Genomics England

Working with the new aggregate VCFs – AggV3

Emily Perry

Research Engagement Manager

10th March 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

Questions

Roel Bevers
Senior Bioinformatician - Research Services

Lisa Murphy
Bioinformatician - Research Services

Magdalena Drożdż
Bioinformatician - Research Services

Agenda

1 Introduction and admin

2 How were the AggV3 multisample VCFs created?

3 Interactive sessions in CloudOS

4 Querying AggV3 in the terminal

5 Using AggV3 in batch jobs

6 Taking data in and out of CloudOS

7 Help and questions

2. How were the AggV3 multisample VCFs created?

DRAGEN 3.7.8

A standardised pipeline to which we realign all our germline genomes

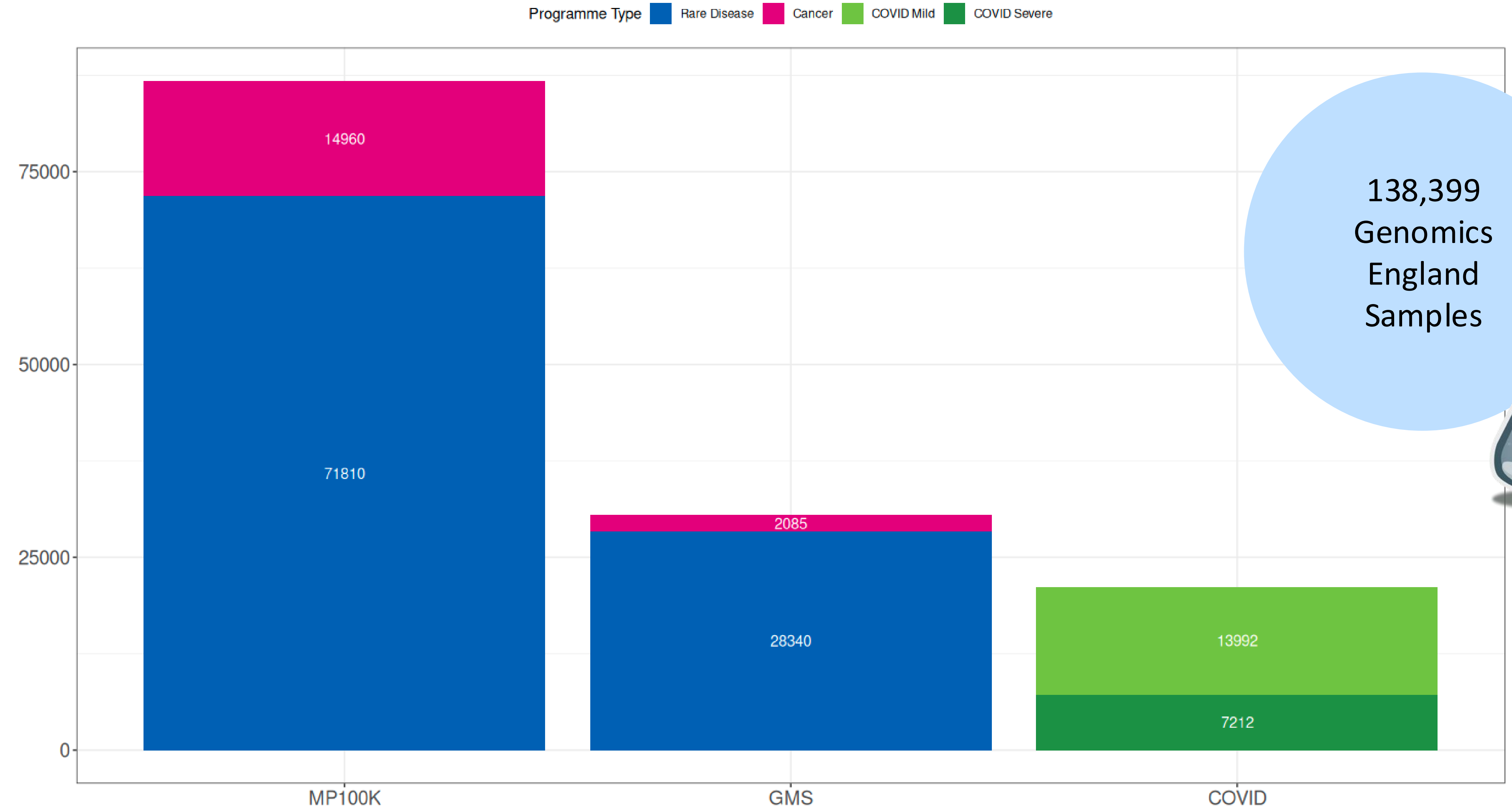
in line with UKBB and AllofUs

AggV3

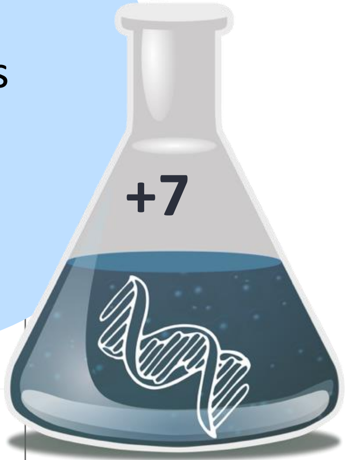
A genomic aggregate onto which we iteratively* add new genomes (**Illumina** collab)

*when numbers make this efficient (i.e. >40k)

Samples in AggV3

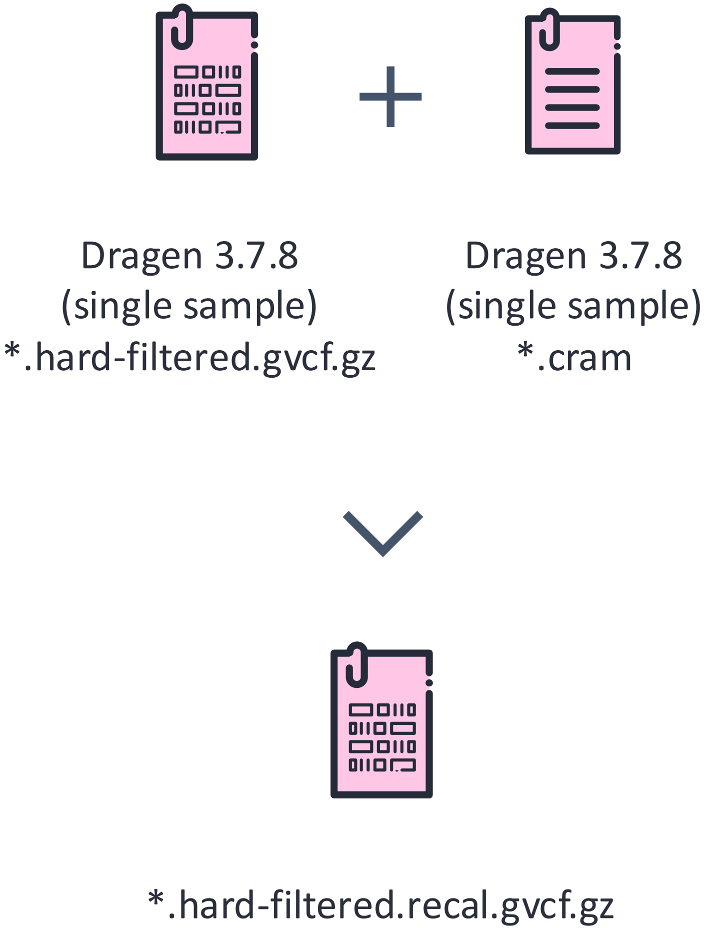


138,399 Genomics England Samples



https://re-docs.genomicsengland.co.uk/aggv3_samples/

Dragen 3.7.8 variant calling



Dragen 3.7.8
(single sample)
*.hard-filtered.gvcf.gz

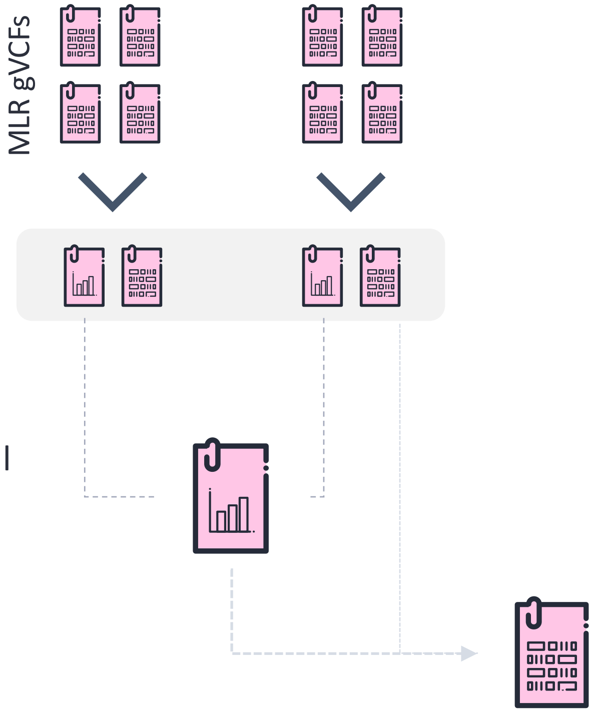
Dragen 3.7.8
(single sample)
*.cram

*.hard-filtered.recal.gvcf.gz

https://re-docs.genomicsengland.co.uk/dragen_378/

Iterative aggregation

Aggregation in batches of 1,000 MLR gVCFs into intermediate ‘**consensus**’ and ‘**cohort**’ files



Batch ‘consensus’ and ‘cohort’ files

Generation of global ‘census’

Generation of msVCF’s per subshard (multiallelic)

https://re-docs.genomicsengland.co.uk/aggv3_data_overview/

Multiallelic and biallelic VCFs

Multiallelic VCFs

```
#CHROM    POS    ID   REF    ALT    QUAL    FILTER    INFO     FORMAT    SAMPLE1    SAMPLE2    SAMPLE3
chr20    123456789    chr20:123456789:G:A,T   G    A,T    29    PASS     .    GT    0/0  1/0   1/2
```

Biallelic VCFs

```
#CHROM    POS    ID   REF    ALT    QUAL    FILTER    INFO     FORMAT    SAMPLE1    SAMPLE2    SAMPLE3
chr20    123456789    chr20:123456789:G:A   G    A    29    PASS     .    GT    0/0  1/0   1/.
chr20    123456789    chr20:123456789:G:T   G    T    29    PASS     .    GT    0/0  0/0   ./1
```

- Missingness is converted to REF when LAD ≥ 4
- Decomposed missing sites are converted to REF

https://re-docs.genomicsengland.co.uk/aggv3_data_overview/#multiallelic-vs-biallelic

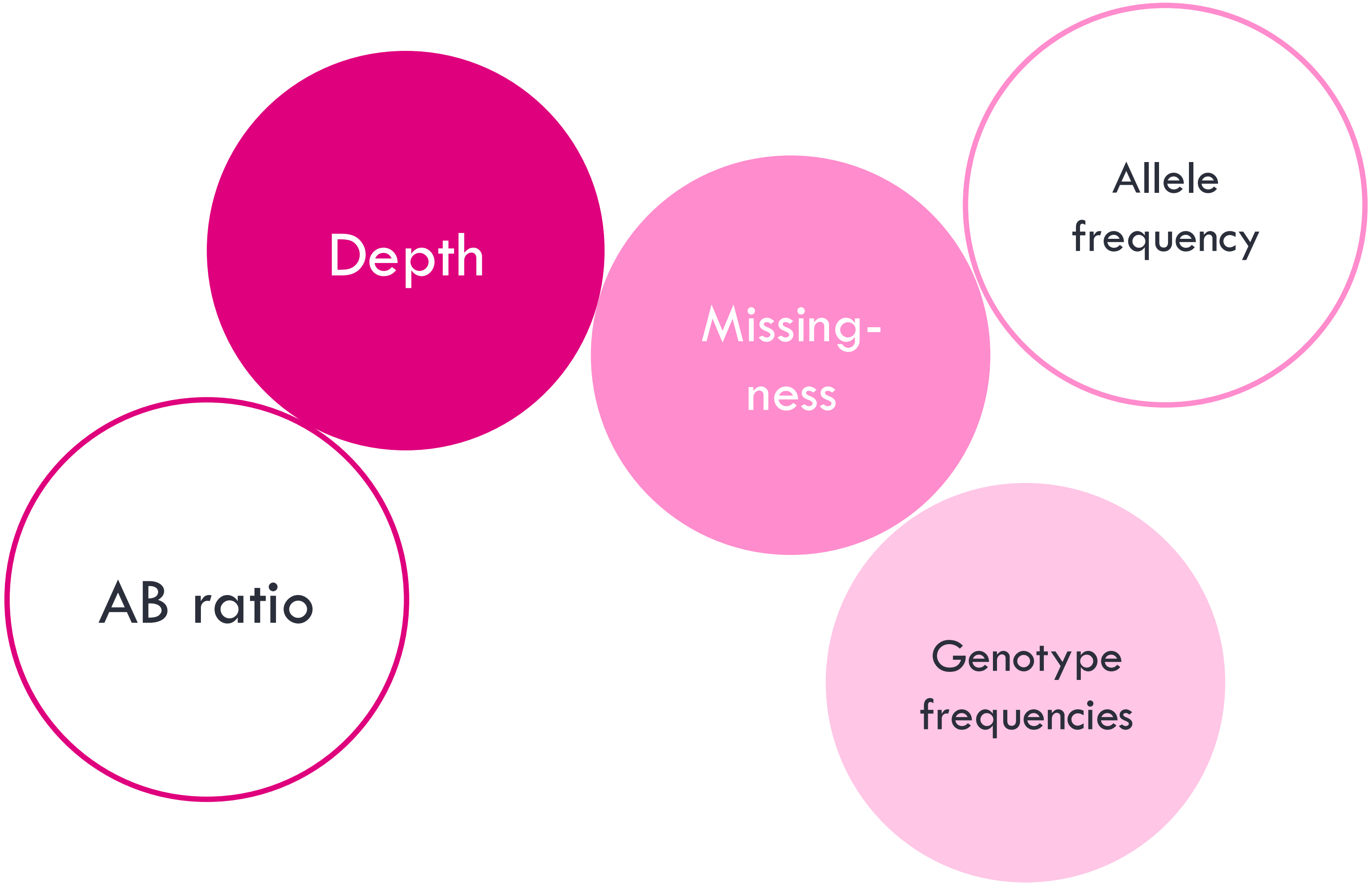
Functional annotation VCFs

- Gene consequences
- Pathogenicity scores
- AlphaMissense
- Splicing predictors
- Known genetic variants
- Global allele frequencies

115

https://re-docs.genomicsengland.co.uk/aggv3_functional_annotation/

Site QC VCFs



Depth

Missing-ness

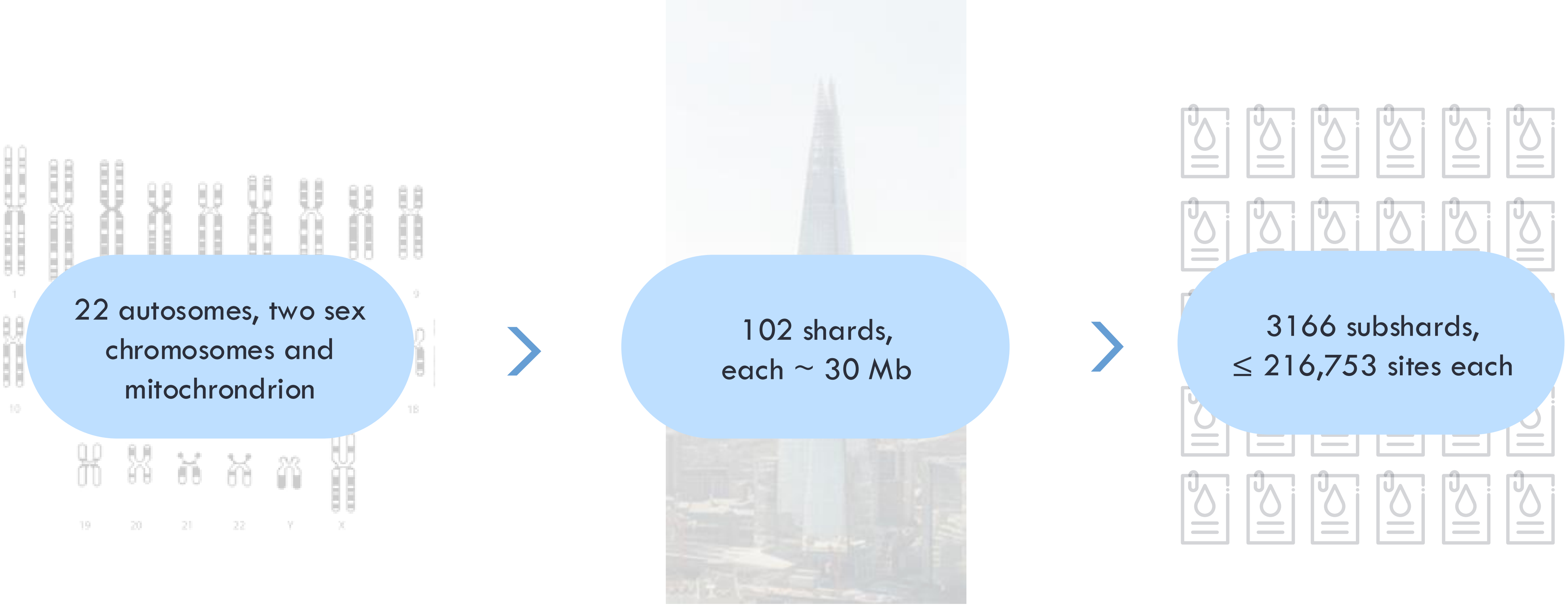
Allele frequency

AB ratio

Genotype frequencies

https://re-docs.genomicsengland.co.uk/aggv3_site_qc/

Sharding



22 autosomes, two sex chromosomes and mitochrondrion

102 shards, each ~ 30 Mb

3166 subshards, ≤ 216,753 sites each

Same shards used across the genotype, functional annotation and site QC VCFs

https://re-docs.genomicsengland.co.uk/aggv3_data_overview/#sharding

AggV3 vs AggV2

More samples

138,399 compared to 78,195, including 100kGP, NHS GMS and COVID-19

Chunks vs shards

Consistent sizing means it’s easier to estimate compute needs of large tasks/pipelines

CloudOS

Only available on CloudOS with no plans to change. Time to get ready for the future Cloud-based RE

Iterative build

We can add more genomes to it as we receive them (>40k at a time to make it viable).

DRAGEN 3.7.8

All genomes aligned and called using the same pipeline - better comparison within RE and to other similar projects

AggGIAB

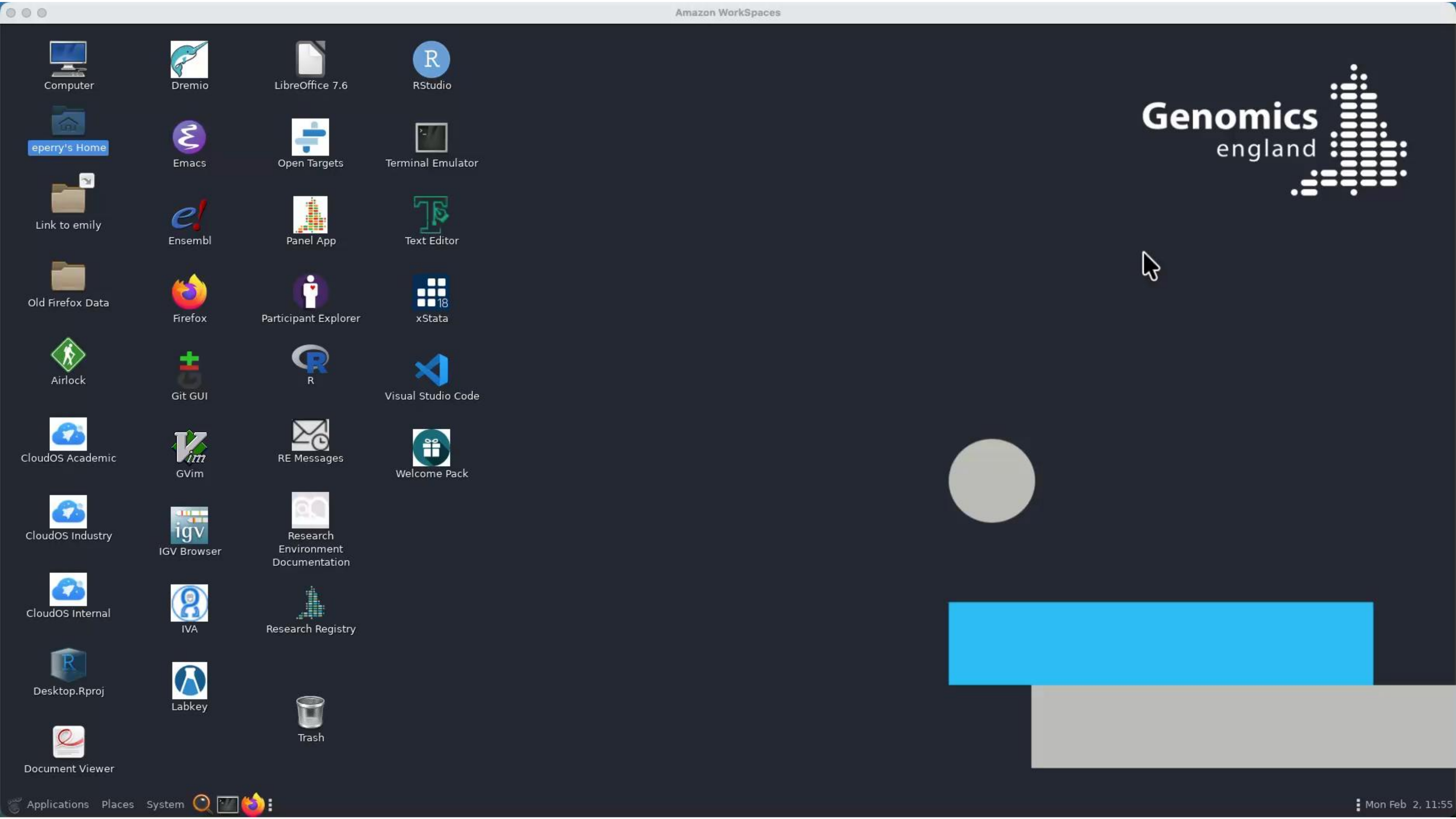
Smaller aggregate for testing your workflows

Coming soon…

In progress:
Population structure and relatedness
HQSNPs

Planned:
Mendellian inconsistencies
UPD cases
Hardy-Weinberg equilibria
Allele frequencies
SiteQC FILTERs

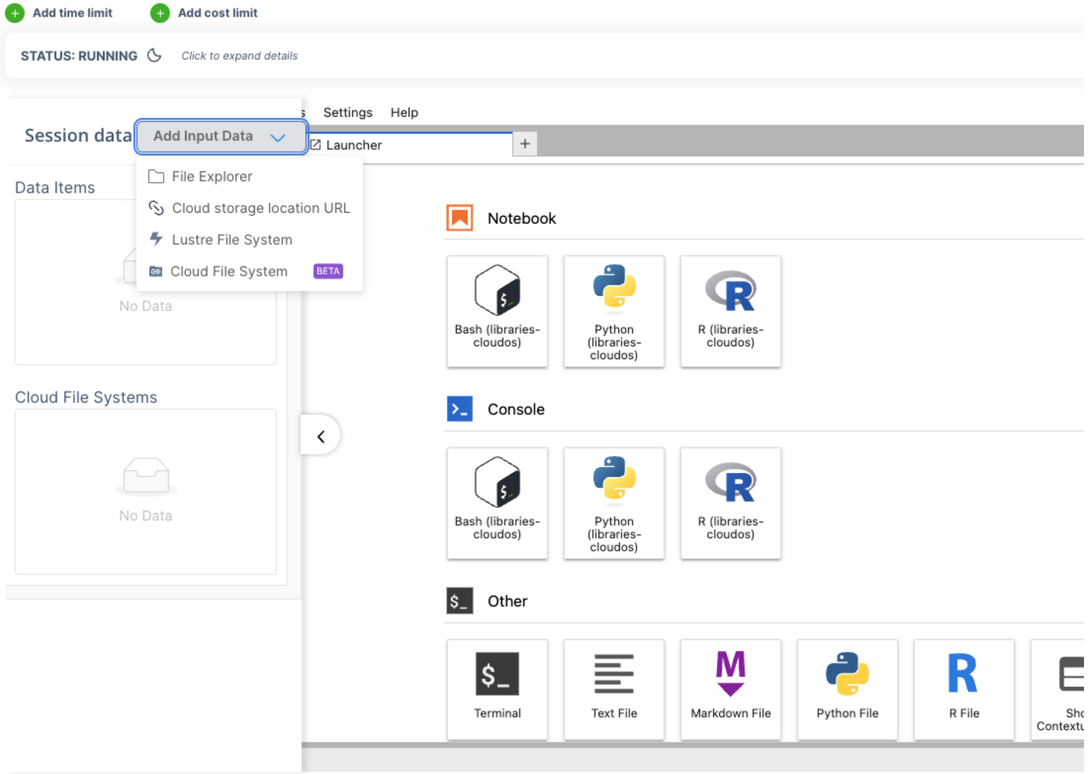
AggV3 data in CloudOS demo


Amazon WorkSpaces
Computer
eperry's Home
Link to emily
Old Firefox Data
Airlock
CloudOS Academic
CloudOS Industry
CloudOS Internal
Desktop.Rproj
Document Viewer
Dremio
Emacs
Ensembl
Firefox
Git GUI
GVim
IGV Browser
IVA
Labkey
LibreOffice 7.6
Open Targets
Panel App
Participant Explorer
R
RE Messages
Research Environment Documentation
Research Registry
Trash
RStudio
Terminal Emulator
Text Editor
xStata
Visual Studio Code
Welcome Pack
Genomics england
Applications
Places
System
Mon Feb 2, 11:55

3. Interactive sessions in CloudOS

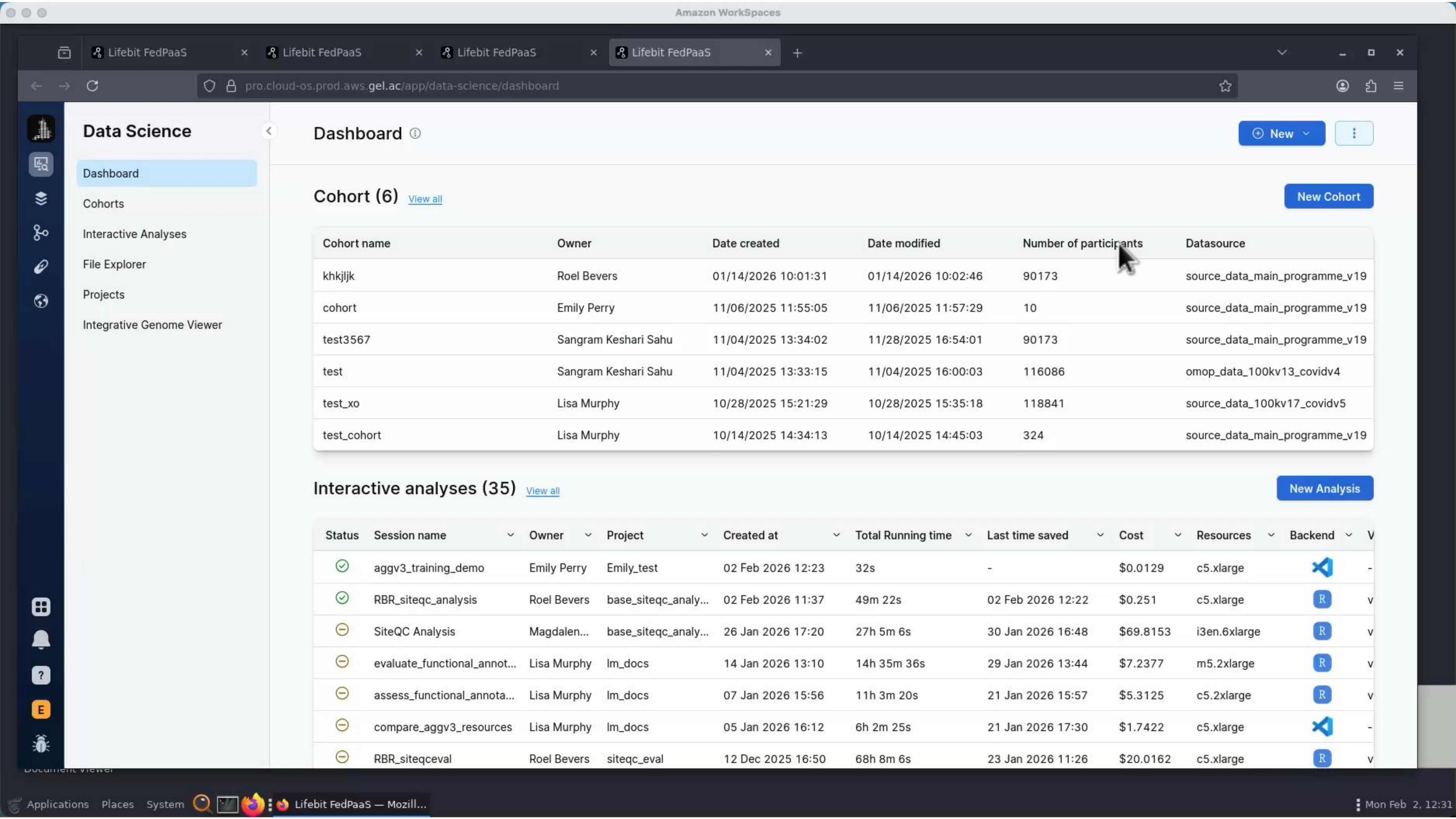
Interactive sessions

- Set time and cost limits
- Choose between VScode, Jupyter and RStudio
- Install the packages you want
- Docker containers and Nextflow pipelines
- Save Snapshots of your session
- Access data programmatically
- Cloud File System for fast linking of large data files



https://lifebit.atlassian.net/wiki/spaces/CD/pages/813105781/Run+a+Jupyter+Notebook+session

Interactive sessions demo



Dashboard

Cohort (6) View all

New Cohort

Cohort name	Owner	Date created	Date modified	Number of participants	Datasource
khkjljk	Roel Bevers	01/14/2026 10:01:31	01/14/2026 10:02:46	90173	source_data_main_programme_v19
cohort	Emily Perry	11/06/2025 11:55:05	11/06/2025 11:57:29	10	source_data_main_programme_v19
test3567	Sangram Keshari Sahu	11/04/2025 13:34:02	11/28/2025 16:54:01	90173	source_data_main_programme_v19
test	Sangram Keshari Sahu	11/04/2025 13:33:15	11/04/2025 16:00:03	116086	omop_data_100kv13_covidv4
test_xo	Lisa Murphy	10/28/2025 15:21:29	10/28/2025 15:35:18	118841	source_data_100kv17_covidv5
test_cohort	Lisa Murphy	10/14/2025 14:34:13	10/14/2025 14:45:03	324	source_data_main_programme_v19

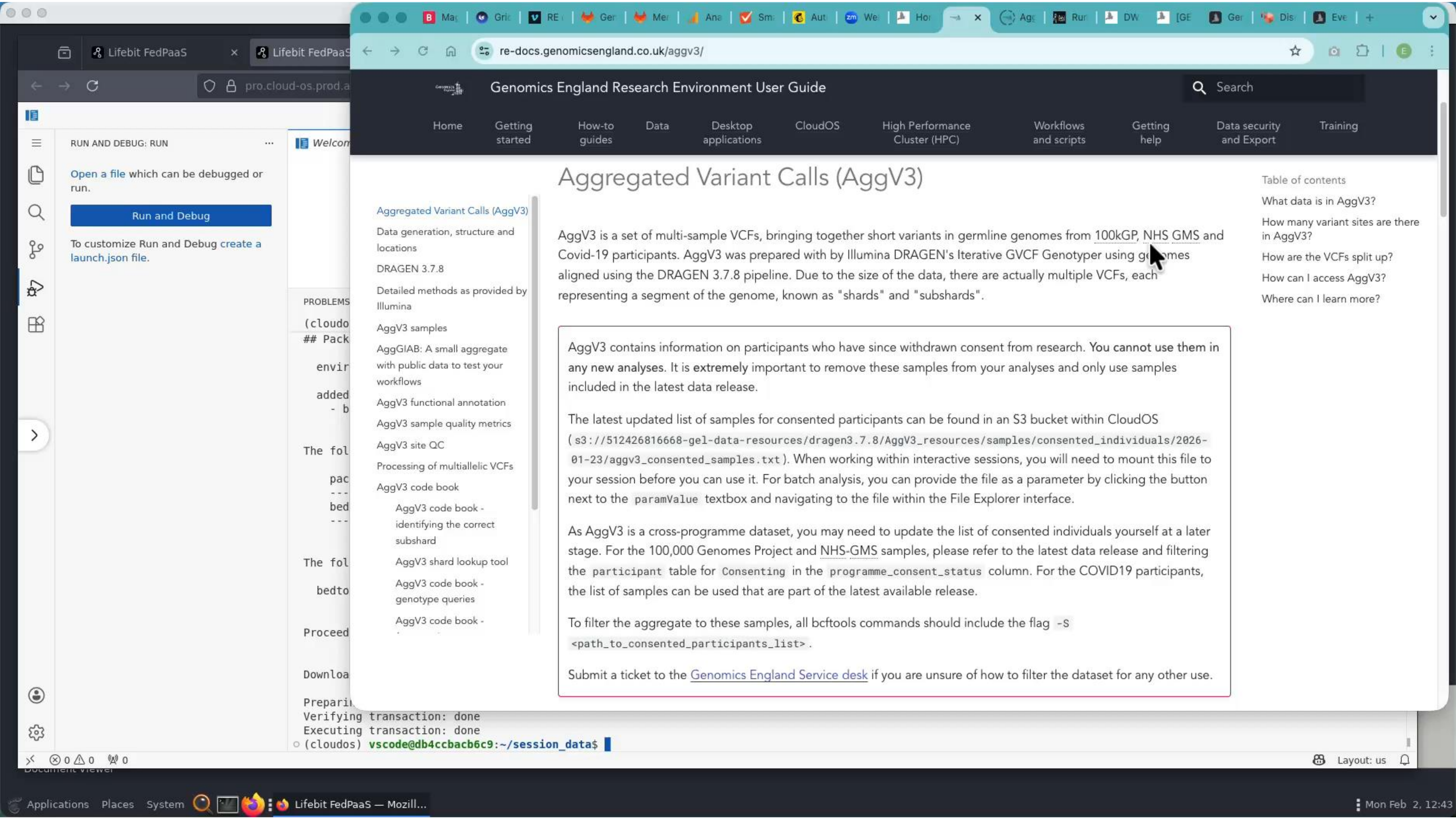
Interactive analyses (35) View all

New Analysis

Status	Session name	Owner	Project	Created at	Total Running time	Last time saved	Cost	Resources	Backend	V
✓	aggv3_training_demo	Emily Perry	Emily_test	02 Feb 2026 12:23	32s	-	$0.0129	c5.xlarge		-
✓	RBR_siteqc_analysis	Roel Bevers	base_siteqc_analy...	02 Feb 2026 11:37	49m 22s	02 Feb 2026 12:22	$0.251	c5.xlarge	R	v
⊖	SiteQC Analysis	Magdalen...	base_siteqc_analy...	26 Jan 2026 17:20	27h 5m 6s	30 Jan 2026 16:48	$69.8153	i3en.6xlarge	R	v
⊖	evaluate_functional_annot...	Lisa Murphy	lm_docs	14 Jan 2026 13:10	14h 35m 36s	29 Jan 2026 13:44	$7.2377	m5.2xlarge	R	v
⊖	assess_functional_annota...	Lisa Murphy	lm_docs	07 Jan 2026 15:56	11h 3m 20s	21 Jan 2026 15:57	$5.3125	c5.2xlarge	R	v
⊖	compare_aggv3_resources	Lisa Murphy	lm_docs	05 Jan 2026 16:12	6h 2m 25s	21 Jan 2026 17:30	$1.7422	c5.xlarge		-
⊖	RBR_siteqceval	Roel Bevers	siteqc_eval	12 Dec 2025 16:50	68h 8m 6s	23 Jan 2026 11:26	$20.0162	c5.xlarge	R	v

4. Querying AggV3 in the terminal

Shards demo



Aggregated Variant Calls (AggV3)

AggV3 is a set of multi-sample VCFs, bringing together short variants in germline genomes from 100kGP, NHS GMS and Covid-19 participants. AggV3 was prepared with by Illumina DRAGEN's Iterative GVCF Genotyper using genomes aligned using the DRAGEN 3.7.8 pipeline. Due to the size of the data, there are actually multiple VCFs, each representing a segment of the genome, known as "shards" and "subshards".

AggV3 contains information on participants who have since withdrawn consent from research. **You cannot use them in any new analyses.** It is **extremely** important to remove these samples from your analyses and only use samples included in the latest data release.

The latest updated list of samples for consented participants can be found in an S3 bucket within CloudOS (`s3://512426816668-gel-data-resources/dragen3.7.8/AggV3_resources/samples/consented_individuals/2026-01-23/aggv3_consented_samples.txt`). When working within interactive sessions, you will need to mount this file to your session before you can use it. For batch analysis, you can provide the file as a parameter by clicking the button next to the `paramValue` textbox and navigating to the file within the File Explorer interface.

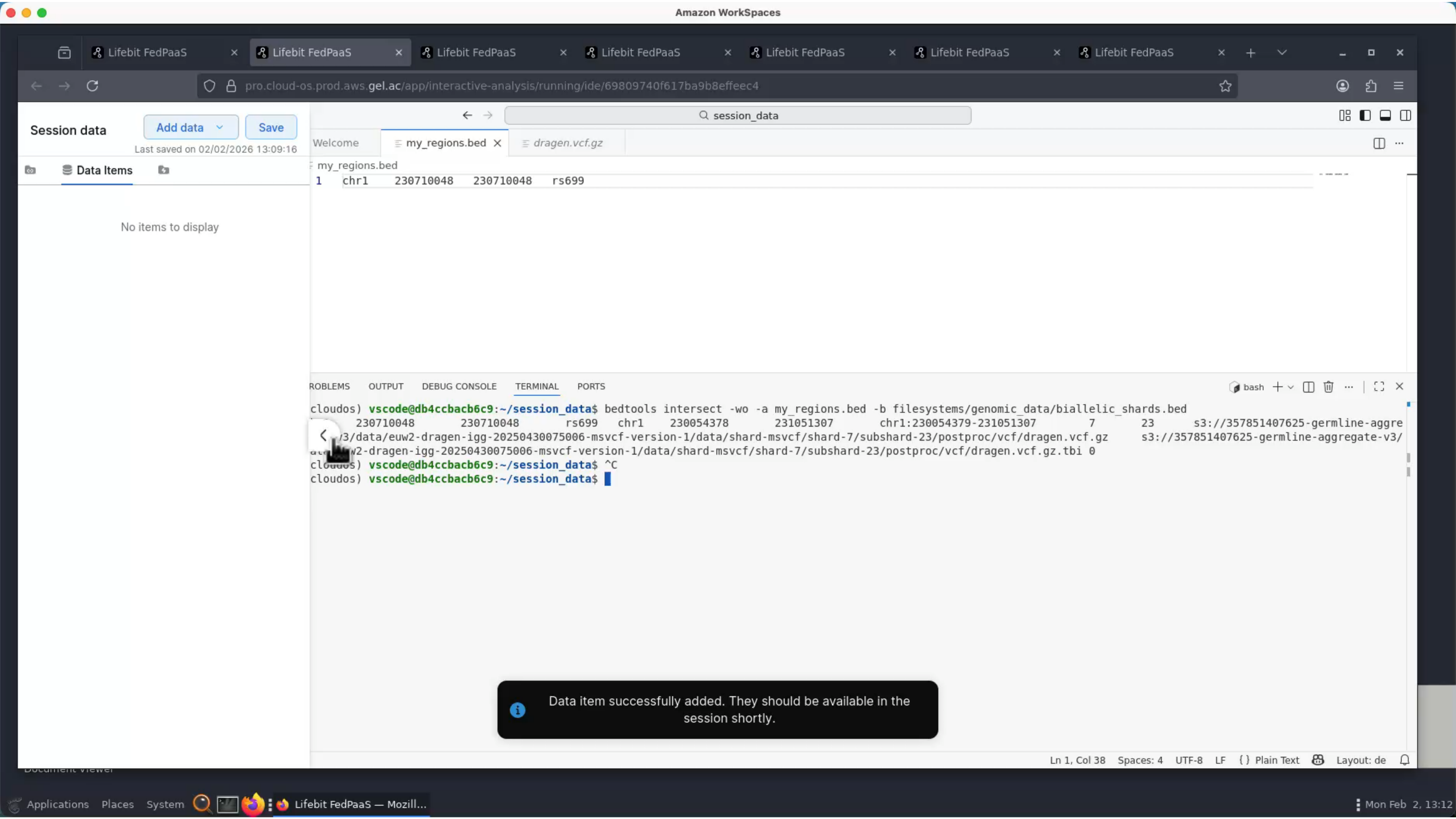
As AggV3 is a cross-programme dataset, you may need to update the list of consented individuals yourself at a later stage. For the 100,000 Genomes Project and NHS-GMS samples, please refer to the latest data release and filtering the `participant` table for `Consenting` in the `programme_consent_status` column. For the COVID19 participants, the list of samples can be used that are part of the latest available release.

To filter the aggregate to these samples, all bcftools commands should include the flag `-S <path_to_consented_participants_list>`.

Submit a ticket to the Genomics England Service desk if you are unsure of how to filter the dataset for any other use.

```
Verifying transaction: done
Executing transaction: done
(cloudos) vscode@db4ccbacb6c9:~/session_data$
```

Genotypes demo



Last saved on 02/02/2026 13:09:16

No items to display

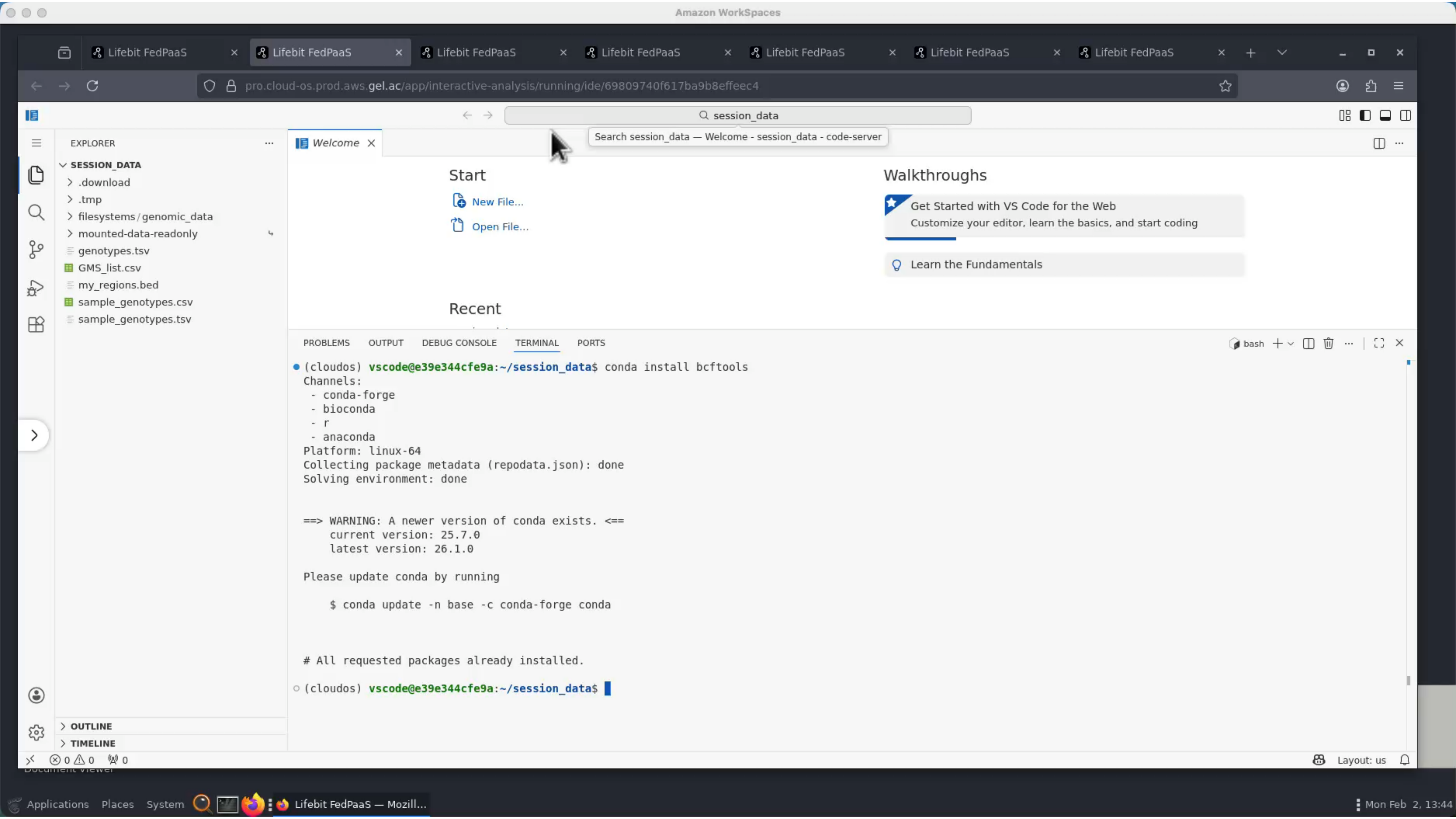
my_regions.bed

```
chr1	230710048	230710048	rs699
```

```
cloudos) vscode@db4ccbacb6c9:~/session_data$ bedtools intersect -wo -a my_regions.bed -b filesystems/genomic_data/biallelic_shards.bed
         230710048        230710048        rs699   chr1    230054378        231051307        chr1:230054379-231051307        7       23      s3://357851407625-germline-aggre
   /3/data/euw2-dragen-igg-20250430075006-msvcf-version-1/data/shard-msvcf/shard-7/subshard-23/postproc/vcf/dragen.vcf.gz      s3://357851407625-germline-aggregate-v3/
   w2-dragen-igg-20250430075006-msvcf-version-1/data/shard-msvcf/shard-7/subshard-23/postproc/vcf/dragen.vcf.gz.tbi 0
cl    s) vscode@db4ccbacb6c9:~/session_data$ ^C
cloudos) vscode@db4ccbacb6c9:~/session_data$
```

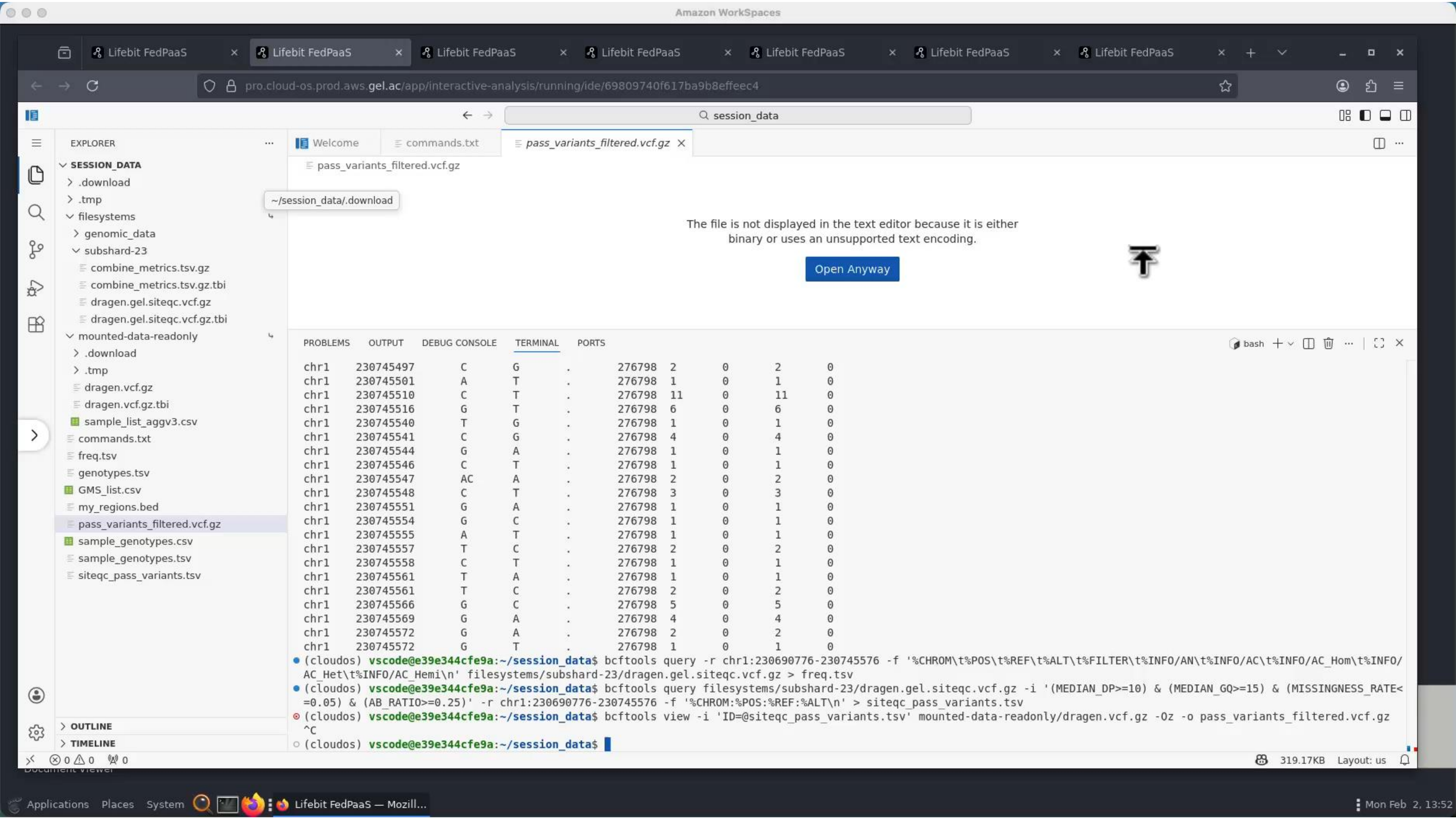
Data item successfully added. They should be available in the session shortly.

QC demo



```
(cloudos) vscode@e39e344cfe9a:~/session_data$ conda install bcftools
Channels:
 - conda-forge
 - bioconda
 - r
 - anaconda
Platform: linux-64
Collecting package metadata (repodata.json): done
Solving environment: done


==> WARNING: A newer version of conda exists. <==
    current version: 25.7.0
    latest version: 26.1.0

Please update conda by running

    $ conda update -n base -c conda-forge conda



# All requested packages already installed.

(cloudos) vscode@e39e344cfe9a:~/session_data$
```

Functional annotation demo



The file is not displayed in the text editor because it is either binary or uses an unsupported text encoding.

```
chr1    230745497       C       G       .       276798  2       0       2       0
chr1    230745501       A       T       .       276798  1       0       1       0
chr1    230745510       C       T       .       276798  11      0       11      0
chr1    230745516       G       T       .       276798  6       0       6       0
chr1    230745540       T       G       .       276798  1       0       1       0
chr1    230745541       C       G       .       276798  4       0       4       0
chr1    230745544       G       A       .       276798  1       0       1       0
chr1    230745546       C       T       .       276798  1       0       1       0
chr1    230745547       AC      A       .       276798  2       0       2       0
chr1    230745548       C       T       .       276798  3       0       3       0
chr1    230745551       G       A       .       276798  1       0       1       0
chr1    230745554       G       C       .       276798  1       0       1       0
chr1    230745555       A       T       .       276798  1       0       1       0
chr1    230745557       T       C       .       276798  2       0       2       0
chr1    230745558       C       T       .       276798  1       0       1       0
chr1    230745561       T       A       .       276798  1       0       1       0
chr1    230745561       T       C       .       276798  2       0       2       0
chr1    230745566       G       C       .       276798  5       0       5       0
chr1    230745569       G       A       .       276798  4       0       4       0
chr1    230745572       G       A       .       276798  2       0       2       0
chr1    230745572       G       T       .       276798  1       0       1       0
(cloudos) vscode@e39e344cfe9a:~/session_data$ bcftools query -r chr1:230690776-230745576 -f '%CHROM\t%POS\t%REF\t%ALT\t%FILTER\t%INFO/AN\t%INFO/AC\t%INFO/AC_Hom\t%INFO/AC_Het\t%INFO/AC_Hemi\n' filesystems/subshard-23/dragen.gel.siteqc.vcf.gz > freq.tsv
(cloudos) vscode@e39e344cfe9a:~/session_data$ bcftools query filesystems/subshard-23/dragen.gel.siteqc.vcf.gz -i '(MEDIAN_DP>=10) & (MEDIAN_GQ>=15) & (MISSINGNESS_RATE<=0.05) & (AB_RATIO>=0.25)' -r chr1:230690776-230745576 -f '%CHROM:%POS:%REF:%ALT\n' > siteqc_pass_variants.tsv
(cloudos) vscode@e39e344cfe9a:~/session_data$ bcftools view -i 'ID=@siteqc_pass_variants.tsv' mounted-data-readonly/dragen.vcf.gz -Oz -o pass_variants_filtered.vcf.gz
^C
(cloudos) vscode@e39e344cfe9a:~/session_data$
```

5. Using AggV3 in batch jobs

Batch Queues

Choose from pre-configured environments

Configure compute resources

vCPUs

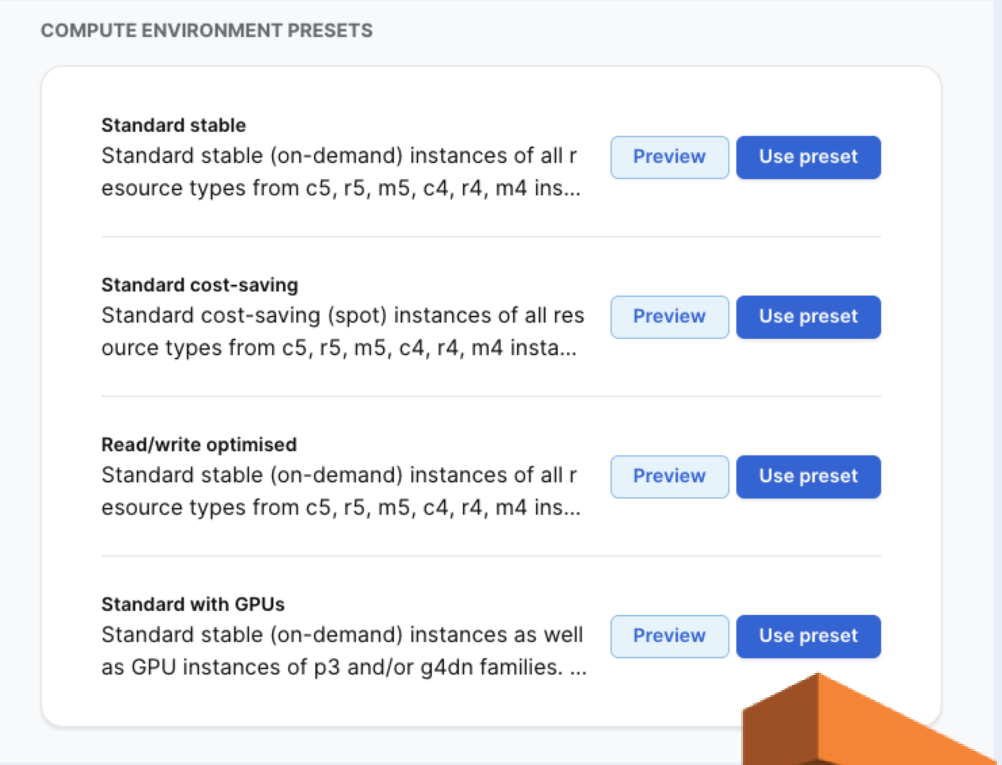
Spot

IOPS

Volume type

On-demand

Optimise for specific tasks



COMPUTE ENVIRONMENT PRESETS

Standard stable
Standard stable (on-demand) instances of all resource types from c5, r5, m5, c4, r4, m4 ins...

Standard cost-saving
Standard cost-saving (spot) instances of all resource types from c5, r5, m5, c4, r4, m4 insta...

Read/write optimised
Standard stable (on-demand) instances of all resource types from c5, r5, m5, c4, r4, m4 ins...

Standard with GPUs
Standard stable (on-demand) instances as well as GPU instances of p3 and/or g4dn families. ...

AWS Batch

https://lifebit.atlassian.net/wiki/spaces/CD/pages/816448106/Create+and+Monitor+Batch+Queues

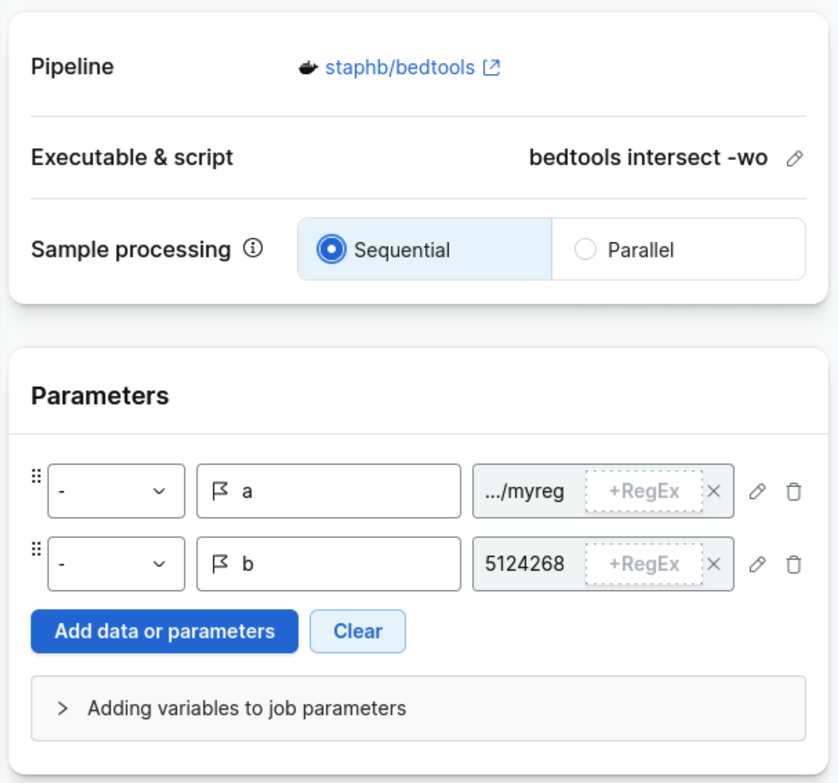
Bash with containers

Find (or build) a container with the package(s) you need

Write a bash script or command to use the packages

parameters – s3 buckets must be added as parameters, not written into your bash script

Can run bash script on a single input or in parallel using an input file



Pipeline: staphb/bedtools

Executable & script: bedtools intersect -wo

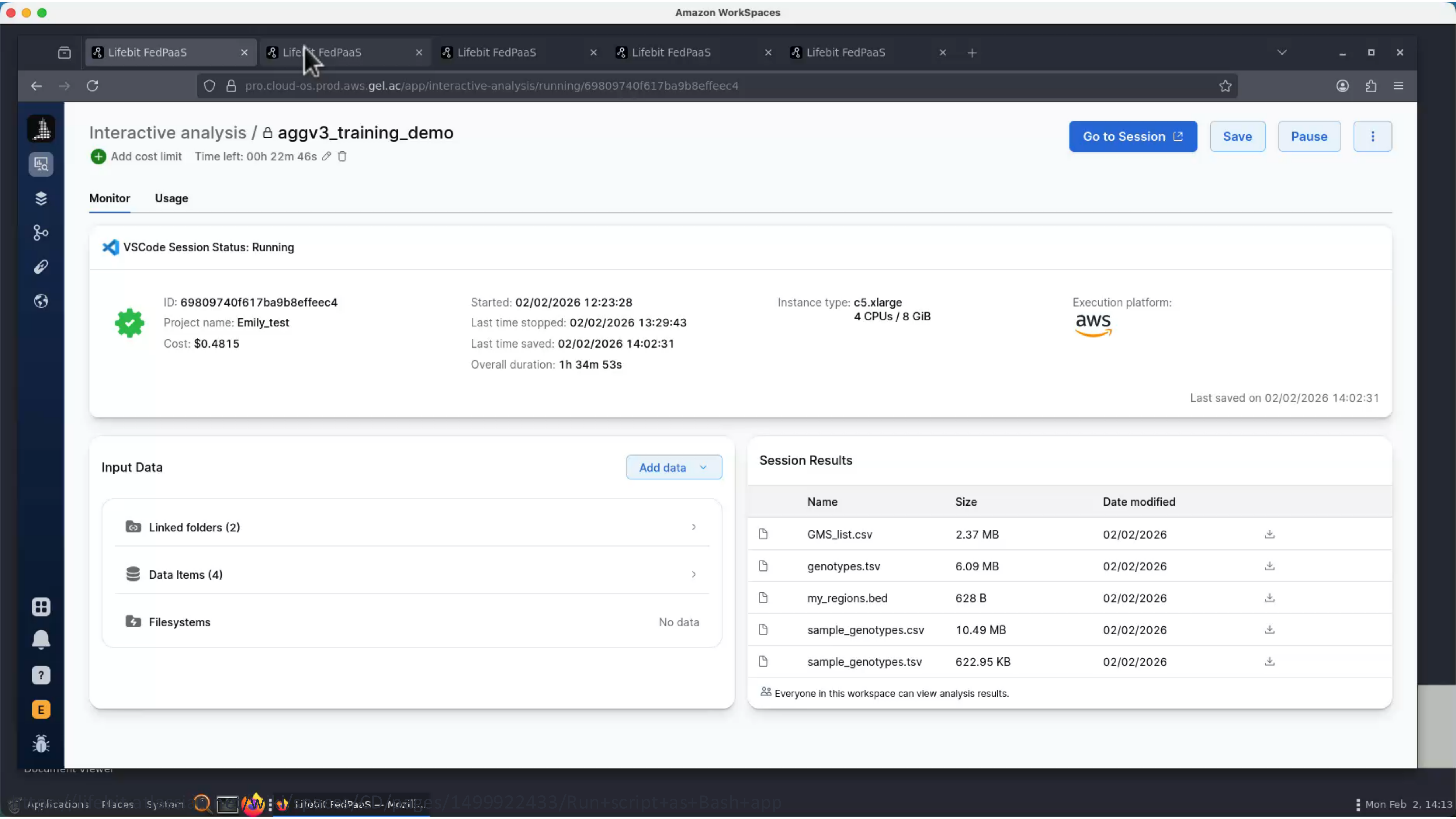
Sample processing: ● Sequential ○ Parallel

Parameters

-	a	.../myreg
-	b	5124268

https://lifebit.atlassian.net/wiki/spaces/CD/pages/1499922433/Run+script+as+Bash+app

Bash jobs demo



Interactive analysis / aggv3_training_demo

Time left: 00h 22m 46s

VSCode Session Status: Running

ID: 69809740f617ba9b8effeec4
Project name: Emily_test
Cost: $0.4815

Started: 02/02/2026 12:23:28
Last time stopped: 02/02/2026 13:29:43
Last time saved: 02/02/2026 14:02:31
Overall duration: 1h 34m 53s

Instance type: c5.xlarge
4 CPUs / 8 GiB

Execution platform: aws

Last saved on 02/02/2026 14:02:31

Input Data

Linked folders (2)

Data Items (4)

Filesystems No data

Session Results

Name	Size	Date modified
GMS_list.csv	2.37 MB	02/02/2026
genotypes.tsv	6.09 MB	02/02/2026
my_regions.bed	628 B	02/02/2026
sample_genotypes.csv	10.49 MB	02/02/2026
sample_genotypes.tsv	622.95 KB	02/02/2026

Everyone in this workspace can view analysis results.

Workflows

Build complex workflows in Nextflow

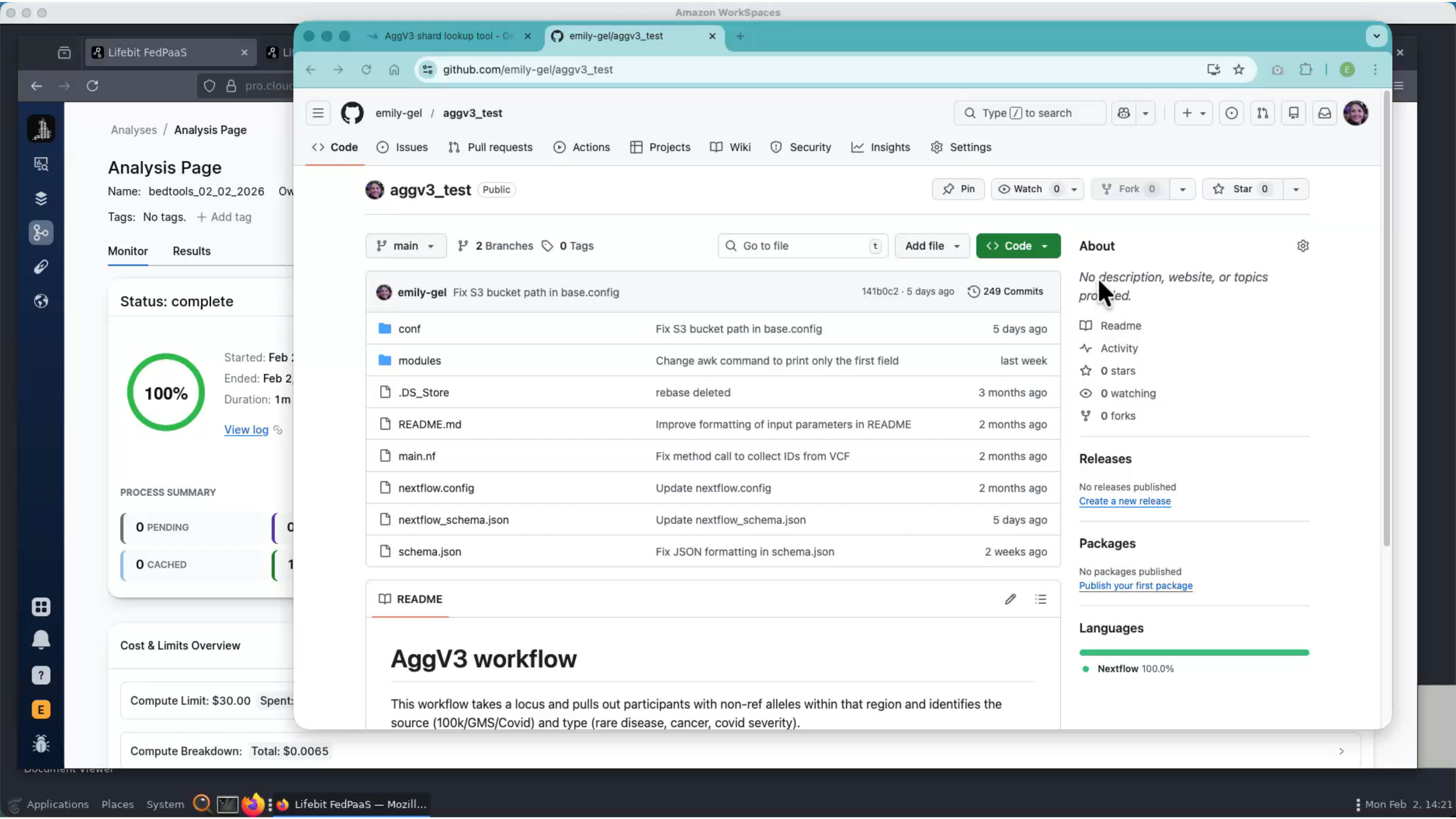
Pull from GitHub into CloudOS

Connect your Github account

Share workflows within your workspace

https://lifebit.atlassian.net/wiki/spaces/CD/pages/813040723/Run+a+pipeline https://training.nextflow.io/2.8.1/

Workflow demo



Amazon WorkSpaces

AggV3 shard lookup tool - Ge | emily-gel/aggv3_test

github.com/emily-gel/aggv3_test

emily-gel / aggv3_test

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

aggv3_test Public

Pin | Watch 0 | Fork 0 | Star 0

main | 2 Branches | 0 Tags | Go to file | Add file | Code

emily-gel Fix S3 bucket path in base.config 141b0c2 · 5 days ago 249 Commits

conf	Fix S3 bucket path in base.config	5 days ago
modules	Change awk command to print only the first field	last week
.DS_Store	rebase deleted	3 months ago
README.md	Improve formatting of input parameters in README	2 months ago
main.nf	Fix method call to collect IDs from VCF	2 months ago
nextflow.config	Update nextflow.config	2 months ago
nextflow_schema.json	Update nextflow_schema.json	5 days ago
schema.json	Fix JSON formatting in schema.json	2 weeks ago

README

AggV3 workflow

This workflow takes a locus and pulls out participants with non-ref alleles within that region and identifies the source (100k/GMS/Covid) and type (rare disease, cancer, covid severity).

About

No description, website, or topics provided.

Readme
Activity
0 stars
0 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Nextflow 100.0%

Lifebit FedPaaS

Analyses / Analysis Page

Analysis Page

Name: bedtools_02_02_2026 Ow

Tags: No tags. + Add tag

Monitor | Results

Status: complete

100%

Started: Feb
Ended: Feb 2
Duration: 1m

View log

PROCESS SUMMARY

0 PENDING
0 CACHED

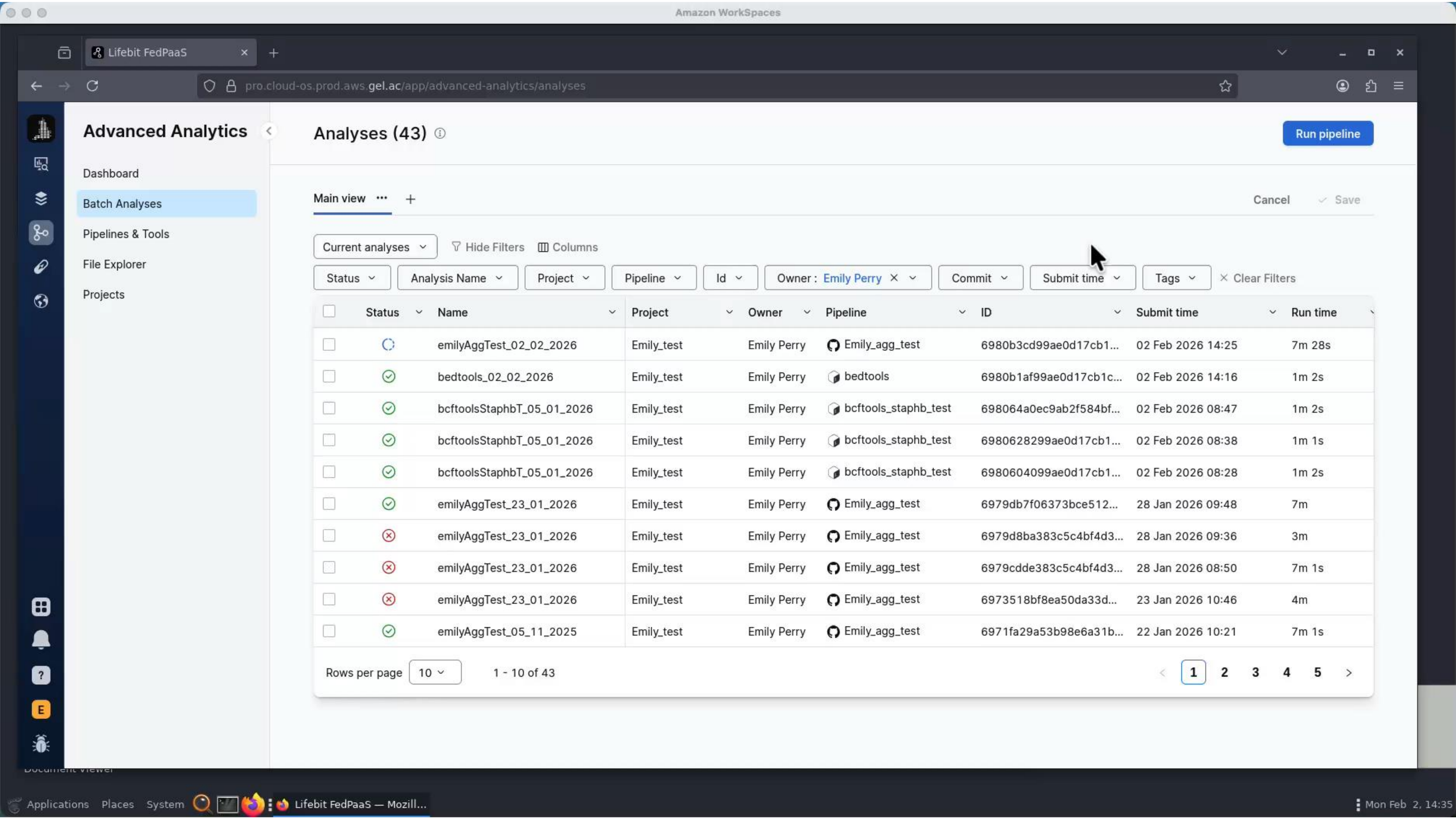
Cost & Limits Overview

Compute Limit: $30.00 Spent:

Compute Breakdown: Total: $0.0065

6. Taking data in and out of CloudOS

Export demo



Analyses (43)

Owner : Emily Perry

☐	Status	Name	Project	Owner	Pipeline	ID	Submit time	Run time
☐	◌	emilyAggTest_02_02_2026	Emily_test	Emily Perry	Emily_agg_test	6980b3cd99ae0d17cb1...	02 Feb 2026 14:25	7m 28s
☐	✓	bedtools_02_02_2026	Emily_test	Emily Perry	bedtools	6980b1af99ae0d17cb1c...	02 Feb 2026 14:16	1m 2s
☐	✓	bcftoolsStaphbT_05_01_2026	Emily_test	Emily Perry	bcftools_staphb_test	698064a0ec9ab2f584bf...	02 Feb 2026 08:47	1m 2s
☐	✓	bcftoolsStaphbT_05_01_2026	Emily_test	Emily Perry	bcftools_staphb_test	6980628299ae0d17cb1...	02 Feb 2026 08:38	1m 1s
☐	✓	bcftoolsStaphbT_05_01_2026	Emily_test	Emily Perry	bcftools_staphb_test	6980604099ae0d17cb1...	02 Feb 2026 08:28	1m 2s
☐	✓	emilyAggTest_23_01_2026	Emily_test	Emily Perry	Emily_agg_test	6979db7f06373bce512...	28 Jan 2026 09:48	7m
☐	⊗	emilyAggTest_23_01_2026	Emily_test	Emily Perry	Emily_agg_test	6979d8ba383c5c4bf4d3...	28 Jan 2026 09:36	3m
☐	⊗	emilyAggTest_23_01_2026	Emily_test	Emily Perry	Emily_agg_test	6979cdde383c5c4bf4d3...	28 Jan 2026 08:50	7m 1s
☐	⊗	emilyAggTest_23_01_2026	Emily_test	Emily Perry	Emily_agg_test	6973518bf8ea50da33d...	23 Jan 2026 10:46	4m
☐	✓	emilyAggTest_05_11_2025	Emily_test	Emily Perry	Emily_agg_test	6971fa29a53b98e6a31b...	22 Jan 2026 10:21	7m 1s

Rows per page 10

1 - 10 of 43

1 2 3 4 5

7. Getting help and questions

Get access to CloudOS

Get in touch via Service Desk

lifebit

https://www.genomicsengland.co.uk/service-desk

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

17/3

21/4

19/5

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

Materials from past training all online

Training sessions

14/4	Building cancer cohorts
12/5	Building rare disease cohorts

Materials from past training all online

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

Feedback

Thank you

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