Running ImmunoVerse on CGC

Guangyuan(Frank) Li, PhD
NYU Grossman School of Medicine
Last update: 2025.10.11

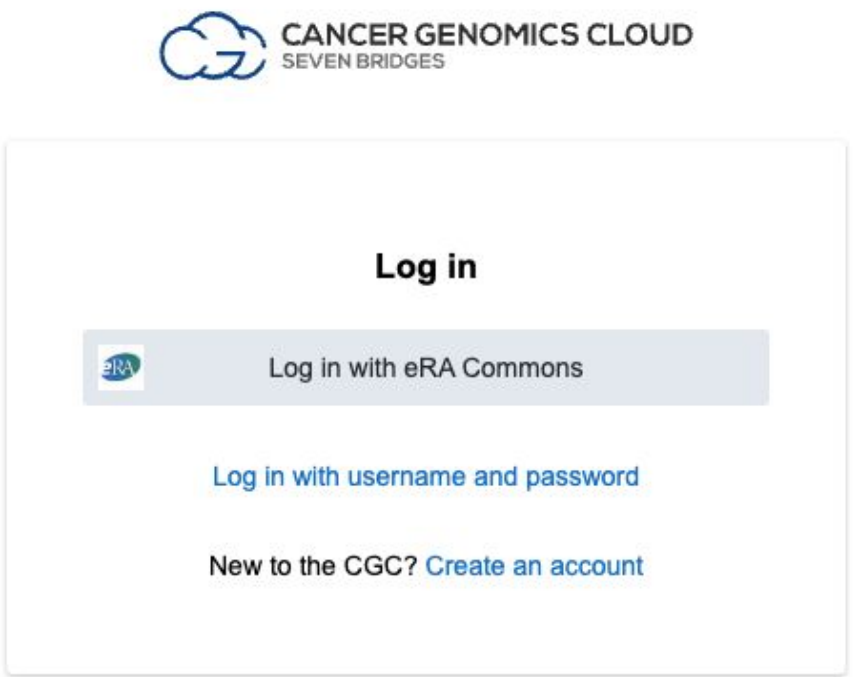
Overview

1. Without the need of writing single line of code, you can simply upload your RNA-Seq fastq.gz files, along with immunopeptidome raw files (if applicable) to the CGC platform, to (a) **profile genetic aberrations**, (b) **generate search space**, (c) **search immunopeptidome**, (d) more to come
2. Following sections (please always read **a** and **b** first, **d/e/f/g** requires **c** or provide your own bam file, **j** should be after **a-i**, **k** is optional you can use other immunopeptidome workflow):
    a. Set things up
    b. Gene expression
    c. Alignment
    d. Splicing and intron retention
    e. Pathogen
    f. Transposable element
    g. Variants
    h. Gene fusion
    i. HLA typing
    j. summarization
    k. Immunopeptidome analysis
    l. Advanced Usage 1 (predicting and evaluating spectra)

Set things up

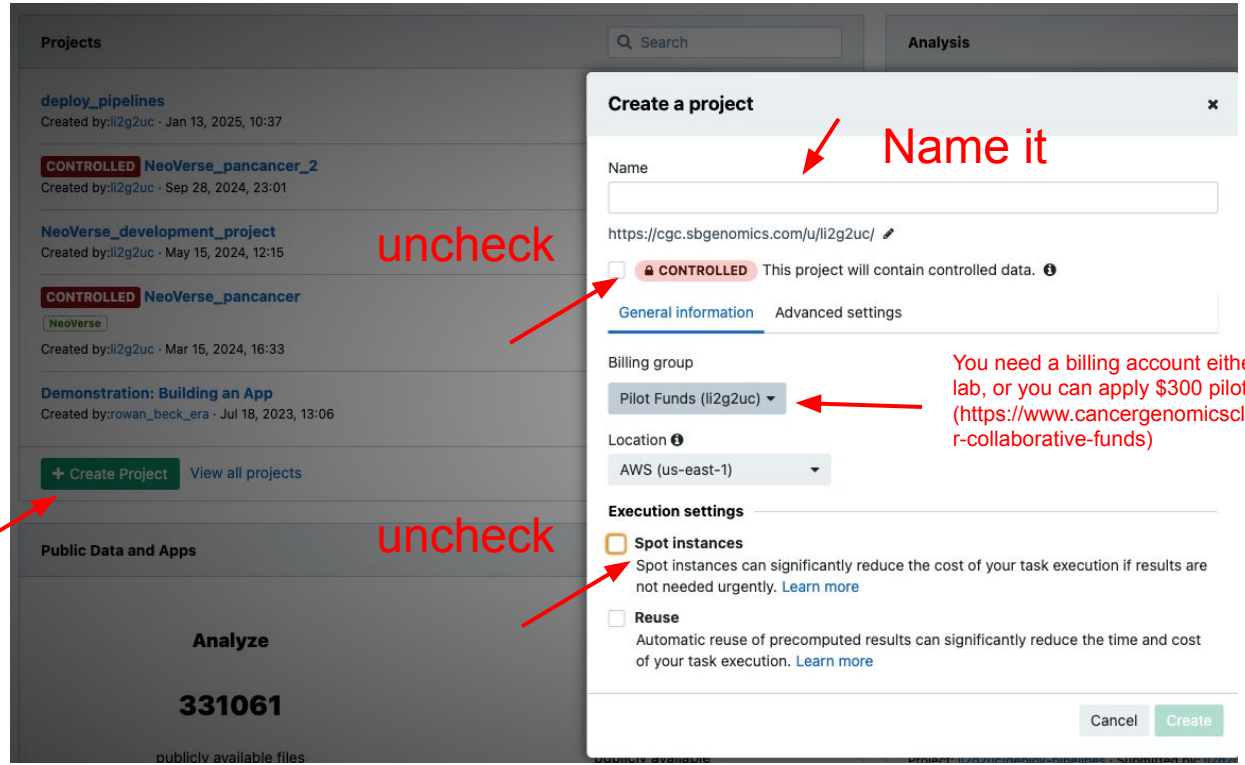
Step 1: Creating Account

1. Visit https://cgc.sbgenomics.com/home
2. eRA common should work just fine, otherwise, please follow the links to create an account using your email

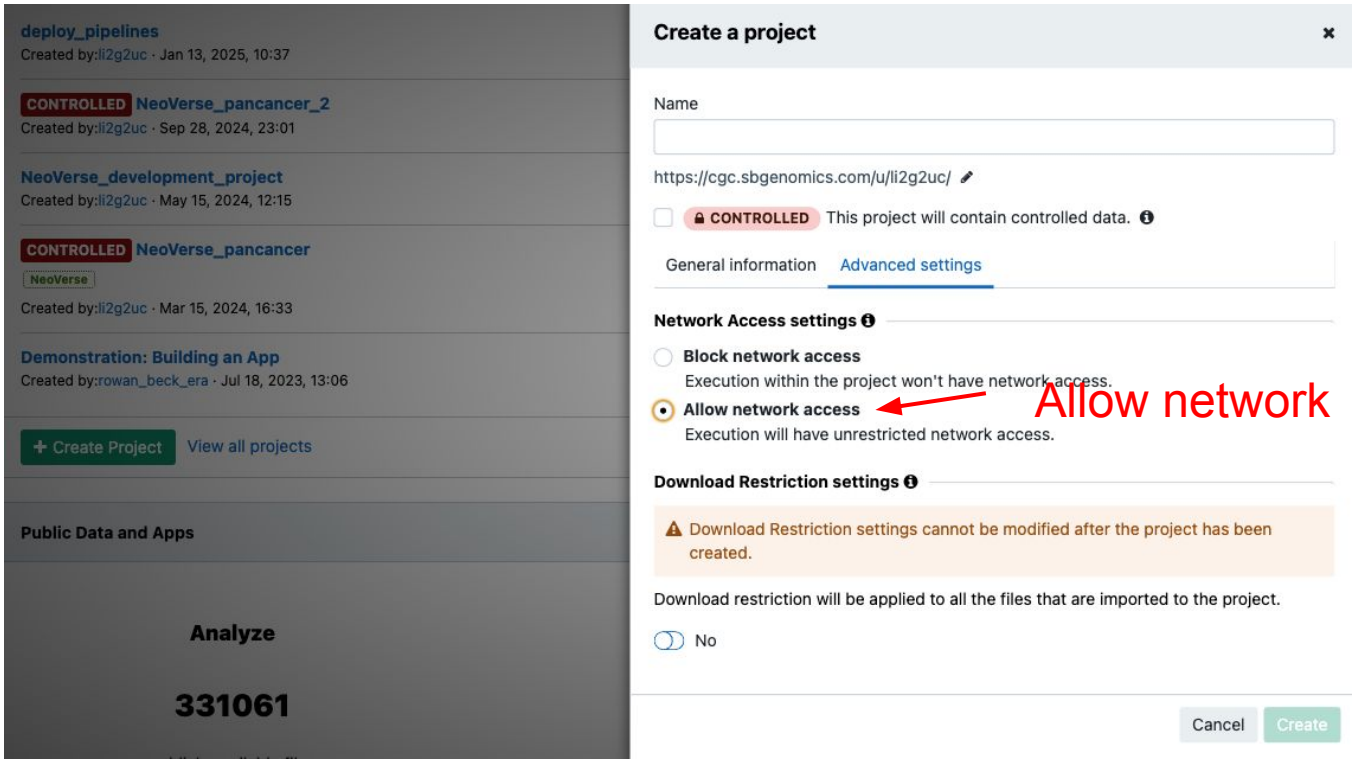


Step 2: Create a project

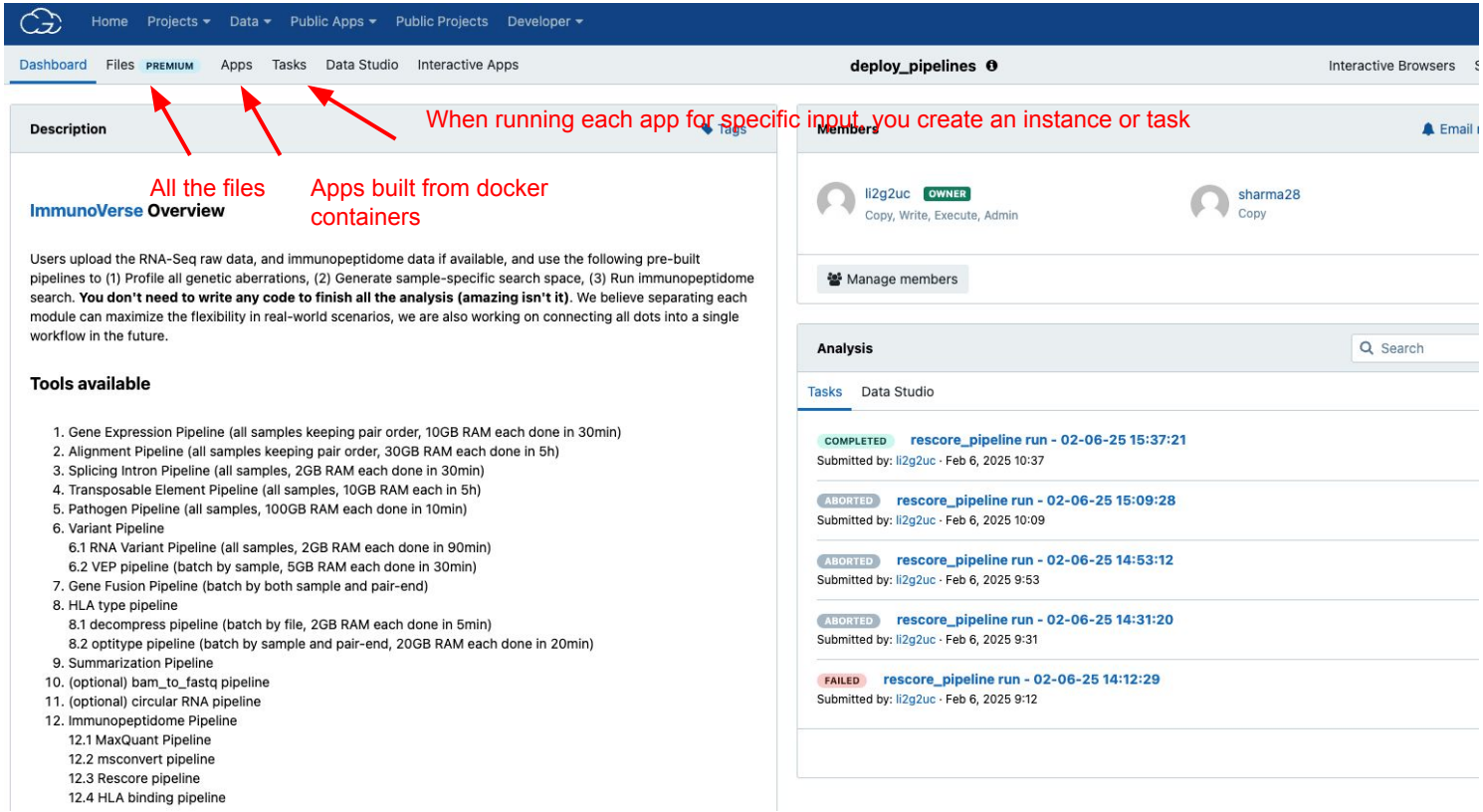
A **project** has all the (1) needed files, (2) workflows and (3) tasks being run



Step 2: Create a project (cont'd)

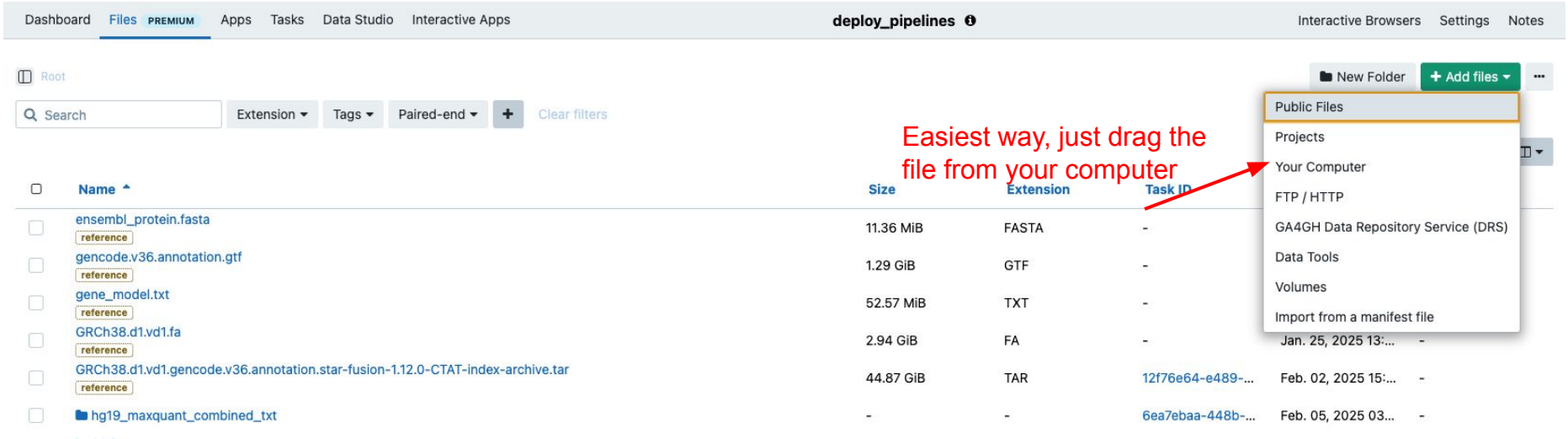


Step 3: Navigate your project

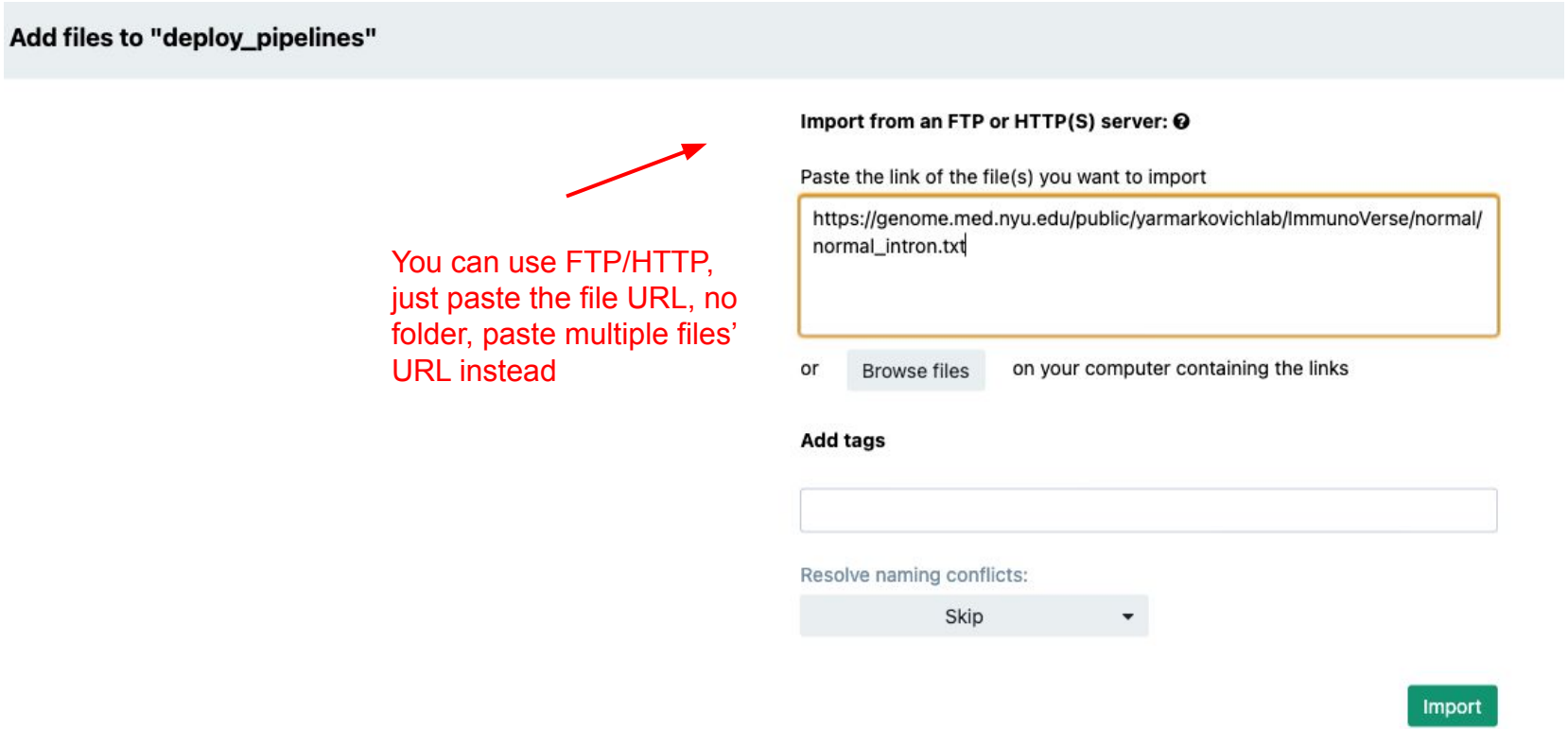


Upload and Download

Upload files

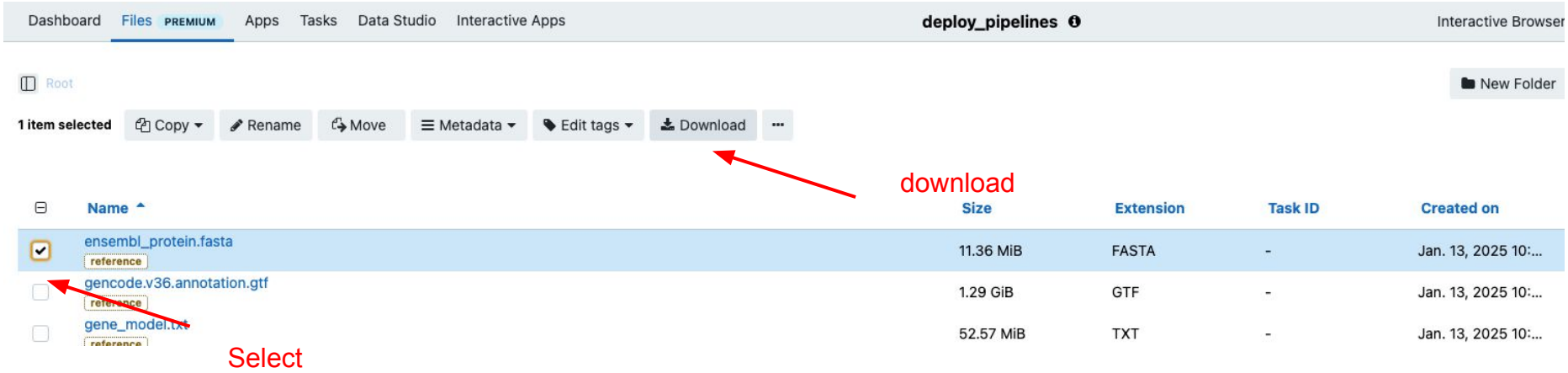


Upload files (cont'd)



You can use FTP/HTTP, just paste the file URL, no folder, paste multiple files' URL instead

Download

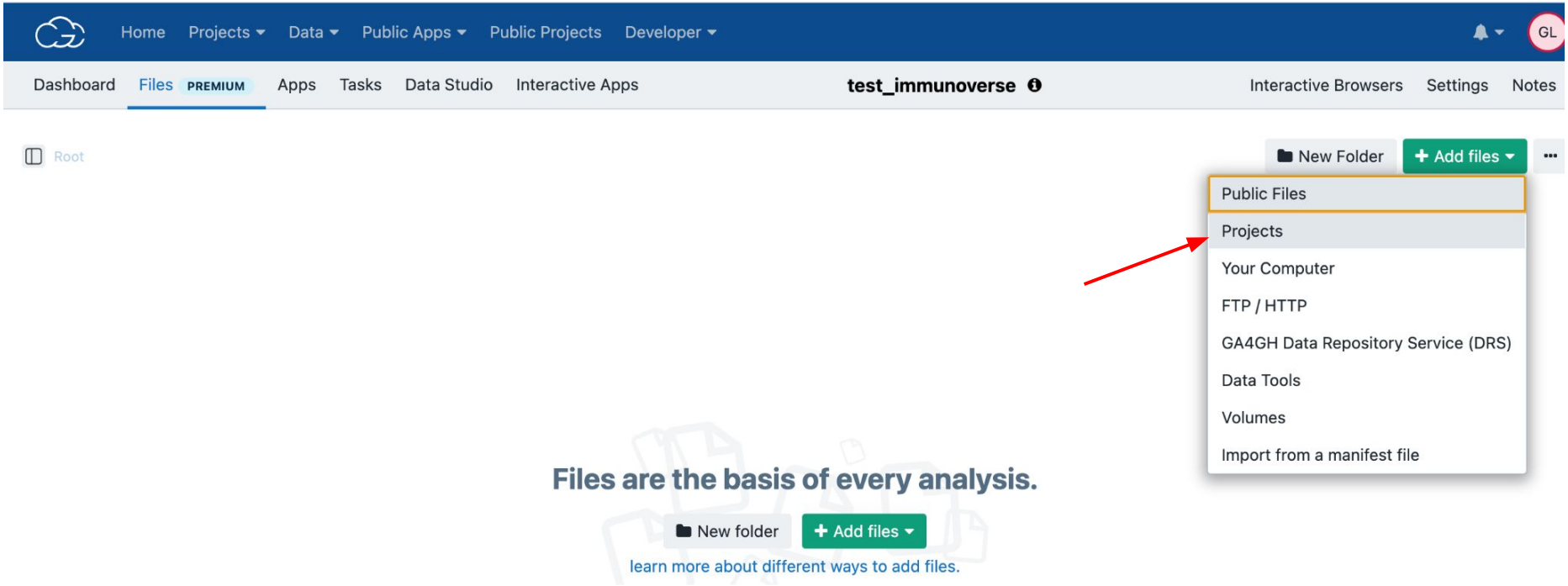

Dashboard
Files PREMIUM
Apps
Tasks
Data Studio
Interactive Apps
deploy_pipelines
Interactive Browser
Root
New Folder
1 item selected
Copy
Rename
Move
Metadata
Edit tags
Download
download
Name
Size
Extension
Task ID
Created on
ensembl_protein.fasta
reference
11.36 MiB
FASTA
-
Jan. 13, 2025 10:...
gencode.v36.annotation.gtf
reference
1.29 GiB
GTF
-
Jan. 13, 2025 10:...
gene_model.txt
reference
52.57 MiB
TXT
-
Jan. 13, 2025 10:...
Select

Now, Let's copy needed files and workflows from my project to your project (step-by-step after this slide)

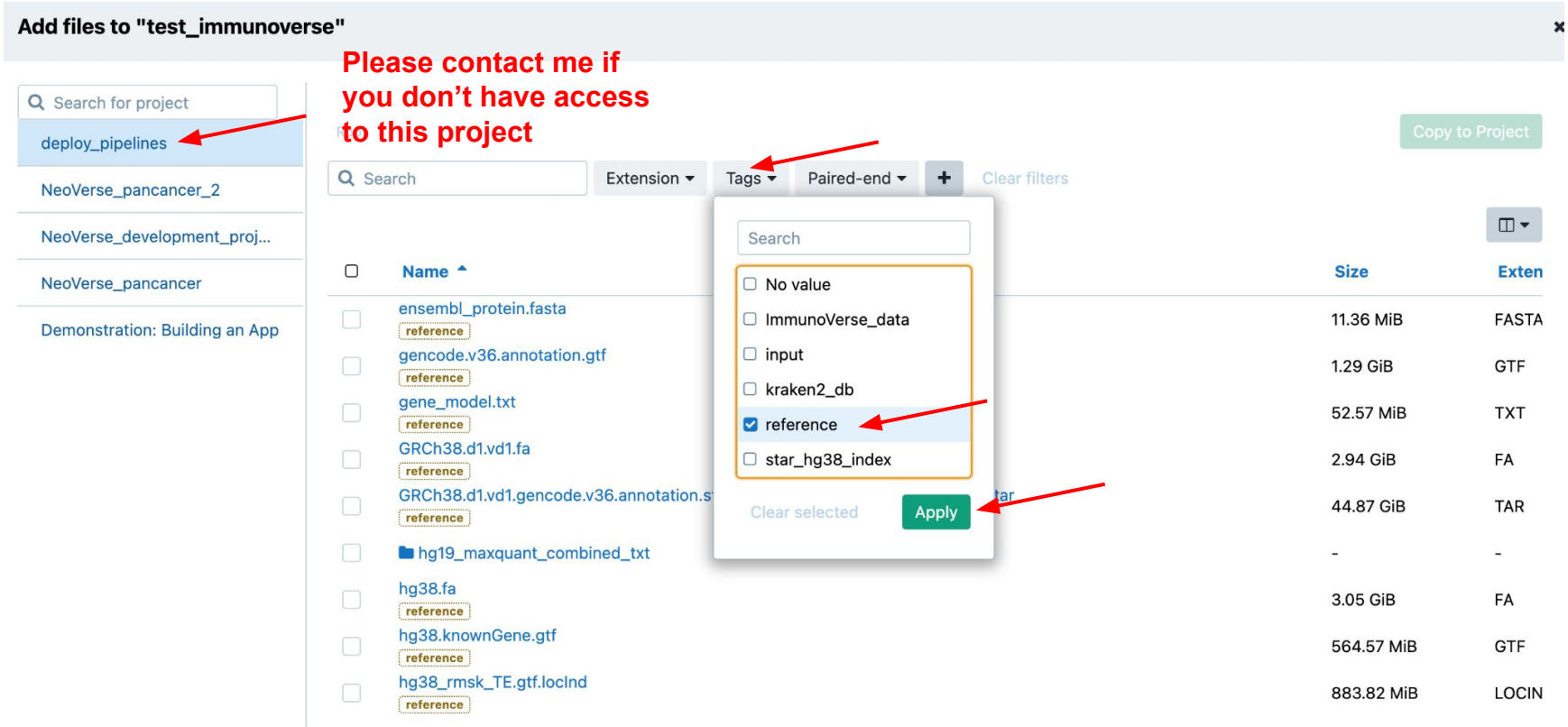
- Transfer all files tagged as **reference** to your project
- Within which we will find files that are further tagged by **ImmunoVerse_data**, **star_hg38_index**, **kraken2_db**, please create three folders named ImmunoVerse_data, star_hg38_index, kraken2_db and moved the corresponding files to these three folders, later, the folder will be passed as an argument in respective workflows
- After step 2, You will notice that two files are present twice, please remove the prefix
  - _1_hg38.fa -> hg38.fa
  - _1_Ensemble_protein.fasta -> Ensemble_protein.fasta
- Copy all the workflows/dockers from my project

Sounds tedious, but you only need to set things up once :)

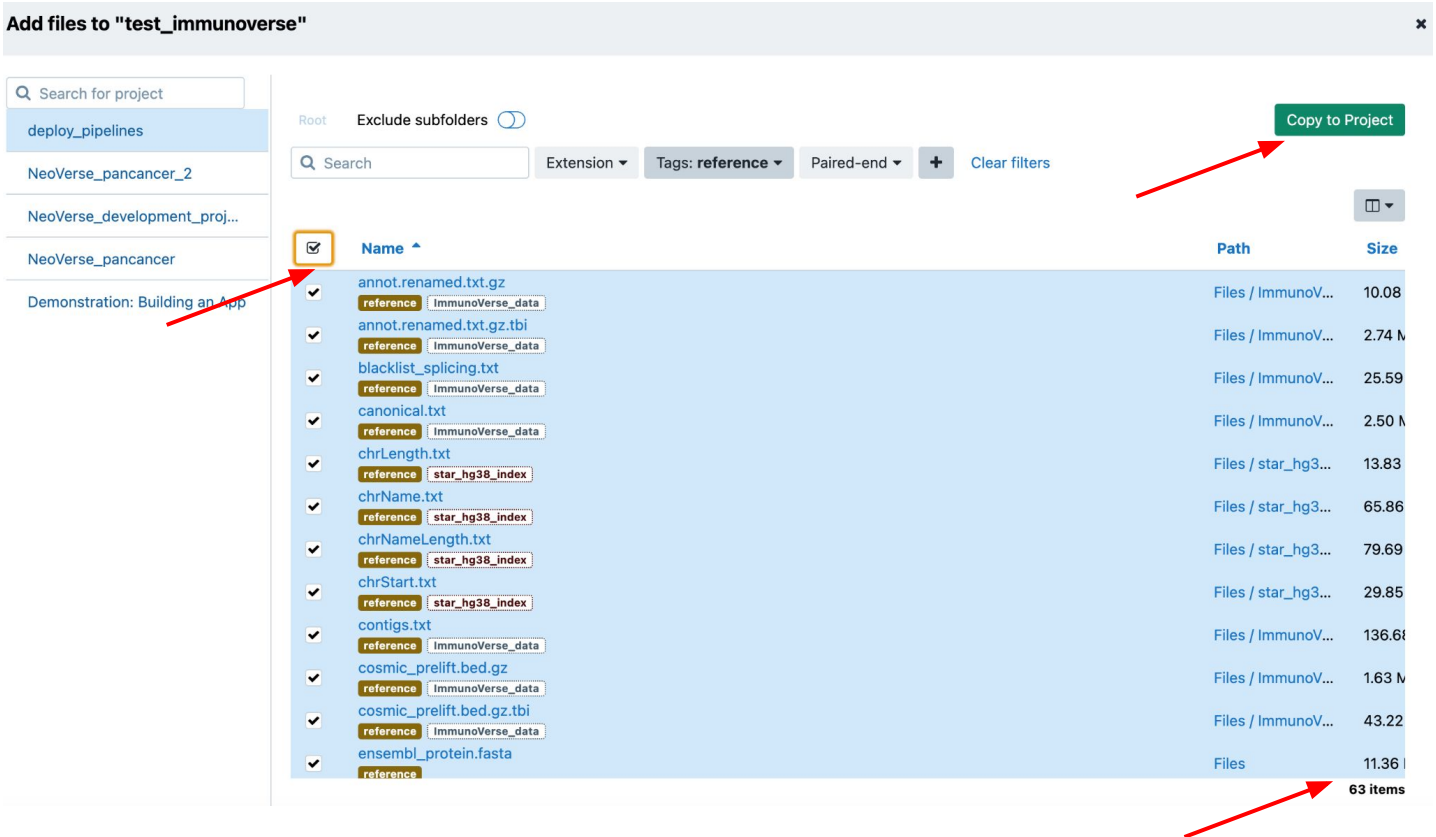
Copy the needed reference files from my project



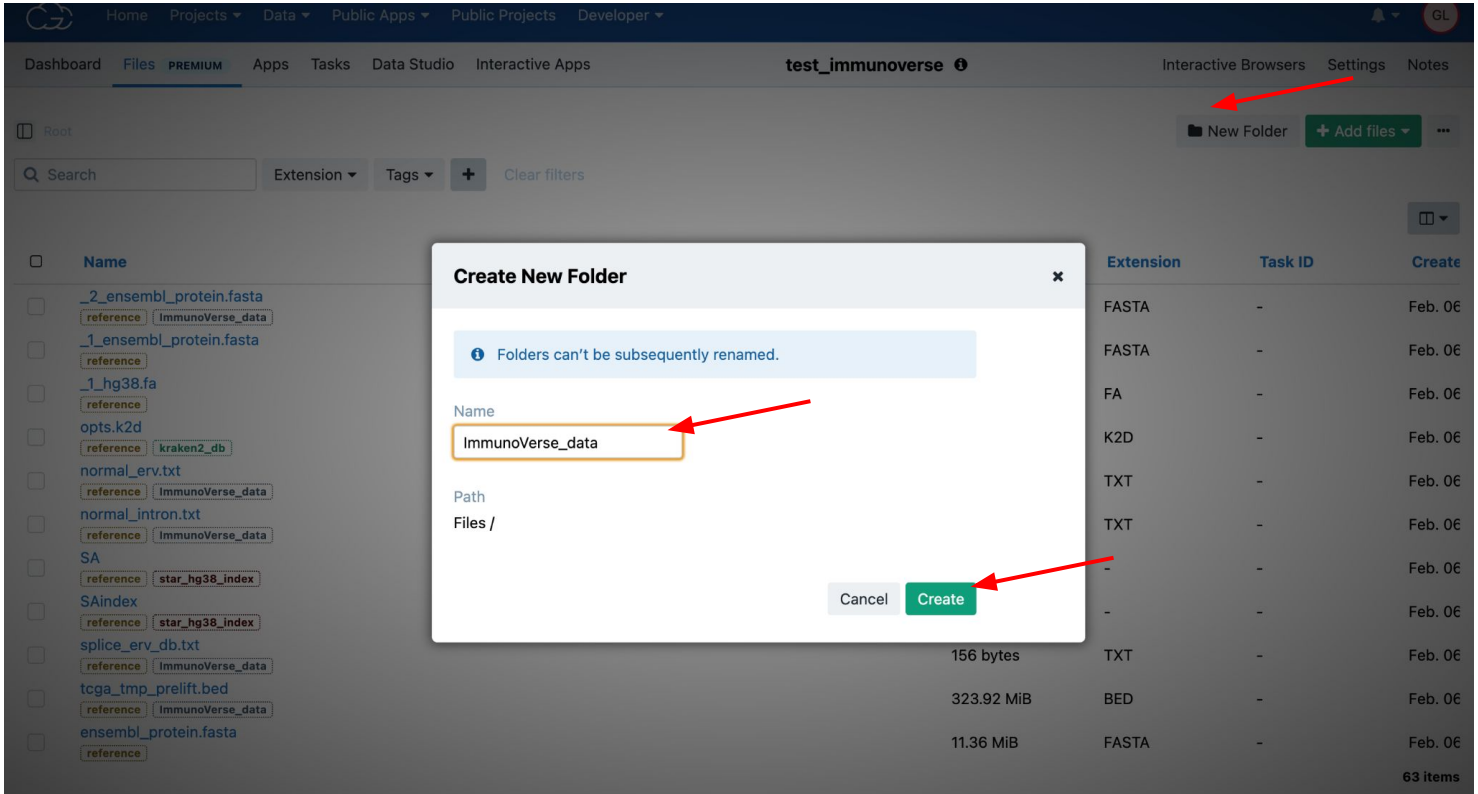
Copy the needed reference files from my project (cont'd)



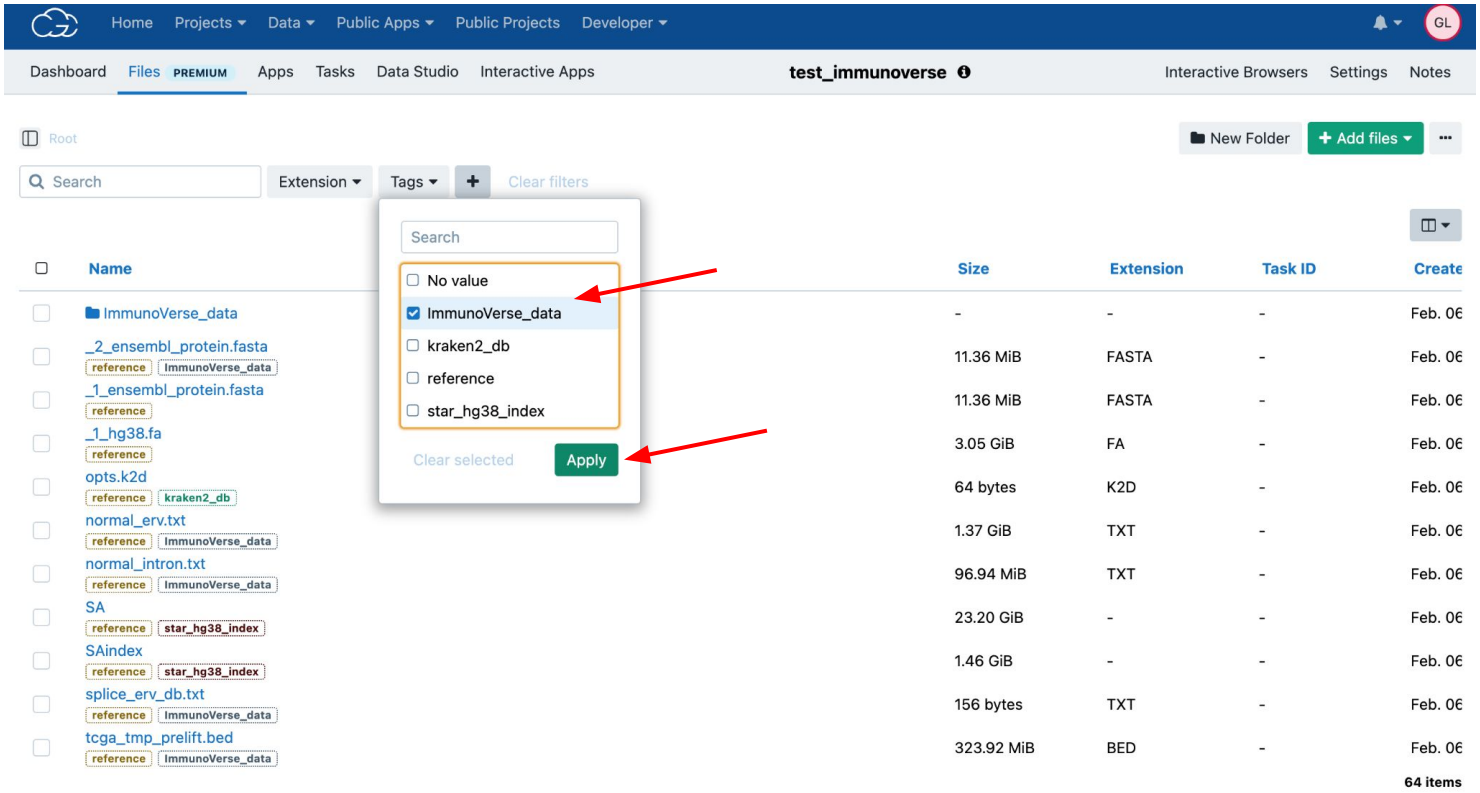
Copy the needed reference files from my project (cont'd)



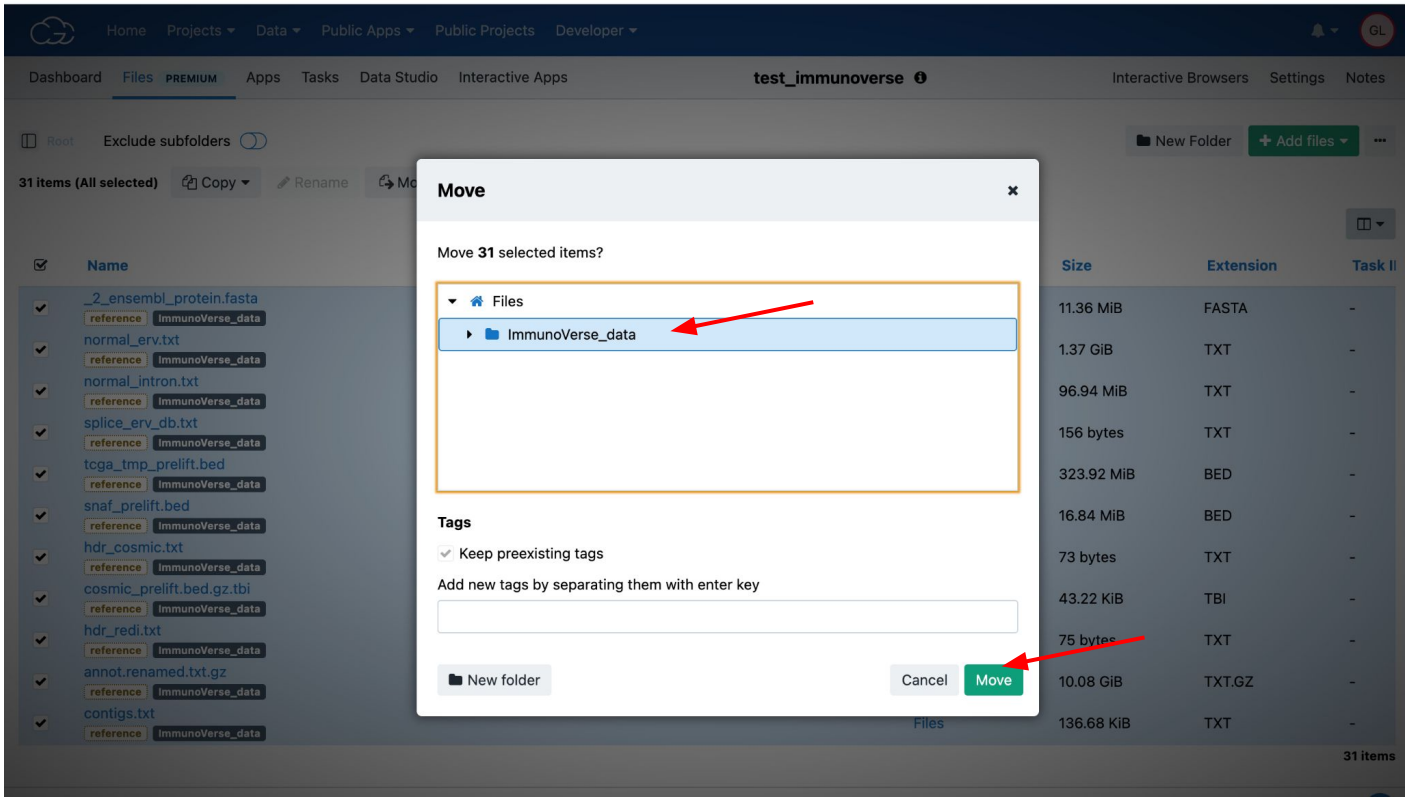
Organize a bit (create a folder named ImmunoVerse_data and move all files tagged as ImmunoVerse_data to this folder)



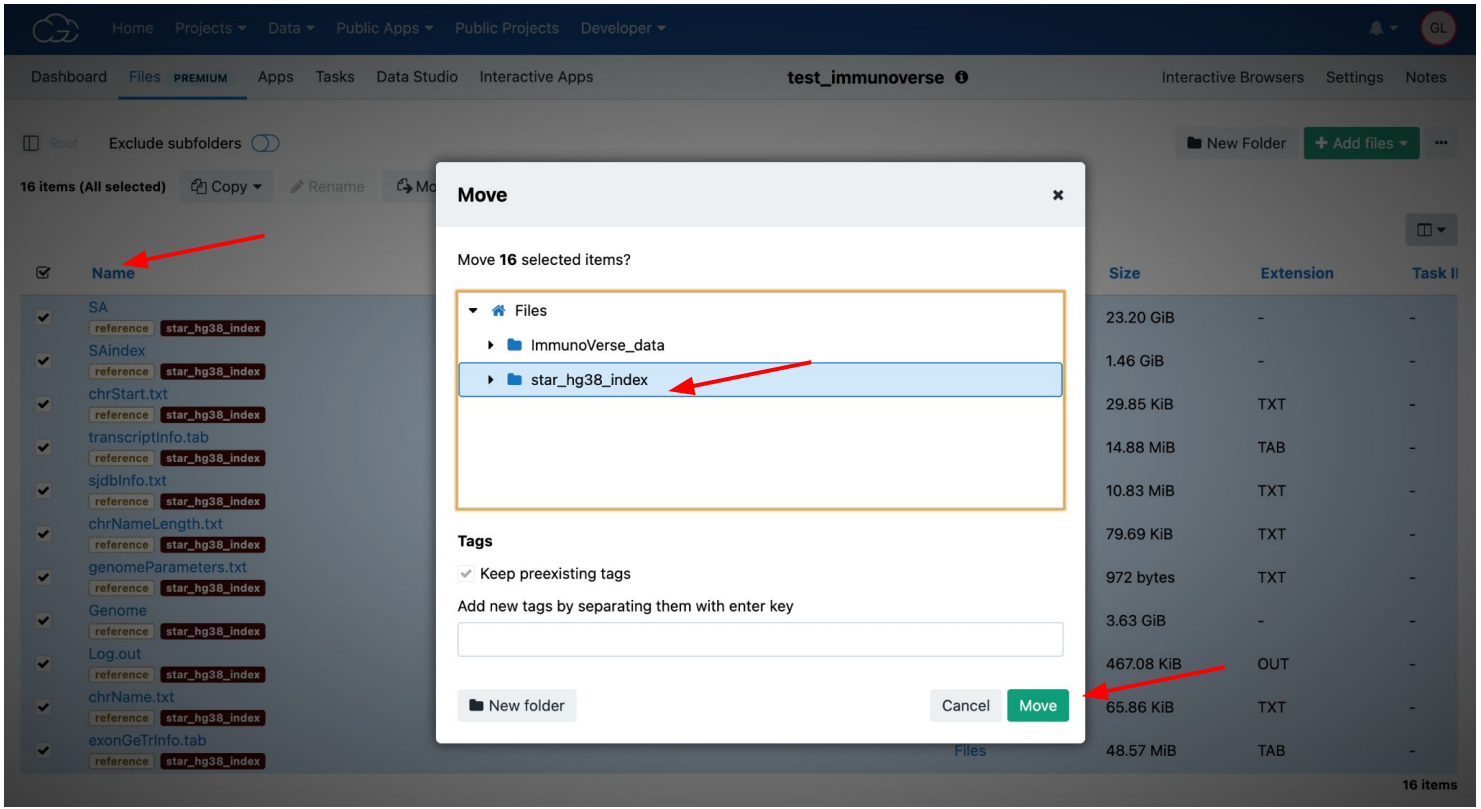
Organize a bit (create a folder named ImmunoVerse_data and move all files tagged as ImmunoVerse_data to this folder)



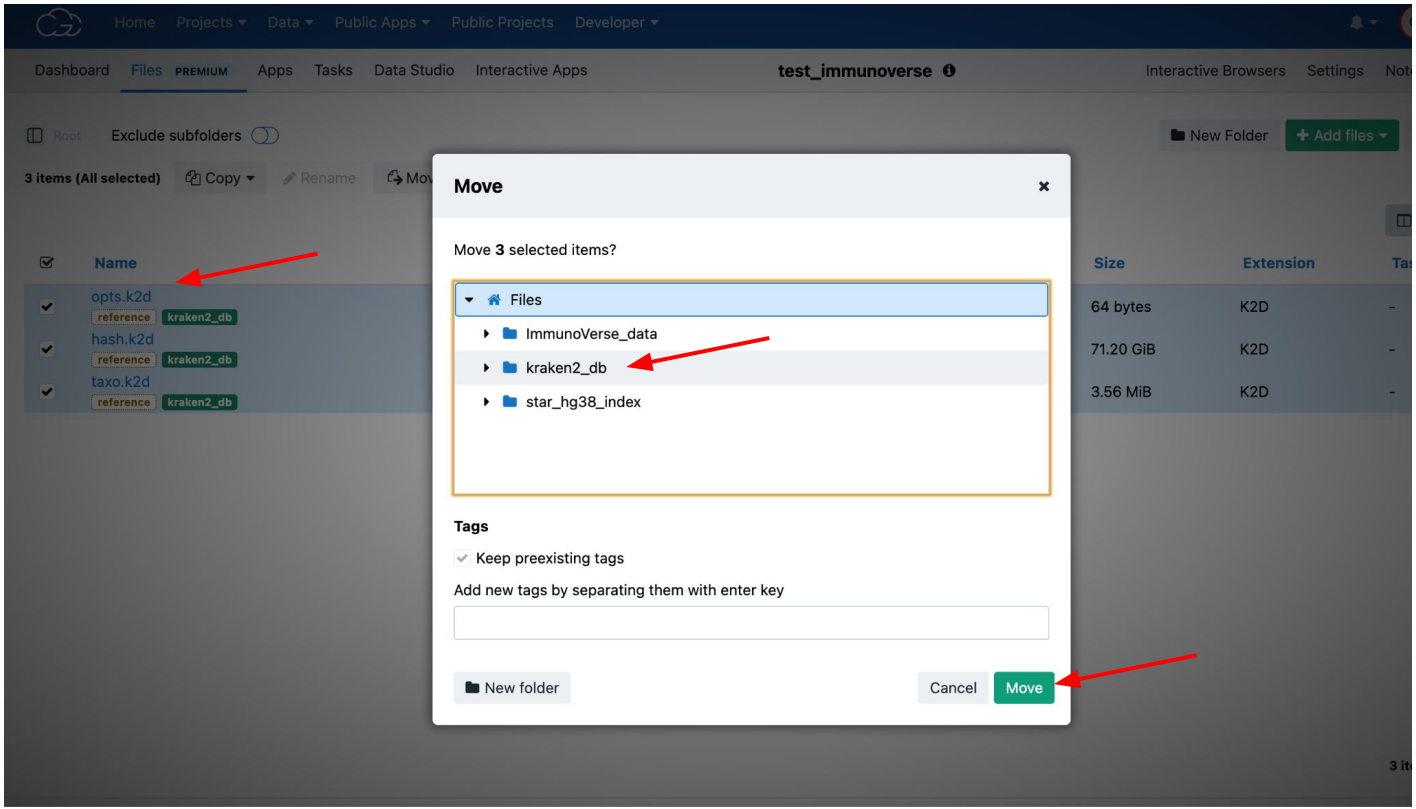
Organize a bit (create a folder named ImmunoVerse_data and move all files tagged as ImmunoVerse_data to this folder)



Organize a bit (create another folder named star_hg_index and move all files tagged as star_hg38_index to this folder)

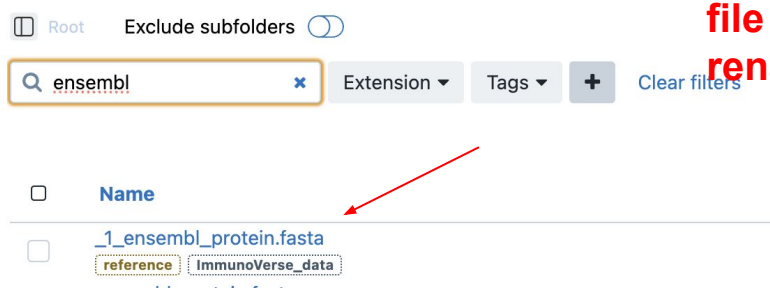


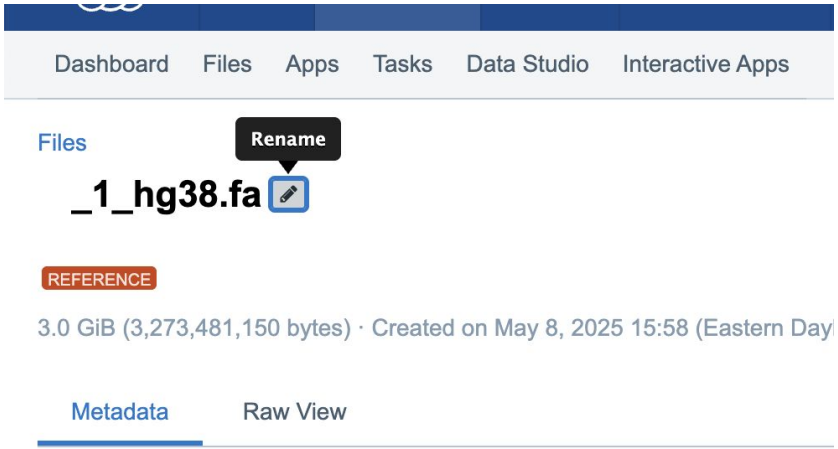
Organize a bit (create a folder named kraken2_db, and move all files tagged as kraken2_db to this folder)

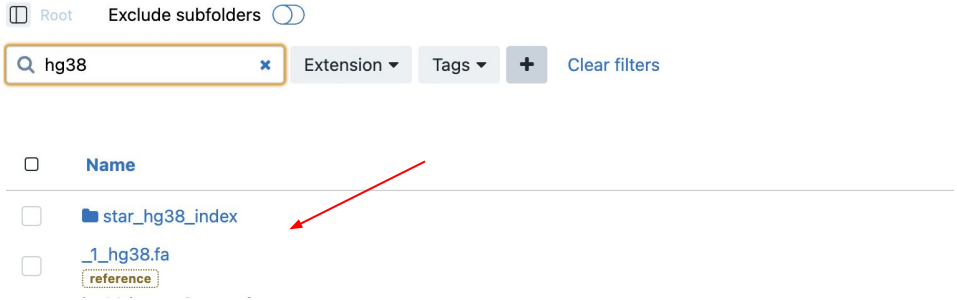


Final clean (if applied to you, may not need)

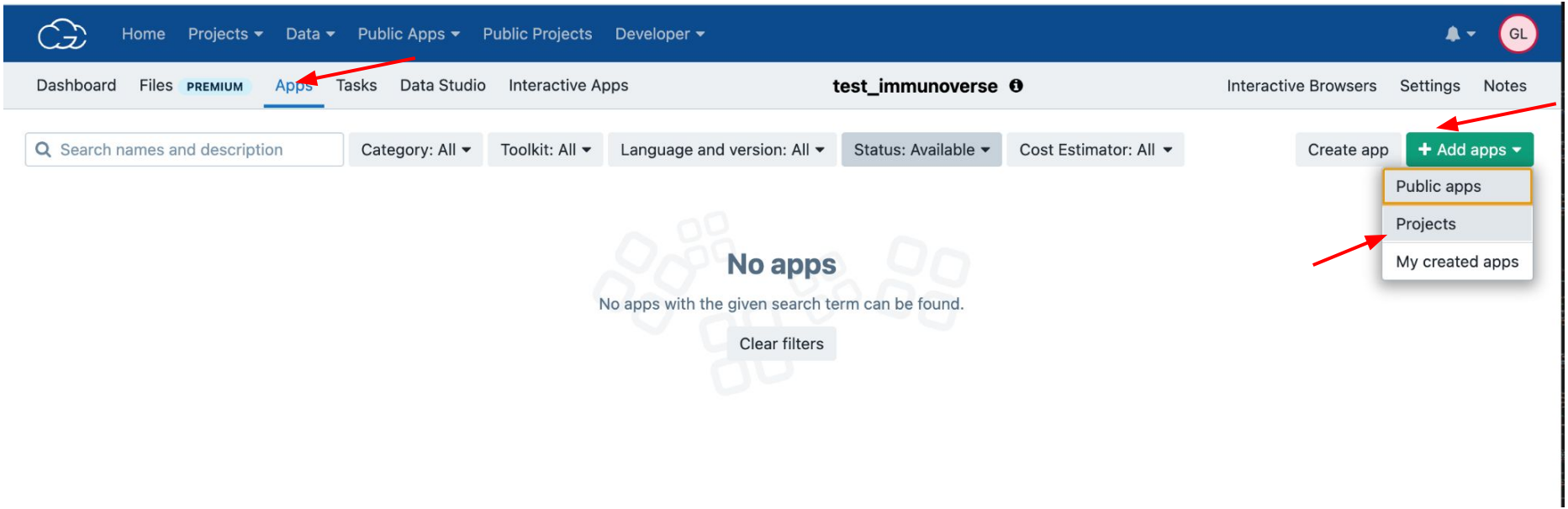
Click the file and rename



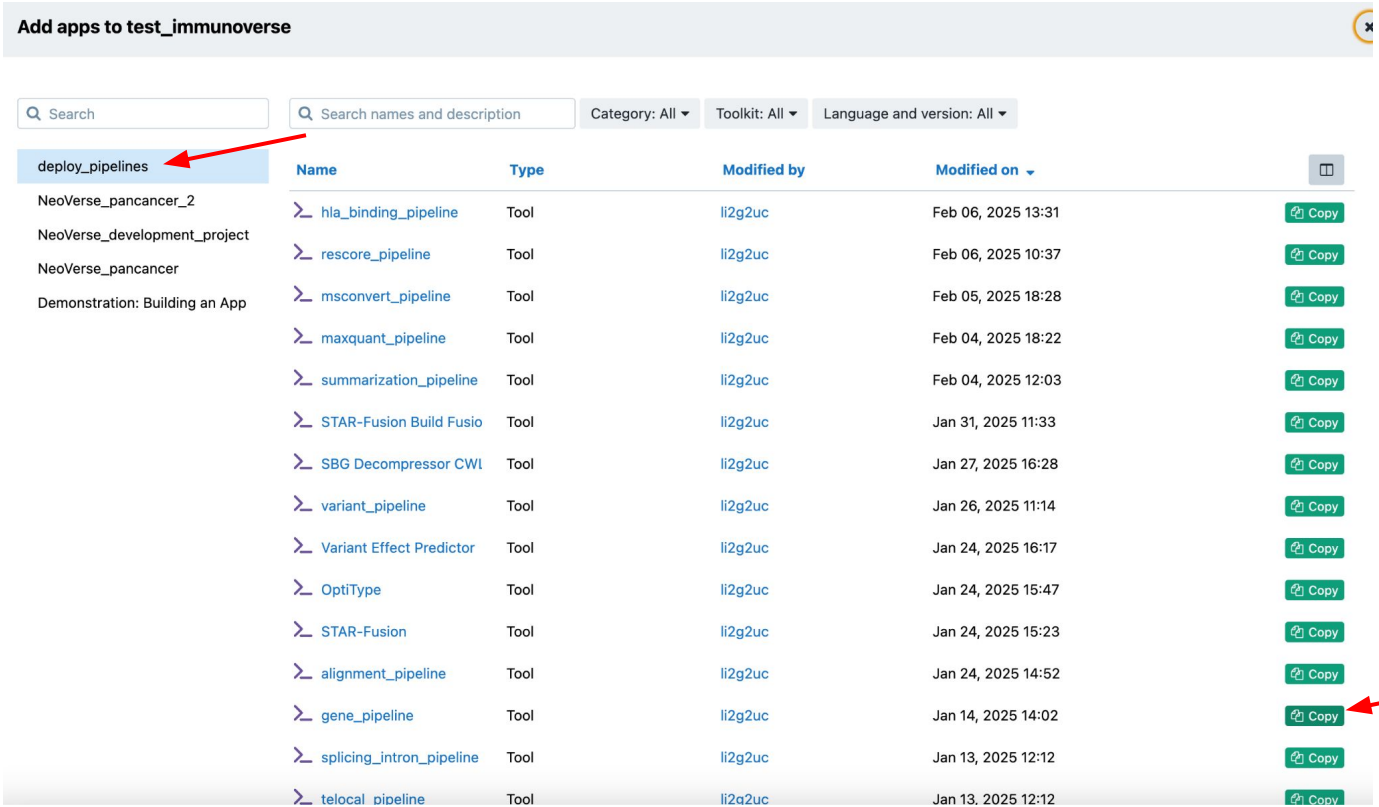




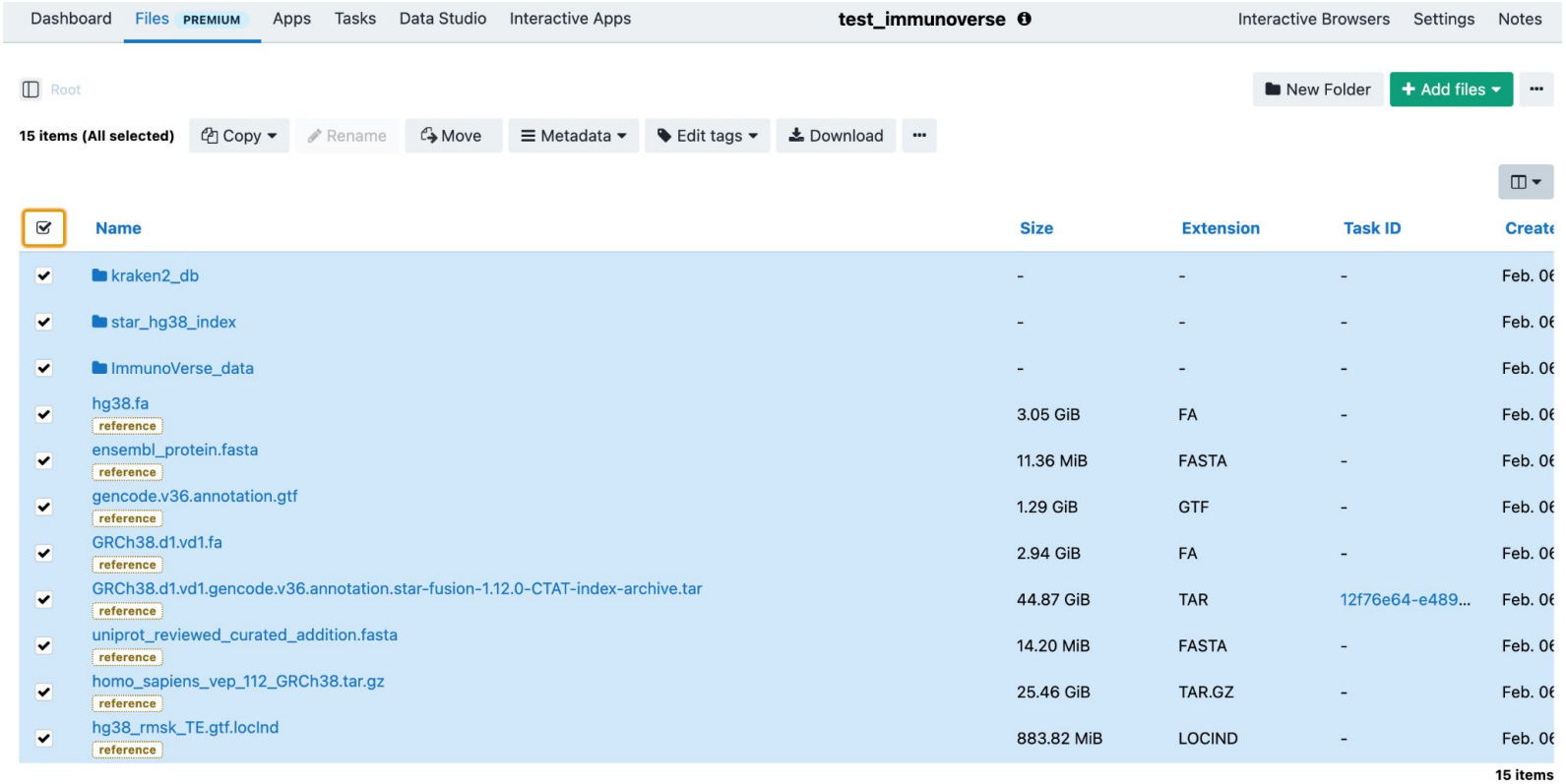
Copy my app/apps



Copy my app/apps



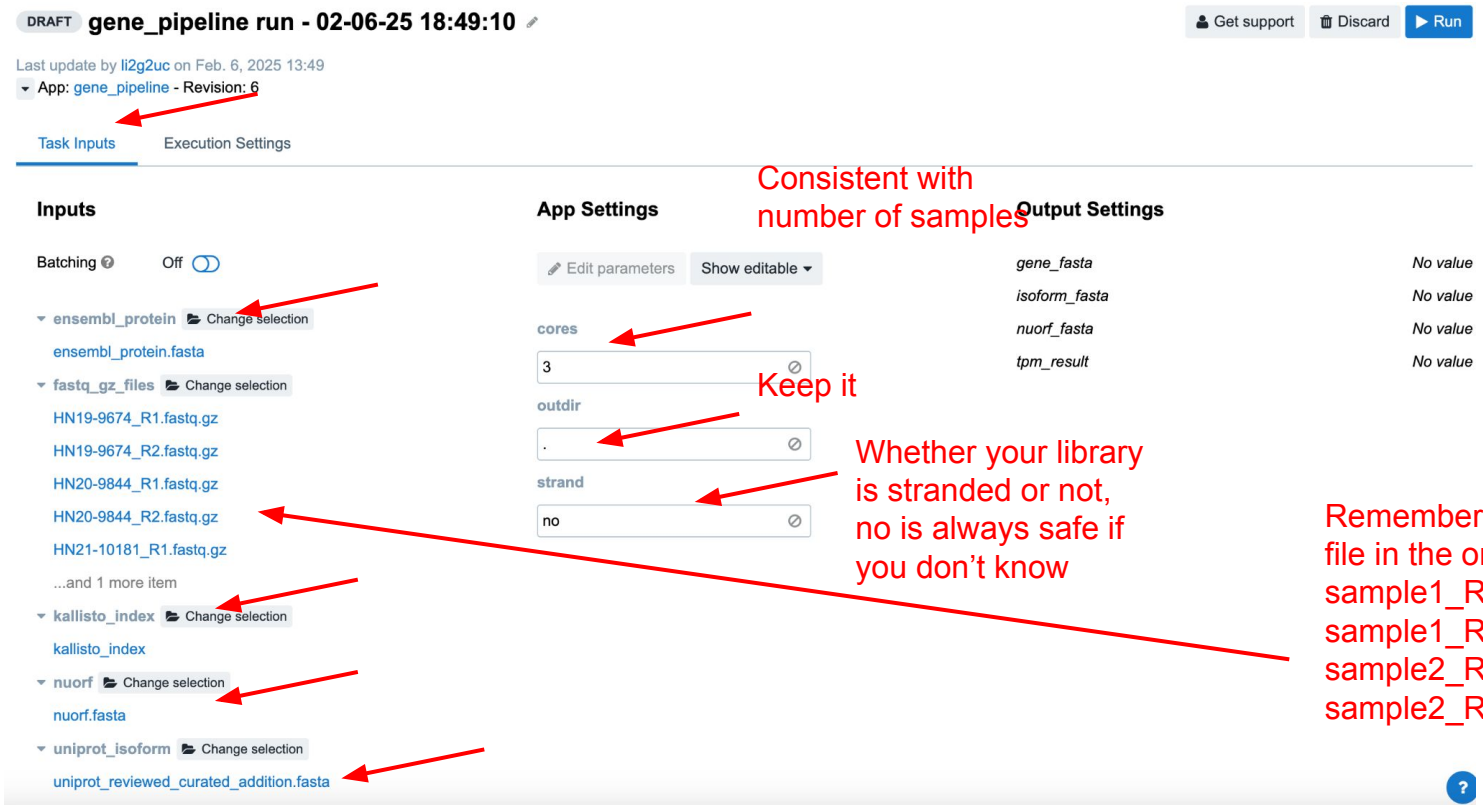
Finished setting things up!!!



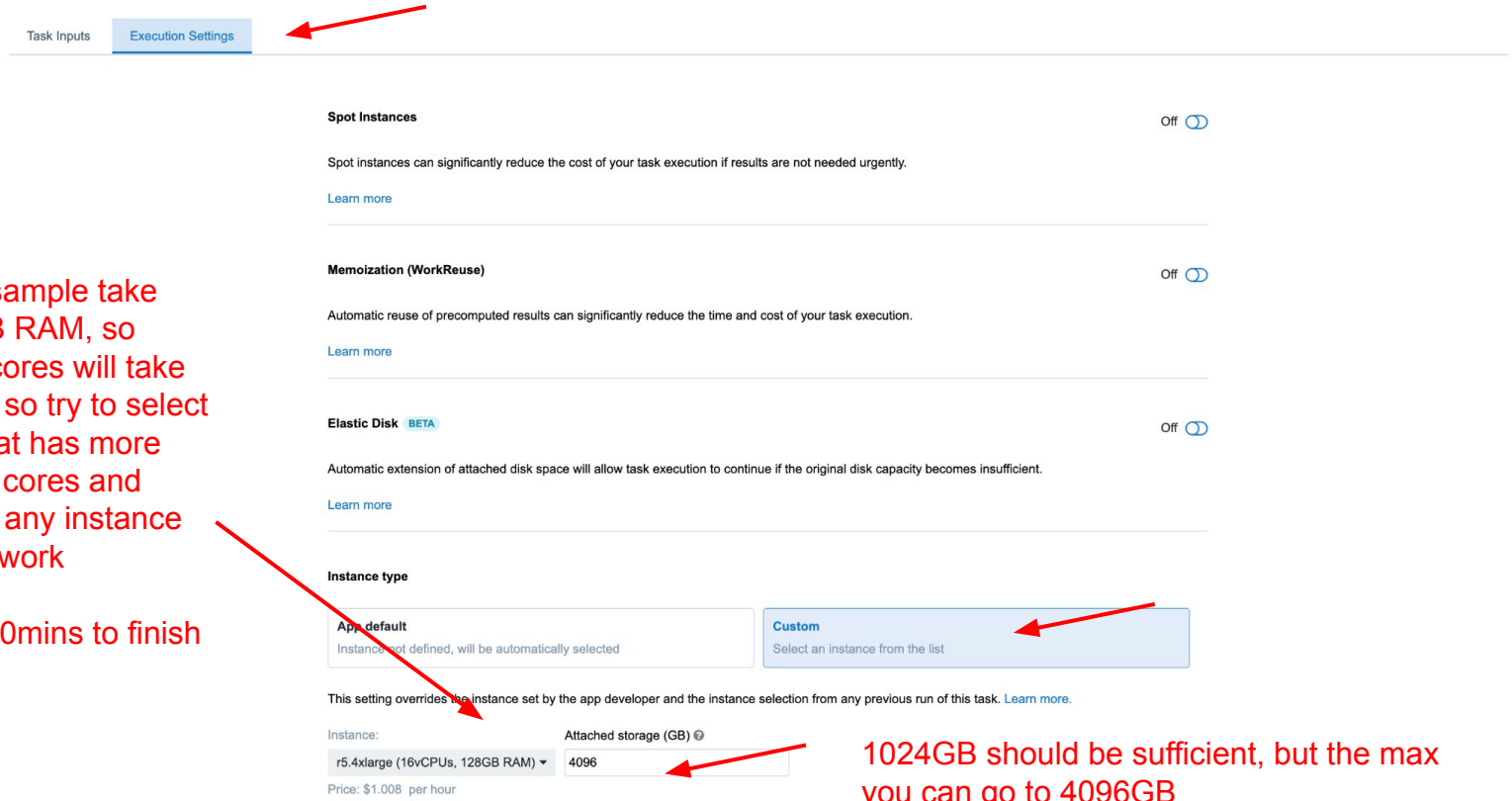
Gene Expression

- For any workflow
  - Select inputs
  - Select reference files
  - Specify parameters
  - Specify running instances
- Each workflow will have slightly different requirement, which I will showcase step by step

Gene expression



Gene expression

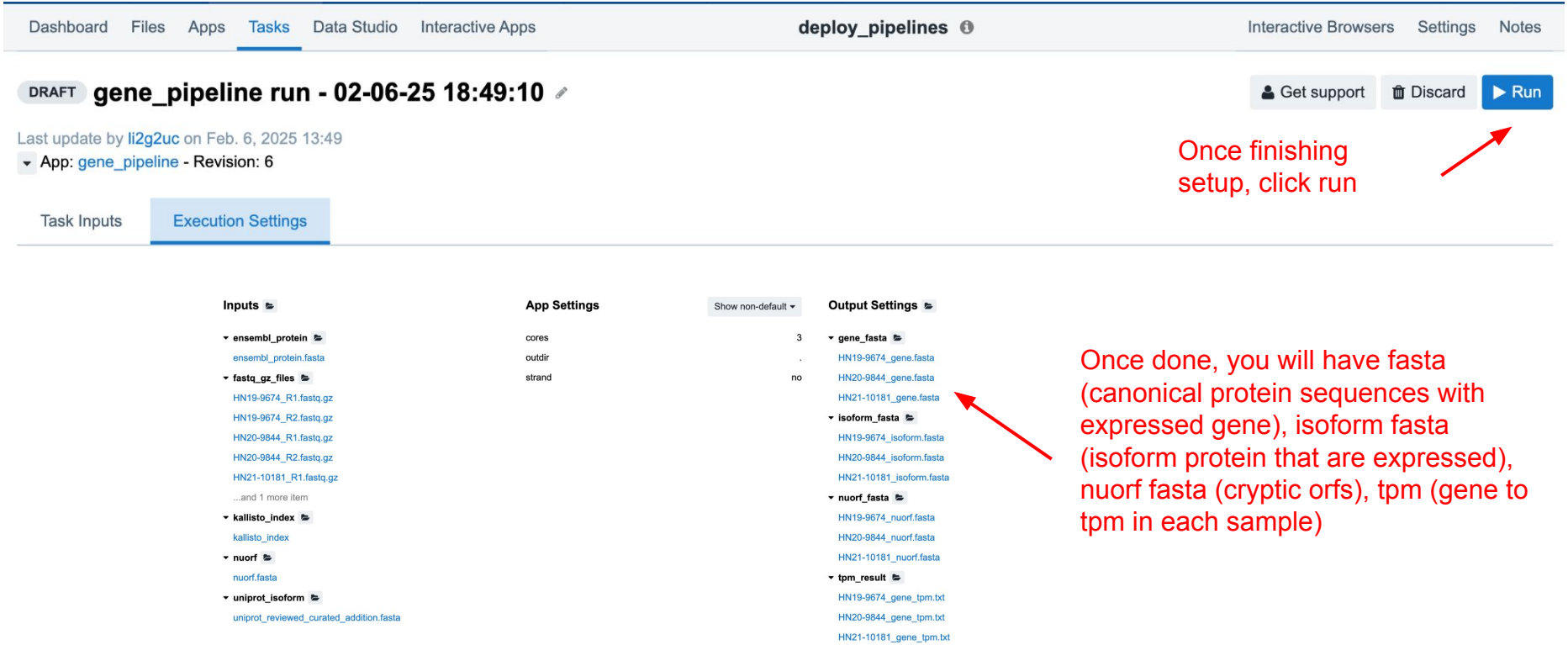


Each sample take ~10GB RAM, so three cores will take 30GB, so try to select one that has more than 3 cores and 30GB, any instance would work

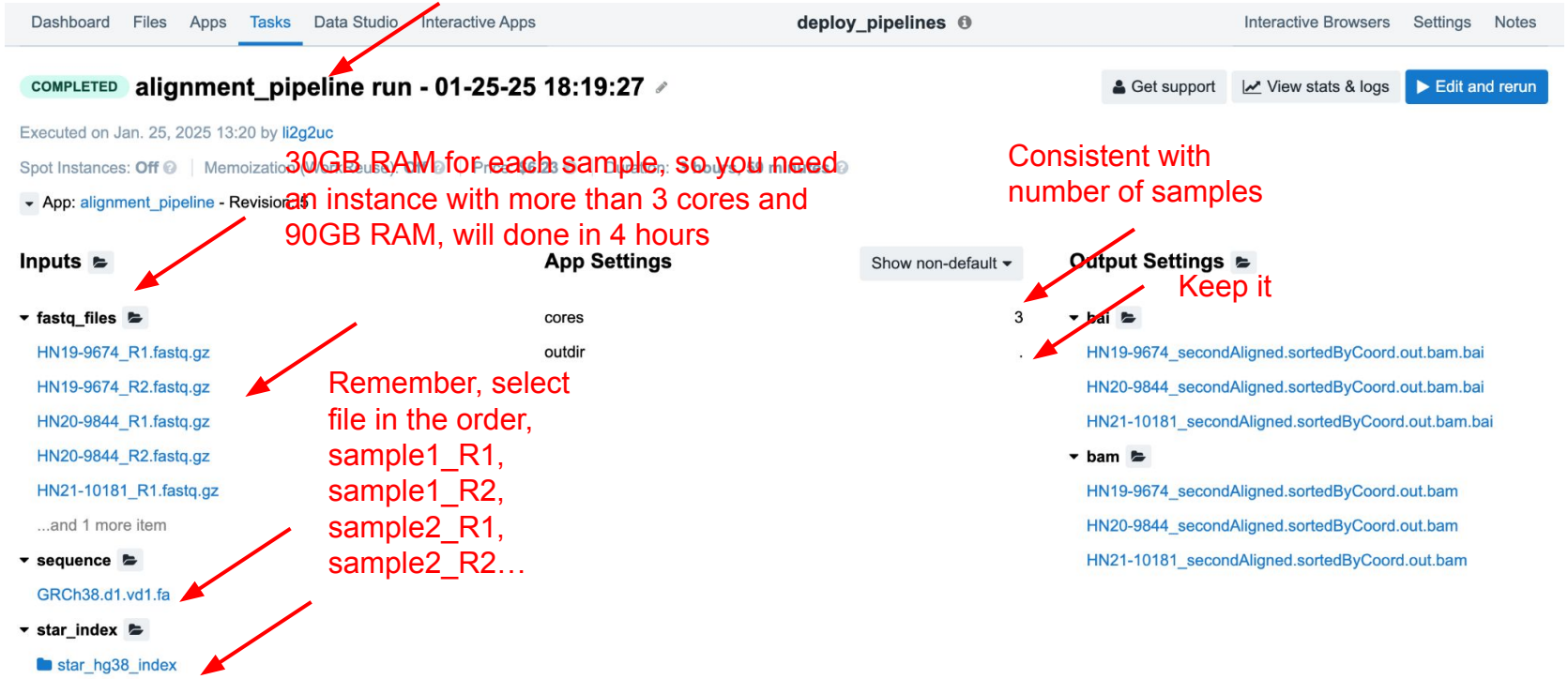
Take 30mins to finish

1024GB should be sufficient, but the max you can go to 4096GB

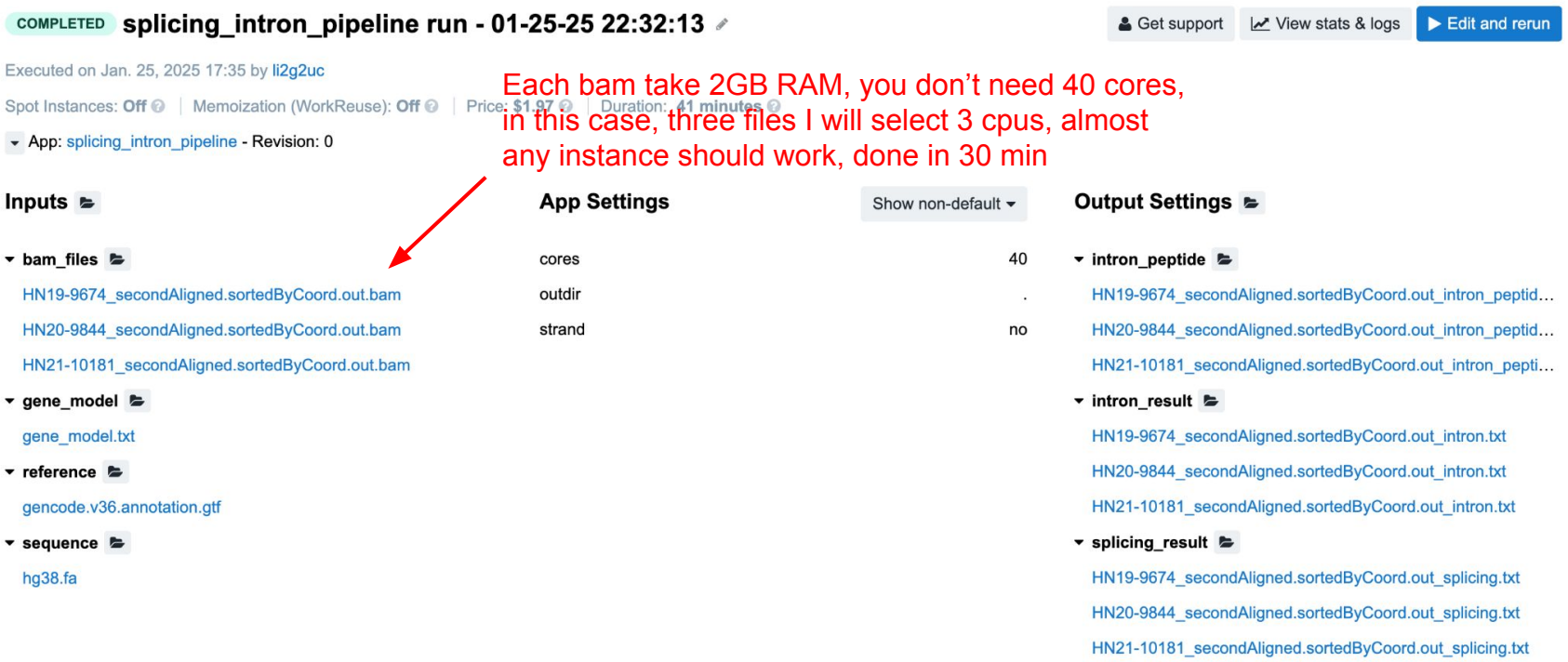
Gene expression



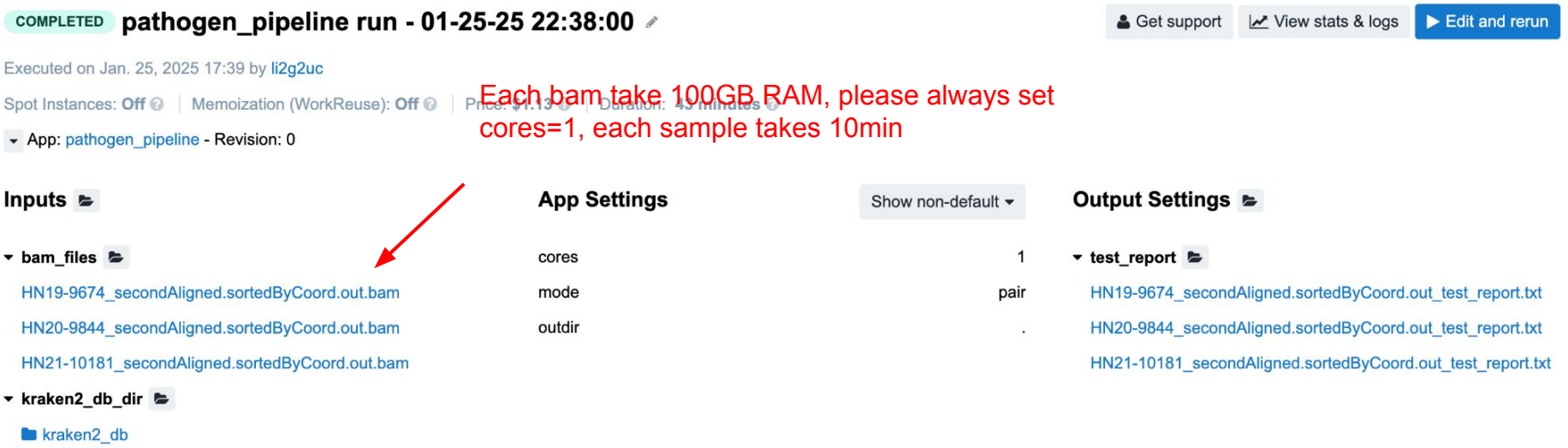
Alignment


Copy this pipeline
Dashboard
Files
Apps
Tasks
Data Studio
Interactive Apps
deploy_pipelines
Interactive Browsers
Settings
Notes
COMPLETED
alignment_pipeline run - 01-25-25 18:19:27
Get support
View stats & logs
Edit and rerun
Executed on Jan. 25, 2025 13:20 by li2g2uc
Spot Instances: Off
App: alignment_pipeline - Revision 3
30GB RAM for each sample, so you need an instance with more than 3 cores and 90GB RAM, will done in 4 hours
Consistent with number of samples
Inputs
App Settings
Show non-default
Output Settings
Keep it
fastq_files
HN19-9674_R1.fastq.gz
HN19-9674_R2.fastq.gz
HN20-9844_R1.fastq.gz
HN20-9844_R2.fastq.gz
HN21-10181_R1.fastq.gz
...and 1 more item
Remember, select file in the order, sample1_R1, sample1_R2, sample2_R1, sample2_R2…
sequence
GRCh38.d1.vd1.fa
star_index
star_hg38_index
cores
3
outdir
.
bai
HN19-9674_secondAligned.sortedByCoord.out.bam.bai
HN20-9844_secondAligned.sortedByCoord.out.bam.bai
HN21-10181_secondAligned.sortedByCoord.out.bam.bai
bam
HN19-9674_secondAligned.sortedByCoord.out.bam
HN20-9844_secondAligned.sortedByCoord.out.bam
HN21-10181_secondAligned.sortedByCoord.out.bam

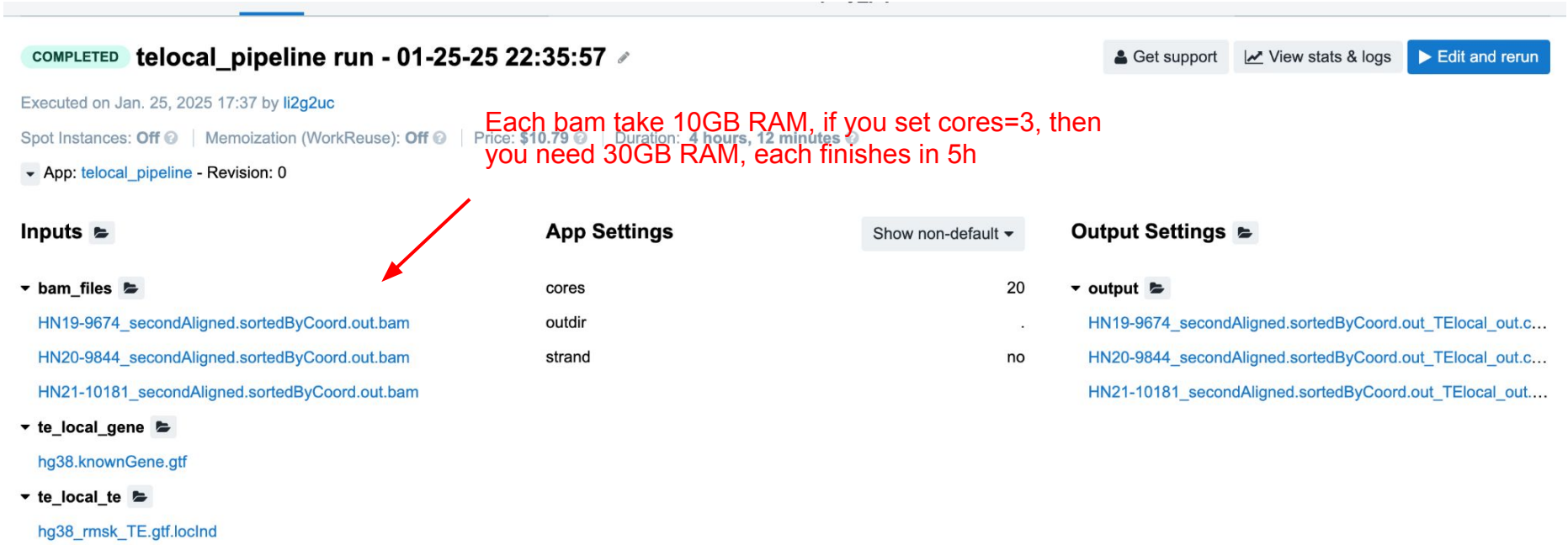
Splicing and intron retention



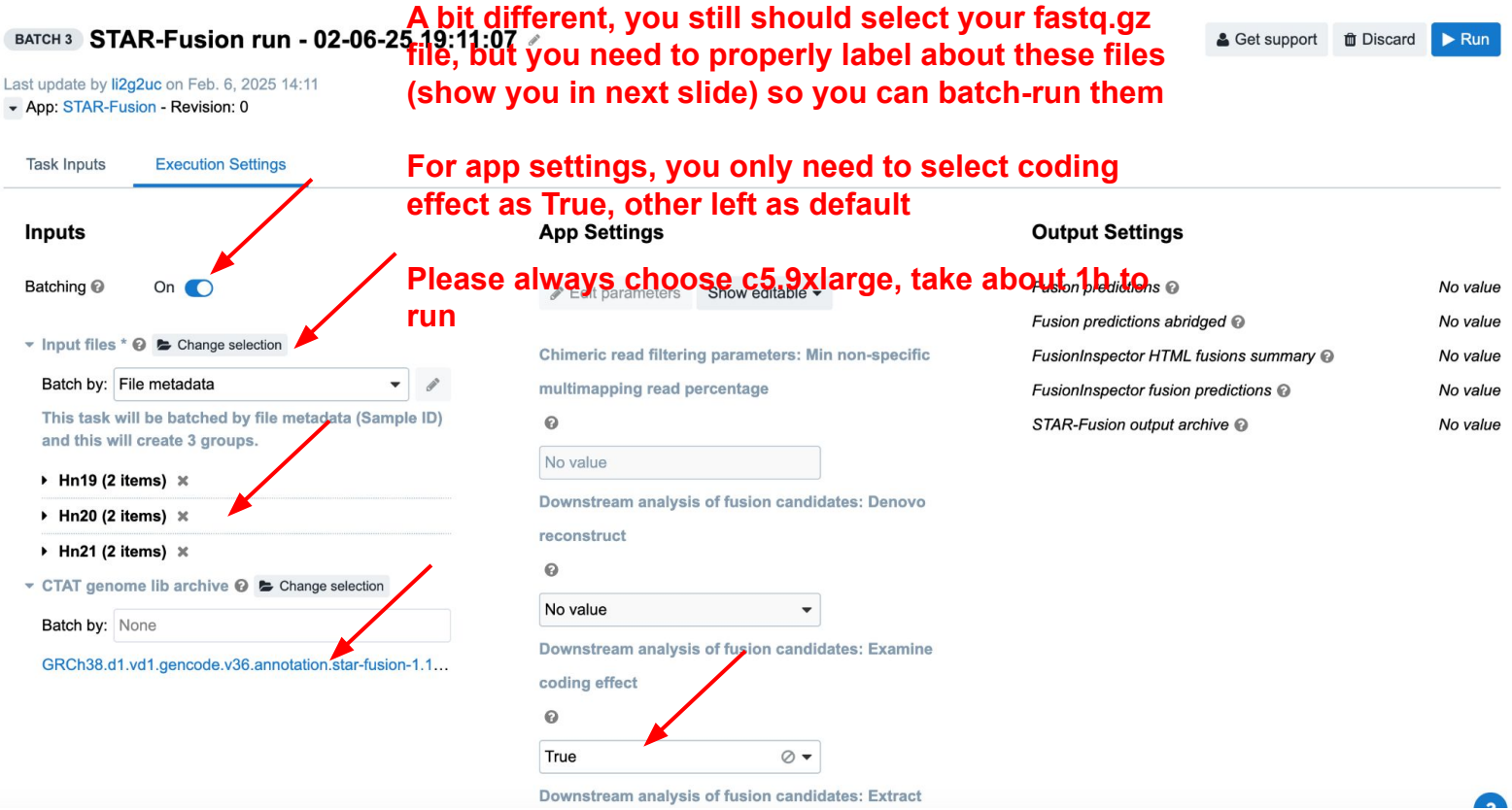
Pathogen


COMPLETED
pathogen_pipeline run - 01-25-25 22:38:00
Get support
View stats & logs
Edit and rerun
Executed on Jan. 25, 2025 17:39 by li2g2uc
Spot Instances: Off
Memoization (WorkReuse): Off
App: pathogen_pipeline - Revision: 0
Each bam take 100GB RAM, please always set cores=1, each sample takes 10min
Inputs
bam_files
HN19-9674_secondAligned.sortedByCoord.out.bam
HN20-9844_secondAligned.sortedByCoord.out.bam
HN21-10181_secondAligned.sortedByCoord.out.bam
kraken2_db_dir
kraken2_db
App Settings
Show non-default
cores 1
mode pair
outdir .
Output Settings
test_report
HN19-9674_secondAligned.sortedByCoord.out_test_report.txt
HN20-9844_secondAligned.sortedByCoord.out_test_report.txt
HN21-10181_secondAligned.sortedByCoord.out_test_report.txt

Transposable element



Gene fusion

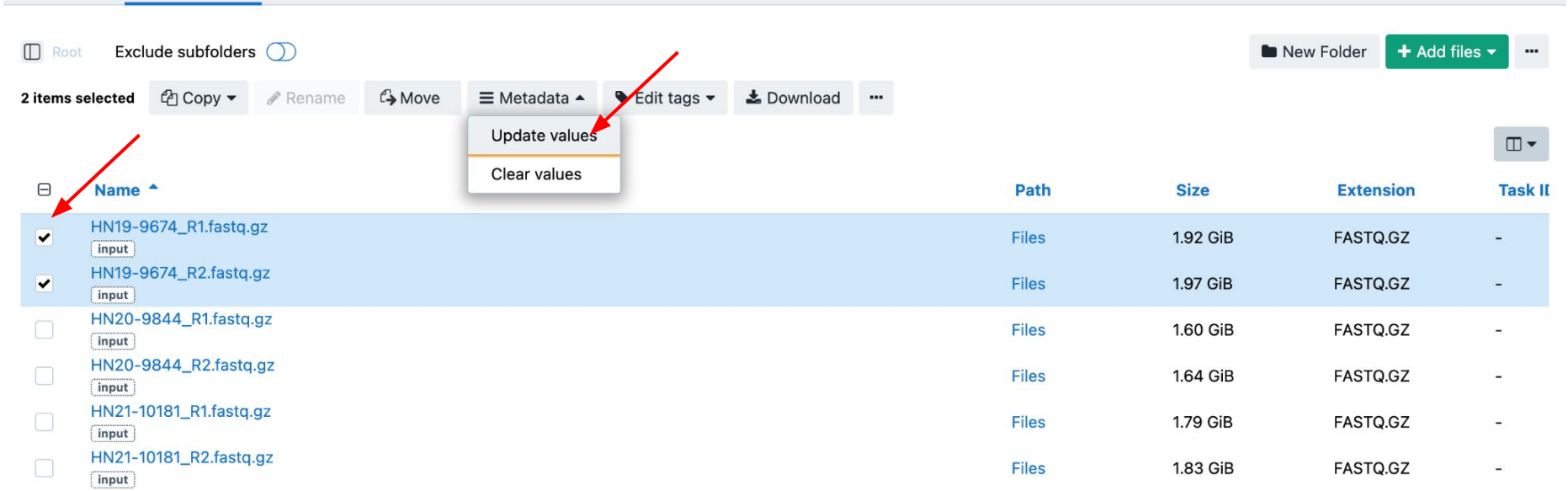


A bit different, you still should select your fastq.gz file, but you need to properly label about these files (show you in next slide) so you can batch-run them

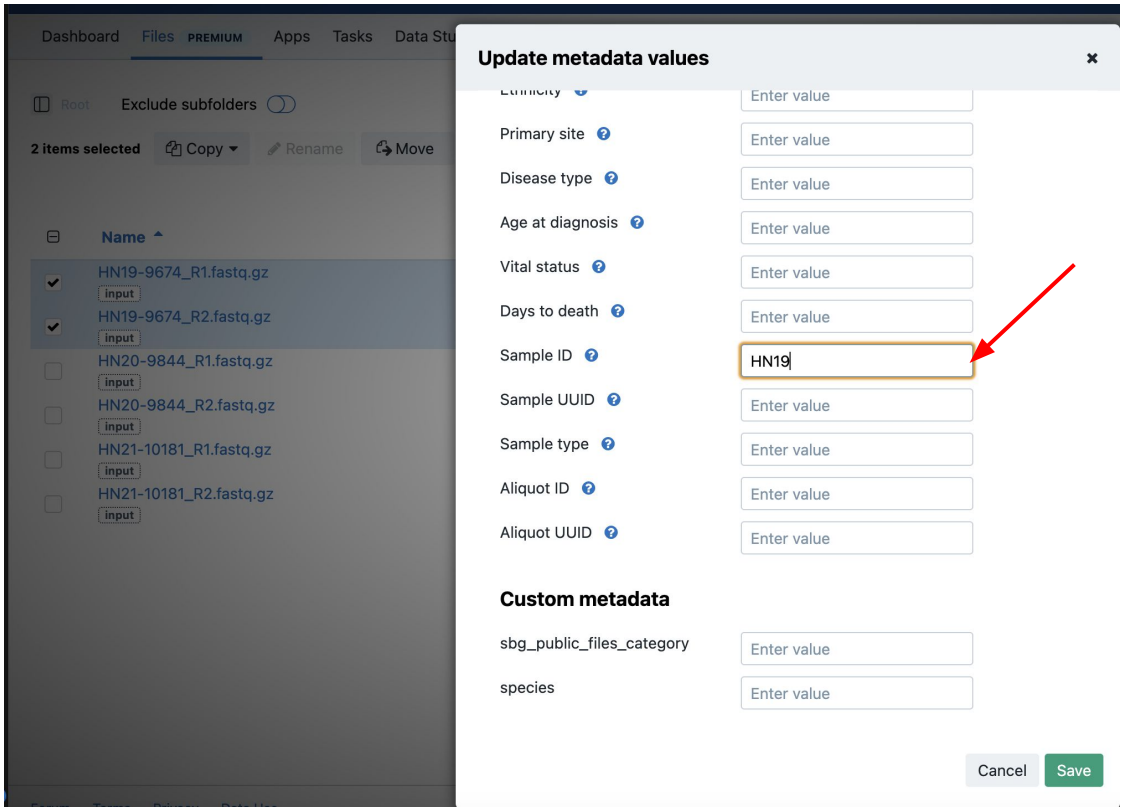
For app settings, you only need to select coding effect as True, other left as default

Please always choose c5.9xlarge, take about 1h to run

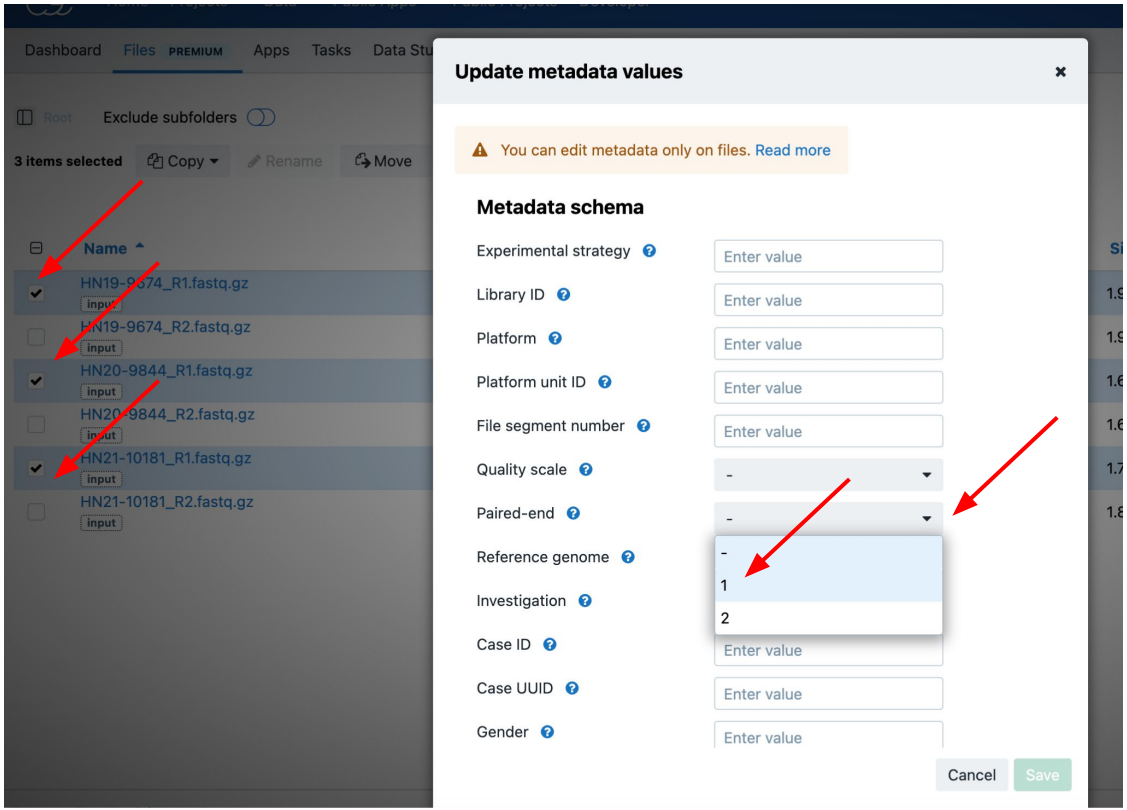
Gene Fusion (label them by sample ID and pair-end)



Gene Fusion (label them by sample ID and pair-end)

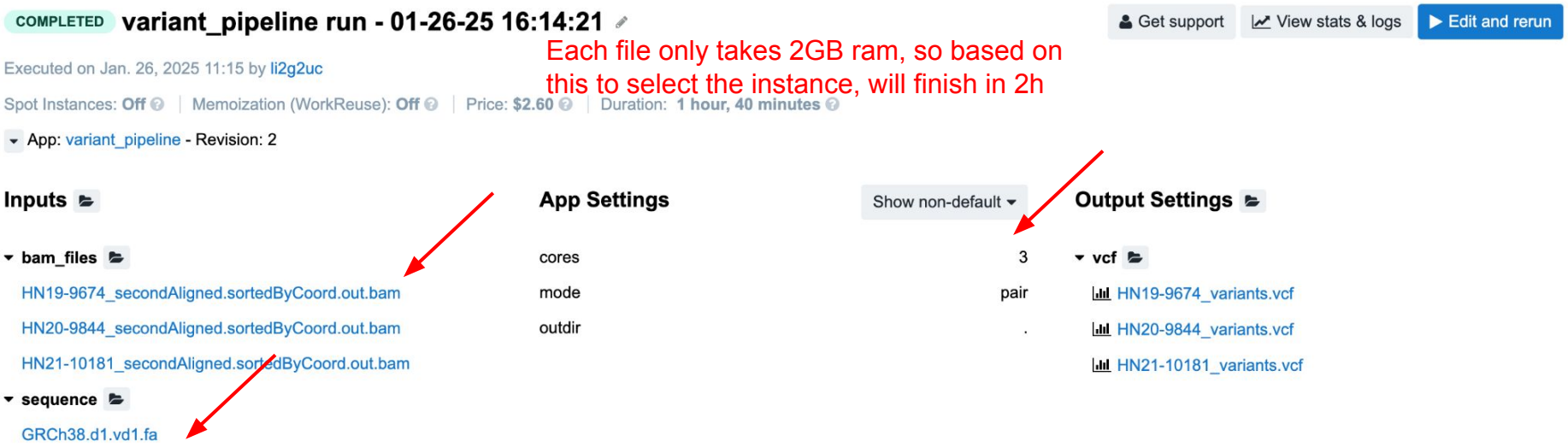


Gene Fusion (label them by sample ID and pair-end)

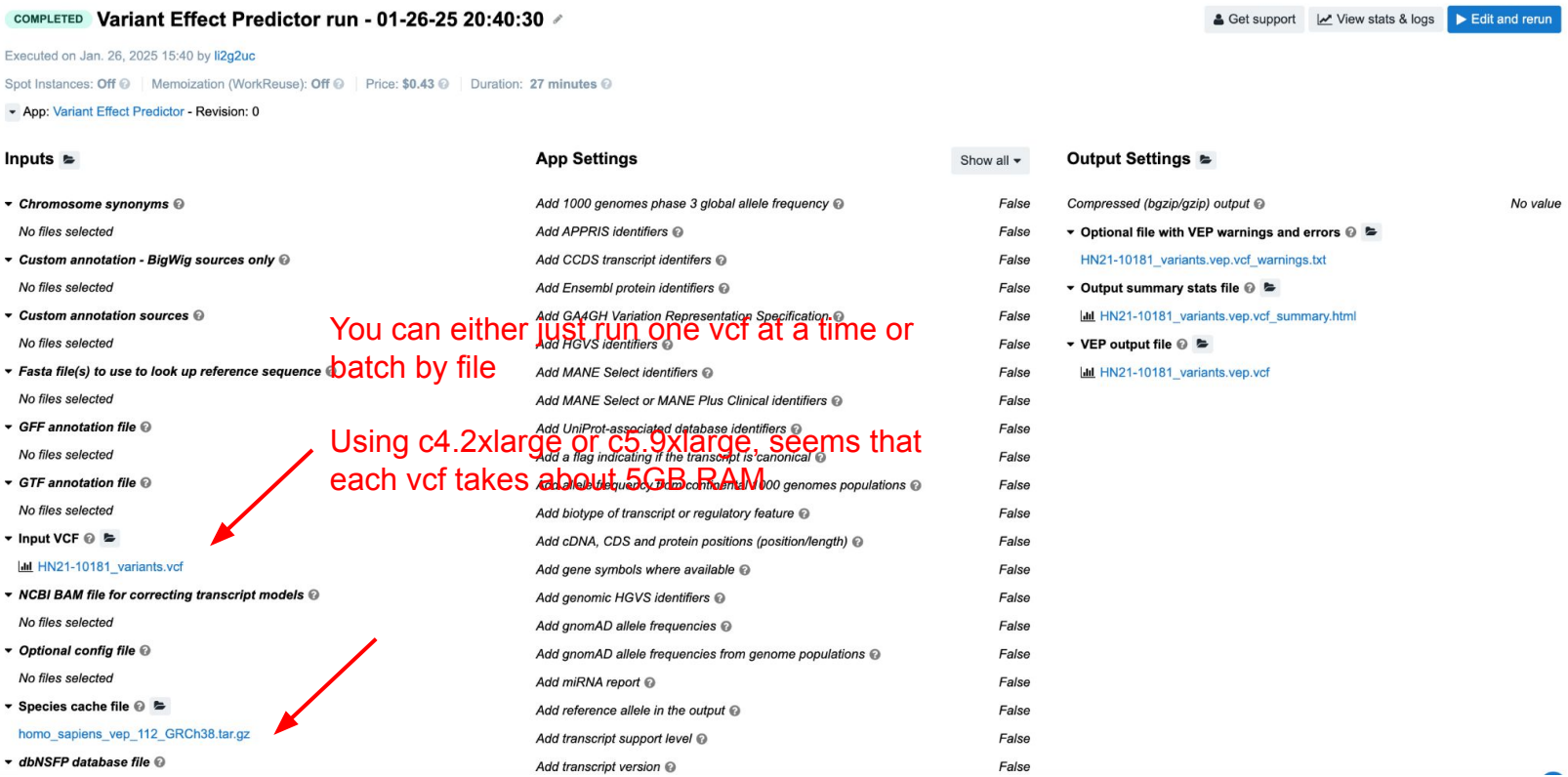


By doing that, when you supply all fastq.gz files, the program will figure out how to pair, and how to parallelize

Variants (step 1 is to get vcf files)



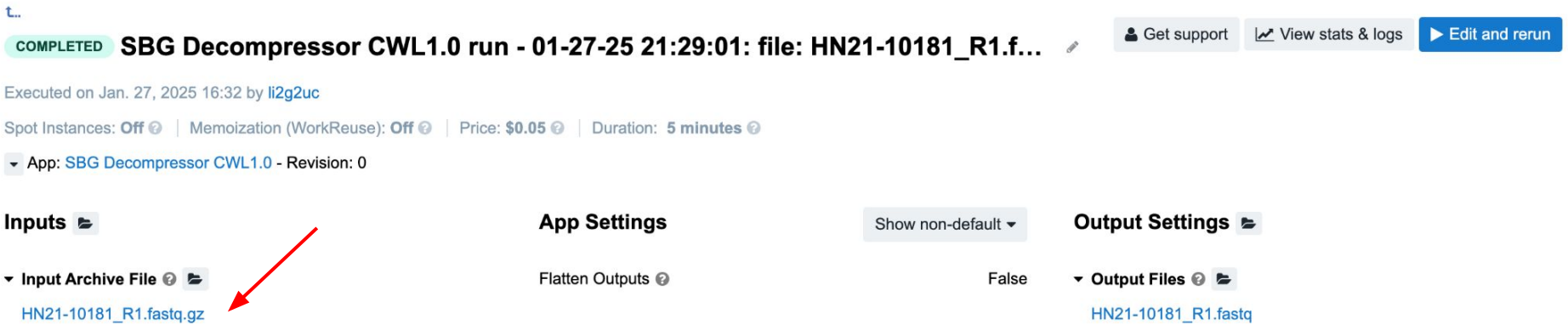
Variants (step 2 is to run variant effect predictor)



You can either just run one vcf at a time or batch by file

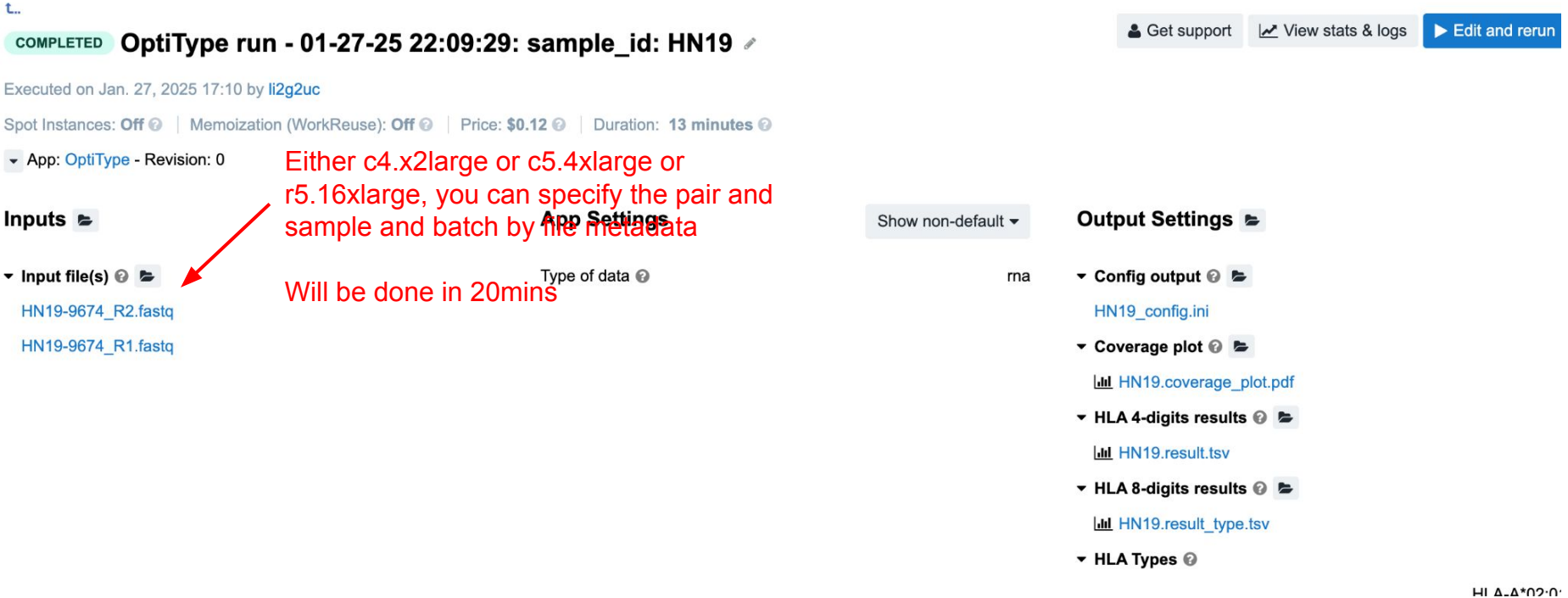
Using c4.2xlarge or c5.9xlarge, seems that each vcf takes about 5GB RAM

HLA typing (step 1 is decompress fastq.gz to fastq)

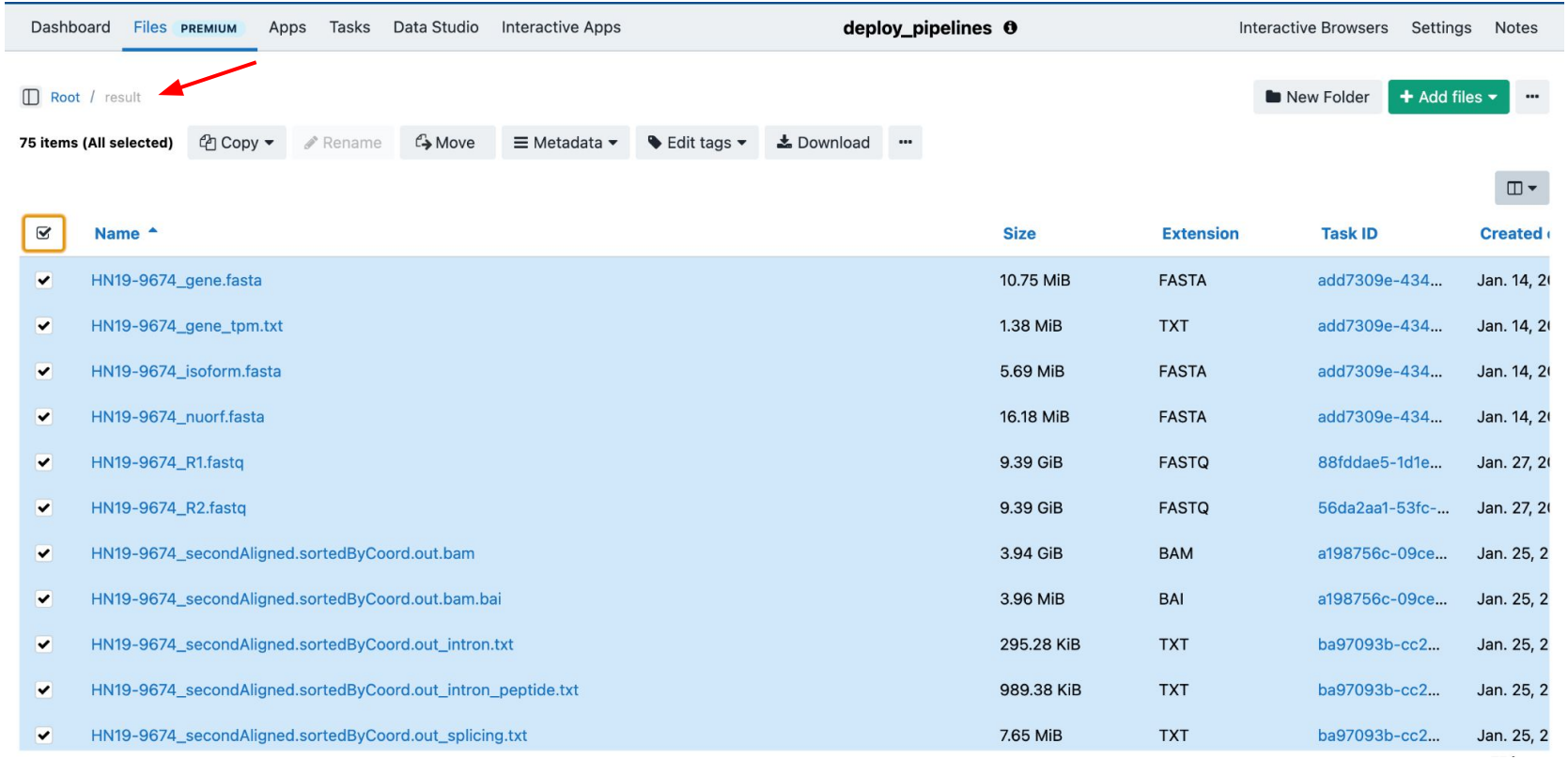


You should batch by file so multiple run can be parallelized, c4.2xlarge should work, will be done in 10 mins

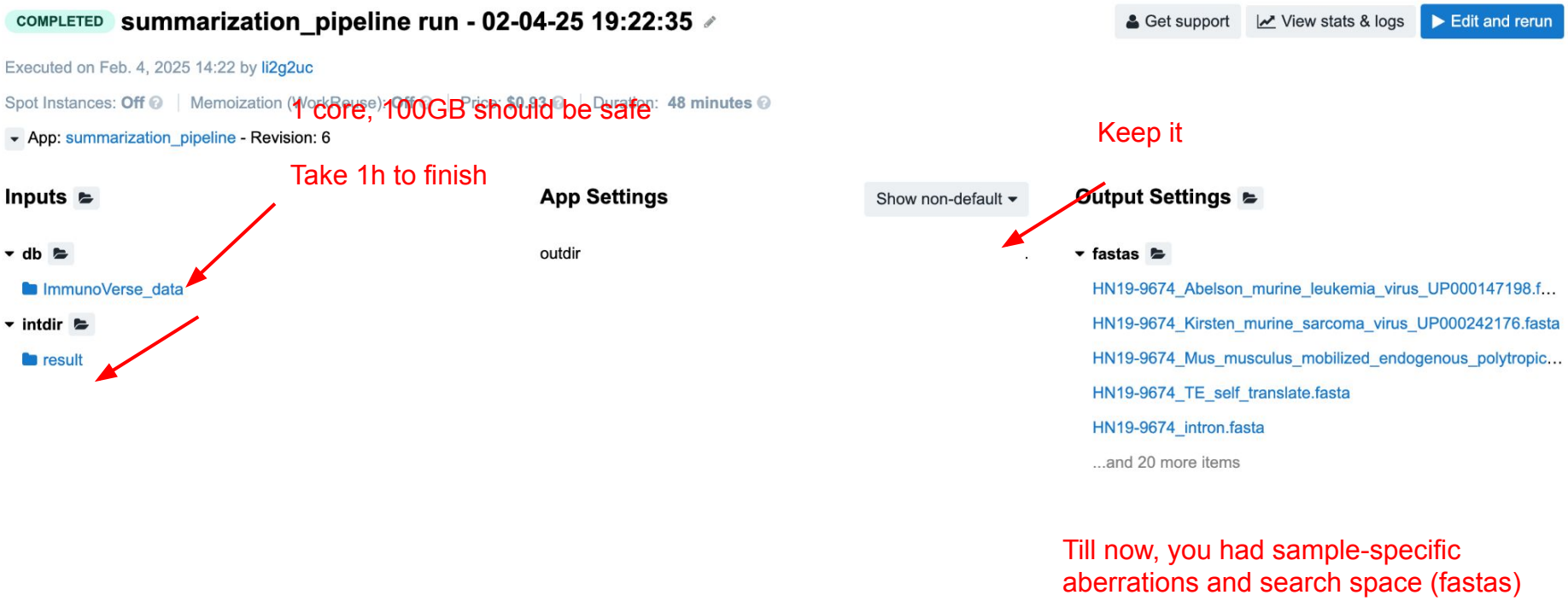
HLA typing (step 1 is decompress fastq.gz to fastq)



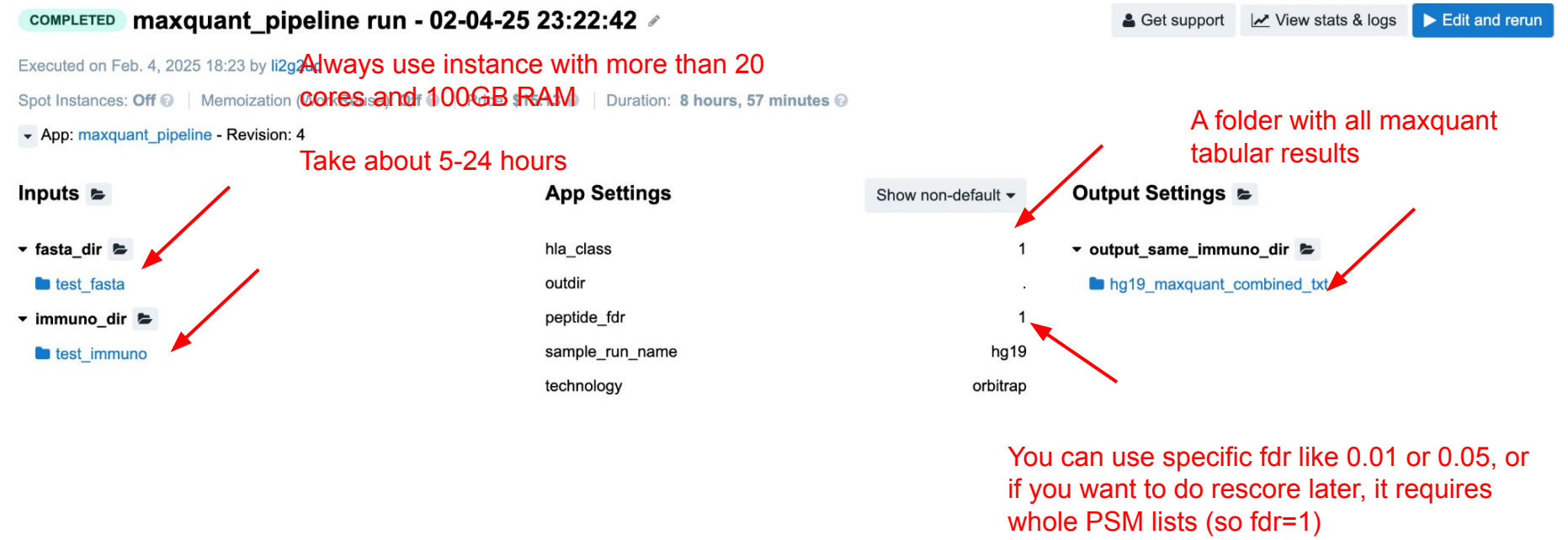
Summarization (step 1 is to put all the outputs generated till this step into a dedicated folder, if we call it **/result** like below)



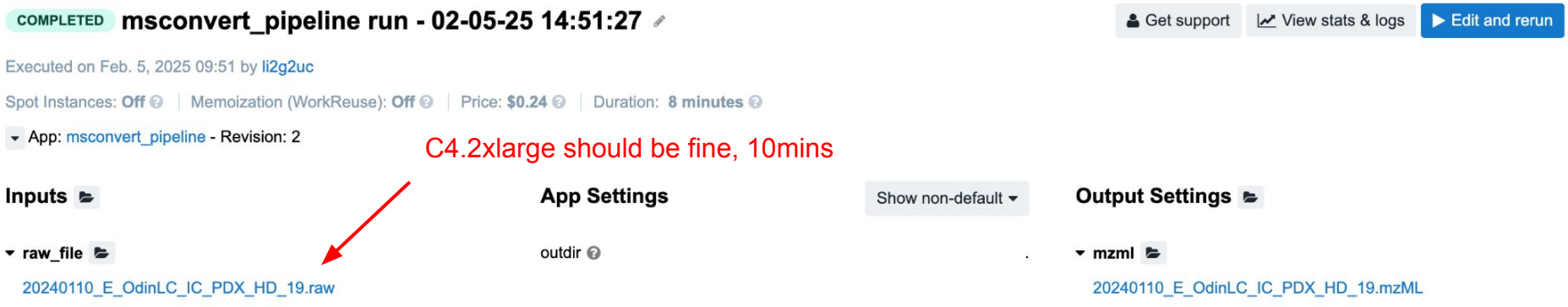
Summarization (step 2 is to run summarization pipeline to get all the search space)



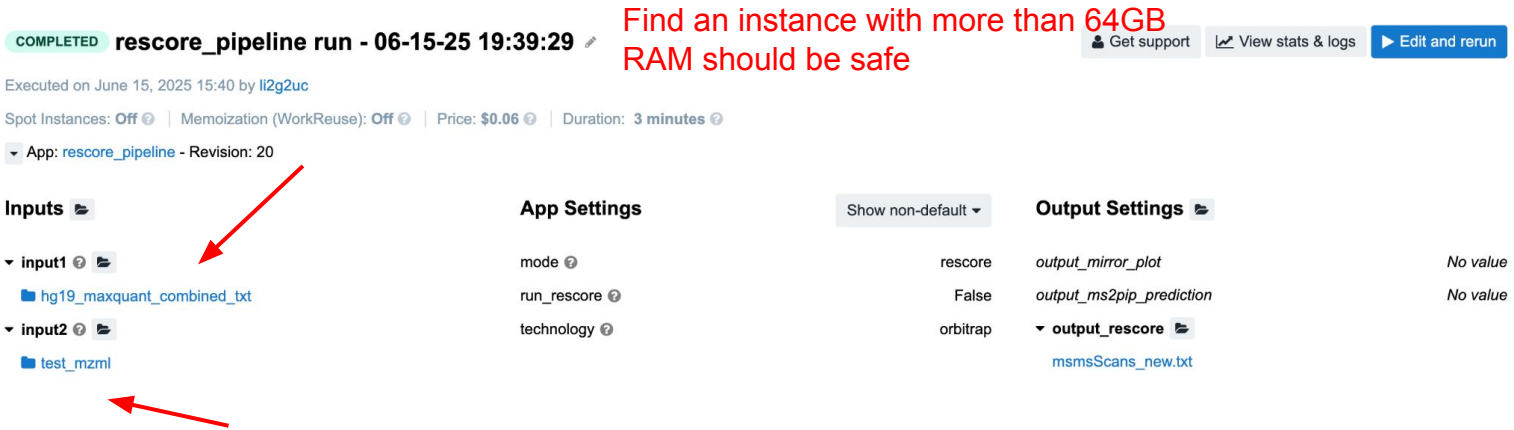
Immunopeptidome analysis (Step 1 is maxquant, you need a folder of all raw files, and a folder of all fasta as search space)



Immunopeptidome analysis (Step 2 is to use msconvert from proteowizard to convert raw to mzml, it is required for rescoring and visualization)



Immunopeptidome analysis (Step 3 is the rescoring step using ms2rescore)



Find an instance with more than 64GB RAM should be safe

Please put the generated mzML files from last step to a folder

Immunopeptidome analysis (Step 4 is the HLA binding prediction)

Just a tab-delimited txt file with two column, raw and hla, raw should be the raw file name, hla format is like below

	A	B	C	D	E	F
1	raw	hla				
2	20240110_E_OdinLC_IC_PDX_HD_19	A*02:01,A*03:01,B*18:01,B*44:02,C*12:03,C*07:04				
3						

Advanced Usage 1

In the manuscript, we showcased that ImmunoVerse wraps MS2PIP spectrum prediction function to evaluate the neoantigen spectra confidence on the fly, this function can be achieved through CGC as well, demonstrated in the following slides

Step 1: predict in silico spectrum for multiple peptides

COMPLETED **rescore_ms2pip run - 06-15-25 20:01:33**

Get support | View stats & logs | ▶ Edit and rerun

Executed on June 15, 2025 16:02 by li2g2uc

Any instance could work, lightweight

Spot Instances: **Off** | Memoization (WorkReuse): **Off** | Price: **$0.05** | Duration: **5 minutes**

App: rescore_ms2pip - Revision: 2

Inputs

input1

peptides.csv

input2

No files selected

App Settings Show non-default

mode: ms2pip_intensity

Output Settings

output_mirror: *No value*

output_ms2pip_prediction

ms2pip_prediction_NEVTTEIRF_2.png

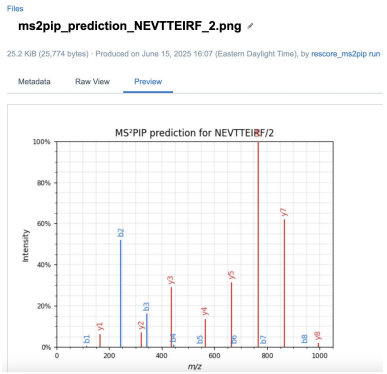
ms2pip_prediction_NEVTTEIRF_2.tsv

ms2pip_prediction_VTYNYPVHY_2.png

ms2pip_prediction_VTYNYPVHY_2.tsv

	A	B	C
1	seq	charge	
2	NEVTTEIRF	2	
3	VTYNYPVHY	2	
4			
5			

csv file



Step 2: Compare experimental and predicted spectra

COMPLETED **rescore_ms2pip run - 06-15-25 20:08:31**

Get support | View stats & logs | ▶ Edit and rerun

Executed on June 15, 2025 16:09 by li2g2uc

Spot Instances: **Off** | Memoization (WorkReuse): **Off** | Price: **$0.05** | Duration: **5 minutes**

App: rescore_ms2pip - Revision: 2

From last step

Inputs

- **input1**
  - ms2pip_prediction_NEVTTEIRF_2.tsv
- **input2**
  - 20240110_E_OdinLC_IC_PDX_HD_19.mzML

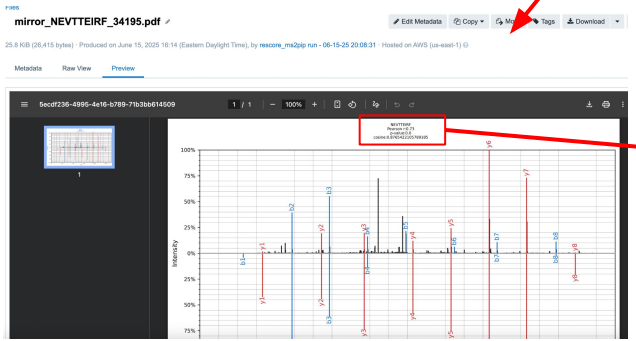
App Settings — Show non-default

mass_analyzer	ITMS
mode	in_silico_spike_in
scan	34195

Output Settings

- **output_mirror**
  - mirror_NEVTTEIRF_34195.pdf
- *output_ms2pip_prediction* — *No value*

You should have mzml file for each raw file



NEVTTEIRF
Pearson r:0.73
p-value:0.0
cosine:0.8765422105789185