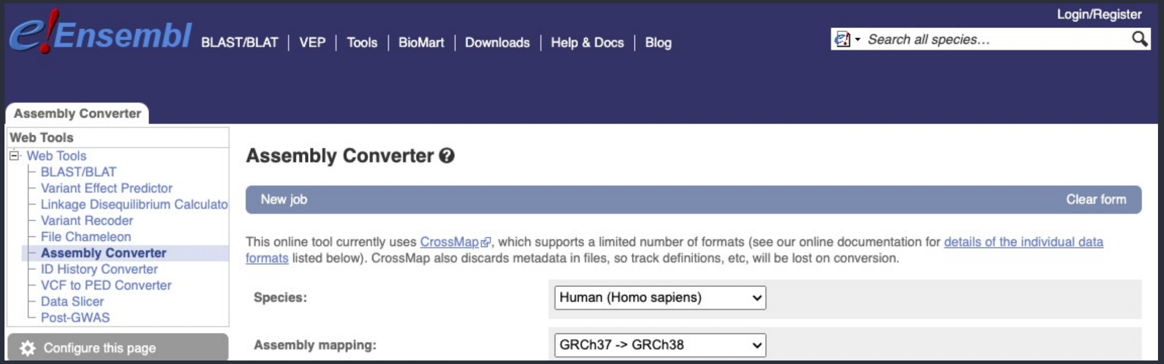
	GRCh37 (hg19)	GRCh38 (hg38)
ZAR1L	13:32,877,837-32,889,481	chr13:32,303,699-32,315,363
ENST00000345108.6:c.931T>C	13:32,878,051	chr13:32,303,914

Converting coordinates between assemblies

Inside the RE:

- Liftover tool on HPC
- Chain files in public_data_resources

Outside the RE:



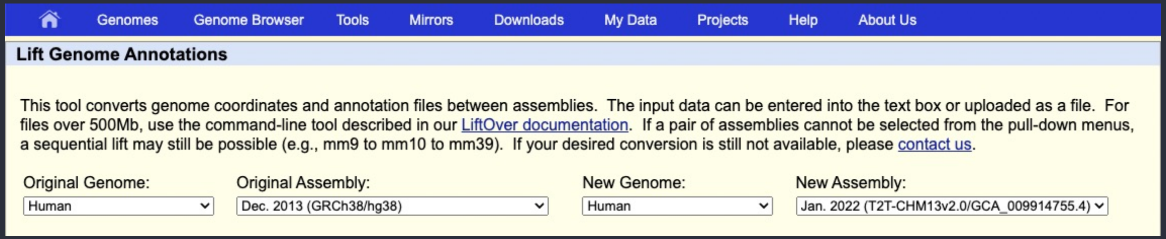
Assembly Converter

This online tool currently uses CrossMap, which supports a limited number of formats (see our online documentation for details of the individual data formats listed below). CrossMap also discards metadata in files, so track definitions, etc, will be lost on conversion.

Species: Human (Homo sapiens)

Assembly mapping: GRCh37 -> GRCh38

https://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter



Lift Genome Annotations

This tool converts genome coordinates and annotation files between assemblies. The input data can be entered into the text box or uploaded as a file. For files over 500Mb, use the command-line tool described in our LiftOver documentation. If a pair of assemblies cannot be selected from the pull-down menus, a sequential lift may still be possible (e.g., mm9 to mm10 to mm39). If your desired conversion is still not available, please contact us.

Original Genome: Human
Original Assembly: Dec. 2013 (GRCh38/hg38)
New Genome: Human
New Assembly: Jan. 2022 (T2T-CHM13v2.0/GCA_009914755.4)

https://genome.ucsc.edu/cgi-bin/hgLiftOver

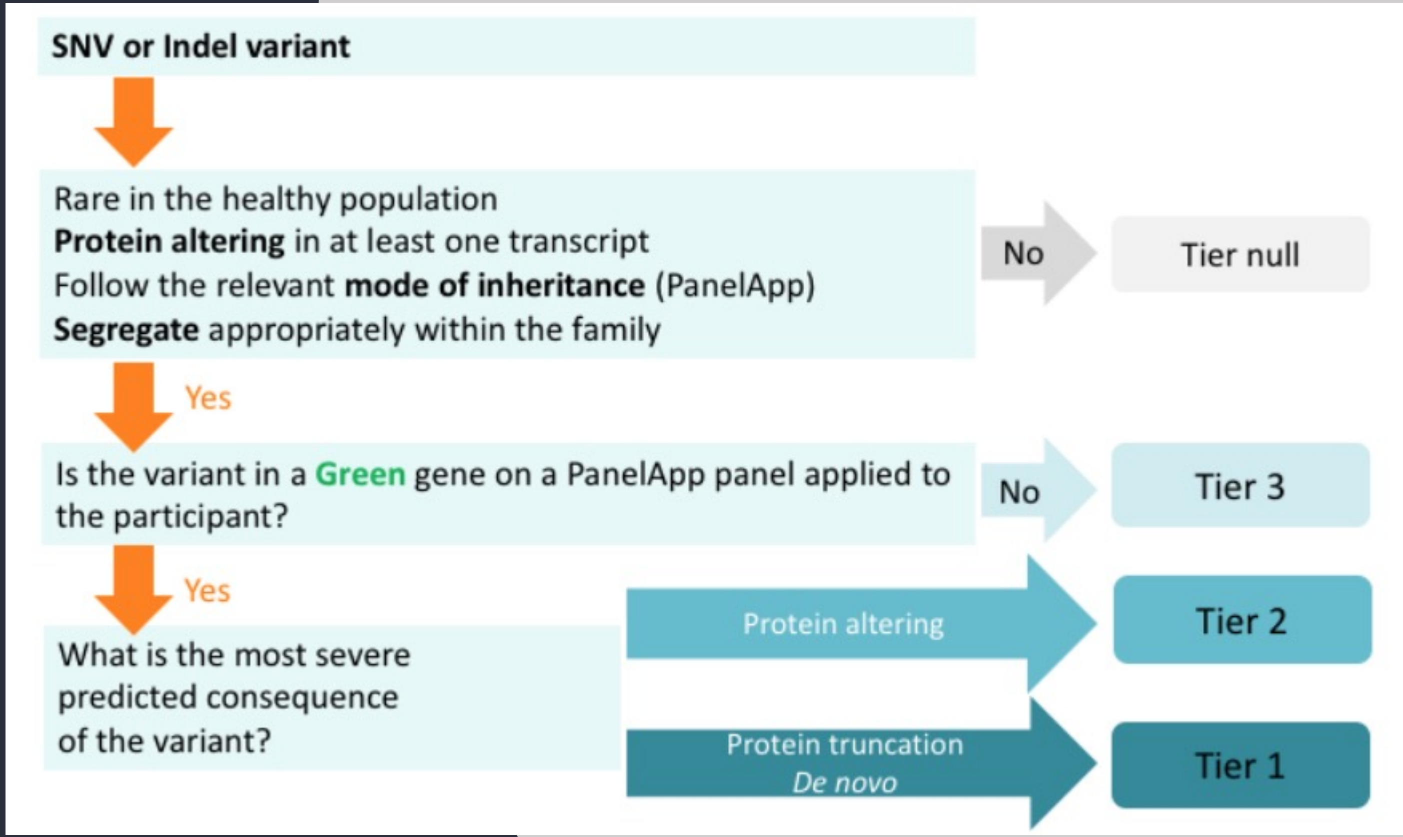
2. LabKey tables of variant genotypes

Labkey

- Participant details and family relationships
- Sample details
- Genomic file locations
- Clinical data
- Bioinformatics analysis results
  - Tiering
  - Exomiser

https://research-help.genomicsengland.co.uk/display/GERE/LabKey

Rare disease tiering



SNV or Indel variant

Rare in the healthy population
Protein altering in at least one transcript
Follow the relevant **mode of inheritance** (PanelApp)
Segregate appropriately within the family

No → Tier null

Yes

Is the variant in a Green gene on a PanelApp panel applied to the participant?

No → Tier 3

Yes

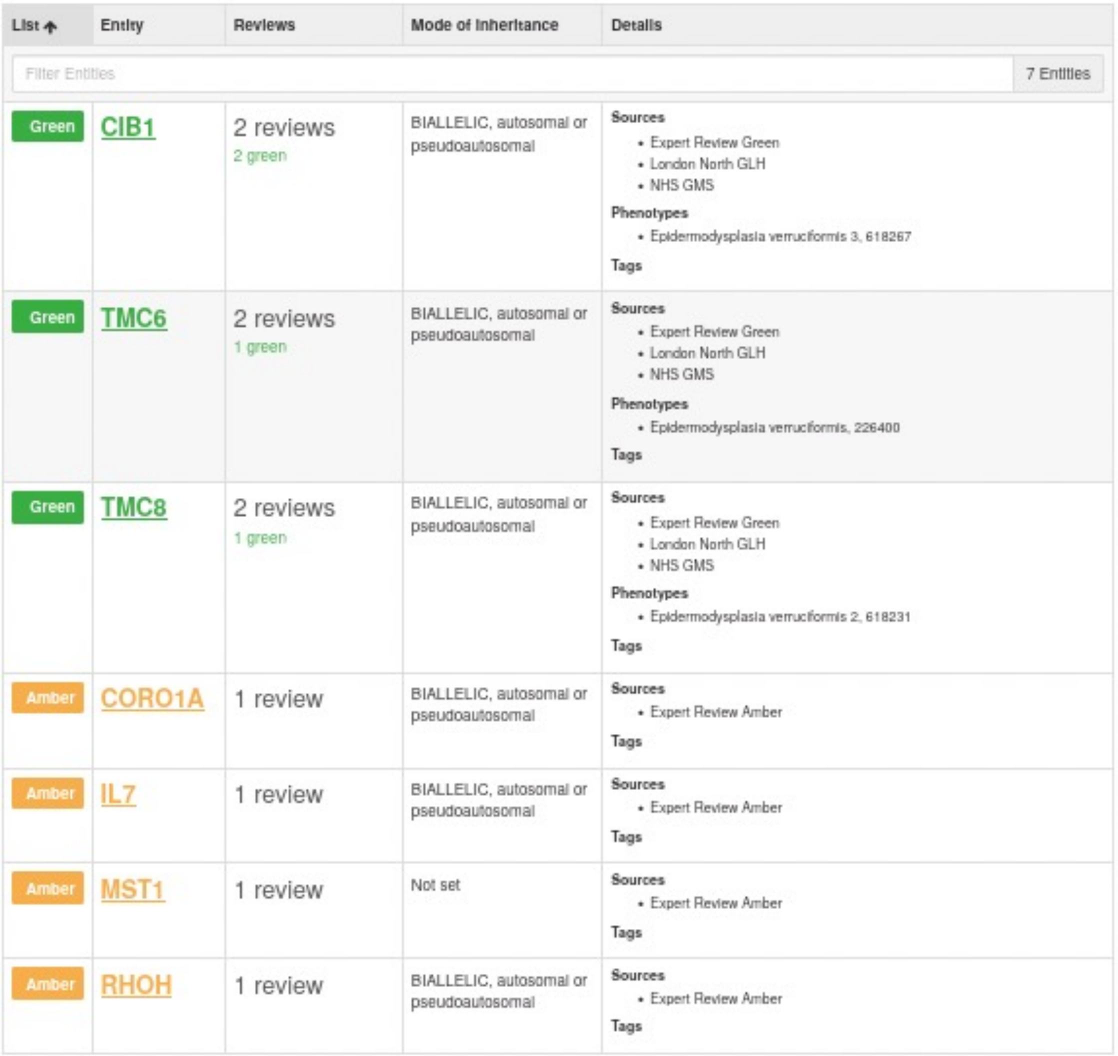
What is the most severe predicted consequence of the variant?

Protein altering → Tier 2

Protein truncation
De novo → Tier 1

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046769

PanelApp



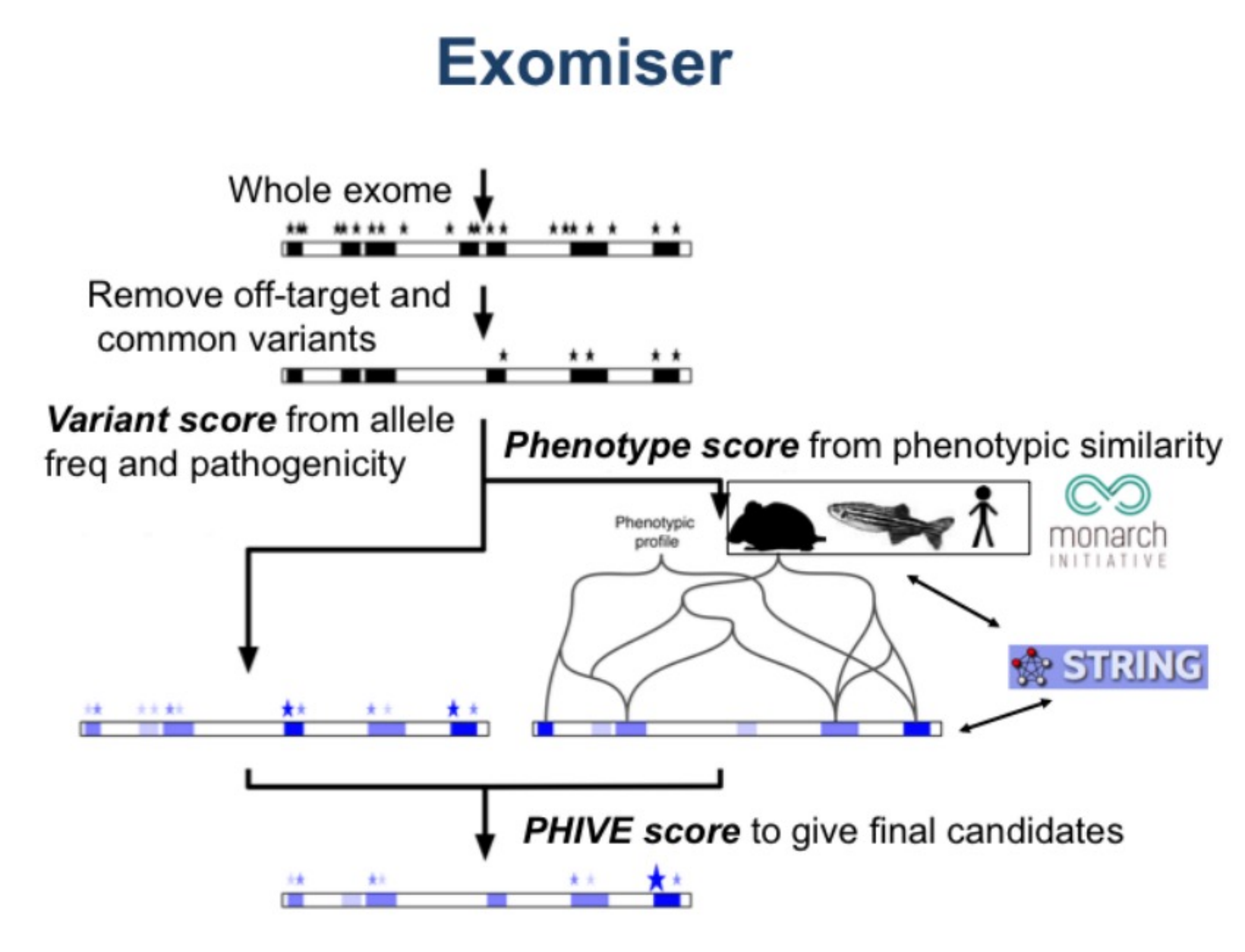
Filter Entities

7 Entities

List	Entry	Reviews	Mode of inheritance	Details
Green	CIB1	2 reviews 2 green	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Green • London North GLH • NHS GMS **Phenotypes** • Epidermodysplasia verruciformis 3, 618267 **Tags**
Green	TMC6	2 reviews 1 green	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Green • London North GLH • NHS GMS **Phenotypes** • Epidermodysplasia verruciformis, 226400 **Tags**
Green	TMC8	2 reviews 1 green	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Green • London North GLH • NHS GMS **Phenotypes** • Epidermodysplasia verruciformis 2, 618231 **Tags**
Amber	CORO1A	1 review	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Amber **Tags**
Amber	IL7	1 review	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Amber **Tags**
Amber	MST1	1 review	Not set	**Sources** • Expert Review Amber **Tags**
Amber	RHOH	1 review	BIALLELIC, autosomal or pseudoautosomal	**Sources** • Expert Review Amber **Tags**

https://panelapp.genomicsengland.co.uk/

Exomiser



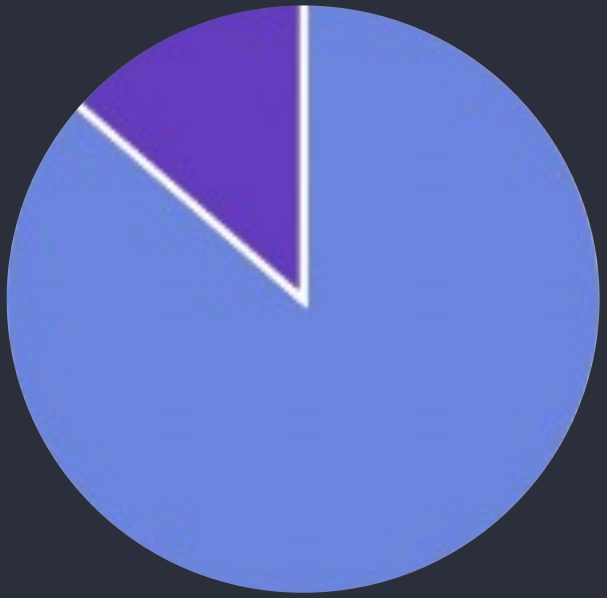
Exomiser

Whole exome
Remove off-target and common variants
Variant score from allele freq and pathogenicity
Phenotype score from phenotypic similarity
Phenotypic profile
monarch INITIATIVE
STRING
PHIVE score to give final candidates

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046618

Exomiser/tiering assembly

Rare disease

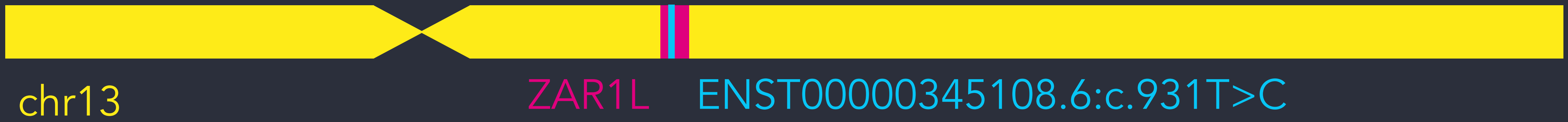


GRCh38 (aka hg38)
GRCh37 (aka hg19)

Cancer

~~Somatic GRCh38~~
~~Germline GRCh38~~
~~Germline GRCh37~~
~~Somatic GRCh37~~

Exomiser/tiering assembly



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

Demo: Variant data in LabKey

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

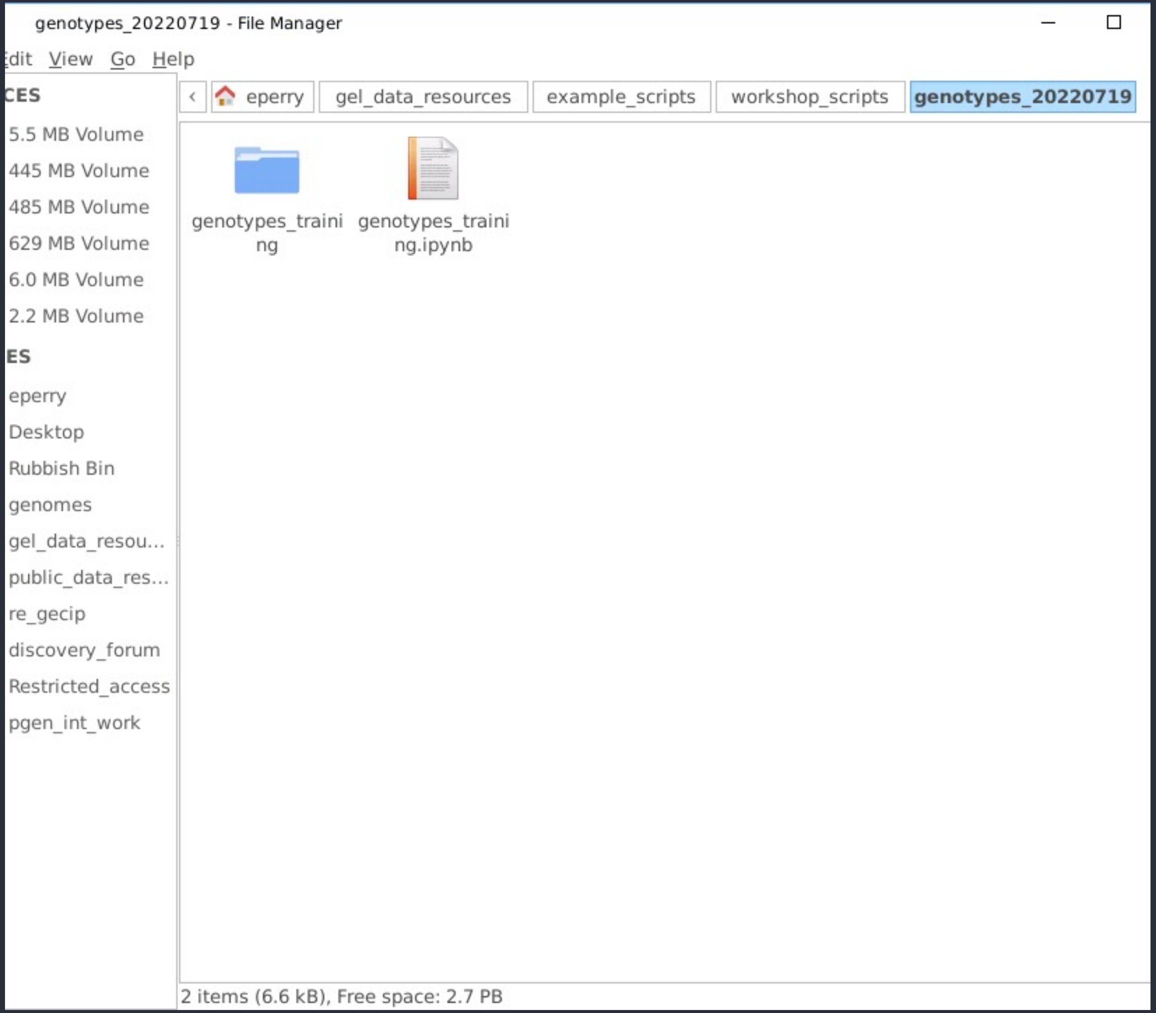
Set up .netrc

- 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 copying and 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/genotypes_20220719



genotypes_20220719 - File Manager

2 items (6.6 kB), Free space: 2.7 PB

Accessing the notebooks

Python

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

Open link in browser

R

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

Demo: Variant data in LabKey API

3. Finding genotypes with IVA

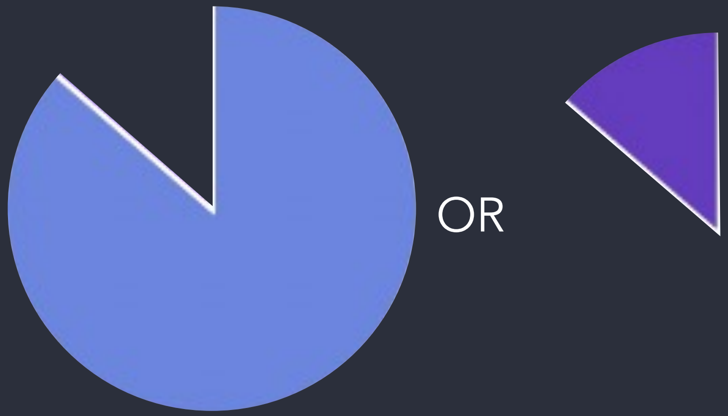
IVA Variant Browser

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

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=45024271

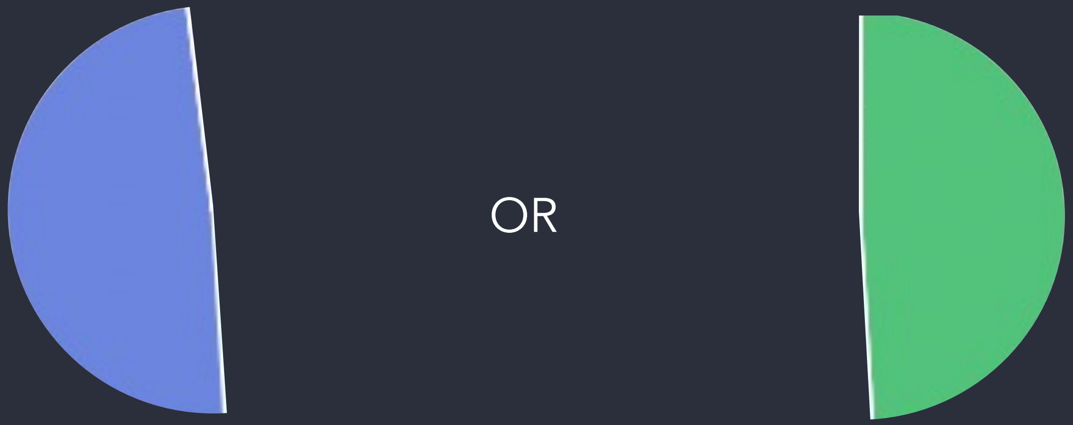
IVA genome assembly

Rare disease



GRCh38 (aka hg38)
GRCh37 (aka hg19)

Cancer



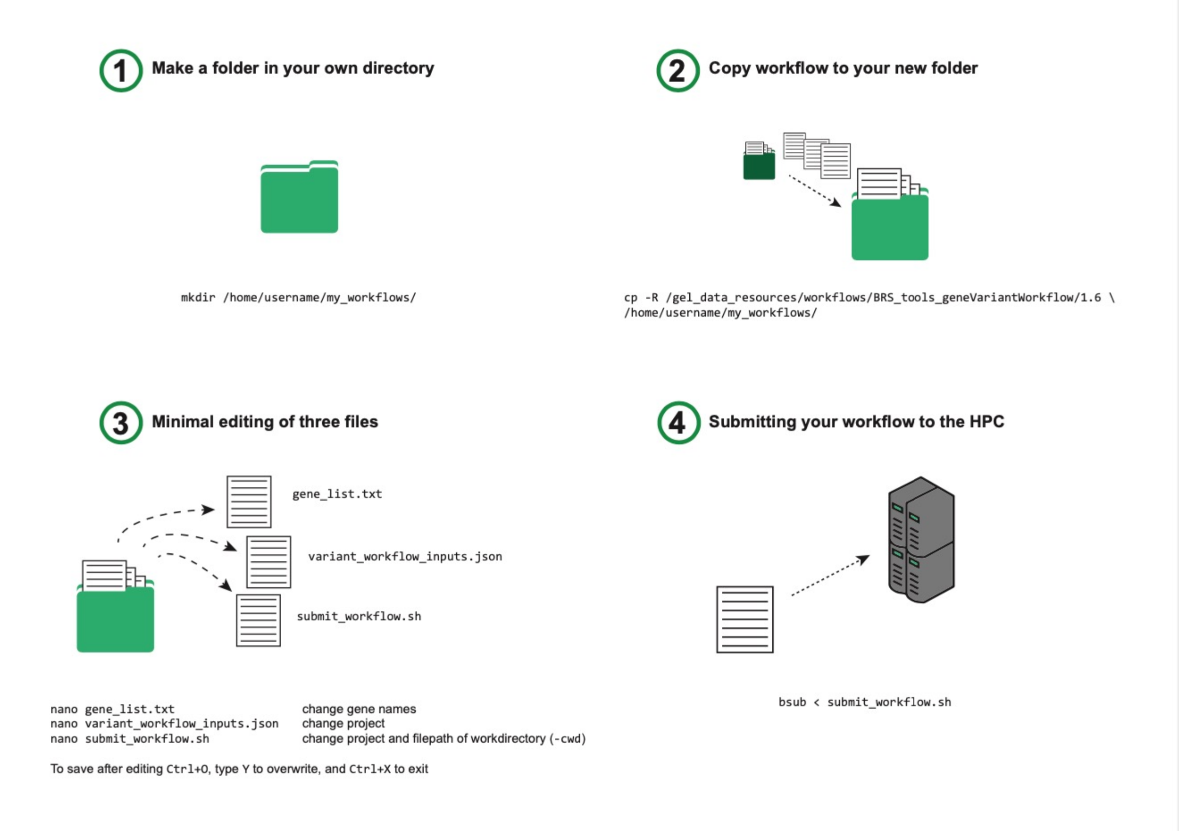
Somatic GRCh38
Germline GRCh38
~~Germline GRCh37~~
~~Somatic GRCh37~~

Demo: Finding variants with IVA

4. The Gene-Variant and SV/CNV workflows

Gene-Variant workflow

- Submit a list of genes or regions
- Find all short variants in these genes/regions
- Get participants with these variants
- Choose somatic/germline, cancer/rare disease



1 Make a folder in your own directory

```
mkdir /home/username/my_workflows/
```

2 Copy workflow to your new folder

```
cp -R /gel_data_resources/workflows/BRS_tools_geneVariantWorkflow/1.6 \
/home/username/my_workflows/
```

3 Minimal editing of three files

```
nano gene_list.txt                    change gene names
nano variant_workflow_inputs.json     change project
nano submit_workflow.sh               change project and filepath of workdirectory (-cwd)
```

To save after editing `Ctrl+O`, type `Y` to overwrite, and `Ctrl+X` to exit

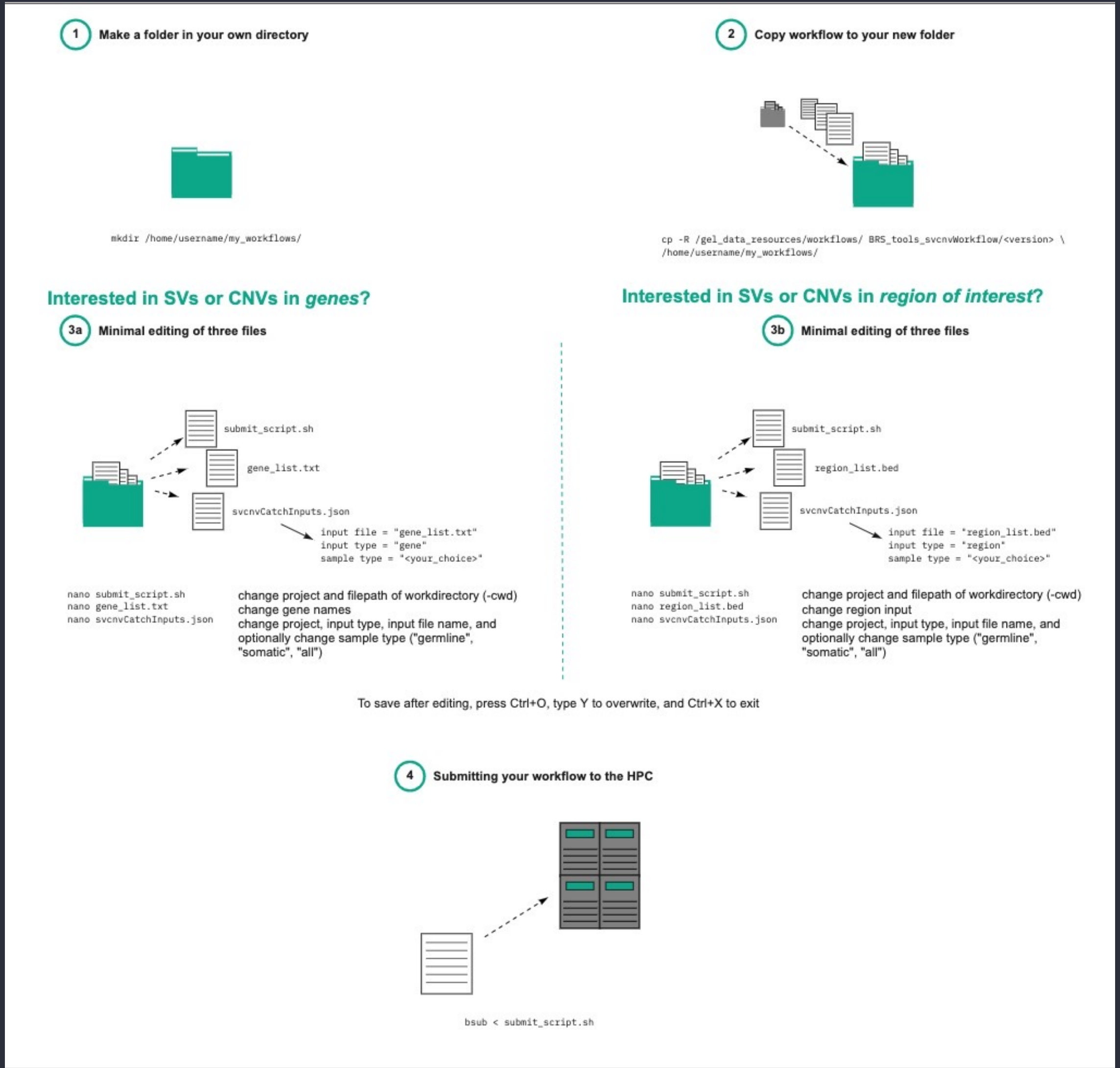
4 Submitting your workflow to the HPC

```
bsub < submit_workflow.sh
```

https://research-help.genomicsengland.co.uk/display/GERE/Gene-Variant+Workflow

SV-CNV workflow

- Submit a list of genes or regions
- Find all SVs/CNVs in these genes/regions
- Get participants with these variants
- Choose somatic/germline, cancer/rare disease



1 **Make a folder in your own directory**

```
mkdir /home/username/my_workflows/
```

2 **Copy workflow to your new folder**

```
cp -R /gel_data_resources/workflows/ BRS_tools_svcnvWorkflow/<version> \
/home/username/my_workflows/
```

Interested in SVs or CNVs in *genes*?

3a **Minimal editing of three files**

```
input file = "gene_list.txt"
input type = "gene"
sample type = "<your_choice>"
```

```
nano submit_script.sh
nano gene_list.txt
nano svcnvCatchInputs.json
```

change project and filepath of workdirectory (-cwd)
change gene names
change project, input type, input file name, and optionally change sample type ("germline", "somatic", "all")

Interested in SVs or CNVs in *region of interest*?

3b **Minimal editing of three files**

```
input file = "region_list.bed"
input type = "region"
sample type = "<your_choice>"
```

```
nano submit_script.sh
nano region_list.bed
nano svcnvCatchInputs.json
```

change project and filepath of workdirectory (-cwd)
change region input
change project, input type, input file name, and optionally change sample type ("germline", "somatic", "all")

To save after editing, press Ctrl+O, type Y to overwrite, and Ctrl+X to exit

4 **Submitting your workflow to the HPC**

```
bsub < submit_script.sh
```

https://research-help.genomicsengland.co.uk/display/GERE/SVCNVworkflow+v2.0.1

Benefits of using workflows

SV/CNV WORKFLOW IS THE ONLY WAY TO GET SV/CNVS ACROSS ALL PARTICIPANTS IN RE

BOTH WORKFLOWS QUERY GENOMES ALIGNED TO GRCH37 AND GRCH38

Workflows and genome assembly

chr13
ZAR1L
ENST00000345108.6:c.931T>C

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

Demo: Running workflows on the HPC

5. Aggregated variant files

AggV2

combined germline gVCFs

Split into 1371 chunks

Release 8

GRCh38

Quality metrics

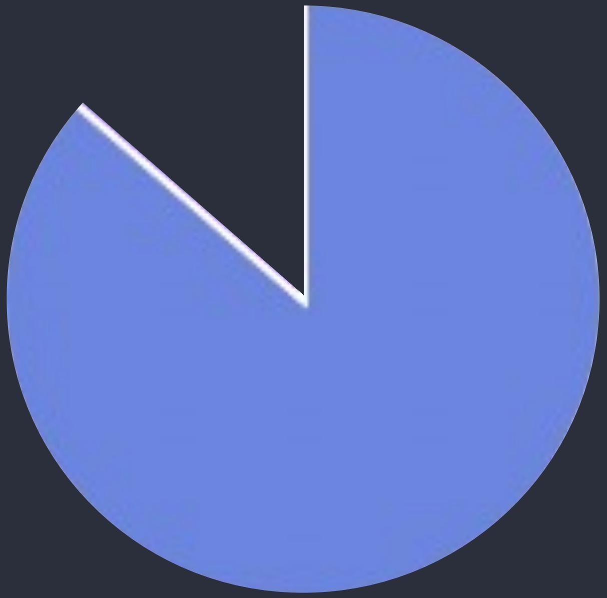
Functional annotation

- No genomes aligned to GRCh37
- Includes participants who have since withdrawn consent – see docs for how to remove

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046780

AggV2 assembly

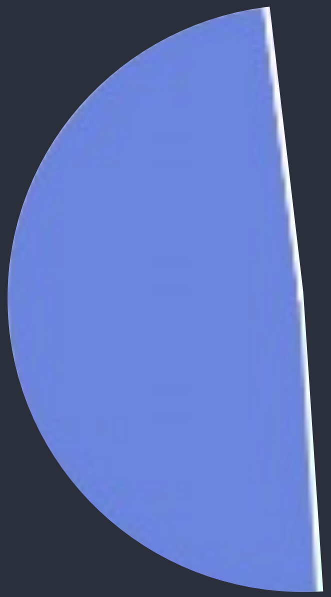
Rare disease



GRCh38 (aka hg38)

~~GRCh37 (aka hg19)~~

Cancer



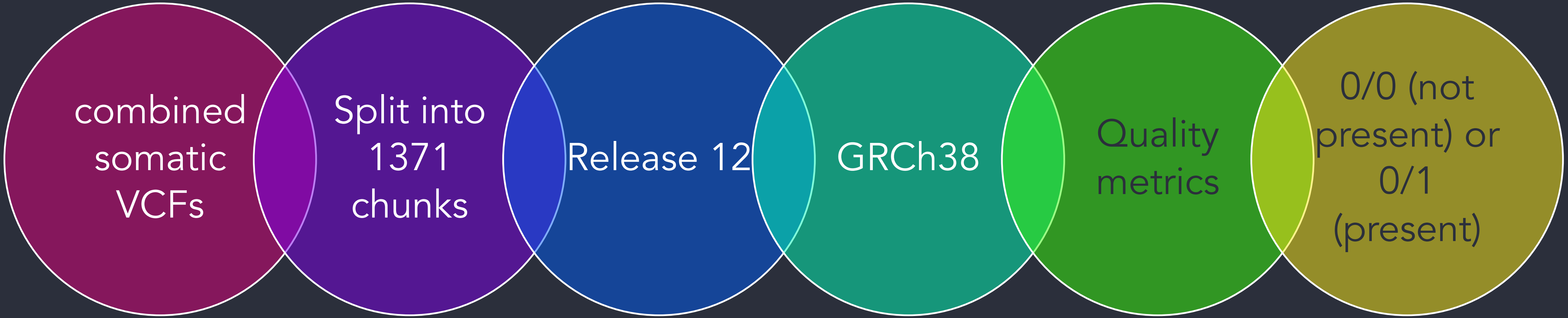
~~Somatic GRCh38~~

Germline GRCh38

~~Germline GRCh37~~

~~Somatic GRCh37~~

somAgg



⚠

- No genomes aligned to GRCh37
- Includes participants who have since withdrawn consent – see docs for how to remove
- Feedback on somAgg

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38046780

somAgg assembly

Rare disease

~~GRCh38 (aka hg38)~~
~~GRCh37 (aka hg19)~~

Cancer

Somatic GRCh38
~~Germline GRCh38~~
~~Germline GRCh37~~
~~Somatic GRCh37~~

Aggregated files code books

- AggV2 and somAgg can be analysed with:
  - Bcftools
  - Split-vep
  - R/Python
  - Bedtools
- Code books available to take you through most common use-cases:
  - AggV2 https://research-help.genomicsengland.co.uk/display/GERE/aggV2+Code+Book
  - somAgg https://research-help.genomicsengland.co.uk/display/GERE/somAgg%3A+Code+Book

Aggregated files chunks

Split into 1371 chunks

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

6. Using bcftools on the HPC

Demo: Using bcftools on the HPC

Summary and comparison of tools

Comparison – access type

Labkey tables: Tiering and Exomiser	Point-and-click access to tables plus API
IVA	Point-and-click
Gene-variant and SV/CNV workflows	Command line
Aggregated VCFs with bcftools	Command line

Comparison – search by

Labkey tables: Tiering and Exomiser	Gene, region or HGVS (Exomiser only)
IVA	Gene, region or rsID
Gene-variant and SV/CNV workflows	Gene or region
Aggregated VCFs with bcftools	Region

Comparison – variants available

Labkey tables: Tiering and Exomiser	Only variants that have passed tiering or exomiser filters
IVA	All variants
Gene-variant and SV/CNV workflows	All variants
Aggregated VCFs with bcftools	All variants present in GRCh38-aligned genomes from release 8 (AggV2) or 12 (somAgg)

Comparison – genome assembly

Labkey tables: Tiering and Exomiser	Variants on both assemblies in the table, with assembly as a column
IVA	Assemblies available as separate datasets
Gene-variant and SV/CNV workflows	By-gene searches both assemblies
Aggregated VCFs with bcftools	Only GRCh38

Comparison – underlying VCFs

Labkey tables: Tiering and Exomiser	Platypus
IVA	Platypus
Gene-variant and SV/CNV workflows	Strelka
Aggregated VCFs with bcftools	Strelka

Key takeaways

Use IVA for a fast overview

Pre-written workflows for gene-based searches

Aggregated VCFs have code-books for common use-cases

Key takeaways

Use IVA for a fast overview

!

Genome Assembly

Aggregated VCFs have code-books for common use-cases

7. Getting help and questions

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

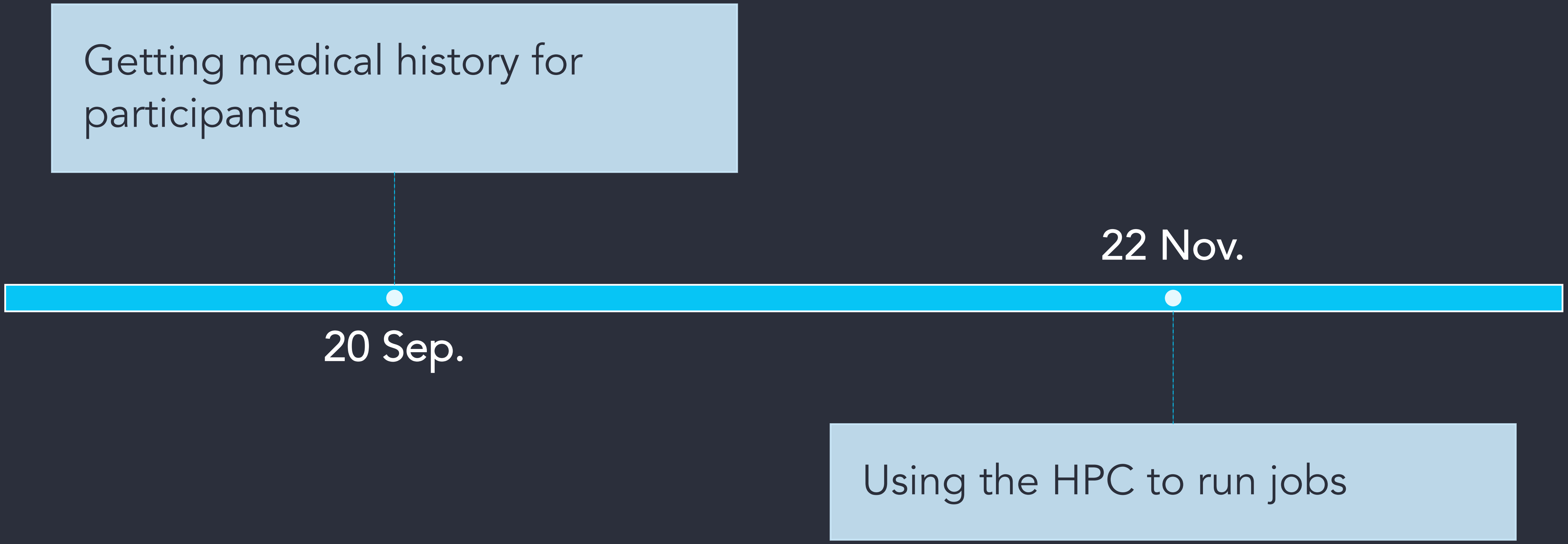
Questions

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



Getting medical history for participants

20 Sep.

22 Nov.

Using the HPC to run jobs

Feedback

Thank you