Genomics England

Introduction to the Genomics England Research Environment

Emily Perry / Roel Bevers

Research Engagement Manager / Senior Bioinformatician – Research Services

15th March 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
- We will record this training and distribute the censored video afterwards

Questions

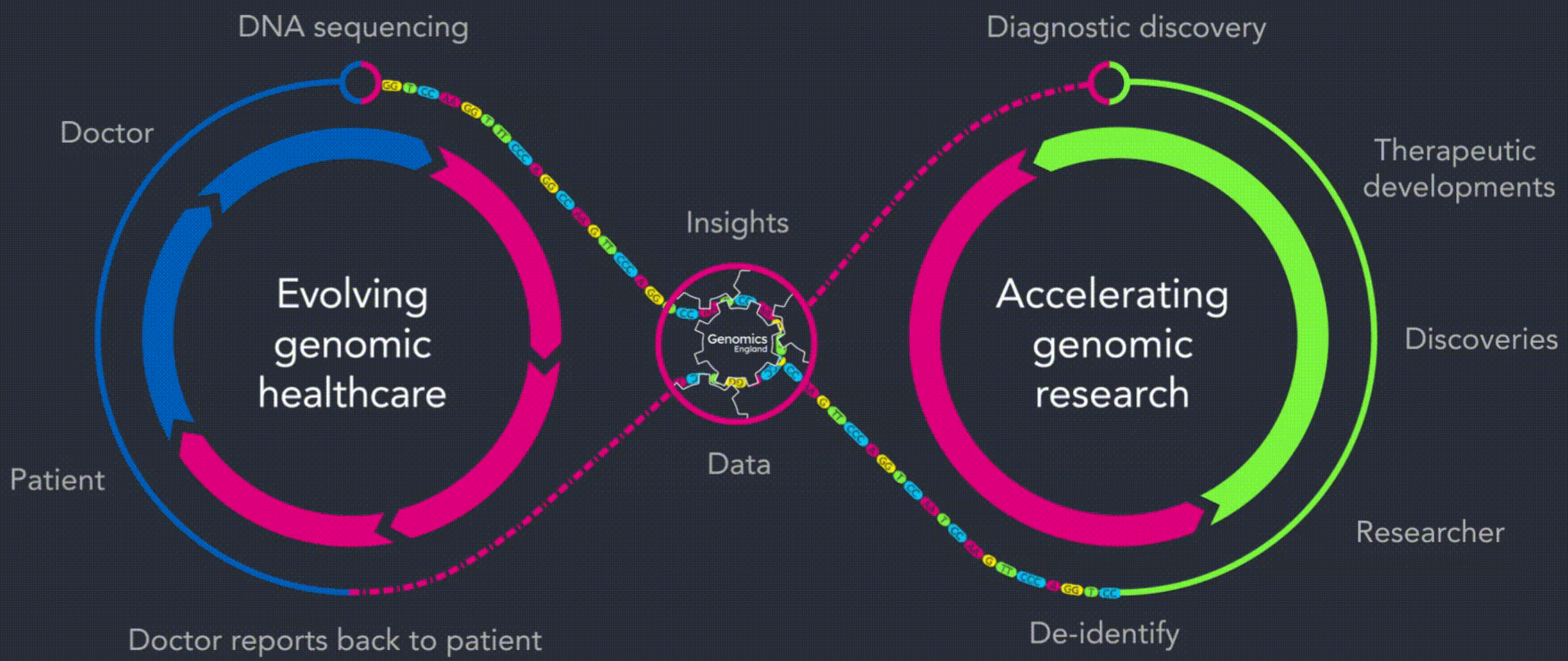
Your microphones are all muted

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

Agenda

1. Introduction and admin
2. Sources and type of data in the Research Environment
3. Tools in the Research Environment
4. Programmatic access to Genomics England data
5. Running command line tools and pipelines using our HPC cluster
6. The Airlock, restricted import and export of data
7. The future: CloudRE and the GMS
8. Getting help
9. Questions



DNA sequencing

Doctor

Evolving genomic healthcare

Patient

Doctor reports back to patient

Insights

Genomics England

Data

Diagnostic discovery

Therapeutic developments

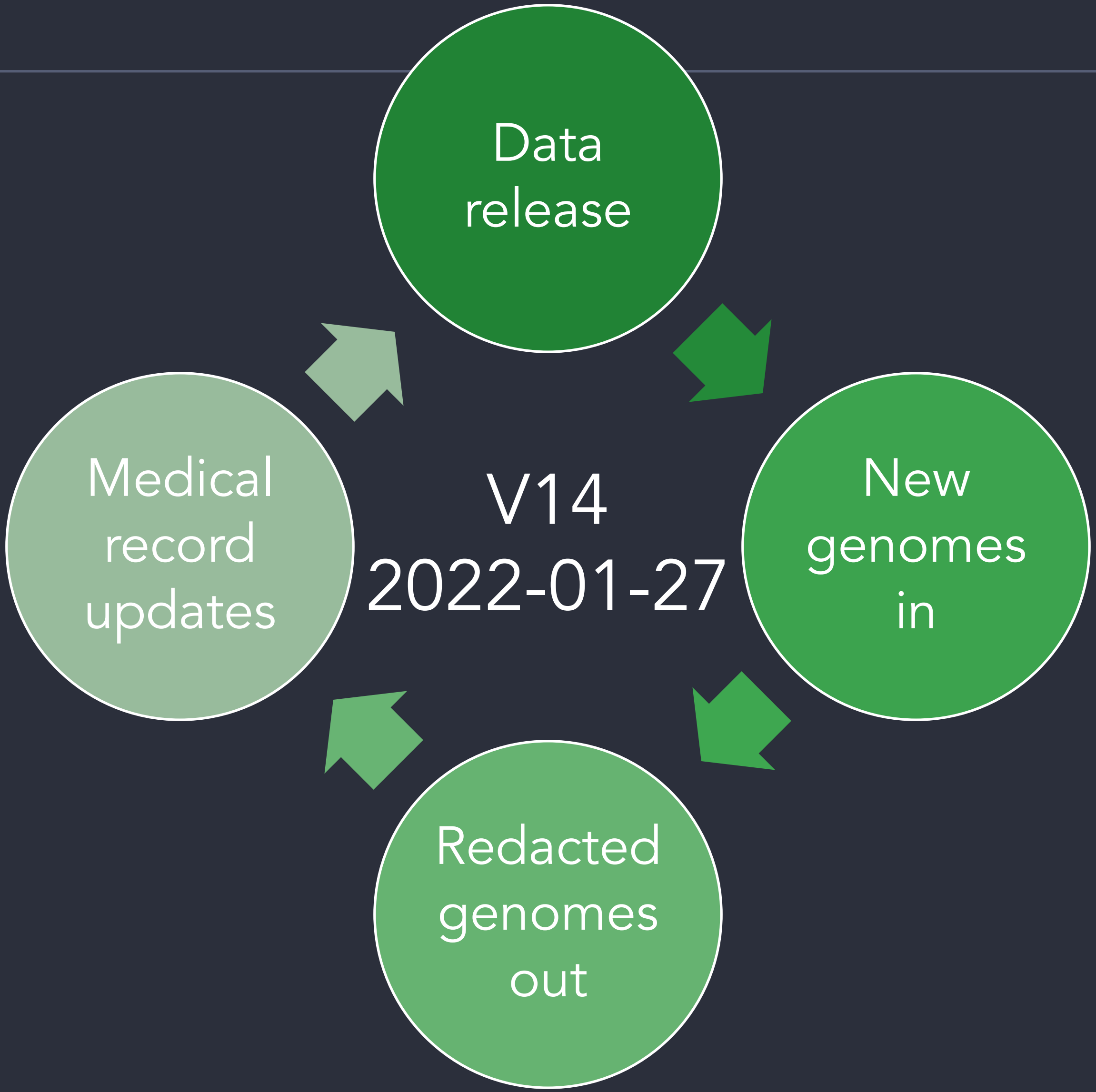
Accelerating genomic research

Discoveries

Researcher

De-identify

Data releases



Data release

New genomes in

Redacted genomes out

Medical record updates

V14
2022-01-27

https://research-help.genomicsengland.co.uk/display/GERE/Main+Programme+Data+Releases

100,000 Genomes Project Data

Genomics

	Cancer	Rare Disease	Total
Participants	17,955	72,955	90,259 + 35K COVID in CloudRE only
Genomes	42,922 Germline + Tumour 30x 100x	75,526 Germline <20% Singleton	118,488

100,000 Genomes Project Data

Genomics

Clinical Data

- HPO terms
  - Rare disease
  - Other conditions
- Tumour staging
- Tumour location
- Histological subtype
- Treatment regimen

- NHS records
  - Hospital Episode Statistics
  - Mental Health Services Data Set

- Mortality data ONS

- Exit questionnaire for rare disease

- COVID-19 status

- Primary Care Data for COVID-19

100,000 Genomes Project Data

Genomics

Clinical Data

Consent

Clinically accredited pipelines

for diagnostics

Lifetime follow-up

+ full retrospective data

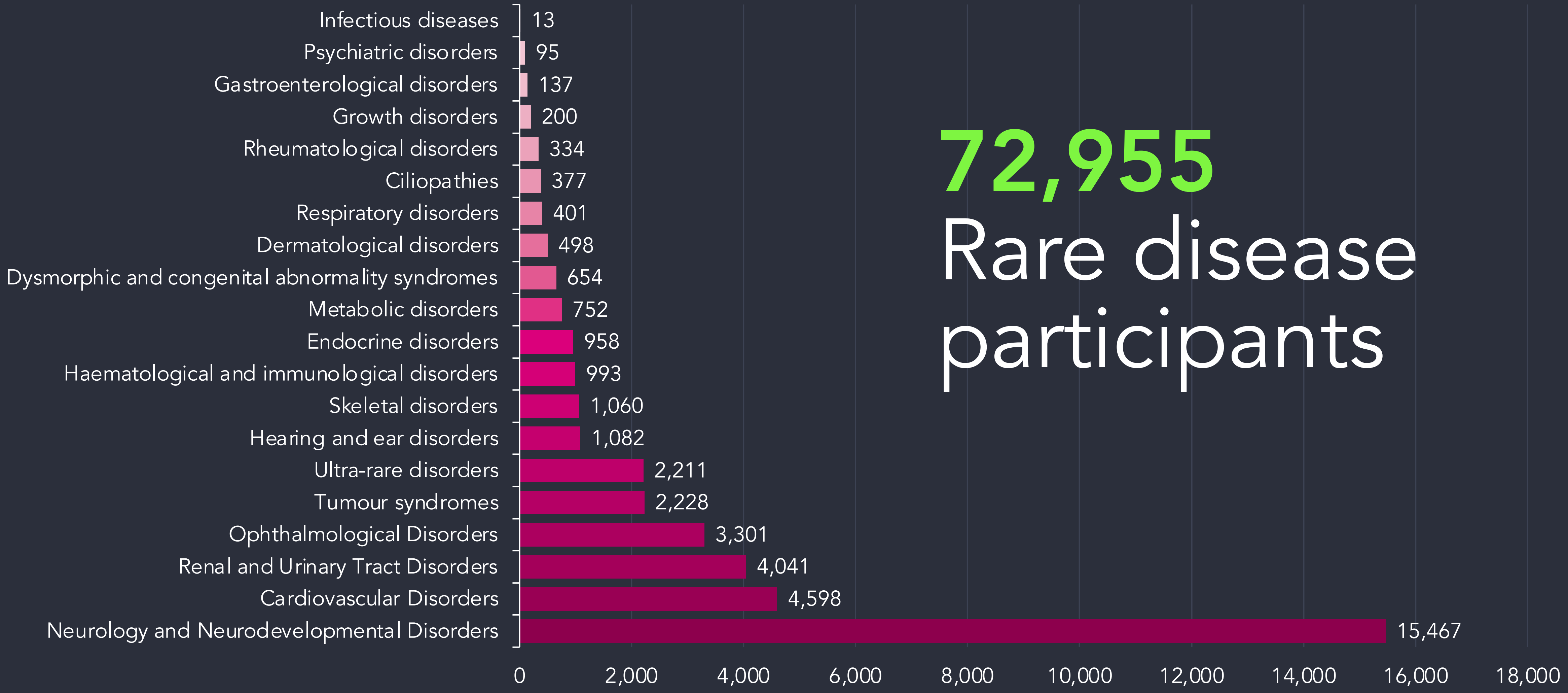
Re-engagement

re-phenotyping

re-sampling

re-cruiting

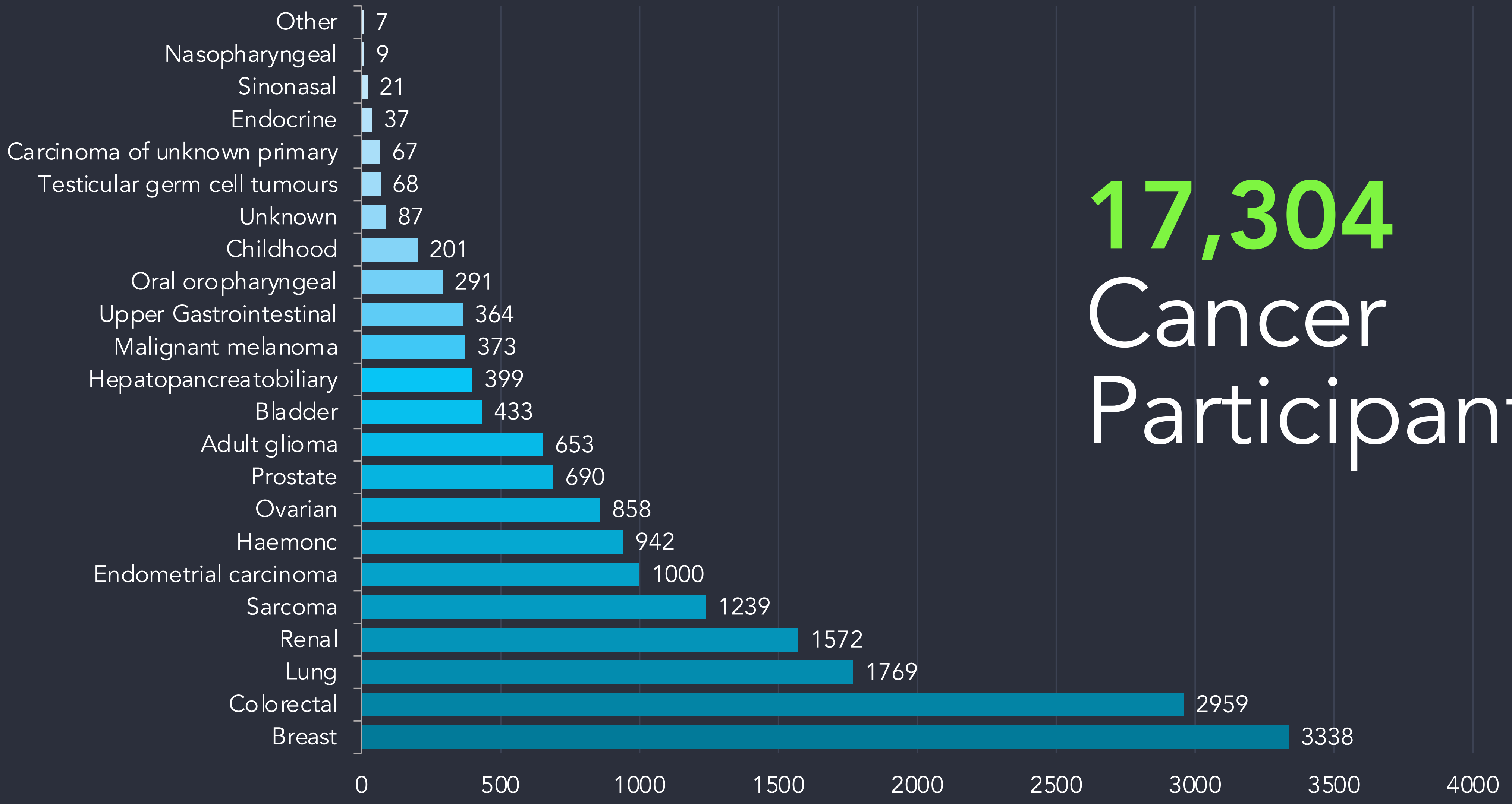
100KGP rare disease participants



72,955
Rare disease participants

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=66846734

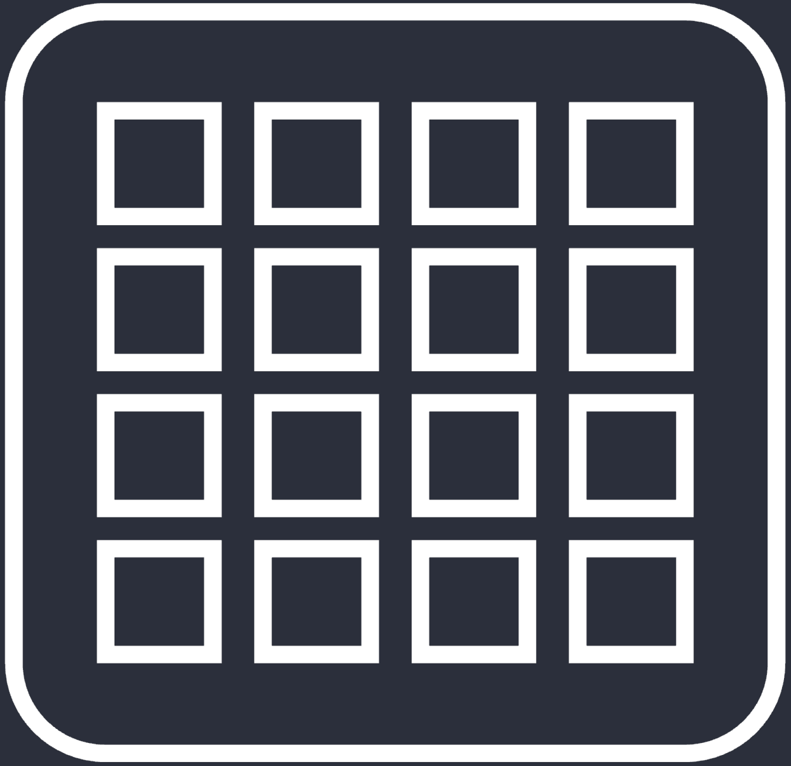
100KGP Cancer participants



17,304
Cancer
Participants

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=66846734

Two REs: RE1.0 and CloudRE



3. Tools in the Research Environment

Participant Explorer
Search for participants by phenotypes or identifiers

IVA
Explore genomic variants and what genotypes GEL participants have for them

IGV
Visualise genomic data

LabKey
Explore the tables of GEL data

https://research-help.genomicsengland.co.uk/pages/viewpage.action?pageId=38047248

Demo: tools in the RE

4. Programmatic access to Genomics England data

Labkey API allows you to:

- Combine data and filters from multiple tables
- Replicate queries exactly:
  - When new data releases come out
  - Between analyses
- 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

Labkey API - Python

```
# Import the needed modules, labkey and pandas (for dataframes)
import labkey
import pandas as pd

# Specify what we are connecting to, and what schema and tables we want
labkey_server = "labkey-embassy.gel.zone"  # The labkey server we are connecting to.
project_name = "main-programme/main-programme_v14_2022-01-27"  # The data we want to access.
context_path = "labkey"
schema_name = "lists"  # The schema we are getting data from.

# Create the SQL query as a string
sql = (
    "SELECT participant.Participant_Id, participant.Programme, sequencing_report.lab_sample_id /
    FROM lists.participant /
    JOIN lists.sequencing_report /
    ON participant.Participant_Id = sequencing_report.Participant_Id /
    WHERE sequencing_report.lab_sample_id between 1018056774 and 1018068634;"
)

# Create an object that will let us connect to the LabKey databases. This does not change.
server_context = labkey.utils.create_server_context(
    labkey_server, project_name, context_path, use_ssl=True
)

# The data are returned and stored in the variable results.
results = labkey.query.execute_sql(server_context, schema_name, sql)

# Data are returned as a dictionary, will all of the table information stored under the key "rows".
# We make a dataframe of all of the table information using pandas.
table_of_data = pd.DataFrame(results["rows"])
```

Change the database version you're accessing

SQL query with standard SQL syntax

Data returned as a dictionary and can be converted to a data-frame

Labkey API - R

```
# Import the labkey library
library(Rlabkey)

# Set the baseURL
labkey.setDefaults(baseUrl= "https://labkey-embassy.gel.zone/labkey/")
Project_name <- "/main-programme/main-programme_v14_2022-01-27"

# Write your SQL query here
query <- “SELECT participant.Participant_Id, participant.Programme, sequencing_report.lab_sample_id
    FROM lists.participant
    JOIN lists.sequencing_report
    ON participant.Participant_Id = sequencing_report.Participant_Id
    WHERE sequencing_report.lab_sample_id between 1018056774 and 1018068634;"

mysql <- labkey.executeSql(
    schemaName="lists",
    colNameOpt = "rname",
    maxRows = 100000000,
    folderPath = project_name,
    sql = query
)
```

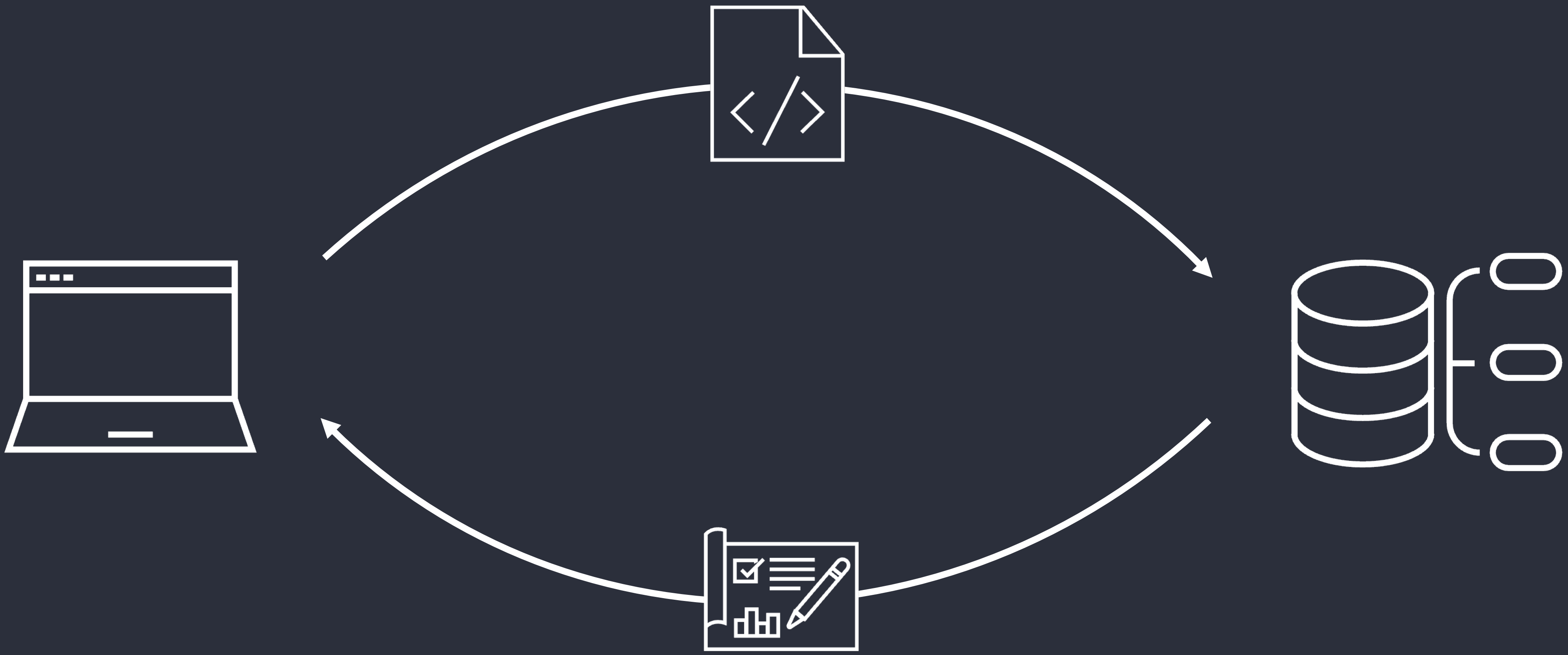
Change the database version you’re accessing

SQL query with standard SQL syntax

Data returned as a data-frame

Demo: running LabKey API scripts

5. Running command line tools and pipelines using our HPC cluster



https://research-help.genomicsengland.co.uk/display/GERE/4.+In-Depth+Guide+to+HPC+Usage

Pre-installed tools on the cluster

APBS
AdapterRemoval
AutoDock-Vina
BCFTools
BEDOPS
BEDTools
BLAST
BLAT
BWA
BamTools
Bio-DB-HTS
BioPerl
Bowtie
CADD
CNVator
CaVEMAN
Canvas
Centrifuge
Circos
Clustal-Omega
EIGENSOFT
FASTX-Toolkit
FASTQC
FlashPCA2
GATK
GD
GISTIC
GMAP-GSNAP
HISAT2
HLA-LA
HTSlib
IGV
IMPUTE2
Jellyfish
KNIME
Kraken
LUMPY
MAFFT
MetaGeneAnnotator
MultiQC
NGS
NextGenMap
OMA
OptiType
OrfM
PHYLIP
PLINK
Pindel
Pysam
Quip
RTG-Tools
SAMtools
SHAPEIT
SPAdes
STAR
SURVIVOR
SVINT
Salmon
Sambama
SeqAn
Trimmomatic
UN-CNVc
VCFtools
VEGAS
VEP

https://research-help.genomicsengland.co.uk/display/GERE/Software+Available+on+the+HPC

Pre-installed tools on the cluster

Velvet
ViennaRNA
alleleCount
Bam-readcount
Cellbase
Cromwell
Cryptsplice
Ea-utils
Fastp
Gvcfgenotyper
Kallisto
liftOver
Locuszoom

meRanTK
minimap2
Mosdept
Ncbi-vdb
New_fugue
Nextflow
Picard
Platypus
Rvtests
Seqtk
Singularity
Snptest
Strelka

Tabix
verifyBamID
Vt
GATE
GCC
GCCcore
LLVM
Ispc
DBD-mysql
GDAL
HDF5
MariaDB
PyTables

ROOT
XML-LibXML
XML-Parser
datamash
ntCDF
Savvy
Shrinkwrap
GDB
Autoconf
Automake
Autotools
Boost
Cmake

Doxygen
Gradle
Junit
LZOM4

https://research-help.genomicsengland.co.uk/display/GERE/Software+Available+on+the+HPC

Demo: using standard tools on the HPC

Creating your own workflows to use on the cluster

- You can incorporate any of the existing tools into your own workflow
- Import scripts via Airlock
- Using containers
  - Singularity
  - Docker
  - Quay.io

https://research-help.genomicsengland.co.uk/display/GERE/7.+Analysis+Scripts+and+Workflows

Ready-made scripts/workflows

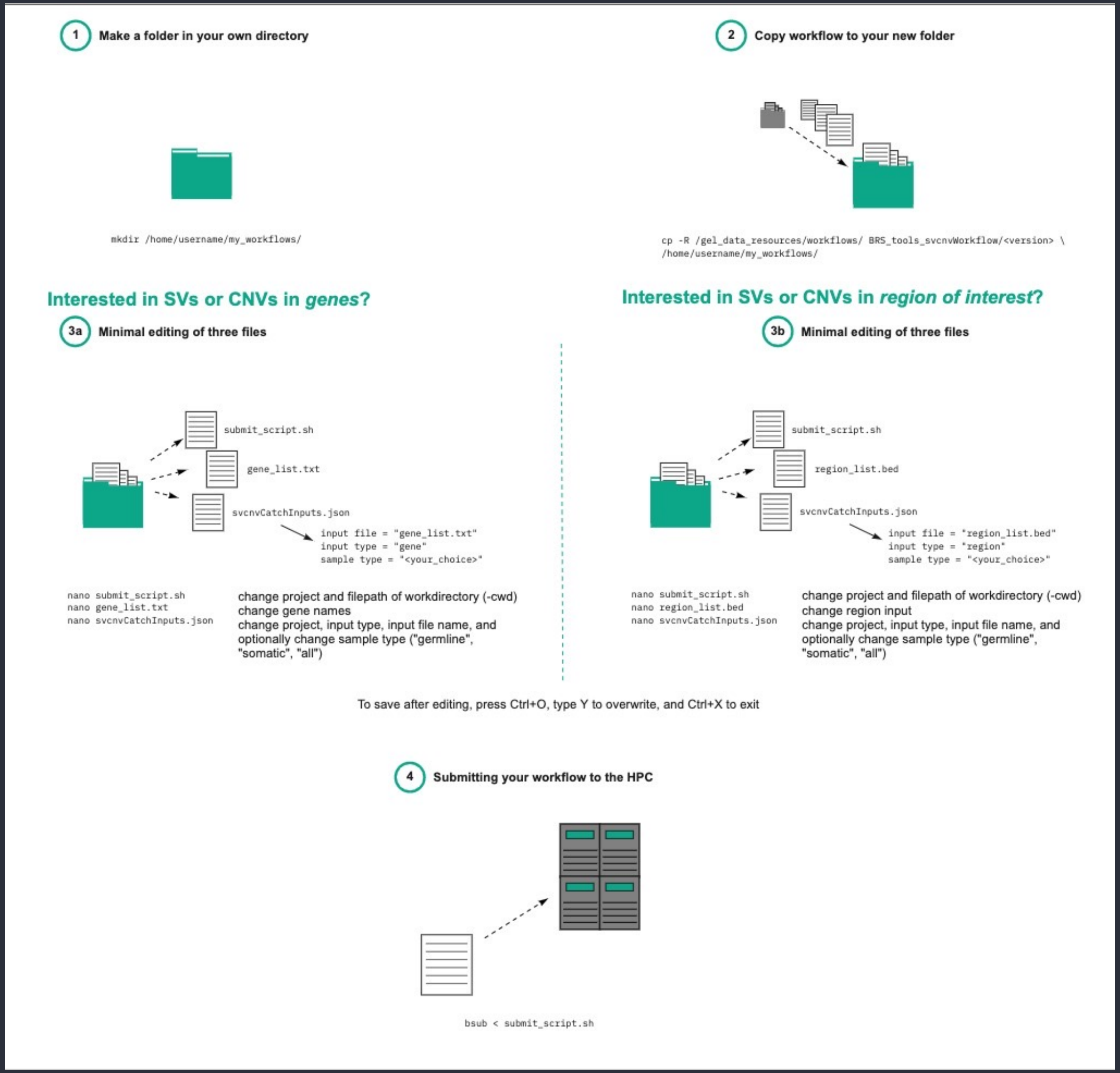
- Extract variants (small or SV; germline or somatic) by coordinate or gene
- Gene centric SNV reports for cancer
- GWAS
- Survival - cancer
- Aggregate variant testing
- Functional variant annotation

https://research-help.genomicsengland.co.uk/display/GERE/7.+Analysis+Scripts+and+Workflows

Demo: running a workflow on the HPC

SV-CNV workflow

- Submit a list of genes or regions
- Find all SVs/CNVs in these genes/regions
- 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")

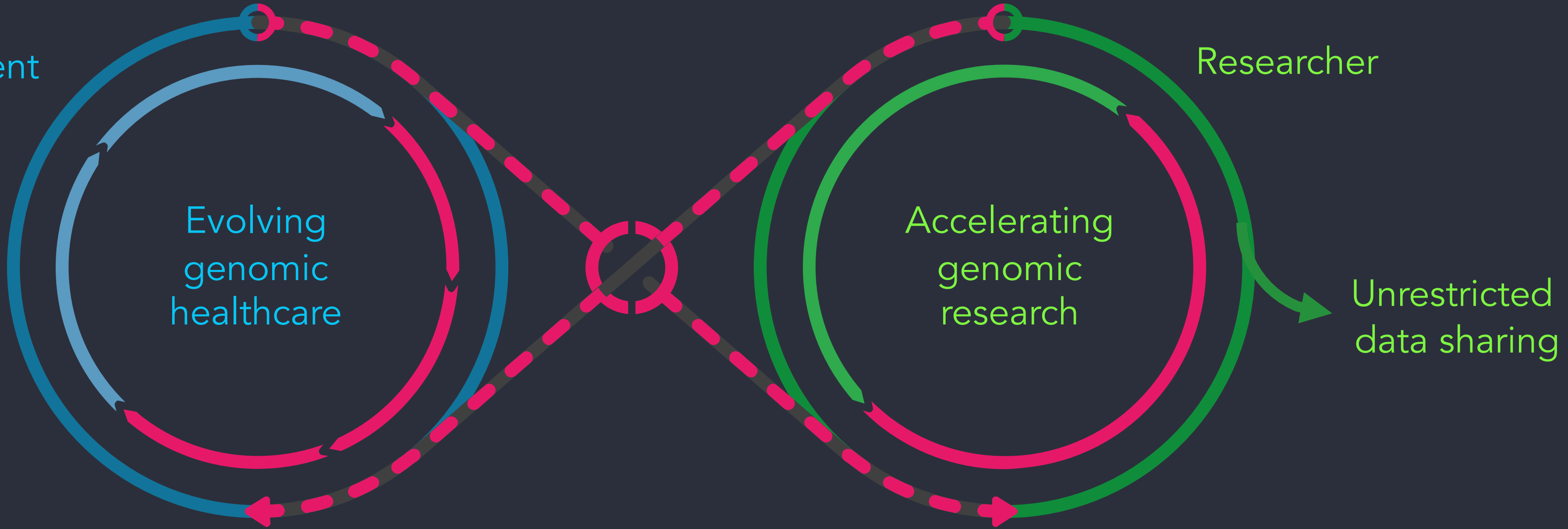
To save after editing, press Ctrl+O, type Y to overwrite, and Ctrl+X to exit

4 Submitting your workflow to the HPC

```
bsub < submit_script.sh
```

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

6. The Airlock, restricted import and export of data


Patient
Evolving genomic healthcare
Researcher
Accelerating genomic research
Unrestricted data sharing

Patients

Healthcare teams

Researchers

https://research-help.genomicsengland.co.uk/display/GERE/5.+The+Airlock

Airlock: what can you export

👍

- Data matching your approved research project
- Aggregate data for groups ≥5 participants

👎

- Data from a project that has not been approved
- Individual data or data that can be otherwise identified

https://research-help.genomicsengland.co.uk/display/GERE/5.+The+Airlock

Demo: exporting data using the Airlock

7. The future: CloudRE

- GEL data
  - 100,000 Genomes Project: rare disease and cancer
  - COVID-19 – severe and mild cases
- Point-and-click tools
  - Cohort browser
  - Running pipelines
- Compute in the Cloud
  - Flexible options based on budget and speed needed
  - Not limited by load
- Bring in data and tools from outside
  - From Github
  - In Containers
  - In S3 buckets

Demo: tools in the CloudRE

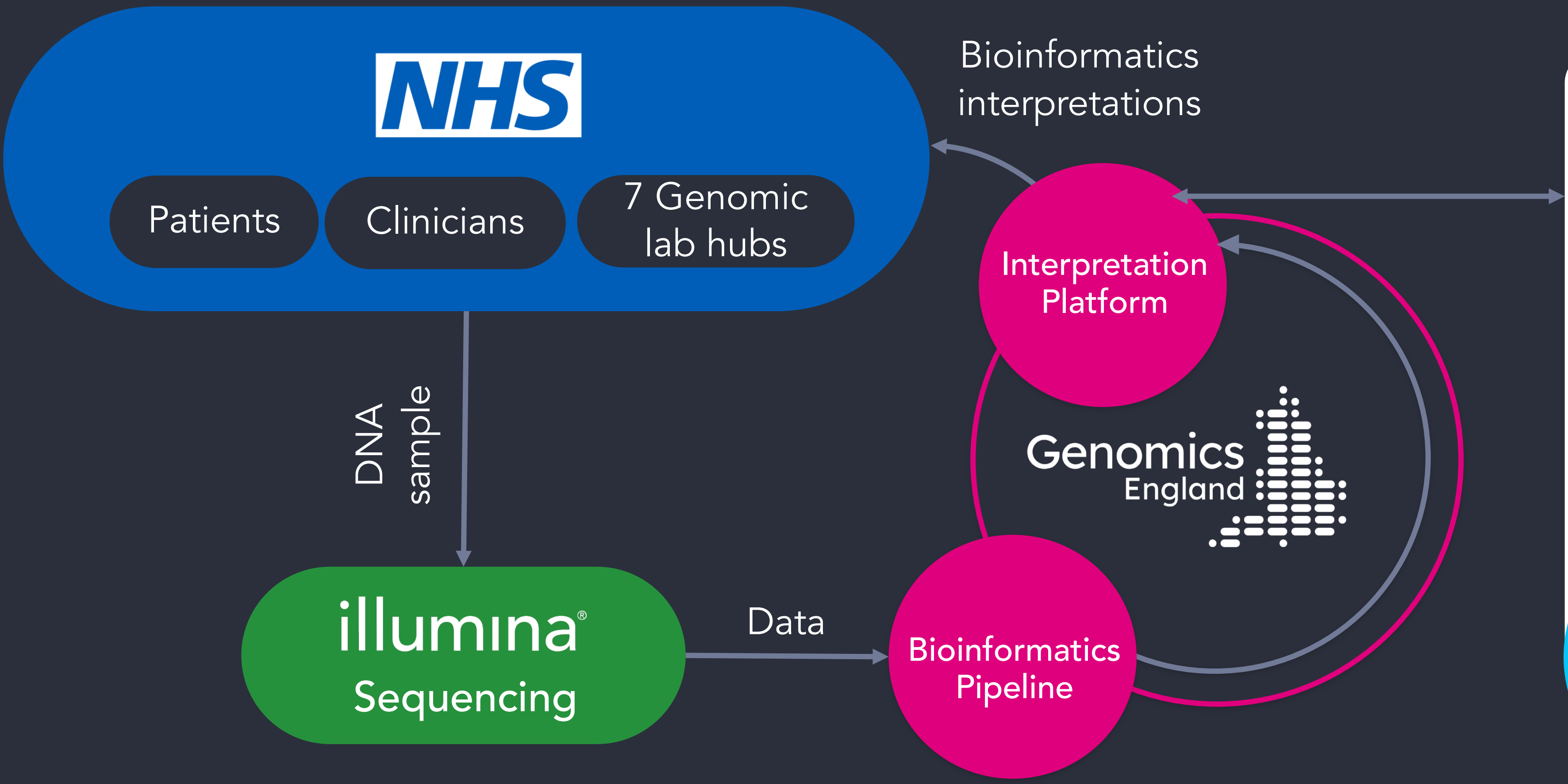
If the CloudRE is for you…

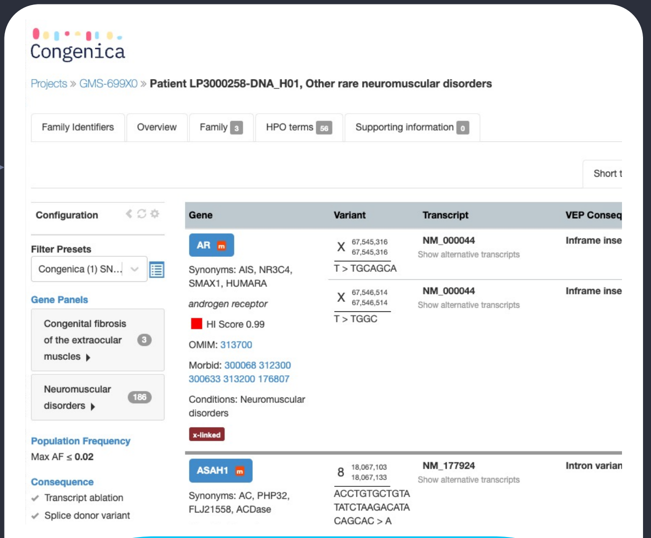
Contact gecip-help@genomicsengland.co.uk

Contact your partnerships director (James/Kate)

- Brief description of why you'd like access
  - Use-case
  - Data
- We will get in touch to discuss feasibility

The future: Genomic Medicine Service





Congenica

Projects » GMS-699X0 » **Patient LP3000258-DNA_H01, Other rare neuromuscular disorders**

Filter Presets

Congenica (1) SN...

Gene Panels

Congenital fibrosis of the extraocular muscles 3

Neuromuscular disorders 186

Population Frequency

Max AF ≤ **0.02**

Consequence

- ✓ Transcript ablation
- ✓ Splice donor variant

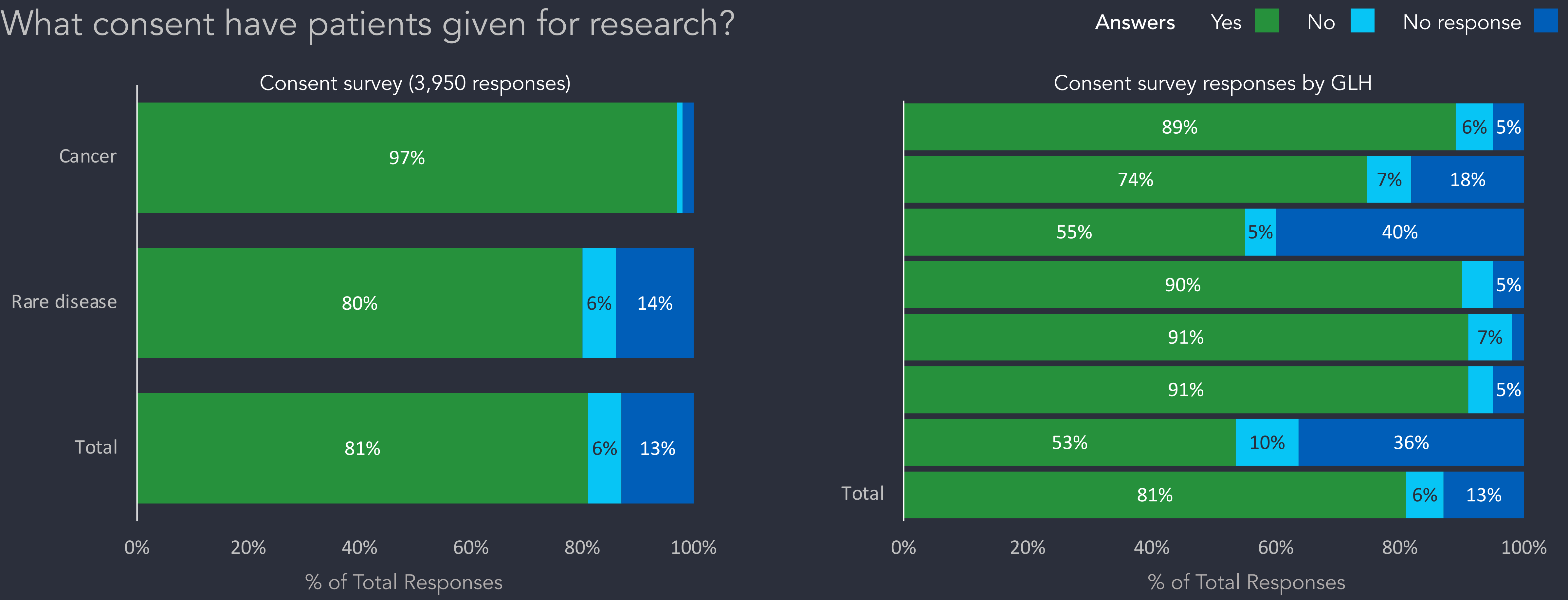
Gene	Variant	Transcript	VEP Conseq
AR m Synonyms: AIS, NR3C4, SMAX1, HUMARA *androgen receptor* 🟥 HI Score 0.99 OMIM: 313700 Morbid: 300068 312300 300633 313200 176807 Conditions: Neuromuscular disorders x-linked	X 67,545,316 67,545,316 T > TGCAGCA	**NM_000044** Show alternative transcripts	**Inframe inse**
	X 67,546,514 67,546,514 T > TGGC	**NM_000044** Show alternative transcripts	**Inframe inse**
ASAH1 m Synonyms: AC, PHP32, FLJ21558, ACDase	8 18,067,103 18,067,133 ACCTGTGCTGTATATCTAAGACATACAGCAC > A	**NM_177924** Show alternative transcripts	**Intron varian**

Congenica

Rare Disease Decision Support

The future: Genomic Medicine Service

What consent have patients given for research?



Consented that data and samples may be used for research, separate to NHS care

8. 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

9. 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

24th May	Building a cohort based on phenotypes and a matching control cohort
22nd July	Finding participants based on genotypes
20th September	Getting medical history for participants
22nd November	Using the HPC to run jobs

Thank you